Xbox 360 System Architecture, 'Anderews, Baker

Xbox 360

- Three CPU cores each with two hardware threads.
- Developers can assign their software threads to specific CPU threads and XAudio2 and XACT can easily be set to run on their own hardware threads.

Georgia Tech College of Computing

Xbox 360 System Block Diagram

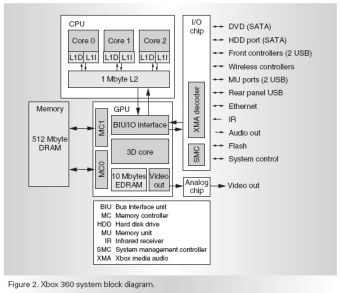


Figure 2. Xbox 360 system block diagram.

Georgia Tech College of Computing

Xbox 360 Architecture

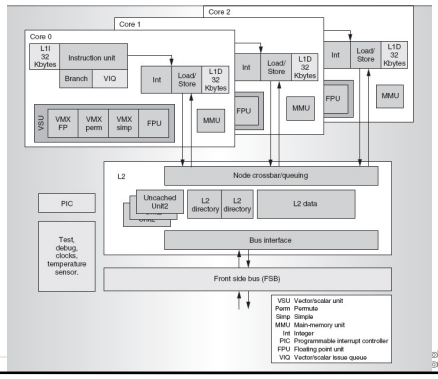
- 3 CPU cores
  - 4-way SIMD vector units
  - 8-way 1MB L2 cache (3.2 GHz)
  - 2 way SMT
- 48 unified shaders
- 3D graphics units
- 512-Mbyte DRAM main memory
- FSB (Front-side bus): 5.4 Gbps/pin/s (16 pins)
- 10.8 Gbyte/s read and write

Xbox 360 vs. Windows

- Xbox 360: Big endian
- Windows: Little endian

http://msdn.microsoft.com/en-us/library/cc308005(VS.85).aspx

Xbox 360 CPU Block Diagram



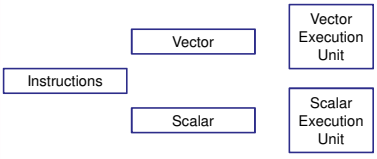
On-chip caches

- L2 cache :
  - Greedy allocation algorithm
  - Different workloads have different working set sizes
- 2-way 32 Kbyte L1 I-cache
- 4-way 32 Kbyte L1 data cache
- Write through, no write allocation
- Cache block size :128B (high spatial locality)

Georgia Tech College of Computing

Core

- 2-way SMT,
- 2 insts/cycle,
- In-order issue
- Separate vector/scalar issue queue (VIQ)



Georgia Tech College of Computing

VMX 128

- Four-way SIMD VMX 128 units:
  - FP, permute, and simple
- 128 registers of 128 bits each per hardware thread
- Added dot product instruction (simplifying the rounding of intermediate multiply results)
- 3D compressed data formats . Use compressed format to store at L2 or memory. 50% of space saving.

Georgia Tech College of Computing

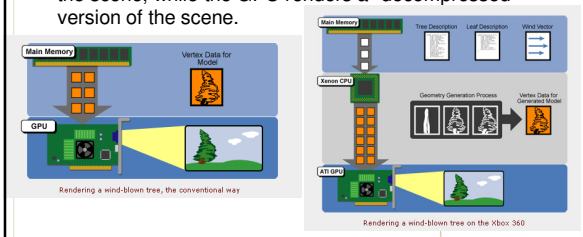
Cache-set-locking

- Threads owns a cache sets until the instructions retires.
- Reduce cache contention.
- Common in Embedded systems
- Use L2 cache as a FIFO buffer: sending the data stream into the GPU

Georgia Tech College of Computing

Procedural Synthesis

- Microsoft refers to this ratio of stored scene data to rendered vertex data as a **compression ratio**, the idea being that main memory stores a "compressed" version of the scene, while the GPU renders a "decompressed" version of the scene.

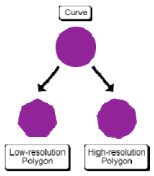


From http://arstechnica.com/articles/paedia/cpu/xbox360-1.ars/2

Georgia Tech College of Computing

Real-time Tessellation

- Tessellation: The process of taking a higher order curve and approximating it with a network of small flat surfaces is called tessellation.
- Traditional GPU: Artist
- Xbox 360: using Xeon
- Real time tessellation
  - Data compression
  - Dynamic Level of Detail (LOD)



From http://arstechnica.com/articles/paedia/cpu/xbox360-1.ars/2

Georgia Tech College of Computing

Stream Optimizations

- 128B cache line size
- Write streaming:
  - L1s are write through, write misses do not allocate in L1
  - 4 uncacheable write gathering buffers per core
  - 8 cacheable, non-sequential write gathering buffers per core
- Read streaming:
  - 8 outstanding loads/prefetches.
  - xDCBT: Extended data cache block touch, brining data directly to L1 , never store L2
  - Useful for non-shared data

Georgia Tech College of Computing

CPU/GPU

- CPU can send 3D compressed data directly to the GPU w/o cache
- Geometry data
- XPS support:
  - (1): GPU and the FSB for a 128-byte GPU read from the CPU
  - (2) From GPU to the CPU by extending the GPU's tail pointer write-back feature.

Georgia Tech College of Computing

Tail Pointer write-back

- Tail pointer write-back: method of controlling communication from the GPU to the CPU by having the CPU poll on a cacheable location, which is updated when a GPU instruction writes an updated to the pointer.
- Free FIFO entry
- System coherency system supports this.
- Reduce latency compared to interrupts.
- Tail pointer backing-store target

Georgia Tech College of Computing

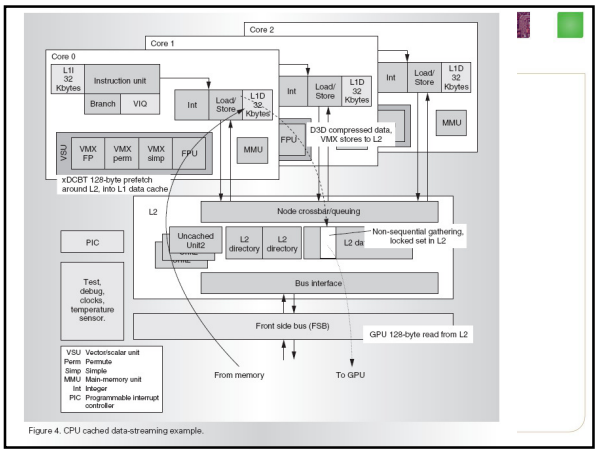
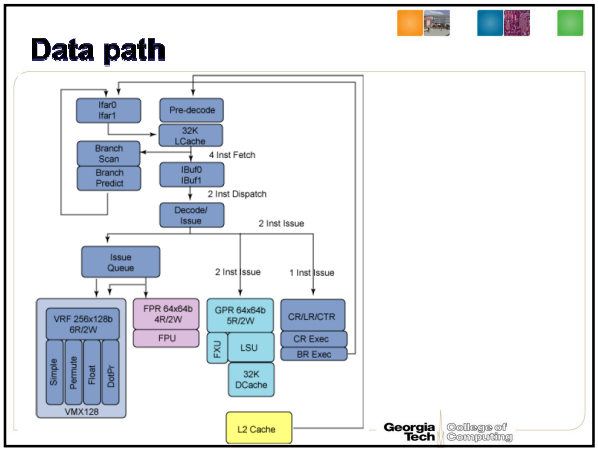
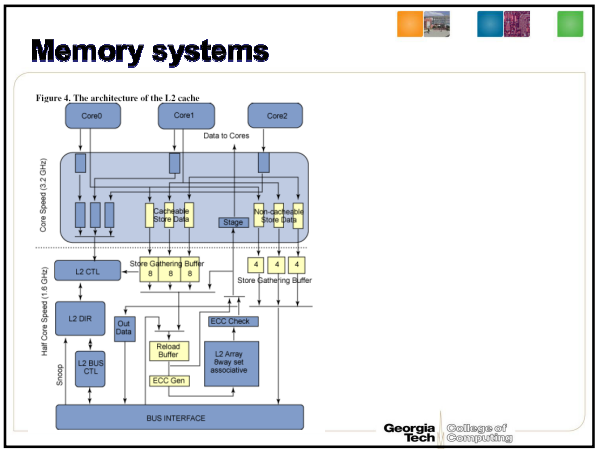


Figure 4. CPU cached data-streaming example.





XMA

- A hardware real-time XMA decoder.
- XMA data read from disk or memory can be rapidly decoded.
- Xbox 360 is capable of decoding hundreds of mono streams simultaneously
- The XMA hardware can also perform looping in hardware for in-memory sounds.
- Most date format in Xbox 360 will be in the XMA format.

Georgia Tech College of Computing