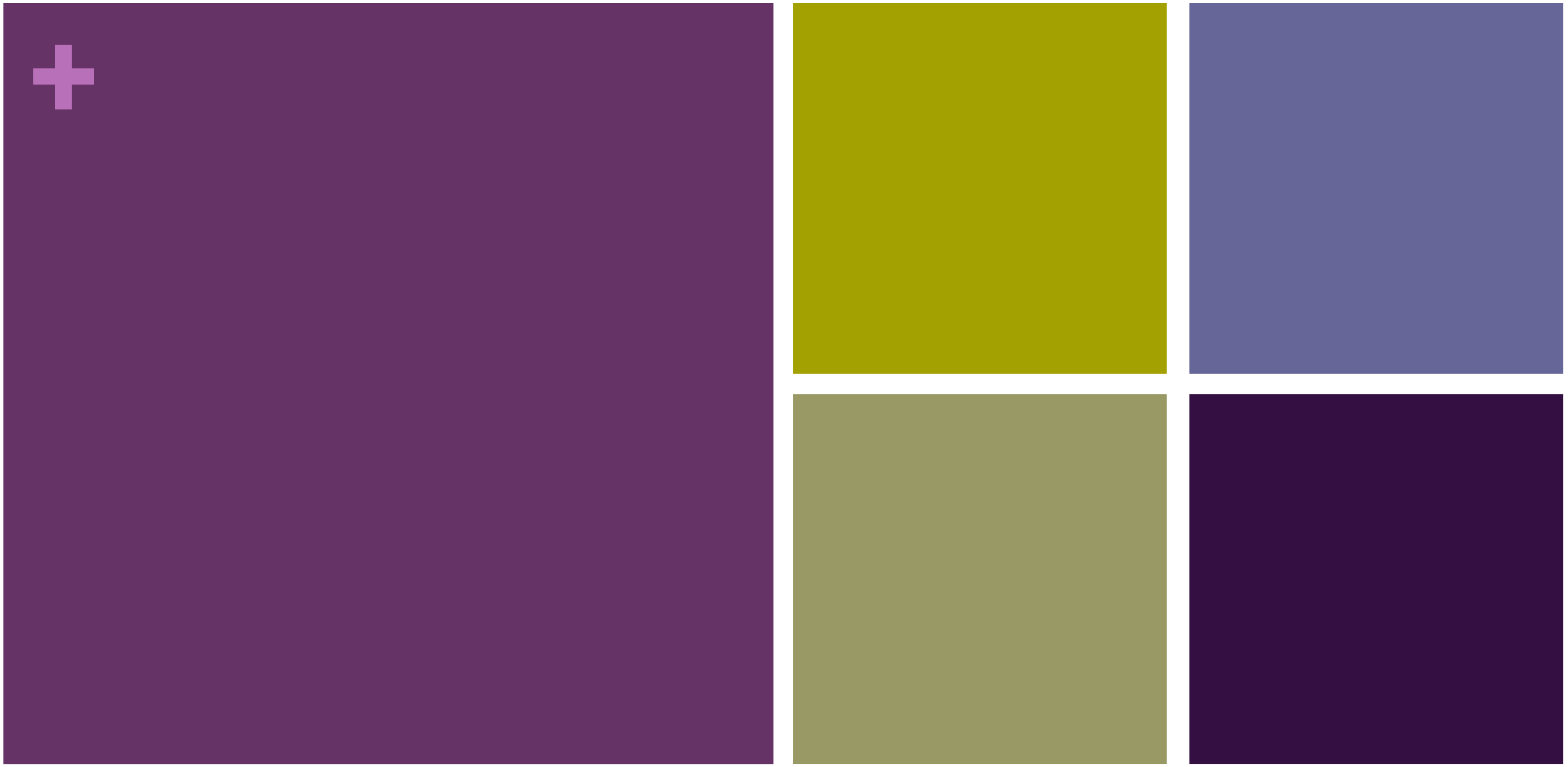


High Performance Computing - HPC

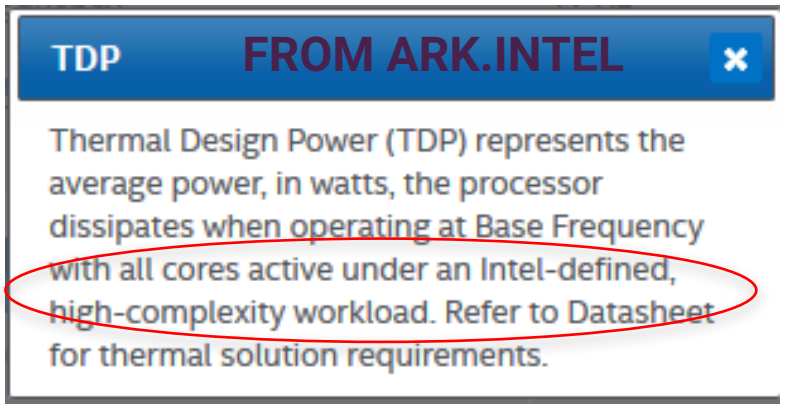
Daniele Cesini – INFN-CNAF
Enrico Mazzoni – INFN-Pisa

Glossary: GFLOPS and TDP

- GFLOPS: Billions of Floating Point Operations per second
  - Max GFLOPS of a system can be calculated using:

$$\text{GFLOPS} = \text{sockets} \times \frac{\text{cores}}{\text{socket}} \times \text{clock} \times \frac{\text{FLOPs}}{\text{cycle}}$$ (Clock in GHz)

- TDP: **Thermal Design Power** is the maximum amount of heat generated by the CPU that the cooling system in a computer is required to dissipate in **typical** operation (*)



(*) From wikipedia

Once upon a time....

The vector machines

- Serial number 001 Cray-1™
  - Los Alamos National Laboratory in 1976
  - $8.8 million
  - 80 MFLOPS scalar, 160/250 MFLOPS vector
  - 1 Mword (64 bit) main memory
  - 8 vector registers
    - 64 elements 64bit each
  - Freon refrigerated
  - 5.5 tons including the Freon refrigeration
  - 115 kW of power
    - 330 kW with refrigeration

(*) Source: Wikipedia

- Serial number 003 was installed at the National Center for Atmospheric Research (NCAR) in 1977 and decommissioned in 1989

Wireless technology inside

CRAY-XMP...Vector MultiProcessor

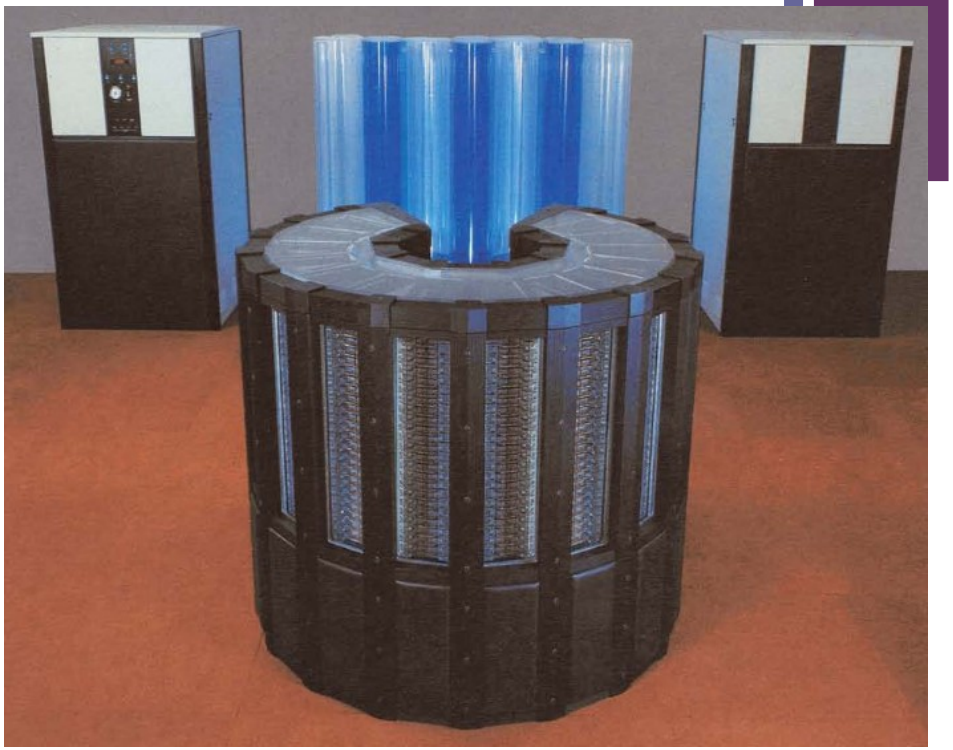
- 1982 CRAY-XMP 2 processors
  - 9.5 ns clock cycle (105 MHz)
  - 2x200MFLOPS
  - 2Mwords (64 bit) = 16MB

CRAY-XMP48 @ CERN in 1984
(*) source:
http://cerncourier.com/cws/article/cern/29150

- 1984 CRAY-XMP four processors
  - 800 MFLOPS
  - 8Mwords
    - 64 MB main memory
  - about US$15 million
    - plus the cost of disks!!!

The Cray-2

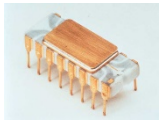
- The **Cray-2** released in 1985
  - 4 processors
  - 250MHz (4.1 ns)
  - 256 Mword (64bit) Main Memory
    - 2 GByte
  - 1.9 GFLOPS
  - 150 - 200 kW
  - Fluorinet cooling
  - 16 sq ft floor space
  - 5500 pounds
  - About $17 million



An inert fluorocarbon liquid circulates in the mainframe cabinet in direct contact with the integrated circuit packages. This liquid immersion cooling technology allows for the small size of the CRAY-2 mainframe and is thus largely responsible for the high computation rates.

(*) http://archive.computerhistory.org/resources/text/Cray/Cray.Cray2.1985.102646185.pdf

Microprocessors



The Intel 4004
1971

The Intel 8080
1974

The Intel 8086
1978

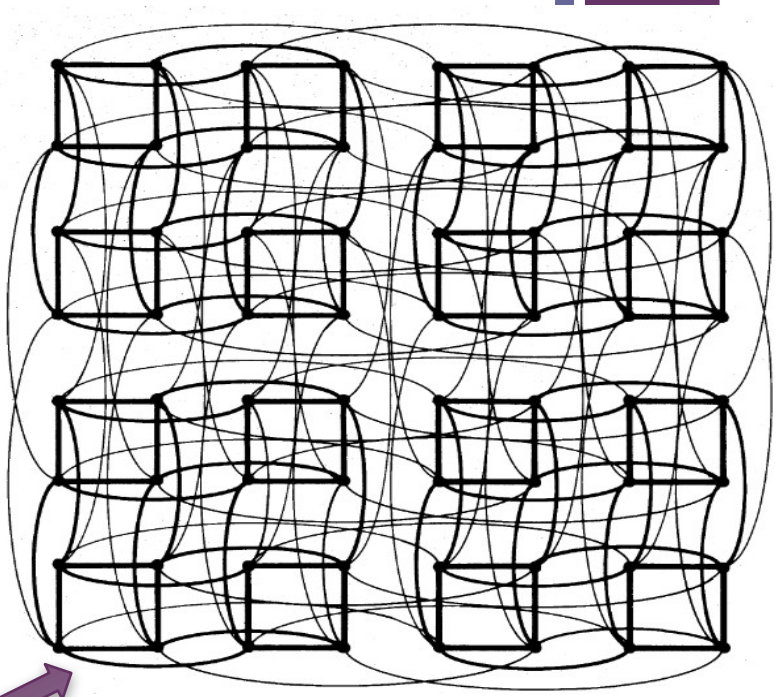
The Intel 80286
1982

The Intel 80486
1989

In the 80s started the replacement of vector supercomputers with Massively Parallel Processors (MPP) and Clusters assembled from a larger number of lower performing microprocessors

The attack of the Killer Micros

Taken from the title of Eugene Brooks' talk "**Attack of the Killer Micros**" at Supercomputing 1990



- Caltech Cosmic Cube
  - By Charles Seitz and Geoffrey Fox in1981
  - 64 Intel 8086/8087 processors
  - 128 kB per processor
  - 6 dimensions hypercube

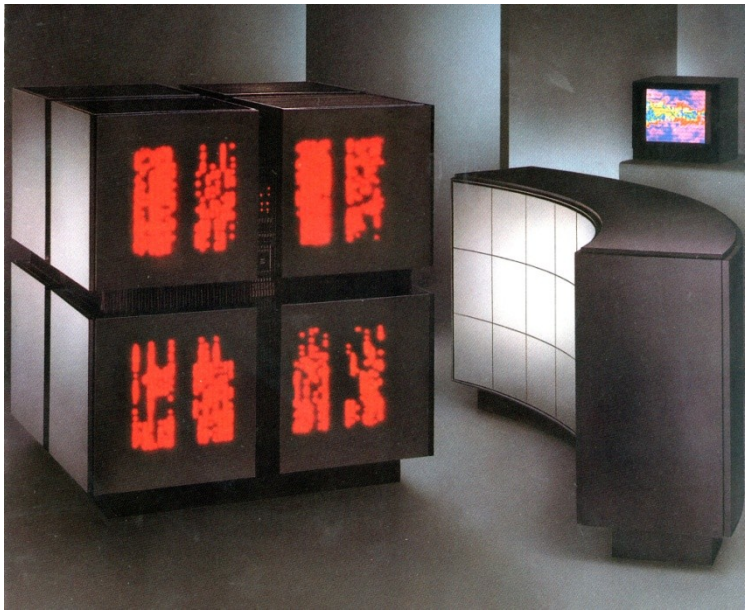
(*)http://calteches.library.caltech.edu/3419/1/Cubism.pdf

Massively Parallel Processor (MPP)

- A single computer with many networked processors
  - Specialized interconnect networks
  - Low latency interconnection
- Up to thousands of processors
- Some examples
  - Connection Machines (CM-1/2/200/5)
  - Intel Paragon
  - ASCI series
  - IBM SP
  - IBM BlueGene

Thinking Machines

- 1985: Thinking Machines introduces the connection Machine CM-1
- Connection Machine CM-200
  - maximum configuration of 65536 1-bit CPUs(!)
  - floating-point unit for every 32 1-bit CPUs
  - A cube composed of 8 cubes
    - each cube contains up to 8096 processors
  - (The curved structure is a Data Vault - a disk array)
  - 40 GFLOPS peak
- 1991: CM-5
  - **Featured in “Jurassic Park”**



(*) Sources:
http://www.corestore.org/cm200.htm
http://www.new-npac.org/projects/cdroms/cewes-1999-06-vol1/nhse/hpccsurvey/orgs/tmc/tmc.html
http://en.wikipedia.org/wiki/Thinking_Machines_Corporation

Intel Paragon MPP

- Launched in 1993
- Up to 2048 (later 4000) Intel i860 RISC microprocessors
  - Connected in a 2D grid
  - Processors @ 50 MHz
- World most powerful supercomputer in 1994
  - Paragon XP/S140
  - 3680 processors
  - 184 GFLOPS peak

(*) Source: Wikipedia

ASCI Red MPP

TERA SCALE

- 1996 At Sandia Laboratories
- Based on the Paragon architecture
- Fastest supercomputer from 1997 to 2000
  - 1.4 TFLOPS (peak) in 1997
    - 9152 cores
  - 3,2 TFLOPS (peak) in 1999
    - 9632 cores
- 1st supercomputer above 1 TFLOPS

(*) Source: Wikipedia

IBM BlueGene/Q MPP

PETA SCALE

- **Trading the speed of processors for lower power consumption**
- System-on-a-chip design. All node components were embedded on one chip
- A large number of nodes
- 5D xTorus interconnect
- Compute chip is an 18 core chip
  - The 64-bit PowerPC A2
  - 4-way simultaneously multithreaded per core
  - 1.6 GHz
  - a 17th core for operating system functions
  - chip manufactured on IBM's copper SOI process at 45 nm.
  - 204.8 GFLOPS and 55 watts per processor
- Up to 20 PFLOPS (peak)
  - 16384 cores

(*) Source: Wikipedia

Clusters

[a cluster is a] parallel computer system comprising an integrated collection of independent nodes, each of which is a system in its own right, capable of independent operation and derived from products developed and marketed for other stand-alone purposes

Dongarra et al. : "High-performance computing: clusters, constellations, MPPs, and future directions", Computing in Science & Engineering (Volume:7 , Issue: 2)

(*) Picture from: http://en.wikipedia.org/wiki/Computer_cluster

Top500.org

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,463,616	174.70	255.75	5,610
5	**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	2,414,592	148.60	200.79	10,096
6	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
7	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371

Nov 2022 List

Frontier

© Wikipedia

- Frontier, or OLCF-5, is the world's first and fastest exascale supercomputer, hosted at the Oak Ridge Leadership Computing Facility (OLCF) in Tennessee, United States
- It is based on the Cray EX and is the successor to Summit (OLCF-4).
- As of March 2023, Frontier is the world's fastest supercomputer.
- Frontier achieved an Rmax of 1.102 exaFLOPS
- Frontier uses 9,472 AMD Epyc 7453s "Trento" 64 core 2 GHz CPUs (606,208 cores) and 37,888 Radeon Instinct MI250X GPUs (8,335,360 cores).

Active	Deployment: Sep. 2021 Completion: May 2022
Operators	Oak Ridge National Laboratory and U.S. Department of Energy
Location	Oak Ridge Leadership Computing Facility
Power	21 MW
Operating system	HPE Cray OS
Space	680 m^2 (7,300 sq ft)
Speed	1.102 exaFLOPS (Rmax) / 1.685 exaFLOPS (Rpeak)[1]
Cost	US$600 million (estimated cost)
Purpose	Scientific research and development
Website	www.olcf.ornl.gov/frontier/

Summit OLCF-4 supercomputer

summit

Sponsors	U.S. Department of Energy
Operators	IBM
Architecture	9,216 POWER9 22-core CPUs 27,648 Nvidia Tesla V100 GPUs[1]
Power	13 MW[2]
Storage	250 PB
Speed	200 petaflops (peak)
Purpose	Scientific research
Web site	www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

600 GB of coherent memory addressable by all CPUs and GPUs

800 GB of non-volatile RAM that can be used as a burst buffer or as extended memory.

Non-blocking fat-tree topology using a dual-rail Mellanox EDR InfiniBand interconnect for both storage and inter-process communications traffic

United States Department of Energy awarded a $325 million contract in November, 2014 to IBM, Nvidia and Mellanox for the construction of Summit and Sierra

- Summit is tasked with civilian scientific research and is located at the Oak Ridge National Laboratory in Tennessee.
- Sierra is designed for nuclear weapons simulations and is located at the Lawrence Livermore National Laboratory in California.
- Summit is estimated to cover the space of two basketball courts and require 136 miles of cabling

Fugaku Supercomputer

Active	From 2021
Sponsors	MEXT
Operators	RIKEN
Location	RIKEN Center for Computational Science (R-CCS)
Architecture	158,976 nodes Fujitsu A64FX CPU (48+4 core) per node Tofu interconnect D
Operating system	Custom Linux-based kernel
Memory	HBM2 32 GiB/node
Storage	1.6 TB NVMe SSD/16 nodes (L1) 150 PB shared Lustre FS (L2)[1] Cloud storage services (L3)
Speed	442 PFLOPS (per TOP500 Rmax), after upgrade; higher 2.0 EFLOPS on a different mixed-precision benchmark
Cost	US$1 billion (total programme cost)[2][3]
Ranking	TOP500: 1, June 2020
Web site	www.r-ccs.riken.jp/en/fugaku
Sources	Fugaku System Configuration

- The supercomputer is built with the Fujitsu A64FX microprocessor.
  - Based on the ARM version 8.2A processor architecture
  - Fugaku was aimed to be about 100 times more powerful than the K computer
    - i.e. a performance target of 1 exaFLOPS
- The initial (June 2020) configuration of Fugaku used 158,976 A64FX CPUs joined together using Fujitsu's proprietary torus fusion interconnect.
- An upgrade in November 2020 increased the number of processors
  - **To reach 442 petaFLOPS**
- **1 Billion $ total cost**

Sunway TaihuLight

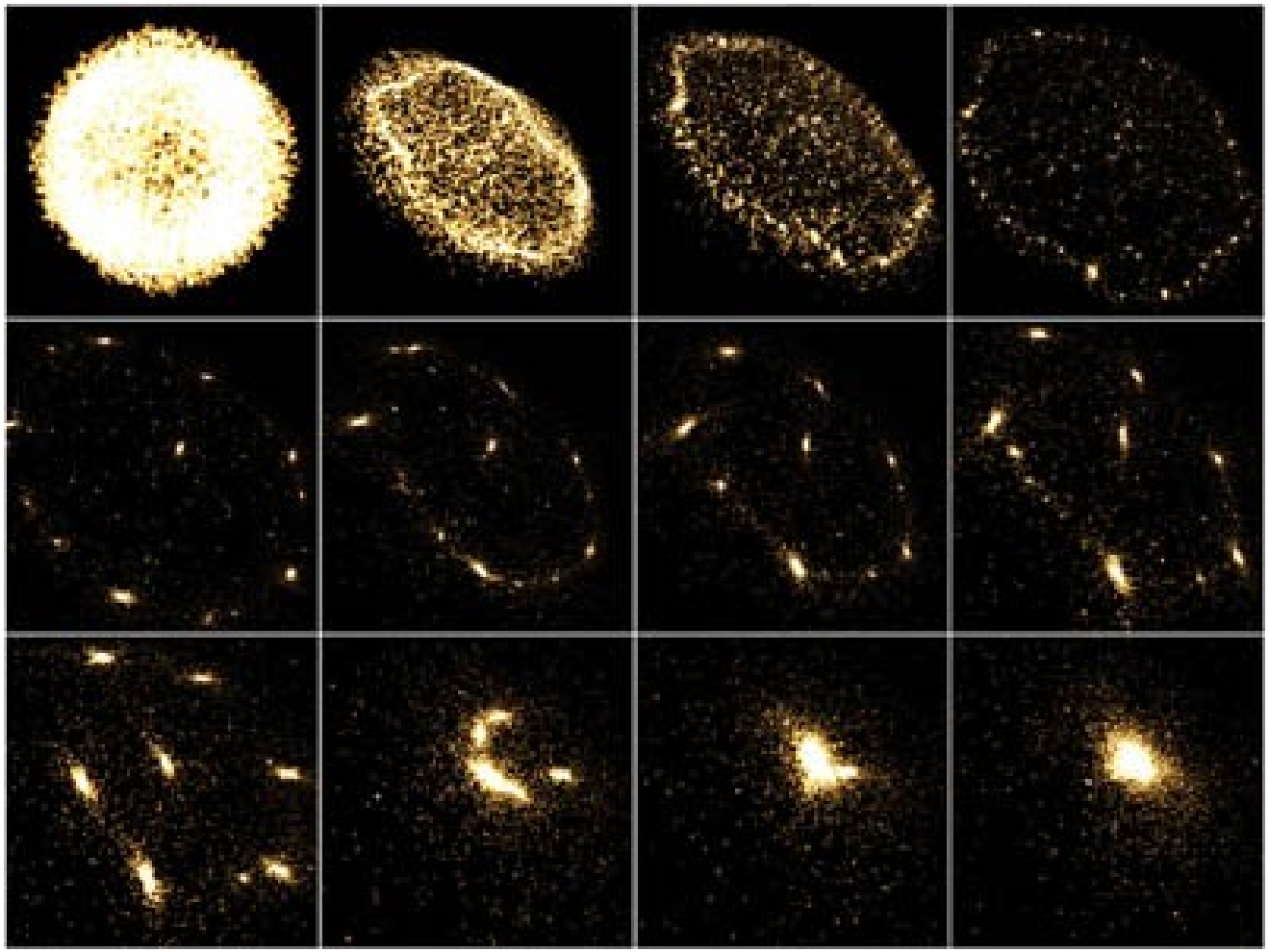
Sunway TaihuLight

Active	June 2016
Operators	National Supercomputing Center in Wuxi
Location	National Supercomputer Center, Wuxi, Jiangsu, China
Architecture	Sunway
Power	15 MW (Linpack)
Operating system	Sunway RaiseOS 2.0.5 (based on Linux)
Memory	1.31 PB (5591 TB/s total bandwidth)
Storage	20 PB
Speed	1.45 GHz (3.06 TFlops single CPU, 105 PFLOPS Linpack, 125 PFLOPS peak)
Cost	1.8 billion Yuan (US$273 million)
Purpose	Oil prospecting, life sciences, weather forecast, industrial design, pharmaceutical research
Web site	http://www.nsccwx.cn/wxcyw/

Sunway TaihuLight

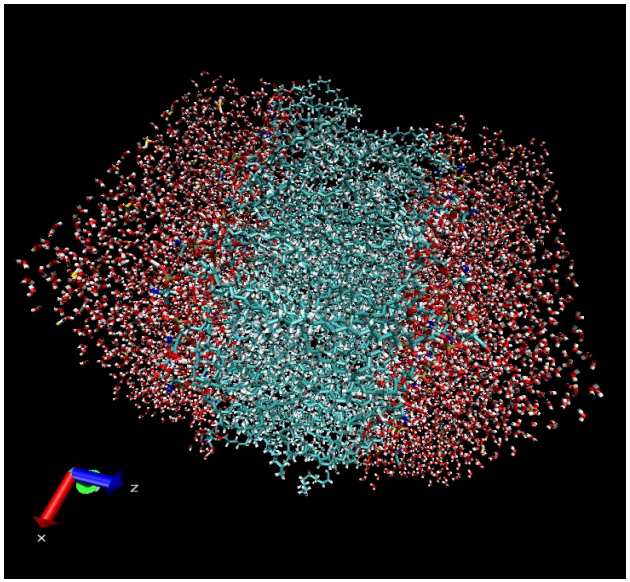
- Architecture
- The Sunway TaihuLight uses a total of 40,960 Chinese-designed SW26010 manycore 64-bit RISC processors based on the Sunway architecture.
- Each processor chip contains 256 processing cores, and an additional four auxiliary cores for system management (also RISC cores, just more fully featured) for a total of 10,649,600 CPU cores across the entire system.
- The processing cores feature 64 KB of scratchpad memory for data (and 16 KB for instructions) and communicate via a network on a chip, instead of having a traditional cache hierarchy.
- http://www.netlib.org/utk/people/JackDongarra/PAPERS/sunway-report-2016.pdf

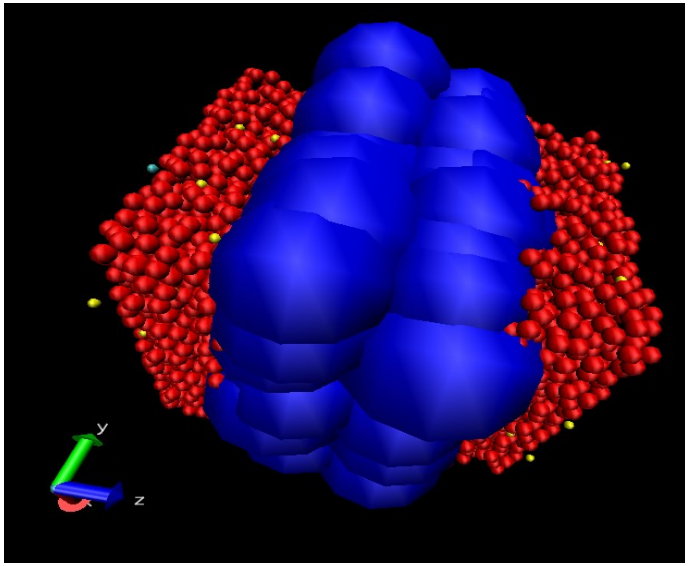
Applications



HPC - Applications

- Molecular Dynamics

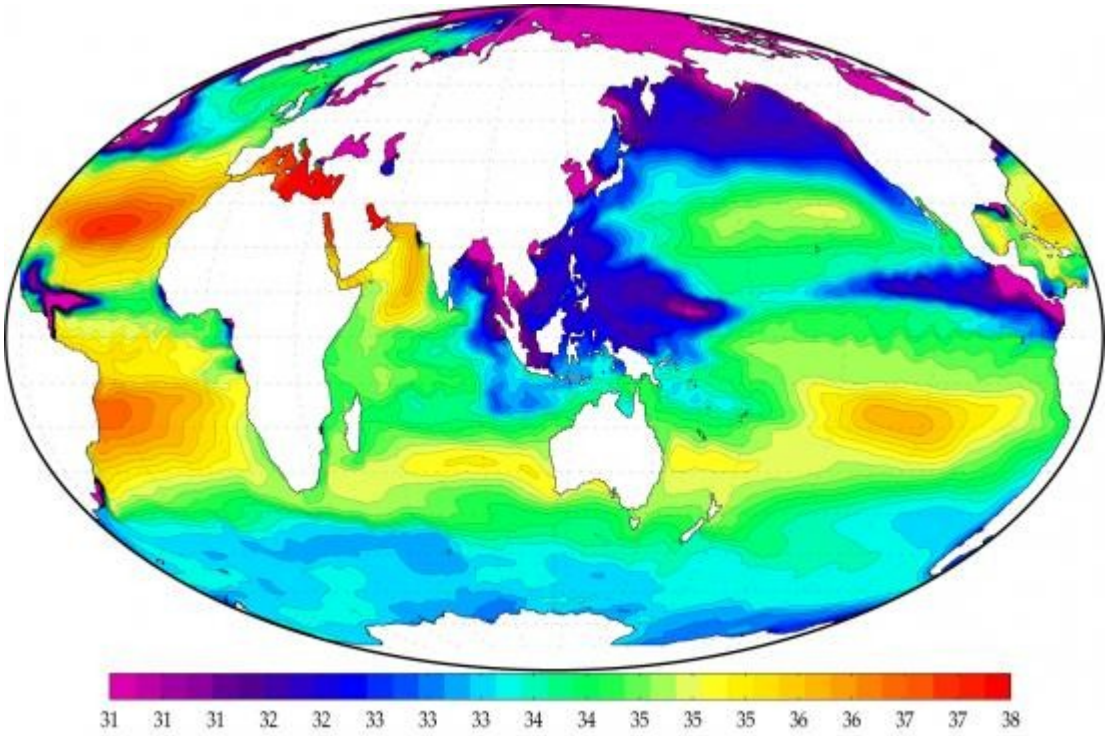


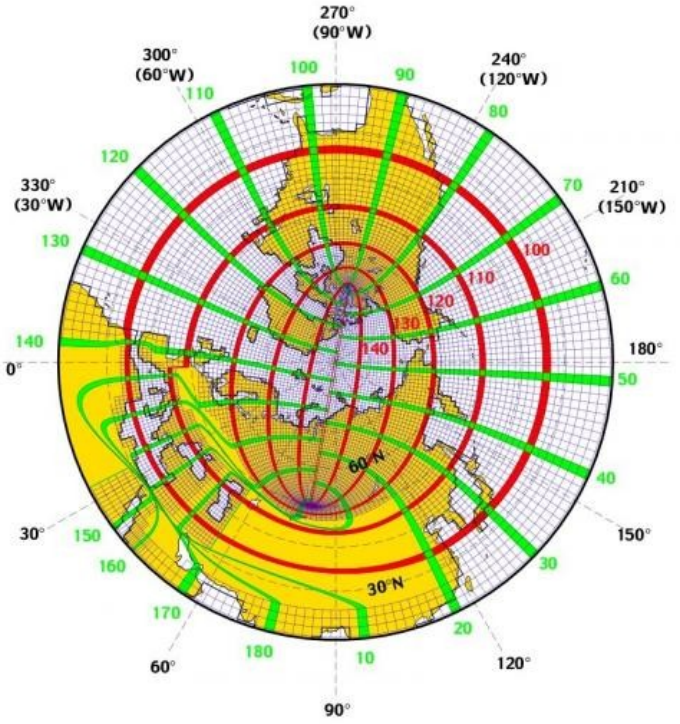


NAMD, Quantum Espresso, Gromacs, Gaussian, etc..

HPC - Applications

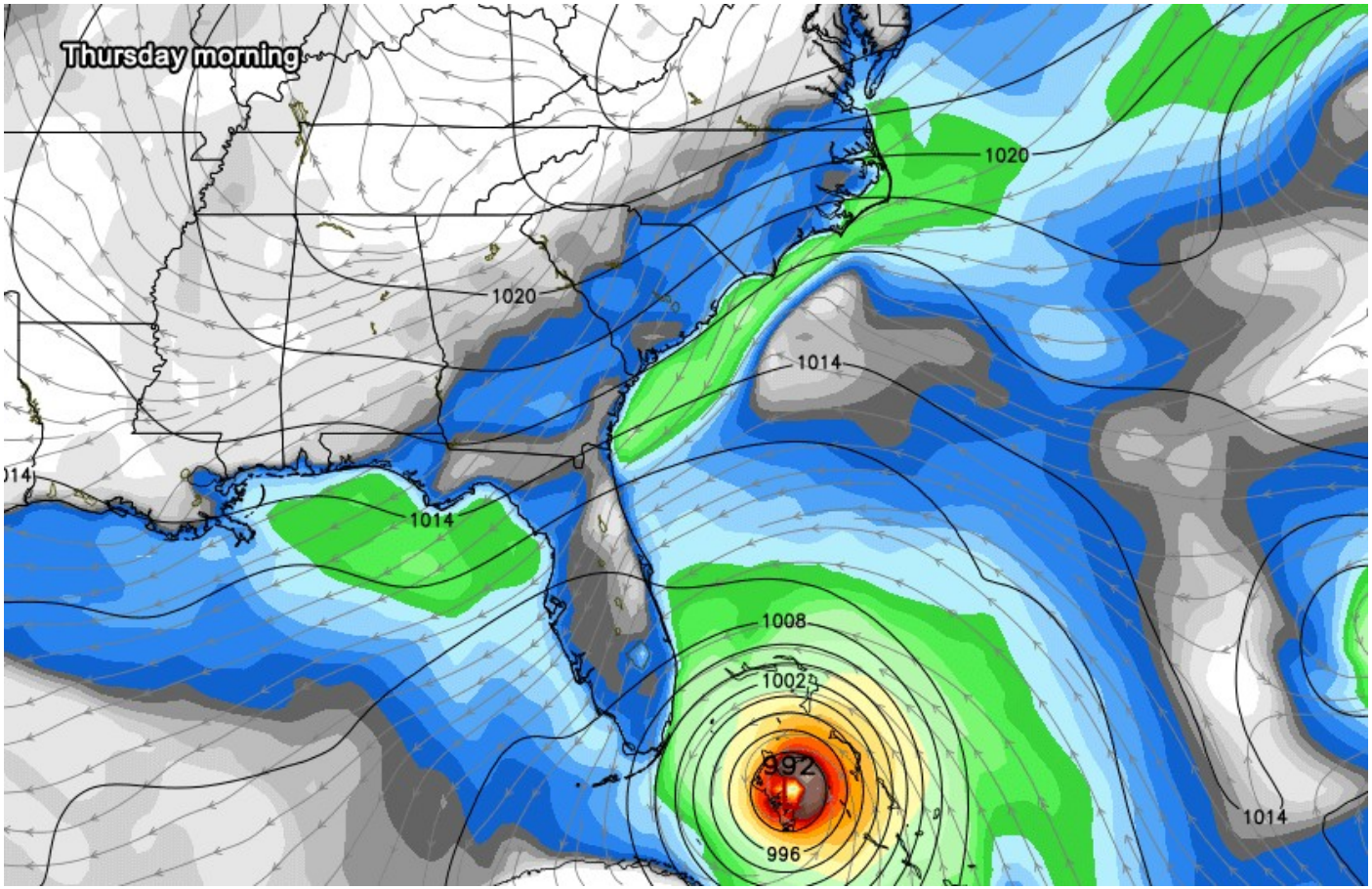
- Earth simulation



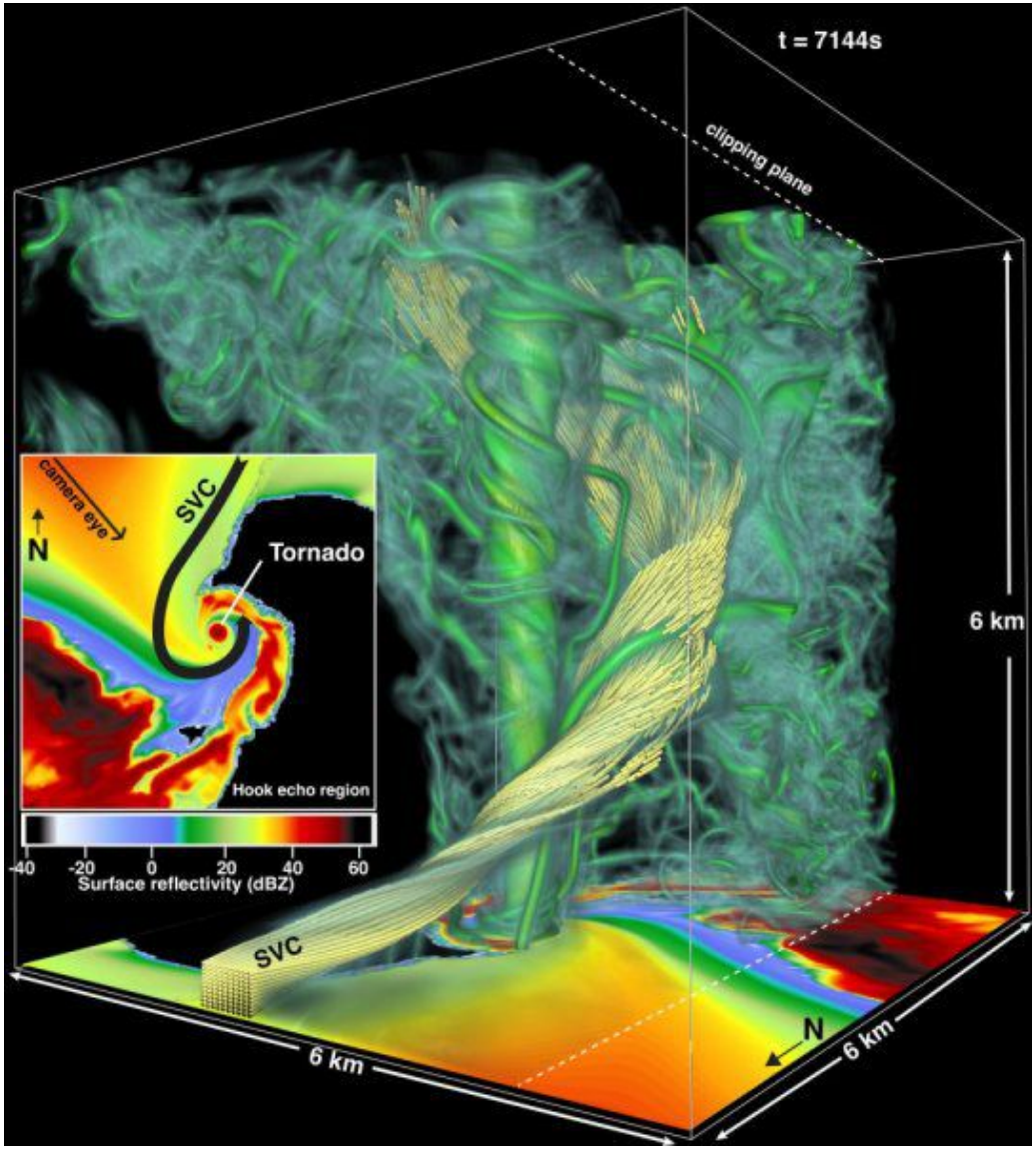


WRF, MM5, GLOBO, NEMO, etc..

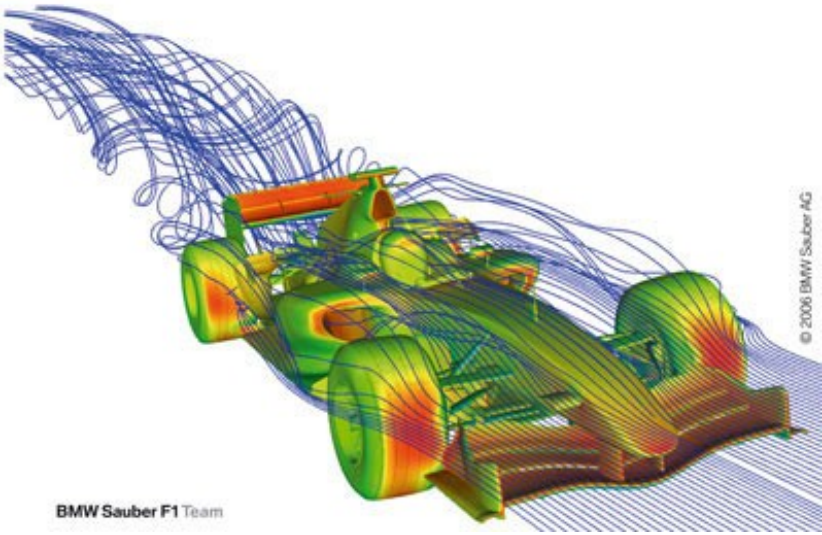
HPC- Applications



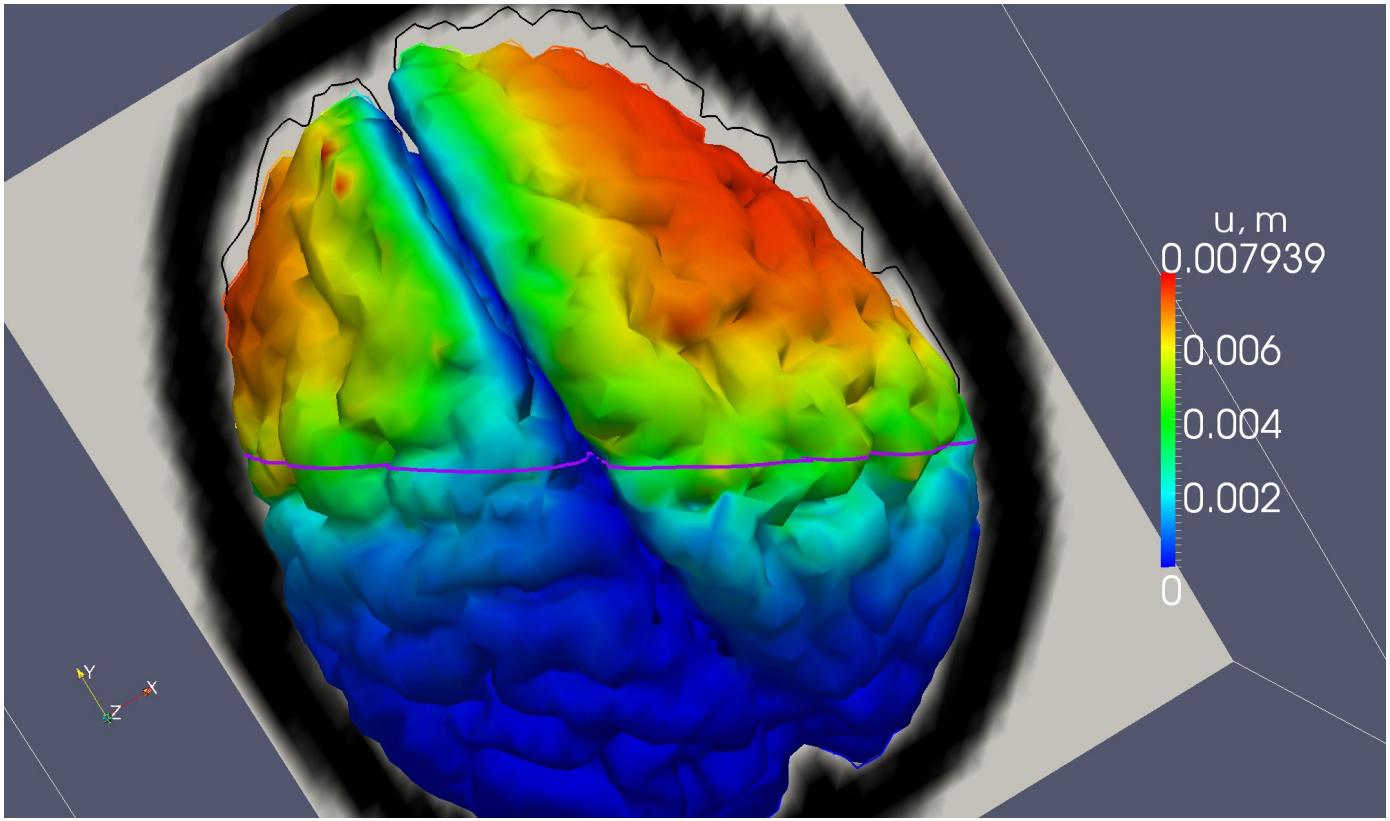
HPC - Applications



- Fluid Dynamics



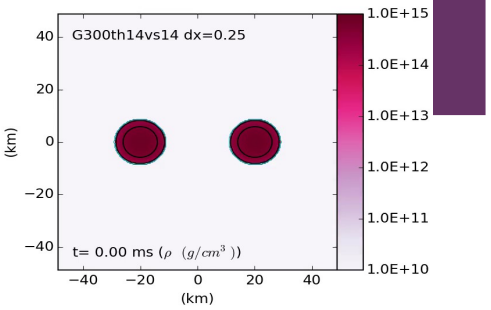
HPC - Applications

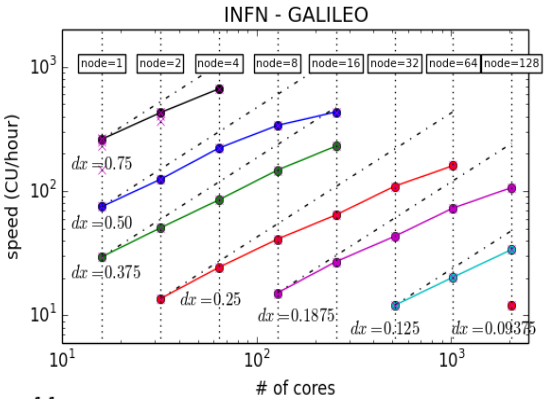


- Brain Simulation

HPC - Applications

- **General relativity**
- **The scientific case: high resolution simulation of inspiral and merger phase of binary neutron stars system**
  - **one of source of the gravitational waves that are the observational target of the LIGO/VIRGO experiment**
- **Computation performed using The Einstein ToolKit**
- **Result obtained on Galileo at CINECA**





HPC – key components

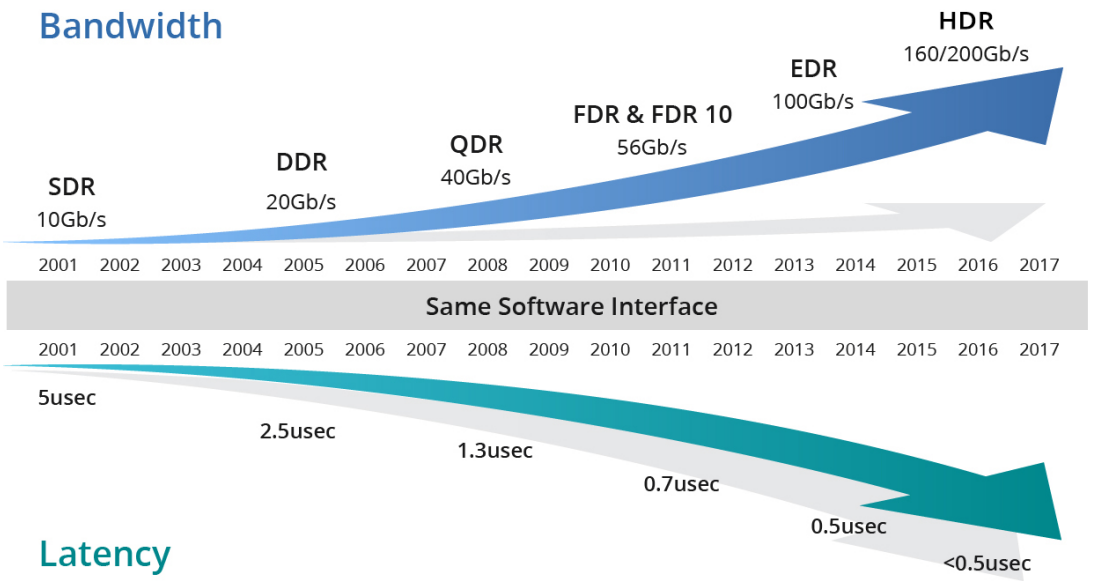
Componenti chiave

- Interconnessione dei nodi
- Condivisione dei dati
- Gestione dei job e contesa delle risorse

Interconnessione dei nodi

- In tutti i campi di applicazione dell'HPC il calcolo in un punto del dominio dipende dai risultati nei punti prossimi vicini
- Questo comporta la necessità di scambio continuo di dati fra i nodi coinvolti nel calcolo
- Quindi la rete deve garantire grande banda passante e bassa latenza, oggi giorno si hanno due possibilità InfiniBand o OmniPath

InfiniBand



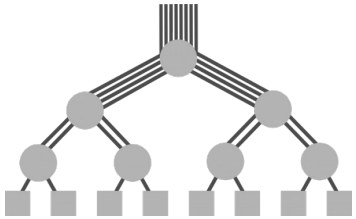
(*) Source: FS

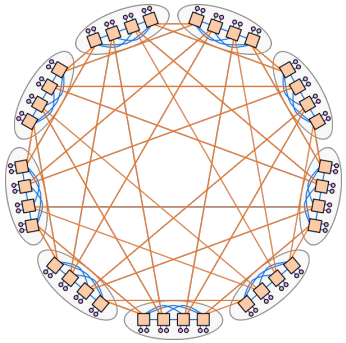
- Nasce nel 1999 dalla fusione di due progetti Future I/O e NextGeneration I/O
- Nello stesso anno nasce Mellanox
- Oggi arriviamo fino a 800Gb/s (XDR)
- La vera forza è l'offload: RDMA, GPUDirect RDMA, SHARP

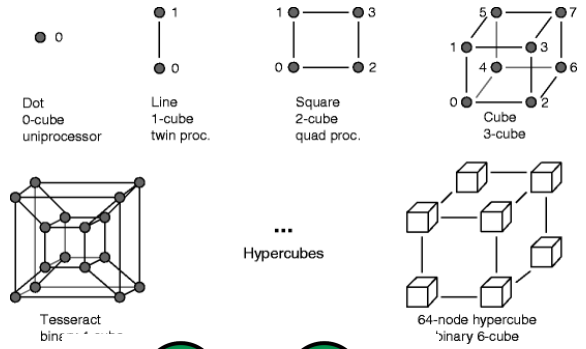
InfiniBand: topologie

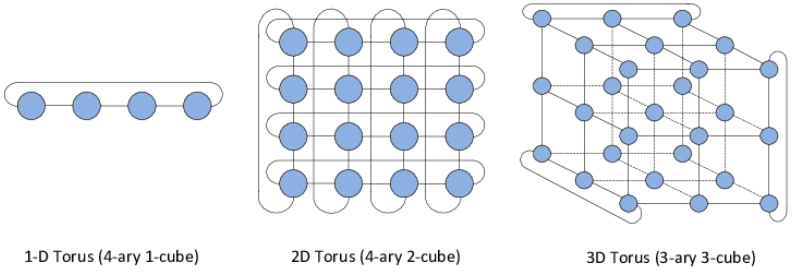
- Fat-Tree: connessione gerarchica multilivello, salendo dalle radici (nodi) al tronco si aumenta il numero di link
- Dragonfly: i nodi sono raggruppati in sottogruppi di piccole dimensioni completamente connessi, i sottogruppi sono collegati fra di loro
- Hypercube
- Torus
- Mesh

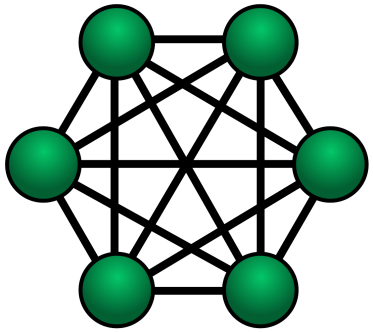
Topology	Scalability	Latency	Bandwidth	Fault Tolerance	Complexity
Fat-Tree	High	Moderate	High	High	High
Dragonfly	Very High	Low	Very High	Moderate	High
Hypercube	Moderate	Low	Moderate	High	Moderate
Torus	High	Low	High	Moderate	High
Mesh	Moderate	Moderate	Moderate	Low	Low











OmniPath?

- Fratello economico di InfiniBand
- Prestazioni simili, latenza leggermente più alta
- Scalabilità (n. di nodi) simile
- Manca la parte di off-load

Condivisione dei dati

- Un filesystem che sia accessibile da tutti i nodi del cluster, ma:
- Sia in grado di reggere all'accesso parallelo delle applicazioni, magari molti processi che accedo a pochi file
- Evitare il lock a file, necessario lock a blocchi
- Esere in grado di scalare le prestazioni

File System Parallelo

Parallel File Systems

- Dati distribuiti fra più server conenssi in rete
- I/O coordinato fra nodi client e server in modo da ottimizzare le prestazioni
- Stripe a livello di file, i singoli blocchi sono scritti su diversi storage gestiti da diversi server
- Non è necessario sapere dove si trovano i file, name space globale
- Avere la possibilità di separare metadati e dati
- Si aumentano le prestazioni aggiungendo sistemi di storage e/o server spesso in modo trasparente all'utenza
- Due possibilità
  - IBM SpectrumScale (GPFS)
  - Lustre

Gestione dei job

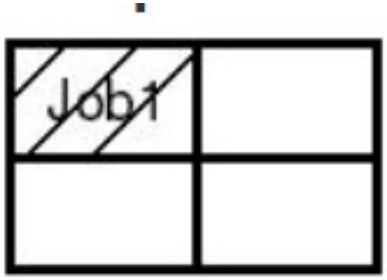
- Tipicamente si hanno molti utenti che condividono la risorsa HPC, quindi:
  - Necessario ottimizzare l'accesso per non sprecare tempo CPU e potenza elettrica
  - Utenti diversi possono avere diritti diversi sul tempo macchina
  - Utenti diversi possono avere priorità di accesso diverse

- Il sistema di batch permette una regolazione da "gentiluomini" di queste varie esigenze
- Ma deve sapere che è un cluster HPC

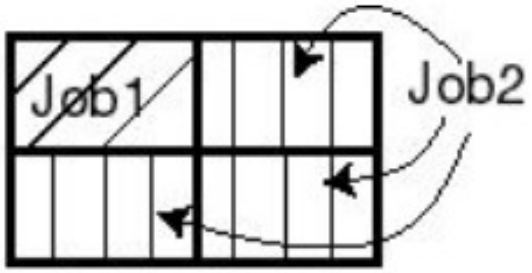
Batch System per HPC

- Deve “concepire” il fatto che un job giri su più nodi
- Quindi saper tener traccia dei nodi assegnati ad un job
- Fornire all’utente le informazioni relative ai nodi assegnati al job
- Possibilmente: meccanismi furbi di gestione delle risorse come reservation e backfill

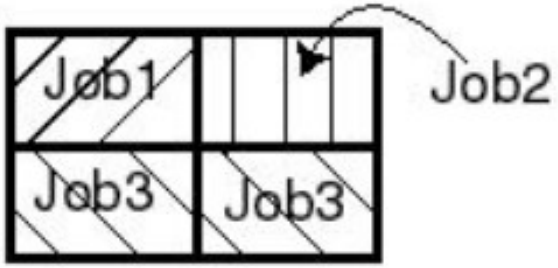
Reservation / Backfill



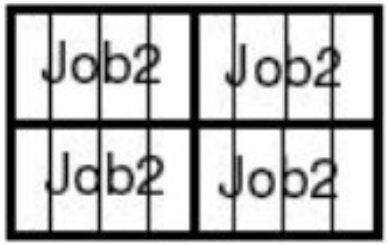
(a) Job1 started at 8:00 am.
Will finish at 10:00 am.



(b) Job2, submitted but can't start since it needs 4 processors.
Remaining 3 reserved by Job2.



(c) At 8:30 am Job3 submitted.
Job3 backfills Job2.



(d) At 10:00 am, Job2 starts.

Batch System per HPC

- Esistono vari Batch System, quelli che vanno per la maggiore nel settore HPC sono
- LSF, di IBM a pagamento
- SLURM, progetto opensource nato dalla collaborazione fra varie entità, nasce esplicitamente per il supporto HPC molti dei più grandi presenti nella Top500 lo usano