www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Supercomputación, Big Data y Computación Cognitiva

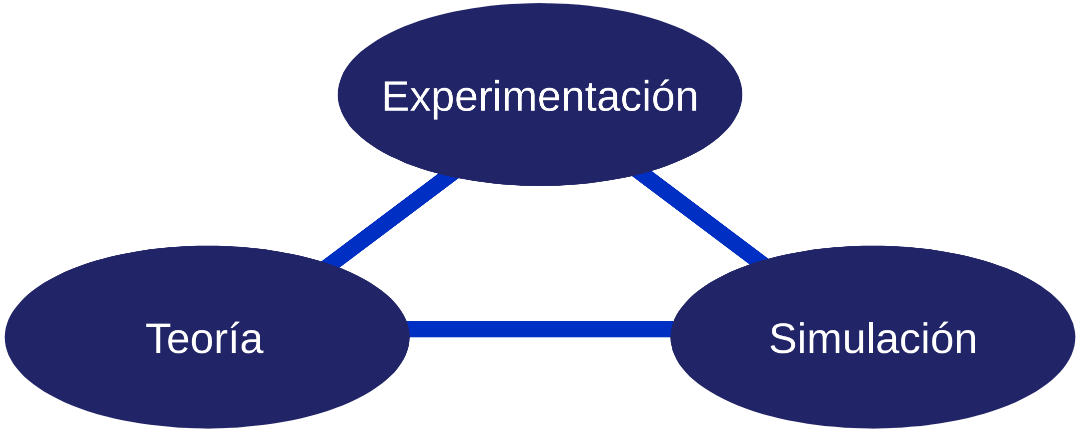
Prof. Mateo Valero

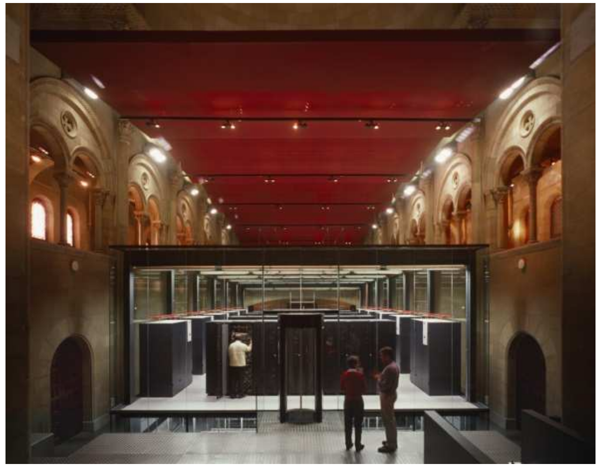
BSC Director

erc

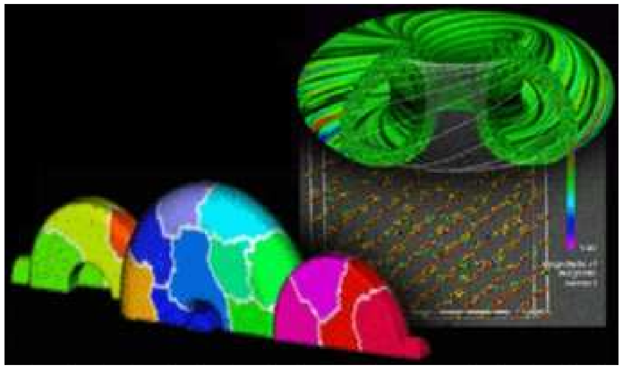
Granada, Abril, 2015

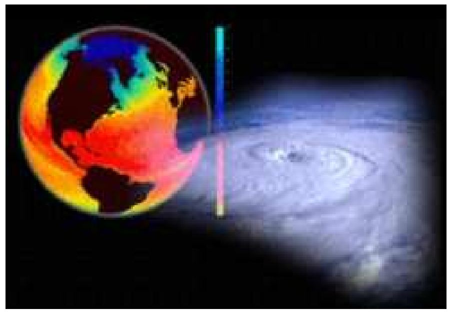
¿Cómo avanza la ciencia hoy?





Simulación = Calcular las formulas de la teoría





CARO

PELIGROSO

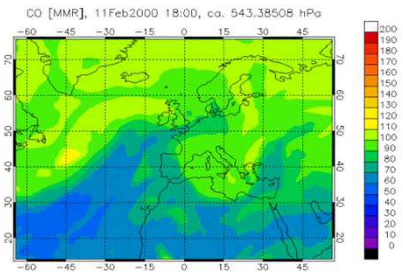
IMPOSIBLE

Simulation on Supercomputers helps to solve scientific, industrial and societal challenges

EXCELENCIA SEVERO OCHOA

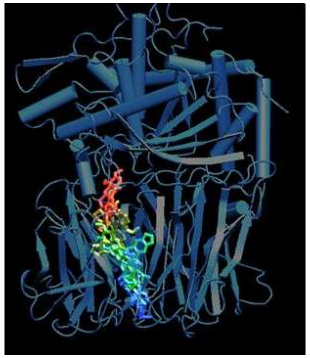
Environment

Climate prediction

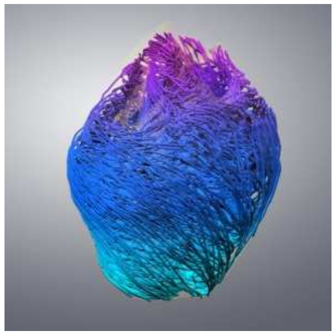


Air quality

Life Sciences

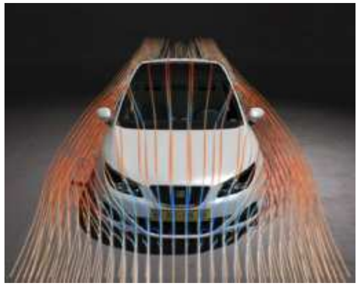


Personalised medicine



The virtual patient

Engineering



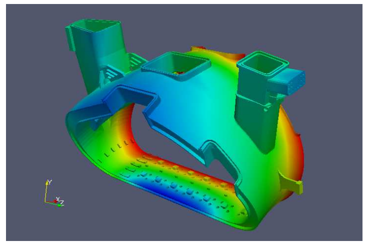
Industrial process improvement

Virtual prototyping

Energy

Oil exploration

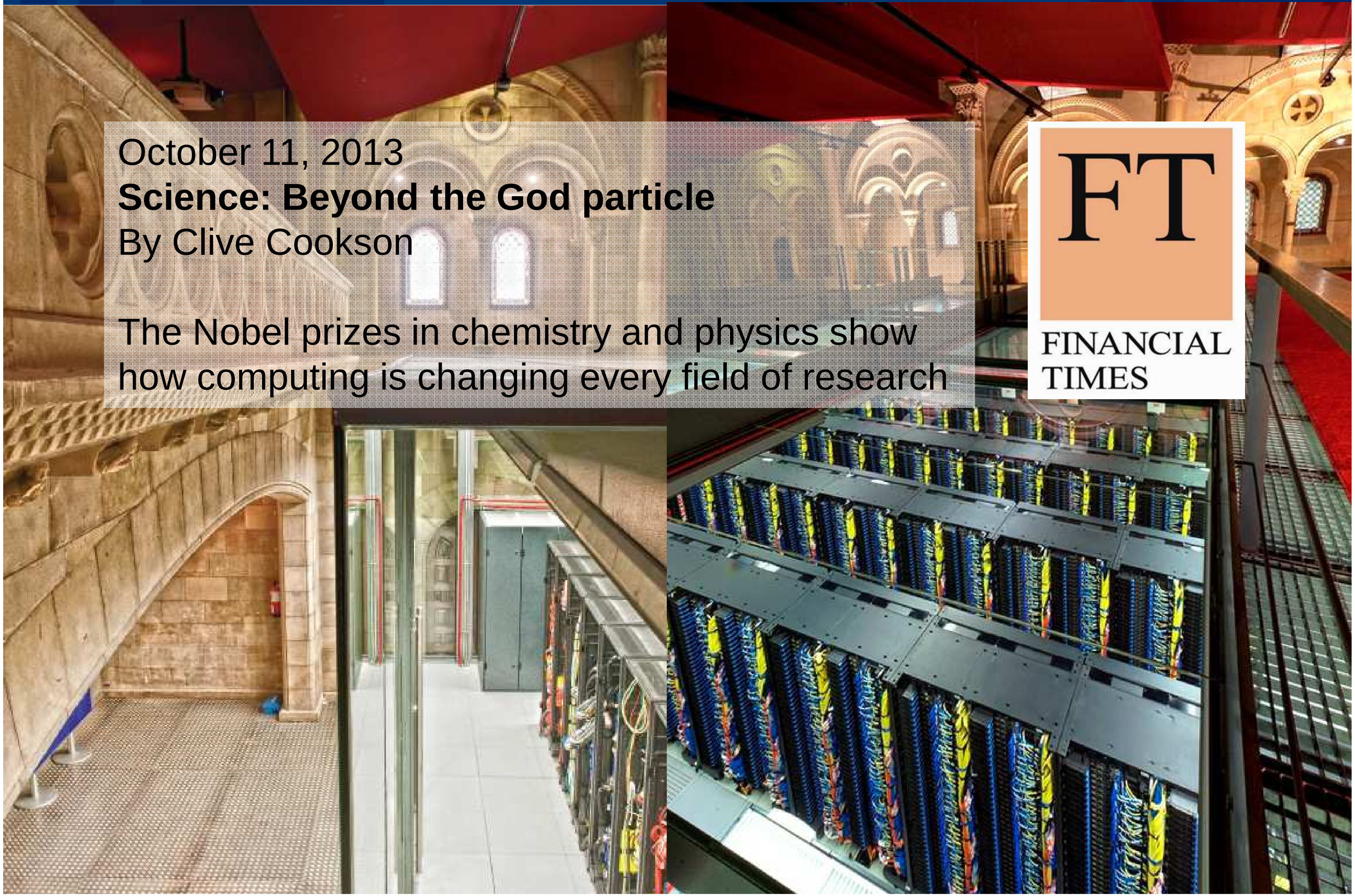
Wind farm design



Fusion (ITER)

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Computers are now an essential part of almost all research



October 11, 2013

Science: Beyond the God particle

By Clive Cookson

The Nobel prizes in chemistry and physics show how computing is changing every field of research

FINANCIAL TIMES

Top10

Novembers 2014 list (Only change number 10)

Rank	Site	Computer	Procs	Rmax	Rpeak	Power	GFlops/W att	Name
1	National Super Computer Center in Guangzhou	TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P	3120000 2736000	33,86	54,90	17,8	1,90	Tianhe-2 (MilkyWay-2)
2	DOE/SC/OAK Ridge National Lab	CRAY XK7, Opteron 6274 16C, 2.20 GHz, Cray Gemini interconnect, NVIDIA K20x	560640 261632	17,59	27,11	8,21	2,14	Titan
3	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	1572864	17,17	20,13	7,89	2,18	Sequoia
4	RIKEN Advanced Institute for Computational Science (AICS)	Fujitsu, K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	705024	10,51	11,28	12,65	0,83	K
5	DOE/SC/Argonne National Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	786432	8,58	10,06	3,94	2,18	Mira
6	CSCS	Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x	115984 73808	6,27	7,79	2,32	2,70	Piz Daint
7	Texas Advanced Computing Center	PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi	462462 366366	5,17	8,52	4,51	1,14	Stampede
8	Forschungszentrum Juelich (FZJ)	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	458752	5,00	5,87	2,30	2,18	JUQUEEN
9	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	393216	4,29	5,03	1,97	2,18	Vulcan
10	Government	Cray XC30, Intel Xeon E5-2660v2 10C 2.2GHz, Aries, NVIDIA K40	72800 62400	3,57	6,13	1,50	2,39	

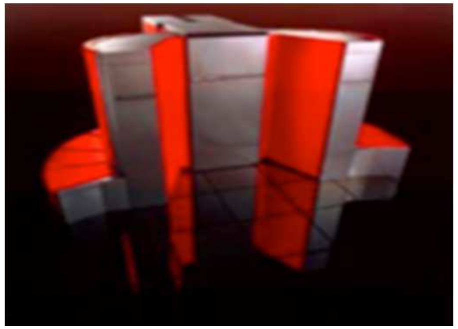
Evolution of the computing power of Supercomputers

FLOP/second (operaciones sobre números reales 64 bits)

E P T G

1000000000000000000

~2018
? ($1x10^7$ processadors



1988
Cray Y-MP (8 processadors)

2008
Cray XT5 (15000 processadors)

1998
Cray T3E (1024 processadors)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The “Formula 1” of Supercomputers today

FLOP/segon (operaciones sobre números reales de 64 bits)

E P T G

1000000000000000000

#1 (55 PF)

Tianhe @ National University of Defense Technology, 54.9 PFlops

Samsung Exynos > 50 Gflops
Prototips MontBlanc @ BSC

#1 Espanya (1PF)
MN3 @ BSC

#1 EU (5PF)

JUQUEEN @ Forschungszentrum Jülich

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

BSC-CNS objectives:

- R&D in Computer, Life, Earth and Engineering Sciences
- Supercomputing services and support to Spanish and European researchers

BSC-CNS is a consortium that includes:

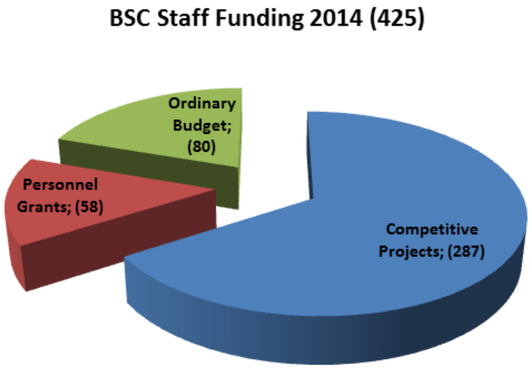
– Spanish Government	51%
– Catalonian Government	37%
– Universitat Politècnica de Catalunya (UPC)	12%

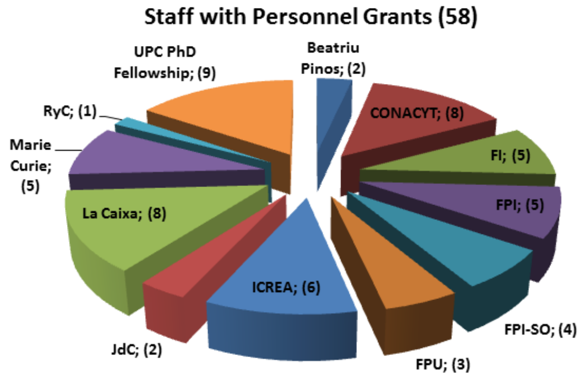
GOBIERNO DE ESPAÑA — MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Generalitat de Catalunya **Departament d'Economia i Coneixement**

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

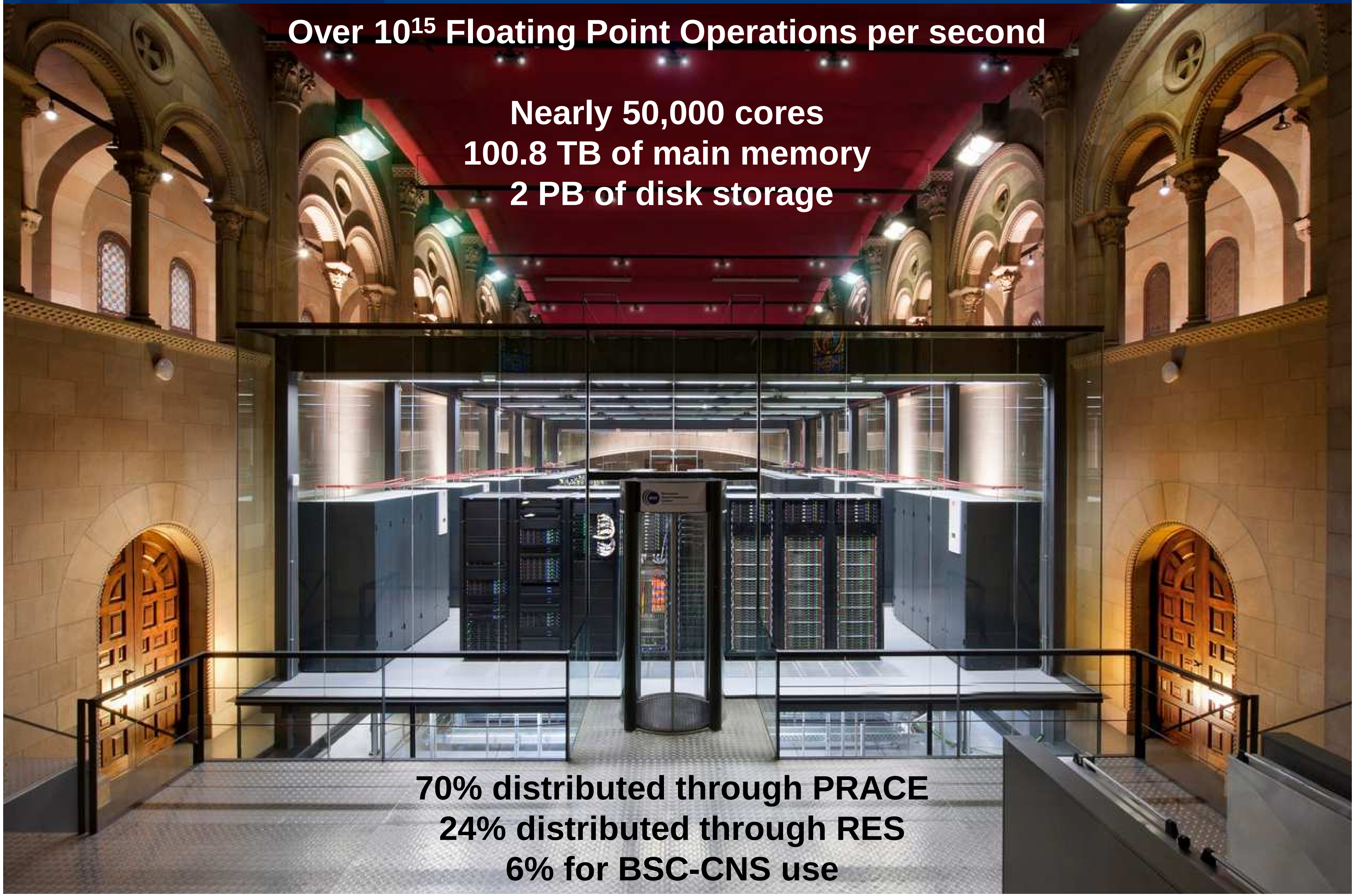
425 people, 41 countries





The MareNostrum 3 Supercomputer

EXCELENCIA SEVERO OCHOA



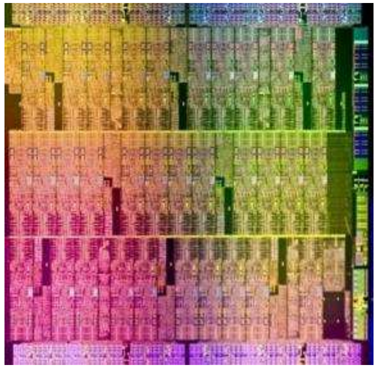
Over 10^{15} Floating Point Operations per second

Nearly 50,000 cores
100.8 TB of main memory
2 PB of disk storage

70% distributed through PRACE
24% distributed through RES
6% for BSC-CNS use

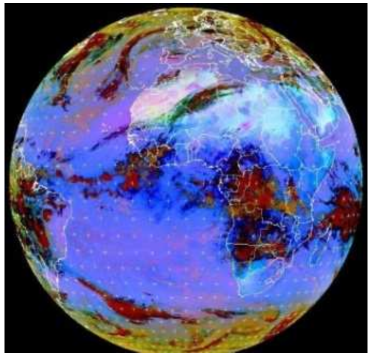
COMPUTER SCIENCES

To influence the way machines are built, programmed and used: programming models, performance tools, Big Data, computer architecture, energy efficiency.



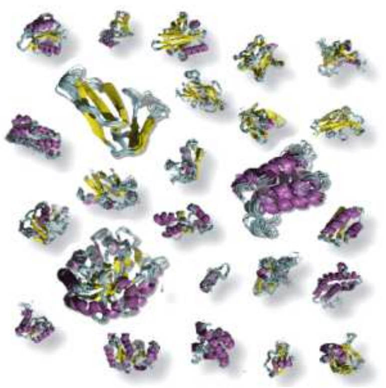
EARTH SCIENCES

To develop and implement global and regional state-of-the-art models for short-term air quality forecast and long-term climate applications.



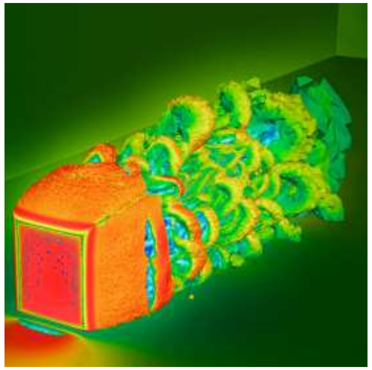
LIFE SCIENCES

To understand living organisms by means of theoretical and computational methods (molecular modeling, genomics, proteomics).



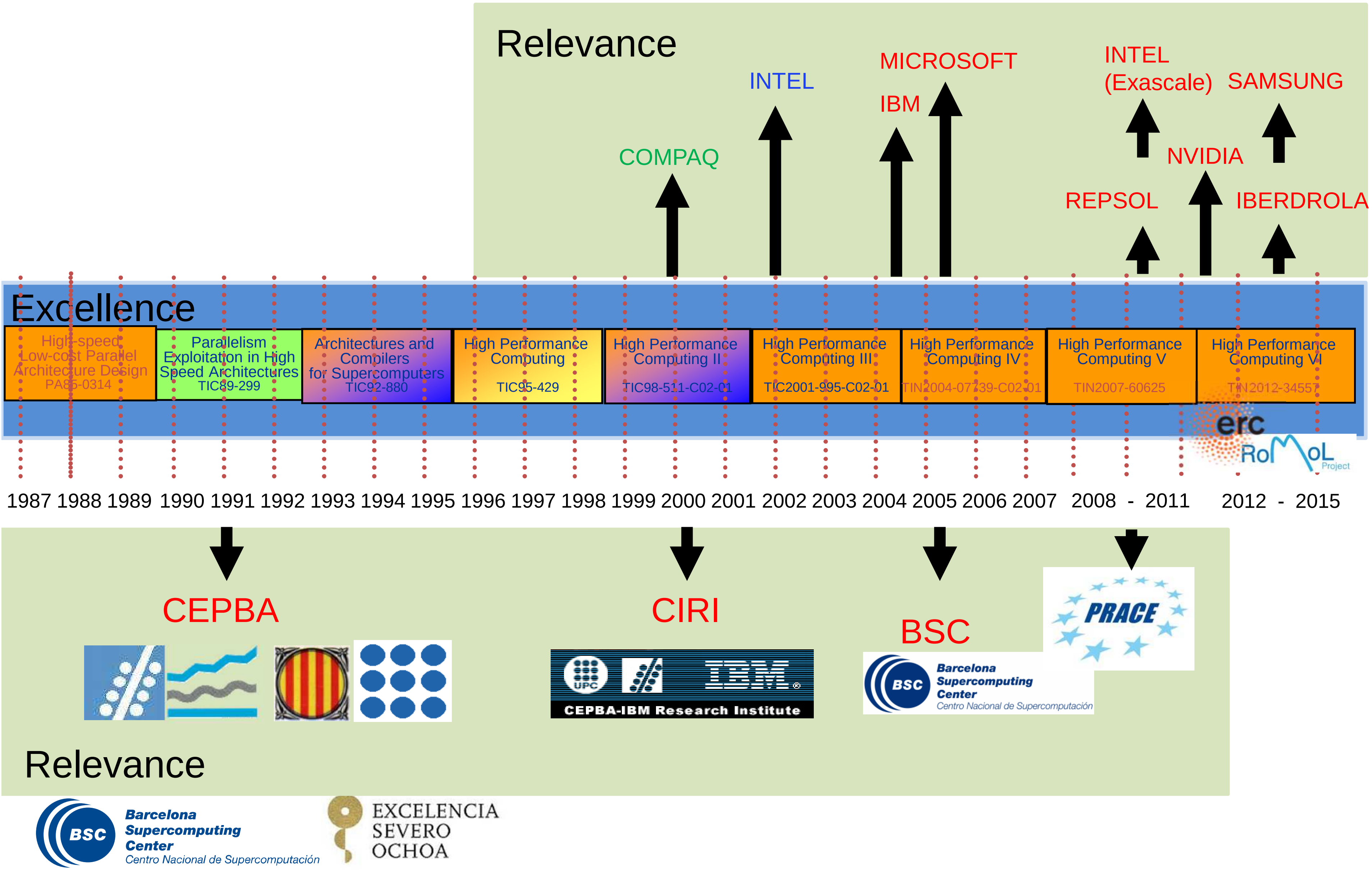
CASE

To develop scientific and engineering software to efficiently exploit super-computing capabilities (biomedical, geophysics, atmospheric, energy, social and economic simulations).

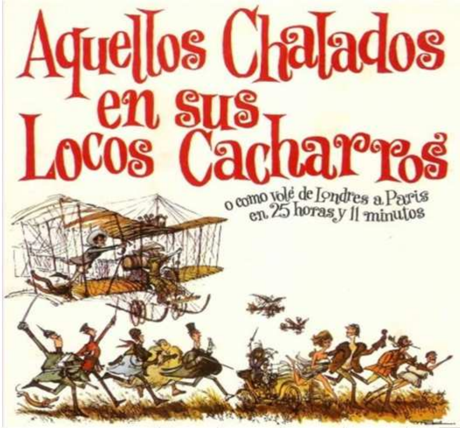


Our Origins.......Plan Nacional de Investigación

High-performance Computing group @ Computer Architecture Department (UPC)



Venimos de muy lejos …



CEPBA-IBM Research Institute

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Parsys Multiprocessor

Parsytec CCi-8D
4.45 Gflop/s

Compaq GS-140
12.5 Gflop/s

Compaq GS-160
23.4 Gflop/s

BULL NovaScale 5160
48 Gflop/s

Maricel
14.4 Tflops, 20 KW

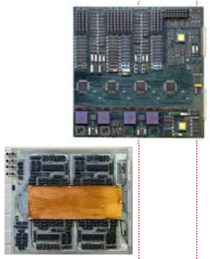
Transputer cluster

Convex C3800

SGI Origin 2000
32 Gflop/s

SGI Altix 4700
819.2 Gflops

SL8500
6 Petabytes



Research prototypes

Connection Machine CM-200
0,64 Gflop/s

IBM RS-6000 SP & IBM p630
192+144 Gflop/s

IBM PP970 / Myrinet
MareNostrum
42.35, 94.21 Tflop/s

1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

Venimos de muy lejos …

Ayto. Barcelona
Uitesa
UPC-EIO

AMES, CIMNE

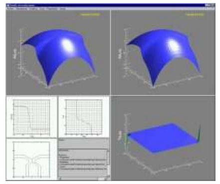
TGI
UPM-DATSI

Hesperia
Neosystem
s
UPC-EIO

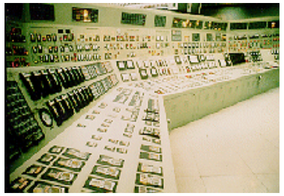
Soler y Palau
CIMNE
CEPBA-UPC

Torres
Soft Greenhouse
CEPBA-UPC

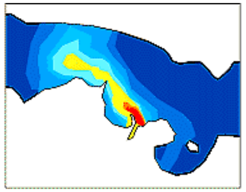
INDO, CEPBA-UPC



Metodos Cuantitativos
Gonfiesa
CESCA, CESGA



Tecnatom, UMA



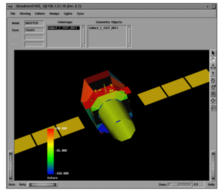
AZTI
UPC-LIM

Iberdrola, Uitesa, UPV

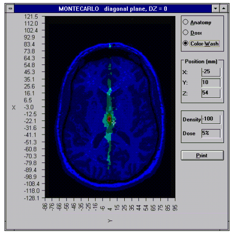
BCN COSIVER
Mides
UPC-EIO

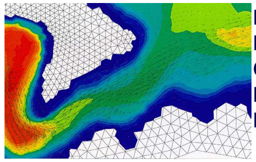
CEPBA
CESCA
UMA
UNICAN
UPM

CASA
Envision
GTD
Intespace
RUS



Ospedali Galliera
Le Molinette
Parsytec
PAC
EDS





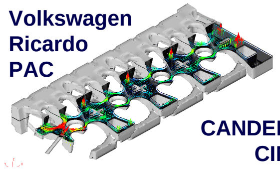
ENEL
EDF
CSR4
Reiter
Kenijoki

Ferrari, Genias, P3C

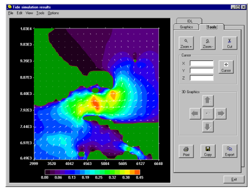
ST Mecanica
DERBI
AUSA
CEPBA-UPC

Intera SP
Intera UK
UPC-DIT
CEPBA-UPC

Volkswagen
Ricardo
PAC

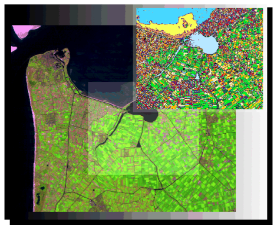


CANDEMAT
CIMNE
CEPBA-UPC

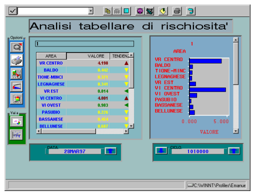


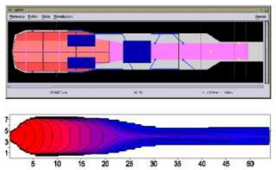
SENER
CIC
UNICAN

Italeco
Geospace
Intecs
Univ. Leiden



Cari Verona
AIS
PAC
Univ.
Cat. Milan

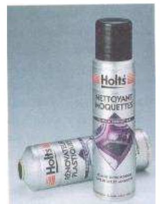


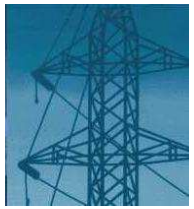


Cristaleria Española
UNICAN
CEPBA-UPC

Inisel Espacio
Infocarto
UPC-TSC
CEPBA-UPC

CEBAL-ENTEC
NEOSYSTEMS





Iberdrola
SAGE
CEPBA-UPC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Scientific mission

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
EXCELENCIA SEVERO OCHOA

What

Influence the way machines are …
… built …
… programmed …
… and used

Why

Our strength …
… critical mass of people …
… holistic/vertical vision/background …
… stable and exploratory paths …
… and co-design approach

How

Through ideas, …
… demonstration, …
… cooperation with manufacturers, …
… and "products"

- Programming environments
- Performance tools
- Resource management
- Architecture, from single core HPC to other areas

- Supercomputing
- BigData
- Multicores real-time
- Mobile and embedded

Performance, productivity, power/energy and reliability

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Objectives

What

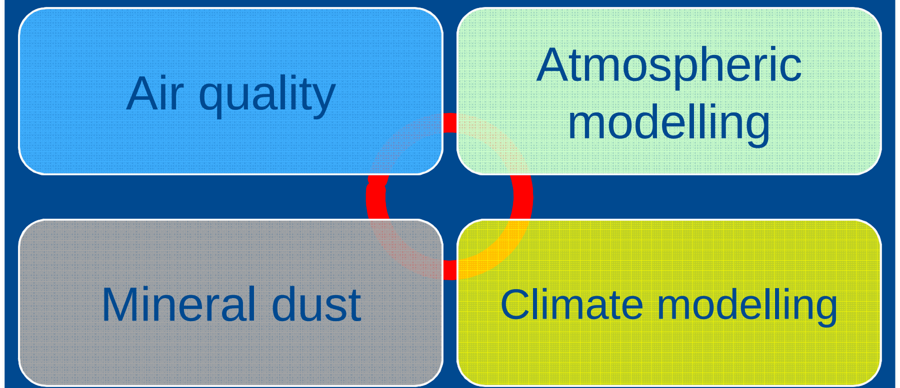
Perform research in Earth sciences for the development, implementation and refinement of global and regional state-of-the-art models for short-term air quality forecast and climate applications

Why

Our strength …
- … operations …
- … research …
- … service …
- … high resolution …

How

Bringing together knowledge in atmospheric dynamics, natural and anthropogenic emissions, improvement of air quality forecasting, transport and dispersion of pollutants in complex terrain, urban air quality, aerosol optical properties, aerosol radiative effects and the feedback between meteorology and air pollution with the advances in the parallelization of air quality model codes



Vision and mission

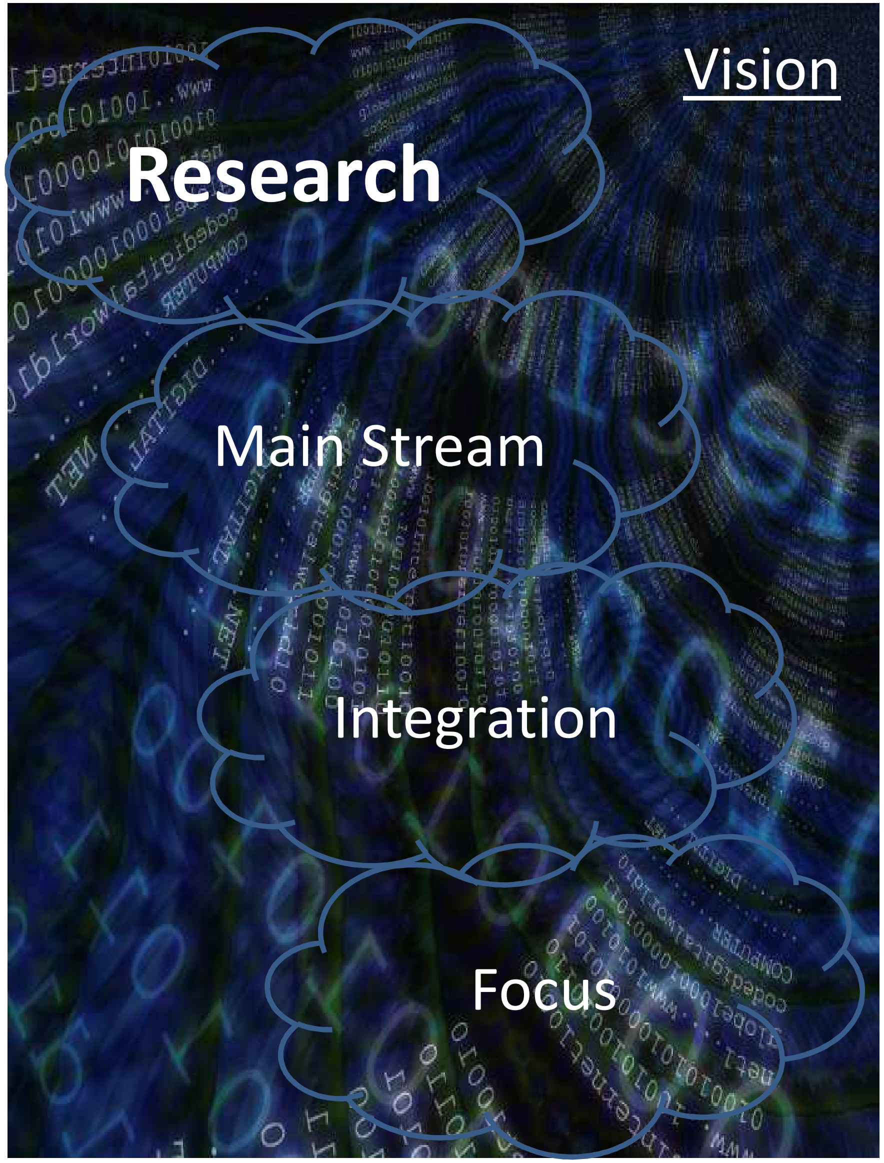
What

The understanding of living organisms by means of theoretical methods.

How

Use computational methods to get information and simulate biological systems …

… with the final goal of explaining biological systems from the basic rules of physics and chemistry



Scientific mission

What

Develop relevant simulation software …

… science

… engineering

Why

Our strength …

… multidisciplinary background

… access to hardware

… co-design approach

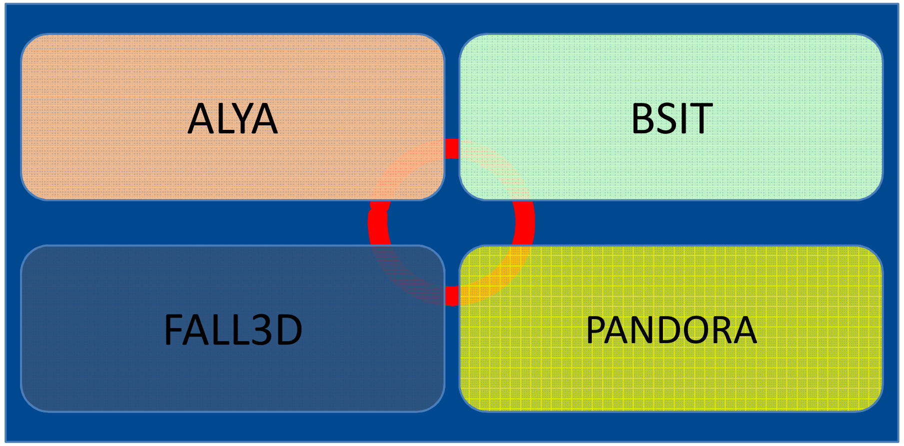
How

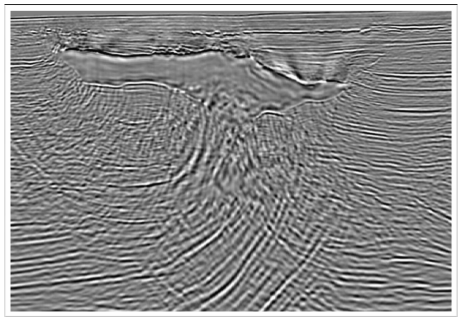
Close contact with industry …

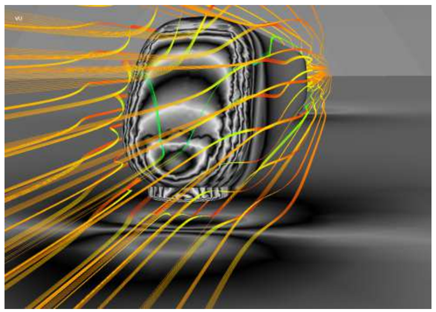
… innovative manufacturing

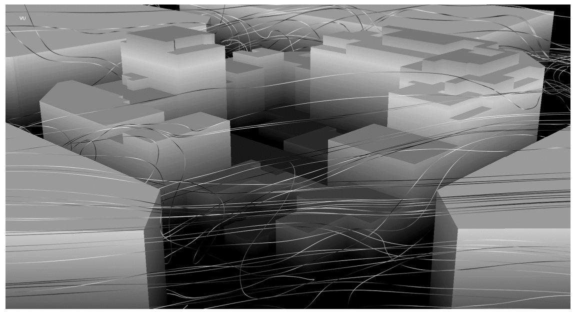
… energy

… pharmaceutical







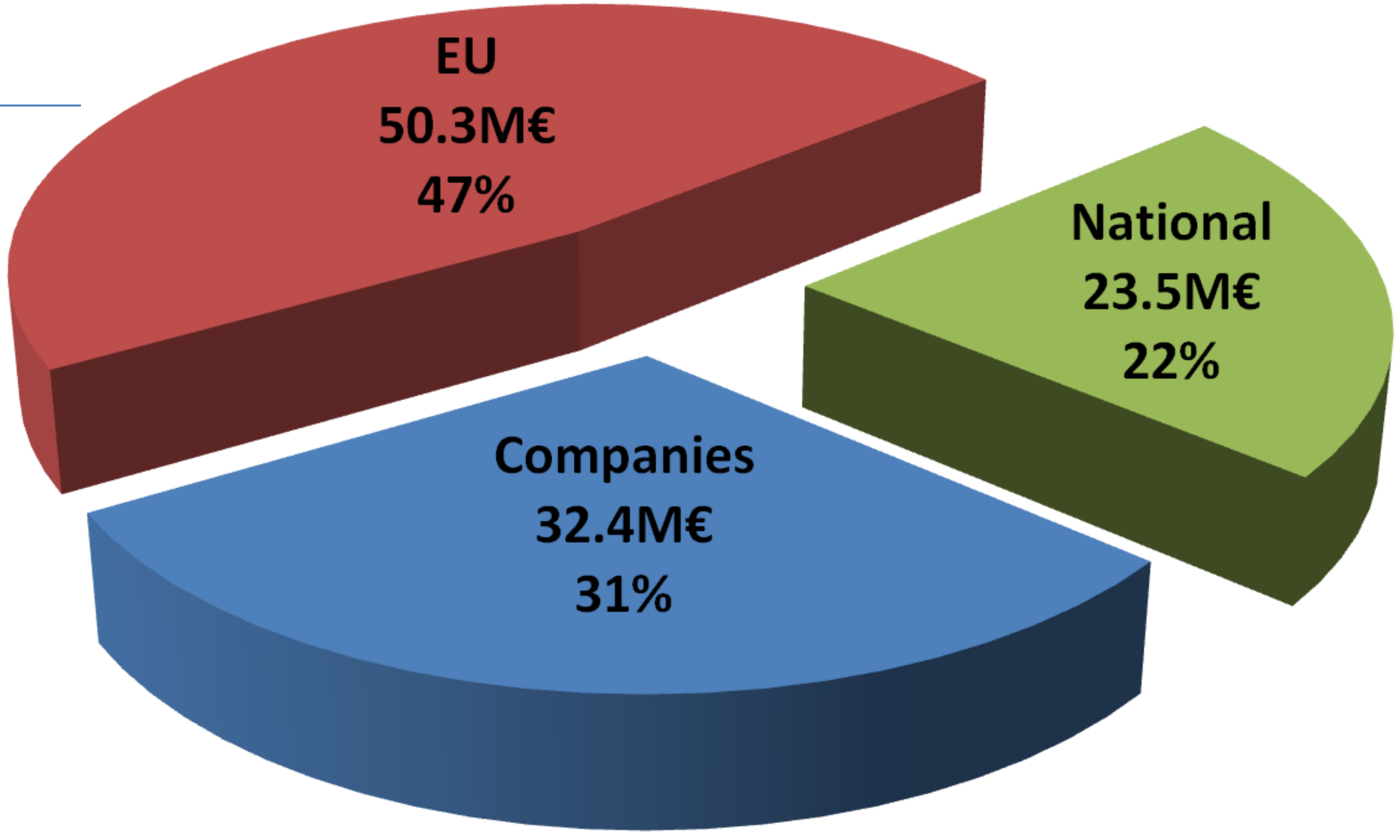


EXCELENCIA SEVERO OCHOA

Over 106M€ in grants and contracts

BSC achieved the 9th greatest return from FP7 of all Spanish institutions according to CDTI report 11/11/2013

Fourteen H2020 projects approved or under negotiation (Grant value: 5.97M€)



- Includes personnel grants and future income for approved projects 21/01/2015
- National income includes ICTS calls

Joint Research Centres with IT Companies

EXCELENCIA SEVERO OCHOA

BSC-IBM Technology Center for Supercomputing

Future challenges for supercomputers including power efficiency and scalability, new programming models, and tools for analysis and optimization of applications

BSC-NVIDIA CUDA Center of Excellence

Training in Parallel Programming using CUDA and StarSs Optimising management of execution resources in multi-GPU environments with GMAC

BSC-Microsoft Research Centre

Analysis of Hadoop workload performance under different software parameters and hardware configurations.
Results available online

Intel-BSC Exascale Lab

Multi-year agreement focussing on optimising efficiency through research into Programming Models, Performance Tools and Applications

Agreement on memory performance in HPC systems with SAMSUNG

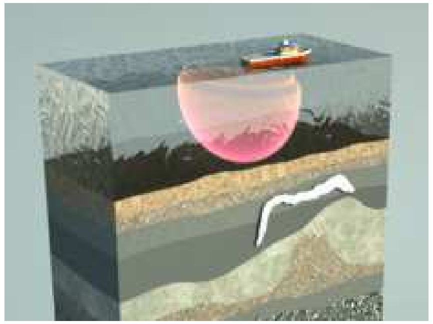
Key Spanish Industrial Partners

EXCELENCIA SEVERO OCHOA

Repsol-BSC Research Center

REPSOL

Research into advanced technologies for the exploration of hydrocarbons, subterranean and subsea reserve modelling and fluid flows



Iberdrola Renovables

IBERDROLA

Design and optimization of wind farms

JUAN YACHT DESIGN
Juan Kouyoumdjian · Naval architecture

INTELLIGENT PHARMA

Telefónica
Telefónica Investigación y Desarrollo

ANAXOMICS
Systems biology solutions

Termo Fluids

FUNDACIÓN BOTÍN

Mind the Gap: 500K €
Nostrum Drug Discovery
Spin off

erc

Proof of Concept:
150K€ PELE

Patents

- dpEDMD: method for drug discovery
- SMUFIN: Somatic Mutations Finder
- miRNA markers for Morbid obesity diseases

Some Strategic Projects

EXCELENCIA SEVERO OCHOA

Severo Ochoa

A multidisciplinary research program to address the complex challenges in the path towards Exascale. A set of key strategic scientific projects and improvements in HR management, training, mobility and communication.

Human Brain Project

10-year FET Flagship research project to simulate human brain and design computers based on its workings

Mont-Blanc

Developing an European Exascale approach Based on embedded power-efficient technology

Personalised Medicine

Combining genomics, proteomics and transcriptomics with computer simulation to support predictive medicine and personalized treatment

Riding on Moore's Law

Optimizing performance, energy consumption and reliability of parallel computer architectures through higher level abstraction (ERC M Valero)

Alya Red

Computational mechanics simulation tools designed for biomedical research. Winner 2012 Science-NSF visualisation challenge

Internationalisation

HPC Infrastructure for Europe's Best Scientists

Over 90 EC Framework Projects

Part of Future Spanish node of ICT-Labs

Bridge between EU and Latin America

Helping to define the future of global HPC

EXCELENCIA SEVERO OCHOA

Enabling the Data Revolution

International Roadmapping

Leadership in Exascale

Contributing to Standardisation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-CNS, PRACE Hosting Member

PRACE

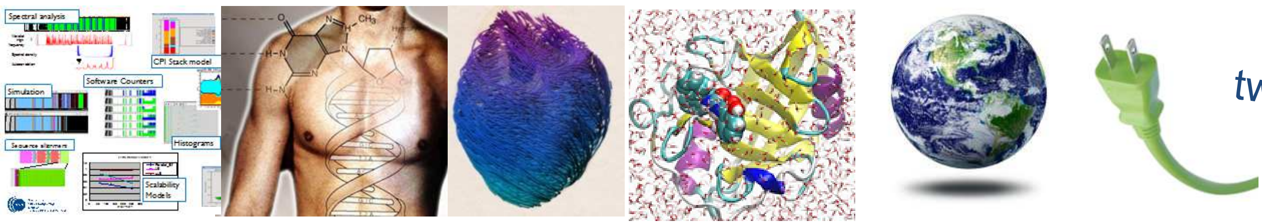
ETP 4 HPC
THE EUROPEAN TECHNOLOGY PLATFORM FOR HIGH PERFORMANCE COMPUTING

BSC-CNS, ETP Founding Member

EXCELENCIA SEVERO OCHOA
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Access to **best HPC resources** for industry and academia

Autonomous EU development of **Exascale technologies**



BSC-CNS leads two CoE proposals, participates in others

Centers of Excellence in **HPC applications**
SME Competence Centers

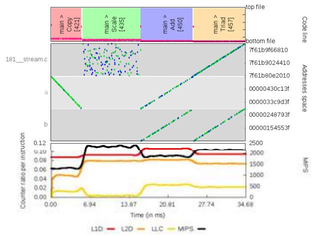
Convergence of HPC, Big Data & Embedded systems

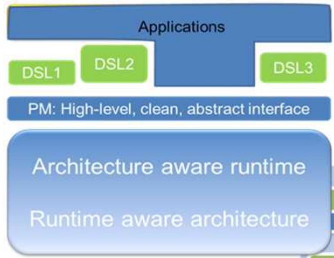
- Maintain leadership and visibility
  - On programming models and performance analytics
  - More platforms, more intelligence
  - More Apps, engage with communities
  - Influence standards

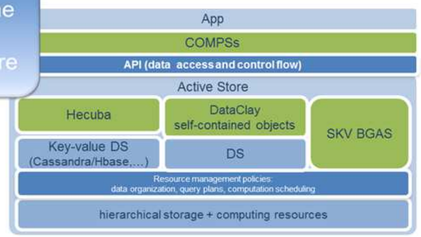
- Push forward
  - Architectures for real-time
  - Convergence of HPC and BigData
  - Runtime-aware multicore architecture
    - RoMoL, Mont-Blanc 3

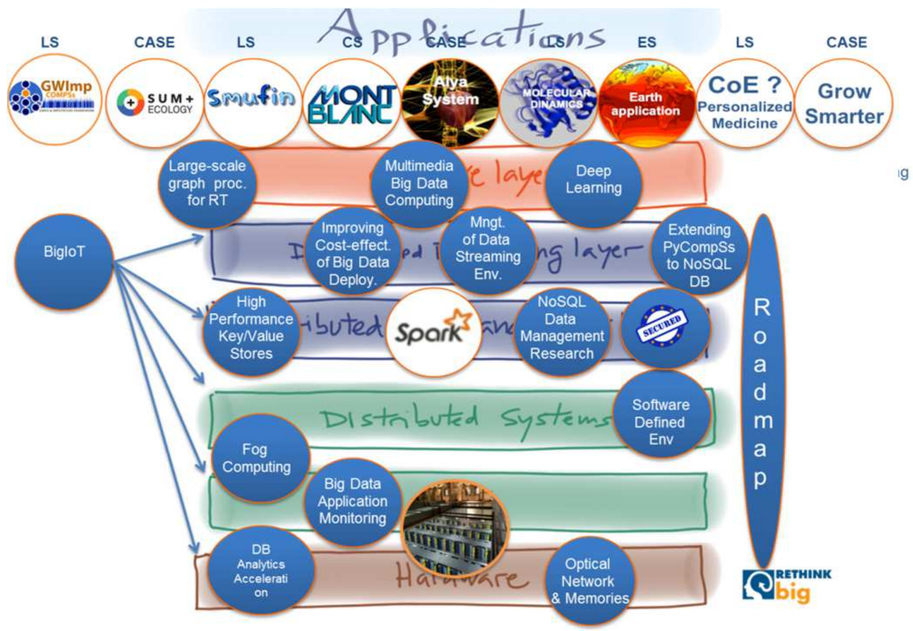
- Further exploration
  - Convergence of embedded and HPC, IoT
  - Algorithmic development in different domains
  - Cognitive computing

OpenMP Forerunner







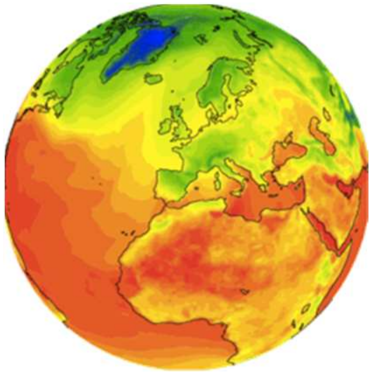


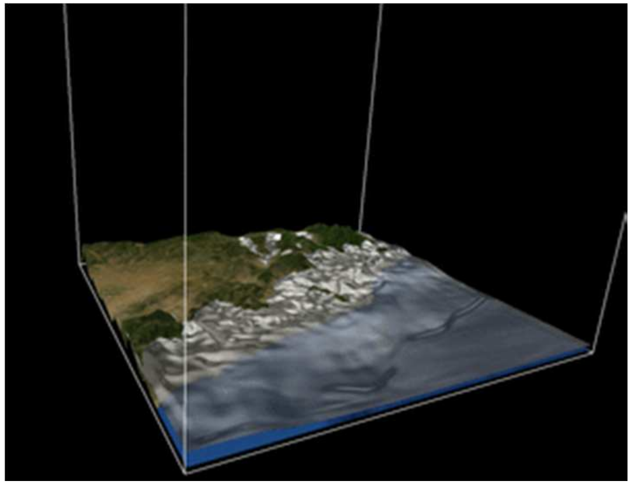
Environmental Forecasting for Services

- Development of the best knowledge of plausible predictions of the direction, scope, speed and intensity of a set of environmental changes.

- Bringing together the human activity and the environment through the most efficient combination of monitoring and modeling.

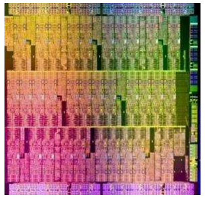
- Focused on the development of modeling solutions to provide weather, climate and air quality information at a global scale for the public and private sectors, with a special interest in the Mediterranean, African, Arctic and South American regions.



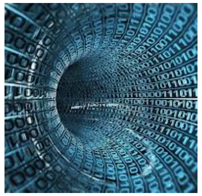


IT&LS convergence for personalized medicine

Computer Sciences



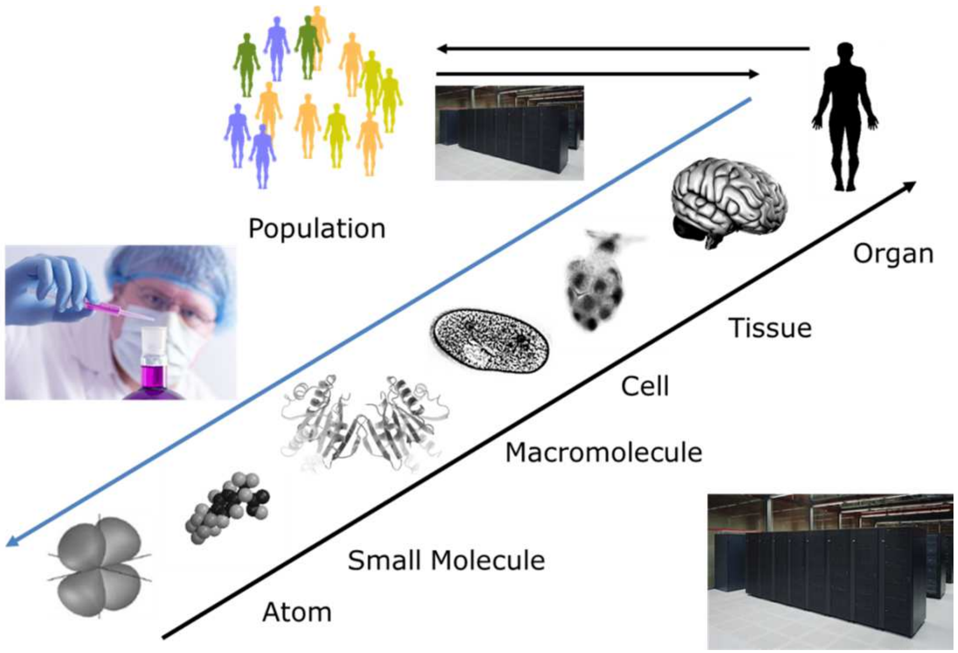
Architecture



Big Data

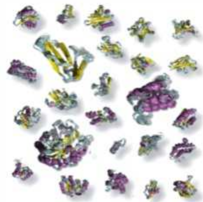
Software

Personalized Medicine Model



Life Sciences

Genomics



Modeling

Chemistry

Solutions for Biotech, Pharma and Clinical

GENOTYPING | GWAS | DIAGNOSIS/ PROGNOSIS | LEAD GENERATION | DRUG REPOSITION | DRUG RESISTANCE

- Integrate Personalized Medicine into the HPC ecosystem
- Implement an HPC platform serving the needs of the Personalized Medicine community

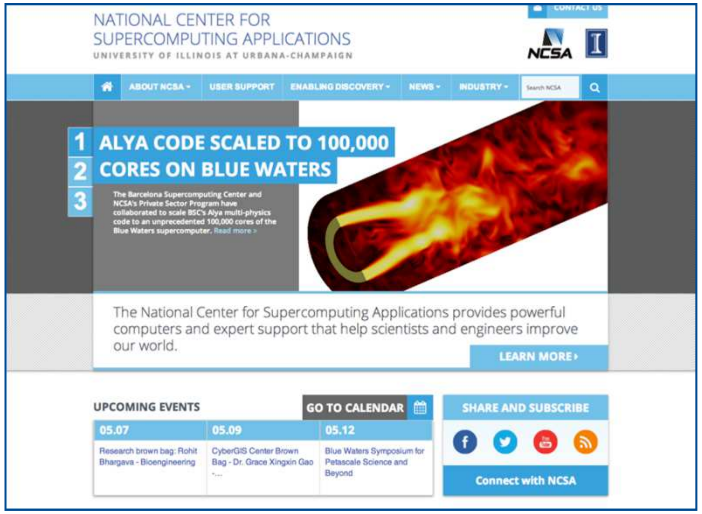
Alya software consolidation as standard for HPC modelling

Alya is the BSC in-house simulation code:

- Coupled multi-scale and multi-physics
- Complex simulation scenarios
- Parallel efficiency in supercomputers

In 2014 it becomes the first in its class, simulating three complex multi-physics problems in 100.000 cores (“2014 Top Supercomputing Achievement” by HPCWire and announced in SC).

The target is 1.000.000 cores in 2016: CODEX project

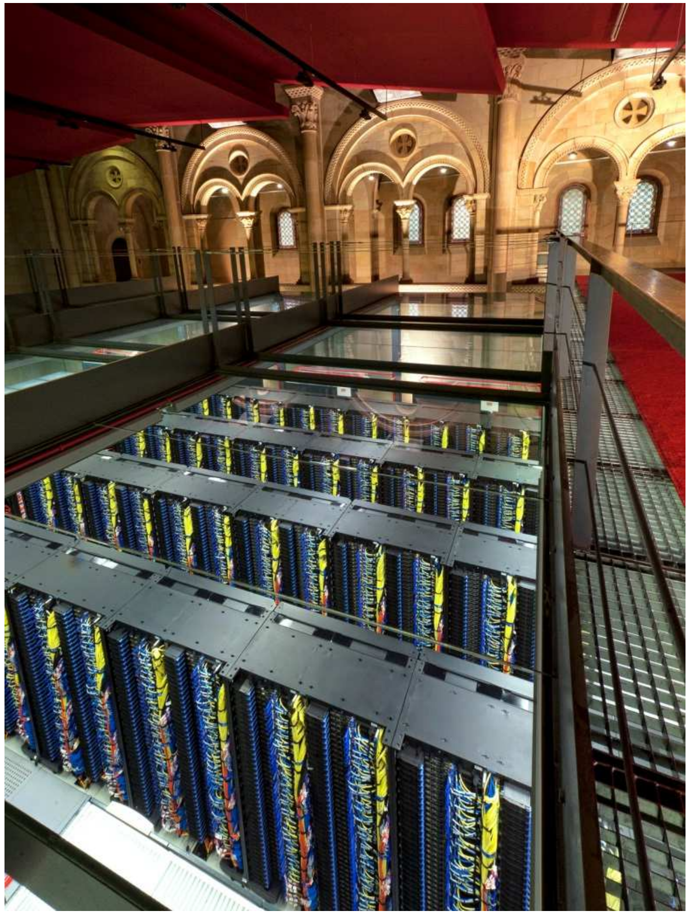


HPC & Data Analytics service convergence

- Massive storage capabilities.

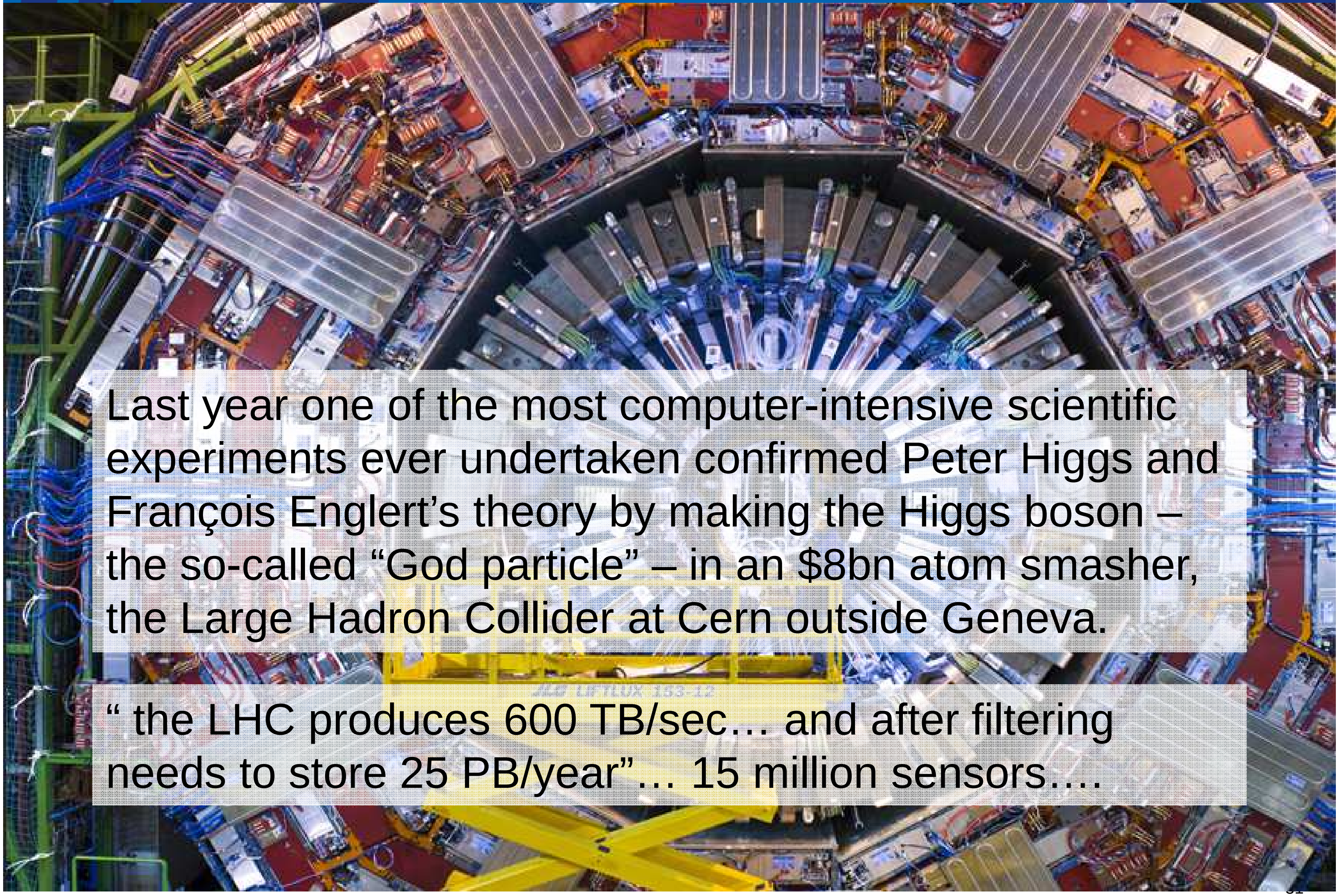
- Workflows definition.

- Adequate User Support skills.



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

THE BIG DATA ERA: DATA GENERATION EXPLOSION



Last year one of the most computer-intensive scientific experiments ever undertaken confirmed Peter Higgs and François Englert’s theory by making the Higgs boson – the so-called “God particle” – in an $8bn atom smasher, the Large Hadron Collider at Cern outside Geneva.

“ the LHC produces 600 TB/sec… and after filtering needs to store 25 PB/year”… 15 million sensors….

Big Data in Biology

TECHNOLOGY FEATURE

THE BIG CHALLENGES OF BIG DATA

As they grapple with increasingly large data sets, biologists and computer scientists uncork new bottlenecks.

Extremely powerful computers are needed to help biologists to handle big-data traffic jams.

BY VIVIEN MARX

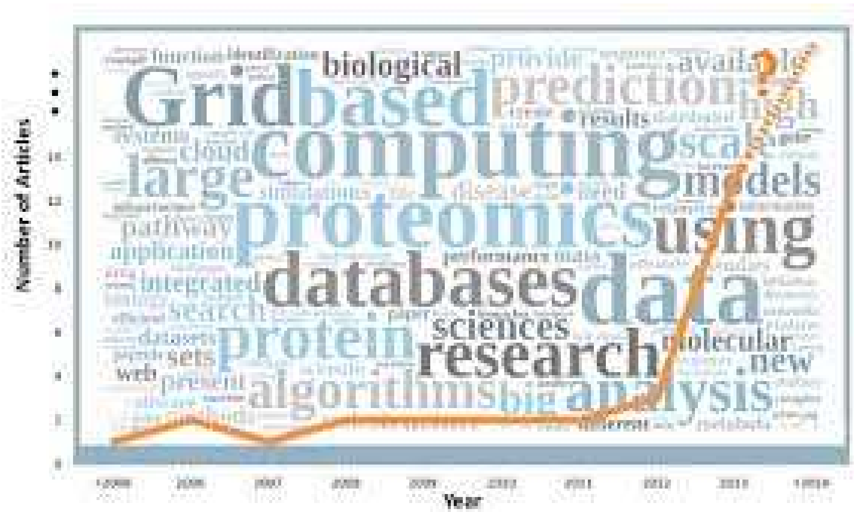
Biologists are joining the big-data club. With the advent of high-throughput genomics, life scientists are starting to grapple with massive data sets, encountering challenges with handling, processing and moving information that were once the domain of astronomers and high-energy physicists[1].

With every passing year, they turn more often to big data to probe everything from the regulation of genes and the evolution of genomes to why coastal algae bloom, what microbes dwell where in human body cavities and how the genetic make-up of different cancers influences how cancer patients fare[2]. The European Bioinformatics Institute (EBI) in Hinxton, UK, part of the European Molecular Biology Laboratory and one of the world's largest biology-data repositories, currently stores 20 petabytes (1 petabyte is 10^{15} bytes) of data and back-ups about genes, proteins and small molecules. Genomic data account for 2 petabytes of that, a number that more than doubles every year[3] (see 'Data explosion').

This data pile is just one-tenth the size of the data store at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. Every year, particle-collision events in CERN's Large Hadron Collider generate around 15 petabytes of data — the equivalent of about 4 million high-definition feature-length films. But the EBI and institutes like it face similar data-wrangling challenges to those at CERN, says Ewan Birney, associate director of the EBI. He and his colleagues now regularly meet with organizations such as CERN and the European Space Agency (ESA) in Paris to swap lessons about data storage, analysis and sharing.

All labs need to manipulate data to yield research answers. As prices drop for high-throughput instruments such as automated

- High resolution imaging
- Clinical records
- Simulations
- Omics

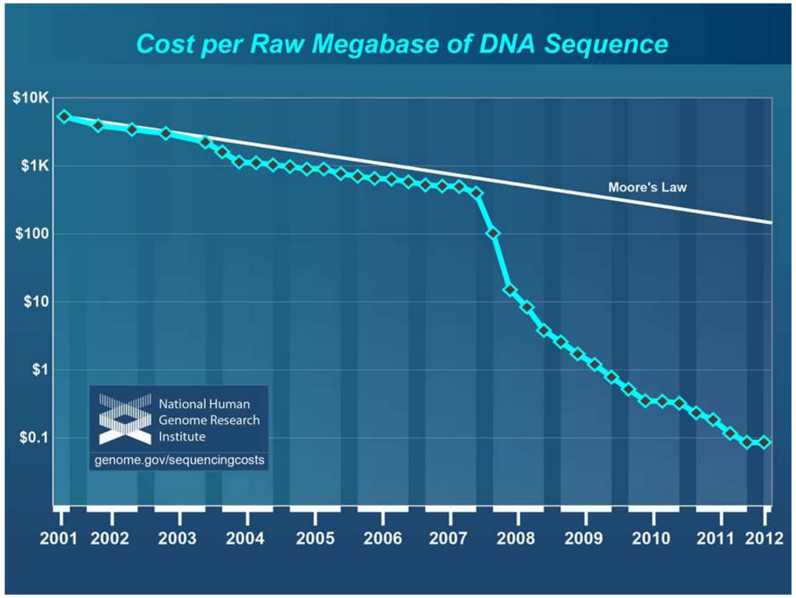


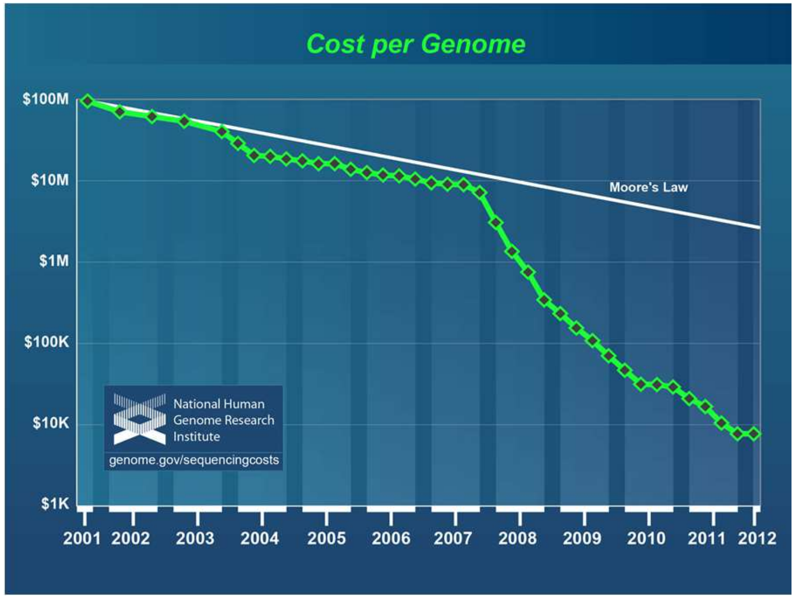
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

RoMoL Project

Sequencing Costs



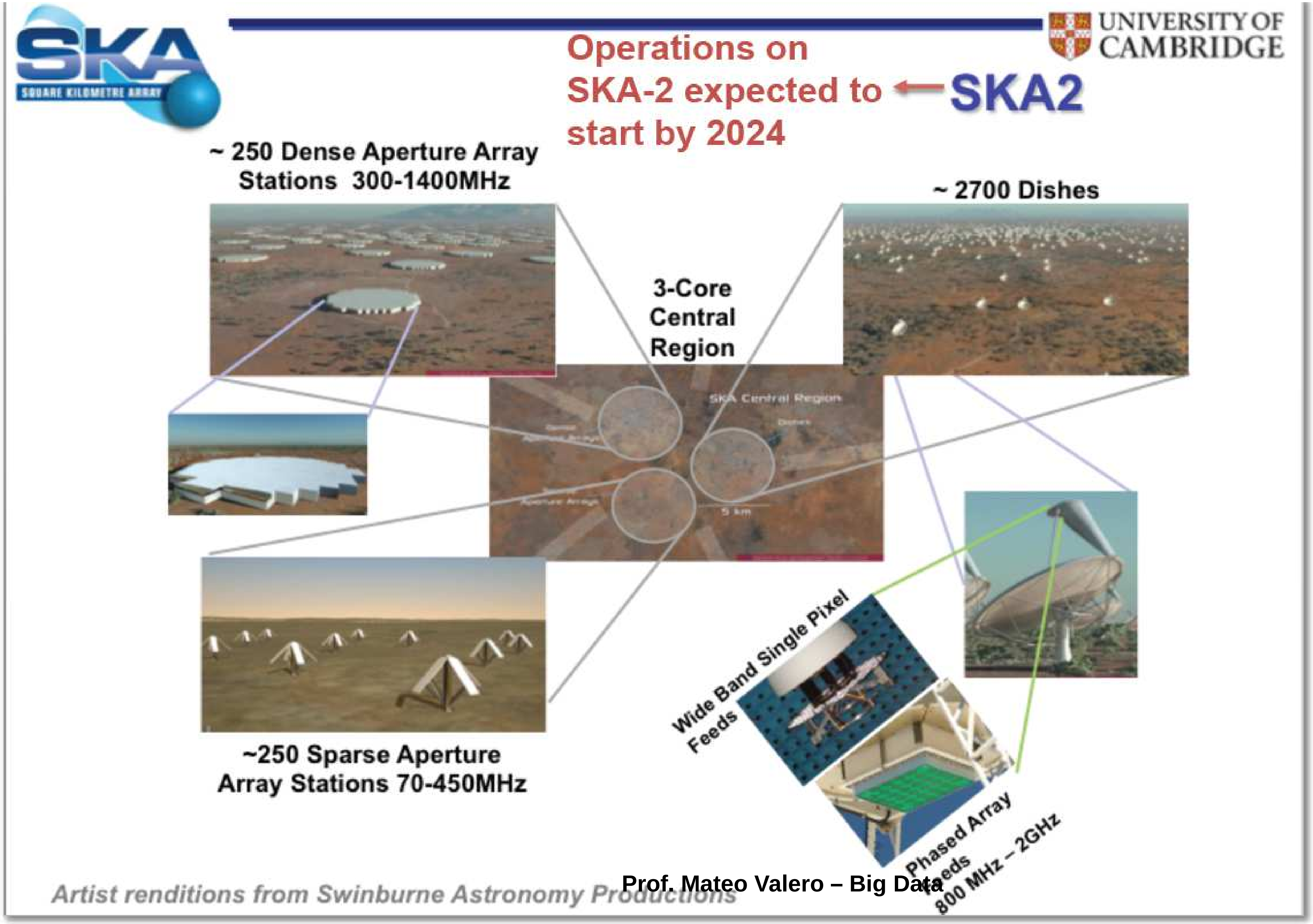


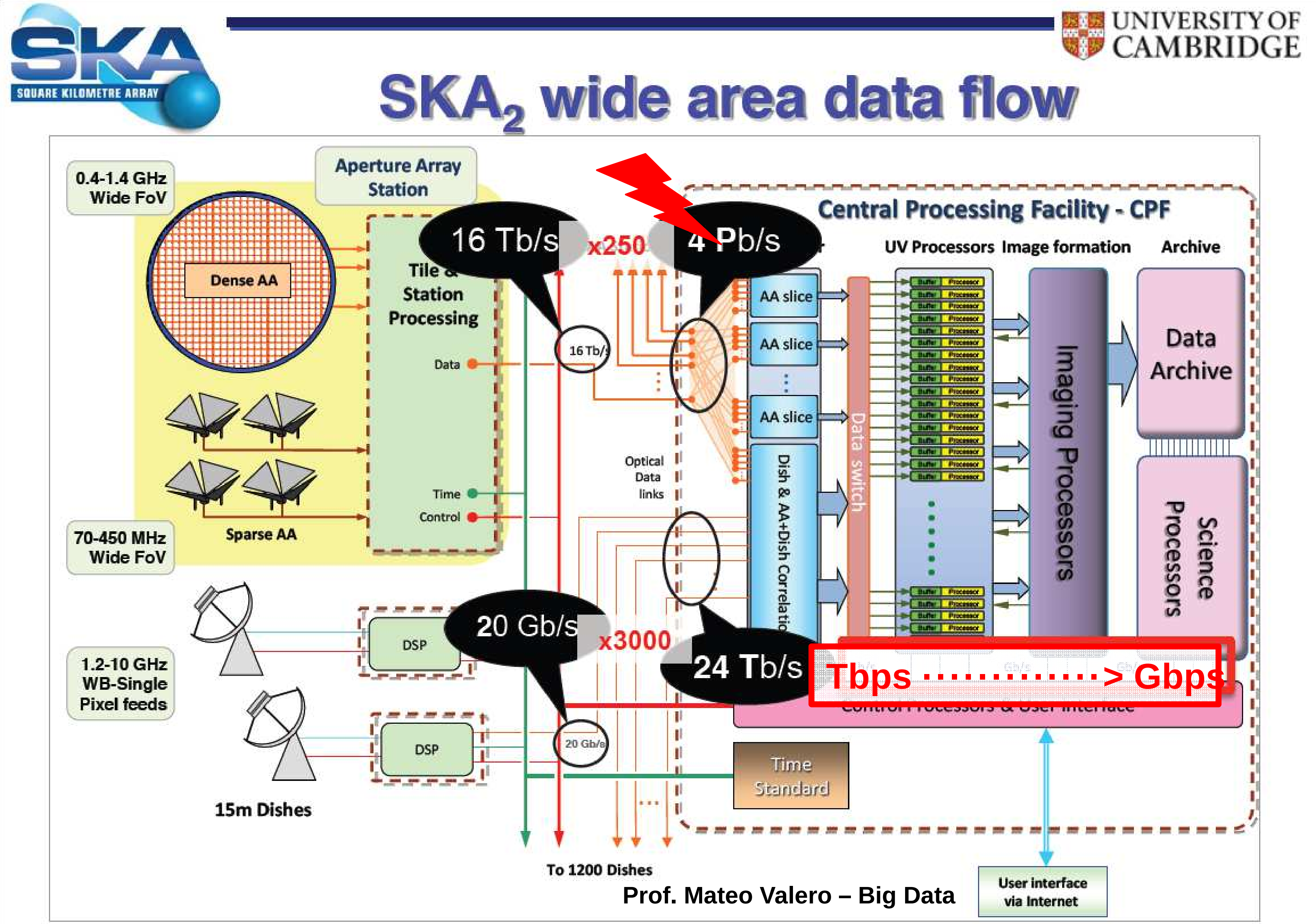
Source: **National Human Genome Research Institute** (NHGRI)
http://www.genome.gov/sequencingcosts/

(1) "Cost per Megabase of DNA Sequence" — the cost of determining one megabase (Mb; a million bases) of DNA sequence of a specified quality
(2) "Cost per Genome" - the cost of sequencing a human-sized genome. For each, a graph is provided showing the data since 2001

In both graphs, the data from 2001 through October 2007 represent the costs of generating DNA sequence using Sanger-based chemistries and capillary-based instruments ('first generation' sequencing platforms). Beginning in January 2008, the data represent the costs of generating DNA sequence using 'second-generation' (or 'next-generation') sequencing platforms. The change in instruments represents the rapid evolution of DNA sequencing technologies that has occurred in recent years.

Square Kilometer Array (SKA)




SKA
SQUARE KILOMETRE ARRAY
UNIVERSITY OF CAMBRIDGE
SKA2 wide area data flow
0.4-1.4 GHz Wide FoV
Aperture Array Station
Dense AA
Tile & Station Processing
Data
Time
Control
Sparse AA
70-450 MHz Wide FoV
16 Tb/s
x250
4 Pb/s
16 Tb/s
Central Processing Facility - CPF
UV Processors
Image formation
Archive
AA slice
AA slice
AA slice
Data switch
Dish & AA+Dish Correlation
Imaging Processors
Data Archive
Science Processors
Optical Data links
1.2-10 GHz WB-Single Pixel feeds
DSP
DSP
20 Gb/s
x3000
24 Tb/s
20 Gb/s
Tbps ·············> Gbps
Control Processors & User interface
Time Standard
15m Dishes
To 1200 Dishes
User interface via Internet

Current infrastructure sagging under its own weight

2013

- **98,000** tweets
- **23,148** apps downloaded
- **400,710** ad requests
- **2000** lyrics played on Tunewiki
- **1,500** pings sent on PingMe
- **208,333** minutes Angry Birds played

In 60 sec today

Internet of Things

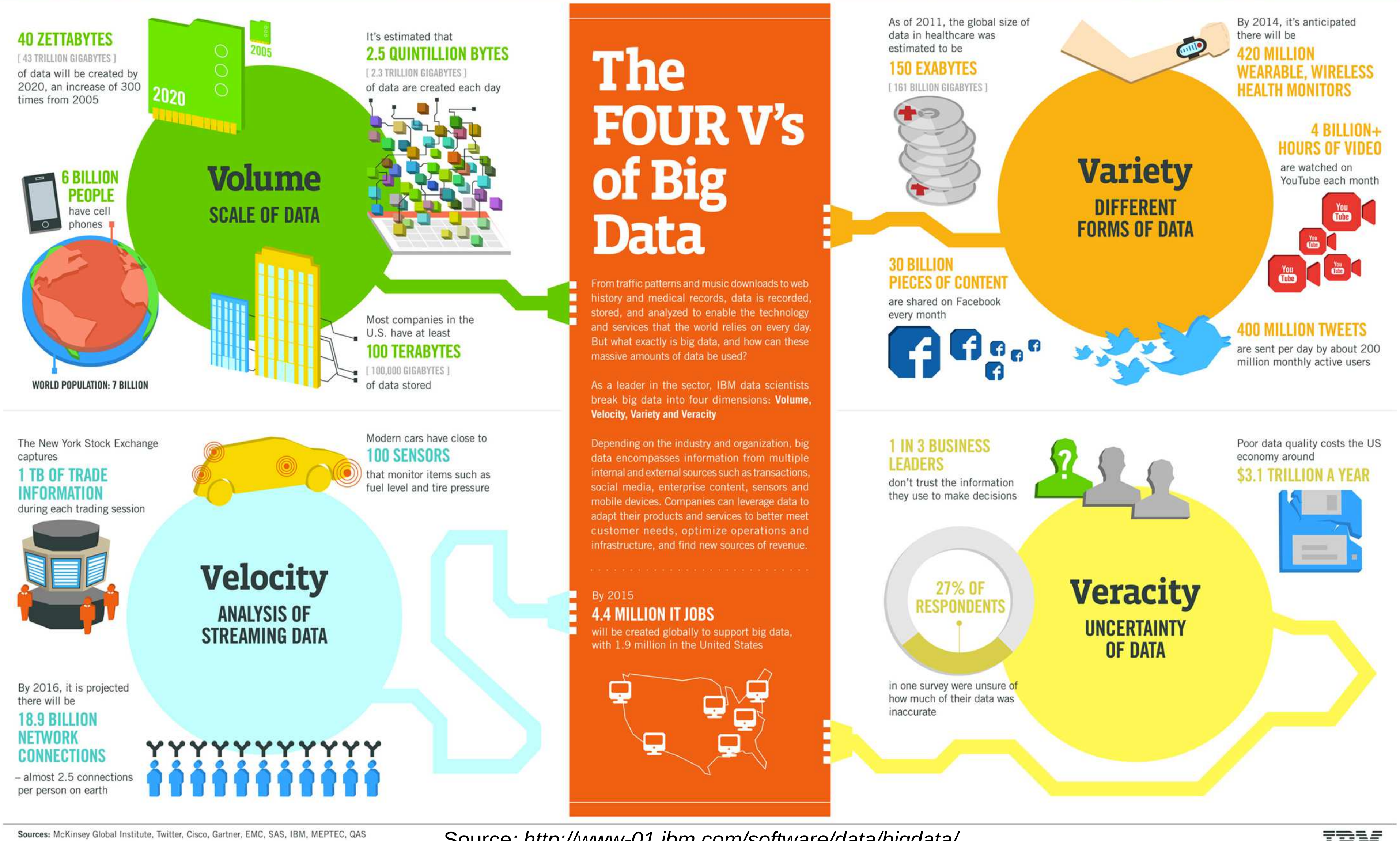
Pervasive Connectivity | Smart Device Expansion | Explosion of Information

By 2020

- **30 Billion** (1) Devices
- **40 Trillion GB** (2) DATA
- **10 Million** (3) Mobile Apps
- **… for 8 Billion** (4)

(1) IDC Directions 2013: Why the Datacenter of the Future Will Leverage a Converged Infrastructure, March 2013, Matt Eastwood ; (2) & (3) IDC Predictions 2012: Competing for 2020, Document 231720, December 2011, Frank Gens; (4) http://en.wikipedia.org

Challenges of data generation



Source*: http://www-01.ibm.com/software/data/bigdata/*

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

RoMoL Project

Internet & Big Data

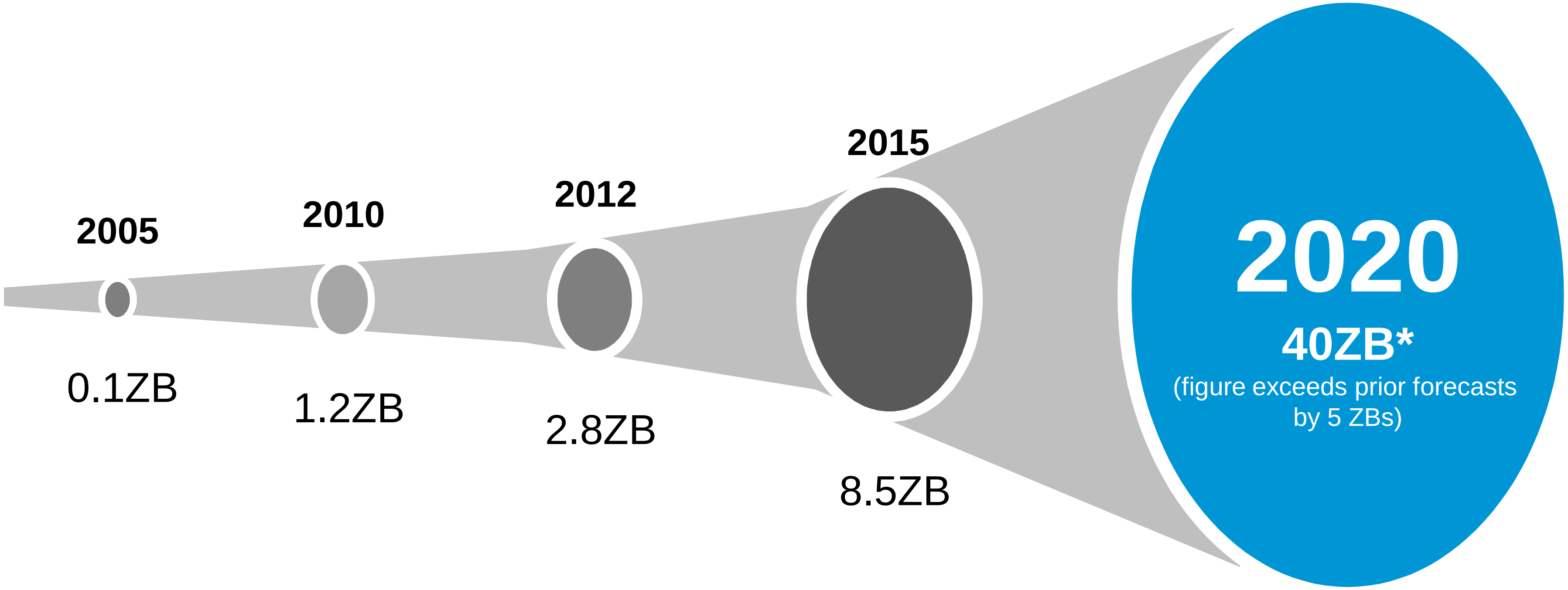
ALMACENAMIENTO DE DATOS
Del byte al yottabyte

El byte es una unidad de información utilizada como un múltiplo del bit que equivale a 8 bits. Se usa comúnmente como unidad de información en dispositivos de almacenamiento de datos

(Múltiplos del byte utilizando los prefijos del Sistema Internacional)

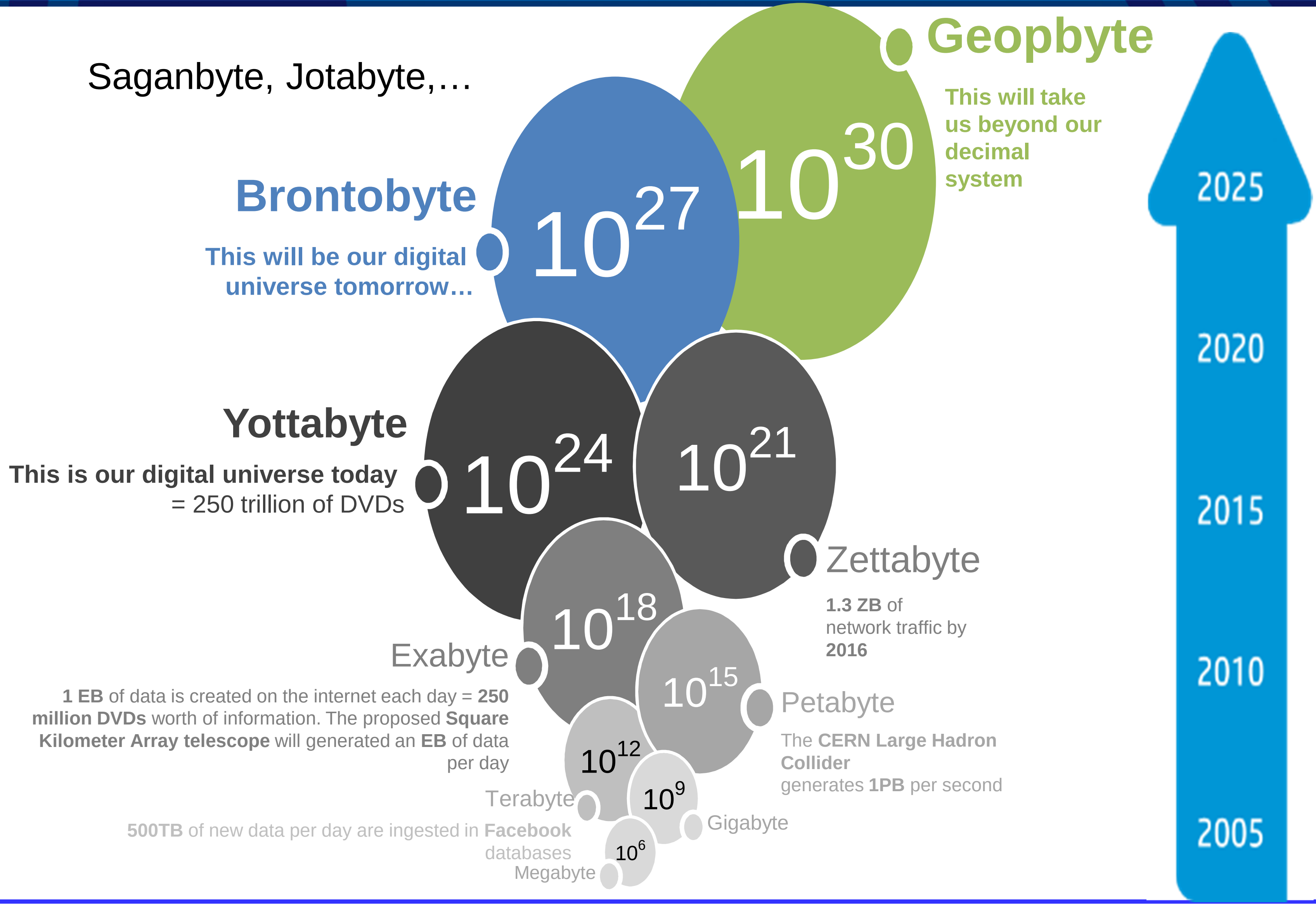
Número de bytes	Múltiplo	Símbolo	Nombre	¿Cuánto cabe? (Equivalencia aproximada)
1	10^0	1 B	Byte	Una letra '1'
10	10^1	10 B		Dos palabras 'usuario' 'descargas'
100	10^2	100 B		Una o dos frases *'En 100 bytes caben entre una y dos frases'*
1.000	10^3	1 kB	Kilobyte	Una historia muy corta
10.000	10^4	10 kB		Una página de enciclopedia
...	10^5	100 kB		Una foto de resolución media
	10^6	1 MB	Megabyte	Una novela
	10^7	10 MB		Dos copias de la obra completa de Shakespeare
	10^8	100 MB		Un estante de 1 m de libros
	10^9	1 GB	Gigabyte	Una furgoneta llena de hojas con texto
	10^{12}	1 TB	Terabyte	La pila de hojas con texto tendría 16 km de altura, los aviones comerciales vuelan a 13 km
	10^{13}	10 TB		La colección impresa de la Biblioteca del Congreso de EE.UU. Entre libros y documentos, 147 millones de archivos
	10^{15}	10 PB	Petabyte	Filmar la vida de una persona (100 años) en alta definición
	10^{18}	10 EB	Exabyte	Entre 1/5 y 1/9 del tráfico anual de internet
	10^{21}	10 ZB	Zetabyte	La mitad del contenido generado en el 2011
	10^{24}	10 YB	Yottabyte	Un billón de terabytes

The Data Deluge



* Source: IDC

How big is big?



The data explosion is transforming science

- Every area of science is now engaged in data-intensive research
- Researchers need
  - Technology to publish and share data in the cloud
  - Data analytics tools to explore massive data collections
  - A sustainable economic model for scientific analysis, collaboration and data curation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BIG DATA TECHNOLOGIES:

Outline

- Big Data Technologies
  - Storing data
  - Processing data
  - Where do we place data?
  - Managing Big Data

Magnetic Tape Memory

- Invented by Eckert & Mauchly for the UNIVAC I, March, 21,1951
  - Model UNISERVO
  - 224 KB of data
  - 1/2 inches of diameter,1200 feets, 128 characters per inch
  - Speed: 100 inches per second, equivalent to 12800 characters per second

- Storage Tech, 2013, T-10000D,
  - 8.5 Terabytes (40.000.000 increase)
  - EBW: 250 Mbytes/second (20.000 increase)
  - Load Time: 10 seconds

Tapes advancing fast

May 2014

IBM and Fujifilm have demoed a **154TB** LTO-size tape cartridge which could come to market in **10 years' time.**

The Sony development involved a **148Gbit/in2** tape density and its own tape design to achieve a **185TB** uncompressed capacity.

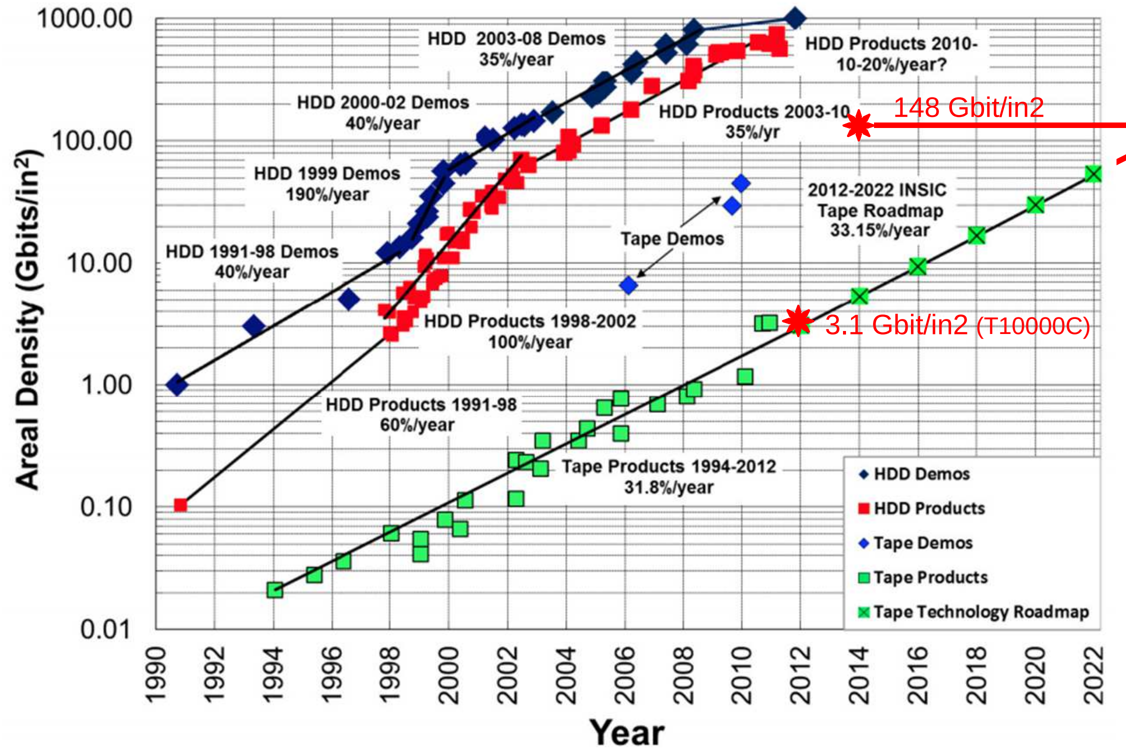
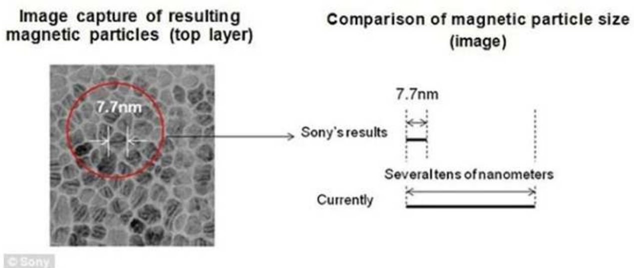


Figure 40: Areal Density of Hard Disk and Tape Laboratory Demonstrations and Products [71].



http://www.insic.org/news/2012Roadmap/12index.html

HDD: Hard Drives Disk

- 1956, IBM 305 RAMAC
- 4 MB, 50x24” disks, 1200 rpm, 100 bits/track
- Intertracks: 0.1 inches, Density: 1000 bits/in2
- 100 ms access , Tubes, 35k$/y rent

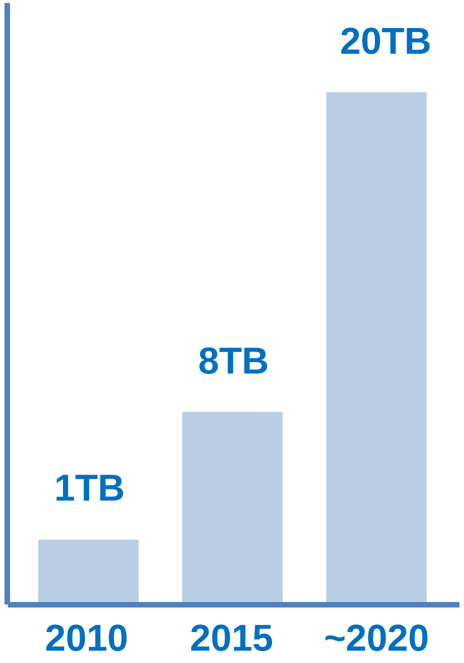
- Year 2013: 4 Terabytes (1.000.000 increase)
- Average access time: few milliseconds (40 to 1)
- Areal: doubling in average every 2/4 years, but not now

- Predicted: 14 Terabytes in 2020 at the cost of $40

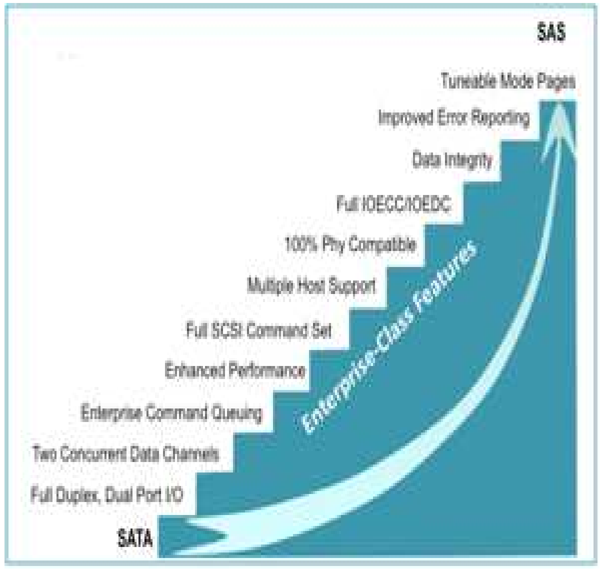
How We Increase 10x and Beyond…

- **Seagate Step up to SAS; SAS Roadmap - Source: SCSI Trade Association**

Capacity

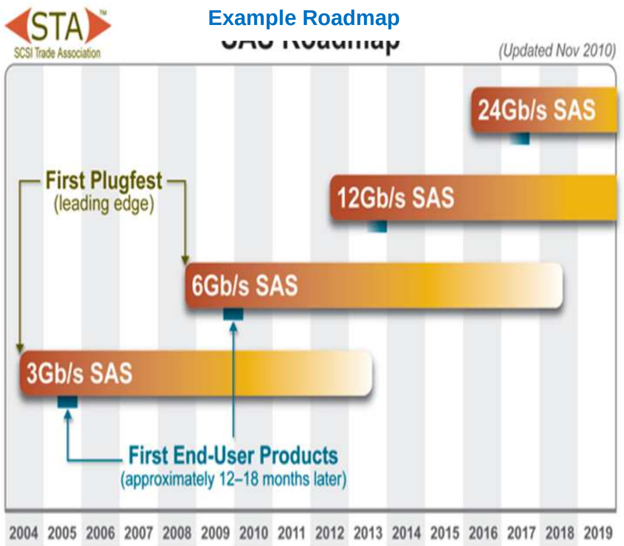


Drive Technologies



Example features for capacity driven usage

Bandwidth and Connectivity



Areal Density

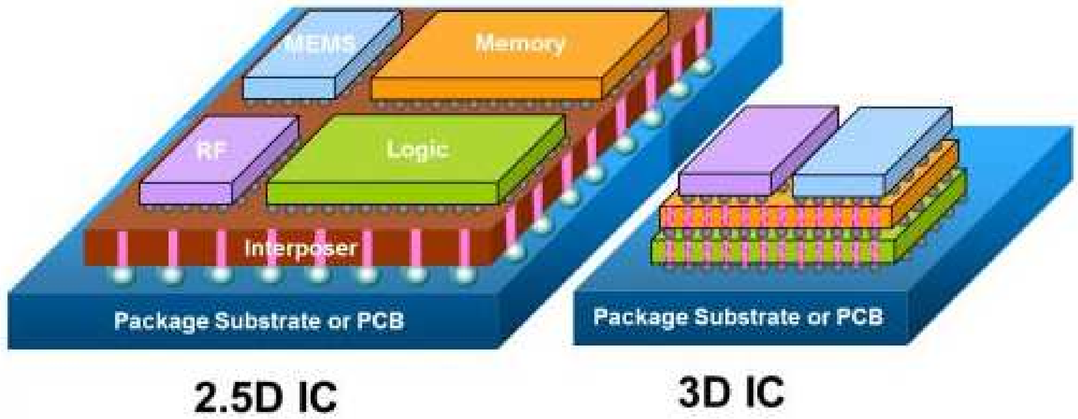
Seagate Storage Effect Web Site

Shingled Magnetic Recording (SMR)	Heat-Assisted Magnetic Recording(HAMR)	Future Candidate(s) Single Molecule Magnets
Projected 25% Areal Density Increase	**Projected 55%** Areal Density Increase	**Envisioned 1,000x** Areal Density Increase
Enables +10TB and above	**Theoretical up to 30TB to 60TB Hard Drive**	**Theoretical up to 3000TB Hard Drive**

Capability computing: Research directions

- Can *"trade"* memory capacity for other metrics of interest – e.g. bandwidth?

- Packaging
  - 2.5D stacking
  - 3D stacking



- Technologies
  - Wide I/O
  - Hybrid memory cube

Emerging (non volatile) Memories

	Traditional Technologies				Emerging Technologies				
			Improved Flash						
	DRAM	SRAM	NOR	NAND	FeRAM	MRAM	PCM	Memristor	NEMS
Cell Elements	1T1C	6T	1T		1T1C	1T1R	1T1R	1M	1T1N
Half pitch (F) (nm)	50	65	90	90	180	130	65	3-10	10
Smallest cell area (F^2)	6	140	10	5	22	45	16	4	36
Read time (ns)	< 1	< 0.3	< 10	< 50	< 45	< 20	< 60	< 50	0
Write/Erase time (ns)	< 0.5	< 0.3	10^5	10^6	10	20	60	< 250	1ns(140ps-5ns)
Retention time (years)	seconds	N/A	> 10	> 10	> 10	> 10	> 10	> 10	> 10
Write op. Voltage (V)	2.5	1	12	15	0.9-3.3	1.5	3	< 3	< 1
Read op. Voltage (V)	1.8	1	2	2	0.9-3.3	1.5	3	< 3	< 1
Write endurance	10^{16}	10^{16}	10^5	10^5	10^{14}	10^{16}	10^9	10^{15}	10^{11}
Write energy (fJ/bit)	5	0.7	10	10	30	1.5×10^5	6×10^3	< 50	< 0.7
Density (Gbit/cm^2)	6.67	0.17	1.23	2.47	0.14	0.13	1.48	250	48
Voltage scaling	Fairly scalable					no	poor	promising	promising
Highly scalable	Major technological barriers				poor		promising	promising	promising

Sources: Chong et al ICCAD09, Eshraghian TVLSI11, ITRS13

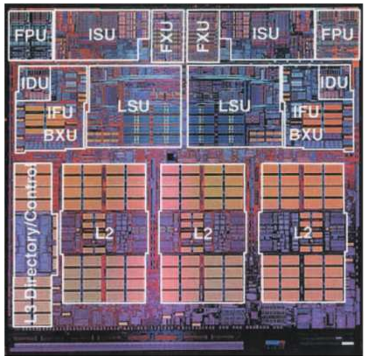
Outline

- Big Data Technologies
  - Storing data
  - Processing data
  - Where do we place data?
  - Managing Big Data

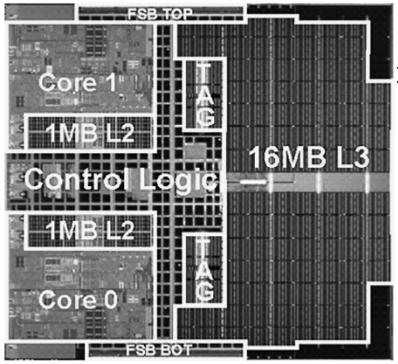
The MultiCore Era

- Moore's Law + Memory Wall + Power Wall

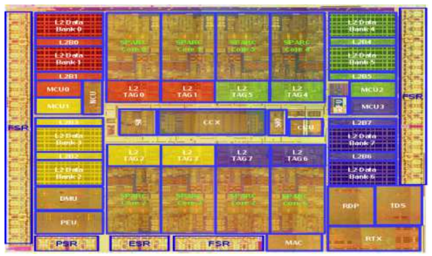
Chip MultiProcessors (CMPs)

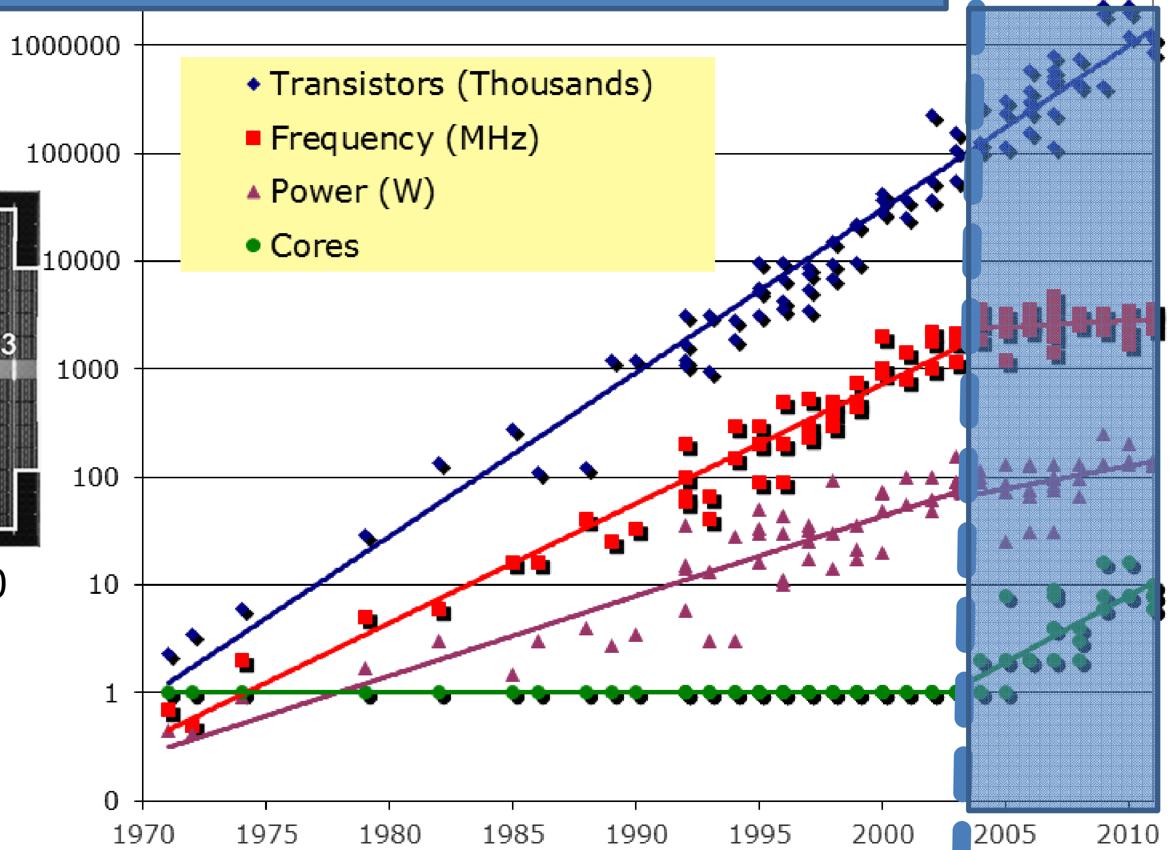


POWER4 (2001)



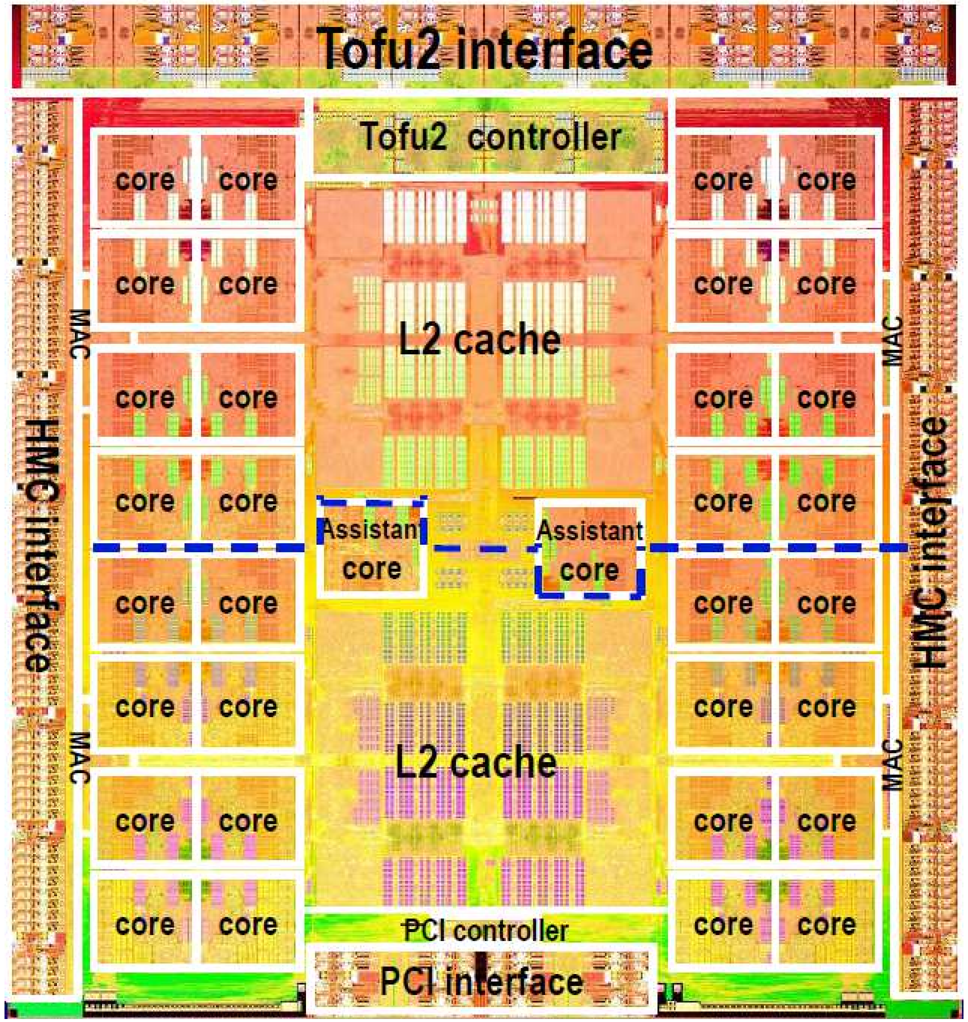
Intel Xeon 7100 (2006)


UltraSPARC T2 (2007)

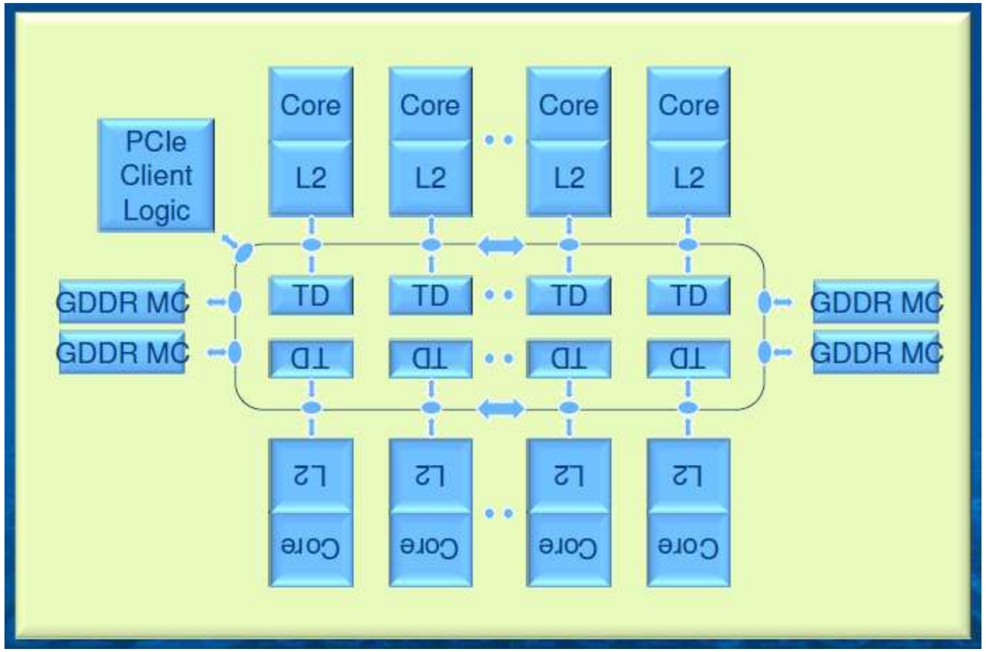


Fujitsu SPARC64 XIfx

- 32 computing cores (single threaded) + 2 assistant cores
- 24MB L2 sector cache
- 256-bit wide SIMD
- 20nm, 3.75M transistors
- 2.2GHz frequency
- 1.1TFlops peak performance
- High BW interconnects
  - HMC (240GB/s x 2 in/out)
  - Tofu2 (125GB/s x 2 in/out)

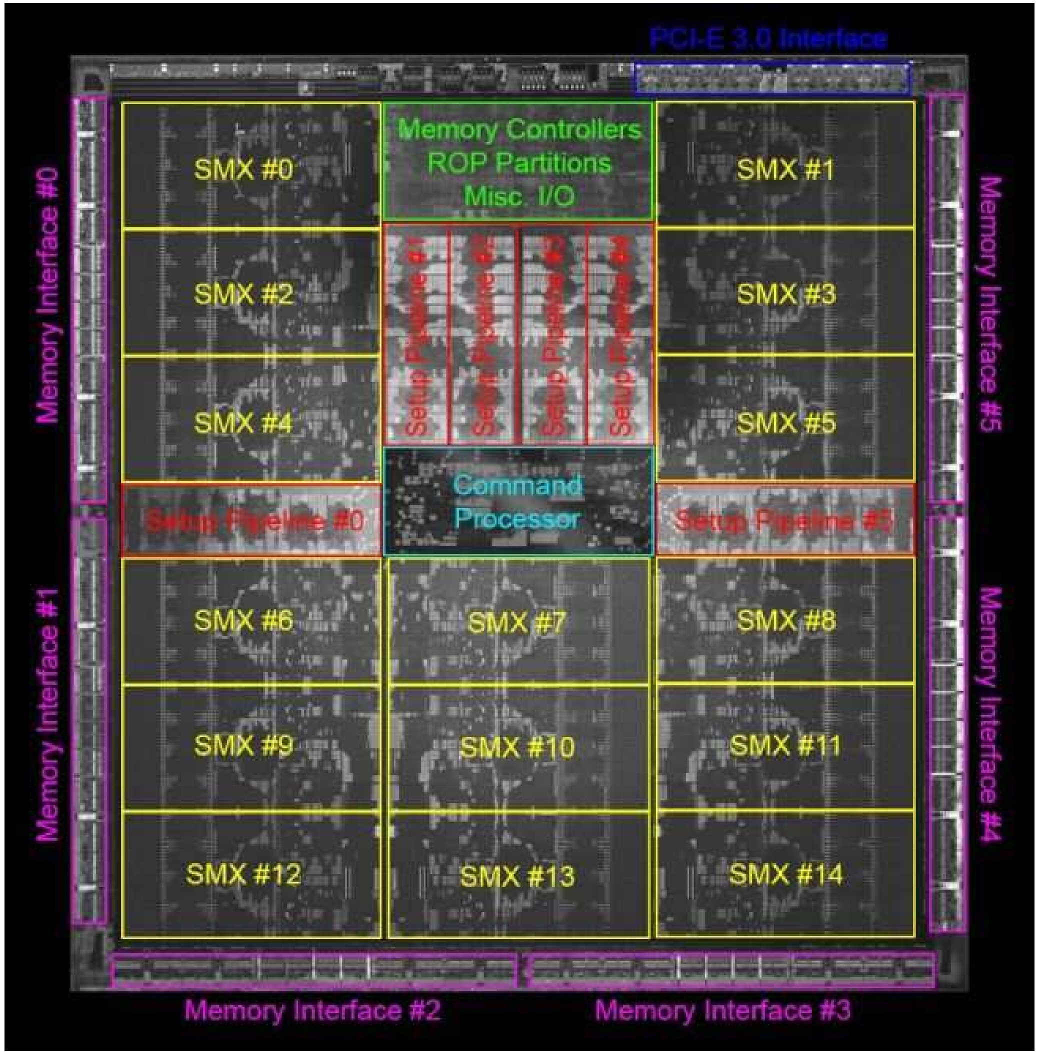


Intel Xeon Phi or
Intel Many Integrated Core Architecture (MIC)

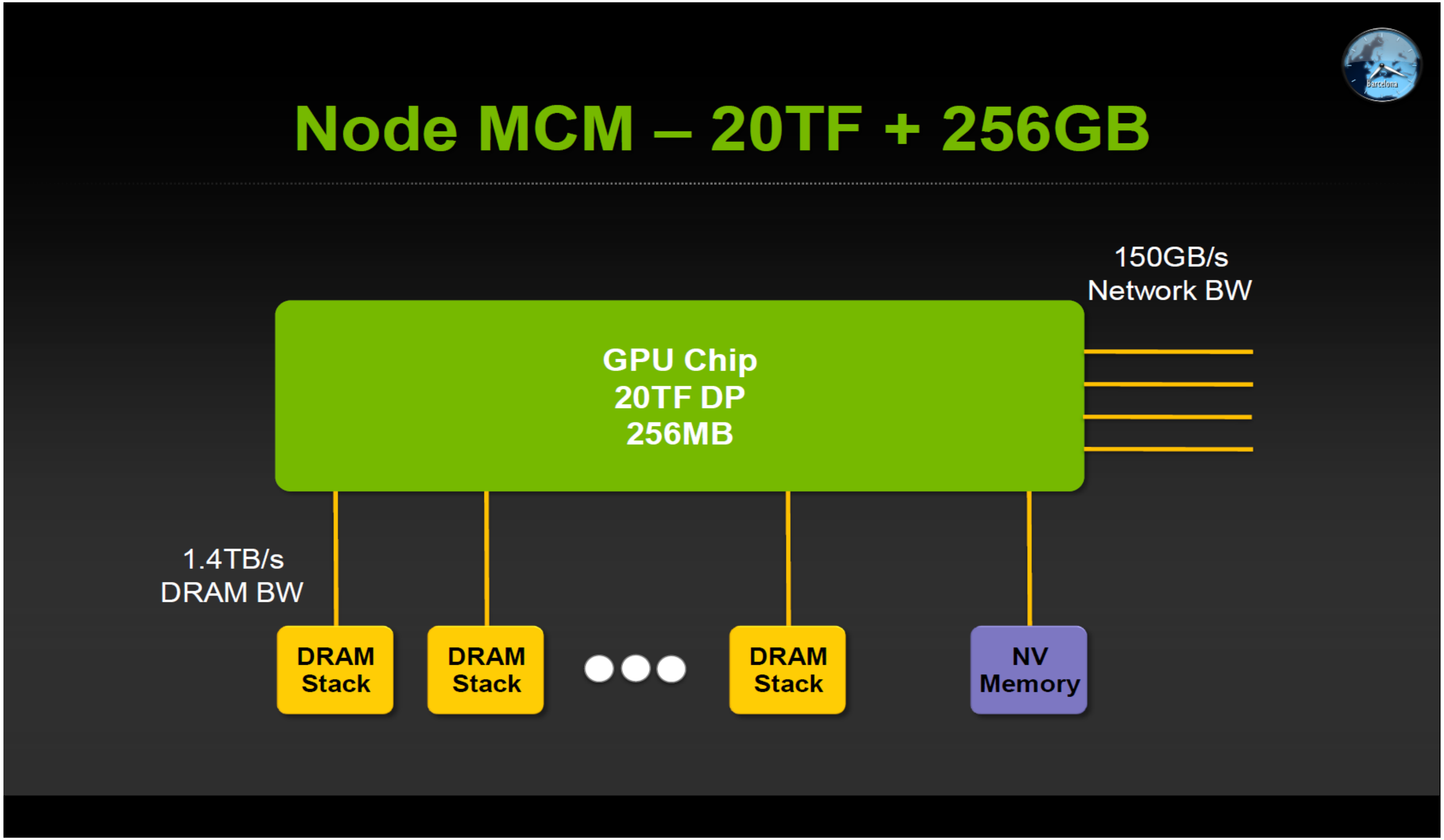


- Knights Corner (2011)
  - Coprocessor, 61x86 cores, 22nm, AVX-512, 4 HTs
  - 1.2TFLOPS (DP), 300W TDP, 4 GFLOPS/W
  - 512KB/core L2 coherent
  - Int Netw: Ring
  - Mem BW: 352GB/s
- Knights Landing (exp 2015)
  - Coprocessor or **host processor**
  - 72 Atom cores, 14nm, AVX512 per core, 4 HTs
  - Up to 16GB of DRAM 3D stacked on-package, 384GB GDDR
  - 3TFLOPS (DP), 200W TDP, 15GFLOPS/W

- DP Performance: 1.43 Tflop
- Mem BW (ECC off): 288 GB/s
- Memory size (GDDR5): 12 GB
- 15 SMX units
  - 192 single-precision CUDA cores
  - 64 double-precision units
  - 32 special function units
  - 32 load/store units
- Six 64-bit memory controllers



Nvidia: Node for the Exaflop Computer



Graph500 vs. Top500

- Top500 defined a benchmark (Linpack) to rate HPC machines upon performance. This benchmark is not suitable to address the characteristics of Big Data applications.

- Linpack:
  - computation bounded
  - focused on floating-point operations
  - Bulk-Synchronous-Parallel model: behavior based on big computation and short communication bursts
  - dense data structures highly organized and coalesced (spatial locality)

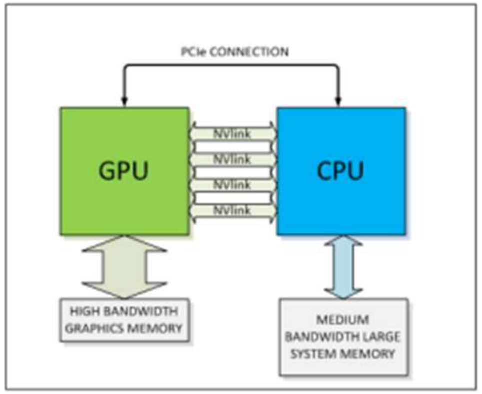
- Graph500:
  - communication bounded
  - focused on integer operations
  - asynchronous spatial uniform communication interleaved with computation
  - larger sparse datasets (very low spatial and temporal locality)

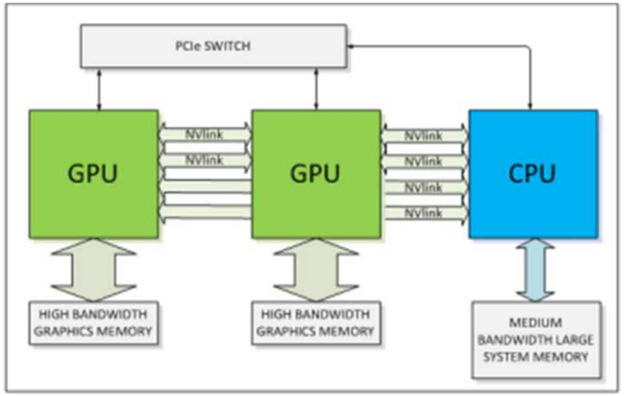
- Green Graph 500 list:
  - Collects performance-per-watt metrics
  - To compare the energy consumption of data intensive computing workloads.

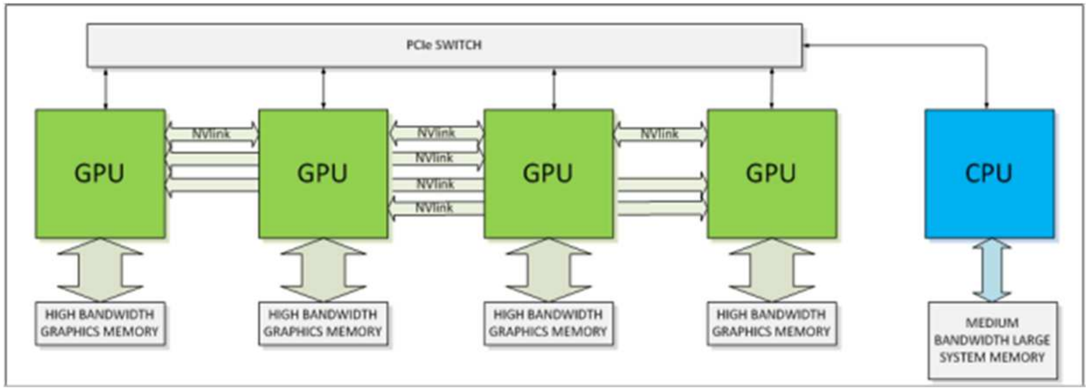
Graph500: graph500.org Top500: top500.org

Accelerators: NVIDIA Pascal (expected 2016)

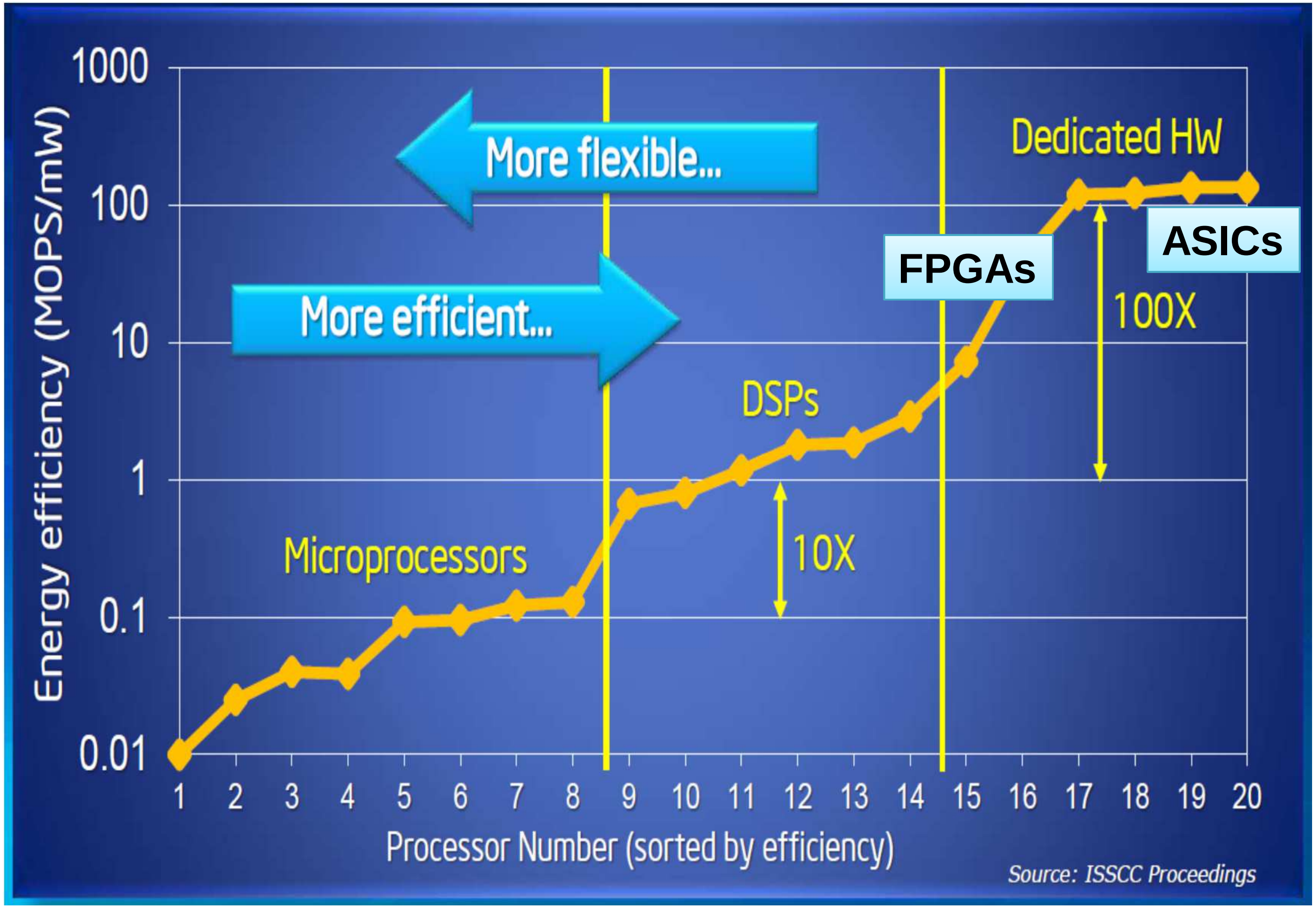
- Aimed to fix data movement bottleneck
- Based on NVlinks
  - chip-to-chip communication approach
  - comprised of bi-directional 8-lane links
  - provide between 80 and 200 GB/s of bandwidth
- This approach is expected to provide 4x speedups w. r. t. current GPU-based designs







Specialization is Everything (?)



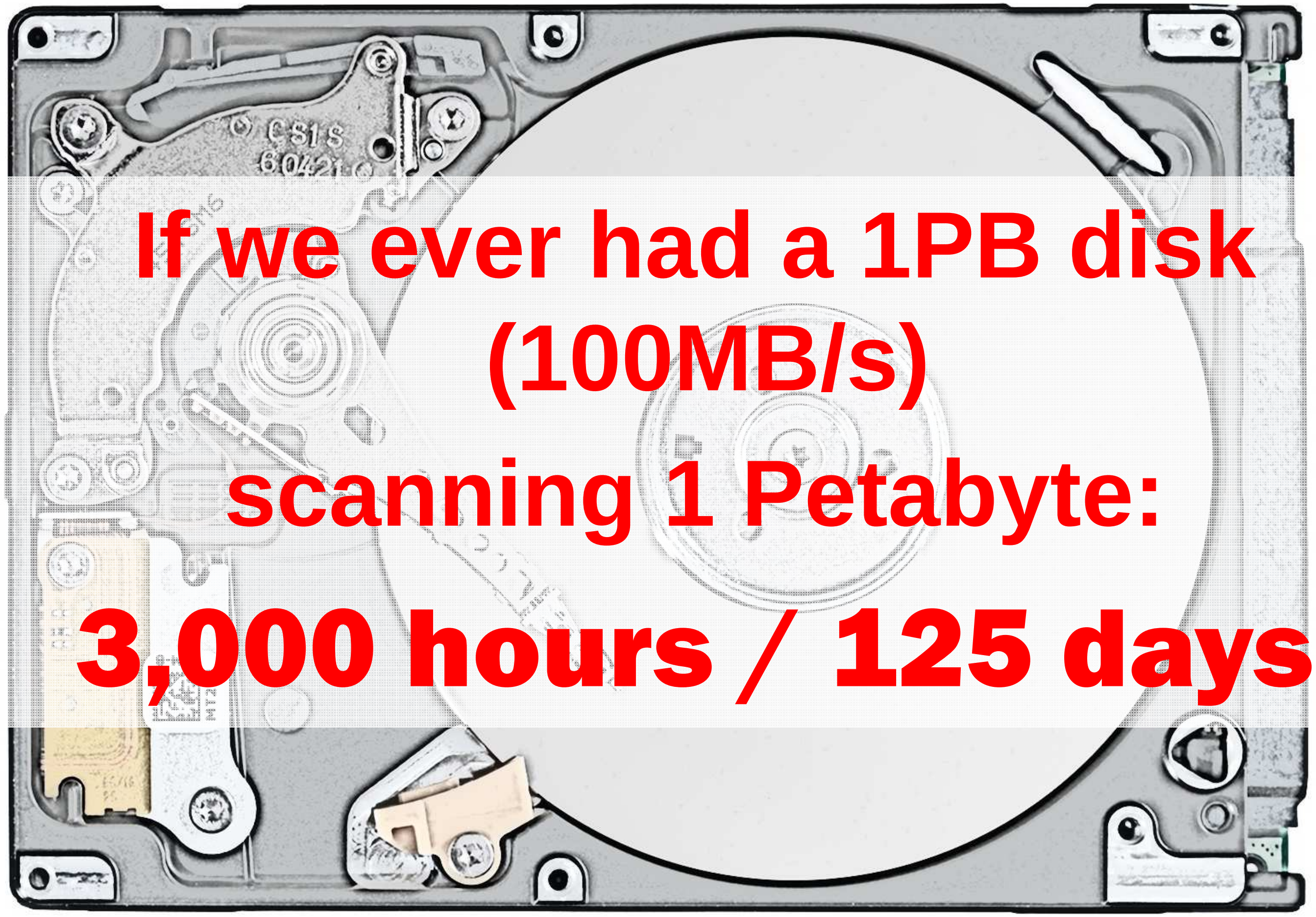
Source: Bob Broderson, Berkeley Wireless group

Outline

- Big Data Technologies
  - Storing data
  - Processing data
  - Where do we place data?
  - Managing Big Data

100.000 subjects suffering different diseases

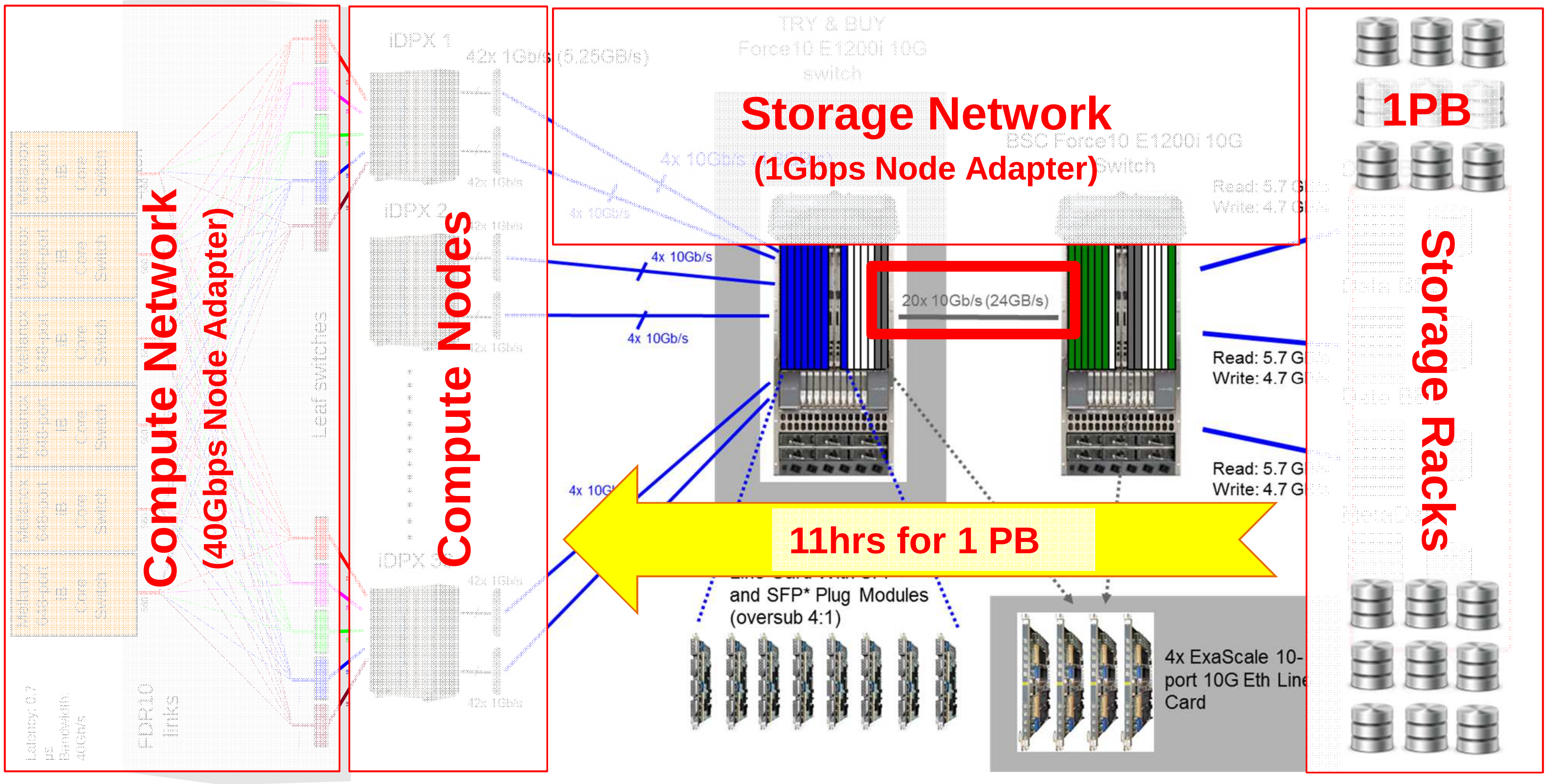
What is the time required to retrieve information?



Supercomputing is about doing things FAST…

What if we want to process the data in 1 hour ?

Evolution of storage architecture for Big Data



Solid State Hybrid (SSHD) Technology

Reduces Costs
Now practical to employ enterprise-class SLC NAND flash

SSHD: Capacity, performance, and value

HDD large capacity + SSD high speed

Adaptive Memory™ technology

- Self-learning software algorithms make HDD SLC NAND flash work together.
- Enables SSD-like performance when accessing most frequently-used files
- Reduces workload and increases reliability of SLC NAND flash

Compute and Communication Energies

- More energy to move data than to compute on it
  - Computation almost feels "free" relative to communication
  - Time will make this worse
- There are two long poles in the communication energy tent
  - Memory
  - Storage
- This is a predicate for this talk

Operation	Energy (pJ)
64-bit integer operation	1
64-bit floating-point operation	20
256 bit on-die SRAM access	50
256 bit bus transfer (short)	26
256 bit bus transfer (1/2 die)	256
Off-die link (efficient)	500
256 bit bus transfer(across die)	1,000
DRAM read/write	16,000
HDD read/write	$\mathcal{O}(10^6)$

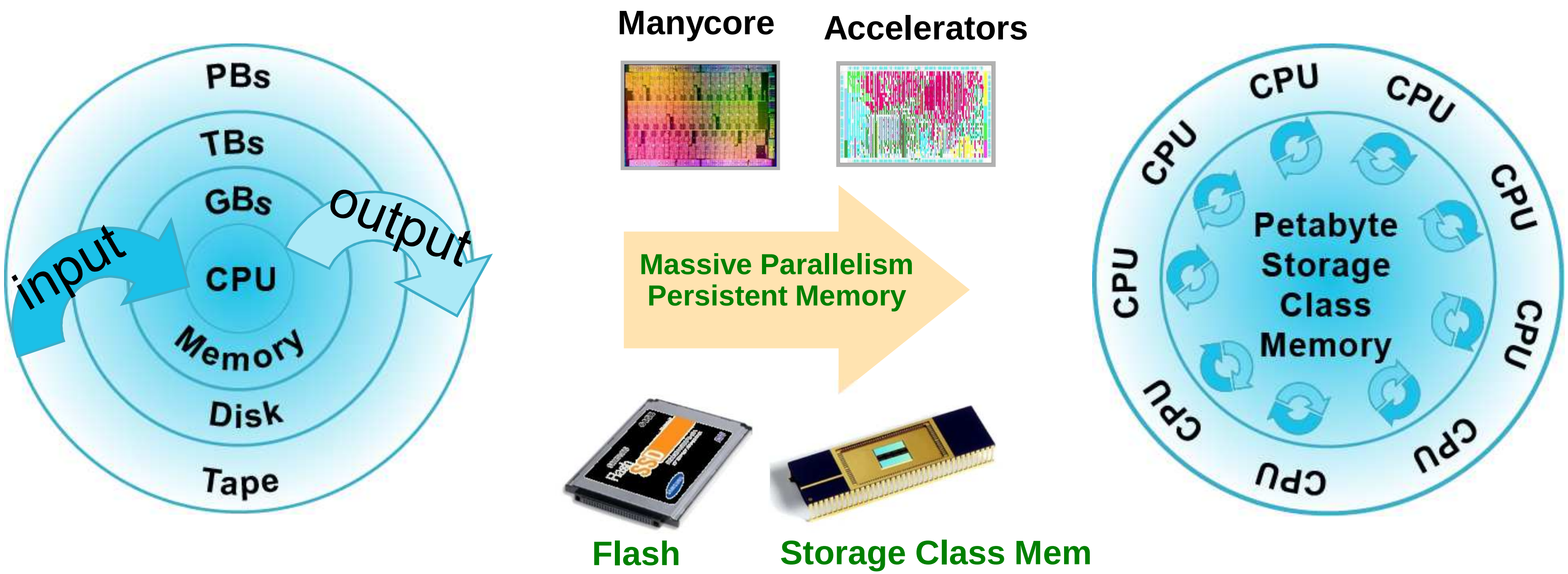
Source: Processors and sockets: What's next? Greg Astfalk / Salishan Conference / April 25, 2013

Old

Compute-centric Model

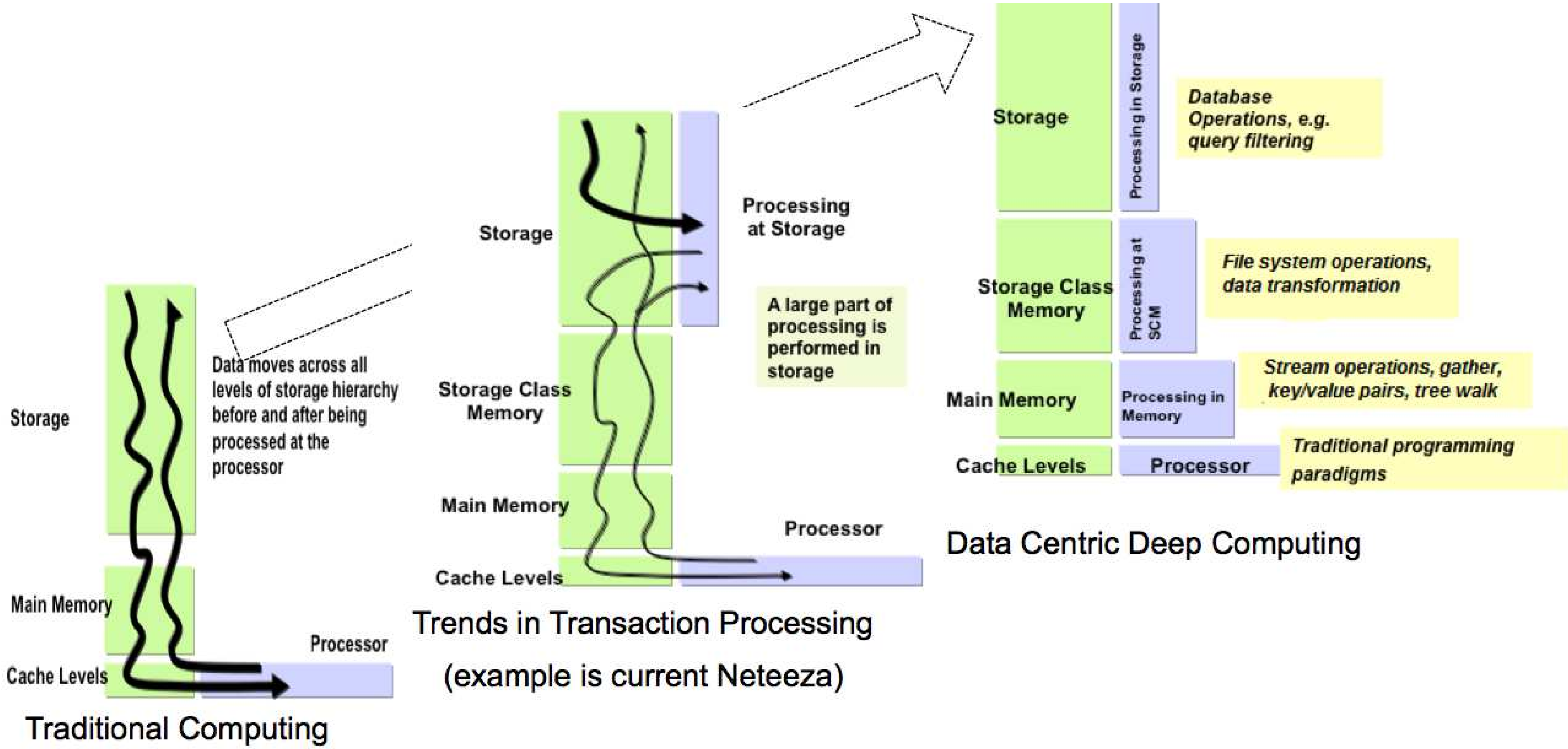
New

Data-centric Model



Source: Heiko Joerg http://www.slideshare.net/schihei/petascale-analytics-the-world-of-big-data-requires-big-analytics

Optimized System Design for Data-Centric Deep Computing



Source: IBM Corporation, 2013

Blue Gene Active Storage (BGAS) Concept

IBM

"How to" guide:

- Remove 512 of 1024 BG/Q compute nodes in rack – to make room for solid state storage
- Integrate 512 Solid State (Flash+) Storage Cards in BG/Q compute node form factor

Flash Storage	2012 Targets
Capacity	2 TB
I/O Bandwidth	2 GB/s
IOPS	200 K

BGAS Rack Targets

Nodes	512
Storage Cap	1 PB
I/O Bandwidth	1 TB/s
Random IOPS	100 Million
Compute Power	104 TF
Network Bisect.	512 GB/s
External 10GbE	512

... scale it like BG/Q.

System Software Environment

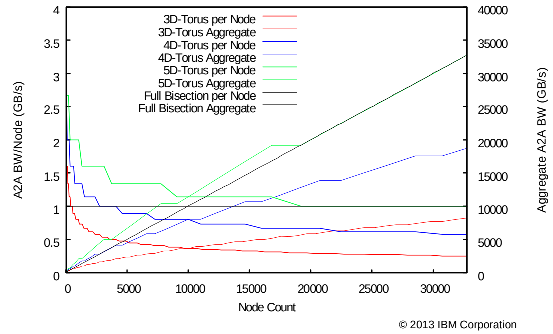
- Linux OS enabling storage + embedded compute
- OFED RDMA & TCP/IP over BG/Q Torus – failure resilient
- Standard middleware – GPFS, DB2, MapReduce, Streams

Active Storage Target Applications

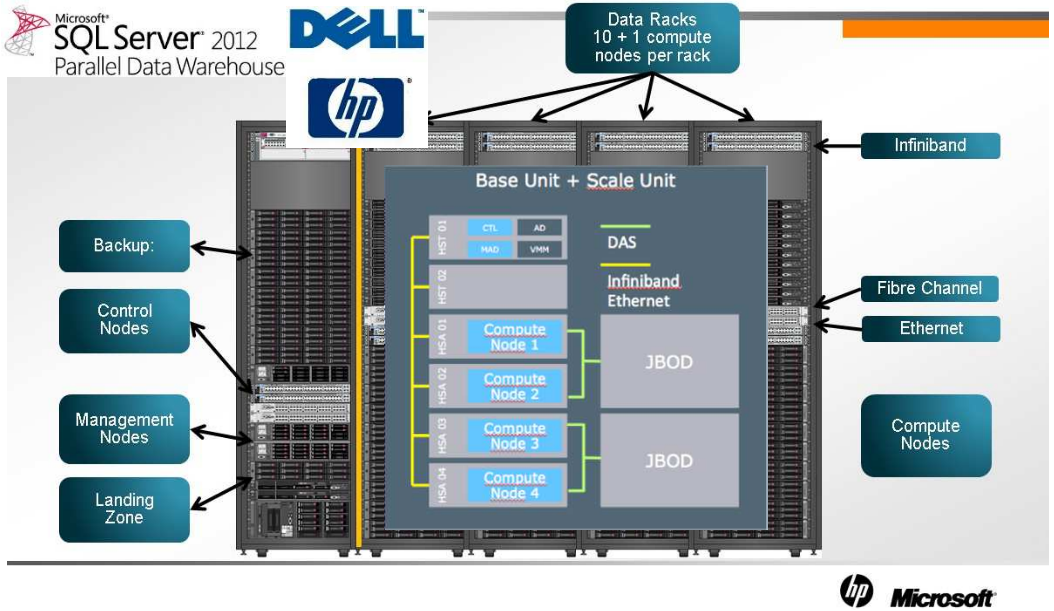
- Parallel File and Object Storage Systems
- Graph, Join, Sort, order-by, group-by, MR, aggregation
- Application specific storage interface

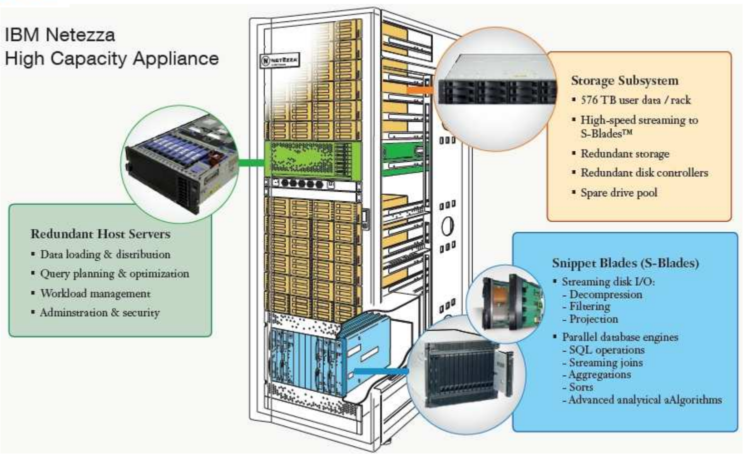
Key architectural balance point: All-to-all throughput roughly equivalent to Flash throughput

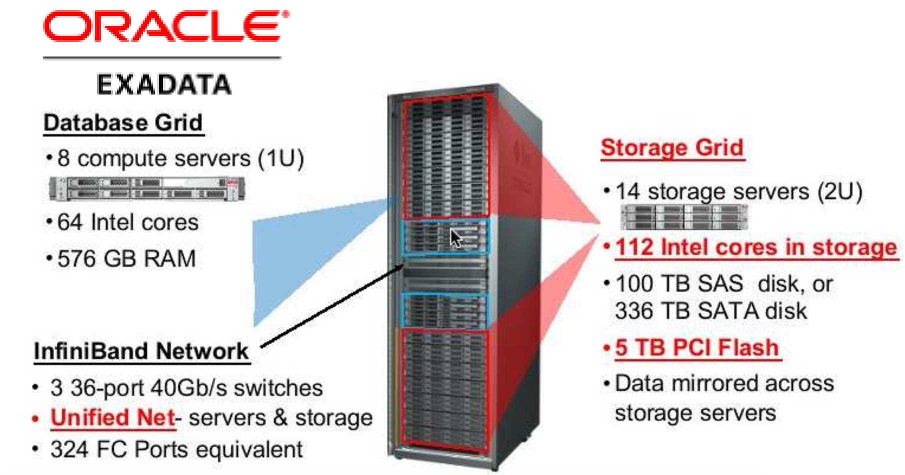
14 Blue Gene Active Storage © 2013 IBM Corporation



HW-SW-Network co-design: Data Appliances







Microsoft Catapult

Doug Burger From Microsoft Research Talks About Project Catapult Which Will Make Bing Twice As Fast
(*Jun 17, 2014)*

http://microsoft-news.com/doug-burger-from-microsoft-research-talks-about-project-catapult-which-will-make-bing-twice-as-fast-video/

Microsoft is planning to replace traditional CPUs in data centers with field-programmable arrays, or FPGAs, processors that Microsoft could modify specifically for use with its own software. These FPGAs are already available in the market and Microsoft is sourcing it from a company called Altera. The FPGAs are 40 times faster than a CPU at processing Bing's custom algorithms.

HP unveils "The Machine"

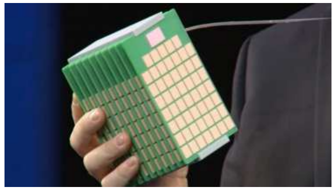
June 11, 2014

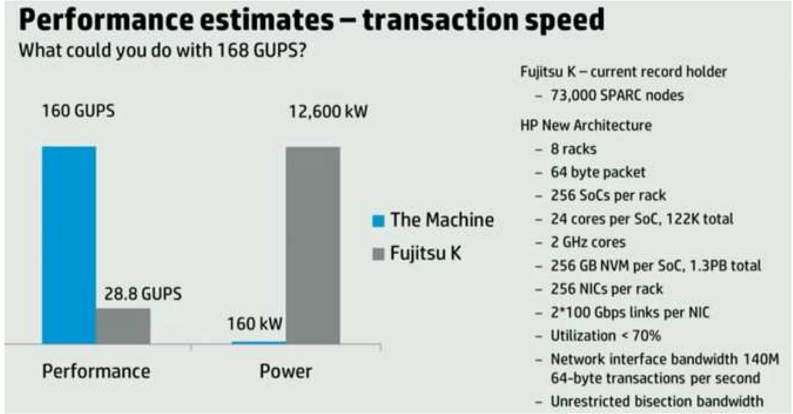
hp next

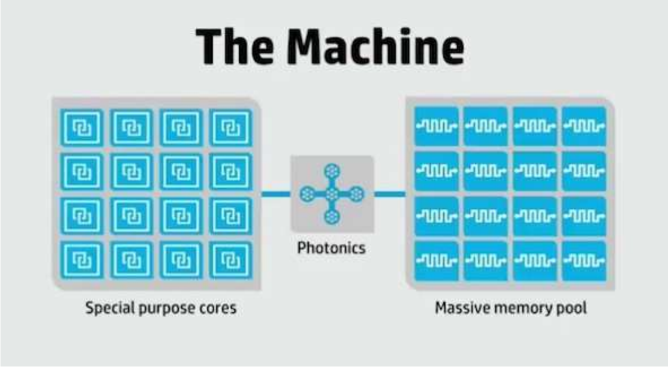
HP unveils "The Machine"

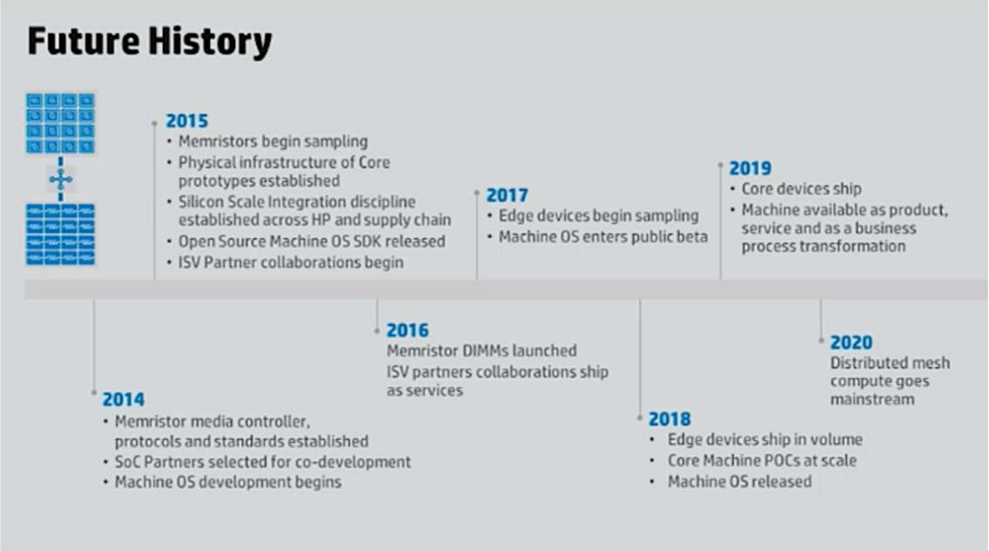
It uses clusters of special-purpose cores, rather than a few generalized cores; photonics link everything instead of slow, energy-hungry copper wires; memristors give it unified memory that's as fast as RAM yet stores data permanently, like a flash drive.

A Machine server could address 160 petabytes of data in 250 nanoseconds; HP says its hardware should be about six times more powerful than an existing server, even as it consumes 80 times less energy. Ditching older technology like copper also encourages non-traditional, three-dimensional computing shapes, since you're not bound by the usual distance limits.







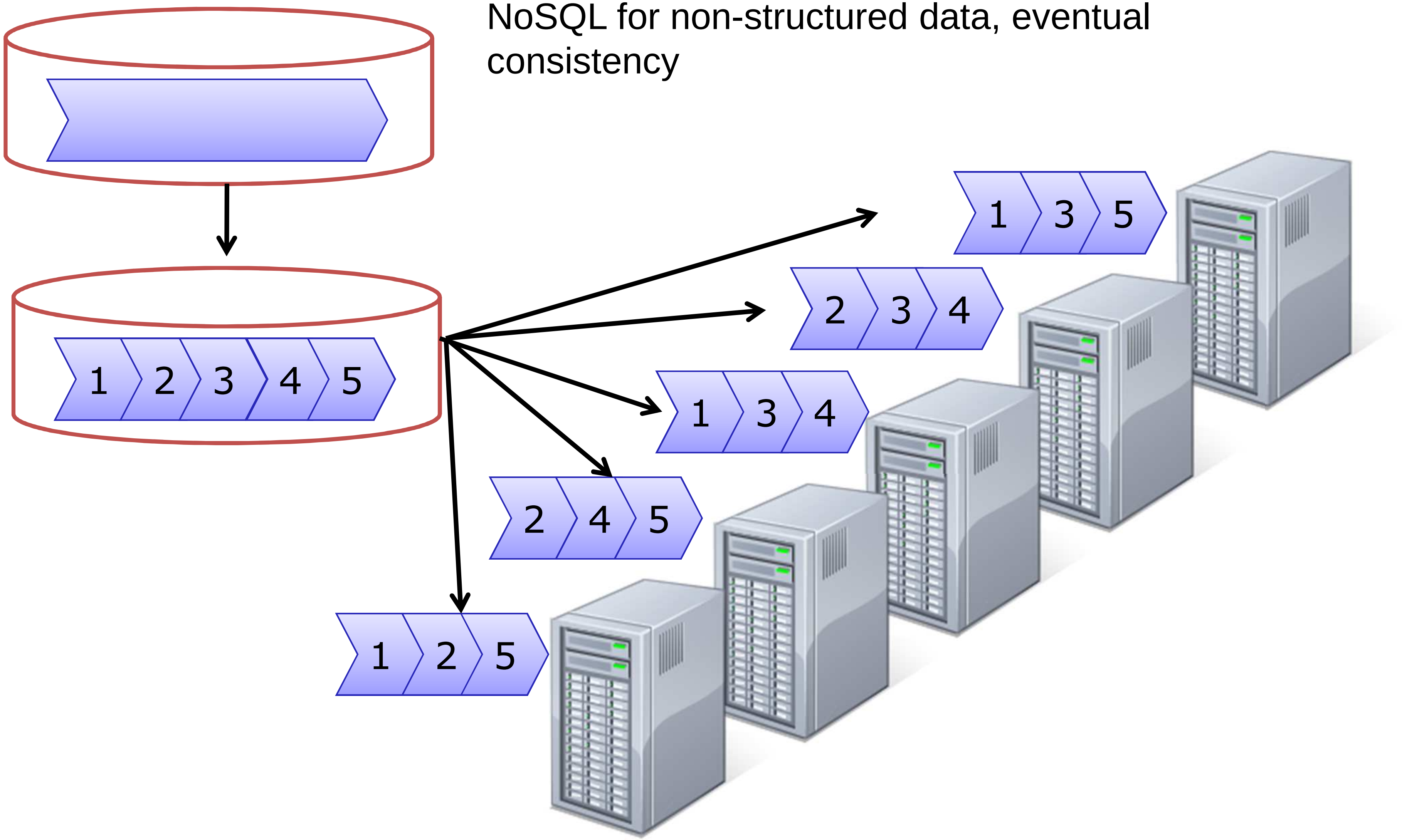


Outline

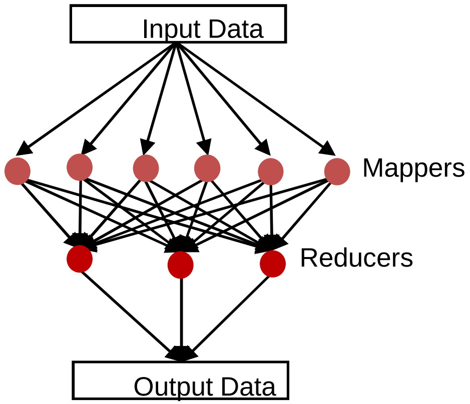
- Big Data Technologies
  - Storing data
  - Processing data
  - Where do we place data?
  - Managing Big Data

Relational databases sometimes not good for scale-out

NoSQL for non-structured data, eventual consistency



- **To meet the challenges: MapReduce**
  - **Programming Model** introduced by Google in early 2000s to support **distributed computing** (special emphasis in fault-tolerance)
    - Ecosystem of big data processing tools
      - open source, distributed, and run on commodity hardware.
- The key innovation of MapReduce is
  - the ability to **take a query** over a data set, **divide it**, and **run it** in parallel over many nodes.



- Two phases
  - **Map phase**
  - **Reduce phase**

Google progressively dropping MapReduce

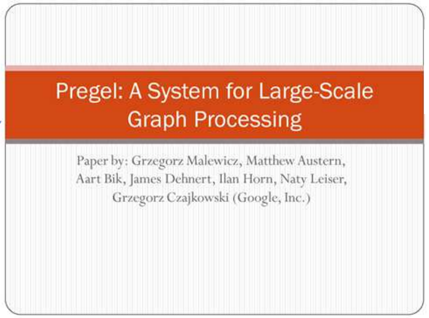
2006
The Apache Hadoop Project brings an opensource MapReduce Implementation, with the team of the Apache Nutch Crawler and the support of Yahoo!

Batch only

Graph Processing

Batch & Stream

2004
introduces
MapReduce



2010
Google announces Pregel used for building incremental reverse indexes instead of MapReduce
Phylosophy: '*think like a vertex*' (event-driven)

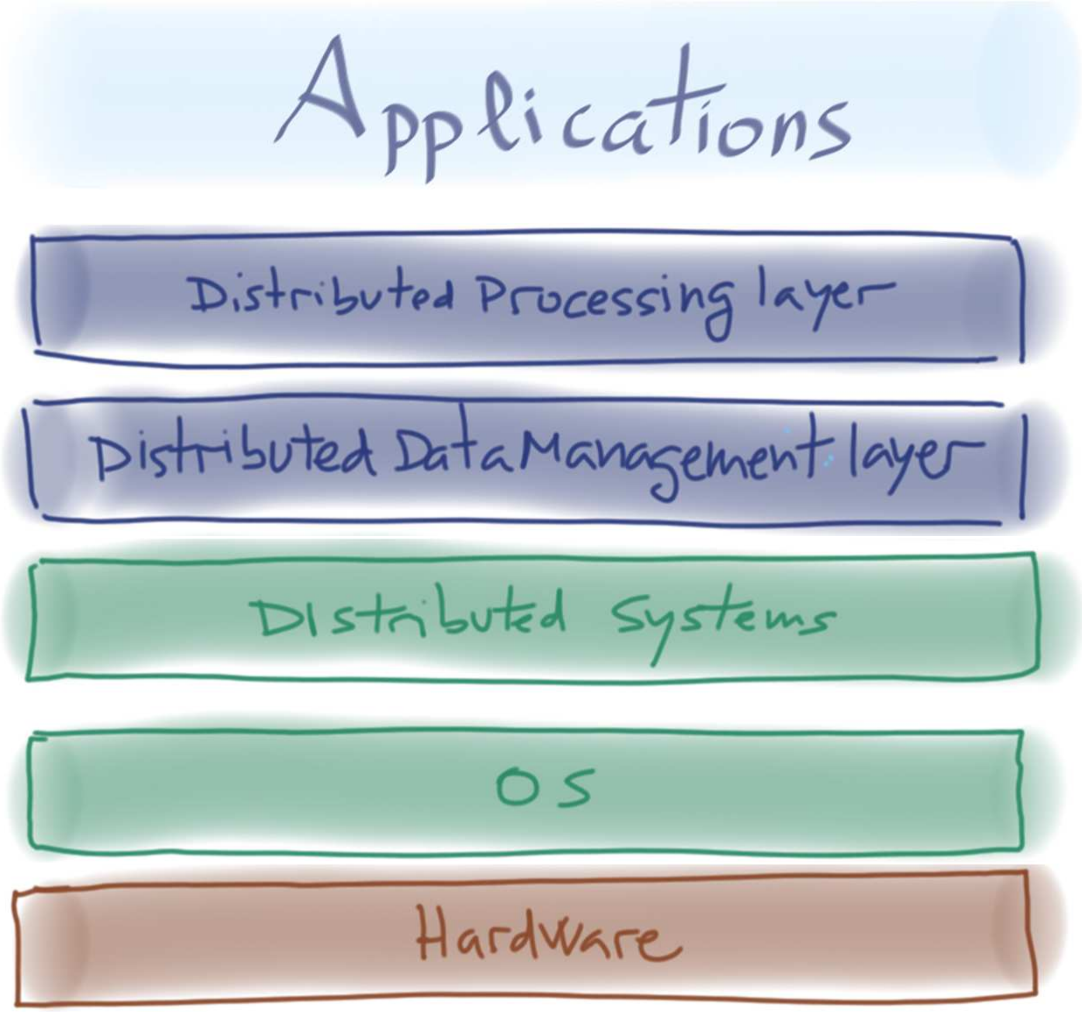
June 25, 2014
Google announces Cloud Dataflow: write code once, run it in batch or stream mode

Cloud Dataflow is a managed service for creating data pipelines that ingest, transform and analyze data in **both batch and streaming modes**.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

RESEARCH IN BIG DATA AT BSC

Abstraction of computer middleware



e.g. Genomics

e.g. MapReduce

e.g. NoSQL DB

e.g. Cloud

e.g. Linux

e.g. FPGAs, GPUs

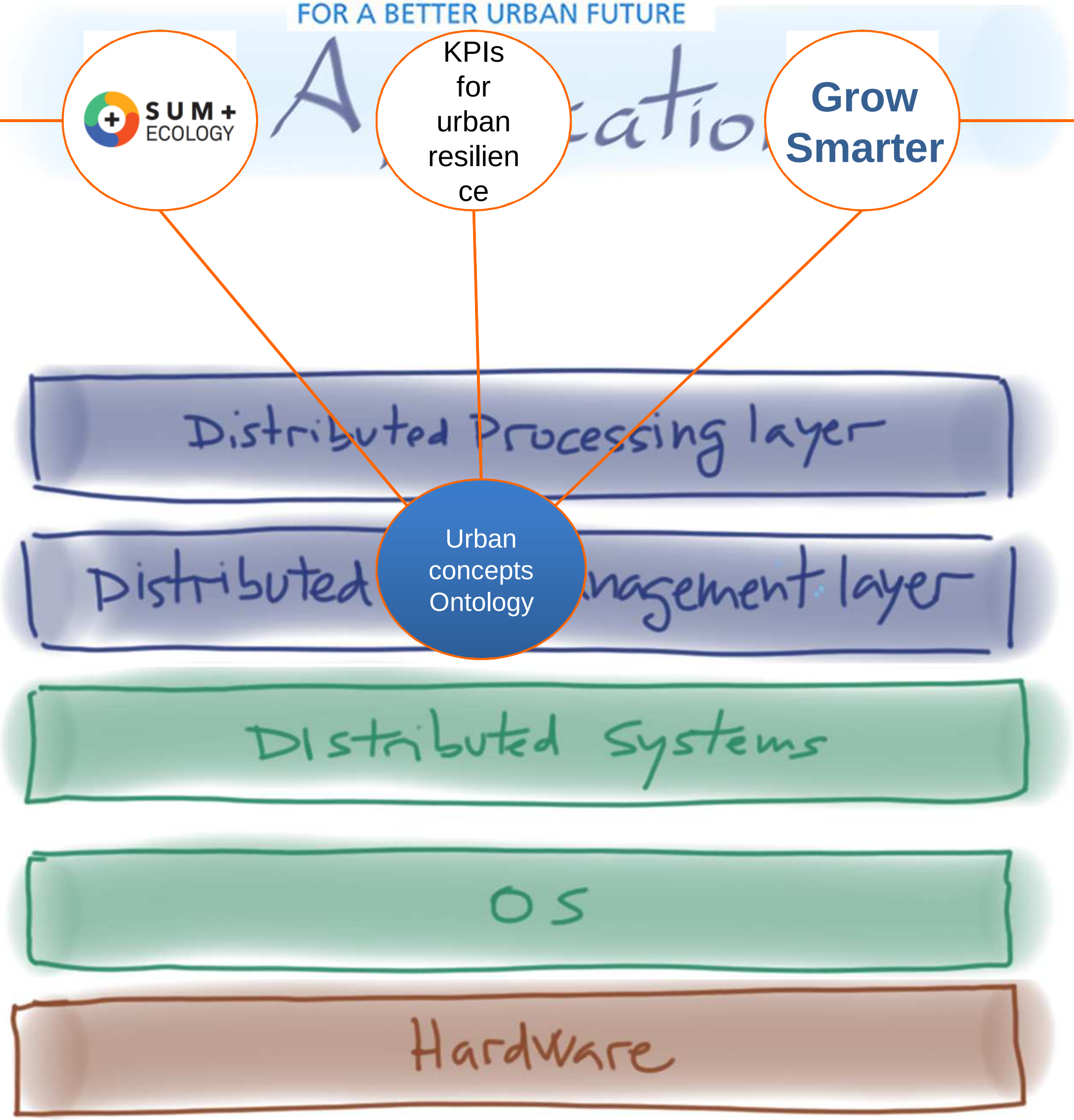
Urban semantics

UN HABITAT
FOR A BETTER URBAN FUTURE

IBM BECOLOGIAN

Models a set of KPI on city mobility, sustainability, energy efficiency,…

Extending an ontology that integrates urban concepts



HORIZON 2020

Stockholm, Köln, Barcelona cities.
Urban Pollution model: to reduce transport emissions by 60%.
Reducing energy consumption in buildings.
Smart lighting and better communication facilities.
Create 1500 jobs.

Ajuntament de Barcelona

Stadt Köln

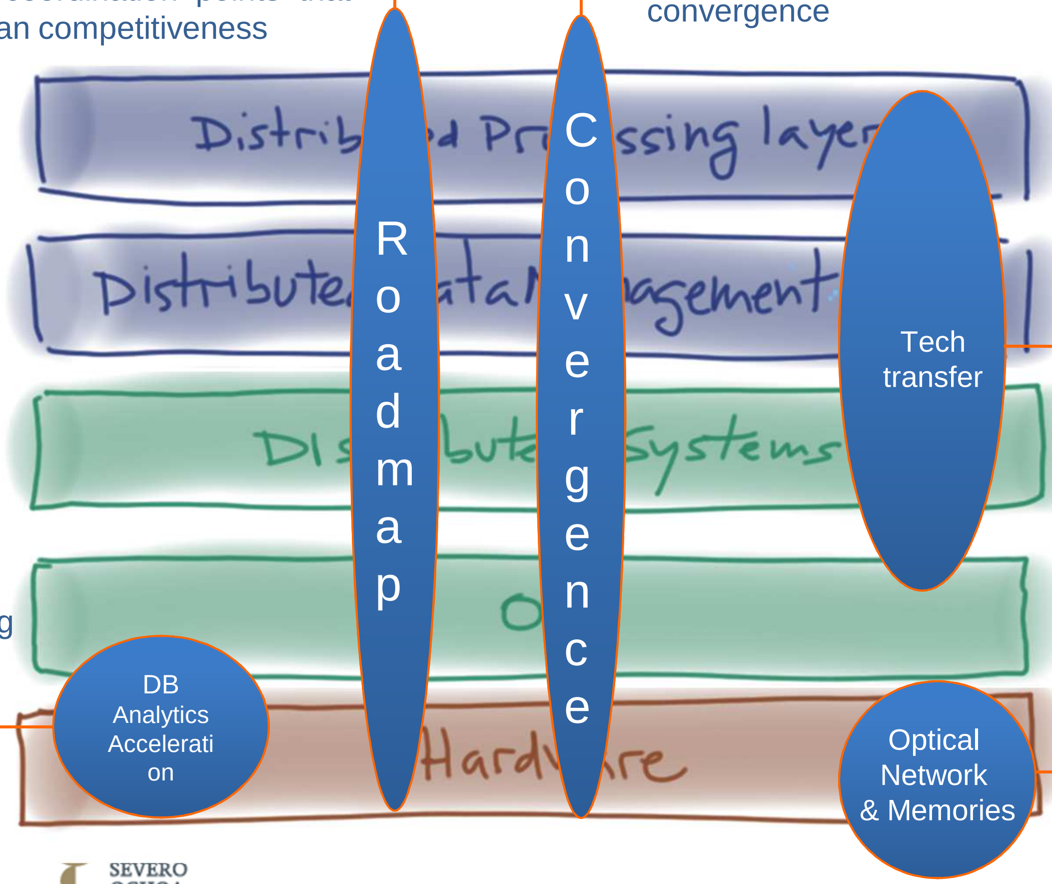
Ongoing Big Data projects - I

RETHINK big

Bring together the key European hardware, networking, and system architects with the key producers and consumers of Big Data to identify the industry coordination points that will maximize European competitiveness

Series of international workshops with participation of worldwide experts to address the challenges of both extreme scale computing and big data focusing on its convergence

BDEC



BIGDATACoE BARCELONA

Technology transfer to SMEs in the Barcelona region

AXLE

Focus on automatic scaling of complex analytics, while addressing the full requirements of real data sets.

New architectures developed using optical interconnections and optical memory

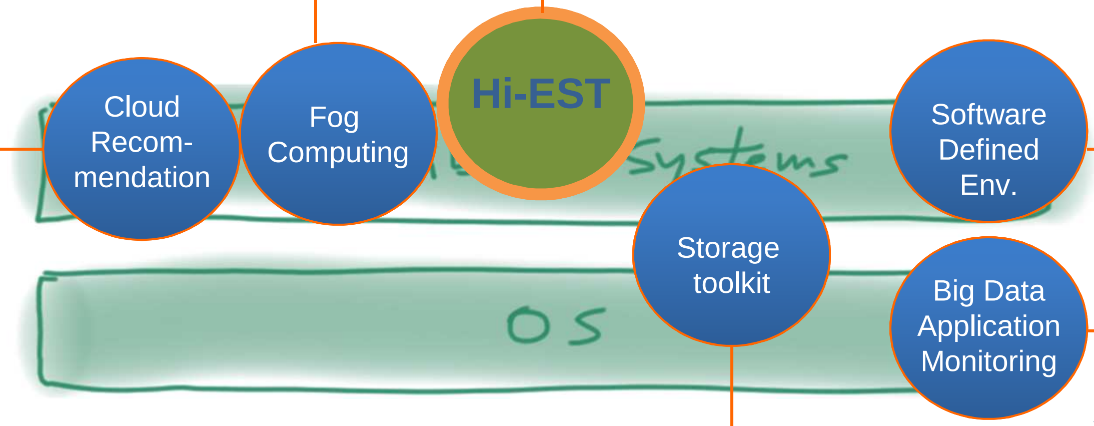
Ongoing Big Data projects – II

Fog distributed platform capable of processing the data close to the end-points, becoming an excellent platform for Internet of Things (IoT)

Holistic Integration of Emerging Supercomputing Technologies
Automatic Optimization of Programmable Data Centers

Enhance a decision support system for the selection of the right set of Clouds
Trade-off between cost, reliability, risks and quality impacts

Building cloud environments embracing hardware and network heterogeneity to host a variety of workloads (JSA-SDN)



Design and implement a set of tools to monitor and to trace network traffic for hadoop applications

Creation of a Software-defined Storage toolkit for Big Data on top of the OpenStack platform

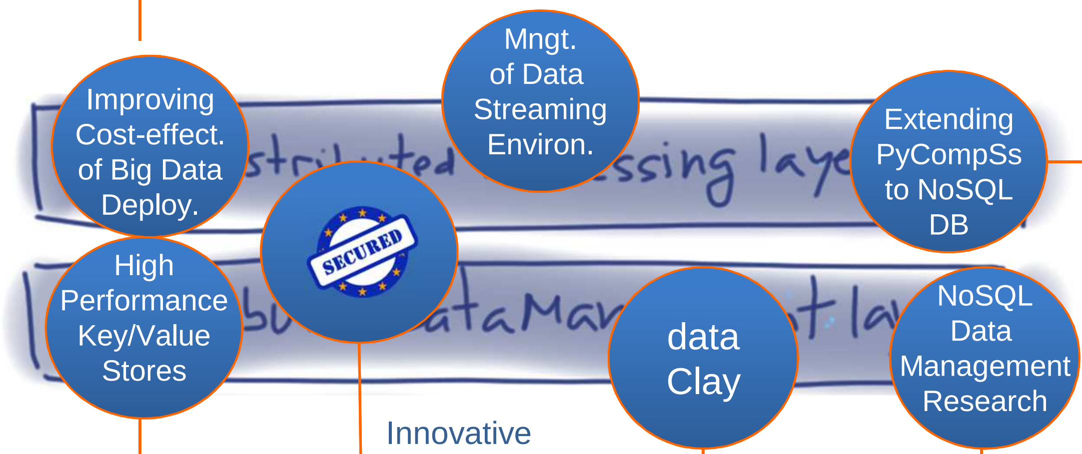
Ongoing Big Data projects - III

Microsoft

Automated characterization of cost-effectiveness of Hadoop deployments (runtime performance vs software and hardware configuration choices)

compose

Explore novel architectures of the emerging **IoT stream processing platforms**, that provide the capabilities of data stream composition, transformation and filtering in real time



Integration of COMPSs runtime with persistent storage and exploration of scheduling policies driven by data locality

EXCELENCIA SEVERO OCHOA

HP

IBM

Explore the use of high performance key/value databases for fast persistent memory technologies

Innovative architecture to achieve protection from Internet threats by offloading execution of security applications into programmable devices at the edge of the network

Self-contained objects library

EXCELENCIA SEVERO OCHOA

Design and implement a software layer to enable NoSQL databases to decouple data organization from data model and provide NoSQL databases with efficient multi-level indexing support

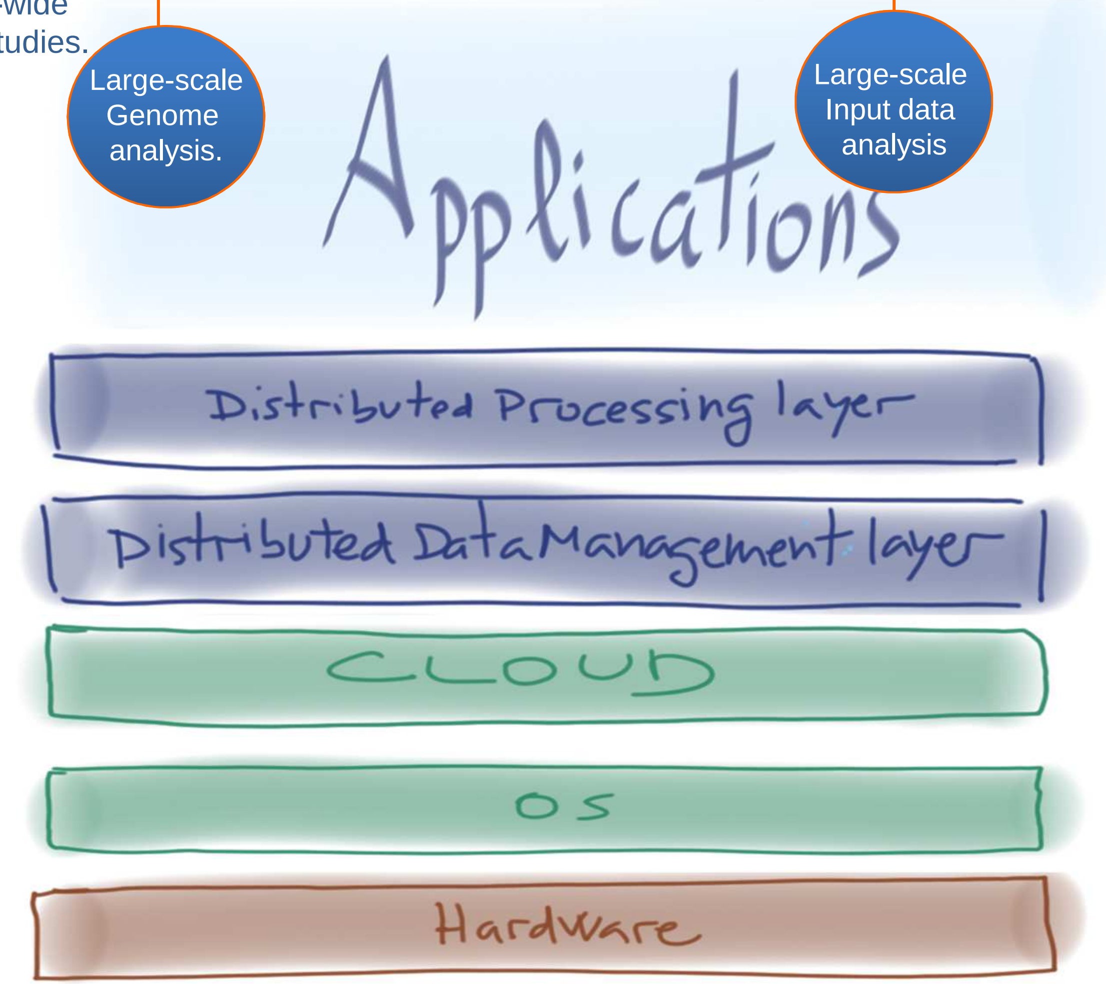
Ongoing Big Data projects - V

GWImp COMPSs
GWAS & IMPUTATION FRAMEWORK

Large-scale Imputation and Genome-wide Association studies.

Smufin

Somatic Mutation Finder on cancer genomes.



Use of statistical imputation theory to infer unknown genetic variants.
Large input data analysis and efficient parallel computation.
Big-data analytics

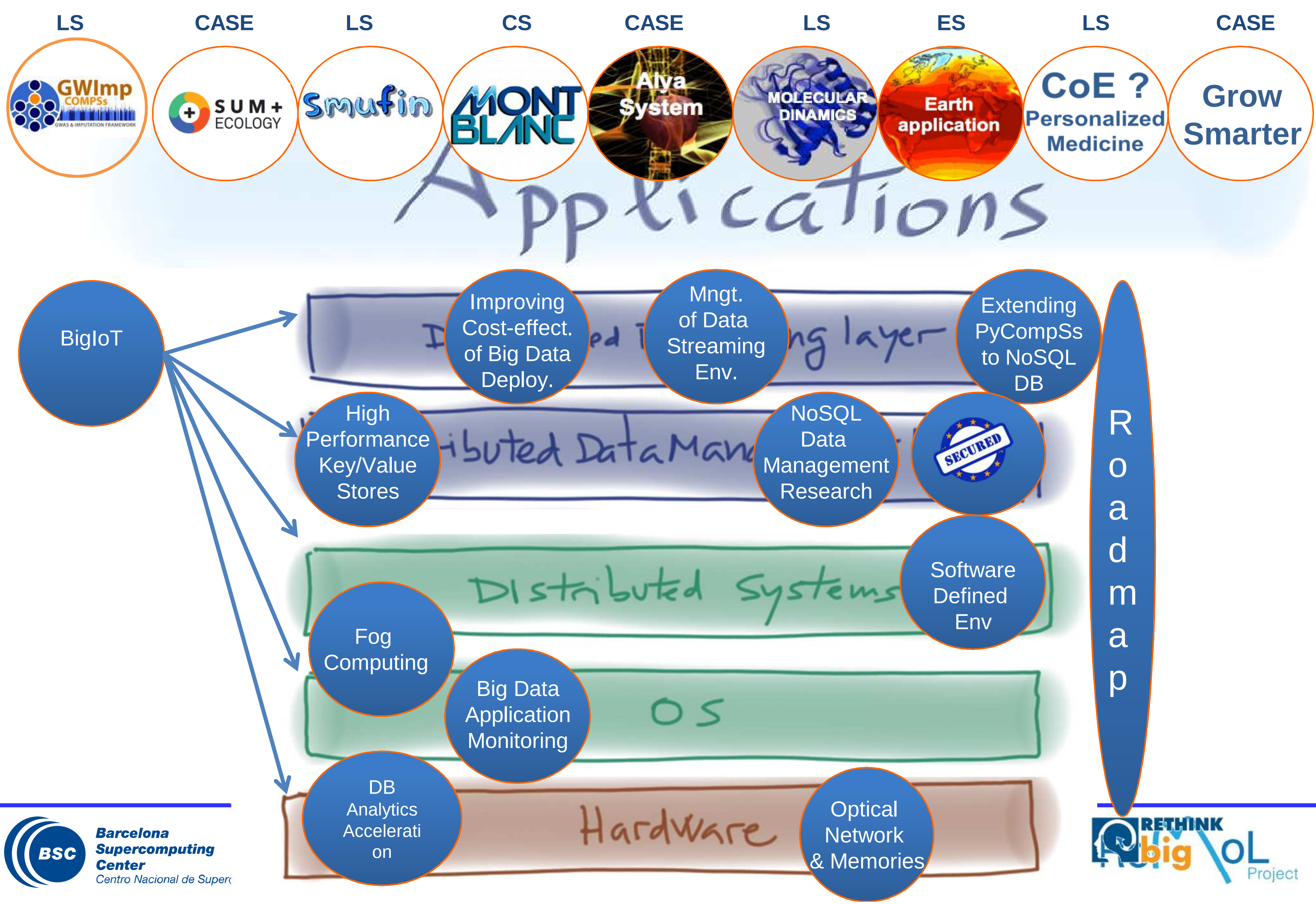
Efficient Big-data genomic workflows deployment in current and future: HPC Clusters, Clouds, etc.

Optimized DNA compression algorithms, such as BWA, Suffix Arrays, FM-Indexing, etc., to compare multi-patients DNA information .
Big-data analytics

New architectures development for Computational Genomics tasks

RoMoL Project

Ongoing Big Data projects :Global picture



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BIG DATA ANALYTICS

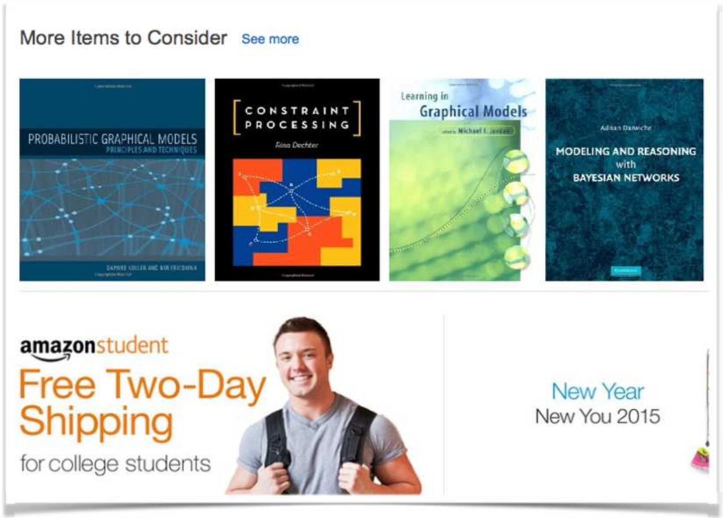
Data Analytics is already influencing our everyday lives

- the **technology is often so subtle** that consumers have no idea that big data is actually helping make their lives easier

Source: www.DesignedInBarcelona.com

Example: Online Shopping

- **Amazon's recommendation engine** uses big data and its database of around 250 million customers to suggest products by looking at previous purchases and other variables.



- Amazon are also developing a new technology which predicts what items you might want and sends it to your nearest delivery hub, (meaning faster deliveries for you)

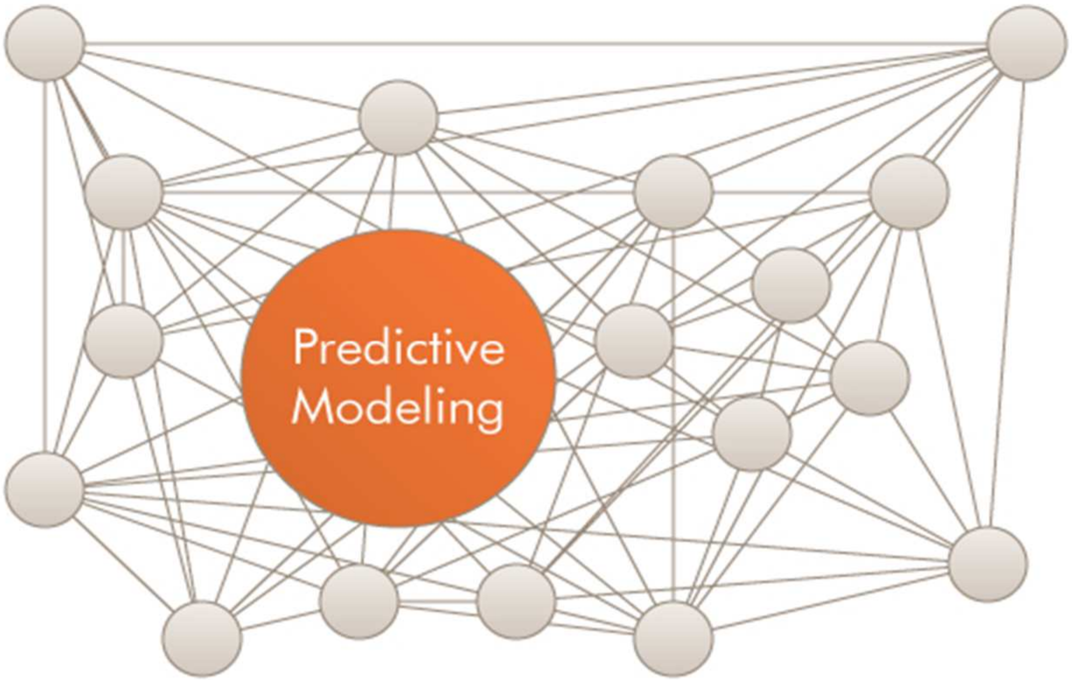
Example: **Target Corporation**, the second-largest retailer in the United States

- In 2012 identified a pregnant teenager before her family knew about her condition.
  - The company can analyze customers' purchasing habits by monitoring credit card data, coupon usage, customer help lines, and emails for specific activities associated with pregnancy.
  - They've identified 25 items that, when purchased in a particular order.

How they do it?

- **To do so, they are using predictive models**

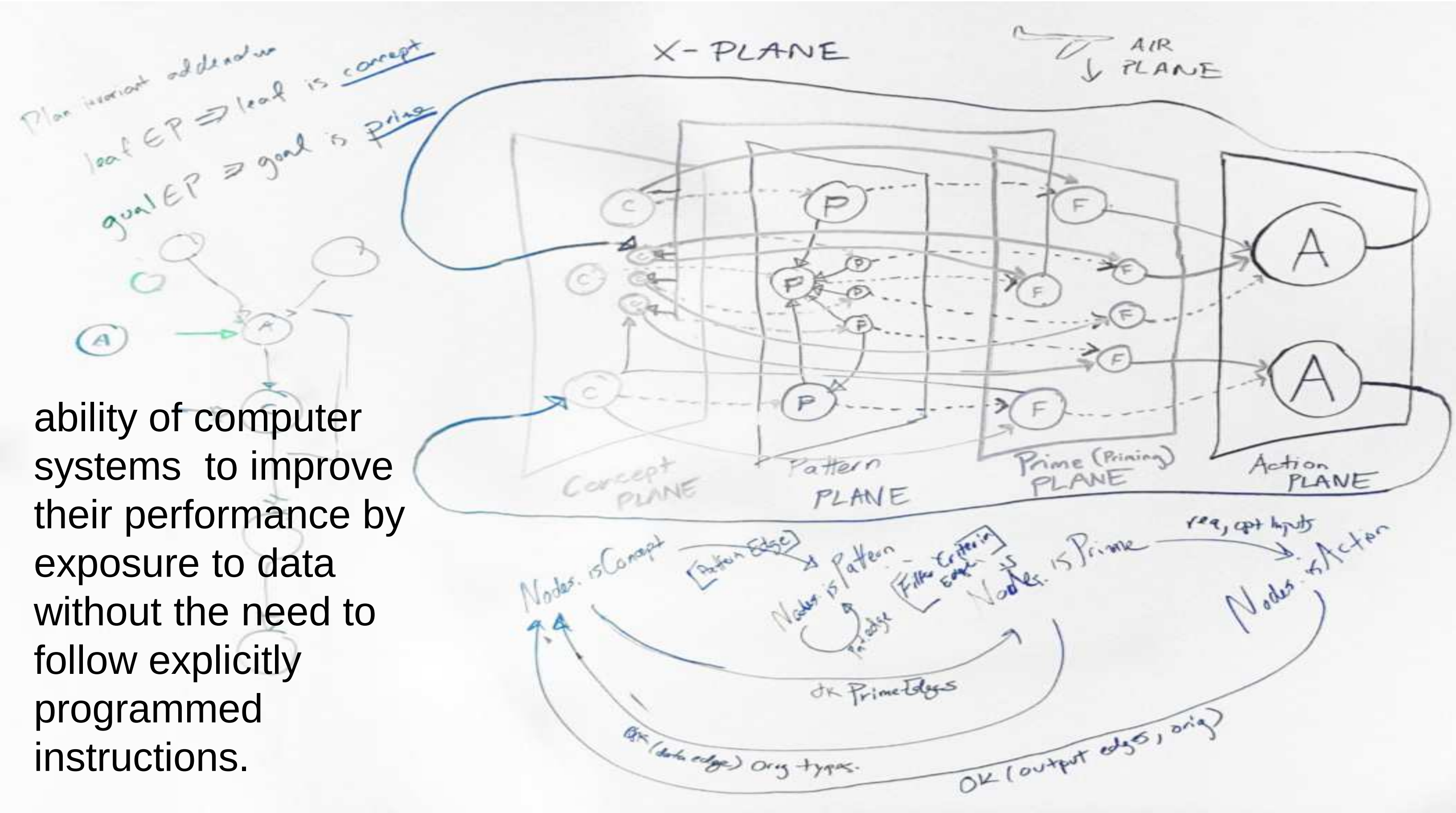
  a collection of mathematical and programming techniques used to determine the probability of future events, analyzing historic and current data to create a model to predict future outcomes.



Today, predictive models form the basis of many of the things that we do online: search engines, computer translation, voice recognition systems, etc.

Source: http://bigsonata.com/wp-content/uploads/2014/07/PredictiveModeling.jpg

This can be achieved with Machine Learning Algorithms

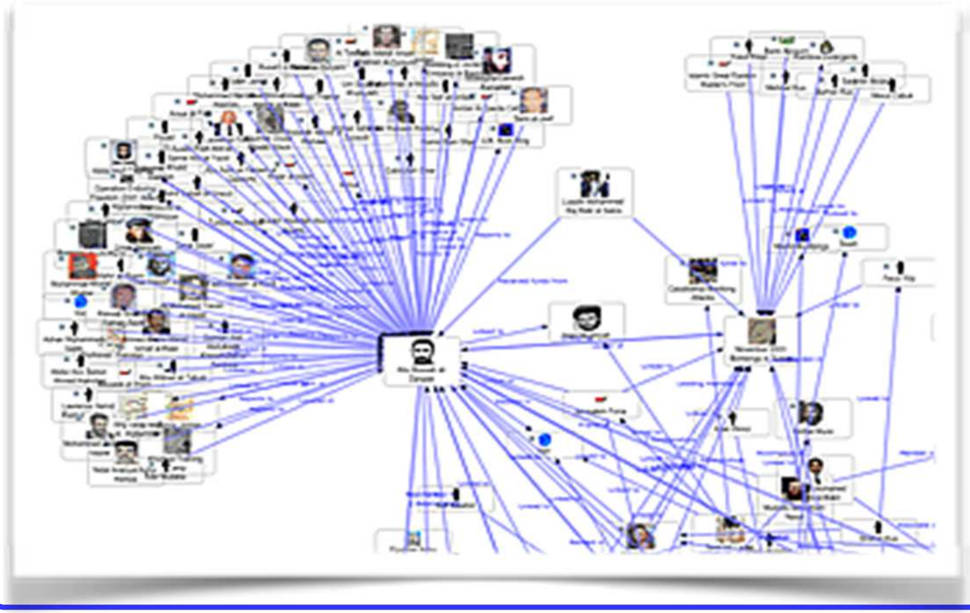


ability of computer systems to improve their performance by exposure to data without the need to follow explicitly programmed instructions.

- **We are now at a turning point of the history of computing**

The **first wave** of computing made **numbers** computable

The **second wave** has **made text and rich media** computable and accessible digitally



- We are in the **next,** **third wave** that will also make **context computable**

Systems that embed predictive capabilities, providing the right functionality and content at the right time, for the right application, by continuously learning about them and predicting what they will need.

For example identify and extract context features such as hour, location, task, history or profile to present an information set that is appropriate for a person at a specific time and place.

New self-learning systems are required

- Today computers require programming, and by definition programming does not allow for alternate scenarios that have not been programmed

- To allow alternating outcomes would require going up a level, creating a self-learning systems

The general idea is that instead of instructing a computer what to do, we are going to simply throw data at the problem and **tell the computer to figure it out itself**.

Cognitive?

- For this purpose the computer software takes functions from the brain like: inference, prediction, correlation, abstraction, … giving to the systems to possibility to do this by themselves.
  - Giving computers a greater ability to understand information, and to learn, to reason, and act upon it

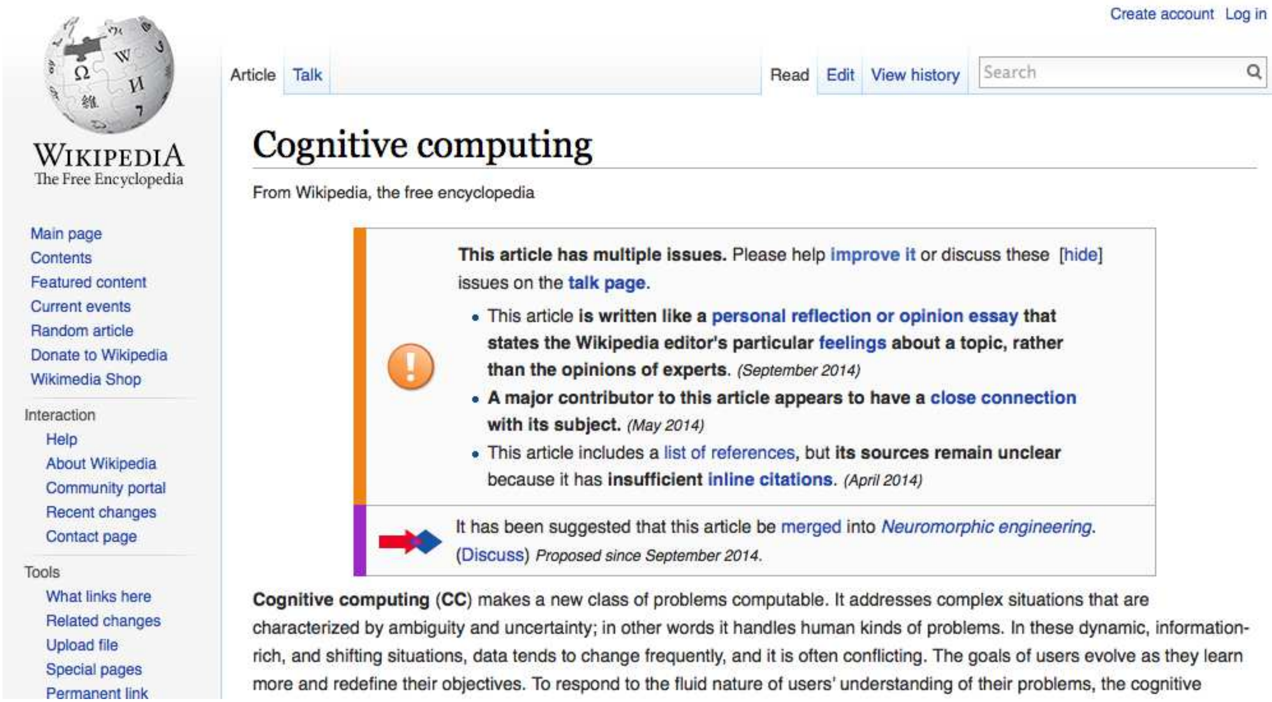
- And here it comes the use of cognitive word to describe this new computing!

Augment our reasoning capabilities

These new systems will raise the potential to augment our reasoning capabilities and empower us to make better informed decisions in order to address complex situations that are characterized by ambiguity and uncertainty.

Cognitive Computing

- Its meaning is not clear yet …
  - The term “cognitive computing” remains a bit confusing since it covers systems that use different analytic approaches.


Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link

Cognitive computing

From Wikipedia, the free encyclopedia

This article has multiple issues. Please help improve it or discuss these issues on the talk page. [hide]

- This article **is written like a personal reflection or opinion essay that states the Wikipedia editor's particular feelings about a topic, rather than the opinions of experts.** *(September 2014)*
- **A major contributor to this article appears to have a close connection with its subject.** *(May 2014)*
- This article includes a list of references, but **its sources remain unclear** because it has **insufficient inline citations.** *(April 2014)*

It has been suggested that this article be merged into *Neuromorphic engineering*. (Discuss) *Proposed since September 2014.*

Cognitive computing (CC) makes a new class of problems computable. It addresses complex situations that are characterized by ambiguity and uncertainty; in other words it handles human kinds of problems. In these dynamic, information-rich, and shifting situations, data tends to change frequently, and it is often conflicting. The goals of users evolve as they learn more and redefine their objectives. To respond to the fluid nature of users' understanding of their problems, the cognitive

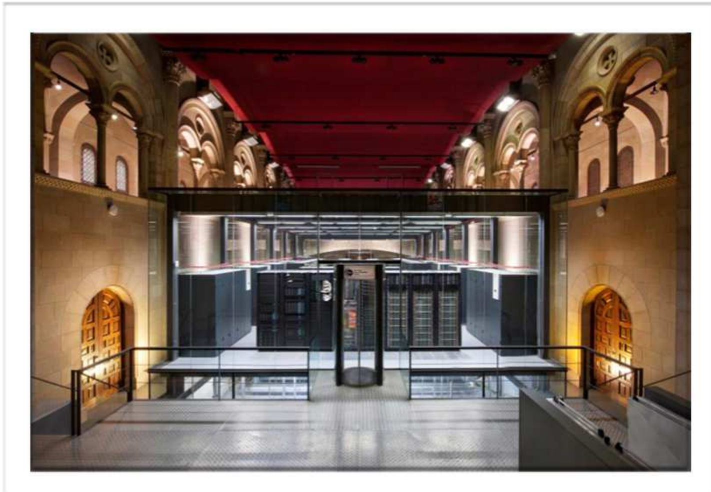
(*) Others use Smart Computing, Intelligent Computing, …

Why Now?

1. Along the explosion of data …

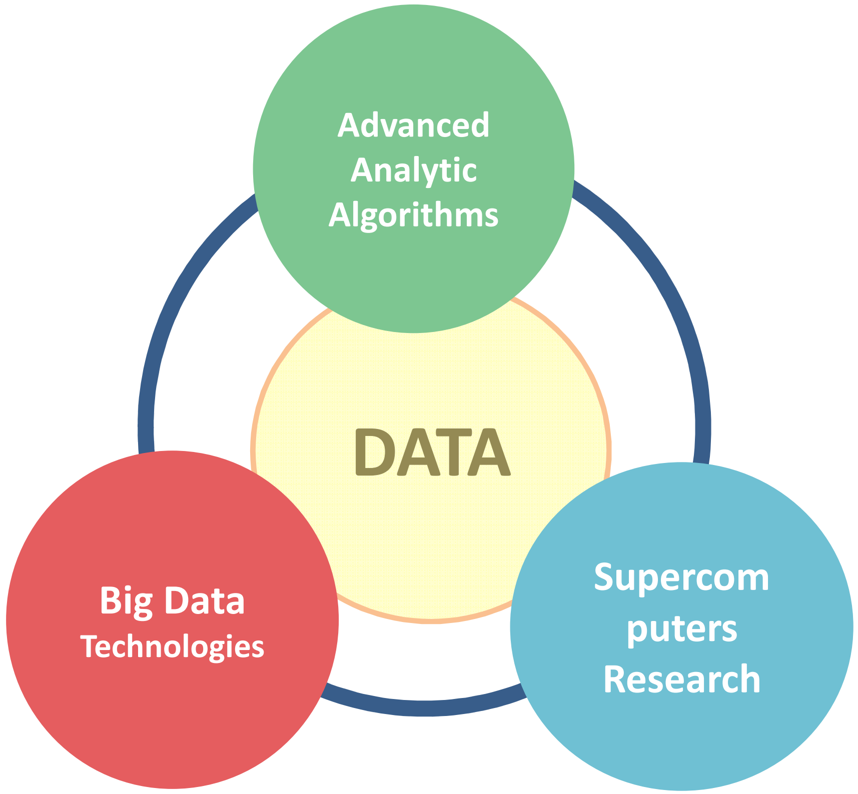
now algorithms can be “trained” by exposing them to large data sets that were previously unavailable.

2. And the computing power necessary to implement these algorithms are now available



Foundational Building Blocks

1. HPC resources
2. Big Data Technologies
3. Cognitive Layer: ML & AI Tech



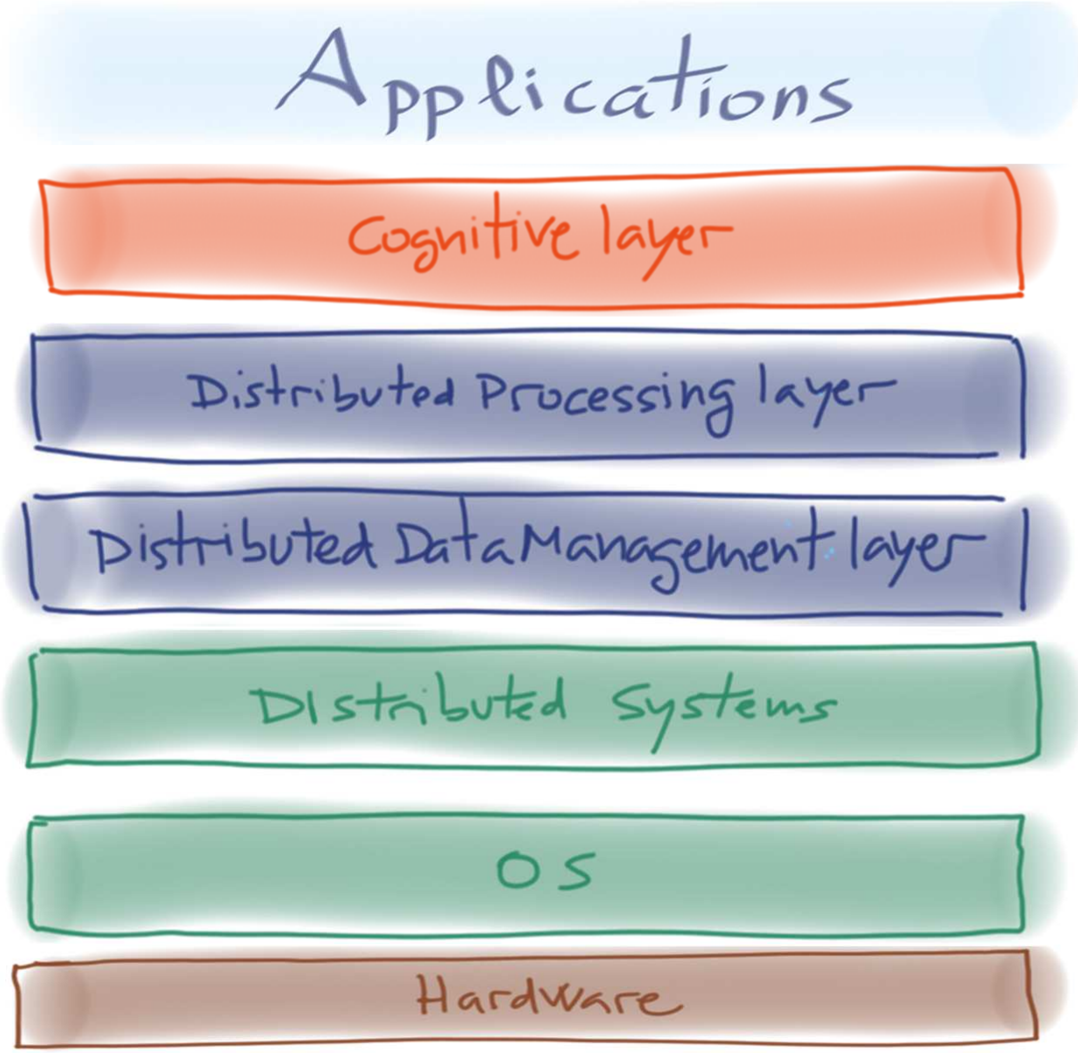
For us "Cognitive Computing" refers to the continuous development of supercomputing systems enabling the convergence of advanced analytic algorithms and big data technologies driving new insights based on the massive amounts of available data

System Middleware: New abstraction layer required

- Systems will have a new cognitive abstraction layer in the software stack

  → Cognitive Layer

offering learning tools, but at the same time, abstracting lower layers to simplify the big data software stack.



Cognitive Layer includes

- Machine Learning algorithms
  - (Neural Networks, SVM, Bayesian methods, …)
- Statistics
  - (regressions, general linear models, decision trees, …)

- Technologies enabled by Artificial Intelligence as

Computer Vision | Speech Recognition | Natural Language Processing

Example: Cognitive Computing is already in business

- In 2011 IBM Watson computer defeated two of Jeopardy' s greatest champions

Since then, Watson supercomputer has become 24 times faster and smarter, 90% smaller, with a 2,400% improvement in performance

Watson Group has collaborated with partners to build 6,000 apps

- The project [Viv](Viv) built by Siri's creators

Booking a flights …

"I want a flight to MWC Barcelona with a return five days later via London."

Just closed on $12.5 M in venture capital funding.

HOW BIG DATA IS ALREADY INFLUENCING OUR EVERYDAY LIVES

The solution analyzed 70,000 scientific articles on p53

#Research

Baylor College of Medicine (Houston, Texas)

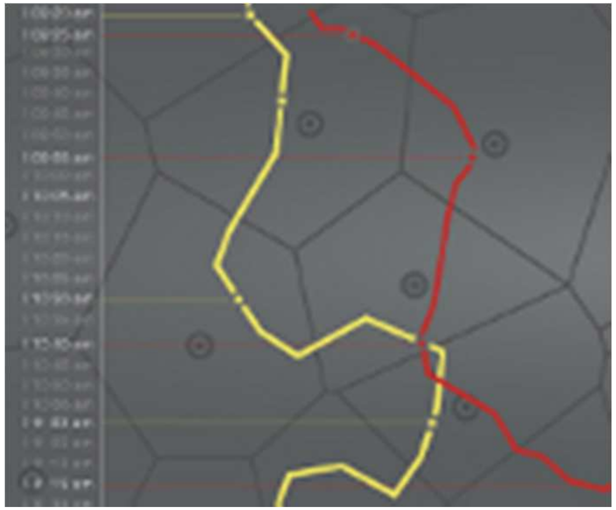
Privacy & Big Data

- Even when real names and other personal information are stripped from big data sets, it is often possible to use just a few pieces of the information to identify a specific person.
  - Example:
    - Data: credit card transactions made by 1.1 million people in 10,000 stores over a three-month period.
    - Results: knowing just four random pieces of information was enough to reidentify 90 percent of the shoppers as unique individuals and to uncover their records.
    - And that uniqueness of behavior combined with publicly available information, like Instagram or Twitter posts, could make it possible to reidentify people's records by name.
    - Source: http://www.sciencemag.org/content/347/6221/536.abstract

Privacy & Big Data

- The old model of anonymity doesn’t seem to be the right model when we are talking about large-scale metadata.

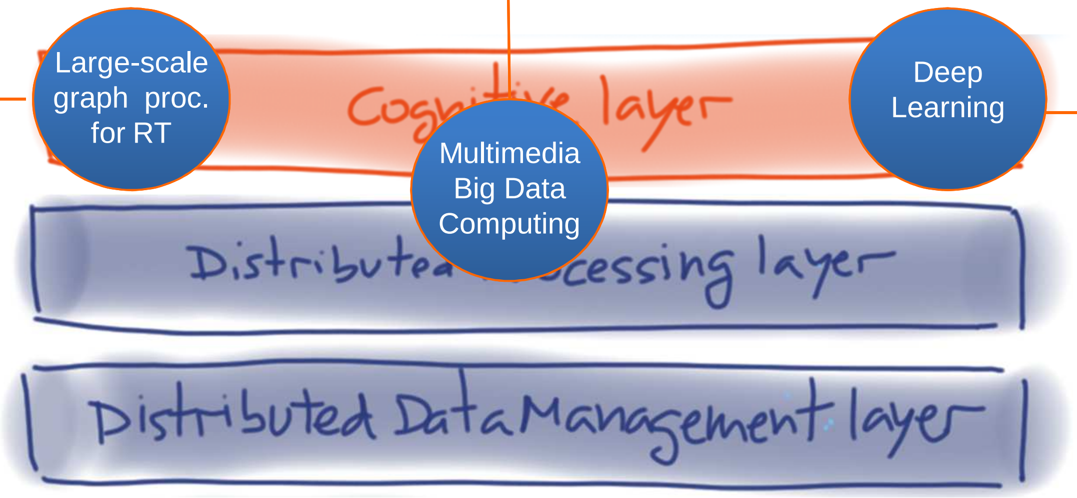
- Example:
  - *Data: 15 months of data from 1.5 million people*
  - *4 points (approximate places and times) are enough to identify 95% of individuals in a mobility database.*
  - *human behavior puts fundamental natural constraints to the privacy of individuals and these constraints hold even when the resolution of the dataset is low; even coarse datasets provide little anonymity.*
  - *Source:* M.I.T. Media Lab
    www.demontjoye.com/projects.html



Ongoing Big Data projects at BSC - IV

Visual concept detection & notation
Scalable indexing of image and video
Multimodal Big Data Analytics
Master Thesis

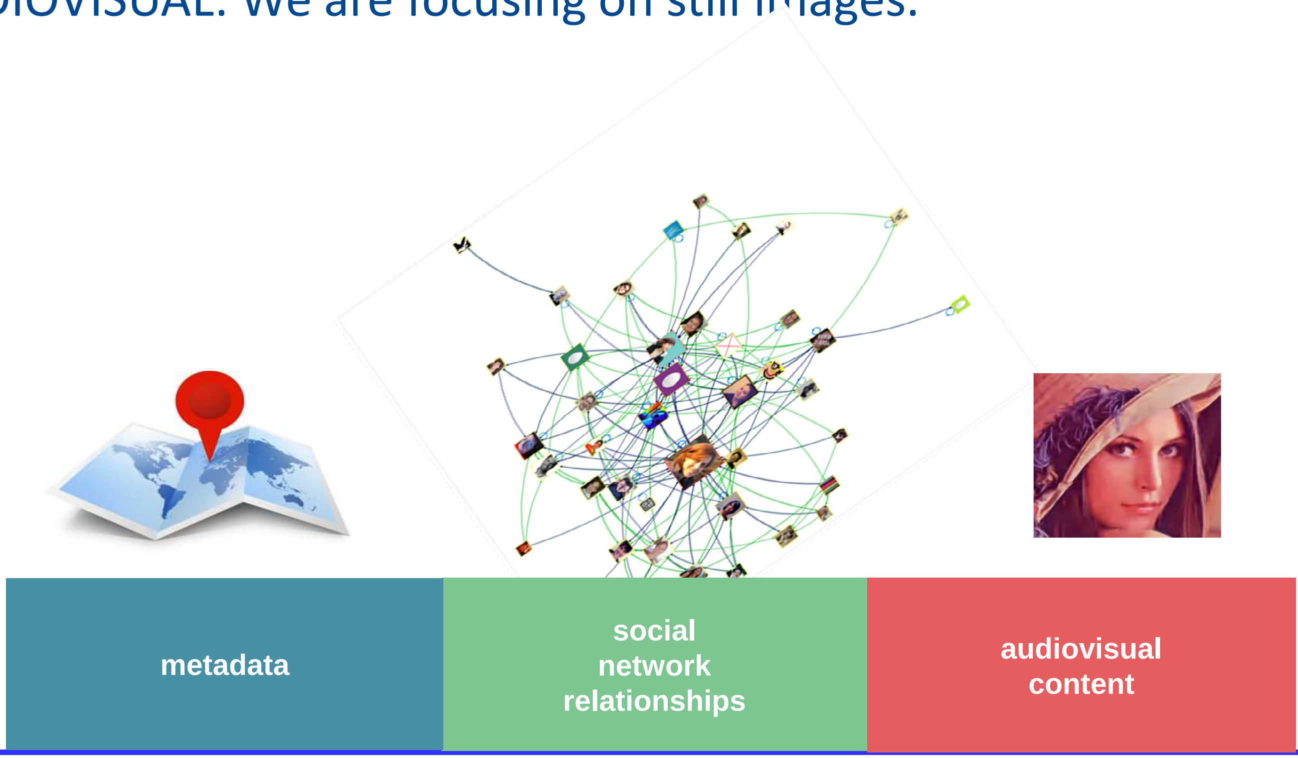
Use of probabilistic Bayesian graph algorithms for real-time applications and its efficient parallel computation



Enhance new deep neural network algorithms, such as auto-encoders of Boltzmann machines to create generative models

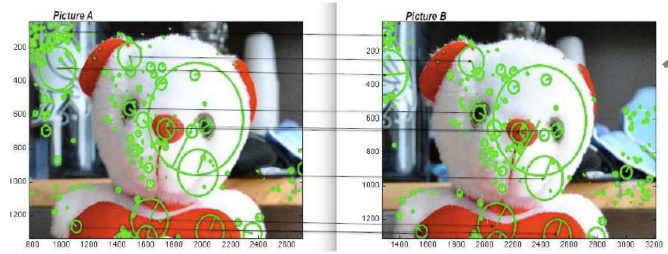
Example of research at BSC: Multimodal Big Data Computing

- The challenge is to work with three kinds of data, at the same time:
  - METADATA: Mainly geolocation, time and user defined tags. Also short descriptions, titles, surrounding text (twitter).
  - SOCIAL NETWORK: Graphs of followers, likes and comments
  - AUDIOVISUAL: We are focusing on still images.

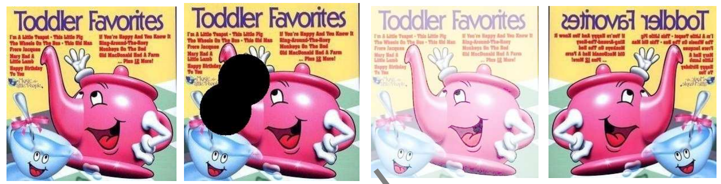


Example of research at BSC: Multimodal Data Computing

- We are analyzing the "CONTENT" of the images

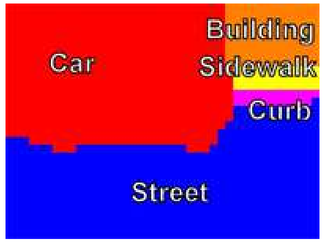


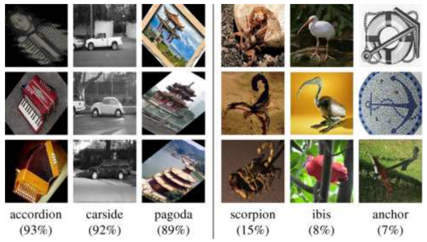
object recognition



scene undrestanding

near replica detection





concept annotation

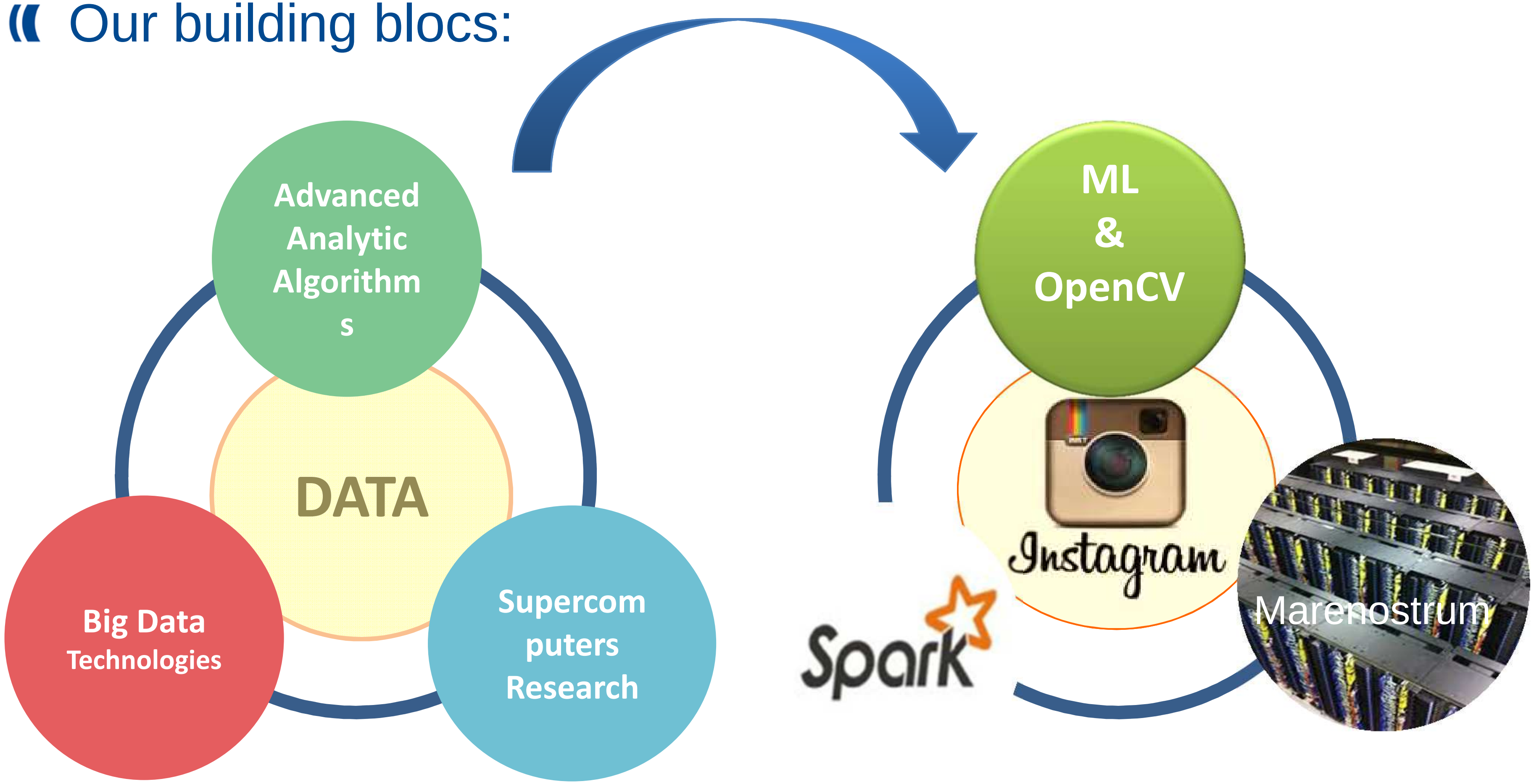
face detection
face verificacion
face identification

image style recogniton

clustering

Example of research at BSC: Multimodal Big Data Computing

- Our building blocs:



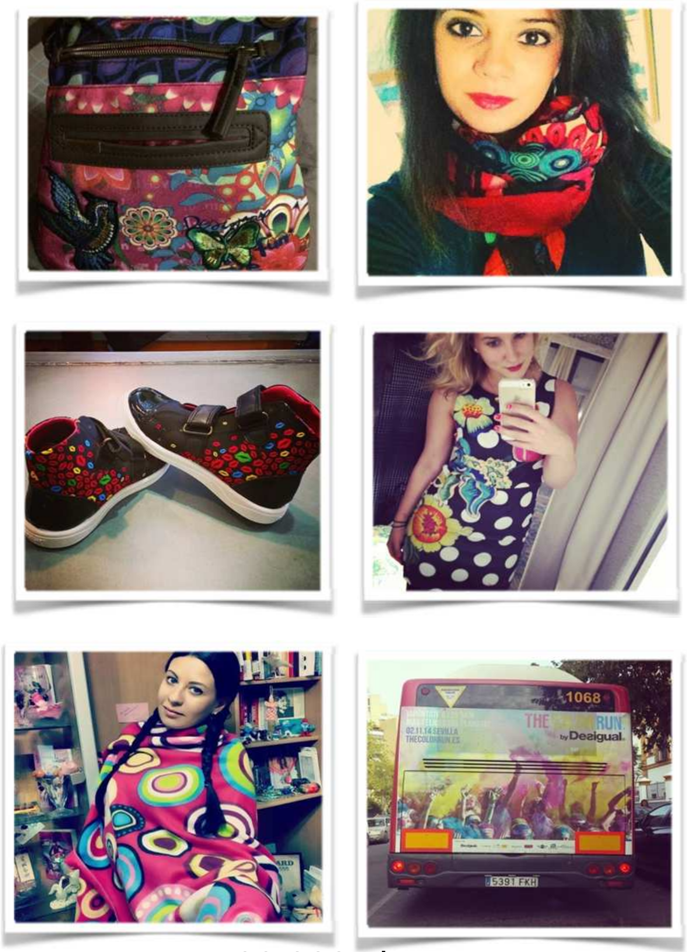
Case Study: Desigual

- Multimodal Data Analytics systems can aid Desigual in better understanding their customers and potential customers through the analysis of social media data sources

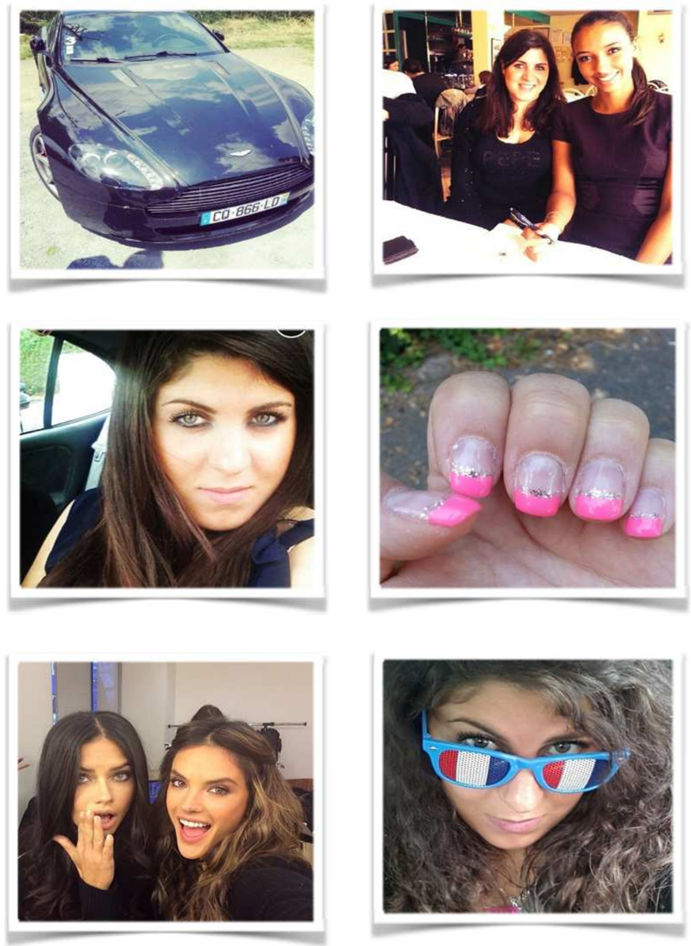
Case Study: Desigual → 2 data sets

#desigual #lavidaeschula
#mydesigual

followers



30.000 photos

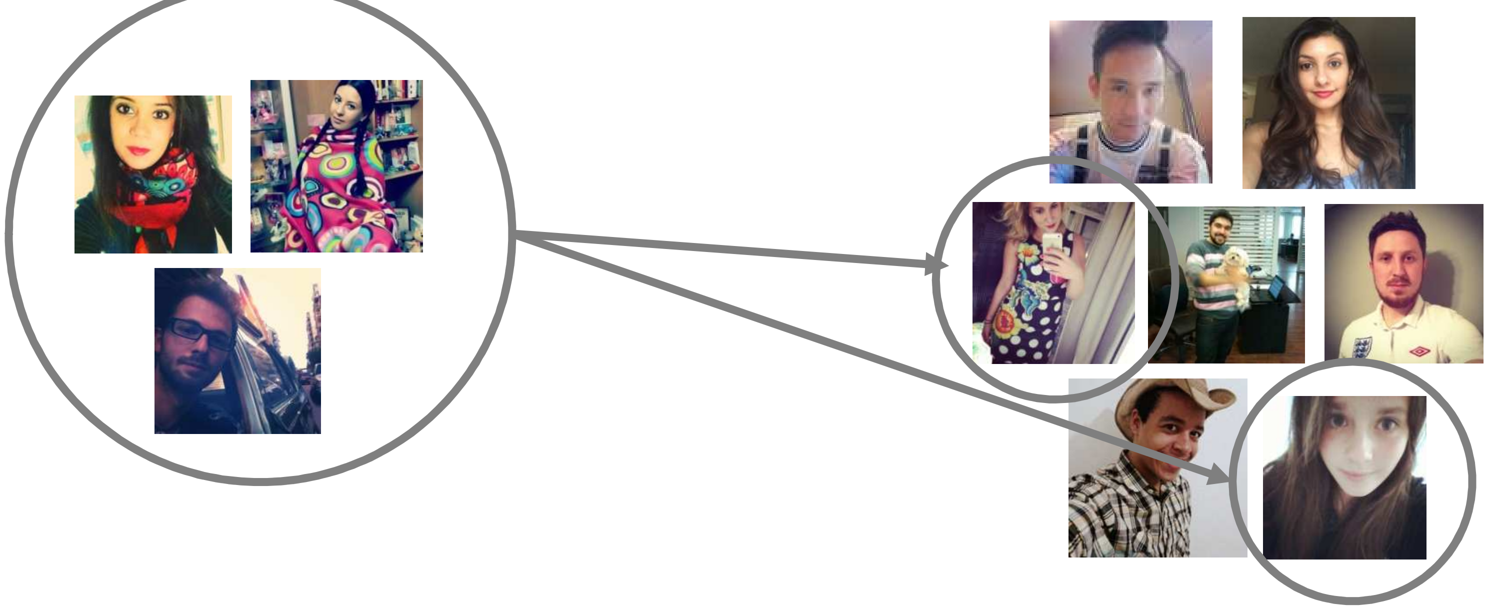


100 photos x 2K followers = 200K Photos (100 GB)

Case Study Desigual: Latent User Attribute Inference

E.g. Predicting Desigual Followers

AGE
GENDER
HOME LOCATION
TRAVEL PATTERNS
LIFESTYLE/CONSUMPTION PATTERNS
…



Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

BIG DATA IMPLICATIONS

Implications of Big Data Analytics

1. **We may be controlled by algorithms** that are likely to predict what we are about to do.
   - Privacy was the central challenge in the second wave era. In the next wave of Cognitive Computing, the challenge will be safeguarding free will.

   - Example: Last week at MWC, Qualcomm announced their latest chip, the Snapdragon 820. The new platform, called Zeroth, is said to anticipate the users actions in advance (deep learning devices)

Implications of Big Data Analytics

2. Big Data Analytics is going to **challenge white collar**, professional knowledge work in the 21st century in the same way that factory automation and the assembly line challenged blue collar labor in the 20th century.

Researchers at Oxford published a study estimating that 47 percent of total US employment is "at risk" due to the automation of cognitive tasks.

THE FUTURE OF EMPLOYMENT: HOW SUSCEPTIBLE ARE JOBS TO COMPUTERISATION?*

Carl Benedikt Frey† and Michael A. Osborne‡

September 17, 2013

Abstract

We examine how susceptible jobs are to computerisation. To assess this, we begin by implementing a novel methodology to estimate the probability of computerisation for 702 detailed occupations, using a Gaussian process classifier. Based on these estimates, we examine expected impacts of future computerisation on US labour market outcomes, with the primary objective of analysing the number of jobs at risk and the relationship between an occupation's probability of computerisation, wages and educational attainment. According to our estimates, about 47 percent of total US employment is at risk. We further provide evidence that wages and educational attainment exhibit a strong negative relationship with an occupation's probability of computerisation.

Keywords: Occupational Choice, Technological Change, Wage Inequality, Employment, Skill Demand

JEL Classification: E24, J24, J31, J62, O33.

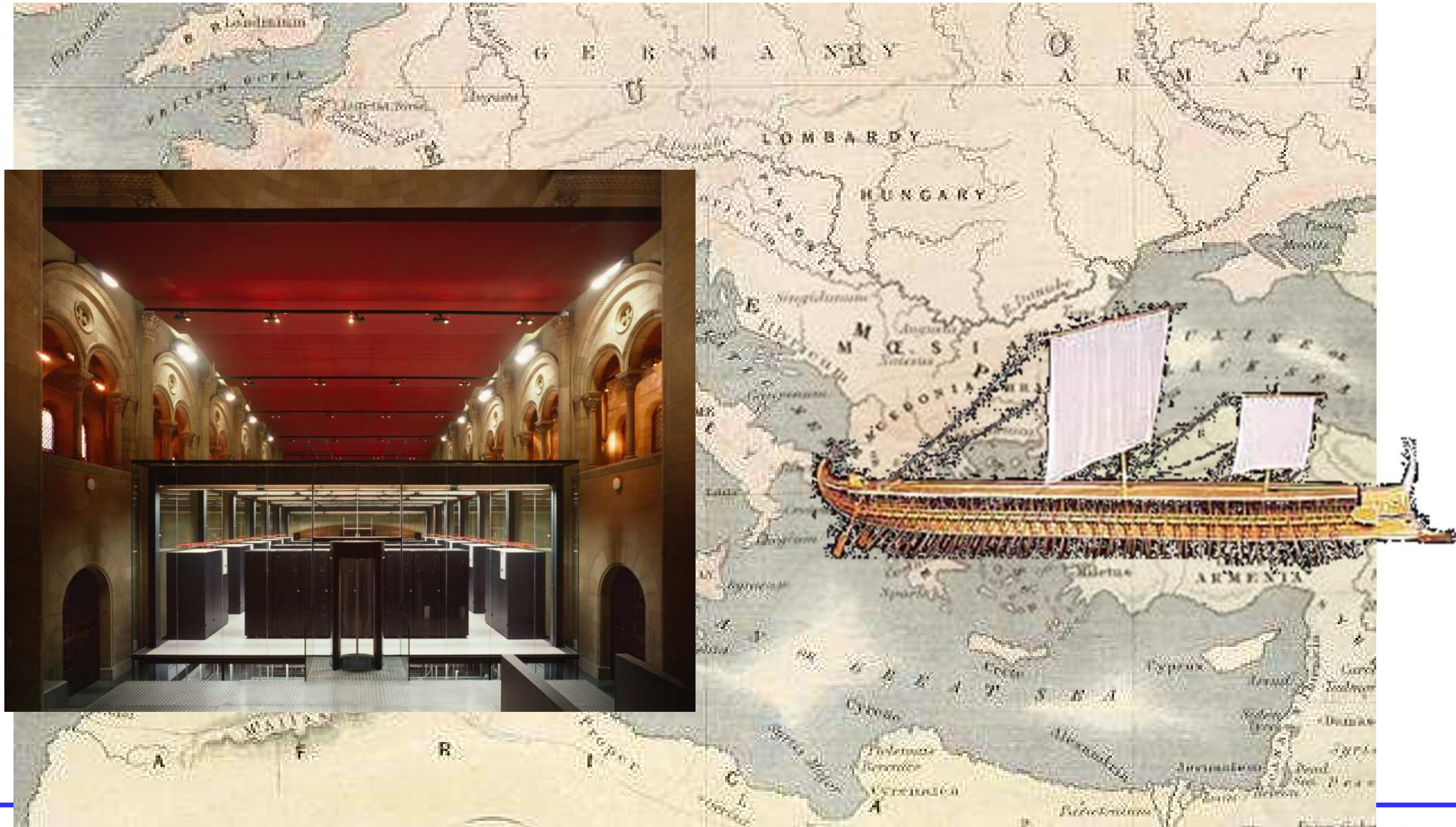
*We thank the Oxford University Engineering Sciences Department and the Oxford Martin Programme on the Impacts of Future Technology for hosting the "Machines and Employment" Workshop. We are indebted to Stuart Armstrong, Nick Bostrom, Eris Chinellato, Mark Cummins, Daniel Dewey, David Dorn, Alex Flint, Claudia Goldin, John Muellbauer, Vincent Mueller, Paul Newman, Seán Ó hÉigeartaigh, Anders Sandberg, Murray Shanahan, and Keith Woolcock for their excellent suggestions.

†Oxford Martin School, Programme on the Impacts of Future Technology, University of Oxford, Oxford, OX1 1PT, United Kingdom, carl.frey@philosophy.ox.ac.uk.

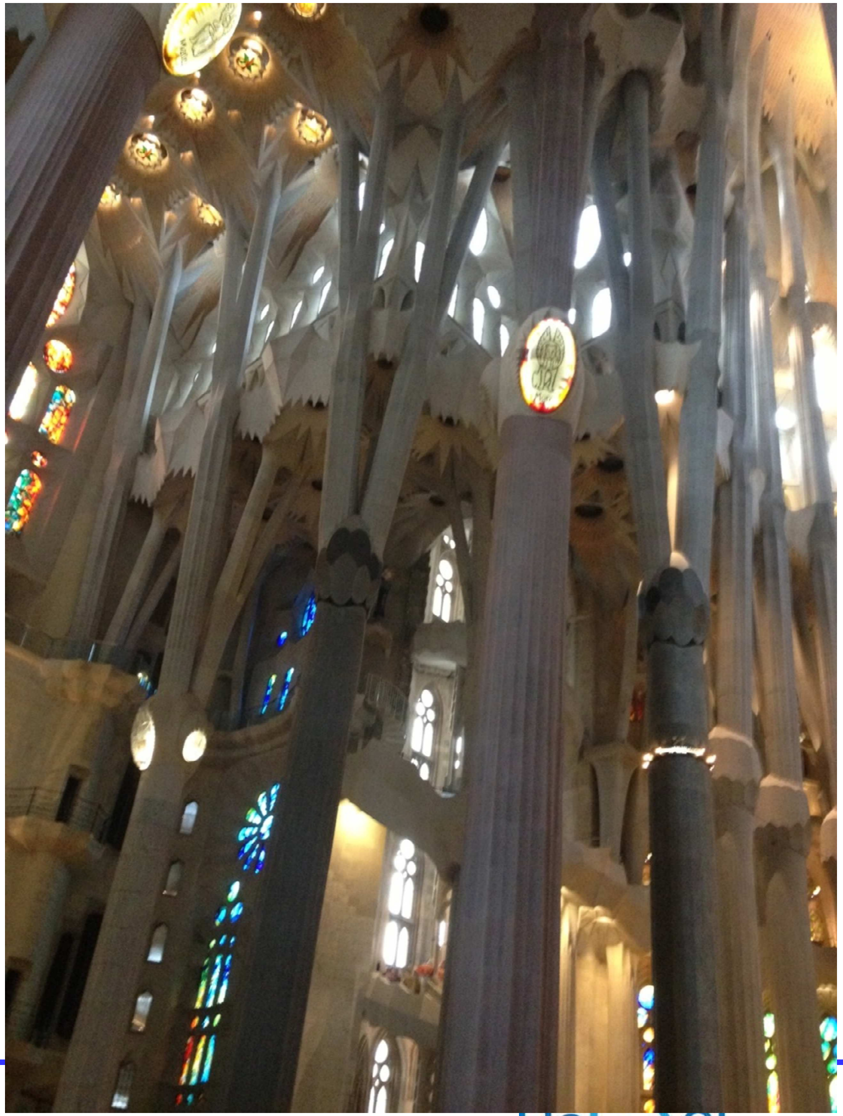
‡Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, United Kingdom, mosb@robots.ox.ac.uk.

1

Navigating the Mare Nostrum



Are we planning to upgrade?.. Negotiating our next site ;)




www.bsc.es
BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación
Thank you!