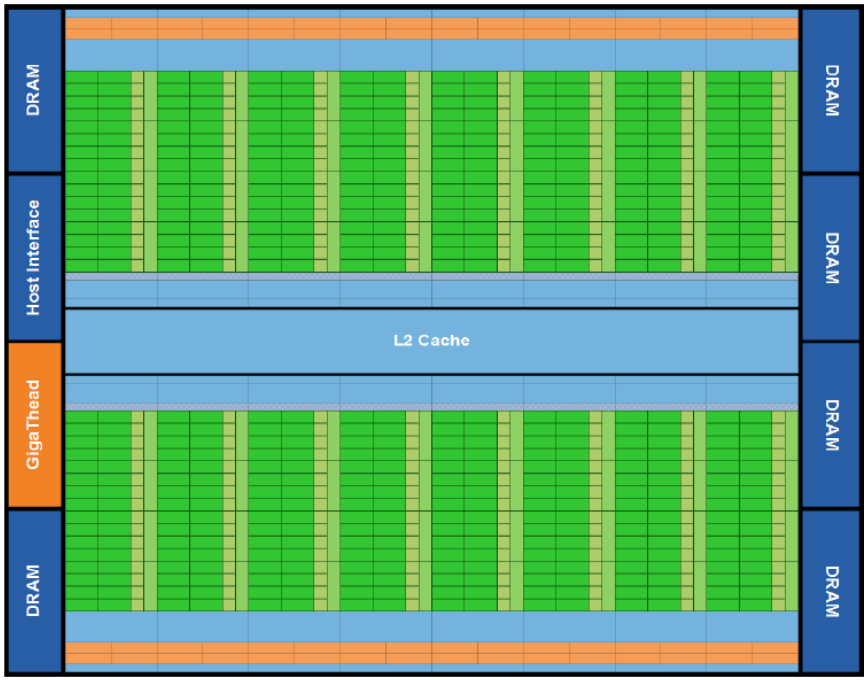
Co-Processor Architectures
Fermi vs. Knights Ferry

Roger Goff
Dell Senior Global CERN/LHC Technologist
+1.970.672.1252 | Roger_Goff@dell.com

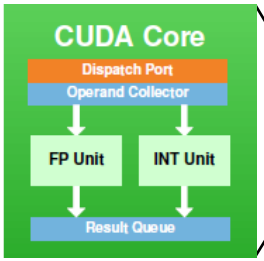
nVidia Fermi Architecture

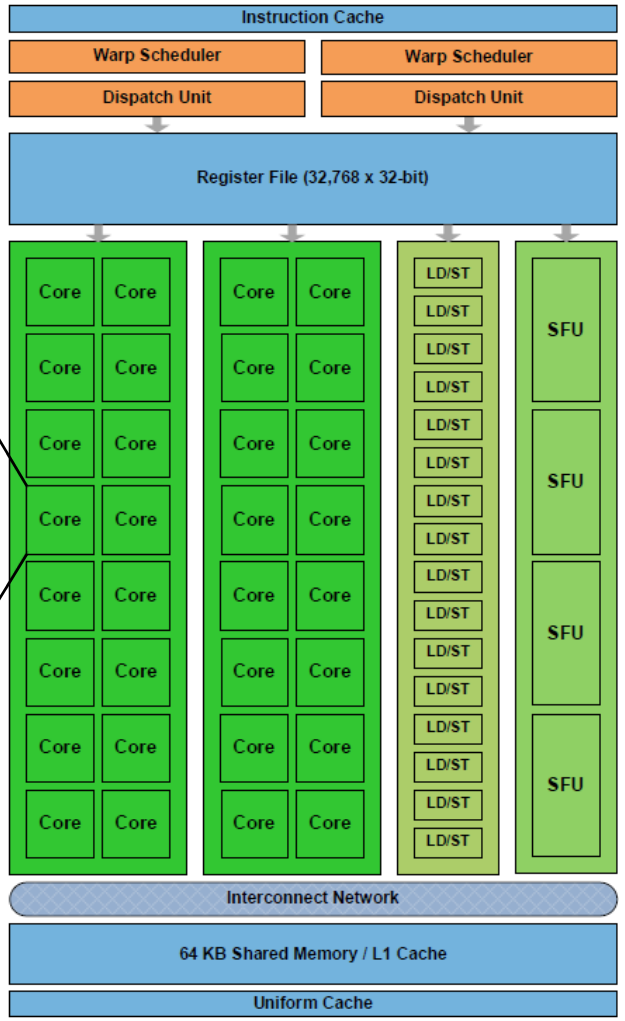
- Up to 512 cores
  - 16 Streaming multiprocessors each with 32 cores @ 1.3GHz
- Parallel DataCache
  - 64 KB Shmem/L1 Cache
  - 768 KB Unified L2 Cache
- Six 64-bit memory partitions
  - 384-bit memory interface
  - Up to 6 GB GDDR5 DRAM
- Up to 16 concurrent kernels
- IEEE floating point math
- ECC memory



Fermi Streaming Multiprocssor Architecture

- 32 Cores
  - 32-bit Integer ALU with 64-bit extensions
  - Full IEEE 754-2008 32-bit and 64-bit precision
- 64 KB Shared Memory/L1 cache
  - 16KB Shmem/48KB cache or 48KB Shmem/16KB L1 cache
- 16 load/store units
- Dual Warp scheduler (dual instruction issue)
- Four Special Function Units (SFUs) for sin, cosine, reciprocal, and square root operations





Comparison to Previous nVidia GPGPUs

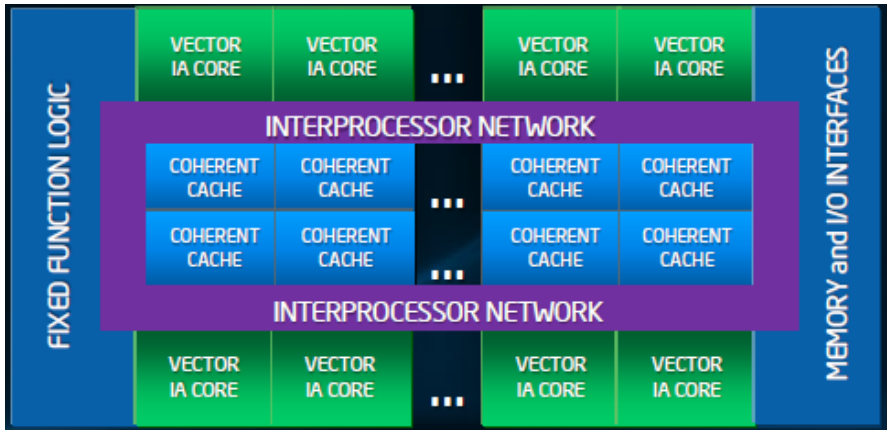
GPU	G80	GT200	Fermi
Transistors	681 million	1.4 billion	3.0 billion
CUDA Cores	128	240	512
Double Precision Floating Point Capability	None	30 FMA ops / clock	256 FMA ops /clock
Single Precision Floating Point Capability	128 MAD ops/clock	240 MAD ops / clock	512 FMA ops /clock
Special Function Units (SFUs) / SM	2	2	4
Warp schedulers (per SM)	1	1	2
Shared Memory (per SM)	16 KB	16 KB	Configurable 48 KB or 16 KB
L1 Cache (per SM)	None	None	Configurable 16 KB or 48 KB
L2 Cache	None	None	768 KB
ECC Memory Support	No	No	Yes
Concurrent Kernels	No	No	Up to 16
Load/Store Address Width	32-bit	32-bit	64-bit

Intel MIC Architecture

Pronounced “Mike”

Many cores with many threads per core

Standard IA programming and memory model



Knights Ferry

- Software development platform
- 1-2GB GDDR5 connected to host memory through PCI DMA operations with virtual addressing
- Intel HPC developer tools

32 Cores @ 1.2 GHz

- ✓ 4 threads/core, 128 total parallel threads
- ✓ 32KB i-cache, 32KB d-cache
- ✓ 256KB coherent L2 cache (8MB total)
- ✓ 512bit vector unit
  - 16 Single precision FLOPs/clock
  - 8 Double precision FLOPS/clock

MIC Programming Environment

- Inherently supports OpenMP.
- Virtual memory environment extends back to host memory.
- Intel Parallel Studio and Cluster Studio support MIC.
- Optimizing performance will take almost as much effort as for CUDA and OpenCL environments.

Knights Corner
1st Production MIC Co-processor

- Second Half 2012
  - Knowns:
    - 50+ cores
    - 22nm manufacturing process
  - Unknowns:
    - Core frequency
    - Size of GDDR5 memory on board
    - ECC support

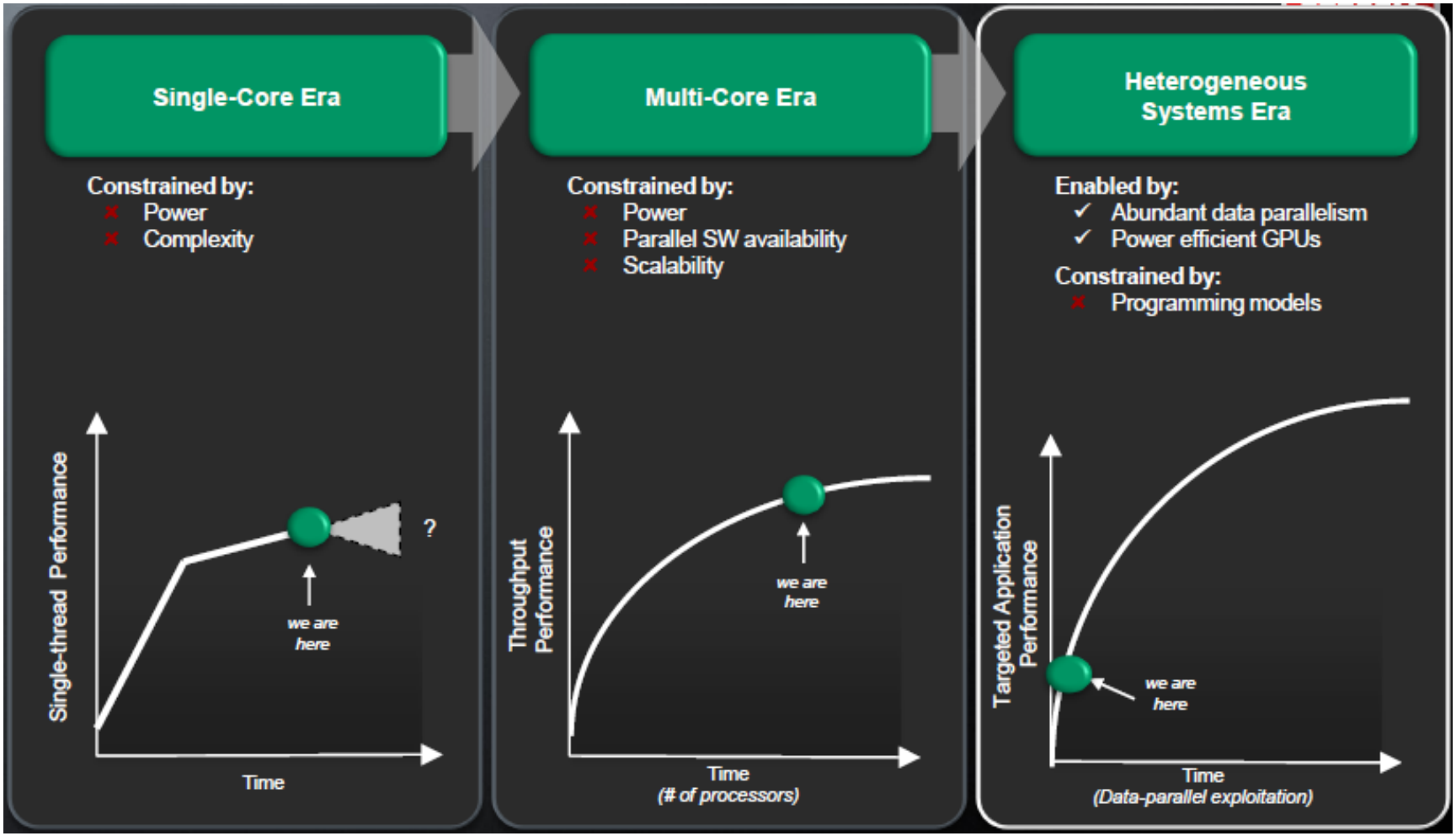
Co-processor Comparison

	AMD Firestream	NVIDIA Fermi	Intel Knights Ferry	Intel Knights Corner Speculation	Intel Knights Corner Speculation2
Cores	1600	512	32*4 threads/core = 128	50*4 threads/core = 200	64*4 threads/core = 256
Core Frequency	700/825 MHz	1.3 GHz	1.2 GHz	1.2 GHz	2 GHz
Thread Granularity	fine	fine	coarse	coarse	coarse
Single Precision Floating Point Capability GFLOPs	2000/2640	1024	614	960	2048
Double Precision Floating Point Capability GFLOPs	400/528	512	307	480	1024
GDDR5 RAM	2/4 GB	3-6 GB	1-2 GB	?	?
L1 cache/processor		64KB (16KB Shmem, 48KB L1 or 48KB Shmem, 16KB L1)	64KB (32KB icache, 32KB dcache)	64KB (32KB icache, 32KB dcache)	64KB (32KB icache, 32KB dcache)
L2 cache/processor		768KB shared L2	8MB coherent total (256KB/core)	12MB coherent total (256KB/core)	16MB coherent total (256KB/core)
programming model		CUDA kernels	posix threads	posix threads	posix threads
virtual memory		no	yes	yes	yes
memory shared with host		no	no	no	no
Software	OpenCL, DirectCompute	C, C++, CUDA, OpenCL, DirectCompute	C, C++, FORTRAN, OpenMP, CUDA, OpenCL, DirectCompute	C, C++, FORTRAN, OpenMP, CUDA, OpenCL, DirectCompute	C, C++, FORTRAN, OpenMP, CUDA, OpenCL, DirectCompute

Co-processor Adoption

- Commercial adoption:
  - Oil & Gas/seismic data processing
  - Financial services
  - Ray tracing
  - Molecular dynamics
  - Commercial applications: MATLAB, ANSYS
- Barriers to adoption
  - Lack of parallel programming skills
  - Immature software development environment & standards
    - CUDA vs. OpenCL vs. OpenMP
    - Waiting for the compiler or libraries to abstract the accelerator
  - Uncertainty of benefit vs. effort
    - Amdahl's law is still the law! Maximum Speedup = $\frac{1}{(1-P)+\frac{P}{N}}$.
    - Huge investment in current codes

AMD "New Era of Processor Performance"



Final Thoughts

1. Co-processors are here to stay, but their architectures will continue to evolve.
2. Programing tools will get easier to use and will further integrate co-processing technology.
3. Further abstraction of the underlying co-processor hardware is necessary to achieve broad adoption.
4. Processors from Intel and AMD will integrate co-processors before the end of the decade.
5. Preparing applications for extreme parallelism will enable users to get the most out of future systems.

Thank you!

Roger Goff
Dell Senior Global CERN/LHC Technologist
+1.970.672.1252 | Roger_Goff@dell.com