Judith Providence

Computer Architecture

CS 654

Outline

- Background/Motivation
- Multi-processors
- Larrabee Architecture
- Performance studies
- Evaluation
- Conclusion

Motivation:Trends Towards Many-core Processors

- Power
- Growth in HPC
- Decrease performance in uniprocessors
    - Limits on Instruction-Level Parallelism
        - Register renaming
        - Branch prediction
        - Jump prediction
        - Memory address Alias Analysis
        - Perfect caches

Larrabee:GPU or CPU?

- GPU
- PCI bus
- Only a minimum amount of memory available
- Only single-precision floating point performance

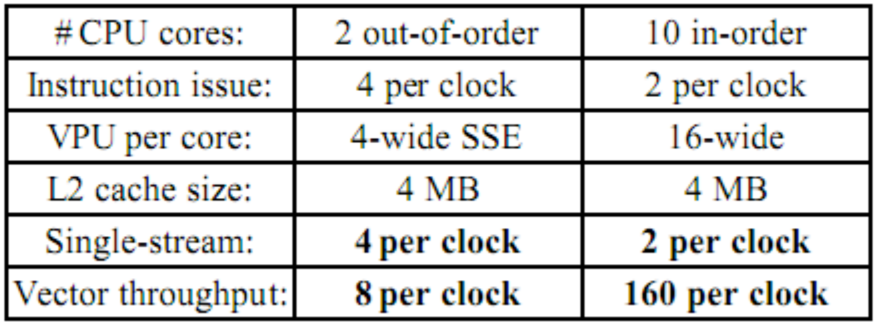
- Larrabee CPU
- It supports 4 threads
- Efficient inter-block communication
  - Ring network for full inter-processor communication
- Each Larrabee core is a complete x86 core that supports
  - Virtual memory and page swapping
  - Fully coherent caches at all levels

Larrabee:CPU

- Larrabee a in-order many-core x86 CPU
- Intel president in 2005 stated: We are dedicating all of our future product development to multi-core designs.
- Multi-core processors vs. many-core processors
- GPU-like capabilities

Motivation for an in-order CPU

- Comparison between a modern out-of-order CPU, the Intel Core2Duo processor, and a in-order test CPU design based on the Pentium processor with a 16-wide VPUs



# 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:	4 MB	4 MB
Single-stream:	**4 per clock**	**2 per clock**
Vector throughput:	**8 per clock**	**160 per clock**

Multi-processors

- Inter-processor Communication

  Inter-processor Ring Network

- Computation

  SIMD vector processing unit, mask register

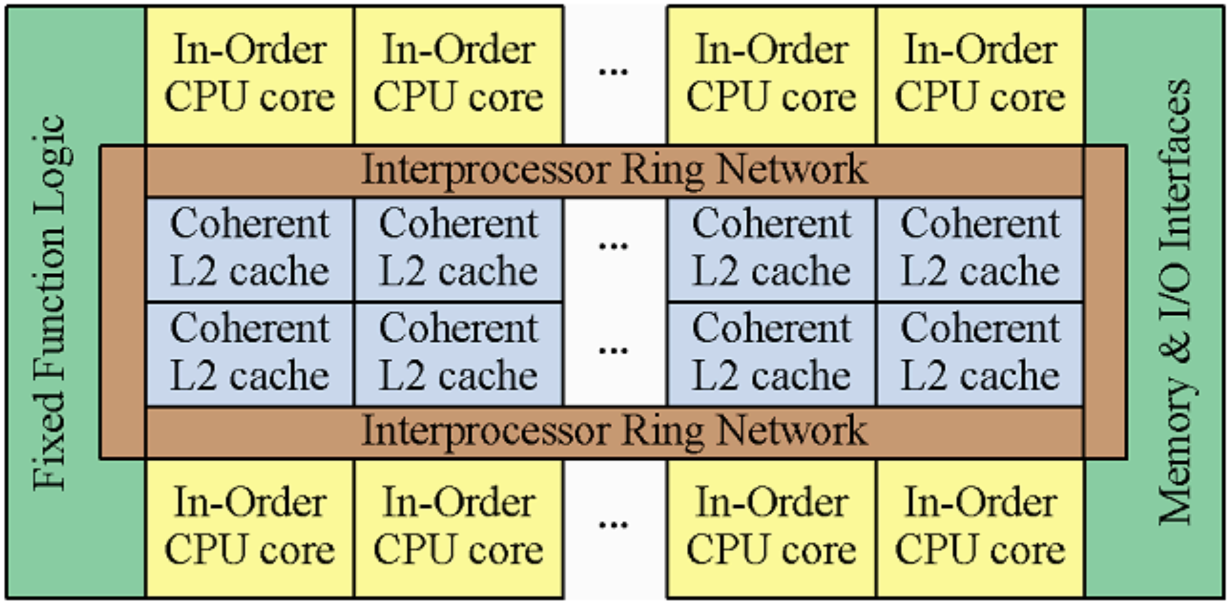
- Shared Memory

  Coherent cached memory hierarchy, MIMD Model

- Synchronization Mechanisms

  Semaphores, Software locks

Larrabee Architecture



Fixed Function Logic

In-Order CPU core | In-Order CPU core | … | In-Order CPU core | In-Order CPU core

Interprocessor Ring Network

Coherent L2 cache | Coherent L2 cache | … | Coherent L2 cache | Coherent L2 cache

Coherent L2 cache | Coherent L2 cache | … | Coherent L2 cache | Coherent L2 cache

Interprocessor Ring Network

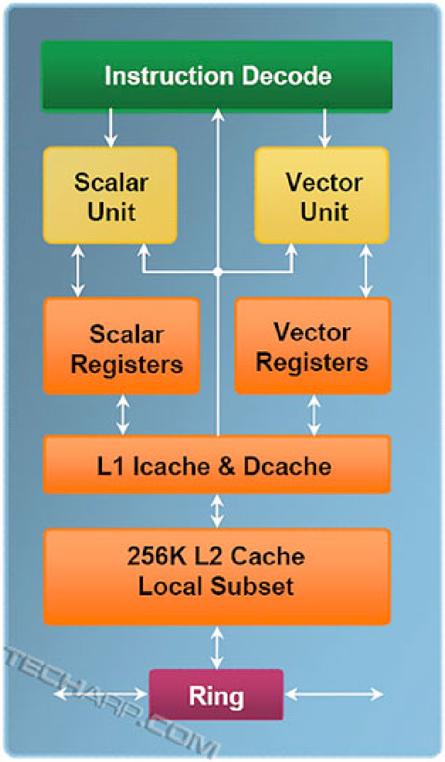
In-Order CPU core | In-Order CPU core | … | In-Order CPU core | In-Order CPU core

Memory & I/O Interfaces

***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.*

Core Design of Larrabee

Larrabee CPU core and associated system blocks: the CPU is derived from the Pentium processor in-order design, plus 64-bit instructions, multi-threading and a wide VPU. Each core has fast access to its 256KB local subset of a coherent 2nd level cache. L1 cache sizes are 32KB for Icache and 32KB for Dcache. Ring network accesses pass through the L2 cache for coherency.

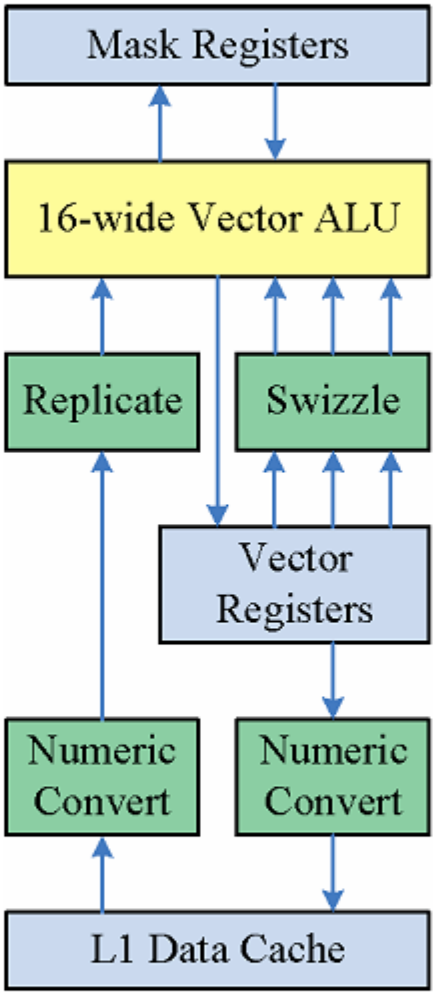


Inter-processor Ring Network

- Bi-directional
- Routing decisions made before messages are placed into the network
- Checks for data sharing
- Provides a path for the L2 cache to access memory
- Allows Fixed Function Logic agents to be accessed by the CPU cores
- Scaling to more than 16 cores

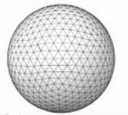
Wide Vector Processing Unit

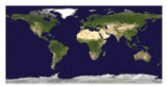
- SIMD
- 16 lanes
- Executes integer and Floating point instructions
- Scatter gather supports a Maximum of 16 elements



Fixed Function Logic Unit

- Used for Graphical tasks
- Larrabee uses software in place of a fixed functional unit for some graphical tasks
- Cores pass commands to the texture unit through the L2 cache
- Texture filter logic
  - would be 12x to 40x longer in software


Sphere with no texture


Texture image

Sphere with texture

Advanced Applications

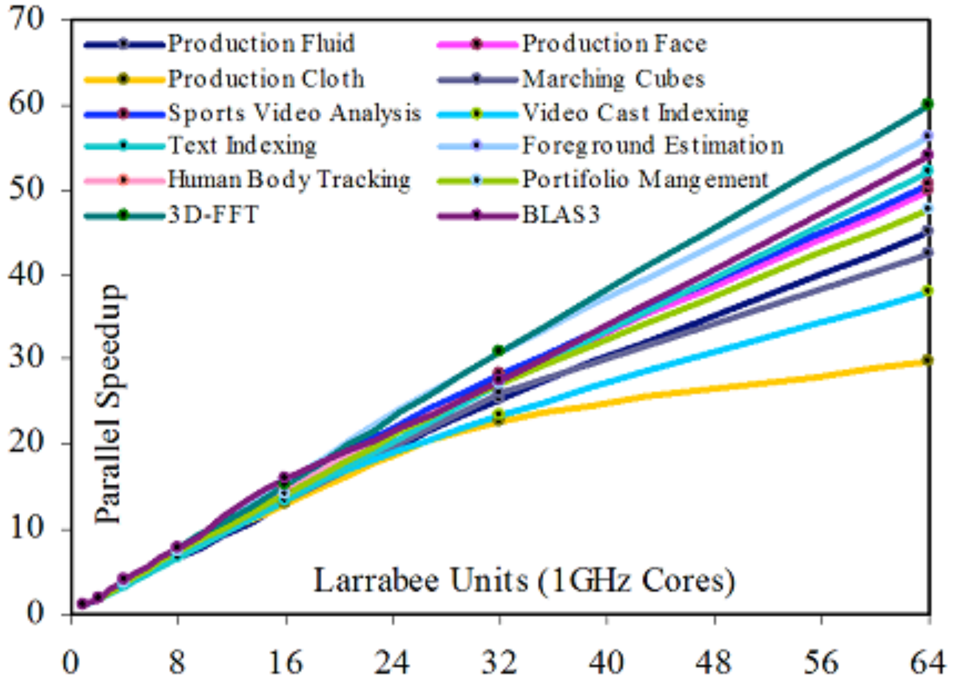
- Larrabee supports irregular data structures
- An efficient scatter-gather support for irregular data structures
- The SIMD vector processing unit can be programmed
- Intel’s auto-vectorization computer technology

Performance Study

- Spectral methods/Dense Linear algebra
- Data is in the frequency domain
- High Performance Kernel-3D-FFT
- Data that are dense matrices or vectors -BLAS-3

High Performance Computing Kernels

- Simulation results are based on Stanford’s PhysBam
- http://physbam.stanford.edu/~fedkiw
- Amdahl’s Law:Speedup $_{\text{maximum}}$ =1/(1-fraction enhanced)



Evaluation of Larrabee for parallel applications

con

- Memory contention
- Lack of error correcting code(ECC) memory, Graphic double data rate
- Shortage of double precision floating point capability

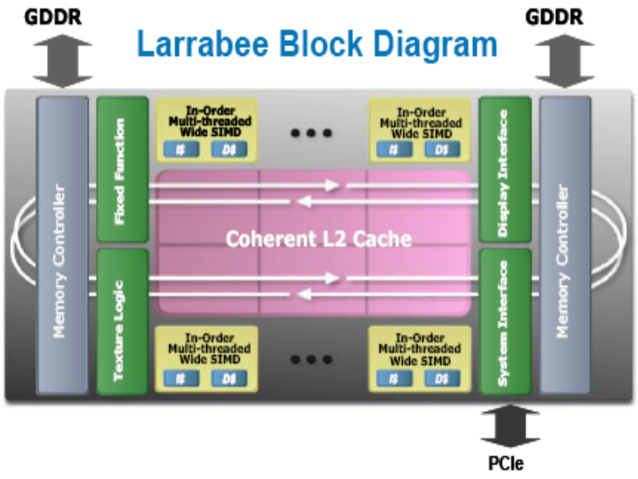


Figure 6: Schematic of the Larrabee many-core architecture

pro

- Load balancing is accomplished by moving processes
- Supports irregular data structures

Conclusion-Relevance of Larrabee for the Future

- Amdahl’s Law - Limitations in parallelism make it difficult to achieve good speedup
- 1965 - Moore’s Law states that the number of transistors on a chip will double about every two years
- Need a Moore’s Law to handle software
- Solution: the establishment of academic communities