MSc Informatics Eng.

2013/14

A.J.Proença

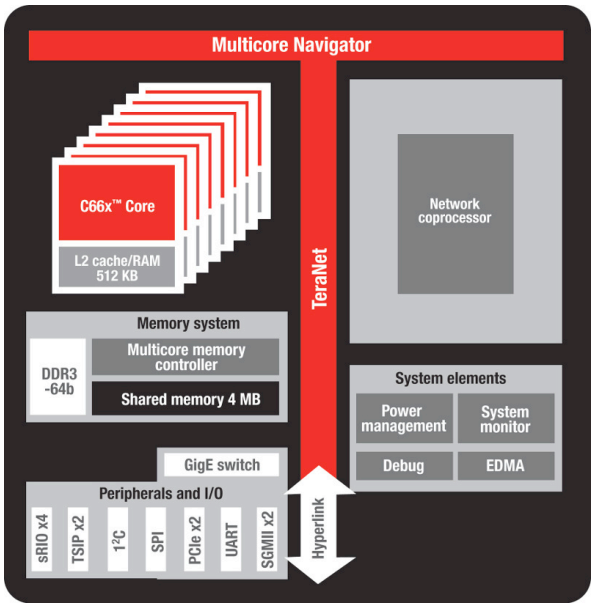
Data Parallelism 2 (*Cell BE, FPGA, MIC, GPU, ...*)

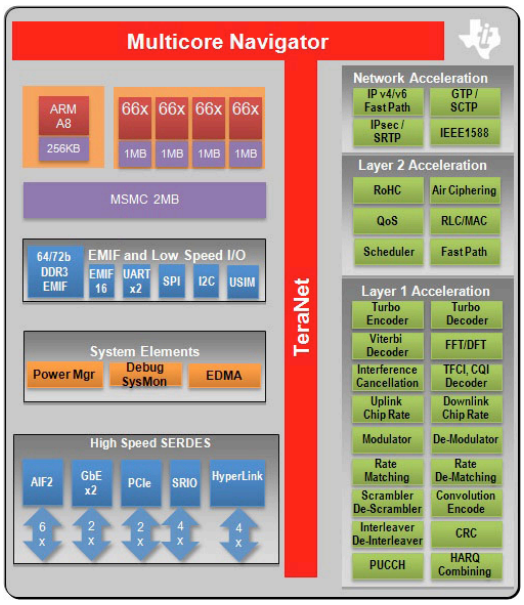
(most slides are borrowed)

Beyond Vector/SIMD architectures

- Vector/SIMD-extended architectures are hybrid approaches
  - mix (super)**scalar + vector** op capabilities on a single device
  - **highly pipelined** approach to reduce memory access penalty
  - **tightly-closed access to shared memory**: lower latency
- Evolution of Vector/SIMD-extended architectures
  - **CPU cores with wider vectors and/or SIMD cores**:
    - DSP VLIW cores with vector capabilities: **Texas Instruments** (...?)
    - PPC cores coupled with SIMD cores: **Cell Broadband Engine** (past...)
    - ARM64 cores coupled with SIMD cores: project Denver/BSC (**NVidia**) (...?)
    - x86 many-cores: **Intel** MIC / Xeon Phi / Knights C/L**, AMD** FirePro**...**
  - **devices with no scalar processor: accelerator devices**
    - CPU-cores + accel devices (disjoint physical memories) => **PCI-E**xpress
    - focus on SIMT/SIMD to hide memory latency: **GPU**-type architecture
    - ISA-free architectures, code compiled to silica: **FPGA**

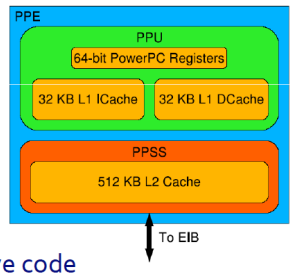
Texas Instruments: Keystone DSP architecture





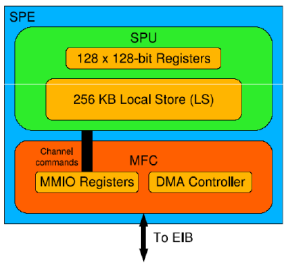
Cell Broadband Engine *(PPE)*

- Heterogeneous multicore processor
  - 1 x Power Processor Element (PPE)
    - 64-bit Power-architecture-compliant processor
    - Dual-issue, in-order execution, 2-way SMT processor
    - PowerPC Processor Unit (PPU)
      - 32 KB L1 IC, 32 KB L1 DC, VMX unit
    - PowerPC Processor Storage Subsystem (PPSS)
      - 512 KB L2 Cache
    - General-purpose processor to run OS and control-intensive code
    - Coordinates the tasks performed by the remaining cores



Cell Broadband Engine *(SPE)*

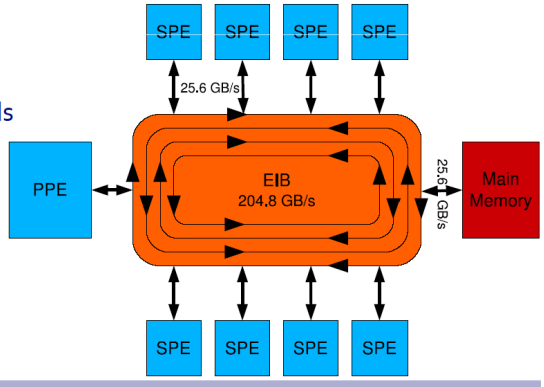
- Heterogeneous multicore processor
  - 8 x Synergistic Processing Element (SPE)
    - Dual-issue, in-order execution, 128-bit SIMD processors
    - Synergistic Processor Unit (SPU)
      - SIMD ISA (four different granularities)
      - 128 x 128-bit SIMD register file
      - **256 KB Local Storage (LS) for code/data**
    - Memory Flow Controller (MFC)
      - Memory-mapped I/O registers (MMIO Registers)
      - DMA Controller: commands to transfer data in and out
    - Custom processors specifically designed for data-intensive code
    - Provide the main computing power of the Cell BE



Cell Broadband Engine *(EIB)*

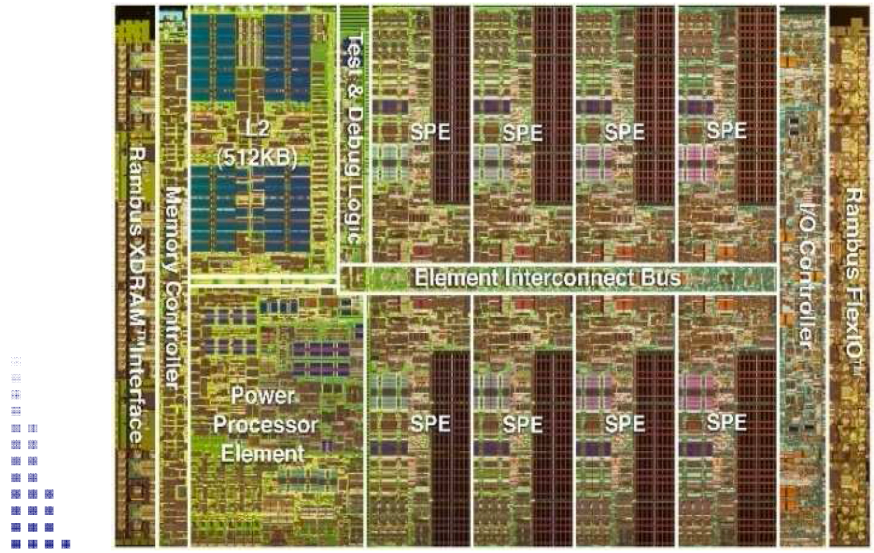
- Element Interconnect Bus (EIB)
  - Interconnects PPE, SPEs, and the memory and I/O interface controllers
    - 4 x 16 Byte-wide rings (2 clockwise and 2 counterclockwise)
  - Up to three simultaneous data transfers per ring
  - Shortest path algorithm for transfers
- Memory Interface Controller (MIC)
  - 2 x Rambus XDR I/O memory channels

    (accesses on each channel of 1-8, 16, 32, 64 or 128 Bytes)
- Cell BE Interface (BEI)
  - 2 x Rambus FlexIO I/O channels



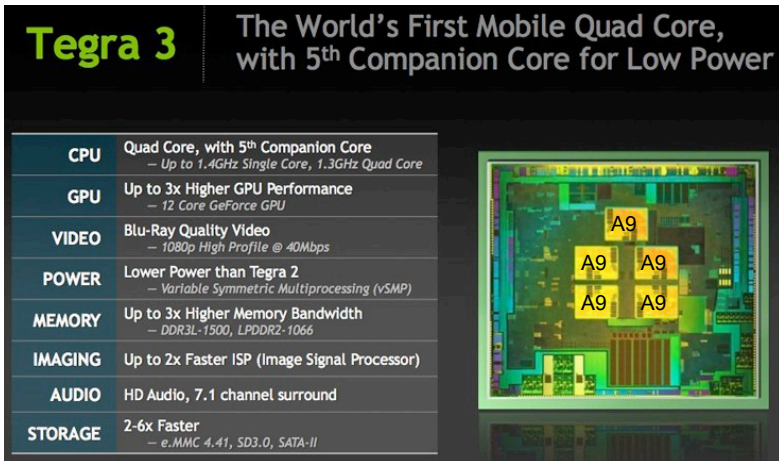
Cell Broadband Engine (chip)

Architecture

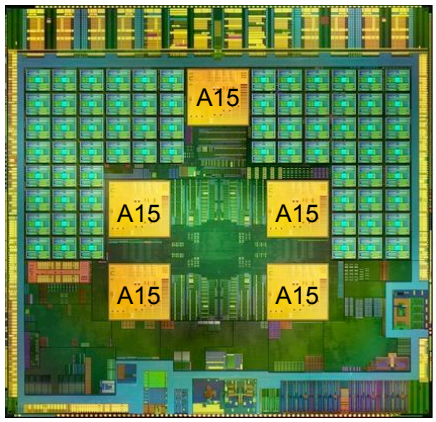


NVidia: pathway towards ARM-64 (1)

- Pick a successful line: Tegra 3, 4, 5, 6 ...



- Replace the 32-bit ARM Cortex A9 by Cortex A15, and add 72 GPUs



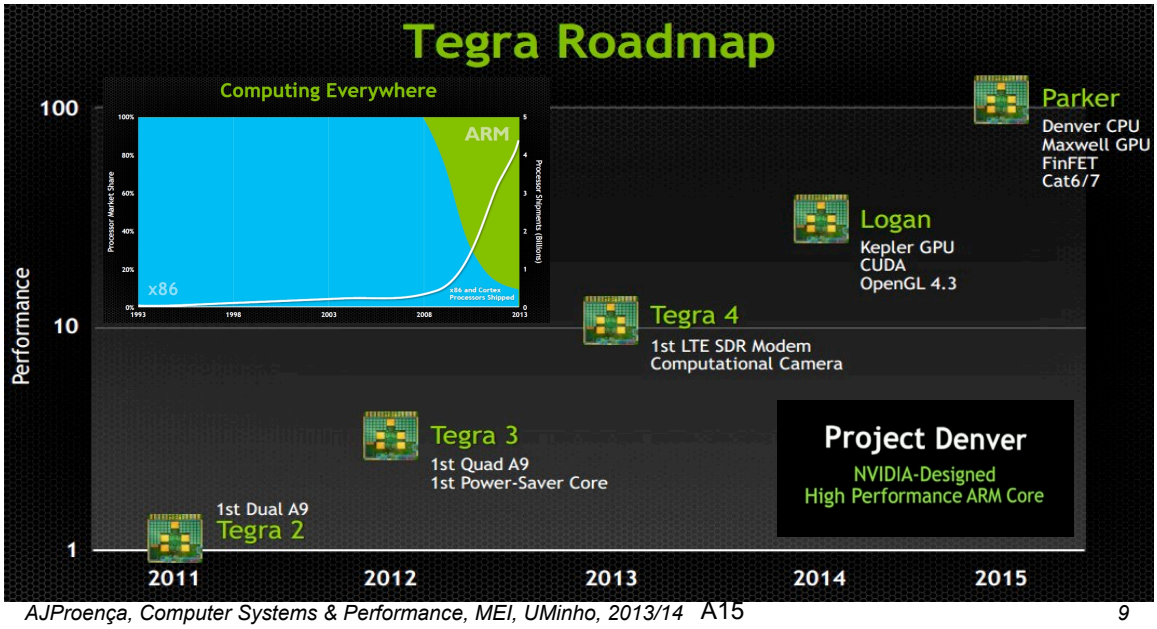
Tegra 4

- Replace the GPUs by Kepler SMX and give support to CUDA 5

  => **Tegra 5**

NVidia: pathway towards ARM-64 (2)

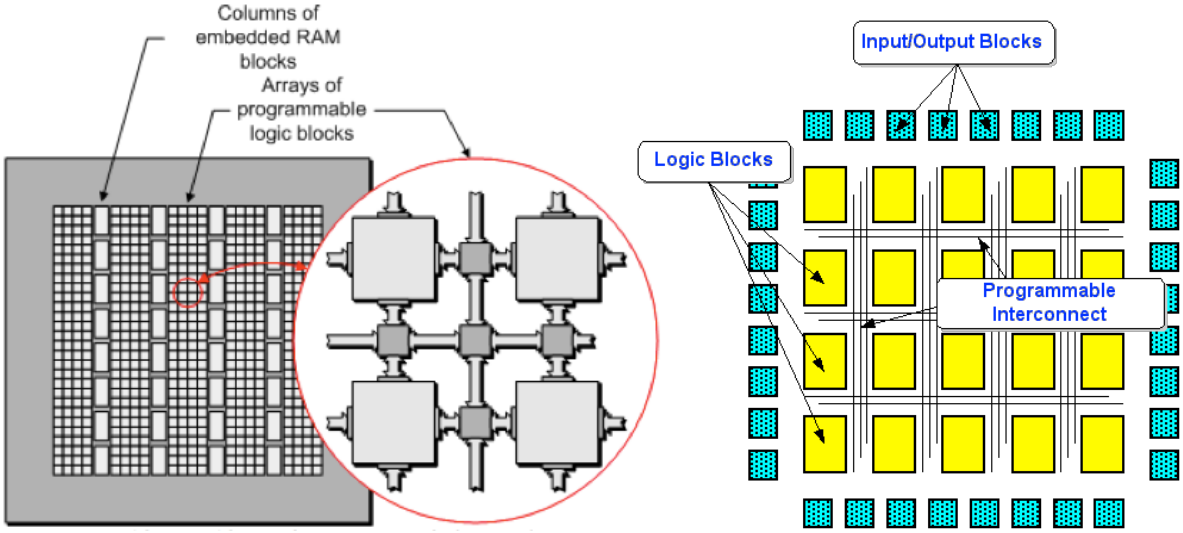
- Replace the 32-bit ARM by a novel 64-bit ARM (*Denver*) and the Kepler SMX by the Maxwell SMX => **Tegra 6**



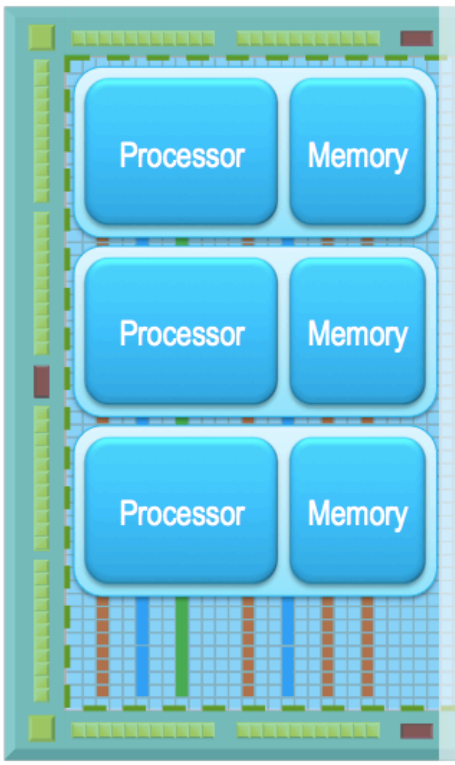
What is an FPGA

Field-Programmable Gate Arrays (FPGA)

A fabric with 1000s of simple configurable logic cells with LUTs, on-chip SRAM, configurable routing and I/O cells



FPGA as a multiple configurable ISA



- Many coarse-grained processors
  - Different Implementation Options
    - Small soft scalar processor
    - or Larger vector processor
    - or Customized hardware pipeline
  - Each with local memory
- Each processor can exploit the fine grained parallelism of the FPGA to more efficiently implement it's "program"
- Possibly heterogeneous
  - Optimized for different tasks
- Customizable to suit the needs of a particular application

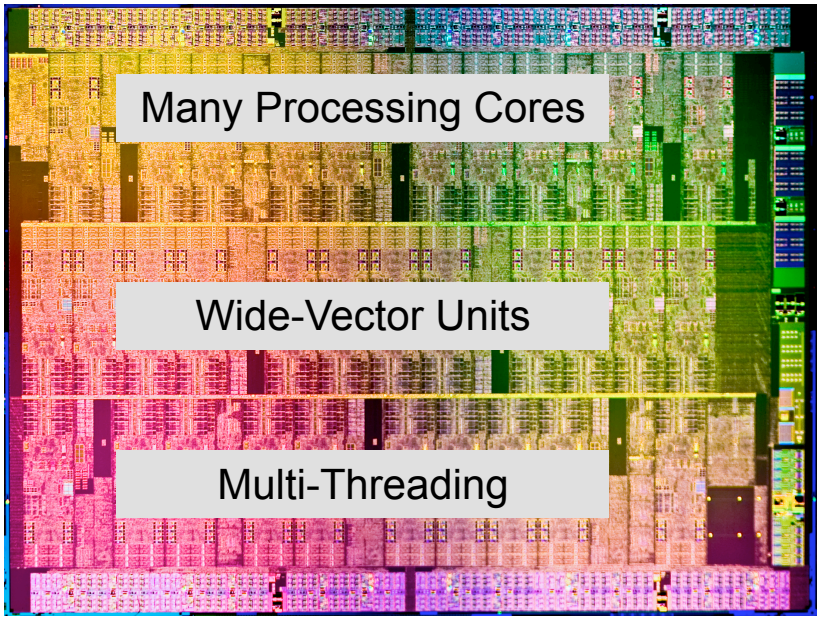
Intel MIC: Many Integrated Core

From:

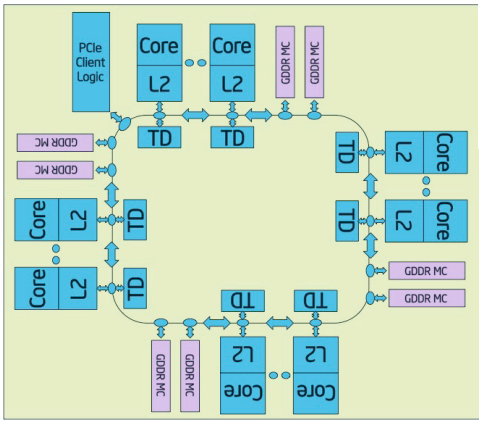
- Larrabee
  (80-core GPU)
- SCC
  (Single-chip Cloud Comp
  24x dual-core tiles)

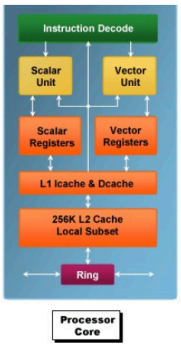
to MIC:

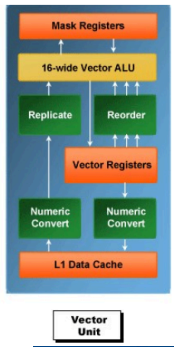
- Knights Ferry
  (pre-production)
- Knights Corner
  (Xeon Phi co-processors
  up to 61 cores)
- Knights Landing
  (2nd generation)

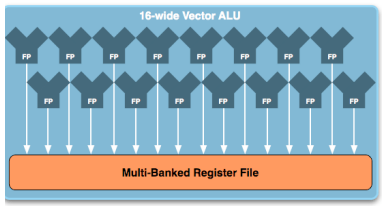


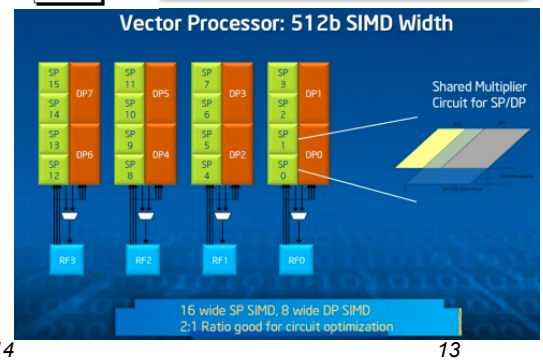
Intel Knights Corner architecture



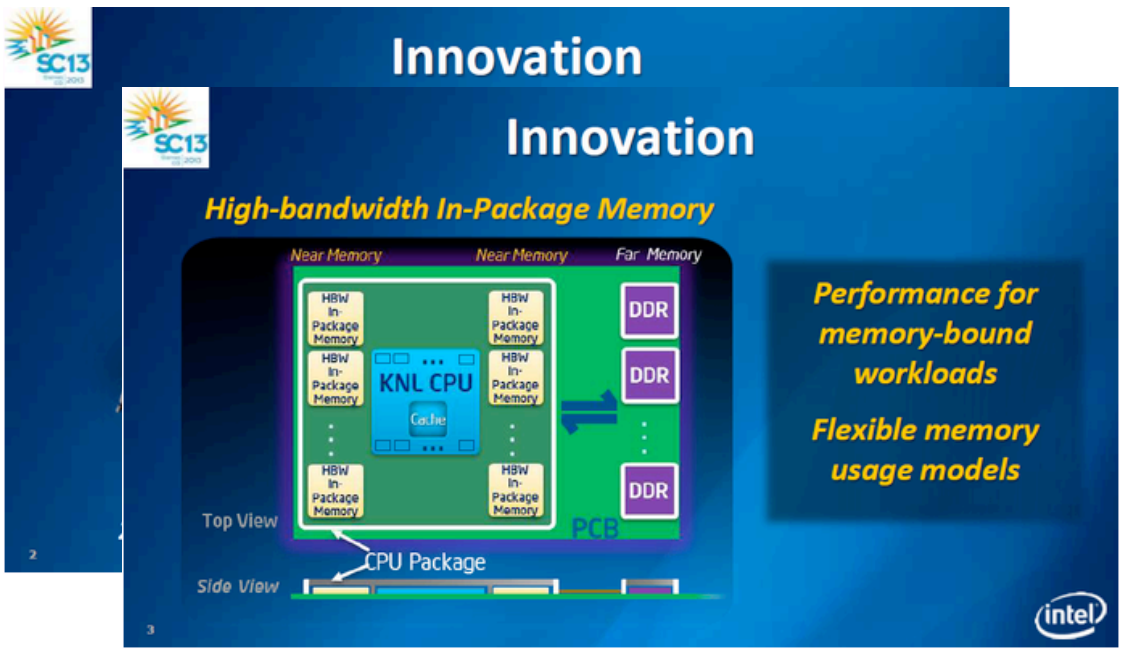




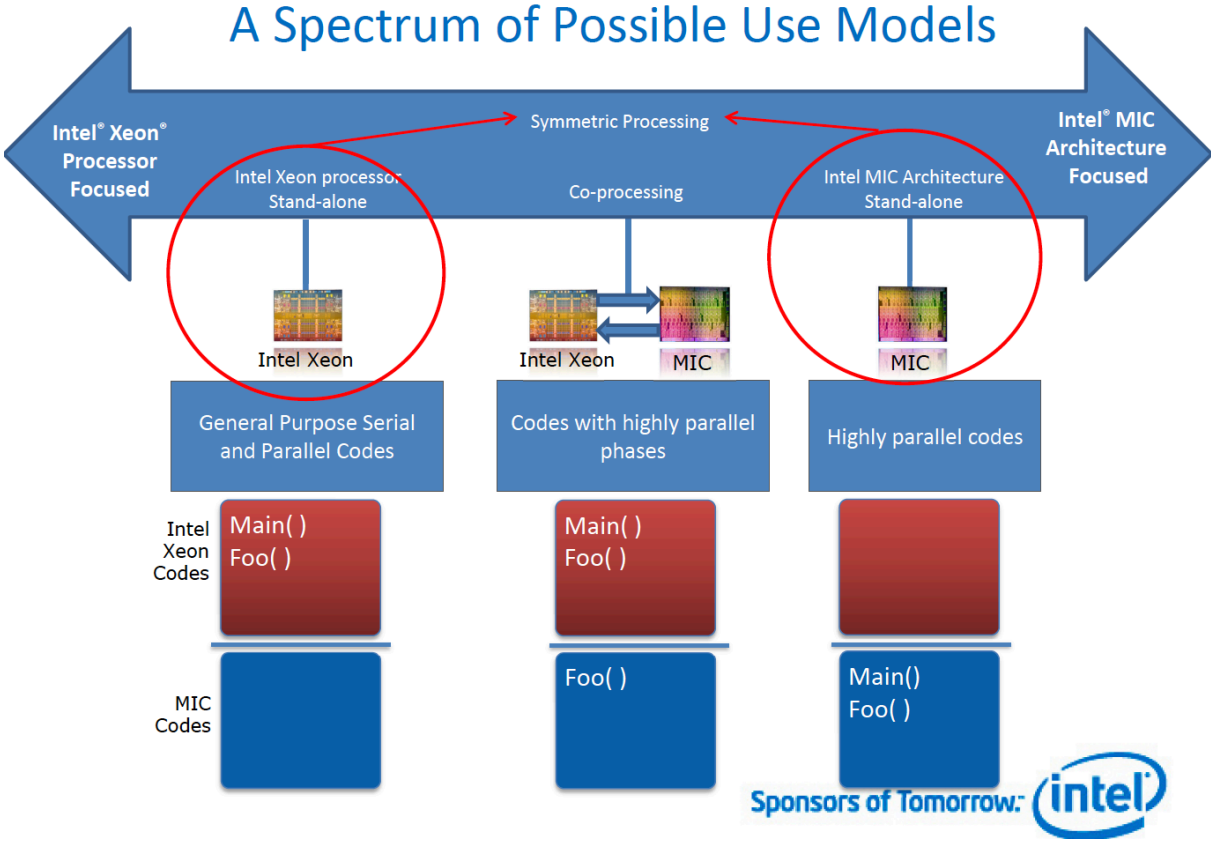




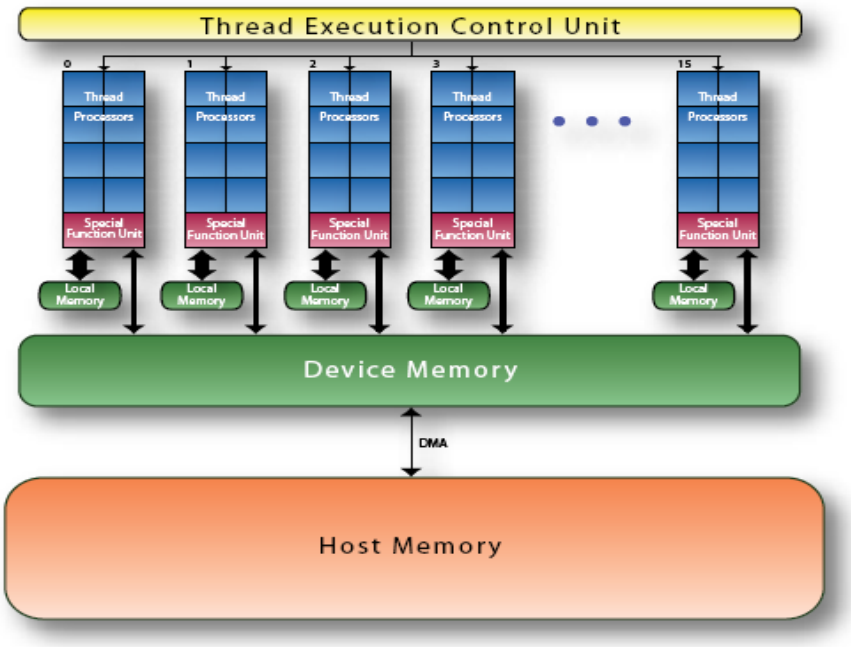
The new Knights Landing architecture



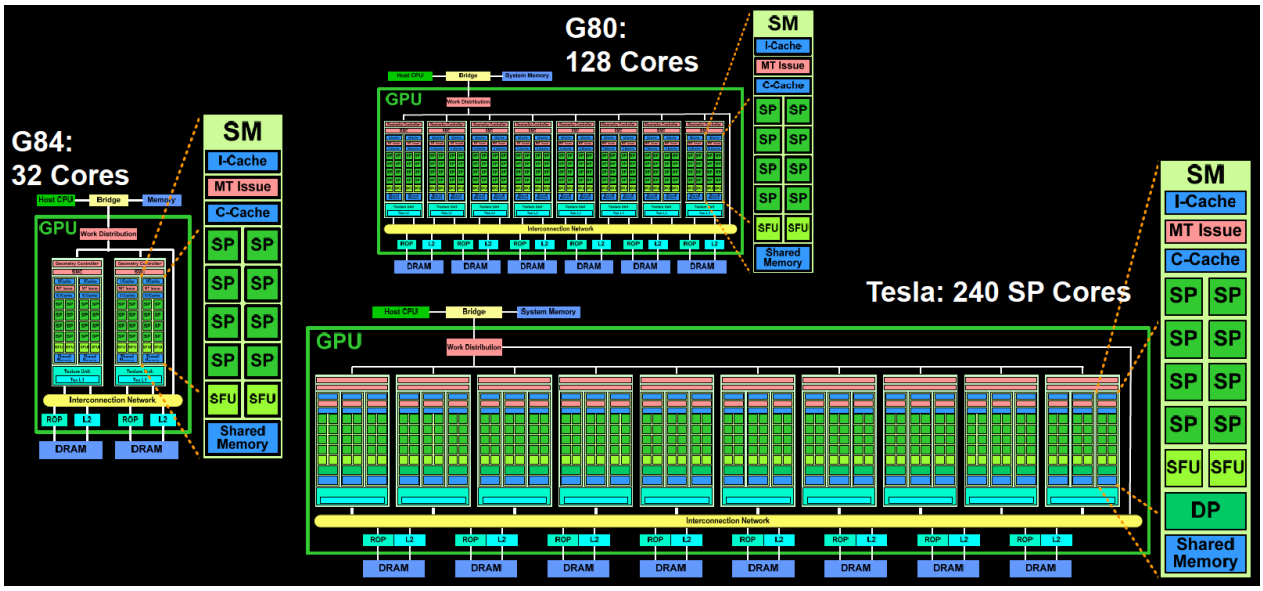
A Spectrum of Possible Use Models



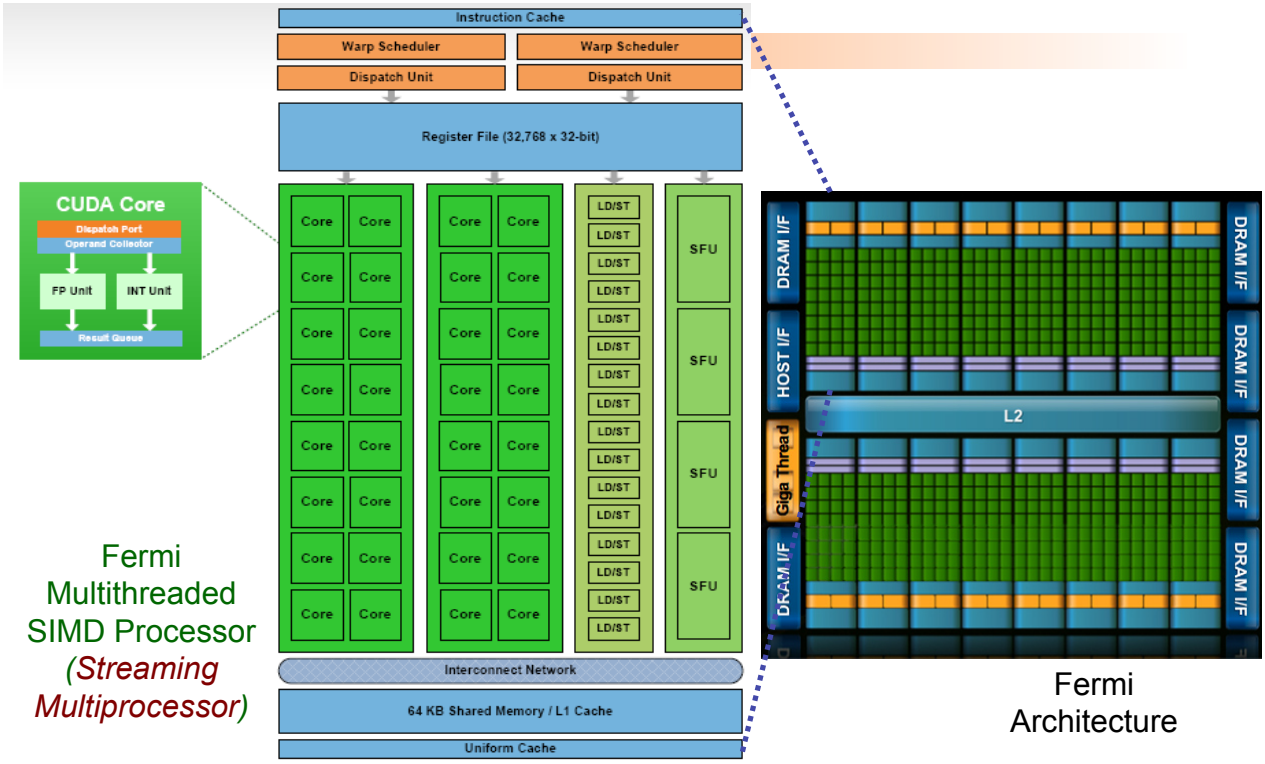
The GPU as a compute device: the G80



NVidia GPU structure & scalability



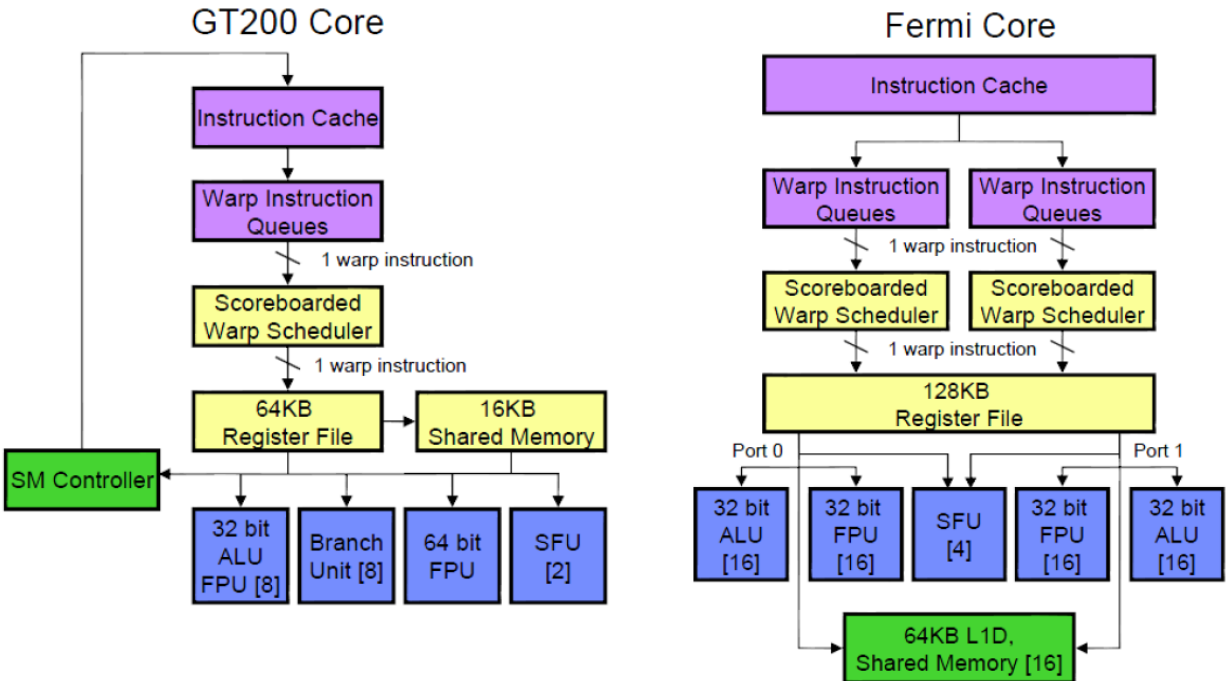
The NVidia Fermi architecture



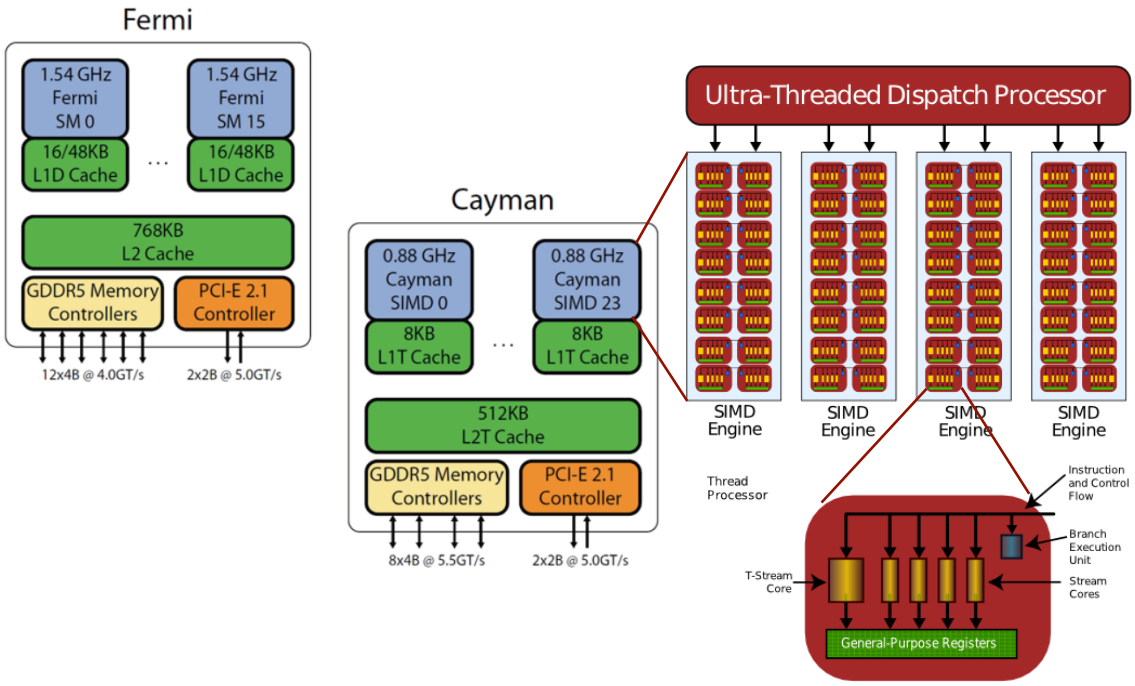
Fermi Multithreaded SIMD Processor *(Streaming Multiprocessor)*

Fermi Architecture

GT200 and Fermi SIMD processor



GPU: NVidia Fermi versus AMD Cayman



Fermi Architecture Innovations

- Each SIMD processor has
  - Two SIMD thread schedulers, two instruction dispatch units
  - 16 SIMD lanes (SIMD width=32, chime=2 cycles), 16 load-store units, 4 special function units
  - Thus, two threads of SIMD instructions are scheduled every two clock cycles
- Fast double precision
- Caches for GPU memory
- 64-bit addressing and unified address space
- Error correcting codes
- Faster context switching
- Faster atomic instructions

Families in NVidia GPU

GPU	G80	GT200	Fermi
Transistors	681 million	1.4 billion	3.0 billion
CUDA Cores	128	240	512
Double-Precision	None	30 FMA ops per clock	256 FMA ops per clock

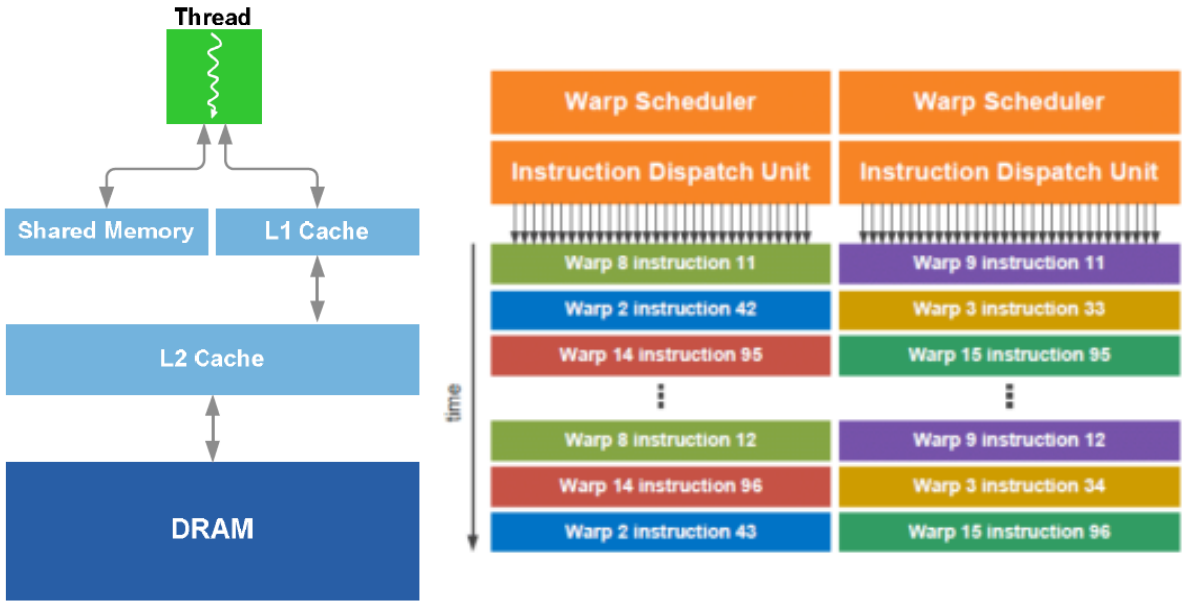
GPU	GT200 (Tesla)	GF110 (Fermi)	GK104 (Kepler)
Transistors	1.4 billion	3.0 billion	3.54 billion
CUDA Cores	240	512	1536
Graphics Core Clock	648MHz	772MHz	1006MHz
Shader Core Clock	1476MHz	1544MHz	n/a
GFLOPs	1063	1581	3090
Texture Units	80	64	128
Texel fill-rate	51.8 Gigatexels/sec	49.4 Gigatexels/sec	128.8 Gigatexels/sec
Memory Clock	2484 MHz	4008 MHz	6008MHz
Memory Bandwidth	159 GB/sec	192.4 GB/sec	192.26 GB/sec
Max # of Active Displays	2	2	4
TDP	183W	244W	195W

ECC Memory Protection	No	No	Yes
Concurrent Kernels	No	No	Up to 16

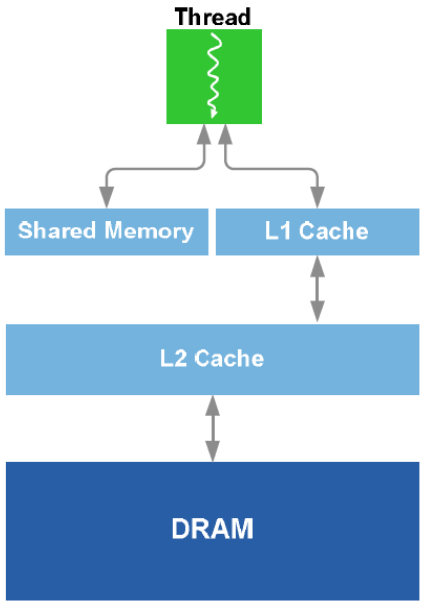
NVIDIA GPU Memory Structures

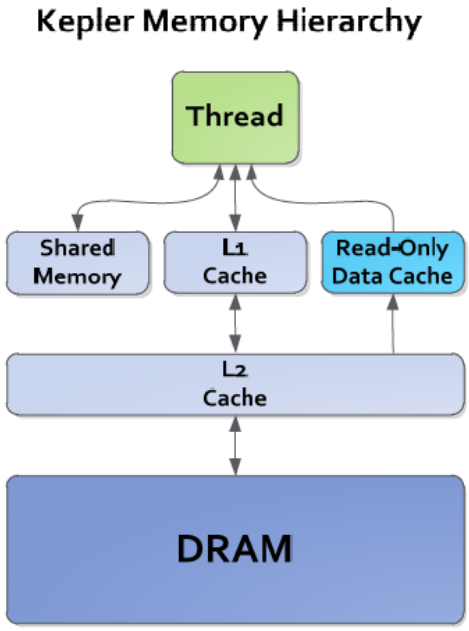
- Each SIMD Lane has private section of **off-chip DRAM**
  - "Private memory" *(Local Memory)*
  - Contains stack frame, spilling registers, and private variables
- Each multithreaded SIMD processor also has local memory *(Shared Memory)*
  - Shared by SIMD lanes / threads within a block
- Memory shared by SIMD processors is GPU Memory *(Global Memory)*
  - Host can read and write GPU memory

Fermi: Multithreading and Memory Hierarchy



From Fermi into Kepler: The Memory Hierarchy

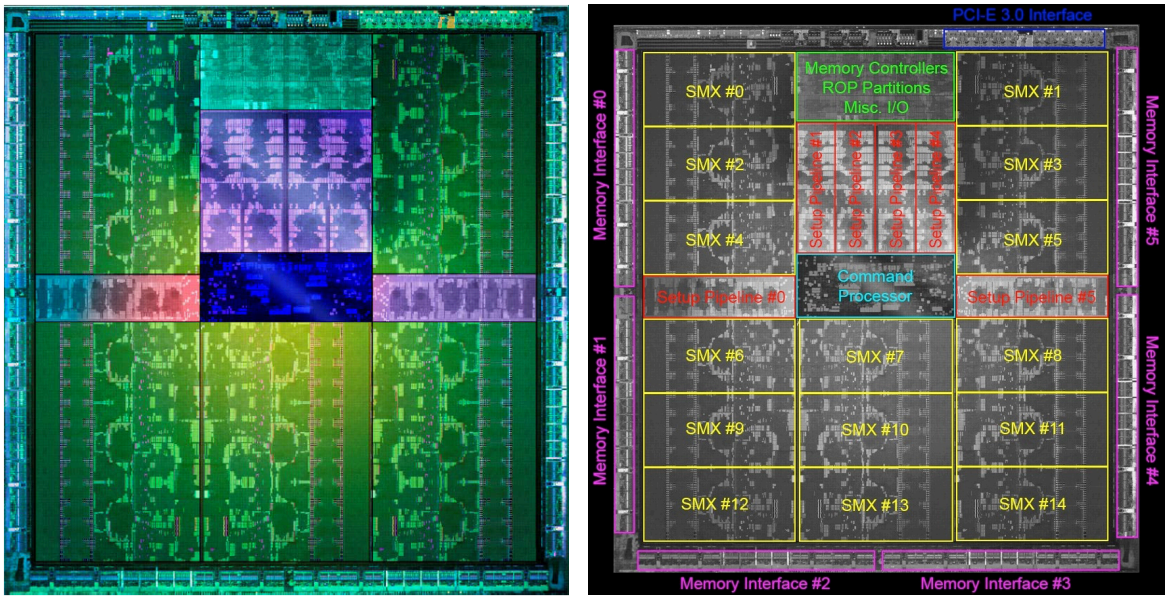




From Fermi into Kepler: Compute capabilities

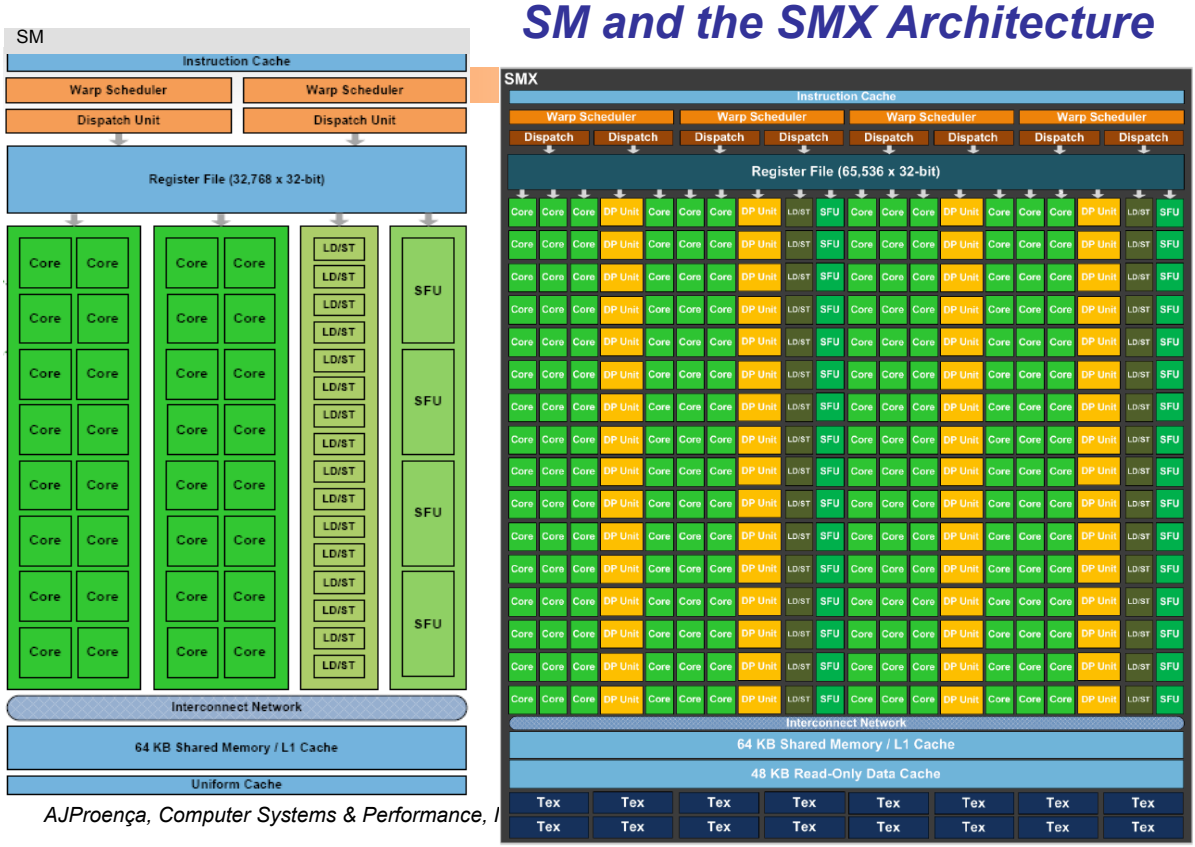
	FERMI GF100	FERMI GF104	KEPLER GK104	KEPLER GK110
Compute Capability	2.0	2.1	3.0	3.5
Threads / Warp	32	32	32	32
Max Warps / Multiprocessor	48	48	64	64
Max Threads / Multiprocessor	1536	1536	2048	2048
Max Thread Blocks / Multiprocessor	8	8	16	16
32-bit Registers / Multiprocessor	32768	32768	65536	65536
Max Registers / Thread	63	63	63	255
Max Threads / Thread Block	1024	1024	1024	1024
Shared Memory Size Configurations (bytes)	16K 48K	16K 48K	16K 32K 48K	16K 32K 48K
Max X Grid Dimension	2^16-1	2^16-1	2^32-1	2^32-1
Hyper-Q	No	No	No	Yes
Dynamic Parallelism	No	No	No	Yes

Kepler GK110 Die & Architecture

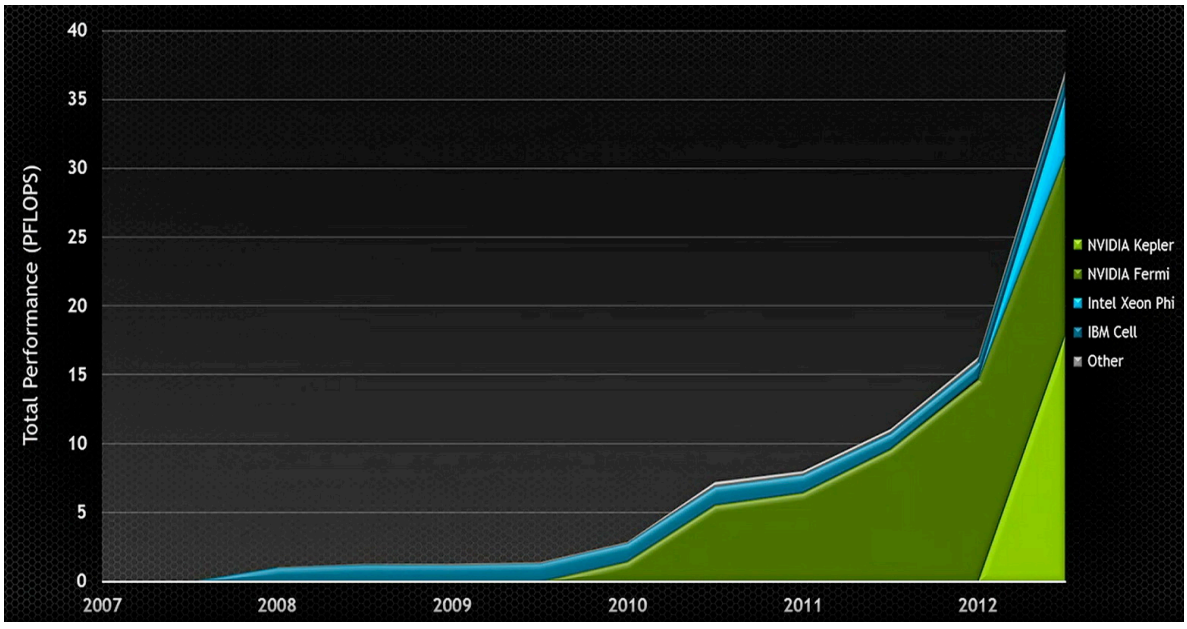


Overview of GK110 Kepler Architecture

From Fermi to Kepler core: SM and the SMX Architecture

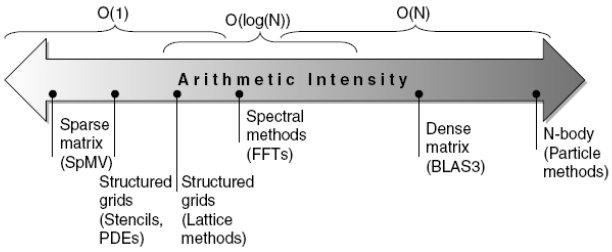


Top500: Performance from Accelerators



Roofline Performance Model

- Basic idea:
  - Plot peak floating-point throughput as a function of arithmetic intensity
  - Ties together floating-point performance and memory performance for a target machine
- Arithmetic intensity
  - Floating-point operations per byte read



Examples

- Attainable GFLOPs/sec Min = (Peak Memory BW × Arithmetic Intensity, Peak Floating Point Perf.)