Computational infrastructure for NGS data analysis

Pablo Escobar

pescobar@cipf.es

Computational infrastructure for NGS

- In NGS we have to process <u>really big amounts of data</u>, which is not trivial in computing terms.

- Big NGS projects require <u>supercomputing</u> infrastructures

Data tsunami is real

- Some disks in our lab.....

Sequencing cost vs IT cost

Sequencing cost goes down....so IT cost goes up

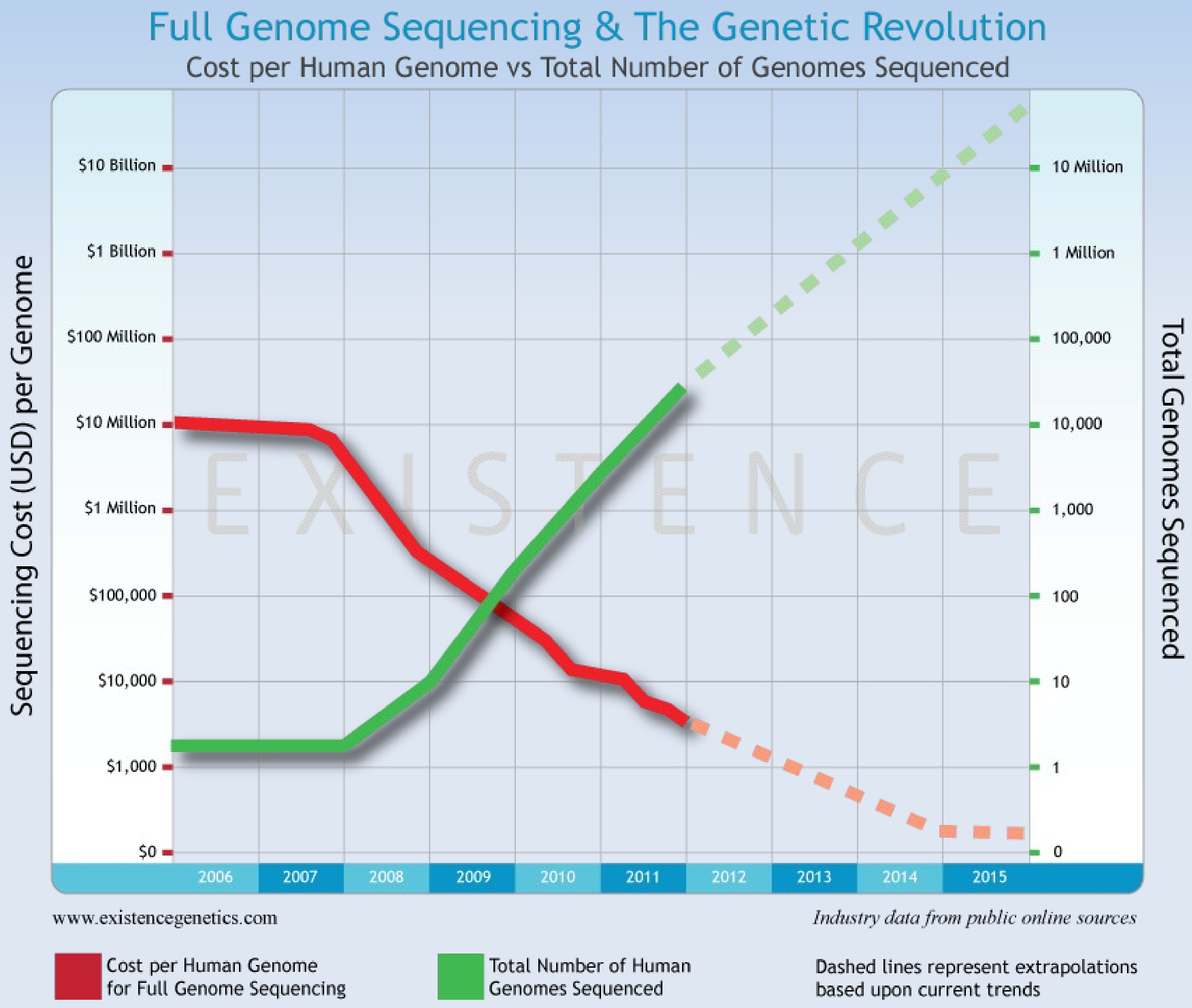


Image source: http://www.existencegenetics.com/fullgenome.php

Computational infrastructure for NGS

These infrastructures are expensive and not trivial to use, we require:

- Conditioned data center (servers room). This is expensive
- Computing cluster:
  - Many computing nodes (servers)
  - High performance and high capacity storage
  - Fast networks (10Gb ethernet, infiniband...)
- Skilled people in computing (sysadmins and developers).
  - In CNAG currently 30 staff - >50% informatics

Computing cluster

- Distributed memory cluster
  - 8 or 12 cores by node
  - x86_64 arch
  - At least 48GB ram per node
- Fast networks
  - 10Gbit
  - Infiniband
- Batch queue system (sge, condor, pbs, slurm)
- Many GPUs tools are being developed, no a bad idea to have some if you plan to use gpu tools

Storage system

- Storage is the most important piece in the IT infrastructure for NGS
- Storage is the most expensive
- Good design is really important. Talk with experts
- Keep in mind the storage scalability.
- Try to keep storage flexible. Changes come fast

Storage System

- Traditional backups are a problem, if even possible.
- Raid is your friend.
- Plan a good data storage policy
- Recommended reading:

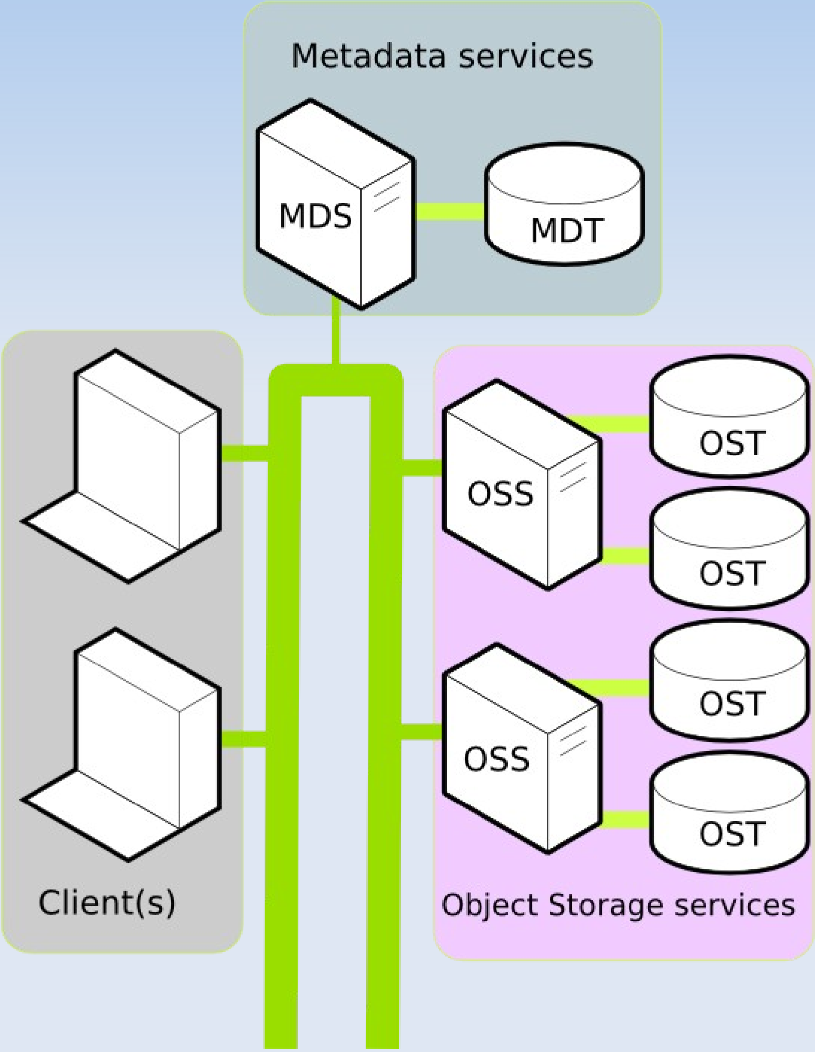
http://www.bioteam.net/wp-content/uploads/2010/03/cdag-xgen-storageForNGS_v3.pdf

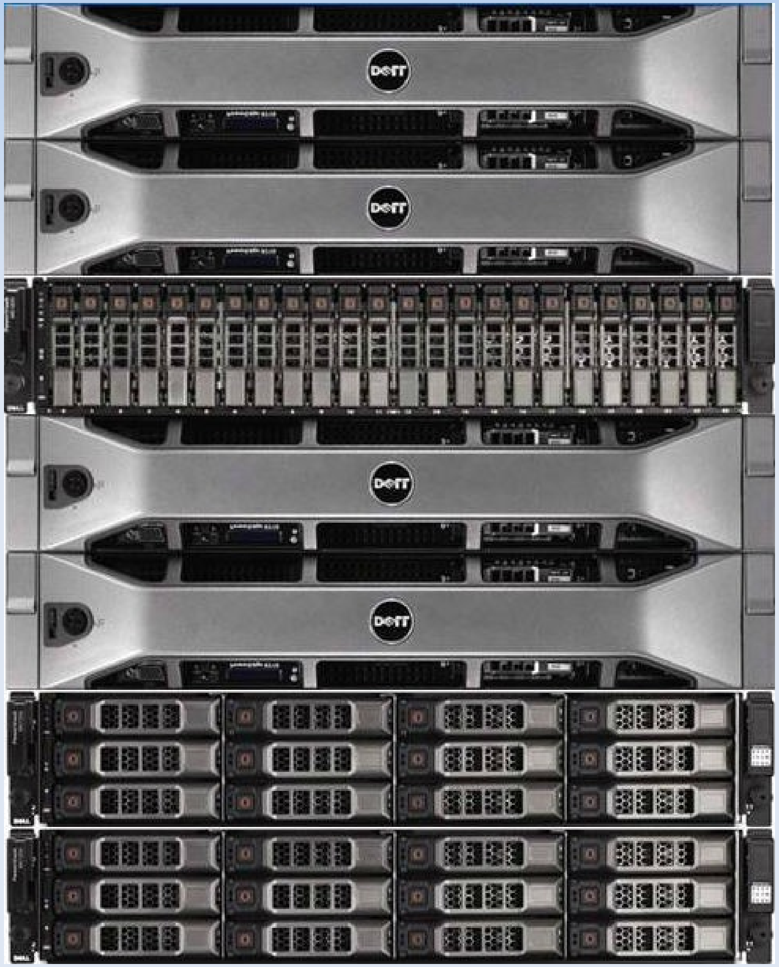
Storage system

- Distributed filesystem for high performance storage
  - Lustre
  - GPFS
  - Ibrix
  - GlusterFS
  - Panasas
  - Isilon

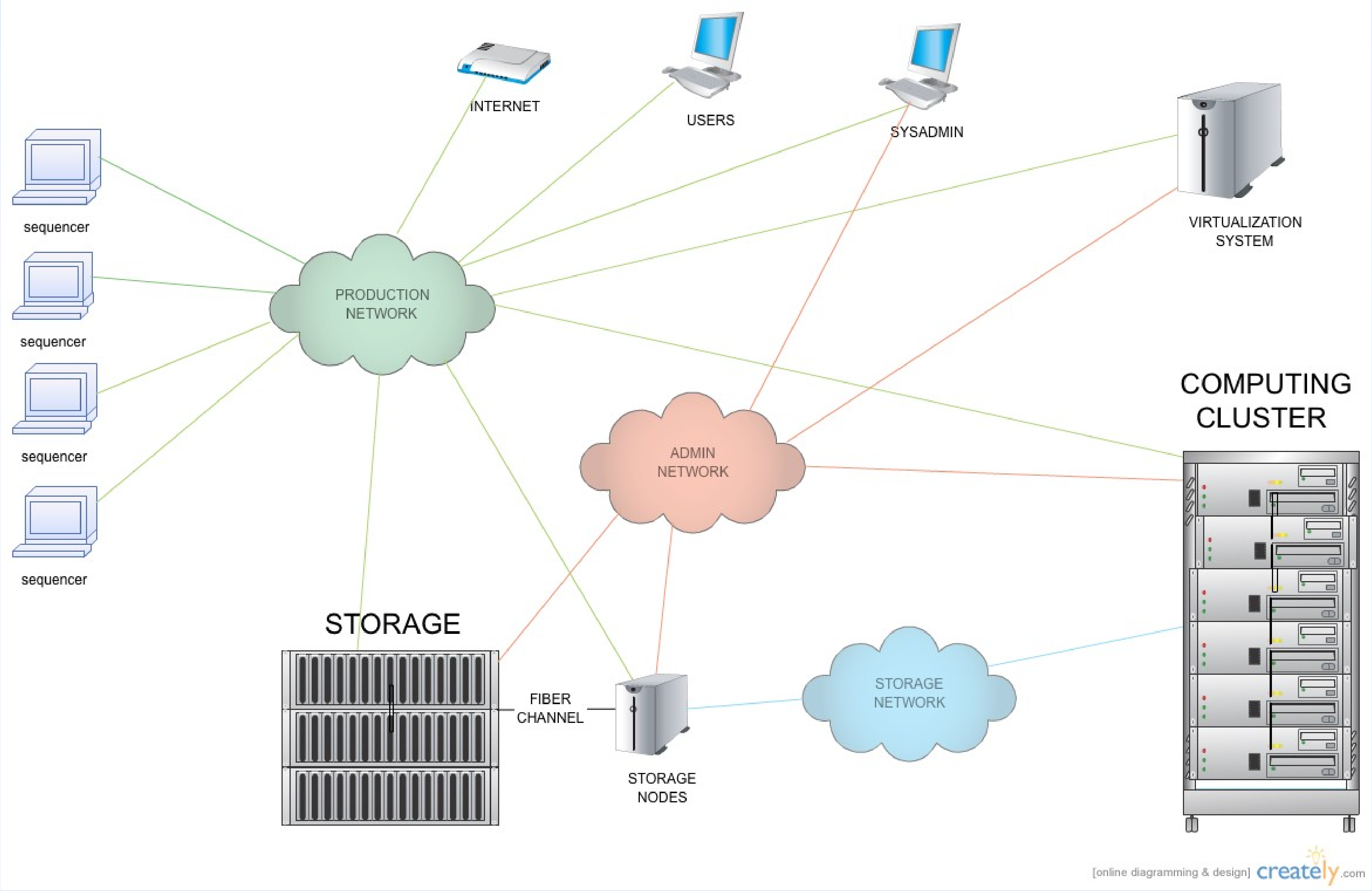
- These filesystems are not trivial to administer
- NFS is not a good option for supercomputing

Distributed filesystem schema





Infrastructure schema



Small infrastructure

- Recommended at least 2 machines
  - 8 or 12 cores each machine.
  - 48Gb ram minimum each machine.
  - BIG local disk. At least 4TB each machine
    - As much local disks as we can afford
- Price range: starting at 8.000€ - 10.000€ (two machines)

Sequencing centers in Spain

Medical Genome Project

- Sequencing Instruments
  - 7 GS-FLX (Roche)
  - 4 SolidTM 5500 (Applied Biosystems)
- Informatics infrastructure
  - 300 core cluster
  - 0,5 petabyte ibrix filesystem

Medical genome project

Storage racks

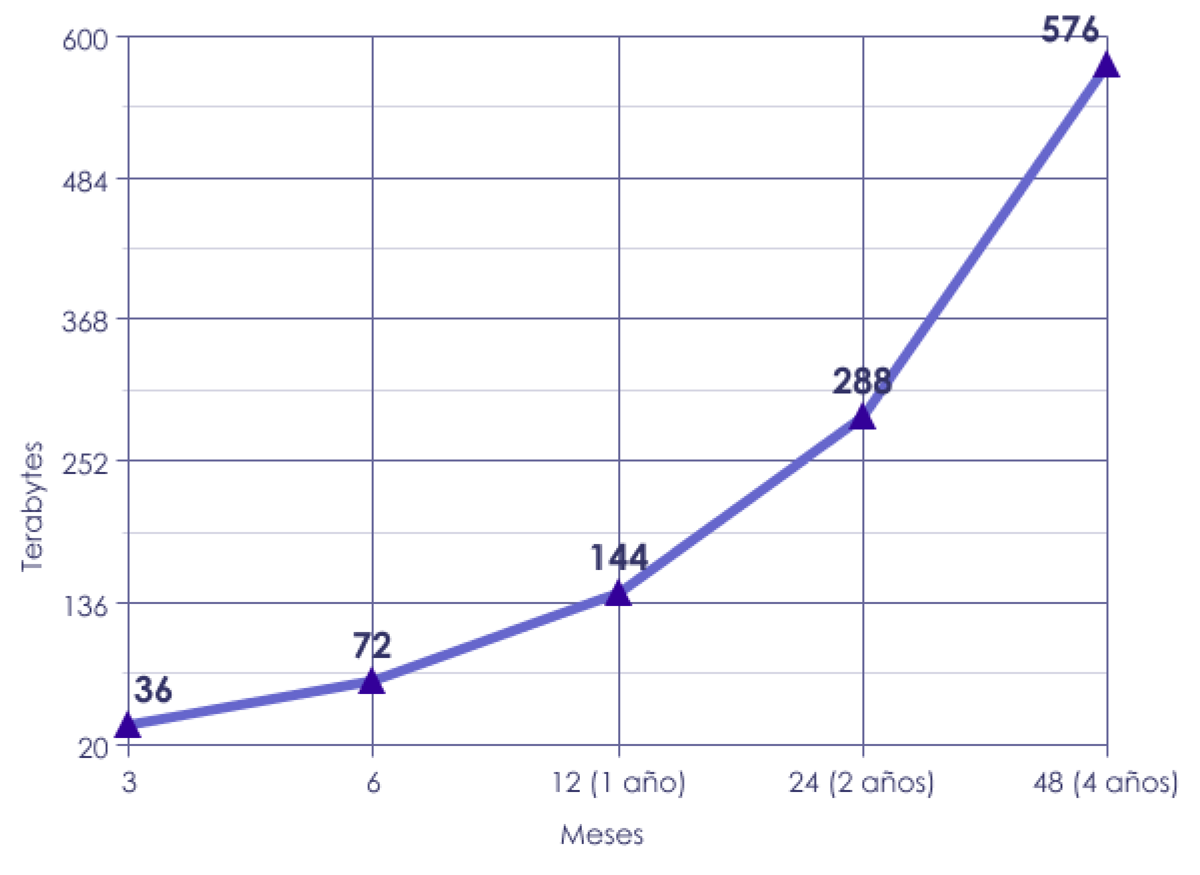
IBRIX filesystem

front-ends

MGP raw data generation

- a solid sequencer run
  - 7 days running
  - Generates around 4TB
- Only the four solid sequencers working full time can generate around 12TB each week.
- 12TB just of raw data. After running bioinformatics analysis more data is generated
- Raw data size grows really fast
  - New sequencer models
  - New reagents

MGP raw data generation



RAW data generation at MGP

Sequencing centers in Spain

CNAG

- Sequencing Instruments
  - 10 Illumina HiSeq2000
- Informatics infrastructure
  - 850 core cluster
  - 1.2 petabyte lustre filesystem (growing to 2PB)
  - 10 x 10 Gb/s link with marenostrum (Barcelona Super Computer 10,240 cores)

CNAG

Informatics

10 x 10 Gb/s

850 core cluster supercomputer

1.2 petabyte hardiscs

Barcelona Super Computer 10,240 cores

BGI - Largest sequencing center in the world

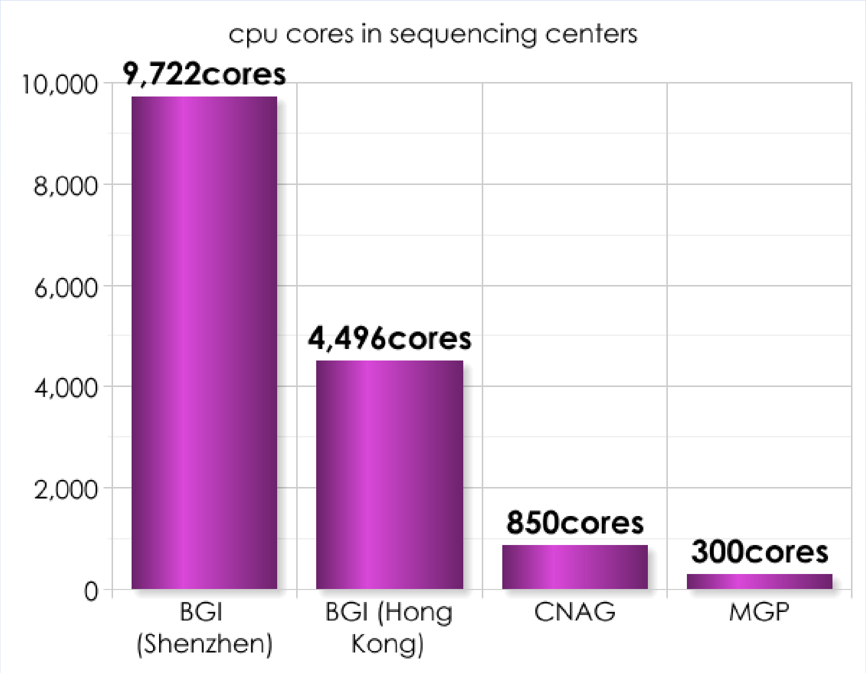
- Sequencing Instruments
  - Illumina HiSeq
  - AB SOLiD System
  - Ion Torrent
- Informatics infrastructure (8 datacenters)
  - 20,576 cores cluster
  - 17PB

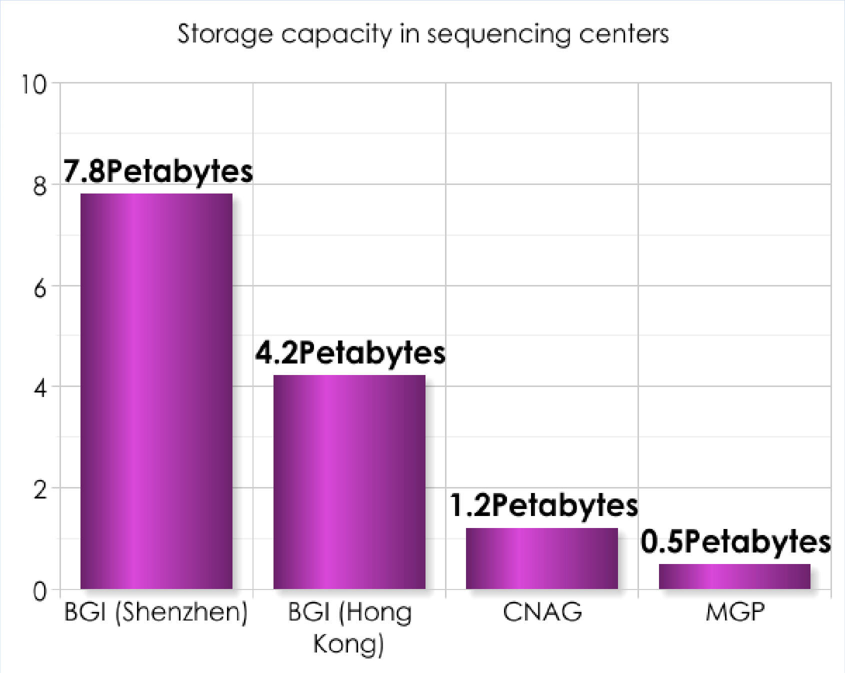
Source: http://www.genomics.cn/en/navigation/show_navigation?nid=4109

Largest sequencing center in the world

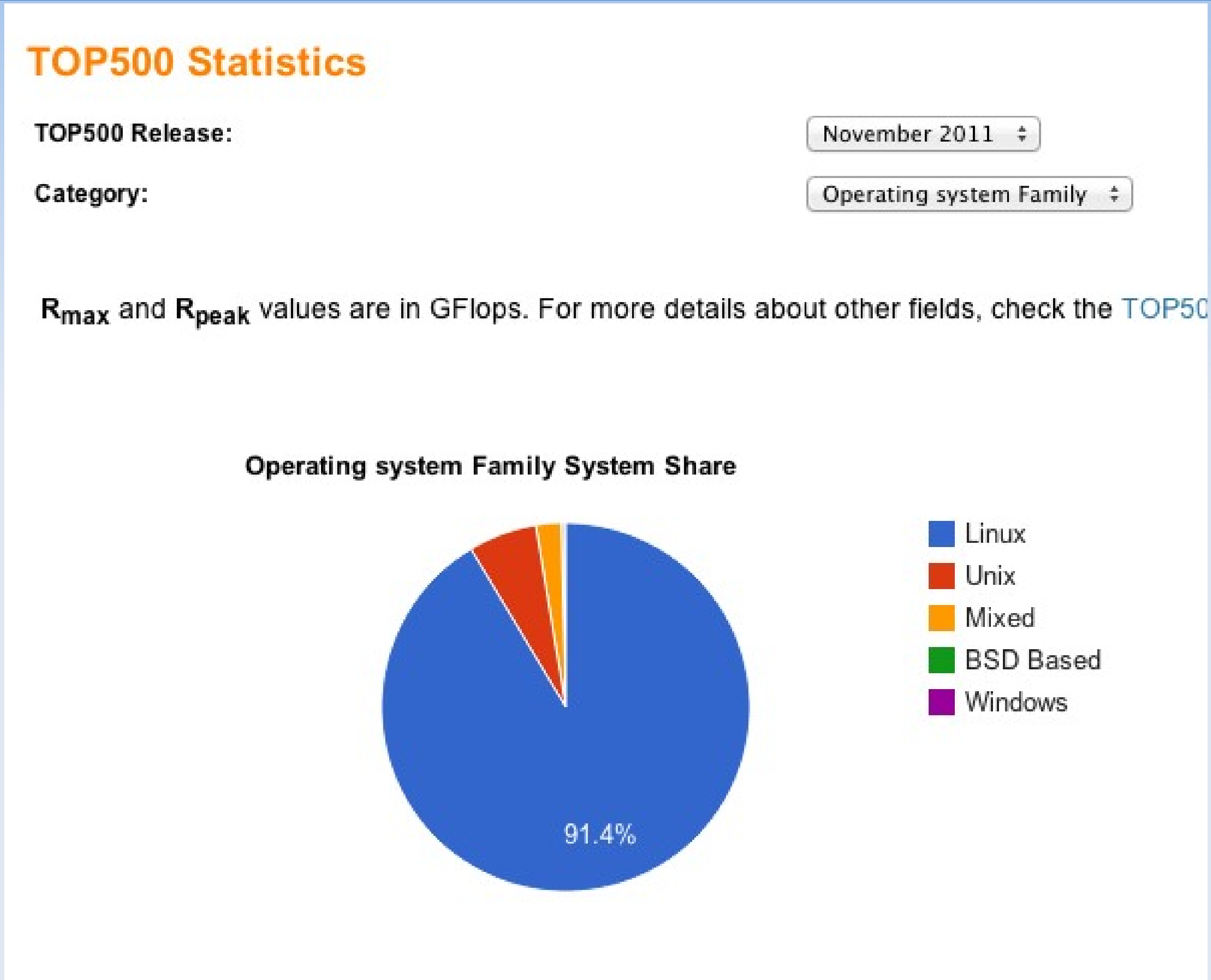
- Beijing Genomics Institute (BGI)

Sequencing center resources





Most used operating system is GNU/LINUX



TOP500 Statistics

TOP500 Release: November 2011

Category: Operating system Family

R_{max} and R_{peak} values are in GFlops. For more details about other fields, check the TOP50

Operating system Family System Share

Source:
http://www.top500.org/stats/list/36/osfam

Alternatives – cloud computing

- Pros
  - Flexibility.
  - You pay what you use.
  - Don´t need to maintain a data center.
- Cons
  - Transfer big datasets over internet is slow.
  - You pay for consumed bandwidth. That is a problem with big datasets.
  - Lower performance, specially in disk read/write.
  - Privacy/security concerns.
  - More expensive for big and long term projects.

That's all Folks!