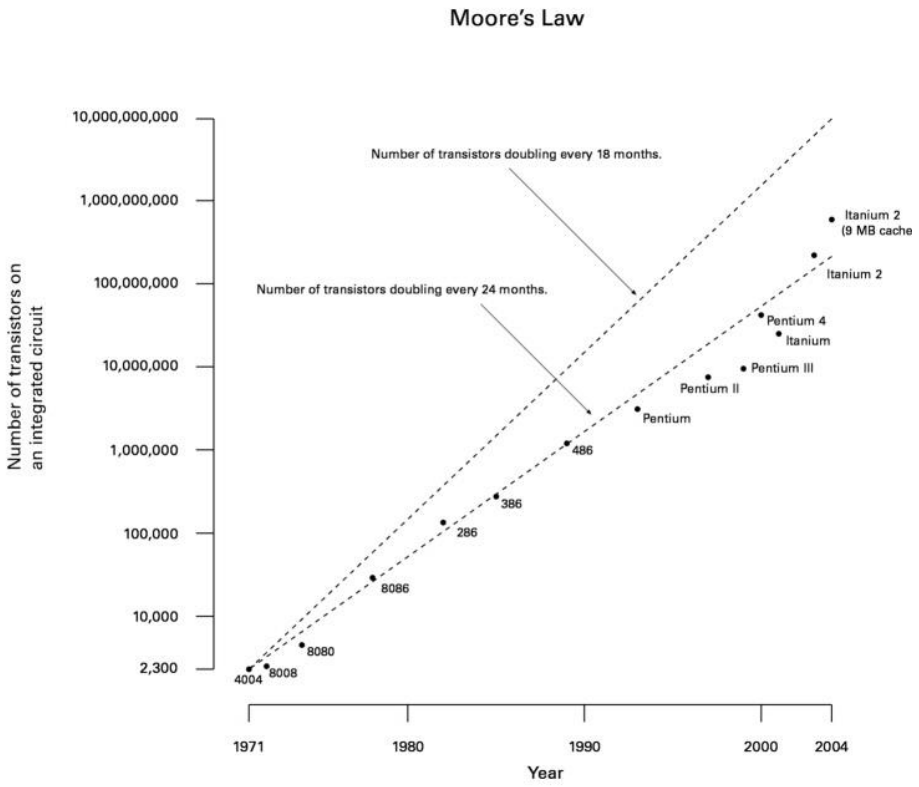
HPC Architecture Trends

Carlo Cavazzoni

Moore's Law

Number of transistors per chip double every 18 month

The true it double every 24 month



- Oh-oh! Huston!

Dennard scaling law (downscaling)

new VLSI gen.

old VLSI gen.

$L' = L / 2$

$V' = V / 2$

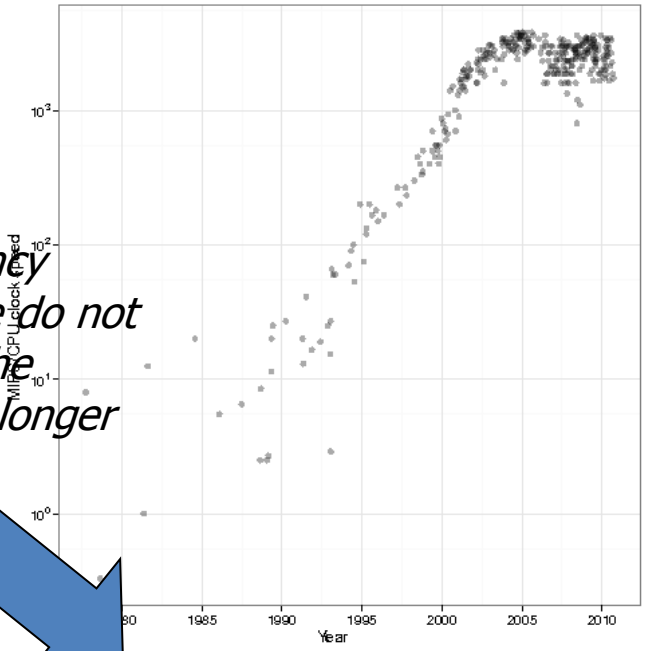
$F' = F * 2$

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

$P' = P$

do not hold anymore!

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



$L' = L / 2$

$V' = \sim V$

$F' = \sim F * 2$

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

$P' = 4 * P$

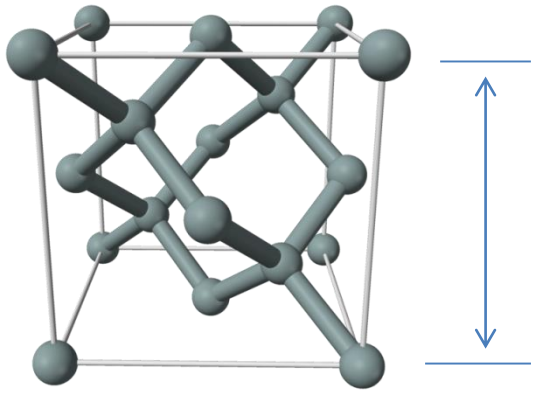
Increase the number of cores to maintain the architectures evolution on the Moore's law

- 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.

The power crisis!

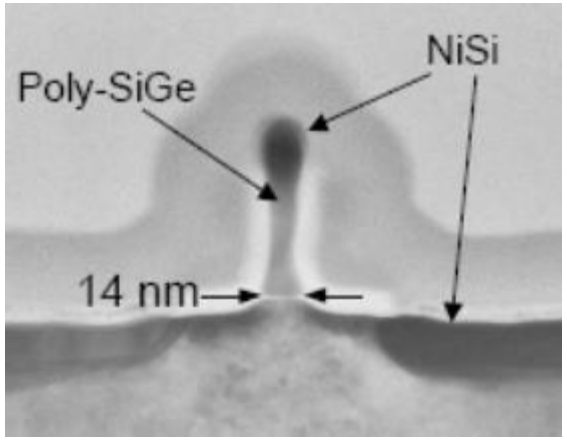
Programming crisis!

The silicon lattice



0.54 nm

Si lattice

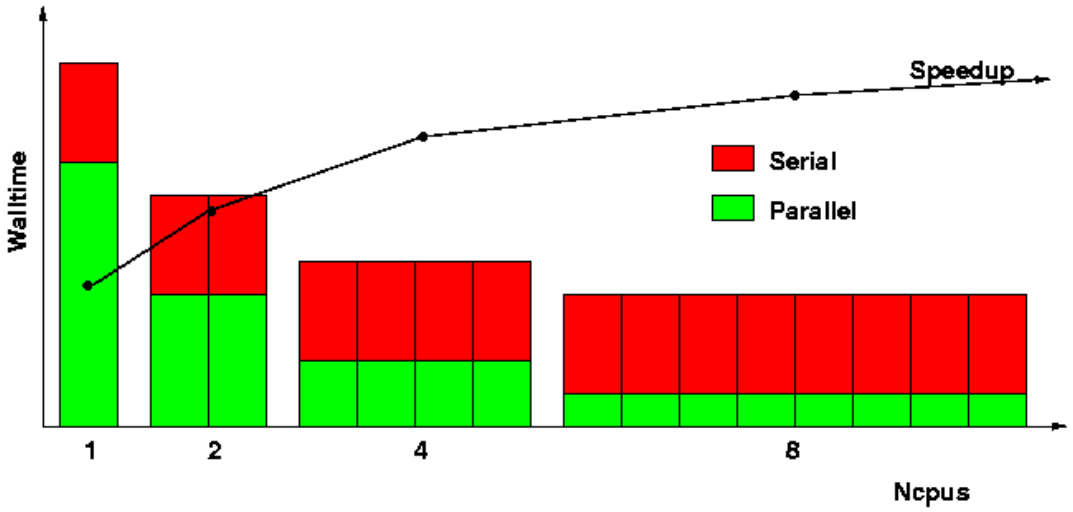


50 atoms!

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).

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

$P = 0.999999$

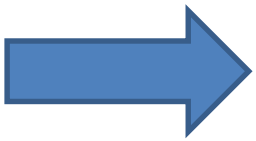
serial fraction= 0.000001

HPC Trends

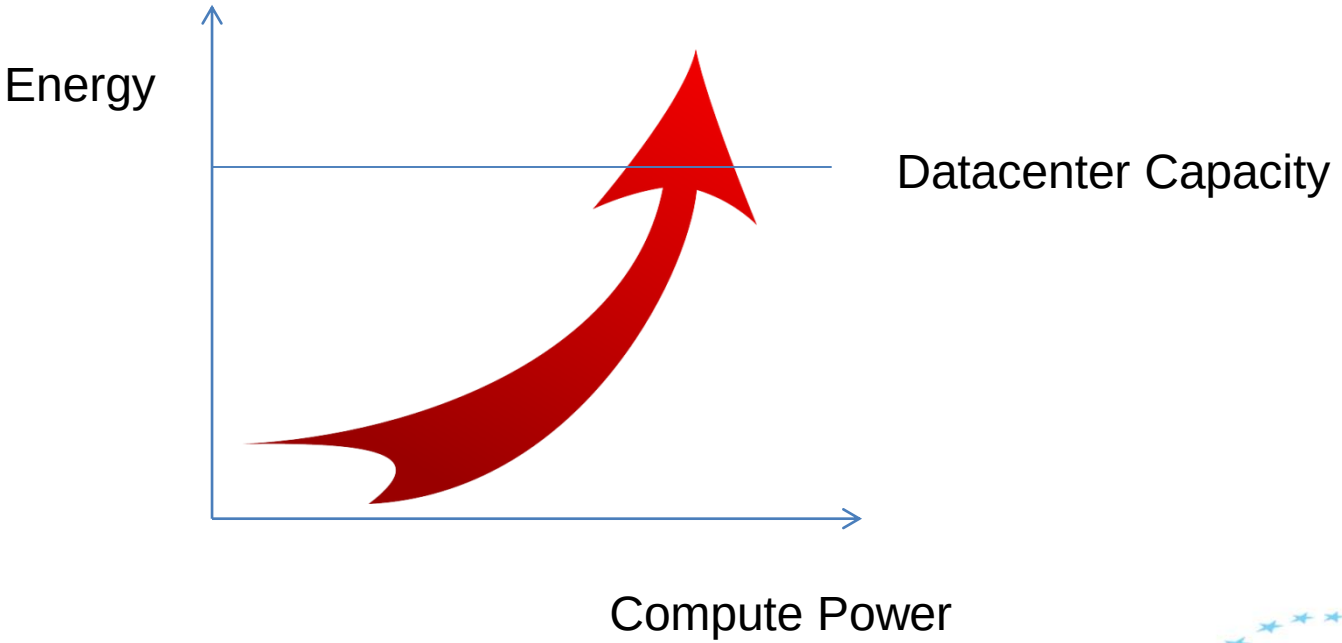
Peak Performance	Exaflops 10^18	Moore law
FPU Performance	Gigaflops 10^9	Dennard law
Number of FPUs	10^9	Moore + Dennard
App. Parallelism	serial fraction 1/10^9	Amdahl's law

Energy trends

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

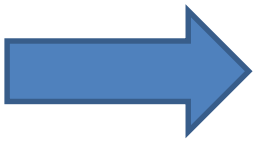


A lot of silicon to maximize single thread performace

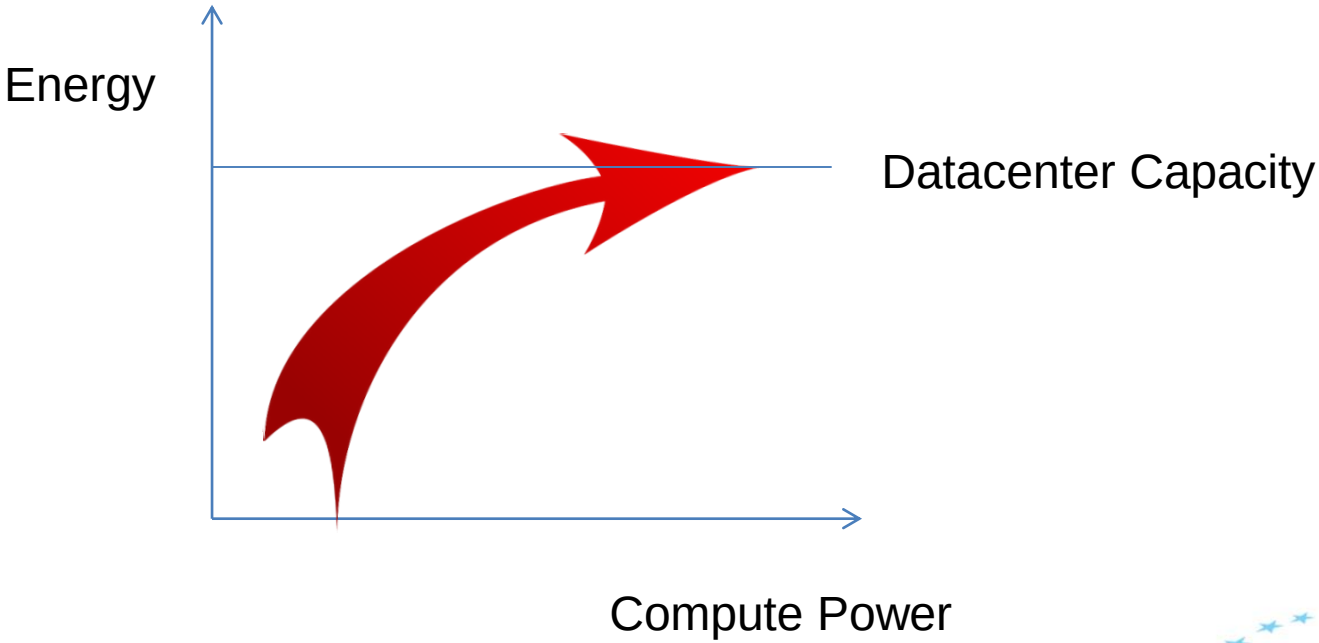


Change of paradigm

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



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



(sub) Exascale architecture

still two model

- Hybrid, but…
- Homogeneus, but…

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)	

New CINECA Tier-0

A1 - April 2016 - 1512 Lenovo NeXtScale Server con processore Intel E5-2697 v4 Broadwell (2PFs) processore E5-2697 v4 con 18 cores e 2,3GHz.

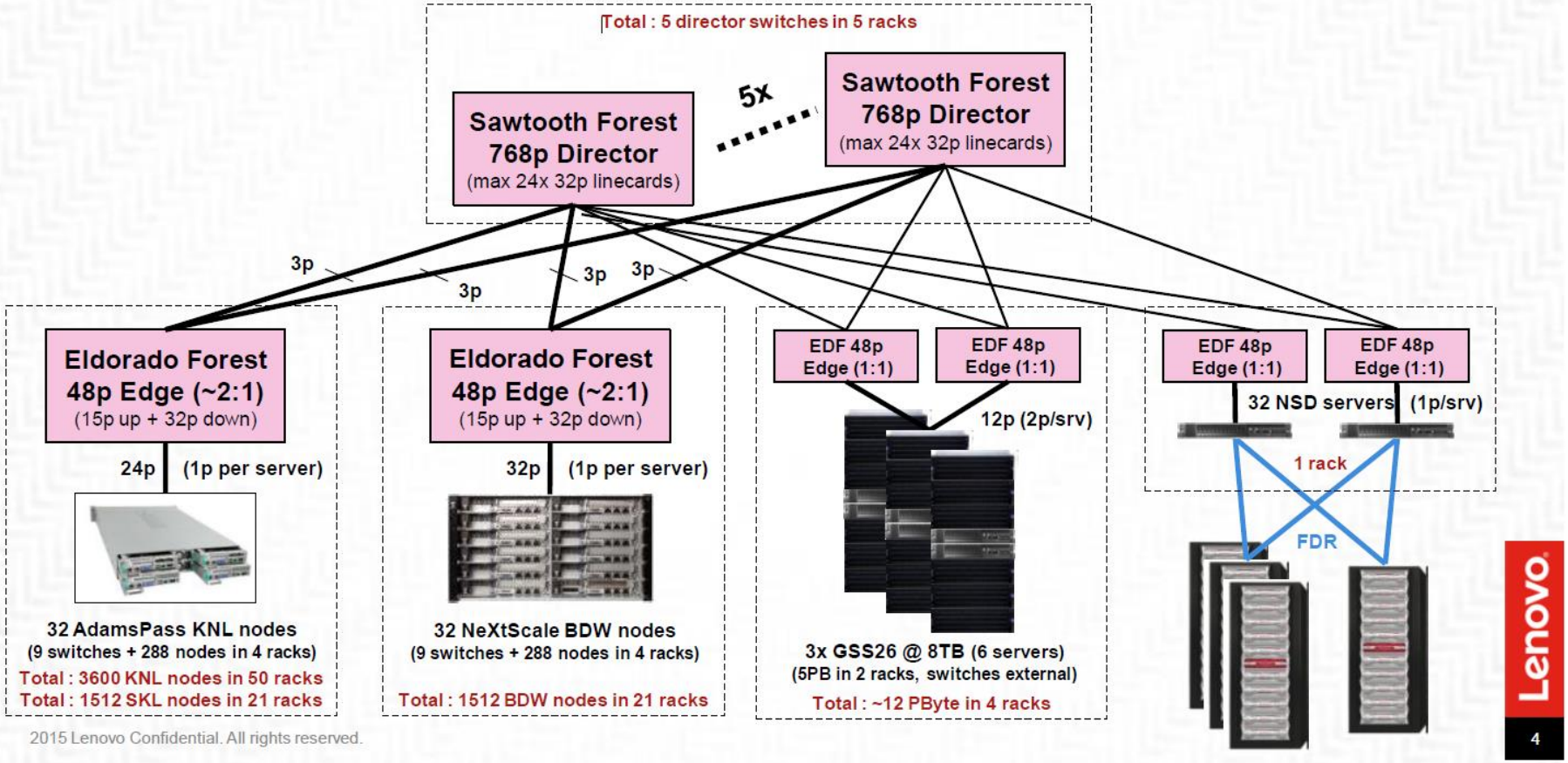
A2 – Sept. 2016 - 3600 KNL (11PFs peak)

A3 – June 2017 - 2300 Lenovo Stark Server con processore Intel E5-2680 SkyLake (7PFs peak)

Intel OmniPath interconnect

System Layout

- CINECA – Omni-Path Fabric Architecture (with 32:15 blocking)



Energy efficiency

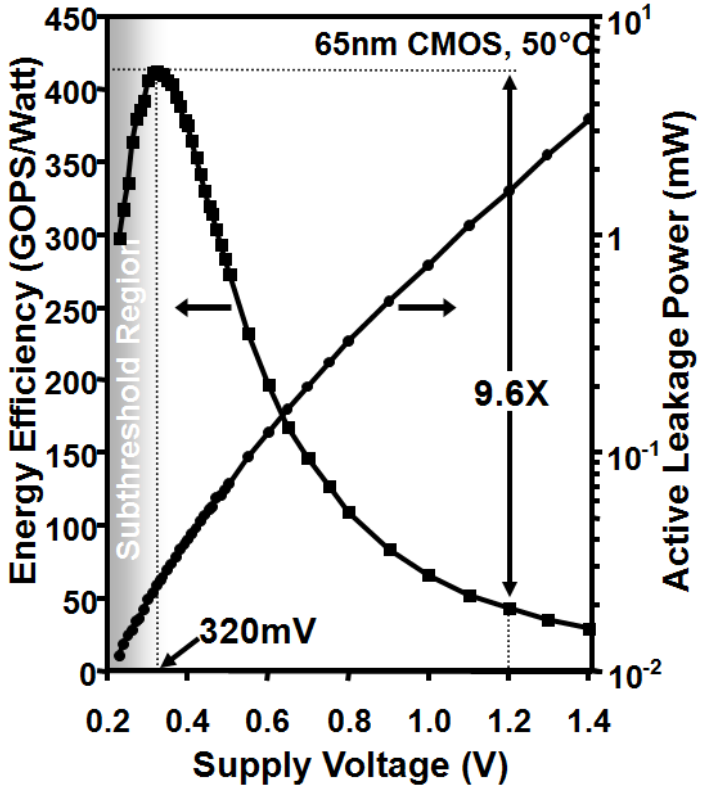
Where power is used:

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

Short term impact on programming models

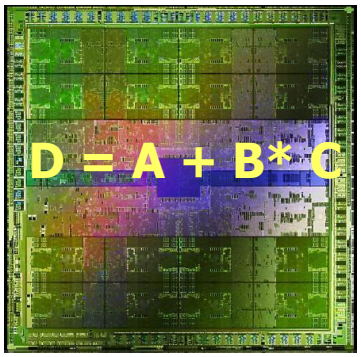
Chip efficiency

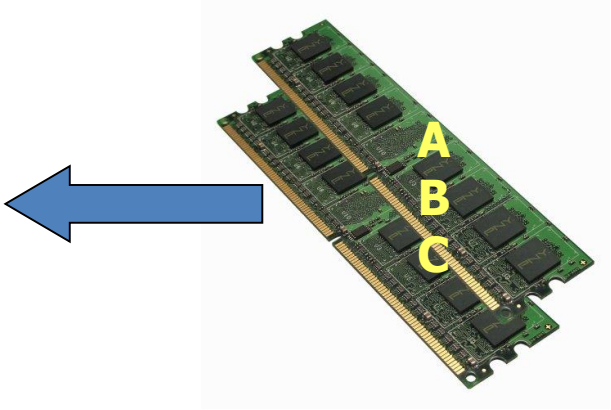
- The efficiency of CMOS transistor against the supply voltage peaks close to the insulator/conductor transition
- Possibility to design a new Near Threshold Voltage (NTV) chip architecture that is able to work at different regime.
- Accommodate the needs of different workloads and meet the requirements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



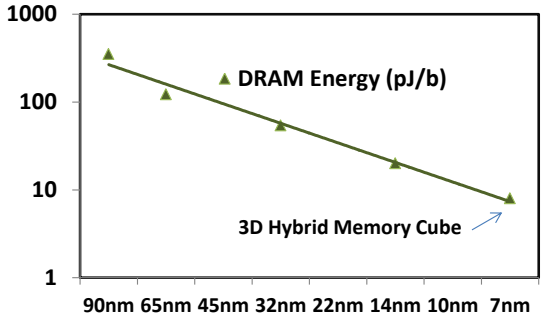
Memory

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



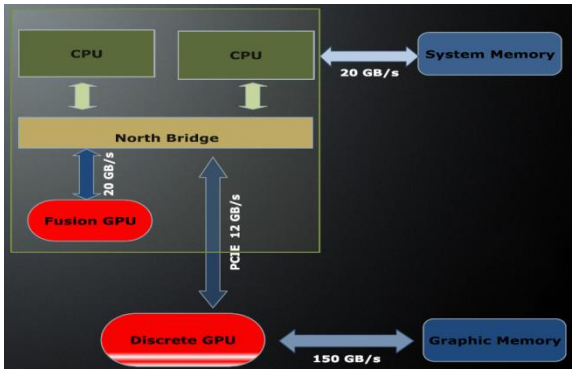
50 pJ/b today
8 pJ/b demonstrated
Need < 2pJ/b

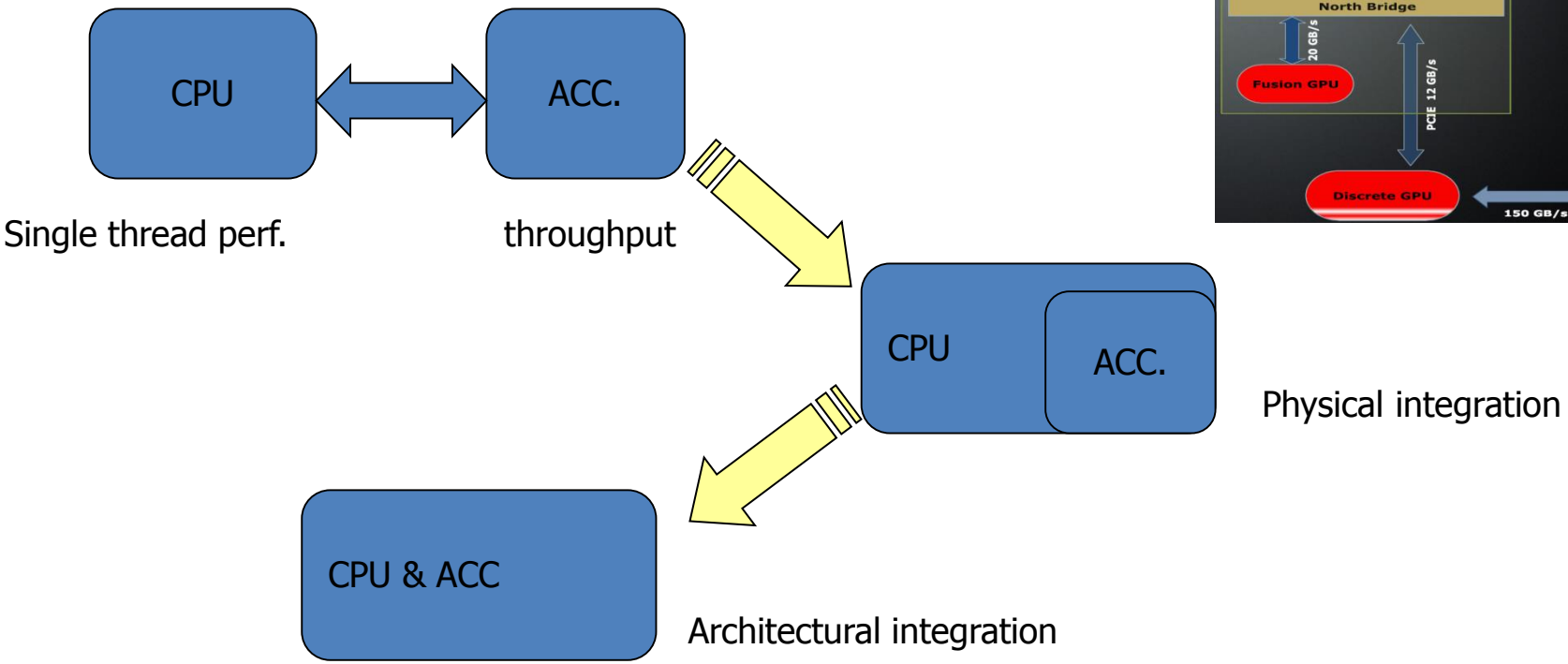
Extrapolating down to 10nm integration, the energy required to move date
Becomes 100x !

We need locality! → Fewer memory per core

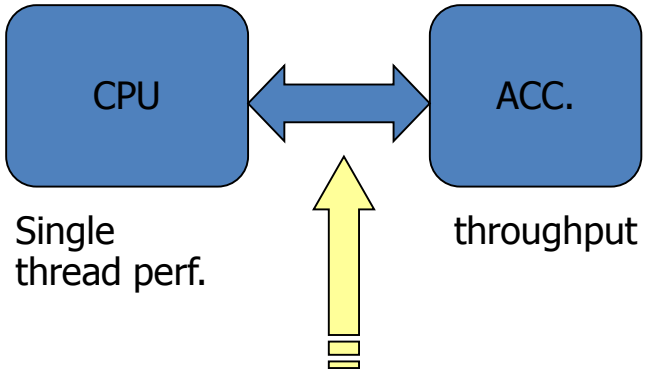
What is an Accelerator.

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)*



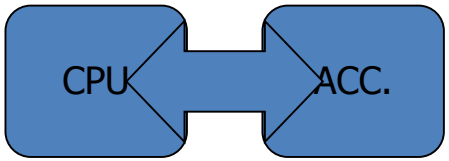


Architecture toward exascale

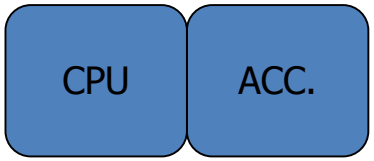


bottleneck

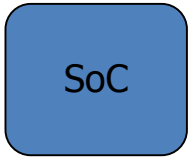
GPU/MIC/FPGA



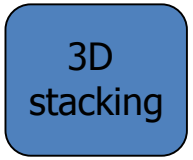
OpenPower
Nvidia GPU



AMD APU
ARM Big-Little



KNL (next Intel PHI)



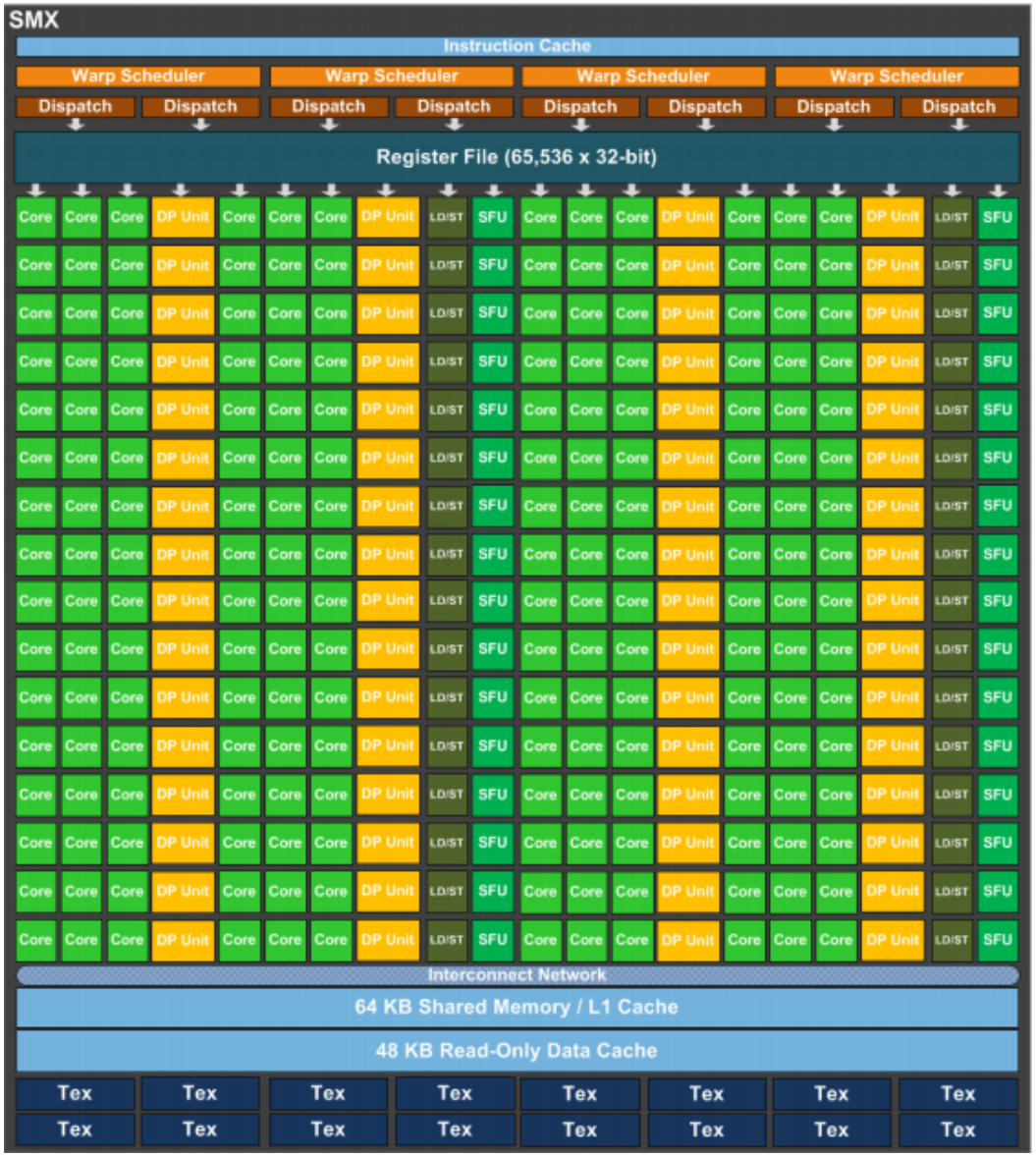
Active memory

Photonic -> platform flexibility
TSV -> stacking

K20 nVIDIA GPU

15 SMX Streaming Multiprocessors

SMX



192 single precision cuda cores

64 double precision units

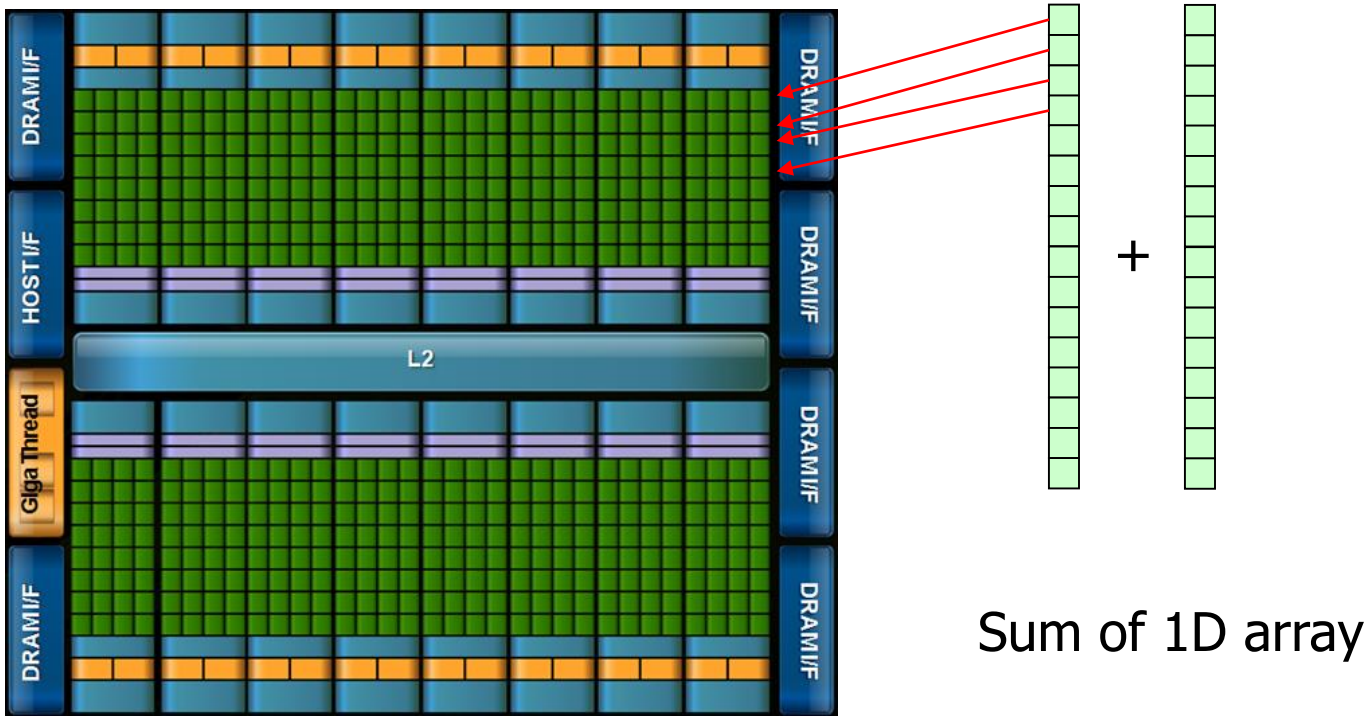
32 special function units

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

Xeon PHI Roadmap

- Knight Landing (KNL) successor of Knight Corner (KNC) processor.

- Throughput x86 solution, based on Sylvermont x86 core, Maximize Flop/watt wrt other x86 solution

- Stand-alone processor (~1.5GHz TDP freq)
- 2, 4 Numa sub-clustering

- 2xAVX512 FPU/core, 32Flop/Clk, peak perf. >= 3TFlops, 200-215watt

- Co-processor version for a later stage

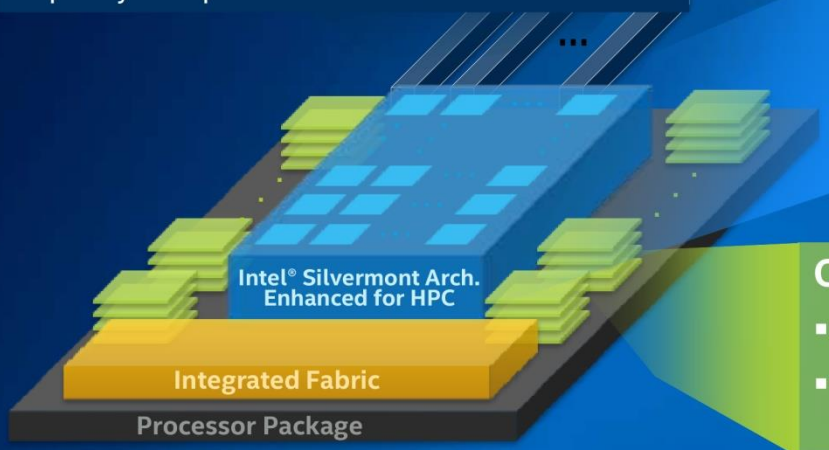
Unveiling Details of Knights Landing

(Next Generation Intel® Xeon Phi™ Products)

2nd half '15
1st commercial systems

3+ TFLOPS[1]
In One Package
Parallel Performance & Density

Platform Memory: DDR4 Bandwidth and Capacity Comparable to Intel® Xeon® Processors



Compute: Energy-efficient IA cores[2]

- *Microarchitecture enhanced for HPC*[3]
- ***3X*** *Single Thread Performance vs Knights Corner*[4]
- *Intel Xeon Processor Binary Compatible*[5]

On-Package Memory:

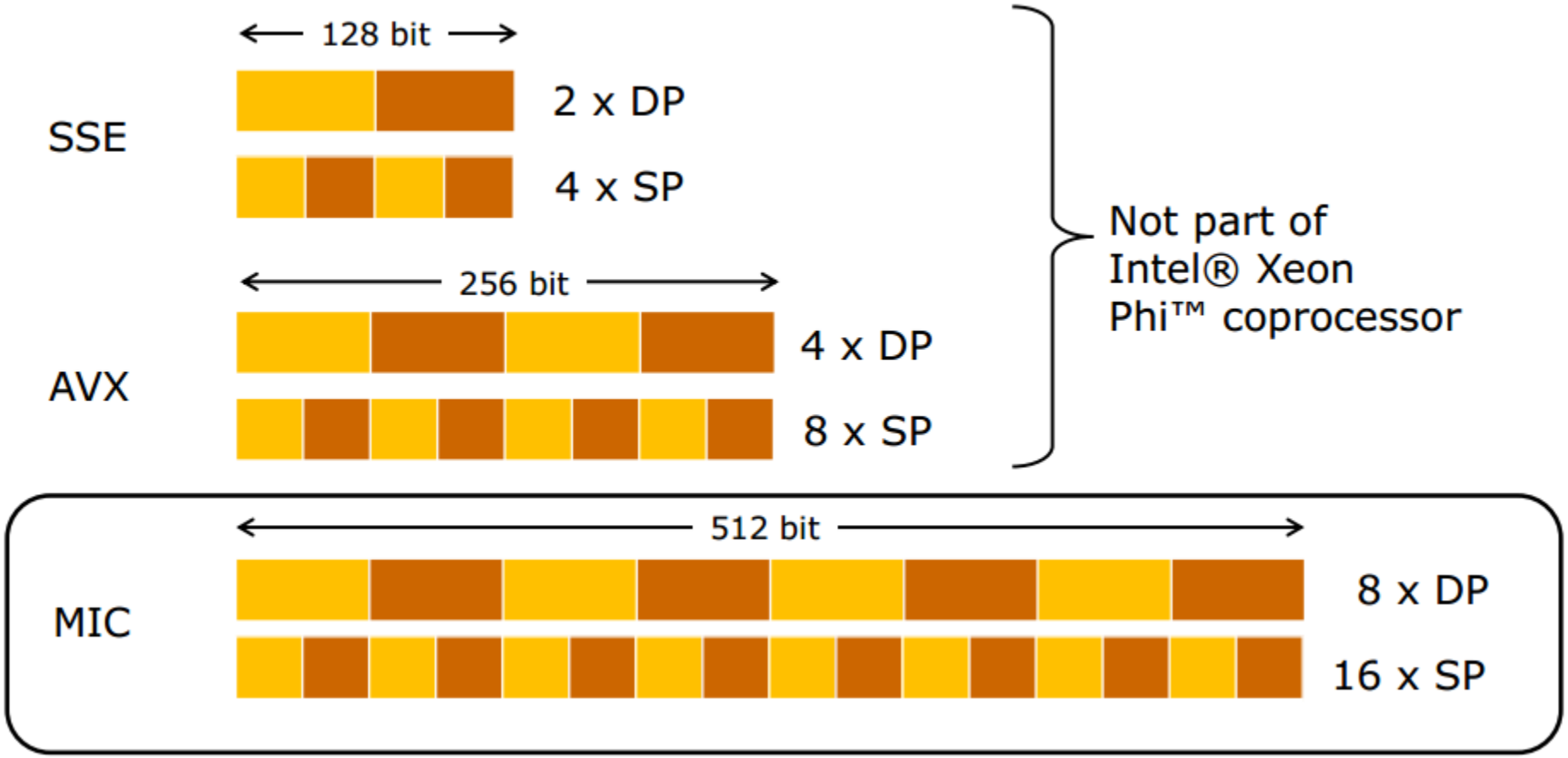
- *up to* ***16GB*** *at launch*
- ***1/3X*** *the Space*[6]
- ***5X*** *Bandwidth vs DDR4*[7]
- ***5X*** *Power Efficiency*[6]

Jointly Developed with Micron Technology

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. [1]Over 3 Teraflops of peak theoretical double-precision performance is preliminary and based on current expectations of cores, clock frequency and floating point operations per cycle. FLOPS = cores x clock frequency x floating-point operations per second per cycle. . [2]Modified version of Intel® Silvermont microarchitecture currently found in Intel® Atom™ processors. [3]Modifications include AVX512 and 4 threads/core support. [4]Projected peak theoretical single-thread performance relative to 1st Generation Intel® Xeon Phi™ Coprocessor 7120P (formerly codenamed Knights Corner). [5] Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX) . [6]Projected results based on internal Intel analysis of Knights Landing memory vs Knights Corner (GDDR5). [7]Projected result based on internal Intel analysis of STREAM benchmark using a Knights Landing processor with 16GB of ultra high-bandwidth versus DDR4 memory only with all channels populated.

Conceptual—Not Actual Package Layout

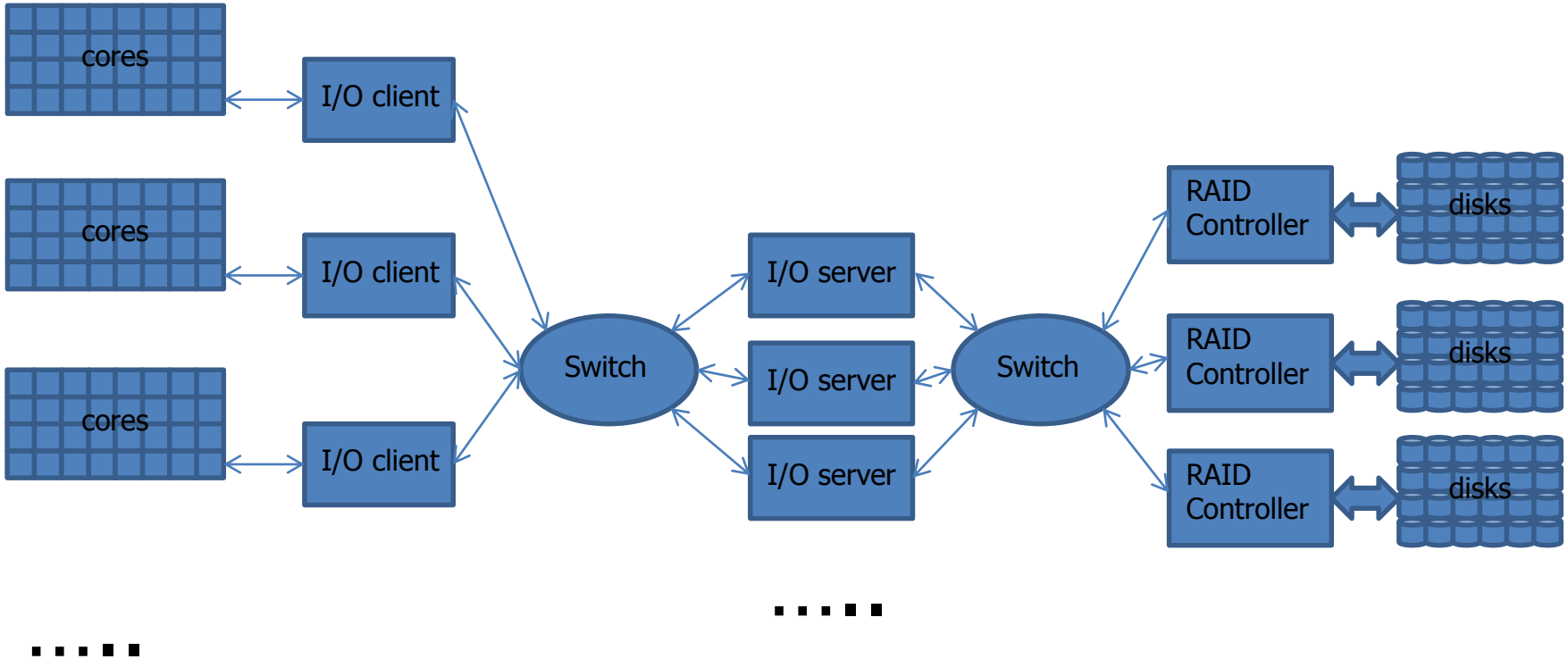
Intel Vector Units



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

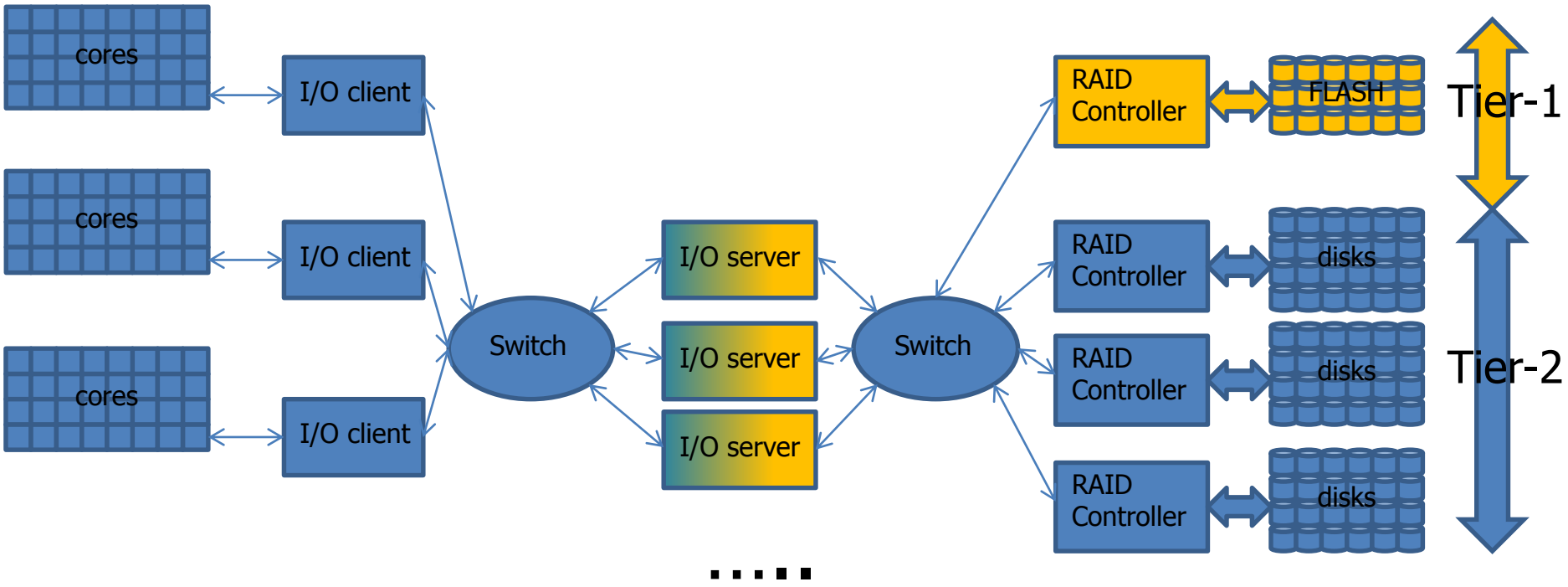
Today



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

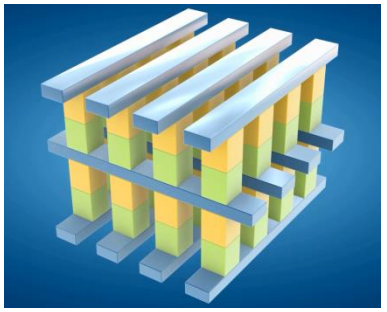
IMPORTANT: I/O subsystem has its own parallelism!

Today-Tomorrow



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

3D Xpoint



- 3D Xpoint is a technology for implement NVRAM by Micron & Intel

- Memory Cell based on Material property not on electron storage.
- No transistor are involved in storing data -> more density.

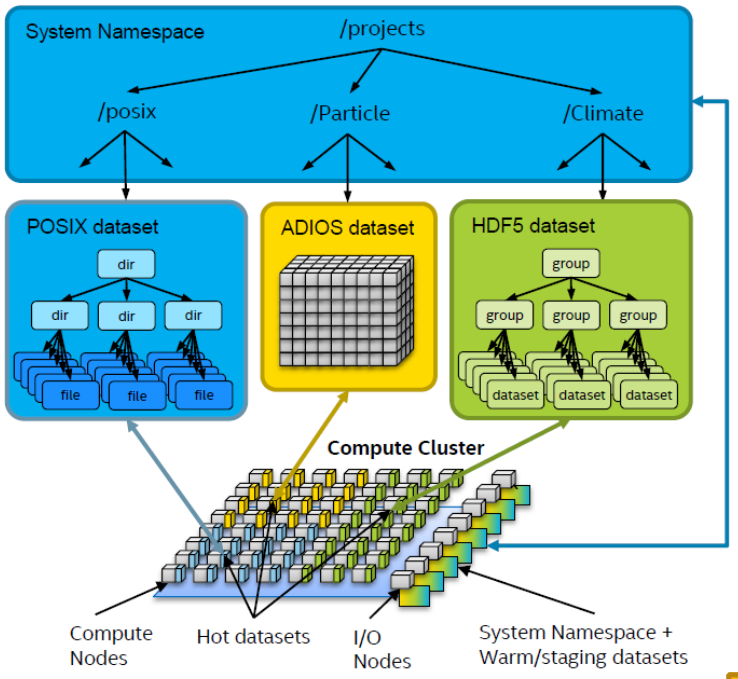
- 1,000 times lower latency and exponentially greater endurance than NAND
- 10 times denser than DRAM (no transistor technology)

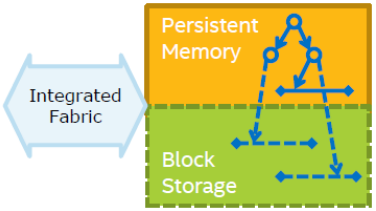
- Based on a three-dimensional arrangement of memory cells,
- allowing the cells to be addressed individually.

NVRAMM enable new Memory tiering

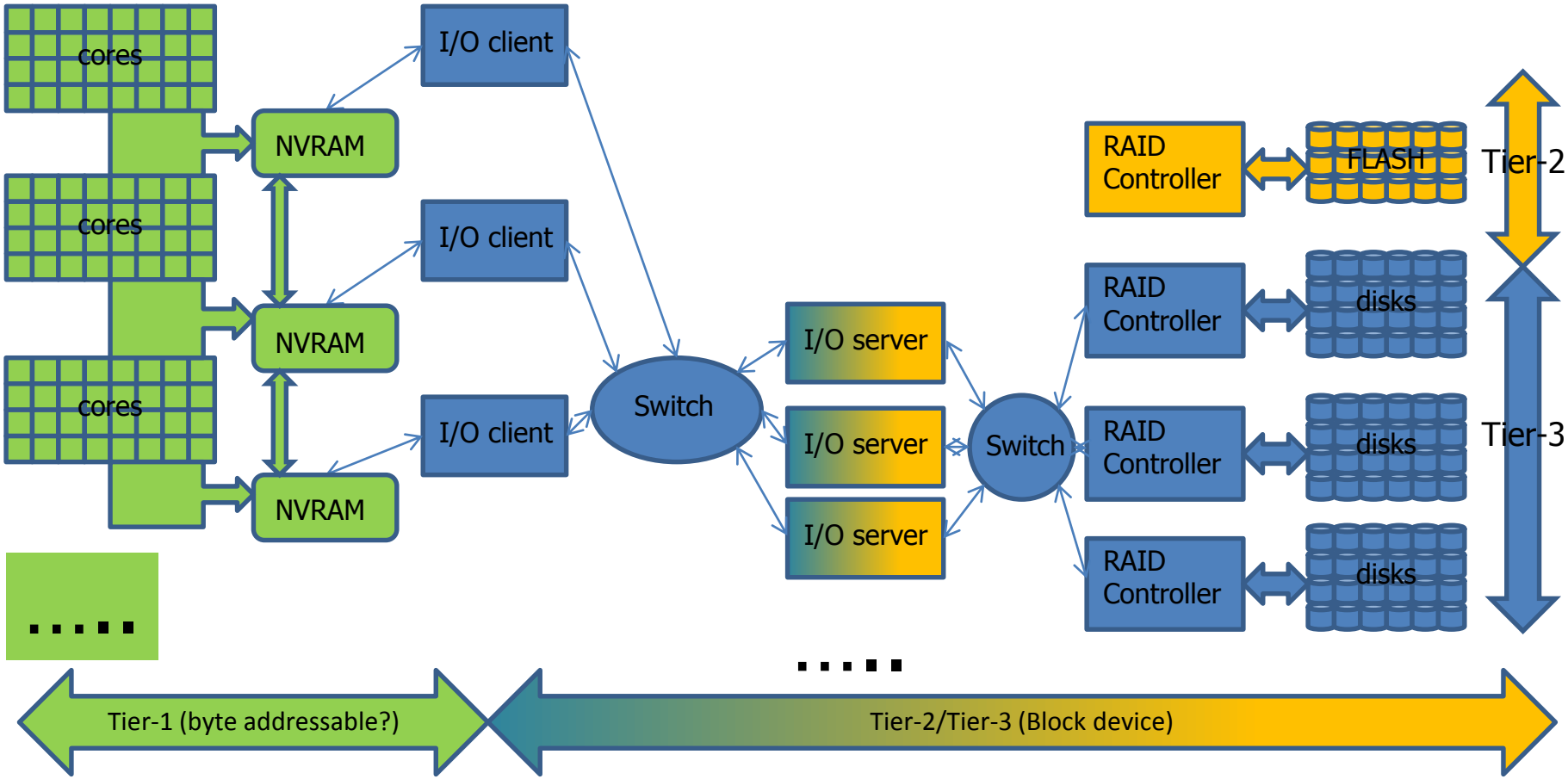
Byte addressable
Speed comparable to DRAMM
Enable new I/O stack
Beyond POSIX block filesystem
Object Storage solutions
Improve system reliability
Helps fault tolerance

- Multiple Schemas
- POSIX*
- Scientific: HDF5*, ADIOS*, SciDB*, …
- Big Data: HDFS*, Spark*, Graph Analytics, …





Tomorrow



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

Applications Challenges

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

QUANTUMESPRESSO

HOME PROJECT DOWNLOAD RESOURCES PSEUDOPOTENTIALS CONTACTS NEWS & EVENTS

SEARCH

Search here...

Forum

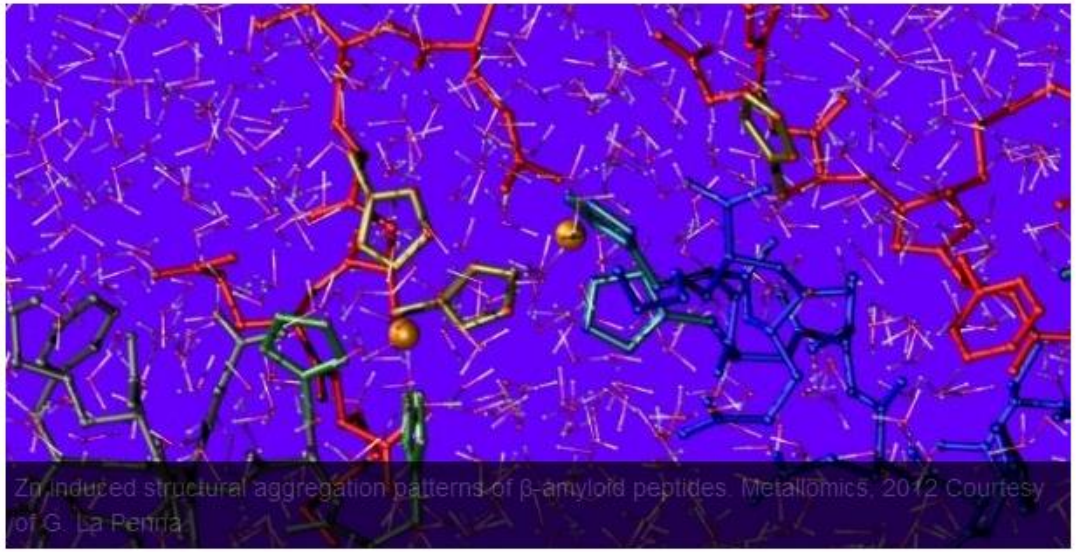
NEWS

16.06.14

THE QUANTUM ESPRESSO PRIZE

The Quantum ESPRESSO Foundation, in collaboration with Eurotech, announces the establishment of *the Quantum ESPRESSO prize for quantum mechanical materials modeling*. The prize, which consists of a diploma and a check of one thousand euros, will be awarded annually in January to recognize outstanding doctoral thesis research in the field of quantum mechanical materials modeling, realized with the help of the Quantum ESPRESSO suite of computer codes. Excellence will be rewarded for both original applications and methodological innovation.

For more information visit http://foundation.quantum-espresso.org/prize



Zn-induced structural aggregation patterns of β-amyloid peptides. Metallomics, 2012 Courtesy of G. La Penna

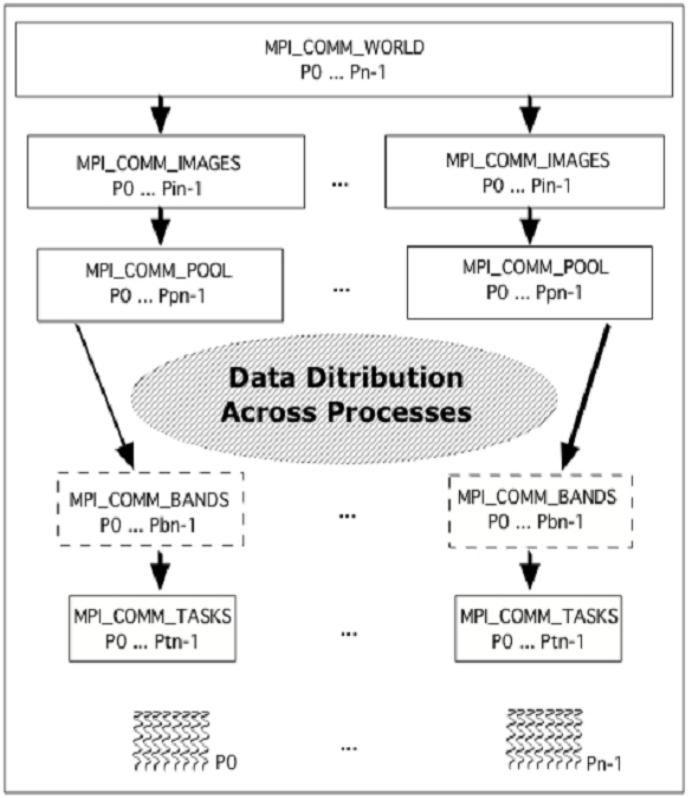
QUANTUM ESPRESSO

is an integrated suite of Open-Source computer codes for electronic-structure calculations and materials modeling at the nanoscale. It is based on density-functional theory, plane waves, and pseudopotentials.

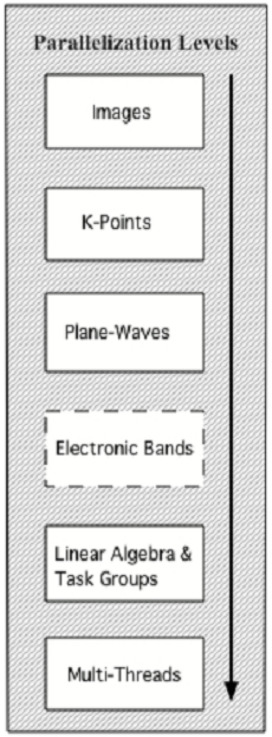
READ MORE >

www.quantum-espresso.org

Scalability
The case of Quantum Espresso



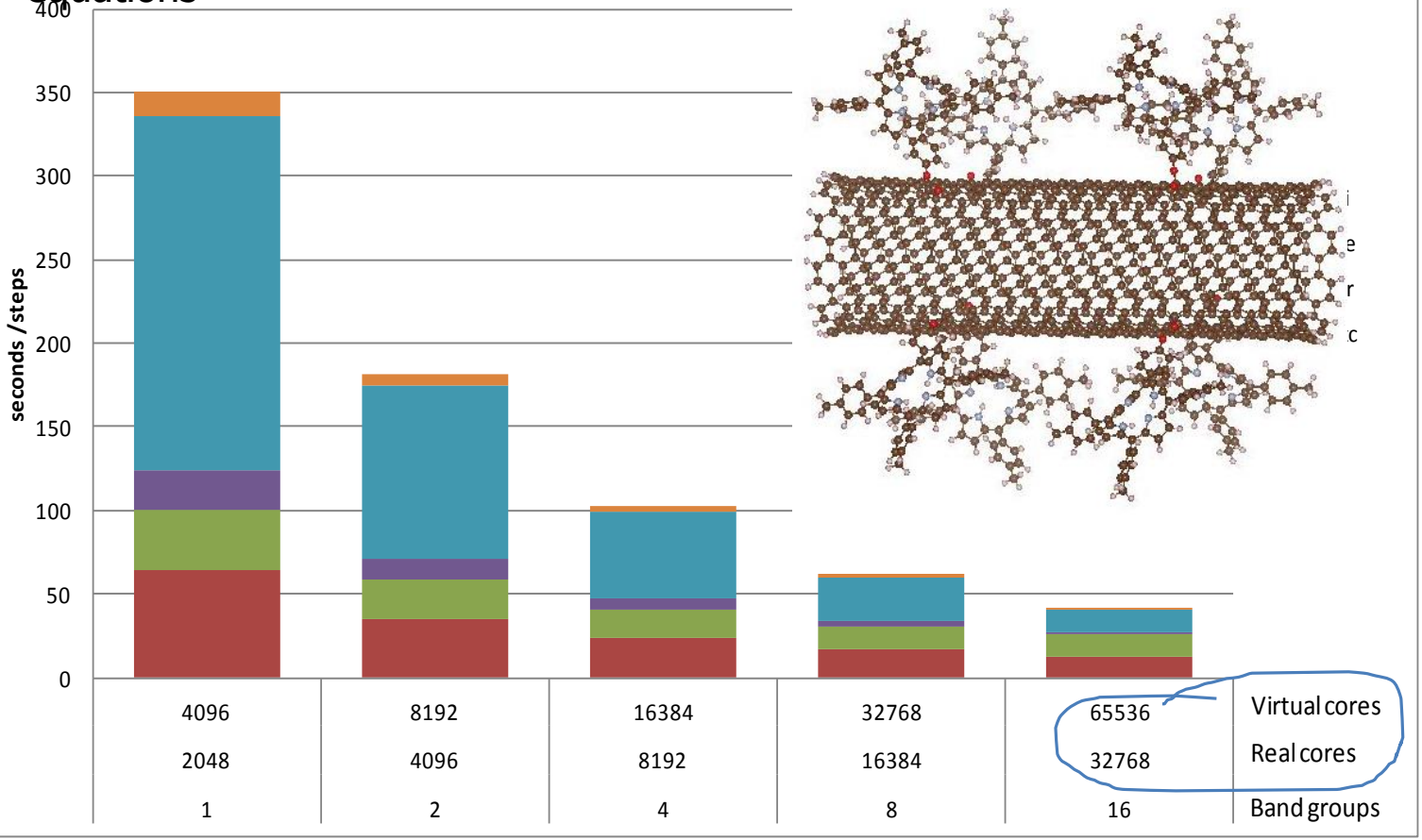
MPI Communicators Hierarchy



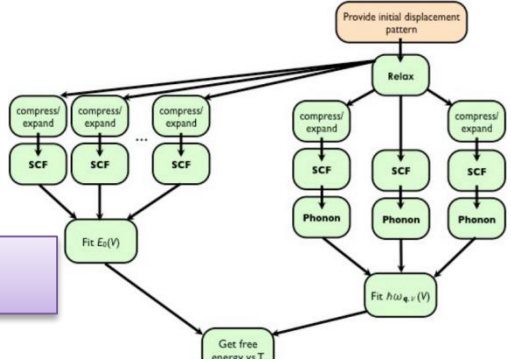
QE parallelization hierarchy

ok for 10^6 CPU cores (Petascale), not enough for 10^9 CPU cores (exascale)

- Ab-initio simulations -> numerical solution of the quantum mechanical equations



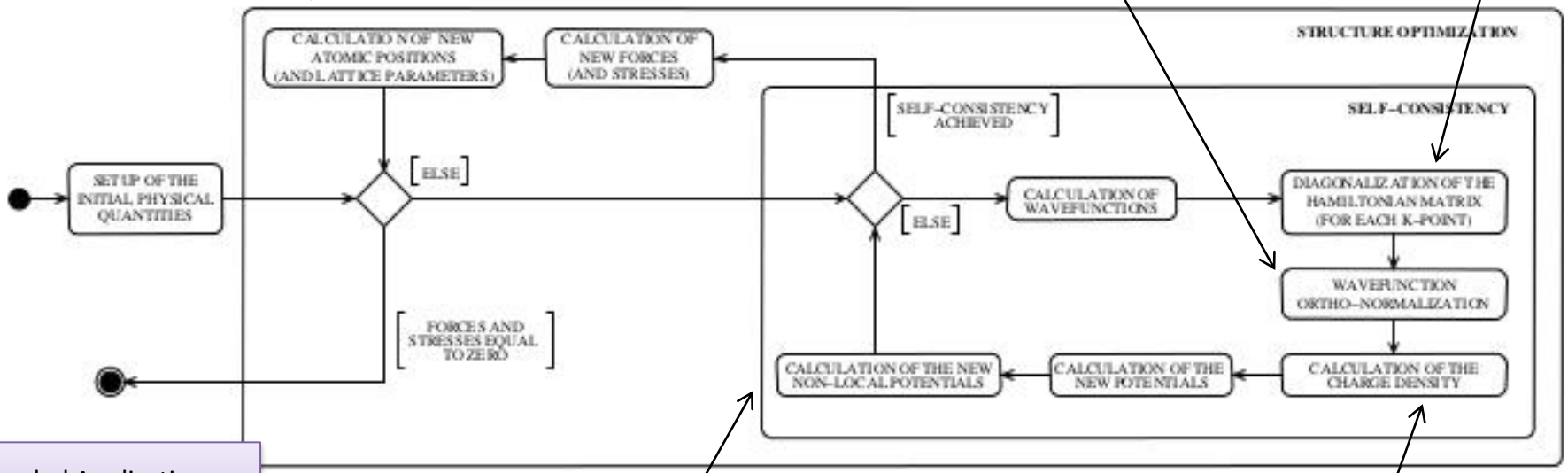
QE evolution



High Throughput / Ensamble Simulations

Communication avoiding

New Algorithm: CG vs Davidson

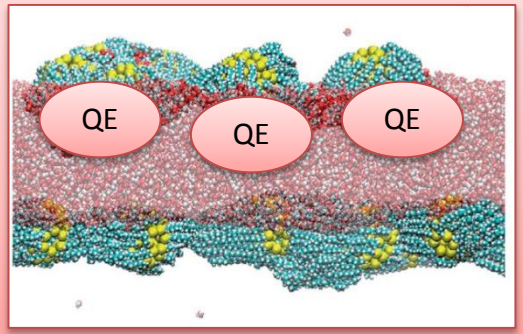


Coupled Application DSL

Task level parallelism

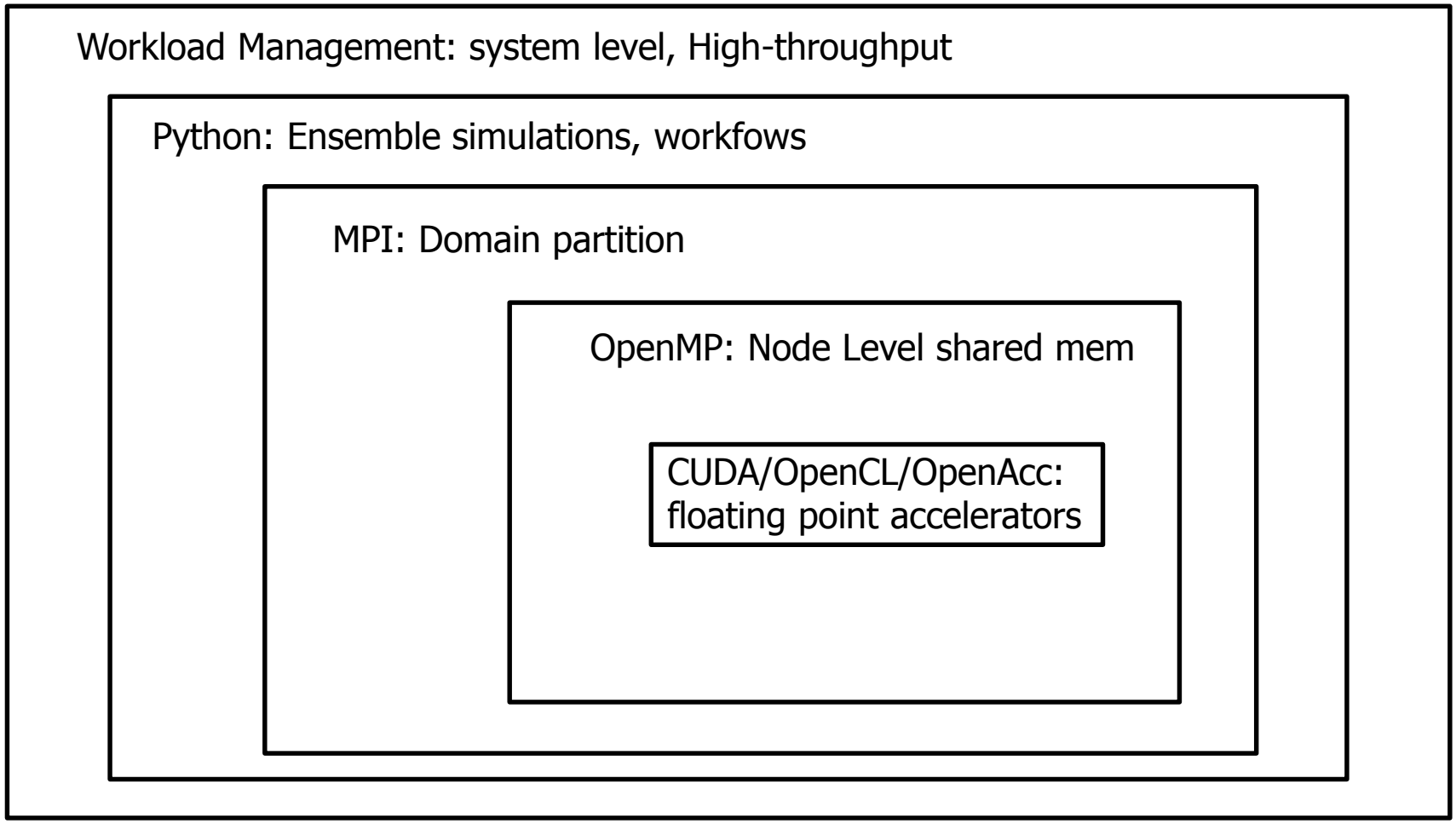
Double buffering

LAMMPS



- Reliability
- Completeness
- Robustness
- Standard Interface

Multi-level parallelism



Conclusions

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