Radeon GPU Architecture and the Radeon 4800 series

Michael Doggett
Graphics Architecture Group
June 27, 2008

Graphics Processing Units

AMD
Smarter Choice

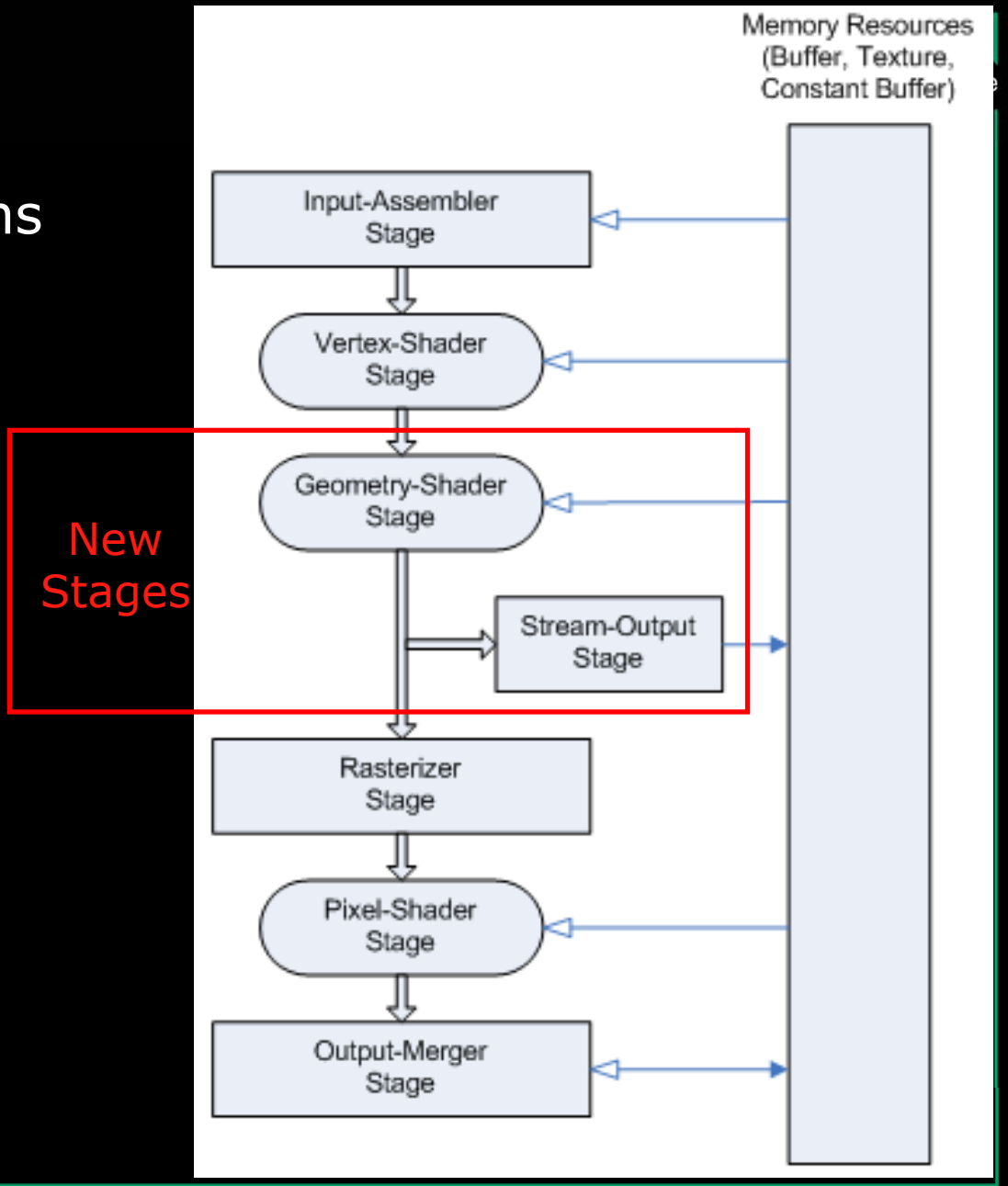
- Introduction
- Radeon 4800
- GPU research

GPU Evolution

- GPU started as a triangle rasterizer
  - Texturing, Z/S and color blending
  - Next added Transformation and Lighting
- Programmable GPU
  - Direct3D 8 – Assembler programming
  - Direct3D 9 – HLSL

Direct3D 10

- Program instantiations
  - Vertex programs
  - Primitive programs
  - Pixel programs
- 1,000 of short programs running pipelined in parallel



Massively Parallel Processor

AMD
Smarter Choice

- All parallel operations are hidden via domain specific API calls
- General Purpose GPU
- ATI GPU Programming APIs
  - Close To the Metal (CTM)
  - Compute Abstraction Layer (CAL)
  - Brook+

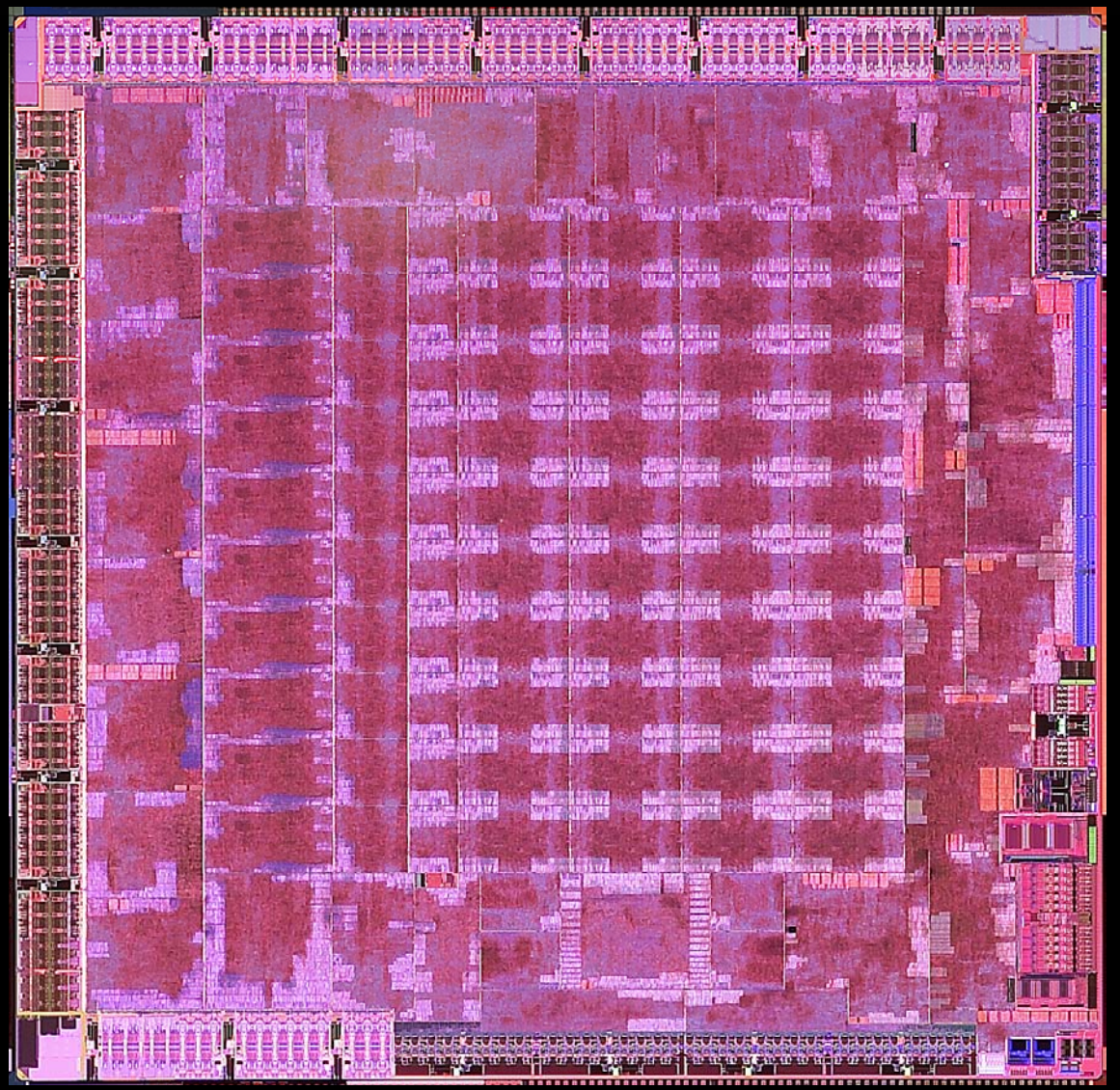
Chip Design Focus Point

AMD
Smarter Choice

CPU	GPU
Lots of instructions little data Out of order exec Branch prediction	Few instructions lots of data SIMD Hardware threading
Reuse and locality	Little reuse
Task parallel	Data parallel
Needs OS	No OS
Complex sync	Simple sync
Latency machines	Throughput machines

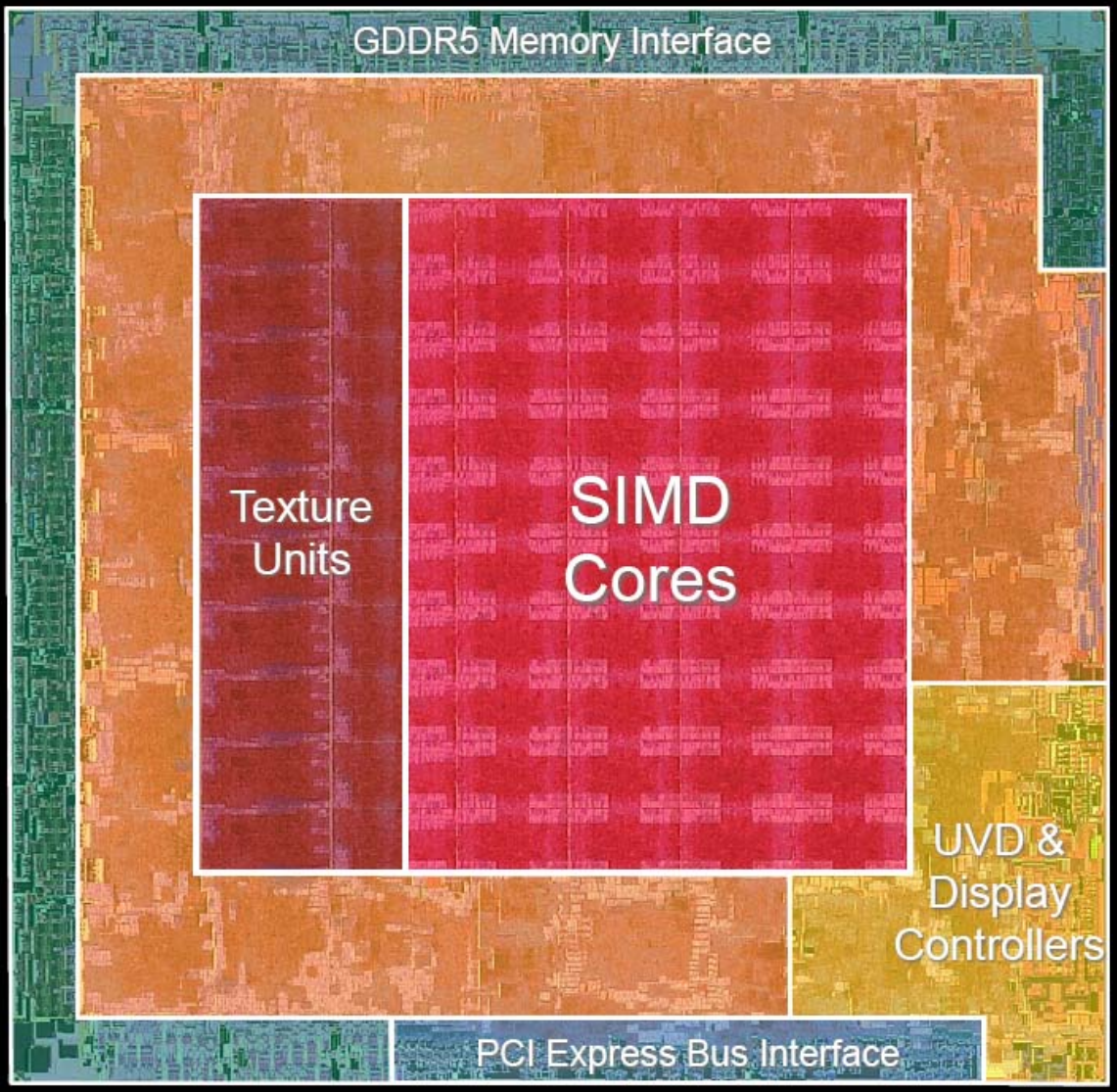
Radeon 4800 Series

- 260mm^2
- 956 MTransistors
- 64 z/stencil
- 40 texture
- 10 SIMDs
  - 800 shaders



Radeon 4800 Series

- 260mm^2
- 956 MTransistors
- 64 z/stencil
- 40 texture
- 10 SIMDs
  - 80 32bit FP Stream Processing Units (SPUs)



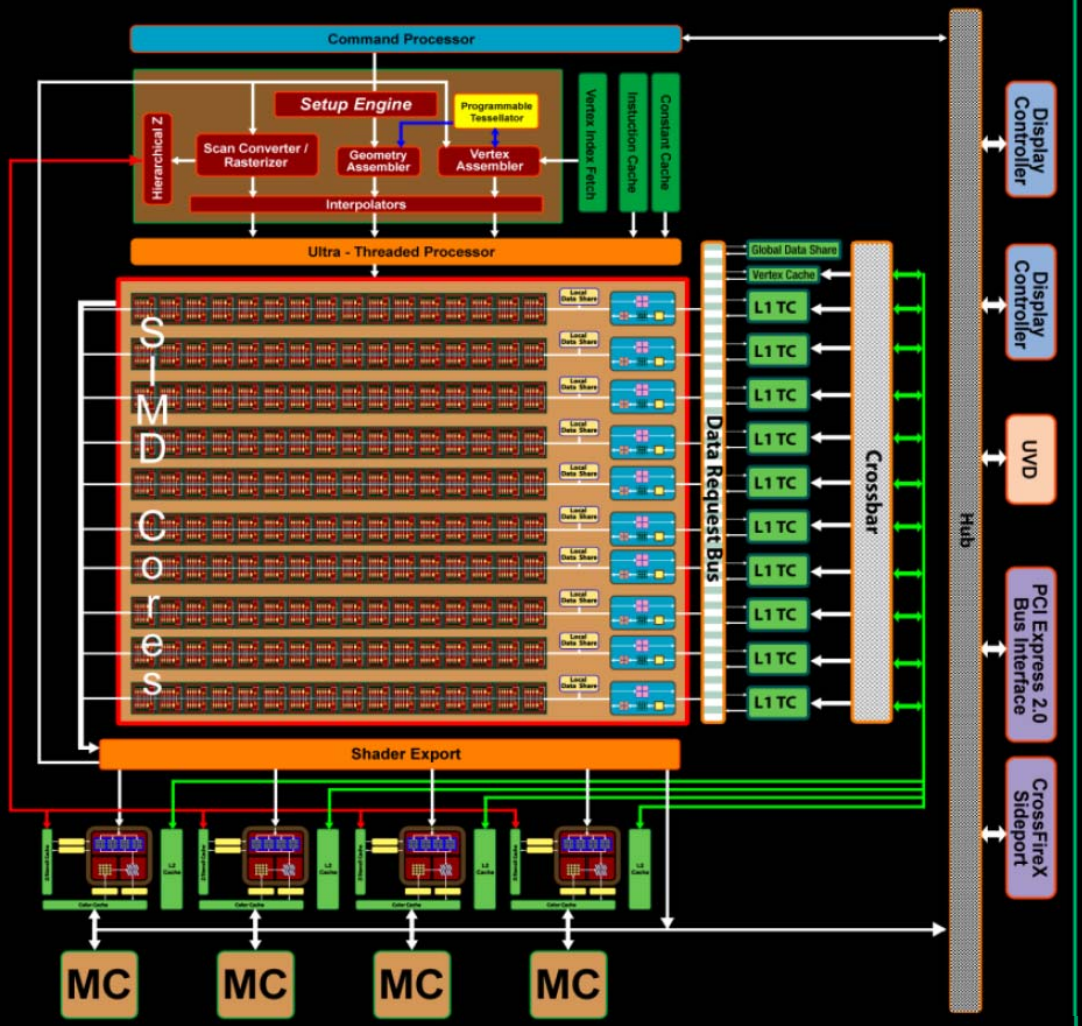
Radeon 4800 Series

- Performance per watt/mm^2/dollar
- ATI Radeon HD 4870
  - USD$299
  - 1.2 teraFLOPS
  - Core clock 750 MHz
  - 512 MB of GDDR5 - 3.6 gigabits/second
  - dual-slot PCI Express 2.0
  - 160 watts
- ATI Radeon HD 4850
  - USD$199
  - 1 teraFLOP
  - Core clock 625 MHz
  - 512 MB of GDDR3 - 2 gigabits/sec
  - single-slot PCI Express 2.0
  - 110 watts.

ATI Radeon™ 4800 Series Architecture

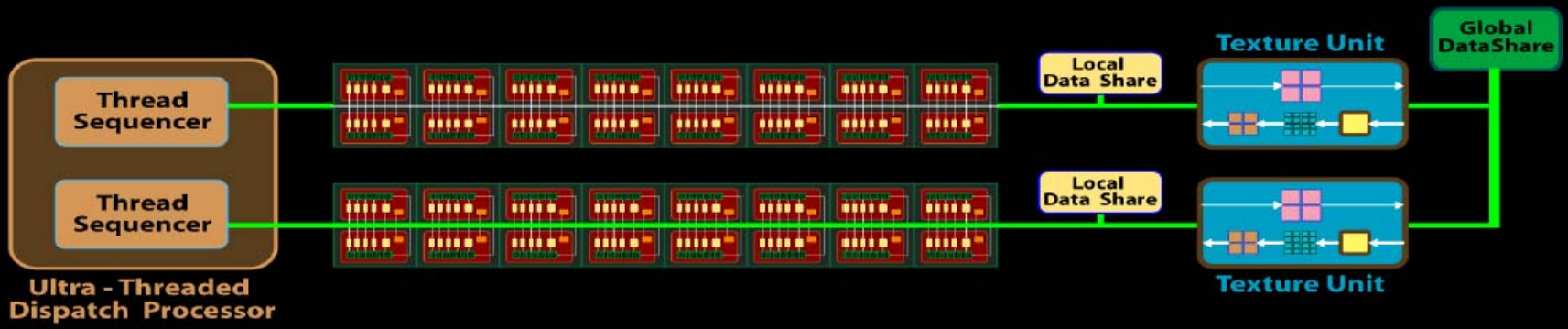
AMD Smarter Choice

- 800 stream processing units
- New SIMD core layout
- New texture cache design
- New memory architecture
- Optimized texture and render back-ends
- Enhanced geometry shader



SIMD Cores

- Each core:
  - Includes 80 scalar stream processing units in total
  - 16KB Local Data Share
    - Persistent allocation of data between wavefronts
  - Has its own control logic and runs from a shared set of threads
  - Has 4 dedicated texture units + L1 cache
  - Communicates with other SIMD cores via 16KB Global Data Share
    - Data sharing between threads running on different SIMDs
- New design allows texture fetch capability to scale with shader power, maintaining 4:1 ALU:TEX ratio

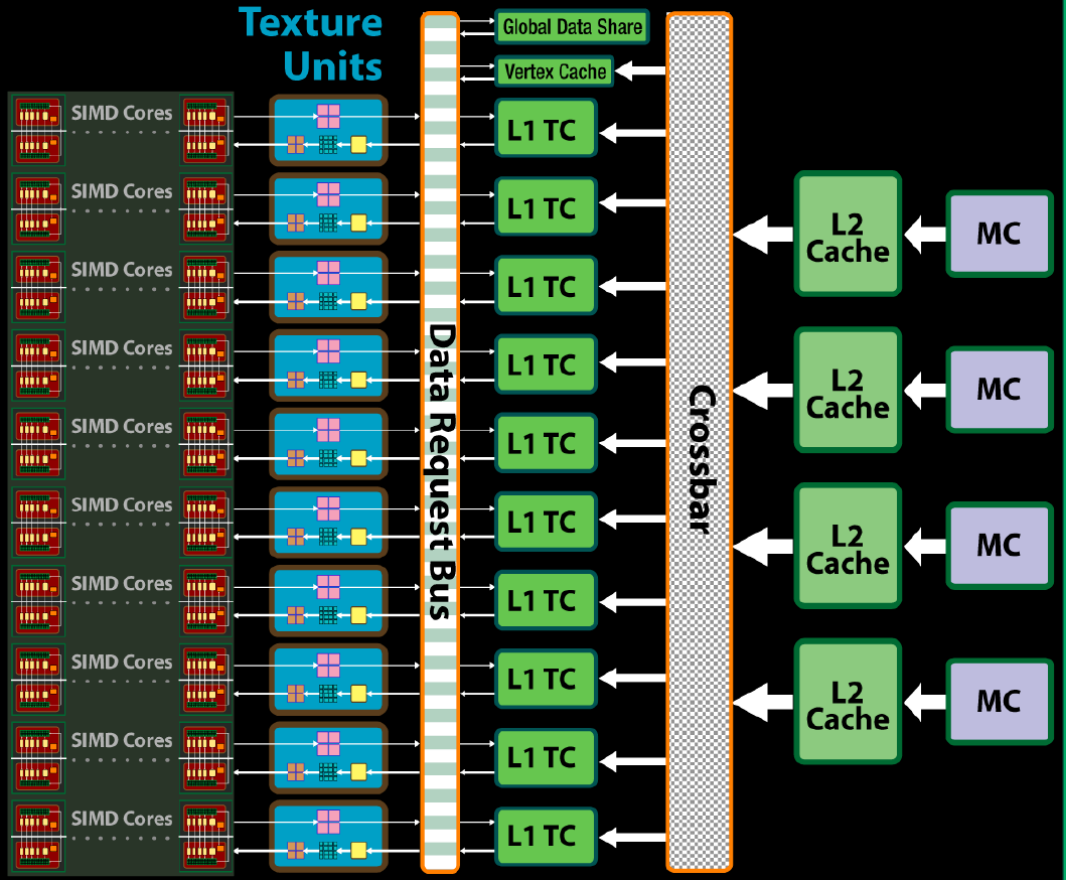


Radeon 4800

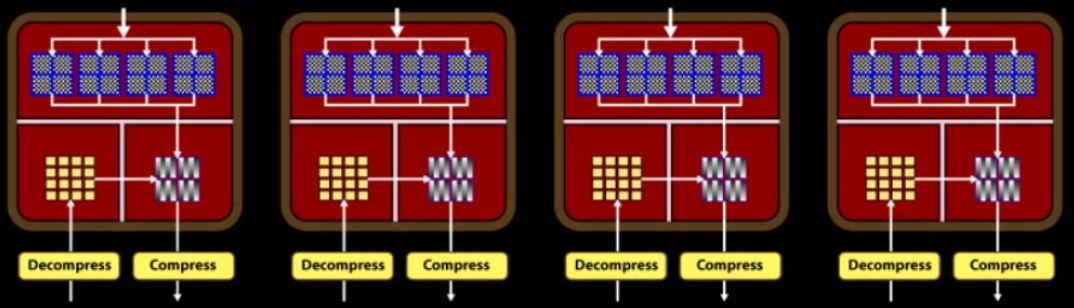
Texture Units

New cache design

- L2s aligned with memory channels
- L1s store unique data per SIMD
  - 2x increase in effective storage per L1
  - 5x increase overall
- Separate vertex cache
- Increased bandwidth
  - Up to 480 GB/sec of L1 texture fetch bandwidth
  - Up to 384 GB/sec between L1 & L2



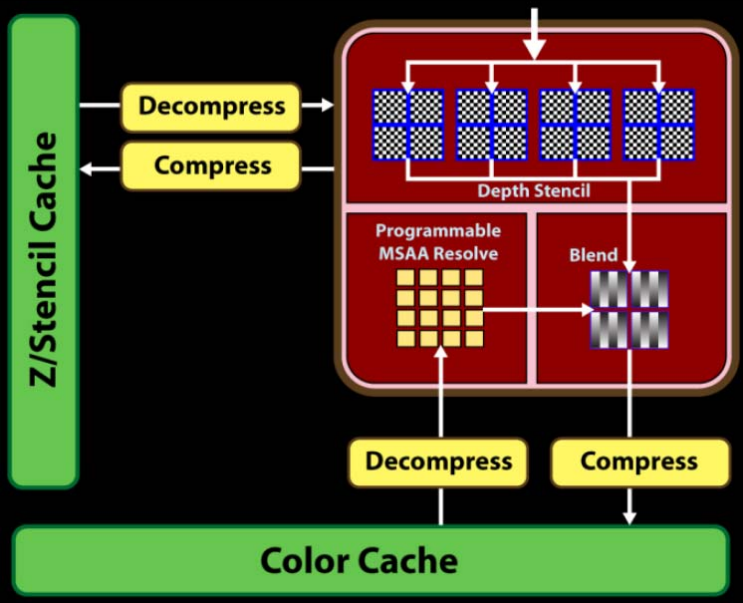
Render Back-Ends



- Focus on improving

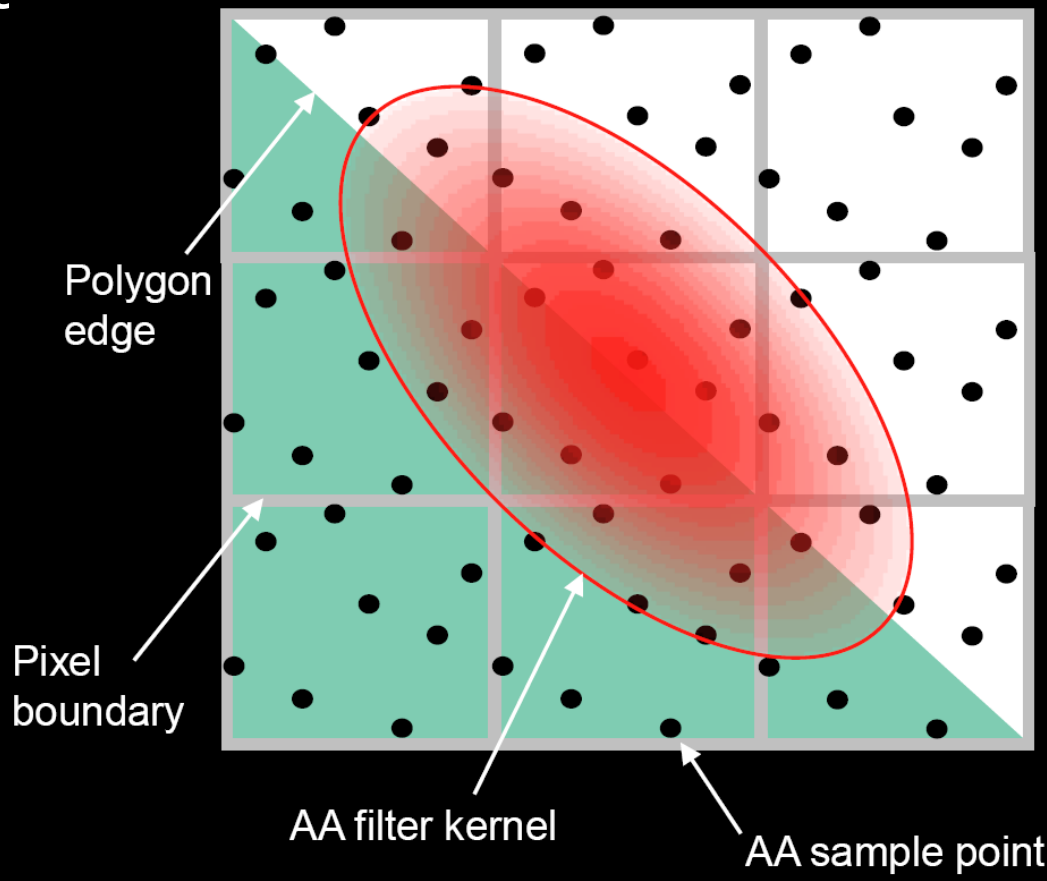
AA performance per mm2

  - Doubled peak rate for depth/stencil ops to64 per clock
  - Doubled AA fill rate for 32-bit & 64-bit color
  - Doubled non-AA fill rate for 64-bit color
- Supports both fixed function (MSAA) and programmable (CFAA) modes



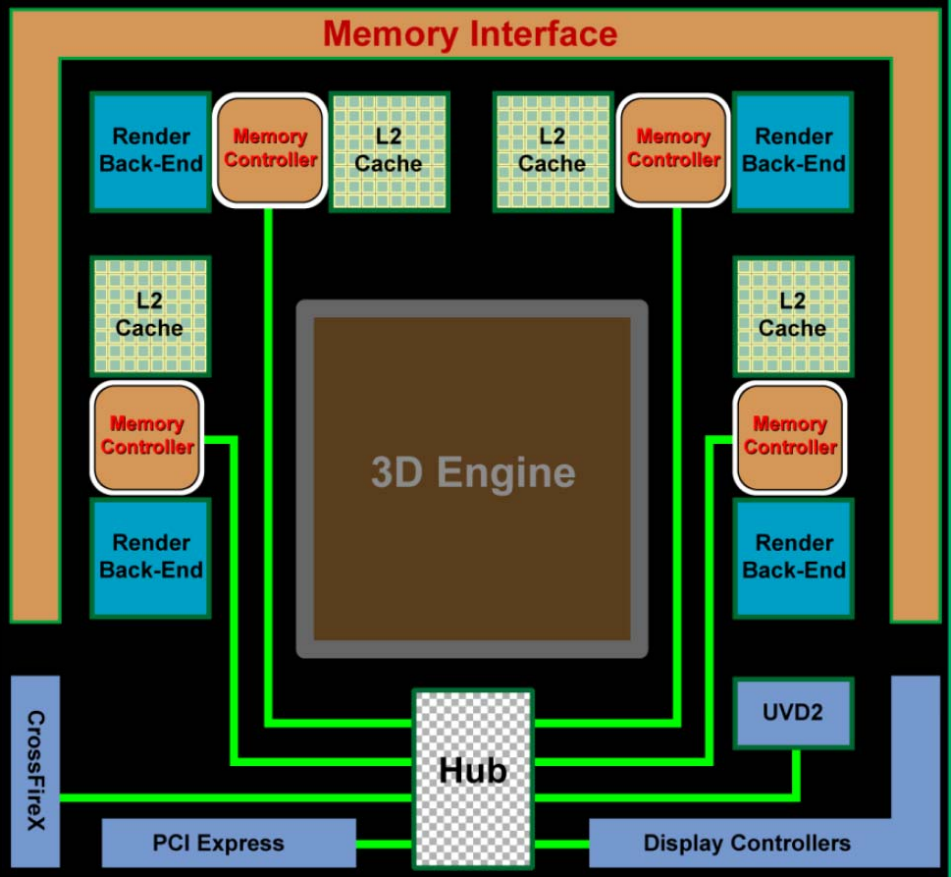
Edge Detect Custom Filter AA

- Enhanced edge-detect filter delivers 12x & 24x CFAA modes
- Avoids blurring by taking additional samples along edges, not across them
- Same memory footprint as 4x & 8x MSAA
- Works with Adaptive AA

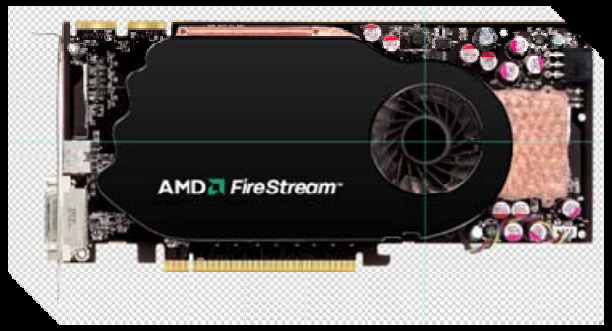


Memory Controller Architecture

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



AMD FireStream 9250

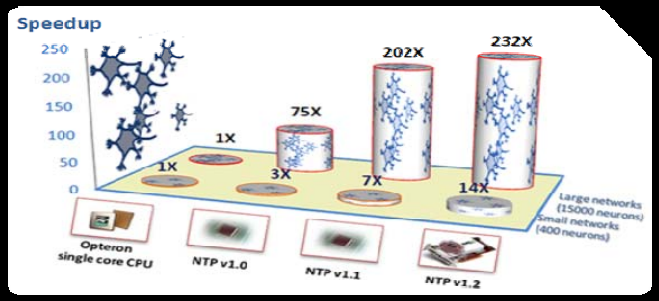


- 8 GFlops/Watt
- 240 GFlops of IEEE 64 bit FP
- OpenCL™ specification proposed by Apple to the Compute working group
  - OpenCL™ is a C-like language that enables programmers to tap into teraFLOPS of compute power on widely available GPU architectures

Building the Ecosystem: AMD Stream Application Successes

AMD Smarter Choice

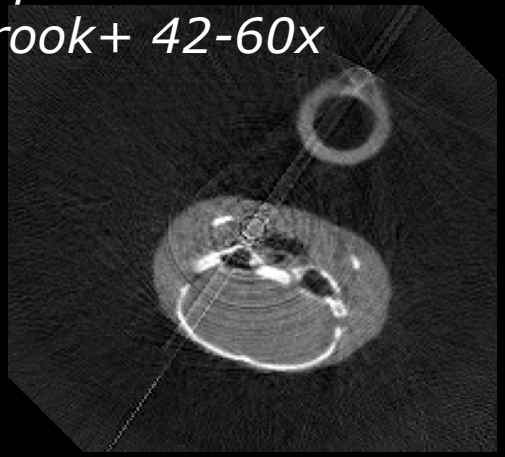
neurala 10-200x



Centre de Physique des Particules de Marseille

Tomographic reconstruction

Brook+ 42-60x



Computer Systems, Inc. MERCURY Challenges Drive Innovation™

ROGUE WAVE SOFTWARE

VizExperts

CAPS innovative tools for a new paradigm

RAPIDMIND

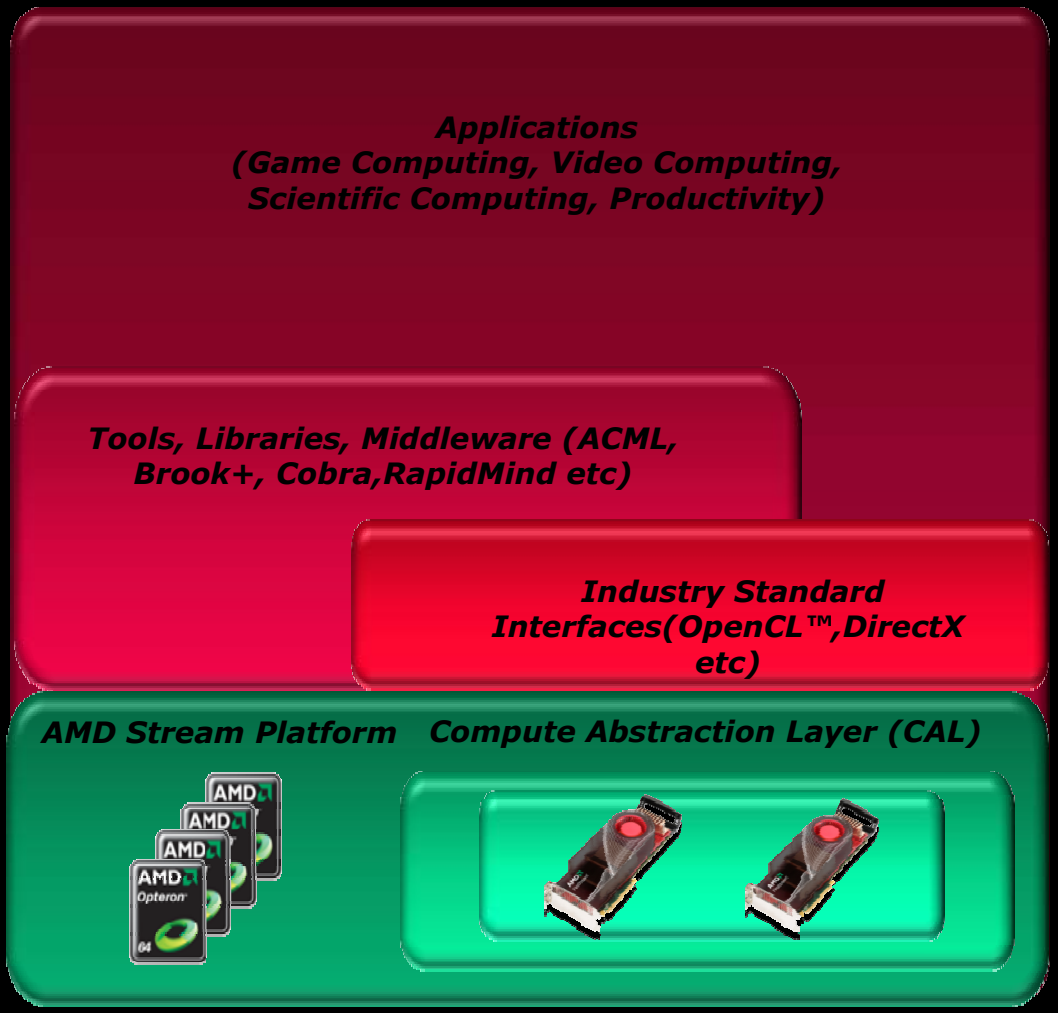
55x speedup* on binomial options pricing

ACCIT New Systems Research

EDA simulation >10x

Radeon 4800

AMD Stream Processing Strategy



Radeon 4800

FireStream Software Development Kit
High-level Development Tools

AMD Smarter Choice

Brook+
- High-level language, C extensions for the GPU
- Based on Brook from Stanford; AMD enhancements will be open-sourced

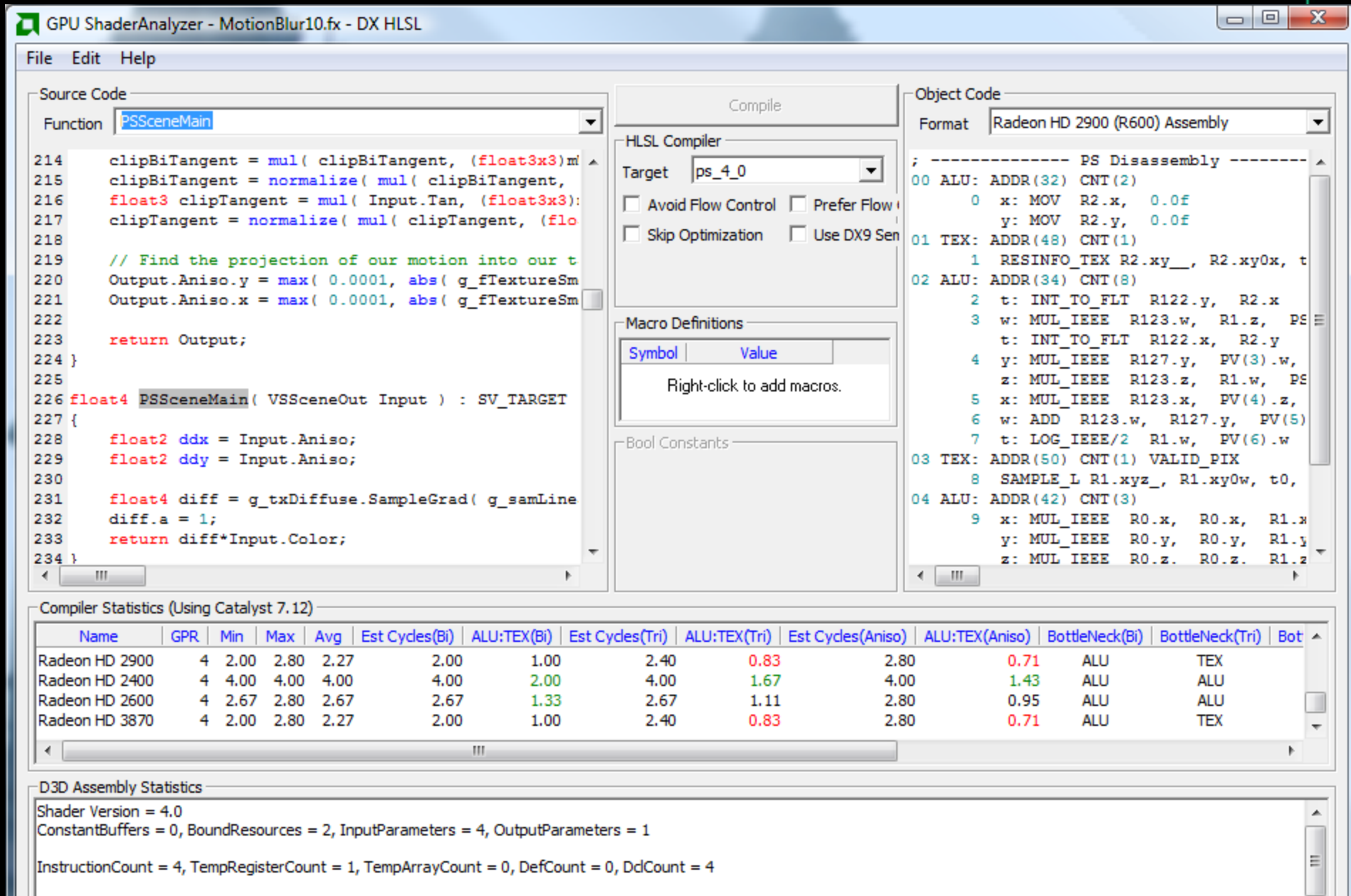
Libraries
- AMD's math library ACML provides GPU-accelerated math functions

Tools
- GPU Shader Analyzer
- AMD Code Analyst

- AMD Compute Abstraction Layer

GPU ShaderAnalyzer

AMD Smarter Choice



GPU research

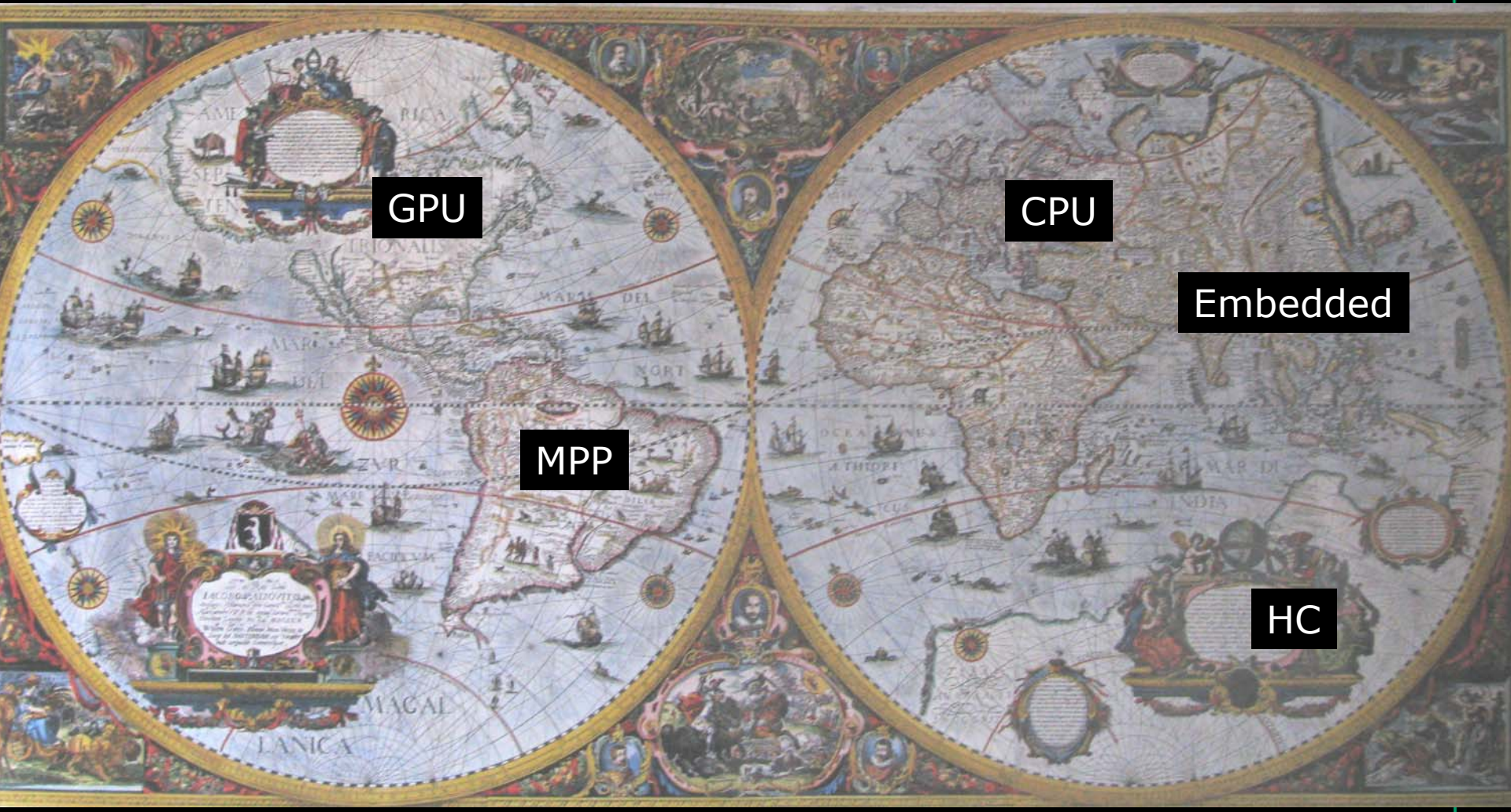
- Compute
  - Game Physics/AI
  - Lots of challenges in Architecture and Software
  - Hetergeneous Computing/Accelerated Computing, CPU + GPU
    - Software layers
  - Applications – Vision, recognition, data mining

GPU research

- Rendering
  - Real-Time rendering, game rendering
  - Tessellation
  - Ray tracing
  - Radiosity
    - Light transfer between surfaces

Processor World Map

AMD
Smarter Choice



Radeon 4800

Conclusion

- GPUs are massively parallel devices
- Radeon 4800 series
  - Teraflop processing power
  - High performance/$
- Brook+ for compute programming

Internships

Fellowships

Questions ?

michael.doggett ‘at’ amd.com

DISCLAIMER

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.

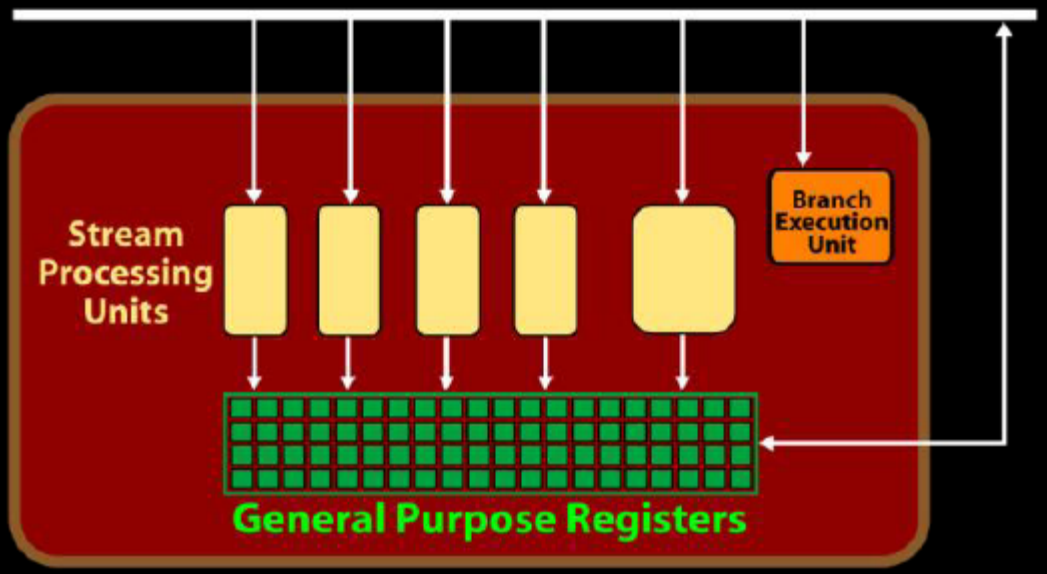
AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY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

Trademark Attribution

AMD, the AMD Arrow logo, ATI, the ATI logo, Radeon and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Other names used in this presentation are for identification purposes only and may be trademarks of their respective owners.

Stream Processing Units

- 40% increase in performance per mm2
- More aggressive clock gating for improved Performance per Watt
- Fast double precision processing(240 GigaFLOPS –2x of GTX280)
- Integer bit shift operations for all units (12.5x Improvement)



Texture Units

Streamlined design

–70% increase in performance/mm2

•More performance

–Double the texture cache bandwidth of the HD 3000 series

–2.5x increase in 32-bit filter rate

–1.25x increase in 64-bit filter rate

–Up to 160 fetches per clock