BIOINforMATics

STAZIONE ZOOLOGICA ANTON DOHRN • Napoli •

HPC systems @SZN A. Dohrn

M. Miralto
M. Sangiovanni

High Performance Computing

High-Performance Computing (HPC) is a term used to describe computing environments which utilize **supercomputers** and **computer clusters** to solve very complex problems .

In first instance, a **supercomputer** is just an extremely powerful computer, performing near the highest operational rate for the current available technology and showing a much higher level of performance compared to a general-purpose computer

Mainframes

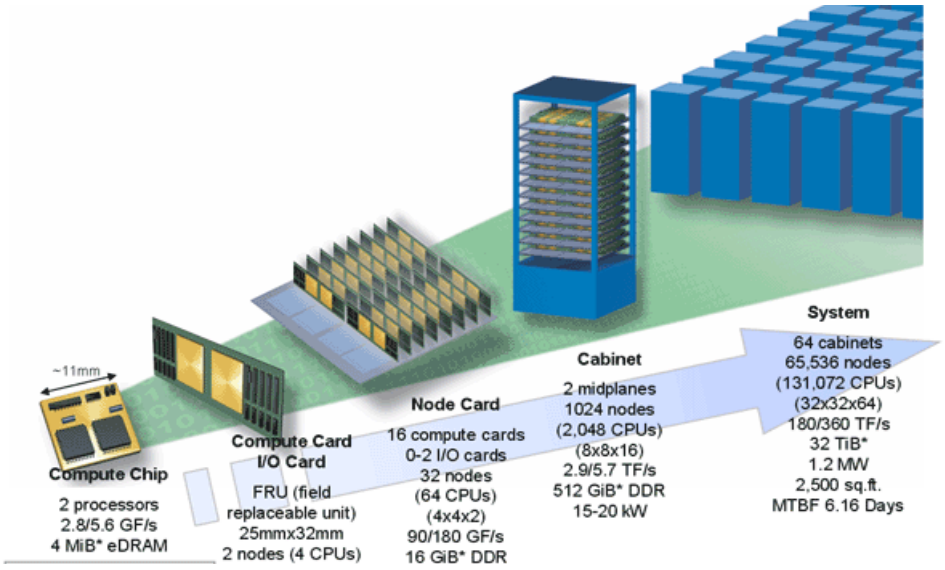
In the 50s-70s of last century, the industry approach to supercomputing was essentially tied to the development of mainframe architectures: a big number of CPUs and storage resources to address big volumes of computation and data

Critical issues:

- Expansive and very complex electronic architectures
- Scalability
- Size
- Fault

Idea: use affordable and easy to obtain hardware to build computer clusters. It is cheaper, much more easy to manage, and, in most of the cases, simple to scale.

What is a HPC cluster

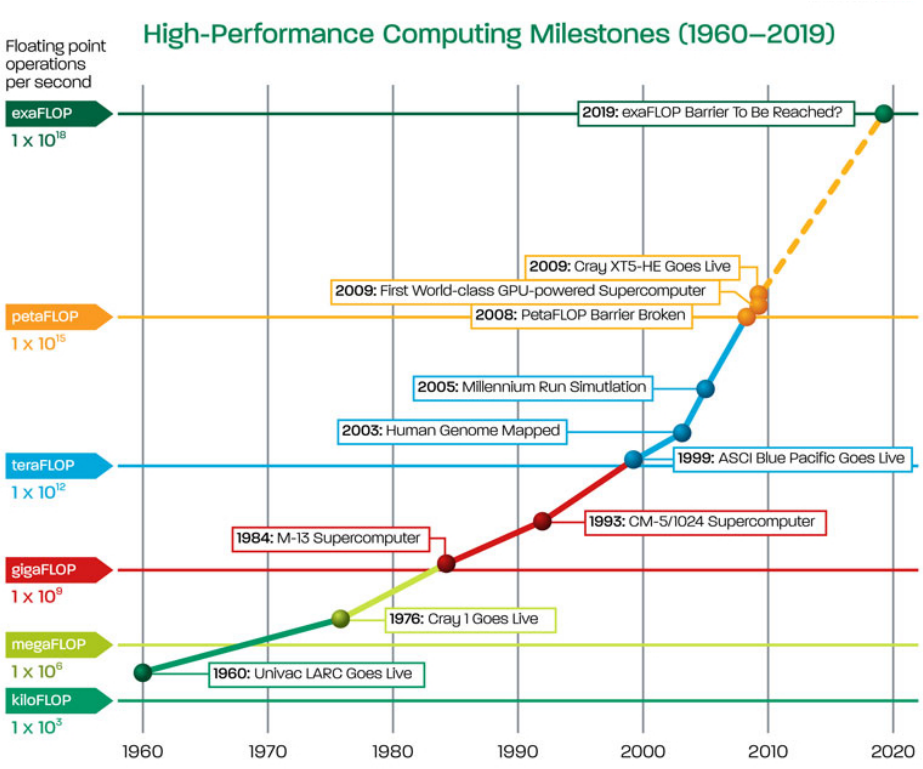


HPC clusters are systems made of a number of (headless) computers, usually referred as "**nodes**", interconnected with **high performance network switches**. Each node has its own CPUs, GPUs, RAM and possibly a small amount of storage.
Node's jobs are coordinated via a software layer.
HPC cluster nodes usally share a **common set of installed software and libraries** and user-dedicated storage.

Nodes can show the same technical specifications (amount of RAM, amount of storage) or be extremely different under all or some of these aspects.

Computational resources of each node are usually managed and allocated to jobs via a **workload manager** or a **Message Passing Interface**, software layers that allows to launch processes cluster-wide, **dispatching the user task to one or more nodes** if resource demands are met.

Supercomputers performances



Performances are measured in **floating-point operations per second** (FLOPS).

We now have machines that reach the **petaFLOP,** with the exaFLOP barrier within reach in two years.

HPC clusters in the world

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	2,282,544	122,300.0	187,659.3	8,806
2	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
3	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	1,572,480	71,610.0	119,193.6	
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61,444.5	100,678.7	18,482
5	National Institute of Advanced Industrial Science and Technology (AIST) Japan	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR Fujitsu	391,680	19,880.0	32,576.6	1,649

The top 5 ranking supercomputers in June 2018 according to Top500 (http://www.top500.org).

Interesting details:

- Cores and CPU architecture
- GPUs (emerging trend)
- Networking
- Power consumption

Green clusters in the world

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	359	**Shoubu system B** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. Advanced Center for Computing and Communication, RIKEN Japan	794,400	857.6	47	18.404
2	419	**Suiren2** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. High Energy Accelerator Research Organization /KEK Japan	762,624	798.0	47	16.835
3	385	**Sakura** - ZettaScaler-2.2, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. PEZY Computing K.K. Japan	794,400	824.7	50	16.657
4	227	**DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States	22,440	1,070.0	97	15.113
5	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	8,806	13.889

The top 5 **green** supercomputers in June 2018 according to Green500 (http://www.green500.org).

Higher ranking means a better Gflops/watts ratio.

The Summit supercomputer, ranking number 5 in this chart, is showing great results mostly due to extensive use of GPUs.

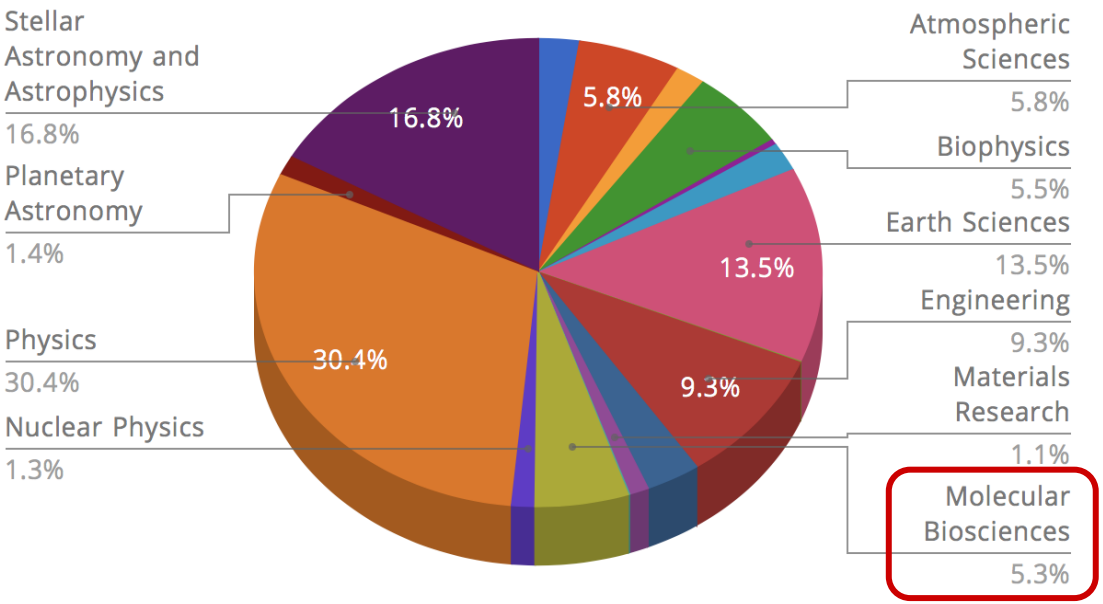
When do we need HPC?

Distribution of workloads by science field for an academic HPC system.

Physics analysis and simulations are the most common tasks, biosciences are still a niche.

Other common applications:

- Military
- Entertainment



Examples of the bioinfo world

High Performance Computing for biology is needed for **storing**, **analysing** and **visualizing** biological **Big Data**. Some examples:

- **Next Generation Sequencing** analysis:
  - Genome/Transcriptome assembly
  - DNA-seq/RNA-seq/Chip-seq
  - Single Cell Analysis
  - Differential Analysis
  - Variant calling
  - Metagenomics/Metatranscriptomics
  - Sequence alignment (BLASt, K-mer based,...)
- **Protein folding and protein dynamics**
- **Systems Biology Genome-Scale Modeling**

Why acquire a HPC cluster?

Key factors for HPC acquisition:

- **Performances**: whenever we are launching **heavy, parallelizable task**, HPC cluster can return a sensible gain on completion times (e.g. a process completion time could scale down linearly while adding more processing cores)
- **Big Volumes**: HPC cluster nodes usually come with a **large amount of RAM** (in the range of HUNDREDS of GB). Big storages (in the range of HUNDREDS of TB) are also attached.
- **Schedulable resources**: multiple users can **request resources at the same time**. Software-based allocation is guaranteed according to different criteria defined by system administrators (e.g. FIFO, Weighted Fair Sharing policies). We will cover this topic more in depth later

HPC facilities @SZN

Two different HPC clusters: **Kraken** and **Falkor**

- **Kraken** (192 cores, 6 TB RAM, ~140 TB Panasas storage)
  - GNU/Linux OS (Ubuntu 14.04)
  - 6 nodes (each node sports 4 CPUs, 8c per CPU, 32c total, 1 TB RAM)
- **Falkor** (128 physical cores, 224 threads, 4 TB RAM ~140 TB Panasas storage)
  - GNU/Linux OS (Debian 9.2)
  - 4 nodes (3 nodes support HT, each node sports 4 CPUs, 8c per CPU, 32c total, 1 TB RAM)

These two clusters share **the same user storage** and **the same authentication database**!

HPC facilities @SZN

Brief history:

- **2014**: 3 nodes + 1 frontend acquired. 1 Panasas High performance storage shelf
- **2015**: 7 more nodes acquired + 2 Panasas High performance storage shelves
- **2016**: Bad faults! MTBF: 15 days
- **September 2016**: HPC cluster operation were stopped while it was being relocated
- **February 2017**: the BIOINforMA group becomes operative. Part of the HPC cluster is relocated and reconfigured to be operational again. 3 nodes were returned for repair.
- **December 2017**: 3 repaired nodes are finally back and configured as the nucleus of a new cluster
- **July 2018**: MTBF is about 6 months!

Available software @SZN

Each HPC cluster provides a set of available software

- Softwares are listed on http://bioinfo.szn.it/tools/
- You can request new software opening a ticket on https://ticketing.bioinfo.szn.it
- Complex software stacks are organized in modules that can be easily loaded to build instantly the software environment you need

Falkor HPC cluster software list

Show 10 entries Search:

NAME	CATEGORY	HOMEPAGE	DESCRIPTION	VERSION	MODULEFILE
Abyss	assembler	http://www.bcgsc.ca/platform/bioinfo/software/abyss	ABySS is a de novo, parallel, paired-end sequence assembler that is designed for short reads. The single-processor version is useful for assembling genomes up to 100 Mbases in size. The parallel version is implemented using MPI and is capable of assembling larger genomes.	2.0.2	module load abyss/2.0-openmpi
AlignGraph	assembler	https://github.com/baoe/AlignGraph	Algorithm for secondary de novo genome assembly guided by closely related references		module load aligngraph/latest
bamtools	formats toolkit	https://github.com/pezmaster31/bamtools	BamTools provides both a programmer's API and an end-user's toolkit for handling BAM files.	2.5.1	module load bamtools/2.5.1
bbmap/bbtools	tool suite	https://jgi.doe.gov/data-and-tools/bbtools/	BBTools is a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. BBTools can handle common sequencing file formats such as fastq, fasta, sam, scarf, fasta+qual, compressed or raw, with autodetection of quality encoding and interleaving.	38.08	module load bbmap/38.08

Workload manager

Simple Linux Utility for Resource Management (**SLURM**)

- It is an **open source**, fault-tolerant, and highly scalable cluster management and job scheduling system for **large and small Linux clusters**.
- It is the workload manager on about 60% of the **TOP 500 supercomputers.**

As a cluster workload manager SLURM:

- it **allocates exclusive and/or non-exclusive access to resources** (compute nodes) to users for some duration of time so they can perform work
- it **provides a framework for starting, executing, and monitoring work** across cluster nodes (usually parallel jobs, but distributed MPI tasks can be managed too)
- it **arbitrates contention for resources** by managing a queue of pending work.

Other computational facilities

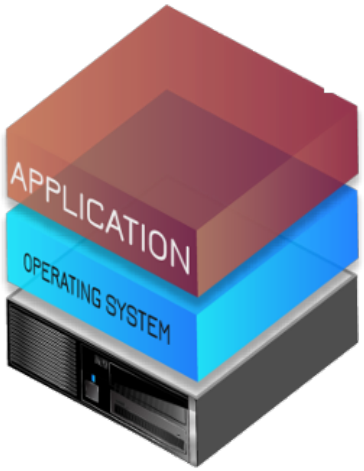
Several other servers and workstations @SZN:

- Most of them are nearing End of Life
- Mostly dedicated to light bioinfo tasks
- In the range of 8-40 cores
- Small amounts of RAM (32-128 GB)

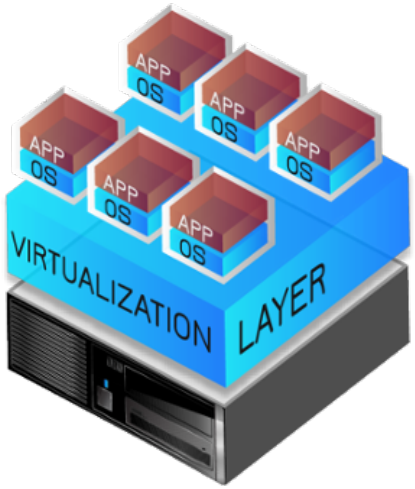
Recently acquired system with 36c/72t and 512GB of RAM and a good amount of local storage:

- Dedicated to web and db services
- Kvm machines
- Docker virtualized environments

Virtualized services



Traditional Server Architecture



Virtualized Server Architecture

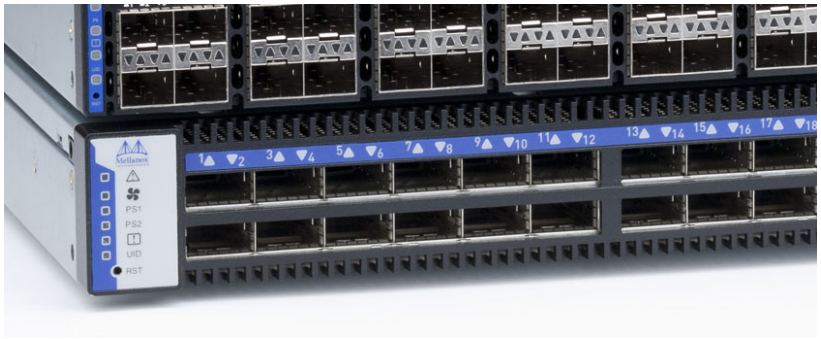
Deploying virtualized services using different technologies

- Docker
- KVM
- ESXI
- Xen

Future plans

SZN infrastructure provides power to face a good number of complex problems.

BUT THERE'S ALWAYS ROOM FOR IMPROVEMENT!



HPC cluster expansion

- Scale to a bigger number of cores: ~800c/1600t
- High density systems: twin chassis or blades
- More efficient systems: keep an eye on power consumption
- GPUs acquisition: a good number of algorithms, even in bioinformatics, can make use of GPUs
- Test Infiniband solutions, move to 100/200/400 Gbit ports

Future plans

SZN infrastructure provides power to face a good number of complex problems.

BUT THERE'S ALWAYS ROOM FOR IMPROVEMENT!

Cloud infrastructure

- Deploy an in-house cloud (compute and storage) infrastructure using state-of-the-art free solutions: Openstack + LXD + Docker)
- Cloud hosted services: migrate services on a resilient infrastructure

Future plans

SZN infrastructure provides power to face a good number of complex problems.

BUT THERE'S ALWAYS ROOM FOR IMPROVEMENT!

Storage infrastructure

- Best-effort storage: dedicated to backup and elastic drives
- High performance storage: parallel distributed filesystems attached to HPC clusters (Panasas systems, BeeGFS)
- Cloud-based object and block storage: flexible and resilient