Larrabee

A Many-Core x86 Architecture for Visual Computing

Seiler, L., Carmean, D., et al. SIGGRAPH'08

Presented by Valentin Pistol

CPS221 Spring 2010

Context

- Graphics market evolving very fast
  - Mostly driven by gaming
  - Trend
    - Real time realistic graphics require more and more computation
    - Low power - big concern for mobile (laptops, phones)
    - Converging CPU and GPU apps
    - Integrate CPU and GPU in a single package and even SoC
    - Driven by increase in transistor count and shrink size
    - Extract parallelism, almost everyone has a graphics card (desktop or mobile)
    - Highly programmable units
    - HPC market, scientific workloads, throughput oriented

Objectives and reality

- Intel Larrabee
  - Many x86 cores, wide vector processor units, some fixed functional logic, software renderer → aims for high performance and flexibility
- Designed by Intel's Hillsboro, Oregon (Nehalem)
- Expected
  - Late 2009 release on 45nm process
  - 2010 shrink to 32nm
- Killed in December '09 by Intel (way behind schedule)
  - 1st generation
  - Platform will be used for multi-core hw/sw research and development

- Sources
  - http://arstechnica.com/hardware/news/2009/12/intels-larrabee-gpu-put-on-ice-more-news-to-come-in-2010.ars

Inside Larrabee

- Hybrid software/hardware GPU
- Software rasterization and interpolation
  - Optimized for particular workload
  - Special purpose equations
  - Parallelizable rasterization and flexible rendering pipeline placement
- Software instruction and thread scheduling (compiler)
  - Dynamic load balancing – e.g. raytracing
- Fixed (hardware) texture unit (with 32K cache)
- Limitations (as of paper prototype)
  - Application sys call porting
  - Application recompilation

Architecture

	Tight Synchronization	Data Path Divergence
Vectors	Good	Bad
Threads	Bad	Fine

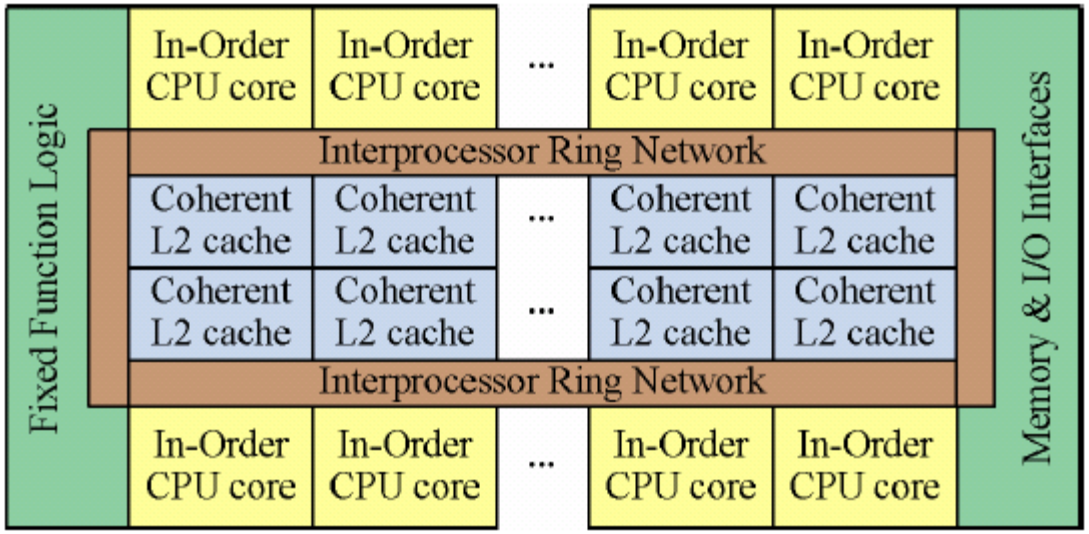


***Figure 1**: Schematic of the Larrabee many-core architecture: The number of CPU cores and the number and type of co-processors and I/O blocks are implementation-dependent, as are the positions of the CPU and non-CPU blocks on the chip.*

Specs

- Simple in-order cores (16+), fully x86 ISA compatible + vector instructions
  - High power efficiency
  - Based on Pentium P54C (intro in ’94 – embedded use)
- High bandwidth ring network (512 bit wide for each direction)
- Shared and coherent cache hierarchy
- L1 $
  - 32K L1I$ + 32K L1D$, per core
- L2 $
  - 256K local L2 cache slice, per core
  - Local is faster (obviously)
  - Special instructions for cache manipulation (eviction hints, prefetch, streams)
- Explicit DMA transfers
- Latency hiding
  - 4-way multithreading per core (interleaved)
  - Cell SPE has 1, PPU has 2
- Main Cell core manages and runs OS
  - “The PPE is the main processor of the Cell BE, and is responsible for running the operating system and coordinating the SPEs. “
  - Larrabee identical cores
- Source:
  - http://www.ibm.com/developerworks/power/library/pa-cellperf/

Rendering Pipeline

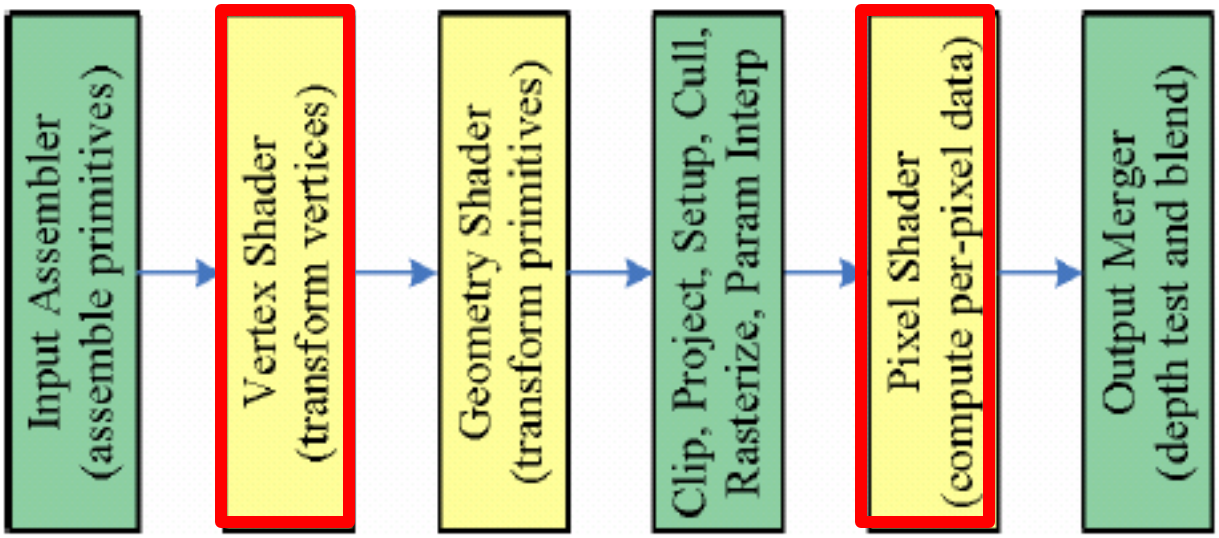
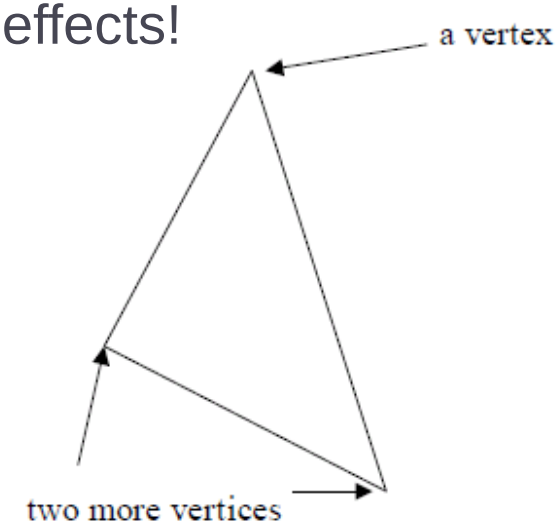


***Figure 2**: Simplified DirectX10 Pipeline: Yellow components are programmable by the user, green are fixed function. Memory access, stream output, and texture filtering stages are omitted.*

Vertex Shaders

- What is a vertex?
  - “A vertex is the corner of the triangle where two edges meet, and thus every triangle is composed of three vertices.” – NVIDIA
- Use?
  - Special effects!



Example: Vertex Data

position:	{X, Y, Z, W}
color:	{Red, Green, Blue, Alpha}
texture1:	{S, T, R, Q}
texture2:	{S, T, R, Q}
.....	
texture-n:	{S, T, R, Q}
fog:	{F}
specularity:	{P}

- Sources:
  - http://www.nvidia.com/object/feature_vertexshader.html
  - http://www.nvidia.com/attach/4049

Pixel Shaders

- “Graphics function that calculates effects on a per-pixel basis.” - NVIDIA
- Use?
  - Incredibly realistic material and lighting effects

- Sources:
  - http://www.nvidia.com/object/feature_pixelshader.html

Performance

- Theoretical SP (single precision)
  - 32 cores × 16 single-precision float SIMD/core × 2 FLOP (fused multiply-add) × 1GHz = **1 TFLOPS – slow!**
  - Comparison?
    - AMD ‘08 ATI Radeon HD4800 series – 1TFLOPS
    - ATI Radeon 4870X2 card Aug ‘08 – 2.4 TFLOPS
    - ATI Radeon HD 5970 (2xGPU) Nov ’09 – 4.6 TFLOPS !!!
    - ATI FirePro V8800 April 7 ’10 – 2.6 TFLOPS
      - 1600 Stream Processors, < 225W
    - NVIDIA GTX480 (Fermi) April ’10 – 1.35 TFLOPS
      - 6 months late, expensive, extremely hot (~ 100C/210F), loud and power hungry
      - System load: 480W vs 367W (Radeon 5870)

- Sources:
  - http://en.wikipedia.org/wiki/FLOPS
  - http://www.amd.com/us/products/workstation/graphics/ati-firepro-3d/v8800/Pages/v8800-specifications.aspx
  - http://techreport.com/articles.x/18682
  - http://zikkir.net/tech/11889
  - http://www.hardocp.com/article/2010/03/26/nvidia_fermi_gtx_470_480_sli_reviewww

Larrabee Programming

- Transparent memory management
  - All memory on Larrabee is shared by all processors
  - But NVIDIA just launched Fermi with coherency and L1/L2 caches…
- Predication
  - Power-efficient – masks don't compute results for unused lanes
- Gather/scatter
  - Limited by cache speed
- Pthreads, OpenMP, Intel TBB support
- Compiler with auto-vectorization
  - How good?
- Tight integration with host
  - Proxy Larrabee I/O functions – read/write/open/close…
- Full C++ support
  - Available on CUDA now
- *"Profile it once it's running, find out which bits need **love**"* – Intel, SIMD Programming Larrabee, GDC 2009

Look to future and questions

- Hybrid CPU/GPU future?
- Simple cores/logic → less errors/faults/bugs → good yield?
- AMD Fusion project
  - AMD Llano samples in H2'10
  - Target notebook market
  - APU (Application Processor Unit)
  - OO 3GHz quad-core CPU and GPU on single die - 32nm
- Intel Clarkdale 3.46GHz – launch Q1 '10
  - Nehalem micro-architecture
  - Two dies on package – 32nm CPU , 45m integrated graphics
- Linear scaling, really?
  - Game engines hard to parallelize
- Feeding cores with enough bandwidth?
  - Memory subsystem very costly and power hungry
  - Bandwidth doesn't scale linearly across technology nodes
- Crysis game benchmark missing? (released Nov '07)
  - Kills all but the very latest GPUs
- Raytracing?
  - Current raytracers 10-20M+rays /s
  - NVIDIA OPTIX Raytracer released Jan '10, supports Fermi

Sources

- http://apcmag.com/amd-offers-detaisl-on-llano-gpucpu-hybrid-for-laptops.htm
- http://arstechnica.com/business/news/2010/02/amd-reveals-fusion-cpugpu-to-challege-intel-in-laptops.ars
- http://www.xbitlabs.com/news/cpu/display/20090728142821_Intel_Clarkdale_3_46GHz_Clock_Speed_32nm_Process_Tech_Launch_in_Q1_2010.html
- http://developer.nvidia.com/object/optix-beta.html

Thanks!

Intel® Desktop Processor Codename Clarkdale

Intel® Desktop processors codename Clarkdale

- 32 nm, 2nd Generation Hi-K process CPU
- 45nm, Hi-K Process, Integrated Graphics

Key Features[2]:

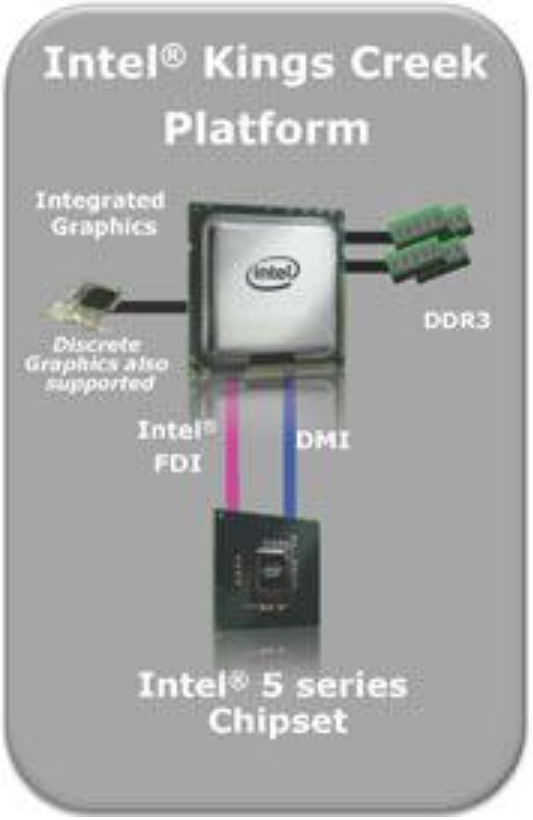
- 32nm Nehalem Microarchitecture (Westmere)
  - Intel® Turbo Boost Technology [1]
  - Intel® Hyper-Threading Technology [2](2 Cores, 4 threads)Up to
- Up to 4MB of Intel® Smart Cache
- Integrated Memory Controller (IMC) – 2ch DDR3, up to 1333
- Integrated Graphics or Discrete graphics support (1x16, 2x8)[4]
- Advanced Encryption Standard (AES) acceleration

Socket:

- LGA1156 Socket (drop-in compatible with Intel® Core™ i7-800 processor series and Intel® Core™ i5-700 processor series)

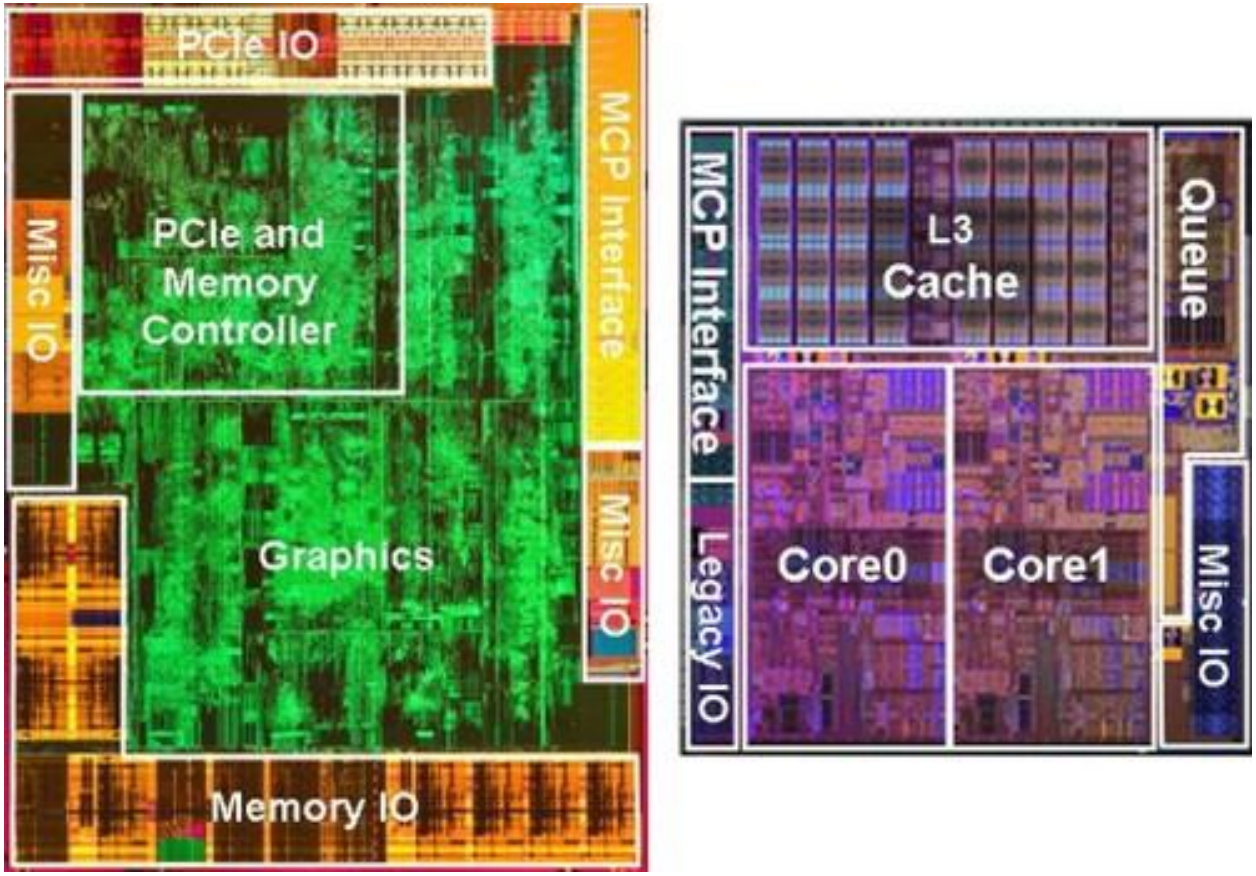
Platform Compatibility:

- Intel® 5 series Chipset



Source: http://www.legitreviews.com/article/1091/2/

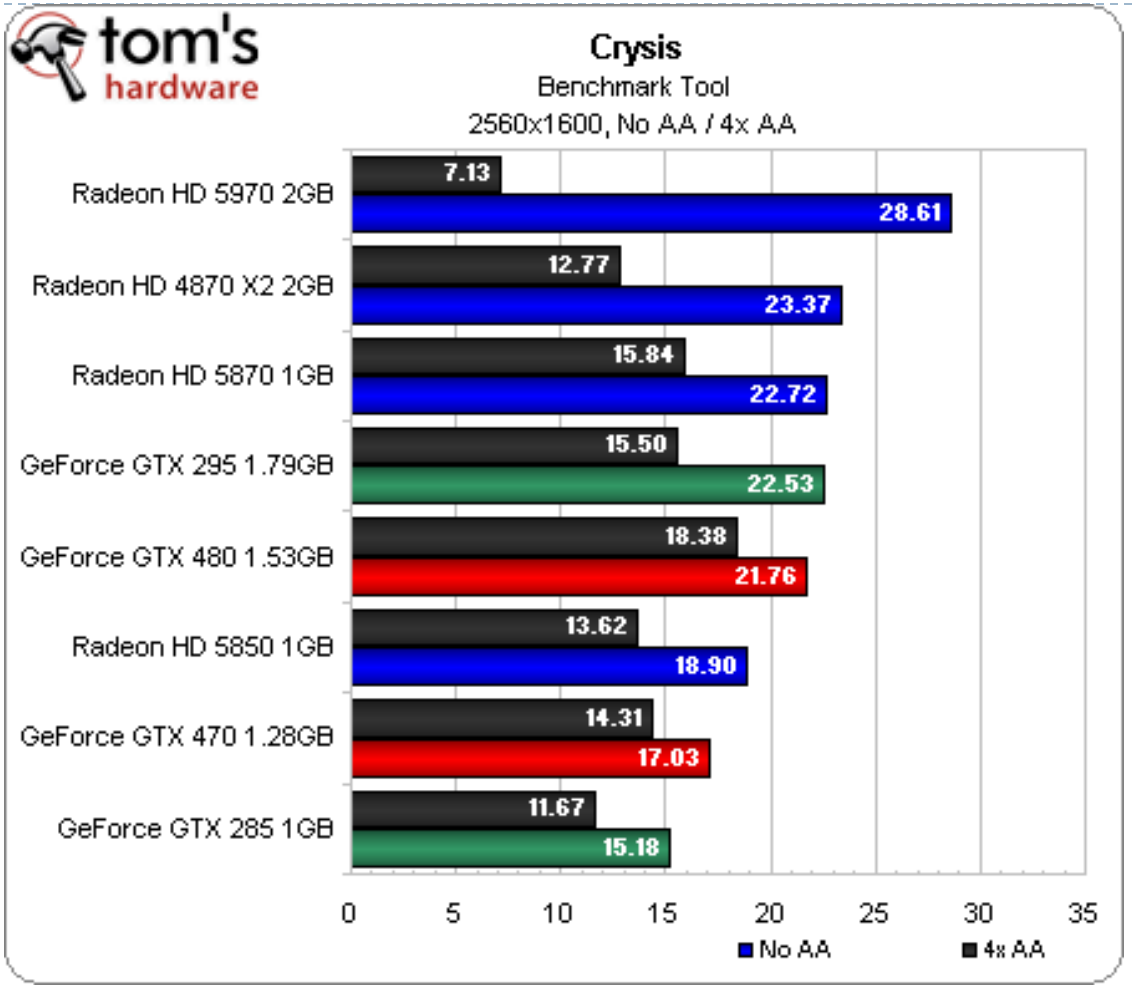
Clarkdale GPU and CPU Dies To Scale



Source: http://hothardware.com/Articles/Intel-Clarkdale-Core-i5-Desktop-Processor-Debuts/

Fermi (GF100) – GTX480

- Source : http://images.bit-tech.net/content_images/2010/03/nvidia-geforce-gtx-480-1-5gb-review/gf100.jpg



- Source: http://www.tomshardware.com/reviews/geforce-gtx-480,2585-10.html

Source: http://upload.wikimedia.org/wikipedia/commons/thumb/e/ec/Glasses_800_edit.png/800px-Glasses_800_edit.png