Energy efficiency and roadmap to exascale

Carlo Cavazzoni

outline

- Roadmap to Exascale
- HPC architecture challanges
- Energy efficiency
- Co processor architecture
- I/O revolution

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

CINECA

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!

$L' = L / 2$

$V' = \sim V$

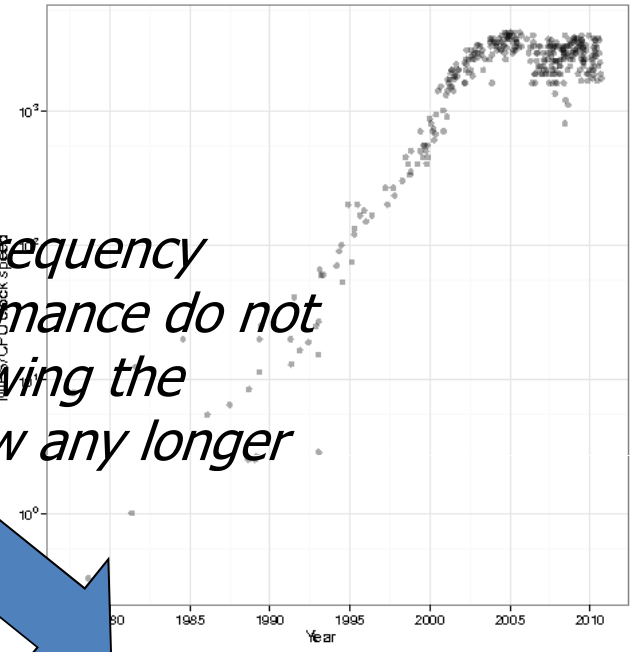
$F' = \sim F * 2$

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

$P' = 4 * P$

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!

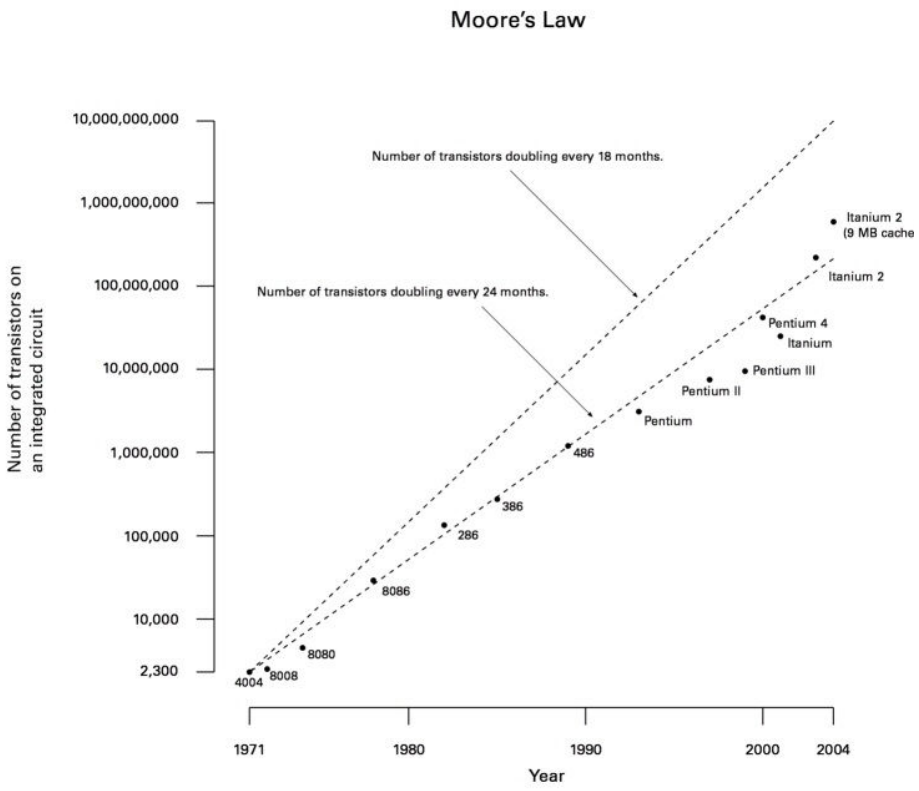
- Now, power and/or heat generation are the limiting factors of the down-scaling
- Supply voltage reduction is becoming difficult, because Vth cannot be decreased any more, as described later.
- Growth rate in clock frequency and chip area becomes smaller.

PRACE CINECA

Moore's Law

Number of transistors per chip double every 18 month

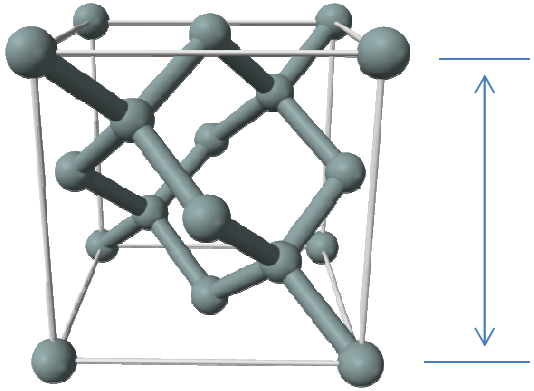
The true it double every 24 month



- Oh-oh! Huston!

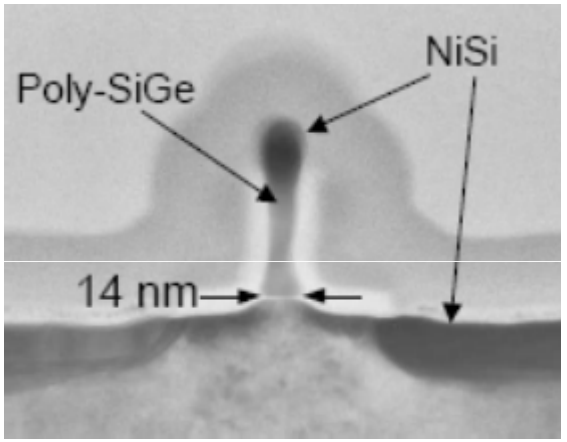
PRACE CINECA

The silicon lattice



0.54 nm

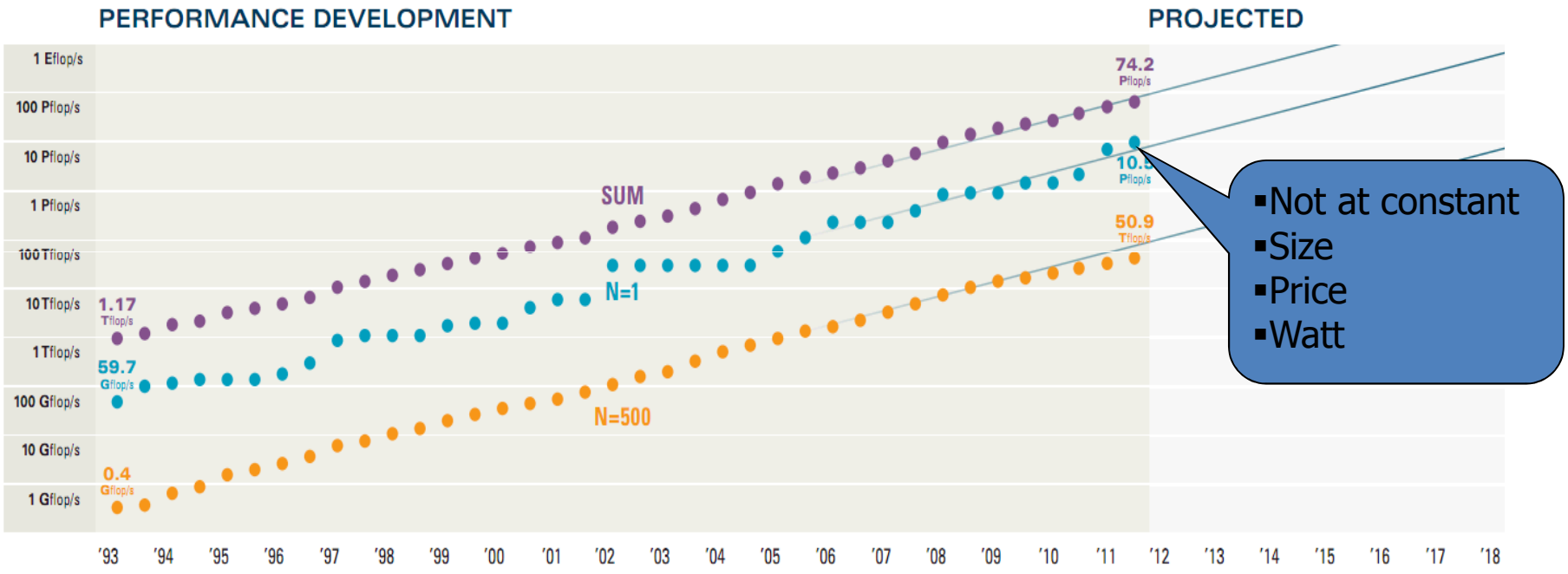
Si lattice



50 atoms!

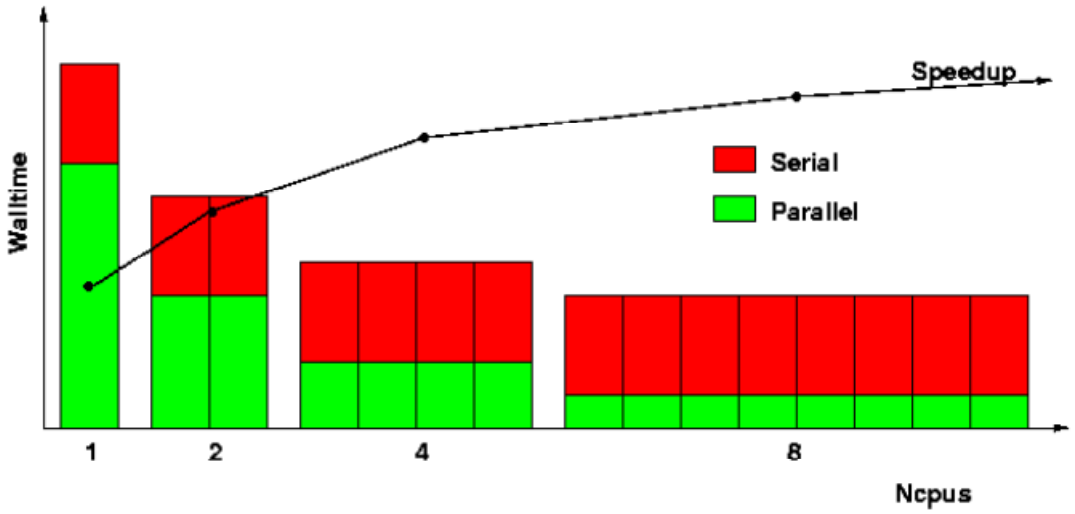
There will be still 4~6 cycles (or technology generations) left until we reach 11 ~ 5.5 nm technologies, at which we will reach downscaling limit, in some year between 2020-30 (H. Iwai, IWJT2008).

PRACE CINECA


PERFORMANCE DEVELOPMENT
PROJECTED
1 Eflop/s
100 Pflop/s
10 Pflop/s
1 Pflop/s
100 Tflop/s
10 Tflop/s
1 Tflop/s
100 Gflop/s
10 Gflop/s
1 Gflop/s
SUM
N=1
N=500
1.17
Tflop/s
59.7
Gflop/s
0.4
Gflop/s
74.2
Pflop/s
50.9
Tflop/s
'93
'94
'95
'96
'97
'98
'99
'00
'01
'02
'03
'04
'05
'06
'07
'08
'09
'10
'11
'12
'13
'14
'15
'16
'17
'18
▪Not at constant
▪Size
▪Price
▪Watt

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).



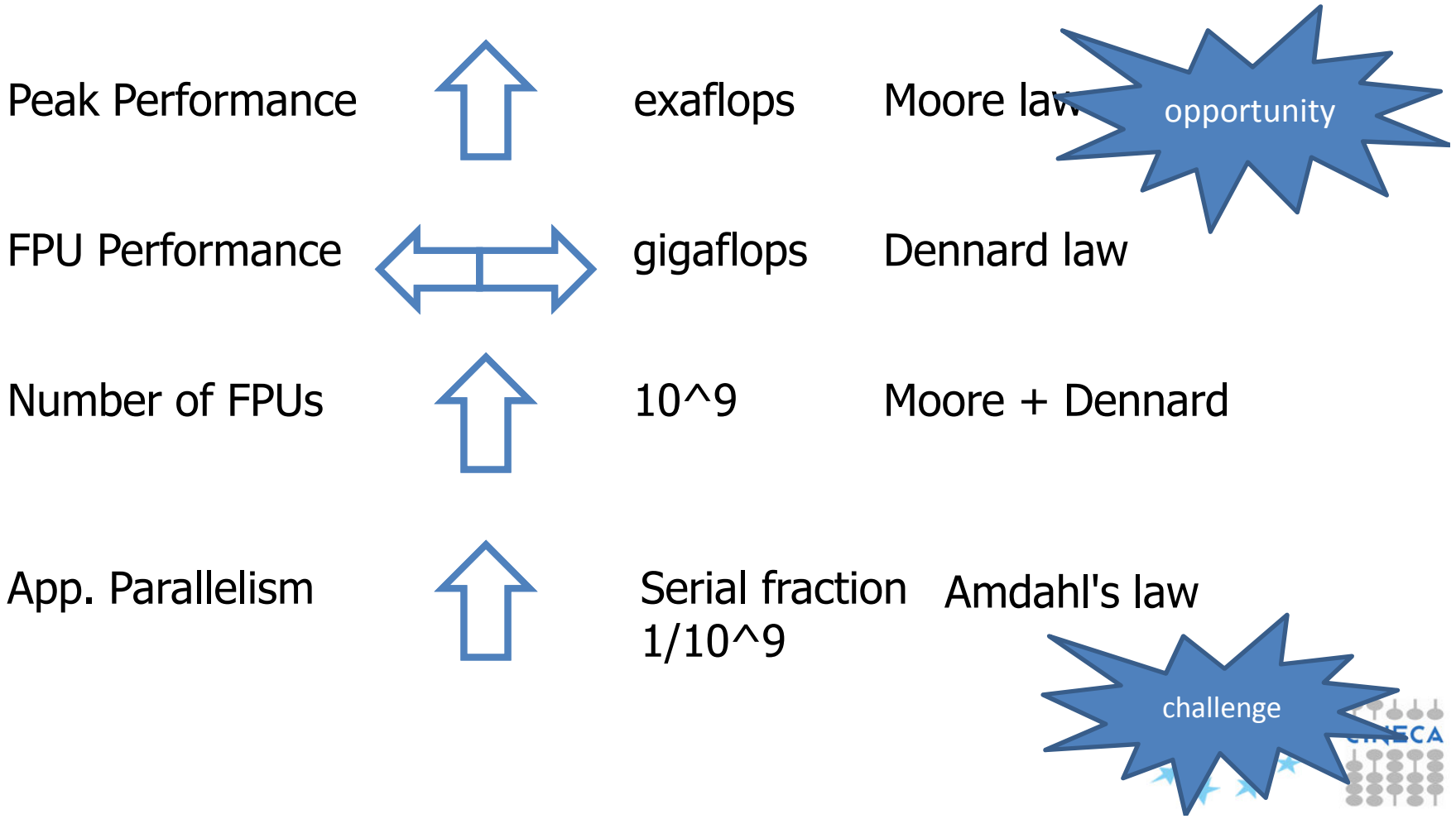
maximum speedup tends to
$1 / (1 - P)$
P= parallel fraction

1000000 core

$P = 0.999999$

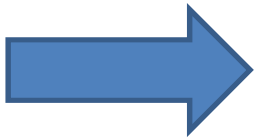
serial fraction= 0.000001

HPC trends
(constrained by the three law)

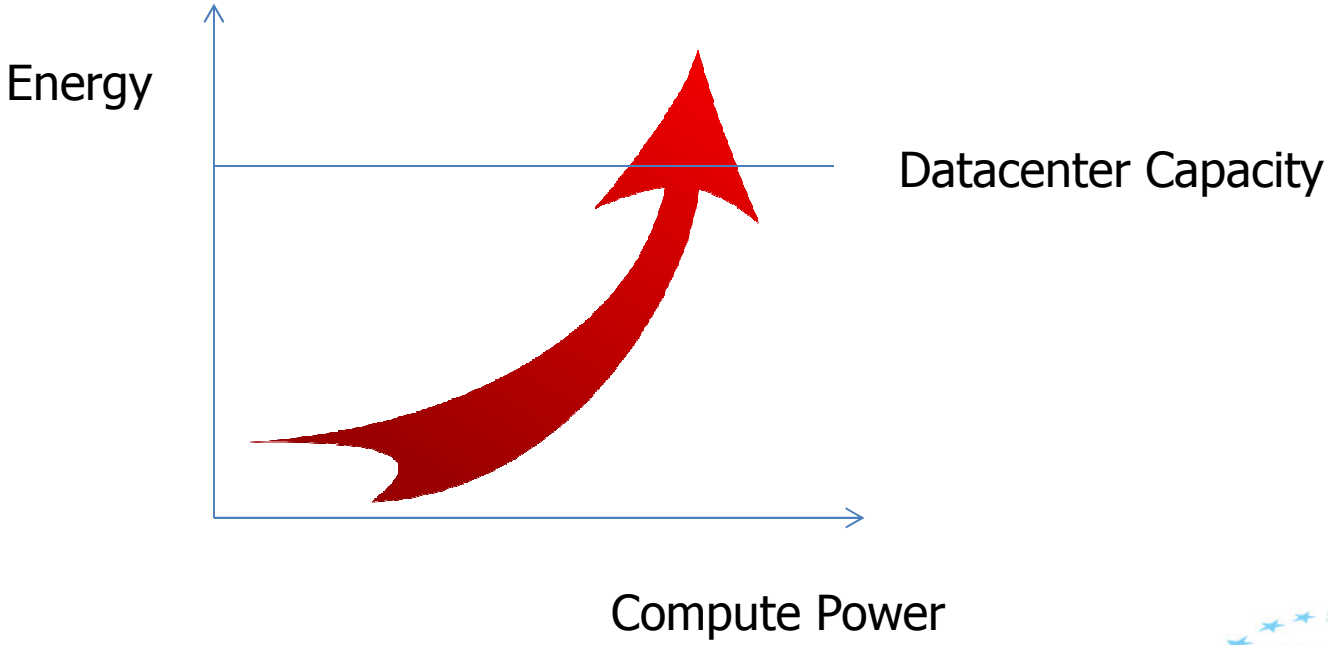


Energy trends

"traditional" RISK and CISC chips are designed for maximum performance for all possible workloads



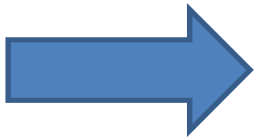
A lot of silicon to maximize single thread performace



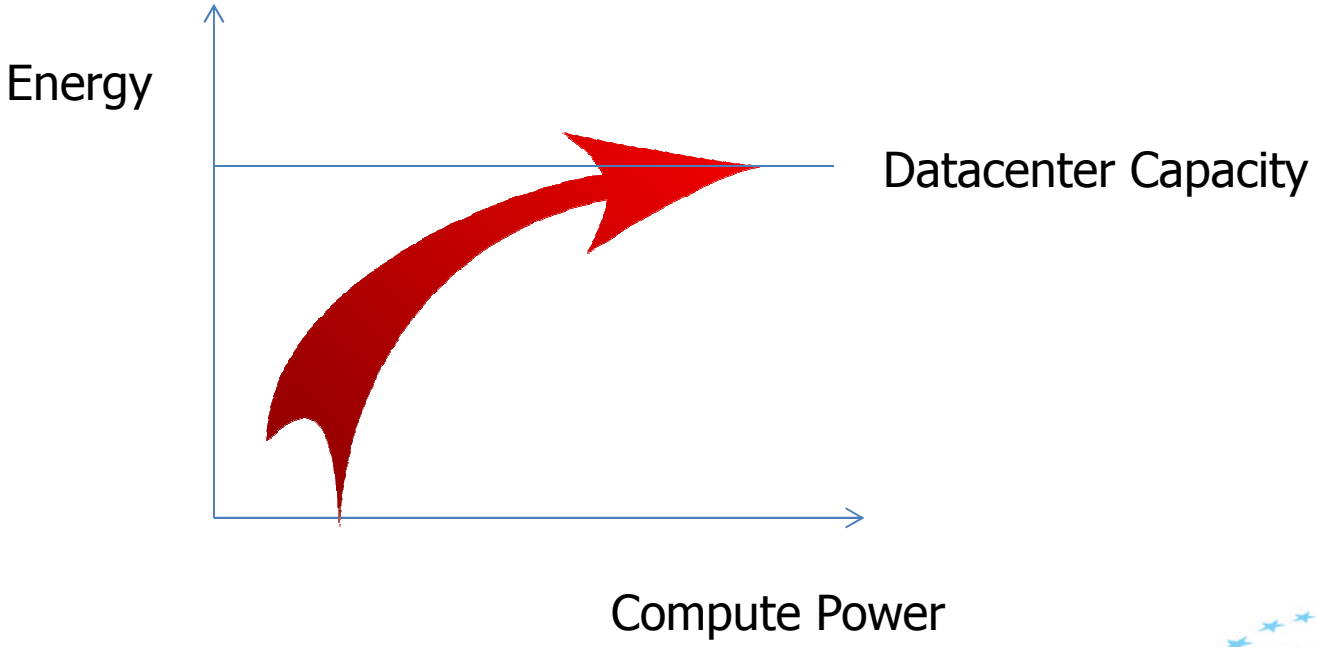
PRACE CINECA

Change of paradigm

New chips designed for maximum performance in a small set of workloads



Simple functional units, poor single thread performance, but maximum throughput



Exascale architecture

two model

- Hybrid
- Homogeneus

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

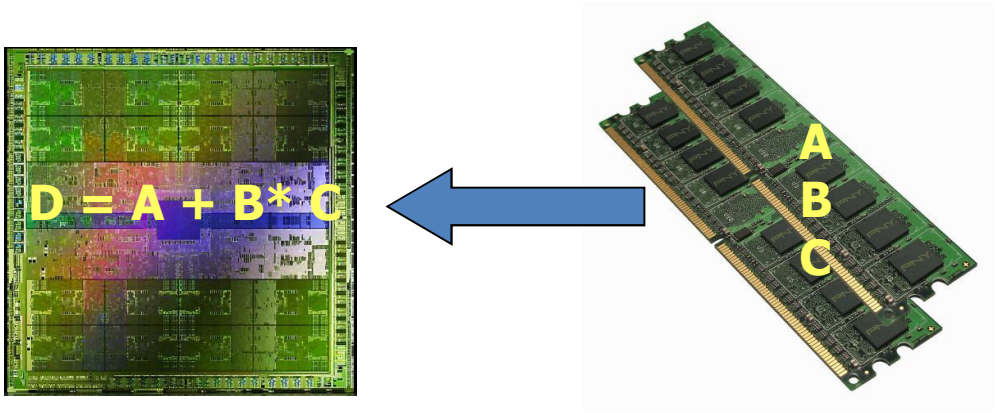
Where power is used:

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

Short term impact on programming models

Memory

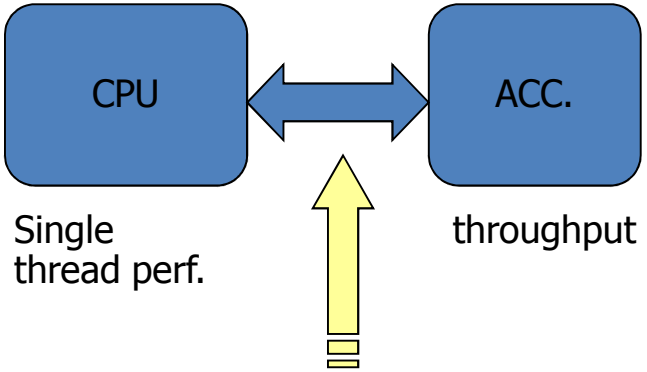
Today the cost of moving operands to compute a 64bit floating-point FMA takes more energy with respect to the FMA operation itself



at 10nm integration, the energy required to move date is expected to becomes 100x !

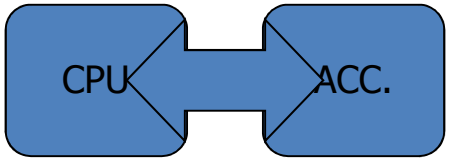
We need locality! → Less “fast” memory per core

Architecture toward exascale

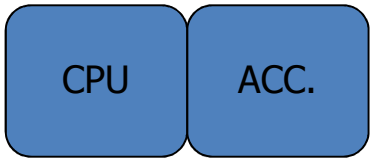


bottleneck

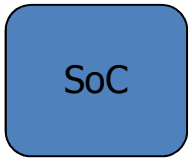
GPU/MIC/FPGA



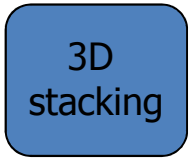
OpenPower
Nvidia GPU



AMD APU
ARM Big-Little



ARM
KNL (next Intel PHI)



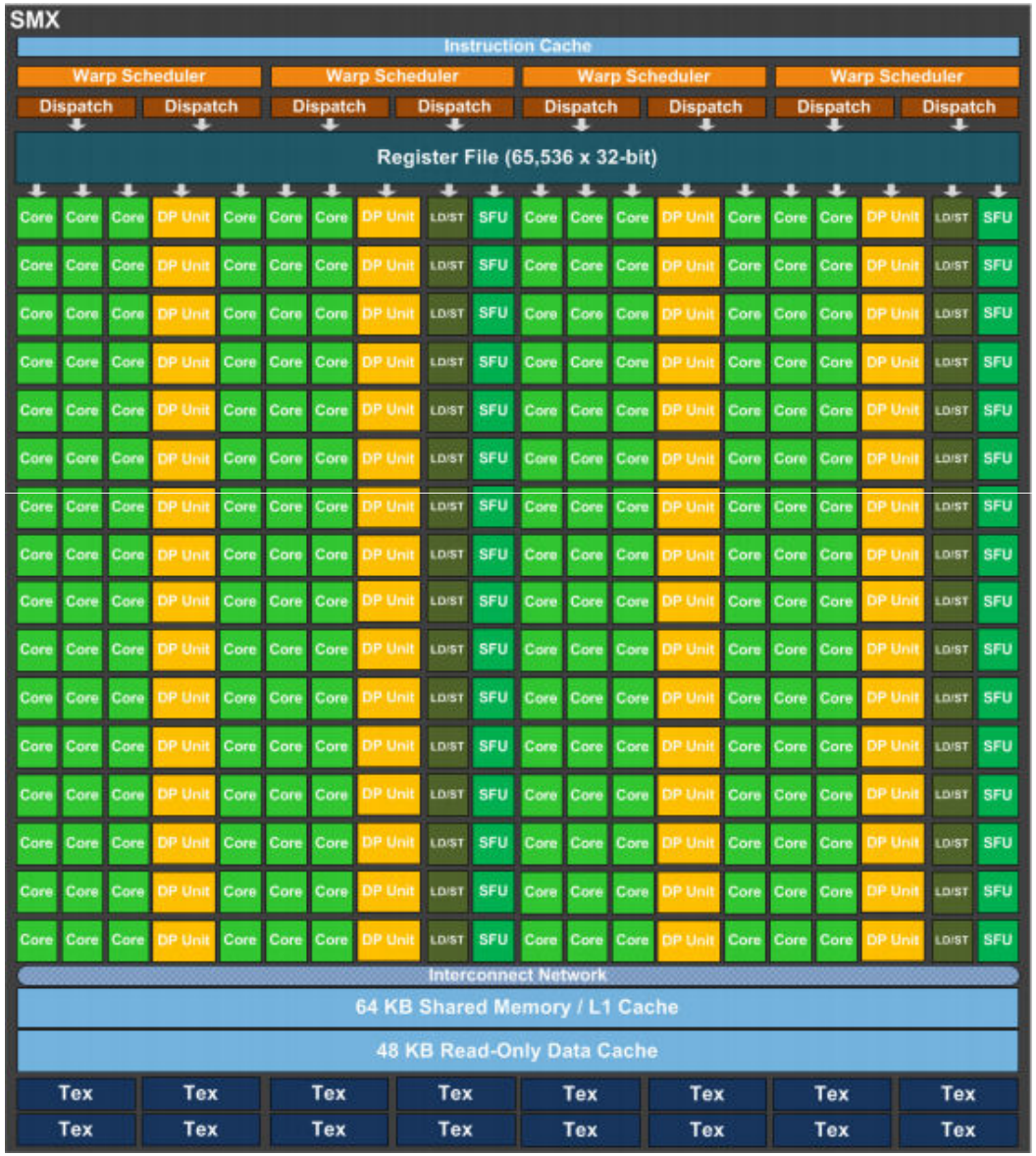
Active memory

Photonic -> platform flexibility
TSV -> stacking

K20 nVIDIA GPU

15 SMX Streaming Multiprocessors

SMX



192 single precision cuda cores

64 double precision units

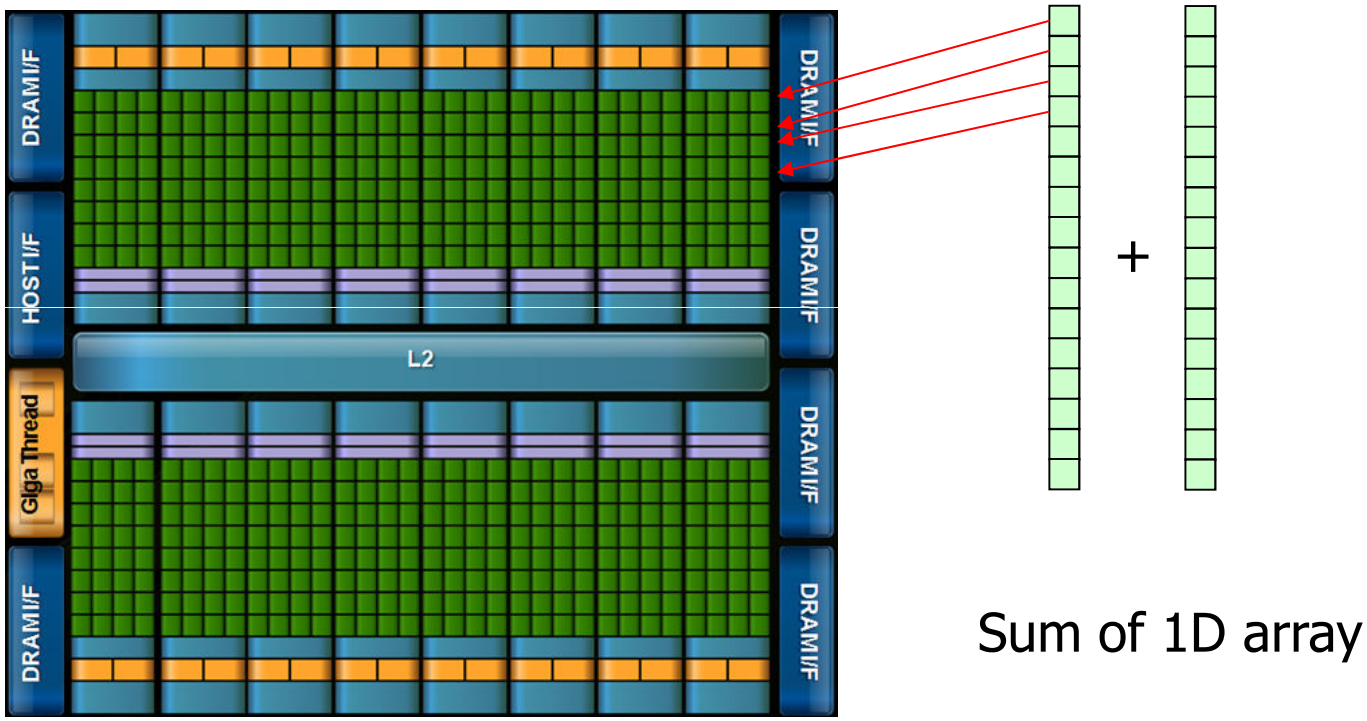
32 special function units

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

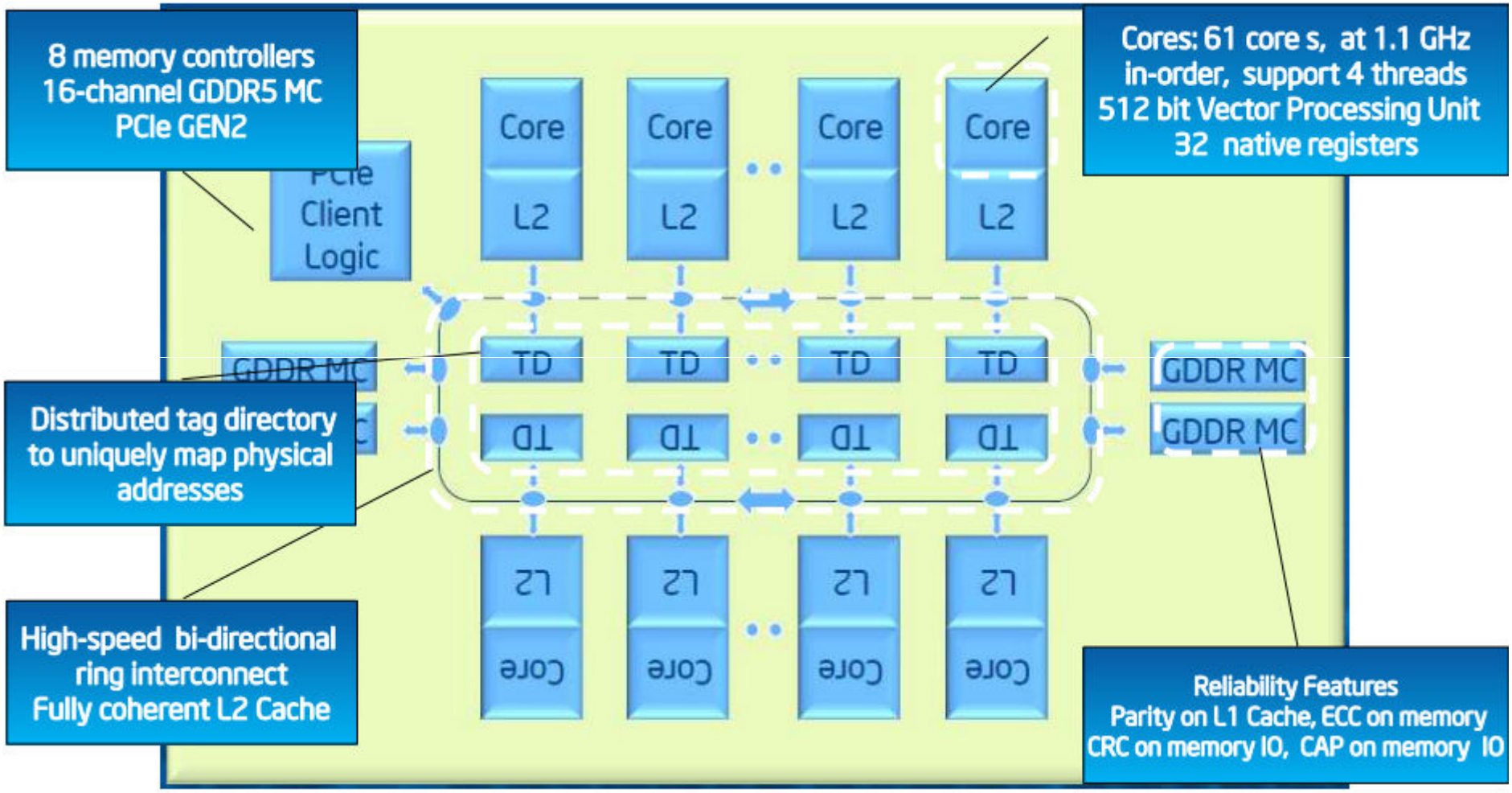
CUDA sample

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

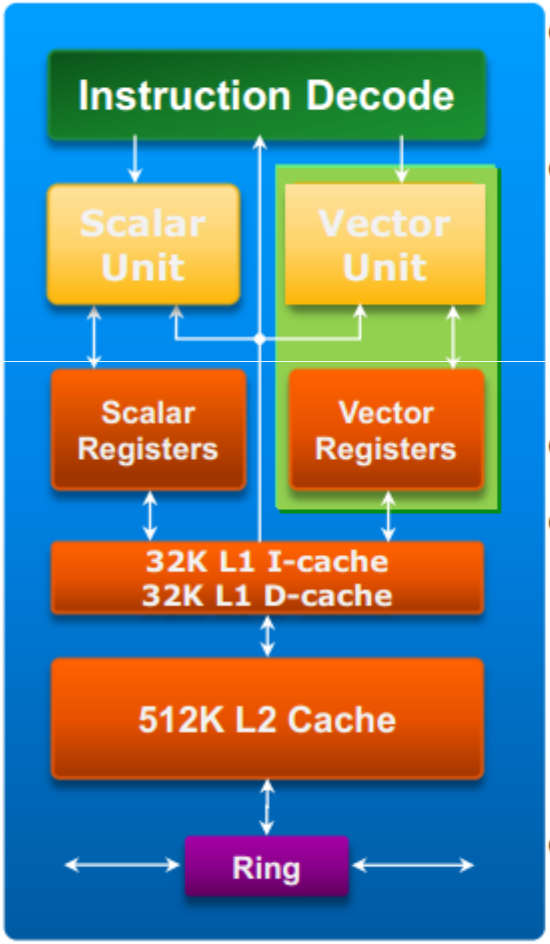
```
__global__void  GPUCode( int* input1, int*input2, int* output, int length) {
               int idx = blockDim.x * blockIdx.x + threadIdx.x;
                if ( idx < length ) {
                      output[ idx ] = input1[ idx ] + input2[ idx ];
               }
}
```

Each thread execute one loop iteration

Intel Xeon PHI Architecture



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

Knights Landing is the codename for Intel's 2nd generation Intel® Xeon Phi™ Product Family, which will deliver massive thread parallelism, data parallelism and memory bandwidth – with improved single-thread performance and Intel® Xeon® processor binary-compatibility in a standard CPU form factor. Additionally, Knights Landing will offer integrated Intel® Omni-Path fabric technology, and also be available in the traditional PCIe* coprocessor form factor.

The following is a list of public disclosures that Intel has previously made about the forthcoming product:

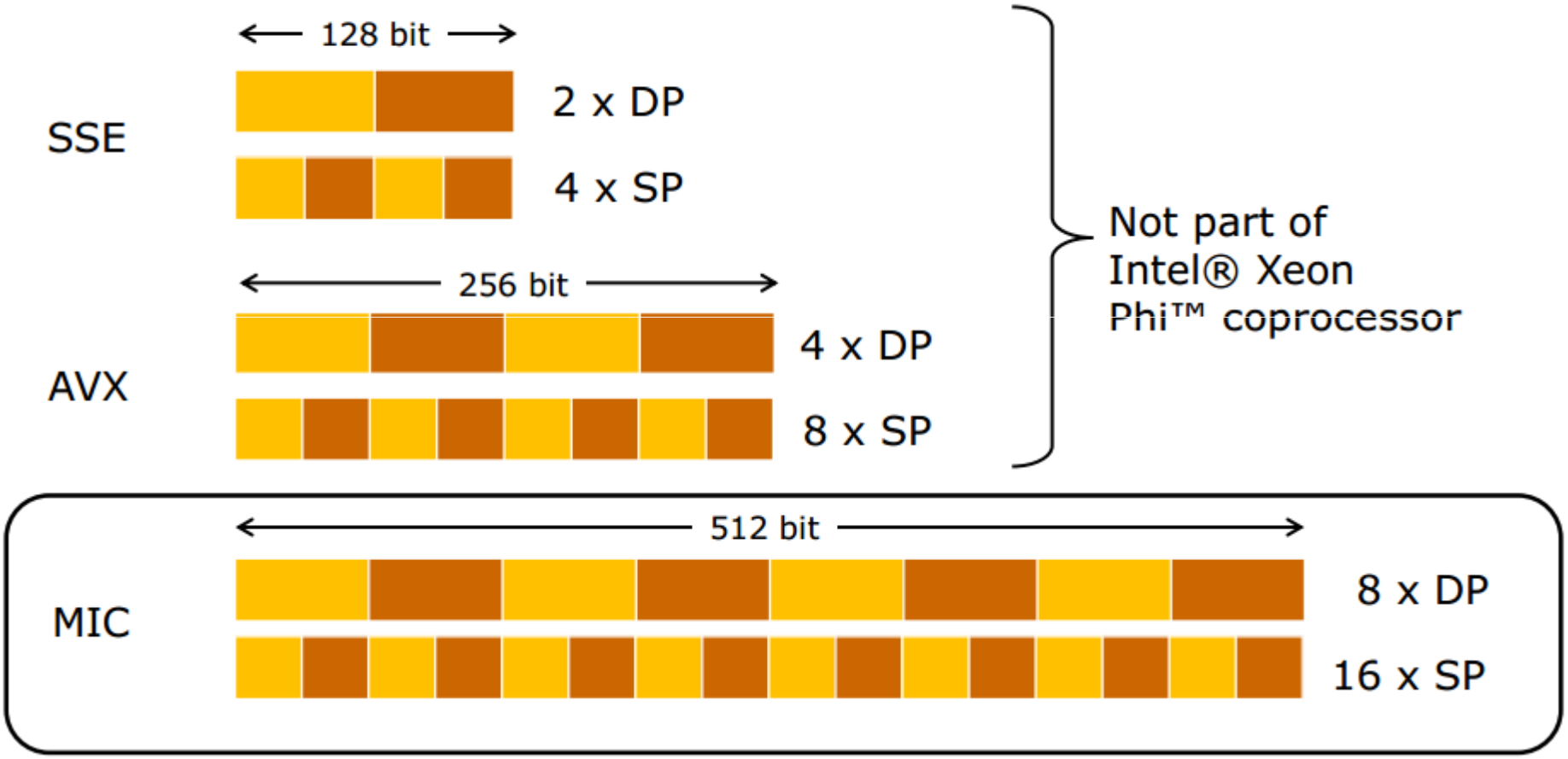
PERFORMANCE

3+ TeraFLOPS of double-precision peak theoretical performance per single socket node[0]

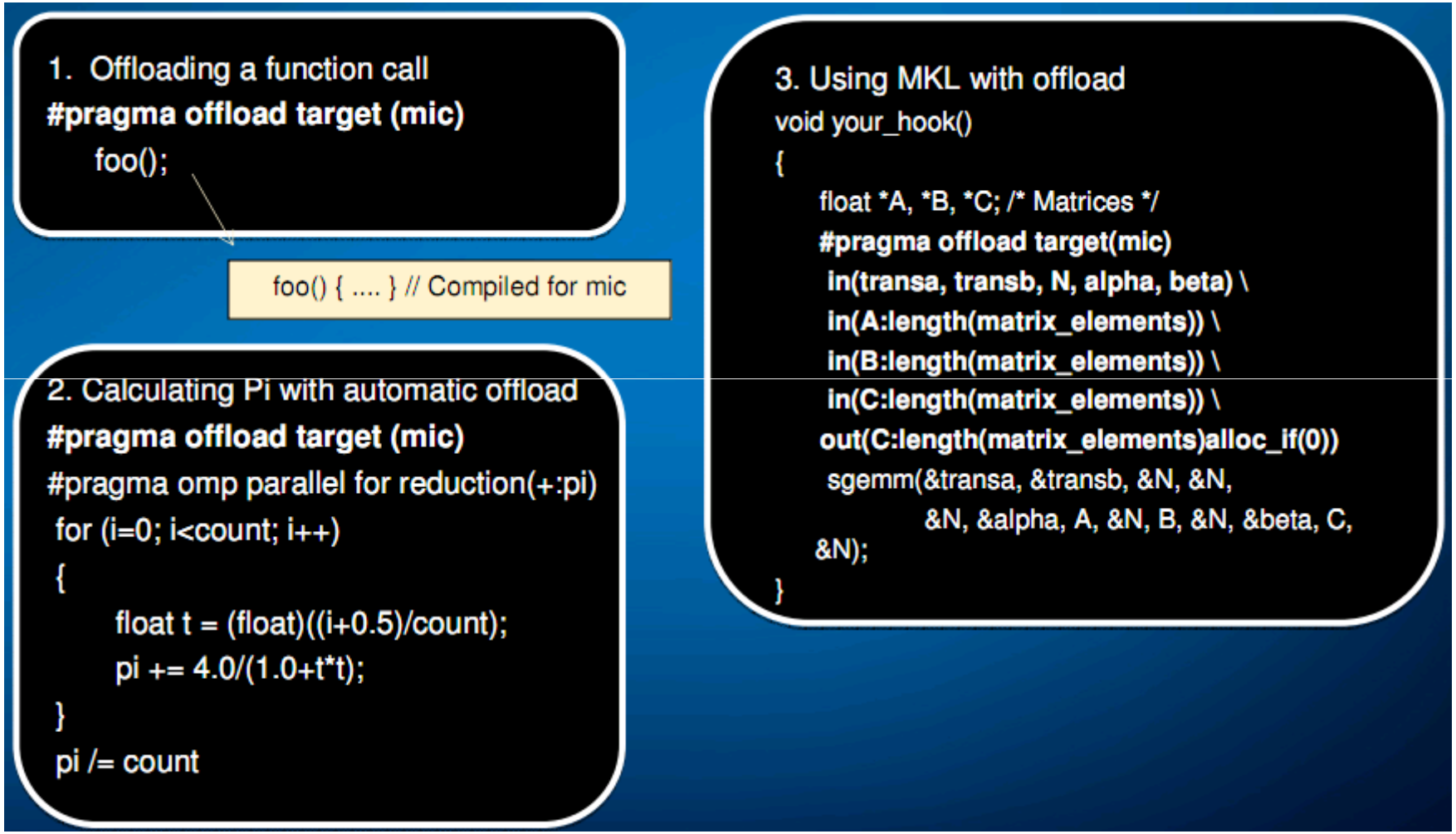
High-performance on-package memory (MCDRAM)	Over 5x STREAM vs. DDR4[1] ⇒ Over 400 GB/s
	Up to 16GB at launch
	NUMA support
	Over 5x Energy Efficiency vs. GDDR5[2]
	Over 3x Density vs. GDDR5[2]
	In partnership with Micron Technology
	Flexible memory modes including cache and flat

https://software.intel.com/en-us/articles/what-disclosures-has-intel-made-about-knights-landing?utm_content=buffer9926a&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Intel Vector Units



Programming MIC

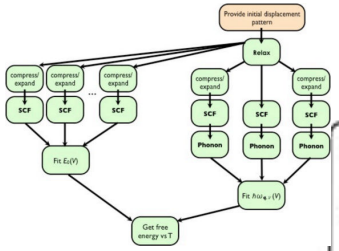


Applications Challenges

- Programming model
- Scalability
- I/O, Resiliency/Fault tolerance
- Numerical stability
- Algorithms
- Energy Awareness/Efficiency

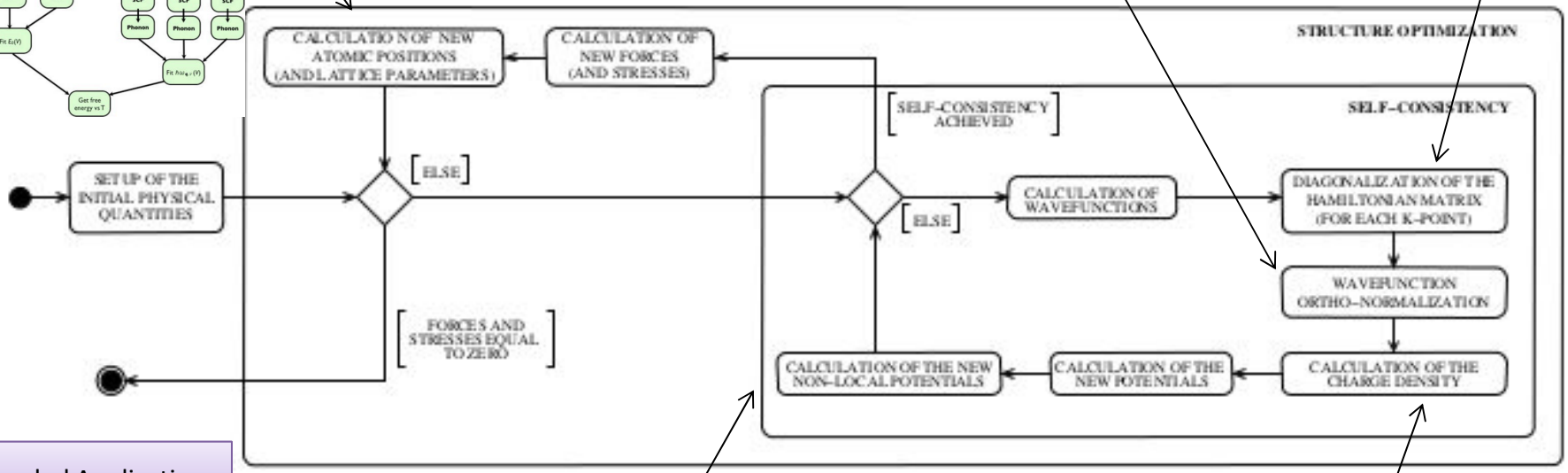
Quantum Espresso toward exascale

High Throughput / Ensamble Simulations



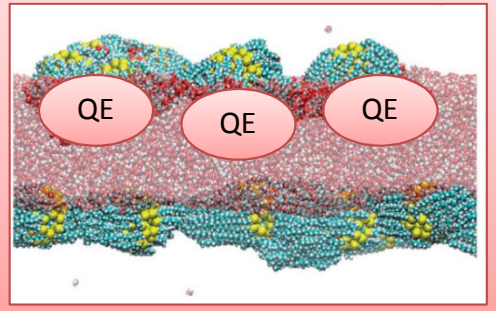
Communication avoiding

New Algorithm: CG vs Davidson



Coupled Application

LAMMPS



Task level parallelism

Double buffering

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

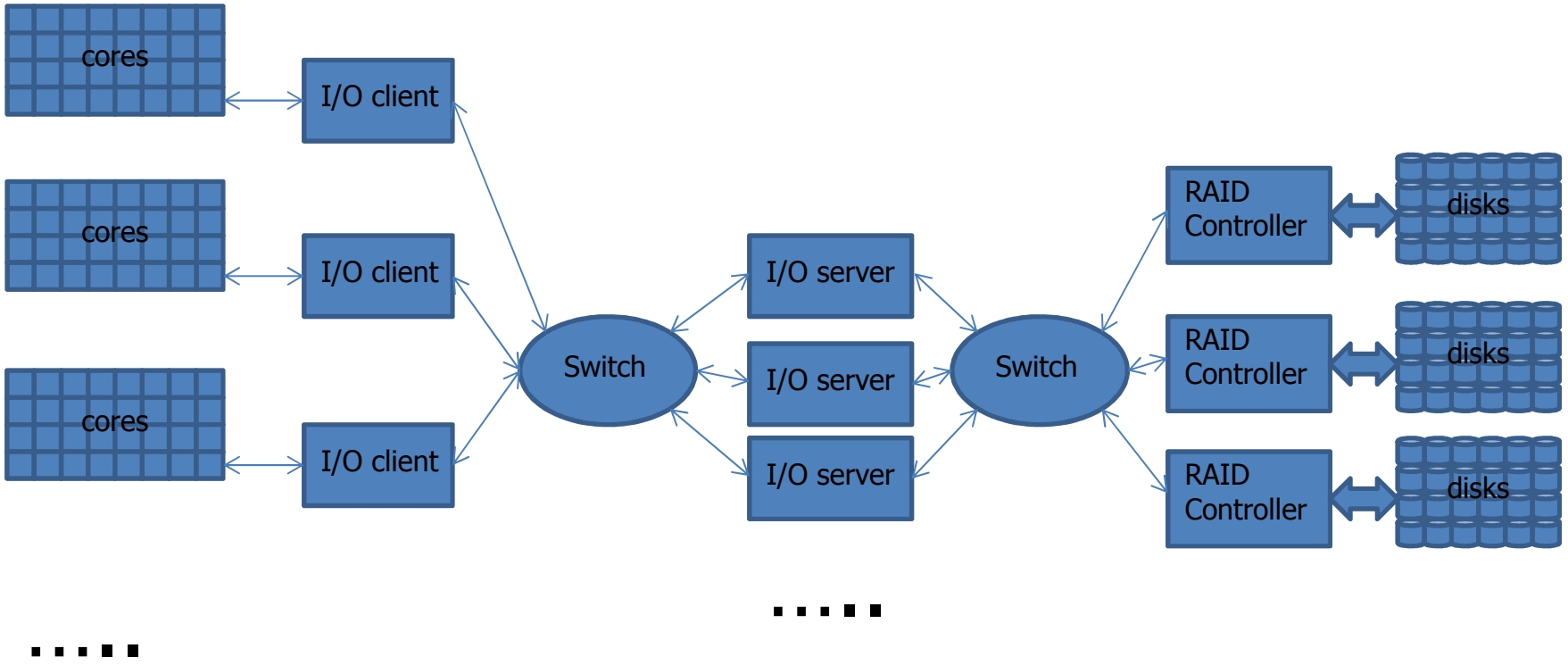
I/O Subsystem

I/O subsystem of high performance computers are still deployed using spinning disks, with their mechanical limitation (spinning speed cannot grow above a certain regime, above which the vibration cannot be controlled), and like for the DRAM they eat energy even if their state is not changed. Solid state technology appear to be a possible alternative, but costs do not allow to implement data storage systems of the same size. Probably some hierarchical solutions can exploit both technology, but this do not solve the problem of having spinning disks spinning for nothing.

I/O Challenges

Today	Tomorrow
100 clients	10K clients
1000 core per client	100K core per clients
3PByte	1Exabyte
3K Disks	100K Disks
100 Gbyte/sec	100TByte/sec
8MByte blocks	1Gbyte blocks
Parallel Filesystem	Parallel Filesystem
One Tier architecture	Multi Tier architecture

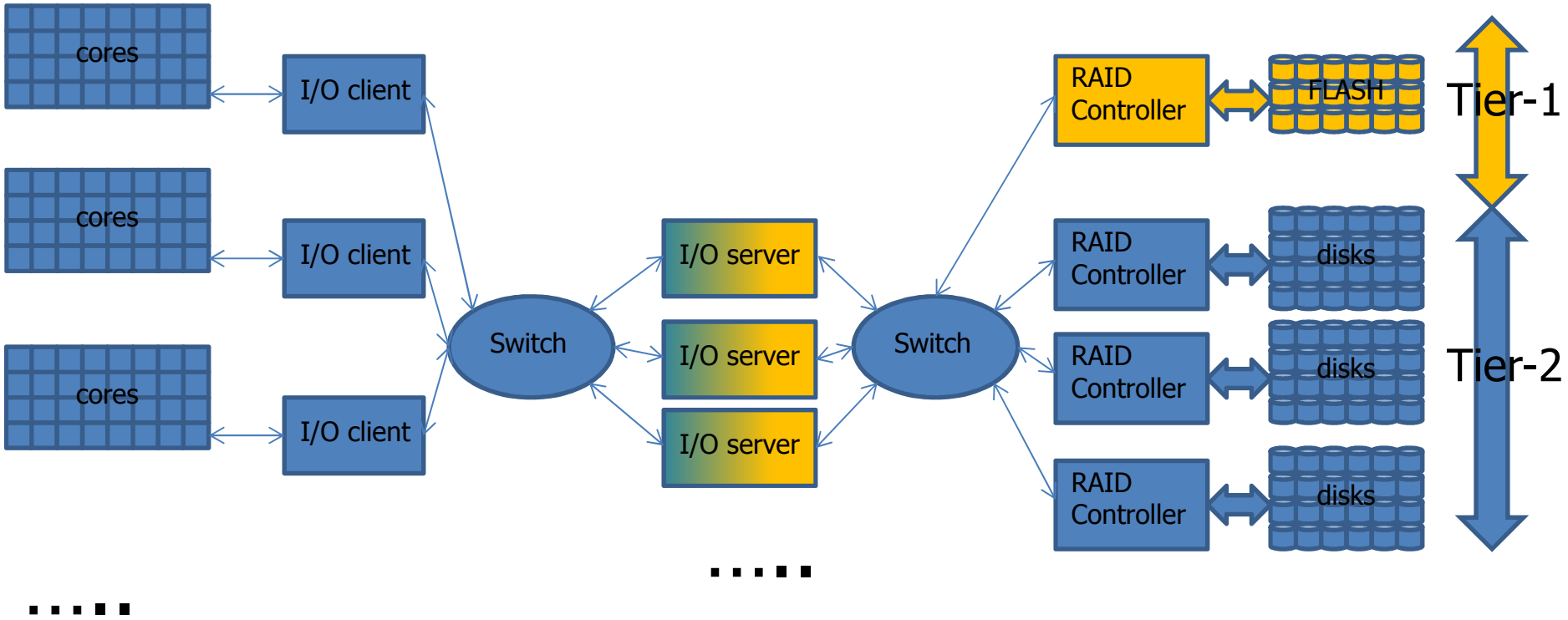
Today



160K cores, 96 I/O clients, 24 I/O servers, 3 RAID controllers

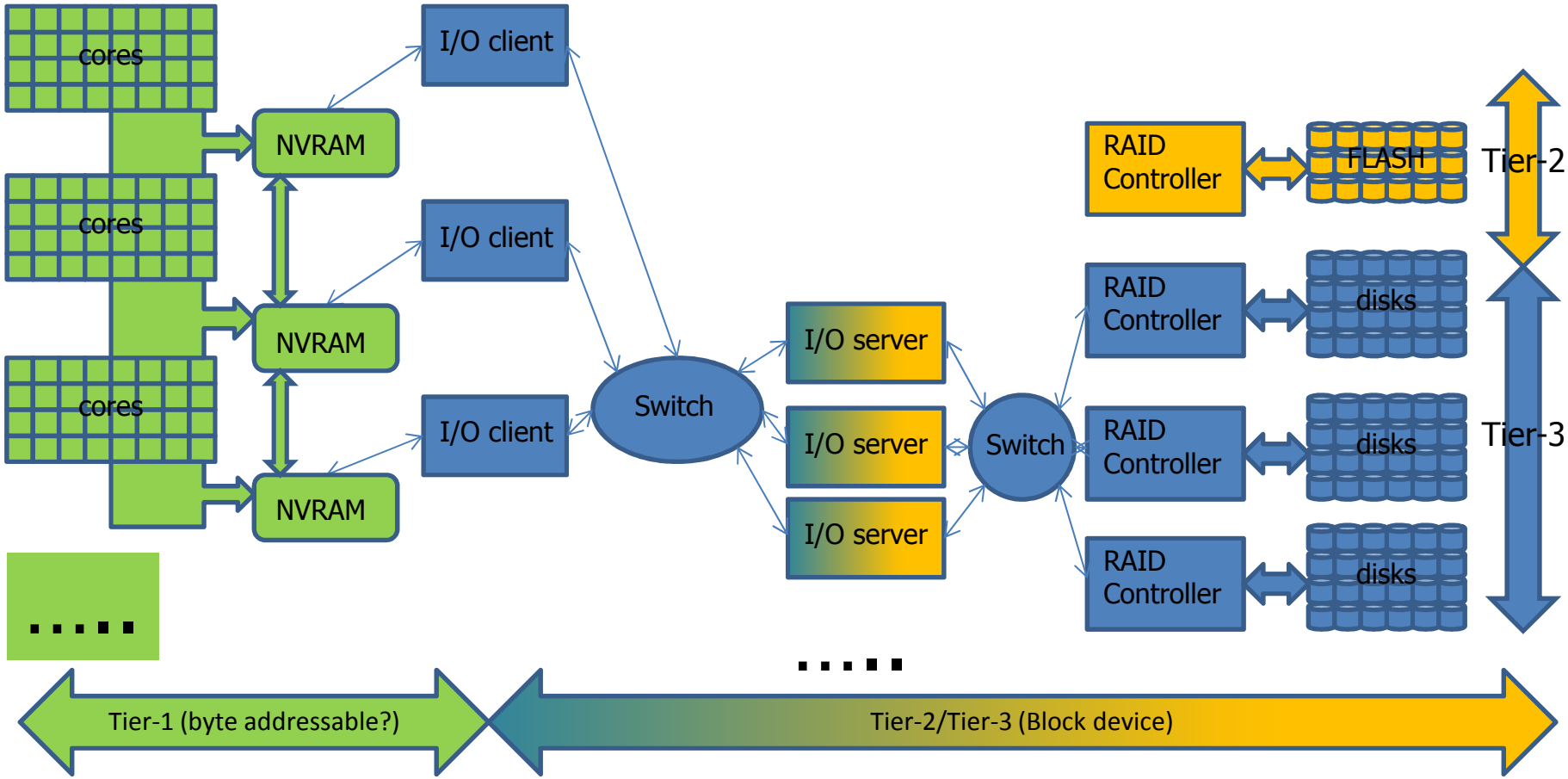
IMPORTANT: I/O subsystem has its own parallelism!

Today-Tomorrow



1M cores, 1000 I/O clients, 100 I/O servers, 10 RAID FLASH/DISK controllers

Tomorrow



1G cores, 10K NVRAM nodes, 1000 I/O clients, 100 I/O servers, 10 RAID controllers

Impact on programming and execution models

DATA:
- Billion of (application) files
- Large (check-point/restart) file

Posix Filesystem:
- low level
- lock/syncronization -> transactional IOP
- low IOPs (I/O operation per second)

Physical supports:
- disk too slow -> archive
- FLASH aging problem
- NVRAM (Non-Volatile RAM), PCM (Phase Change Memory), not ready

Middlewere:
- Library HDF5, NetCDF
- MPI-I/O
- Each layer has its own semantics

Conclusions

- Exascale Systems, will be there
- Power is the main architectural constraints
- Exascale Applications?
- Yes, but…
- Concurrency, Fault Tolerance, I/O …
- Energy awareness