Intel Larrabee

Varun Sampath
Penn CIS 565 Spring 2011

Image: http://www.intel.com/pressroom/archive/releases/20100531comp.htm

Agenda

- Goals of the project
- Architecture
- Rendering pipeline
- Performance
- Past, Present, and Future

Goals

- Performance per watt and unit area of highly parallel workloads

	Core 2 Duo	Test CPU design
# CPU cores:	2 out-of-order	10 in-order
Instruction issue:	4 per clock	2 per clock
VPU per core:	4-wide SSE	16-wide
L2 cache size:	4MB	4MB
Single-stream:	**4 per clock**	**2 per clock**
Vector throughput:	**8 per clock**	**160 per clock**

Table: [L. Seiler et al. 2008]

Goals

- More programmable than typical GPUs
  - “completely programmable” rendering pipeline
  - Should “just work” on CPU code

Microarchitecture

- Intel Pentium CPU (1992)
  - Primary and secondary pipelines
  - In-order
    - New CPU designs increase performance 1.5-1.7x, die area 2-3x, and power consumption 2-2.5x
- Extend it
  - 64-bit extensions
  - 4-way SMT
  - 32KB L1 I$ and 32KB L1 D$
  - 256KB slice of L2 (fully coherent)
- New cache control instructions
  - Prefetch into L1 or L2
  - Cache line evict hints

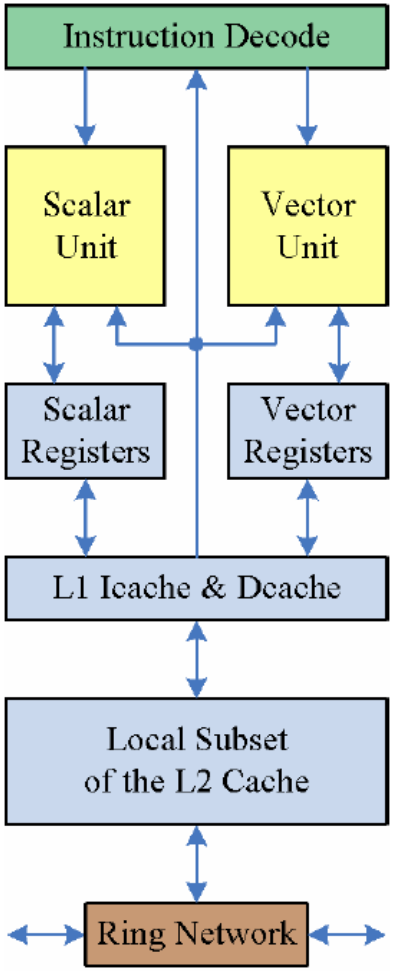


Image: [L. Seiler et al. 2008]

Microarchitecture

- VPU
  - Integer, single and double-precision support
  - 1/3 area of core
  - 512-bit (16-wide)
    - 88% utilization if 16 fragment shaders process one component at a time
  - New vector instructions (LRBni)
  - L1 can be used as a source operand in a vector operation for free
  - Predication
    - 8 16-bit mask registers for vector operations
    - Note: CUDA supports predication too

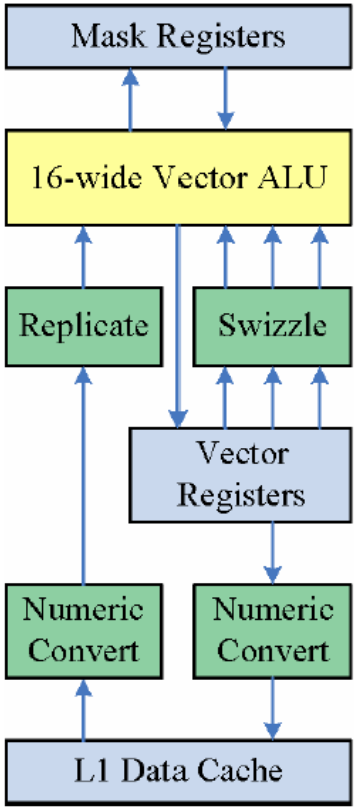
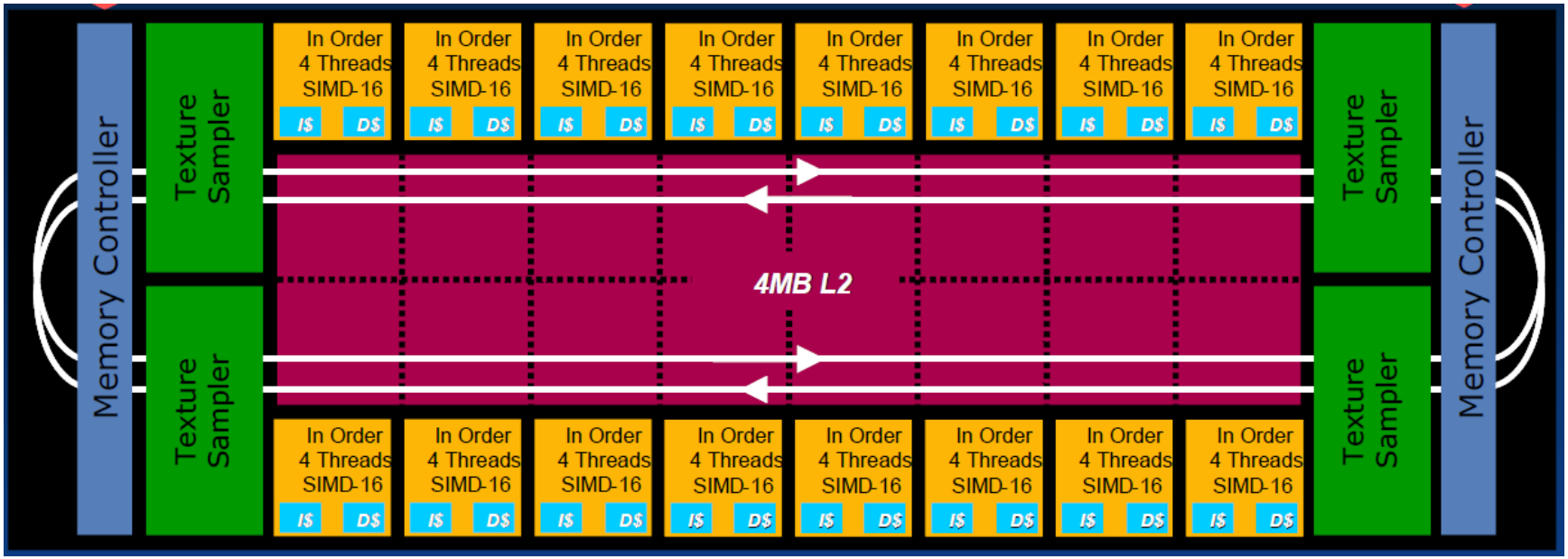


Image: [L. Seiler et al. 2008]

Larrabee



- Ring bus
  - 512-bits wide per direction
- Fixed function texture sampler
  - 12-40x faster than software texture filter

Image: [Forsyth 2010]

G80

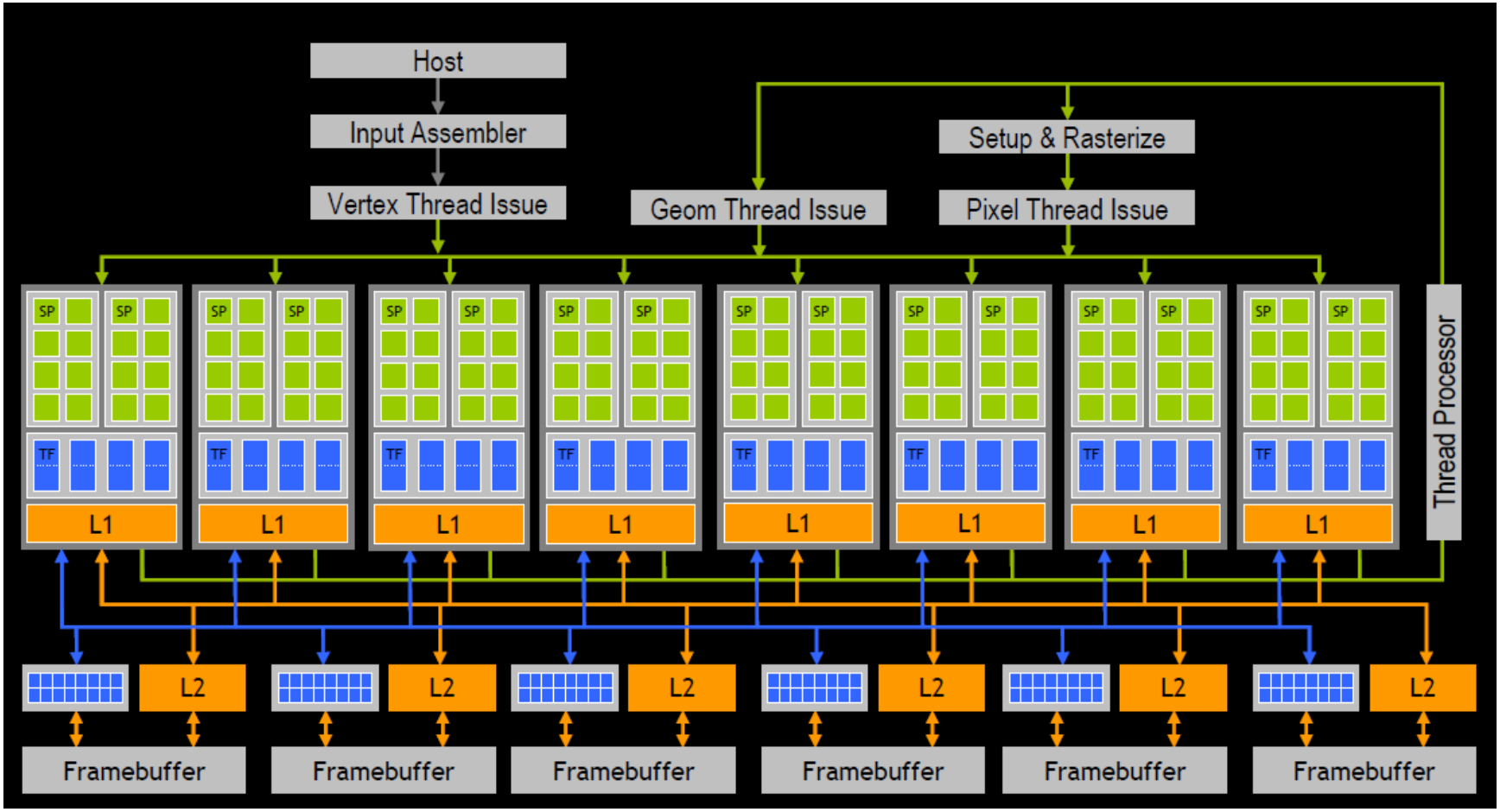


Image: [Luebke 2008 http://s08.idav.ucdavis.edu/luebke-nvidia-gpu-architecture.pdf]

Code

```
;if (d > 0)
;{e+=d*dot(c.xyz,a.xyz+b.xyz)}
vcmpps_gt kT, vD, [ConstZero]{1to16}
kortest kT, kT
jz skip_all_this
vaddps vTempx, vAx, vBx
vaddps vTempy, vAy, vBy
vaddps vTempz, vAz, vBz
vmulps  vT, vTempx, vCx
vmaddps vT, vTempy, vCy
vmaddps vT, vTempz, vCz

vmaddps vE{kT}, vD, vT
skip_all_this:
```

Code: [Forsyth 2010]

Gather and Scatter

- Can read or write to non-contiguous memory in a single vector instruction
  - **`vgather v1{k2},[rax+v3]`**
- Limited by L1 (can typically load one cache line per clock)
- Helpful for Arrays of Structures → Structure of Arrays

Code: [Forsyth 2010]

Software Rendering

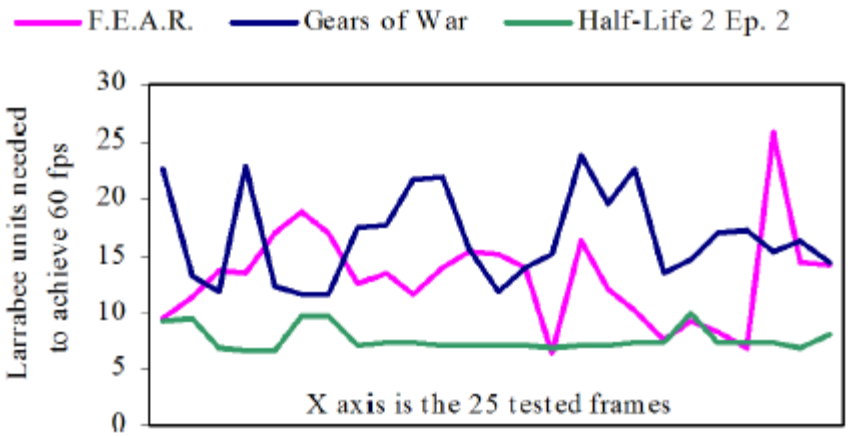
- The only fixed function hardware is the texture sampler
- Why?
  - Don’t need new hardware for every API revision
  - Pipeline can be reconfigured for different applications
  - Resources can be reallocated

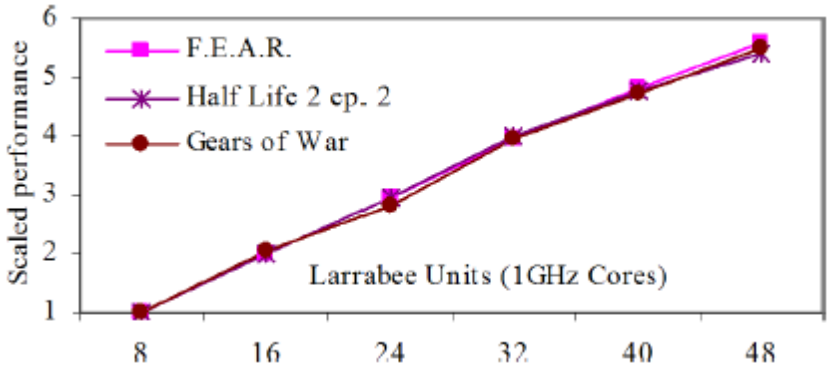
Software Rendering

- “sort-middle” i.e. tiled renderer
  - One core handles geometry processing for set of primitives
  - Resulting triangles are rasterized by that core
  - Triangles that intersect tiles of fragments are binned for that tile
    - Tiles sized to fit in core’s L2 cache subset
  - Each tile assigned to a core for fragment shading
- Chosen primarily to minimize software locks
  - Benefits bandwidth and load balancing too

Graphics Performance

Half Life 2 ep. 2	F.E.A.R.	Gears of War
1600x1200 4 sample	1600x1200 4 sample	1600x1200 1 sample
25 frames (1 in 30)	25 frames (1 in 100)	25 frames (1 in 250)
Valve Corp.	Monolith Productions	Epic Games Inc





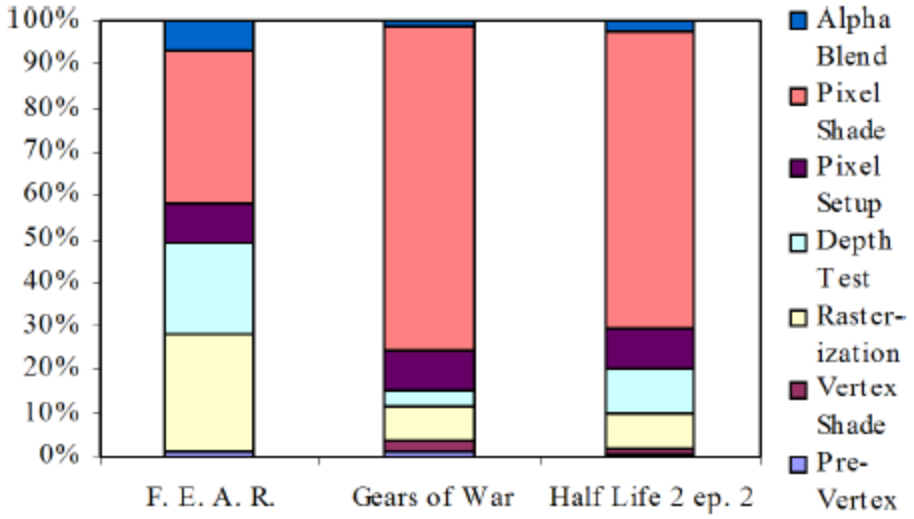


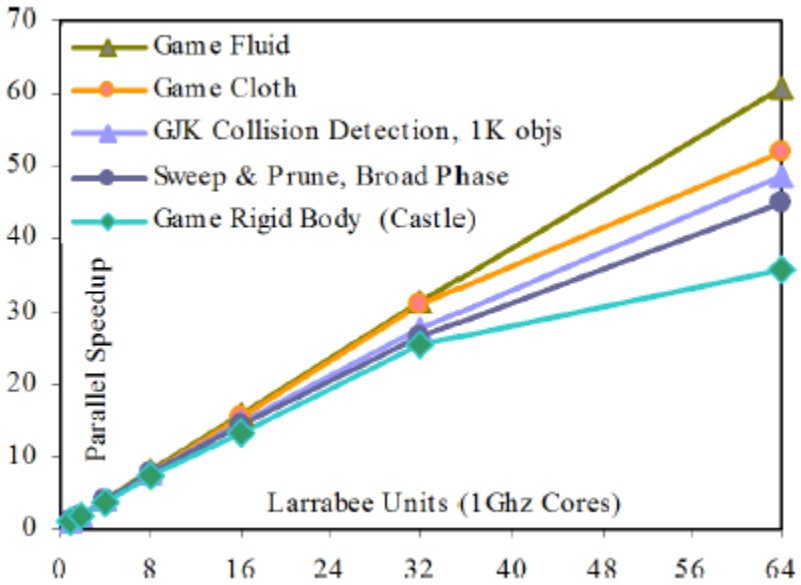
Image: [L. Seiler et al. 2008]

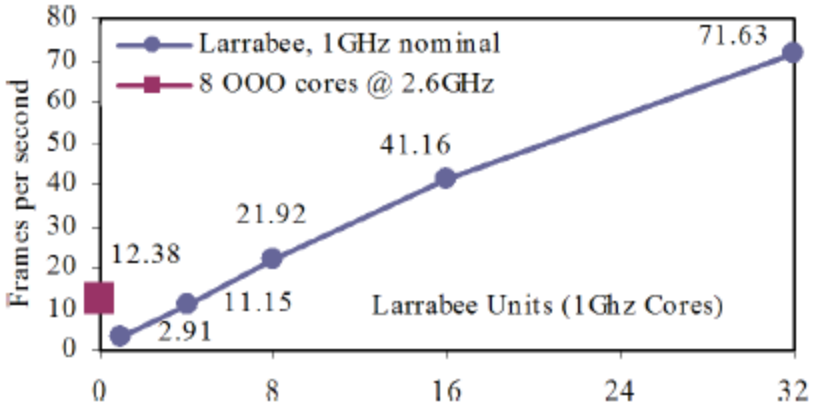
General Programming Model

- Larrabee C/C++ compiler
  - Recompile most applications without modification
  - Auto-vectorization
- P-threads, or alternatively Larrabee Native task scheduling API
- OpenMP
- All of the Intel parallel development tools
- Under the hood
  - Pre-emptive multitasking OS with virtual memory

Performance

- 1 Larrabee Unit = 16 * 2 FP operations (fused multiply-add) * 1GHz = 32 GFLOPS
- 32 Larrabee Units → 1 TFLOP





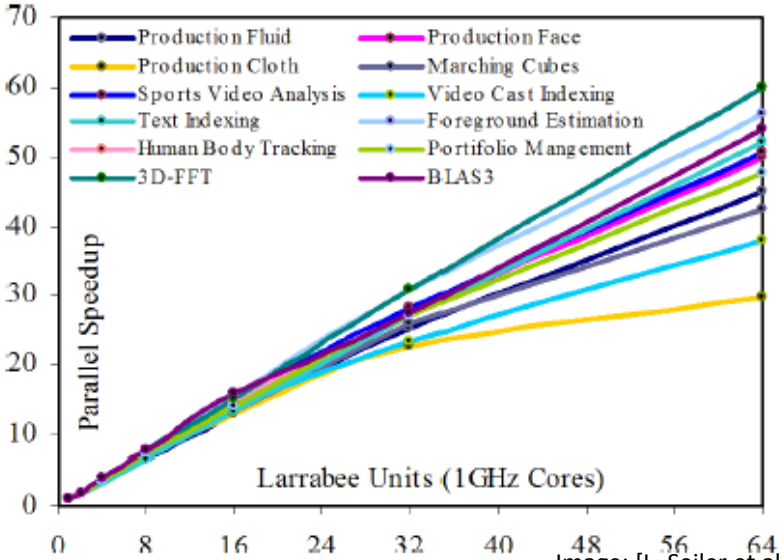
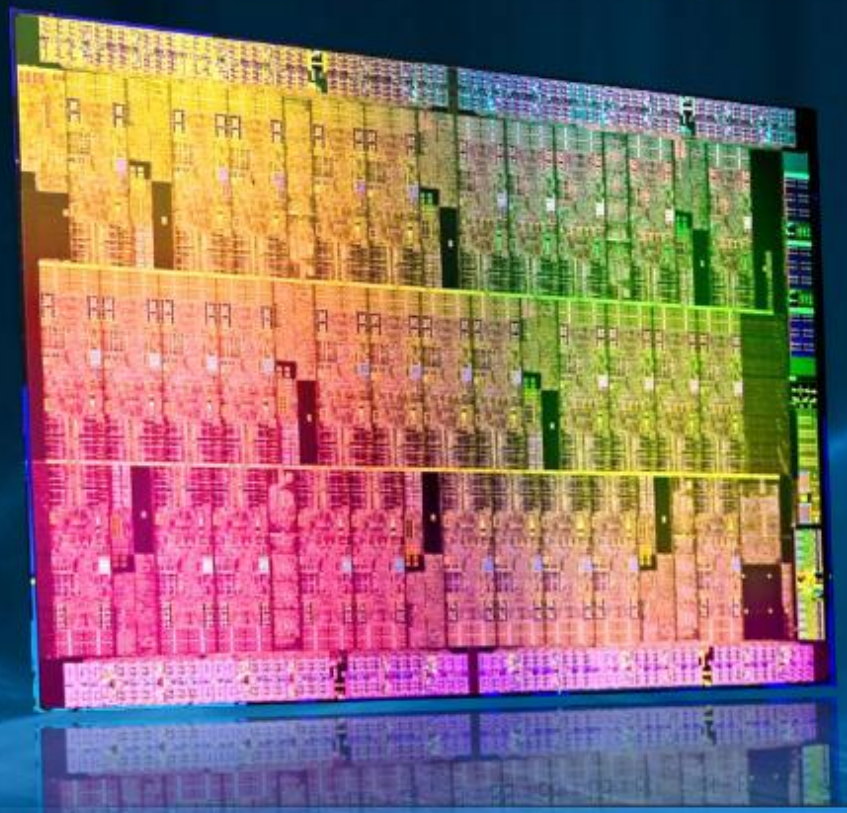


Image: [L. Seiler et al. 2008]

Where is it?

- 8/4/2008: Intel's Larrabee Architecture Disclosure: A Calculated First Move
- 12/4/2009: Intel Cancels Larrabee Retail Products, Larrabee Project Lives On
- 5/25/2010: Intel Kills Larrabee GPU, Will Not Bring a Discrete Graphics Product to Market

From Research to Realization. Announcing...



Intel® Many Integrated Core Architecture

The Newest Addition to the Intel Server Family.
Industry's First General Purpose Many Core Architecture

intel

Slide: [Skaugen 2010]

Knights Ferry

- Software development platform
- Growing availability through 2010
- 32 cores, 1.2 GHz
- 128 threads at 4 threads / core
- 8MB shared coherent cache
- 1-2GB GDDR5
- Bundled with Intel HPC tools

Software development platform for Intel® MIC architecture

intel

Slide: [Skaugen 2010]

References

- [L. Seiler et al. 2008] Larry Seiler, Doug Carmean, Eric Sprangle, Tom Forsyth, Michael Abrash, Pradeep Dubey, Stephen Junkins, Adam Lake, Jeremy Sugerman, Robert Cavin, Roger Espasa, Ed Grochowski, Toni Juan, and Pat Hanrahan. 2008. Larrabee: a many-core x86 architecture for visual computing. *ACM Trans. Graph.* 27, 3, Article 18 (August 2008), 15 pages. DOI=10.1145/1360612.1360617 http://doi.acm.org/10.1145/1360612.1360617

- [Forsyth 2010] Forsyth, T. (2010, January 6). *The Challenges of Larrabee as a GPU.* Retrieved April 4, 2011, from Stanford EE Computer Systems Colloquium: http://www.stanford.edu/class/ee380/Abstracts/100106.html

- [Skaugen 2010] Skaugen, K. (2010, May 31). *Petascale to Exascale: Extending Intel's HPC Committment.* Retrieved April 4, 2011, from Intel: http://download.intel.com/pressroom/archive/reference/ISC_2010_Skaugen_keynote.pdf