Overview

- 3-way symmetric multiprocessor
  - Each CPU core is a specialized PowerPC chip running @ 3.2 GHz with custom vector floating point unit (VMX128)
  - 64 KB L1 cache (Harvard Architecture) per core
  - Shared 1 MB L2 cache for all three cores
  - 21.6 GB/sec Bandwidth Front Side Bus (10.8 GB/s write & 10.8 GB/s read)
- ATI custom GPU running @ 500 MHz
  - 48 parallel combined vector & scalar shader ALUs
  - 10 MB Embedded DRAM (EDRAM)
- 512 MB unified Main Memory @ 700 MHz
- 20 GB Hard Disk
General Architecture Overview

CPU – Xenon (1)

- 64 bit PowerPC architecture
  - Supports all PowerPC ISA
  - Custom Vector Multimedia Extension (VMX) SIMD instruction set specialized for graphic workloads
- Shared 8-way set associative 1 MB Cache L2
- Per core features:
  - 2 per cycle in-order issuance & decoupled Vector/Scalar issue queue
  - 64 KB split L1 cache:
    - 2-way set associative 32 KB Instructions Cache
    - 4-way set associative 32KB Data Cache, store-through, non-allocation on write miss, with non-blocking access so a cache miss doesn’t hold up a subsequent hit (hit under miss)
CPU – Xenon (2)

- Per core features (cont...):
  - Capable of handling up to two simultaneous threads of execution
  - Pipeline Execution Units:
    - Branch Unit (4KB Branch History Table)
    - Integer Unit (Fixed Point Unit)
    - Load/Store Unit
    - Scalar FPU
  - 3 VMX128 Units:
    - Floating Point Unit, Permute Unit, Simple Unit
    - 128 128-bit Vector Registers

Xenon Architecture
XPS – XBox Procedural Synthesis (1)

- Traditional 3D Graphics Rendering (PC)

XPS – XBox Procedural Synthesis (2)

- Technique for fast data streaming in which the CPU acts as a data decompressor
- Makes optimal use of system bandwidth and main memory by dynamically generating lower-level geometry data from statically stored higher-level geometry data
  a. Information that represents the objects on screen is stored in a compressed way in main memory (e.g. higher order curves as Bezier curves)
  b. Data is sent to CPU
  c. CPU decompresses and generates vertex data
  d. Vertex Data is sent to GPU
  e. GPU renders image with given vertex data
Write Streaming Mode Cache

- CPU can produce decompressed vertex data at a faster rate than the GPU can consume.
- L2 cache used to store data waiting to be consumed by GPU
- The programmer can reserve certain space on the cache to act as a FIFO output queue, to store outgoing data
- The data generation thread feeds its vertex data output directly into this FIFO queue (bypassing L1 cache)
- GPU reads that vertex data directly from this queue using a modified DMA protocol

Read Streaming Mode Cache

- Data can be fed directly into the L1 cache or into the registers (bypassing the L2 cache or both caches in the case of registers)
- Useful to provide a way of preventing streaming media data that won’t be reused, from polluting the caches
Xenon Pipeline (1)

- Difficult instructions implemented through microcode (at dispatch they are cracked and converted into multiple micro-ops)
- 23 stages deep (plus VMX128)
- 4 Stages Fetch
- 4 Stages Branch Prediction
- 3 Stages Decode
- 3 Stages Issuance & Dependency Check
- Load/Store instructions of 9 cycles latency
- VMX operations from 4 to 14 cycles latency depending on operation

Xenon Pipeline (2)
Conclusions

- Improvements:
  - First console to include 3 core multiprocessor design
  - XPS (Xbox Procedural Synthesis)
  - Cache Read & Write Streaming
  - Programmers challenge: Parallel Programming
    - Thread Level Parallelism
    - Branch Hints

References

- XBOX 360 System Architecture; Jeff Andrews & Nick Baker (Microsoft)
  http://www.hotchips.org/archives/hc17/
- Application-customized CPU design: The Microsoft Xbox 360 CPU story; Jeffrey Brown (IBM)
- Inside the Xbox 360, part I: procedural synthesis and dynamic worlds; Jon Stokes (ars technica)
  http://arstechnica.com/articles/paedia/cpu/xbox360-1.ars/1
- Inside the Xbox 360, Part II: the Xenon CPU; Jon Stokes (ars technica)
  http://arstechnica.com/articles/paedia/cpu/xbox360-2.ars/1