Exascale and transprecision computing

Andrew Emerson
Cineca |www.hpc.cineca.it

But first an introduction to Cineca - Where we are

Three locations, with the head office in Casalecchio di Reno (nr Bologna).

Cineca Description

- A consortium (non-profit) formed from 70 Italian Universities, 6 National Research Institutes and the Italian Ministry of Education and Research.
- Founded in 1969 as a centre for providing supercomputer resources (CINECA= *Consorzio Interuniversitario per il Calcolo Automatico dell'Italia Nord Orientale*), its activities now also include services for public administration, health and for the private sector.

High Performance Computing

- HPC Department at Cineca is called SCAI (SuperComputing Applications and Innovation).
- Manages and provides HPC resources and training for the Italian and European Research community.
- Participates also in many projects funded by the European Commission, e.g. OPRECOMP (Open Transprecision Computing).

Training

HPC training events 2017			
COURSE NAME	Bologna	Rome	Milan
Programming paradigms for GPU devices		01/03 March	
Introduction to modern Fortran	18/21 September	13/16 March	22/25 May
High Performance Molecular Dynamics	27/29 September	05/07 April	
HPC Numerical Libraries	10/12 April		
Debugging and Optimization of Scientific Applications	27/29 November	19/21 April	
Introduction to R for data analytics	26/27 April		
Scientific Visualization for Computational Chemistry	02/04 May		
Introduction to Scientific and Technical Computing in C	25/27 October	03/05 May	24/26 January

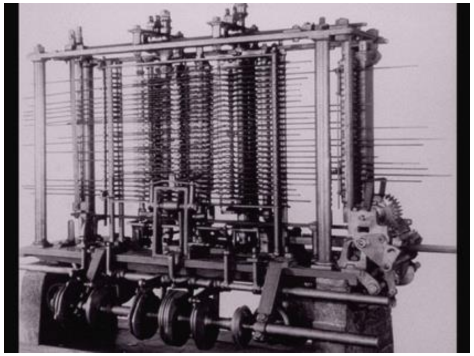
Back to the talk. But what is Exascale?

First we must start with the question:

“What is a supercomputer?”

Supercomputers are defined as the most powerful computers available in a given period of time.

Powerful is meant in terms of execution speed, memory capacity and accuracy of the machine.

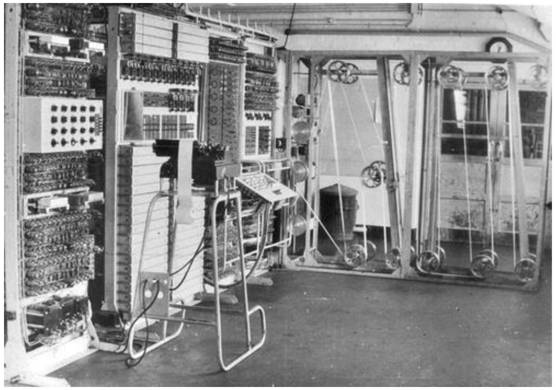


Supercomputer:"*new statistical machines with the mental power of 100 skilled mathematicians in solving even highly complex algebraic problems"..*

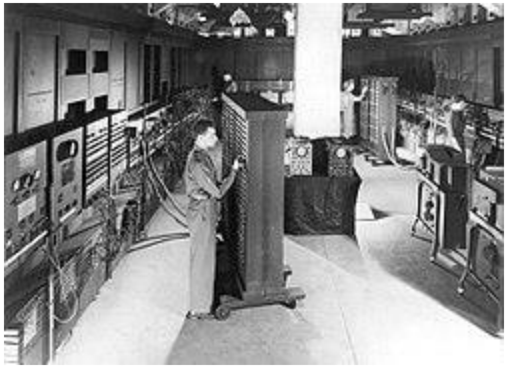
New York World, march 1920

to describe the machines invented by Mendenhall and Warren, used at Columbia University's Statistical Bureau.

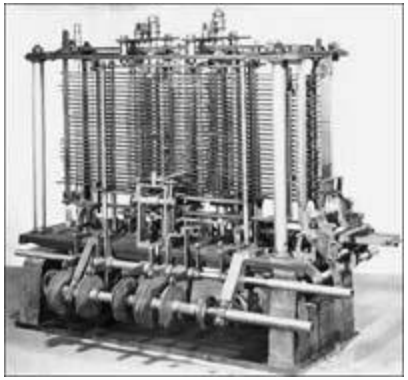
The first computers



COLUSSUS, Bletchley Park, UK (first programmable computer)



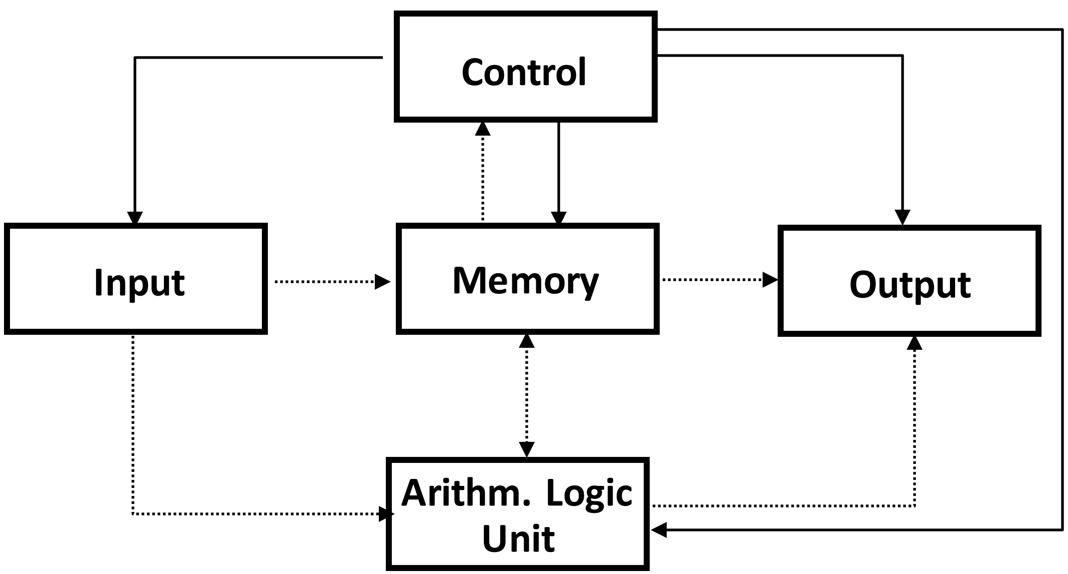
ENIAC - first electronic computer

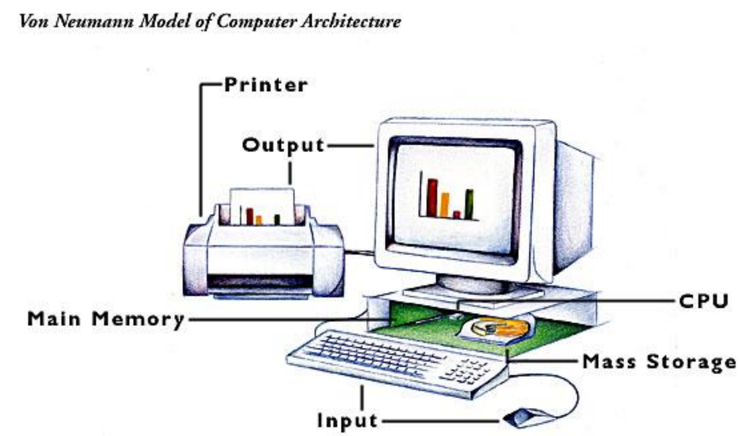


Analytical Engine (Babbage)

How do they work? - it starts from the von Neumann Model

Conventional Computer





.... **Data**

———— **Control**

Instructions are processed sequentially

1. A single instruction is loaded from memory (**fetch**) and decoded
2. Compute the addresses of operands
3. Fetch the operands from memory;
4. Execute the instruction ;
5. Write the result in memory (**store**).

Processor speed: Clock Cycle and Frequency

The instructions of all modern processors need to be *synchronised* with a timer or *clock*.

The *clock cycle* τ is defined as the time between two adjacent pulses of oscillator that sets the time of the processor.

The number of these pulses per second is known as clock speed or clock frequency, generally measured in GHz (gigahertz, or billions of pulses per second). **In principle the higher the frequency the faster the processor.**

The clock cycle controls the synchronization of operations in a computer: All the operations inside the processor last a multiple of τ.

Processor	τ (ns)	freq (MHz)
CDC 6600	100	10
Cyber 76	27.5	36
IBM ES 9000	9	111
Cray Y-MP C90	4.1	244
Intel i860	20	50
PC Pentium	< 0.5	> 2 GHz
Power PC	1.17	850
IBM Power 5	0.52	1.9 GHz
IBM Power 6	0.21	4.7 GHz

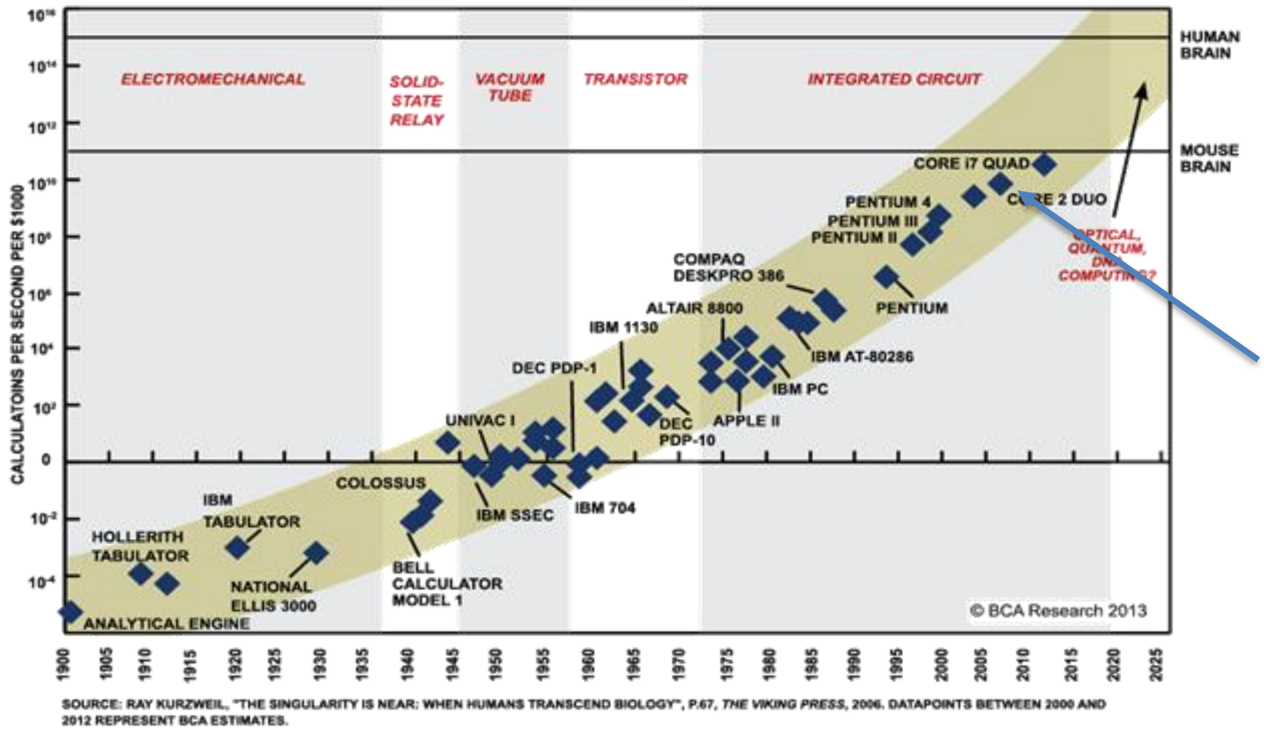
Increasing the clock frequency:

The **speed of light** sets an upper limit to the speed with which electronic components can operate .

Propagation velocity of a signal in a vacuum: **300,000 Km/s = 30 cm/ns**

Heat dissipation problems inside the processor. Power consumption varies as the square or cube of the clock frequency.

Moore’s Law



Tri-Gate 3D transistor (2011, Intel, e.g.22nm Ivy Bridge)

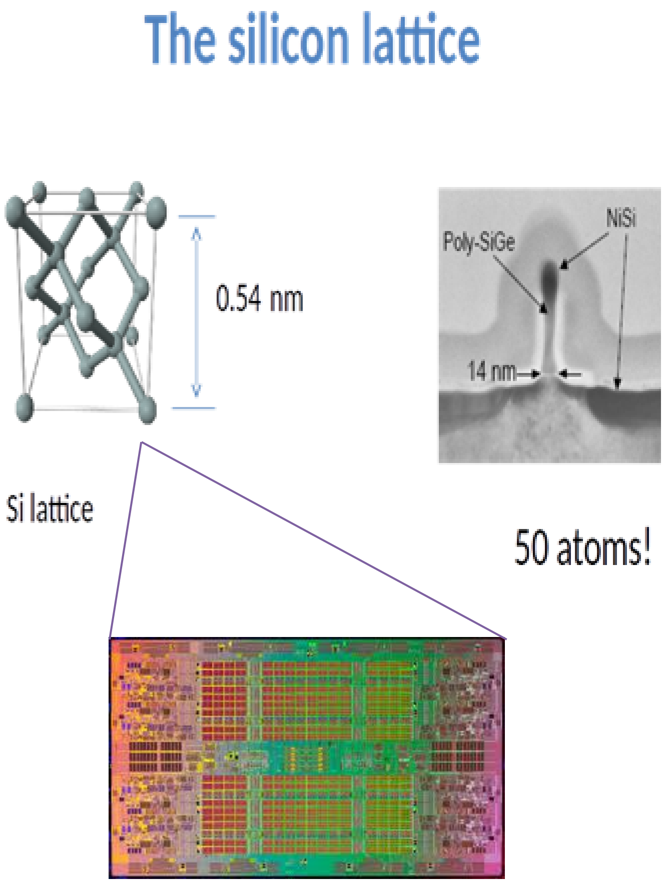
Empirical law which states that the complexity of devices (number of transistors per square inch in microprocessors) doubles every 18 months..

Gordon Moore, INTEL co-founder, 1965

The end of Moore's Law?

There is some debate as to whether Moore's Law still holds (probably not) but will undeniably fail for the following reasons:

- Minimum transistor size
  - Transistors cannot be smaller than single atoms. Most chips today use 14nm fabrication technology, although IBM in 2015 demonstrated a 7nm chip.
- Quantum tunnelling
  - As transistors get smaller quantum effects such as tunnelling get more important and can cause current leakage.
- Heat dissipation and power consumption
  - Increasingly difficult to remove heat and keep power levels within reasonable limits. Partially offset by multi-core chips.



Increase in transistor numbers does not necessarily mean more CPU power - software usually struggles to make use of the available hardware threads.

Concepts of Parallelism

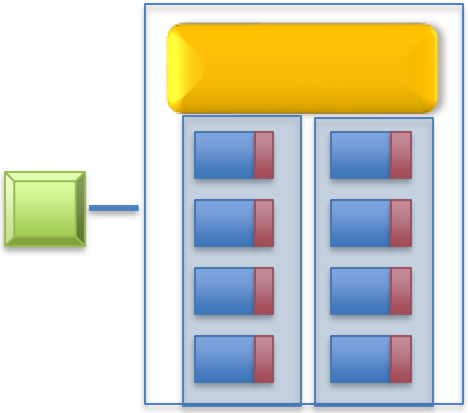
It has been recognised for some time that *serial computing* cannot bring the increases in performance required in HPC applications.

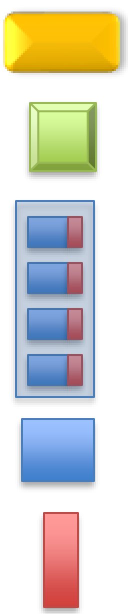
The key is to introduce *parallelism* which can be present at many levels:

- Instruction level (e.g. fma = fused multiply and add).
- Vector processing (e.g. data parallelism)
- Hyperthreading (e.g. 4 hardware threads/core for Intel KNL, 8 for PowerPC).
- Cores / processor (e.g. 18 for Intel Broadwell)
- Processors (or sockets) / node - often 2 but can be 1 (KNL) or >2/
- Processors + accelerators (e.g. CPU+GPU)
- Nodes in a system

To reach the maximum (*peak*) performance of a parallel computer, all levels of parallelism need to be exploited.

Parallel computers - the basic design (NUMA - Non-Uniform Memory Access)





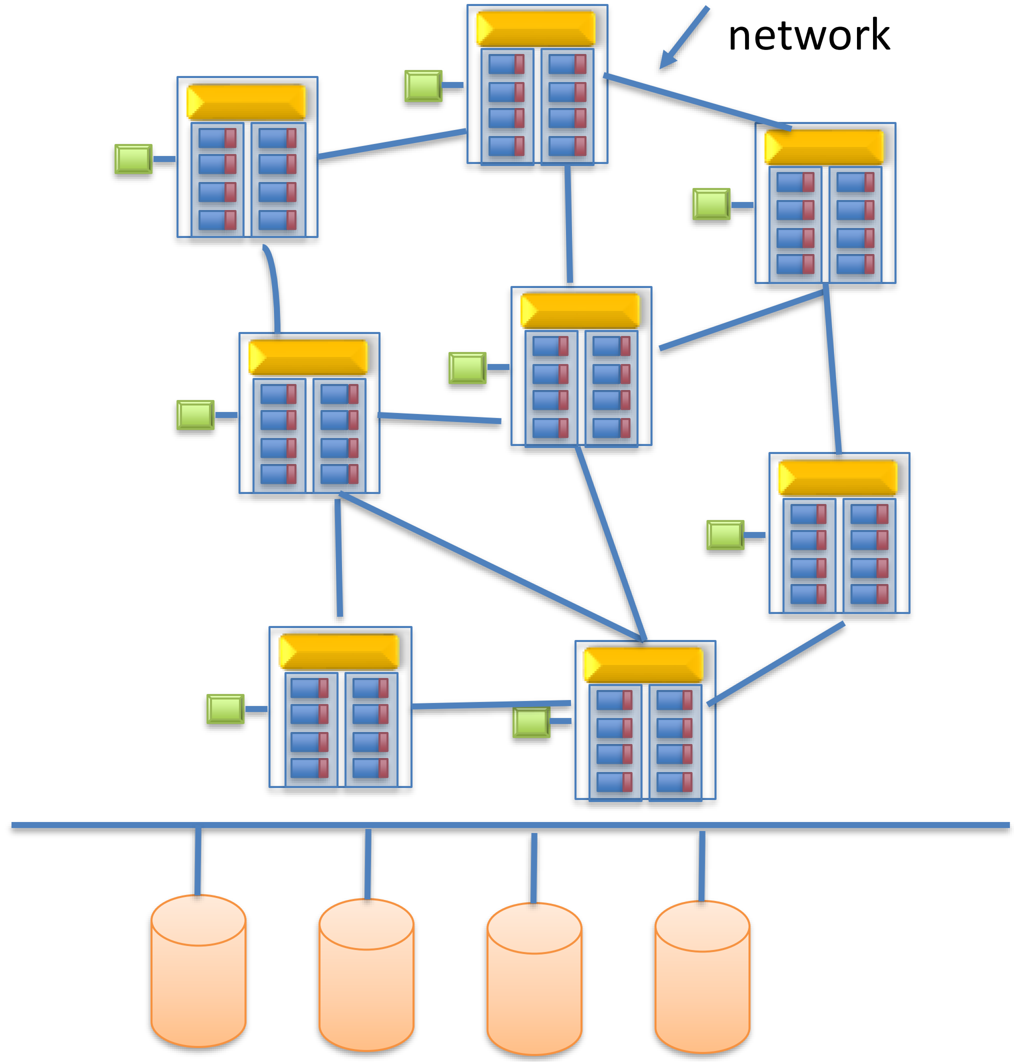
memory

accelerator (e.g. GPU)

processor

core

vector unit



parallel filesystem

Which factors drive the evolution in HPC architecture?

- The first (super) computers were mainly used by defence organisations and the US Govt (esp. Department of Energy) still makes significant investments. Later they were used for scientific research in a small number of centres.
- But the high cost of the dedicated components, and the fact that HPC is not a strong market, has caused a shift into using commodity or off-the-shelf devices such as processors, disks, memories, networks, etc.
- This shift has had a number of consequences:
  - Some manufactures have changed business or no longer make supercomputers (e.g. SUN microsystems).
  - Other supercomputer vendors (e.g. CRAY and SGI) no longer make microprocessors so the market is dominated by one or two brands, i.e. Intel and, to a lesser extent, IBM PowerPC.
  - Since microprocessors were not designed for HPC, the programmer must work harder to get maximum performance.
- But porting has become easier as only a few processor types are available and Linux has replaced all the other operating systems. It is now also possible for smaller organisations such as university departments to run small clusters.

Supercomputer evolution in Cineca

1969: CDC 6600 1st system for scientific computing
1975: CDC 7600 1st supercomputer
1985: Cray X-MP / 4 8 1st vector supercomputer
1989: Cray Y-MP / 4 64
1993: Cray C-90 / 2 128
1994: Cray T3D 64 1st parallel supercomputer
1995: Cray T3D 128
1998: Cray T3E 256 1st MPP supercomputer
2002: IBM SP4 512 1 Teraflops
2005: IBM SP5 512
2006: IBM BCX 10 Teraflops
2009: IBM SP6 100 Teraflops
2012: IBM BG/Q 2 Petaflops
2016: Lennovo (Marconi) 13 Pflops
2018: Lennovo (Marconi) 20 Pflops

TOP500 list November 2017

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	361,760	19,590.0	25,326.3	2,272
4	Japan Agency for Marine-Earth Science and Technology Japan	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz ExaScaler	19,860,000	19,135.8	28,192.0	1,350
5	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
6	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
7	DOE/NNSA/LANL/SNL United States	**Trinity** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	979,968	14,137.3	43,902.6	3,844
8	DOE/SC/LBNL/NERSC United States	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	622,336	[illegible]	[illegible]	[illegible]
9	Joint Center for Advanced High Performance Computing Japan	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Fujitsu	556,104	[illegible]	[illegible]	[illegible]
10	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660

List published twice a year (June and November) listing the top 500 most powerful computer systems.
Dominated in recent years by Japan and China, although US expected to advance with the CORAL procurement.
CINECA currently in 14th place but may rise due to upgrade.
Not included yet 18.6 Pflop GPU cluster owned by ENI.

14	CINECA Italy	**Marconi Intel Xeon Phi** - CINECA Cluster, Lenovo SD530, Intel Xeon Phi 7250 68C 1.4GHz/Platinum 8160, Intel Omni-Path Lenovo	314,384	7,471.1	15,372.0

CINECA Marconi

TOP500 List June 2018

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,282,544	122,300.0	187,659.3	8,806
2	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
3	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/NNSA/LLNL United States	1,572,480	71,610.0	119,193.6	
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	32,576.6	1,649
6	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272
7	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209

The US claimed the top spot in the Top 500 in June 2018 with Summit, a 190 Pflops IBM P9 cluster with Volta Nvidia GPUs. The high performance mainly due to the powerful Nvidia GPUs.

Also 3rd place with Sierra, a similar P9 cluster + GPUs.

Cineca 18th with Marconi but top in Italy is a GPU cluster of ENI (13th).

HPC Evolution - The Energy cost

The main driver in the last 10 years has been the need to reduce the power consumption.

This has already led to multi-core processors which are now multi-core (4,6,8, 10 and increasing) but of low frequency (rarely above 2.5 GHz) as power consumption varies exponentially with GHz.

To increase overall performance, but keep energy costs down, overall parallelism must be increased.

Two possible solutions have been proposed for low energy clusters:

1. Large homogenous clusters
2. Hybrid clusters with different processors and accelerators

The Homogenous solution - IBM Bluegene

Homogeneous cluster containing very large number of low-power cores (tens or hundreds of thousands).

Common HPC solution for a number of years but unpopular with users since applications needed to be very highly parallel (at least use upto 1024 cores) and had poor I/O performance.

Also non-Linux OS on compute nodes.

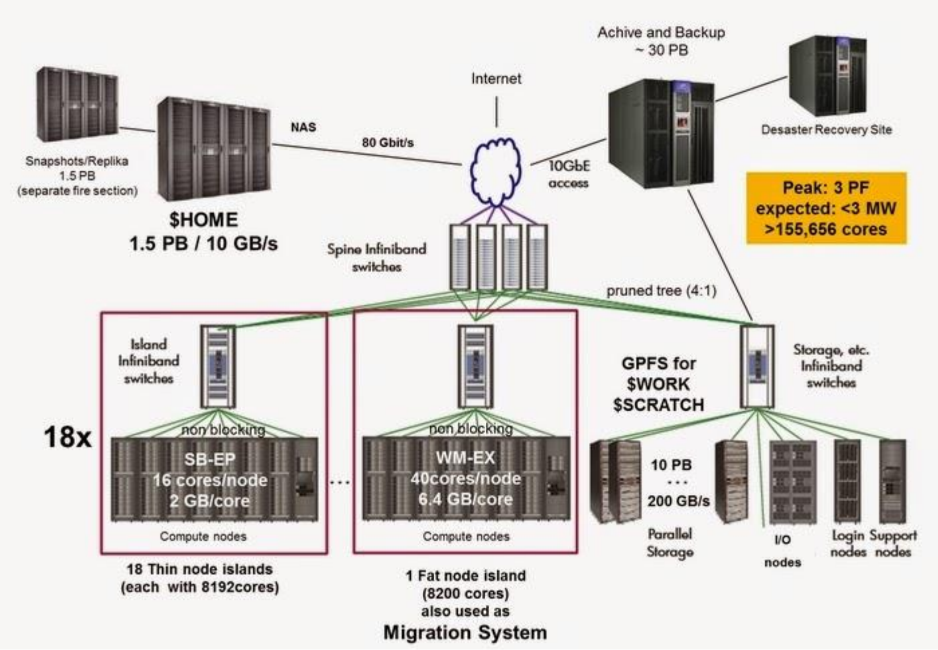
The most recent example is the Bluegene range but this architecture has been discontinued by IBM (although still appears in the TOP500).

IBM BG/Q system (Fermi)

- 163840 cores at 1.6GHz
- 10240 nodes (16 cores/node)
- 16 GB/node
- 5D Torus network
- 2 Pflops performance

Heterogenous, hybrid clusters

- The most common solution nowadays is a hybrid architecture consisting of different types of processor, accelerator or other devices in “islands” or partitions.
- Might be difficult to manage but more flexible since different application-types are likely to fit.
- Cineca Marconi is a hybrid system with 3 types of Intel processors (Broadwell, KNL, Skylake) in 3 partitions (A1, A2 and A3).
- Other clusters may have “fat” or “thin” nodes depending on memory available (e.g. LRZ in Munich).



Intel Xeon PHI

Low power device range based on Intel's Many Integrated Core (MIC) technology.

Large number of low frequency Pentium cores (e.g. 1.0 GHz) loosely connected on a chip with onboard memory.

The first commercially available Xeon Phi device Knight's Corner (KNC) could be used only as an accelerator.

Although runs standard FORTRAN and C/C++, difficult to obtain good performance.

Many application developers did not optimise codes for KNC.

Intel Knight's Corner

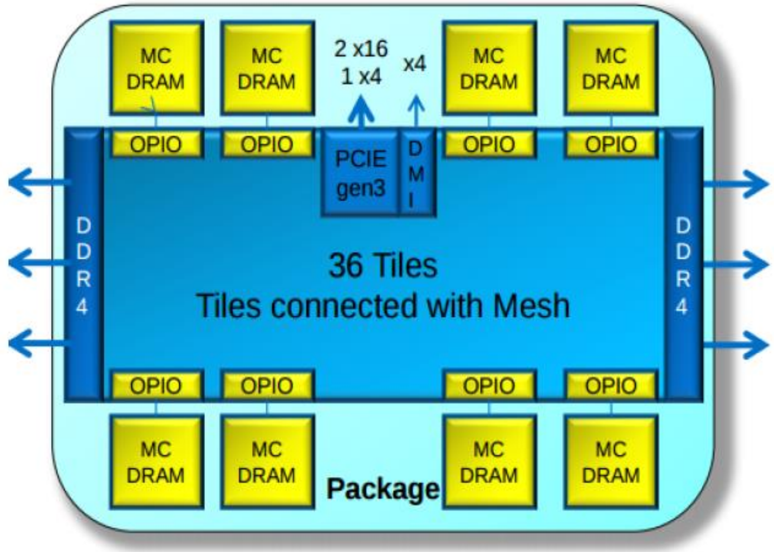
- 61 cores, 1.0-1.2 GHz
- 8-16 Gb RAM
- 512 bit vector unit
- 1-2 Tflops
- ring topology of cores
- With compiler option, runs standard FORTRAN or C (i.e. no CUDA or OpenCL necessary) and MPI.

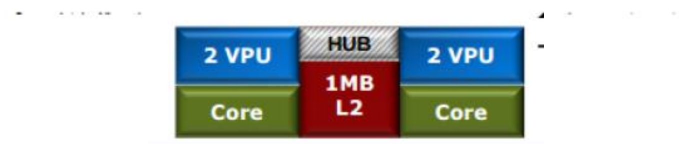
Intel Xeon PHI - Knight's Landing (KNL)

Major upgrade to KNC:

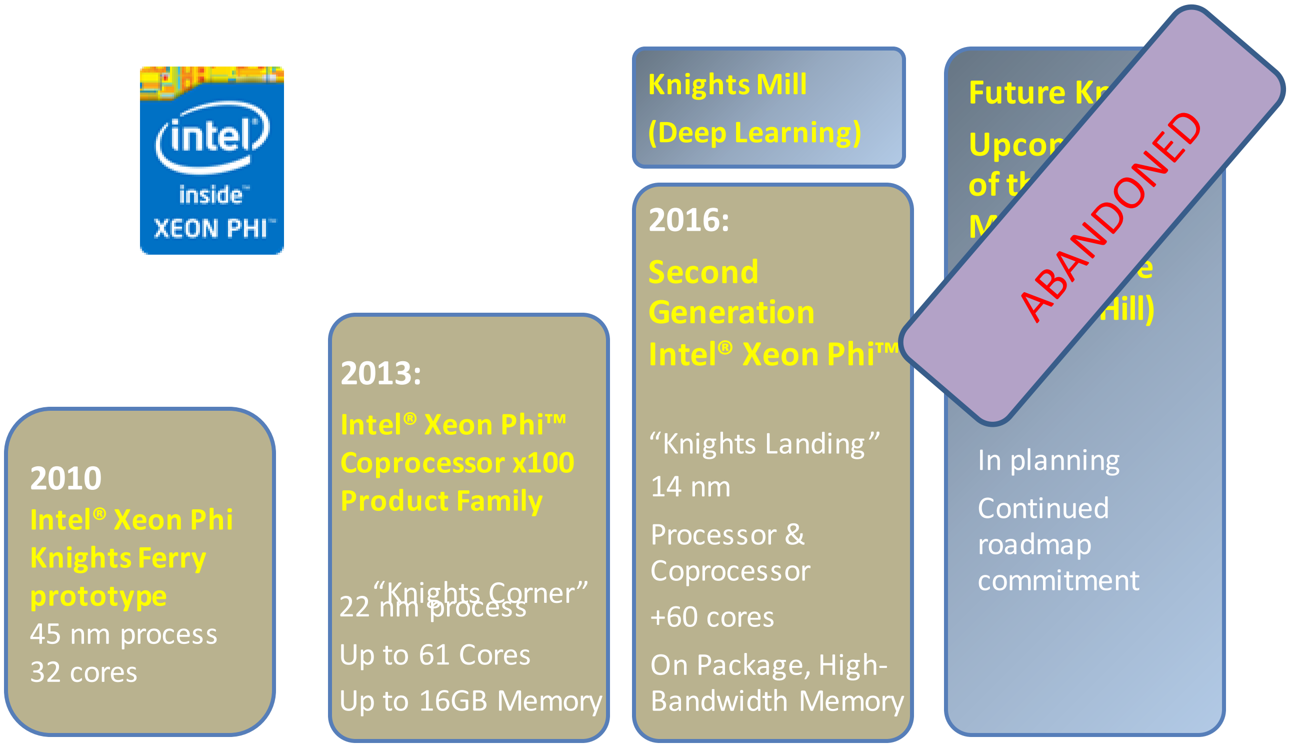
- Standalone, self-boot CPU.
- Upto 72 Silvermont-based cores (1.4 GHz)
- 4 threads, 2 AVX512 vector units/core (i.e. 272 threads in total).
- 2D Mesh interconnect
- 16 GB MCDRAM (High Bandwidth Memory) 400 Gb/s.
- Intel OmniPath on chip.
- 3 Tflops (DP) peak per package.

Binary compatible with other Intel processors but recommended to recompile to allow use of extended vector units.





Marconi A2 partition consists of 3600 nodes with a total performance of 13 Pflops.



*Per Intel's announced products or planning process for future products

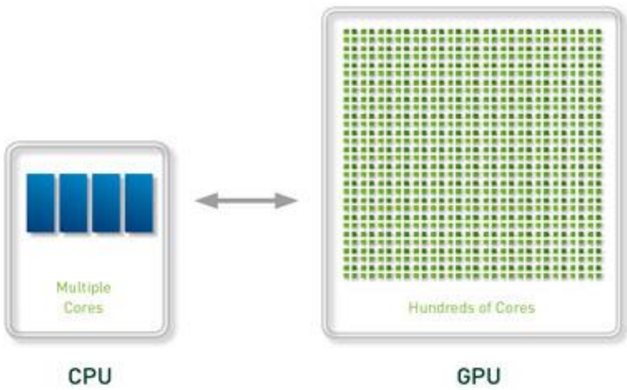
In November 2017 Intel announced that the Xeon Phi line would be abandoned due to "market and customer needs" (Intel).
Probably means that KNL had no market outside HPC.

Intel Xeon Phi lessons

Although no future for Intel Xeon Phi, the experience has been useful:

- with 68 cores/node you need only a few KNL nodes in order to test parallel scaling (i.e. speedup as no. of cores increase)
- emphasised the need for multi-threaded applications, good for testing programming paradigms which use this model (e.g mixed MPI+OpenMP).
- KNLs have been good for testing new high bandwidth memories (i.e. MCDRAM).

General Purpose GPUs, or simply GPUs (Graphical Processing Units), are devices designed for graphics processing which are used in non-graphical applications.



Became popular in HPC in 2006-2007 when Nvidia introduced CUDA, greatly simplifying the programming of such devices.

The design is different to a standard CPU, being based on hundreds or thousands of *stream processors*.

Used as an *accelerator*, attached to a conventional CPU (e.g. Intel). The idea is that parallel sections of the application are *offloaded* to the GPU, while the CPU handles the sequential sections. In certain circumstances can give large speed increases, compared to non accelerated code.

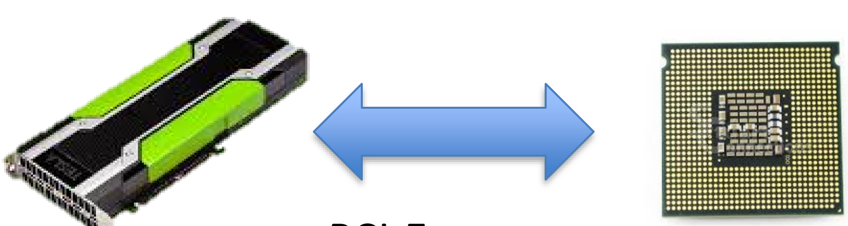
GPGPUs

Plus points:

- Can accelerate applications many times (e.g. 2x,3x even 20x or more)
- Performance in Flops/Watt and price in Flops/$ often much better than conventional CPUs.

Difficulties:

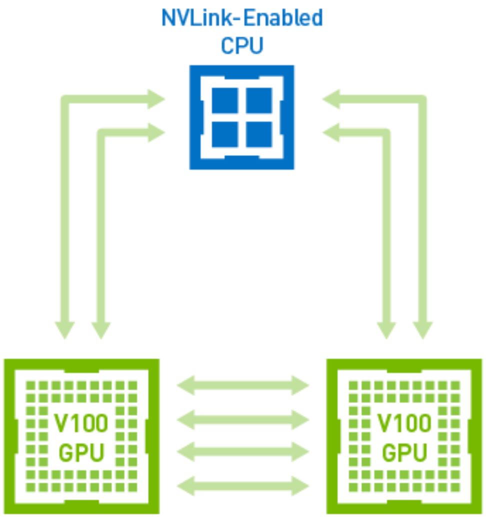
- Need to use CUDA (a C dialect) to get best performance, although other methods are becoming available (OpenACC, FORTRAN etc.).
- Porting to GPUs requires effort -> some applications do not have CUDA ports.
- PCI-e bus (connection) to the CPU is quite slow and device memory is limited (e.g. 16 Gb)



PCI-Express
(30 GB/s)

GPGPU advances

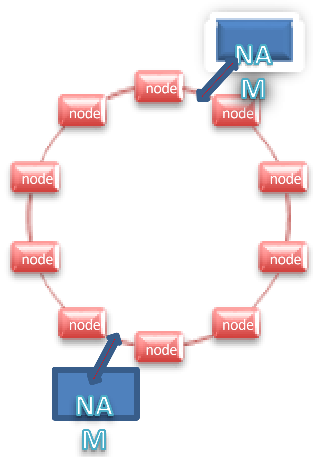
- Latest Nvidia devices (Pascal P100 and Volta V100) can use Nvlink which is upto 10X the speed of PCIe.
- Allows also fast GPU - GPU connections.
- Unified memory simplifies memory management of applications.
- Nvidia GPUs are becoming important in DEEP learning applications.



New memory technologies

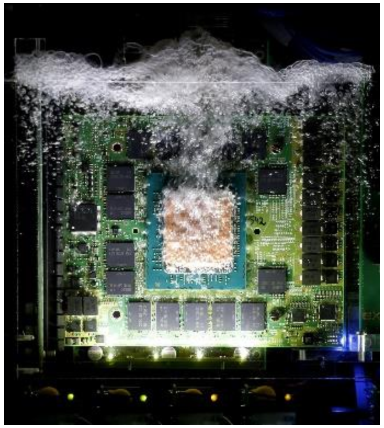
There have been rapid advances in disk and memory technologies, prompted by increasing needs in data storage:

- NVRAM (Non-Volatile Random Access Memory)
  - e.g. Flash memory. Retains information even when power switched off. Current use is for Solid State Disks (SSD) to replace “spinning disks” (i.e. conventional disks). SSD storage has particularly low latencies.
- HBM (High Bandwidth Memory)
  - e.g. MCDRAM in KNL, shared memory in Nivida P100, V100. High bandwidths (400-850 Gb/s) compared to standard DDR4 memory (e.g. 100 Gb/s). May need code changes to use.
- NAM (Network Attached Memory)
  - Memory attached directly to the network, rather than passing through the processor. Capable also of limited processing. Idea is to enable *near data computing*. Still in the research stage for HPC.



Cooling technologies

Need to reduce (and possibly re-use) waste heat has led to innovative cooling technologies such as hot water cooling instead of air and even immersing components in heat-removing solvents.



DEEP GreenICE Booster

SuperMUC uses 40 percent less energy than would be required by an equivalent air-cooled system.

The Road to Exascale

Multi-core, accelerated clusters with innovative cooling designs have brought petaflop class machines to everyday HPC users.

What do we need to go the next step, i.e. to have a computer able to reach *1 Exaflop*?

..the US Department of Energy (DoE) performed a survey

The DoE 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

but widely overoptimistic – still nowhere near Exascale (2018)

Exascale

In 2018 we still do not have a computer capable of 1 Exaflop - now expected around 2022-2025.

The trends are fairly clear:

– Large number of parallel processes or threads.
– With very high parallelism, hardware failures more problematic.
– Memory and I/O bandwidth not increasing at the same rate as concurrency.
– Memory/core decreasing.

Exascale prototypes

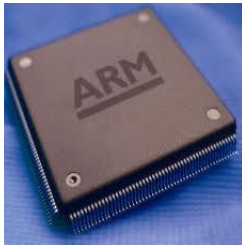
DAVIDE

IBM Power8 +GPU (P100) cluster with innovative cooling design.

Currently in pre-production at Cineca.

Mont Blanc

Prototype hardware and clusters based on ARM chips (Barcelona Supercomputing Centre).



But energy efficiency remains the main problem

- Energy consumption for most computers in the TOP500 is of the order of a few Gflops/W (Green Top500 Shoubu system B =17 Gflops/W for a peak of 840 Tflops).
- Scaling upto 1 Exaflop gives hundreds of MW power (~60 MW for Shoubu), i.e. impractical (and unethical) energy consumption.

In Europe 1 MWh costs between 30-50 Euros*

Also:

"This year [2018], electricity use at Bitcoin mining data centres is likely to exceed that of all Iceland's homes" (840 gigawatt hours)

(BBC, http://www.bbc.com/news/technology-43030677)

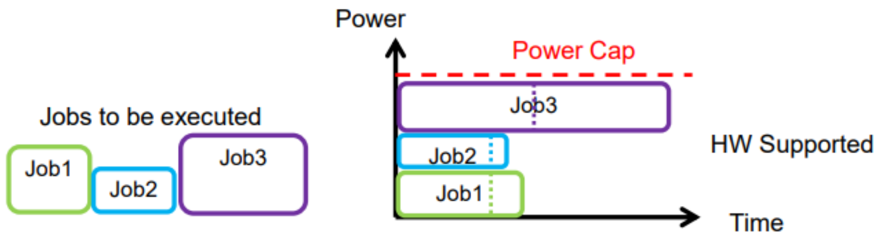
* The European Power sector in 2017, https://sandbag.org.uk/wp-content/uploads/2018/01/EU-power-sector-report-2017.pdf

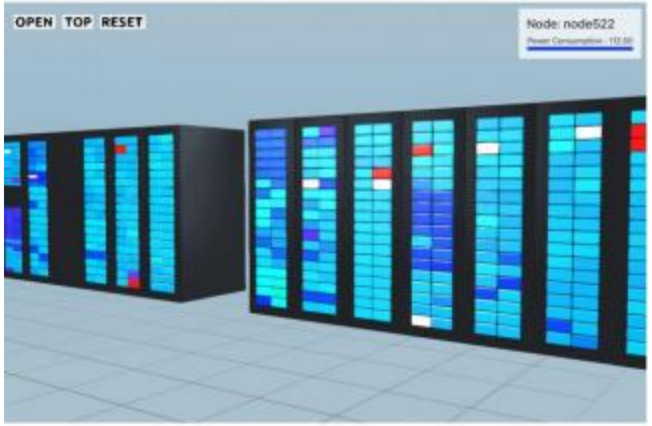
Measuring energy consumption

Considerable research efforts are being directed towards measuring the energy efficiency of applications, using sensors installed in the hardware.

Uses of energy measurements include:

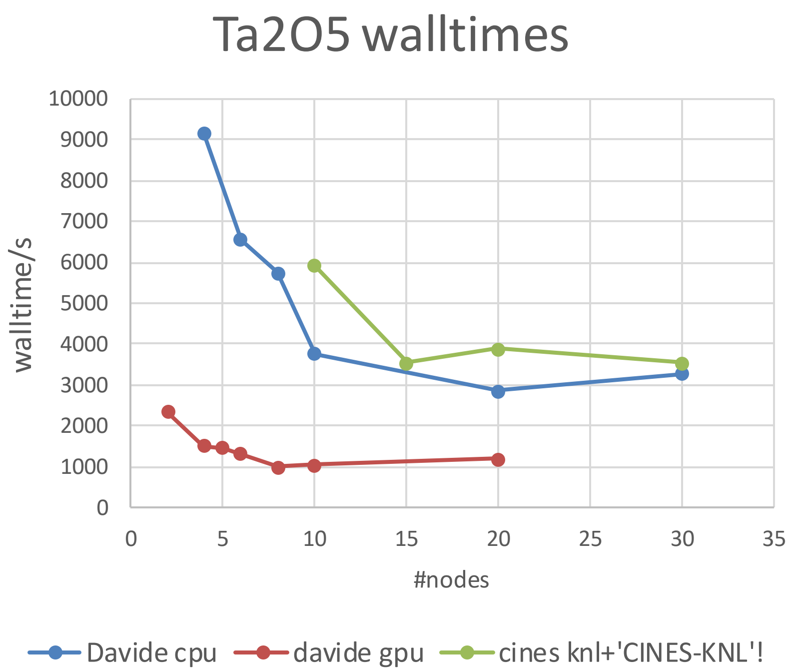
- Charging users based on energy consumption, rather than just wall time, to encourage good usage of energy resources.
- With fine-grain management possible to see which periods of a program's execution require less/more energy -> possible to cycle up/down the processor as a result.
- System managers can see which nodes are overheating or faulty or can deploy *power capping* based on estimates on what power batch jobs will use.

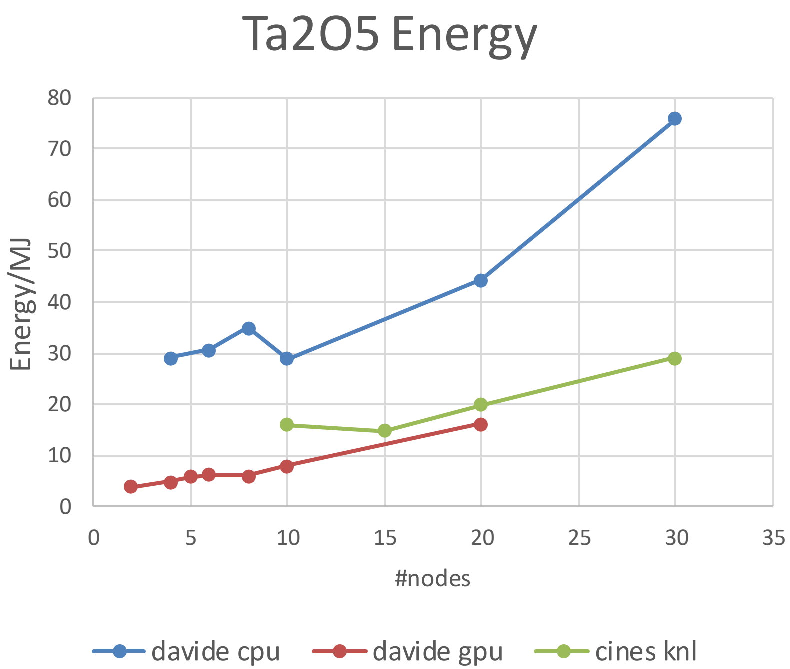




CINECA works closely with the group of Andrea Bartolini (UniBo) for energy measurements.

Energy benchmarks for Quantum Espresso





Tests in a PRACE project have shown GPUs to be more energy efficient than CPUs alone (Power8) or Intel KNLs.

Major Exascale Initiatives in US and Europe

CORAL (United States)

- Major US procurement program for pre-exascale computers based on different technologies.
- Systems range from 100-200 Pflops, expandable to 1 exaflop.
- Computer systems include:
  1. Sierra (LLNL), Power +Nvidia GPUs.
  2. Summit (Oak Ridge), Power9 +GPUs.
  3. Aurora (Argonne), originally Cray +Intel Knights Hill but now delayed to 2022-2023 (probably CPU+GPU).

European HPC Initiative

- Declaration signed on March 23 2017 signed by 7 countries (France, Germany, Italy, Luxembourg, the Netherlands, Portugal and Spain) supporting exascale computing by 2022-2023.
- Plans include the development of a European Processor.

SUMMIT (Oak Ridge National Laboratory)

SYSTEM CHARACTERISTICS	
Sponsor	US Department of Energy
Vendor	IBM
Architecture	9216 POWER9 22-core CPUs 27,648 Nvidia Tesla V100 GPUs
Storage	250PB
POWER	***15 MW***

Each node has over 500GB of coherent memory (high-bandwidth memory plus DDR4 SDRAM) which is addressable by all CPUs and GPUs plus 800GB of non-volatile RAM that can be used as a burst buffer or as extended memory. (Wikipedia)

SUMMIT

"For some AI applications, researchers can use less calculations than flops, potentially ***quadrupling*** *Summit's performance to* ***exascale levels****, or more than a billion billion calculations per second."*

(SUMMIT Website)

SUMMIT

“For some AI applications, researchers can use ***less calculations*** *than flops, potentially* ***quadrupling*** *Summit’s performance to* ***exascale levels****, or more than a billion billion calculations per second.”*

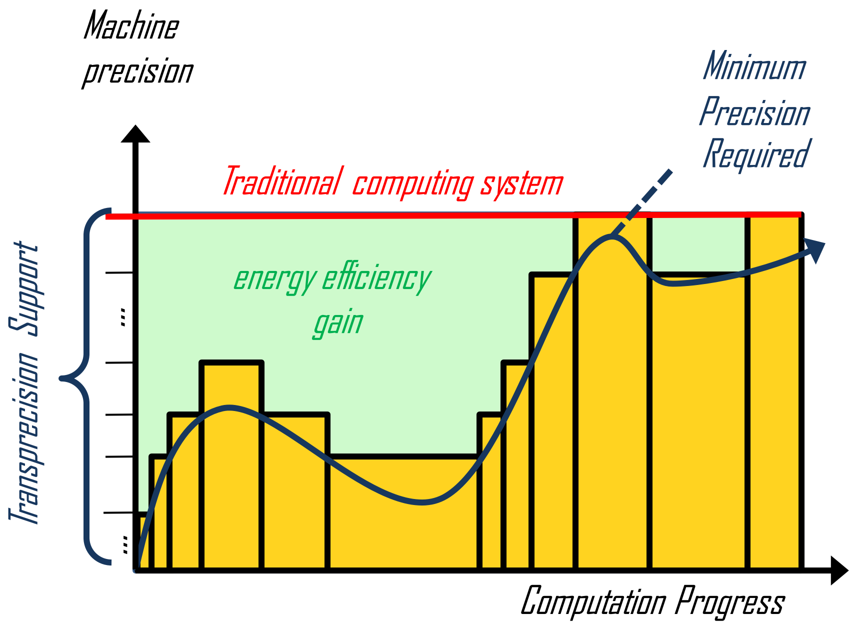
(SUMMIT Website)

Transprecision as a disruptive technology for Exascale

disruptive innovation - a major change in the status quo, as opposed to sustainable innovation which advances by incremental improvements.

Transprecision

- Applications are traditionally written in single or double precision but many problems do not require such accuracy (esp. Deep learning but also in other fields).
- The aim of projects such as Oprecomp, for example, is to apply transprecision principles in applications, runtime systems and hardware design to reduce energy consumption.



How can transprecision save energy?

For floating point variables reducing the precision (i.e. reducing the number of bits to store data) results in *lower memory usage*, speeding-up transfers and improving cache coherency.

For parallel systems where many data may be transferred between compute units (e.g. via message passing), significant speed-ups can be observed. The improved cache usage can also be important.

Assuming the result with reduced precision is acceptable, the increase in processing speed results in ***lower energy*** requirements.

Transprecision and HPC systems

How can we use transprecision on HPC systems?

Common HPC hardware usually only supports single (FP32) or double (FP64) variables.

In fact for Intel x86 variables (single or double) are stored as FP80 unless they are in the *vector unit* .

Half precision implementations are available for:

- Vector variables on x86 with Intel compilers;
- Nvidia GPUs, e.g. P100 or V100.
- ARM processors with GCC (but ARM rarely used for HPC).

Unless special hardware is used, any other precision must be emulated in software.

What happens when we lower the variable precision?

The effects will be algorithm-dependent, but the increase in error due to reduced precision could lead to problems in convergence or inaccurate results.

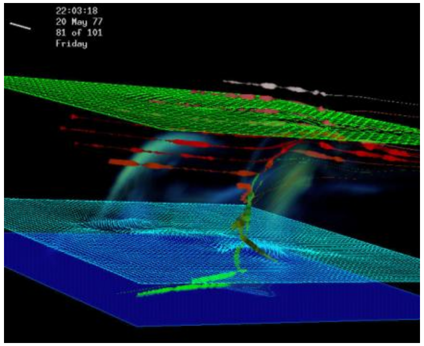
To investigate whether reduced precision can be used in programs the Oprecomp project has designed a benchmark suite ("micro-benchmarks") in the fields of Deep Learning, Big-data and Data Analytics and HPC and Scientific Computing.

The benchmark suite contains representative kernels of important algorithms (e.g. PageRank, SVM, FFT, etc) and the idea is to run each benchmark with various variable precisions and measure the performance and accuracy of the results.

To test variables other than single, double or half (GPU), the FloatX library can be used to represent arbitrary precisions for testing accuracy.

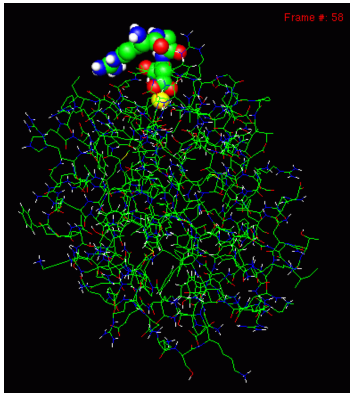
HPC and Scientific applications

Preliminary results obtained by Oprecomp for some benchmarks in the Deep Learning and Big Data-Data Analytics fields have shown that the kernels can be run at much lower precision with little loss in accuracy.



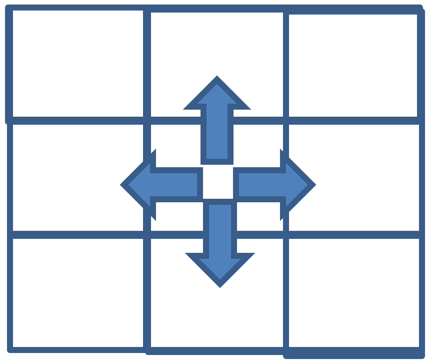
Scientific computing traditionally uses the highest precision (i.e. double or even quad) so this field could provide a stern test of transprecision.

In the following we present preliminary results for the stencil microbenchmark.



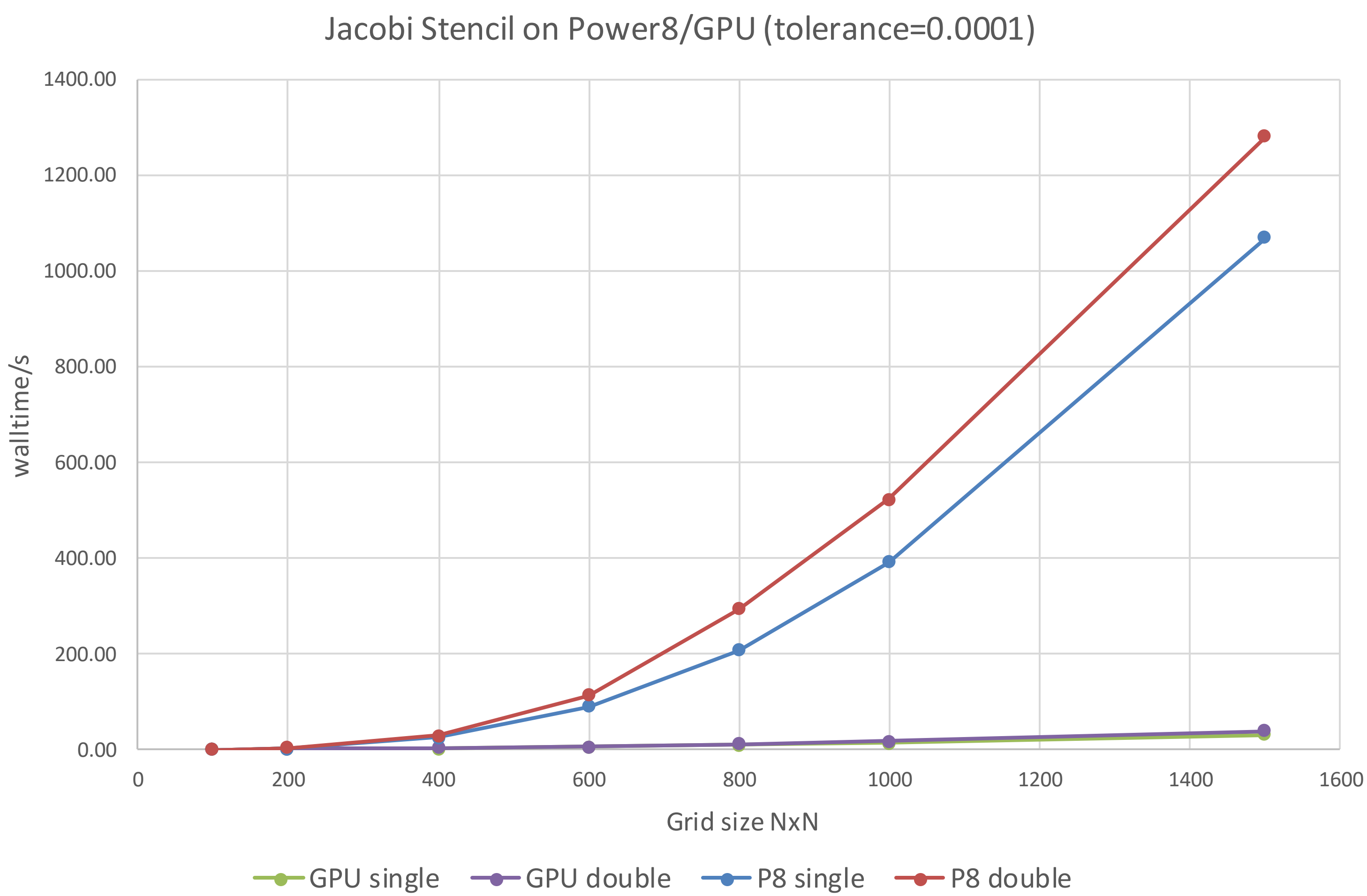
Jacobi Stencil

- Algorithm common in HPC programs for solving, for example, partial differential equations.
- Original C code with OpenMP directives tested on Intel Haswell and Power8 in single and double precision for a fixed convergence tolerance (1E-4).
- For all grid sizes tested, the calculations converged with a number of iterations independent on the precision (i.e. single or double).
- Code then converted to CUDA with C++ templates to allow variable types to be changed easily.



```
A(i,j)=0.25*(A(i-1,j)
+A(i+1,j)+A(i,j+1)+A(i,j-1))
```

Stencil Performance on P8/GPU



Acceleration on GPU P100 w.r.t to Power8 performances, but identical results.

Stencil with CUDA half precision (FP16)

Nvidia GPUs with compute capability >= 5.3 can perform half precision (FP16) data operations on the device.

Half datatypes with CUDA intrinsics allow conversion between half and float representations.

Calculations re-run on the Cineca Power8/P100 cluster (DAVIDE).

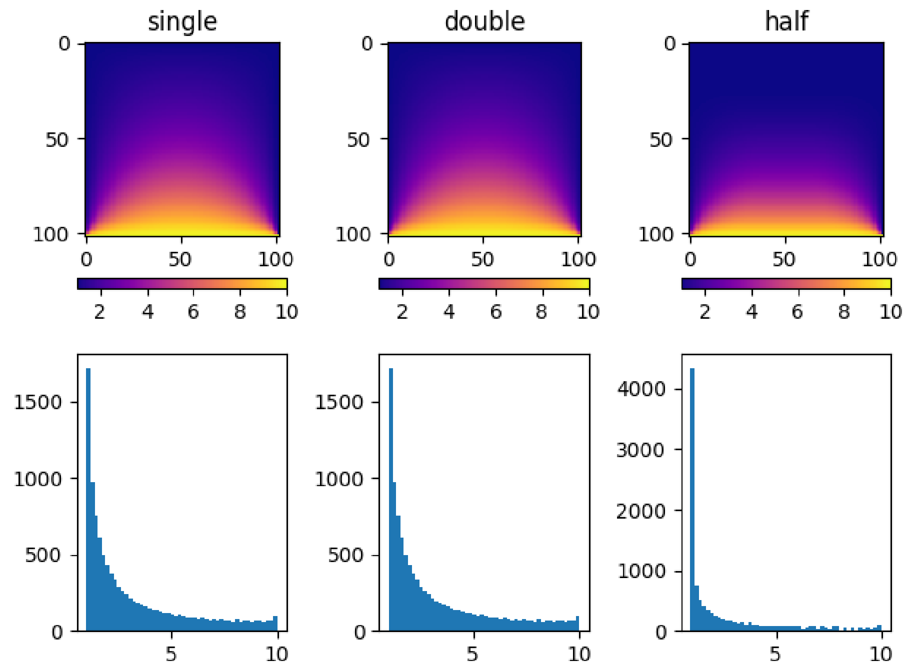
```
template<>
__global__
void stencil_sum<half>(half *grid,half *grid_new, const int nx,
const int ny)
{
 int index=blockIdx.x * blockDim.x +threadIdx.x; // global
thread id

const half hq =__float2half(0.25);

 if (index<nx*ny) {
   half kleft = grid[k-1];
   half kright = grid[k+1];
   half kdown = grid[k-ny2];
   half kup = grid[k+ny2];

#if __CUDA_ARCH__ >= 530
 half temp1 = __hadd(kleft,kright);
 half temp2 = __hadd(kdown,kup);
 half temp3  = __hmul(hq,__hadd(temp1,temp2));
 grid_new[k]=temp3;
```

For performance reasons, strictly speaking the FP16 variables should be packed into FP32 vectors, but with the non-contiguous memory access of the Jacobi stencil this is non-trivial.

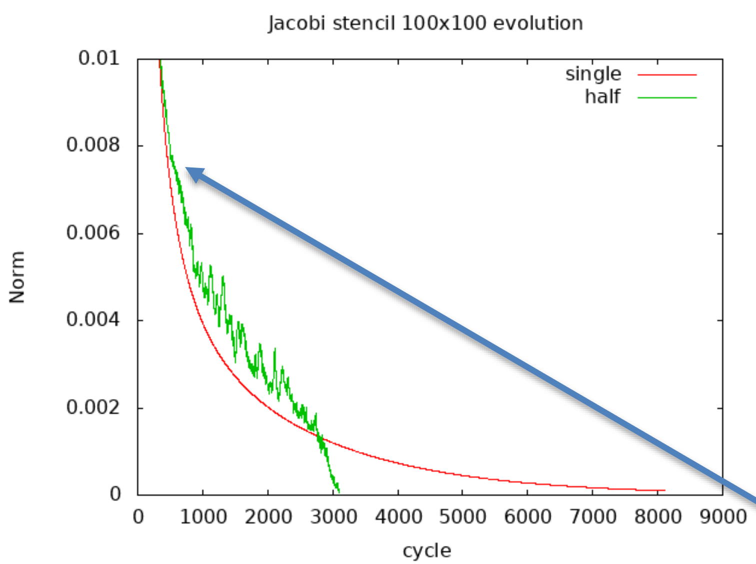
Stencil with half precision CUDA (FP16)

Convergence achieved with a significantly lower number of iterations.

Visualization and histograms of final grid, show different results between half and single or double.



Final configuration of 100x100 grid with tolerance=1e^{-4}



This type of algorithm is very sensititive to the variable precision. in our examples we clearly cannot use half precision

Current effort devoted to dynamically switching precision (e.g. from half to single) as convergence is approached.

Transprecision stencil legacy

- We have not yet done a full scan of all grid sizes (within the available memory) and tolerances but, at least in most cases, running stencil in single instead of double does not affect the accuracy.
- CUDA half precision on the other hand is not sufficient in most (if not all) cases and leads to inaccurate results.
- Two activities being currently pursued:
  1. Dynamically switching from half to single (or double) as convergence is approached;
  2. Using the FloatX library to test half or other precisions on standard CPUs.
- Preliminary results for FloatX suggest that also on CPUs half precision is too inaccurate.

Other scientific microkernels - FFT

Fast Fourier Transform (FFT)

- Algorithm performing DFT (Discrete Fourier Transform) used for example in transforming between time and frequency domains. Often the time determining kernel of many scientific codes (e.g. classical Molecular Dynamics).
- DFT is an N^2 problem but FFT allows an Nlog N solution.

$$\boldsymbol{X_k} = \sum_{\boldsymbol{n}=0}^{\boldsymbol{N}-1} \boldsymbol{x_n} \boldsymbol{e}^{-\frac{\boldsymbol{i}2\boldsymbol{\pi k n}}{\boldsymbol{N}}} \quad \boldsymbol{k} = 0, \dots, \boldsymbol{N} - 1$$

Transprecision work has just started but the error is found to be strongly dependent on the precision (e.g. single vs double). Now experimenting with FloatX to control more finely the effects of variable precision.

Summary

- HPC hardware has changed rapidly in a relative brief timeframe - from monolithic serial computers, to heterogeneous, massively parallel clusters with a wide range of different devices and memories. The current trend is for increasing parallelism and lower memory/core.
- Machine and Deep learning applications are strong drivers in HPC evolution, but progress towards Exascale is being strongly constrained by energy consumption.
- Incremental improvements in software and porting are insufficient for making significant differences in performance. Transprecision could represent *a disruptive technology* capable of increasing performance whilst maintaining energy requirements.
- Work with the Jacobi stencil though suggests that reducing the precision lower than single does not maintain accuracy. More sophisticated, dynamic approaches are probably required.