AMD
Smarter Choice

GPU Architectures and Futures – Some thoughts

Raja Koduri
CTO, Graphics, AMD

Agenda

AMD
Smarter Choice

- A programmer view of the graphics pipeline
- Anatomy of a modern GPU
- State of the Art in GPU accelerated rendering and other computations
- Some future thoughts

Modern graphics pipeline…

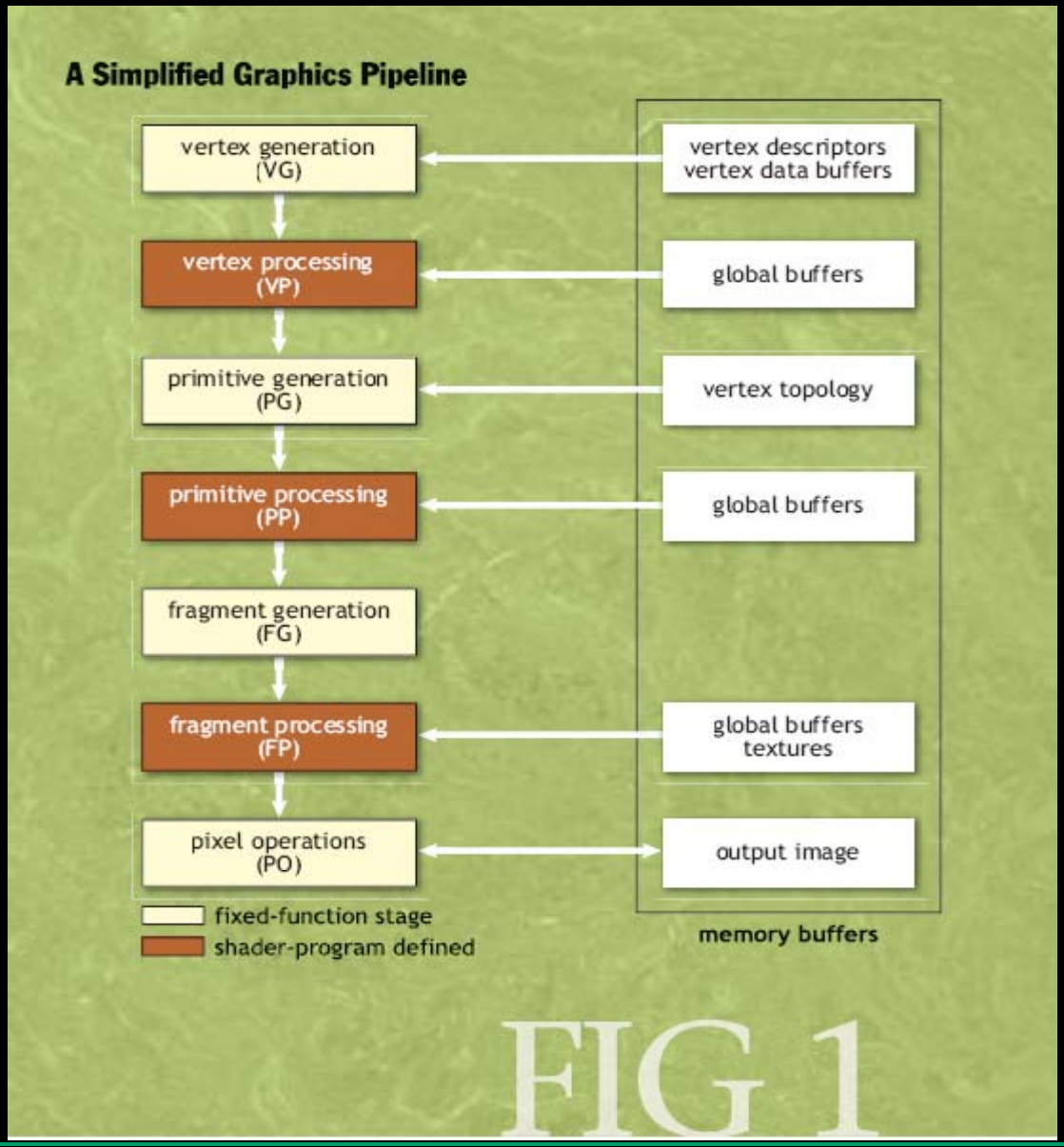
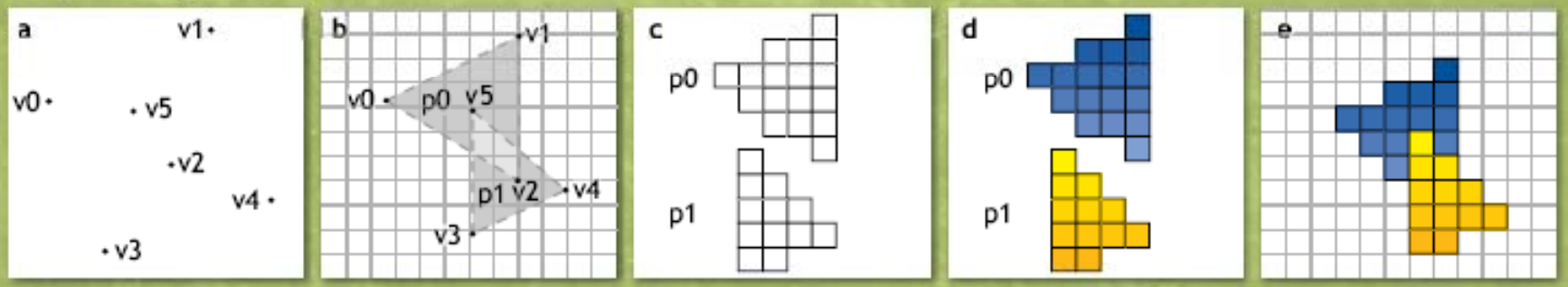


Fig 1 Courtesy of ACM Queue Mar/Apr 2008 – GPUs A closer look by Mike Houston and Kavyon Fatahalian

AMD
Smarter Choice

Graphics Pipeline Operations



(A) Six vertices from the VG output stream define the scene position and orientation of two triangles.

(B) Following VP and PG, the vertices have been transformed into their screen-space positions and grouped into two triangle primitives, *p0* and *p1*.

(C) FG samples the two primitives, producing a set of fragments corresponding to *p0* and *p1*.

(D) FP computes the appearance of the surface at each sample location.

(E) PO updates the output image with contributions from the fragments, accounting for surface visibility. In this example, *p1* is nearer to the camera than *p0*. As a result *p0* is occluded by *p1*.

FIG 2

Fig 2 Courtesy of ACM Queue Mar/Apr 2008 – GPUs A closer look by Mike Houston and Kavyon Fatahalian

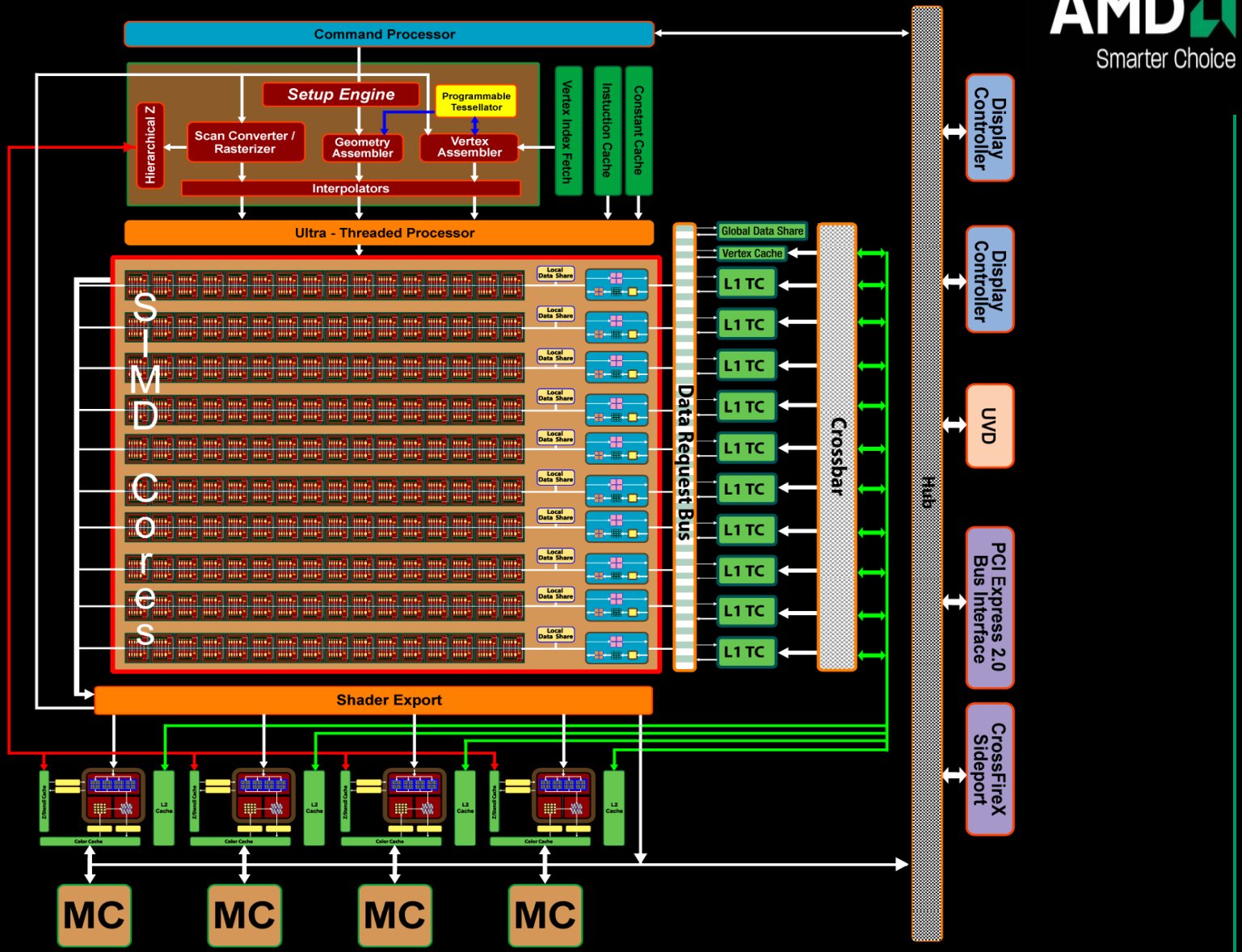
Doom - 1993


UAC
SOFTPEDIA
www.softpedia.com
97
AMMO
123%
HEALTH
2 3 4
5 6 7
ARMS
193%
ARMOR
BULL 400 / 400
SHEL 97 / 100
ROKT 6 / 100
CELL 20 / 600

Doom3 - 2004

AMD
Smarter Choice

Anatomy of a Modern GPU (Radeon 4800 Series)



AMD Smarter Choice

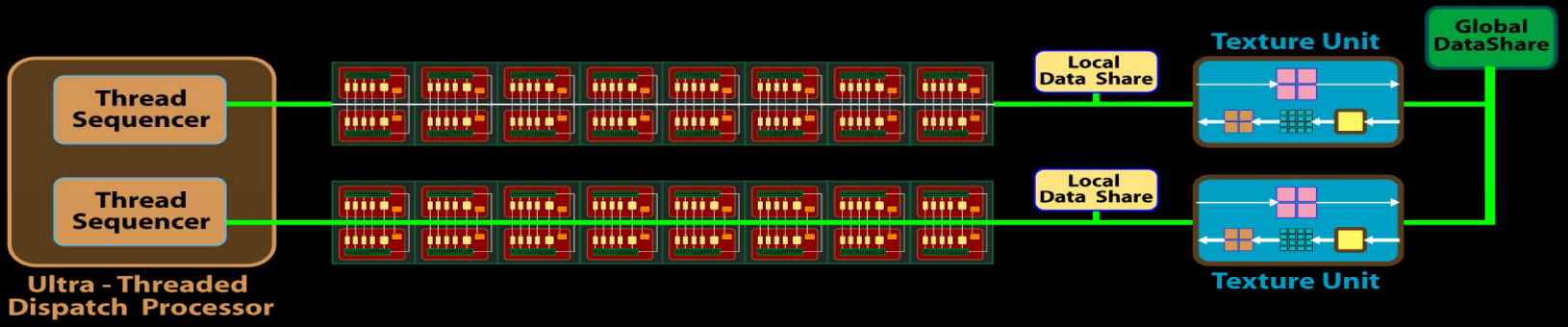
Speeds and Feeds

- Radeon 4800 series (Available June'08)
- 55 nm TSMC process
- 260 mm2 die size
- 110-160 Watts power
- 1.2 TeraFLOPS of compute
- Double precision FP support
- Under $0.25 per GigaFLOP (including memory)
- >8 GFlops-per-watt
- >100 GB/sec memory bandwidth

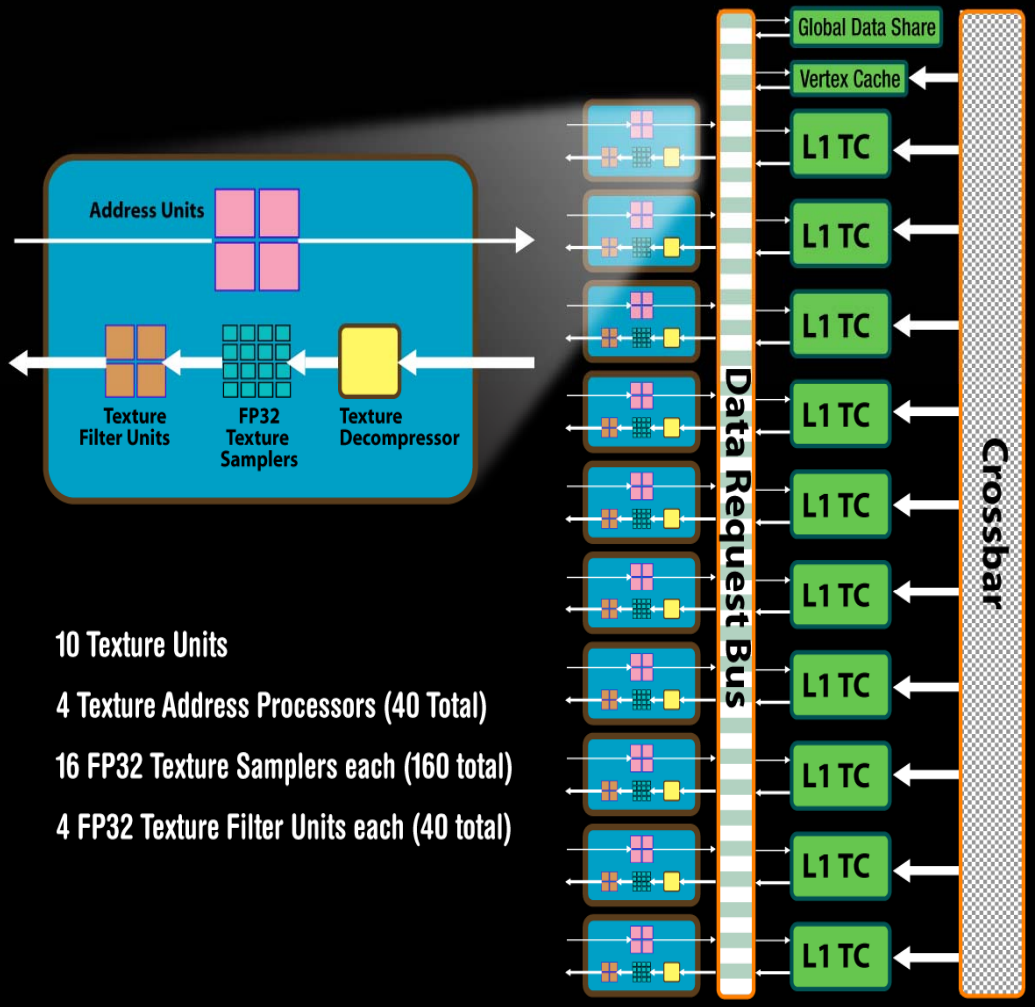
SIMD Core

AMD Smarter Choice

- Each core:
  - Includes 80 scalar stream processing units in total + 16KB Local Data Share
  - 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
- New design allows texture fetch capability to scale with shader power, maintaining 4:1 ALU:TEX ratio



Texture Units



10 Texture Units

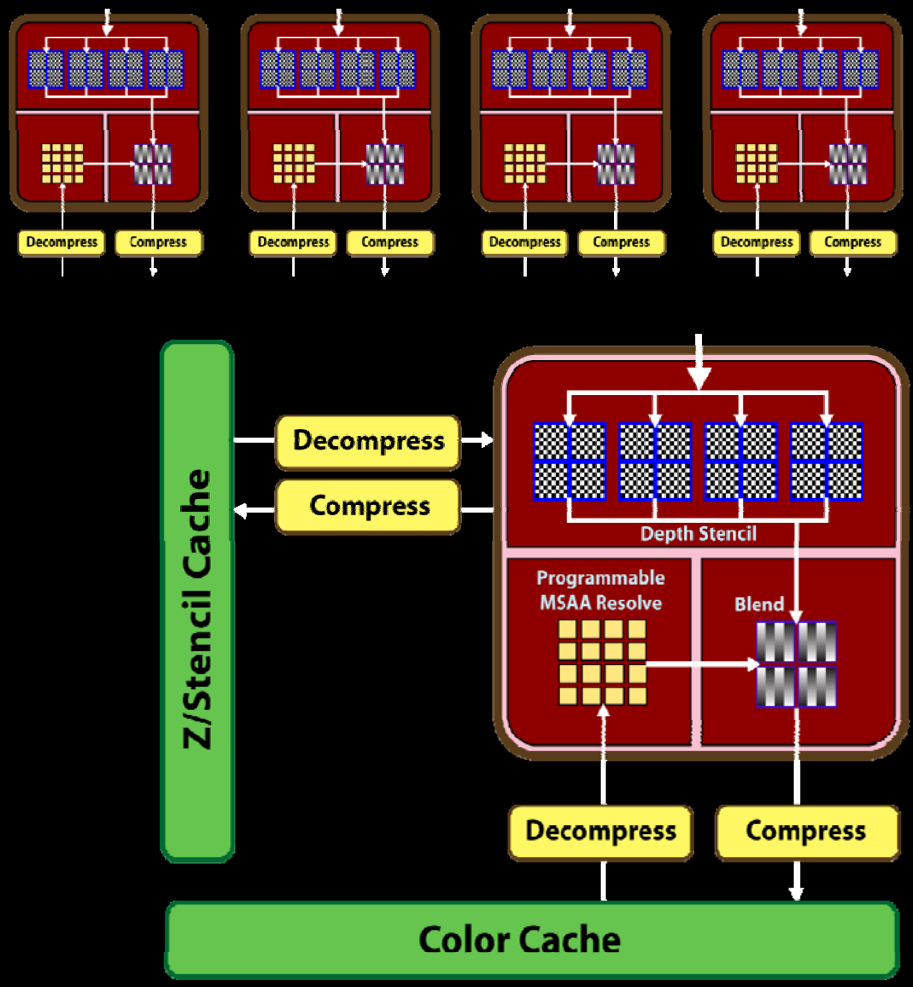
4 Texture Address Processors (40 Total)

16 FP32 Texture Samplers each (160 total)

4 FP32 Texture Filter Units each (40 total)

- These are essentially the main fetch units
- Read-only caches
- Filtering and decompression are the key fixed functions

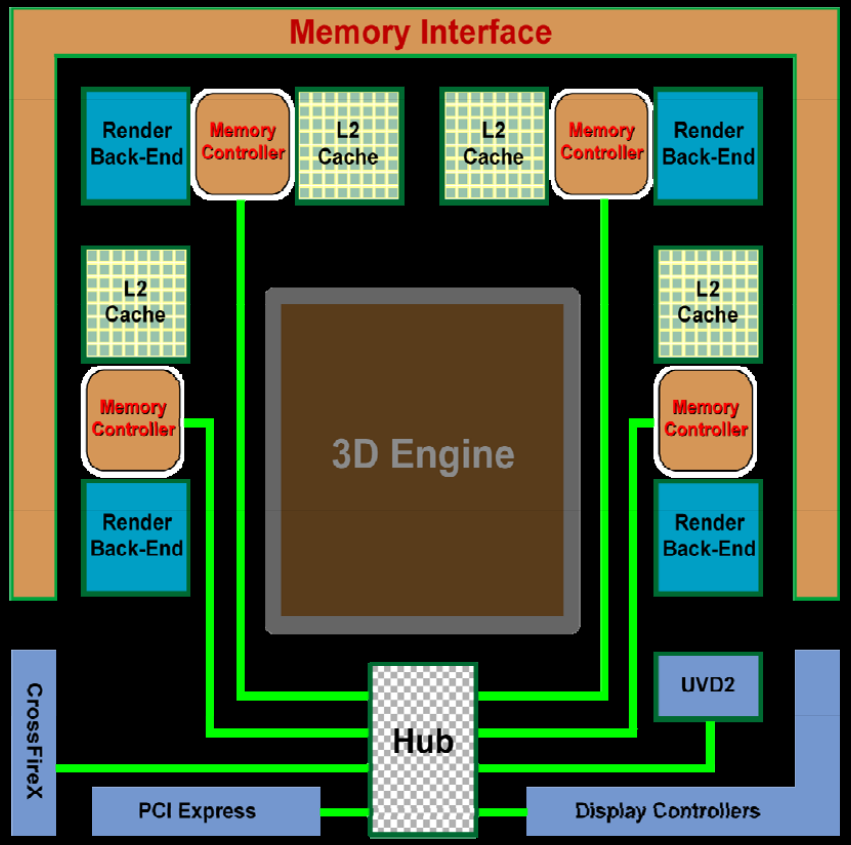
Render Backends



- Depth testing and frame buffer compositing operations are done here
- R/W caches optimized for triangle rasterization based access patterns
- Compression and Decompression to save bandwidth

Memory controller

- Controllers distributed around periphery of chip, adjacent to primary bandwidth consumers
- Memory tiling & 256-bit interface allows reduced latency, silicon area, and power consumption
- Hub handles relatively low bandwidth traffic
  - PCI Express, CrossFireX interconnect, UVD2, display controllers, intercommunication)



AMD
Smarter Choice

So, what can you achieve with a modern GPU?

State of the Art – 2k Boston BioShock

AMD
Smarter Choice

Recent Evolution of Rendering

AMD Smarter Choice

- For the last decade, real-time shading has significantly increased in quality
- Great strides in lighting and shadowing techniques
  - Normal mapping is ubiquitous in games
  - Inverse displacement mapping (such as POM) is present in the latest top graphics-rich games

Parallax occlusion mapping in Crysis

(Crytek)

Typical Production Goals for Games Now

AMD
Smarter Choice

- Cinematographic quality rendering
- Dynamic light and shadows
- MultiGPU & Multicore CPU support
- Huge game levels
- Target GPU from SM 2.0 to SM 4.0 (DirectX10)
- High Dynamic Range
- Dynamic environment (breakable)
- Developing game and engine together

Introducing…
The Froblins

Rendering and computational elements of Froblins

AMD
Smarter Choice

- Dynamic pathfinding AI computations on GPU
- Massive crowd rendering with LOD management
- Tessellation for high quality close-ups and stable performance
- HDR lighting and post-processing effects with gamma-correct rendering
- Terrain system
- Cascade shadows for large-range environments
- Advanced global illumination system

Dynamic and Engaging World Through AI

AMD Smarter Choice

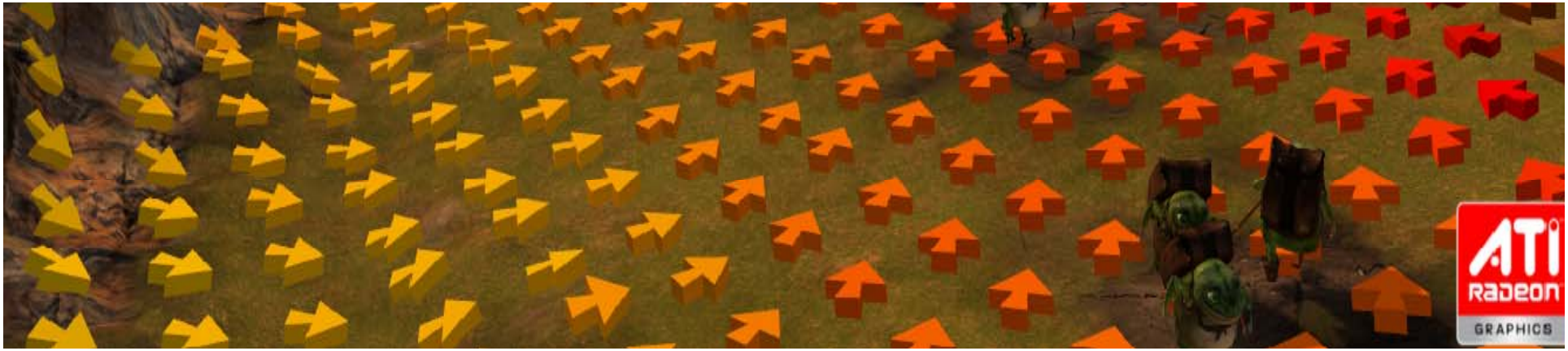
- Global path finding along with local avoidance and individual decisions
- Crowd movement is more stable and realistic solved on a global scale
  - Crowd dynamics similar to fluids
  - Agents “flow” towards the closest goal, along the path of least resistance
  - Froblins follow correct paths and do not “get stuck”

Pathfinding on GPU

Numerically solve a 2nd order PDE (partial differential equation) on GPU with iterative approach (Eikonal solver)

- Environment as a cost field

Applicable to many general algorithms and areas (Google, anyone?)



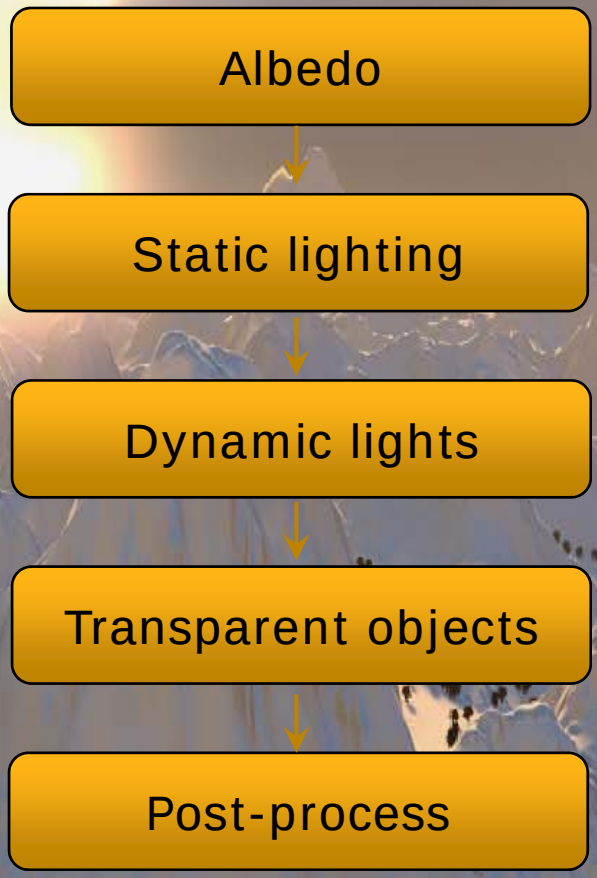
Where Do All the Flops Go?
Dynamic Character Animation Logic, Of Course

AMD Smarter Choice

- Sophisticated animation system for froblins is managed on GPU
- Logic controlled via a complex shader – 3200 instructions for each froblin
  - Really takes advantage of the high powered 800 stream processors

Global Illumination Using DirectX10.1

- Complex solution to compute global illumination
  - With photon mapping and radiance estimation
- High quality global illumination evaluated at run-time using spherical harmonics based lighting
  - 3rd order SH lighting using FP16
  - **260 instructions** per pixel to evaluate



Performance on ATI Radeon HD 4800

AMD Smarter Choice

- Staggering polygon count at interactive rates (>20fps)
  - From 900 polygons -> 6K -> 1.6M at the closest tessellated level of details
  - Up to 18M triangles per frame at fast interactive rates
  - 6M-8M triangles on average at 20-25 fps
- Full high quality lighting and shadowing solution
  - Rendering all objects into multiple shadow maps more than doubles polygon count per frame
- Rendering and simulating high quality detailed 3K froblins at 22 fps on average

AI Performance Statistics

AMD Smarter Choice

- All modes render with 4X MSAA HDR and post-processing
- Simulate behavior and render > 65K agents at 30 fps
  - AI simulation for 65K agents alone – 45 fps
  - Rendering and simulating 65K agents – 31 fps

    Rendering 9,800,000 polygons each frame!
- AI simulation executes with high efficiency resulting in *911 billion* flops = 0.9 teraflops!

Several challenges still exist for photorealism in real-time

AMD Smarter Choice

- Complex environments
  - Data sets keeps exploding in size

Per-Frame	Shrek	Typical Game
Content size	100 M polygons	200-500K polygons
Animation quality	350 bones skinned on the CPU	32-40 bones skinning on GPU
Rendering time	8000 sec a frame on a Pentium IV	0.015 sec or less a frame

A frame from the Shrek [Yee04, PDI/Dreamworks]

- Efficient real-time global illumination techniques
- Transparency and translucency
- Volumetric Effects and participating media effects

Realistic characters

AMD
Smarter Choice

AMD
Smarter Choice

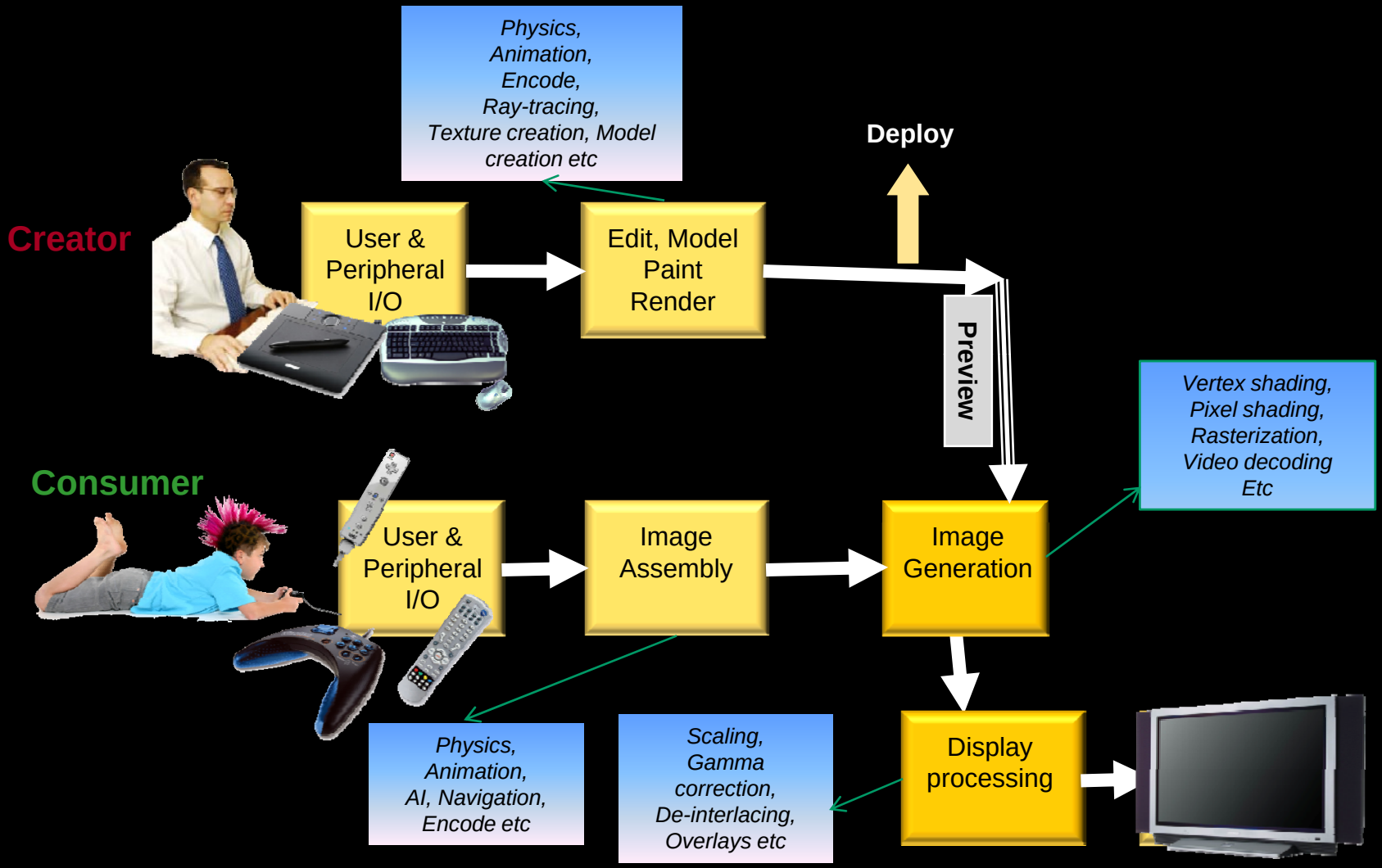
GPU Futures

GPU Visual Futures – 2011 and beyond

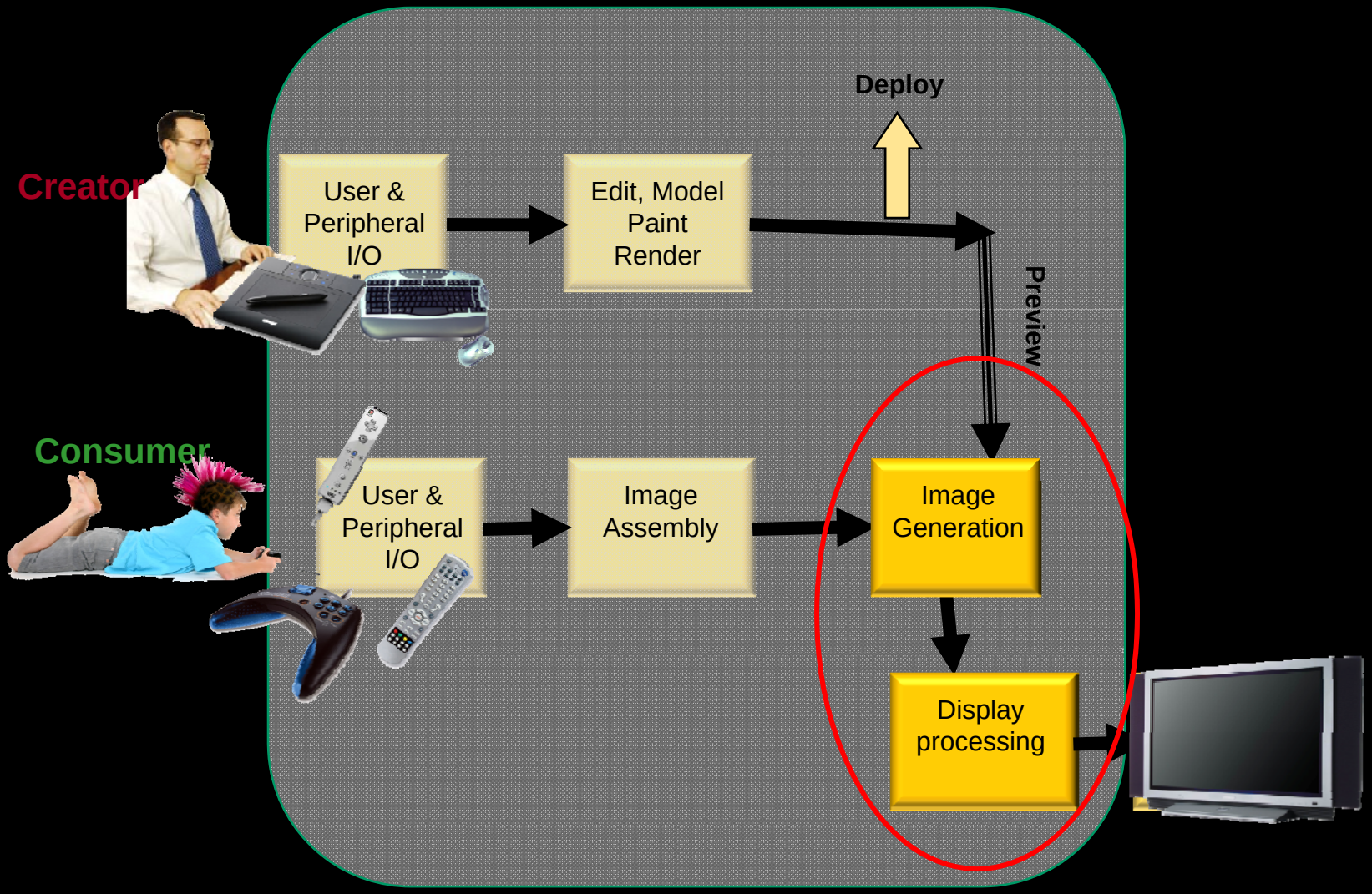
AMD
Smarter Choice

- Content creation pipeline acceleration is critical
- The GPU architecture evolves into a powerful content computing processor

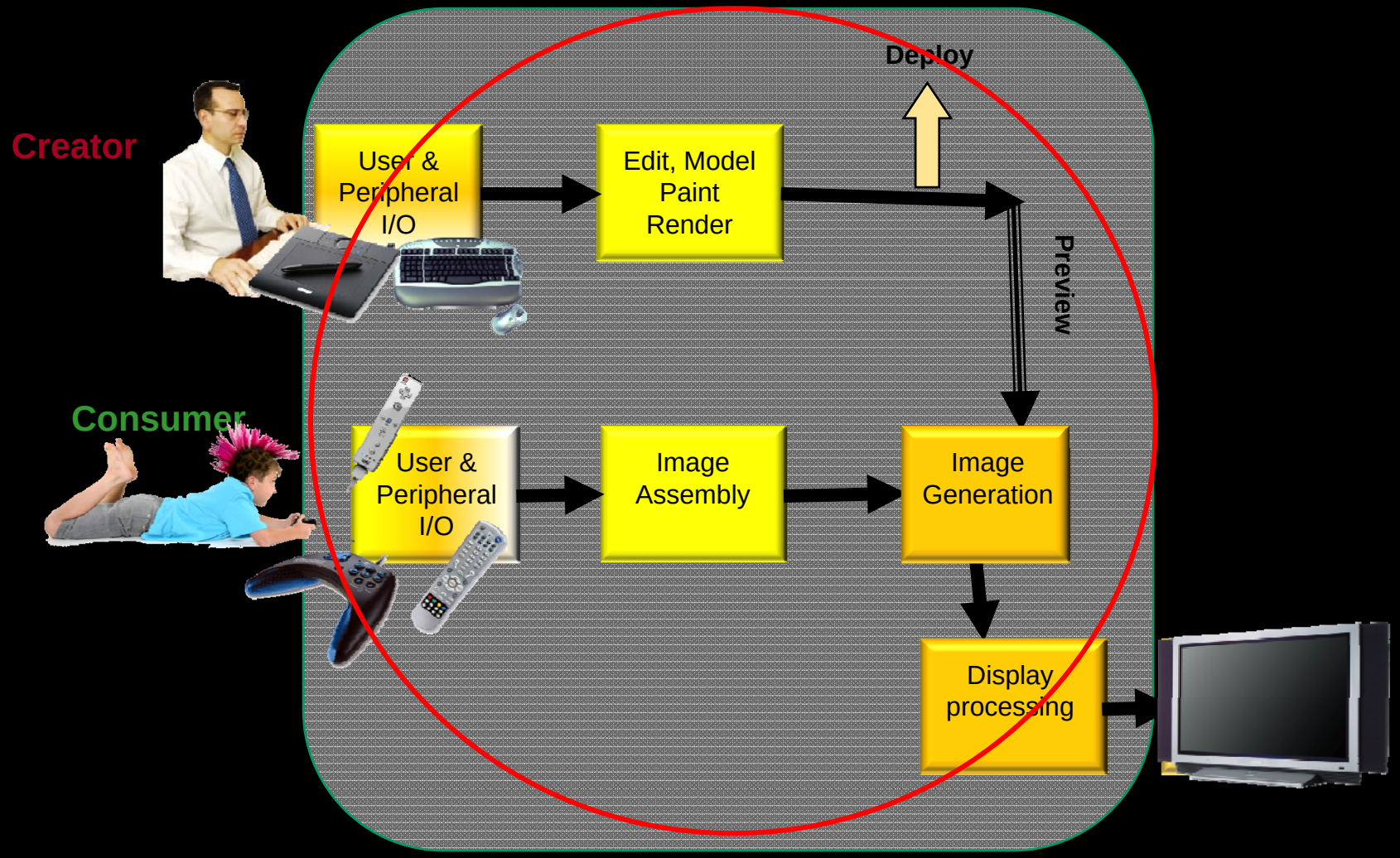
The Content Pipeline



GPU Focus Today



Future GPU – Accelerating the content pipeline



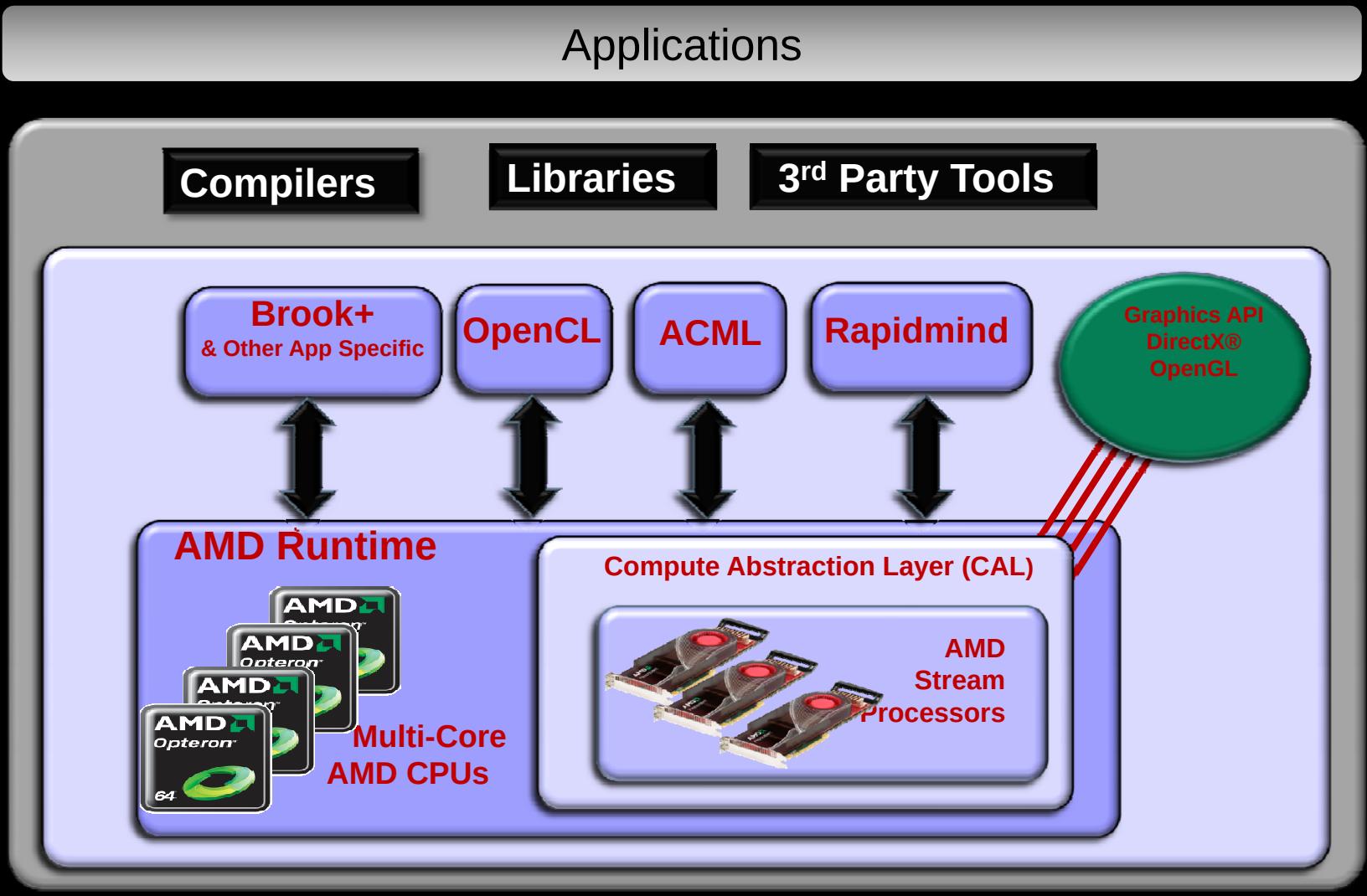
AMD
Smarter Choice

- GPU is very a attractive data parallel compute process as well

- New software models necessary to leverage the compute capabilities of the GPU

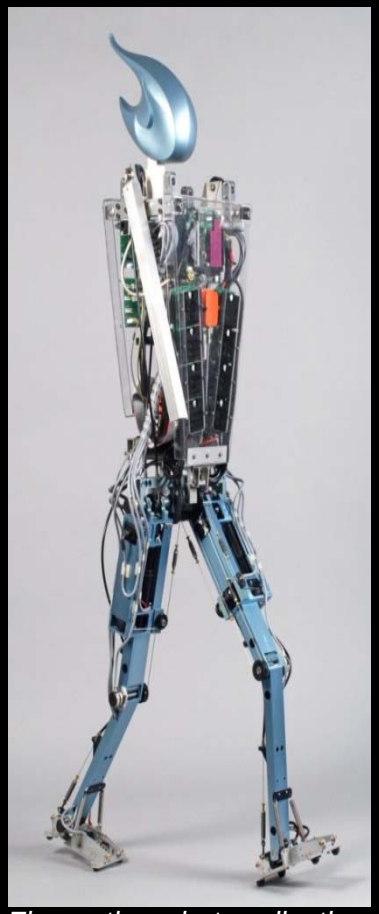
AMD Stream SDK *Software Development Stack*

AMD
Smarter Choice



Future compute platform thoughts

Models for the Future of Computers



Flame, the robot, walks the way humans walk. (Credit: Image courtesy of Delft University of Technology)

Human brain image licensed under Creative Commons 3.0 Unreported license. See http://en.wikipedia.org/wiki/Image:Brain_090407.jpg

Human Brain

© 2006 Stanford Racing Team Web design by ToTheWeb LLC
http://cs.stanford.edu/group/roadrunner/images/graphics/junior-3.jpg

Autonomous Vehicle

F-16 Fighting Falcon image is a U.S. government work and is in the public domain. See http://en.wikipedia.org/wiki/Image:F-16_Fighting_Falcon.jpg

Fighter Jet

Heterogeneous Systems

- Parallel signal processing of sensor data
- Parallel Classification and Recognition
  Vision, Sound, Contact, Patterns, Object
- Serial Cognitive and Reasoning
- Parallel Search, Scan, Sort

AMD Smarter Choice

Advantages

Optimized for execution of sequential program

- Complex Pipelines to achieve max frequency
- Out Of Order, Super Scalar to achieve max ILP
- Branch Predictors, Speculative execution
- Register Renaming
- Large Caches optimized for low latency access

Multi-Core enables parallel multi-tasks/threads

- Improved user response
- Background task such as OS chores, virus scan, etc

AMD
Smarter Choice

Disadvantage

Fine grain sharing of work between cores/caches
- OS Overhead – spawn, communication, fork
- Finding large number of task to Multi-task is hard

Embarrassing parallel apps
- Limited Speed up (ALU & Bandwidth)
- Limited Power/Flop advantage

AMD
Smarter Choice

Advantages

- Optimized for structured parallel execution
  - Extensive ALU counts & Memory Bandwidth
  - Cooperative multi-threading hides latency
- Shared Instruction Resources
- Fixed function units for parallel workloads dispatch
- Extensive exploitation of Locality

AMD
Smarter Choice

Disadvantages

Ineffective for Single threaded execution

- Divergence
  - Parallelism lost opportunity
  - Loss of efficiency
- Upload/download of data cost
- Requires large workload to fully utilize
- Small Caches are optimized for locality/throughput

Building Blocks - Best of Both

AMD Smarter Choice

Heterogeneous Cores (Mixed CPU/GPU)

- Lowest Power & Highest Performance Solution
- Serial Single threaded – Leverage Fast OOCs
- Task Parallel – Leverage multi-CPU or GPU cores
- Data Parallel - Leverage GPU or Application Specific Cores

Building Blocks - Best of Both (Con'd)

AMD Smarter Choice

Unified Memory Architectures

- Remove large data copies of workload and results
  - Reduce power consumption per computation
  - Enable small job offload
- Remove OS Overhead and latency of communication
- Fast Synchronization Primitives
- Increasing capable memory systems & BW

Software Solution (Real Challenge)

AMD
Smarter Choice

Development of Software often exceeds that of hardware

- Determine and implement long term solutions
- Enable highly coherent machines with heterogeneous accelerators
- Enable control of memory hierarchies for system scaling
- Communication/Messaging ➔ Producer/Consumer relationships

Simplifying Programming Model

- **B**uild on emerging multi-core models
- Enable the Masses to Programming with parallel abilities
- Enable Application Specific Libraries
- Enable open standards

Disclaimer and Attribution

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

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.

ATTRIBUTION