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

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

The power crisis!

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.
The silicon lattice

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).

Si lattice

50 atoms!
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).

$$\text{maximum speedup tends to } \frac{1}{1 - P}$$

$P = \text{parallel fraction}$

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

opportunity

challenge
Energy trends

“traditional” RISK and CISC chips are designed for maximum performance for all possible workloads.

A lot of silicon to maximize single thread performance.

![Graph showing the relationship between Datacenter Capacity and Compute Power.](image)
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

Compute Power

Energy

Datacenter Capacity
(sub) Exascale architecture

still two model

Hybrid, but...
Homogeneous, but...

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>System peak</td>
<td>10 Tera</td>
<td>2 Peta</td>
<td>200 Petaflop/sec</td>
<td>1 Exaflop/sec</td>
</tr>
<tr>
<td>Power</td>
<td>~0.8 MW</td>
<td>6 MW</td>
<td>15 MW</td>
<td>20 MW</td>
</tr>
<tr>
<td>System memory</td>
<td>0.006 PB</td>
<td>0.3 PB</td>
<td>5 PB</td>
<td>32-64 PB</td>
</tr>
<tr>
<td>Node performance</td>
<td>0.024 TF</td>
<td>0.125 TF</td>
<td>0.5 TF</td>
<td>7 TF</td>
</tr>
<tr>
<td>Node memory BW</td>
<td>25 GB/s</td>
<td>0.1 TB/sec</td>
<td>1 TB/sec</td>
<td>0.4 TB/sec</td>
</tr>
<tr>
<td>Node concurrency</td>
<td>16</td>
<td>12</td>
<td>O(100)</td>
<td>O(1,000)</td>
</tr>
<tr>
<td>System size (nodes)</td>
<td>416</td>
<td>18,700</td>
<td>50,000</td>
<td>5,000</td>
</tr>
<tr>
<td>Total Node Interconnect BW</td>
<td>1.5 GB/s</td>
<td>150 GB/sec</td>
<td>1 TB/sec</td>
<td>250 GB/sec</td>
</tr>
<tr>
<td>MTTI</td>
<td>day</td>
<td>O(1 day)</td>
<td></td>
<td>O(1 day)</td>
</tr>
</tbody>
</table>
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
CINECA – Omni-Path Fabric Architecture (with 32:15 blocking)

System Layout

- Eldorado Forest 48p Edge (~2:1)
  - 24p (1p per server)
  - 15p up + 32p down
  - 32 AdamsPass KNL nodes
    - 9 switches + 288 nodes in 4 racks
    - Total: 3600 KNL nodes in 50 racks
    - Total: 1512 SKL nodes in 21 racks

- Eldorado Forest 48p Edge (~2:1)
  - 32p (1p per server)
  - 15p up + 32p down
  - 32 NeXtScale BDW nodes
    - 9 switches + 288 nodes in 4 racks
    - Total: 1512 BDW nodes in 21 racks

- Sawtooth Forest 768p Director
  - Max 24x 32p linecards
  - 5 director switches in 5 racks

- EDF 48p Edge (1:1)
  - 12p (2p/srv)

- EDF 48p Edge (1:1)
  - 3x GSS26 @ 8TB (6 servers)
    - 6PB in 2 racks, switches external
    - Total: ~12 PByte in 4 racks

2015 Lenovo Confidential. All rights reserved.
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.

Extrapolating down to 10nm integration, the energy required to move data becomes 100x!

We need locality! Fewer memory per core.

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

CPU<br>Single thread perf.<br>

ACC.<br>throughput<br>

bottleneck<br>

GPU/MIC/FPGA<br>

OpenPower Nvidia GPU<br>

AMD APU<br>ARM Big-Little<br>

KNL (next Intel PHI)<br>

Active memory<br>

Photonic -> platform flexibility<br>TSV -> stacking

SoC
K20 nVIDIA GPU

15 SMX Streaming Multiprocessors
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 independent instruction per warp
Accelerator/GPGPU

Sum of 1D array
CUDA sample

```c
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)

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

Compute: Energy-efficient IA cores
- Microarchitecture enhanced for HPC
- 3X Single Thread Performance vs Knights Corner
- Intel Xeon Processor Binary Compatible

On-Package Memory:
- up to 16GB at launch
- 5X Bandwidth vs DDR4

Joinly Developed with Micron Technology

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. Modifications include AVX512 and 4 threads/core support.
3 Projected peak theoretical single-thread performance relative to 2nd Generation Intel® Xeon Phi™ Coprocessor 7150P (formerly codenamed Knights Corner).
4 Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX).
5 Projected result based on internal Intel analysis of Knights Landing memory vs Knights Corner (DDR5).
6 Projected result based on internal Intel analysis of STREAM benchmark using a Knights Landing processor with 16GB of ultra high-bandwidth versus 800MHz memory only, with all channels populated.

Conceptual—Not Actual Package Layout
Intel Vector Units

SSE
- 128 bit
- 2 x DP
- 4 x SP

AVX
- 256 bit
- 4 x DP
- 8 x SP

MIC
- 512 bit
- 8 x DP
- 16 x SP

Not part of Intel® Xeon Phi™ coprocessor
### I/O Challenges

<table>
<thead>
<tr>
<th><strong>Today</strong></th>
<th><strong>Tomorrow</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>100 clients</td>
<td>10K clients</td>
</tr>
<tr>
<td>1000 core per client</td>
<td>100K core per clients</td>
</tr>
<tr>
<td>3PB</td>
<td>1EB</td>
</tr>
<tr>
<td>3K Disks</td>
<td>100K Disks</td>
</tr>
<tr>
<td>100 Gbyte/sec</td>
<td>100TByte/sec</td>
</tr>
<tr>
<td>8MByte blocks</td>
<td>1Gbyte blocks</td>
</tr>
<tr>
<td>Parallel Filesystem</td>
<td>Parallel Filesystem</td>
</tr>
<tr>
<td>One Tier architecture</td>
<td>Multi Tier architecture</td>
</tr>
</tbody>
</table>
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 implementing NVRAM by Micron & Intel.

- Memory Cell based on material properties, not on electron storage.
- No transistors 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
SEARCH

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.

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.

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

www.quantum-espresso.org
Scalability
The case of Quantum Espresso

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

<table>
<thead>
<tr>
<th>Virtual cores</th>
<th>Real cores</th>
<th>Band groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>65536</td>
<td>32768</td>
<td>16</td>
</tr>
<tr>
<td>32768</td>
<td>16384</td>
<td>8</td>
</tr>
<tr>
<td>16384</td>
<td>8192</td>
<td>4</td>
</tr>
<tr>
<td>8192</td>
<td>4096</td>
<td>2</td>
</tr>
<tr>
<td>4096</td>
<td>2048</td>
<td>1</td>
</tr>
</tbody>
</table>

seconds/steps

- $4096$ 2048 $1$
- $8192$ 4096 $2$
- $16384$ 8192 $4$
- $32768$ 16384 $8$
- $65536$ 32768 $16$
QE evolution

High Throughput / Ensemble Simulations

- New Algorithm: CG vs Davidson
- Communication avoiding

Coupled Application DSL

LAMMPS

- Task level parallelism
- Double buffering

- Reliability
- Completeness
- Robustness
- Standard Interface
Multi-level parallelism

Workload Management: system level, High-throughput

Python: Ensemble simulations, workflows

MPI: Domain partition

OpenMP: Node Level shared mem

CUDA/OpenCL/OpenAcc: floating point accelerators
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