10th Advanced School on PARALLEL COMPUTING

Energy Efficiency and Roadmap to Exascale

Carlo Cavazzoni – c.cavazzoni@cineca.it
SuperComputing Applications and Innovation Department

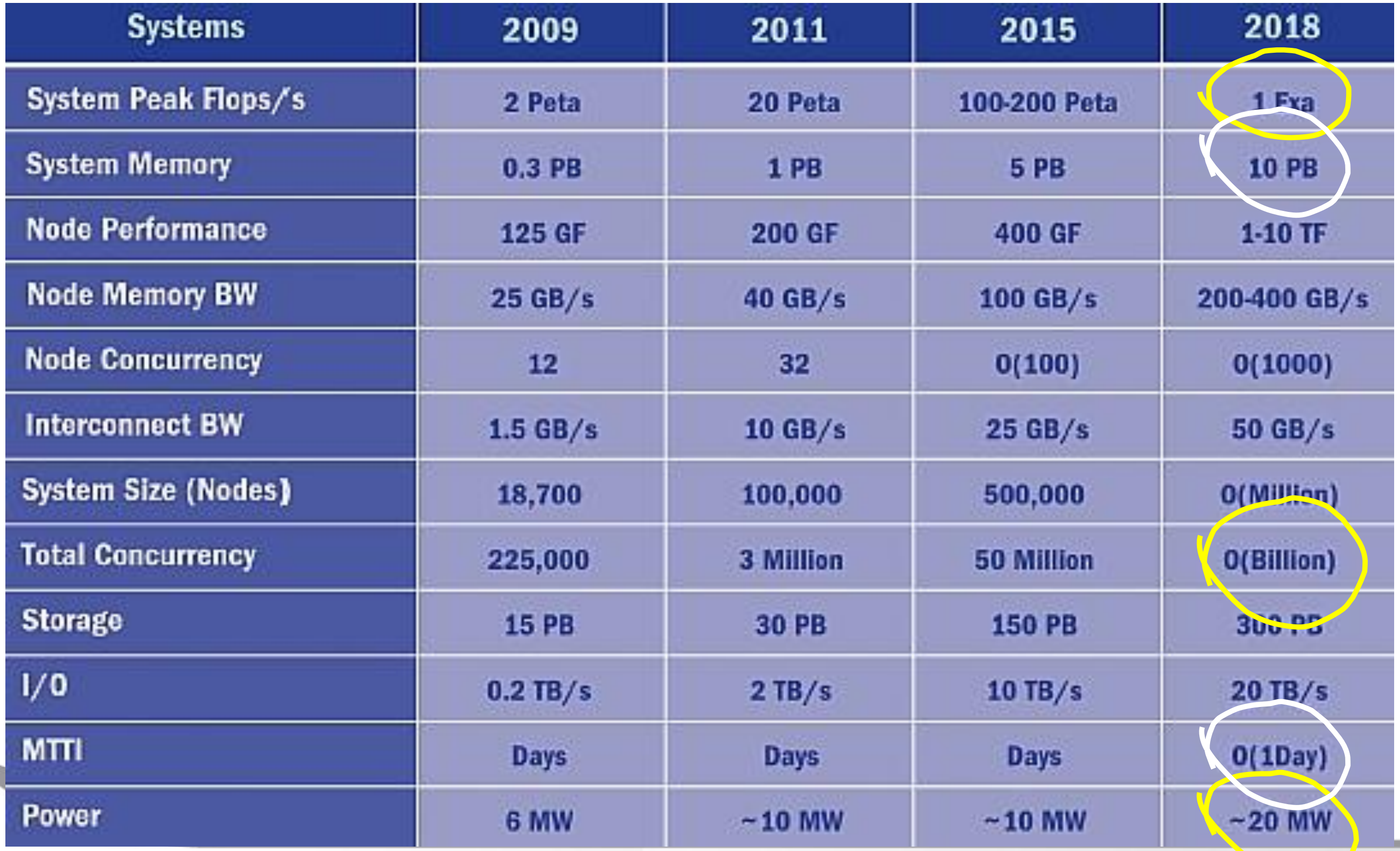
CINECA

outline

- Roadmap to Exascale
- HPC architecture challanges
- Energy efficiency
- Co processor architecture

Roadmap to Exascale

(architectural trends)



Systems	2009	2011	2015	2018
System Peak Flops/s	2 Peta	20 Peta	100-200 Peta	1 Exa
System Memory	0.3 PB	1 PB	5 PB	10 PB
Node Performance	125 GF	200 GF	400 GF	1-10 TF
Node Memory BW	25 GB/s	40 GB/s	100 GB/s	200-400 GB/s
Node Concurrency	12	32	O(100)	O(1000)
Interconnect BW	1.5 GB/s	10 GB/s	25 GB/s	50 GB/s
System Size (Nodes)	18,700	100,000	500,000	O(Million)
Total Concurrency	225,000	3 Million	50 Million	O(Billion)
Storage	15 PB	30 PB	150 PB	300 PB
I/O	0.2 TB/s	2 TB/s	10 TB/s	20 TB/s
MTTI	Days	Days	Days	O(1Day)
Power	6 MW	~10 MW	~10 MW	~20 MW

Dennard scaling law (downscaling)

new VLSI gen.

old VLSI gen.

$L' = L / 2$

$V' = V / 2$

$F' = F * 2$

$D' = 1 / L^2 = 4D$

$P' = P$

do not hold anymore!

$L' = L / 2$

$V' = \sim V$

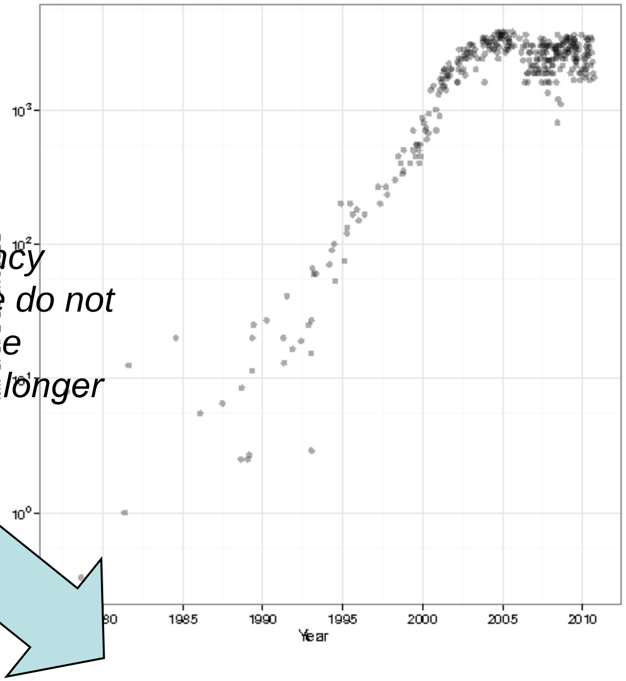
$F' = \sim F * 2$

$D' = 1 / L^2 = 4 * D$

$P' = 4 * P$

The power crisis!

The core frequency and performance do not grow following the Moore's law any longer



Increase the number of cores to maintain the architectures evolution on *the Moore's law*

Programming crisis!

- Now, power and/or heat generation are the limiting factors of the down-scaling
- Supply voltage reduction is becoming difficult, because Vth cannot be decreased any more, as described later.
- Growth rate in clock frequency and chip area becomes smaller.

Economic and market law

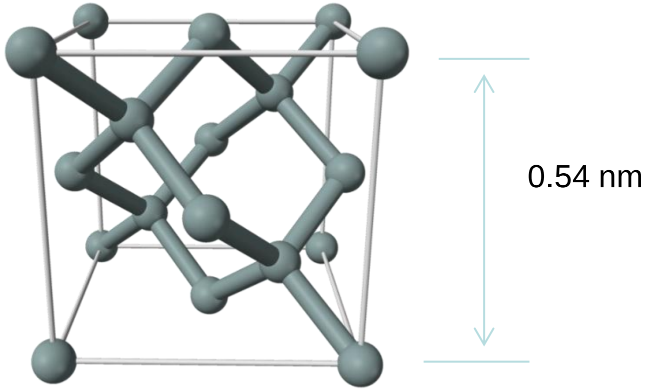
Stacy Smith, **Intel's chief financial officer**, later gave some more detail on the *economic benefits of staying on the Moore's Law race.*

The cost per chip "is going down more than the capital intensity is going up," Smith said, suggesting Intel's profit margins should not suffer because *of heavy capital spending. "This is the economic beauty of Moore's Law."*
And Intel has a good handle on the next production shift, shrinking circuitry to 10 nanometers. Holt said the company has test chips running on that *technology. "We are projecting similar kinds of improvements in cost out to 10 nanometers," he said.*
So, despite the challenges, Holt could not be induced to say there's any looming end to Moore's Law, the invention race that has been a key driver *of electronics innovation since first defined by Intel's co*-founder in the mid-19060s.

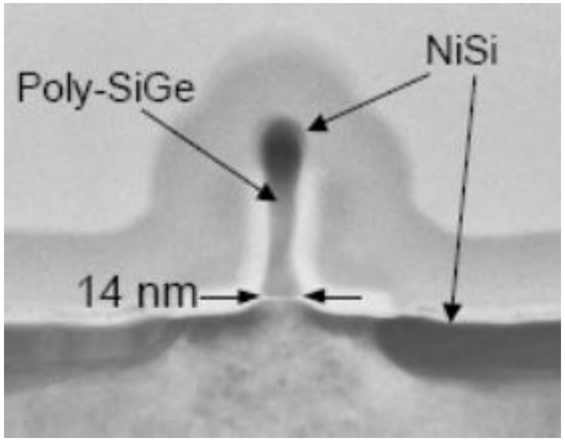
From WSJ

It is all about the number of chips per Si wafer!

But!



Si lattice

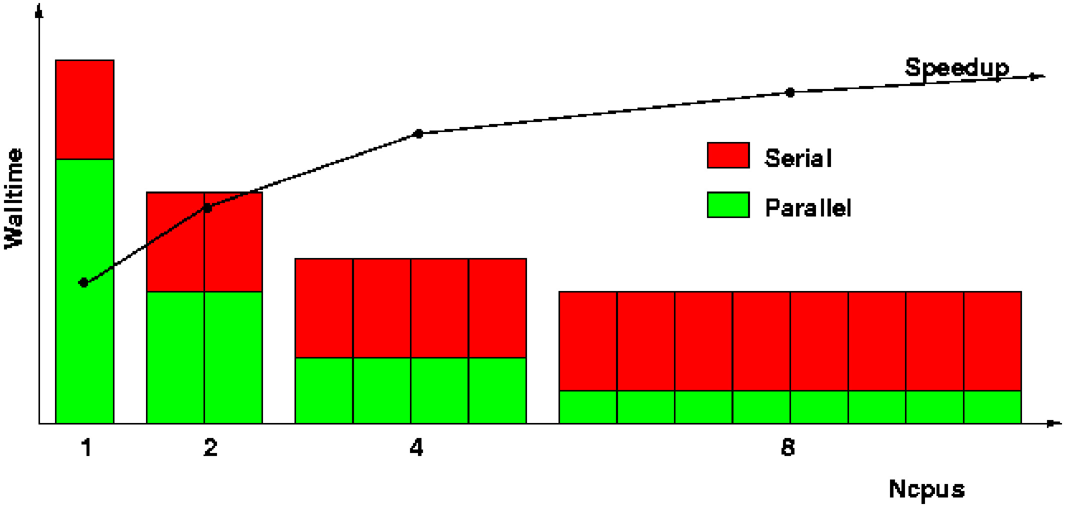


300 atomi!

There will be still 4~6 cycles (or technology generations) left until
we reach 11 ~ 5.5 nm technologies, at which we will reach downscaling limit, in some year between 2020-30 (H. Iwai, IWJT2008).

Amdahl's law

In a massively parallel context, an upper limit for the scalability of parallel applications is determined by the fraction of the overall execution time spent in non-scalable operations (Amdahl's law).



maximum speedup tends to
1 / (1 − *P*)
P= parallel fraction

1000000 core

P = 0.999999

serial fraction= 0.000001

Architectural trends

(constrained by the three law)

Peak Performance	↑	Moore law
FPU Performance	⟺	Dennard law
Number of FPUs	↑	Moore + Dennard
App. Parallelism	↑	Amdahl's law

Architectural trends (estimates)

		2020 estimates
Number of cores	⇧	10^9
Memory x core	⇩	100Mbyte or less
Memory BW/core	⇩	500GByte/sec (socket)
Memory hierachy	⇧	Reg, *L1, L2, L3, …*

Chip Architecture

Strongly market driven → Mobile, Tv set, Screens
Video/Image processing

Intel → New arch to compete with ARM
Less Xeon, but PHI

ARM → Main focus on low power mobile chip
Qualcomm, Texas inst. , Nvidia, ST, ecc
new HPC market, server maket

NVIDIA → GPU alone will not last long
ARM+GPU, Power+GPU

Power → Embedded market
Power+GPU, only chance for HPC

AMD → Console market
Still some chance for HPC

(sub) Exascale architecture

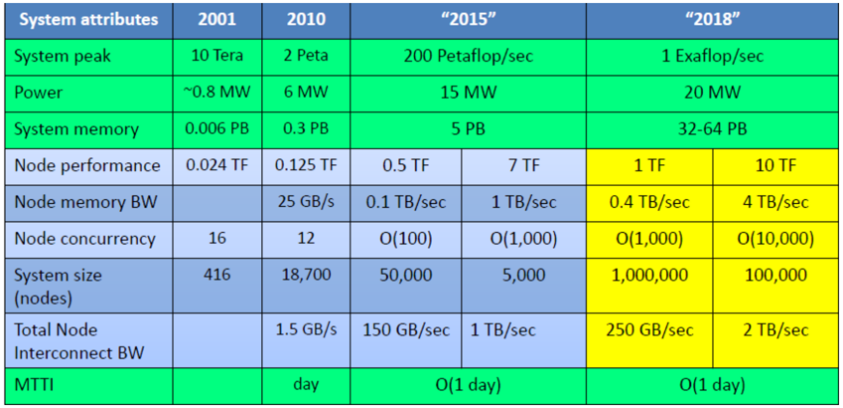
still two model

Hybrid, *but…*

Homogeneus, *but…*

What 100PFlops system we will see … my guess

IBM (hybrid) Power8+Nvidia GPU
Cray (homo/hybrid) with Intel only!
Intel (hybrid) Xeon + MIC
Arm (homo) only arm chip, *but…*
Nvidia/Arm (hybrid) arm+Nvidia
Fujitsu (homo) sparc high density low power
China (homo/hybrid) with Intel only
Room for AMD console chips



System attributes	2001	2010	“2015”		“2018”	
System peak	10 Tera	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	~0.8 MW	6 MW	15 MW		20 MW	
System memory	0.006 PB	0.3 PB	5 PB		32-64 PB	
Node performance	0.024 TF	0.125 TF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW		25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	16	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	416	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW		1.5 GB/s	150 GB/sec	1 TB/sec	250 GB/sec	2 TB/sec
MTTI		day	O(1 day)		O(1 day)	

Energy Efficiency

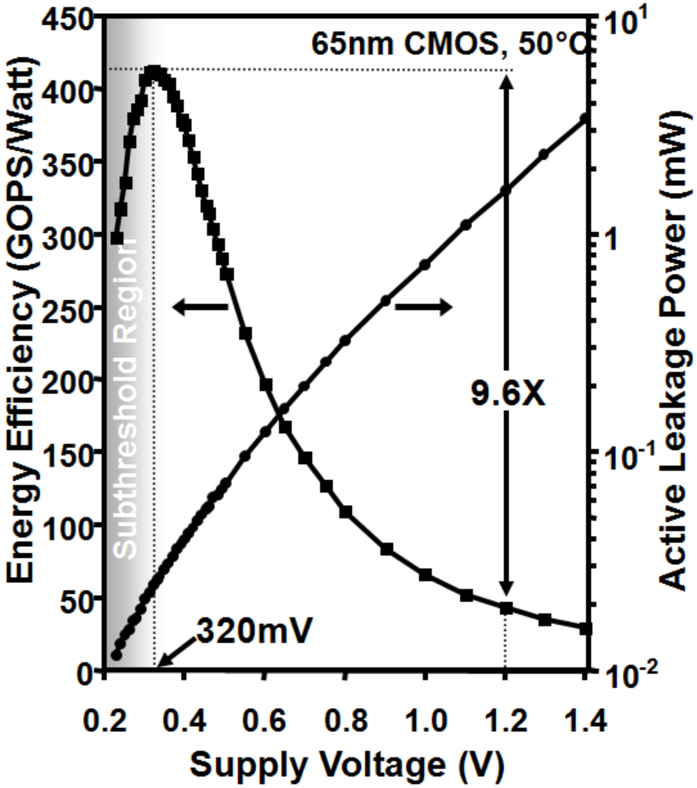
Where power is used:

1) CPU/GPU silicon
2) Memory
3) Network
4) Data transfer
5) I/O subsystem
6) Cooling

Short term impact on programming models

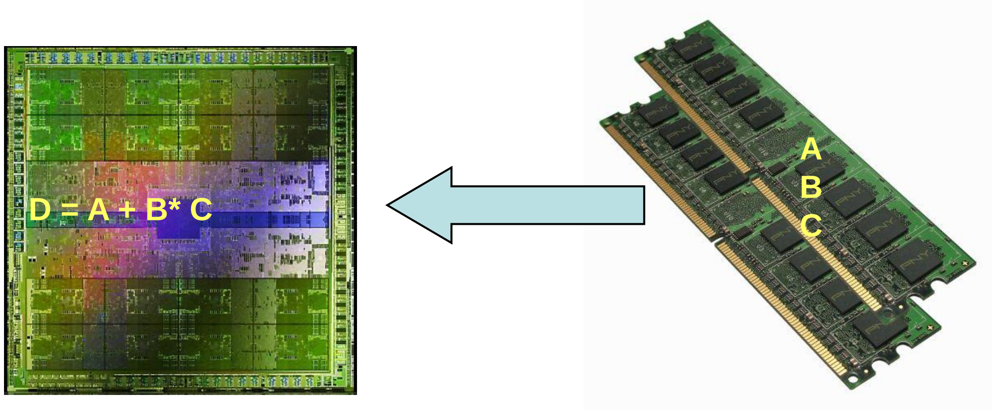
Chip efficiency

- The efficiency of CMOS transistor against the supply voltage peaks close to the insulator/conductor transition
- Possibility to design a new Near Threshold Voltage (NTV) chip architecture that is able to work at different regime.
- Accommodate the needs of different workloads and meet the requirements in term of efficiency.

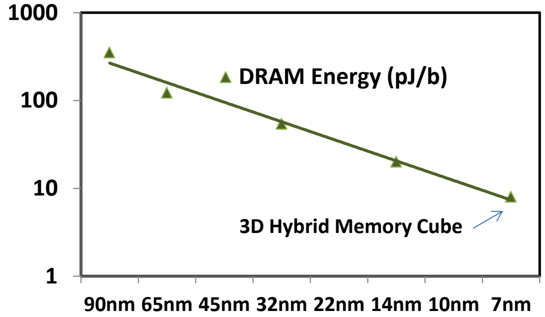


Memory

Today (at 40nm) moving 3 64bit operands to compute a 64bit floating-point FMA takes 4.7x the energy with respect to the FMA operation itself



DRAM energy scales, but not enough



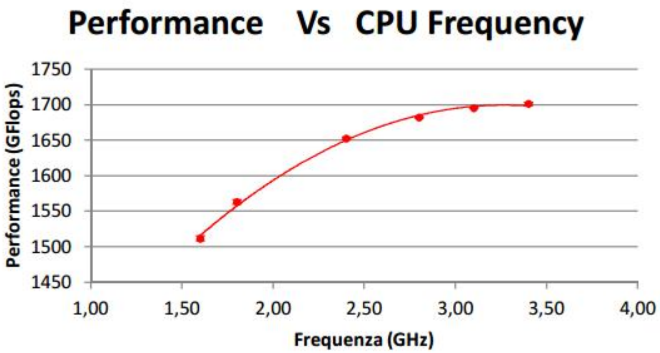
50 pJ/b today
8 pJ/b demonstrated
Need < 2pJ/b

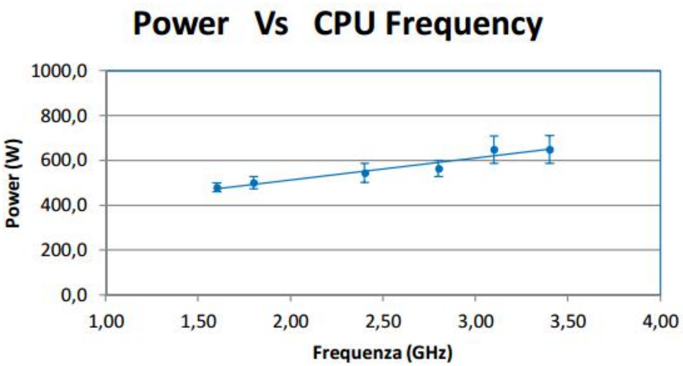
Extrapolating down to 10nm integration, the energy required to move date
Becomes 100x !

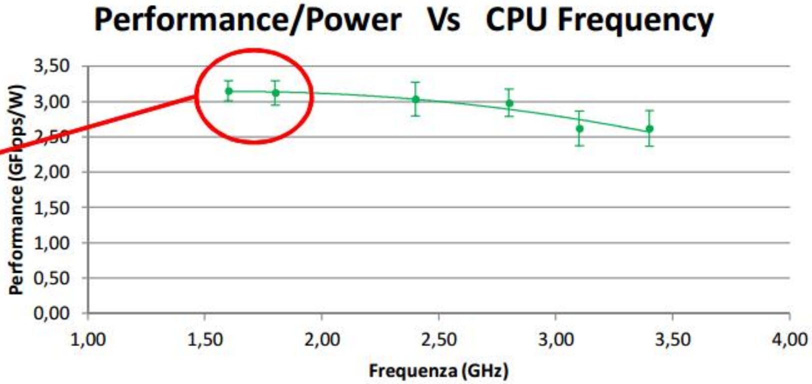
We need locality! → Fewer memory per core

EURORA Benchmarks

HPL Benchmark Results


Performance Vs CPU Frequency

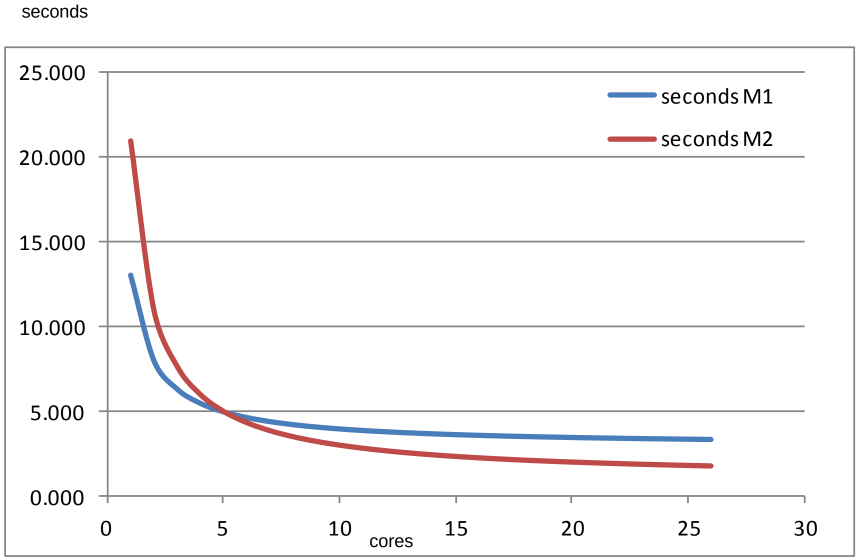

Power Vs CPU Frequency

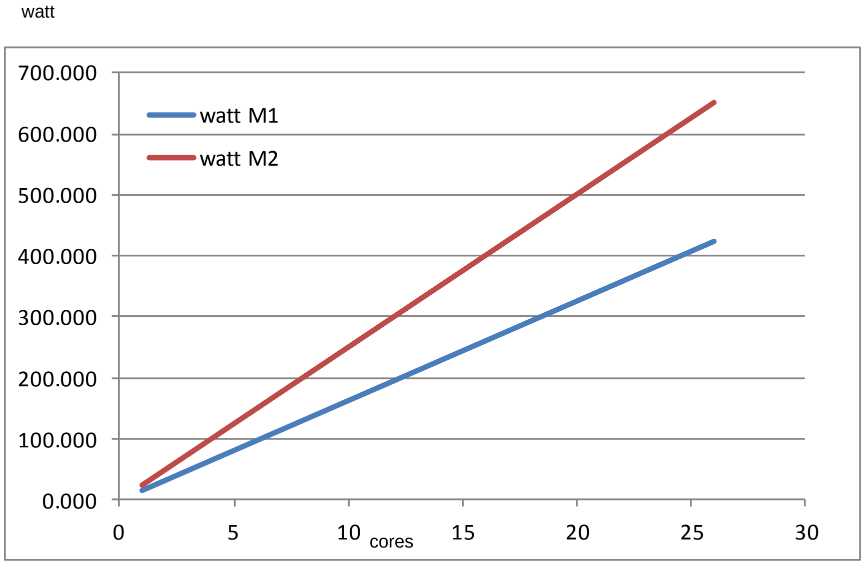

Performance/Power Vs CPU Frequency

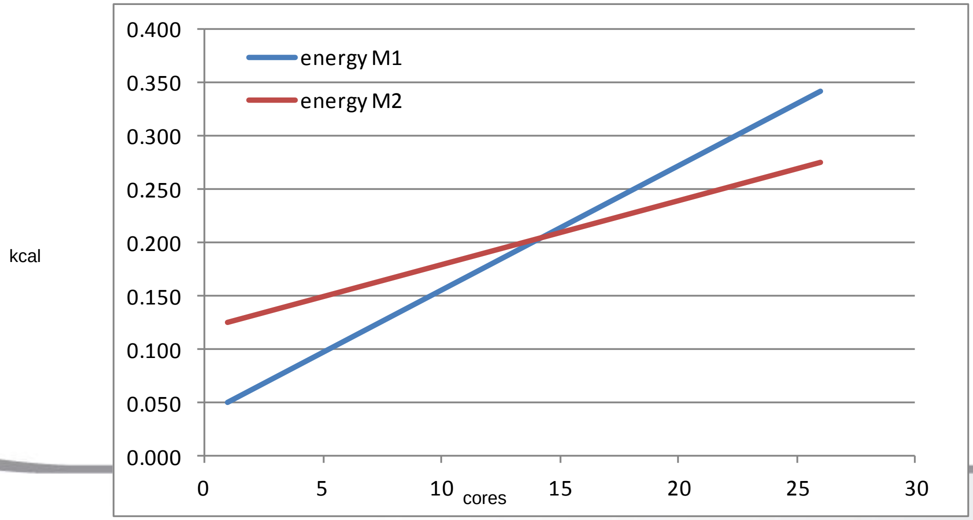
3.15 ± 0.12 !!

Energy measurments

(howto)







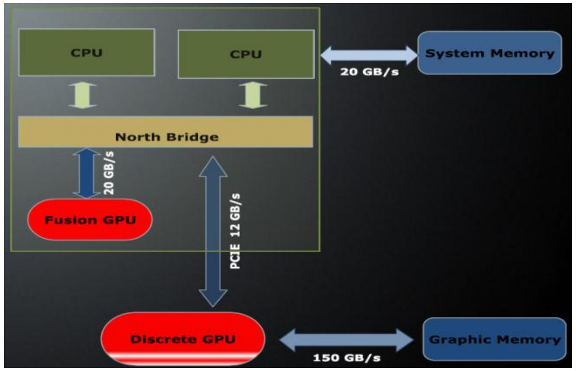
Impact on programming and execution models

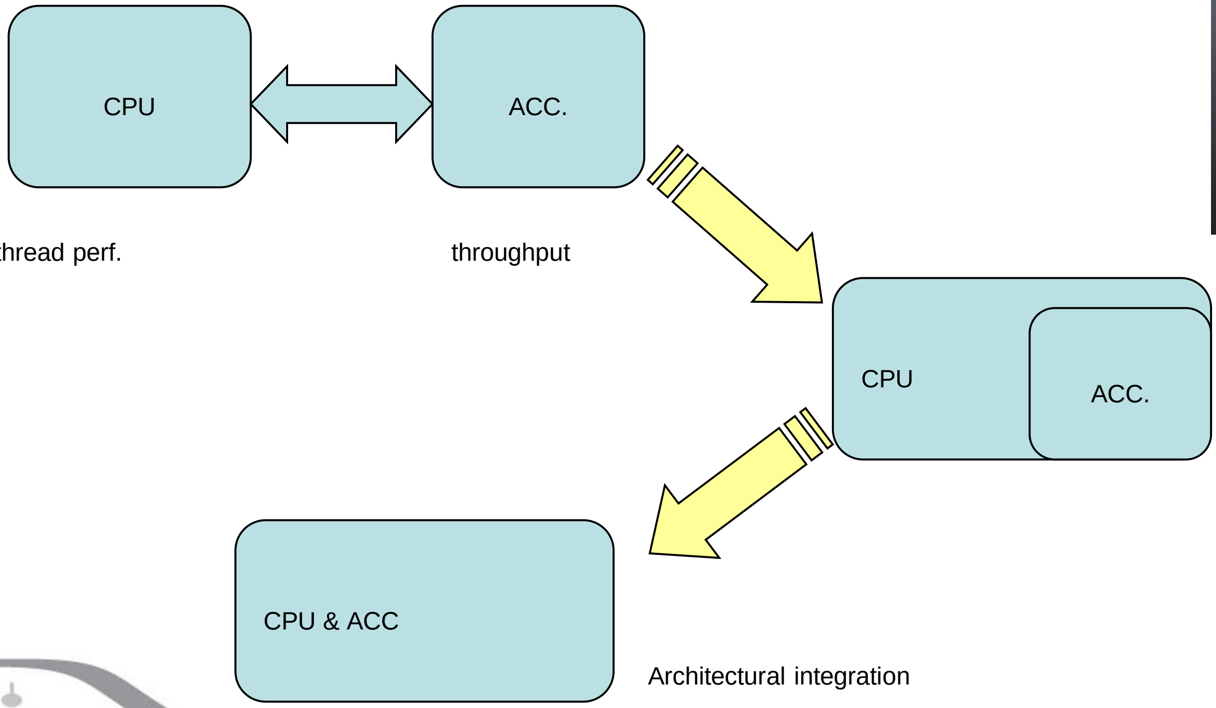
- 1. Event driven tasks (EDT)
  - a. Dataflow inspired, tiny codelets (self contained)
  - b. Non blocking, no preemption
- 2. Programming model:
  - a. Express data locality with hierarchical tiling
  - b. Global, shared, non-coherent address space
  - c.Optimization and auto generation of EDTs
- 3. Execution model:
  - a. Dynamic, event-driven scheduling, non-blocking
  - b. Dynamic decision to move computation to data
  - c.Observation based adaption (self-awareness)
  - d. Implemented in the runtime environment

Accelerators architecture

What is an Accelerator.

A set (one or more) of very simple execution units that can perform few operations (with respect to standard CPU) with very high efficiency. When combined with full featured CPU (CISC or RISC) *can accelerate the “nominal” speed of a system. (Carlo Cavazzoni)*





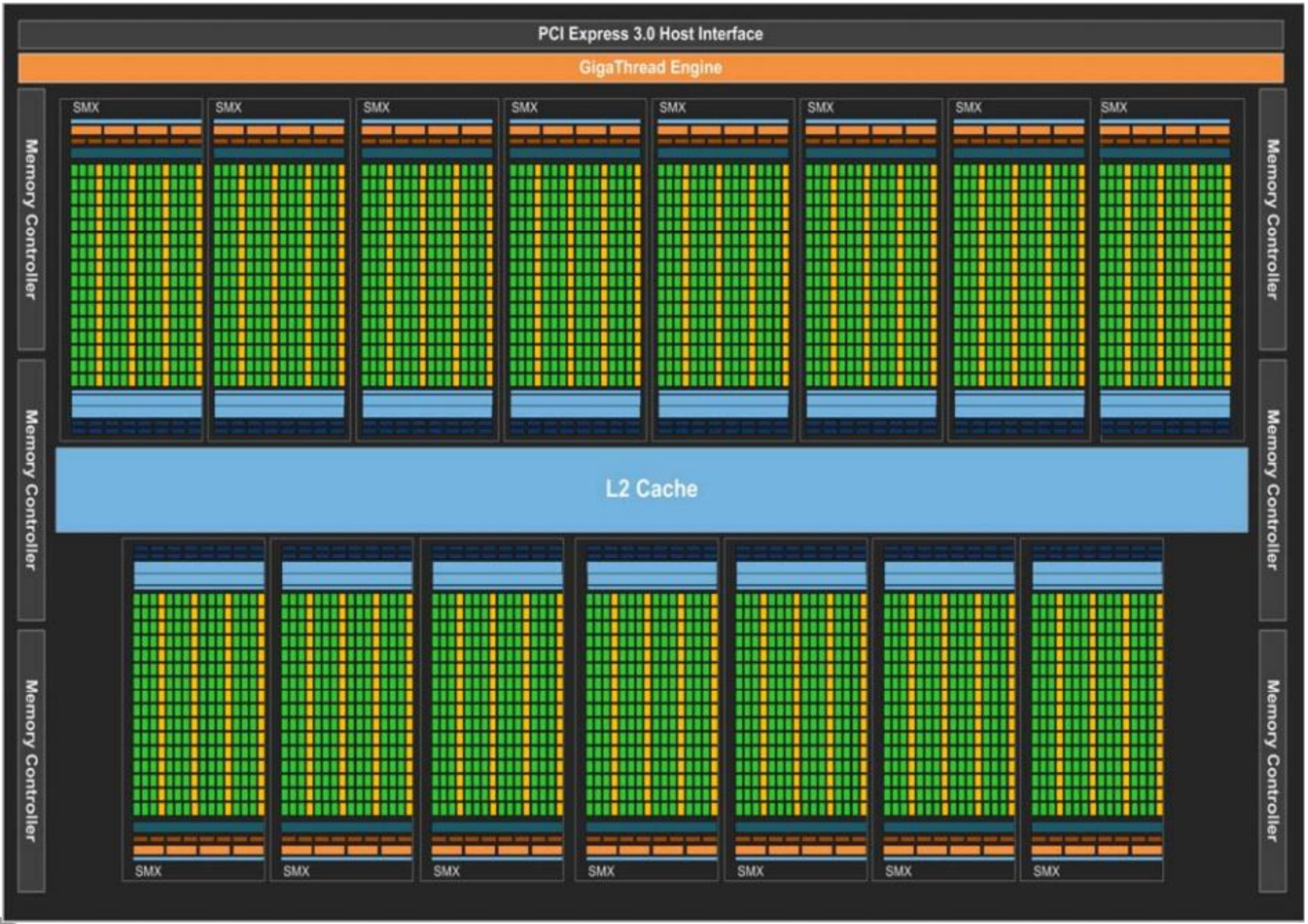
Single thread perf.

throughput

Physical integration

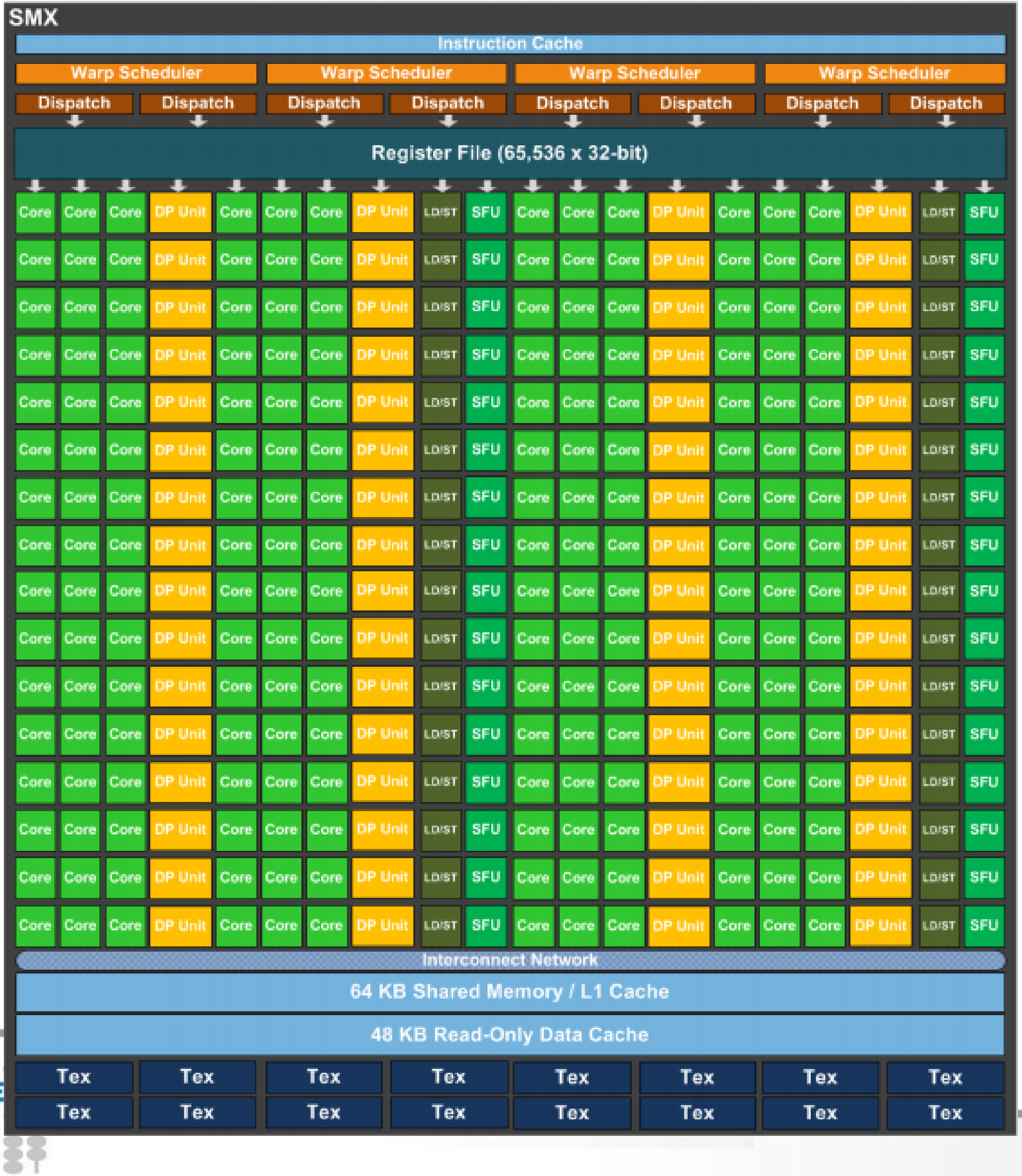
Architectural integration

K20 nVIDIA GPU



15 SMX Streaming Multiprocessors

SMX



192 single precision cuda cores

64 double precision units

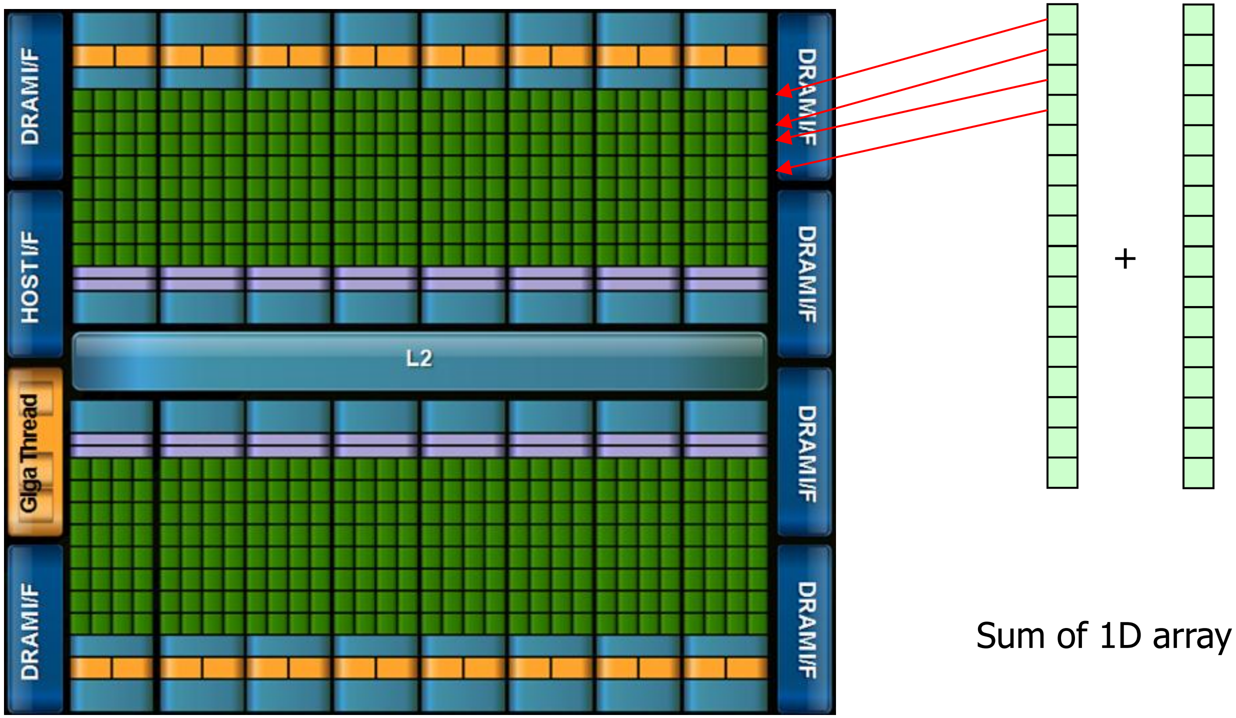
32 special function units

32 load and store units

4 warp scheduler
(each warp contains 32 parallel
Threads)

2 indipendent instruction per warp

Accelerator/GPGPU



Sum of 1D array

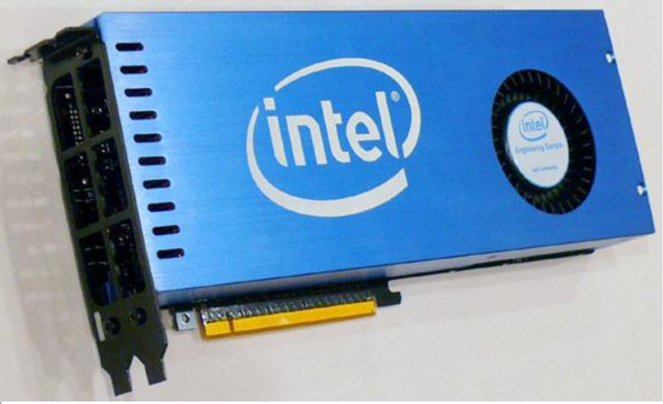
CUDA sample

```
 void  CPUCode( int* input1, int* input2, int* output, int length) {
                for ( int  i = 0; i < length; ++i ) {
                      output[ i ] = input1[ i ] + input2[ i ];
               }
 }
```

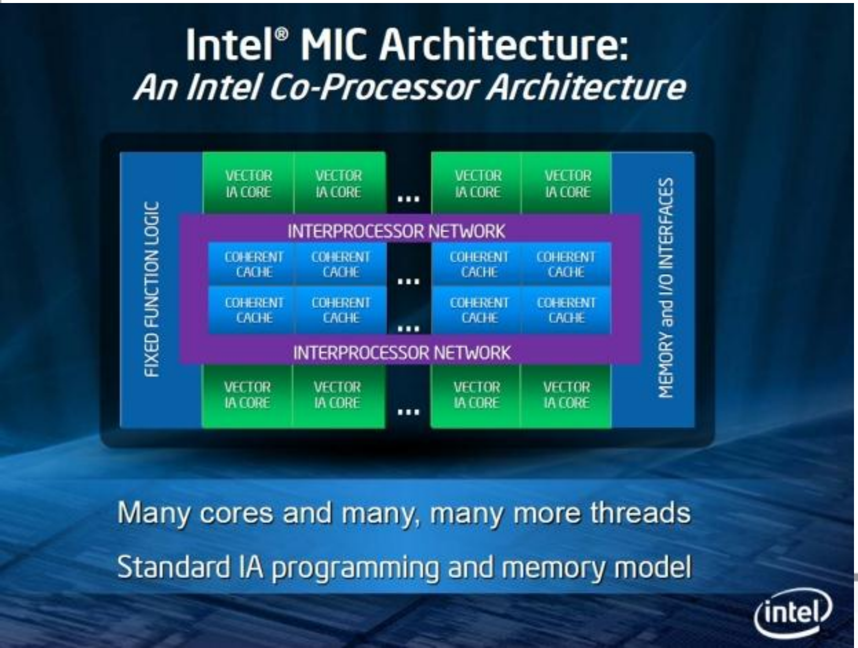
```
__global__void  GPUCode( int* input1, int*input2, int* output, int length) {
               int idx = blockDim.x * blockIdx.x + threadIdx.x;
                if ( idx < length ) {
                      output[ idx ] = input1[ idx ] + input2[ idx ];
               }
}
```

Each thread execute one loop iteration

Intel MIC



Up to 61 Intel® Architecture cores
1.1 GHz
244 threads
Up to 8 GB memory
up to 352 GB/s bandwidth
512-bit SIMD instructions
Linux* operating system, IP addressable
Standard programming languages and tools
Over 1 TeraFlop/s double precision peak performance

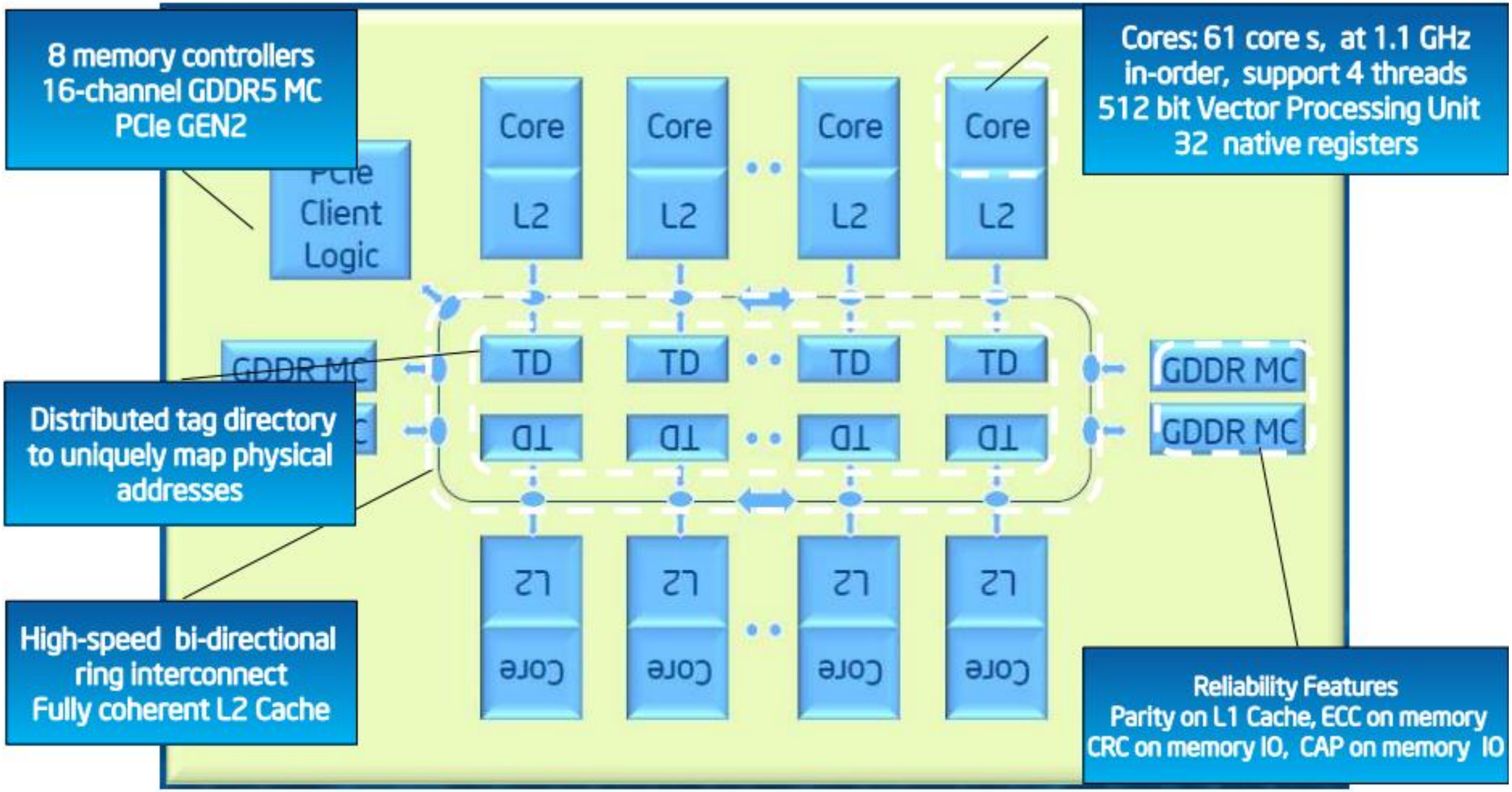


Intel® MIC Architecture:
An Intel Co-Processor Architecture

Many cores and many, many more threads

Standard IA programming and memory model

MIC Architecture



8 memory controllers
16-channel GDDR5 MC
PCIe GEN2

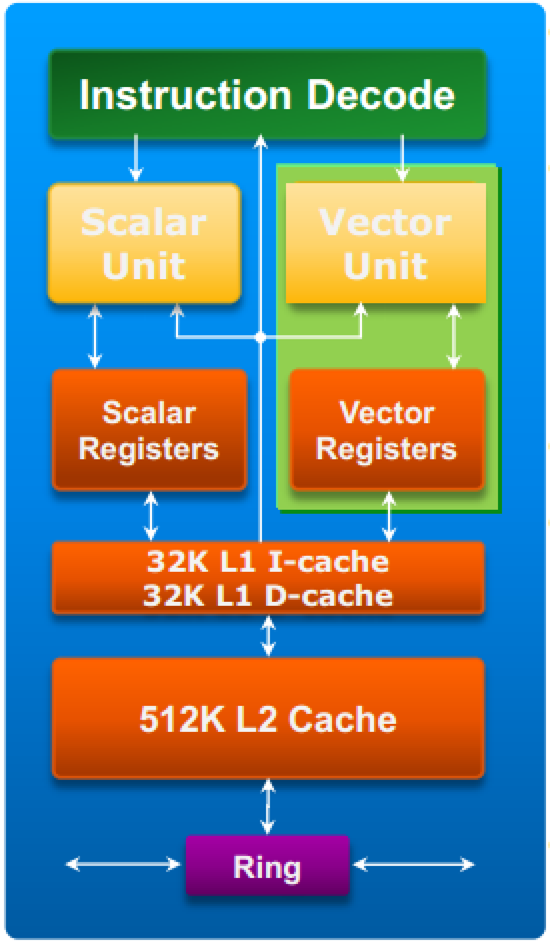
Cores: 61 core s, at 1.1 GHz
in-order, support 4 threads
512 bit Vector Processing Unit
32 native registers

Distributed tag directory to uniquely map physical addresses

High-speed bi-directional ring interconnect
Fully coherent L2 Cache

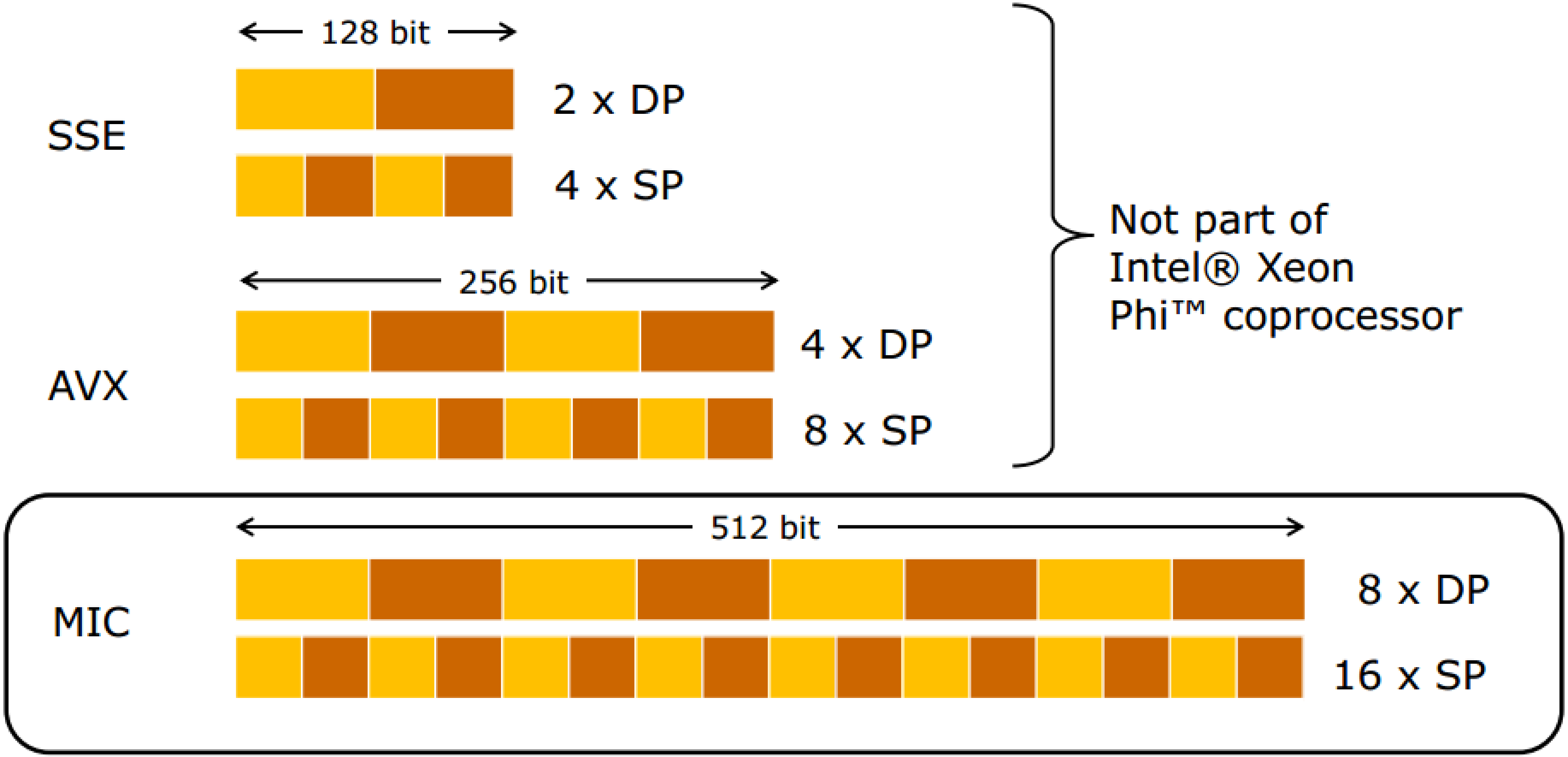
Reliability Features
Parity on L1 Cache, ECC on memory
CRC on memory IO, CAP on memory IO

Core Architecture



- 60+ in-order, low-power Intel® Architecture cores in a ring interconnect
- Two pipelines
  - Scalar Unit based on Pentium® processors
  - Dual issue with scalar instructions
  - Pipelined one-per-clock scalar throughput
- SIMD Vector Processing Engine
- 4 hardware threads per core
  - 4 clock latency, hidden by round-robin scheduling of threads
  - Cannot issue back-to-back inst in same thread
- Coherent 512 KB L2 Cache per core

Intel Vector Units



Not part of Intel® Xeon Phi™ coprocessor