Barcelona Supercomputing Center
Centro Nacional de Supercomputación

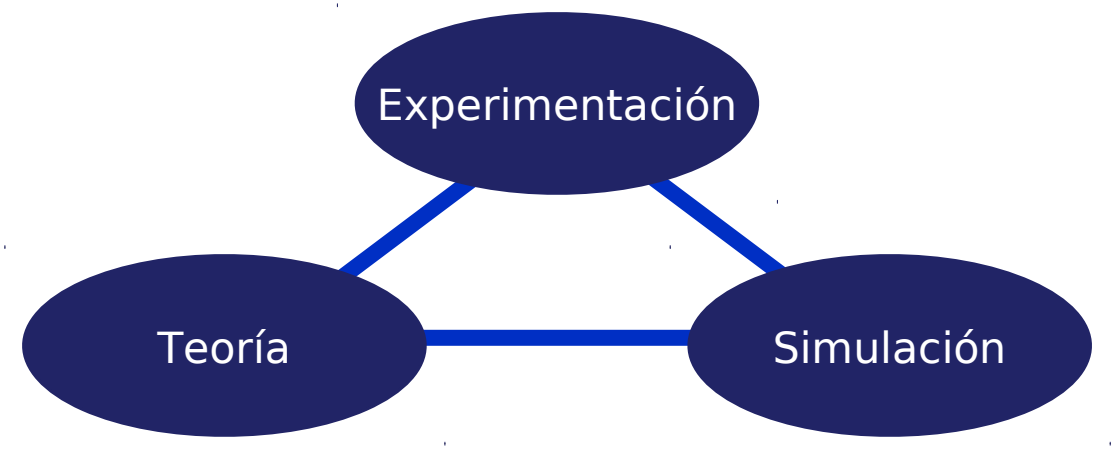
“Futuros supercomputadores Exaflop”

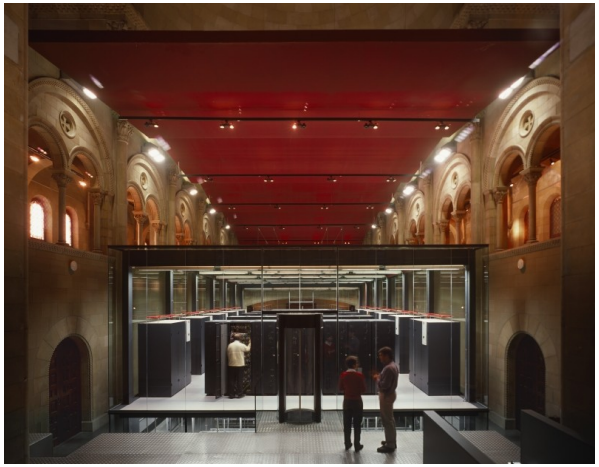
La Supercomputación al Servicio de Investigadores e Innovadores

Mérida, 27 Abril, 2010

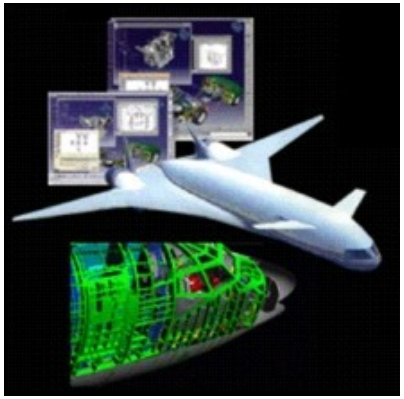
Prof. Mateo Valero
Director

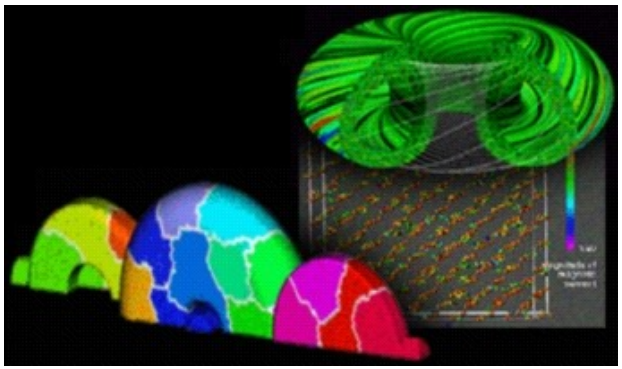
¿Cómo avanza la ciencia hoy?

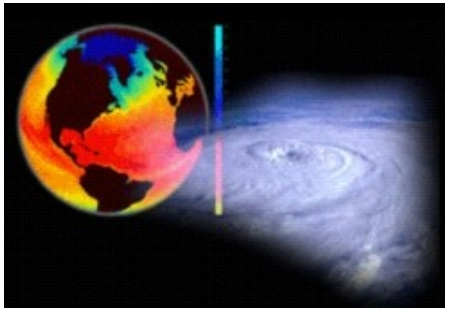




Simulación = Calcular las fórmulas de la teoría



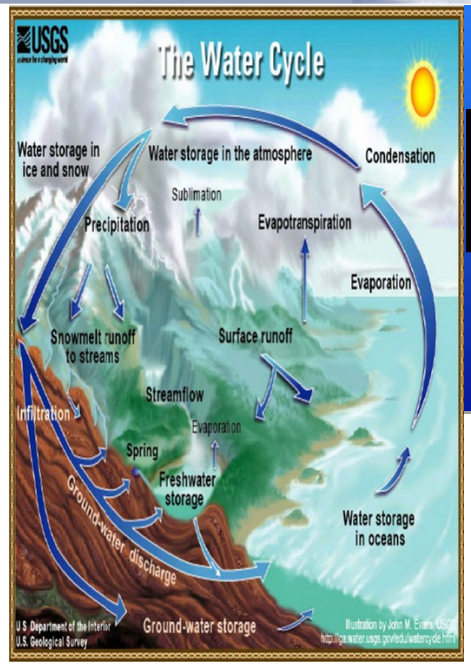


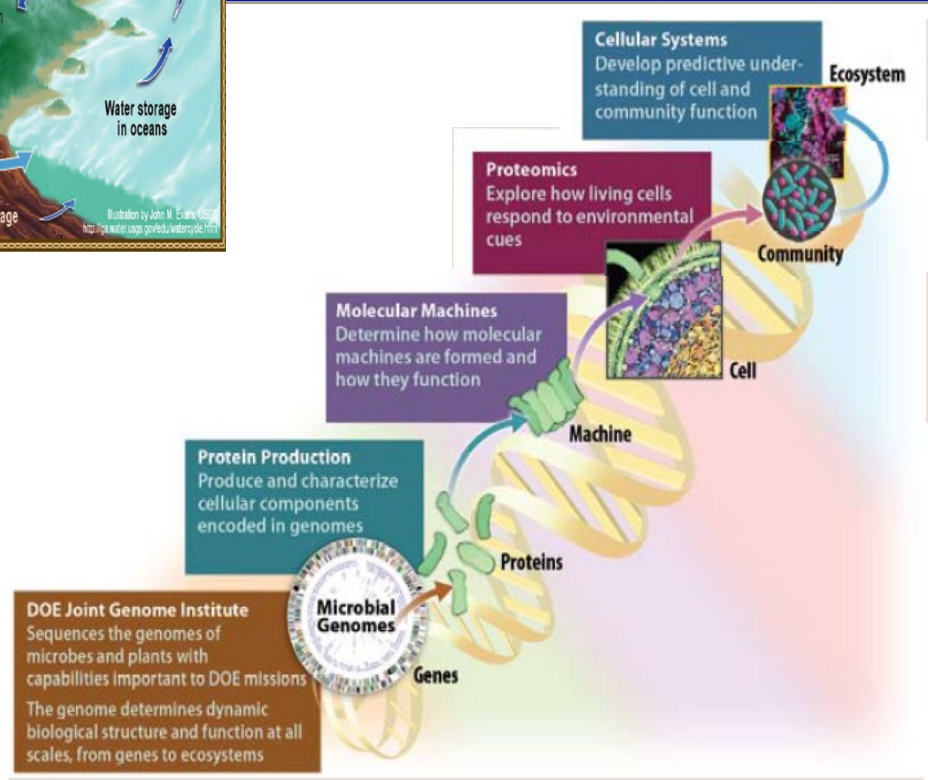


CARO PELIGROSO IMPOSIBLE

Grand Challenge problems

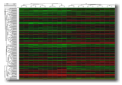
- Systems biology –
  - Model & simulation leading to predictive models with clinical or environmental impact
- Sustainable Systems –
  - Taking into account multi-scale nature - Models are linked to experimental data – providing corroboration of experiments
- Turbulence & Chaos –
  - Characterize boundary layer effects and their impact on global solution and stability
- Environmental
  - Global Warming/Climate Change
  - Energy
  - Water
  - Biodiversity and land use
  - Chemicals, toxics and heavy metals
  - Air pollution
  - Waste management
  - Stratospheric ozone depletion
  - Oceans & fisheries
  - Deforestation



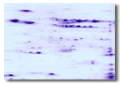


Multi-Scale Patient-Specific Data

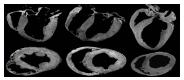
Genetic Variability



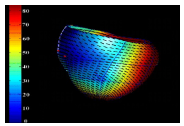
Gene Expression Profiling



Protein Expression Profiling

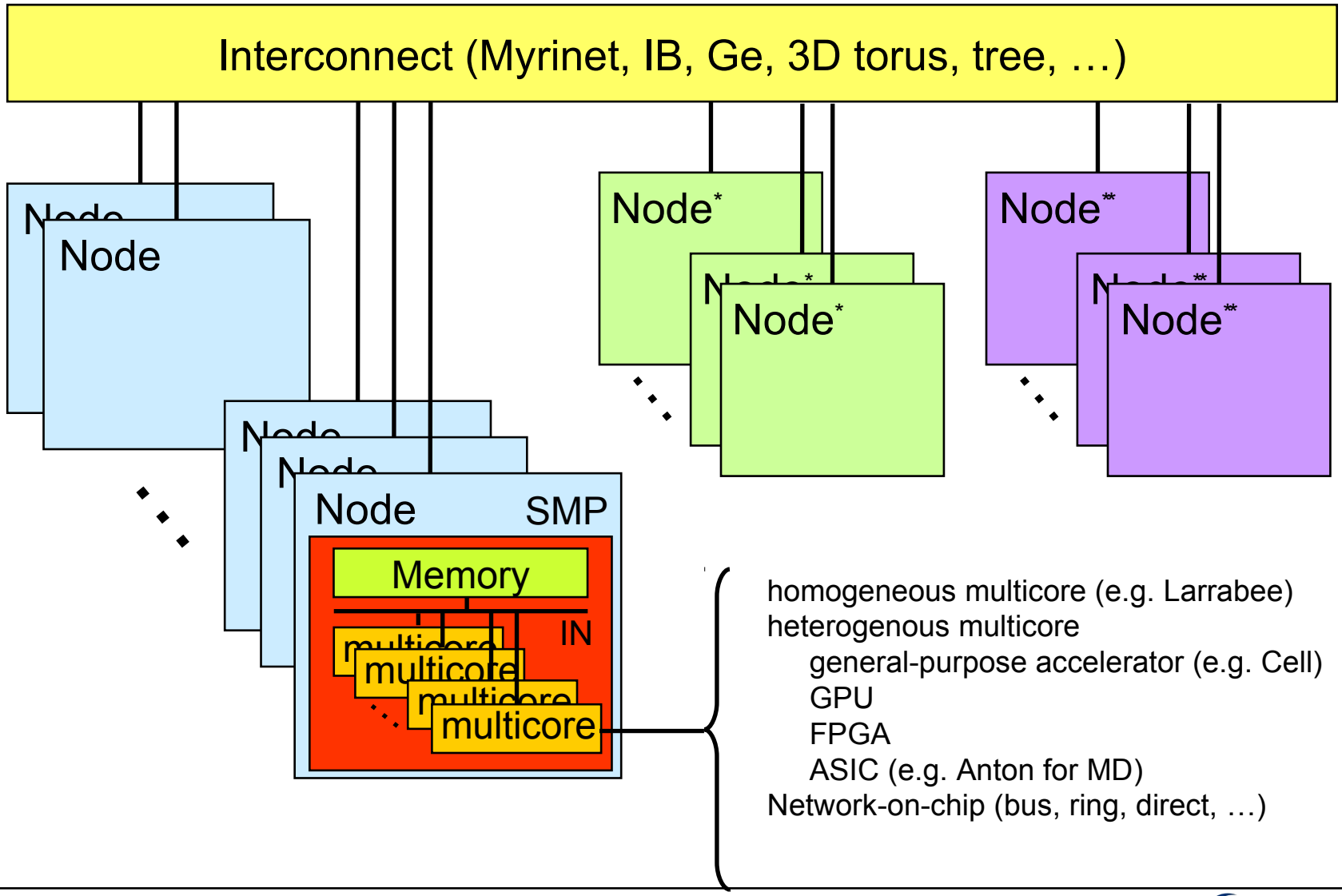


Multi-Modal Imaging



Data Analysis And Modeling

Hybrid SMP-cluster parallel systems



¿Qué son los supercomputadores?

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron Six Core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
2	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.50
3	National Institute for Computational Sciences/University of Tennessee United States	Kraken XT5 - Cray XT5-HE Opteron Six Core 2.6 GHz / 2009 Cray Inc.	98928	831.70	1028.85	
4	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution / 2009 IBM	294912	825.50	1002.70	2268.00
5	National SuperComputer Center in Tianjin/NUDT China	Tianhe-1 - NUDT TH-1 Cluster, Xeon E5540/E5450, ATI Radeon HD 4870 2, Infiniband / 2009 NUDT	71680	563.10	1206.19	
6	NASA/Ames Research Center/NAS United States	Pleiades - SGI Altix ICE 8200EX, Xeon QC 3.0 GHz/Nehalem EP 2.93 Ghz / 2009 SGI	56320	544.30	673.26	2348.00
7	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution / 2007 IBM	212992	478.20	596.38	2329.60
8	Argonne National Laboratory United States	Blue Gene/P Solution / 2007 IBM	163840	458.61	557.06	1260.00
9	Texas Advanced Computing Center/Univ. of Texas United States	Ranger - SunBlade x6420, Opteron QC 2.3 Ghz, Infiniband / 2008 Sun Microsystems	62976	433.20	579.38	2000.00
10	Sandia National Laboratories / National Renewable Energy Laboratory United States	Red Sky - Sun Blade x6275, Xeon X55xx 2.93 Ghz, Infiniband / 2009 Sun Microsystems	41616	423.90	487.74	

TOP500® SUPERCOMPUTER SITES

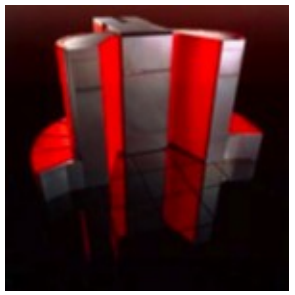
www.top500.org

HPC hierarchy in current Top 10

Rank	Machine name	Reported peak (TF)	Reported max (TF)	Efficiency	Power (MW)	# total cores	# cores	# nodes	# chip/node	#cores/chip	# ops/core	Core frequency (GHz)	Computed peak (GF)	Total memory (TB)	Memory/node (GB)	TF/TB ratio
1	ORNL Jaguar	2331	1759	0,75	6,95	224162	224256	18680	2	6	4	2,6	2332262	299	16	0,13
							62656	7832	2	4	4	2,1	526310	63	8	0,12
2	LANL Roadrunner	1375,7	1042	0,76	2,34	122400	103680	6480	2	8	4	3,2	1327104	104	8	0,08
							12960	3240	2	2	2	1,8	46656		16	
3	NICS/UT Kraken XT5	1028,85	831,7	0,81		98928	99072	8256	2	6	4	2,6	1030349	132	16	0,13
4	FZJ JUGENE BlueGene/P	1002,701	825,5	0,82	2,26	294912	294912	73728	1	4	4	0,85	1002701	147	2	0,15
5	Tianhe-1 NUTD	1206,19	563,1	0,47		71680	20480	2560	2	4	4	2,53	207258	82	32	0,06
							51200	2560	1	20	32	0,75	1228800			
6	NASA Ames Pleiades	673,259	544,3	0,81	2,34	56320	47104	5888	2	4	4	3	565248	75	8	0,11
							9216	1152	2	4	4	2,93	108012		24	
7	LLNL Blue Gene/L	596,38	478,2	0,80	2,32	212992	212992	106496	1	2	4	0,7	596378	32	0,5	0,05
8	ANL Blue Gene/P	557,06	450,3	0,81	1,26	163840	163840	40960	1	4	4	0,85	557056	82	2	0,15
9	TACC ranger	579,38	433,2	0,75	2	62976	62976	3936	4	4	4	2,3	579379	126	32	0,22
10	Red Sky Sandia	487,74	423,9	0,87		41616	41600	5200	2	4	4	2,93	487552	83	16	0,17

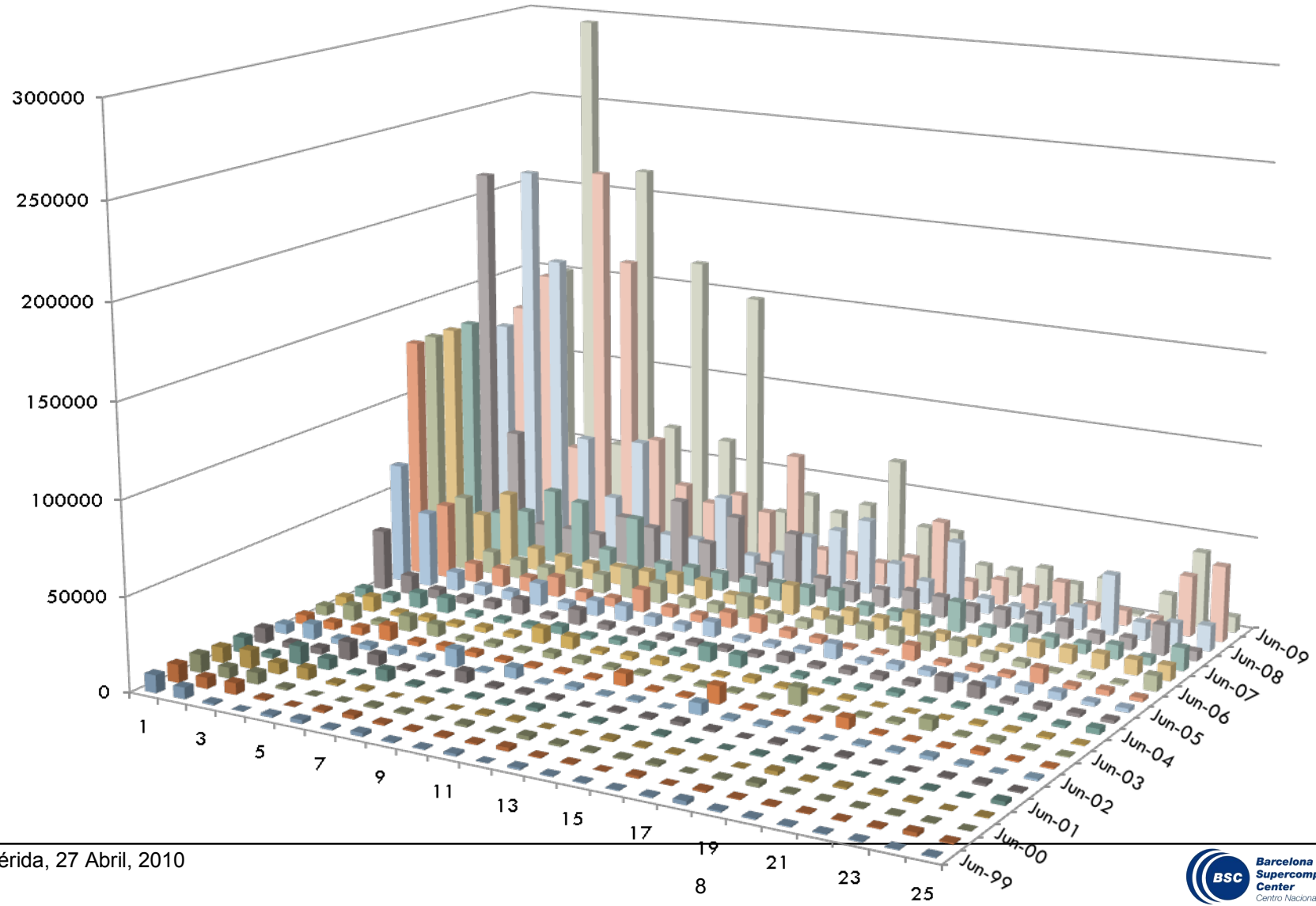
Based on November 2009 list

Looking at the Gordon Bell Prize

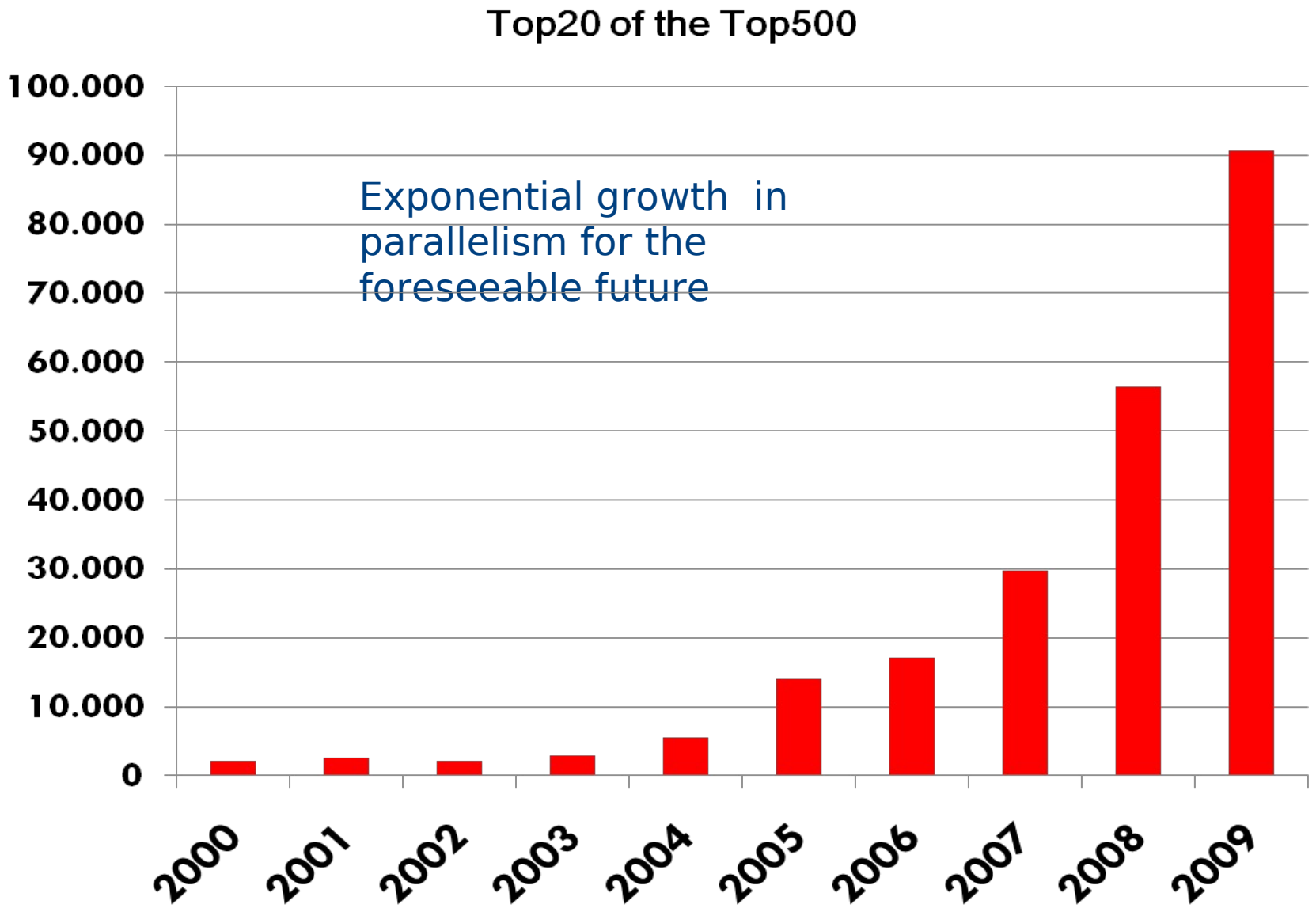


- 1 GFlop/s; 1988; Cray Y-MP; 8 Processors
  - Static finite element analysis
- 1 TFlop/s; 1998; Cray T3E; 1024 Processors
  - Modeling of metallic magnet atoms, using a variation of the locally self-consistent multiple scattering method.
- 1 PFlop/s; 2008; Cray XT5; 1.5×10^5 Processors
  - Superconductive materials
- 1 EFlop/s; ~2018; ?; 1×10^7 Processors (10^9 threads)

Cores in the Top25 Over Last 10 Years



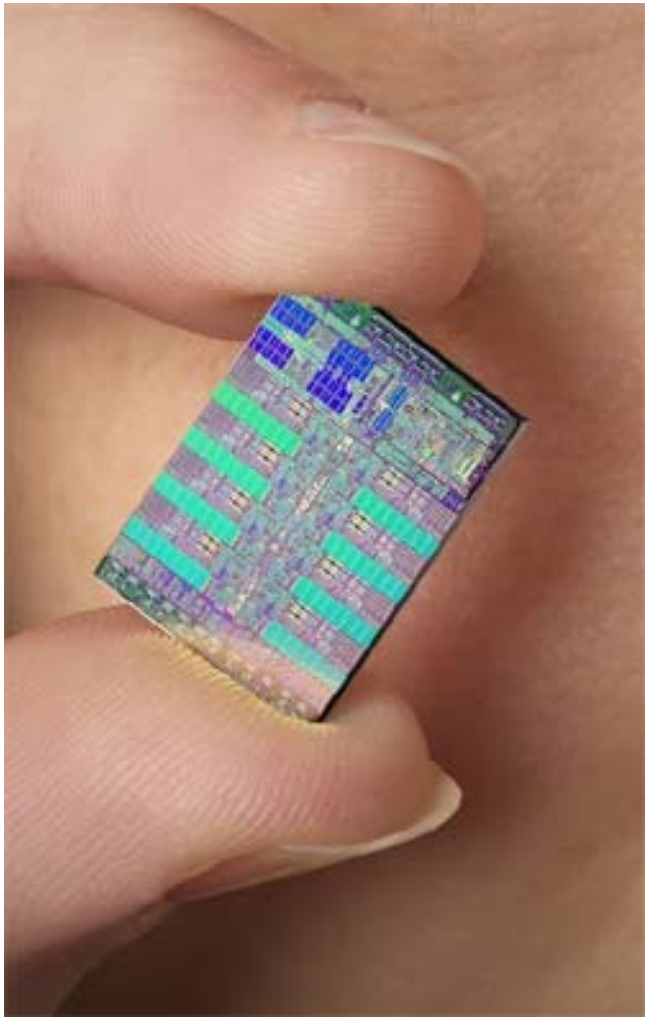
Average Number of Cores Per Supercomputer



Jaguar @ ORNL: 1.75 PF/s

- Cray XT5-HE system
- Quad-core AMD Opteron processors running at 2.6 GHz, 224,162 cores.
- Power: 6.95 Mwatts
- 300 terabytes of memory
- 10 petabytes of disk space
- 240 gigabytes per second disk bandwidth
- Cray's SeaStar2+ interconnect network.

The CELL/B.E. chip



235 Mtransistors
235 mm²

Roadrunner supercomputer
at
Los Alamos National Laboratory

IBM breaks 1 Petaflop barrier at Los Alamos

System Highlights ...

- ✓1st to break the Petaflop barrier
- ✓Fastest machine in USA
- ✓Largest contributor to Top500 aggregate performance with 1.026 of 11.7 Petaflops (8.7%)
- ✓**Third most power efficient system (QS22s at Fraunhofer and IBM Germany are #1 and #2)**

Source: www.top500.org

Site: DOE/NNSA/LANL

System Name: QS22/LS21

System Configuration: IBM BladeCenter cluster of 17 Connected Units (CUs) for a total 3060 nodes dual socket 1.8 GHz Opteron (dual core) LS21 blades plus 6120 nodes dual socket 3.2 GHz PowerXCell 8i (8 SPU + 1 PPU cores) QS22 blades. InfiniBand Interconnect. 280 racks total.

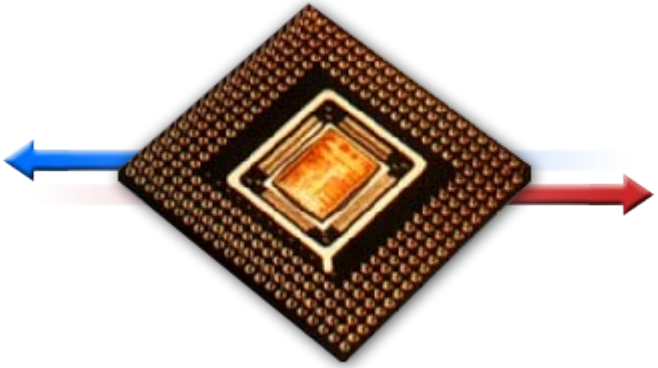
Cores: 122,400
Rmax: 1,026,000 GF= 1.026 PF
Nmax: 2236927
Rpeak: 1,375,776 GF
Power: 2345 kW
Mflops/Watts: 437 Mflops/W

Increasing CPU Performance:

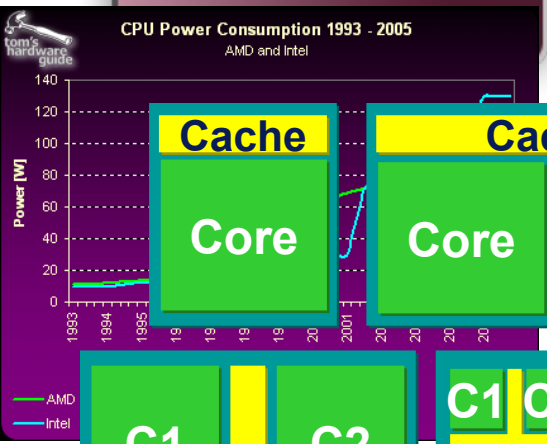
A Delicate Balancing Act

reasing the number of gates into a tight knot and decreasing the cycle time of the process

Lower Voltage



Increase Clock Rate & Transistor Density



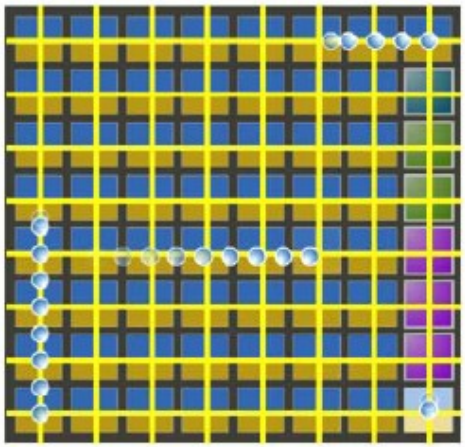
We have seen increasing number of gates on a chip and increasing clock speed.

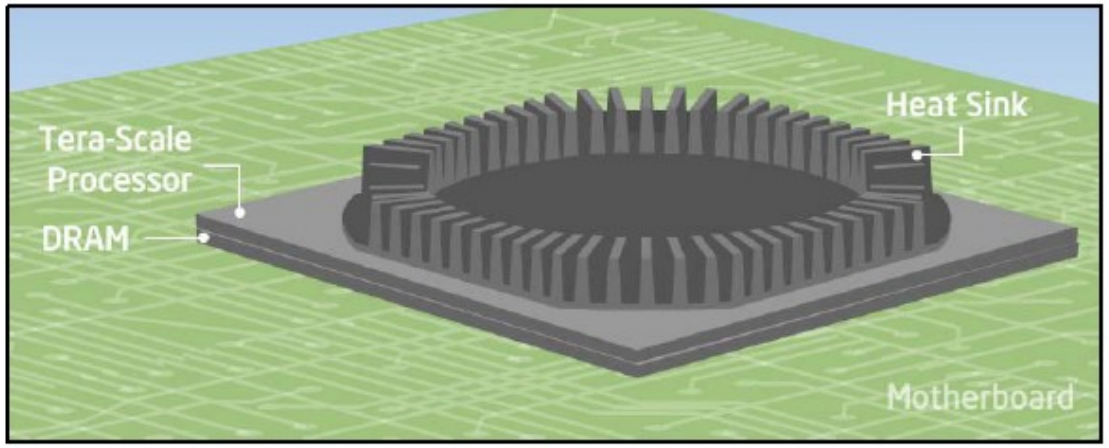
Heat becoming an unmanageable problem, Intel Processors > 100 Watts

We will not see the dramatic increases in clock speeds in the future

However, the number of gates on a chip w continue to i

Increasing chip performance: Intel´s Petaflop chip





The key technologies of this first Tera-scale Research Prototype are a mesh interconnect (left) and support for 3D stacked memory (above).

- 80 processors in a die of 300 square mm.
- Terabytes per second of memory bandwidth
- Note: The barrier of the Teraflops was obtained by Intel in 1991 using 10.000 Pentium Pro processors contained in more than 85 cabinets occupying 200 square meters ☺
- This will be possible in 3 years from now

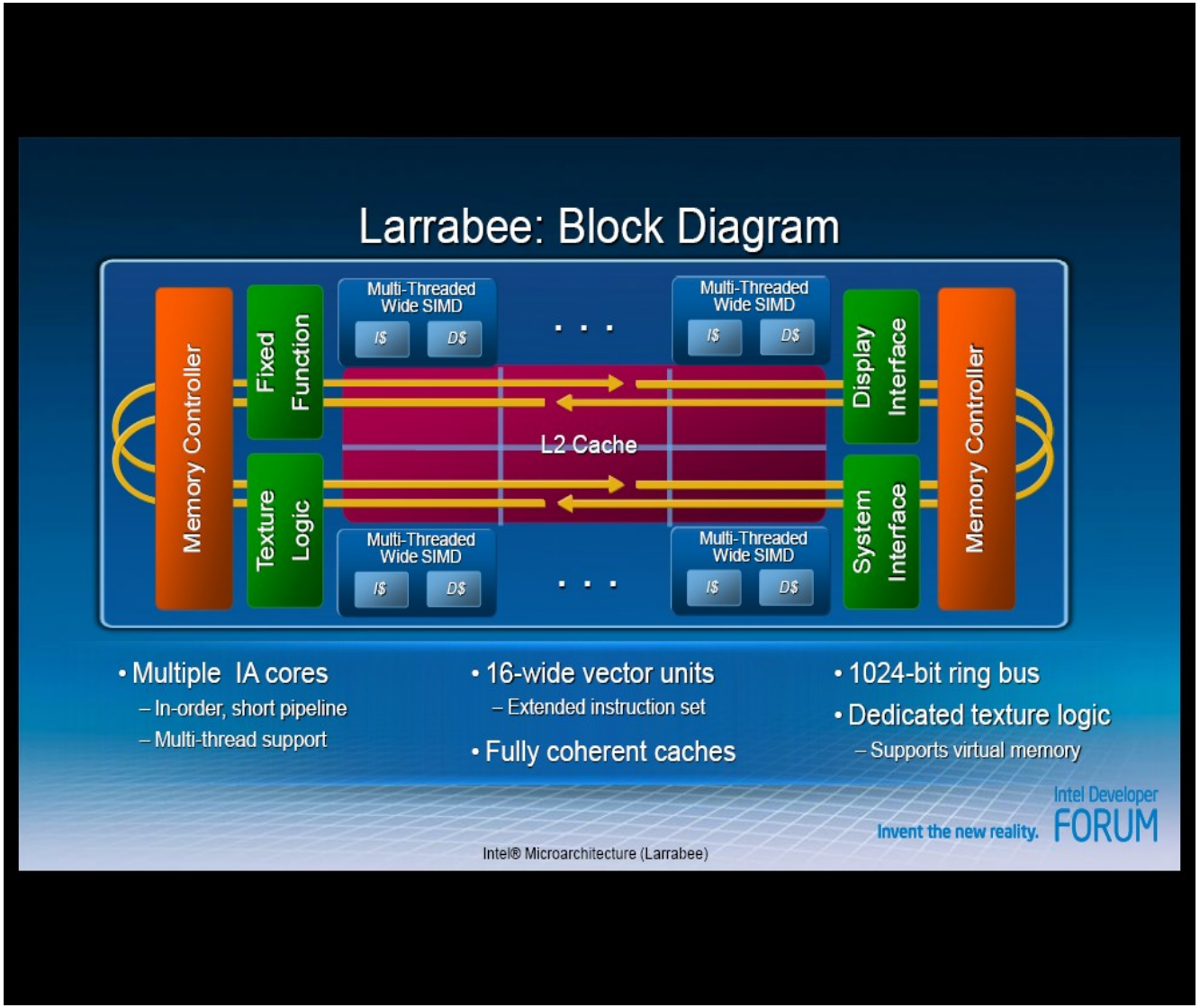
Thanks to Intel

Intel/UPC

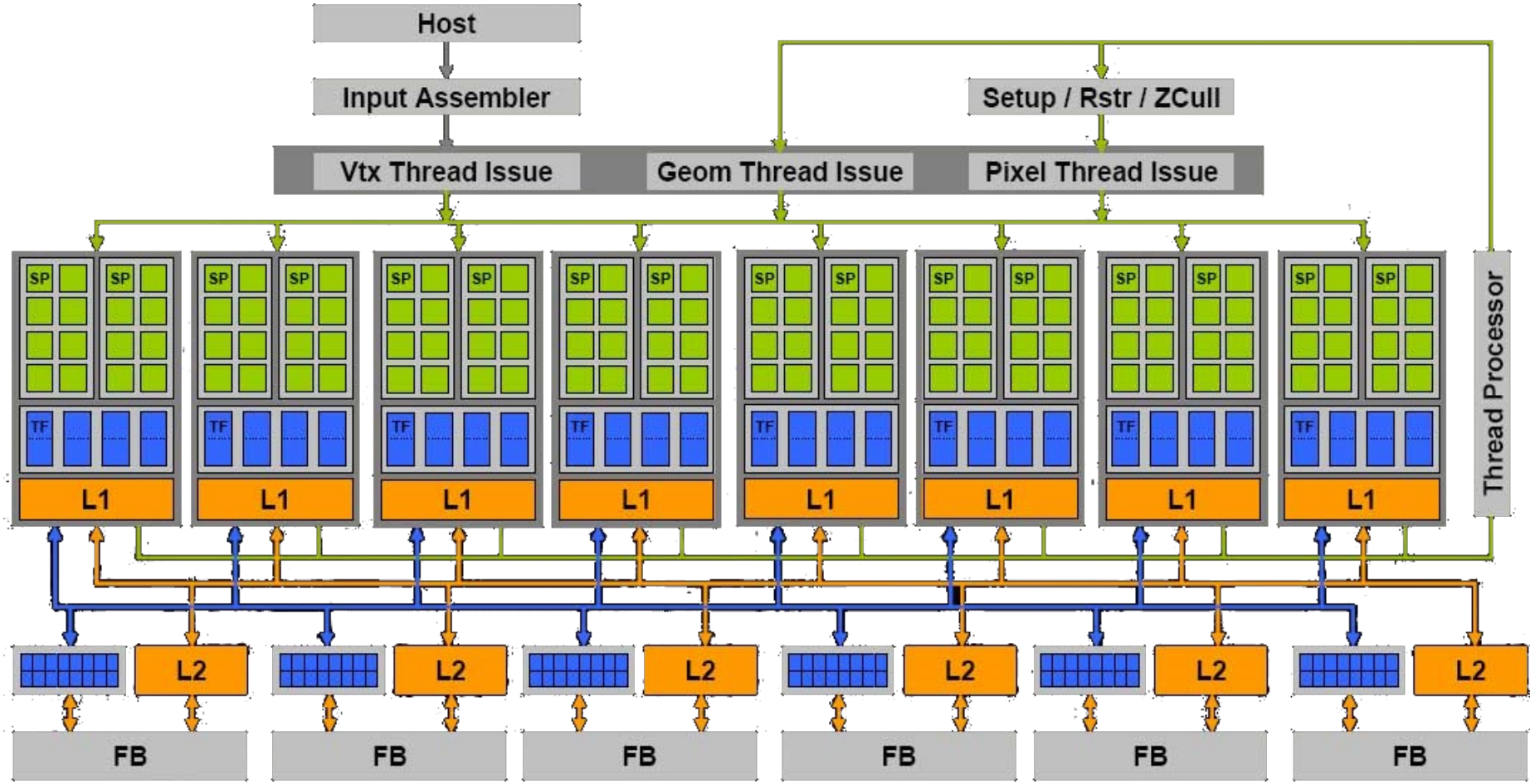
Since 2002 (Roger Espasa, Toni Juan)

40 People

Microprocessor Development (Larrabee x86 many core)



GPUs


Host
Input Assembler
Setup / Rstr / ZCull
Vtx Thread Issue
Geom Thread Issue
Pixel Thread Issue
SP
TF
L1
L2
FB
Thread Processor

NVIDIA Fermi Architecture

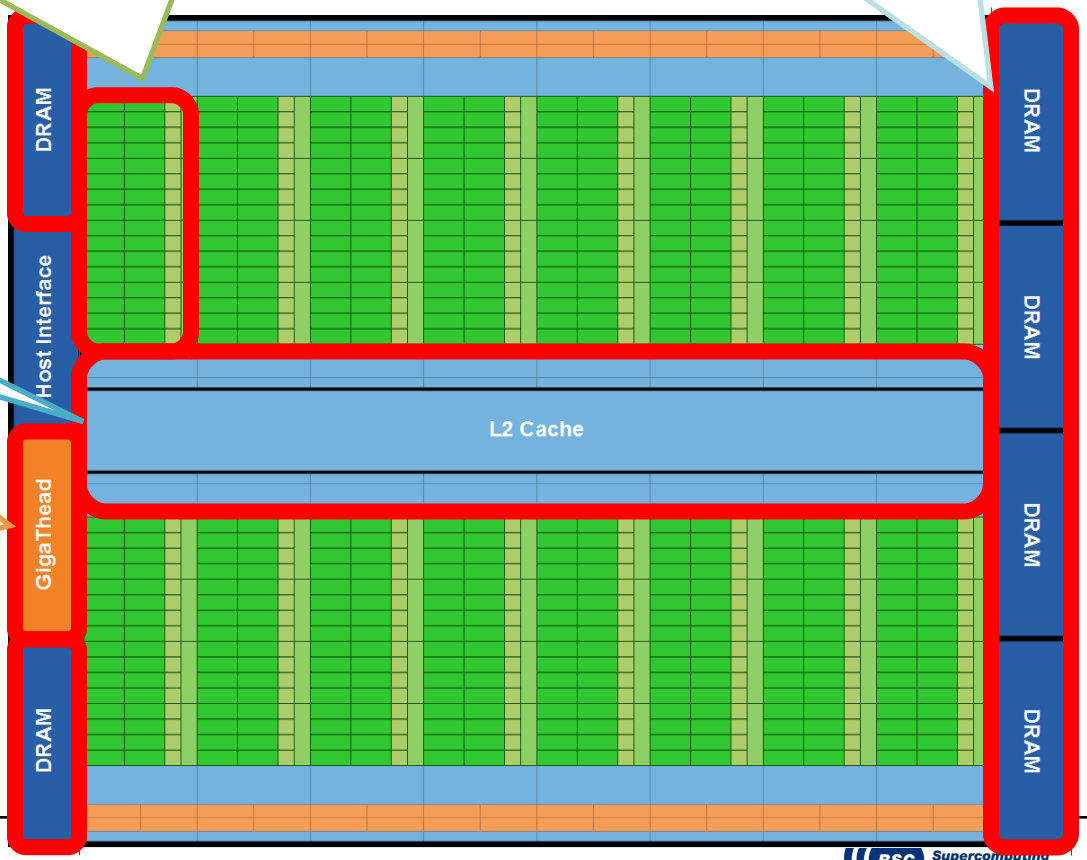
16 Streaming- Multiprocessors (512 cores) execute Thread Blocks

620 Gigaflops

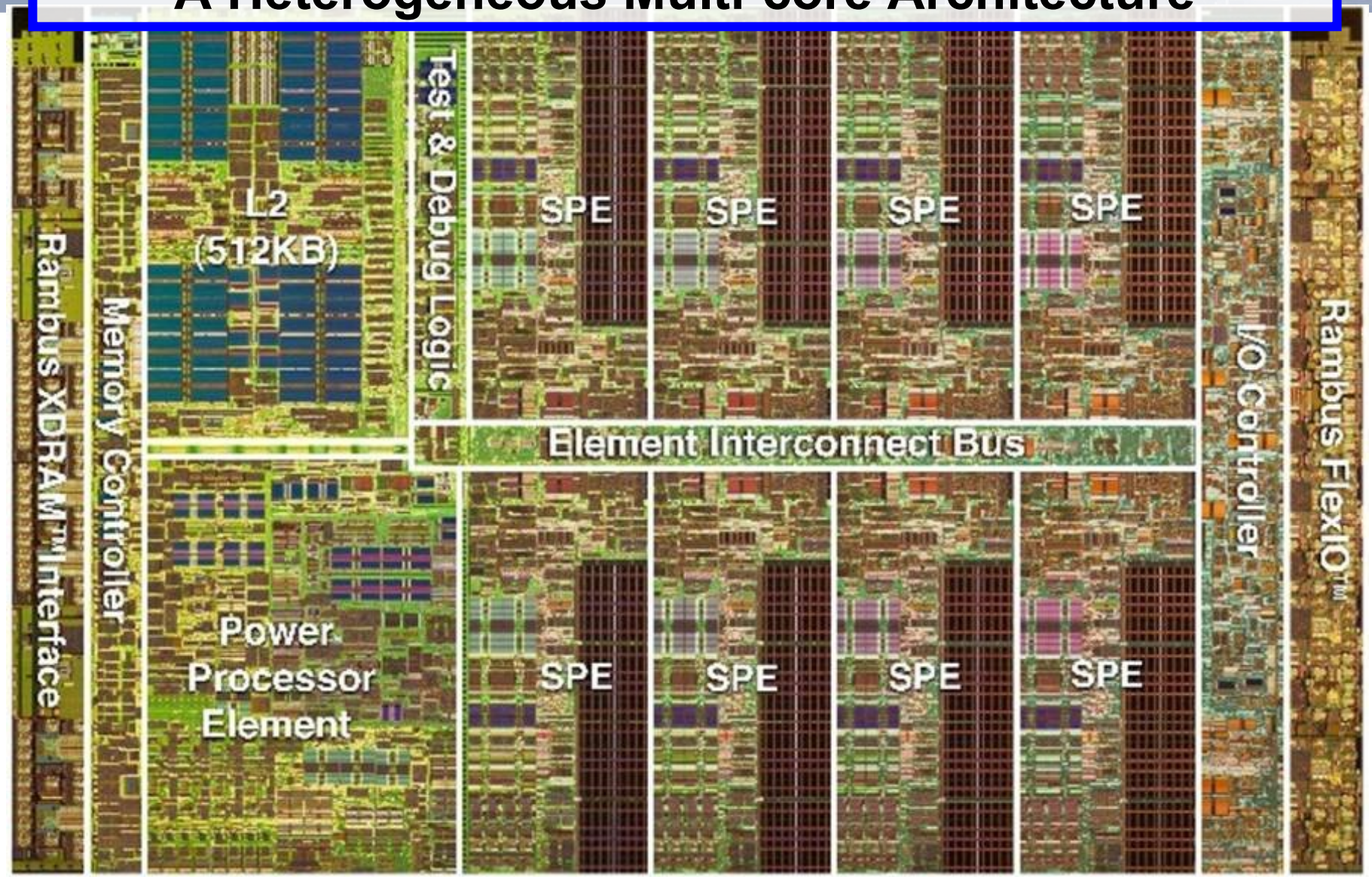
Wide DRAM interface provides 12 GB/s bandwidth

Unified 768KB L2 cache serves all threads

GigaThread hardware scheduler assigns Thread Blocks to SMs



Cell Broadband Engine ™: A Heterogeneous Multi-core Architecture



** Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc.*

Increasing chip performance: ASICs

- Special-purpose system for molecular dynamics (MD) simulations
- 512 custom-designed ASICs interconnected by a specialized high-speed three-dimensional torus network

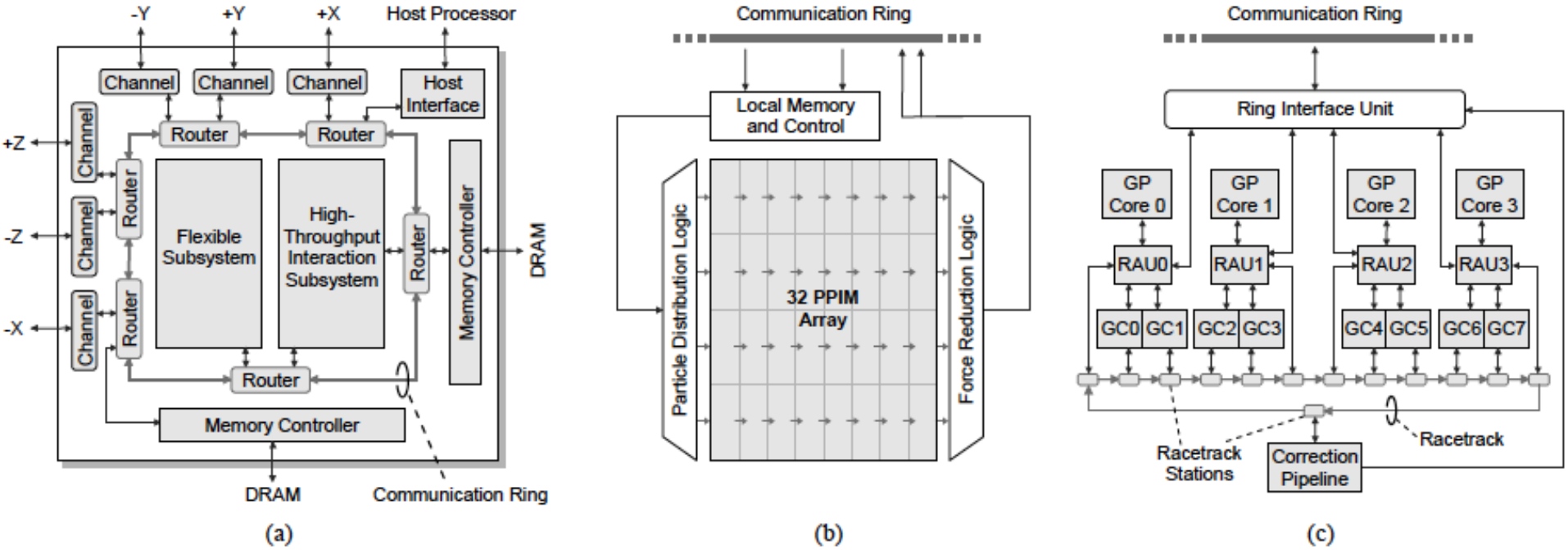
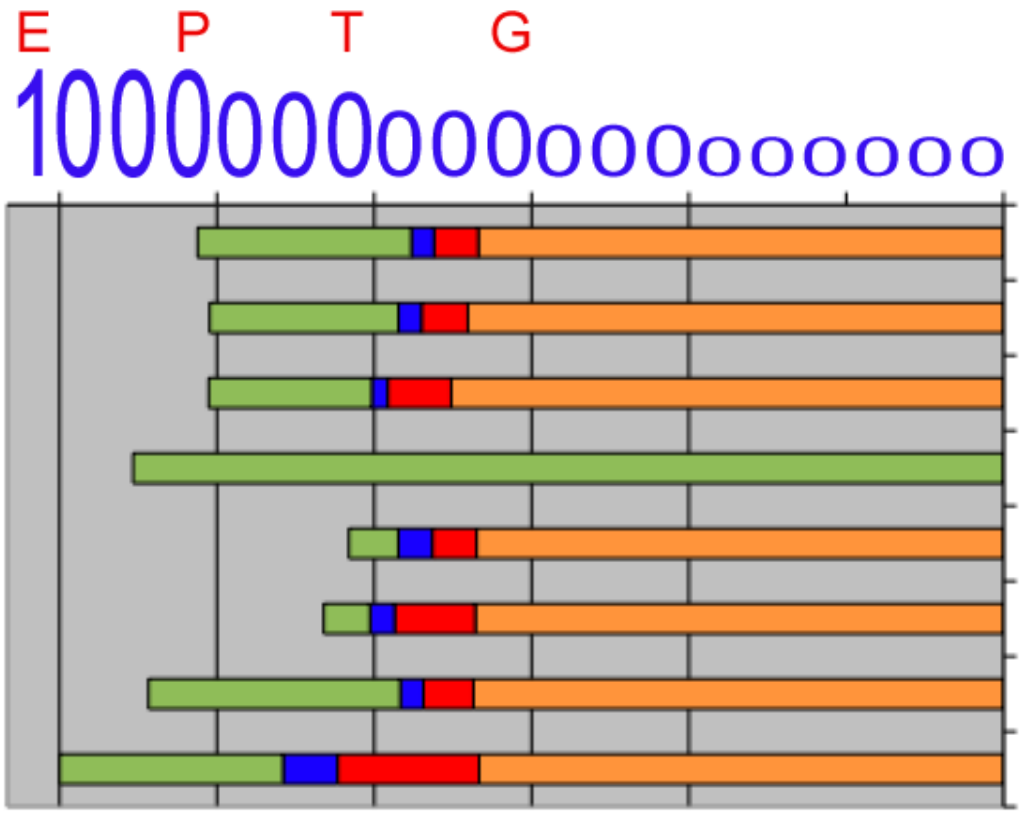


Fig. 2. (a) Anton ASIC. (b) High-throughput interaction subsystem. (c) Flexible subsystem.

Evolution towards Exaflop supercomputers



	cores	nodes	chips/node	cores/chip	ops/core	GHz	GFlops
Personal supercomputer	240	10	4	6	4	2,8	2688
Personal sup. accelerator	120	10	2	6	4	2,8	1344
	10240		2	512	1	0,648	6636

- 20PF/s, 1.6 PB Memory
- 96 racks, 98,304 nodes
- 1.6 M cores (1 GB/core)
- 50 PB Lustre file system
- 6.0 MW power

BSC-CNS e iniciativas a nivel internacional: IESP

Improve the world's simulation and modeling capability by improving the coordination and development of the HPC software environment

Build an international plan for developing the next generation open source software for scientific high-performance computing

Where We Are Today:

- SC08 (Austin TX) meeting to generate interest
- Funding from DOE's Office of Science & NSF Office of Cyberinfratructure
- US meeting (Santa Fe, NM) April 6-8, 2009
  - 65 people
- NSF's Office of Cyberinfrastructure funding
- European meeting (Paris, France) June 28-29, 2009
  - 70 people
  - Outline Report
- Asian meeting (Tsukuba Japan) October 18-20, 2009
  - Draft roadmap
  - Refine Report
- SC09 (Portland OR) BOF to inform others
  - Public Comment
  - Draft Report presented
- Oxford meeting, April 2010

Nov 2008

Apr 2009

Jun 2009

Oct 2009

Nov 2009

Apr 2010

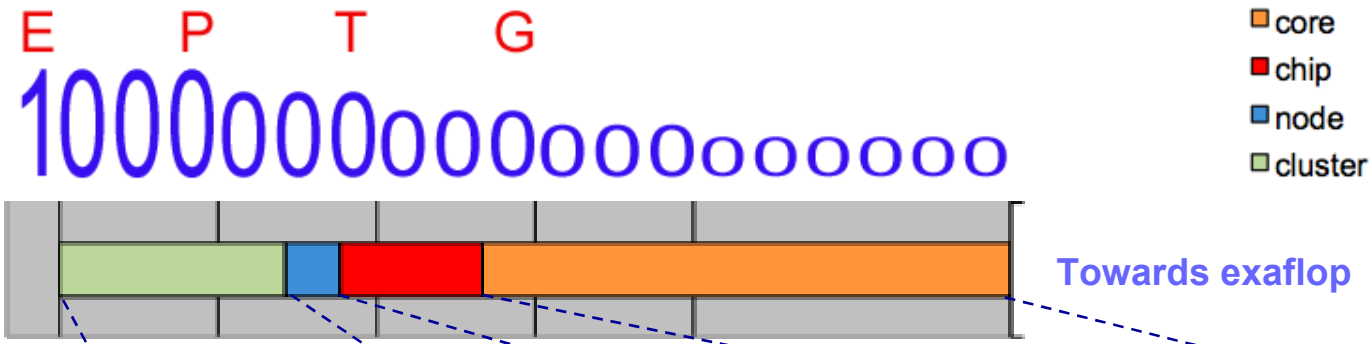
Systems Scaling Projections

Begin Full System Delivery (Yr)	2004	2007	2012	2015	2019
Design Parameters	**BG/L**	**BG/P**	**25PF**	**300PF**	**1200PF**
Cores / Node	2	4	8-24	32-64-128	96-128-500
Clock Speed (GHz)	0.7	0.85	1.6-4.1	2.3-4.8	2.8-6.0
Flops / Clock / Core	4	4	8-32	8-32	16-64
Nodes / Rack	1024	1024	100-1024	256-1024	256-1024
Racks / Full System Config	64	72	128-350	128-400	256-400
MB RAM/core	256	512	1024-4096	1024-4096	1024-4096
Total Power	2.5MW	4.8MW	8MW-20MW	20MW-50MW	30MW-80MW
Flops / Node (GF)	5.6	14	128-640	640-2000	2000-6000
Flops / Rack (TF)	5.7	14	200-400	400-1200	1600-4800
LB Concurrency	5.E+05	1.E+06	10E6-64E6	100E6-1E9	1E9-10E9
Full System					
Total Cores (Millions)	0.13	0.3	.3M-1.5M	1M-50M	4M-200M
Total RAM (TB)	33.6	151	2,000-4,400	3,000-10,000	5,000-50,000
Total Racks	64	72	128-350	128-400	256-400
Peak Flops System (PF)	**0.37**	**1**	**25**	**300**	**1200**

Factors that Necessitate Redesign

- **Steepness of the ascent from terascale to petascale to exascale**
- Extreme parallelism and hybrid design
  - Preparing for million/billion way parallelism
- Tightening memory/bandwidth bottleneck
  - Limits on power/clock speed implication on multicore
  - Reducing communication will become much more intense
  - Memory per core changes, byte-to-flop ratio will change
- Necessary Fault Tolerance
  - MTTF will drop
  - Checkpoint/restart has limitations
  - **Software infrastructure does not exist today**

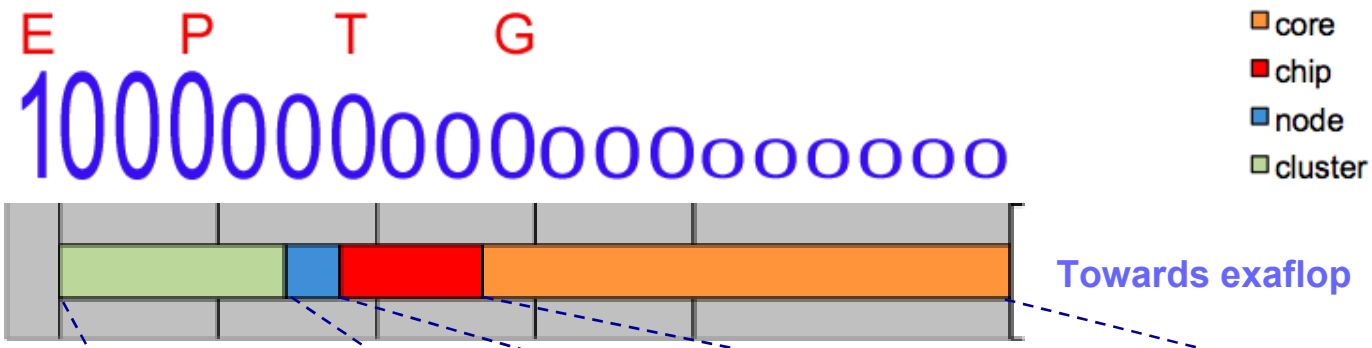
Holistic approach …



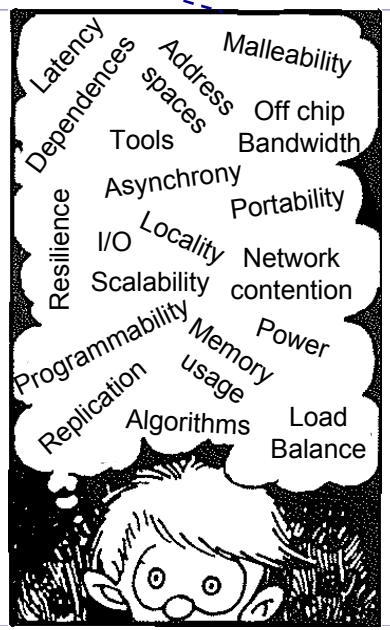
Applications
Performance tools
Programming model
Load balancing
Interconnection
Processor/node architecture

Thanks to Jesus Labarta

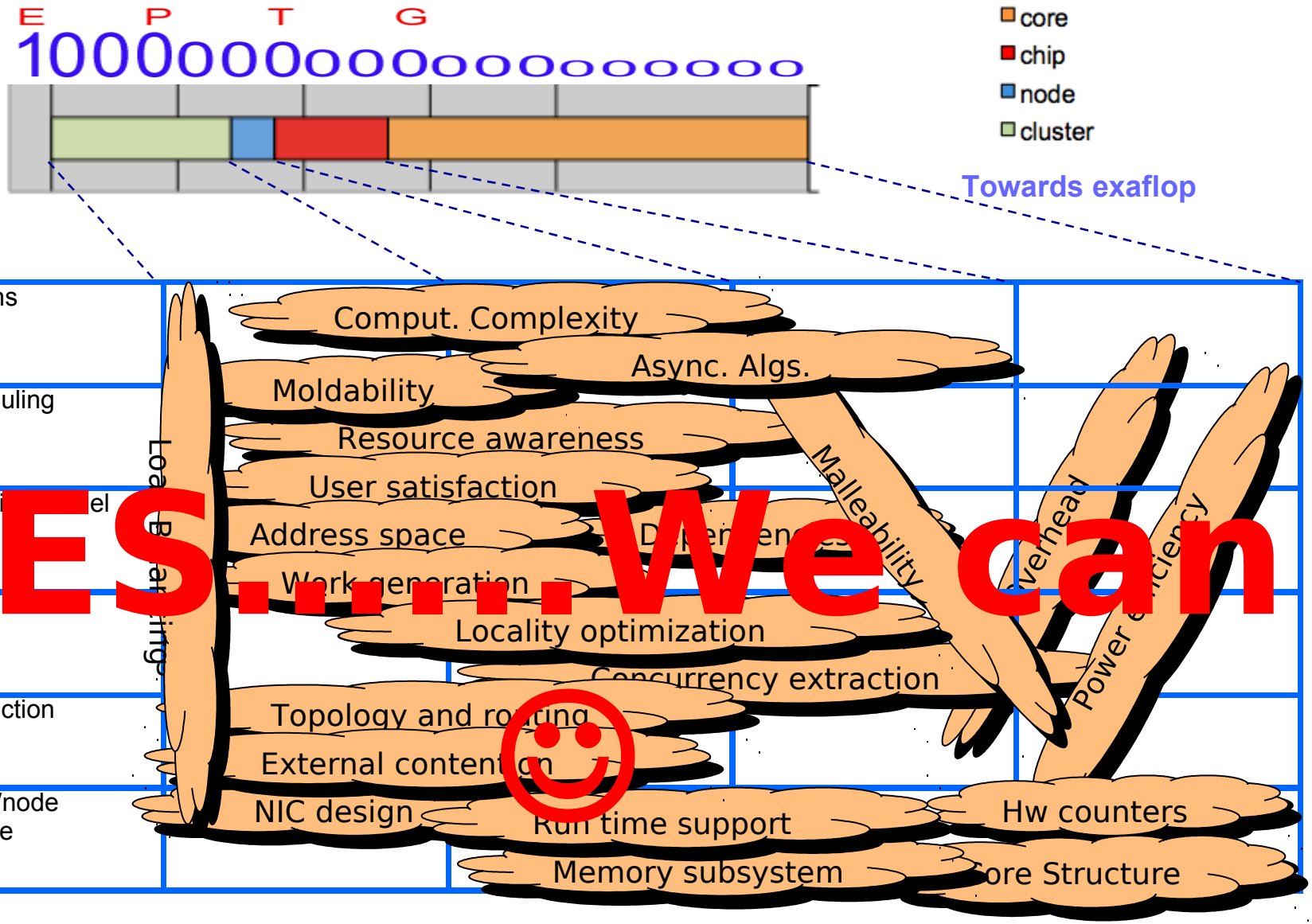
The holistic approach …



Applications
Performance tools
Programming model
Load balancing
Interconnection
Processor/node architecture



Holistic approach …



Challenges ⟷ Approach (a personal view)

- Variability
  - Everywhere, huge
- Efficiency
  - Performance and power reasons
  - Avoid overkills
- Memory
  - Logical and physical structure
  - Bandwidth and latency
- Resilience
  - Impossible to run an app without errors happening halfway
- Constraint: Power
- Programmability:
  - Don Grice: “we can do the hardware but if it can not be programmed …” (approx.)

Jesús Labarta

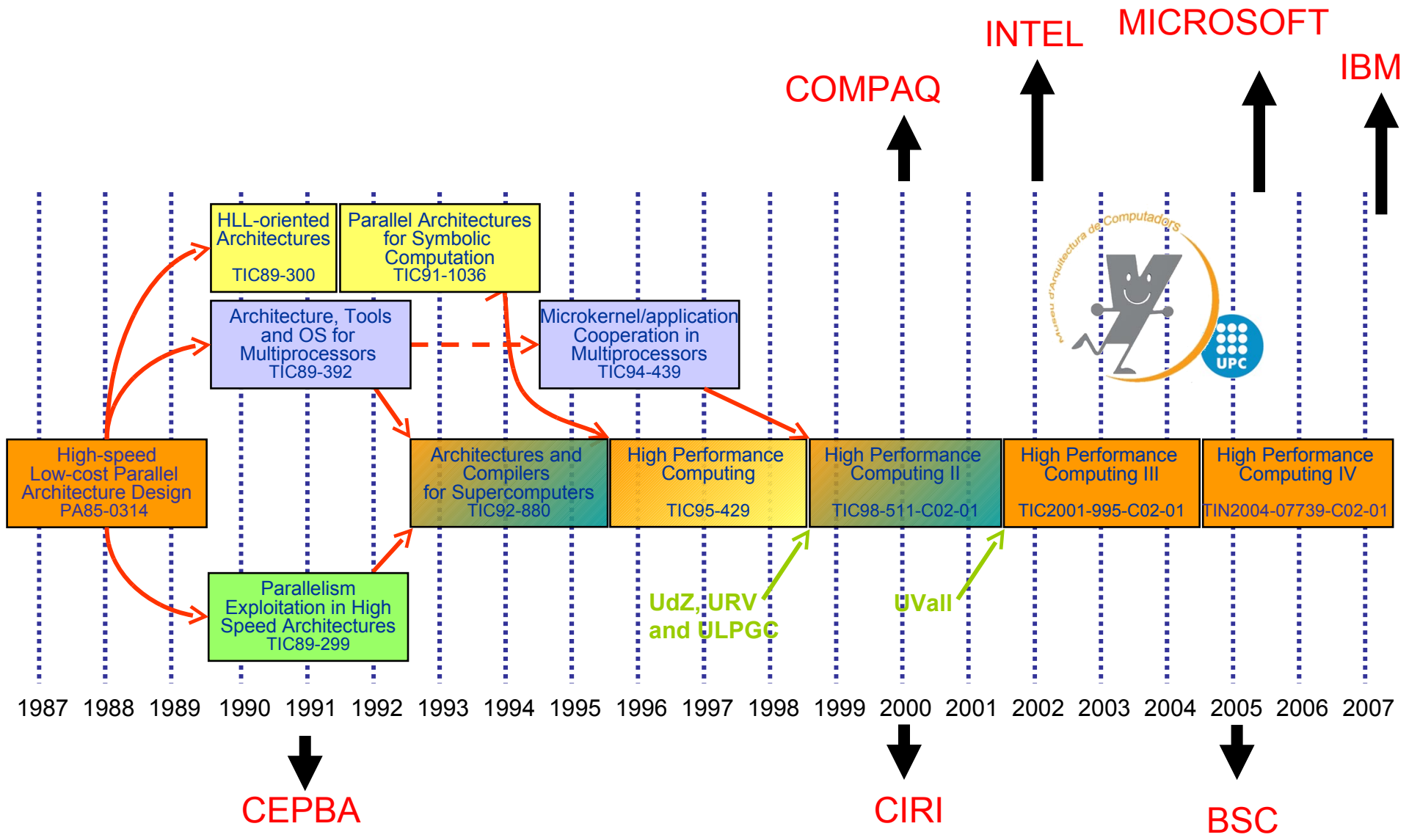
- Programming model
  - Machine independent.
  - What, not how
  - Smooth migration path
- Runtime/Execution model
  - Data access awareness.
    - Locality scheduling, minimize Bandwidth
  - Asynchrony/dataflow
  - Automatic Load balance
  - Malleability
- Algorithms
  - Asynchrony, overlap
  - Minimize bandwidth
- Resilience
  - From recovery to tolerance
- Holistic:
  - Applications: Co-design vehicles
  - Between system software layers and architecture
- Tools:
  - Fly with instruments
  - The importance of detail

Numerical Libraries

- Cat I: Uniquely exascale
  - Fault oblivious, Error tolerant software
  - Smart (AI based) algorithms
- Cat II: Exascale + trickle down
  - Async methods
    - Overlap data and computation
  - Algorithms that minimize communications
  - Self-adapting
- Cat III
  - Autotuning based software
  - Standardization activities
  - Architectural aware algorithms/libraries
  - Energy efficient algorithms
  - Mixed arithmetic
  - Hybrid and hierarchical based algorithms (eg linear algebra split across multi-core and gpu,)

Our origins.....

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



DAC (UPC): High Performance Computing

Computer architecture:
- Superscalar and VLIW
- Hardware multithreading
- Design space exploration for multicore chips and Hw accelerators
- Transactional memory (Hw, Hw-assisted)
- SIMD and vector extensions/units

Benchmarking, analysis and prediction tools:
- Tracing scalability
- Pattern and structure identification
- Visualization and analysis
- Processor, memory, network, system

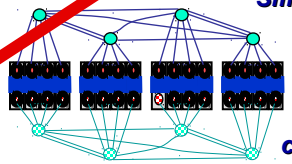
The Grid

Programming models:
- Scalability of MPI and UPC
- OpenMP for multicore, SMP and ccNUMA
- DSM for clusters
- CellSs, streaming
- Transactional Memory
- Embedded architectures

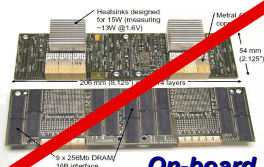
Large cluster systems

Future Petaflop systems

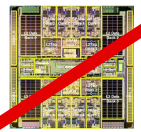
Small DMM



cc-NUMA



On-board SMP



Chip

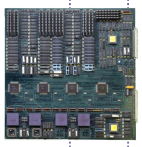
Grid and cloud computing:
- Programming models
- Resource management
- I/O for Grid

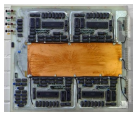
Operating environments:
- Autonomic application servers
- Resource management for heterogenous workloads
- Coordinated scheduling and resource management
- Parallel file system scalability

Algorithms and applications:
- Numerical
- Signal processing

Venimos de muy lejos …

Transputer cluster





Research prototypes

Convex C3800

Connection Machine CM-200
0,64 Gflop/s

Parsys Multiprocessor

Parsytec CCi-8D
4.45 Gflop/s

Compaq GS-140
12.5 Gflop/s

Compaq GS-160
23.4 Gflop/s

SGI Origin 2000
32 Gflop/s

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

BULL NovaScale 5160
48 Gflop/s

SGI Altix 4700
819.2 Gflops

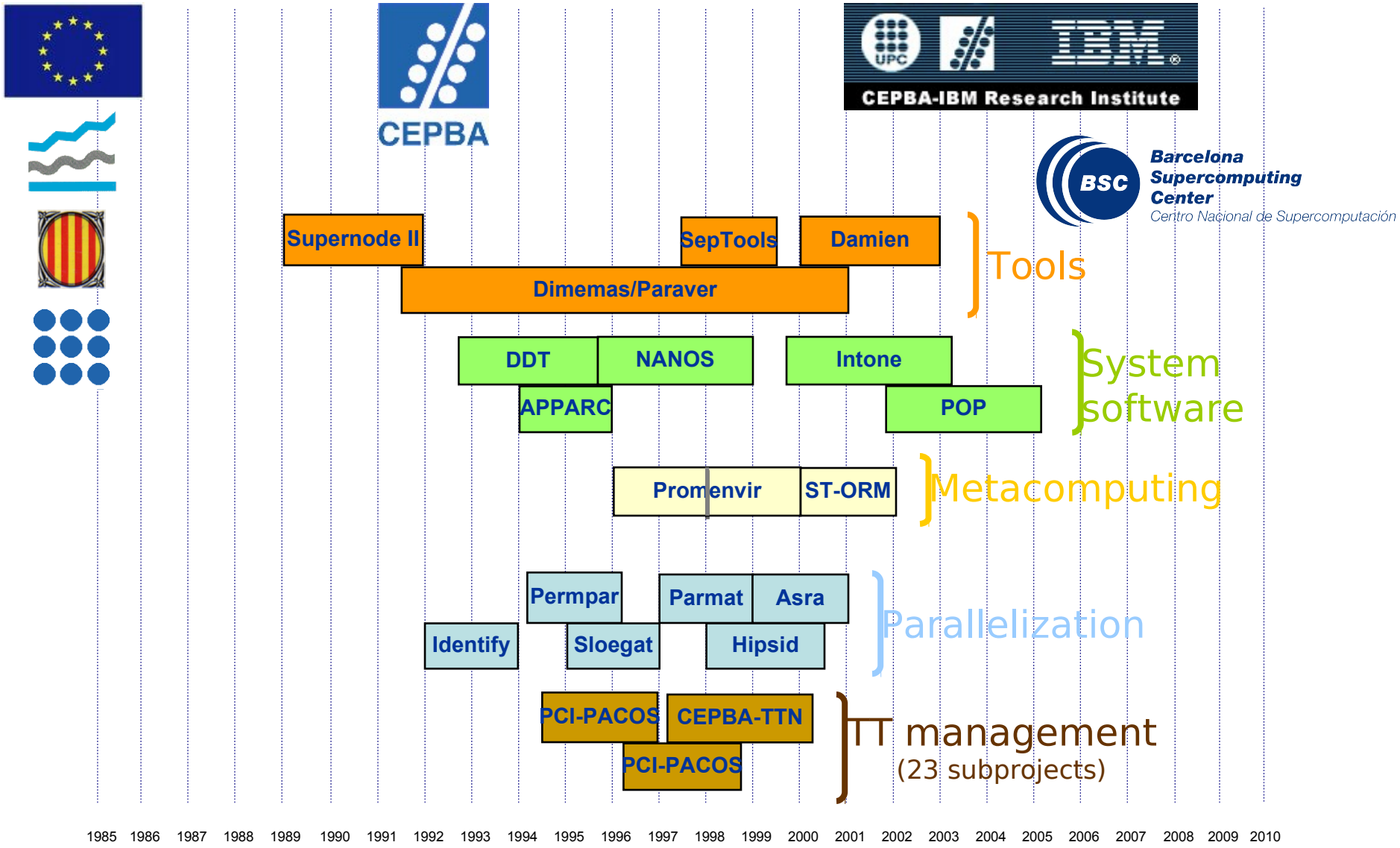
IBM PP970 / Myrinet
MareNostrum
42.35, 94.21 Tflop/s

Maricel
14.4 Tflops, 20 KW

SL8500
6 Petabytes

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.....

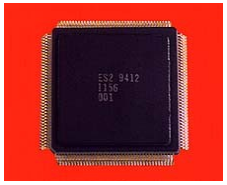


Venimos de muy lejos …

Ayto. Barcelona
Uitesa
UPC-EIO

AMES, CIMNE

TGI
UPM-DATSI

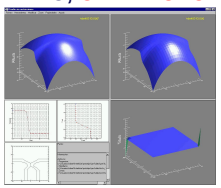


Hesperia
Neosystem
s
UPC-EIO

Soler y Palau
CIMNE
CEPBA-UPC

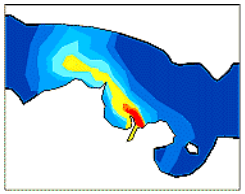
Torres
Soft Greenhouse
CEPBA-UPC

INDO, CEPBA-UPC



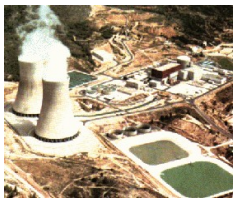
Metodos Cuantitativos
Gonfiesa
CESCA, CESGA

Tecnatom, UMA



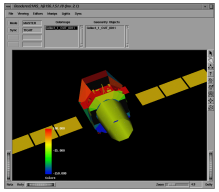
AZTI
UPC-LIM

Iberdrola, Uitesa, UPV



BCN COSIVER
Mides
UPC-EIO

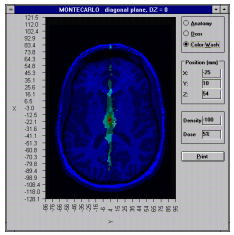
CEPBA
CESCA
UMA
UNICAN
UPM

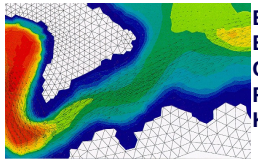


CASA
Envision
GTD
Intespace
RUS

ST Mecanica
DERBI
AUSA
CEPBA-UPC

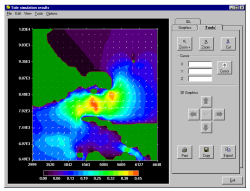
Ospedali Galliera
Le Molinette
Parsytec
PAC
EDS





ENEL
EDF
CSR4
Reiter
Kenijoki

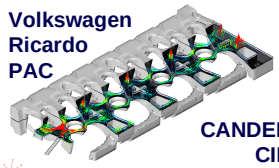
Ferrari, Genias, P3C



SENER
CIC
UNICAN

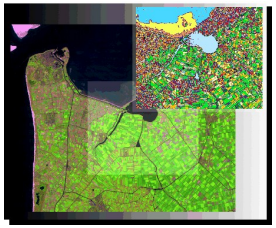
Intera SP
Intera UK
UPC-DIT
CEPBA-UPC

Volkswagen
Ricardo
PAC

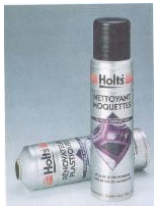


CANDEMAT
CIMNE
CEPBA-UPC

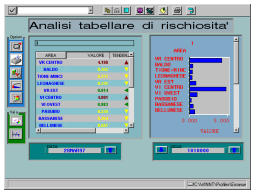
Italeco
Geospace
Intecs
Univ. Leiden

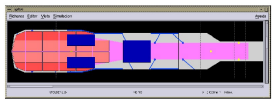


CEBAL-ENTEC
NEOSYSTEMS



Cari Verona
AIS
PAC
Univ.
Cat. Milan





Cristaleria Española
UNICAN
CEPBA-UPC

Inisel Espacio
Infocarto
UPC-TSC
CEPBA-UPC

Iberdrola
SAGE
CEPBA-UPC

The BSC-CNS

- The BSC-CNS mission:
  - Investigate, develop and manage technology to facilitate the advancement of science

- The BSC-CNS objectives:
  - R&D in Computer Sciences, Life Sciences and Earth Sciences.
  - Supercomputing support to external research

- BSC-CNS is a consortium that includes :
  - the Spanish Government (MICINN)
  - the Catalonian Government (DIUE)
  - the Technical University of Catalonia (UPC)
  - the National Council for Scientific Research (CSIC)

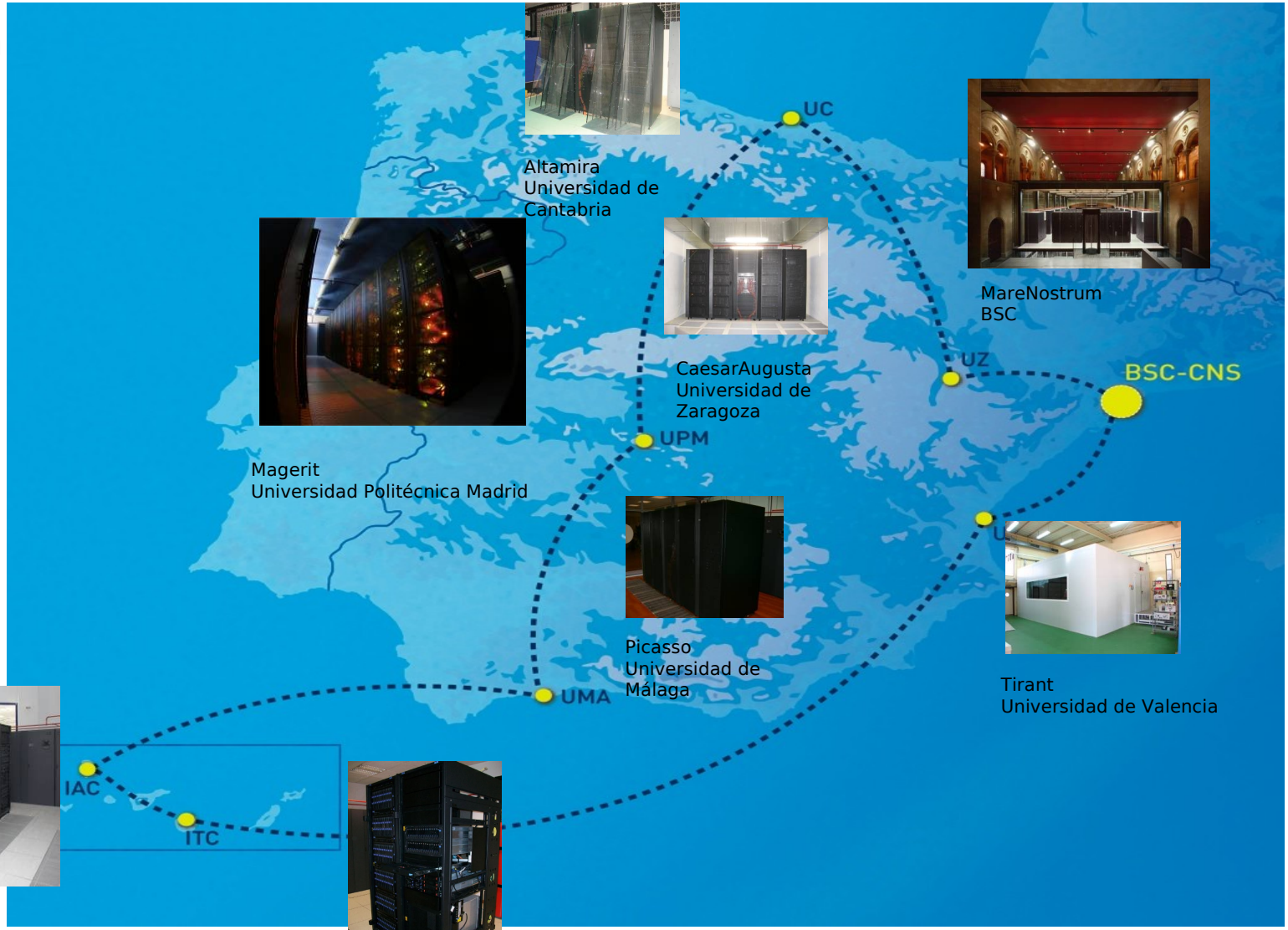
Location


IBM

TOP500 MareNostrum's Evolution

List	World Position	Europe Position
November 2004	4	1
June 2005	5	1
November 2005	8	1
June 2006	11	3
November 2006	5	1
June 2007	9	1
November 2007	13	3
June 2008	26	8
November 2008	40	10
June 2009	60	19
Novembre 2009	76	22

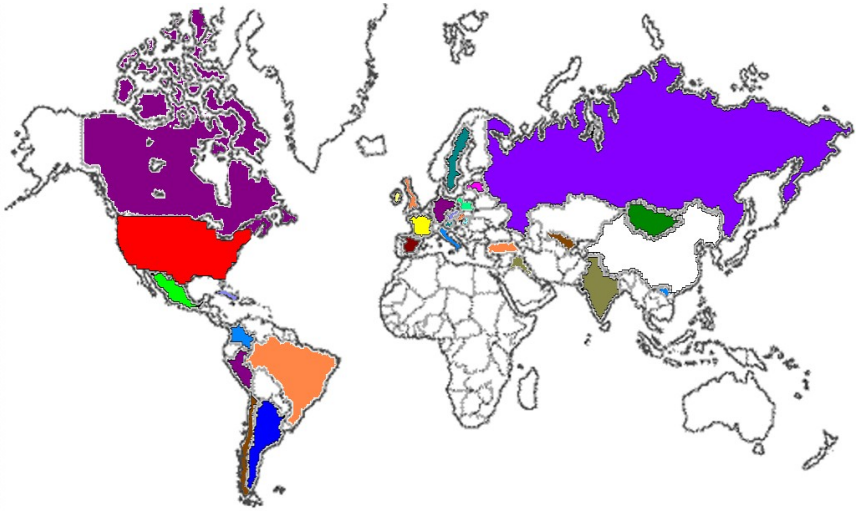
Suport: Red Española de Supercomputación



The BSC-CNS

300 people from 24 different countries

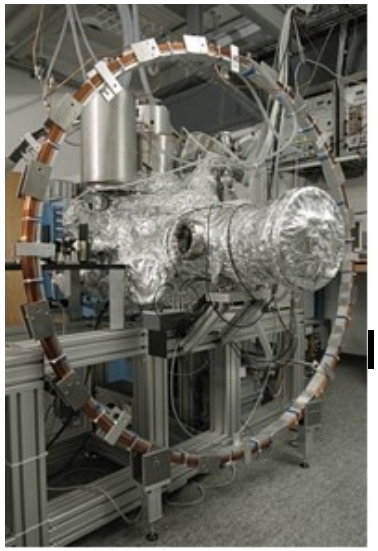
(Argentina, Belgium, Brazil, Bulgaria, Canada, Colombia, China, Cuba, France, Germany, India, Iran, Ireland, Italy, Jordania, Lebanon, Mexico, Pakistan, Poland, Russia, Serbia, Spain, Turkey, UK, USA)



BSC-CNS: sinergia con infraestructuras

- CNS es un complemento fundamental para las infraestructuras científicas experimentales

ICFO

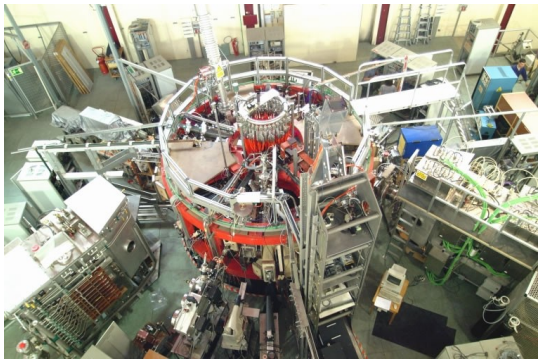


IAC

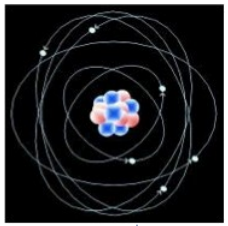
Sincroton

IRB

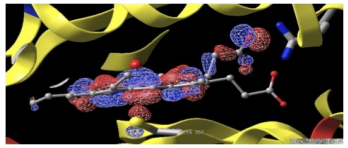
CIEMAT(TJ II)



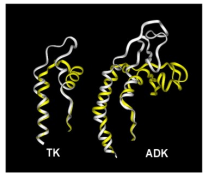
BSC-CNS: Ciencias de la Vida



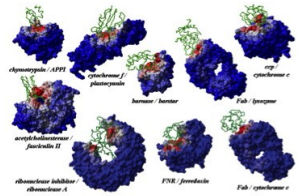
Atomic (and electronic) modeling of protein biochemistry and biophysics



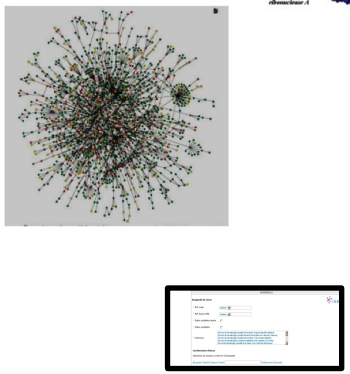
Micro and mesoscopic modeling of macromolecules. Drug Design



Identification of the structural bases of protein-protein interaction

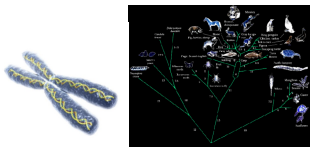


Protein-protein interaction networks
Systems biology



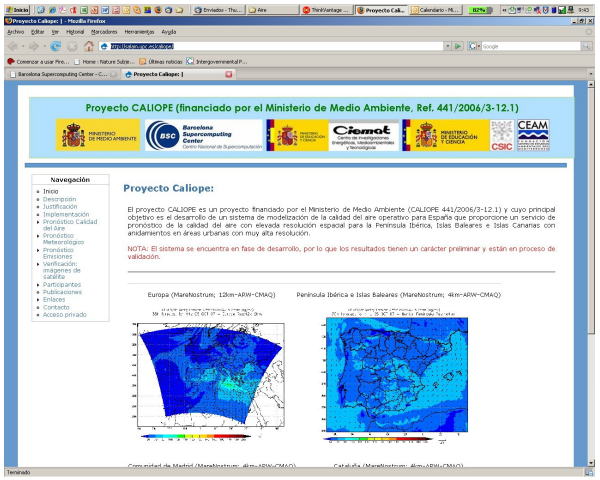
Web services, applications, databases

Analysis of genomes and networks to model diseases, systems and evolution of organisms

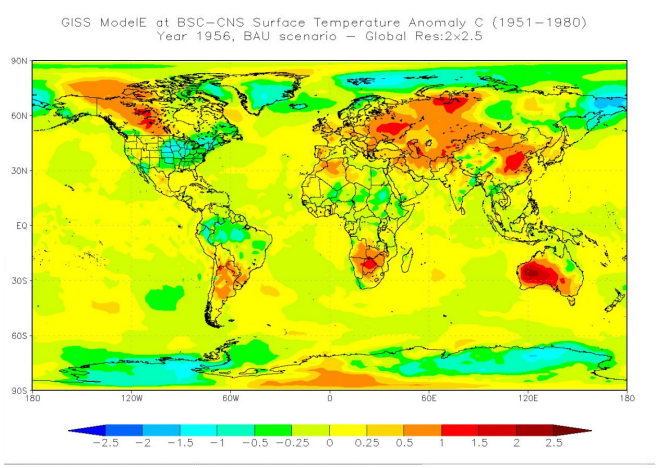


BSC-CNS: Ciencias de la Tierra

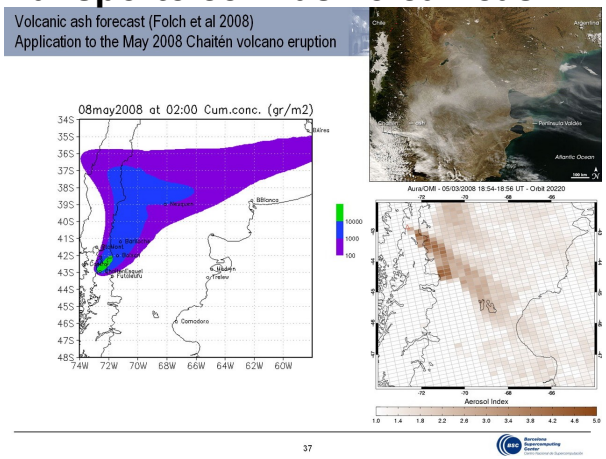
Pronóstico de Calidad del Aire



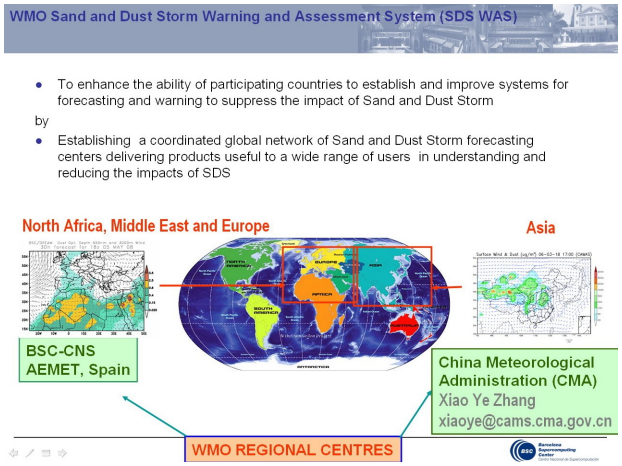
Cambio climático



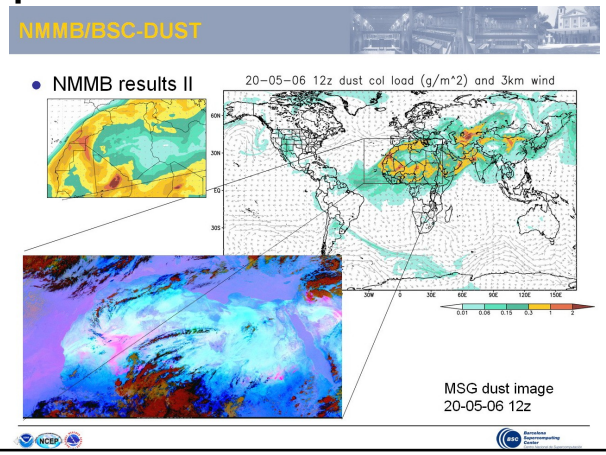
Transporte cenizas volcánicas



Transporte polvo mineral



Desarrollo modelo global de polvo mineral



Transferencia de tecnología: Estudios de impacto ambiental

Mérida, 27 Abril, 2010

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-CNS: Ciencias de los Computadores

Computer architecture:

- Superscalar and VLIW
- Hardware multithreading
- Design space exploration for multicore chips and Hw accelerators
- Transactional memory (Hw, Hw-assisted)
- SIMD and vector extensions/units

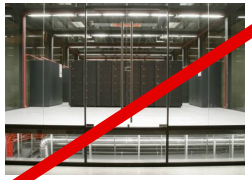
Benchmarking, analysis and prediction tools:

- Tracing scalability
- Pattern and structure identification
- Visualization and analysis
- Processor, memory, network, system

The Grid

Programming models:

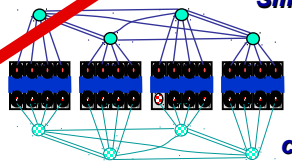
- Scalability of MPI and UPC
- OpenMP for multicore, SMP and ccNUMA
- DSM for clusters
- CellSs, streaming
- Transactional Memory
- Embedded architectures



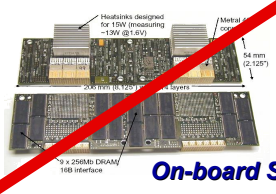
Future Exaflop systems

Large cluster systems

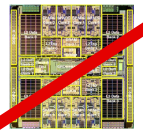
Small DMM



cc-NUMA



On-board SMP



Chip

Grid and cluster computing:

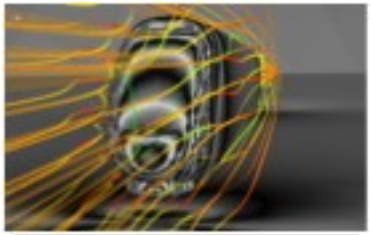
- Programming models
- Resource management
- I/O for Grid

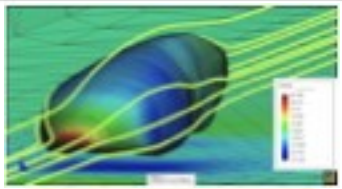
Operating environments:

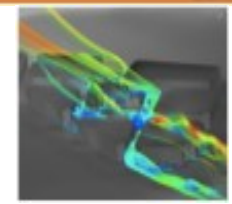
- Autonomic application servers
- Resource management for heterogenous workloads
- Coordinated scheduling and resource management
- Parallel file system scalability

BSC-CNS: Aplicaciones Científicas y de Ingeniería

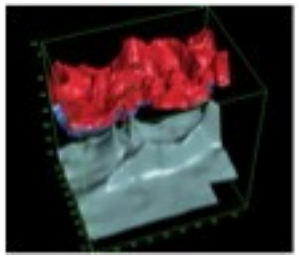
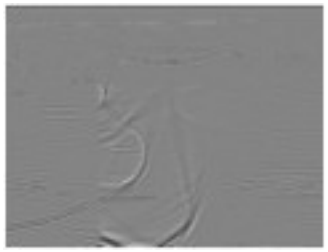
Computational Fluid Dynamics



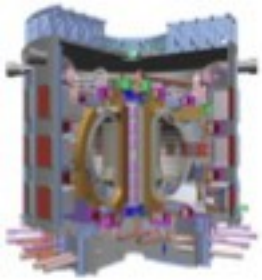
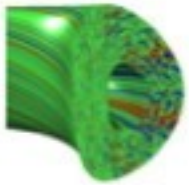




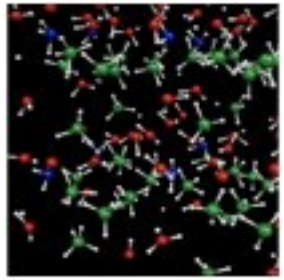
Geophysics

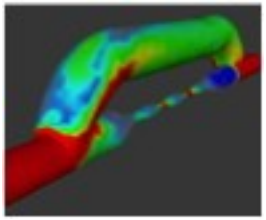

ITER: Plasma physics

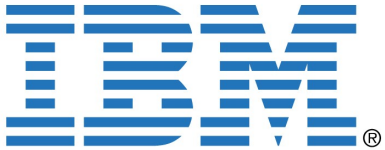
Ab-initio Molecular Dynamics



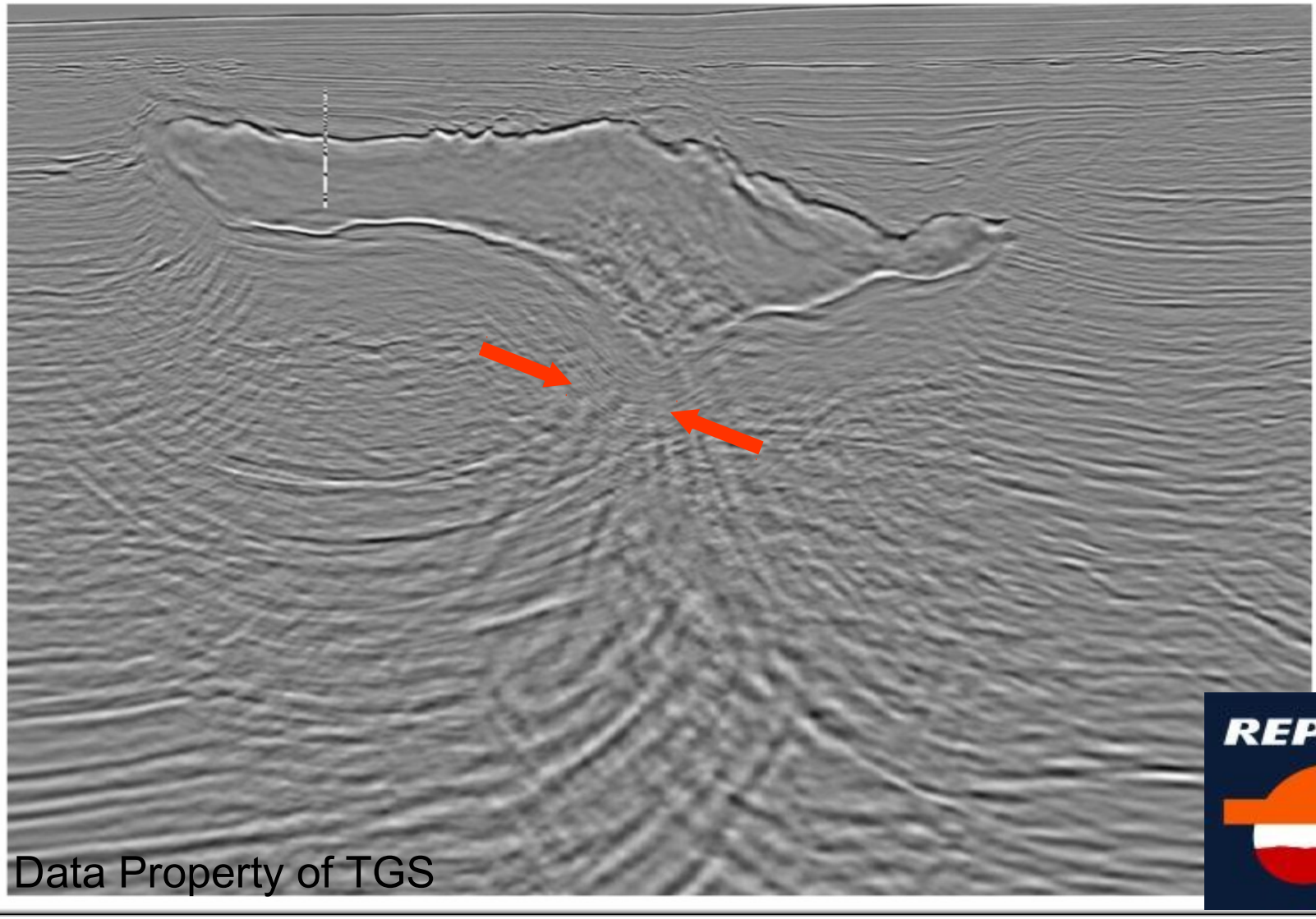
Bio-mechanics

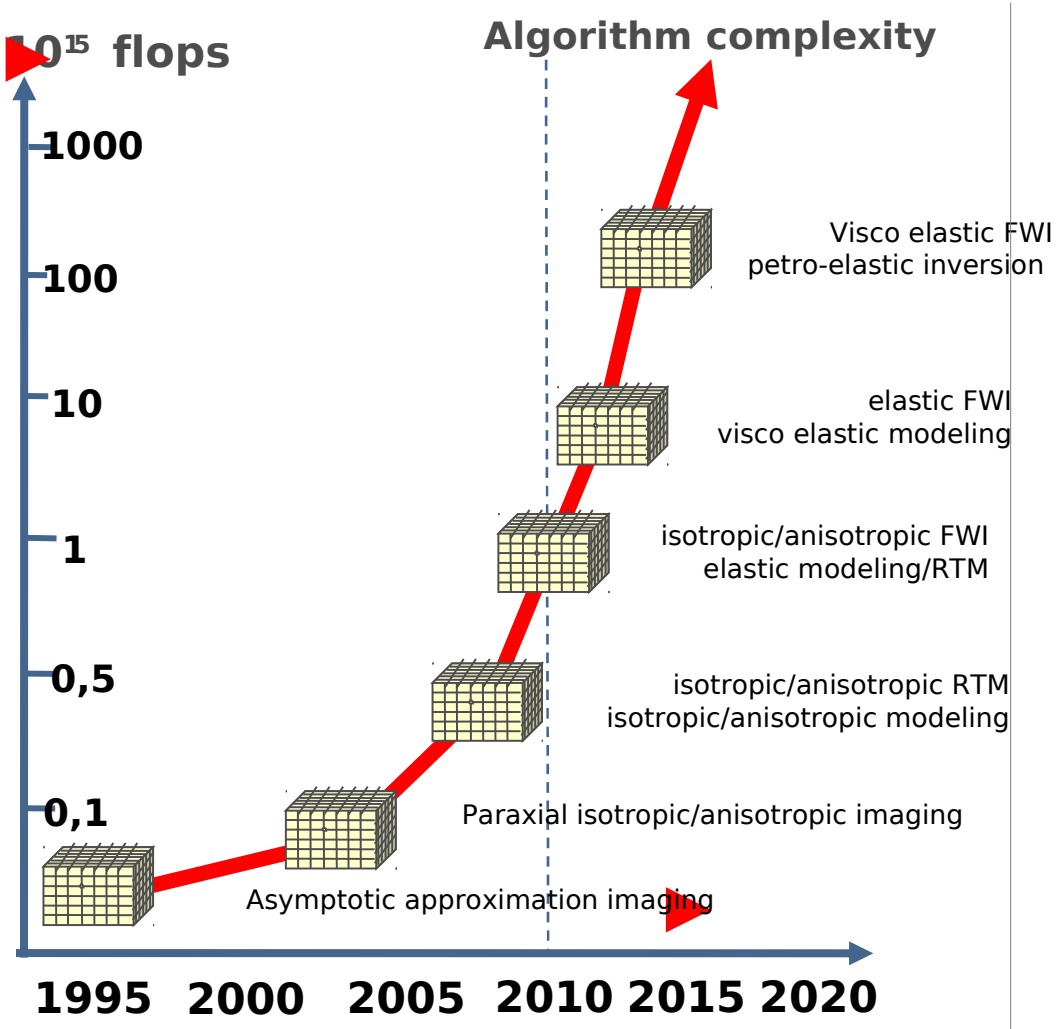

BSC-CNS: transferencia de tecnología a empresas



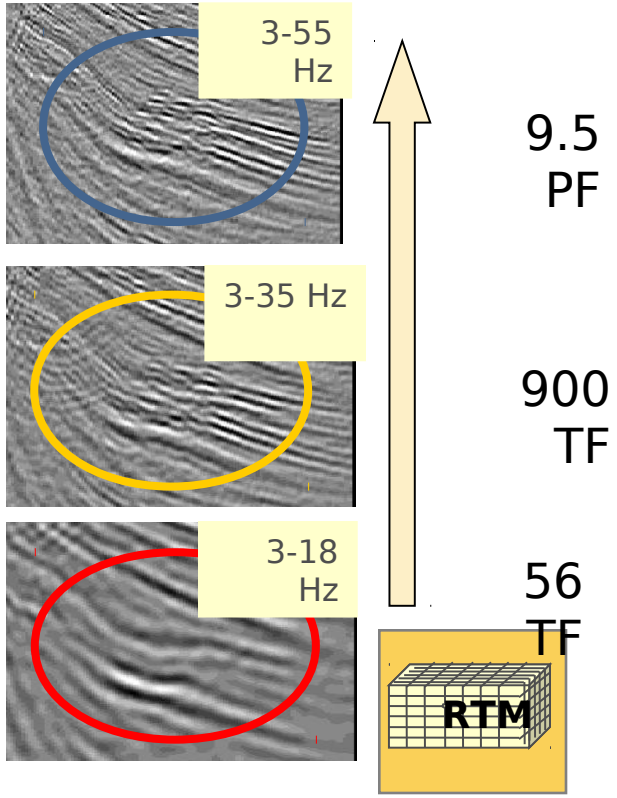
Kaleidoscope: WEM vs. RTM



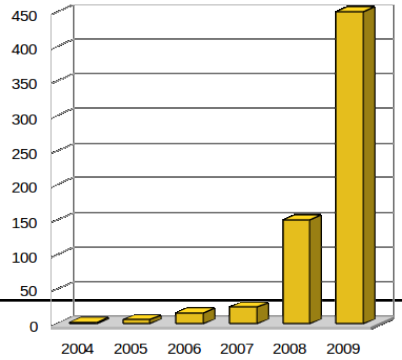
Industrial challenges in the Oil & Gas industry: Depth Imaging roadmap



Algorithmic complexity Vs. corresponding computing power



Substained performance for different frequency content over a 8 day processing duration



HPC Power
PAU (TF)
courtesy

European research projects: Overall picture

BSC-CNS: atractor de multinacionales

- Creado en Enero del 2008, a partir de una colaboración previa inciada en el 2006
  - Microsoft Research Cambridge y Redmond

BSC Microsoft Research Centre

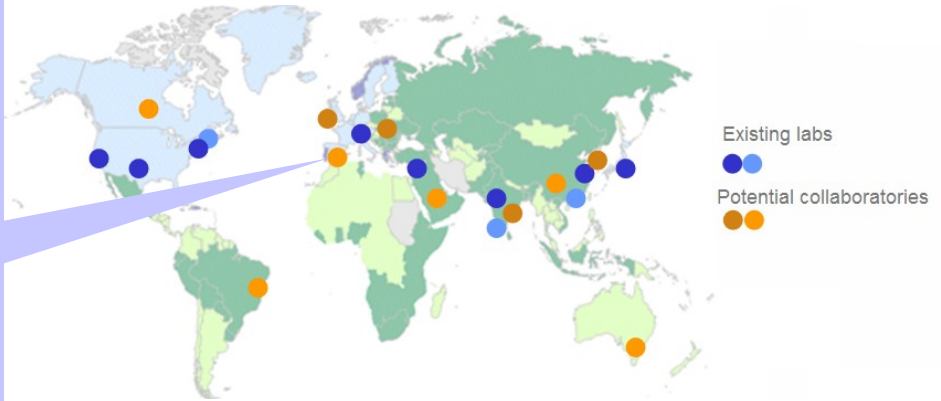
- Soporte de la arquitectura a la ejecución eficiente de programas paralelos en arquitecturas multicore
  - Transational Memory
  - Advanced architectures for mobile devices
  - Object-oriented computer architecture
- Personal dedicado:
  - 3 investigadores senior y más de 20 estudiantes de doctorado

BSC-CNS: Atractor de Multinacionales

IBM Collaboratory

Establish an IBM leading-edge applied research facility in Barcelona

- focus on solutions-driven systems and microprocessor architecture research to enable a ***Smarter Planet.***
- Emphasis on innovative parallel and heterogeneous computing architectures.
- applied in selected strategic domains



a Global Initiative

Strategic Application Domains

Life science and health care

- IBM has strong focus on health care
- BSC has long dedication in Life Sciences
- Barcelona is the capital of BioCat, BioRegió de Catalunya
- Leverage IBM Research advances in the biomedical area

Finance and banking applications

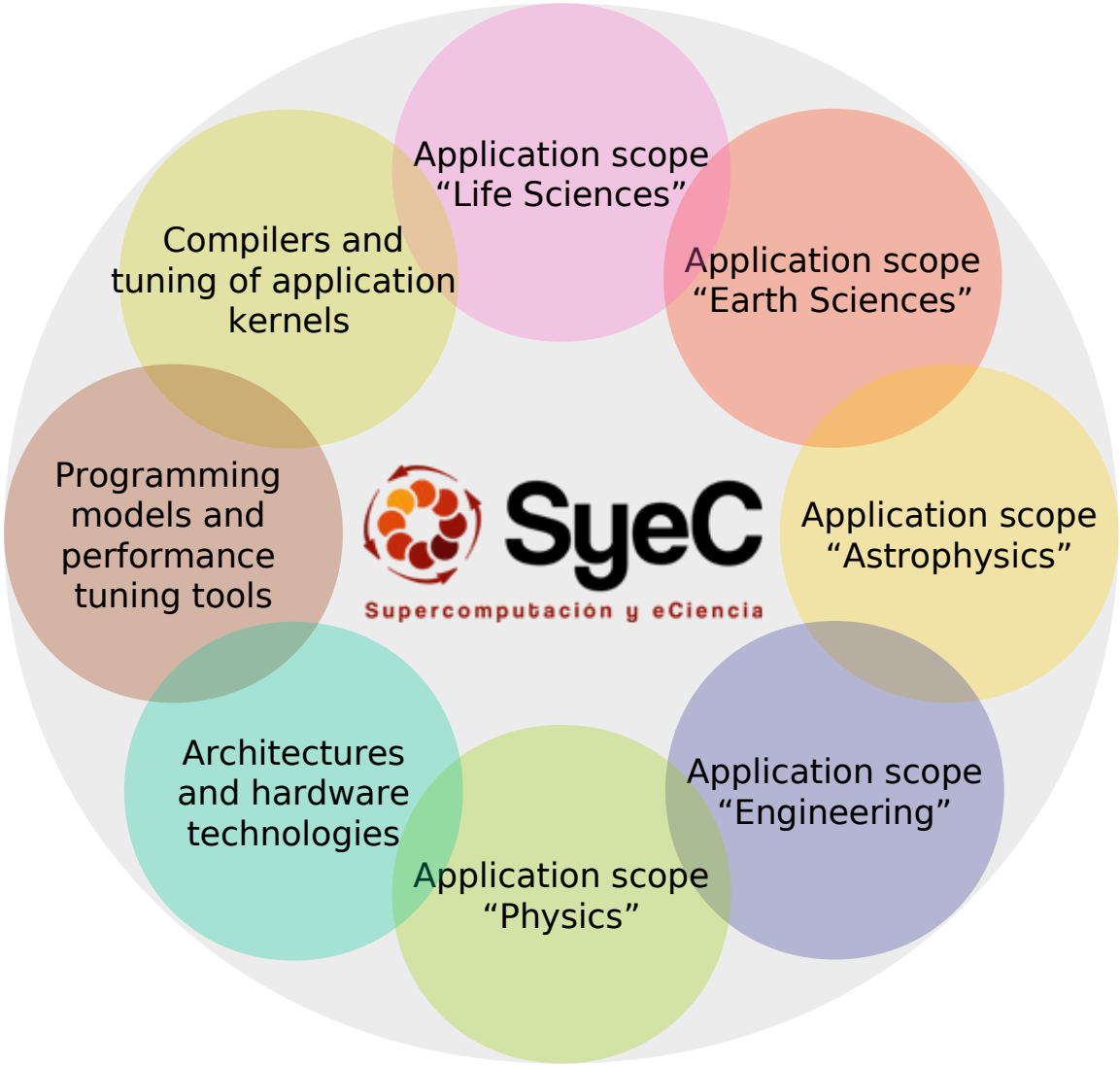
- Increasing importance of the digital economy in the Smarter Planet
- Long expertise in Spain on banking and finance
- Strong relationships between IBM Spain and local financial institutions
- Leverage IBM Research advances in business analytics and modeling

Smarter cities

- Technology innovations that infuse new intelligence into the infrastructure
- Objective of making cities smarter and more efficient
- Of interest to metropolitan areas such as Barcelona
- Leverage IBM Research advances in modeling complex systems

BSC-CNS: vertebrador de la investigación en supercomputación en España

- Supercomputación y eCiencia
  - 22 grupos de élite
  - Más de 120 investigadores seniors
  - Más de 300 estudiantes de doctorado



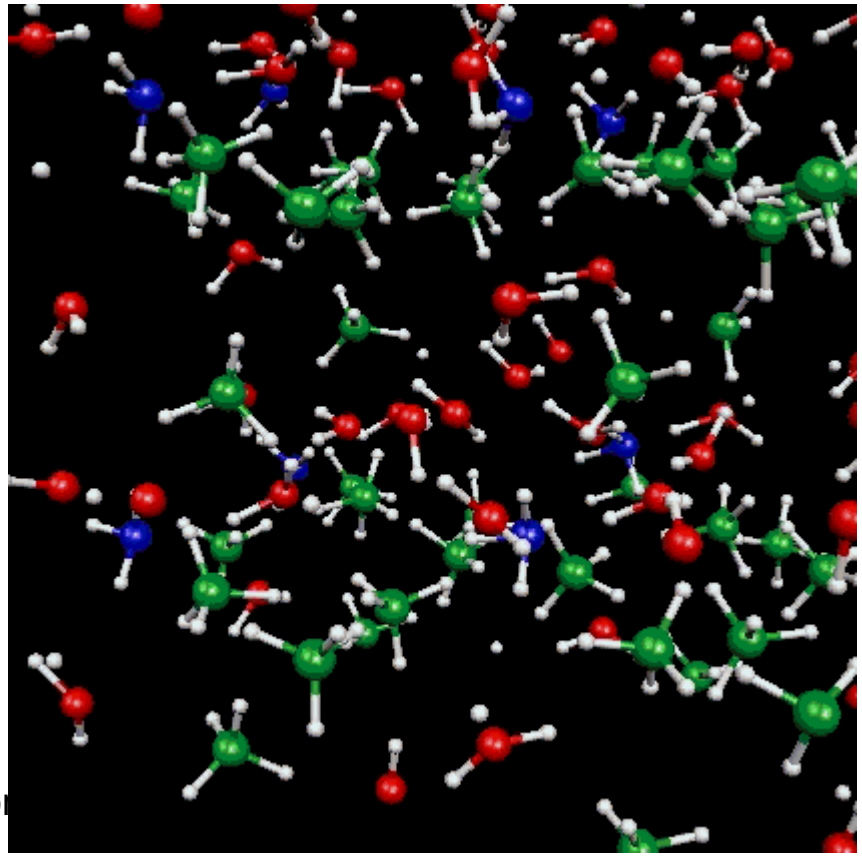
Consolider

PROGRAMA ingenio 2010

SIESTA project

- Ab-inito DFT molecular dynamics code BSC working on its development

- Example: Neptune Mid-layer
  - H_2O+ NH_3+ CH_4
  - Temperature
    - 1500 ºK – 2500 ºK
  - Pressure
    - 0.15 GPa – 60 GPa
  - Simulated time
    - 10 ps equivalent to 20,000 molecular dynamic steps
  - Number of atoms
    - 1269 atoms (100 processors, 2007)
    - Now, more than 500.000 atoms using mo than 1000 processors)



Objective

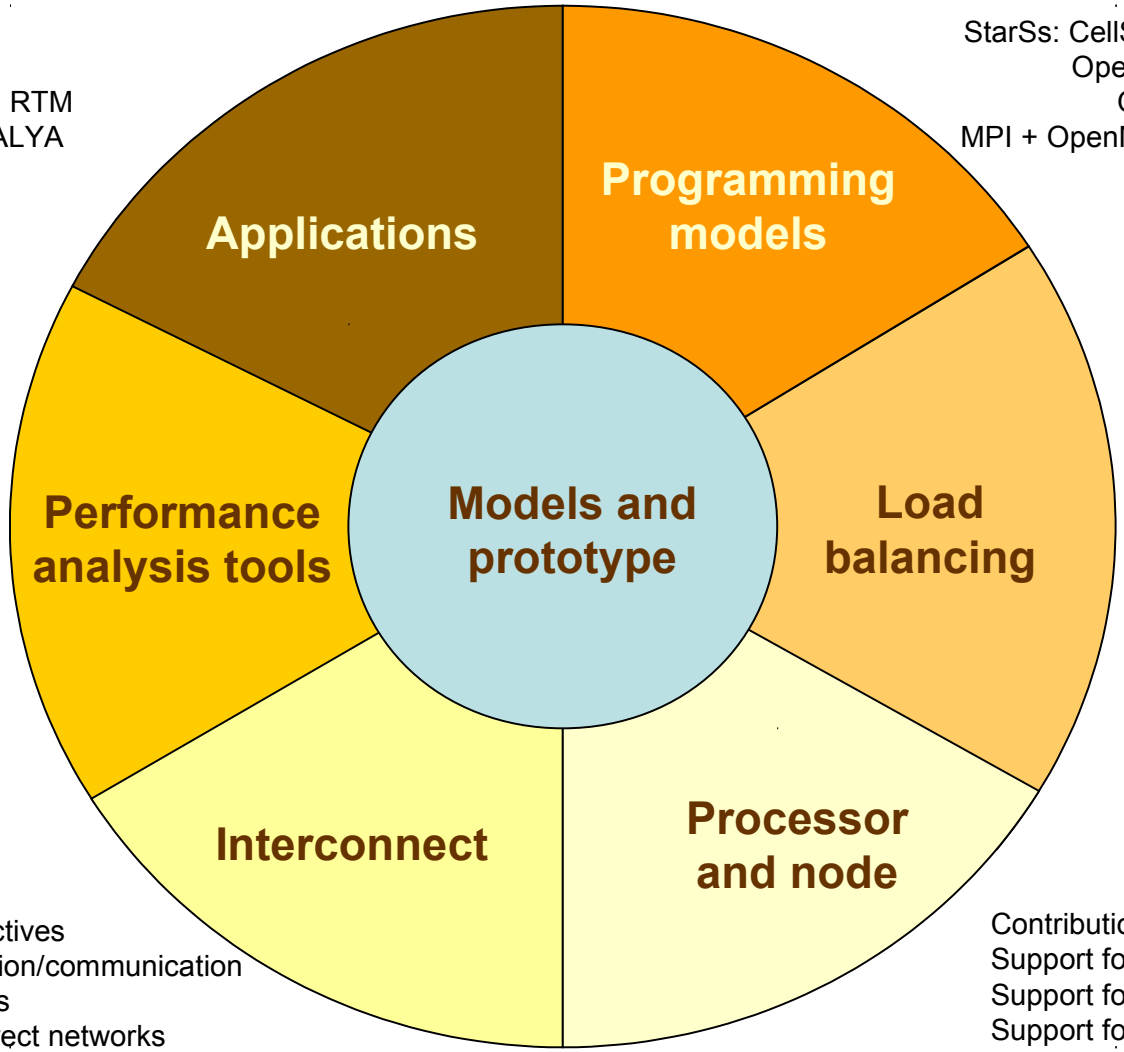
- Design a 10+ Petaflops Supercomputer for 2010-11
- Barcelona Supercomputing Center – Centro Nacional de Supercomputación IBM Cooperation
- Spanish position within PRACE

MareIncognito: Project structure

4 relevant apps:
- Materials: SIESTA
- Geophisics imaging: RTM
- Comp. Mechanics: ALYA
- Plasma: EUTERPE

General kernels

StarSs: CellSs, SMPSs
OpenMP@Cell
OpenMP++
MPI + OpenMP/StarSs



Automatic analysis
Coarse/fine grain prediction
Sampling
Clustering
Integration with Peekperf

Coordinated scheduling:
Run time,
Process,
Job
Power efficiency

Contention, Collectives
Overlap computation/communication
Slimmed Networks
Direct versus indirect networks

Contribution to new Cell design
Support for programming model
Support for load balancing
Support for performance tools
Issues for future processors

Categories for Runtime systems

- Category I: Uniquely Exascale
  - Load balance (including tolerance to noise and temporary shortage of resources (i.e. as a result of faults))
  - Hierarchical execution models and scheduling
  - Scale/optimize Comms: MPI, routing, comm. schedule,...
- Category II: Exascale plus trickle down
  - Asynchrony, overlap
  - Mem. Mgmt. & Locality scheduling
  - Heterogeneity: scheduling
- Category III: Primarily Sub-exascale.
  - Fine grain mechanisms @ node level (for thread mgmt & synch support)

Back to Babel?

Book of Genesis

"Now the whole earth had **one language** and the same words" …

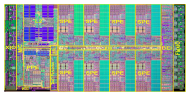
…"Come, let us make bricks, and burn them thoroughly. "…

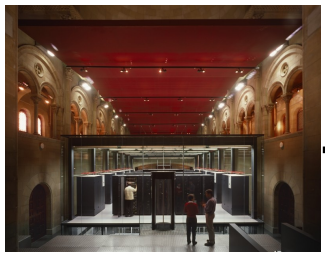
…"Come, let us build ourselves a city, and a tower with its top in the heavens, and let us make a name for ourselves"…

And the LORD said, "Look, they are one people, and they have all one language; and this is only the beginning of what they will do; nothing that they propose to do will now be impossible for them. Come, let us go down, and **confuse their language** there, so that they will not understand one another's speech."

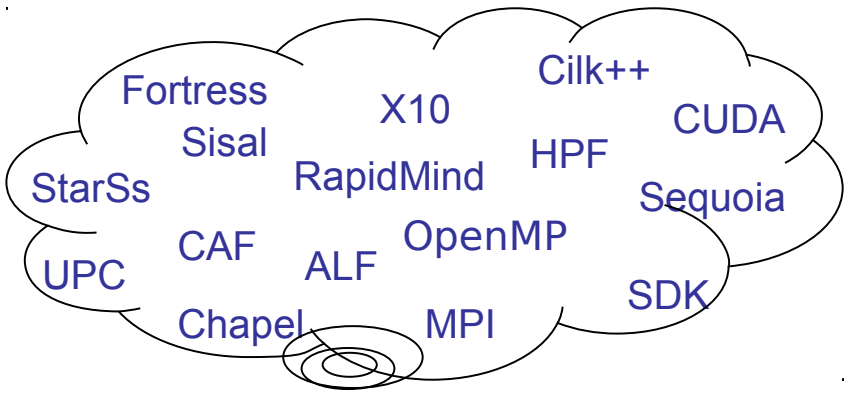
The computer age

Fortran & MPI



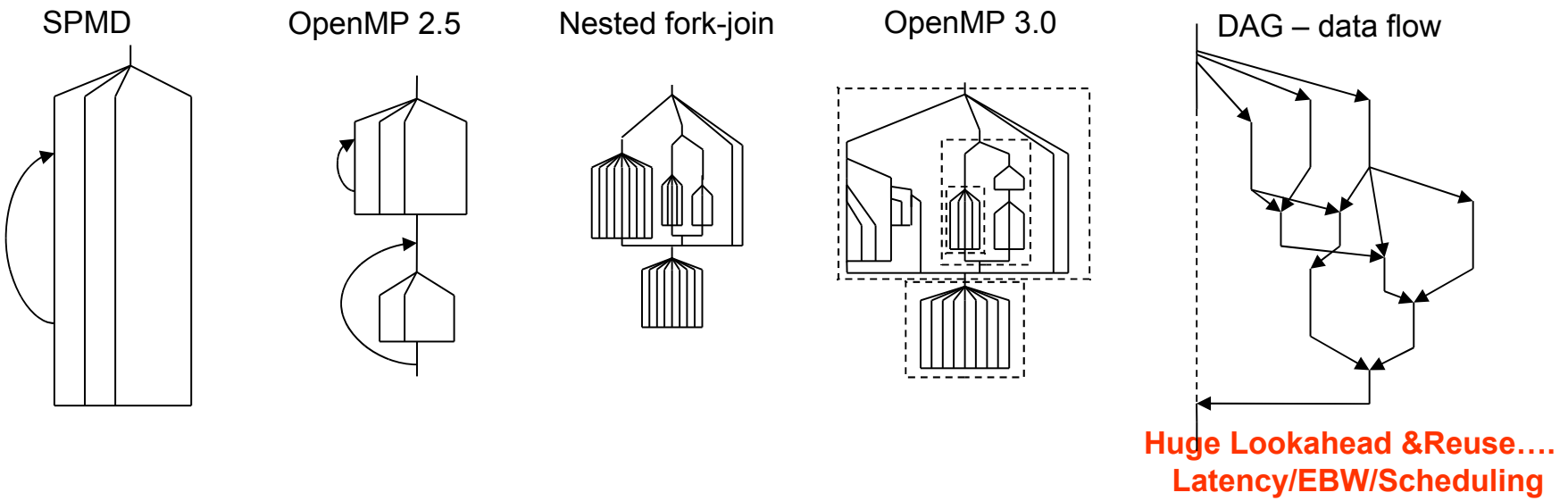


++



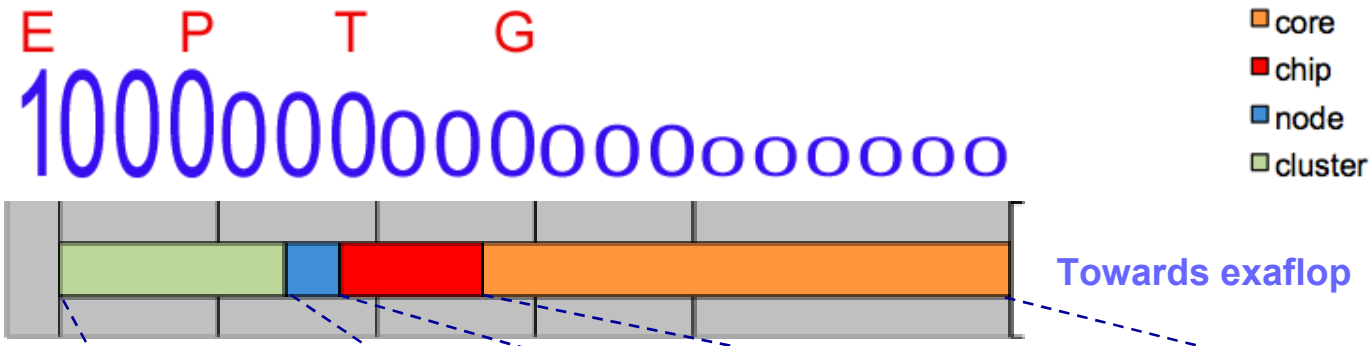
Different models of computation …….

- The dream for automatic parallelizing compilers not true …
- … so programmer needs to express opportunities for parallel execution in the application



- And … asynchrony (MPI and OpenMP too synchronous):
  - Collectives/barriers multiply effects of microscopic load imbalance, OS noise,…

The holistic approach …



Applications	
Performance Tools	
Programming Model	Merge … level … A…
Load Balancing	
Interconnection	*Detect data pr…*
Processor/node architecture	*Efficient sup… … thread…*

Latency

Bandwidth

Parallelism

Scheduling

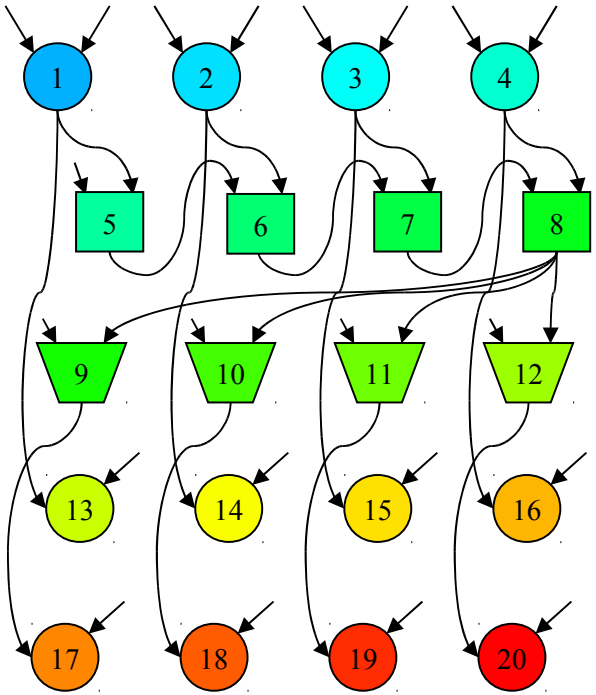
Power

StarSs: … generates task graph at run time …

```
#pragma css task input(A, B) output(C)
void vadd3 (float A[BS], float B[BS],
            float C[BS]);
#pragma css task input(sum, A) output(B)
void scale_add (float sum, float A[BS],
                float B[BS]);
#pragma css task input(A) inout(sum)
void accum (float A[BS], float *sum);
```

```
for (i=0; i<N; i+=BS)              // C=A+B
   vadd3 ( &A[i], &B[i], &C[i]);
...
for (i=0; i<N; i+=BS)             // sum(C[i])
   accum (&C[i], &sum);
...
for (i=0; i<N; i+=BS)             // B=sum*E
   scale_add (sum, &E[i], &B[i]);
...
for (i=0; i<N; i+=BS)             // A=C+D
   vadd3 (&C[i], &D[i], &A[i]);
...
for (i=0; i<N; i+=BS)             // E=C+F
   vadd3 (&C[i], &F[i], &E[i]);
```

Task Graph Generation

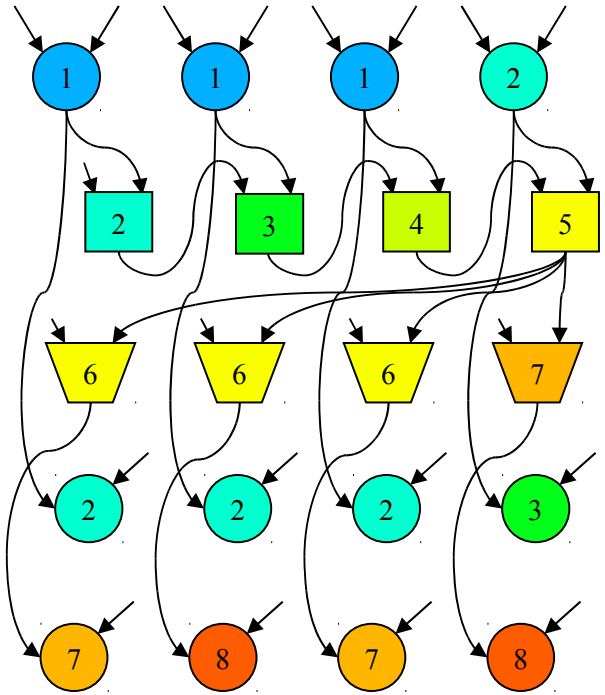


StarSs: … and executes as efficient as possible …

```
#pragma css task input(A, B) output(C)
void vadd3 (float A[BS], float B[BS],
            float C[BS]);
#pragma css task input(sum, A) output(B)
void scale_add (float sum, float A[BS],
                float B[BS]);
#pragma css task input(A) inout(sum)
void accum (float A[BS], float *sum);
```

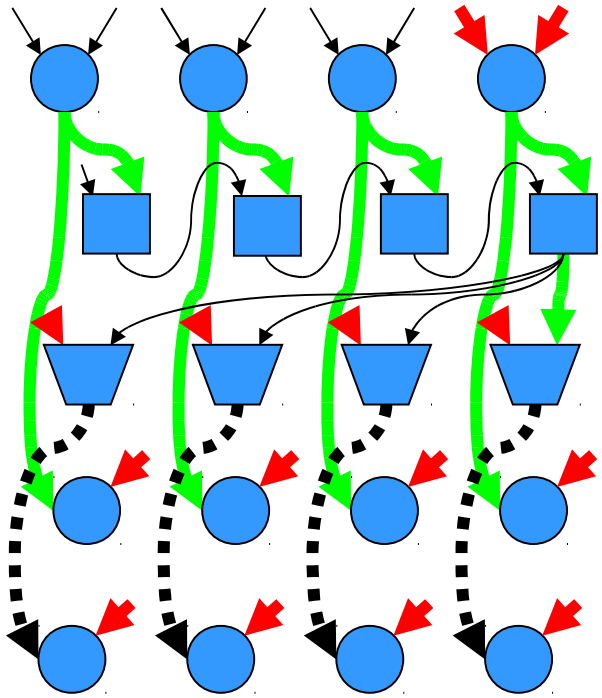
```
for (i=0; i<N; i+=BS)              // C=A+B
   vadd3 ( &A[i], &B[i], &C[i]);
...
for (i=0; i<N; i+=BS)             // sum(C[i])
   accum (&C[i], &sum);
...
for (i=0; i<N; i+=BS)             // B=sum*E
   scale_add (sum, &E[i], &B[i]);
...
for (i=0; i<N; i+=BS)             // A=C+D
   vadd3 (&C[i], &D[i], &A[i]);
...
for (i=0; i<N; i+=BS)             // E=C+F
   vadd3 (&C[i], &F[i], &E[i]);
```

Task Graph Execution



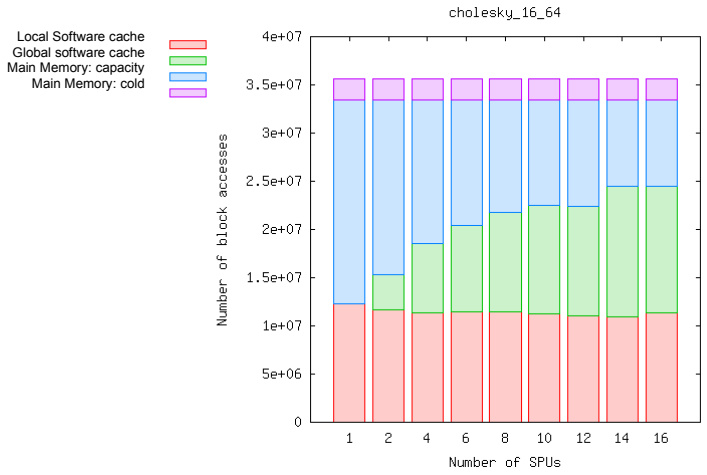
StarSs: … benefiting from data access information

- **Flat global address space seen by programmer**
- Flexibility to dynamically traverse dataflow graph “optimizing”
  - Concurrency. Critical path
  - Memory access
- Opportunities for
  - Prefetch
  - Reuse
  - Eliminate antidependences (rename)
  - Replication management

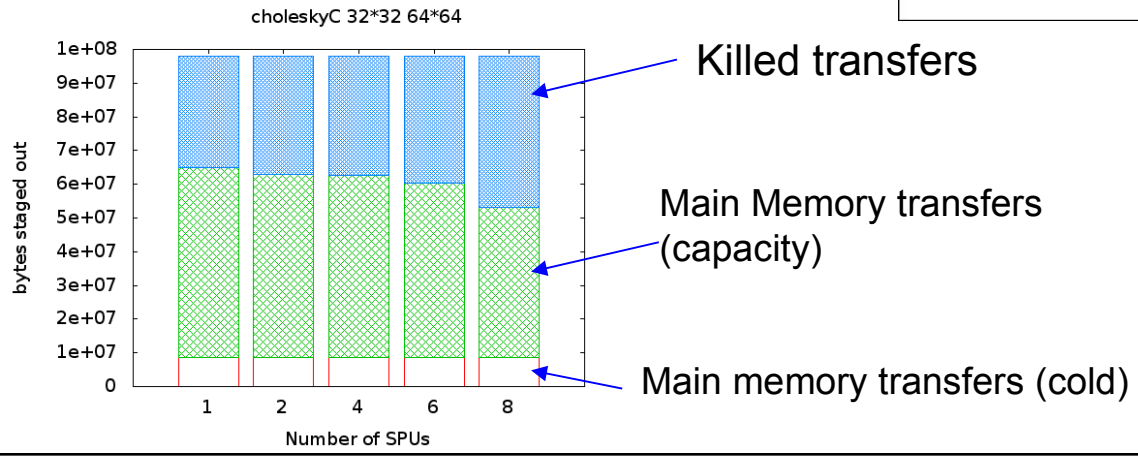


Benefiting from data access information

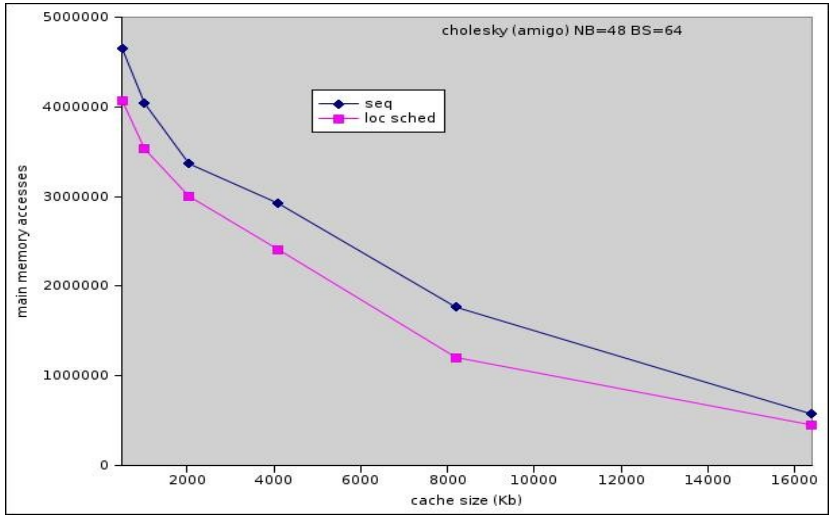
- Software cache
  - Local vs. global across SPEs

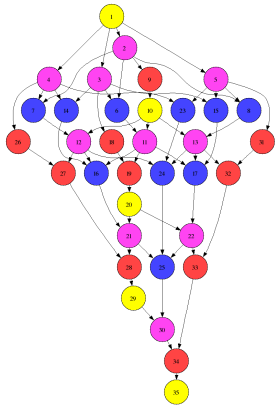


- Lazy renaming (virtual registers)



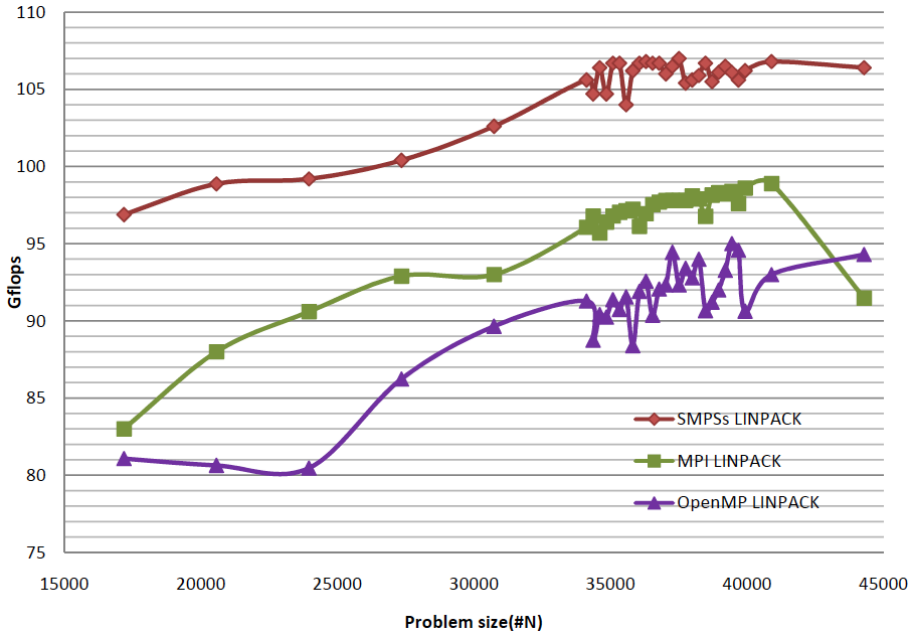
- Schedule to maximize reuse
  - Graph traversal optimizing locality
  - Data reuse vs. critical path

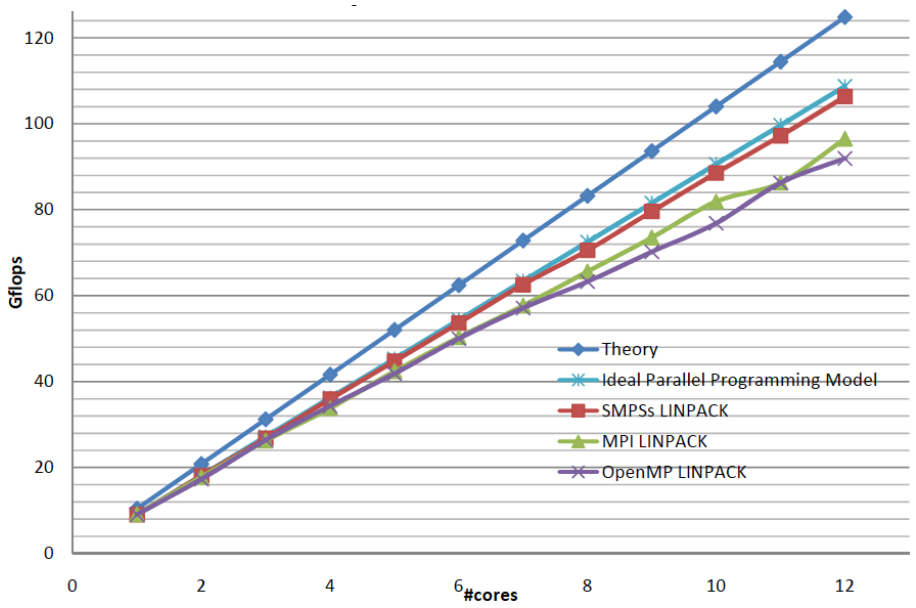




SMPSs: an implemention for SMP and multicores

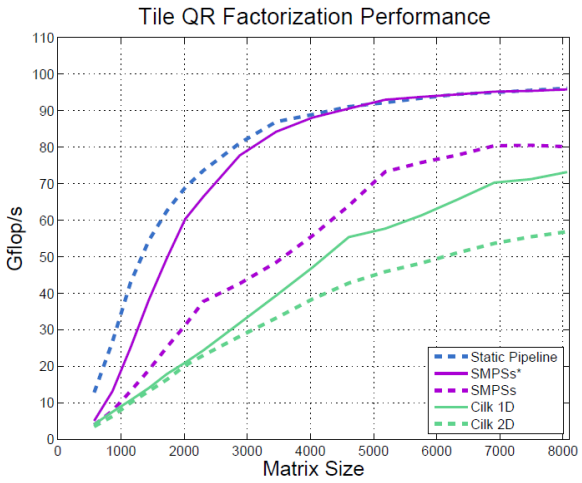
- HPL Linpack: comparison of SMPSs, OpenMP and MPI on a dual socket 6-core

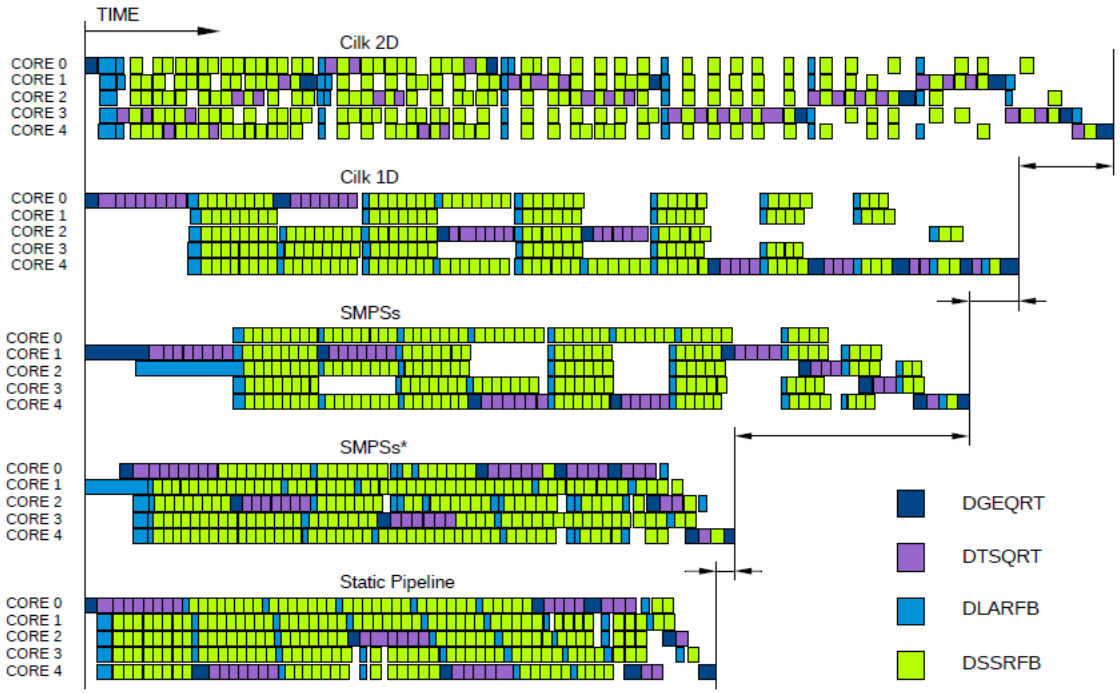




SMPSs: e.g. QR factorization

- Run on quad-socket quad-core Intel Tigerton
- Performance comparable to static, hand-written scheduling

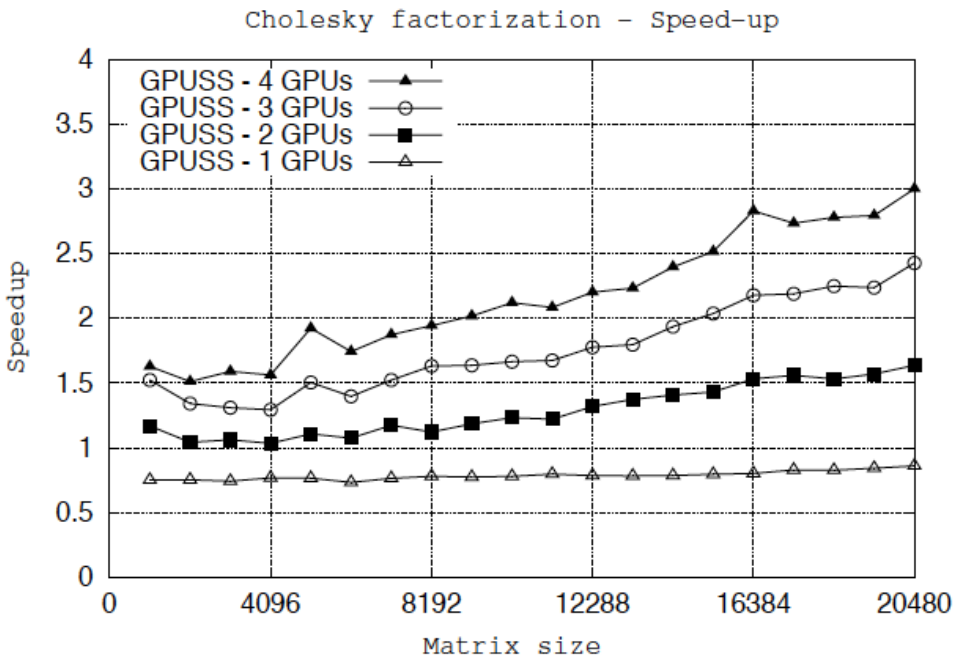




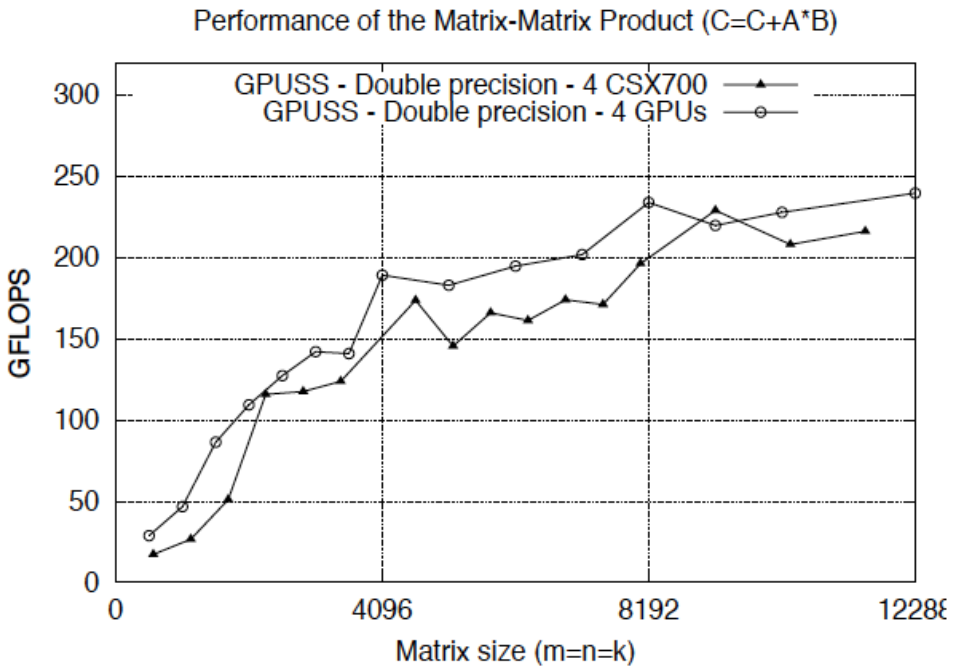
* Kurzak et al, “Scheduling Linear Algebra Operations on Multicore Processors”, LAPACK Working Note 213

StarSs: implementations for GPU and FP coprocessors

- Tesla s1070 with 4 x GT200



- ClearSpeed 4 x CSX700

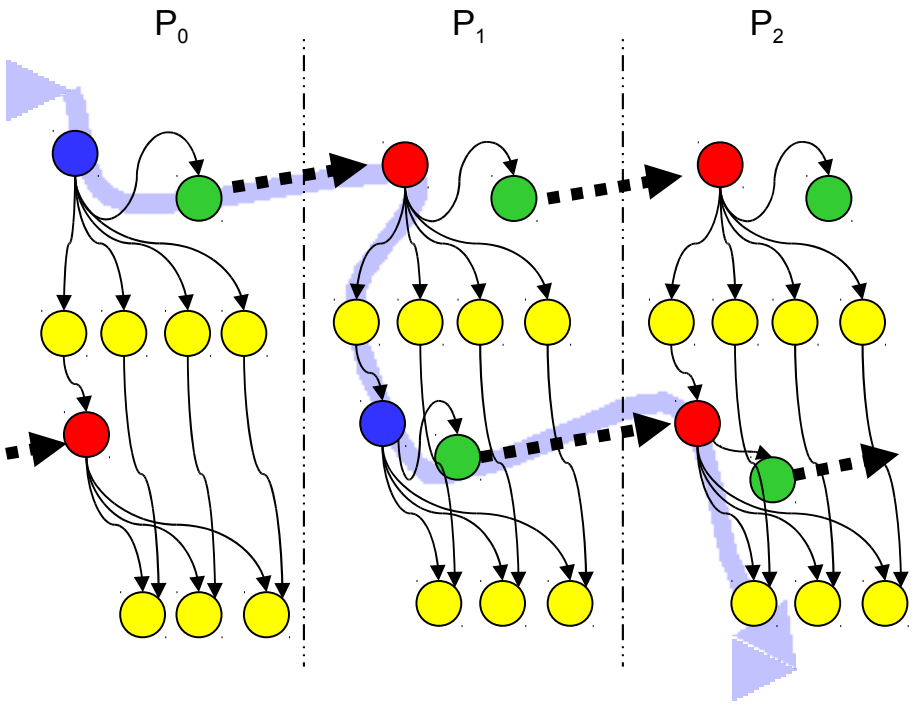


aboration with E. Quintana, F. Igual and R. Mayo from U. Jaume I (Castellon). An Extension of the StarSs Programming for Platforms with Multiple GPUs. Europar-2009.

Hybrid MPI/SMPSs: Linpack example

- Overlap communication/computation
- Extend asynchronous data-flow execution to outer level
- Automatic lookahead

```
…
for (k=0; k<N; k++) {
   if (mine) {
      Factor_panel(A[k]);
      send (A[k])
   } else {
      receive (A[k]);
      if (necessary) resend (A[k]);
   }
   for (j=k+1; j<N; j++)
      update (A[k], A[j]);
…
```

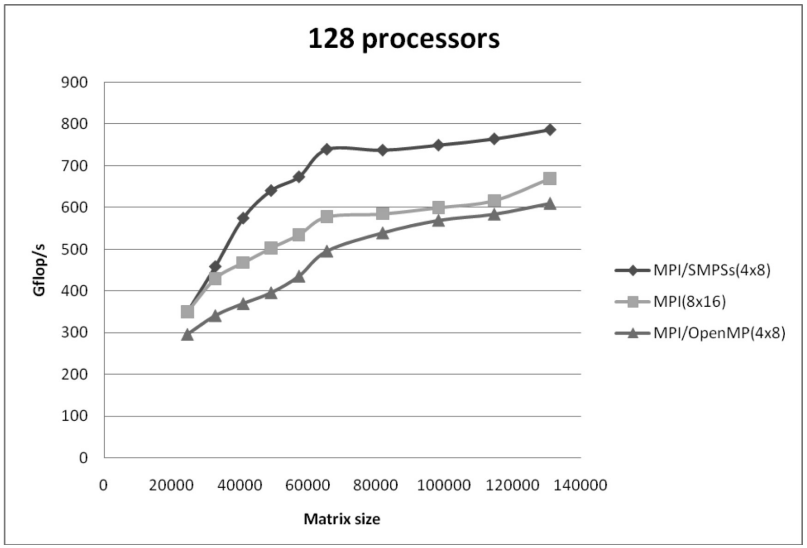


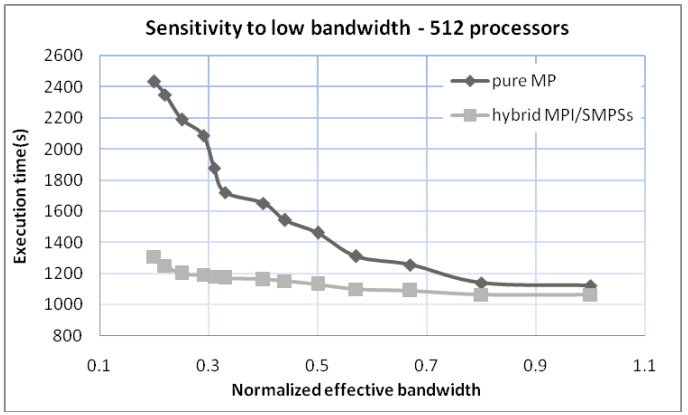

```
#pragma css task inout(A[SIZE])
void Factor_panel(float *A);
#pragma css task input(A[SIZE]) inout(B[SIZE])
void update(float *A, float *B);
```

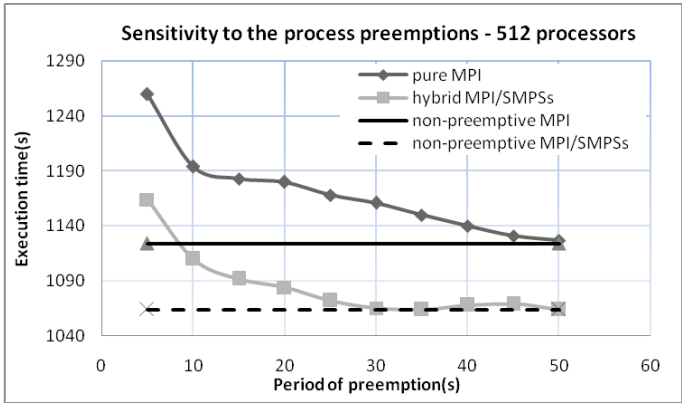
```
#pragma css task input(A[SIZE])
void send(float *A);
#pragma css task output(A[SIZE])
void receive(float *A);
#pragma css task input(A[SIZE])
void resend(float *A);
```

Hybrid MPI/SMPSs: Green Linpack

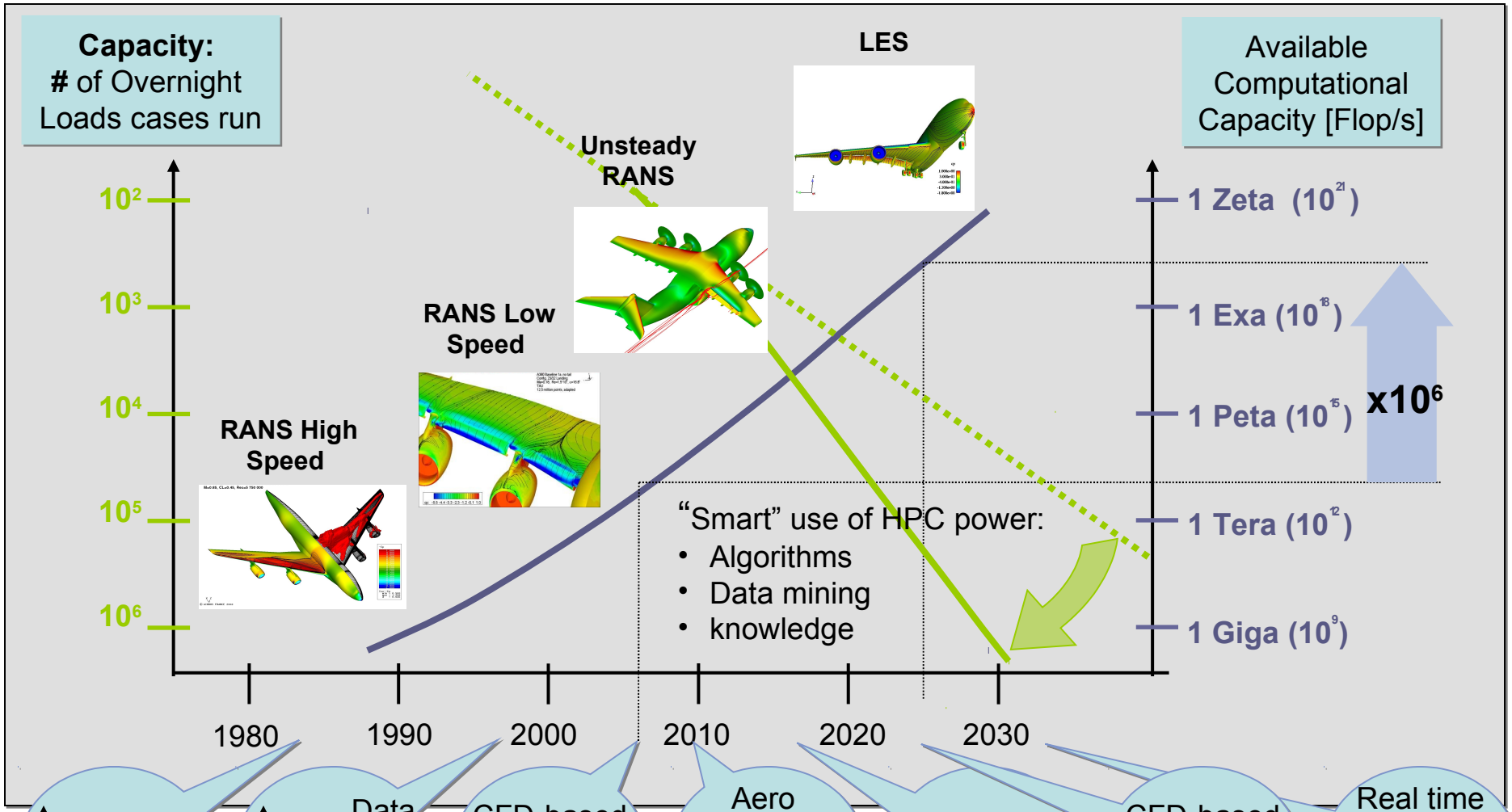
- Performance
  - Higher at smaller problem sizes
  - Improved load balance (less processes)
  - Higher IPC
  - Overlap communication/computation
- Tolerance to bandwidth and OS noise





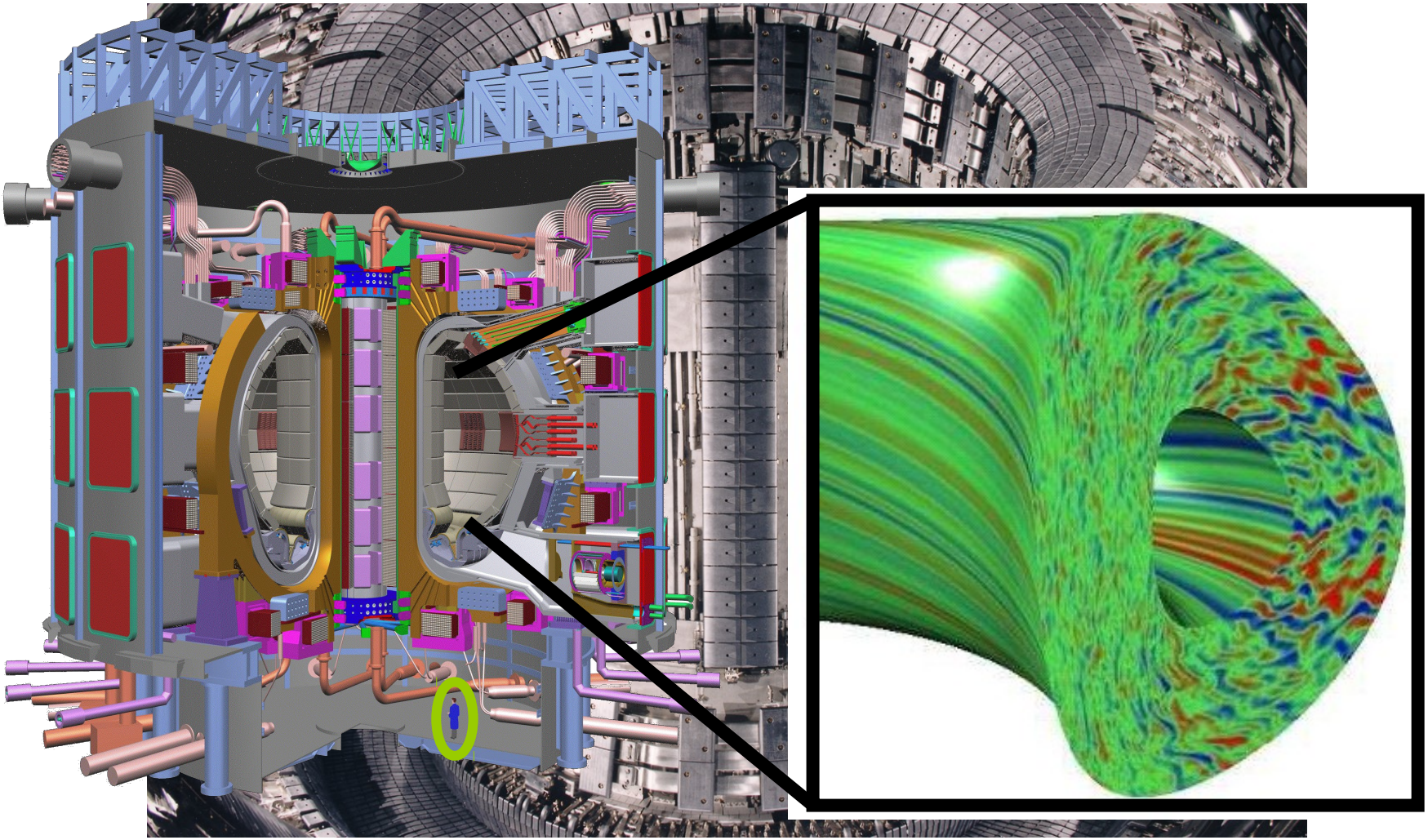


High Performance Computing as key-enabler



Courtesy AIRBUS France

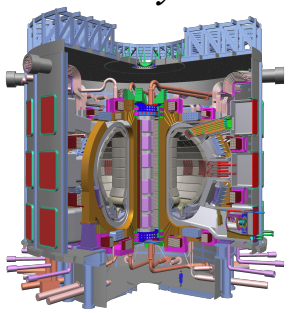
Diseño del ITER



TOKAMAK (JET, Oxford)

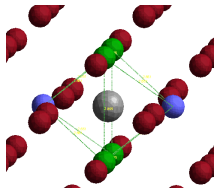
Material ageing: PWI, radiation damage

m^3 10years



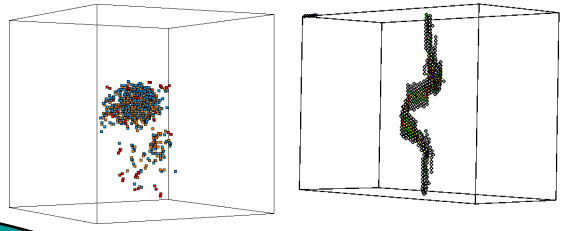
nm^3 ps

Ab initio

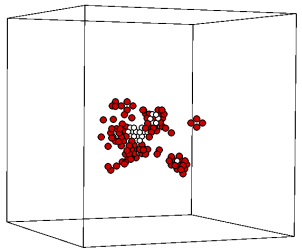


10-30 nm^3 ns

Siesta Molecular dynamics

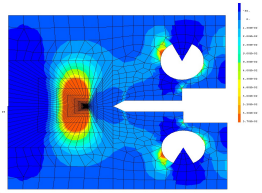


s - h



Finite elements

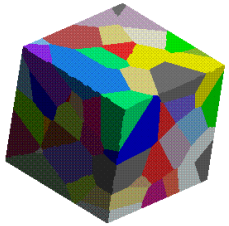
cm^3



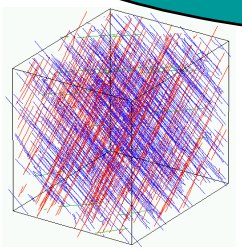
Alya

Multi-Scale Simulation

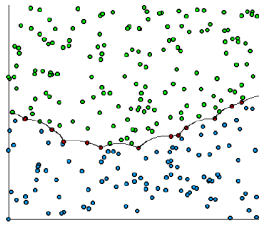
Micro-macro

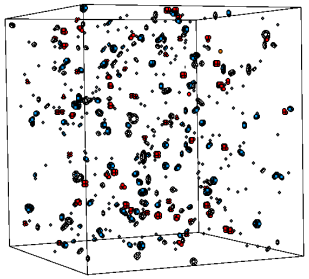


μm^3



Dislocations Dynamics



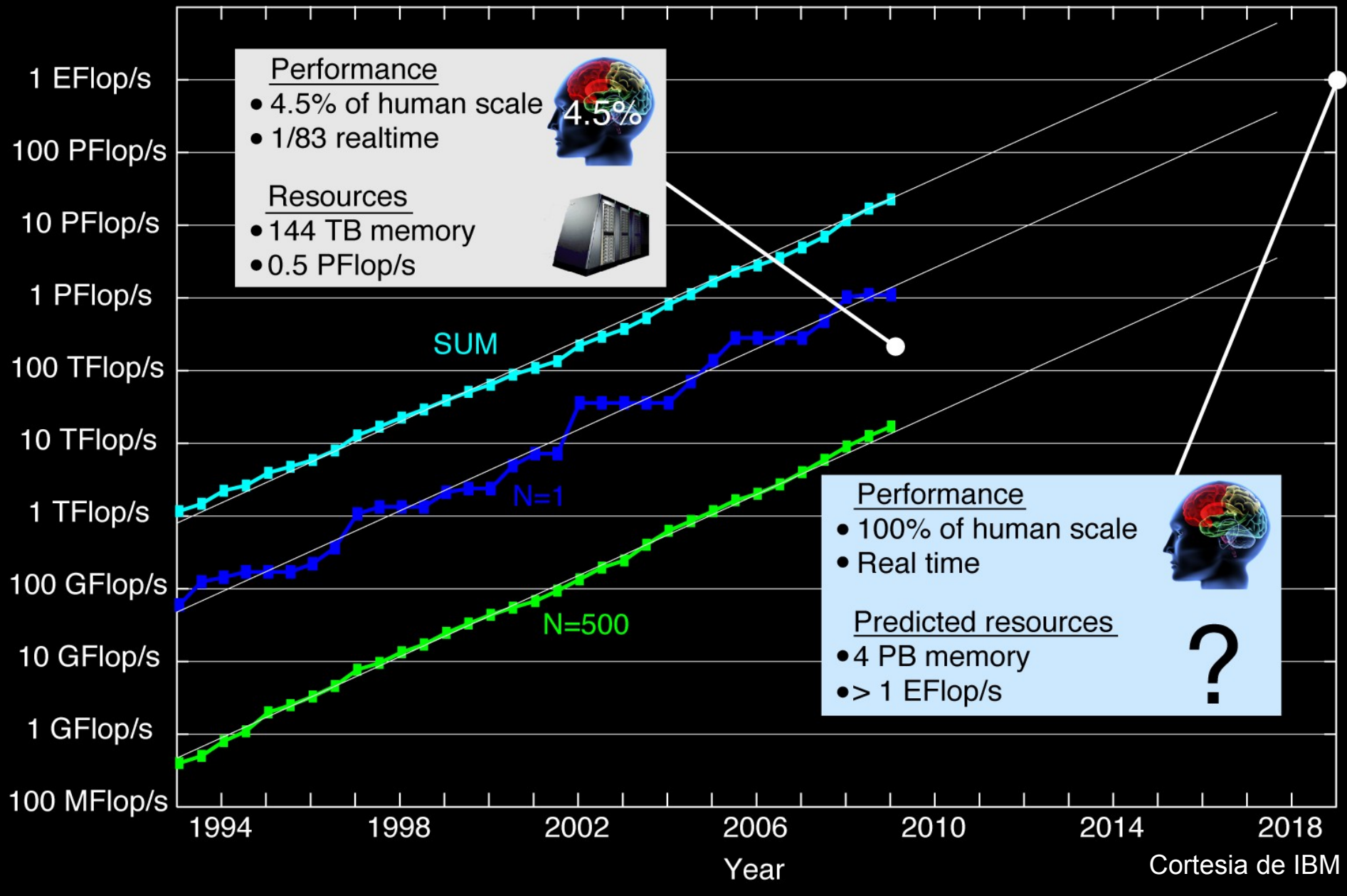


30-100 nm^3 h-year

MonteCarlo Statistical Modelling

Mesoscale

Supercomputación, teoría y experimentación



Cortesia de IBM

Weather, Climate and Earth Sciences: Roadmap

2009

Resolution : 80 km

Memory: ≈ 10 GB

Storage: ≈ 8 TB

NEC-SX9 48 vector procs: ≈ 40 days run

2015

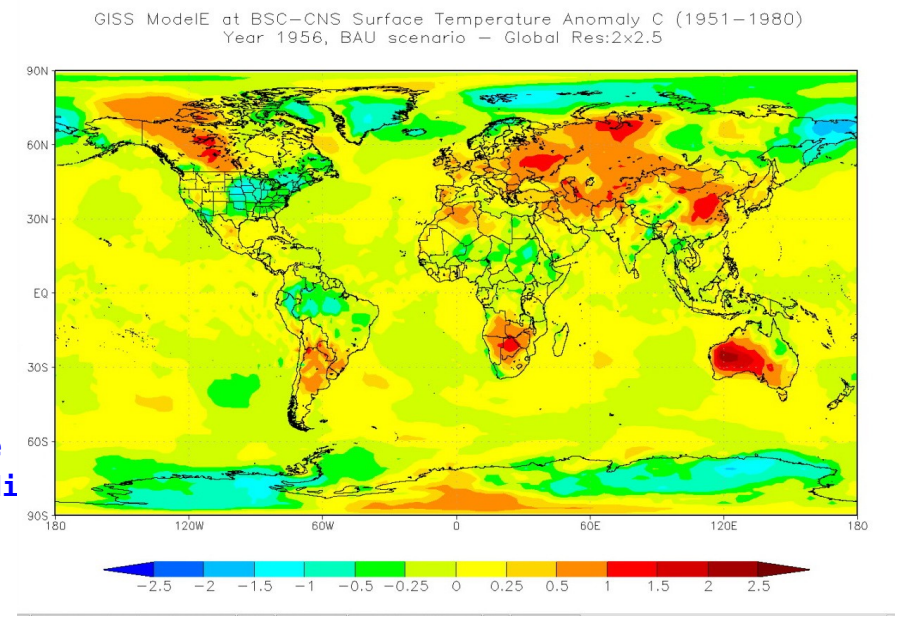
Resolution : 20 km

MemSory: ≈ 3,5 TB

Storage: ≈ 180 TB

High resolution model with complete carbon cycle

Challenges: data viz and post-processing, data di



2020

Resolution : 1 km

Memory: ≈ 4 PB

Storage: ≈ 150 PB

Higher resolution with global cloud resolving model

Challenges: data sharing, transfer memory management, I/O management

Master on HPC

Kernel

- Supercomputer Architectures
- Methods and Algorithms for Parallel Programming
- Optimization and Parallelization of Numerical Simulations

Free-options

- High performance Computational Mechanics
- Performance tuning and analysis tools
- Data Mining 2
- Seminar on Supercomputing I, II, III

Applications

- Computational Astrophysics
- Bionformatics
- Earth Sciences
- Applications of Computational Astrophysics
- Applications of Bionformatics
- Applications of Earth Sciences

THESIS

Barcelona Computing Week, July 5-9, 2010

Programming and Tuning Massively Parallel Systems

- Instructors:
  - Wen-mei Hwu, University of Illinois
  - David B. Kirk, NVIDIA Corporation
- Audience:
  - Three parallel tracks specially designed for beginners, advanced and teachers profiles
- Programming Languages:
  - CUDA, OpenCL, OpenMP, StarSs
- Numerical Methods:
  - FFT, Graph, Tiling, Grid, Montecarlo, FDTD, Sparse matrices…
- Hands-on Labs:
  - Afternoon labs with teaching assistants for each audience/level
- Case Studies:
  - Molecular Dynamics, MRI Reconstruction, RTM Stencil, Dense Linear Systems…

B C W

PUMPS Summer School

Barcelona Supercomputing Center

Universitat Politècnica de Catalunya

Barcelona , July 5-9, 2010

http://bcw.ac.upc.edu

http://bcw.ac.upc.edu

NVIDIA

HiPEAC COMPILATION ARCHITECTURE

SEVENTH FRAMEWORK PROGRAMME

HiPEAC COMPILATION ARCHITECTURE

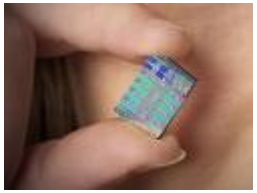
Sixth International Summer School on Advanced Computer Architecture and Compilation for Embedded Systems

11-17 July, 2010 Terrassa (Barcelona), Spain

- "HiPEAC Summer School" : **one week summer school for computer architects and compiler builders**
- 12 courses in 4 course slots and an evening activities
- Keynotes and invited talks
- Poster session

	Lecturer	Course title
Slot 1	Michael Scott	Transactional Memory
	Vivek Sarkar	Multicore Programming Models and their Compilation Challenges
	Andreas Herkersdorf	Application-Specific (MP)SoC Architectures for Internet Networking
Slot 2	David Brooks	Variation-Aware Processor Design
	Derek Chiou	Fast and Accurate Computer System Simulators
	Scott Mahlke	Compilation for Multicore Processors
Slot 3	Dan Sorin	Fault Tolerant Computer Architecture
	Donatella Sciuto	FPGA-based reconfigurable computing
	Steven Hand	System Virtualization
Slot 4	Theodore Ts'o	File Systems and Storage Technologies
	Per Stenström and Andrzej Brud	How to transform research results into a business
	Mahmut Kandemir	Embedded Systems: A Software Perspective

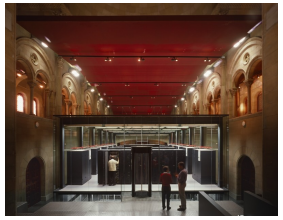
Education for Parallel Programming



I ♥ many-core programming

I ♥ multi-core programming

We all ♥ massive parallel prog.



Multicore-based pacifier

I ♥ games

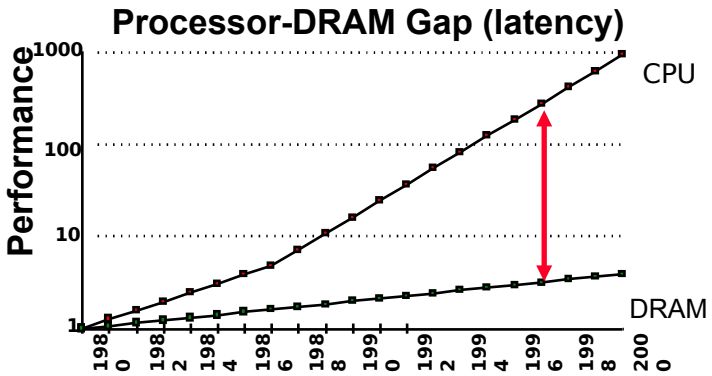
Navigating the Mare Nostrum

Are we planning to upgrade?

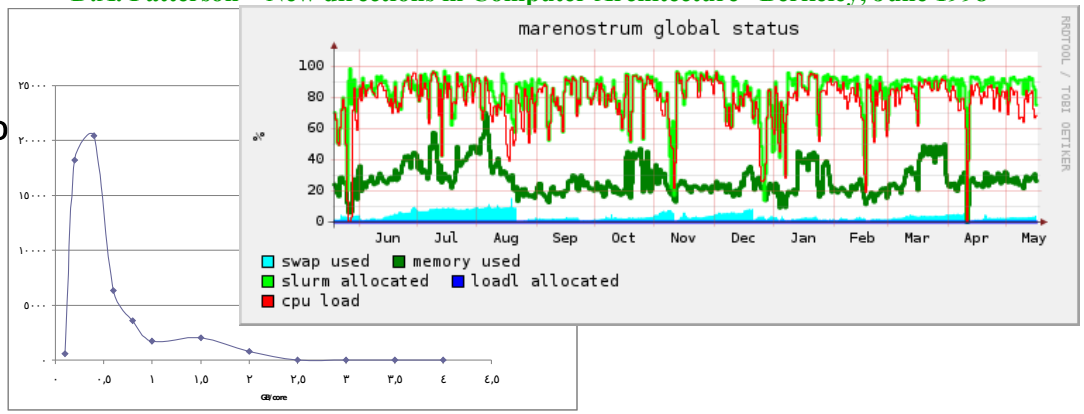
- Negotiating our next site ;)

Memory: more than a wall

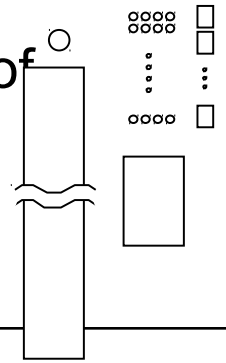
- Performance:
  - Latency
  - bandwidth
- Cost
- Power
- Capacity
  - Real usage (< 40%
  - Accelerator model (→ 2x cost)



D.A. Patterson " New directions in Computer Architecture" Berkeley, June 1998



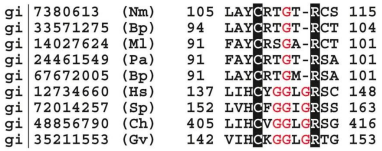
- Main component/nightmare of programming model

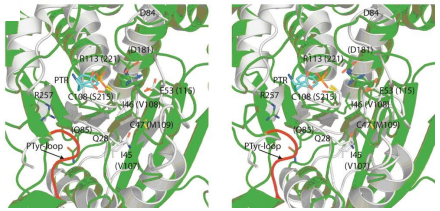


From sequences to structures : HPC Roadmap

200 0 → **201 1** → **2015 and beyond**

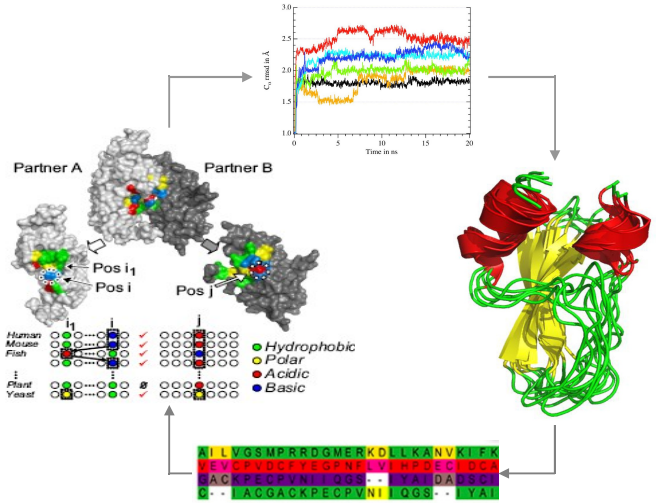
Grand Challenge
GENCI/CCRT



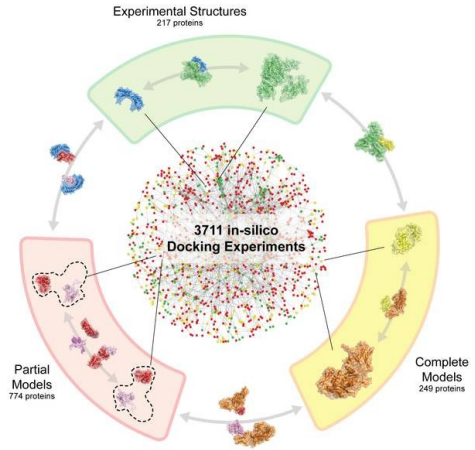


Proteins 69 (2007) 415

Identify all protein sequences using public resources and metagenomics data, and systematic modelling of proteins belonging to the family (Modeller software).



Improving the prediction of protein structure by coupling new bio-informatics algorithm and massive molecular dynamics simulation approaches.



Systematic identification of biological partners of proteins.

Computations using more and more sophisticated bio-informatical and physical modelling approaches ⇒ Identification of protein structure and function

1 family	1 family	1 family
5.10^3 cpu/~week	**5.10^4 cpu/~week**	**~ 10^4*KP cpu/~week**
		CSP : proteins structurally characterized ~ 10^4
# 25 Gb of storage	# 5 Tb of storage	# 5*CSP Tb of storage
500 Gb of memory	5 Tb of memory	5*CSP Tb of memory