Graphics Processing Unit (GPU) Memory Hierarchy

Presented by Vu Dinh and Donald MacIntyre

Agenda

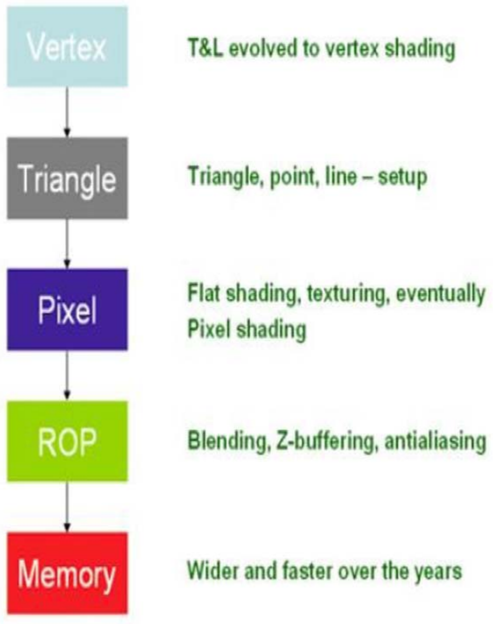
- Introduction to Graphics Processing
- CPU Memory Hierarchy
- GPU Memory Hierarchy
- GPU Architecture Comparison
    - NVIDIA
    - AMD (ATI)
- GPU Memory Performance
- Q&A

Brief Graphics Processing History

- Graphics Processing has evolved from single hardware pipelined units and are now highly programmable pipelined units.
- Over time tasks have been moved from the CPU to the GPU

Graphics pipelines for last 20 years

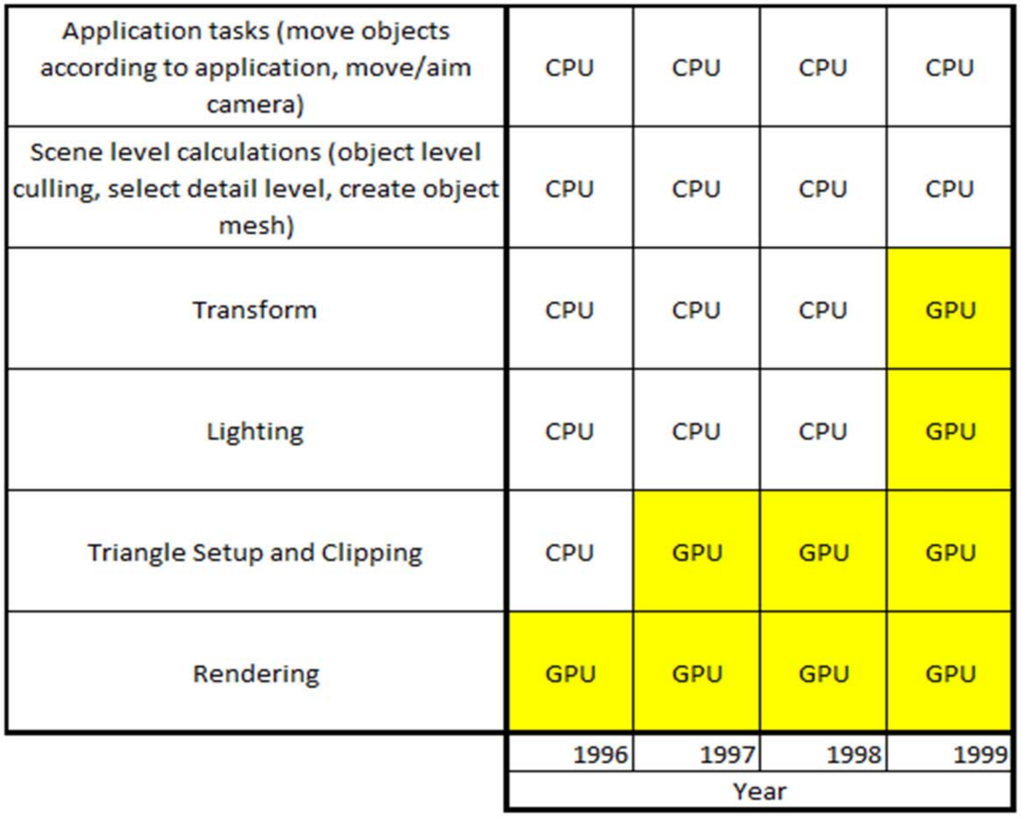
Processor per function


Vertex
T&L evolved to vertex shading
Triangle
Triangle, point, line – setup
Pixel
Flat shading, texturing, eventually Pixel shading
ROP
Blending, Z-buffering, antialiasing
Memory
Wider and faster over the years

Timeline

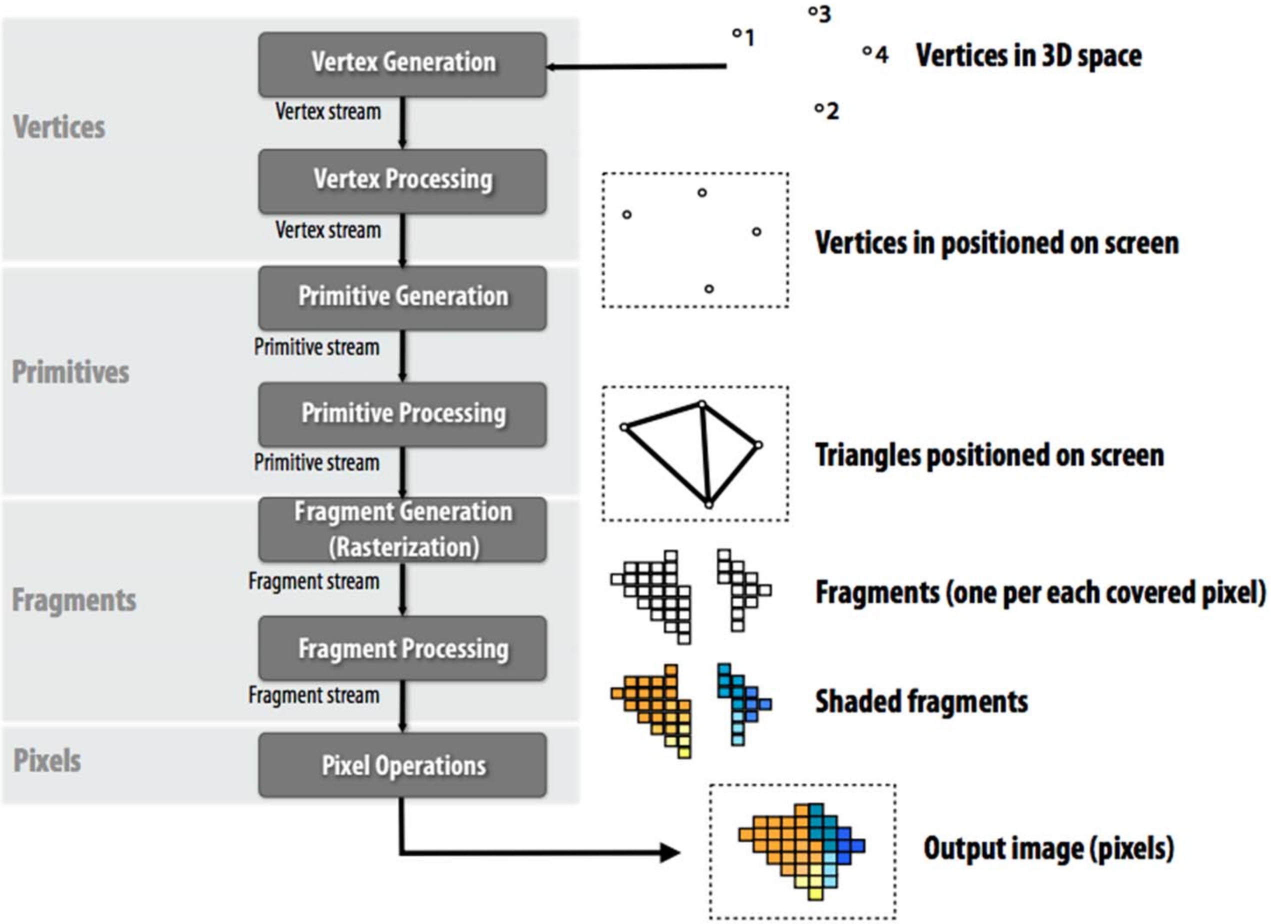
- 1980s
  - Discrete Transistor-Transistor Logic (TTL) frame buffer with graphics processed by CPU
- 1990s
  - Introduction of GPU pipeline - CPU tasks began to be moved to GPU
- 2000s
  - Introduction of Programmable GPU Pipeline
- 2010s
  - GPUs becoming general purpose and also utilized for high performance parallel computations

Movement of Tasks from CPU to GPU



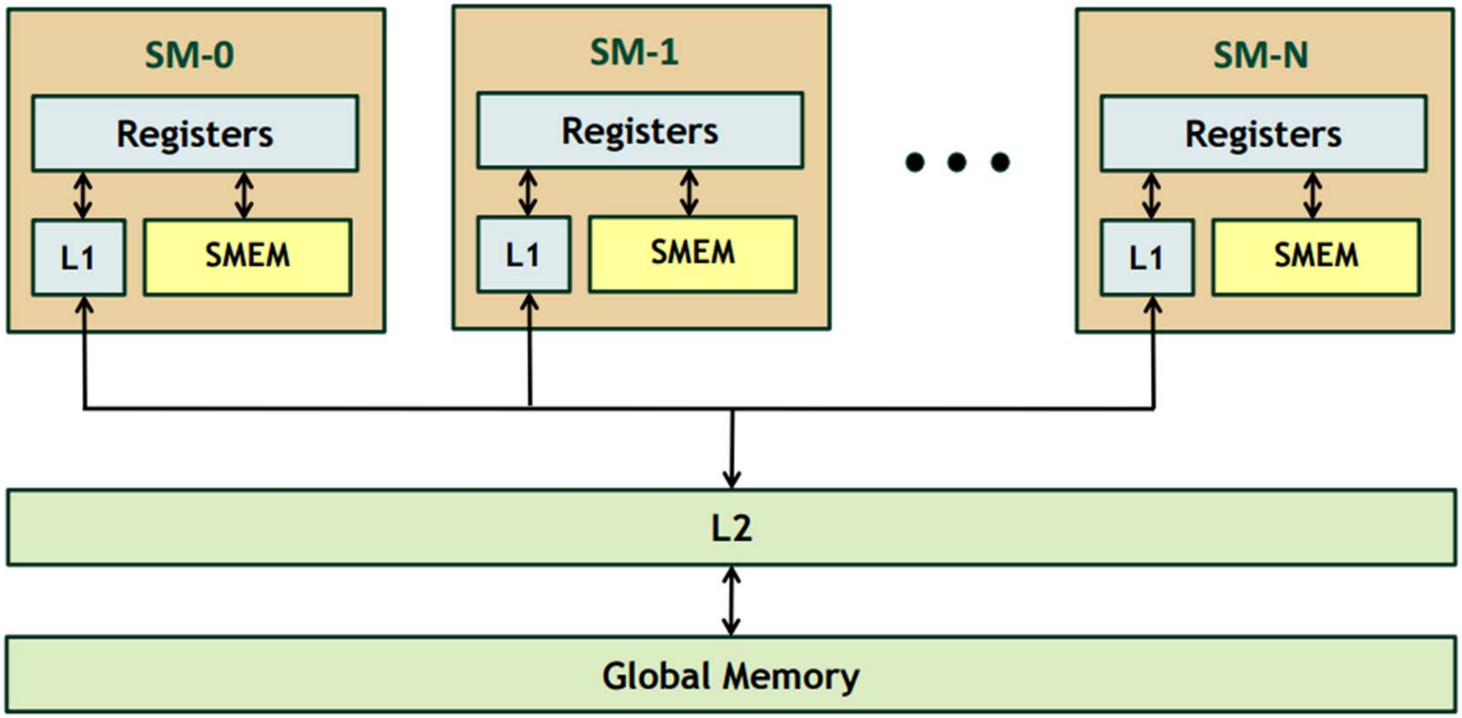
Application tasks (move objects according to application, move/aim camera)	CPU	CPU	CPU	CPU
Scene level calculations (object level culling, select detail level, create object mesh)	CPU	CPU	CPU	CPU
Transform	CPU	CPU	CPU	GPU
Lighting	CPU	CPU	CPU	GPU
Triangle Setup and Clipping	CPU	GPU	GPU	GPU
Rendering	GPU	GPU	GPU	GPU
	1996	1997	1998	1999
	Year			

Introduction to Graphic Processing



CPU Memory Hierarchy

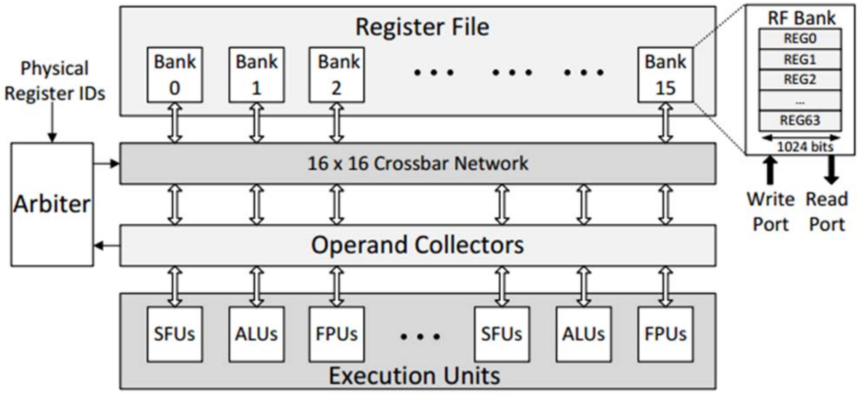
NVIDIA Fermi Memory Hierarchy



GPU Memory Hierarchy

Streaming Multiprocessors (SM) Register Files

- Large and Unified Register File (32768 Registers)
- 16 SMs (128KB Register File per SM), 32 Cores per SM
  -> 2MB across the chip
- 48 warps (1,536 threads per SM)
  -> 21 Registers/Thread
- Multi-Banked Memory
- Very high bandwidth (8,000 GB/s)
- ECC protected

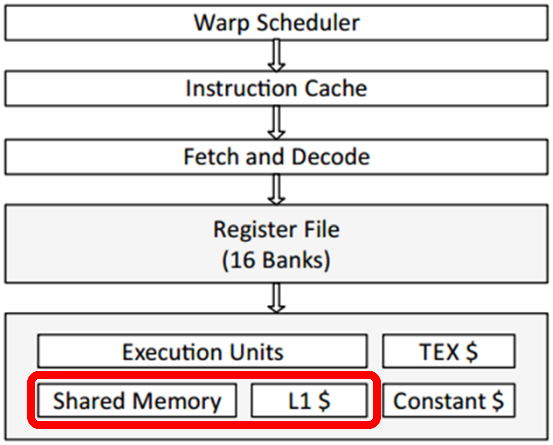


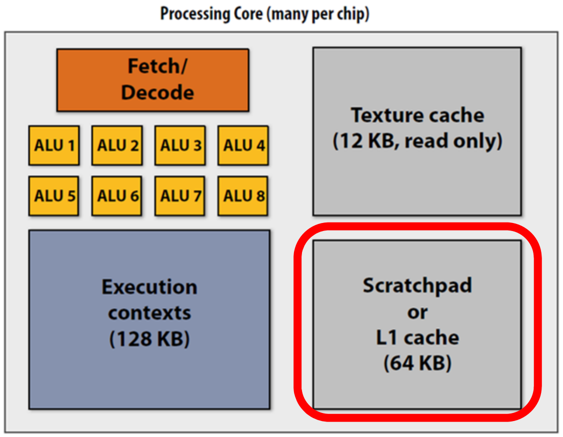
GPU Memory Hierarchy (Cont.)

Shared/L1 Memory

- Configurable 64KB Memory
- 16KB shared / 48 KB L1

  OR 48KB shared / 16KB L1
- Shared Multi-Threads & L1 Private
- Shared Memory Multi-Banked
- Very low latency (20-30 cycles)
- High bandwidth (1,000+ GB/s)
- ECC protected

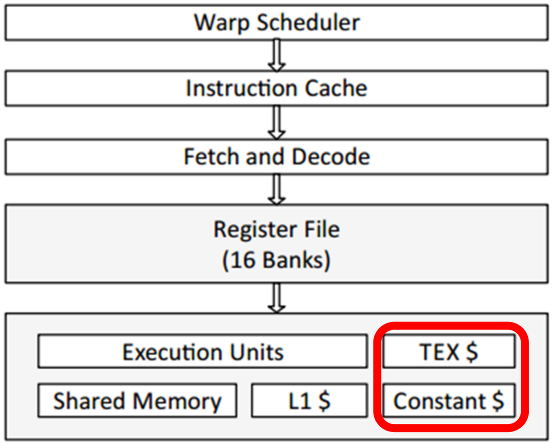




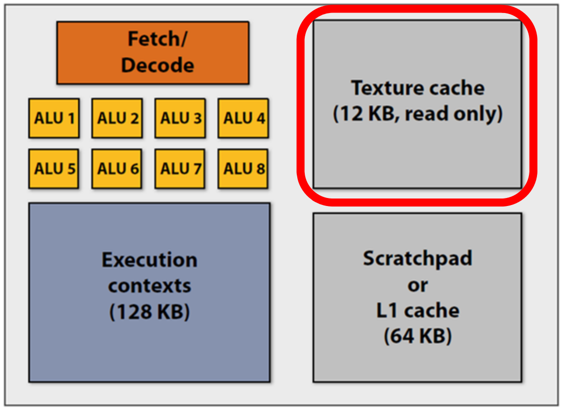
GPU Memory Hierarchy (Cont.)

Texture & Constant Cache

- 64 KB read-only constant cache
- 12 KB texture cache
- Texture cache memory throughput (GB/s): 739.63
- Texture cache hit rate (%): 94.21



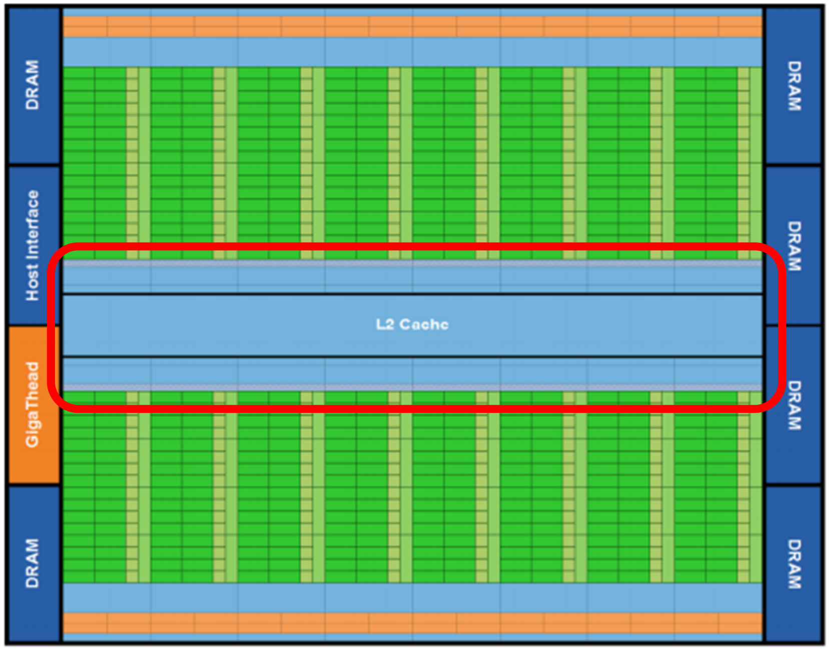
Processing Core (many per chip)



GPU Memory Hierarchy (Cont.)

L2 Cache

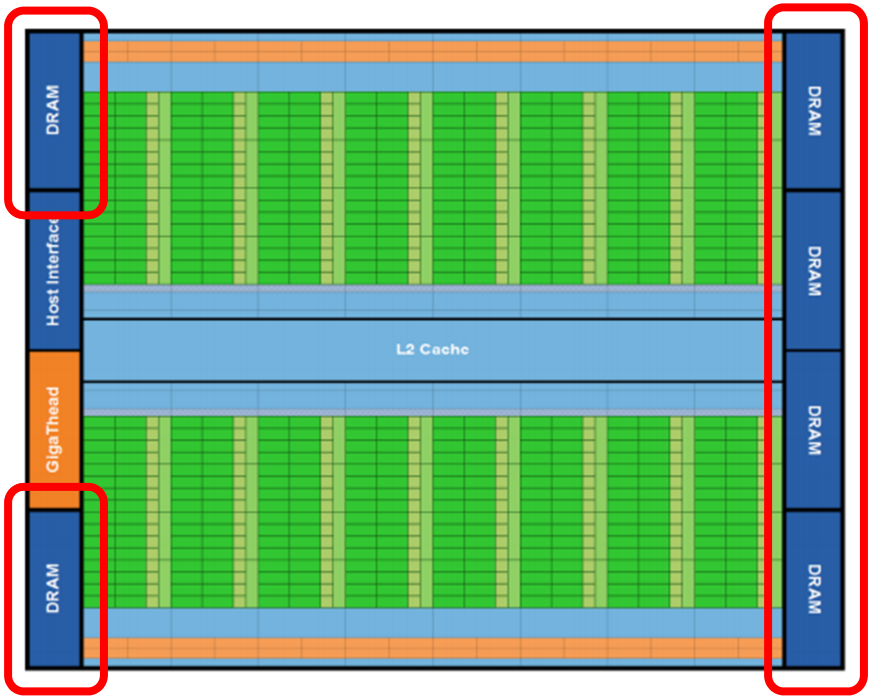
- 768KB Unified Cache
- Shared among SMs
- ECC protected
- Fast Atomic Memory Operations



GPU Memory Hierarchy (Cont.)

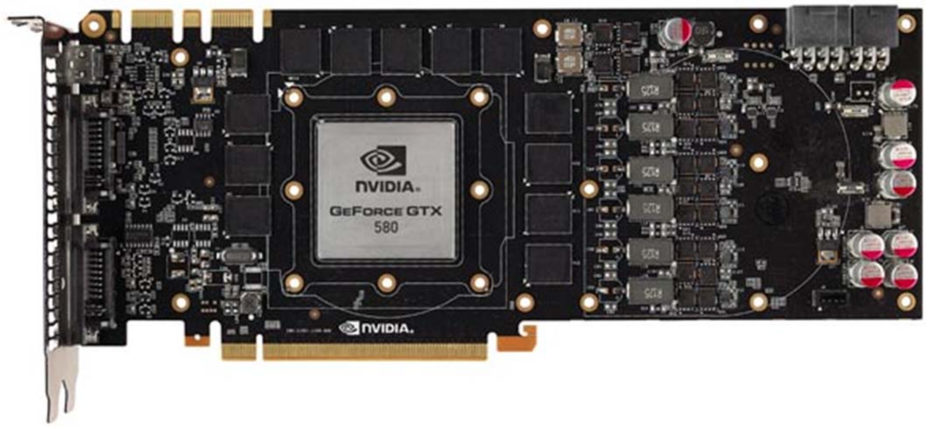
Main Memory (DRAM)

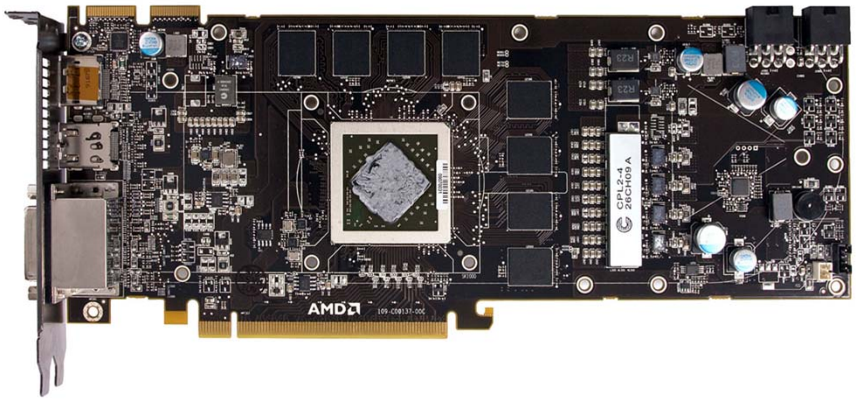
- Accessed by GPU and CPU
- Six 64-bit DRAM channels
- Up to 6GB GDDR5 Memory
- Higher latency (400-800 cycles)
- Throughput: up to 177 GB/s



Different GPU Memory Hierarchies

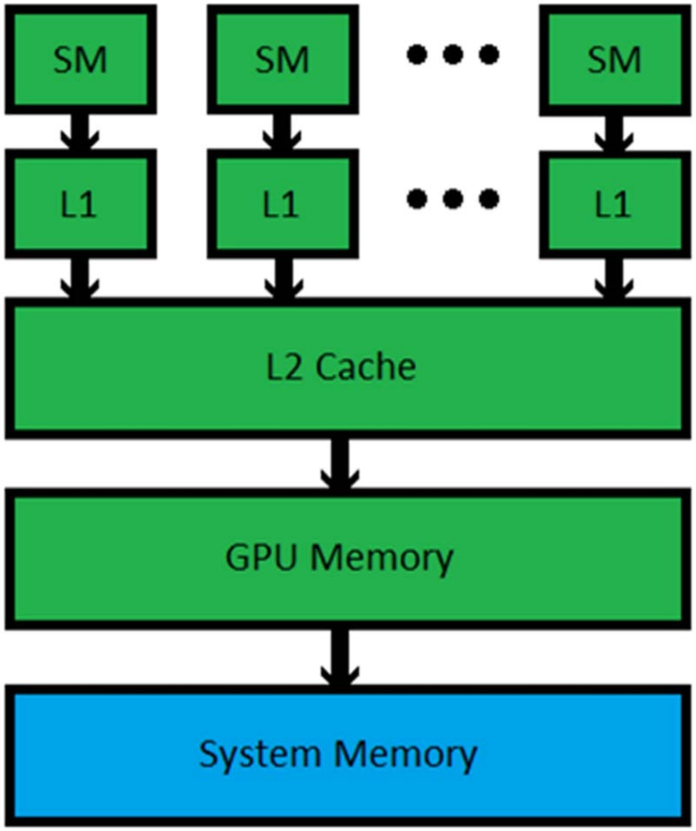
- NVIDIA GeForce GTX 580
- AMD Radeon HD 5870





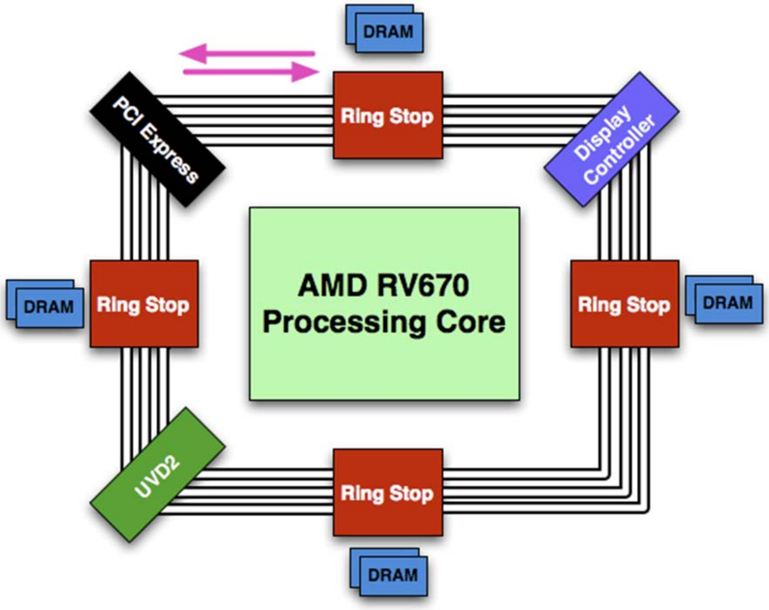
GPU Memory Architecture NVIDIA - Fermi

- On board GPU memory → high bandwidth DDR5 768 MB to 6GB
- L2 shared cache → 512-768 KB high bandwidth
- L1 cache → one for each streaming multiprocessor



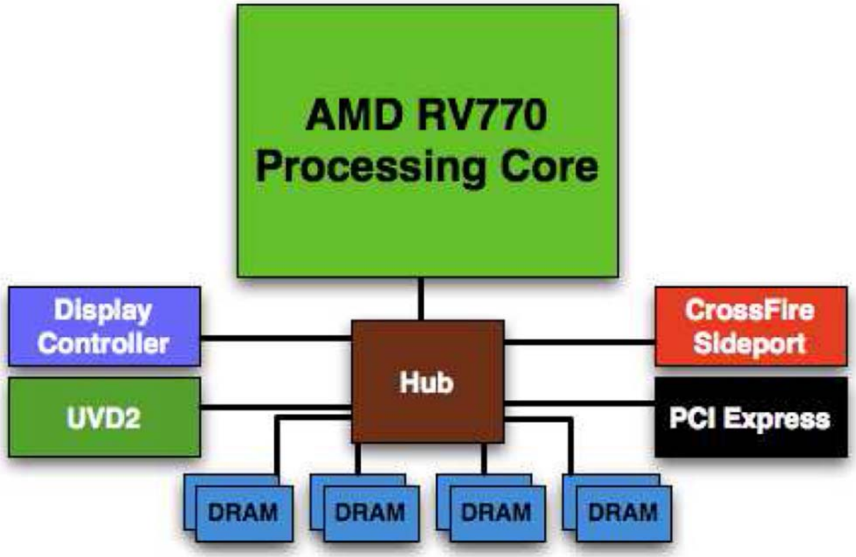
GPU Memory Architecture - AMD Ring

- Mid 2000s design, used to increase memory bandwidth
- To increase bandwidth requires a wider bus
- Ring bus was an attempt to avoid long circuit paths and their propagation delays
- Two 512-bit links for true bi-directional operation
- Delivered 100 GB/s of internal bandwidth



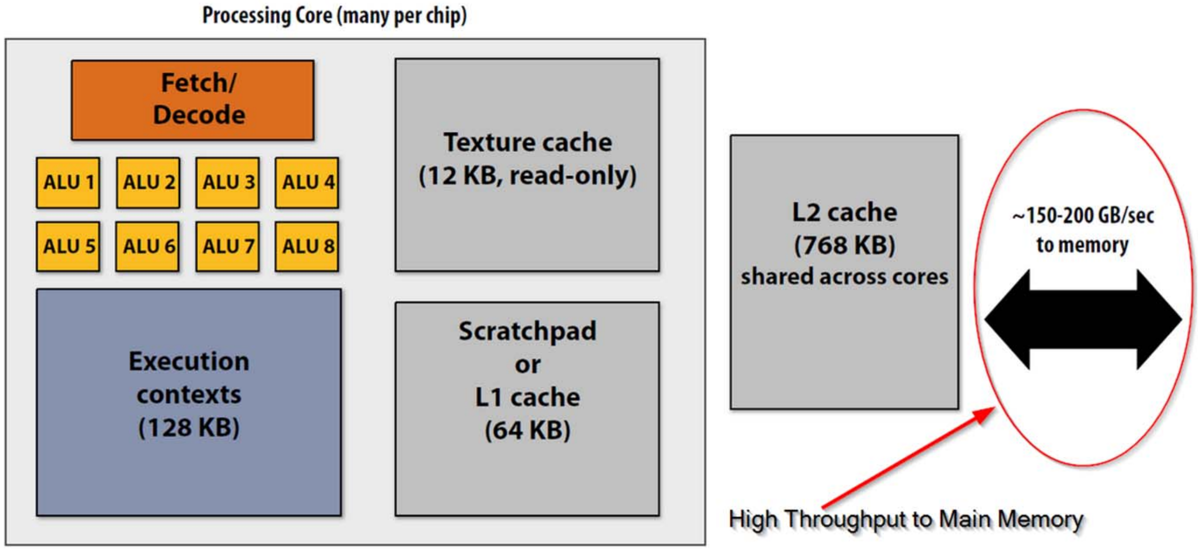
GPU Memory Architecture - AMD Hub

- Ring bus wasted power → all nodes got data even if they did not need it
- Switched hub approach reduces power and latency since data is sent point to point
- AMD increased internal bus width to 2k bits wide
- Maximum bandwidth was 192 GB/s



GPU Bandwidth

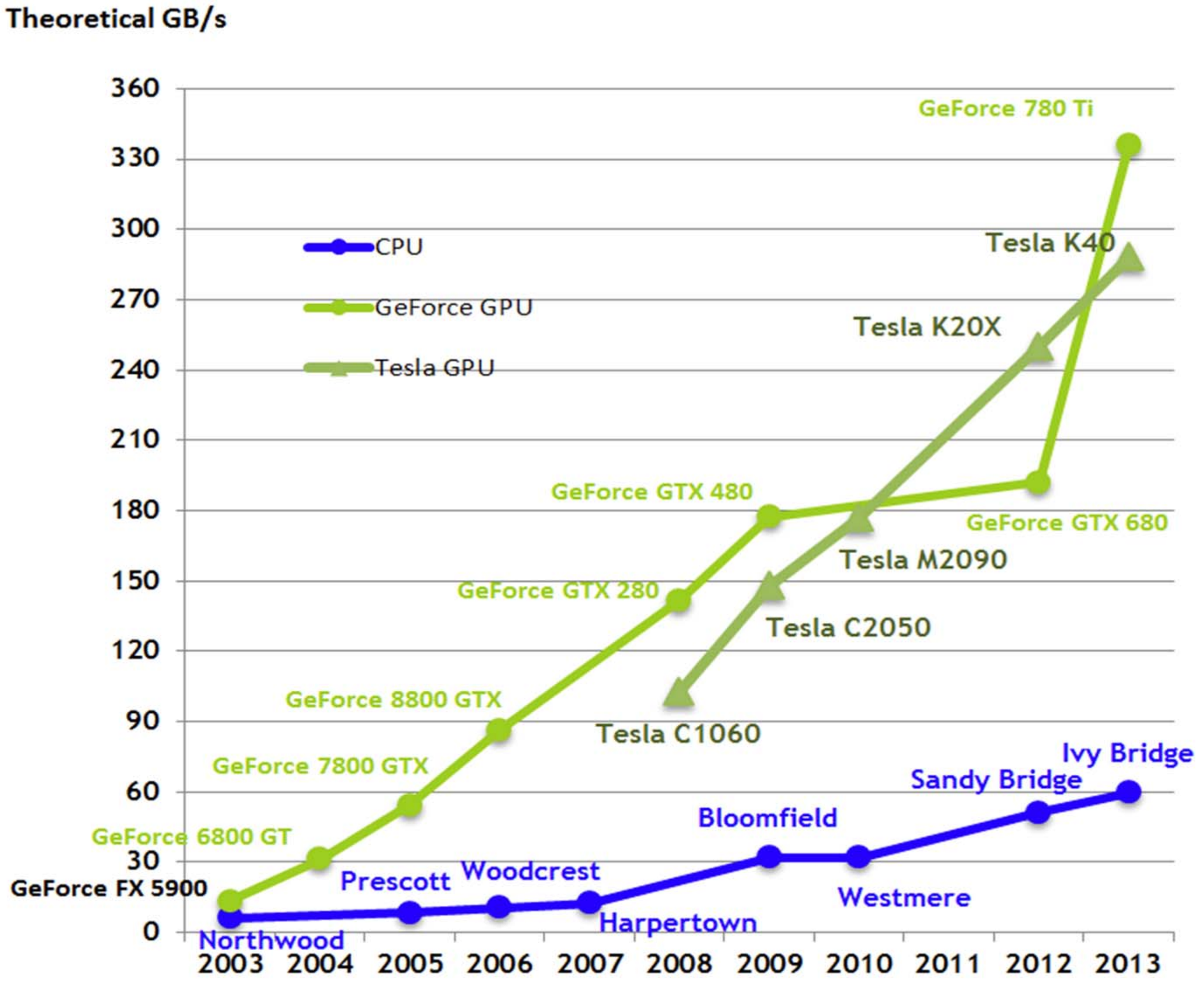
- High bandwidth between main memory is required to support multiple cores
- GPUs have relatively small cache
- GPU memory systems are designed for data throughput with wide memory buses
- Much larger bandwidth than typical CPUs typically 6 to 8 times


Processing Core (many per chip)
Fetch/ Decode
ALU 1
ALU 2
ALU 3
ALU 4
ALU 5
ALU 6
ALU 7
ALU 8
Execution contexts (128 KB)
Texture cache (12 KB, read-only)
Scratchpad or L1 cache (64 KB)
L2 cache (768 KB) shared across cores
~150-200 GB/sec to memory
High Throughput to Main Memory

GPU Bandwidth (Cont.)

- Bandwidth Use Techniques
    - Avoid fetching data whenever possible
        - Share/reuse data
        - Make use of compression
        - Perform math calculations instead of fetching data when possible → math calculations are not limited by memory bandwidth

GPU vs. CPU Bandwidth Growth



GPU Latency

- Big register files
- Dedicated shared memory (configurable)
- Multi-banked memory

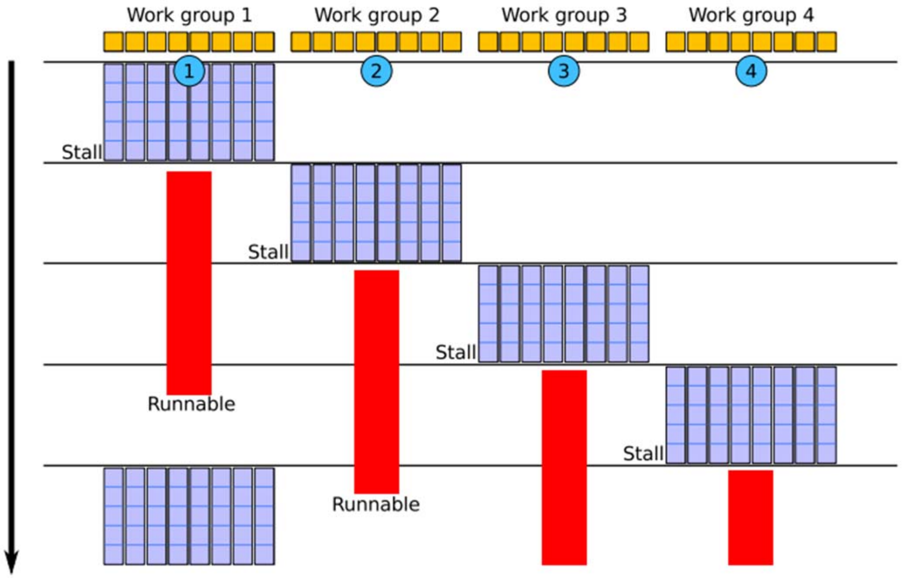
↓

- Reuse data in dedicated memories
- Focus on parallelism

GPU Latency (Cont.)

Latency Hiding

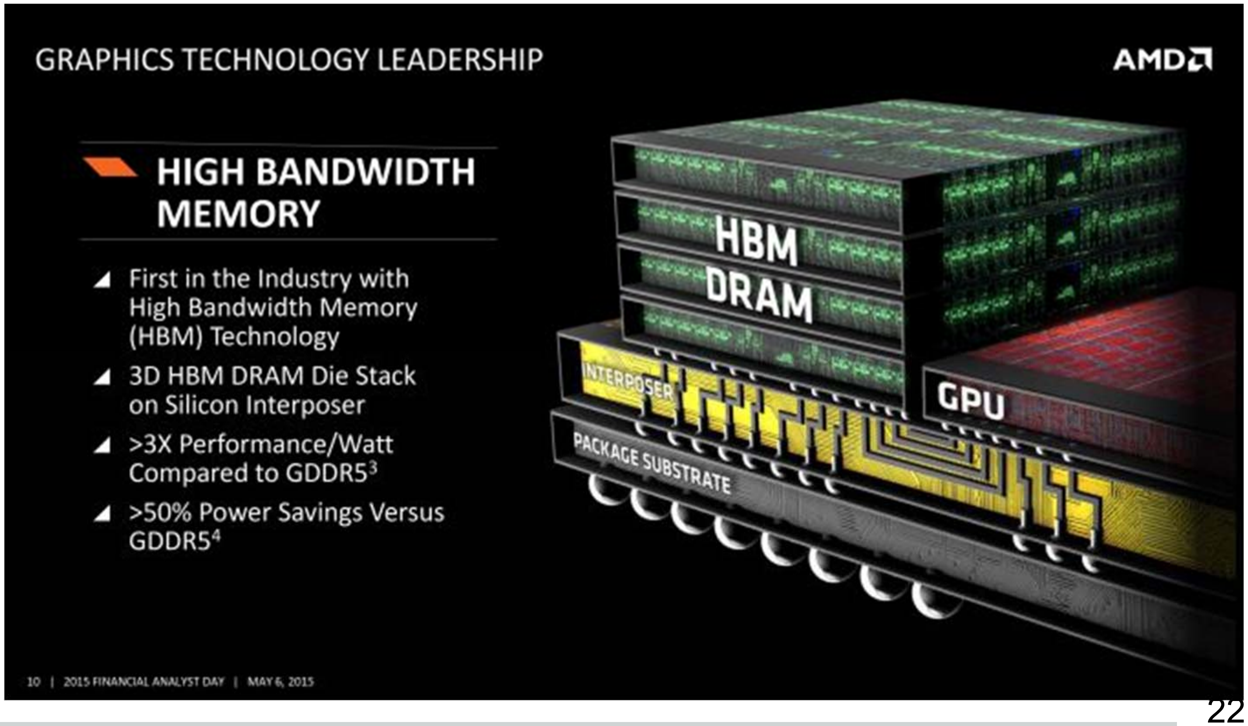
- 1,536 threads per SM (48 warps)
- 32 threads per warp (SIMT)
- 1000 cycles memory access stall
- Switch to another group to hide latency



Future of GPU Memory

- New manufacturing process → High Bandwidth Memory
- Stacking DRAM dies on top of each other thus allowing for close proximity between DRAM and processor

- Allows for very high bandwidth memory bus
- Due to stacking will be harder to cool



GRAPHICS TECHNOLOGY LEADERSHIP

AMD

HIGH BANDWIDTH MEMORY

- First in the Industry with High Bandwidth Memory (HBM) Technology
- 3D HBM DRAM Die Stack on Silicon Interposer
- >3X Performance/Watt Compared to GDDR5[3]
- >50% Power Savings Versus GDDR5[4]

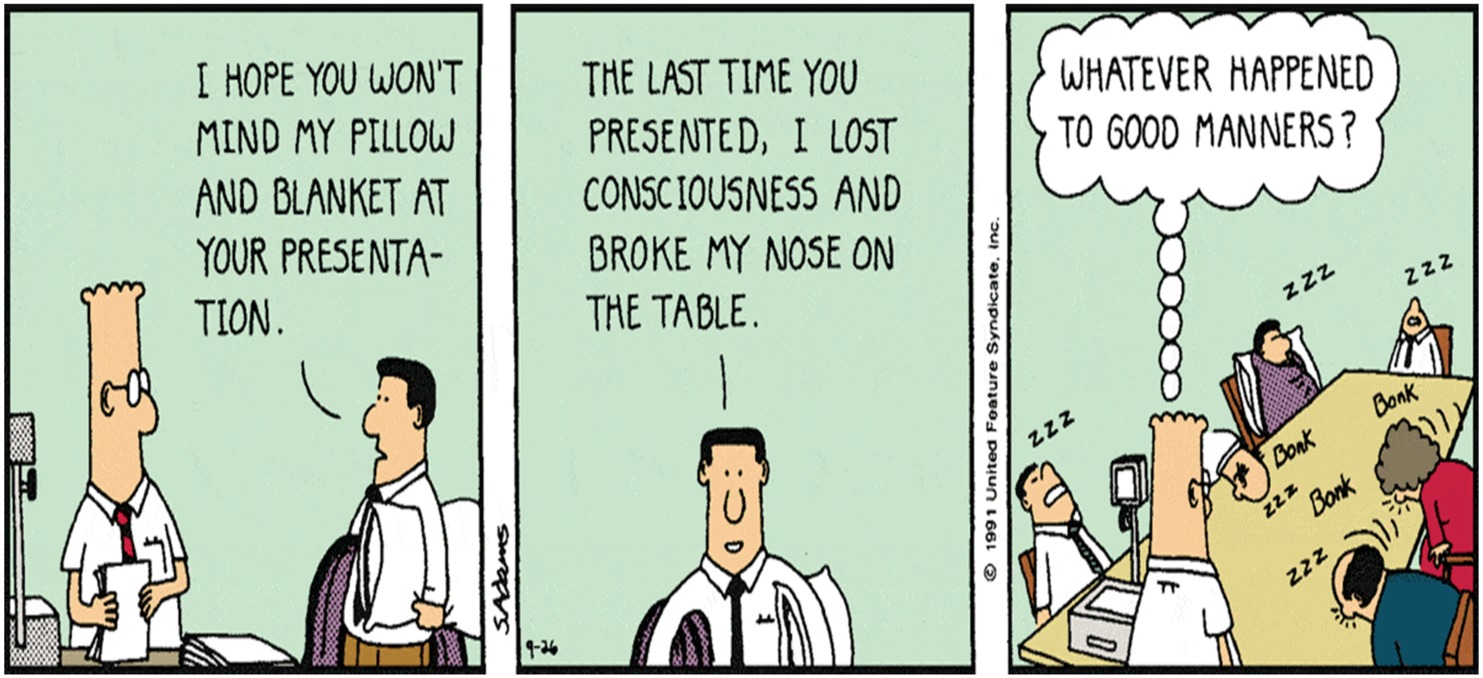
10 | 2015 FINANCIAL ANALYST DAY | MAY 6, 2015

References

- Fatahalian, Kayvon. “The GPU Memory Hierarchy”. Carnegie Mellon University.
- Cao Young. “GPU Memory II”. Virginia Tech.
- McClanahan Chris. “History and Evolution of GPU Architecture”. Georgia Tech.
- “CUDA Memory and Cache Architecture”. Supercomputing Blog.
- “Radeon X1800 Memory Controller”. ATI.

Q&A

Thank you!


I HOPE YOU WON'T MIND MY PILLOW AND BLANKET AT YOUR PRESENTATION.
S.Adams
9-26
THE LAST TIME YOU PRESENTED, I LOST CONSCIOUSNESS AND BROKE MY NOSE ON THE TABLE.
© 1991 United Feature Syndicate, Inc.
WHATEVER HAPPENED TO GOOD MANNERS?
zzz
zzz
zzz
Bonk
Bonk
Bonk
zzz
zzz