HPCS 2014

CINECA

Overview of Intel Xeon Phi

F. Salvadore - Cineca

Contents

HPCS 2014

CINECA

(1) Overview of Intel Xeon Phi

(2) Programming models:

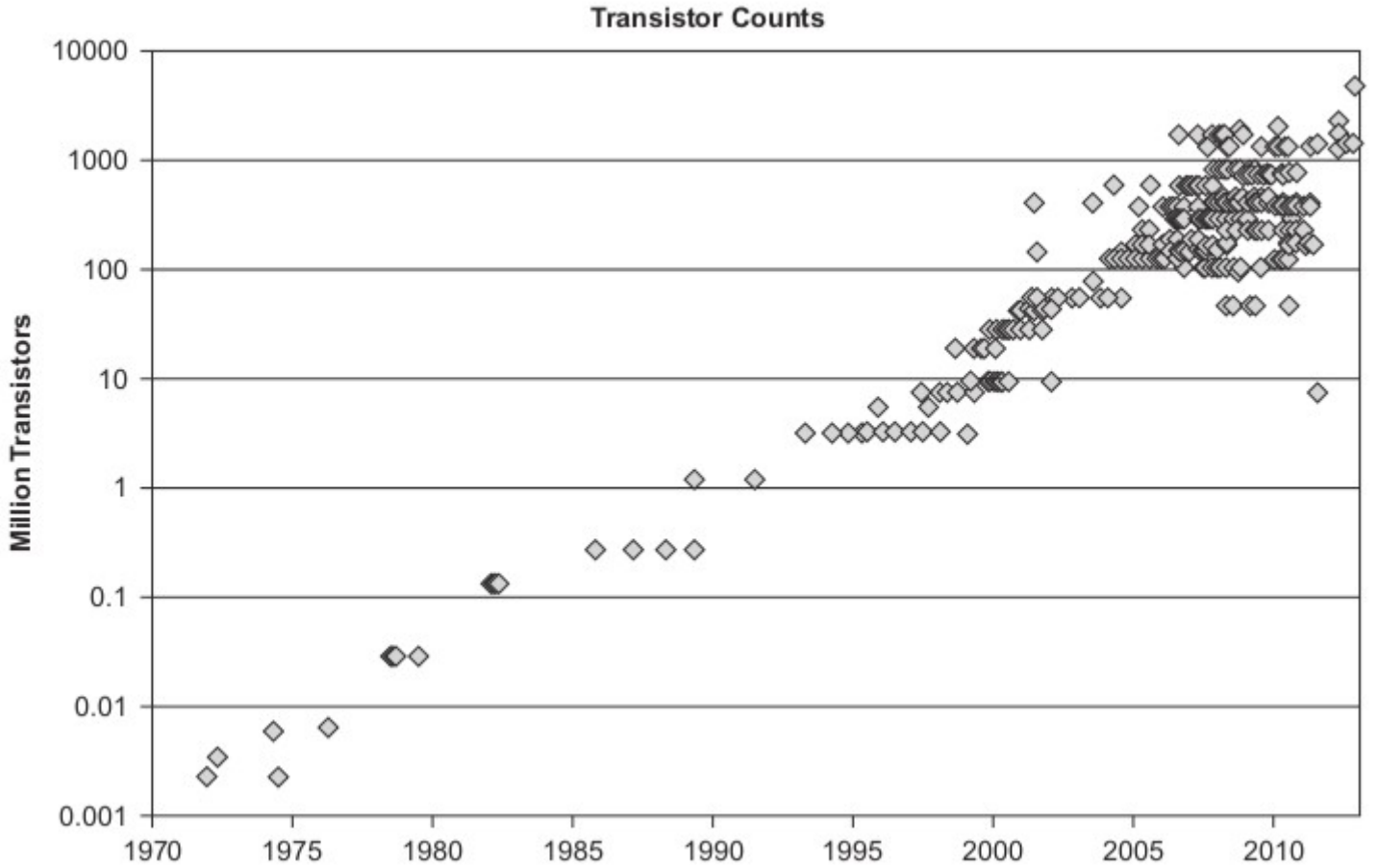
  (a) native mode

  (b) offload mode

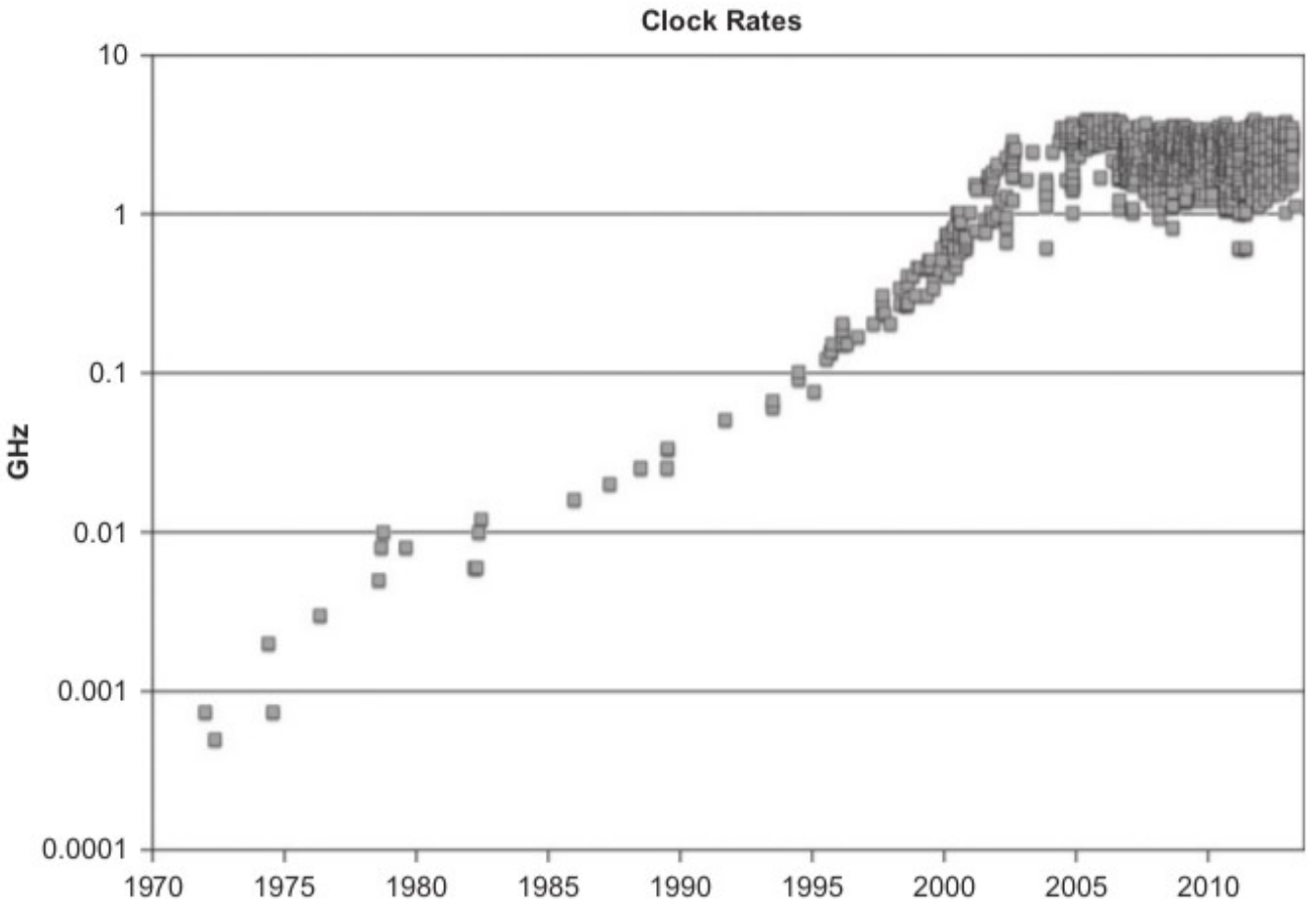
(3) Libraries: MKL

(4) Optimization hints

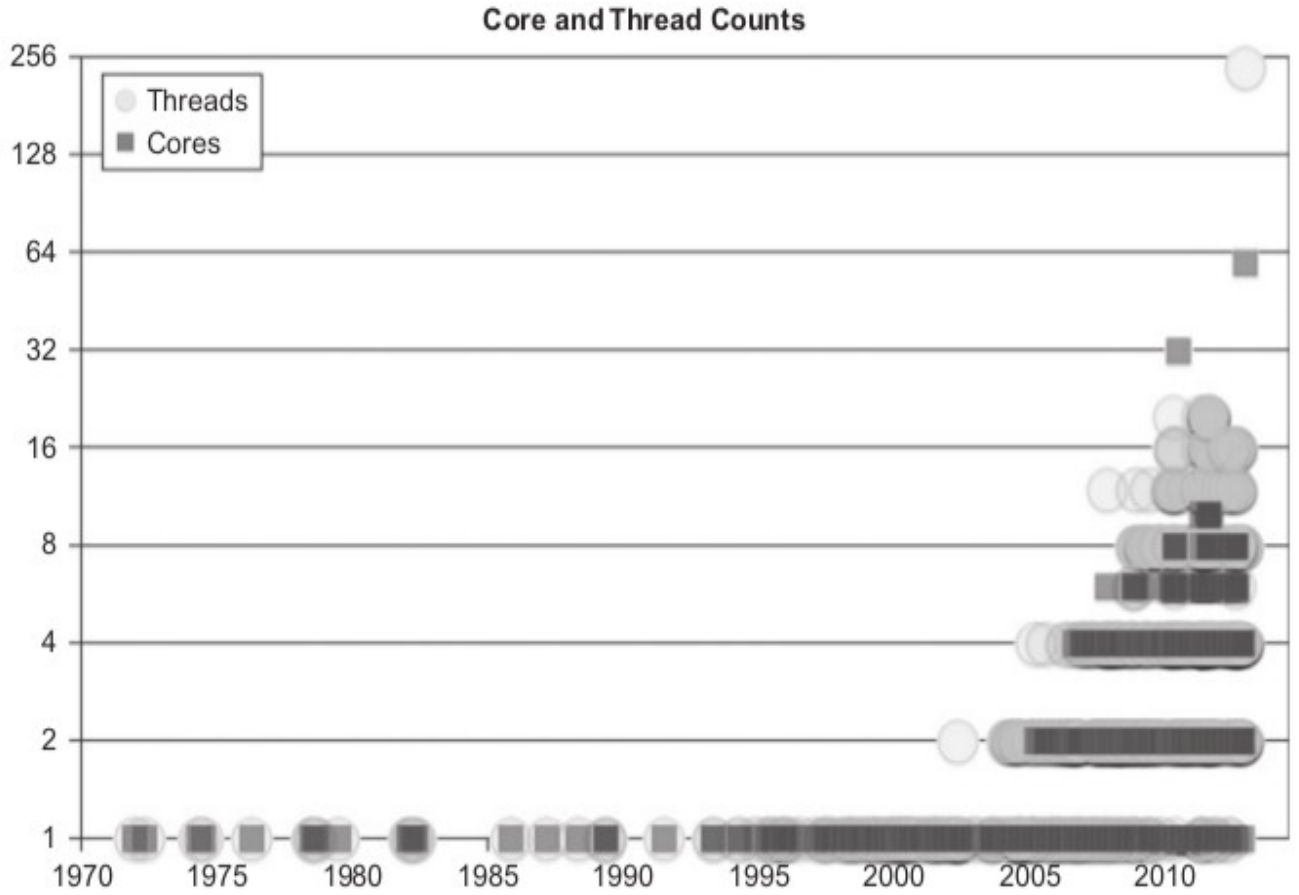
Trends: transistors



Trends: clock rates



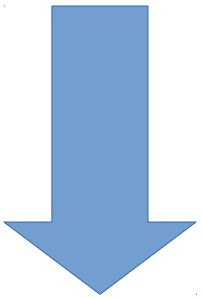
Trends: cores and threads



Trends: summarizing...

- The number of transistors increases
- The power consumption must not increase
- The density cannot increase on a single chip

Solution



- Increase the number of cores

Xeon Phi very basic features

- Let me introduce you

What is Xeon Phi?

HPCS 2014

CINECA

- 7100 / 5100 / 3100 Series available
- 5110P:
  - Intel Xeon Phi clock: 1053 MHz
  - 60 cores in-order
  - ~ 1 TFlops/s DP peak performance (2 Tflops SP)
  - 4 hardware threads per core
  - 8 GB DDR5 memory
  - 512-bit SIMD vectors (32 registers)
  - Fully-coherent L1 and L2 caches
  - PCIe bus (rev. 2.0)
  - Max Memory bandwidth (theoretical) 320 GB/s
  - Max TDP: 225 W

MIC vs GPU *naïve* comparison

HPCS 2014

CINECA

- The comparison is naïve
  - MICs and GPUs are two different types of devices!

System	K20s	5110P
# cores	2496	60 (*4)
Memory size	5 GB	8 GB
Peak performance (SP)	3.52 TFlops	2 TFlops
Peak performance (DP)	1.17 TFlops	1 TFlops
Clock rate	0.706 GHz	1.053 GHz
Memory bandwidth	208 GB/s (ECC off)	320 GB/s

Terminology

- **MIC** = Many Integrated Cores is the name of the architecture
- **Xeon Phi** = Commercial name of the Intel product based on the MIC architecture
- **Knight's corner,** Knight's landing, Knight's ferry are development names of MIC architectures
- We will oftern refer to the CPU as **HOST** and Xeon Phi as **DEVICE**

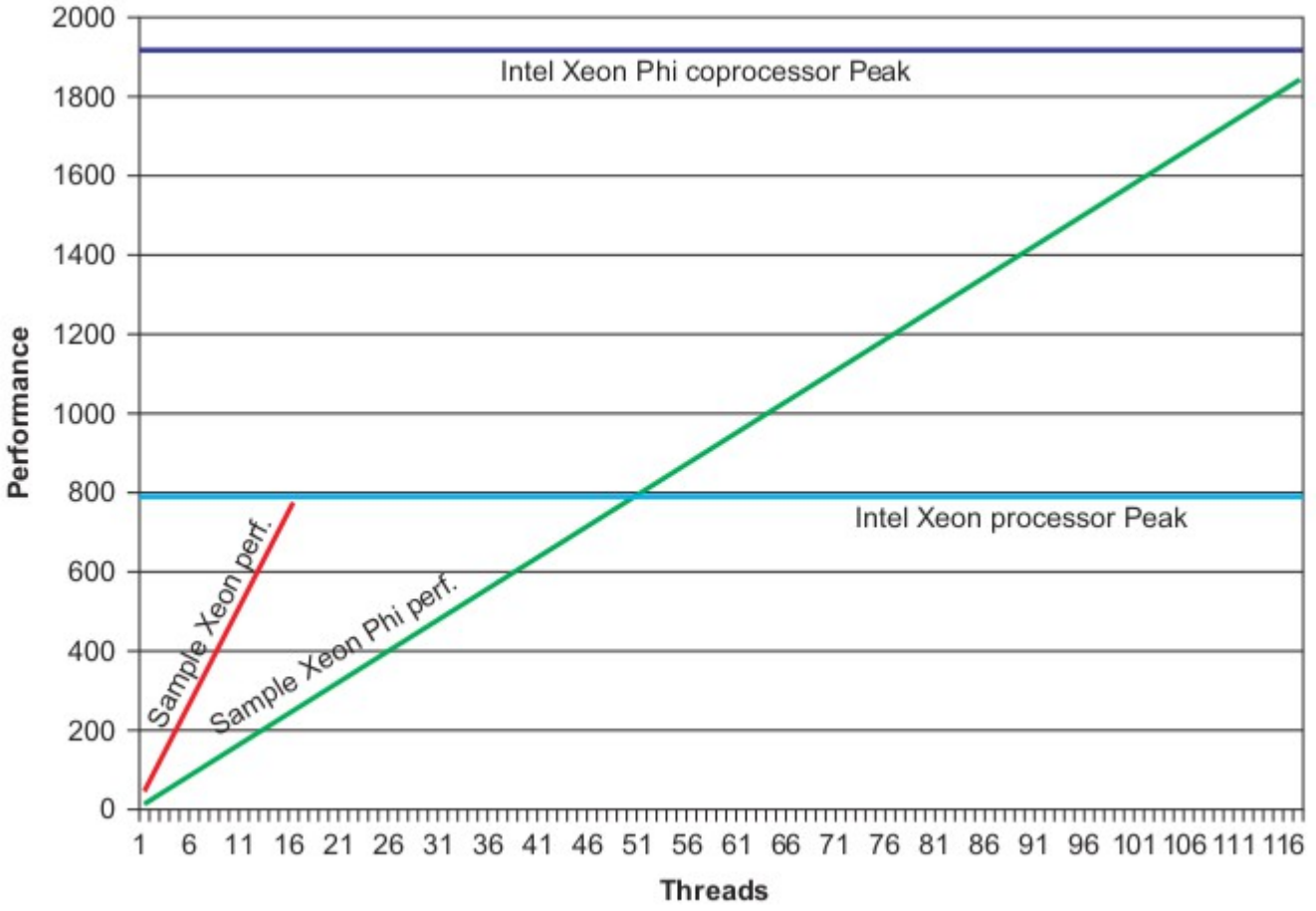
Is it an accelerator?

- YES: It can be used to “accelerate” hot-spots of the code that are highly parallel and computationally extensive
- In this sense, it works alongside the CPU
- It can be used as an accelerator using the “offload” programming model
- An important bottleneck is represented by the communication between host and device (through PCIe)
- Under this respect, it is very similar to a GPU

Is it an accelerator? / 2

- NOT ONLY: the Intel Xeon Phi can behave as a many-core X86 node.
  - Code can be compiled and run “natively” on the Xeon Phi platform using MPI + OpenMP
- The bottleneck is the scalability of the code
  - Amdahl Law
- Under this respect, the Xeon Phi is completely different from a GPU
  - This is way we often call the Xeon Phi “co-processor” rather than “accelerator”

Many-core performances

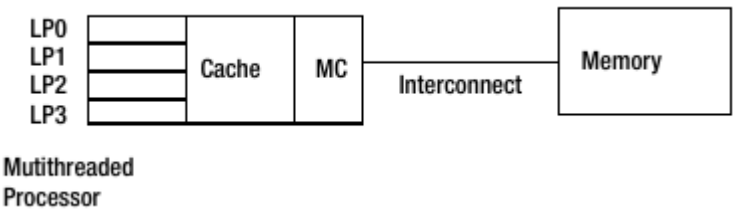


Architecture key points/1

HPCS 2014

CINECA

- Instruction Pipelining
  - Two independent pipelines arbitrarily known as the U and V pipelines
  - (only) 5 stages to cope with a reduced clock rate, e.g. compared to the Pentium 20 stages
  - In-order instruction execution
- Manycore architecture
  - Homogeneous
  - 4 hardware threads per core



Architecture key points/2

- Interconnect: bidirectional ring topology
  - All the cores talk to one another through a bidirectional interconnect
  - The cores also access the data and code residing in the main memory through the ring connecting the cores to memory controller
- Given eight memory controllers with two GDDR5 channels running at 5.5 GT/s
  - Aggregate Memory Bandwidth = 8 memory controllers × 2 channels × 5.5 GT/s × 4 bytes/transfer = 352 GB/s
- System interconnect
  - Xeon Phi are often placed on PCIe slots to work with the host processors

Architecture key points/3

- Cache:
  - L1: 8-ways set-associative 32-kB instruction and 32-kB data
  - L1 access time: 3 cycles
  - L2: 8-way set associative and 512 kB in size (unified)Interconnect: bidirectional ring topology
- TLB cache:
  - L1 data TLB supports three page sizes: 4 kB, 64 kB, and 2 MB
  - L2 TLB
  - If one misses L1 and also misses L2 TLB, one has to walk four levels of page table, which is pretty expensive

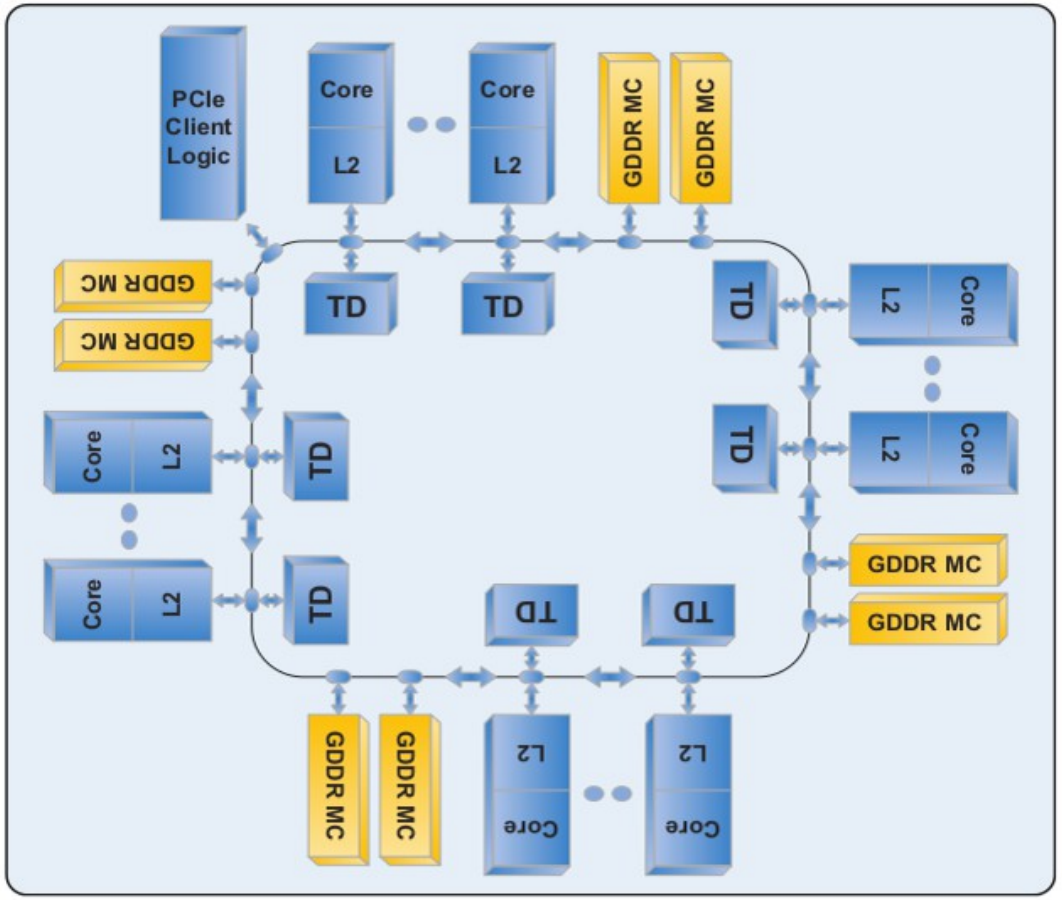
Architecture key points/4

- The VPU (vector processing unit) implements a novel instruction set architecture (ISA), with 218 new instructions compared with those implemented in the Xeon family of SIMD instruction sets.
- The VPU is fully pipelined and can execute most instructions with four-cycle latency and single-cycle throughput.
- Each vector can contain 16 single-precision floats or 32-bit integer elements or eight 64-bit integer or double-precision floating point elements.

Architecture key points/5

- Each VPU instruction passes through one or more of the following five pipelines to completion:
  - Double-precision (DP) pipeline: Used to execute float64 arithmetic, conversion from float64 to float32, and DP-compare instructions.
  - Single-precision (SP) pipeline: Executes most of the instructions including 64-bit integer loads. This includes float32/int32 arithmetic and logical operations, shuffle/broadcast, loads including loadunpack, type conversions from float32/int32 pipelines, extended math unit (EMU) transcendental instructions, int64 loads, int64/float64 logical, and other instructions.
  - Mask pipeline: Executes mask instructions with one-cycle latencies.
  - Store pipeline: Executes the vector store operations.
  - Scatter/gather pipeline: Executes the vector register read/writes from sparse memory locations.
- Mixing SP and DP computations is expensive!

Architecture sketch/1



Architecture sketch/2