www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

El estado del arte del Big Data & Data Science: La revolución de los datos

Prof. Mateo Valero
BSC Director

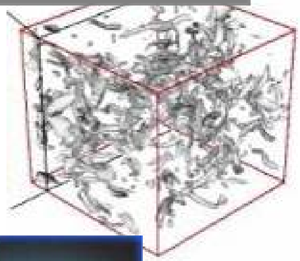
Big Data: de la investigación científica a la gestión empresarial
Fundación Ramón Areces
Madrid, 3 de Julio de 2014

Evolution over time of the research paradigm

- In the last millennium, science was empirical:
  - description of natural phenomena

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

- A few centuries ago opens the theoretical approach
  - using models, formulas and generalizations.

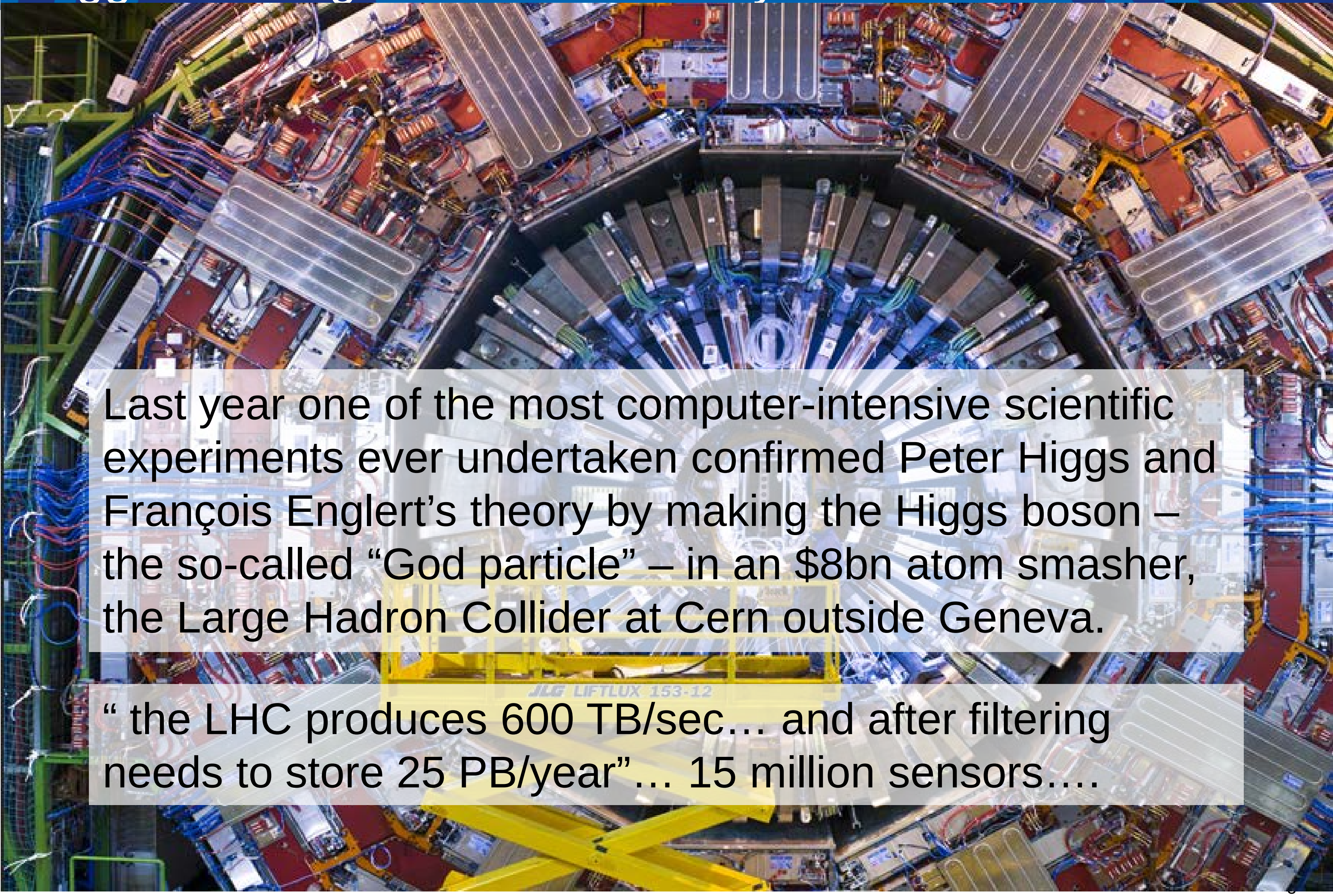


- In recent decades appears computational science
  - simulation of complex phenomena

- Now focuses on the exploration of big data (eScience)
  - unification of theory, experiment and simulation
  - capture massive data using instruments or generated through simulation and processed by
  - computer
  - knowledge and information stored in computers
  - Scientists analyse databases and files on data infrastructures

Jim Gray, National Research Council, http://sites.nationalacademies.org/NRC/index.htm;
Computer Science and Telecommunications Board, http://sites.nationalacademies.org/cstb/index.htm.

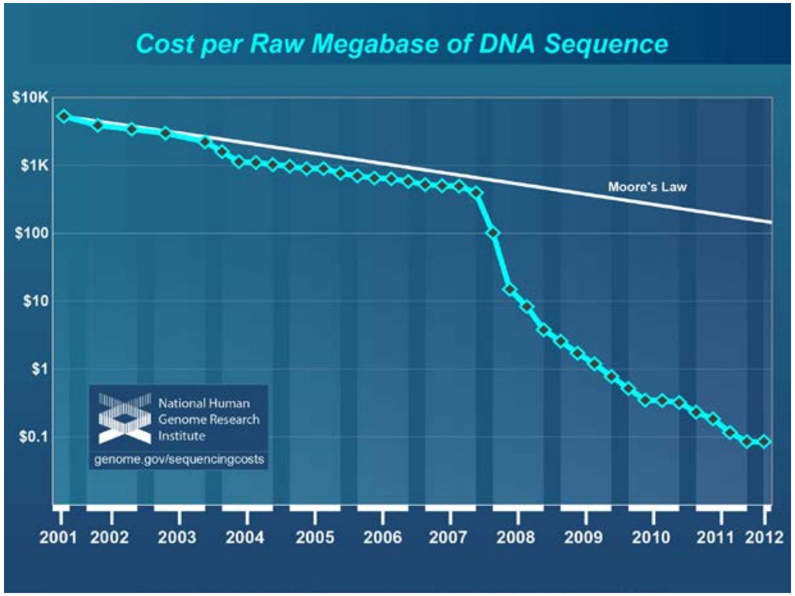
Higgs and Englert’s Nobel for Physics 2013

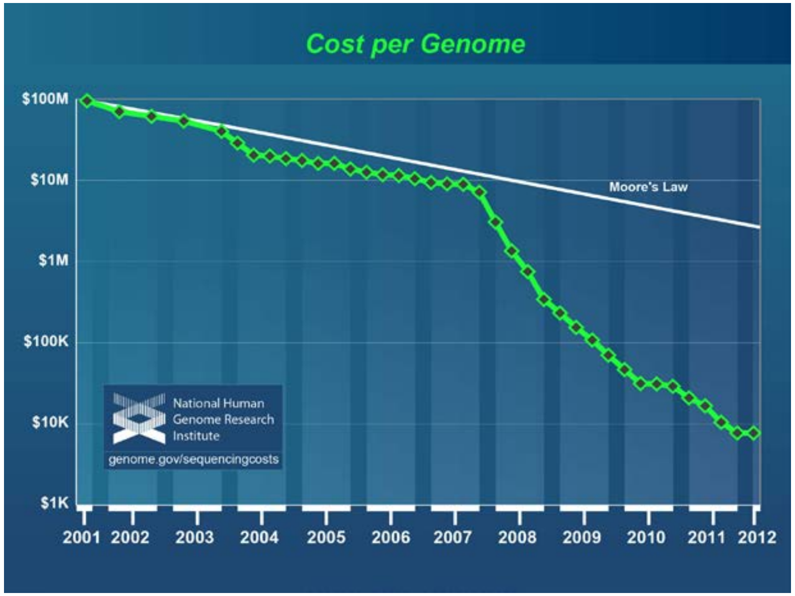


Last year one of the most computer-intensive scientific experiments ever undertaken confirmed Peter Higgs and François Englert’s theory by making the Higgs boson – the so-called “God particle” – in an $8bn atom smasher, the Large Hadron Collider at Cern outside Geneva.

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

Sequencing Costs



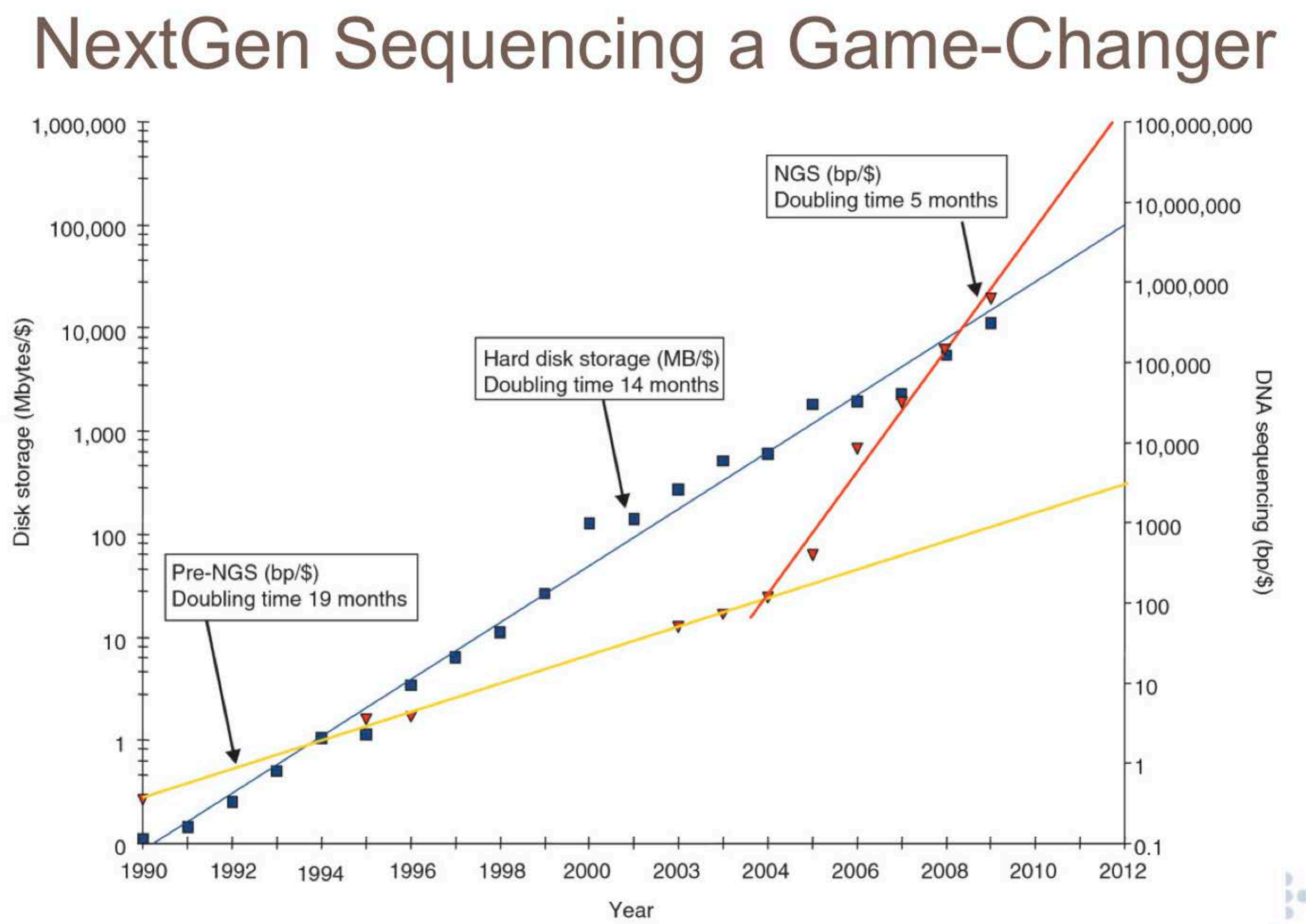


Source: **National Human Genome Research Institute** (NHGRI)
http://www.genome.gov/sequencingcosts/

(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.

Sequencing Costs



We have an issue…. Need for faster ways of processing data, not storing everything…

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

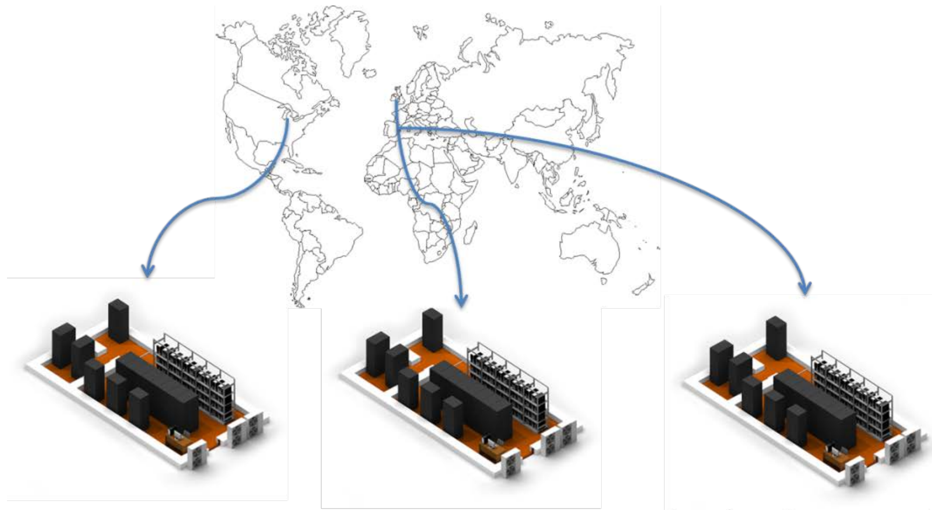
RoMoL Project

World-class genomic consortiums

european genome-phenome archive

International Cancer Genome Consortium

The Cancer Genome Atlas
Understanding genomics to improve cancer care



University of Chicago
NIH Trusted Partner
Holds TCGA Data

EBI
Data Manager for 1KGP
Hosts EGA

Barcelona Supercomputer Center
EGA Partner

THE UNIVERSITY OF CHICAGO MEDICINE

EMBL-EBI

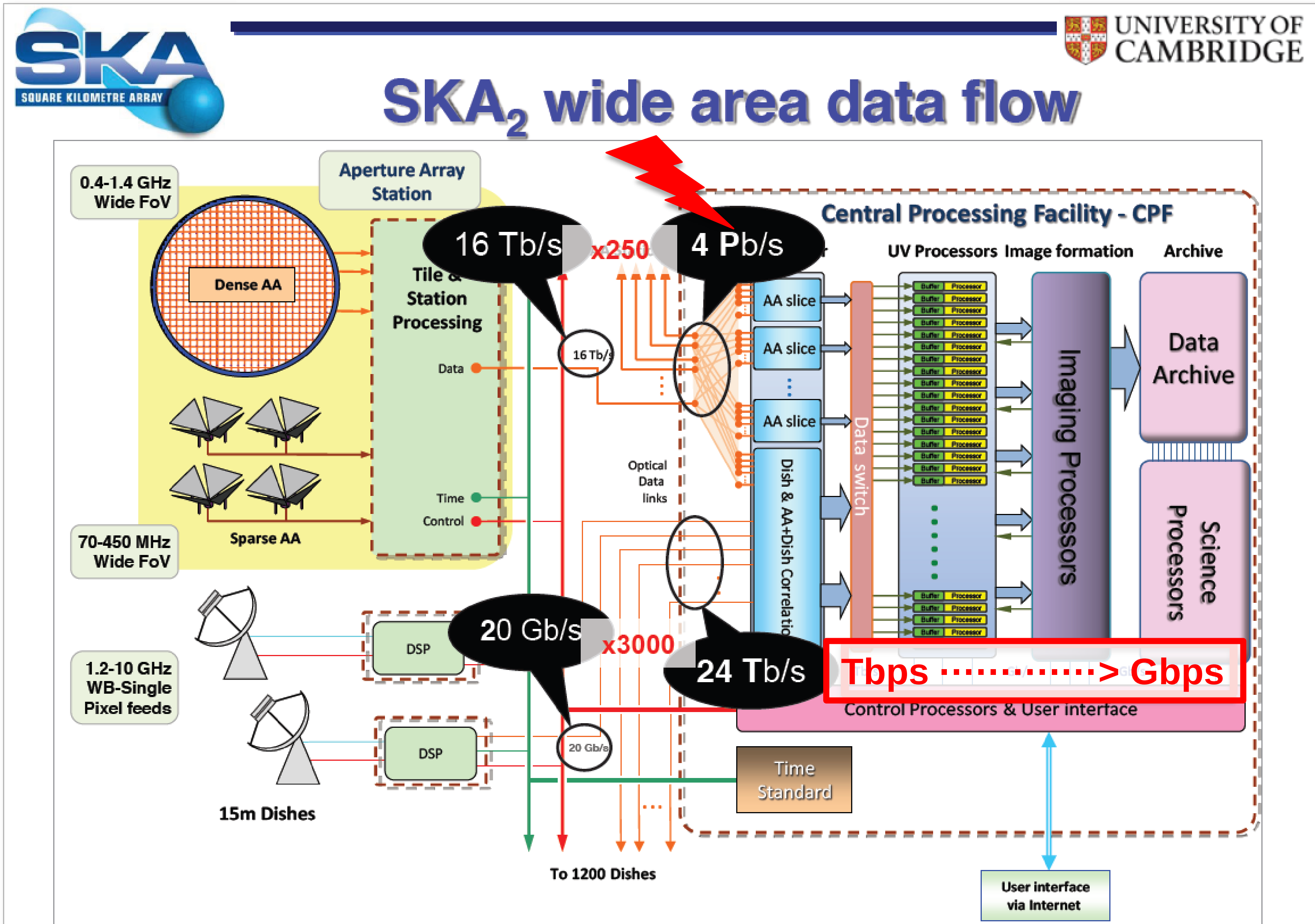
EXCELENCIA SEVERO OCHOA
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

- European Genome-phenome Archive (EGA) repository will allow us to explore datasets from numerous genetic studies.

- Pan-cancer will produce rich data that will provide a major opportunity to develop an integrated picture of differences across tumor lineages.

Square Kilometer Array (SKA)

Square Kilometer Array (SKA)



A New Era of Information Technology

Current infrastructure sagging under its own weight

2013

- **98,000** tweets
- **23,148** apps downloaded
- **400,710** ad requests
- **In 60 sec today**
- **2000** lyrics played on Tunewiki
- **1,500** pings sent on PingMe
- **208,333** minutes Angry Birds played

Internet of Things

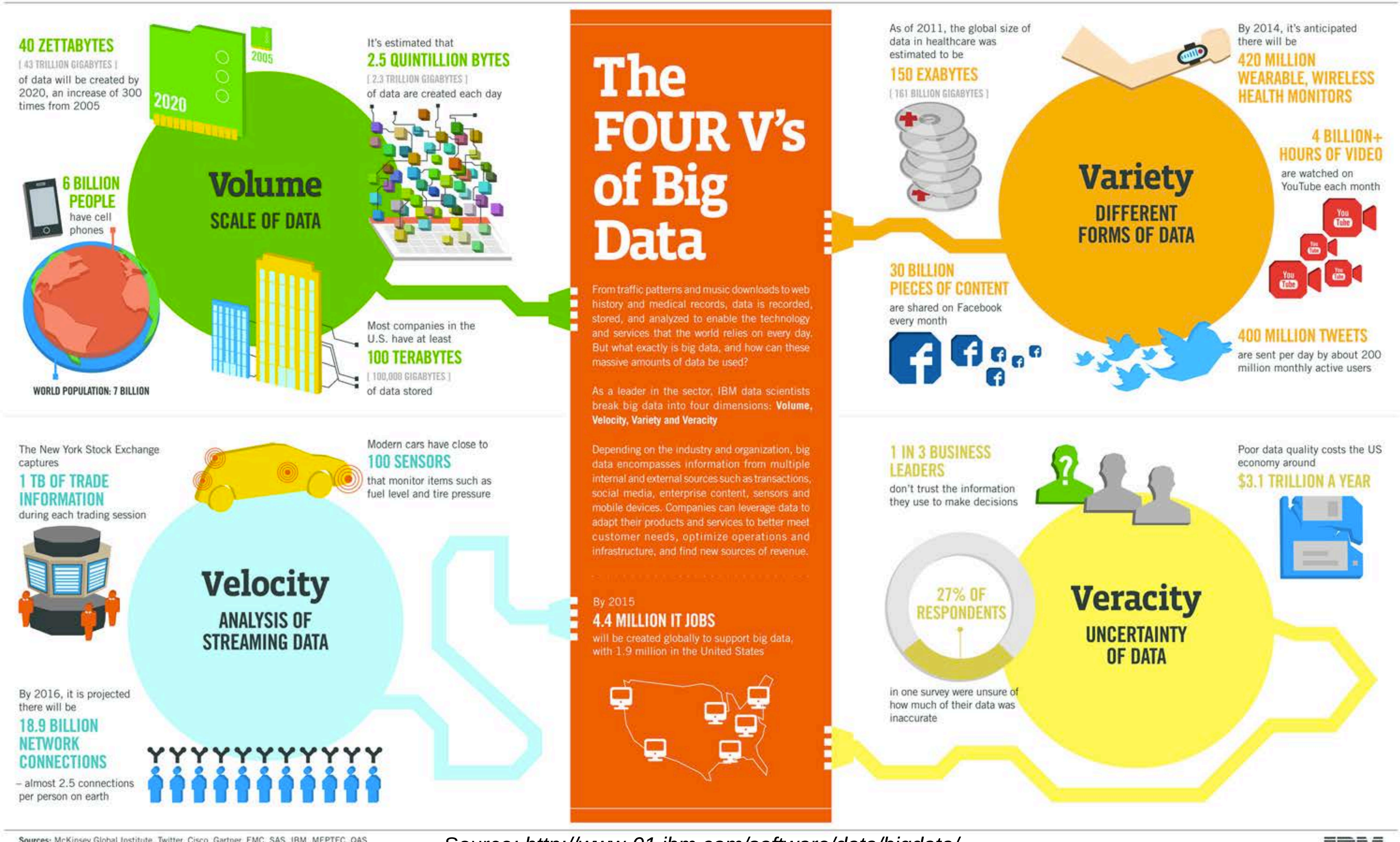
Pervasive Connectivity | Smart Device Expansion | Explosion of Information

By 2020

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

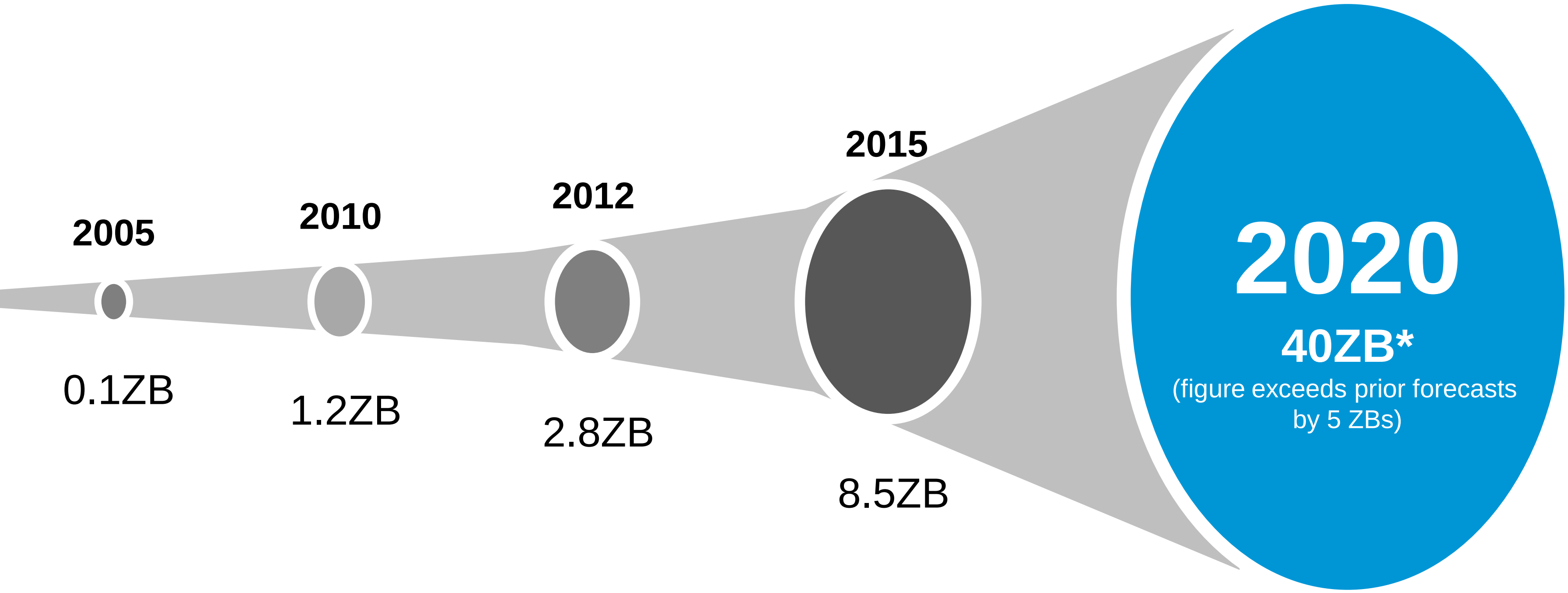
(1) IDC Directions 2013: Why the Datacenter of the Future Will Leverage a Converged Infrastructure, March 2013, Matt Eastwood ; (2) & (3) IDC Predictions 2012: Competing for 2020, Document 231720, December 2011, Frank Gens; (4) http://en.wikipedia.org

Challenges of data generation



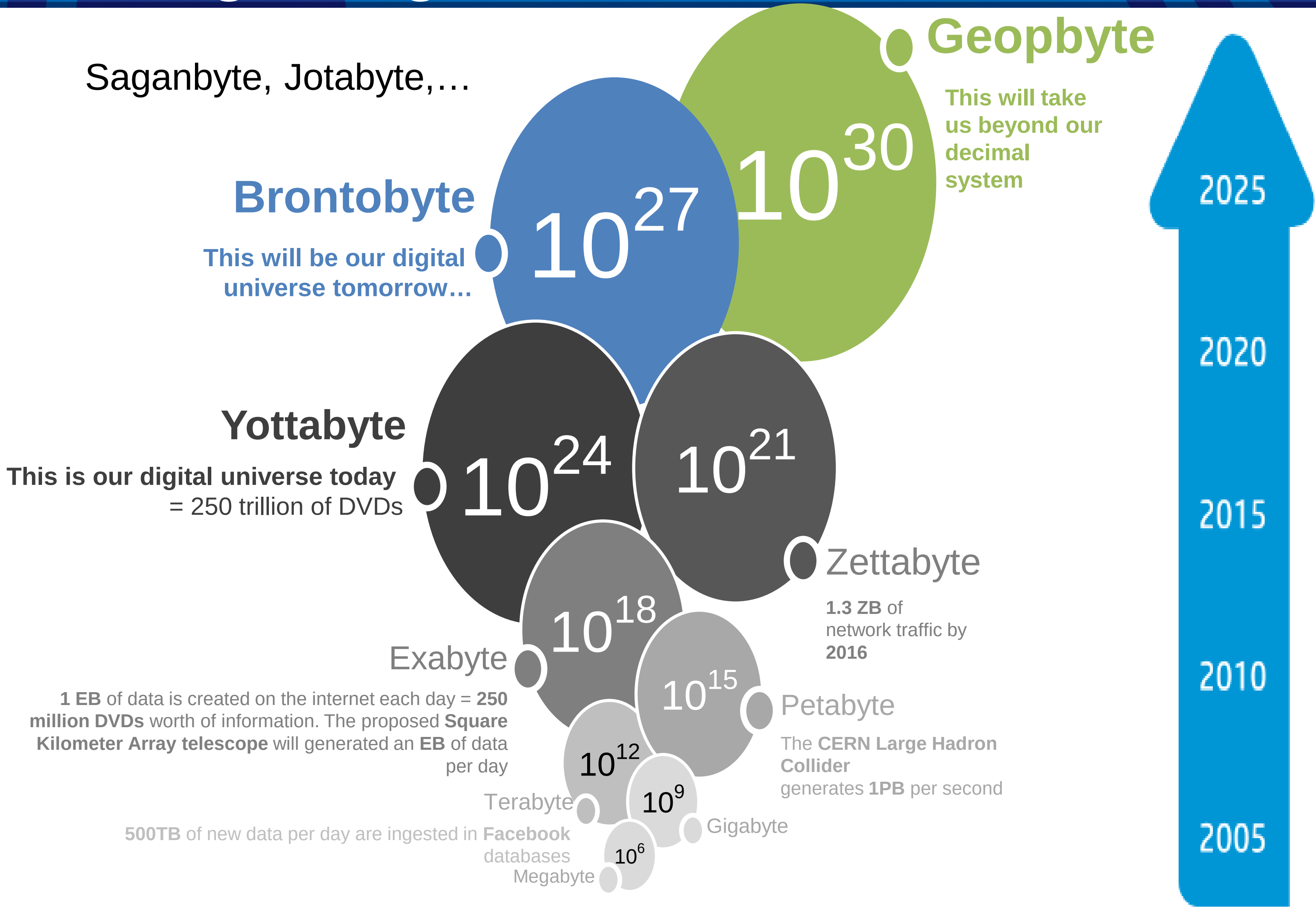
Source: *http://www-01.ibm.com/software/data/bigdata/*

The Data Deluge



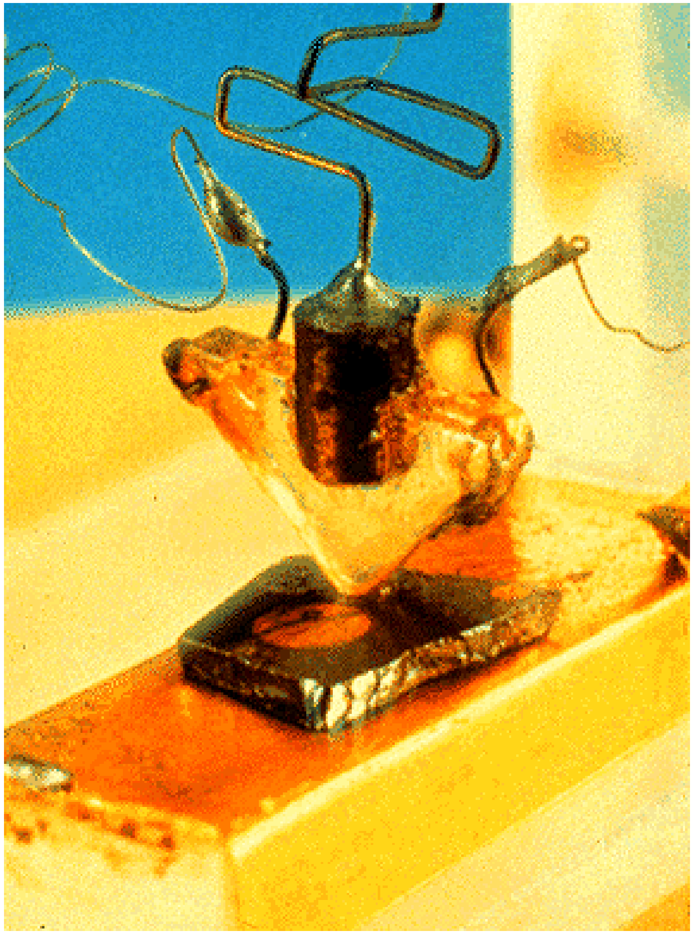
* Source: IDC

How big is big?



Technological Achievements

- Transistor (Bell Labs, 1947)
  - DEC PDP-1 (1957)
  - IBM 7090 (1960)
- Integrated circuit (1958)
  - IBM System 360 (1965)
  - DEC PDP-8 (1965)
- Microprocessor (1971)
  - Intel 4004



Outline

- The Big Data era: data generation explosion
- Storing data
- Processing data
- Where do we place data?
- How we use data?
- Research at BSC

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

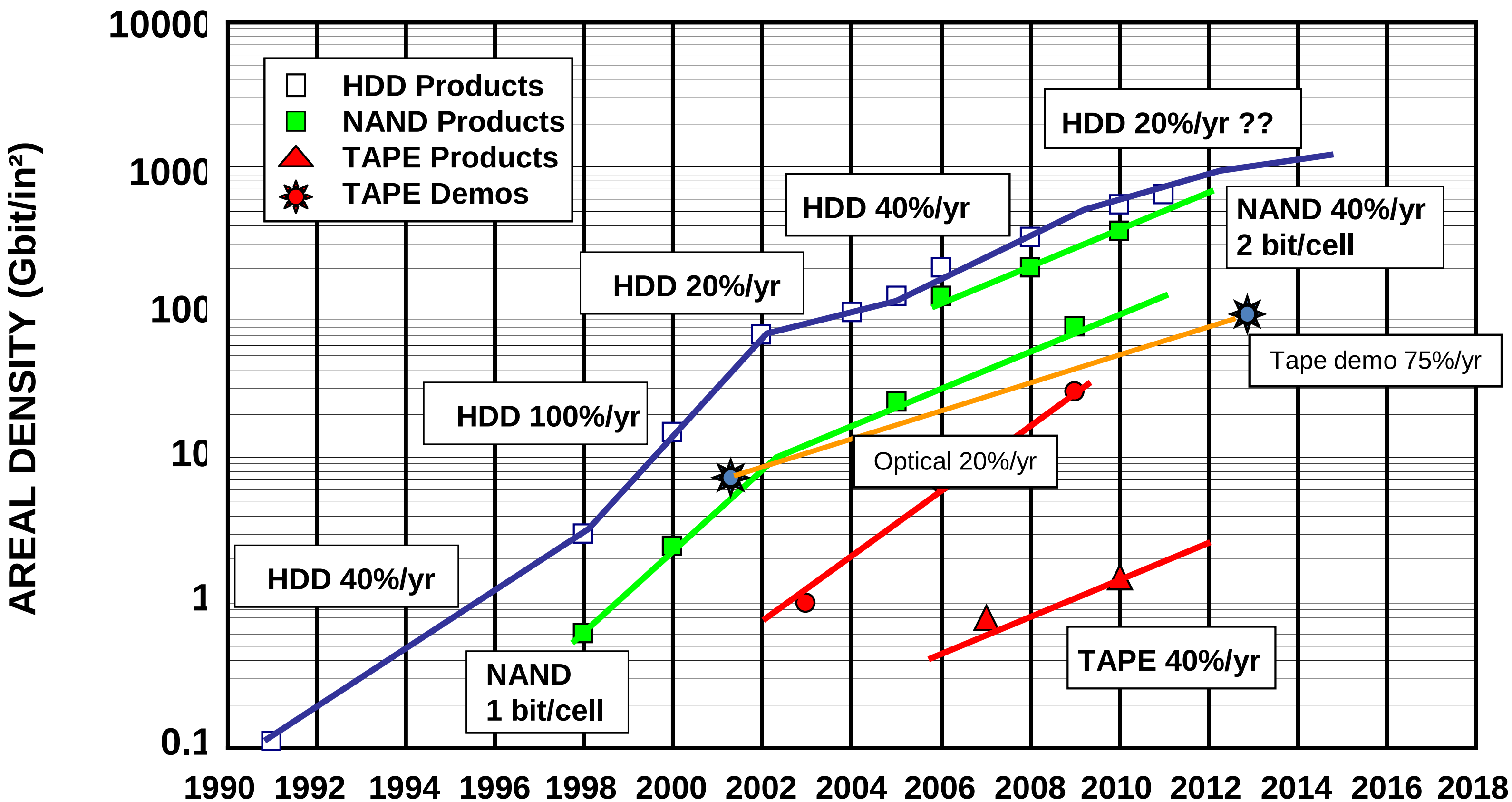
HDD: Hard Drives Disk

- 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

Storage Device Density Landscape



Source: Decad, Fontana,Hetzler
– IBM Journal

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

RoMoL Project

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.

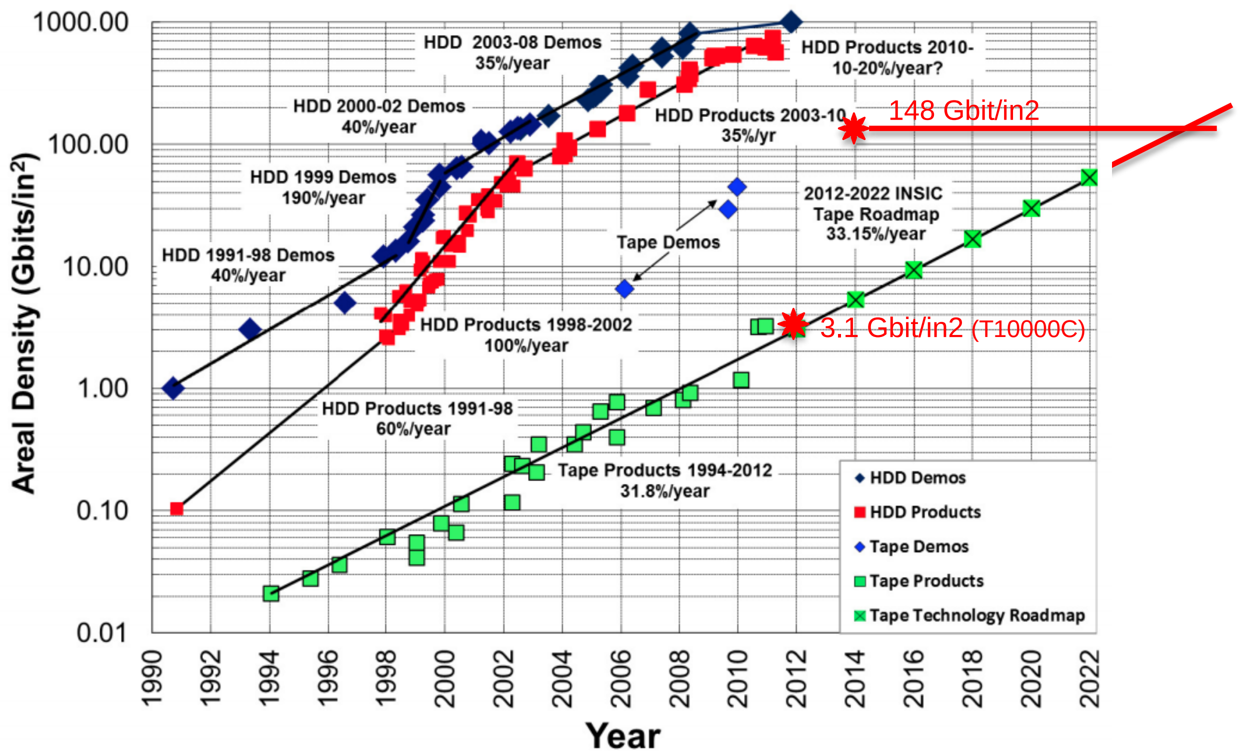
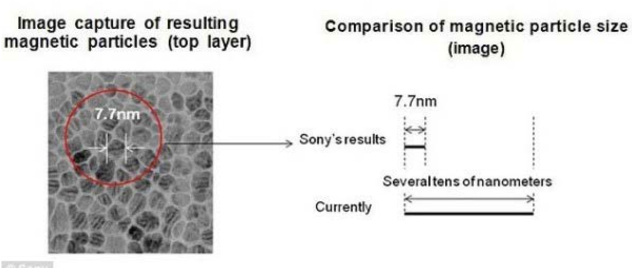


Figure 40: Areal Density of Hard Disk and Tape Laboratory Demonstrations and Products [71].



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

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

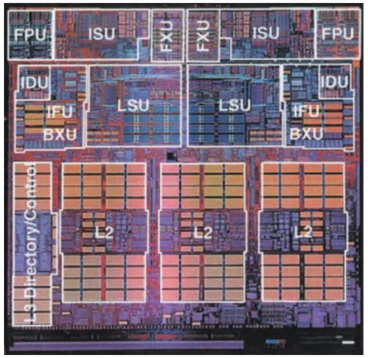
RoMoL Project

Outline

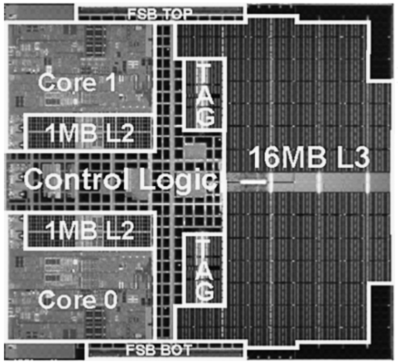
- The Big Data era: data generation explosion
- Storing data
- Processing data
- Where do we place data?
- How we use data?
- Research at BSC

- Moore's Law + Memory Wall + Power Wall

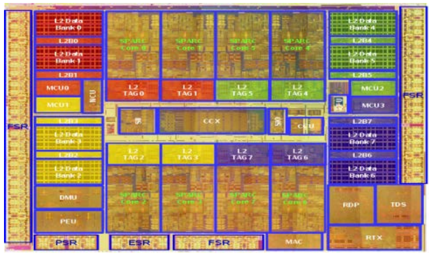
Chip MultiProcessors (CMPs)



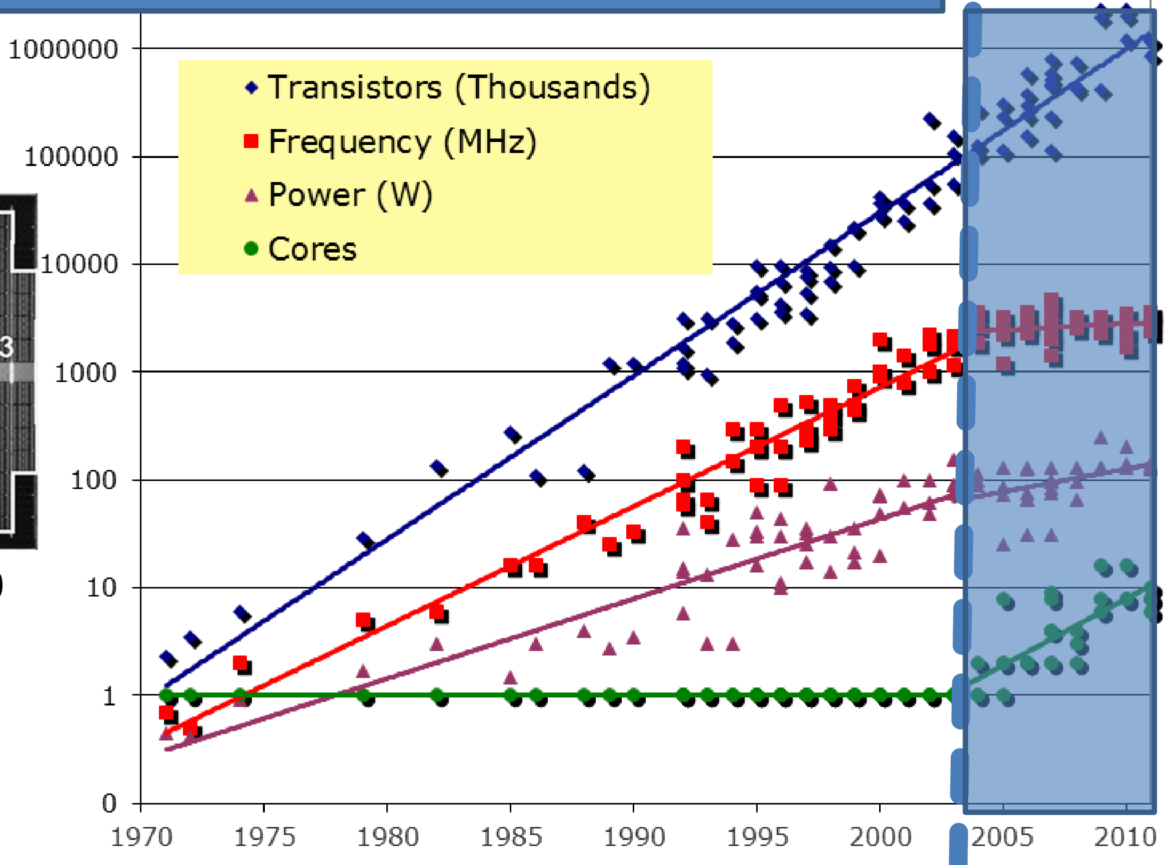
POWER4 (2001)



Intel Xeon 7100 (2006)



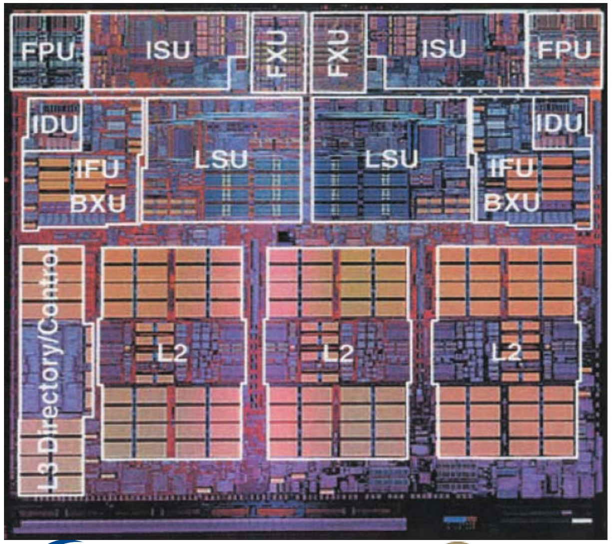
UltraSPARC T2 (2007)



How are the Multicore architectures designed?

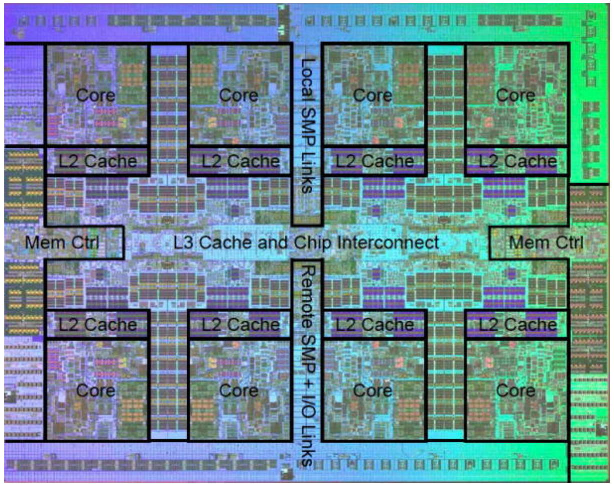
IBM Power4 (2001)

- 2 cores, ST
- 16 Gflops/s
- L1: 64KB+32KB/core
  L2: 1.41MB /chip
  L3: 32MB off-chip
- 115W TDP
- 10GB/s mem BW



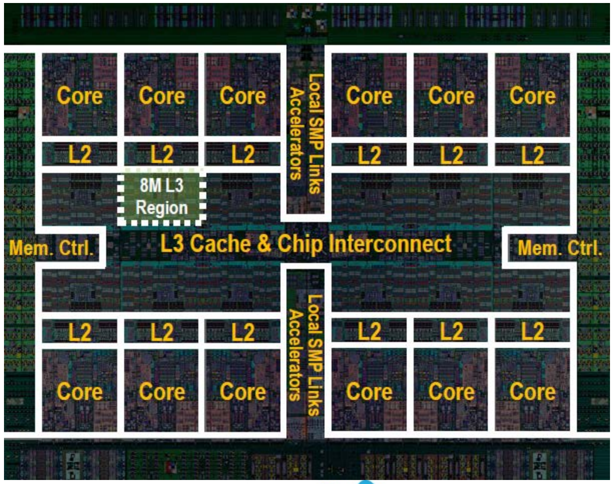
IBM Power7 (2010)

- 8 cores, SMT4
- 128 Gigaflops/s
- L1: 32KB+32KB/core
  L2: 256KB/core
  L3·: 32 MB in total (on-chip)
- 170W TDP
- 100GB/s mem EBW



IBM Power8 (2014)

- 12 cores, SMT8
- 400 Gigaflops/s
- L1:64KB+32KB/core
  L2: 512KB/core
  L3:49,96 MB in total (on-chip)
- 250W TDP
- 410GB/s mem BW

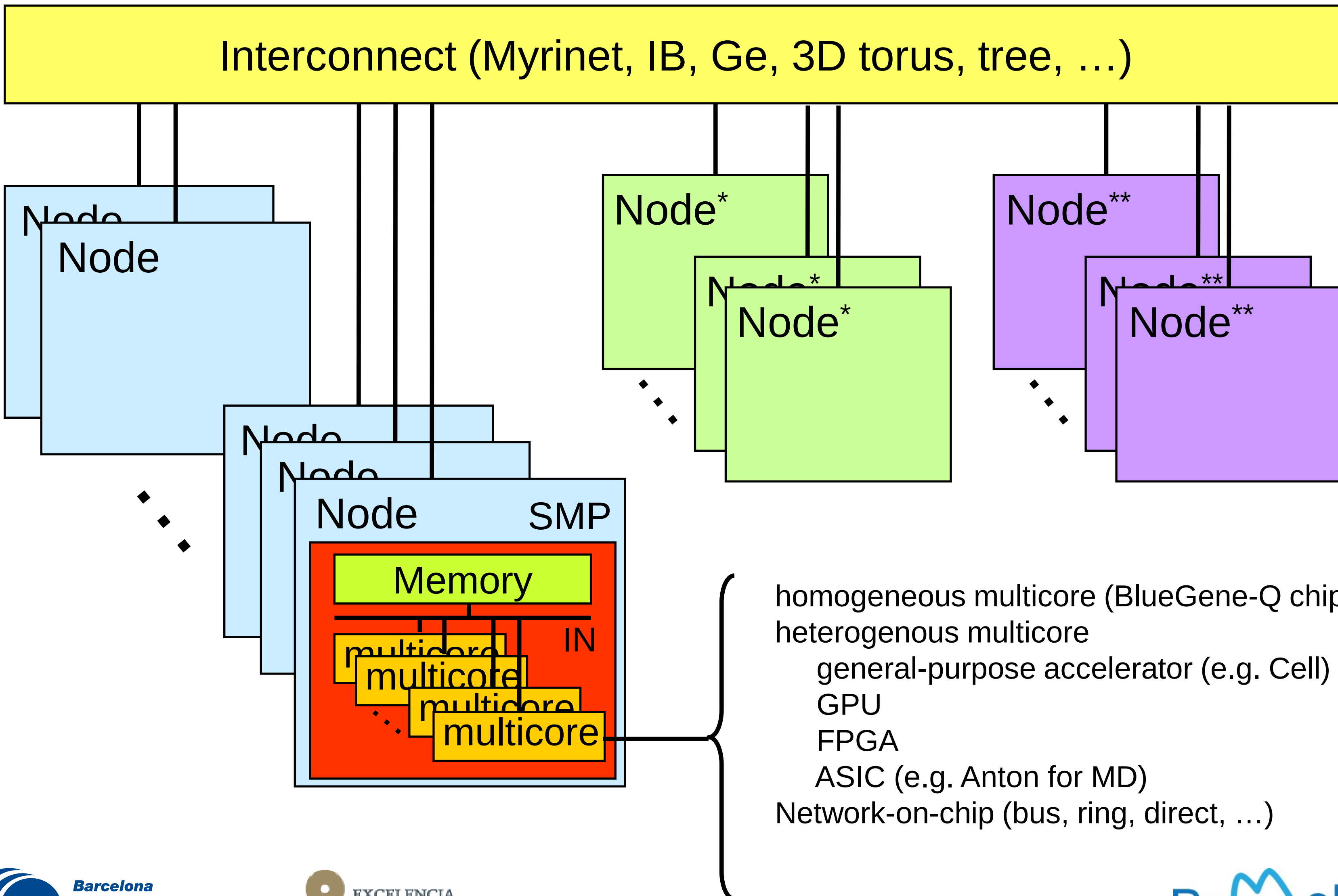


Top10

TOP500 JUNE 2014

Rank	Site	Computer	Procs	Rmax	Rpeak	Power	GFlops/Watt	Name
1	National University of Defense Technology	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 E5-2697v2 12C 2.7GHz, Aries	225984	3,14	4,88			

Parallel Systems



homogeneous multicore (BlueGene-Q chip)
heterogenous multicore
- general-purpose accelerator (e.g. Cell)
- GPU
- FPGA
- ASIC (e.g. Anton for MD)

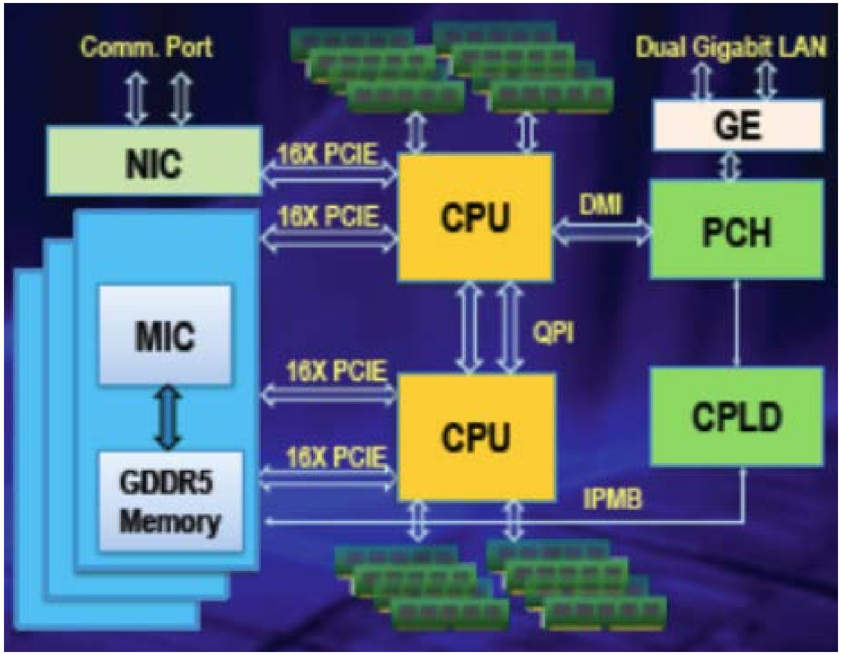
Network-on-chip (bus, ring, direct, …)

Tianhe-2

- Compute node with:
  - 2 IvyBridge Xeon sockets (12 cores each) with 88 GB memory
  - 3 Xeon Phi sockets with 8 GB memory for a total of 3.120.000 cores

- IvyBridge socket: 8 flops/cycle per core * 12 cores/socket * 2.2 GHz = 211.2 Gflop/s peak
- Xeon Phi socket: 16 flops/cycle per core * 57 cores/socket * 1.1 GHz = 1.003 Tflop/s peak

- On a node there are 2 IvyBridge * 0.2112 Tflop/s + 3 Phi * 1.003 Tflop/s → 3.431 Tflop/s per node



Evolution of the computing power of Supercomputers

FLOP/second (operaciones sobre números reales 64 bits)

E P T G

1000000000000000000

~2018
? (1x10[7] processadors

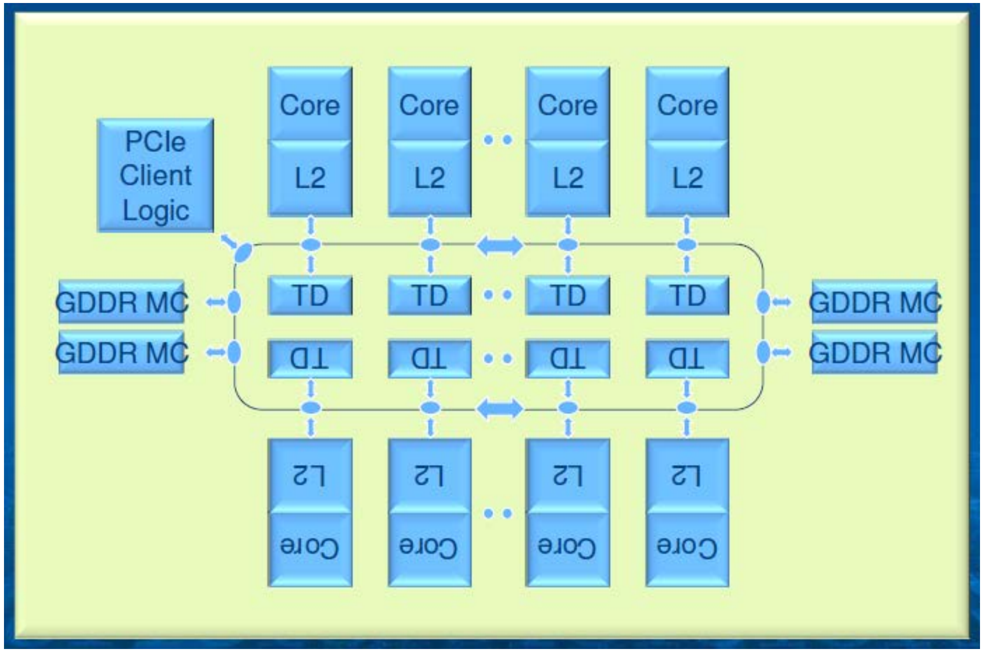
1988
Cray Y-MP (8 processadors)

2008
Cray XT5 (15000 processadors)

1998
Cray T3E (1024 processadors)

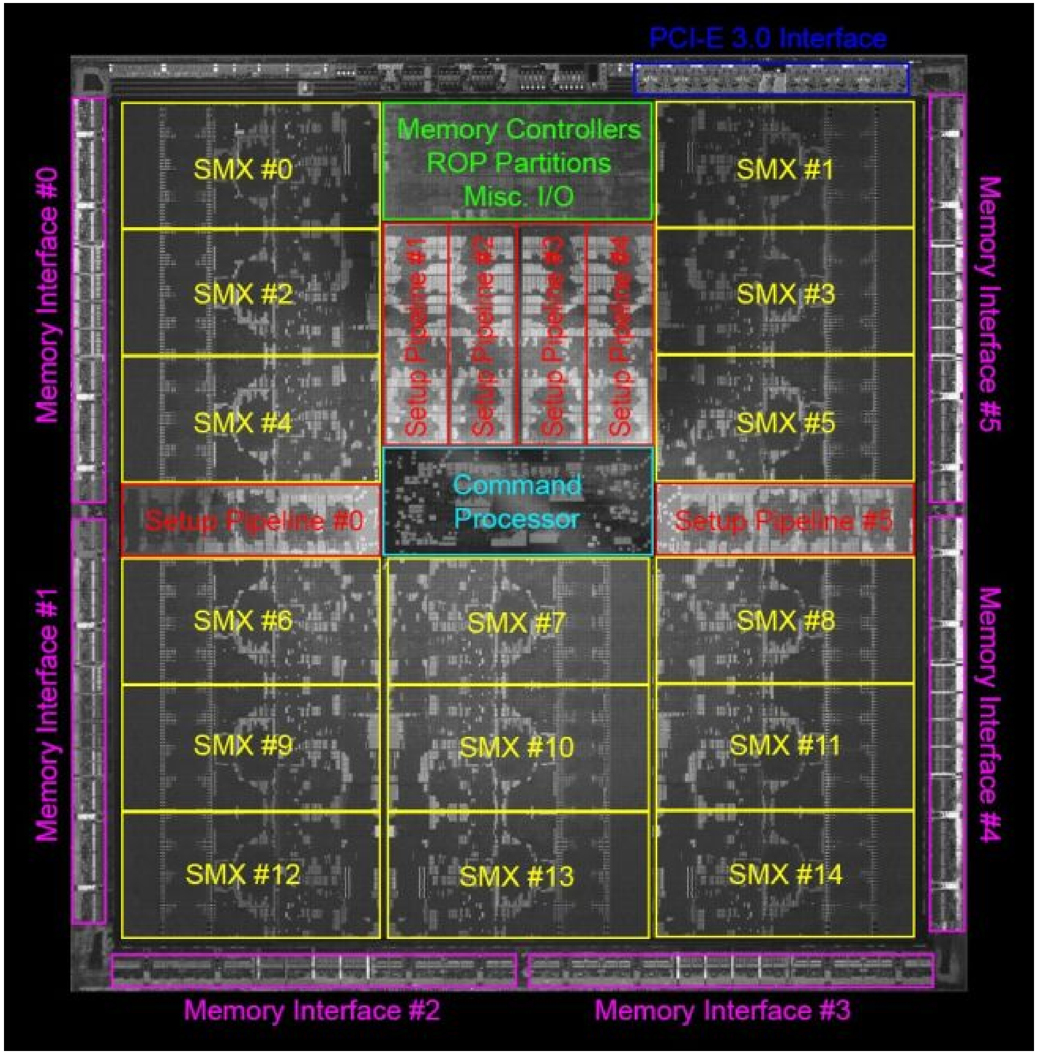
Intel Xeon Phi or Intel Many Integrated Core Architecture (MIC)

- 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

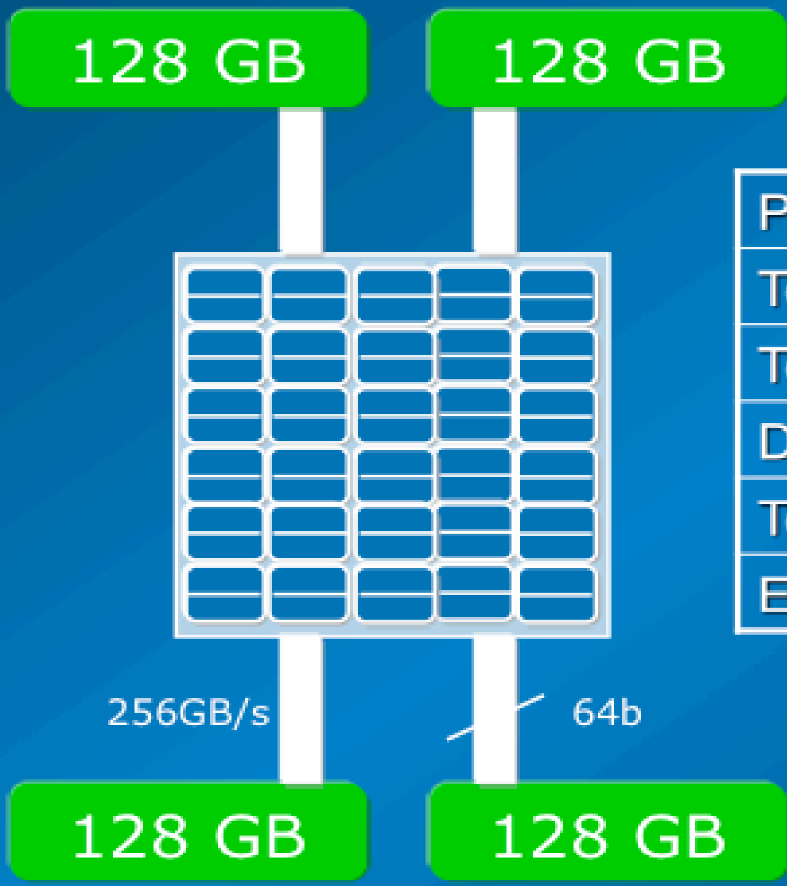


Accelerators: NVIDIA Kepler GK110 GPU (2014)

- 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



Processor node



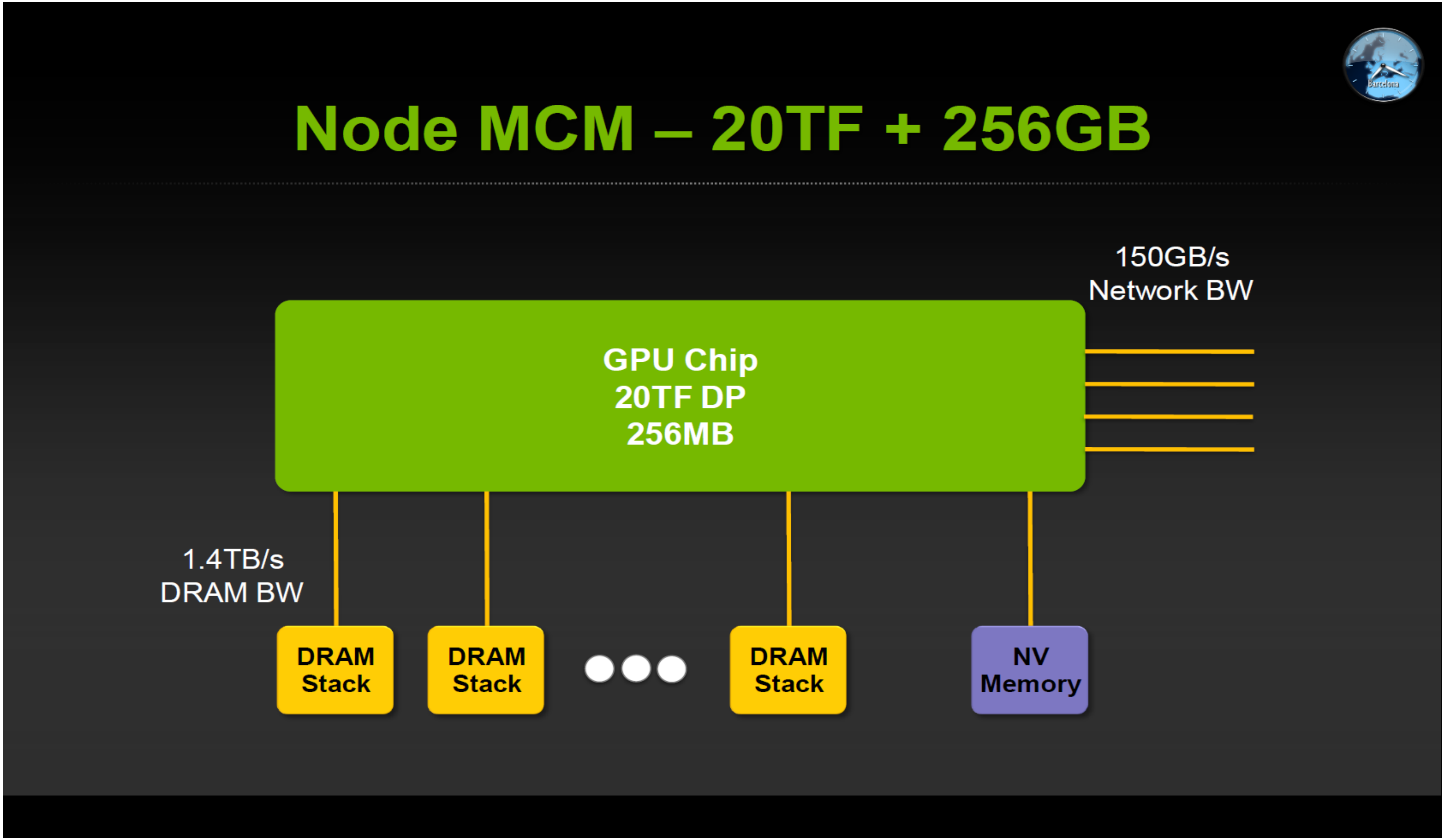
Peak performance	10.6 TF
Total DRAM Capacity	512GB
Total DRAM BW	1TB/s (0.1B/FLOP)
DRAM Power	800 W*
Total Power	1100 - 1400W
Energy efficiency	9.5 - 8 GF/Watt

110-140 MW for Exascale

*Assumes 5% Vdd scaling each technology generation
140 pJ energy consumed per accessed bit

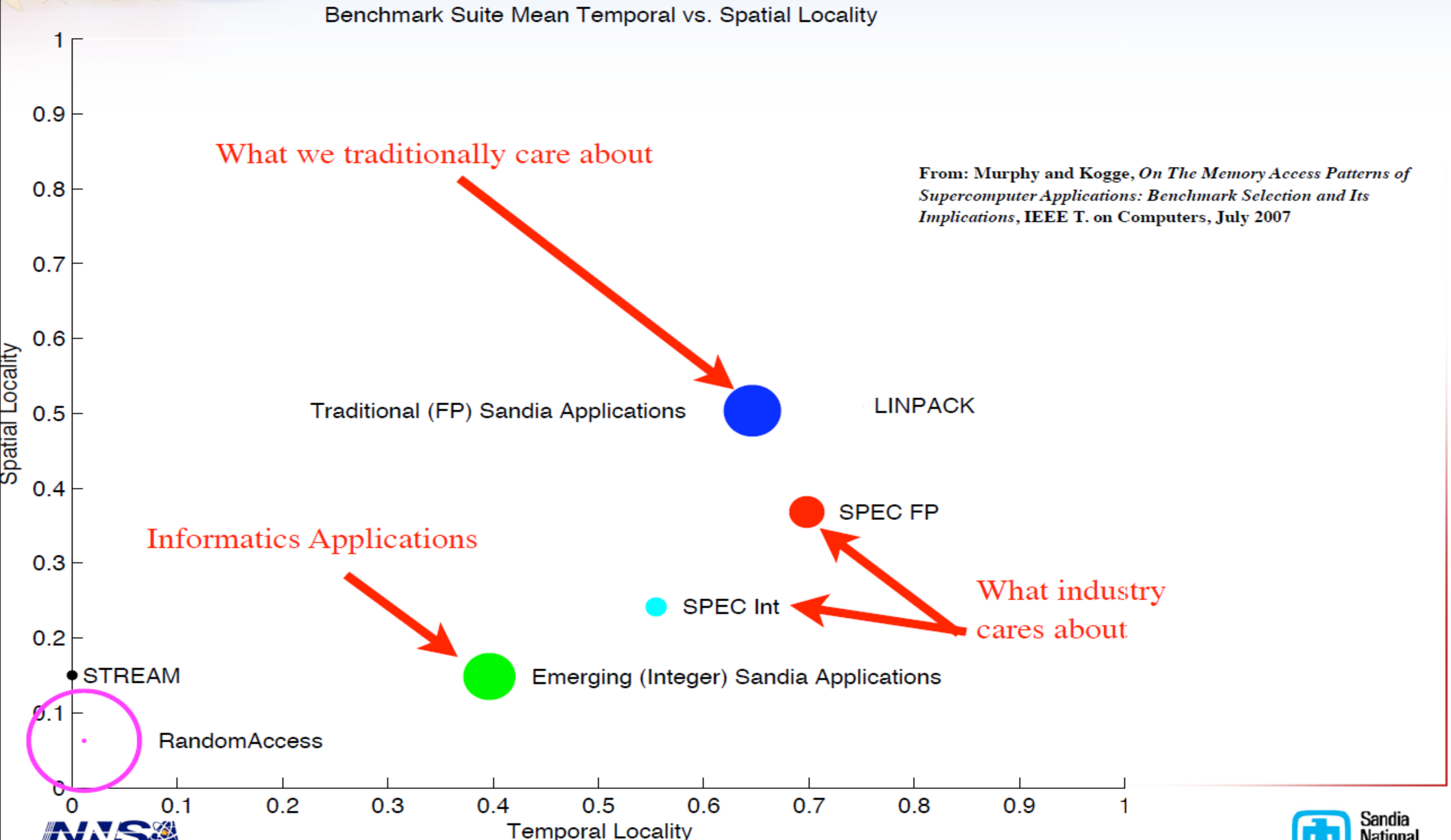
Thanks to S. Borkar, Intel

Nvidia: Node for the Exaflop Computer



Thanks Bill Dally

Not All Applications are Floating Point Oriented



From: Murphy and Kogge, *On The Memory Access Patterns of Supercomputer Applications: Benchmark Selection and Its Implications*, IEEE T. on Computers, July 2007

Graph 500 Project

- HPC benchmarks and performance metrics do not provide useful information on the suitability of supercomputing systems for data intensive applications.
- Graph 500 establishes a set of large-scale benchmarks for these applications.
- Steering Committee: 50 international HPC experts from academia and industry.
- Graph 500 is developing benchmarks to address three application kernels:
  - Concurrent search
  - Optimization (single source shortest path)
  - Edge-oriented (maximal independent set)
- Also they are addressing five graph-related business areas:
  - Cybersecurity
  - Medical Informatics
  - Data Enrichment
  - Social Networks
  - Symbolic Networks.

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.

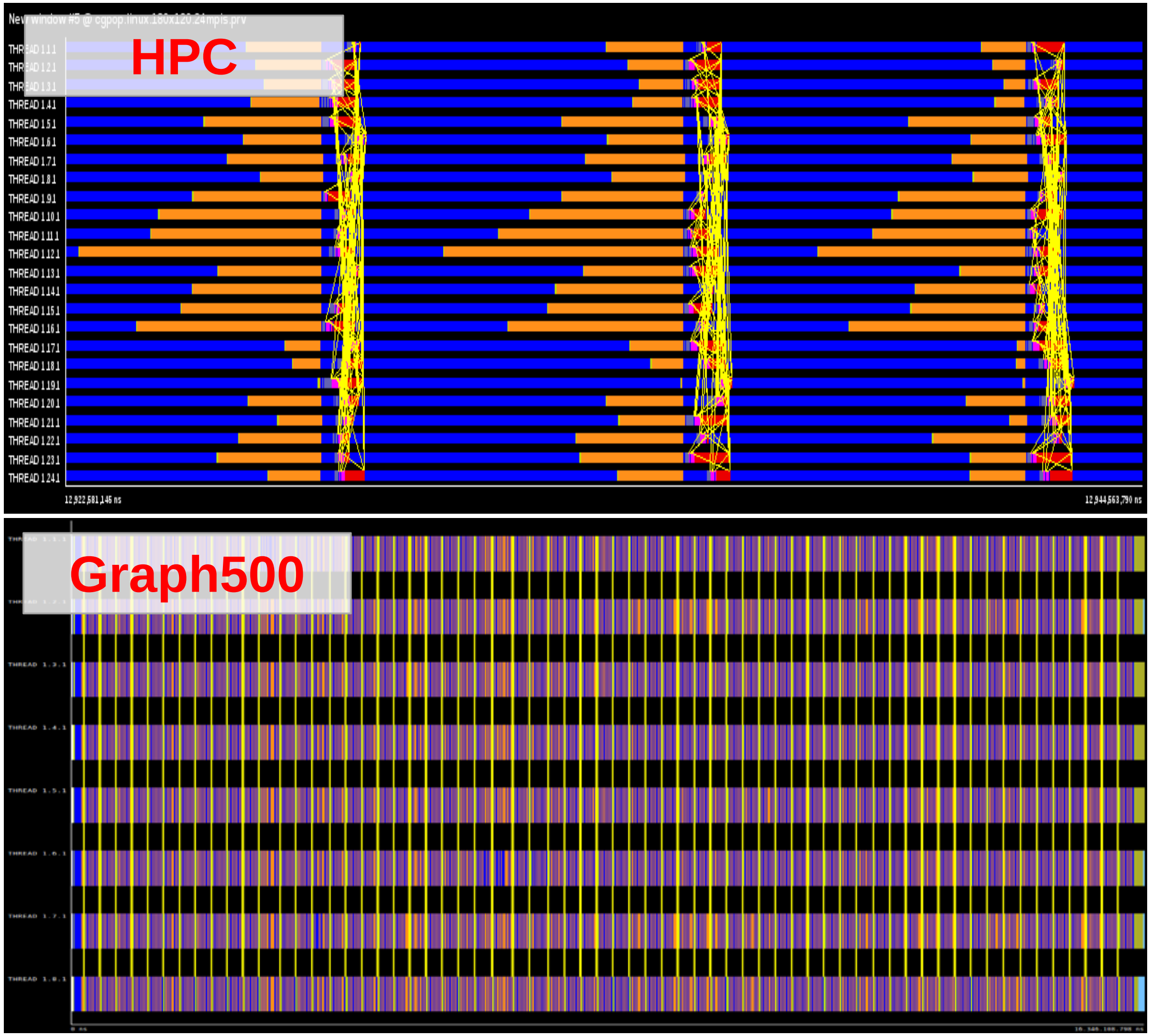
- Linpack:
  - computation bounded
  - focused on floating-point operations
  - Bulk-Synchronous-Parallel model: behavior based on big computation and short communication bursts
  - dense data structures highly organized and coalesced (spatial locality)

- 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

Communication pattern changes



Graph500 and Top500 Lists: June 2014

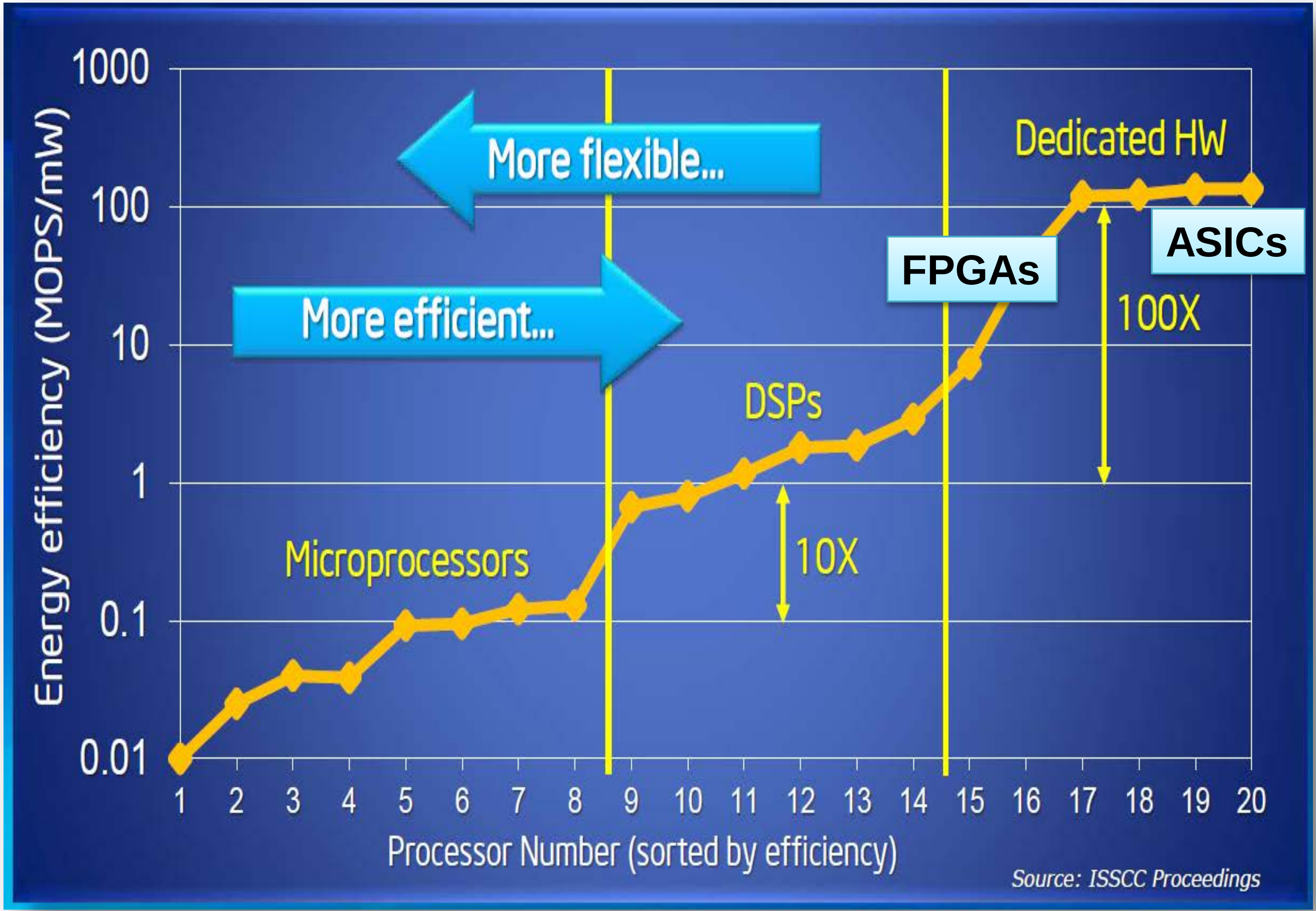
Graph500

#	Machine	Site	Cores	Problem size	GTEPS
1	K computer (Fujitsu - Custom supercomputer)	RIKEN Advanced Institute for Computational Science (AICS)	524,288	40	17,977.1
2	Sequoia (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Lawrence Livermore National Laboratory	1,048,576	40	16,599
3	Mira (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Argonne National Laboratory	786,432	40	14,328
4	JUQUEEN (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Forschungszentrum Juelich (FZJ)	262,144	38	5,848
5	Fermi (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	CINECA	131,072	37	2,567

Top500

#	Machine	Site	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)
1	Tianhe-2 (TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, Intel Xeon Phi 31S1P)	National Super Computer Center in Guangzhou	3,120,000	33,862.7	54,902.4
2	Titan (Cray XK7 , Opteron 6274 16C 2.200GHz, NVIDIA K20x)	Oak Ridge National Laboratory	560,640	17,590.0	27,112.5
3	Sequoia (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Lawrence Livermore National Laboratory	1,572,864	17,173.2	20,132.7
4	K computer (Fujitsu - SPARC64 VIIIfx 2.0GHz)	RIKEN Advanced Institute for Computational Science (AICS)	705,024	10,510.0	11,280.4
5	Mira (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Argonne National Laboratory	786,432	8,586.6	10,066.3

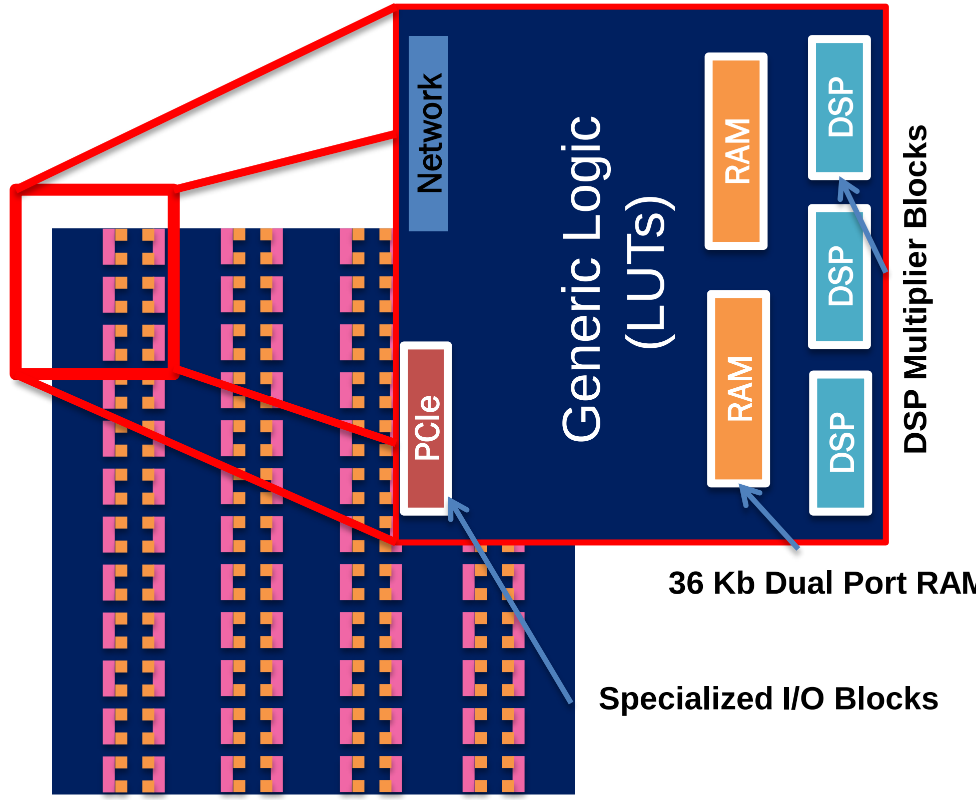
Specialization is Everything (?)



Source: Bob Broderson, Berkeley Wireless group

What are FPGAs?

- FPGA: Field Programmable Gate Array
- Generic sea of programmable logic and interconnect
- Program into specialized computing & networking circuits
- Can be reconfigured in as little as 100 ms



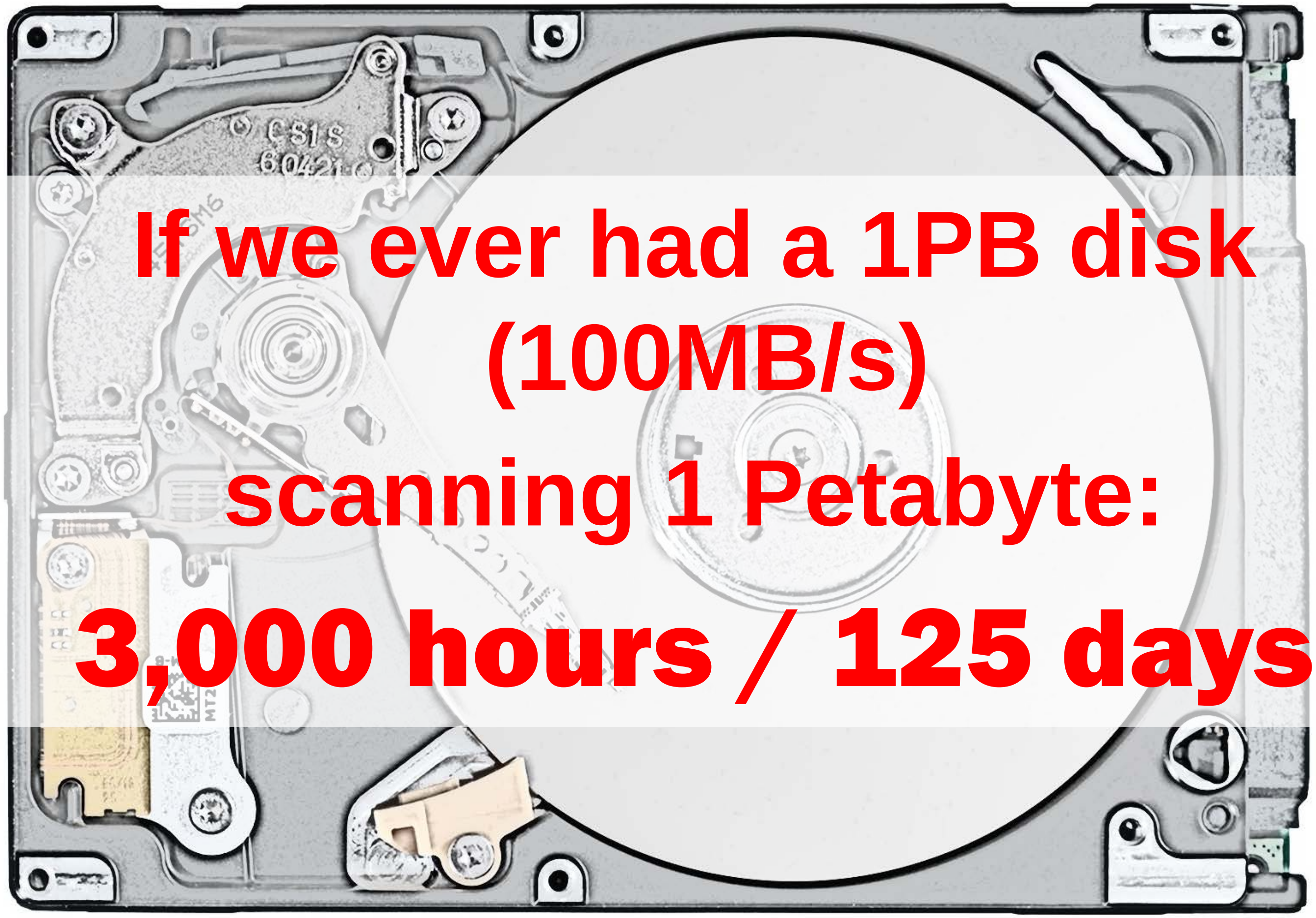
Outline

- The Big Data era: data generation explosion
- Storing data
- Processing data
- Where do we place data?
- How we use data?
- Research at BSC

BioInformatics, Big Data and Supercomputing

100.000 subjects suffering different diseases

What is the time required to retrieve information?



If we ever had a 1PB disk (100MB/s)

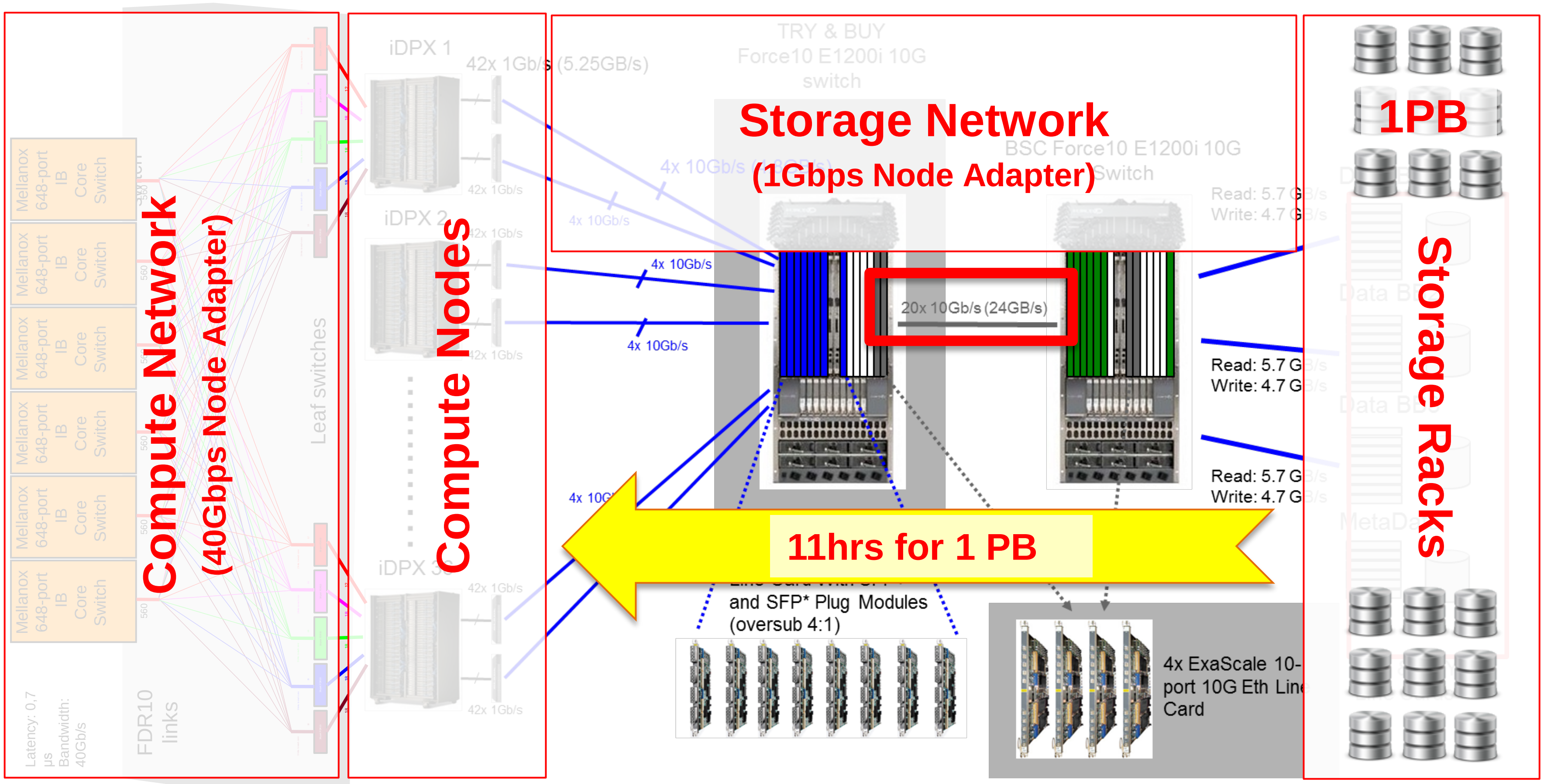
scanning 1 Petabyte:

3,000 hours / 125 days

Supercomputing is about doing things FAST…

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

Evolution of the storage architecture for Big Data



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

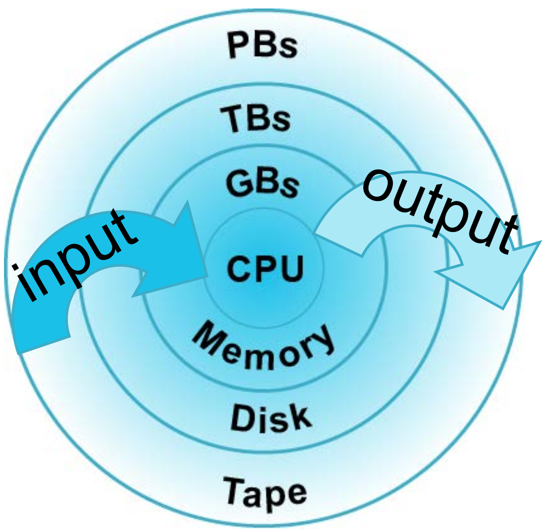
Operation	Energy (pJ)
64-bit integer operation	1
64-bit floating-point operation	20
256 bit on-die SRAM access	50
256 bit bus transfer (short)	26
256 bit bus transfer (1/2 die)	256
Off-die link (efficient)	500
256 bit bus transfer(across die)	1,000
DRAM read/write	16,000
HDD read/write	$\mathcal{O}(10^6)$

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

Paradigm shift

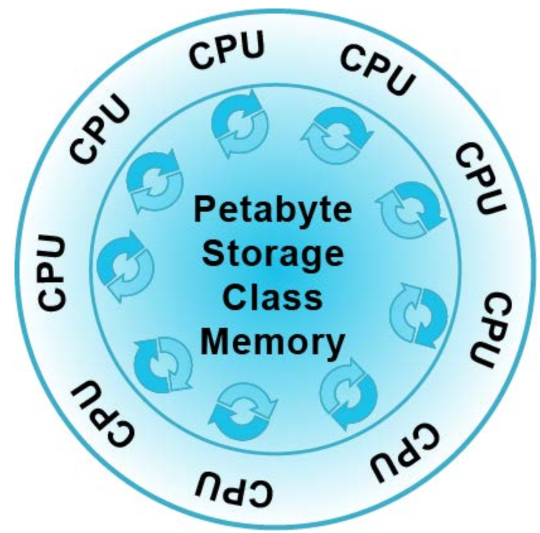
Old

Compute-centric Model

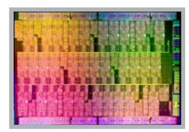


New

Data-centric Model



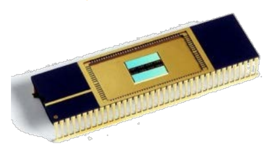
Manycore



Accelerators

Massive Parallelism
Persistent Memory

Flash



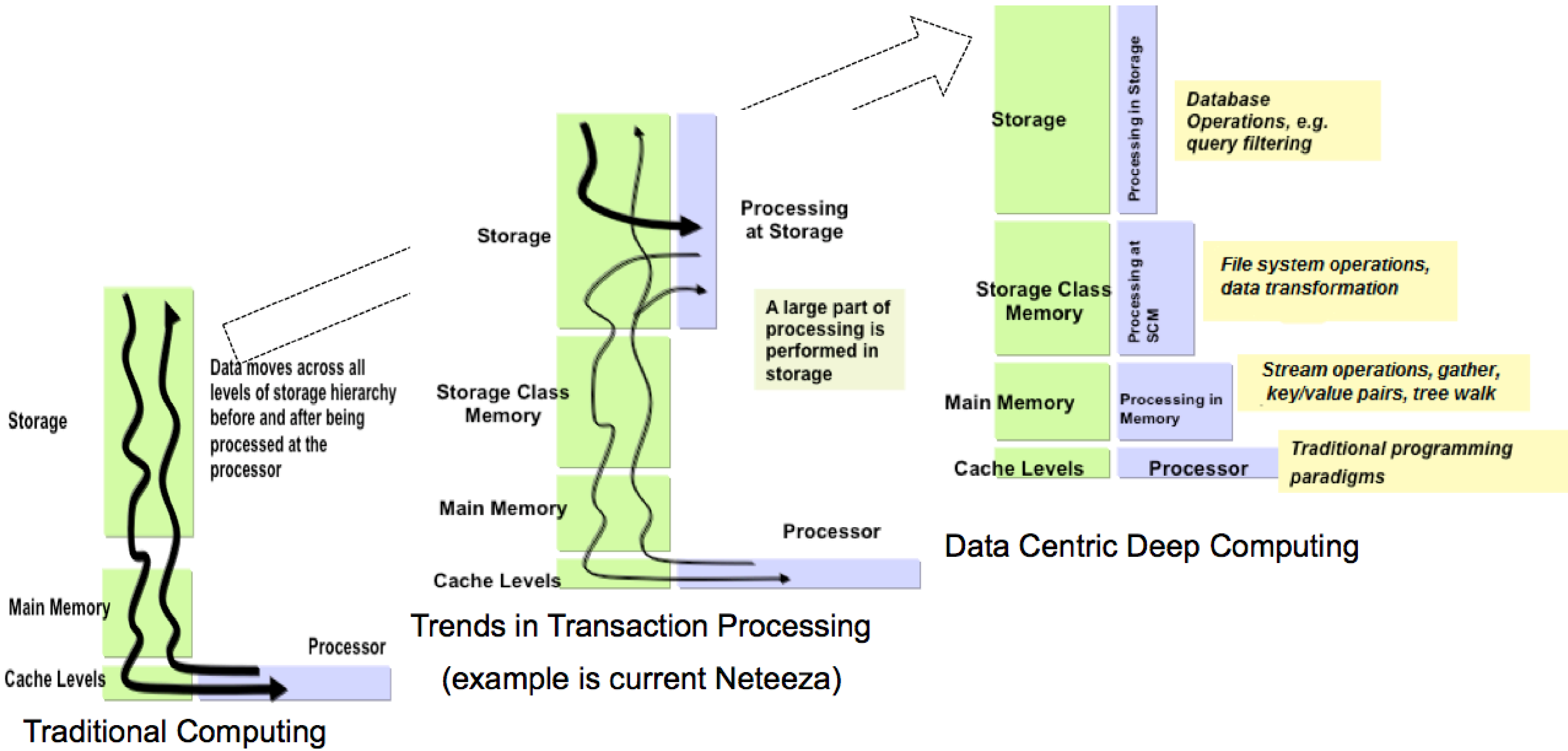
Storage Class Mem

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

Future HW for Big Data: Non-volatile memory

- **New technologies (STTMRAM, CBRAM, RRAM, …)**
- **More density**
- **Non-volatility**
  - good for throwing away the file abstraction, address non-volatile memory directly
- **Replacement for DRAMs**
  - Endurance problem
- **Large influence on software**
  - Data base systems, File systems

Optimized System Design for Data-Centric Deep Computing



Source: IBM Corporation, 2013

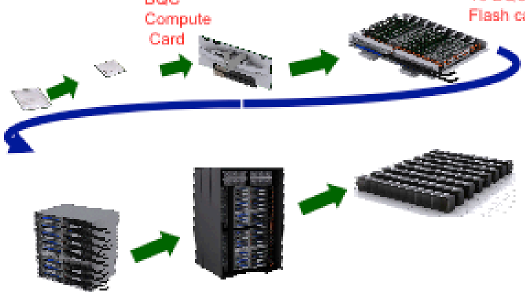
Solutions for Supercomputing

Blue Gene Active Storage (BGAS) Concept

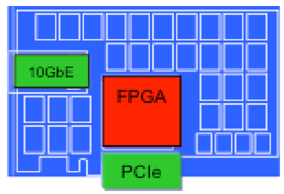
IBM

"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

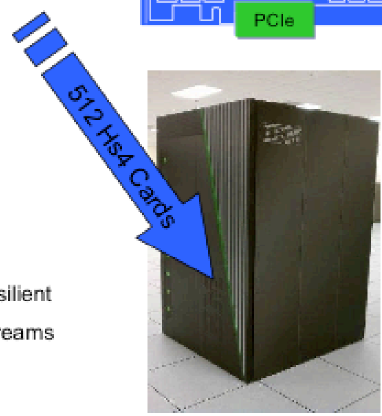
Standard BG/Q Compute Fabric



PCIe Flash Board



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

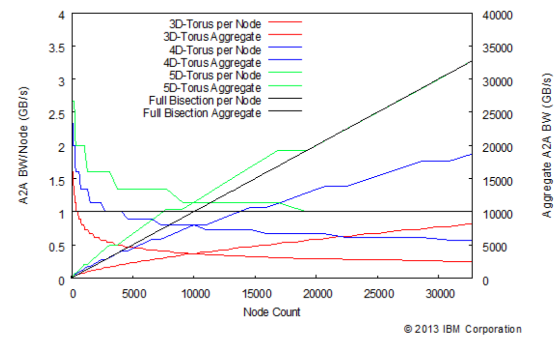
… scale it like BG/Q.

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

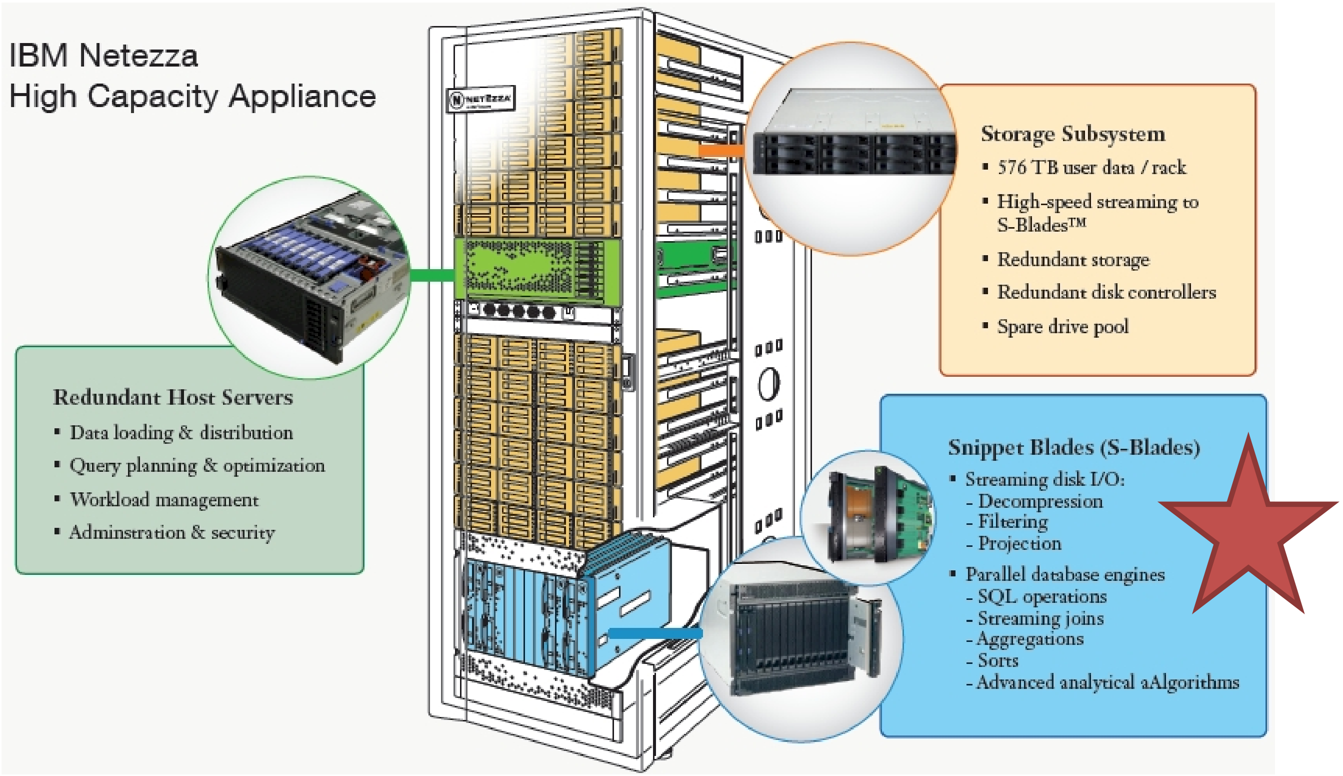


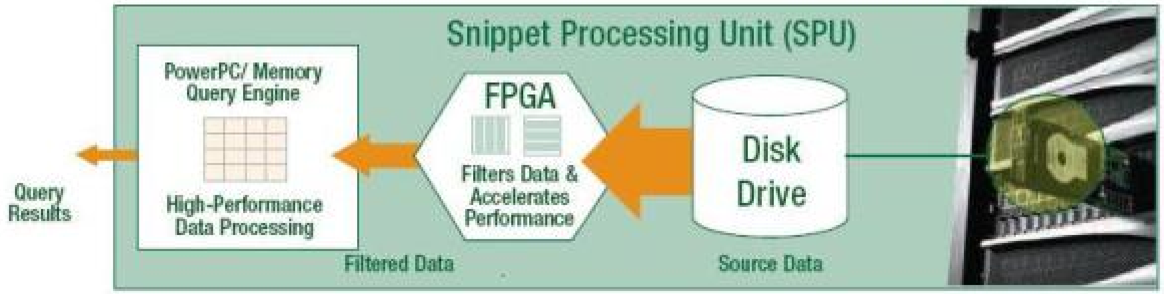
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

RoMoL Project

IBM Netezza Appliance (>50TB)





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.

FE: Feature Extraction

Query: “FPGA Configuration”

Document

FE: Feature Extraction

~4K Dynamic Features

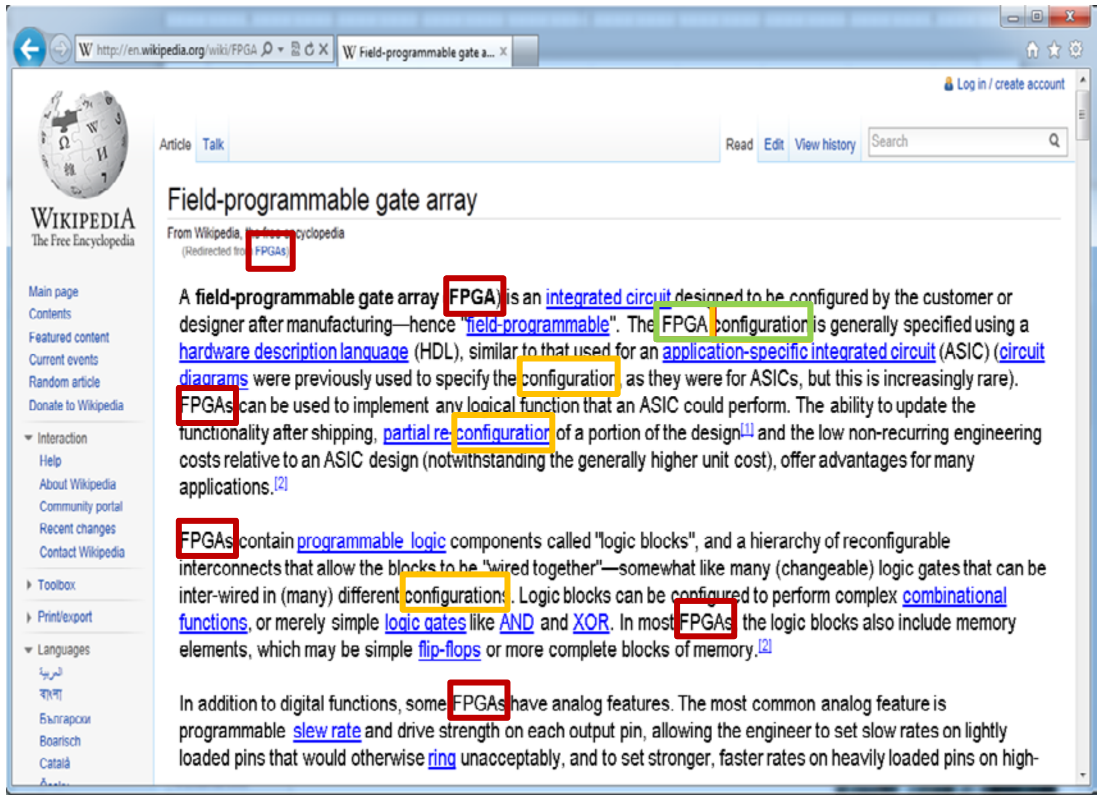
FFE: Free-Form Expressions

~2K Synthetic Features

MLS: Machine Learning Scoring

Score

NumberOfOccurrences_0 = 7 NumberOfOccurrences_1 = 4 NumberOfTuples_0_1 = 1


Field-programmable gate array

From Wikipedia, the free encyclopedia

(Redirected from FPGAs)

A **field-programmable gate array** (**FPGA**) is an integrated circuit designed to be configured by the customer or designer after manufacturing—hence "field-programmable". The FPGA configuration is generally specified using a hardware description language (HDL), similar to that used for an application-specific integrated circuit (ASIC) (circuit diagrams were previously used to specify the configuration, as they were for ASICs, but this is increasingly rare). FPGAs can be used to implement any logical function that an ASIC could perform. The ability to update the functionality after shipping, partial re-configuration of a portion of the design[1] and the low non-recurring engineering costs relative to an ASIC design (notwithstanding the generally higher unit cost), offer advantages for many applications.[2]

FPGAs contain programmable logic components called "logic blocks", and a hierarchy of reconfigurable interconnects that allow the blocks to be "wired together"—somewhat like many (changeable) logic gates that can be inter-wired in (many) different configurations. Logic blocks can be configured to perform complex combinational functions, or merely simple logic gates like AND and XOR. In most FPGAs, the logic blocks also include memory elements, which may be simple flip-flops or more complete blocks of memory.[2]

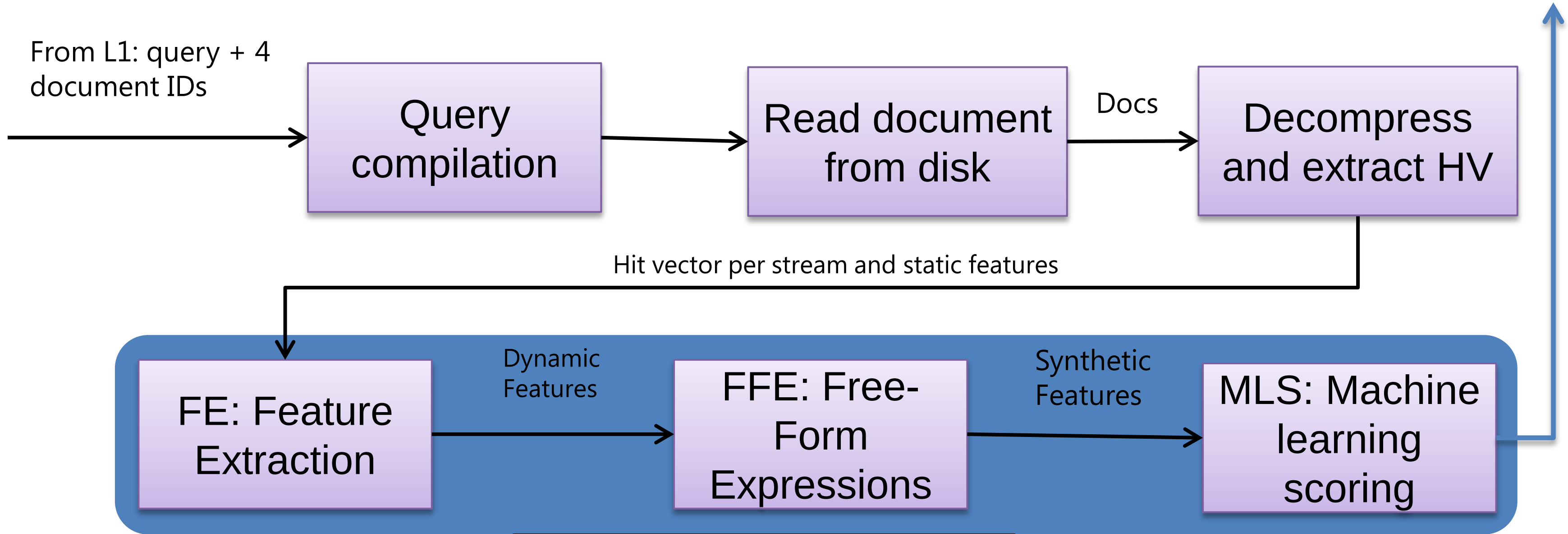
In addition to digital functions, some FPGAs have analog features. The most common analog feature is programmable slew rate and drive strength on each output pin, allowing the engineer to set slow rates on lightly loaded pins that would otherwise ring unacceptably, and to set stronger, faster rates on heavily loaded pins on high-

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

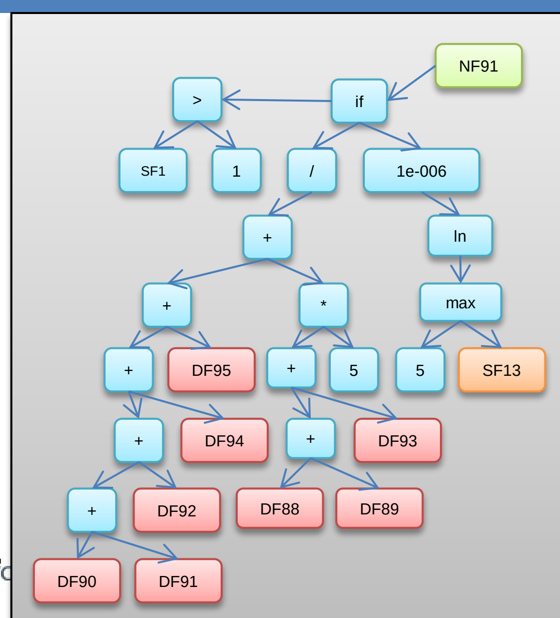
RoMoL Project

Bing RaaS Overview



	Position	Term
S0	5	3
S0	12	4
S0	99	2
S0	107	3
S0	109	3
S0	7	1
	42	**3**
S1	43	7

NumOccurrences_1_3 = 1

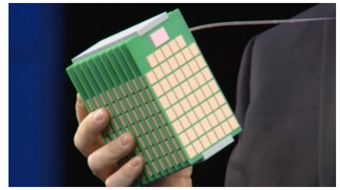


Classification algorithms produce score per document

HP unveils “The Machine”

June 11, 2014

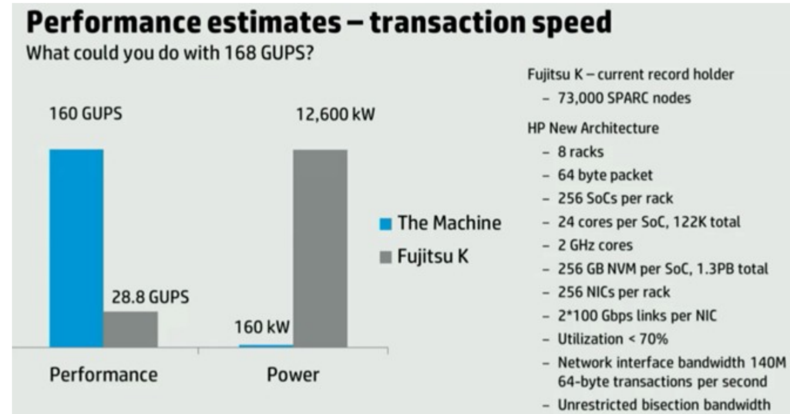
hp next

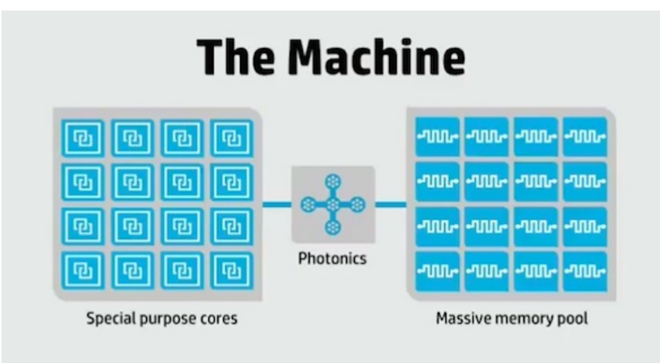


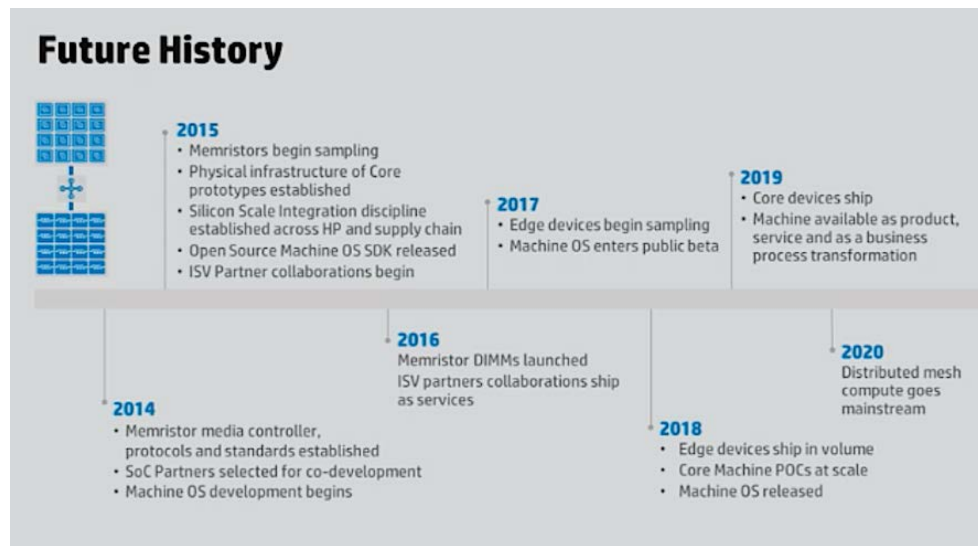
HP unveils “The Machine”

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.







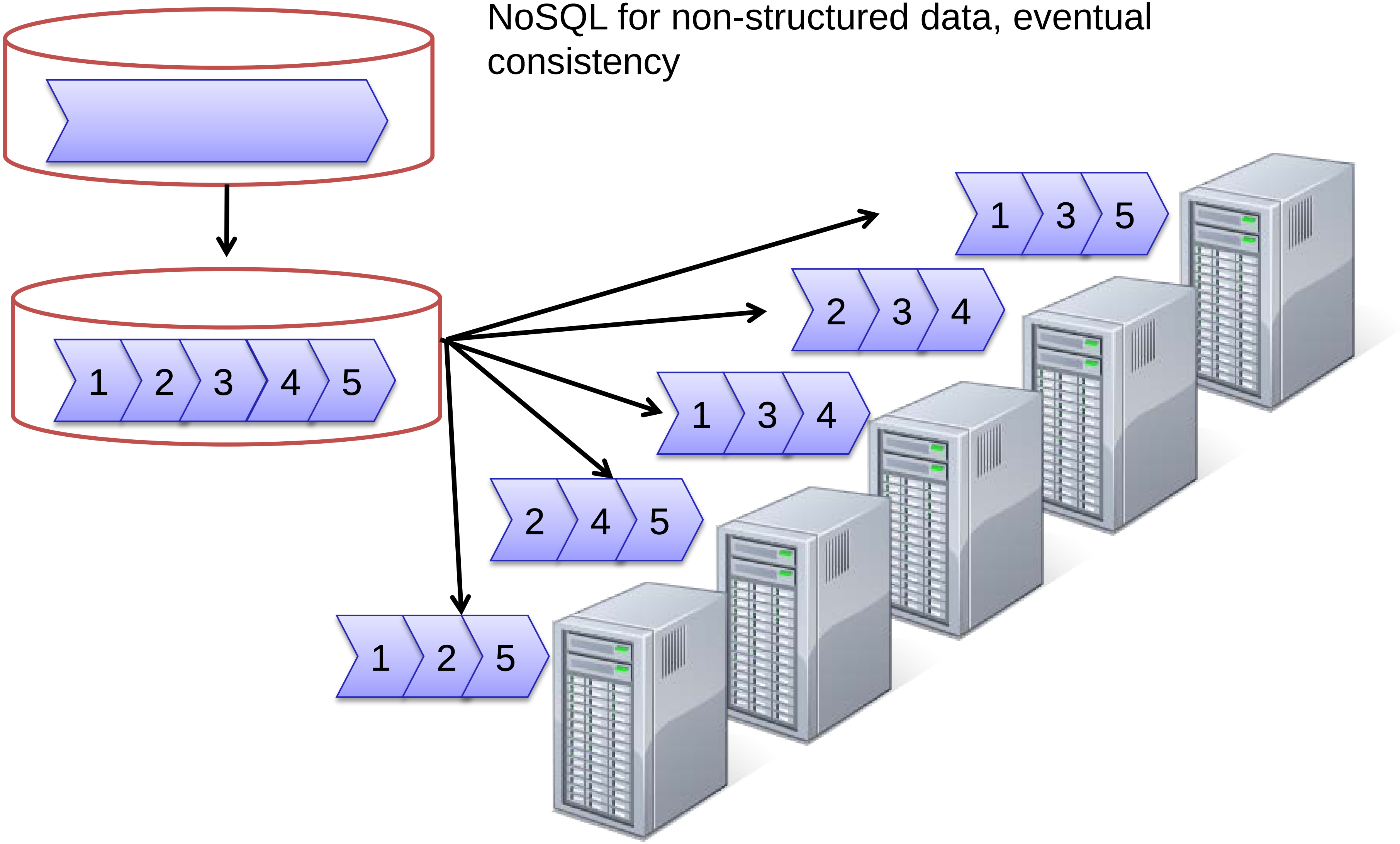
Source: http://www.engadget.com/2014/06/11/hp-the-machine/

Outline

- The Big Data era: data generation explosion
- Storing data
- Processing data
- Where do we place data?
- How we use data?
- Research at BSC

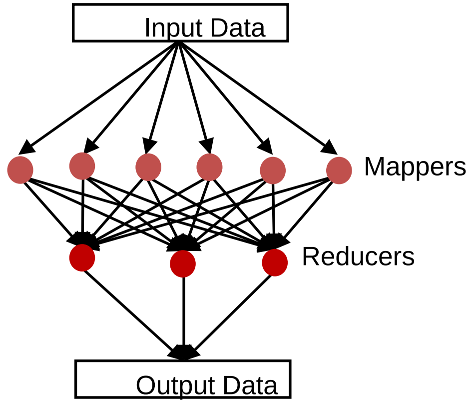
Relational databases sometimes not good for scale-out

NoSQL for non-structured data, eventual consistency



Programming model

- **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**



MapReduce: example

hamlet.txt

to be or not

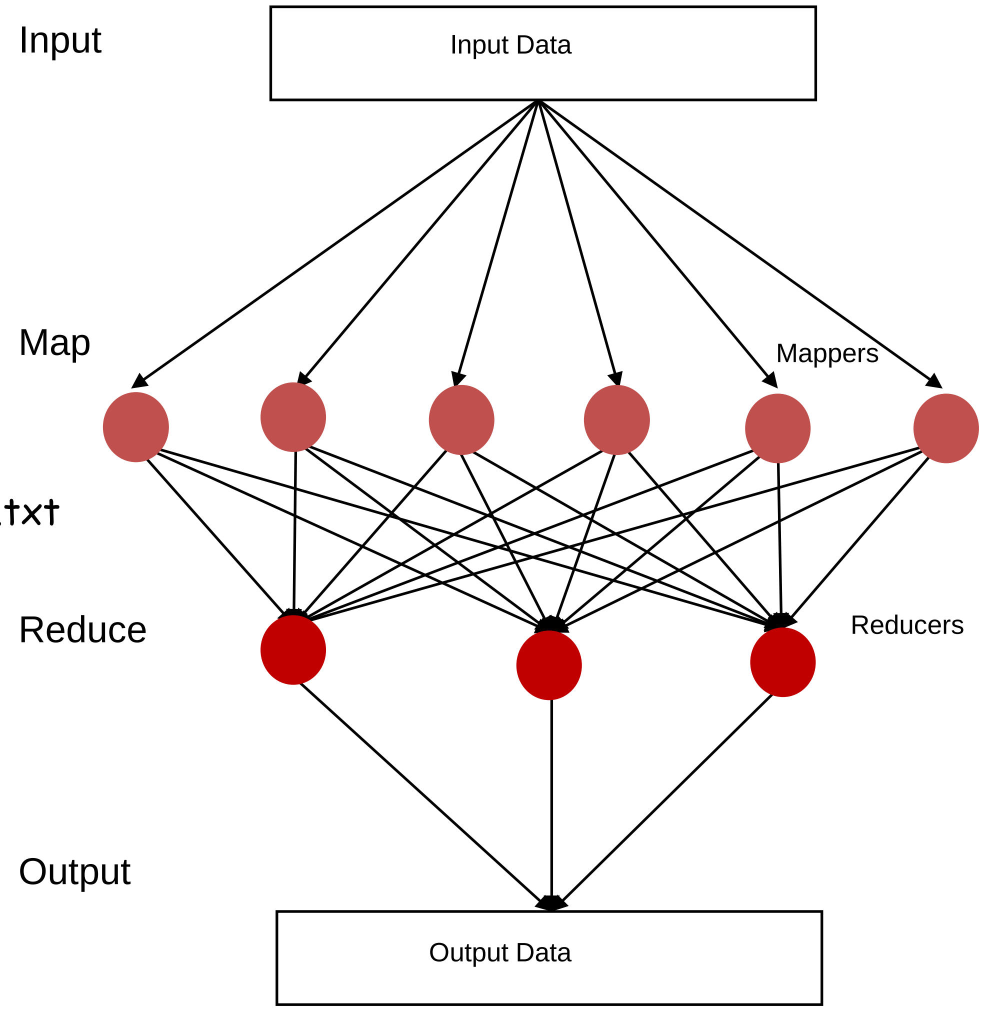
12th.txt

be not afraid
of greatness

to, hamlet.txt
be, hamlet.txt
or, hamlet.txt
not, hamlet.txt

be, 12th.txt
not, 12th.txt
afraid, 12th.txt
of, 12th.txt
greatness, 12th.txt

afraid, (12th.txt)
be, (12th.txt, hamlet.txt)
greatness, (12th.txt)
not, (12th.txt, hamlet.txt)
of, (12th.txt)
or, (hamlet.txt)
to, (hamlet.txt)

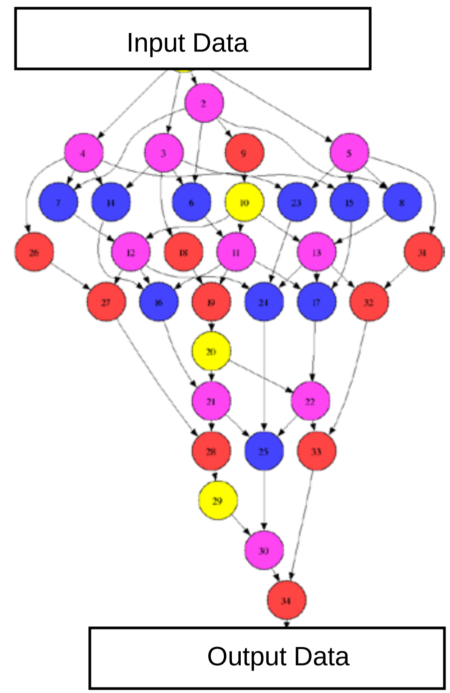


Limitations of MapReduce as a Programming model?

- MapReduce is great but not every one is a MapReduce expert

  "I am a python expert but ...."

- There is a class of algorithms that cannot be efficiently implemented with the MapReduce programming model



- Different programming models deal with different challenges
  - **pyCOMPSs from BSC**
  - **SPARK from Berkeley**

Google progressively dropping MapReduce

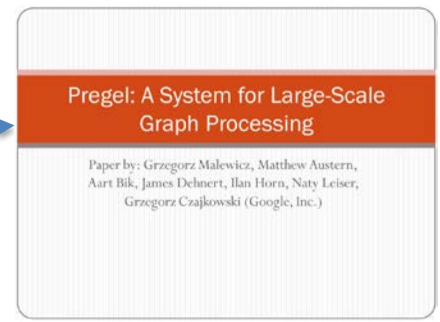
2006
The Apache Hadoop Project brings an opensource MapReduce Implementation, with the team of the Apache Nutch Crawler and the support of Yahoo!

Batch only

Graph Processing

Batch & Stream

2004
introduces
MapReduce


Pregel: A System for Large-Scale Graph Processing

Paper by: Grzegorz Malewicz, Matthew Austern, Aart Bik, James Dehnert, Ilan Horn, Naty Leiser, Grzegorz Czajkowski (Google, Inc.)

APACHE GIRAPH

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

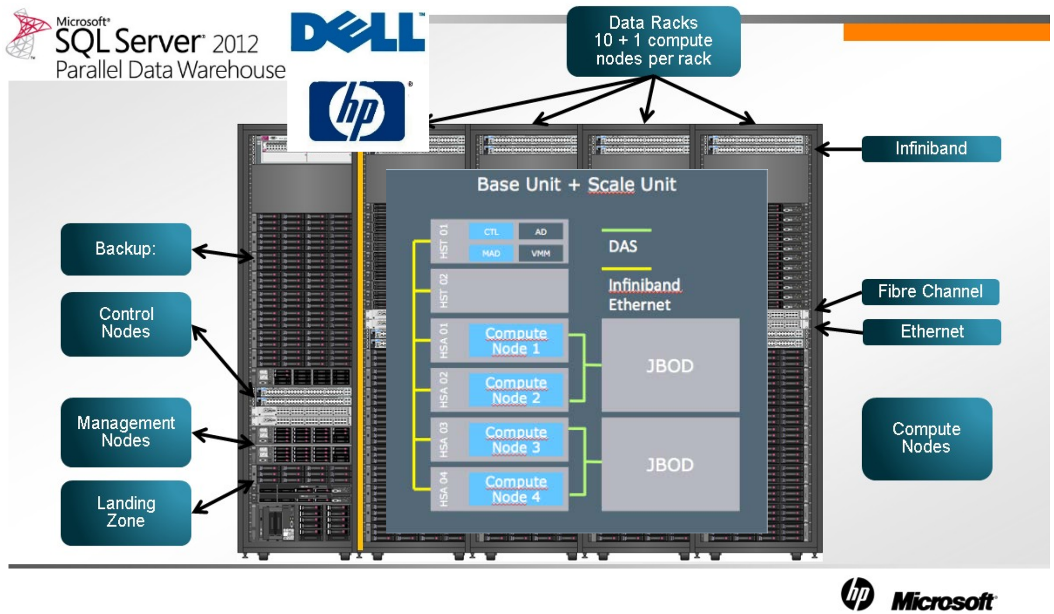
June 25, 2014
Google announces Cloud Dataflow: write code once, run it in batch or stream mode

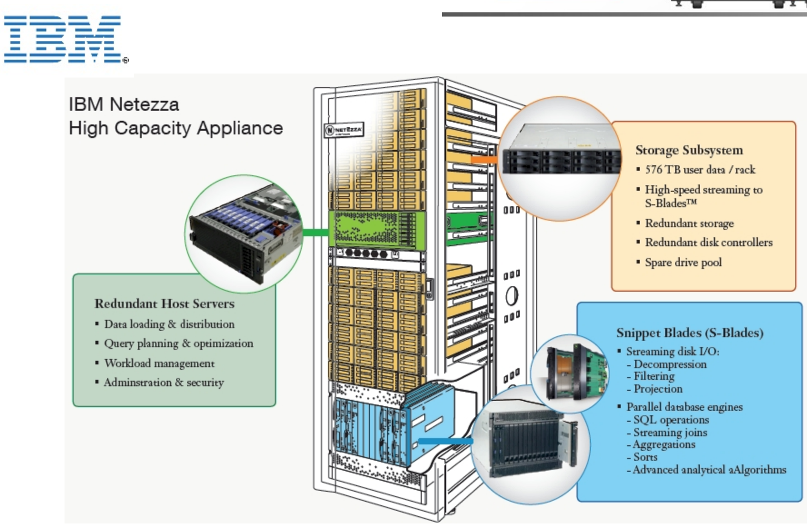
Cloud Dataflow is a managed service for creating data pipelines that ingest, transform and analyze data in **both batch and streaming modes**.

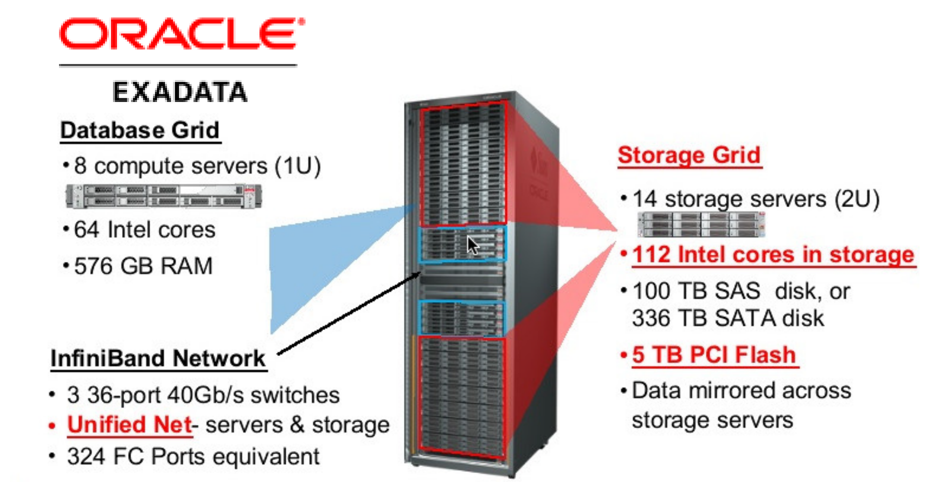
What happens if SW does not consider HW

- **Terasort contest: sorting 100TB data**
  - Number 1: Hadoop
    - 2100 nodes, 12 cores per node, 64 Gb per node
      - 24.000 cores
      - 134 Tb memory
    - Time: 4300 segs
    - Cost in Amazon: $ 8.800
  - Number 2: Tritonsort
    - 52 nodes, 8 cores per node, 24 Gb
      - 416 cores
      - 1,2 Tb memory
    - Time: 8300 segs and 6400 segs
    - Cost in Amazon: $ 294 and 226
  - Hadoop is easy to program, but needs 57X more cores, 100X more memory, and only gets 2X performance

HW-SW-Network co-design: Data Appliances

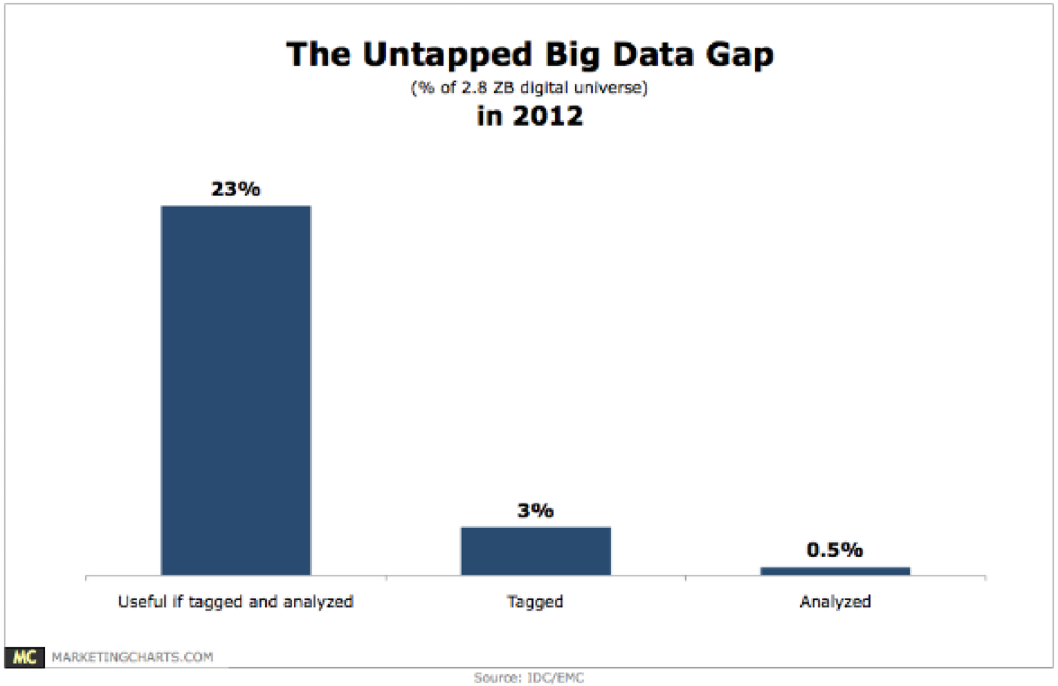






BigData needs Intelligence!

- Deep Learning: Building high-level abstractions
- Watson: Reasoning



Barcelona Supercomputing Center
Centro Nacional de Supercomputación
EXCELENCIA SEVERO OCHOA
RoMoL Project

Deep Learning: generating high level abstractions

- Toronto Face Database (TFD)
  - Automatic recognition of Facial Expressions

(b) Disgust

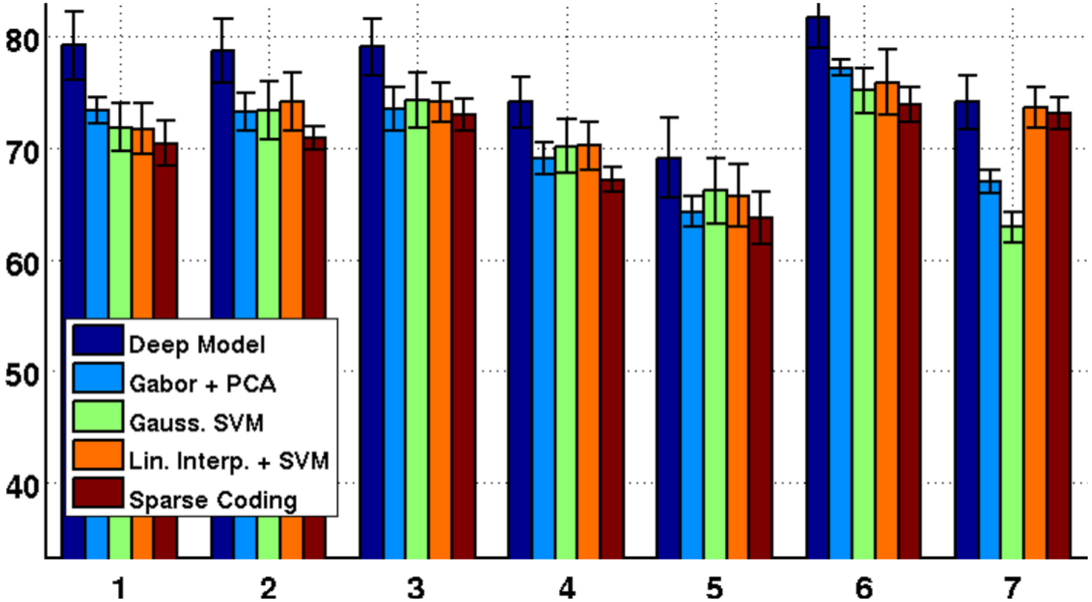
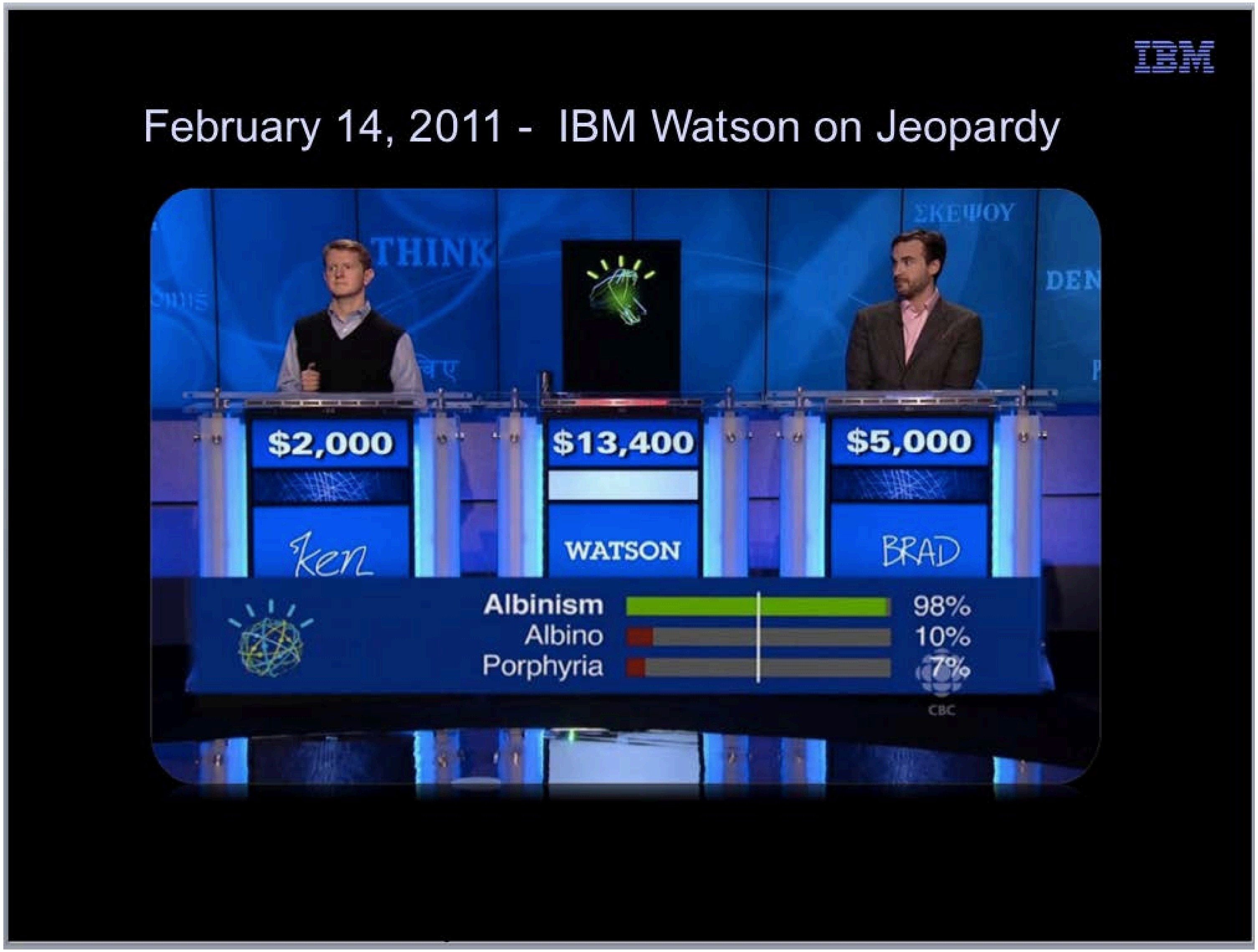


Fig. 15. Expression recognition accuracy on the TFD dataset when both training and test labeled images are subject to 7 types of occlusion.

Ranzato, M., Mnih, V., Susskind, J. and Hinton, G. E. Modeling Natural Images Using Gated MRFs. IEEE Trans. Pattern Analysis and Machine Intelligence, to appear.

Watson: Calculating Vs Learning



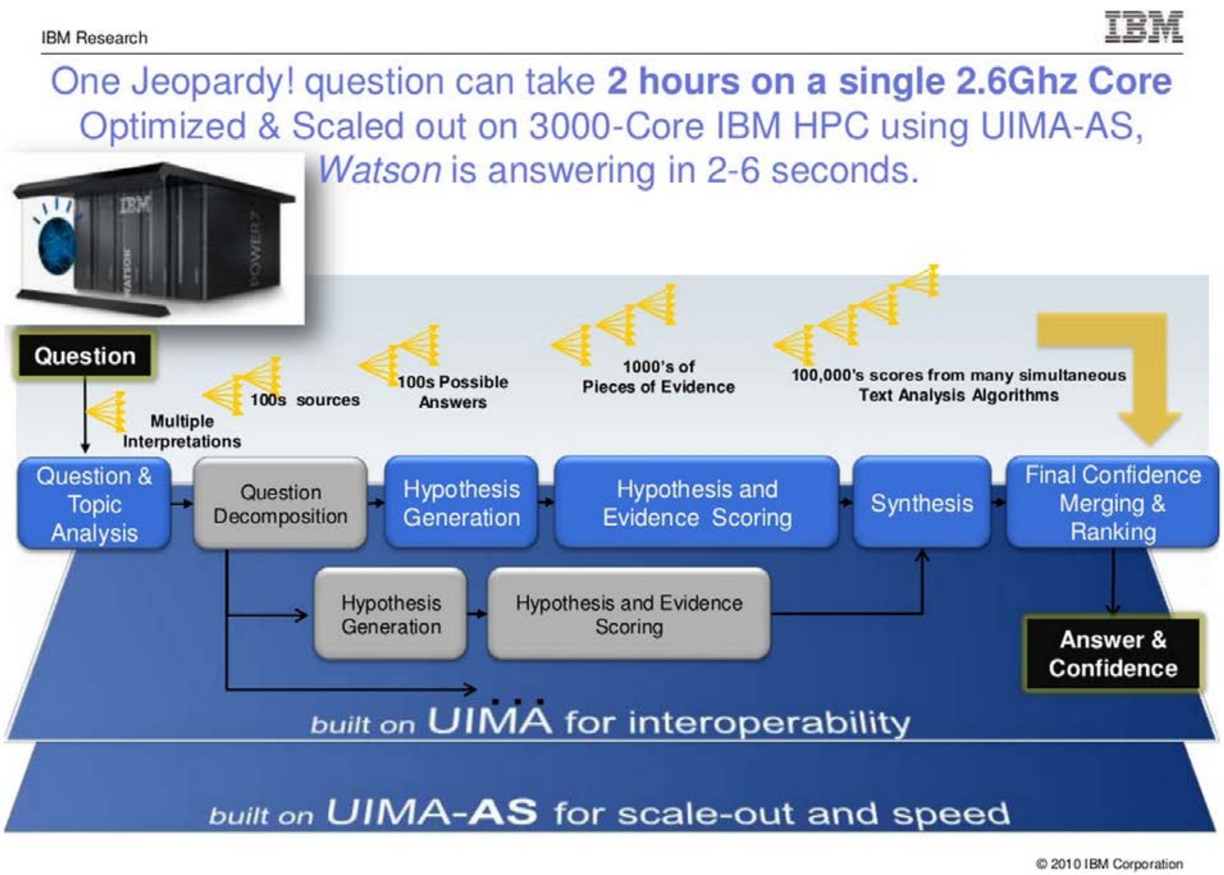
Source: Courtesy of IBM

63

What is behind Watson?

DeepQA: Massively Parallel Probabilistic Evidence-Based Architecture

For each question, generates and scores many hypotheses using a combination of 1000's of **Natural Language Processing, Information Retrieval, Machine Learning and Reasoning Algorithms**. These gather, evaluate, weigh and balance different types of evidence to deliver the answer with the best evidence support it can find



IBM Watson -- How to replicate Watson hardware and systems design for your own use in your basement
By Tony Pearson (ibm.co/Pearson)
http://tinyurl.com/pt6mdfu

To determine how big IBM Watson needed to be, the IBM Research team ran the DeepQA algorithm on a single core. It took 2 hours to answer a single Jeopardy question! Let's look at the performance data:

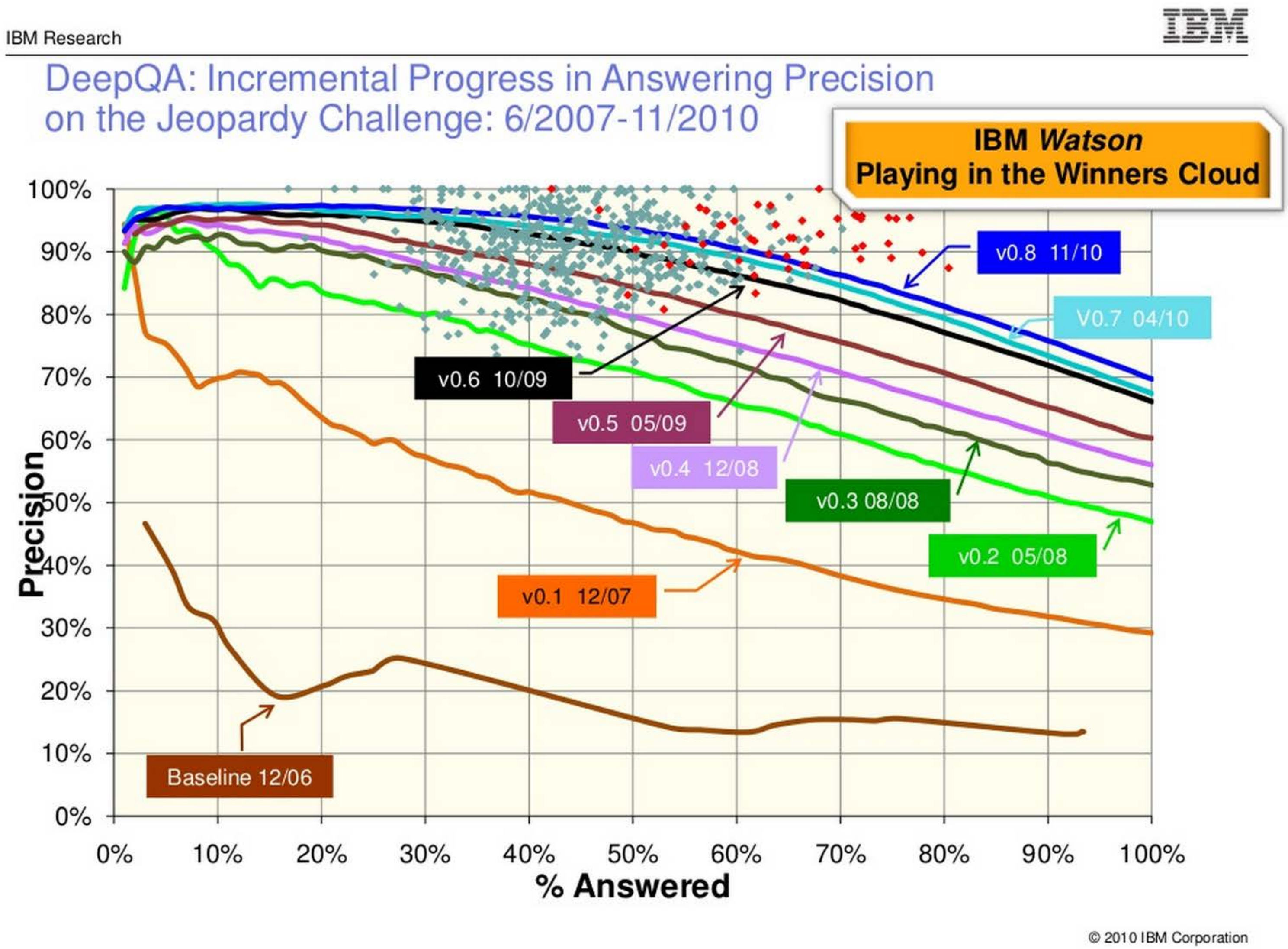
Element	Number of cores	Time to answer one Jeopardy question
Single core	1	2 hours
Single IBM Power750 server	32	< 4 minutes
Single rack (10 servers)	320	< 30 seconds
IBM Watson (90 servers)	2,880	< 3 seconds

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

RoMoL Project

Software tuning key to exploit hardware!



Outline

- The Big Data era: data generation explosion
- Storing data
- Processing data
- Where do we place data?
- How we use data?
- Research at BSC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-CNS objectives:

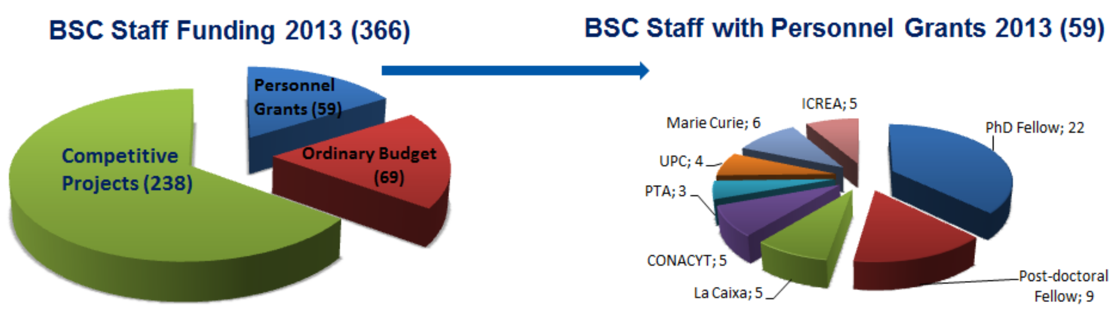
- R&D in Computer, Life, Earth and Engineering Sciences
- Supercomputing services and support to Spanish and European researchers

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

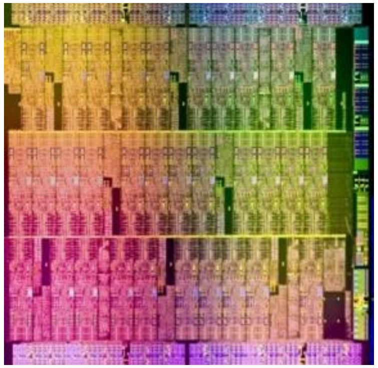
+400 people, 45 countries



In 2013, a total of 100 PhD students received top-up funding from the BSC

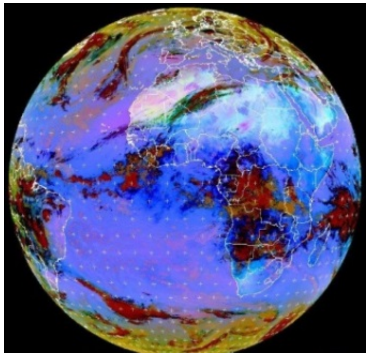
Mission of BSC Scientific Departments

COMPUTER SCIENCES



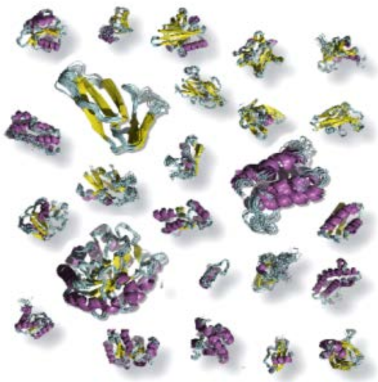
To influence the way machines are built, programmed and used: programming models, performance tools, Big Data, computer architecture, energy efficiency

EARTH SCIENCES



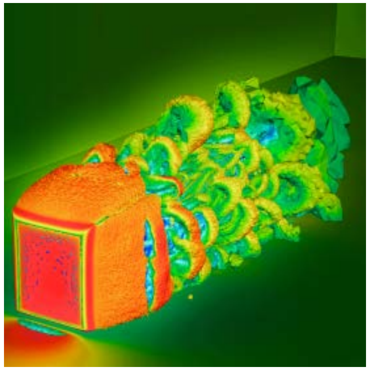
To develop and implement global and regional state-of-the-art models for short-term air quality forecast and long-term climate applications

LIFE SCIENCES



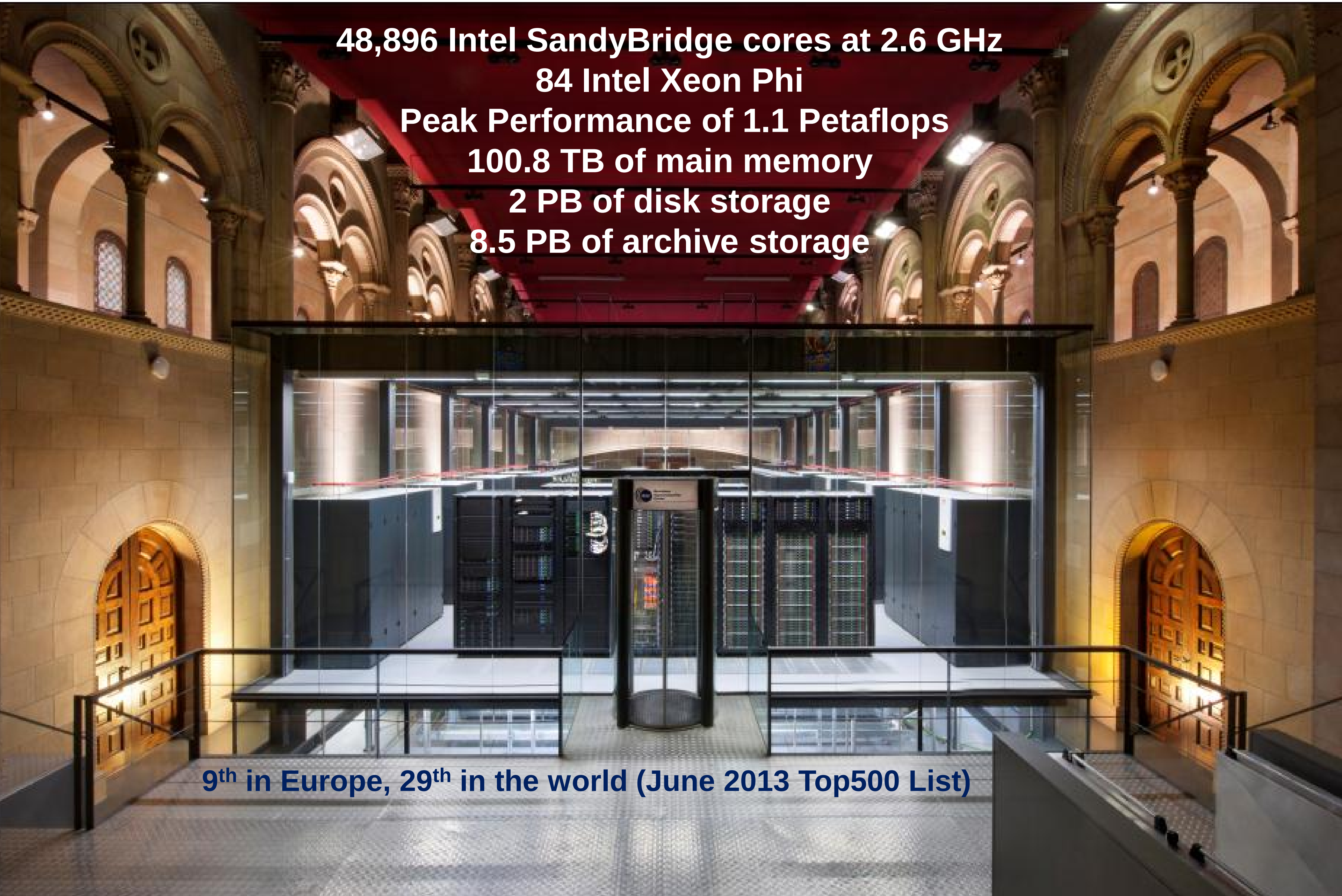
To understand living organisms by means of theoretical and computational methods (molecular modeling, genomics, proteomics)

CASE



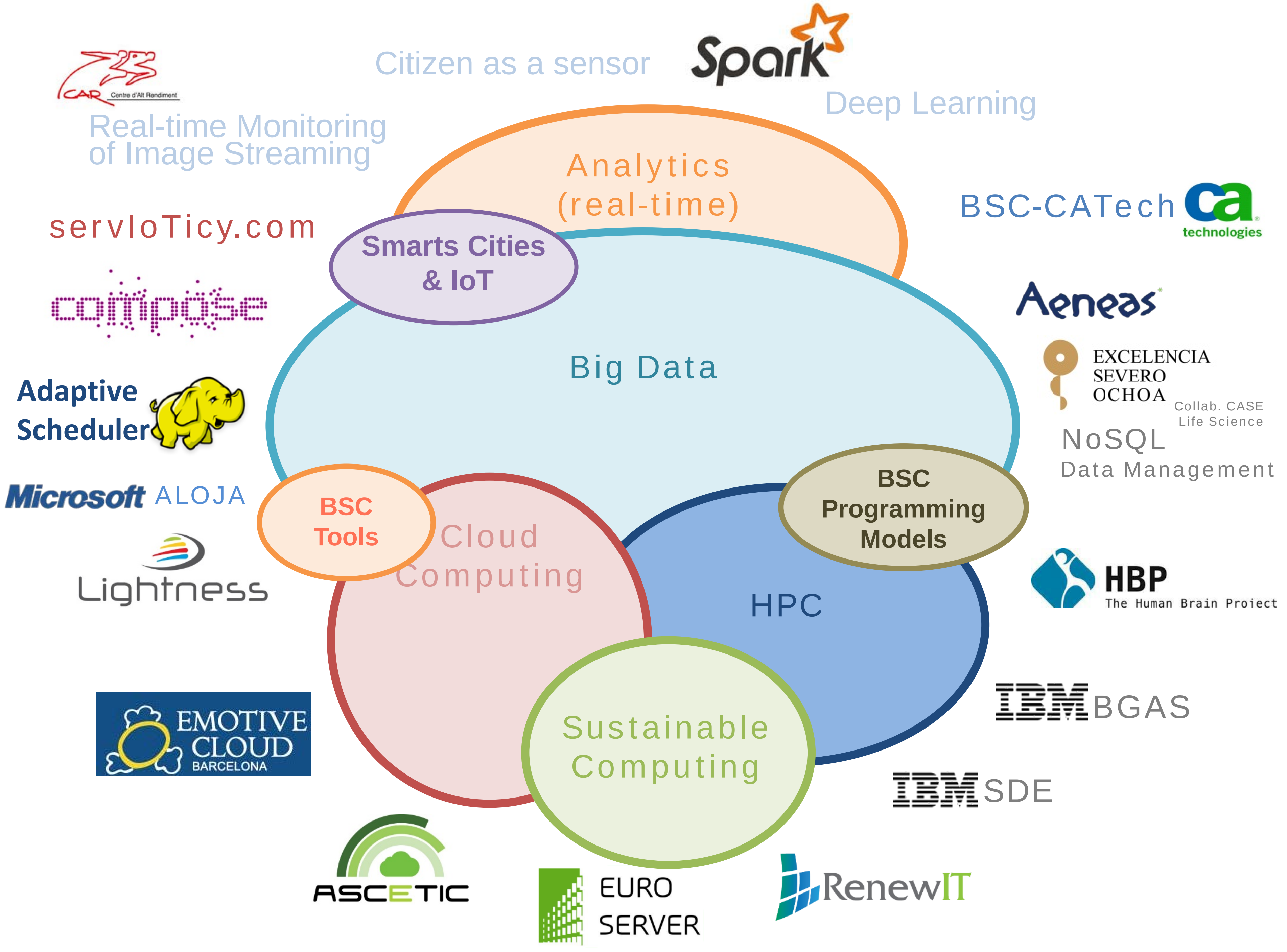
To develop scientific and engineering software to efficiently exploit super-computing capabilities (biomedical, geophysics, atmospheric, energy, social and economic simulations)

MareNostrum 3



48,896 Intel SandyBridge cores at 2.6 GHz
84 Intel Xeon Phi
Peak Performance of 1.1 Petaflops
100.8 TB of main memory
2 PB of disk storage
8.5 PB of archive storage

9th in Europe, 29th in the world (June 2013 Top500 List)



CAR Centre d'Alt Rendiment
Citizen as a sensor
Spark
Deep Learning
Real-time Monitoring of Image Streaming
Analytics (real-time)
BSC-CATech
ca technologies
servIoTicy.com
Smarts Cities & IoT
compose
Aeneas
Big Data
EXCELENCIA SEVERO OCHOA
Collab. CASE Life Science
Adaptive Scheduler
NoSQL
Data Management
Microsoft ALOJA
BSC Tools
Cloud Computing
BSC Programming Models
Lightness
HPC
HBP
The Human Brain Project
IBM BGAS
EMOTIVE CLOUD BARCELONA
Sustainable Computing
IBM SDE
ASCETIC
EURO SERVER
RenewIT

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank you !