Materials modelling and the challenges of petascale and exascale

Andrew Emerson
Cineca Supercomputing Centre,
Bologna,
Italy

CAPACITIES

e-infrastructure

The project MMM@HPC is funded by the 7th Framework Programme of the European Commission within the Research Infrastructures with grant agreement number RI-261594.

26/09/2013

Contents

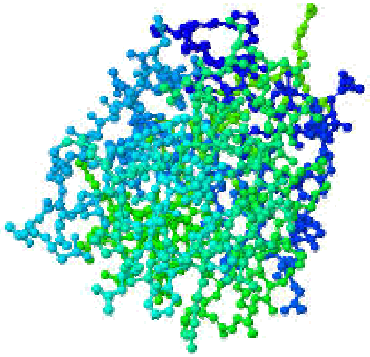
- Introduction to HPC
- HPC and the MMM@HPC project
- Petascale computing
- The Road to Exascale
- Observations

High Performance Computing (HPC). What is it ?

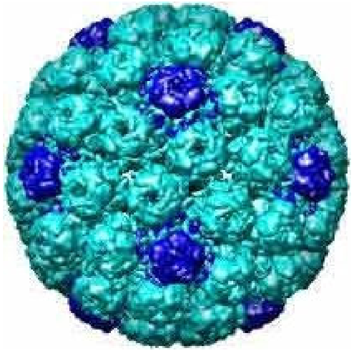
High-performance computing (HPC) is the use of parallel processing for running advanced application programs efficiently, reliably and quickly. The term applies especially to systems that function above a teraflop or 10^{12} floating-point operations per second.
(http://searchenterpriselinux.techtarget.com/definition/high-performance-computing)

A branch of computer science that concentrates on developing supercomputers and software to run on supercomputers. A main area of this discipline is developing parallel processing algorithms and software: programs that can be divided into little pieces so that each piece can be executed simultaneously by separate processors.
(WEBOPEDIA)

Advances due to HPC, e.g. Molecular dynamics

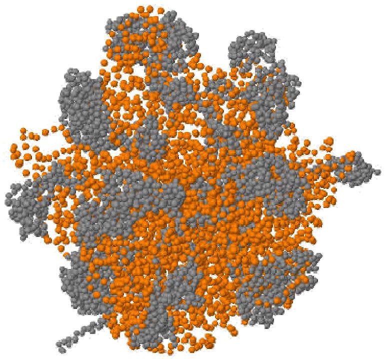


early 1990s. *Lysozyme, 40k atoms*

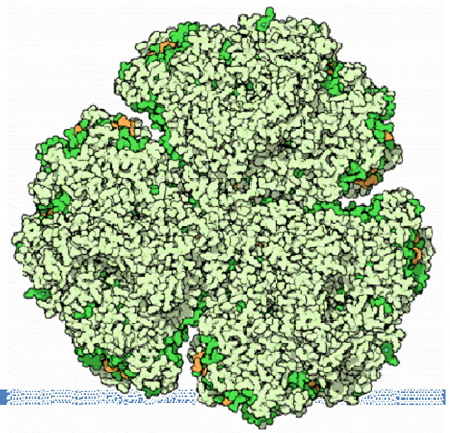


2006. Satellite tobacco mosaic virus (STMV). 1M atoms, 50ns

2008. Ribosome. 3.2M atoms, 230ns.



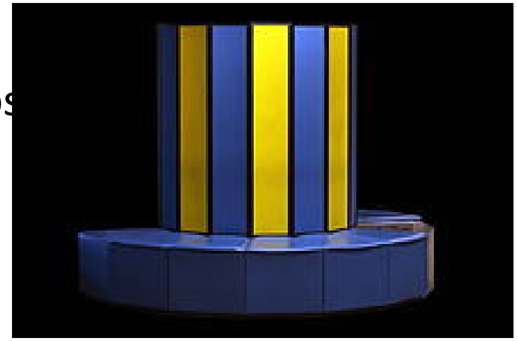
2011. *Chromatophore, 100M atoms (SC 2011)*



High Performance Computing

Cray-1 Supercomputer (1976)
80MHz , Vector processor → 250Mflops

Cray XMP (1982)
2 CPUs+vectors, 400 MFlops



“FERMI”, Bluegene/Q
168,000 cores
2.1 Pflops

High Performance Computing

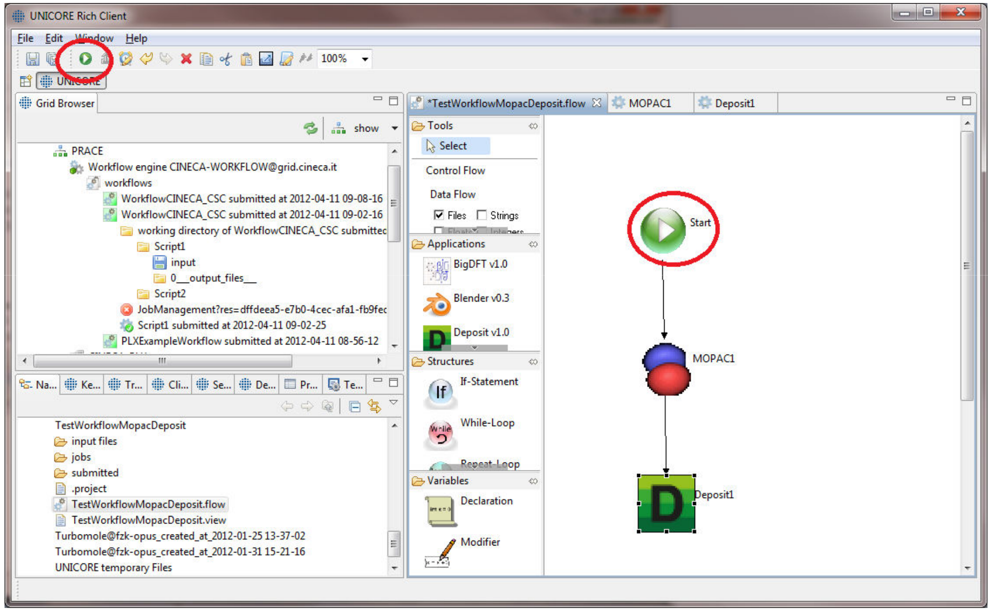
- For the application programmer HPC means introducing/modifying/optimizing
  - Scalar performance
  - Vectorisation
  - I/O usage
  - Memory usage and access
  - Parallelisation

From personal experience, HPC is not getting any easier.

Case Study - MMM@HPC Project

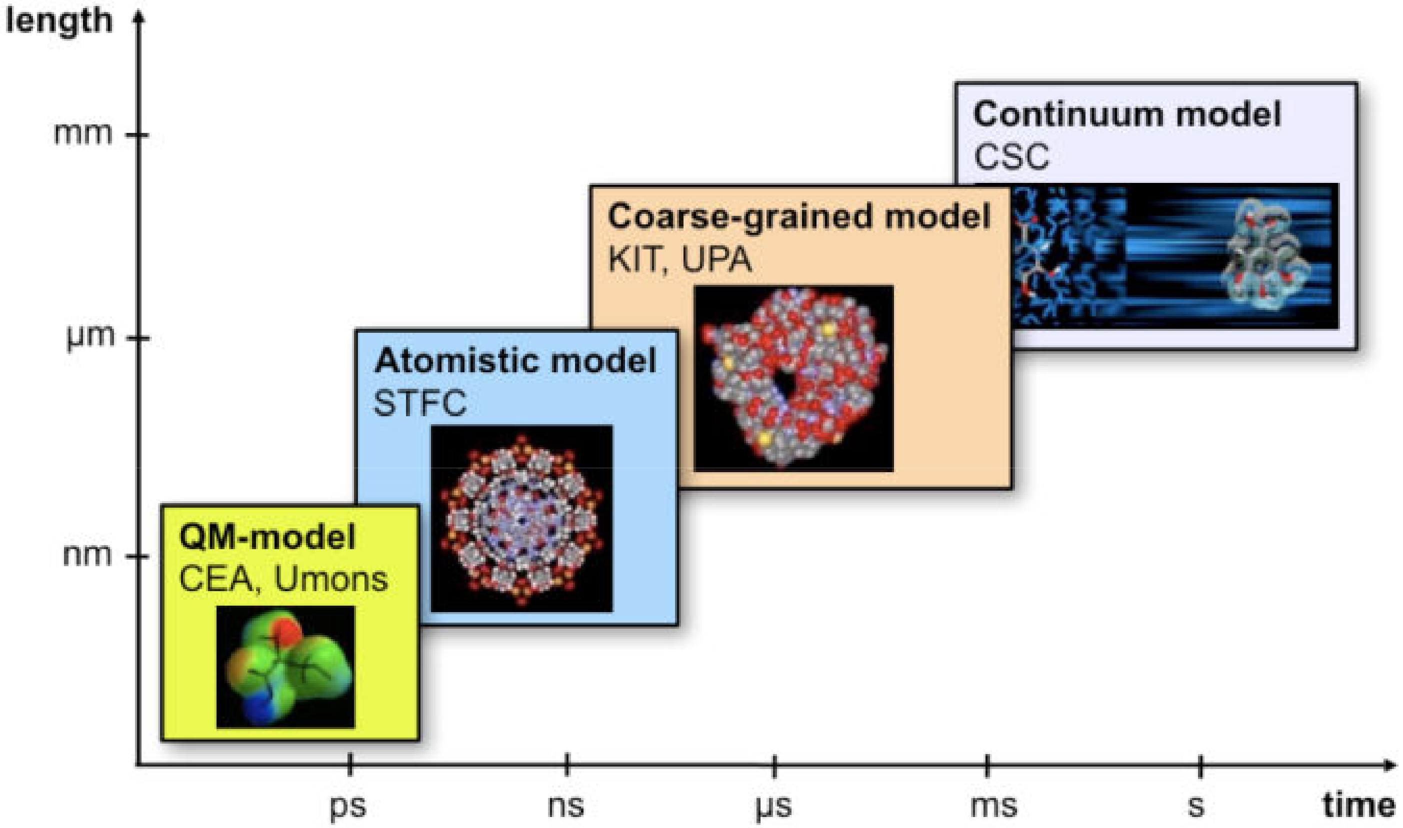
How do researchers use HPC resources ?

Example MMM@HPC project



The project MMM@HPC is funded by the 7th Framework Programme of the European Commission within the Research Infrastructures with grant agreement number RI-261594.

MMM@HPC Project



The aim of the project is to understand and design new materials for devices via simulations at different length and time scales.
Particular challenge since a wide variety of application codes used and on different architectures.

Key Applications of MMM@HPC

Application	Level of theory	Program models for parallel version (see main text)	Typical parallel scalability as used in project(*)
MOPAC (c)	Quantum Mechanics	Serial version only	-
Turbomole (c)	Quantum *Mechanics*	Global Arrays	low
ADF (c)	Quantum Mechanics	MPI	low
BigDFT	Quantum Mechanics	MPI and MPI/OpenMP	high
DL_POLY	Classical molecular dynamics	MPI	medium
LAMMPS	Classical molecular dynamics	MPI and OpenMP	medium
Elmer	Finite element	MPI	medium - high

(c) indicates a commercial code. (*) Parallel scalability is given in terms of the number of cores giving maximum performance such that low= <100 cores, medium=100-1000 cores, high > 1000 cores.

Petascale Computing

MMM@HPC

The Petascale

- Already with us.
- To achieve petaflop performances require very high parallelism but power consumption and heat dissipation are important engineering constraints.
- Stategies
  - Use many low, power cores (e.g BlueGene)
  - Accelerators (e.g. GPUs, MIC) with high performance, low power consumption
  - Sometimes both approaches used together resulting in hybrid structures.

Petascale Computing

MMM@HPC

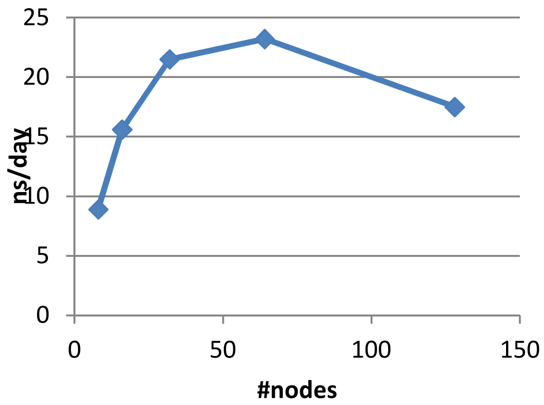
Features of European Petascale machines

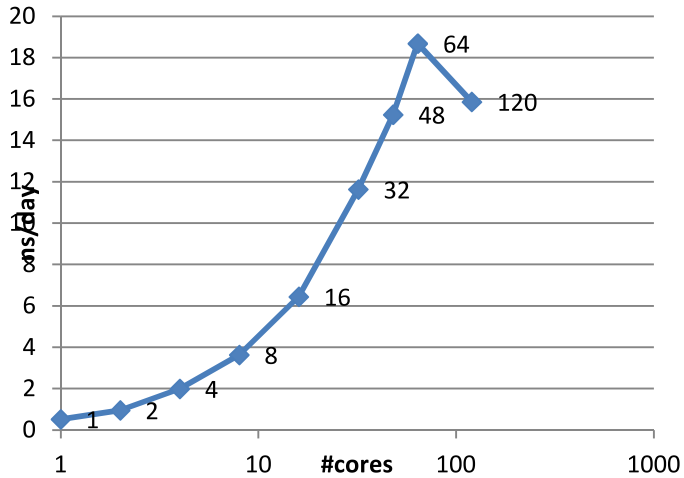
Supercomputer	Hardware (# total cores)	Minimum scaling requirements (PRACE Tier-0)
JUQUEEN (Juelich, Germany)	Bluegene/Q	8192
CURIE (CEA, France)	Bull Cluster (Hybrid)	512 (thin nodes), 2048 (fat nodes)
HERMIT (HLRS, Germany)	Cray XE6	2048
SuperMUC (LRZ, Germany)	IBM Dataplex (~155k)	4096
FERMI (CINECA, Italy)	Bluegene/Q (~163k)	2048
Mare Nostrum (BSC, Spain)	IBM Dataplex	1024

Petascale Computing

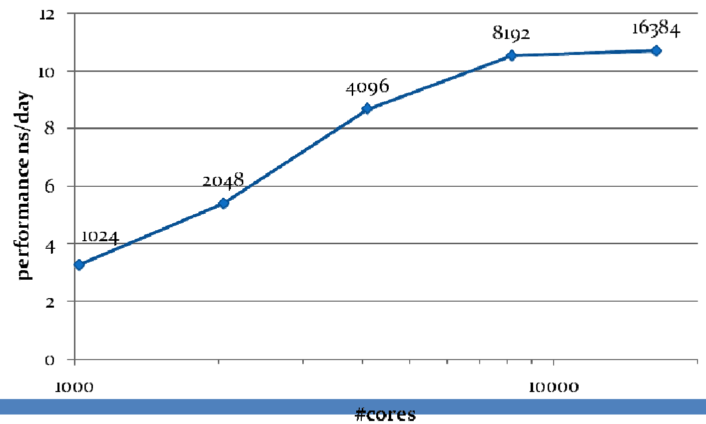
MMM @HPC

The problem is that most codes stop scaling:

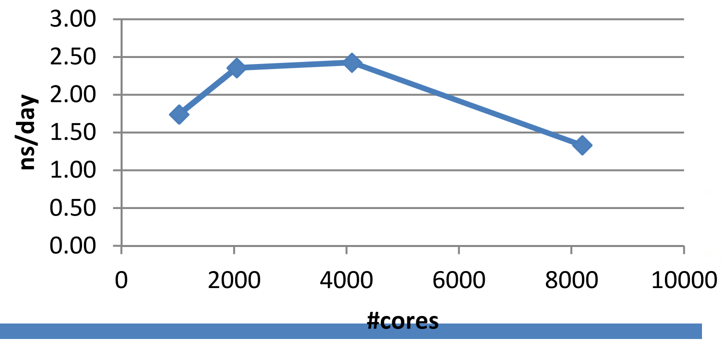




GROMACS BG/P scaling for d.kv12 membrane (1.8M atoms) on Jugene BG/P



GROMACS BG/P scaling for SPC water (0.5M molecules)



CINECA

Why do MD programs stop scaling ?

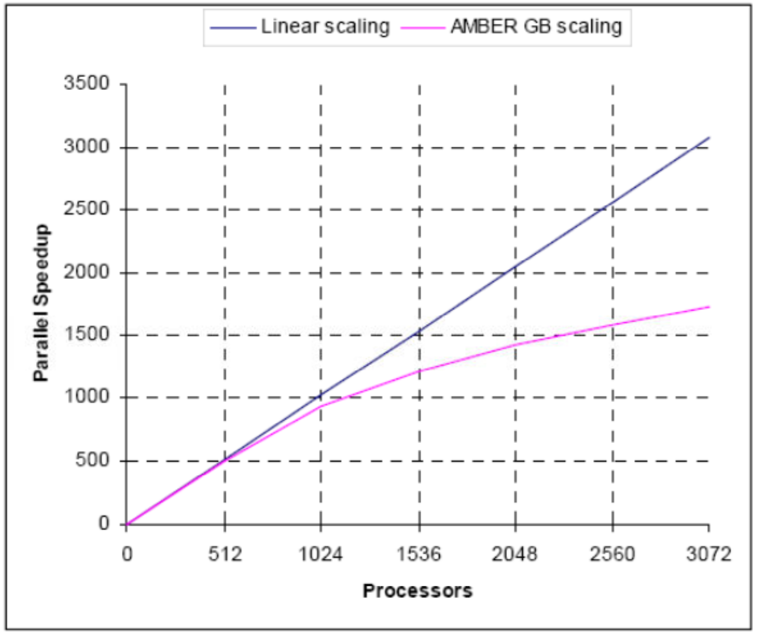


Figure 1. Parallel scaling of AMBER on Blue Gene. The experiment is with an implicit solvent (GB) model of 120,000 atoms (Aon benchmark).

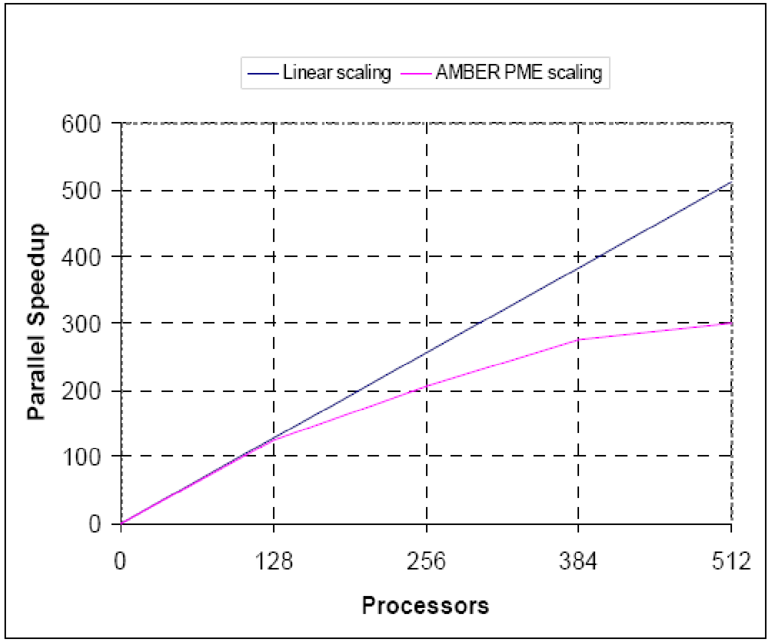


Figure 2. Parallel scaling of AMBER on Blue Gene. The experiment is with an explicit solvent (PME) model of 290,000 atoms (Rubisco).

Life Sciences Molecular Dynamics Applications on the IBM System Blue Gene Solution: Performance Overview, http://www-03.ibm.com/systems/resources/systems_deepcomputing_pdf_lsmdabg.pdf

Petascale Computing

Petascale challenges for projects such as MMM@HPC

- Parallel scalability
- Effort needed for GPU-enabling
- For Quantum mechanics (and often materials science) applications
  - Low memory/node
  - I/O performance

The Exascale

European Exascale Software Initiative (EESI)

The objective of this Support Action, co-funded by the European Commission is to build a European vision and roadmap to address the challenges of the new generation of massively parallel systems composed of millions of heterogeneous cores which will provide multi-Petaflop performances in the next few years and Exaflop performances in 2020.

A key consideration is of power efficiency: according to a DARPA study power consumption should not exceed 20MW.

The Challenge of Exascale

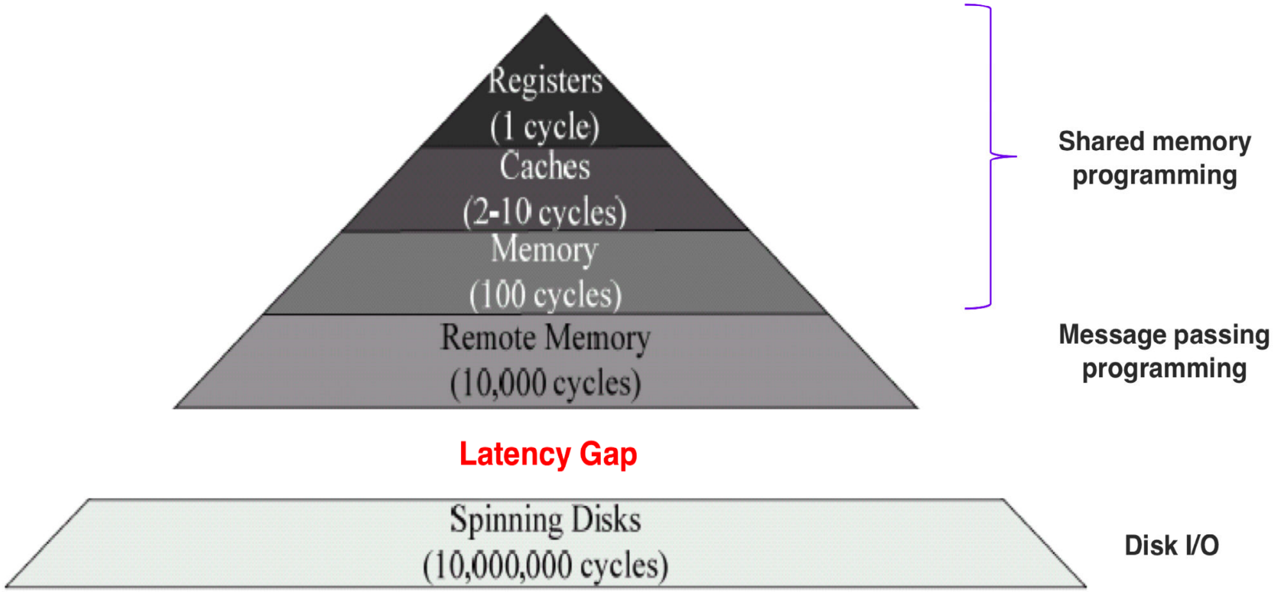
Some key findings of EESI

- Very many nodes, with many cores/node (perhaps thousands). Millions or billions of threads.
- High total system memory (petabytes) but with memory/core lower than is commonly found on present systems (e.g. lower by a factor of 10).

The Challenge of Exascale

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

The Challenge of Exascale



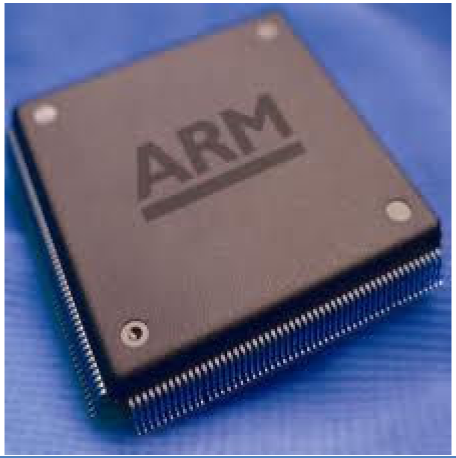
- The I/O subsystem is not keeping the pace with CPU
- Checkpointing will not be possible
- Reduce I/O
- On the fly analysis and statistics
- Disk only for archiving
- Scratch on non volatile memory ("close to RAM")

The Challenge of Exascale

MONT BLANC

The aim of the Mont Blanc project is to confront the problem of energy efficiency in Exascale systems by designing HPC systems based on low power components used in embedded systems and mobile devices such as ARM processors.

One objective is to design system using 30x less power than current systems.



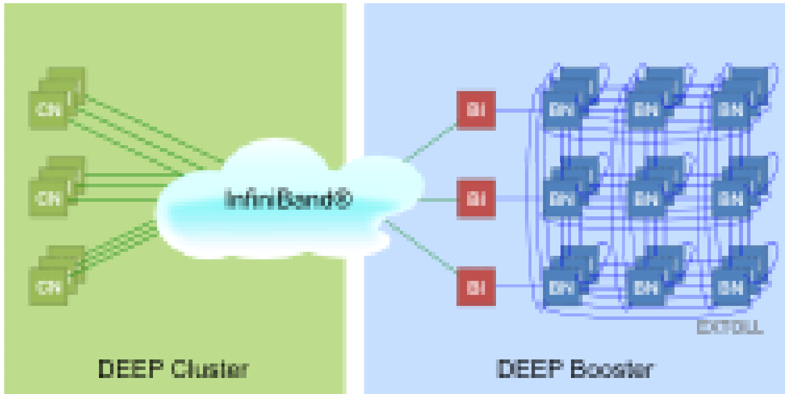
http://www.montblanc-project.eu/

The Challenge of Exascale

DEEP (Dynamical Exascale Entry Platform)

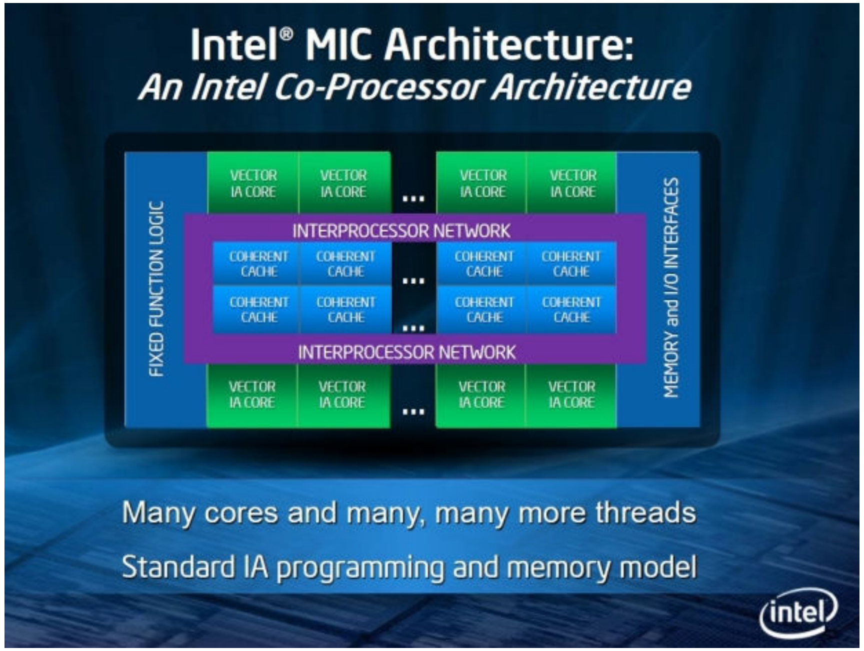
DEEP is an Exascale project funded by the EU 7th framework programme. The main goal is to develop a novel, Exascale-enabling supercomputing platform.

Prototype based on multi-core cluster linked to a "booster" part based on Intel's MIC technology.



Cluster-booster comm handled by Parastation MPI OmpSs to ease application deployment

Intel Many Integrated Core (MIC) Technology



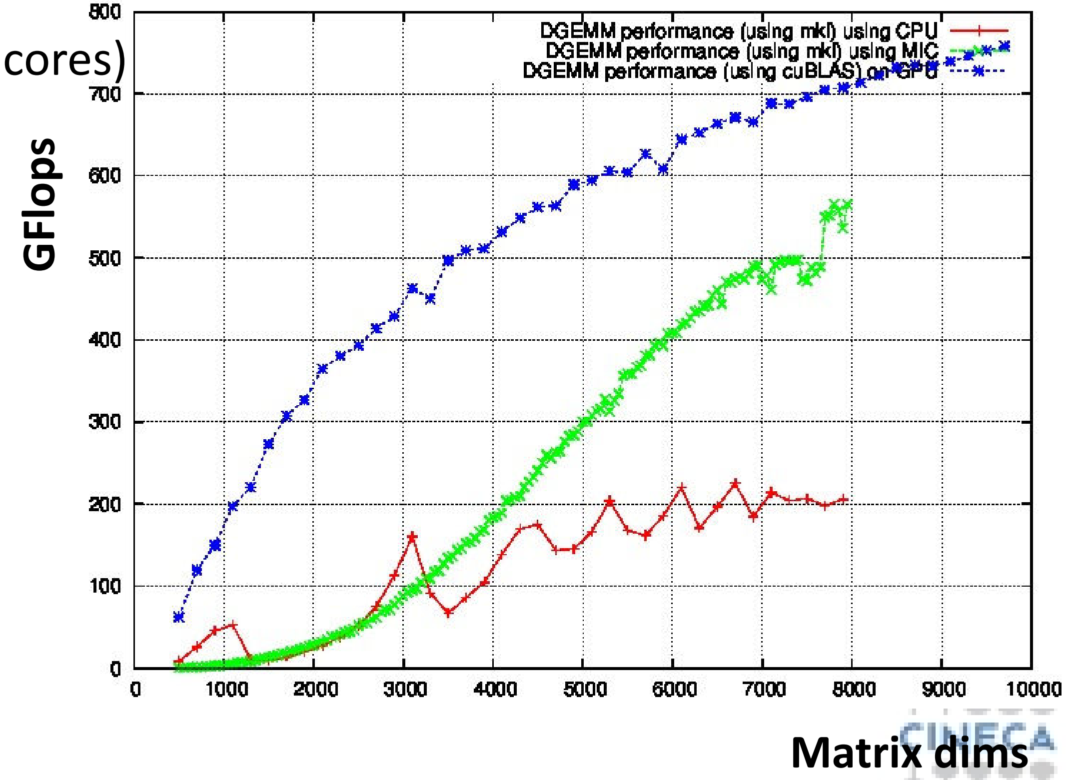
- Basically a “cluster on a chip” .
- Main advantage is that whatever works well on Intel chips works well on MIC as well (vectorization, many threads, cache use and so on).
- Performance/watt claimed to be better than GPUs.

NVIDIA vs Intel-Phi

Benchmark performed on EURORA (PRACE prototype of 64 nodes)

- 32 nodes:
  - 2 Xeon SandyBridge (2*8 cores)
  - 2 GPU NVIDIA K20
- 32 nodes:
  - 2 Xeon SandyBridge (2*8 cores)
  - 2 Intel Phi
- Matrix multiplication using BLAS (mkl and cuBLas)

- Red line: all 16 cores on the node
- Green line: one of the two MICs of the node
- Blue line: one of the two GPUs of the node



Perspectives

- Are we going to reach Exascale with current technologies?
- Look at trends from the top500 list.

HPC today, Successful technologies
Leading technologies path (Swim Lanes)

MMM@HPC

Multicore — **Multicore**: maintain complex cores, and replicate (SPARC64, x86, Power7): #4 K computer, #9 SuperMUC, #13 Power7

BlueGene — **Manycore/embedded:** use many simpler, low power cores from embedded (PowerPC in Bluegene): #3 Sequoia, #5 Mira, #7 Juqueen, #8 Vulcan, #12 FERMI

Hybrid — **Hybrid with accelerators**: Performance obtained using highly specialised processors from gaming/graphics market (accelerators: NVIDIA, Cell, IntelPhi): #1 Tianhe-2, #2 Titan, #6 Stampede, #10 Tianhe-1

John Shalf, NERSC

HPC today, Top12

MMM @HPC

Type	Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
Hybrid	1	National University of Defense Technology China	Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3120000	33862.7	54902.4	17808
Hybrid	2	DOE/SC/Oak Ridge National Laboratory United States	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560640	17590.0	27112.5	8209
BlueGene	3	DOE/NNSA/LLNL United States	Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1572864	17173.2	20132.7	7890
Multicore	4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12660
BlueGene	5	DOE/SC/Argonne National Laboratory United States	Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786432	8586.6	10066.3	3945
Hybrid	6	Texas Advanced Computing Center/Univ. of Texas United States	Stampede - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462462	5168.1	8520.1	4510
BlueGene	7	Forschungszentrum Juelich (FZJ) Germany	JUQUEEN - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458752	5008.9	5872.0	2301
BlueGene	8	DOE/NNSA/LLNL United States	Vulcan - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393216	4293.3	5033.2	1972
Multicore	9	Leibniz Rechenzentrum Germany	SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM	147456	2897.0	3185.1	3423
Hybrid	10	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 NUDT	186368	2566.0	4701.0	4040
Multicore	11	Total Exploration Production France	Pangea - SGI ICE X, Xeon E5-2670 8C 2.600GHz, Infiniband FDR SGI	110400	2098.1	2296.3	2118
BlueGene	12	CINECA Italy	Fermi - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	163840	1788.9	2097.2	822

HPC perspectives

Big questions for the next three years

- Multicore
  - IBM cancels “Blue Waters” contract
  - Maybe multicores with complex cores are nearing the end of the line?
- Manycores/embedded:
  - BlueGene is the last of the line
  - Maybe there will be no more large scale systems of this class?
- Hybrid/accelerated
  - More and more systems like these in TOP500
  - Efficient in power consumption

Prediction: all Top10 systems in 2015 will belong to the Hybrid category

John Shalf, NERSC

Conclusions

- Observations
  - Multiscale modelling requires powerful computer resources to run multiple applications in an HPC environment.
  - Adapting many disparate applications to Petascale computer systems which require very high parallelism, but have low memory/core and low I/O bandwidth, is proving to be quite labour intensive. Regardless of application parallelism, not all program inputs can be petascaled.
  - Considerable investment in Exascale now and in the near future, with power consumption as key criterion. Although challenging, clear effort to port applications in tandem with hardware advances.
  - Hybrid systems with GPU or MIC technology becoming more common.