Barcelona Supercomputing Center
Centro Nacional de Supercomputación

"Supercomputing for the Future, Supercomputing from the Past"

Onassis Foundation
Lectures on Computer Science
Keraklion, Crete, July 21-25, 2008

Heraklion, Crete, July 25th, 2008

Prof. Mateo Valero
Director

Talk outline

- Supercomputing from the past
  - Architecture evolution
  - Applications and algorithms
- Supercomputing for the future
  - Technology trends
  - Multidisciplinary top-down approach
- Conclusions

30th List: The TOP10

	Manufacturer	Computer	Rmax [TF/s]	Installation Site	Country	Year	#Cores
1	IBM	BlueGene/L eServer Blue Gene	478.2	DOE/NNSA/LLNL	USA	2007	212,992
2	IBM	JUGENE BlueGene/P Solution	167.3	Forschungszentrum Juelich	Germany	2007	65,536
3	SGI	SGI Altix ICE 8200	126.9	New Mexico Computing Applications Center	USA	2007	14,336
4	HP	Cluster Platform 3000 BL460c	117.9	Computational Research Laboratories, TATA SONS	India	2007	14,240
5	H[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	13,728
6	Sandia[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	26,569
7	Cray	Jaguar Cray XT3/XT4	101.7	DOE/ORNL	USA	2007	23,016
8	IBM	BGW eServer Blue Gene	91.29	IBM Thomas Watson	USA	2005	40,960
9	Cray	Franklin Cray XT4	85.37	NERSC/LBNL	USA	2007	19,320
10	IBM	New York Blue eServer Blue Gene	82.16	Stony Brook/BNL	USA	2007	36,864

Plenty of room for research!

31th List: The TOP10

	Manufacturer	Computer	Rmax [TF/s]	Installation Site	Country	Power [MW]	#Cores
1	IBM	Roadrunner BladeCenter QS22/LS21	1,026	DOE/NNSA/LANL	USA	2.35	122,400
2	IBM	BlueGene/L eServer Blue Gene Solution	478.2	DOE/NNSA/LLNL	USA	2.33	212,992
3	IBM	Intrepid Blue Gene/P Solution	450.3	DOE/ANL	USA	1.26	163,840
4	Sun	Ranger SunBlade x6420	326	TACC	USA	2.00	62,976
5	Cray	Jaguar Cray XT4 QuadCore	205	DOE/ORNL	USA	1.58	30,976
6	IBM	JUGENE Blue Gene/P Solution	180	Forschungszentrum Juelich (FZJ)	Germany	0.50	65,536
7	SGI	Encanto SGI Altix ICE 8200	133.2	New Mexico Computing Applications Center	USA	0.86	14,336
8	HP	EKA Cluster Platform 3000 BL460c	132.8	Computational Research Laboratories, TATA SONS	India	1.60	14,384
9	IBM	Blue Gene/P Solution	112.5	IDRIS	France	0.32	40,960
10	SGI	SGI Altix ICE 8200EX	106.1	Total Exploration Production	France	0.44	10,240

31st List / June 2008 page 1

IBM continues to lead the TOP20 with 10 system. There was a great deal of activity in the Top20 with 14new, upgraded or improved benchmark entries.

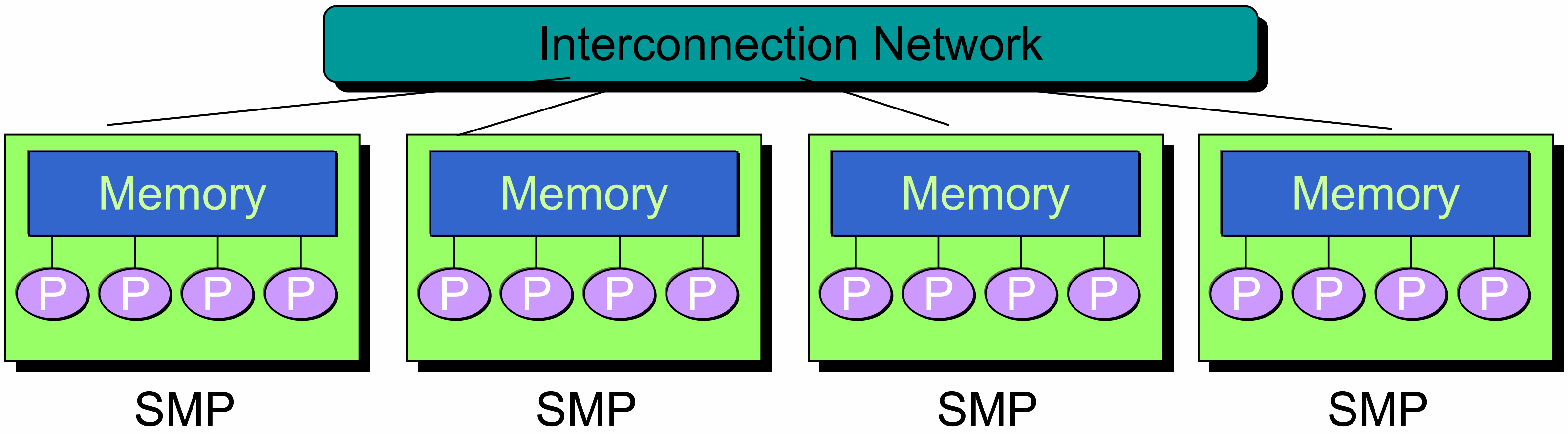
#	Ven-dor	Rmax TFlops	Installation	
1	**IBM**	1026	**DOE/NSSA/LANL** (QS22/LS21)	**New**
2	**IBM**	478.2	**DOE/NSSA/LLNL** (104 racks BlueGene/L)	
3	**IBM**	450.3	**Argonne Natl Lab** (40 racks Blue Gene/P)	**New**
4	Sun	326	**Texas Adv Comp Center** (QC Opteron)	**New**
5	Cray	205	**Oak Ridge NL** (XT4 QC Opteron)	**New**
6	IBM	180	**FZJ Juelich** (16 racks Blue Gene/P)	**Better Bmk**
7	SGI	133.2	**New Mexico CAC** (Altix Clovertown)	**Better Bmk**
8	HP	132.8	**TATA Research Lab** (Clovertow	**Better Bmk**
9	**IBM**	112.5	**IDRIS** (10 racks Blue Gene/P)	**New**
10	SGI	102.8	**Total Exploration** (Altix Quad Core Xeon)	**New**

#	Ven-dor	Rmax TFlops	Installation	
11	HP	102.8	**Swedish Govt** (Clovertown)	
12	Cray	102.2	**Sandia – Red Storm** (XT3 Opteron)	
13	**IBM**	92.96	**EDF R&D** (8 rack Blue Gene/P)	**New**
14	**IBM**	91.29	**BlueGene at Watson** (20 racks BlueGene/L)	
15	Cray	85.368	**NERSC/LBNL** (XT4 Opteron)	
16	Hitachi	82.984	**T2K Open SC-Japan** (QC Opteron)	**New**
17	**IBM**	82.16	**Stony Brook / BNL** (18 racks BlueGene/L)	
18	**IBM**	80.32	**ECMWF** (Power 575, p6)	**New**
19	**IBM**	80.32	**RZG/Max Planck/IPP** (Power 575, p6)	**New**
20	Appro	76.46	**Univ of Tsukuba** (QC Opteron)	**New**

Source: www.top500.org

Hybrid SMP-cluster parallel systems

- Most modern high-performance computing systems are clusters of SMP nodes (performance/cost trade-off)



- Programming models allow to specify:
  - How computation is distributed?
  - How data is distributed and how is it accessed?
  - How to avoid data races?

Per Stenström

IBM breaks 1 Petaflop barrier with hybrid configuration 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)**

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
Nmax: 2236927
Rpeak: 1,375,776 GF
Power: 2345 kW
Mflops/Watts: 437 Mflops/W

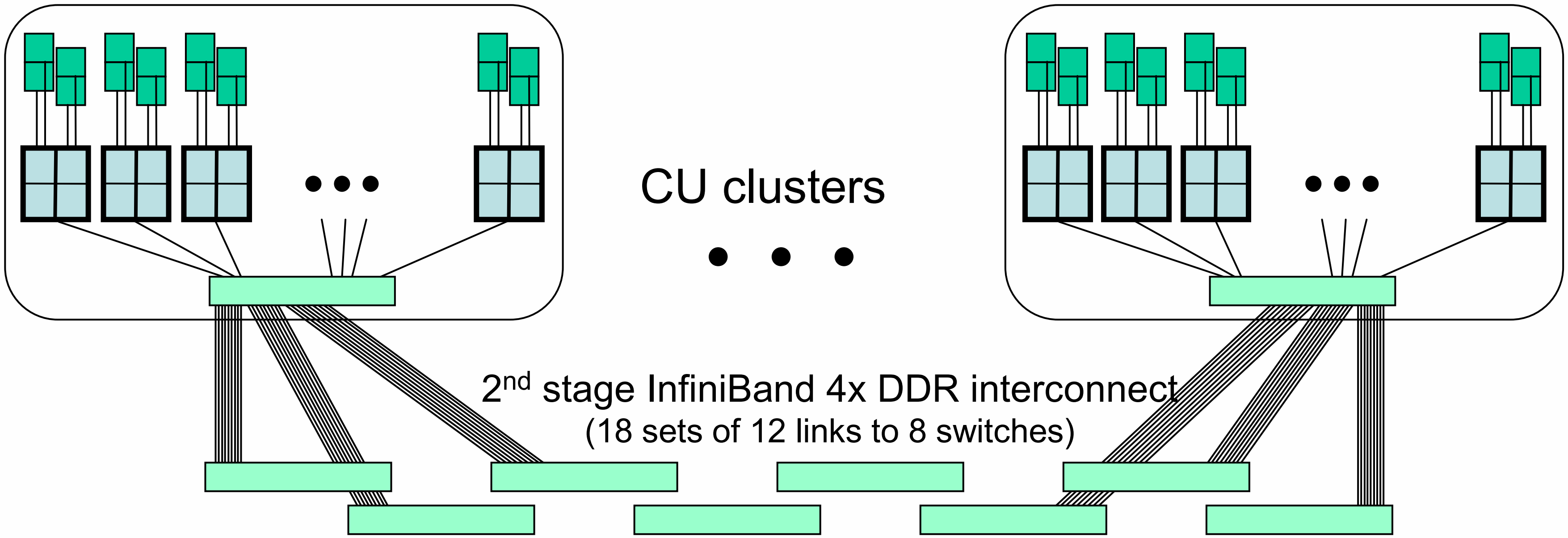
Source: www.top500.org

Roadrunner Final System

"Connected Unit" cluster
192 Opteron nodes
(180 w/ 2 dual-Cell blades
connected w/ 4 PCIe x8 links)

~7,000 dual-core Opterons
- ~50 TeraFlop/s (total)

~13,000 eDP Cell chips
- 1.4 PetaFlop/s (Cell)



2nd Generation IB 4X DDR

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BlueGene/P

Blue Gene/P continues Blue Gene's leadership performance in a space-saving, power-efficient package for the most demanding and scalable high-performance computing applications

System
72 Racks

Rack
32 Node Cards
1024 chips, 4096 procs

Cabled 8x8x16

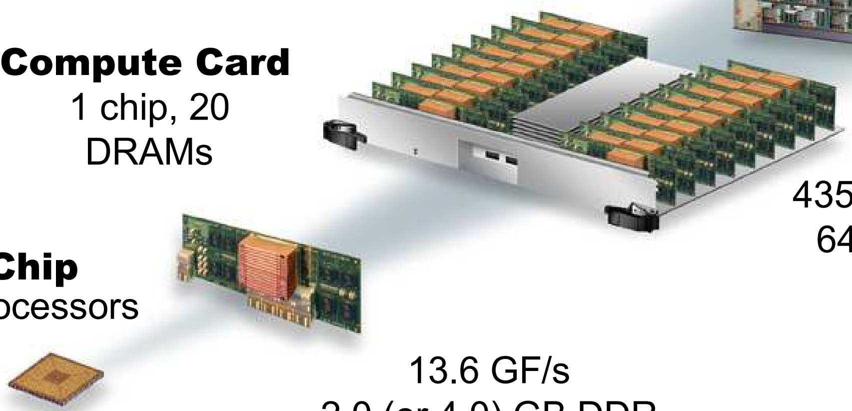
Final System:1 PF/s,144 TB
212992 processors

14 TF/s
2 TB

Node Card
(32 chips 4x4x2)
32 compute, 0-1 IO cards

Compute Card
1 chip, 20
DRAMs

Chip
4 processors



435 GF/s
64 GB

13.6 GF/s
2.0 (or 4.0) GB DDR
Supports 4-way SMP

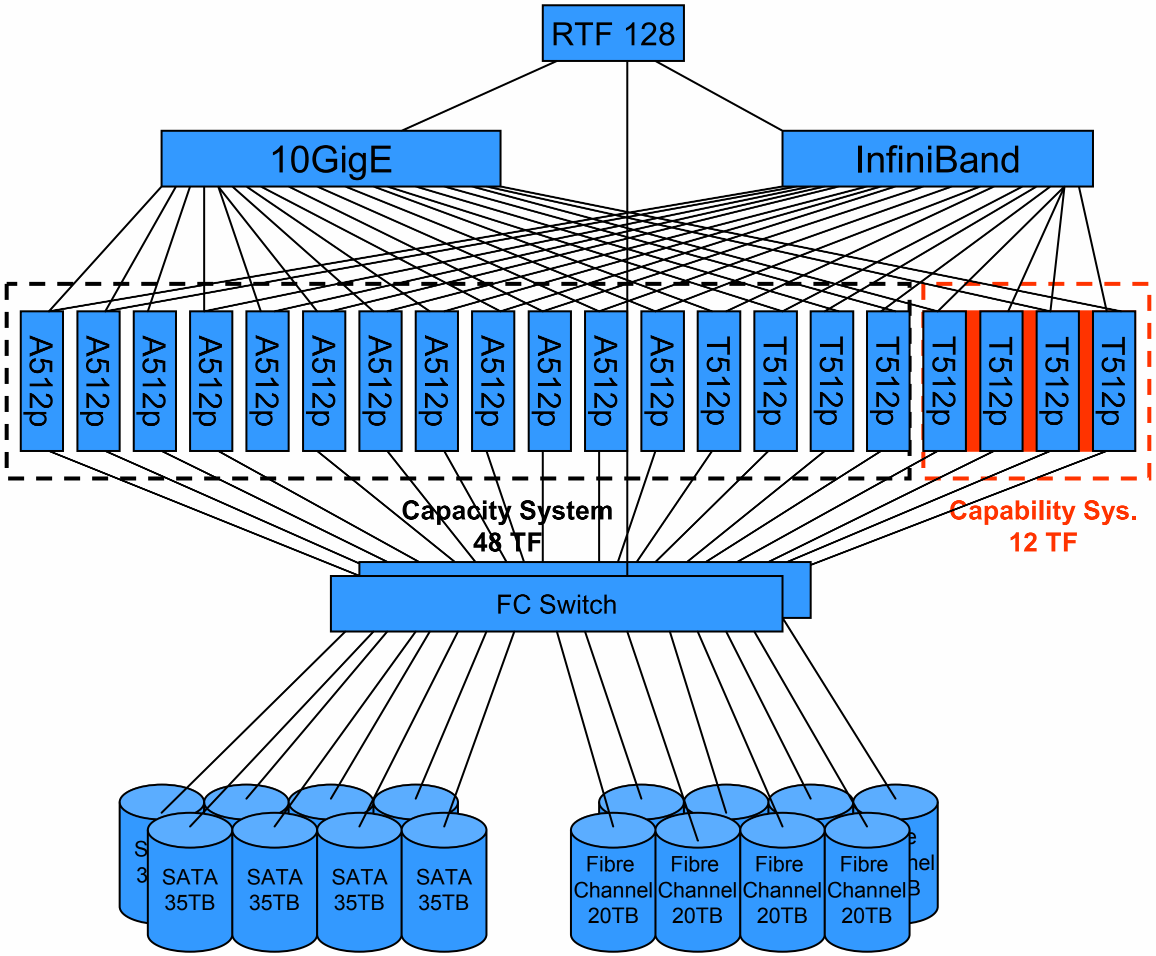
13.6 GF/s
8 MB EDRAM

Front End Node / Service Node
JS21 / Power5
Linux SLES10

HPC SW
Compilers
GPFS
ESSL
Loadleveler

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Columbia configuration



Front End
- 128p Altix 3700 (RTF)

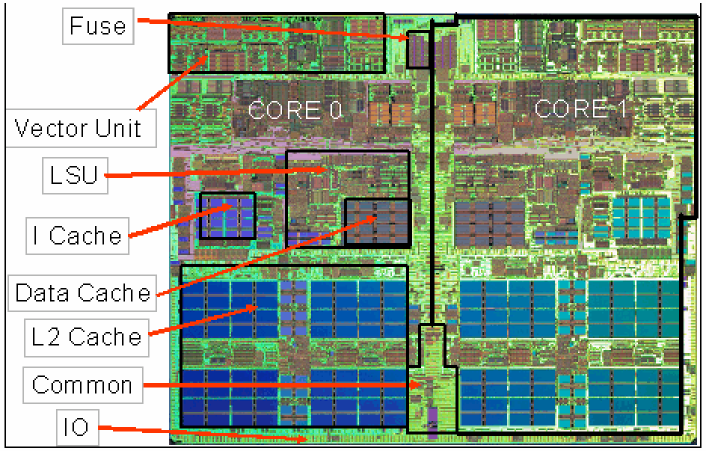
Networking
- 10GigE switch 32-port
- 10GigE cards (1 per 512p)
- InfiniBand switch (288 port)
- InfiniBand cards (6 per 512p)
- Altix 3700 2BX 2048 Numalink Kits

Compute Node (single sys image)
- Altix 3700 (A) 12x512p
- Altix 3700 BX2 (T) 8x512p

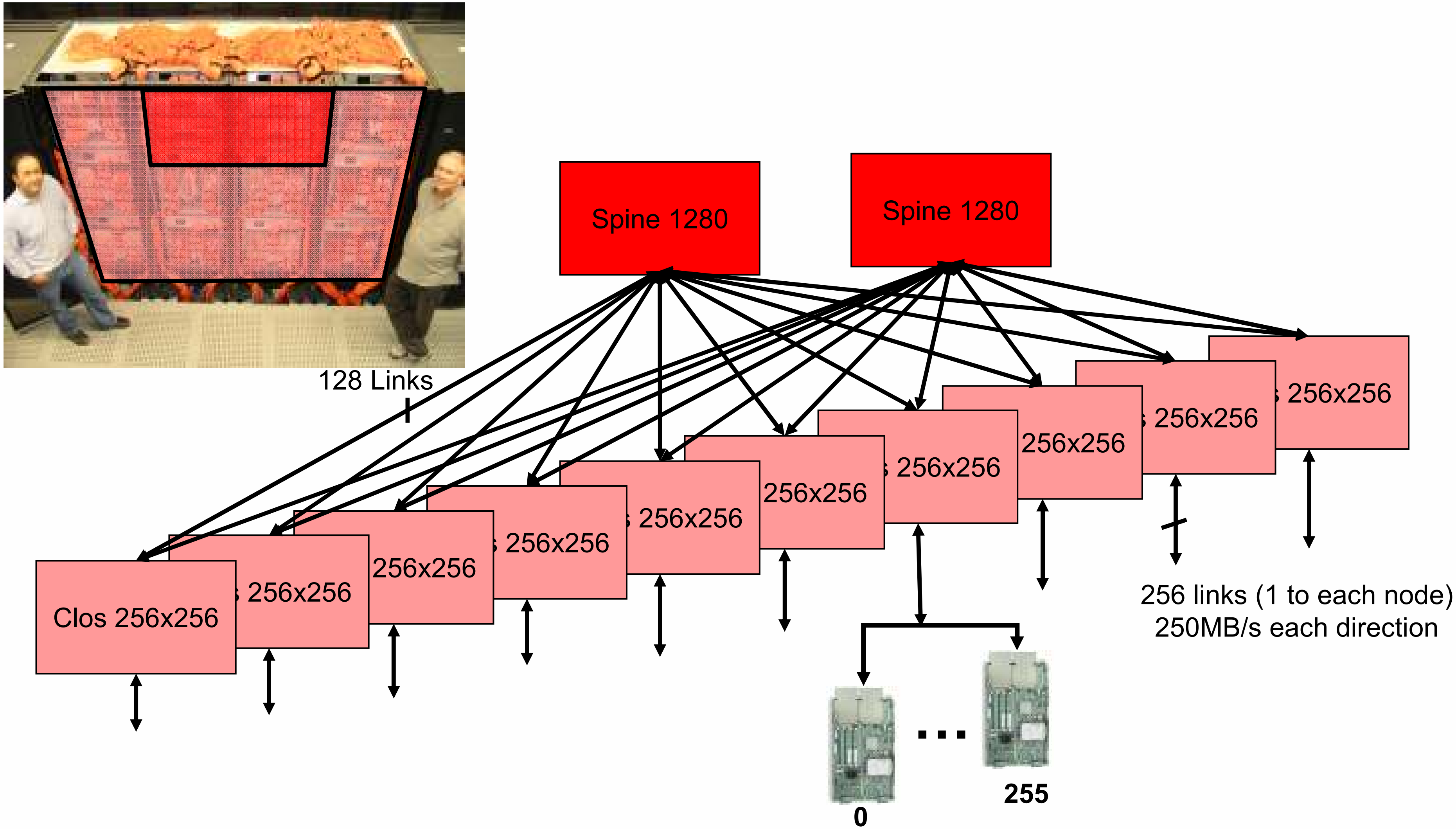
Storage Area Network
- Brocade switch 2x128 port

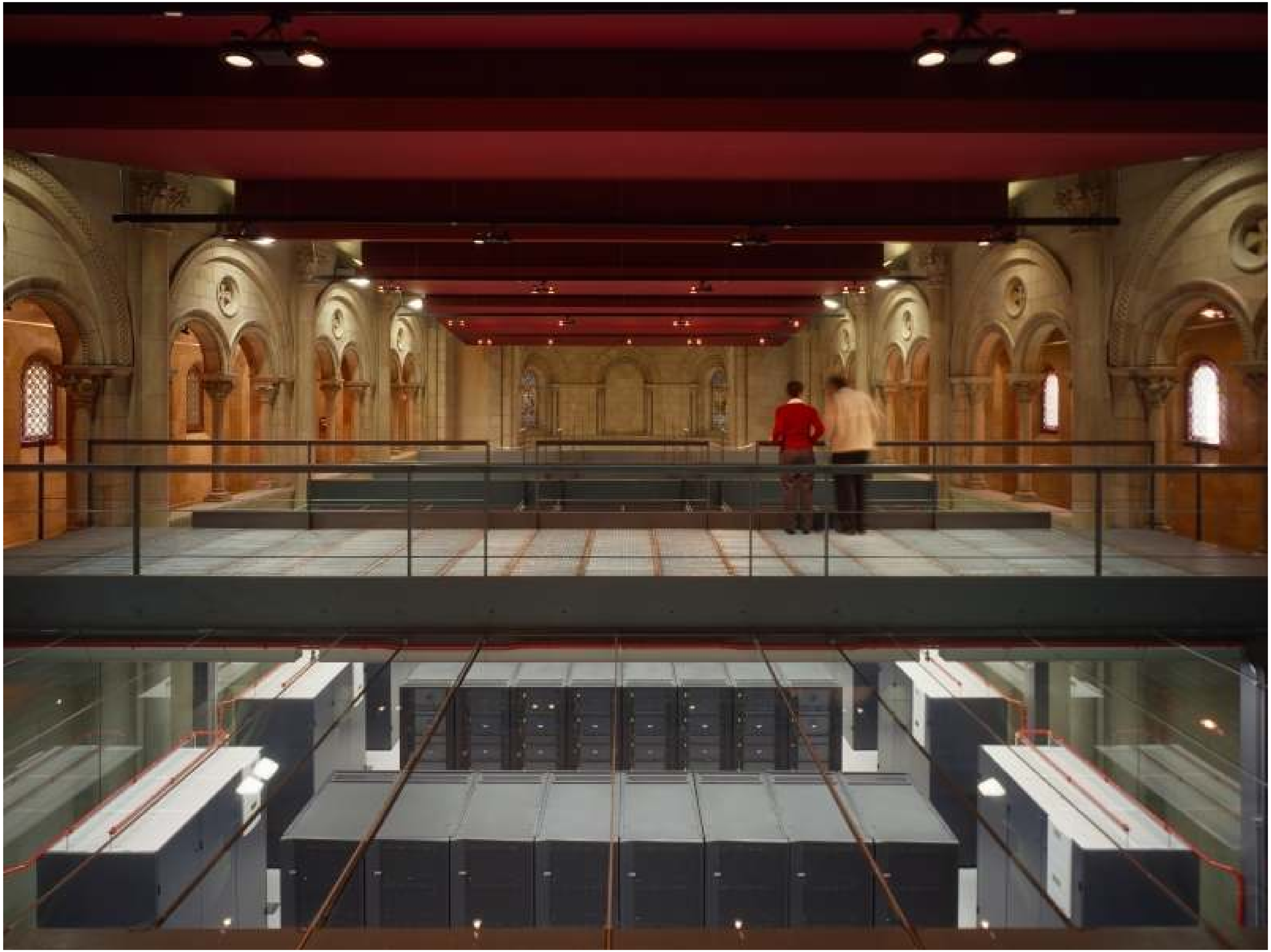
Storage (440 TB)
- FC RAID 8x20 TB (8 racks)
- SATARAID 8x35TB (8 racks)

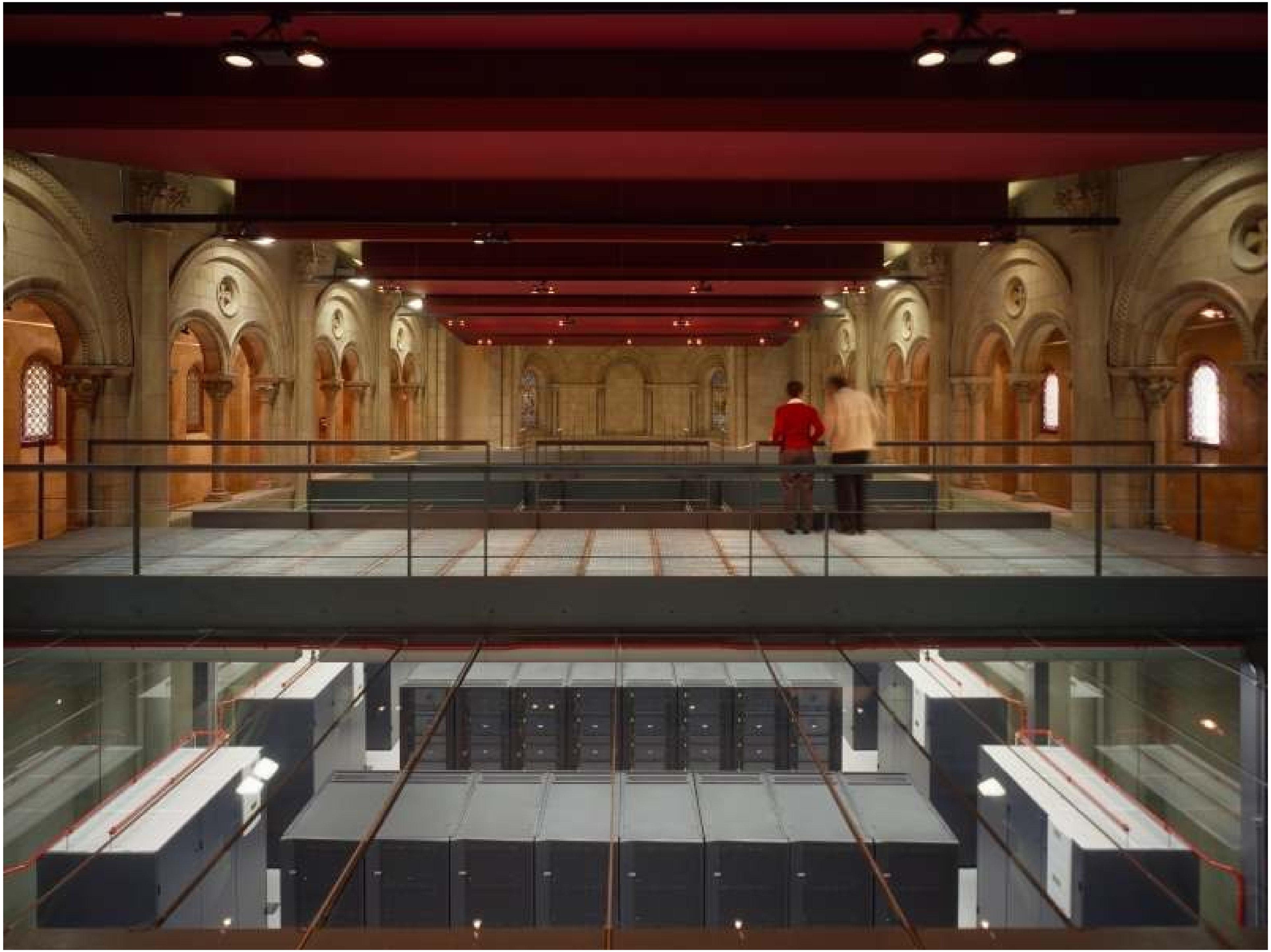
Processors, Blades, BladeCenters and Racks

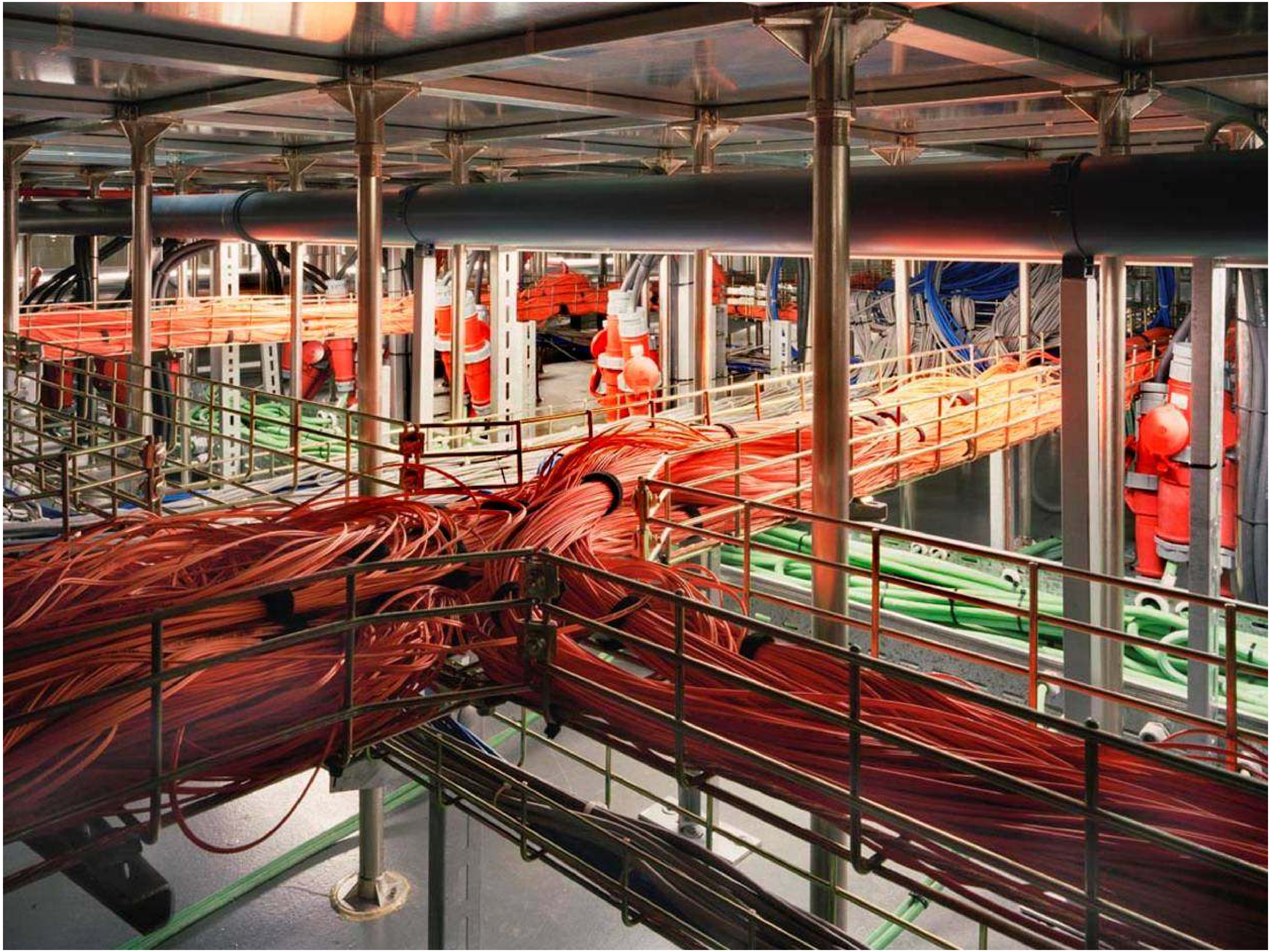


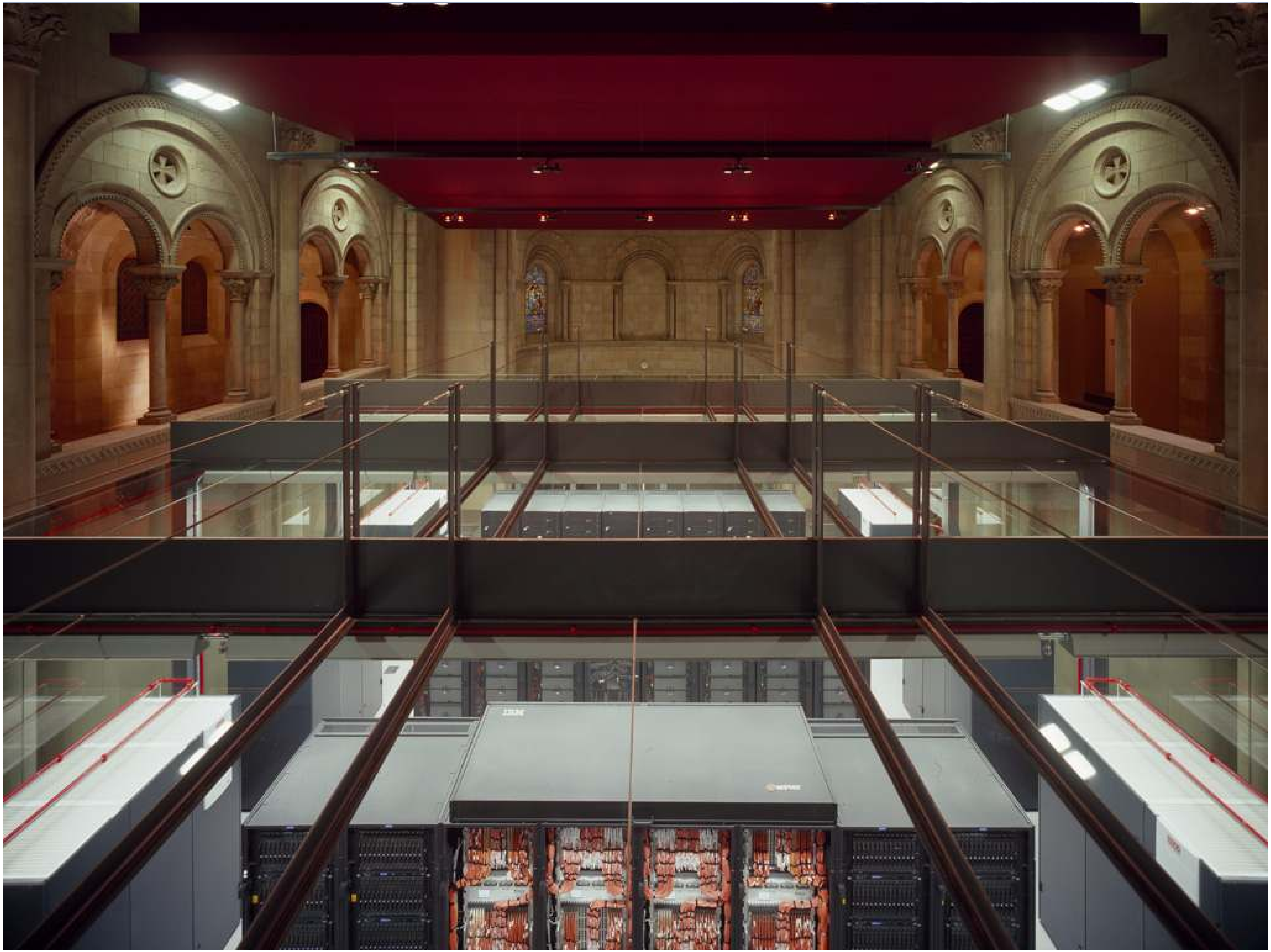
Interconnection network: Myrinet

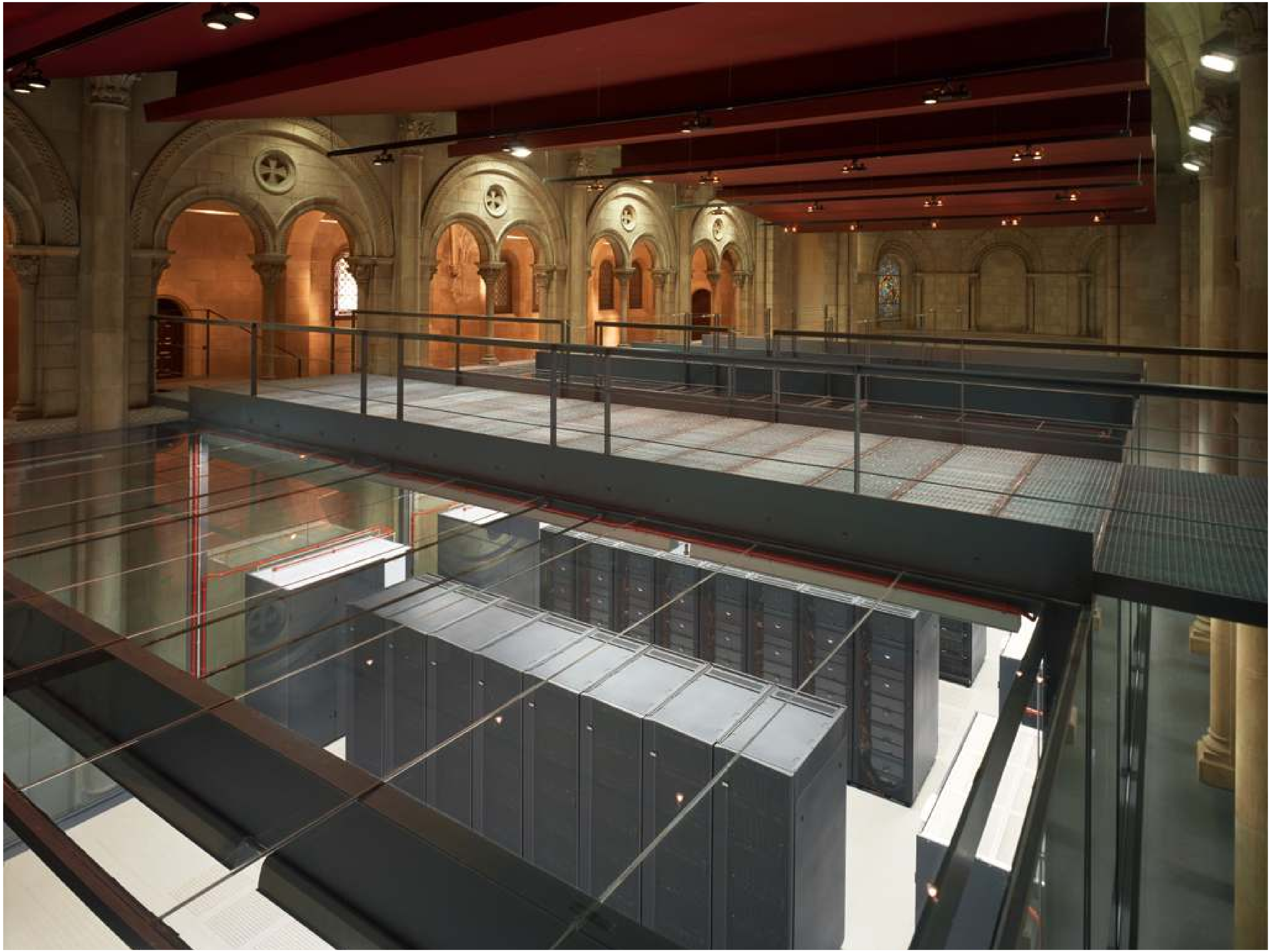


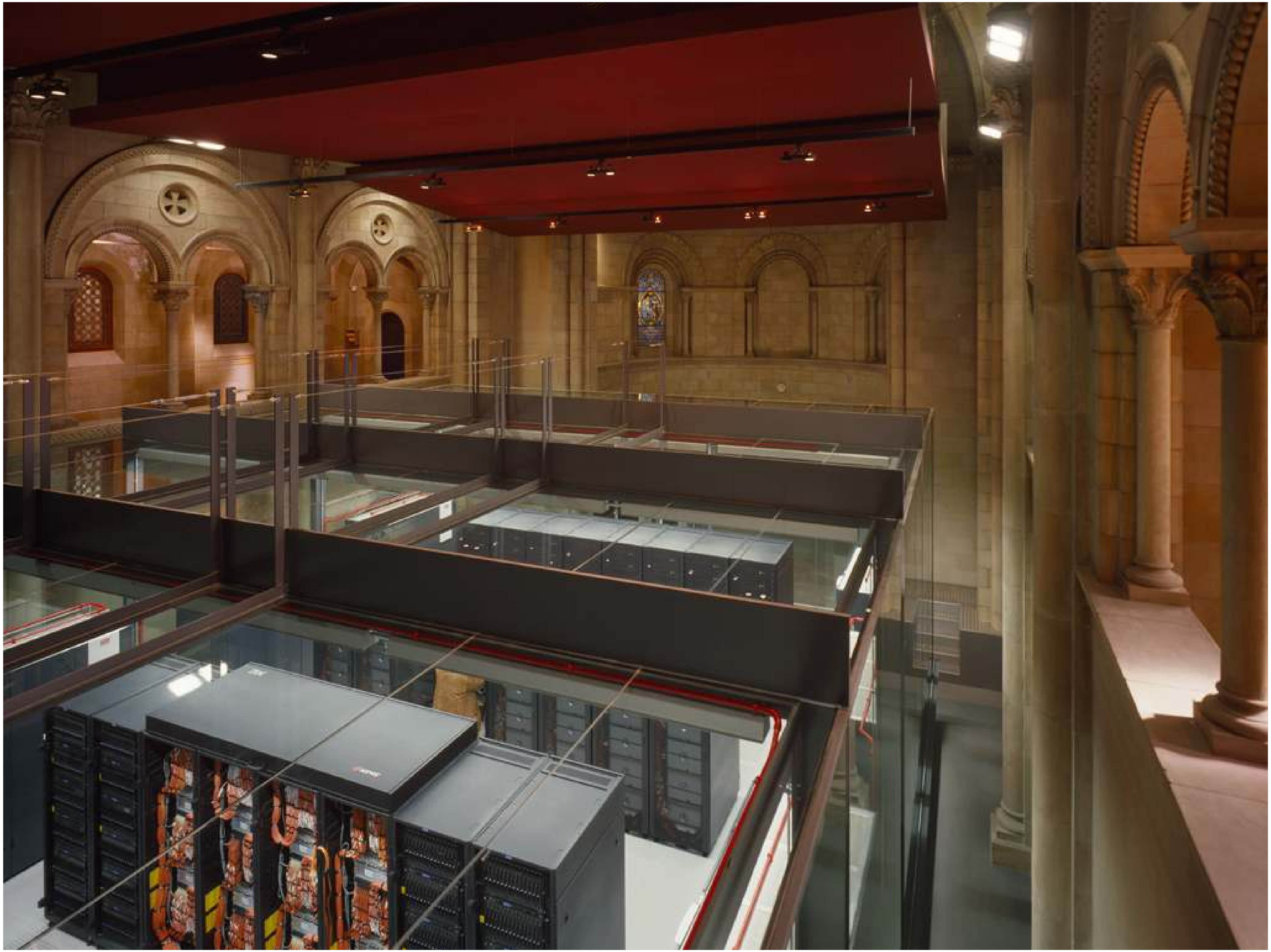


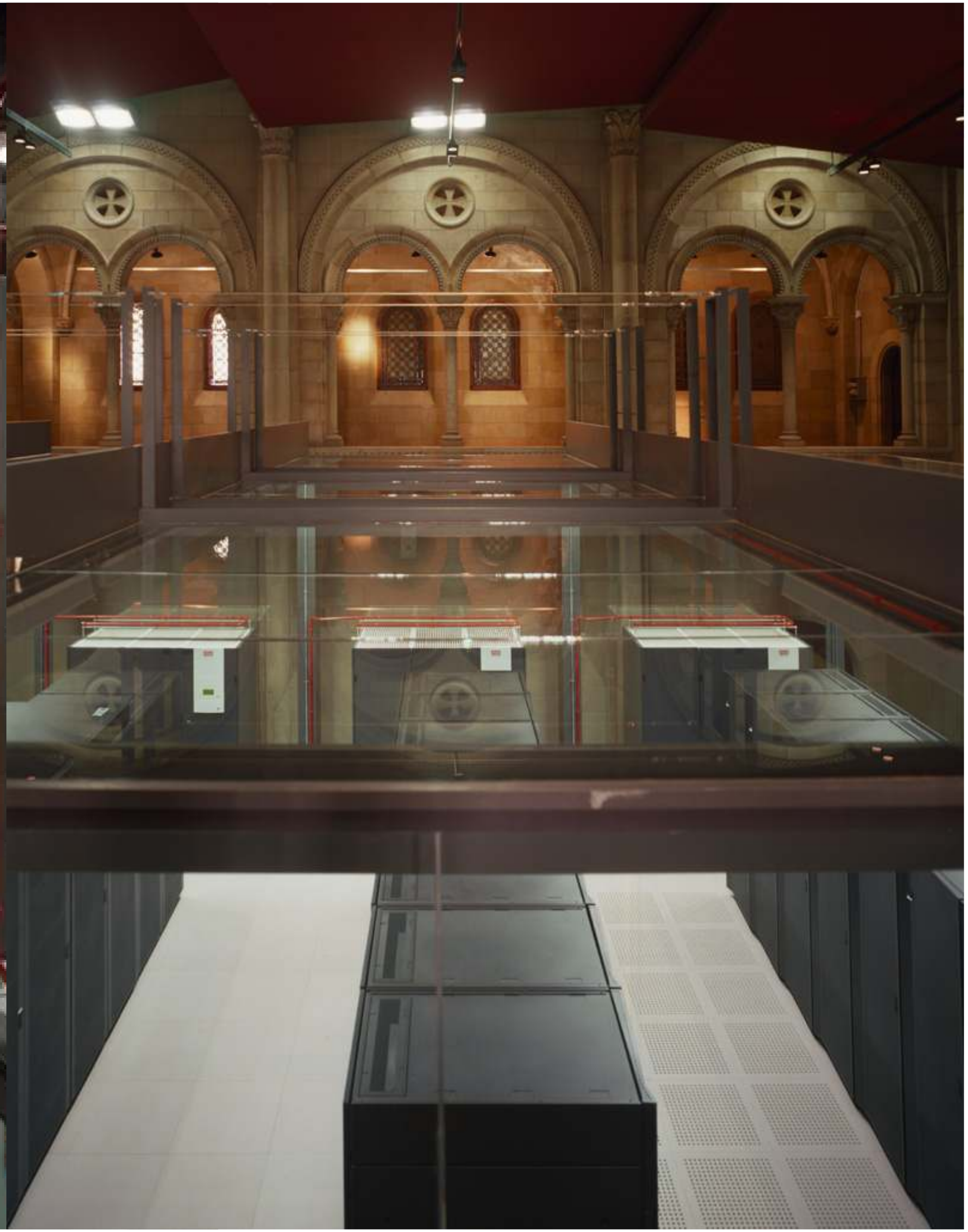




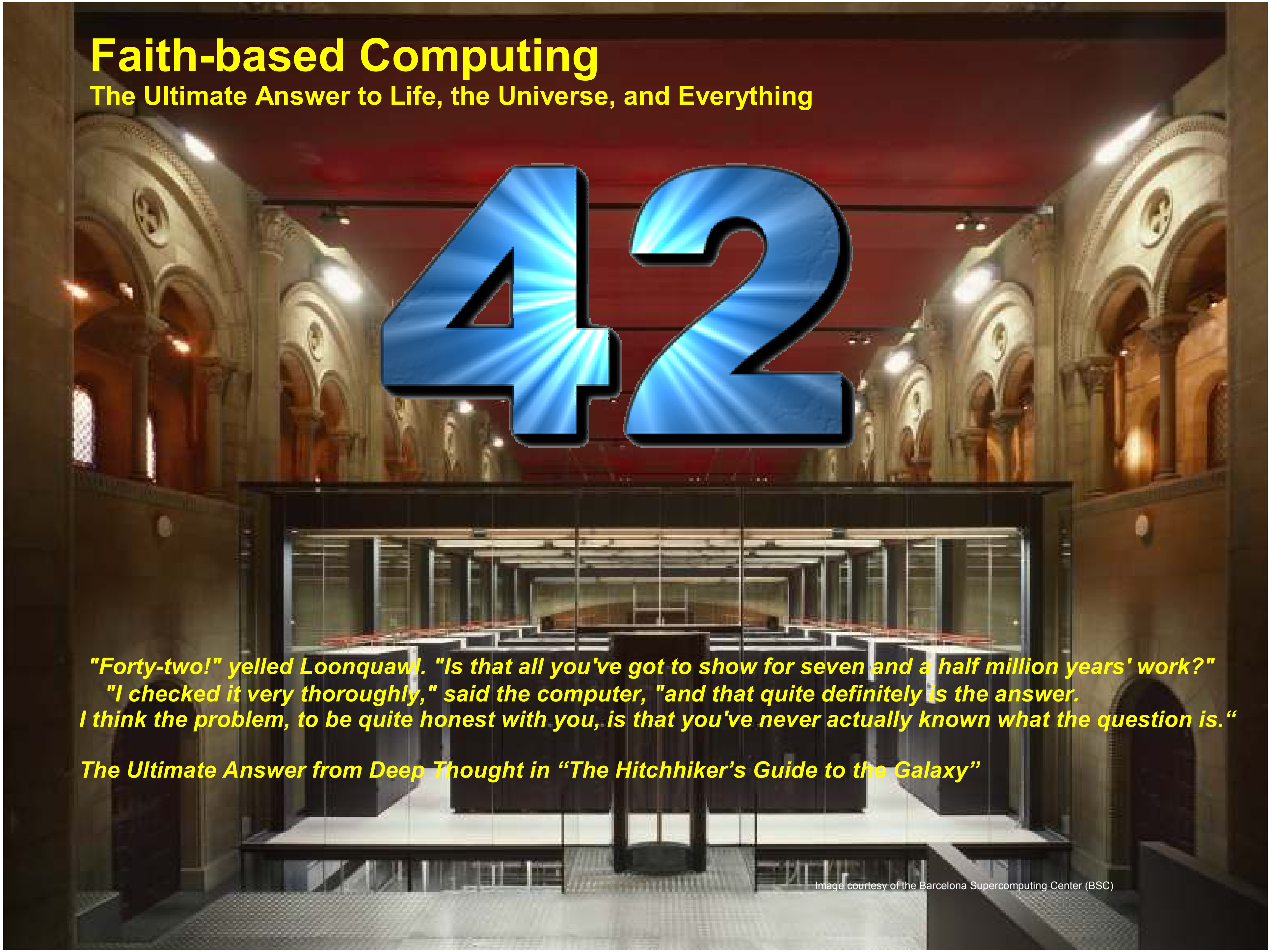






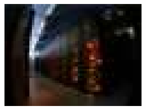



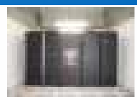

IBM

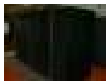

Faith-based Computing
The Ultimate Answer to Life, the Universe, and Everything
42
"Forty-two!" yelled Loonquawl. "Is that all you've got to show for seven and a half million years' work?"
"I checked it very thoroughly," said the computer, "and that quite definitely is the answer.
I think the problem, to be quite honest with you, is that you've never actually known what the question is."
The Ultimate Answer from Deep Thought in "The Hitchhiker's Guide to the Galaxy"
Image courtesy of the Barcelona Supercomputing Center (BSC)

Red Española de Supercomputación


MareNostrum

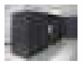

Magerit


CesarAugusta


LaPalma


Altamira


Tirant


Picasso

MareNostrum

Processor:	10240 PowerPC 970 2.3 GHz
Memory:	20 TBytes
Disk:	280 + 90 TBytes
Network:	Myrinet, Gigabit, 10/100
System:	Linux

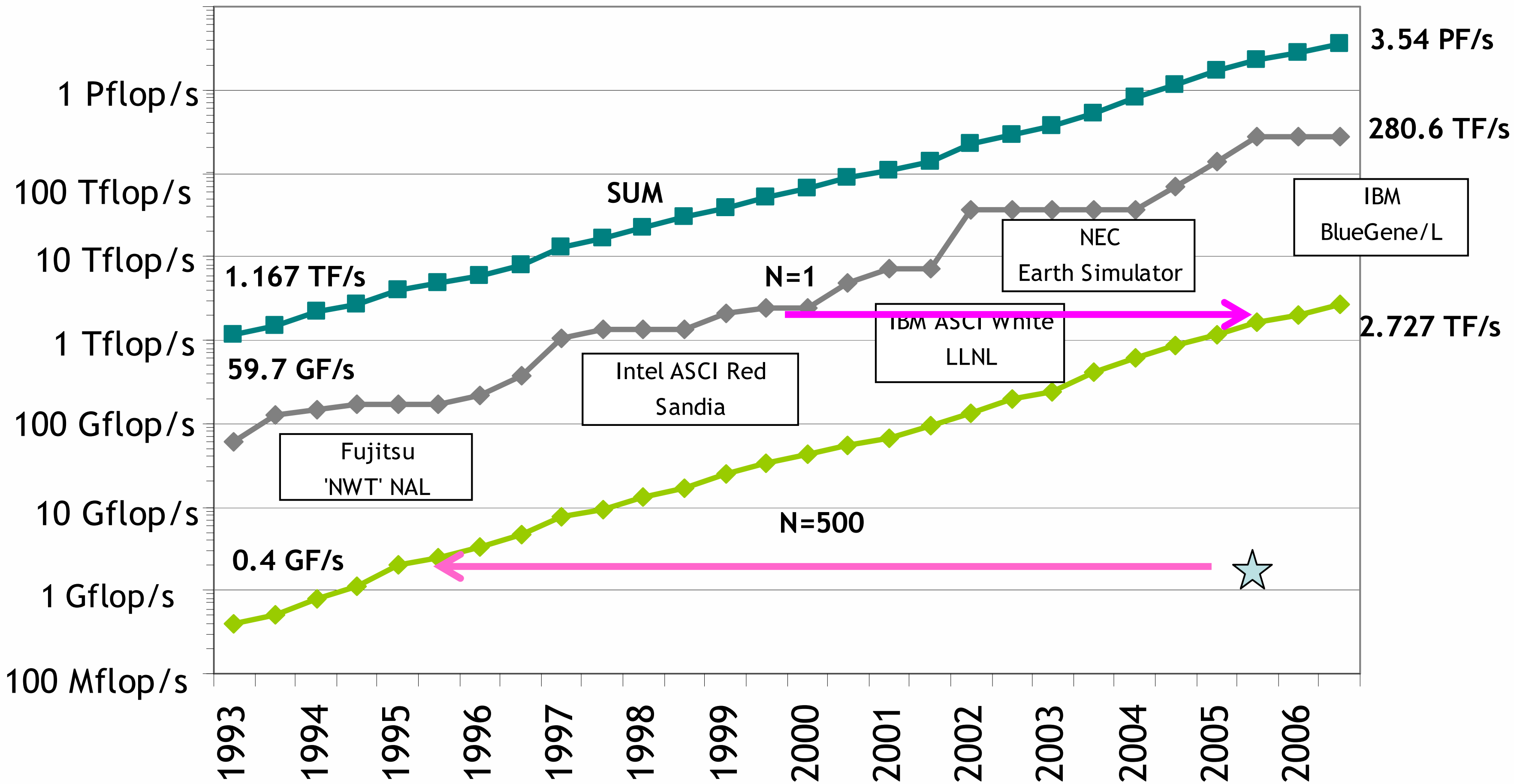
UPM

Processor:	2408 PowerPC 970 2.2 GHz
Memory:	4.7 TBytes
Disk:	63 + 47 TBytes
Network:	Myrinet, Gigabit, 10/100
System:	Linux

IAC, UMA, UC, UZ, UV

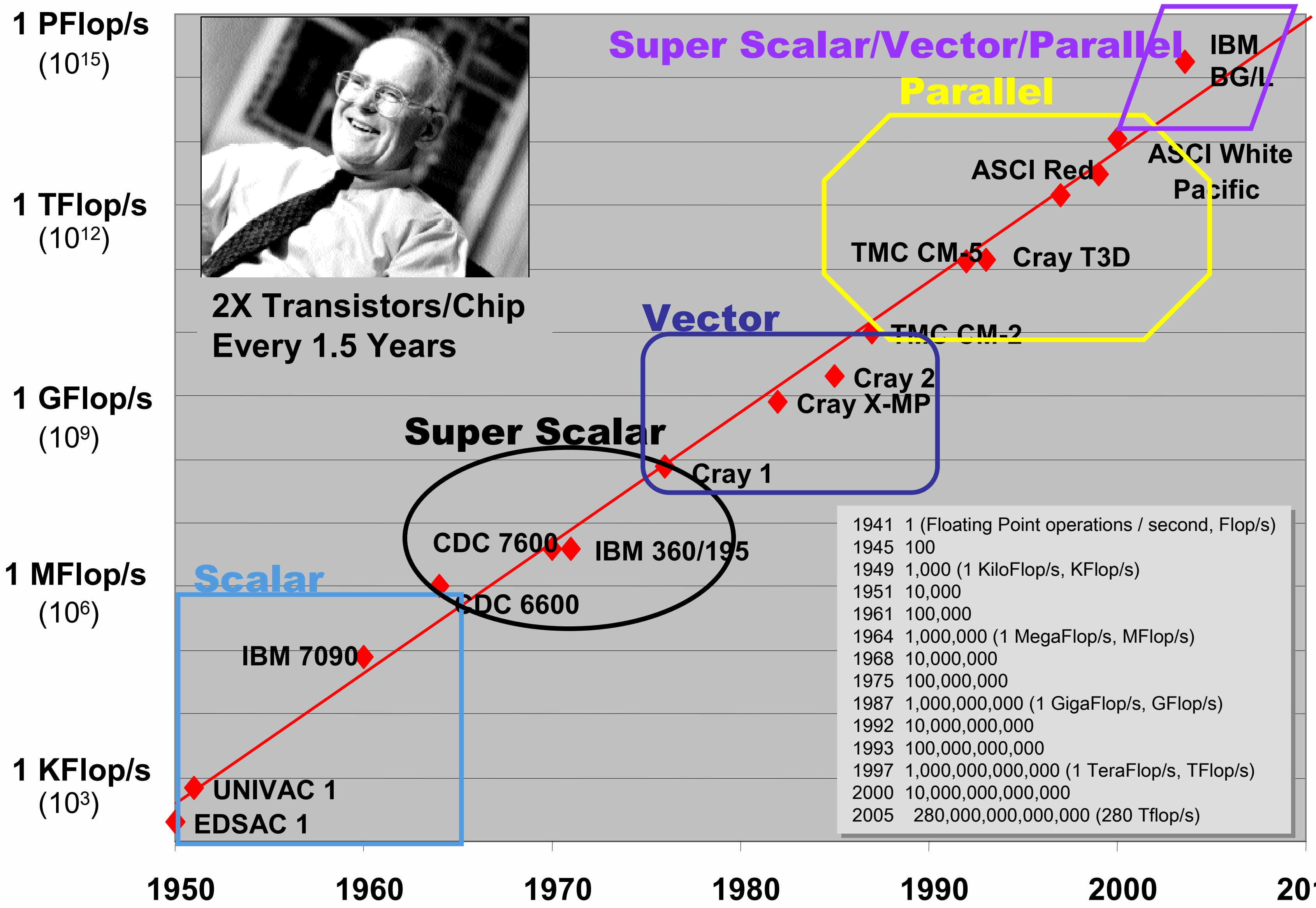
Process:	512 PowerPC 970 2.2 GHz
Memory:	1 TByte
Disk:	14 + 10 TBytes
Network:	Myrinet, Gigabit, 10/100
System:	Linux

Performance development



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

A growth-factor of a billion in performance in a career



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Those who gave their lives in the search for parallelism

Alliant, American Supercomputer, Ametek, AMT, Astronautics, BBN Supercomputer, Biin, CDC, Chen Systems, CHOPP, Cogent, Convex (now HP), Culler, Cray Computers, Cydrome, Dennelcor, Elexsi, ETA, E & S Supercomputers, Flexible, Floating Point Systems, Gould/SEL, IPM, Key, KSR, MasPar, Multiflow, Myrias, Ncube, Pixar, Prisma, SAXPY, SCS, SDSA, Supertek (now Cray), Suprenum, Stardent (Ardent+Stellar), Supercomputer Systems Inc., Synapse, Thinking Machines, Vitec, Vitesse, Wavetracer.

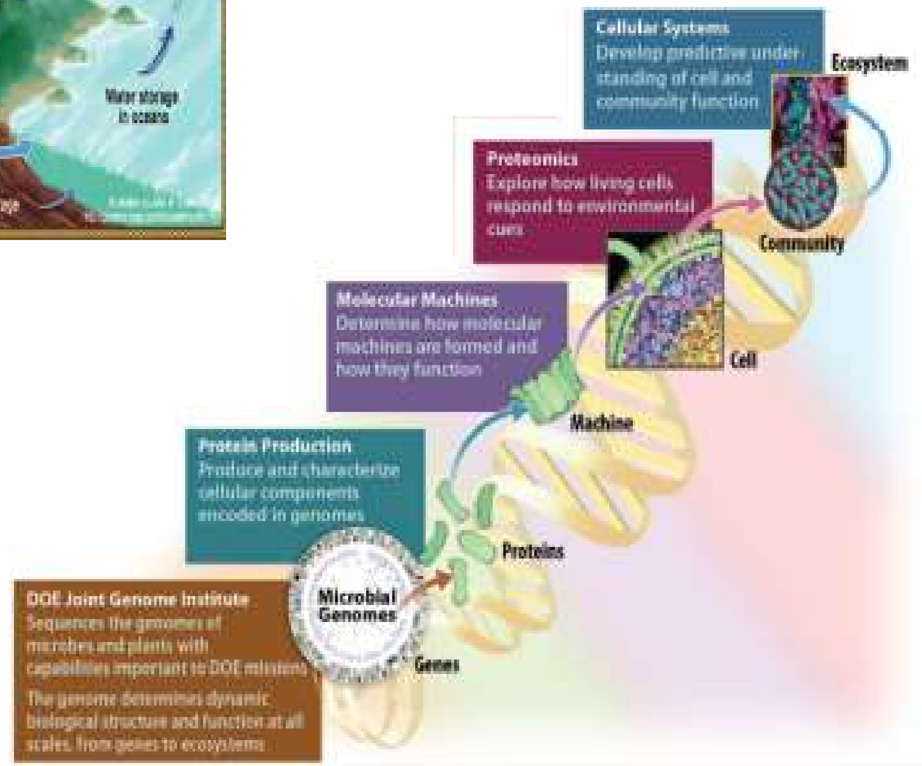
PACT'98 Gordon Bell

Talk outline

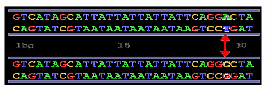
- **Supercomputing from the past**
  - Architecture evolution
  - **Applications and algorithms**
- Supercomputing for the future
  - Technology trends
  - Multidisciplinary top-down approach
- Conclusions

Grand challenge problems

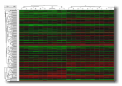
- Systems biology –
  - Model & simulation leading to predictive models with clinical or environmental impact
- Sustainable Systems –
  - Taking into account multi-scale nature - Models are linked to experimental data – providing corroboration of experiments
- Turbulence & Chaos –
  - Characterize boundary 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



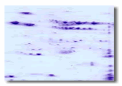
Multi-Scale Patient-Specific Data



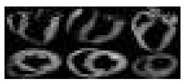
Genetic Variability



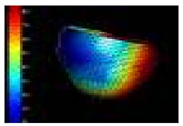
Gene Expression Profiling



Protein Expression Profiling



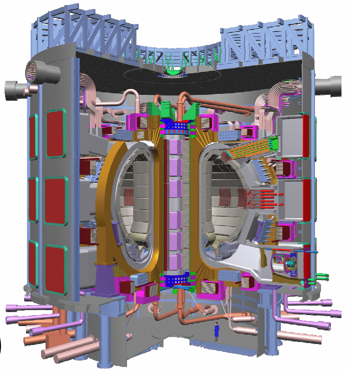
Multi-Modal Imaging



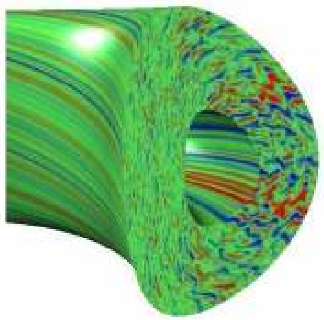
Data Analysis And Modeling

ITER design

- Supercomputing is mandatory for ITER design
- The most computing demanding problems for ITER design
  - **Plasma turbulent transport** (Gyro-kinetic codes)
  - **Plasma Wall Interaction** (DFT+MD+MC+DDD+FE codes)



- Problems generally amenable to parallelisation
  - Gryo-kinetic codes tested till 10^4 processors

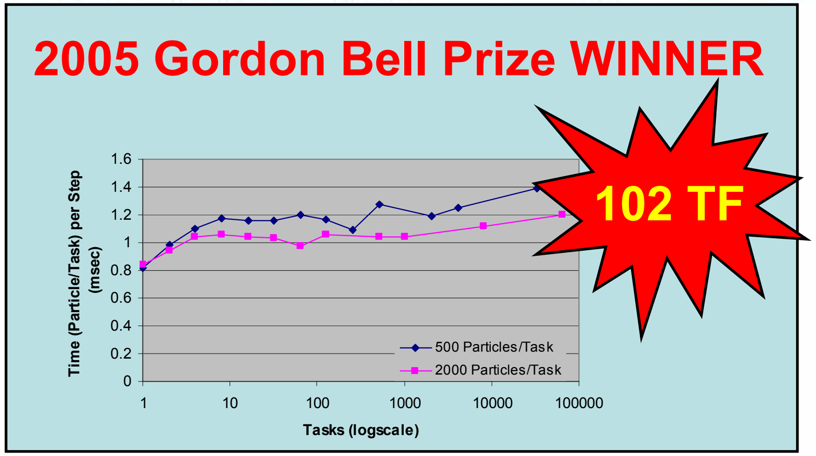


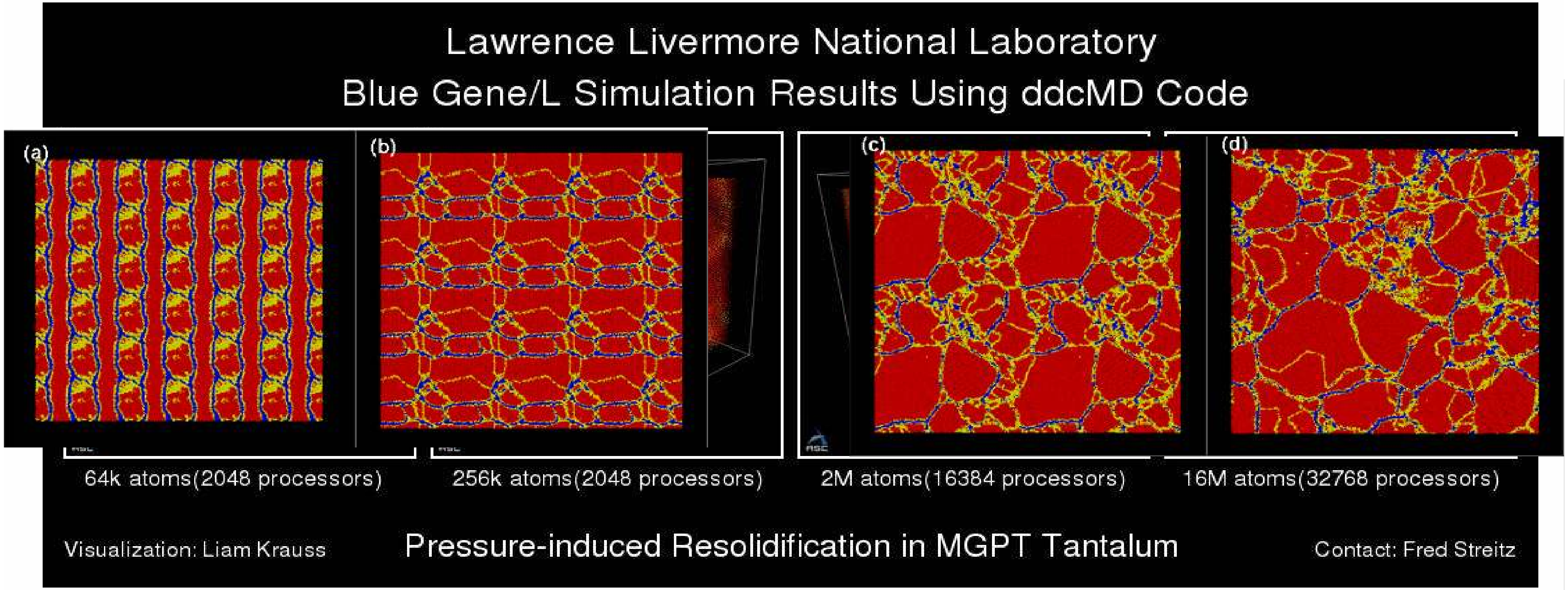
- With a 100 TFlops state-of-the-art machine
  - Gyro-kinetic modelling of JET reactor (tokamak) in days
  - Stellarators are more challenging, but could be simulate
  - ITER needs at least a 10+PFlops machines

Airbus 380 Design

BlueGene/L supports solidification understanding

- Nucleation is initiated at multiple independent sites in each sample cell
- Growth of solid grains initiates independently, but soon leads to grain boundaries which span the simulation cell
- The ddcMD team is currently using 131,072 CPUs of BG/L for unprecedented five hundred million atom MGPT simulations





Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Talk outline

- Supercomputing from the past
  - Architecture evolution
  - Applications and algorithms
- **Supercomputing for the future**
  - **Technology trends**
  - Multidisciplinary top-down approach
- Conclusions

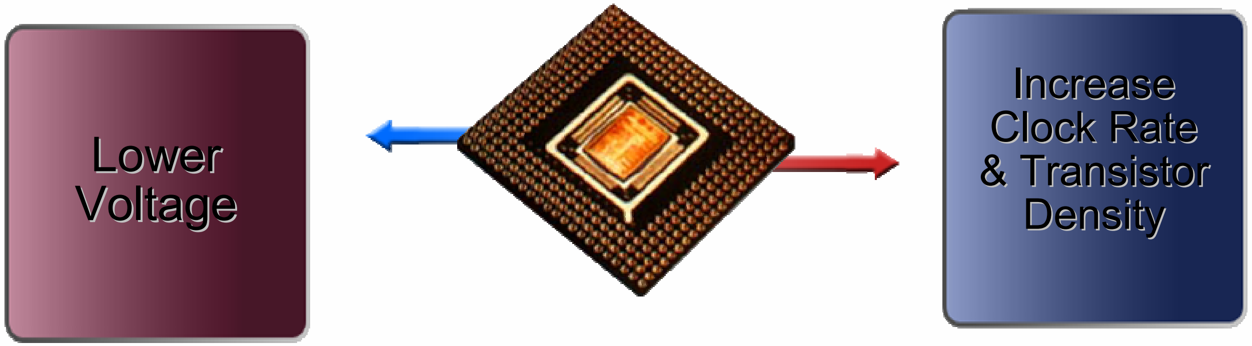
Technology Outlook

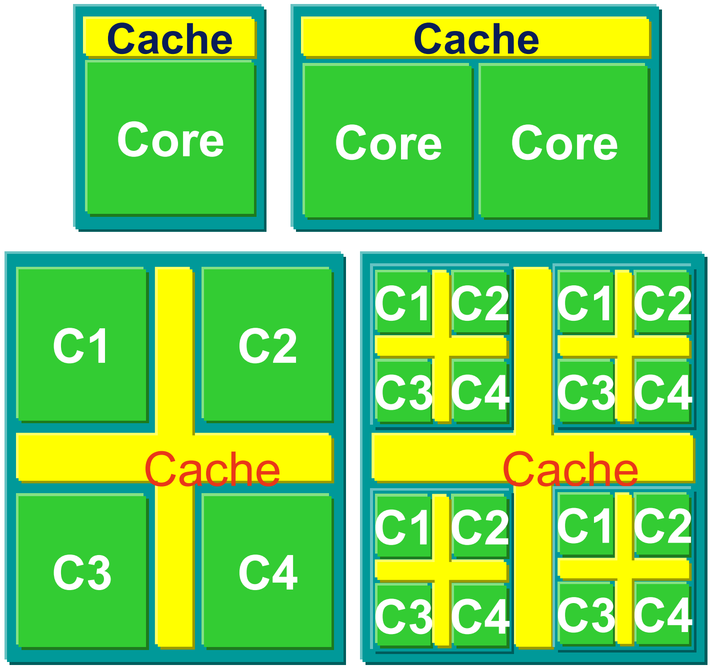
High Volume Manufacturing	2004	2006	2008	2010	2012	2014	2016	2018
Technology Node (nm)	90	65	45	32	22	16	11	8
Integration Capacity (BT)	2	4	8	16	32	64	128	256
Delay = CV/I scaling	0.7	~0.7	>0.7	Delay scaling will slow down				
Energy/Logic Op scaling	>0.35	>0.5	>0.5	Energy scaling will slow down				
Bulk Planar CMOS	High Probability				Low Probability			
Alternate, 3G etc	Low Probability				High Probability			
Variability	Medium			High		Very High		
ILD (K)	~3	<3	Reduce slowly towards 2-2.5					
RC Delay	1	1	1	1	1	1	1	1
Metal Layers	6-7	7-8	8-9	0.5 to 1 layer per generation				

Shekhar Borkar, Micro37, P

Increasing CPU performance: a delicate balancing act

Increasing the number of gates into a tight knot and decreasing the cycle time of the processor





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

Heat becoming an unmanageable problem, Intel Processors > 100 Watts

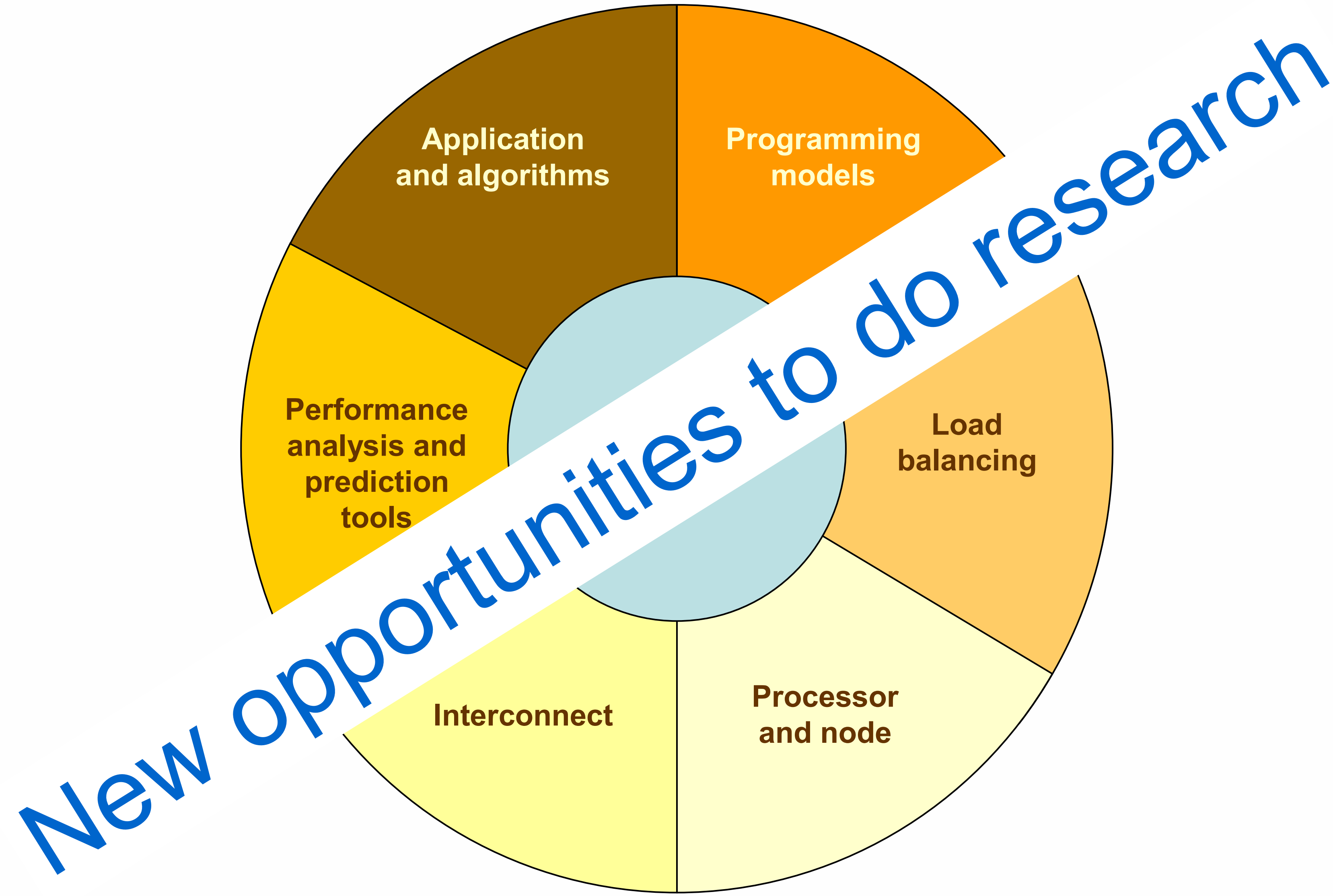
We will not see the dramatic increases in clock speeds in the future.

However, the number of gates on a chip will continue to increase.

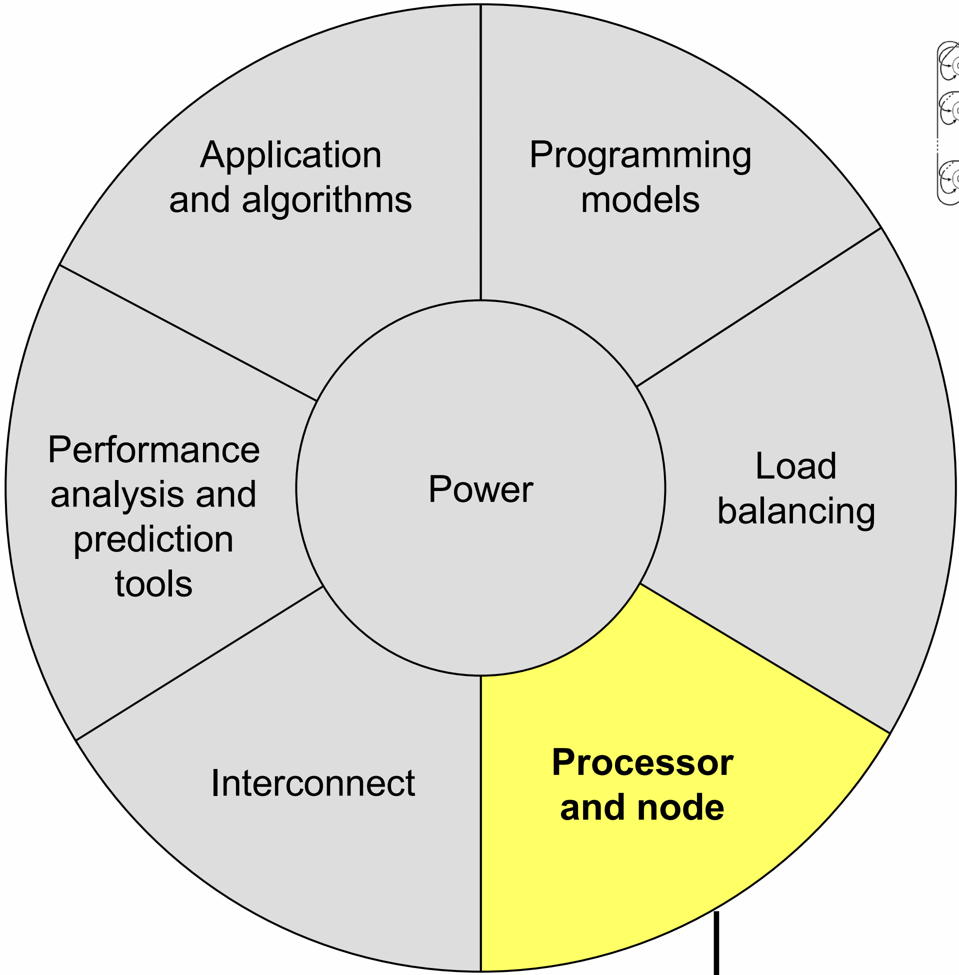
Talk outline

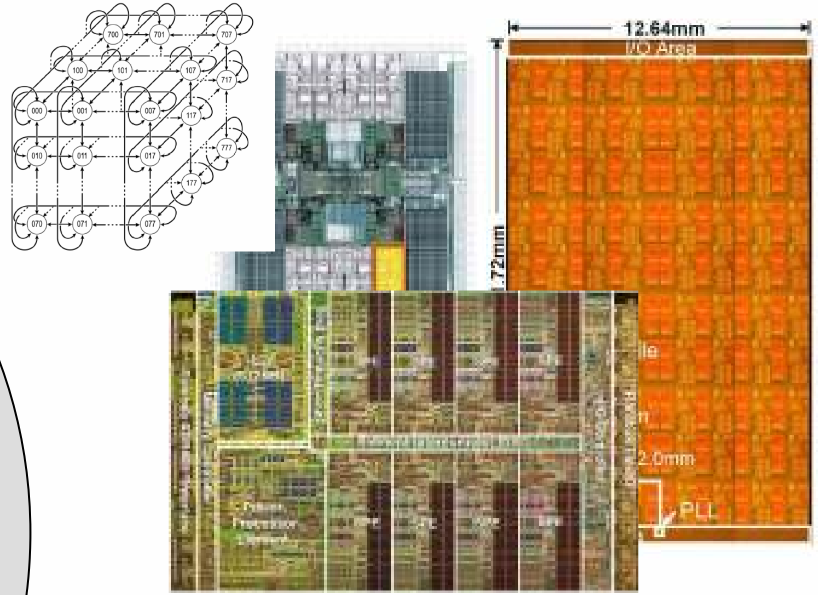
- Supercomputing from the past
  - Architecture evolution
  - Applications and algorithms
- **Supercomputing for the future**
  - Technology trends
  - **Multidisciplinary top-down approach**
- Conclusions

Multidisciplinary top-down approach



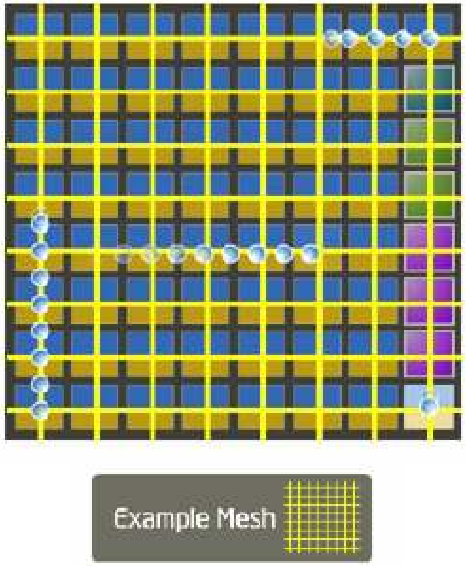
Multidisciplinary top-down approach

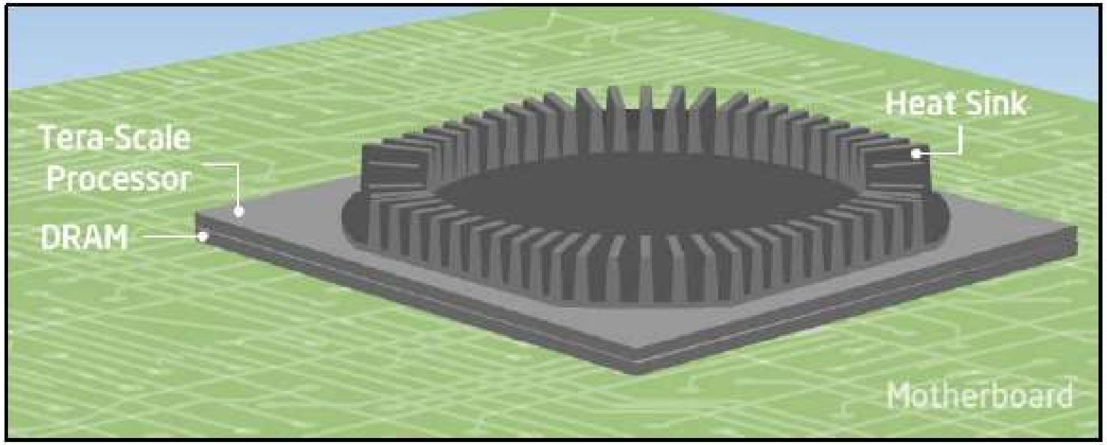




- ASIC
- Homogeneous multicore
- Accelerators
  - GPGPUs
  - FPGAs
  - Cell/B.E.

Intel´s Petaflop chip

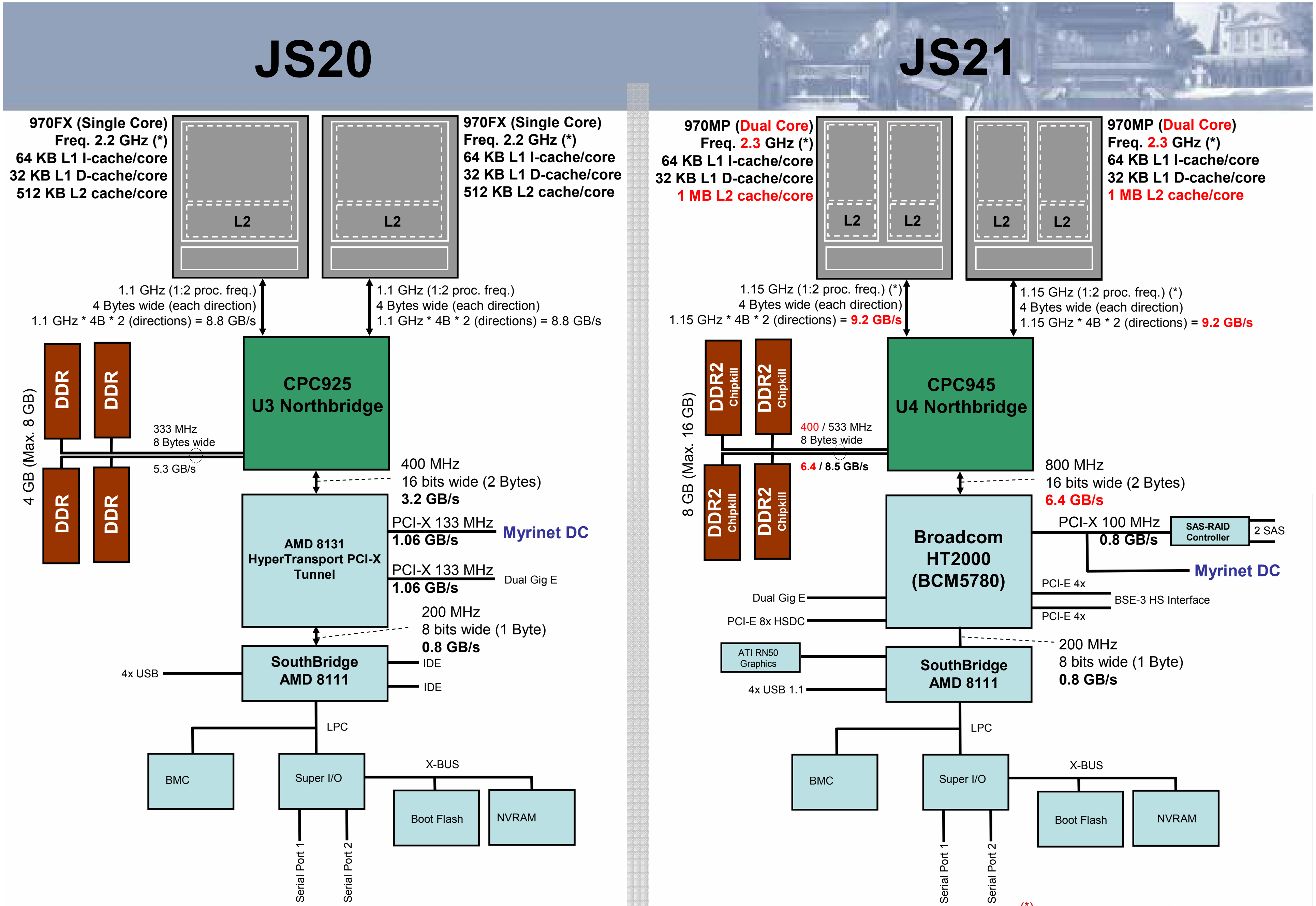




The key technologies of this first Tera-scale Research Prototype are a mesh interconnect (left) and support for 3D stacked memory (above).

- 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

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



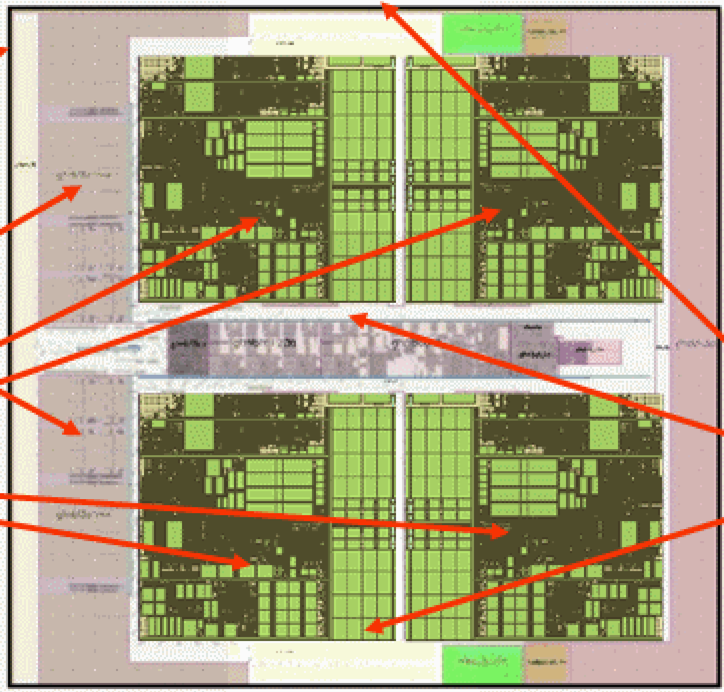
(*) Processor freq 2.3 GHz on BladeCenterE

AMD's Next Generation Processor Technology

Native quad core die

Expandable shared L3 cache

IPC enhanced CPU cores



Ideal for 65nm SOI and beyond

Enhanced Direct Connect Architecture and Northbridge

- 32B instruction fetch
- Improved branch prediction
- Out-of-order load execution
- Up to 4 DP FLOPS/cycle
- Dual 128-bit SSE dataflow
- Dual 128-bit loads per cycle
- Bit Manipulation extensions (LZCNT/POPCNT)
- SSE extensions (EXTRQ/INSERTQ, MOVNTSD/MOVNTSS)

- Four ungangable x16 HyperTransport™ links (up to 5.2GT/sec)
- Enhanced crossbar
- Next-generation memory support
- FBDIMM *when appropriate*
- Enhanced power management and RAS

Ranger System Summary

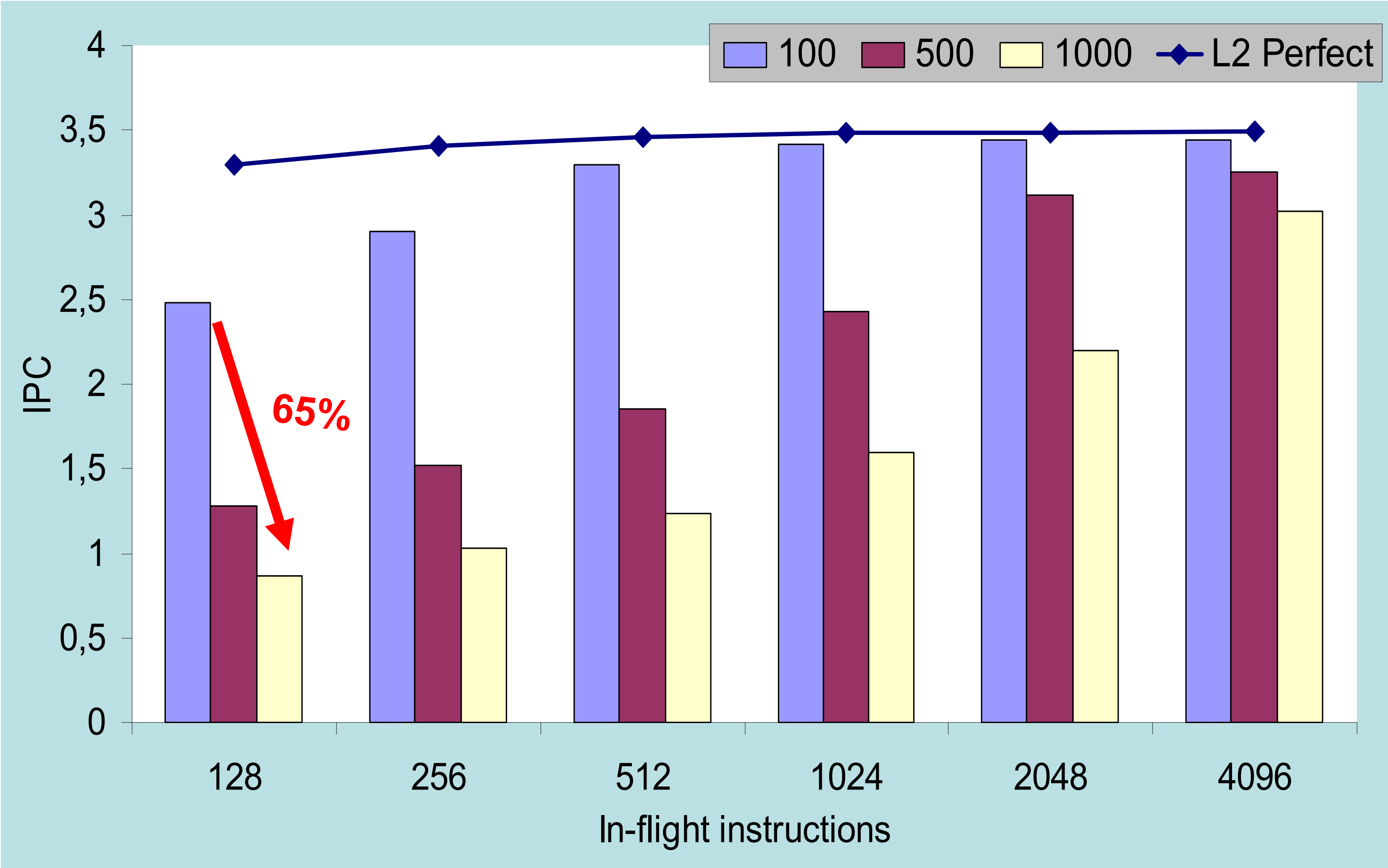
- Compute power - 504 Teraflops
  - 3,936 Sun four-socket blades
  - 15,744 AMD Opteron "Barcelona" processors
    - Quad-core, 2.0 GHz, four flops/cycle (dual pipelines)
- Memory - 125 Terabytes
  - 2 GB/core, 32 GB/node
  - 132 GB/s aggregate bandwidth
- Disk subsystem - 1.7 Petabytes
  - 72 Sun x4500 "Thumper" I/O servers, 24TB each
  - ~72 GB/sec total aggregate bandwidth
  - 1 PB in largest /work filesystem
- Interconnect - 10 Gbps / 2.3 μsec latency
  - Sun InfiniBand-based switches (2) with 3456 ports each
  - Full non-blocking 7-stage Clos fabric
  - Mellanox ConnectX IB cards

TACC

Ranger: All Racks & Power In Place

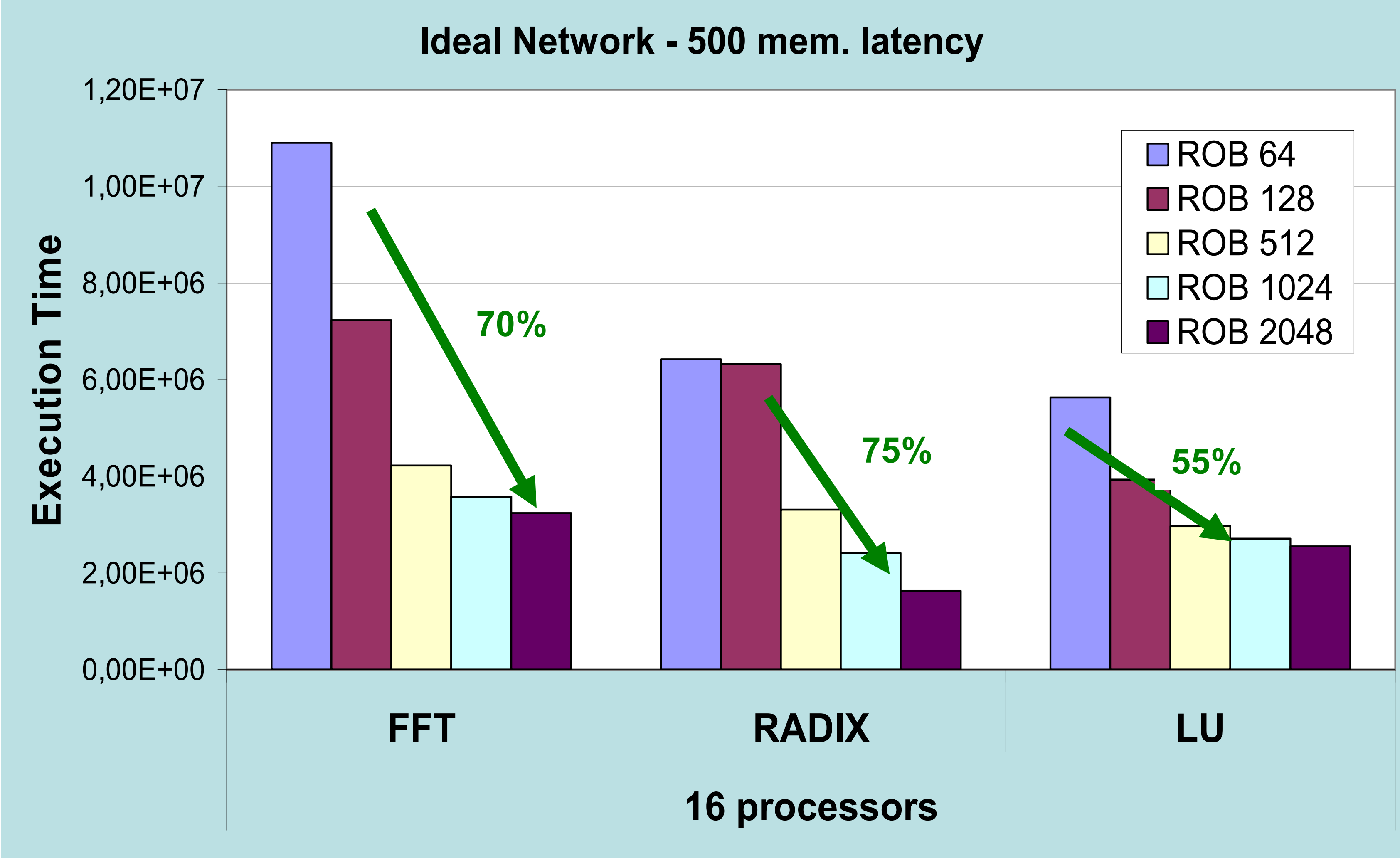
TACC

Kilo-Instruction Processors: hitting the memory wall



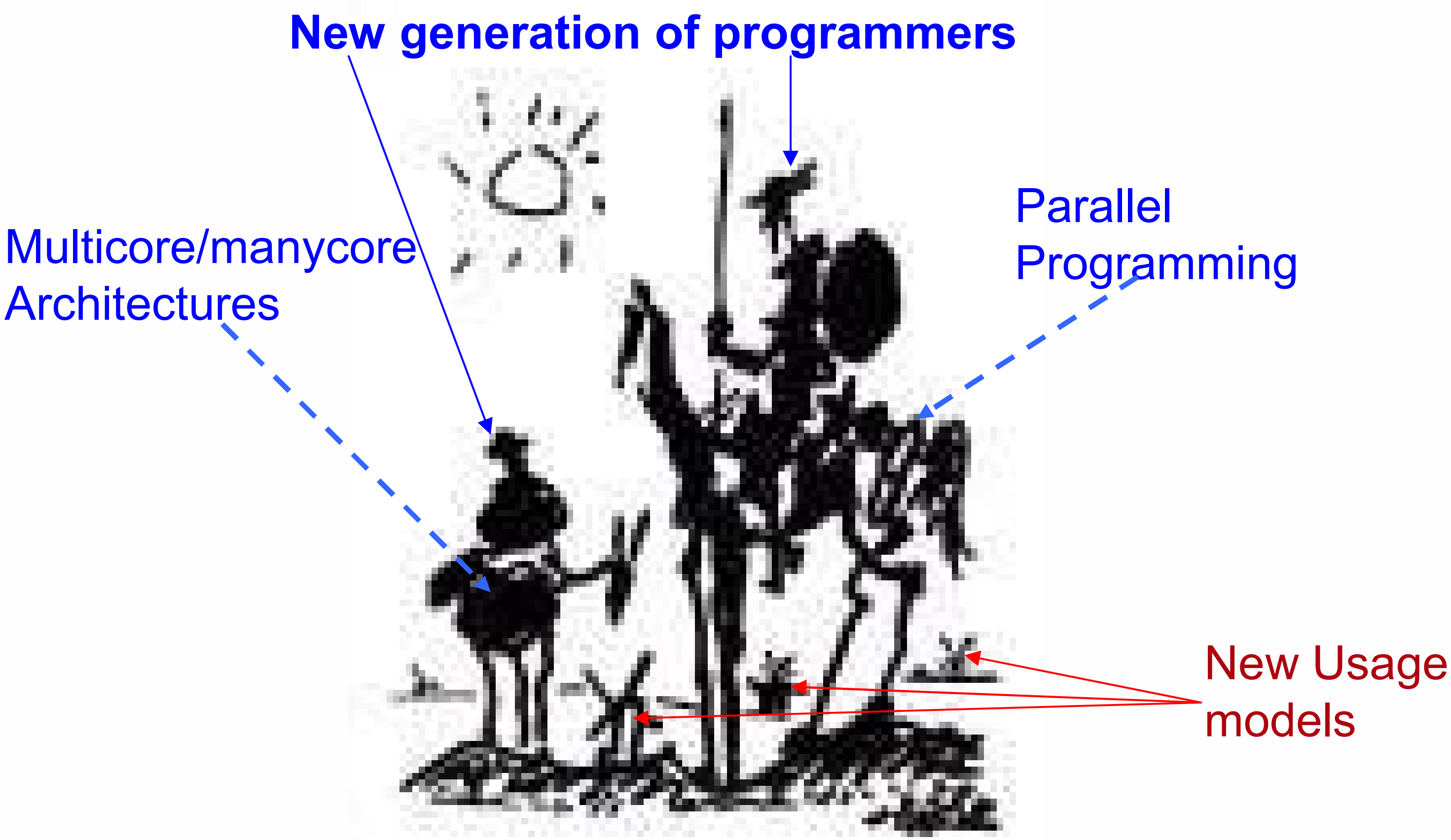
4-way, out-of-order processor - SpecFP 2000 benchmarks, from [Cri00]

Kilo-Instruction Multiprocessors



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

You will see…. in 400 years from now people will get crazy

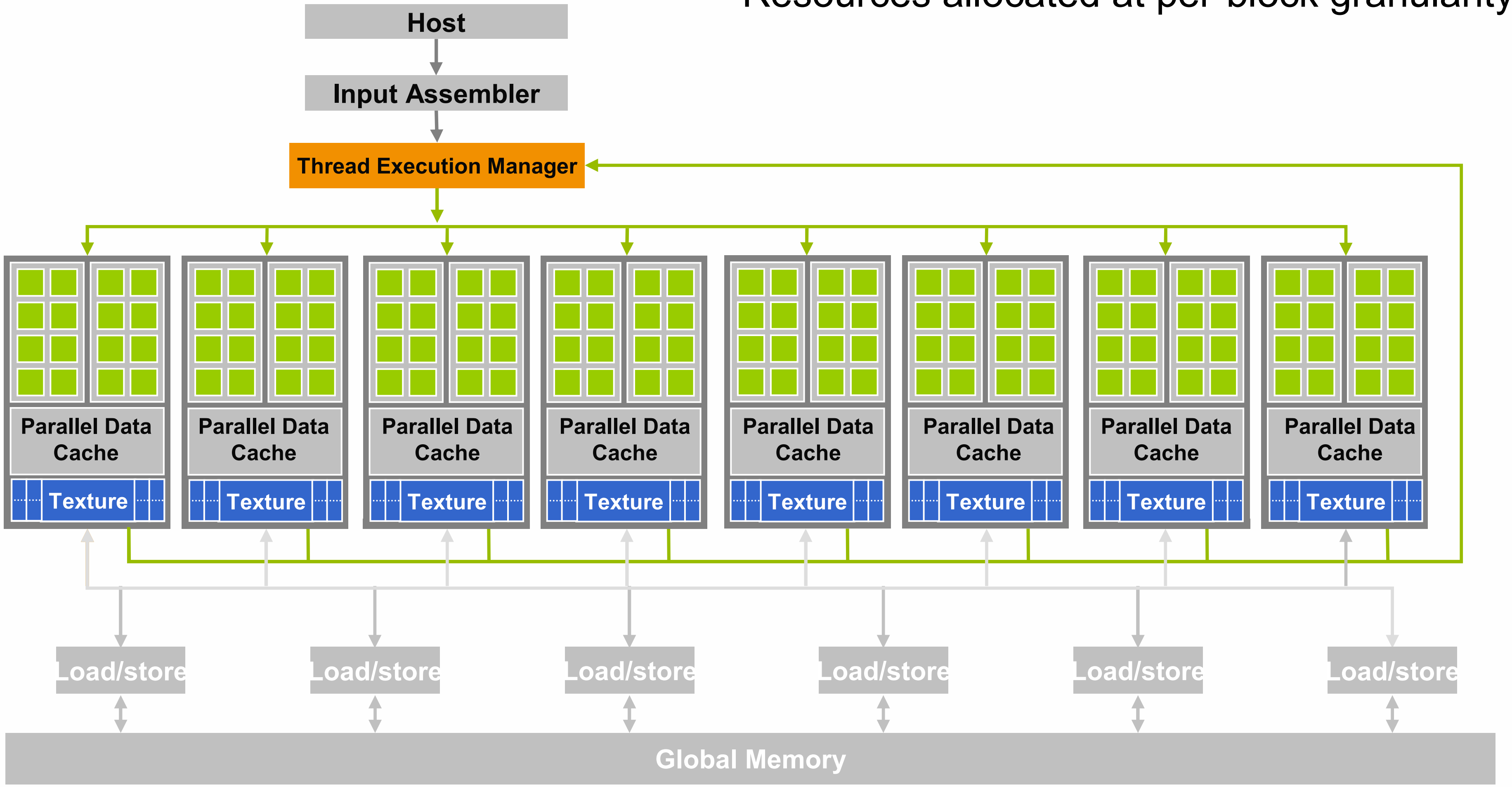


We have parallel systems today (Servers, HPC), but can we replace the “Big cores” with many small core that will run in parallel?

Dr. Avi Mendelson. Keynote at ISC-2007

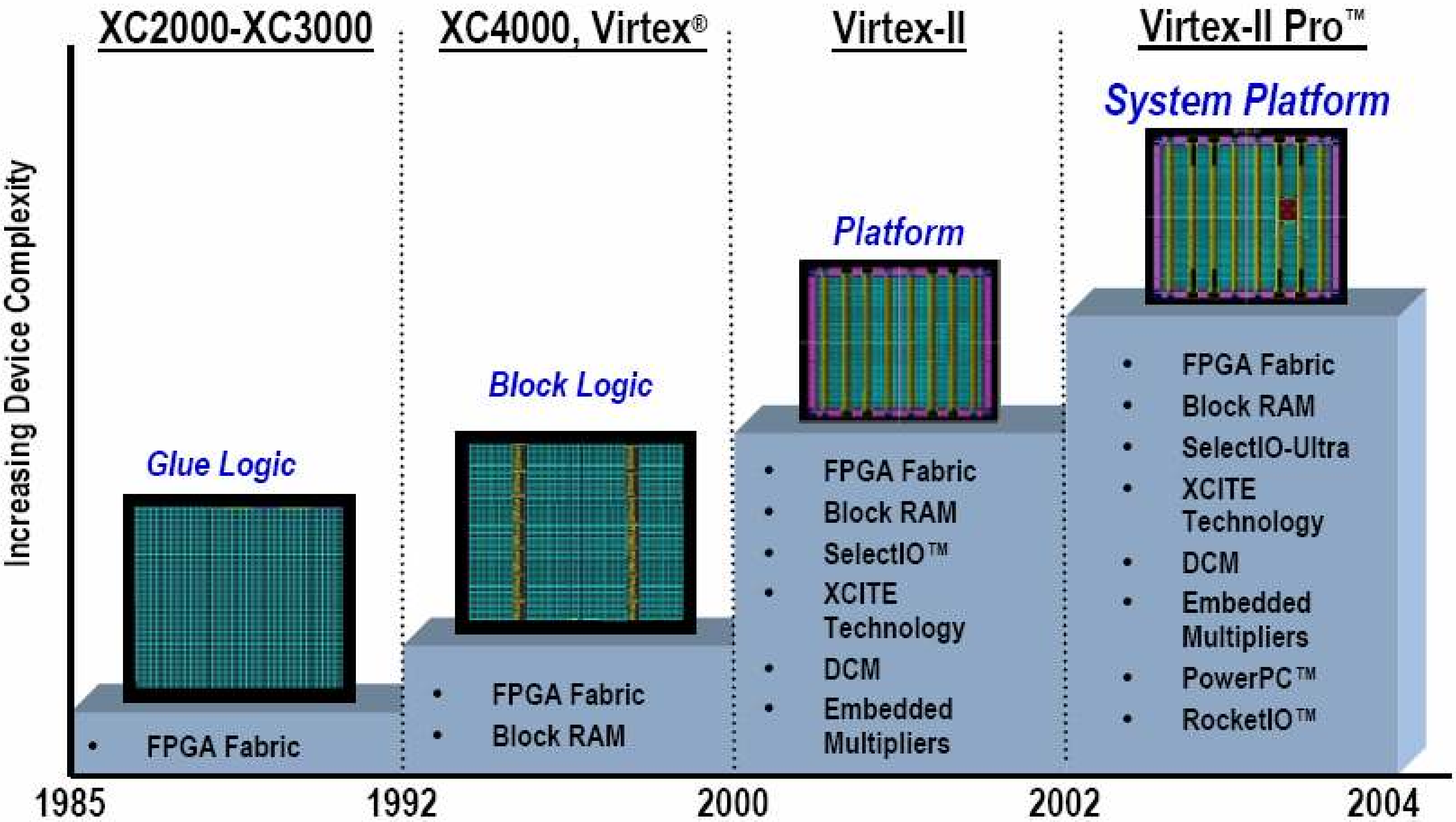
GeForce 8800 GPU Computing

Up to 12,288 active threads, 86.4 GB/s DRAM BW,
16 Streaming MP, 367 GFLOPS, 768 MB DRAM, 8GB/s PCIe
Resources allocated at per-block granularity



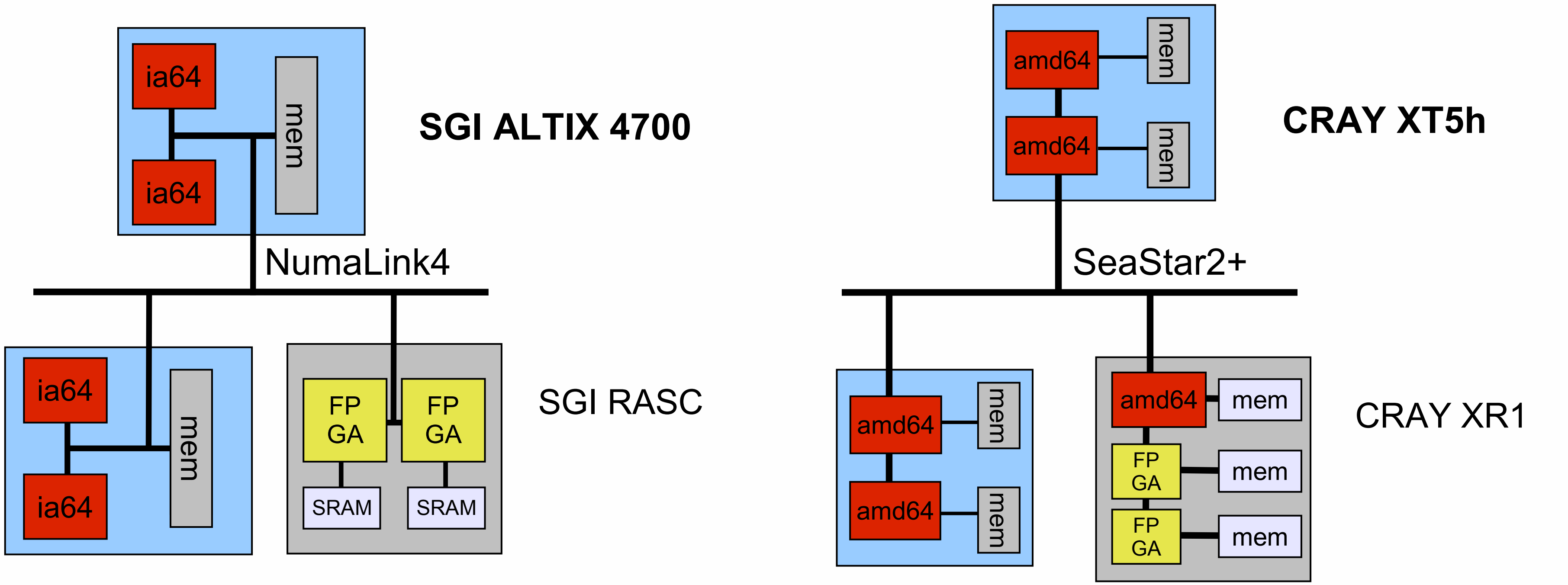
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The Evolution of Programmable Logic

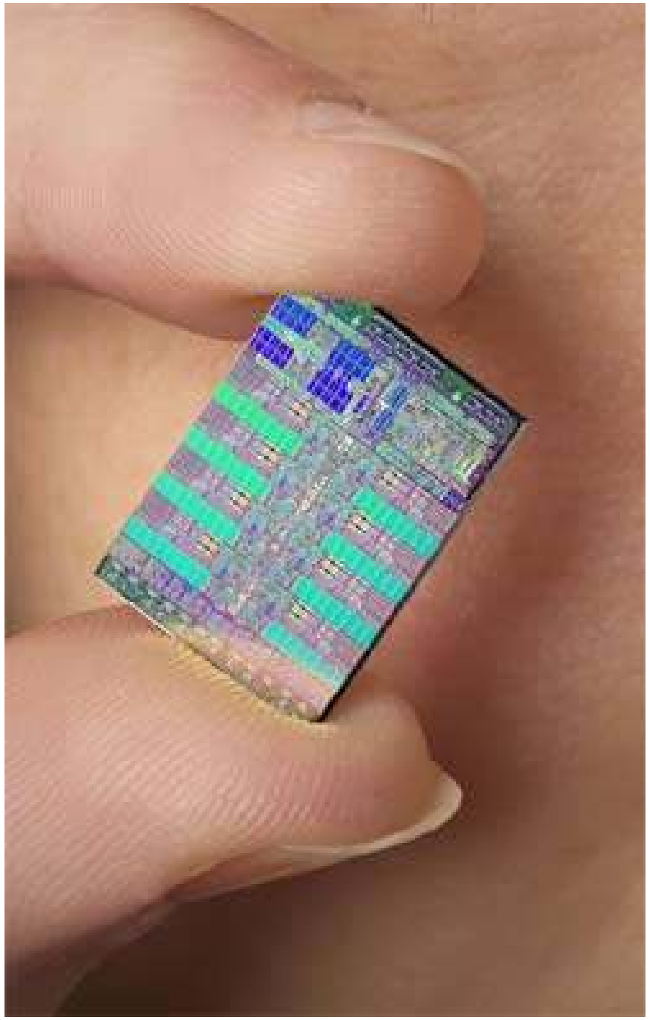


Heterogeneous Architectures Emerging

- New integrated architectures for HPC (Cray XT5h, SGI Altix 350/4700, SRC MAP, etc.)
- Socket plug-in modules (HyperTransport, FSB)
- Which system architecture to choose?



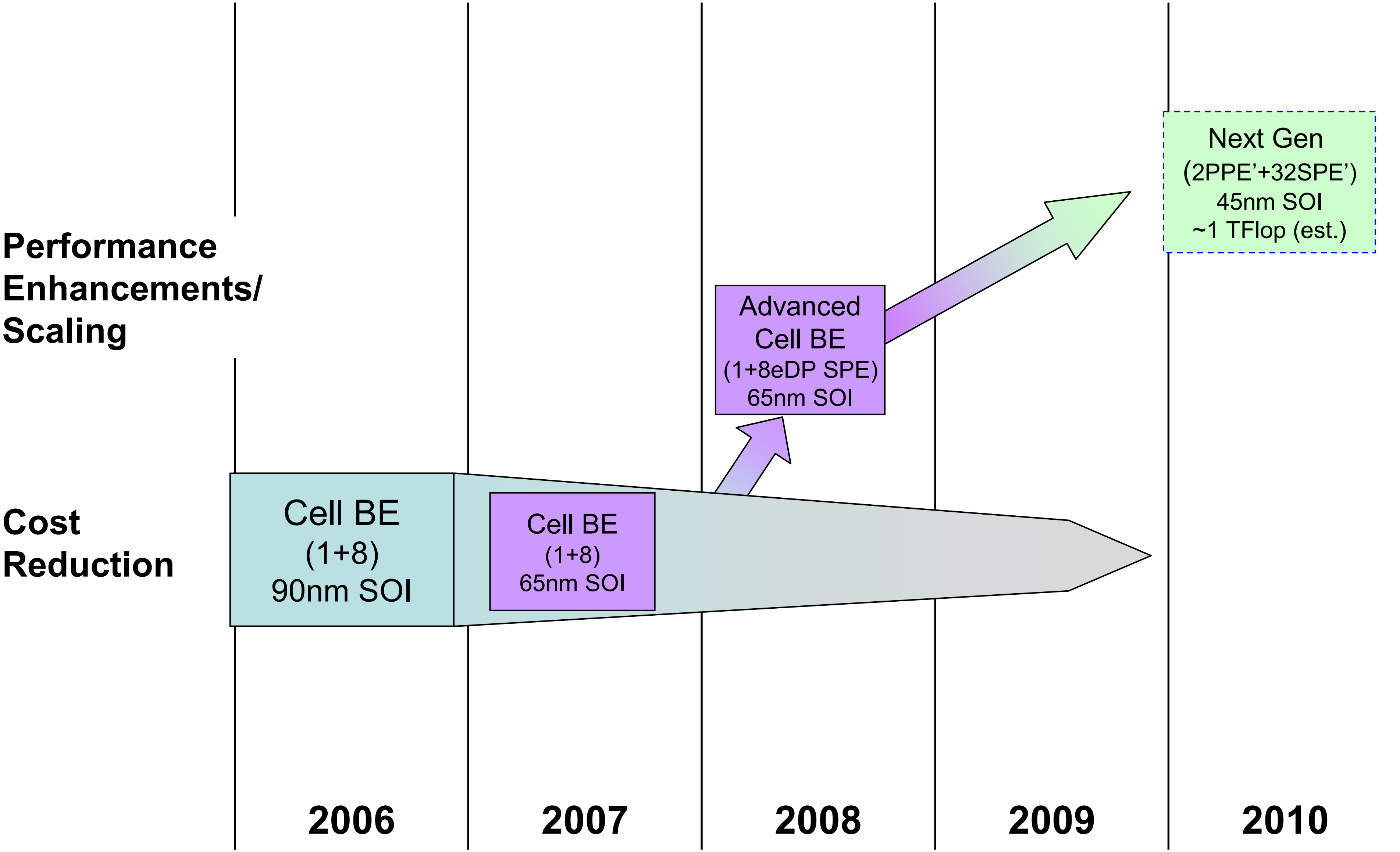
The CELL/B.E. chip



235 Mtransistors
235 mm^2

Roadrunner supercomputer
at
Los Alamos National Laboratory

Cell Broadband Engine Architecture™ Roadmap



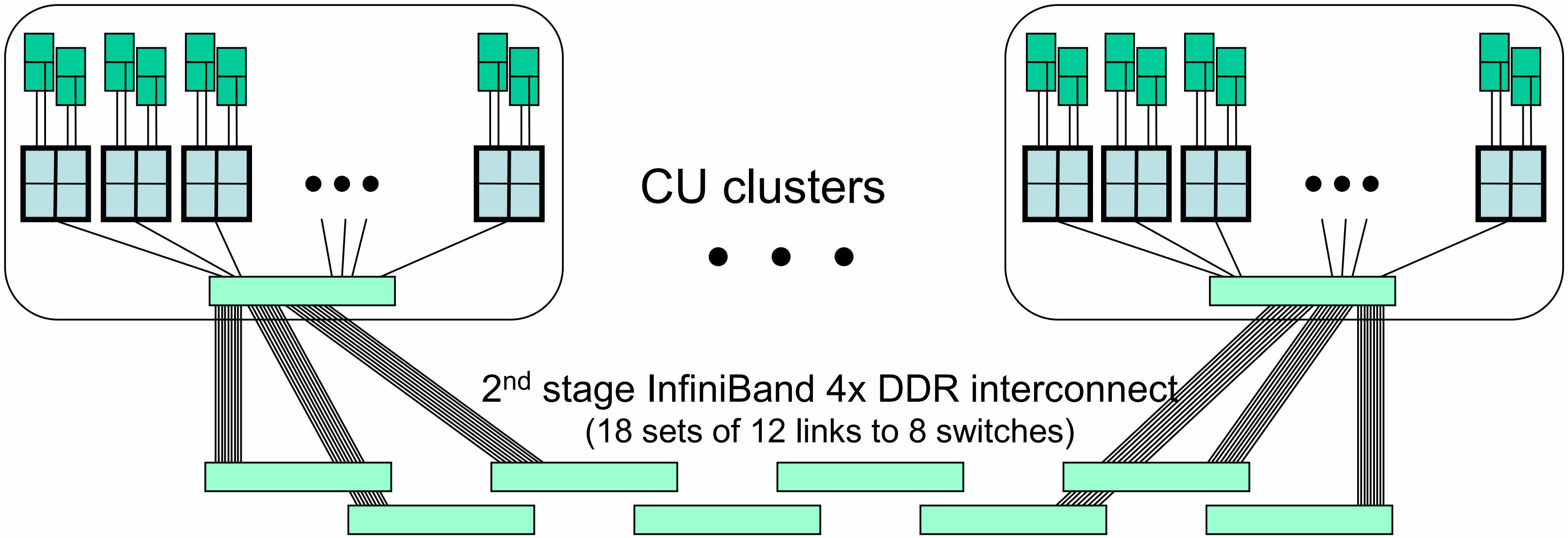
All future dates and specifications are estimations only; Subject to change without notice. Dashed outlines indicate concept designs.

Roadrunner Final System

“Connected Unit” cluster
192 Opteron nodes
(180 w/ 2 dual-Cell blades
connected w/ 4 PCIe x8 links)

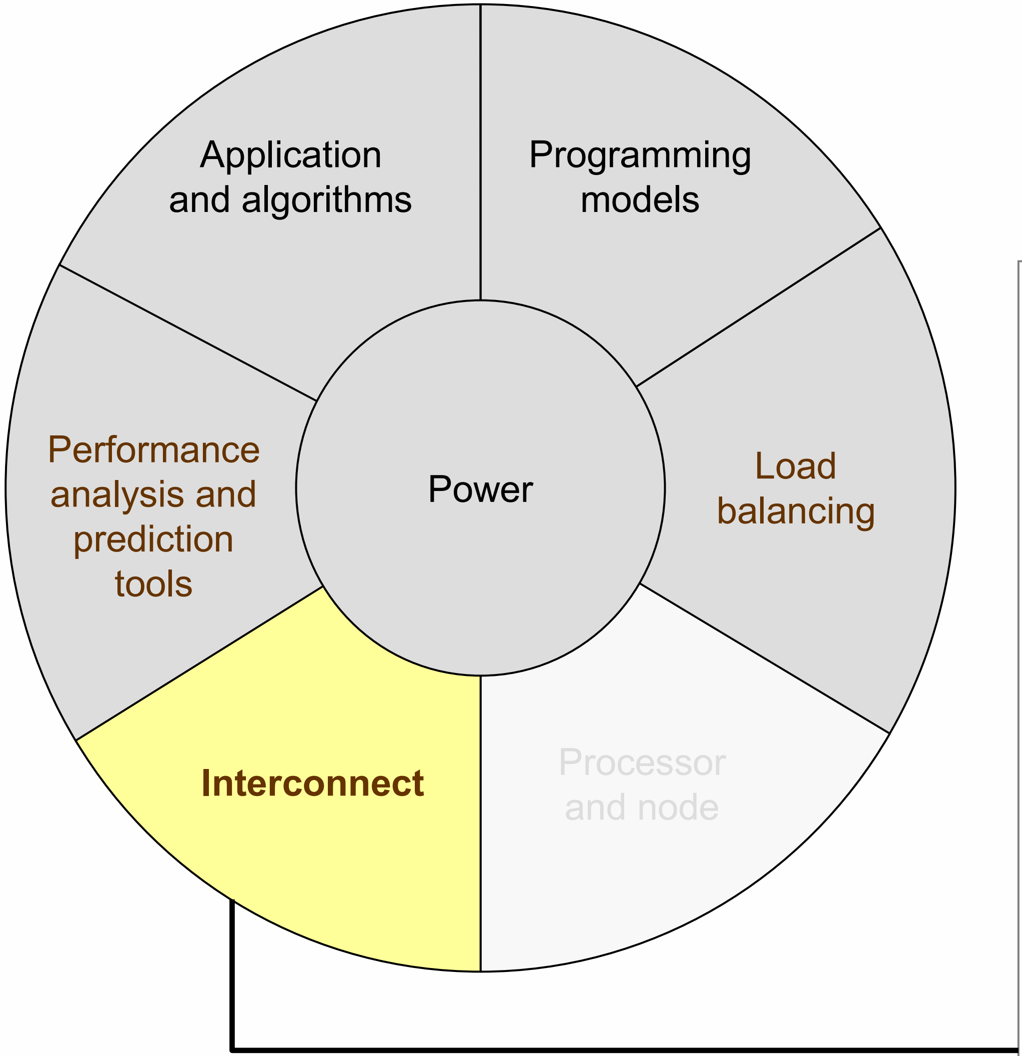
~7,000 dual-core Opterons
- ~50 TeraFlop/s (total)

~13,000 eDP Cell chips
- 1.4 PetaFlop/s (Cell)



2nd Generation IB 4X DDR

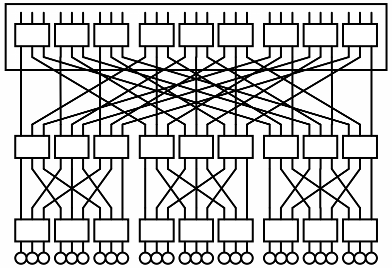
Multidisciplinary top-down approach



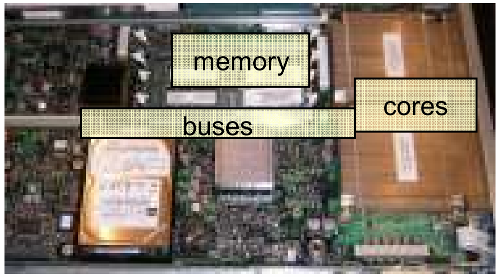
- Importance of the different networks in a Supercomputer
- Communication patterns from the applications
- Latency and bandwidth
- Overlapping Communication and Computation
- Multipath routing
- Optical interconnects

Network integration

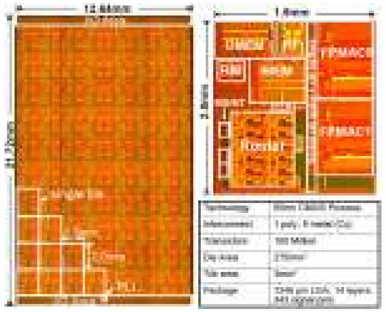
- Between nodes
  - Infiniband, Myrinet, ...
  - 3D Torus



- Inside a node
  - Buses to memory



- Network on Chip
  - Buses: CellBE
  - Direct topologies: Intel's 80 core Polaris

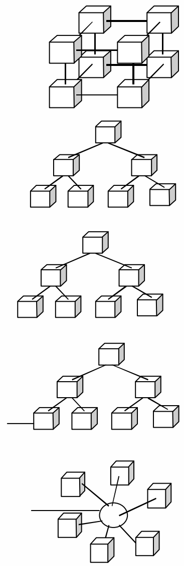


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

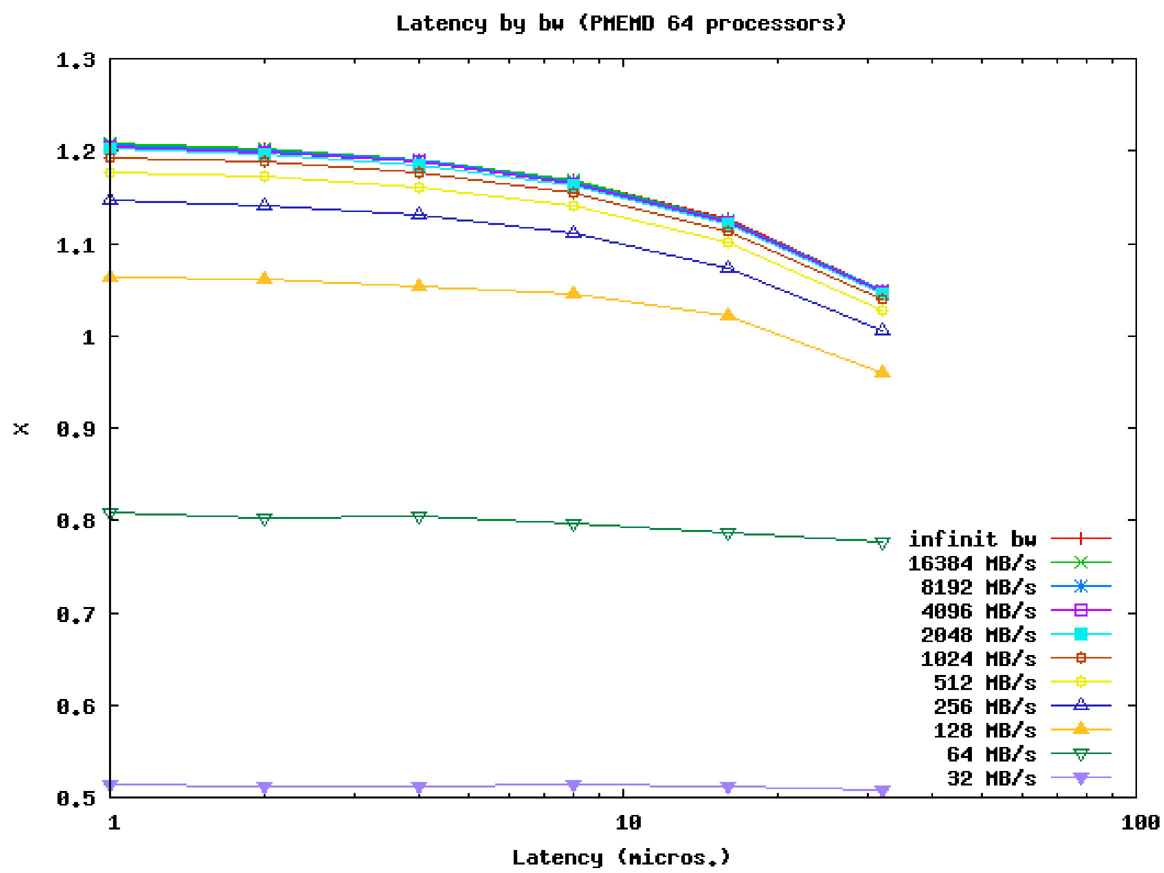
Supercomputer networks

- In the last November Top10 list
  - 4 BlueGenes with 3D Torus Networks
  - 3 Cray XT4 also with 3D Torus Networks
  - 3 Xeon platforms with Infiniband

- 5 independent networks in BlueGene
  - 3D torus: point-to-point
  - Collective network: global operations
  - Global barriers and interrupts
  - Gbit ethernet: file I/O and host interface
  - Control network: boot, monitoring and diagnostics

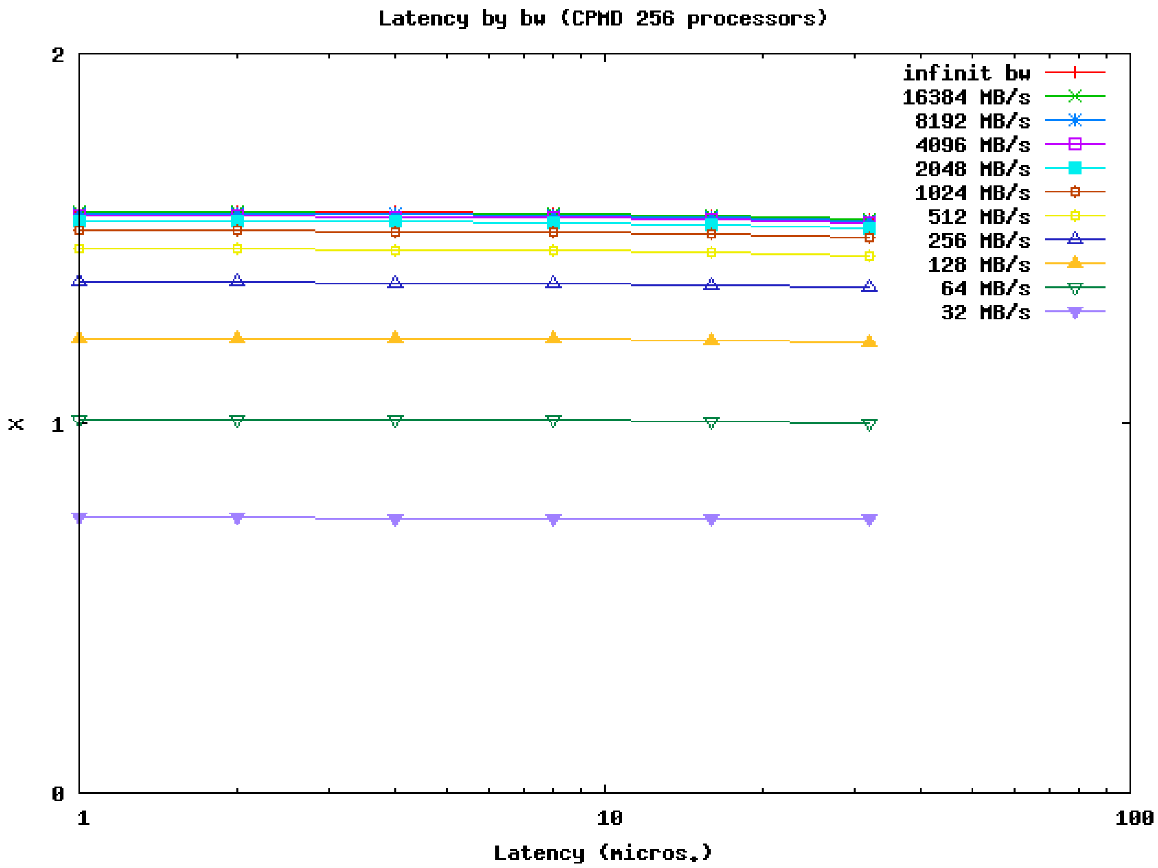


Scientific workloads and network parameters



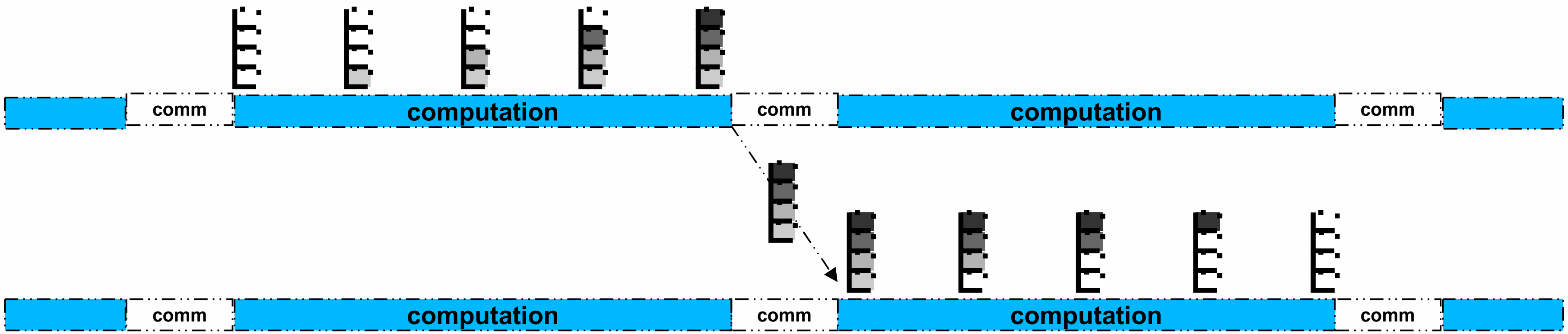
- Low impact of latency (5-10%), compared to bandwidth (-50% to 20%)
- Amber execution, 64 tasks; simulations with different bw and latency

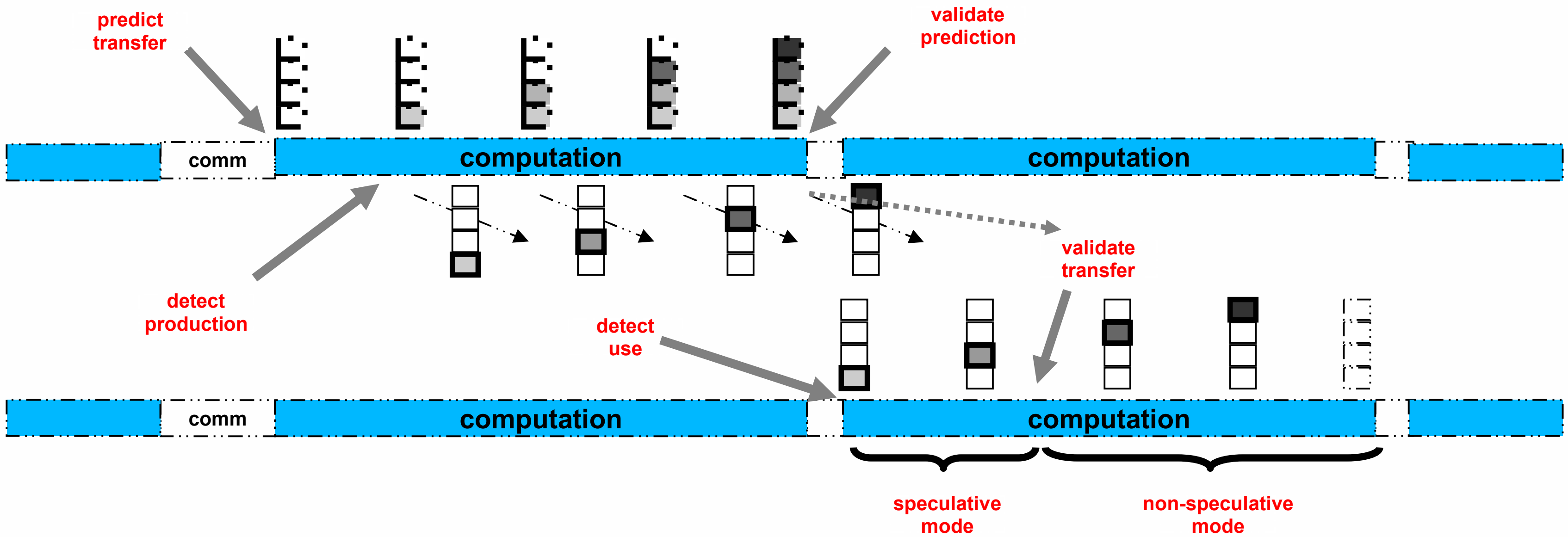
Scientific workloads and network parameters (II)



- No impact of latency, only bandwidth is relevant
- CPMD execution, 256 tasks; simulations with different bw and latency

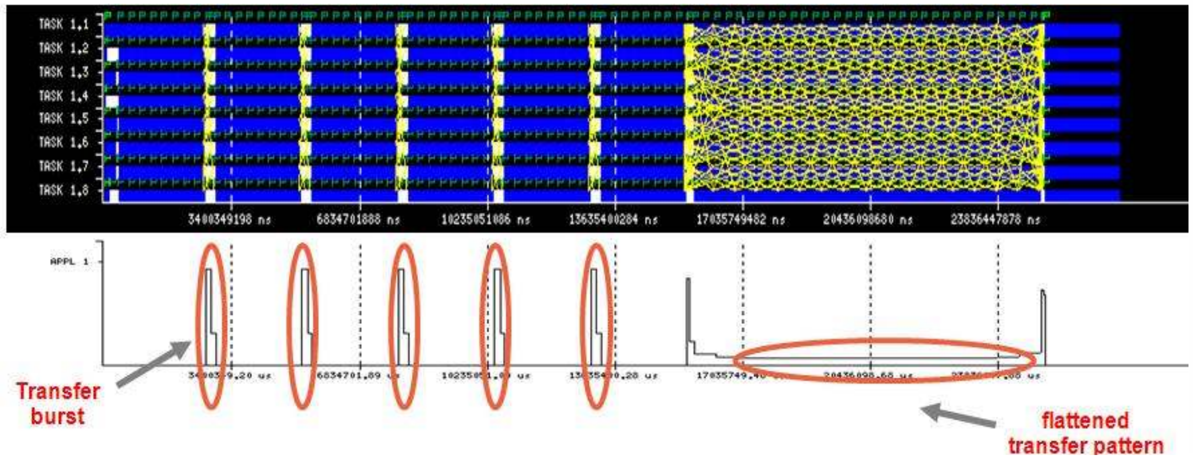
Speculative dataflow





Barcelona Supercomputing Center
Centro Nacional de Supercomputación

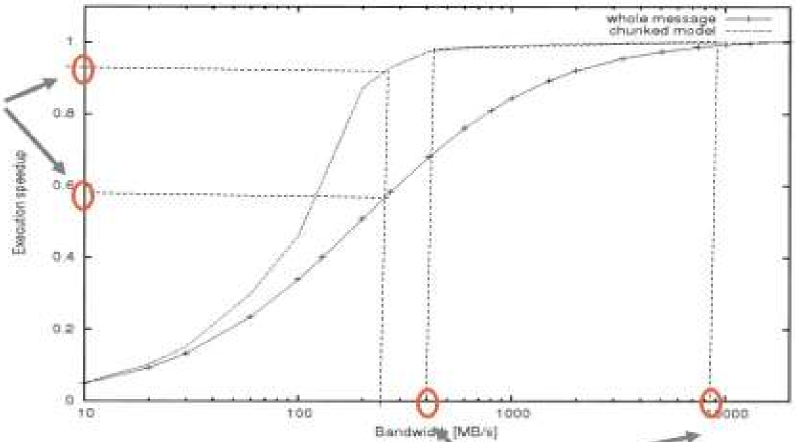
Effects on bandwidth



flattening
communication pattern

thus

speedup of 1.6 for
the same network bandwidth



for the same execution time
20 times lower needed bandwidth

reducing
bandwidth requirements

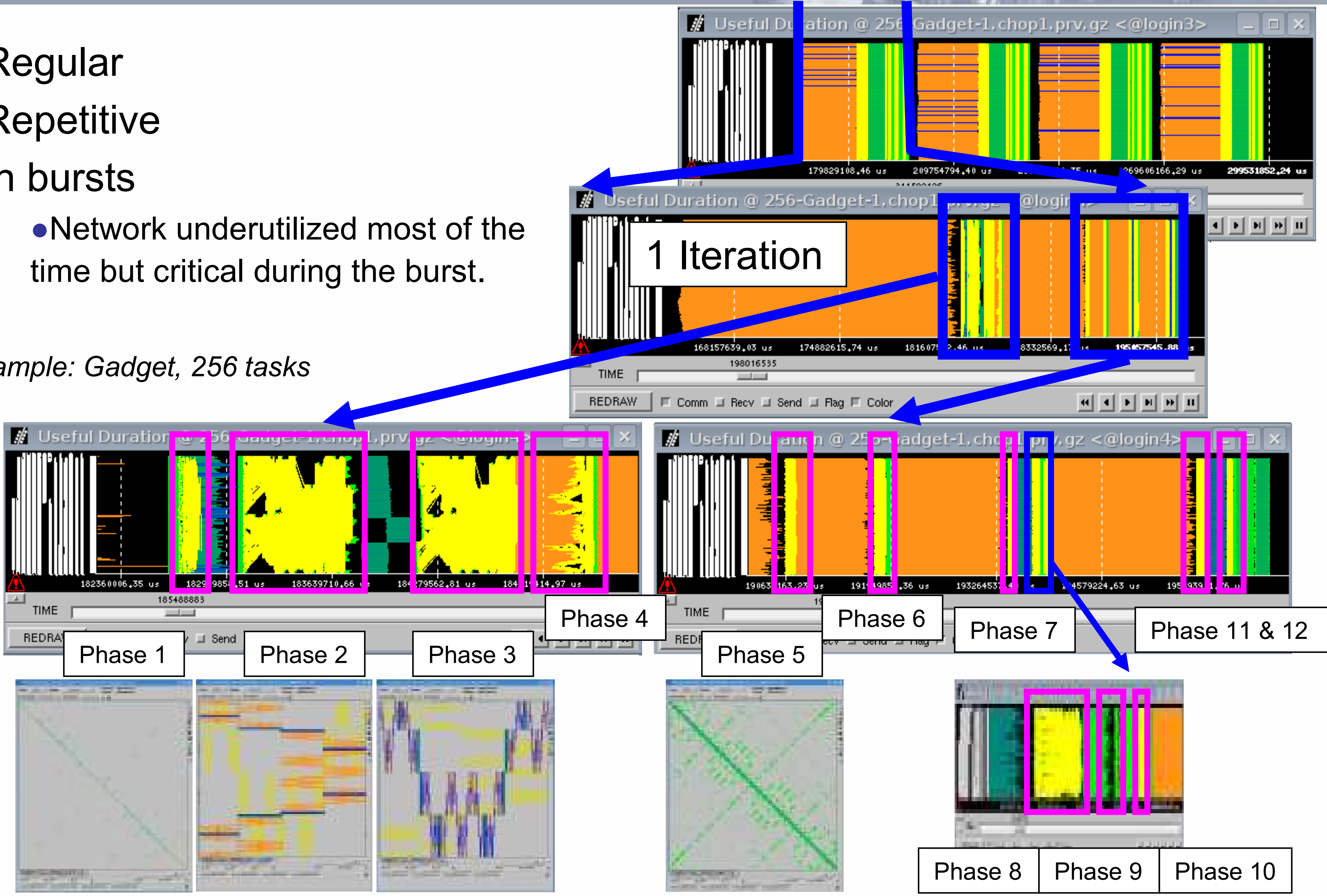
*simulation on application with
ring communication pattern

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Scientific applications communication patterns

- Regular
- Repetitive
- In bursts
  - Network underutilized most of the time but critical during the burst.

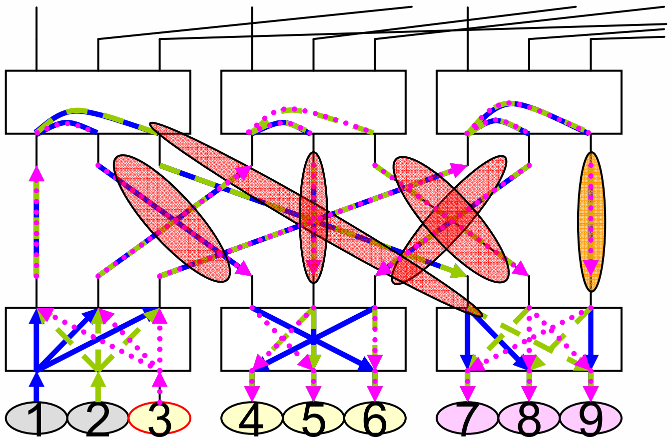
Example: Gadget, 256 tasks



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Better routes, better mapping

Routes without information of Comm.Pattern

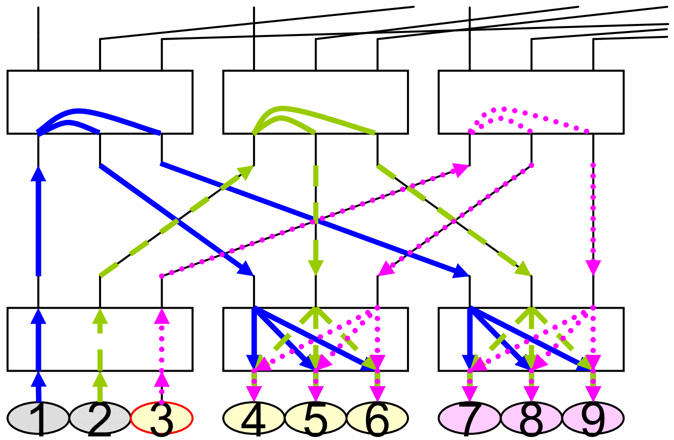


18 conflicting communications!

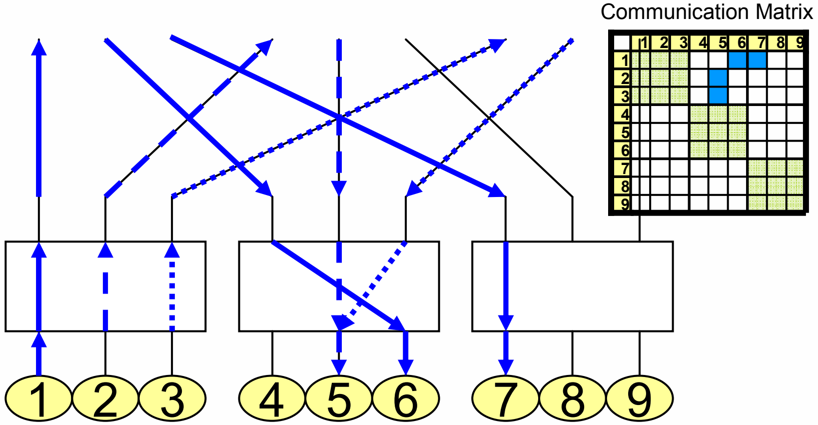
Better Routing

Heuristic Routing

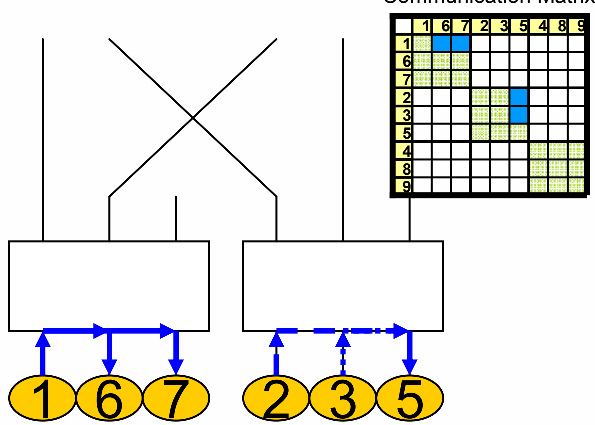
Based on Comm. Pattern



9 Links, No conflicts



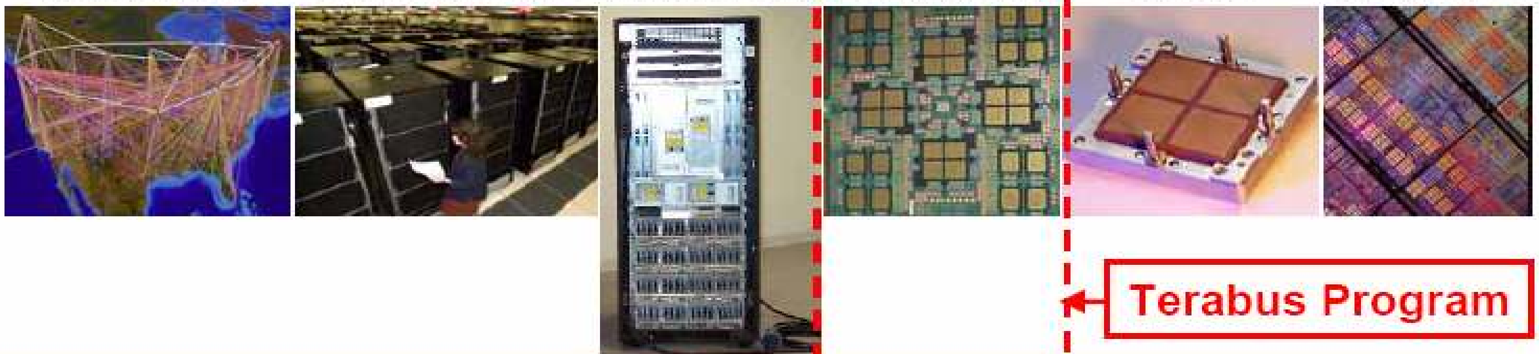
Better Mapping (METIS)



Evolution of Optical interconnects

Time of Commercial Deployment (Copper Displacement):

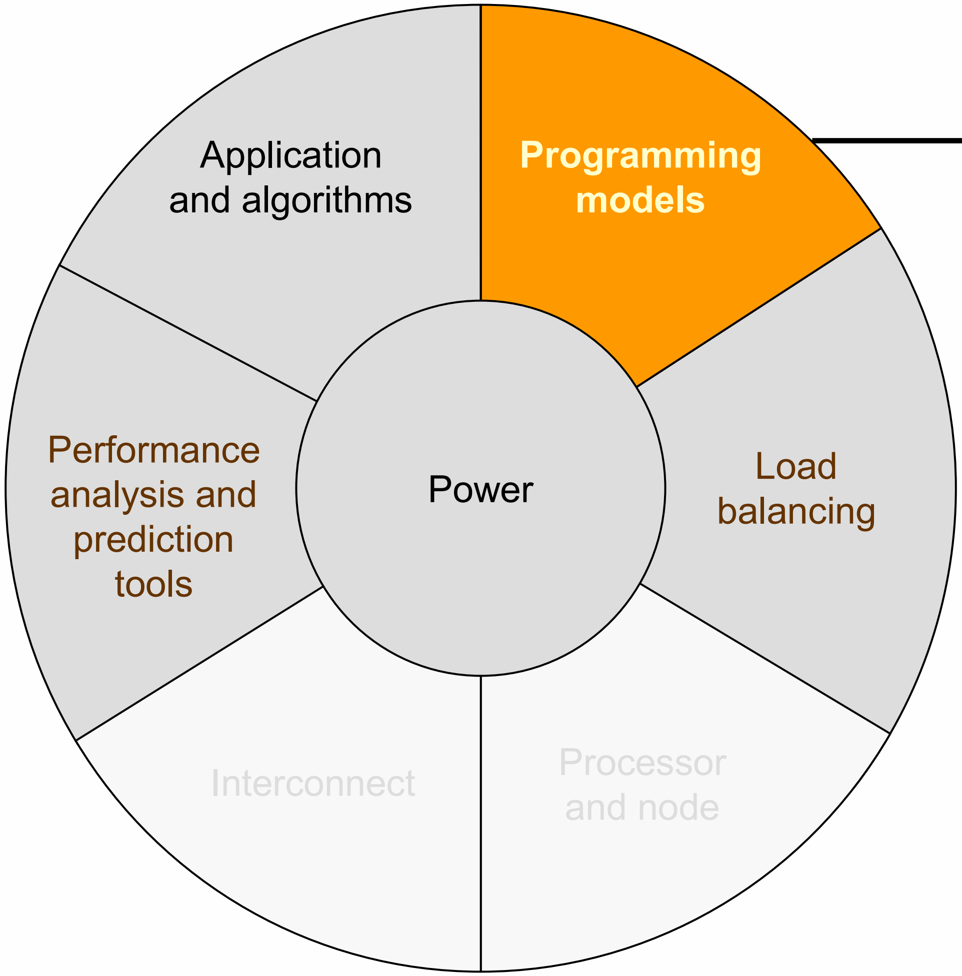
1980's	1990's	2000's	> 2010		
WAN, MAN	**LAN**	**System**	**Board**	**Module**	**Chip**
metro,long-haul	campus, enterprise	intra/inter-rack	module-module	chip-chip	on-chip buses



Terabus Program

Distance	10's – 100's km	10m – 2km	<10 intra <100 inter	< 1 m	< 10 cm	< 20 mm
# of lines	singles	tens	100's-1000's	1000's	10000's	100,000's
Cost ($/Gb/s)	1000			1		10^{-5}
Power (mW/Gb/s)	500			5		0.5
Density (Gb/s/mm²)	10^{-3}			10		1000

Multidisciplinary top-down approach



- Massive burden on programmers
- Back to Babel

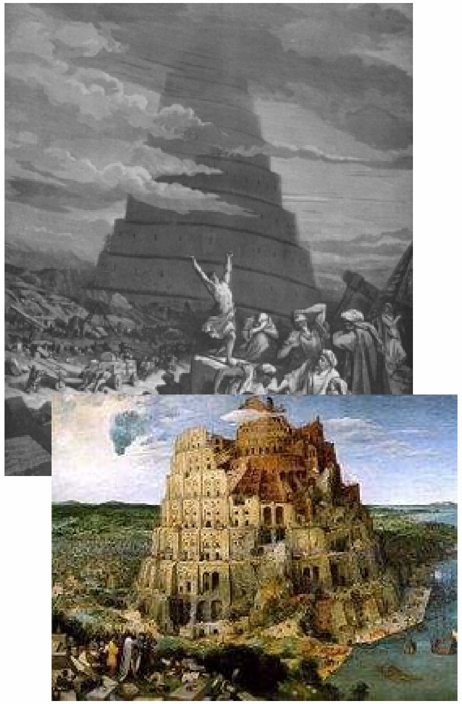
Back to Babel?

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"…

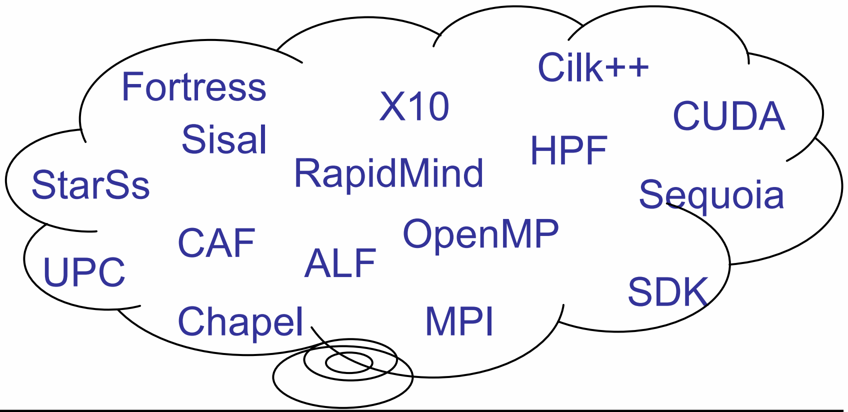


And the LORD said, "Look, they are one people, and they have all one language; and this is only the beginning of what they will do; nothing that they propose to do will now be impossible for them. Come, let us go down, and confuse their language there, so that they will not understand one another's speech."

The computer age

Fortran & MPI

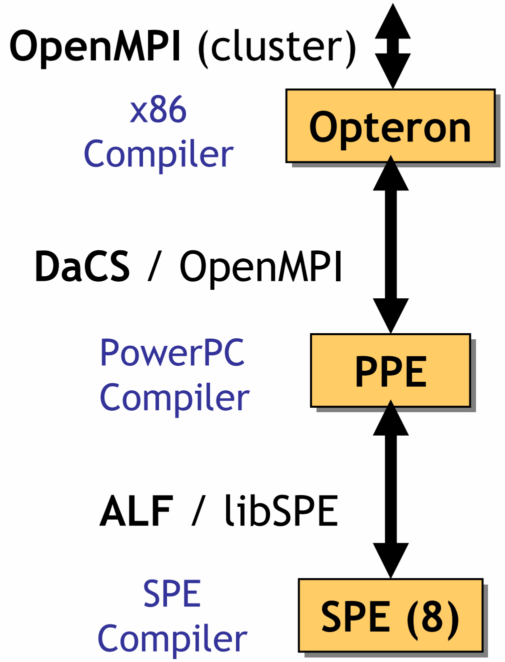
++



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

A simple case ☺: the Cell/B.E.

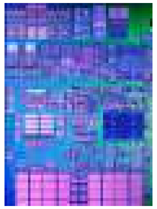
- Libraries
  - libSPE, DaCS, ALF, …
  - Complete modification of your code



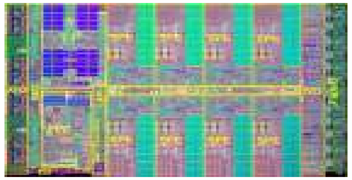
- Follow the standards (i.e. OpenMP)
  - Software cache (runtime/compiler)
  - Tiling and prefetching (compiler)
  - What about performance?

- New programming models
  - CellSs
  - Proof-of-concept implementations that may influence standards

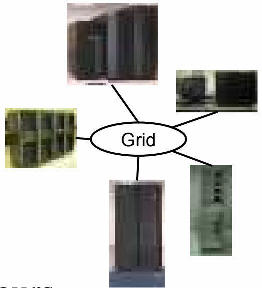
A scaled view of architectures and programming models



≅



≅



ns ➔ 100 useconds ➔ minutes/hours

Mapping of concepts:

Instructions	➔ Block operations	➔ Full binary
Functional units	➔ SPUs	➔ machines
Fetch &decode unit	➔ PPE	➔ home machine
Registers (name space)	➔ Main memory	➔ Files
Registers (storage)	➔ SPU memory	➔ Files

Cell/Grid/SMP Superscalar (StarSs)

standard sequential programming: “easy”

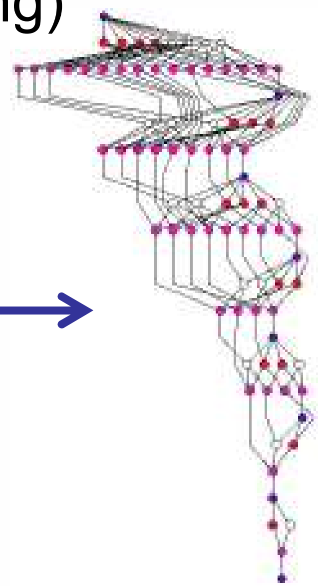
“decent” performance

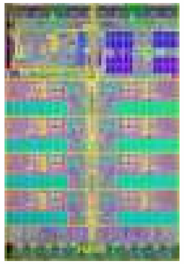
Portable. One language, multiple run times

Propose new programming models: CellSs

- Simple programming model for the Cell/B.E. …
  - allows easy porting of applications
  - oriented towards the exploitation of functional parallelism from a sequential application with annotated functions
- … and a runtime system
  - dynamically exploits functional parallelism (true dependences)
  - removes false dependences (renaming)

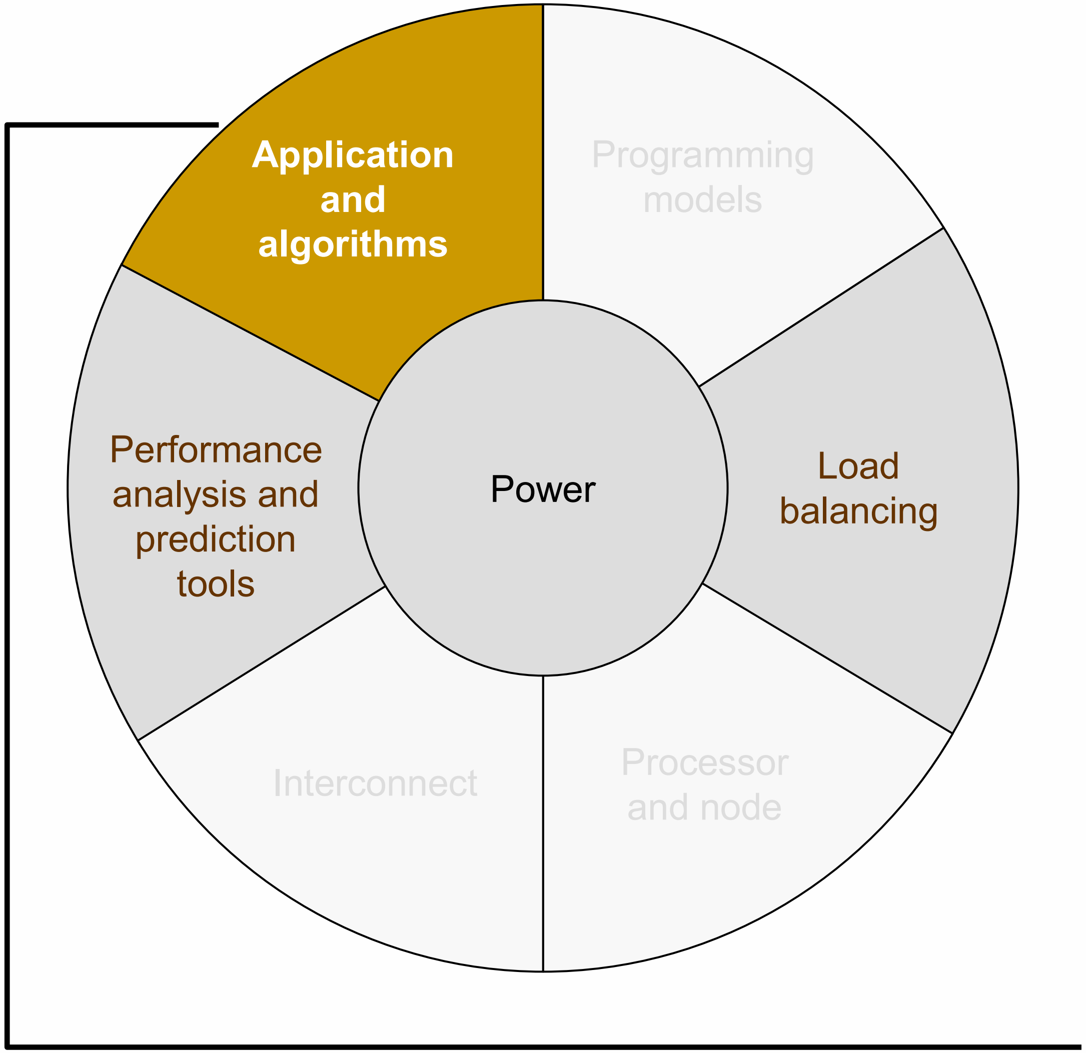
```
#pragma css task inout(diag[B][B]) highpriority
void lu0(float *diag);
#pragma css task input(diag[B][B]) inout(row[B][B])
void bdiv(float *diag, float *row);
#pragma css task input(row[B][B],col[B][B])
                 inout(inner[B][B])
void bmod(float *row, float *col, float *inner);
#pragma css task input(diag[B][B]) inout(col[B][B])
void fwd(float *diag, float *col);
```

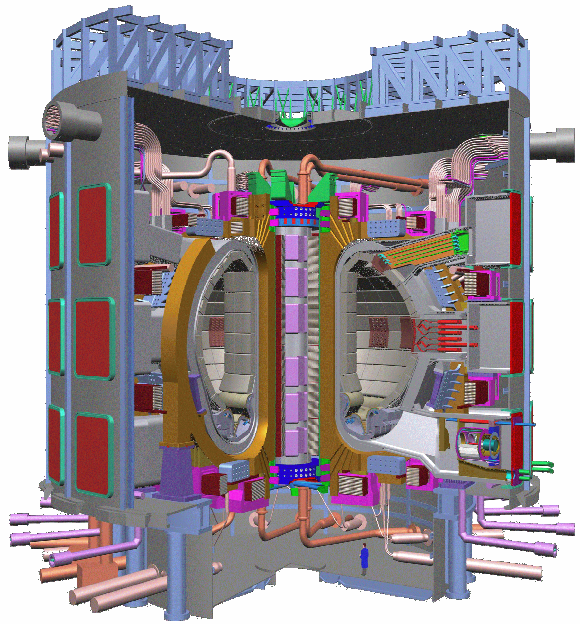




Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Multidisciplinary top-down approach





- Improving applications
- Improving algorithms and methods

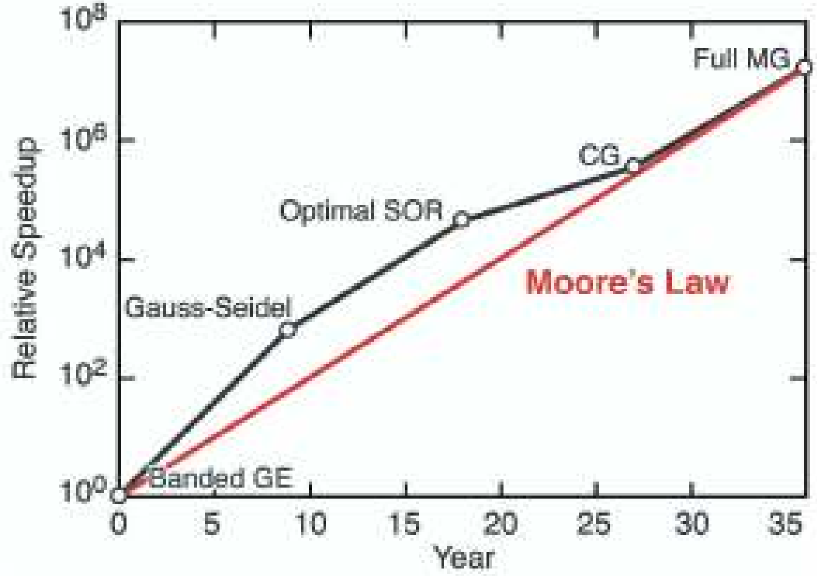
Algorithm kernels

- Traditional Numerical Kernels continue
  - Sparse Linear algebra
  - Dense Linear Algebra
    - BLAS
    - Linear systems
    - Eigenvalues
  - Discretization methods (FD, FE, FV, BE)
  - FFT and other transforms
  - Random number generation

- Algorithm improvement in the last 20 years similar to Moore's Law

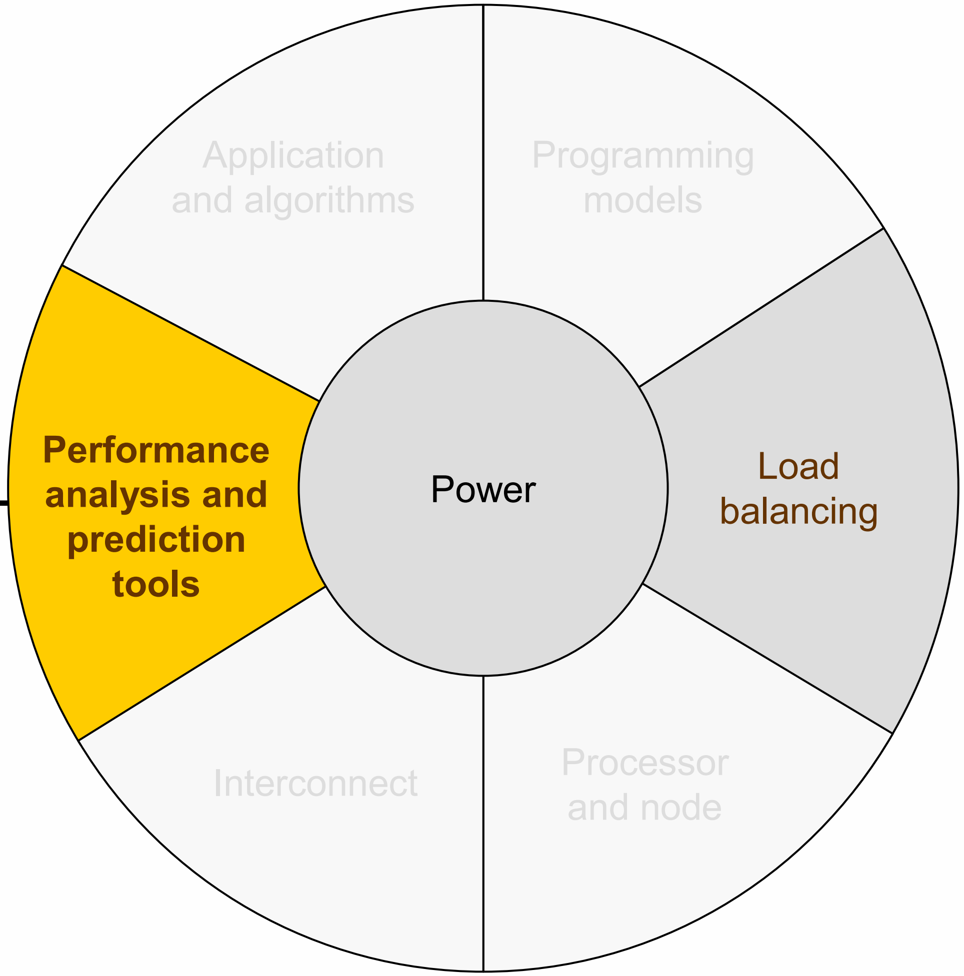
- Emphasis on
  - Memory bandwidth, QoS,..
  - Asynchronism, data flow

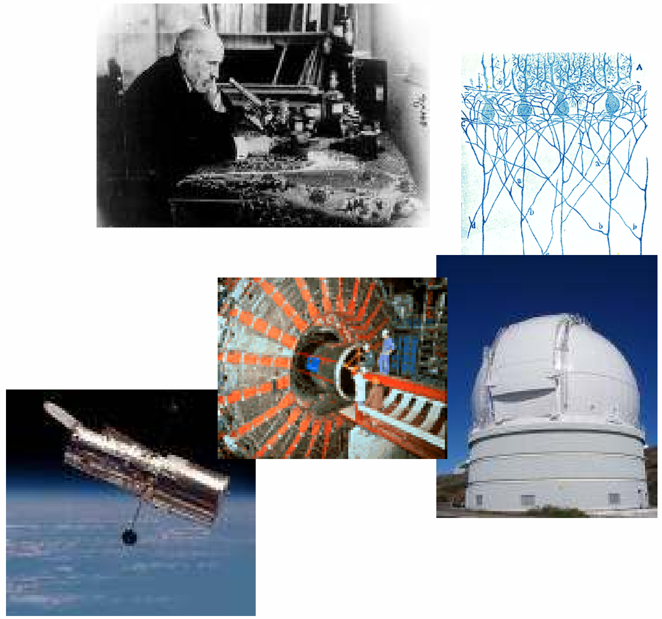
Method	Storage	Flops
GE (banded)	n^5	n^7
Gauss-Seidel	n^3	$n^5 \log n$
Optima l SOR	n^3	$n^4 \log n$
CG	n^3	$n^{3.5} \log n$
Full MG	n^3	n^3



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Multidisciplinary top-down approach



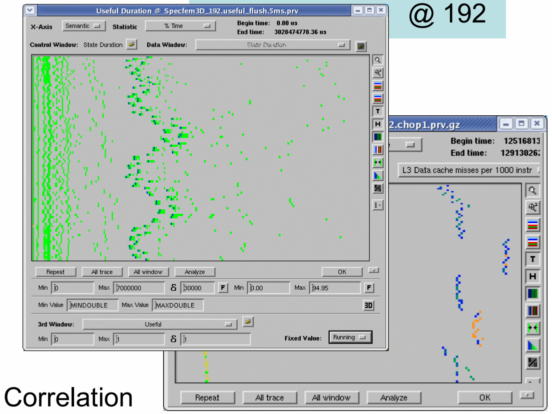


- Site, application and system optimization
- Scalability and intelligence

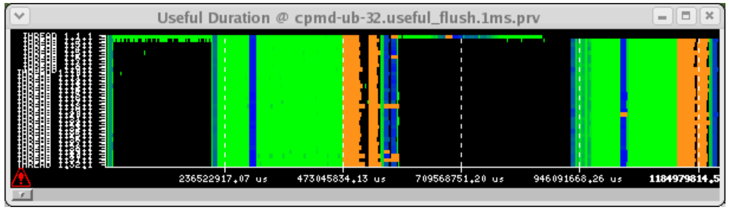
The need of performance analysis tools: Who

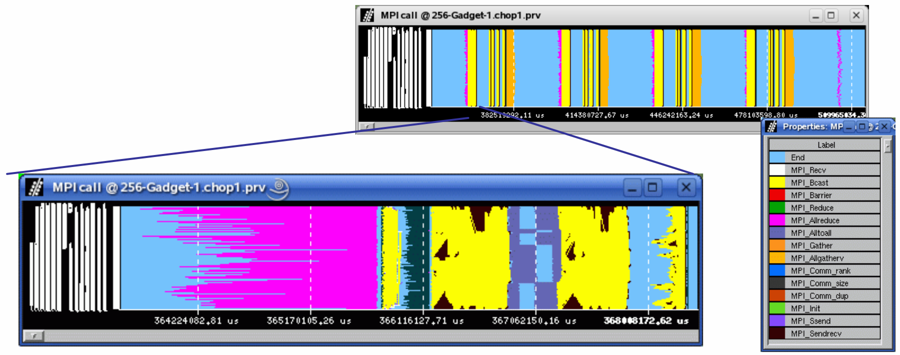
- Users, application developers
  - To confirm assumed behavior (very often reality is different from preconceived)
  - Provide expectations of impact to be used for decision support
    - New machines
    - Tuning efforts → potential rewards
- Operations
  - To plan and ensure proper resource utilization
- System developers
  - To understand global impact of proposed features

Duration SPECFEM3D @ 192



Correlation
load imbalance - L2 misses





CEPBA-Tools environment

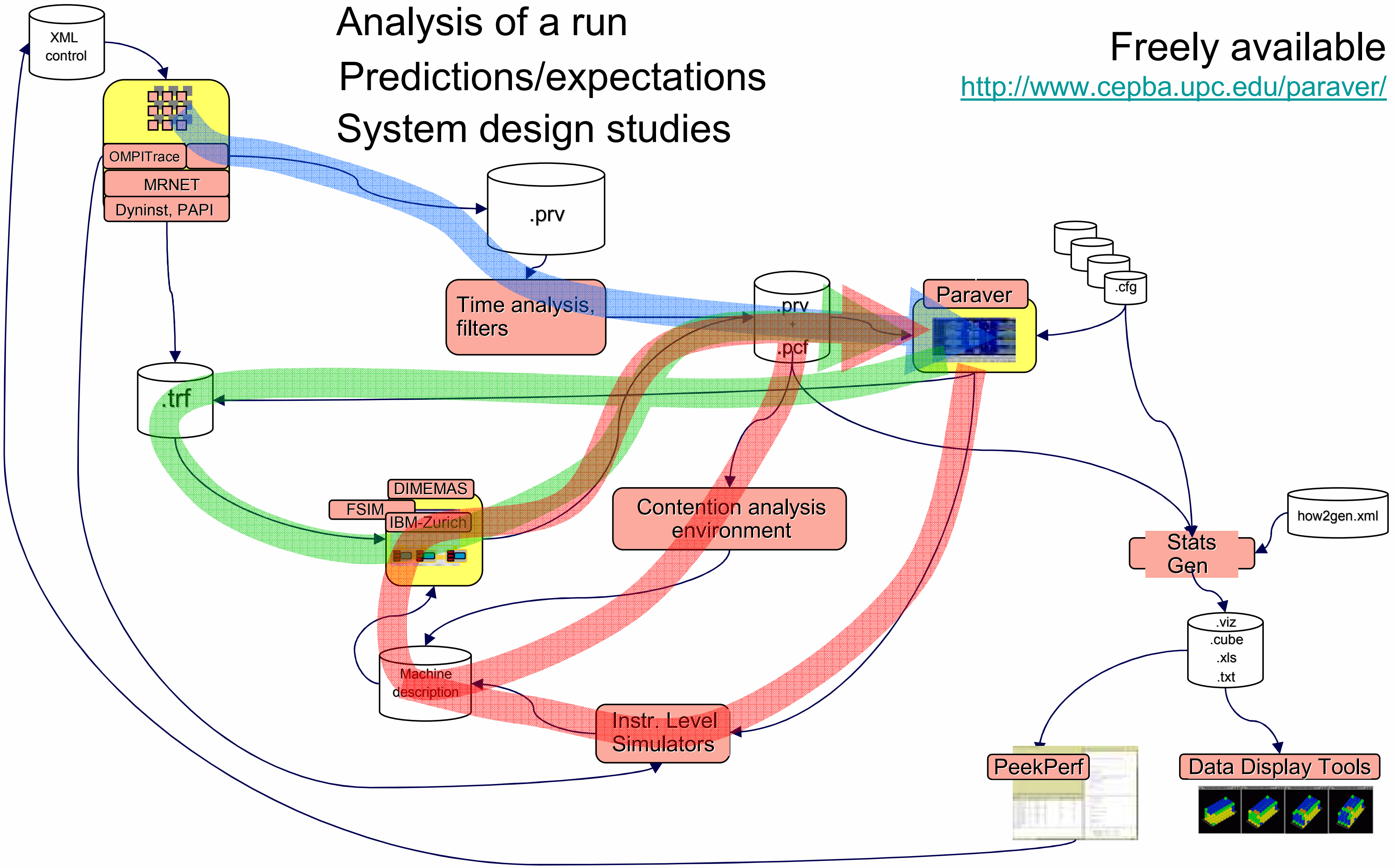
Analysis of a run

Predictions/expectations

System design studies

Freely available

http://www.cepba.upc.edu/paraver/

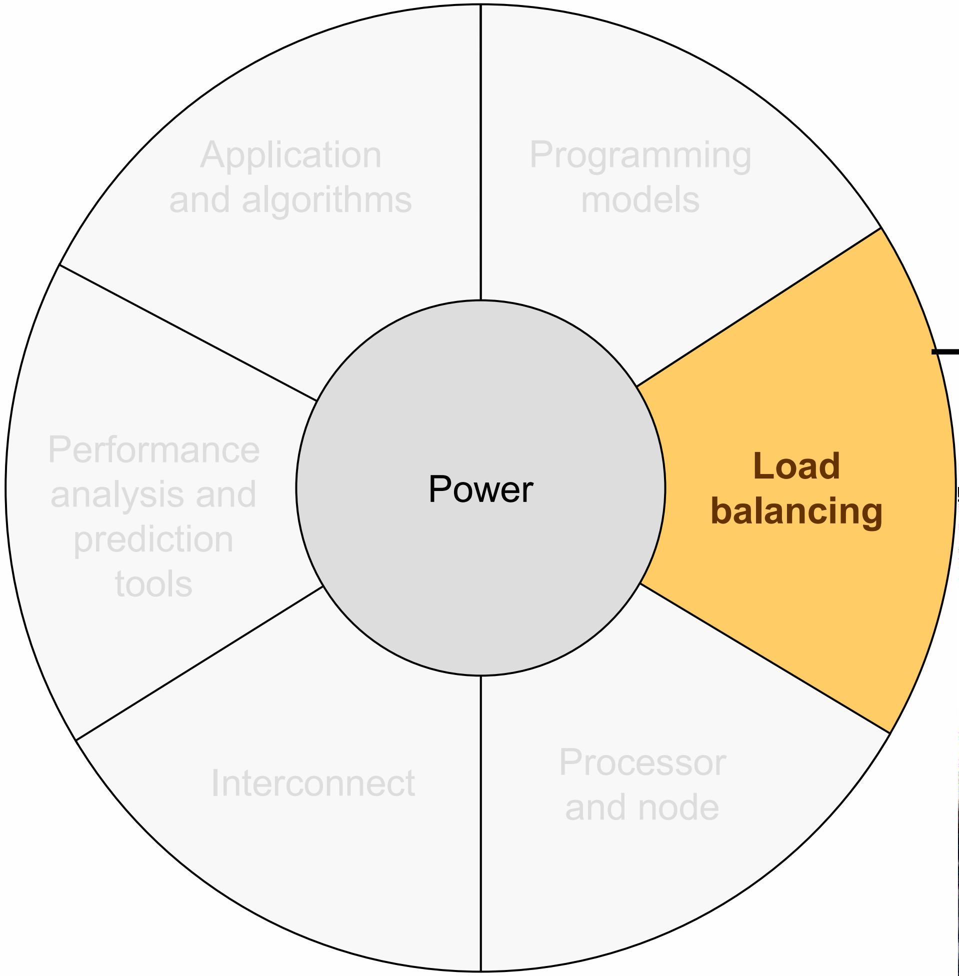


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Performance analysis tools: issues

- Scalability
  - Dynamic range: from long term behavior @ 10K cores to detailed impact of cache or core microarchitecture.
  - Handling huge amounts of data.
- Intelligence
  - sumarizing / Datamining → useful information (leading to right decisions)
  - Models

Multidisciplinary top-down approach

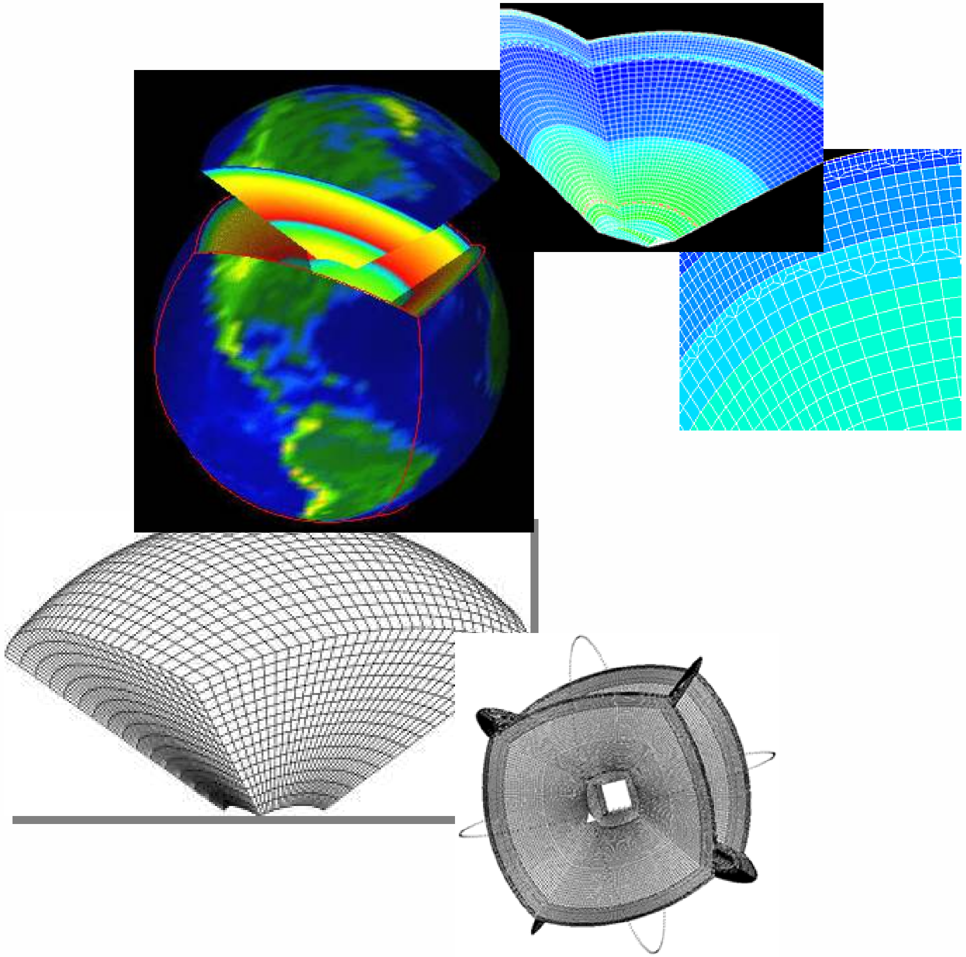


- Load balancing in my application? No.
- SMT/CMP priorities/QoS

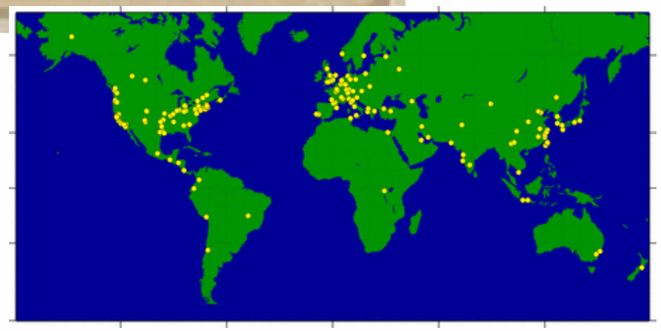
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Specfem3D: a “true” story

- Should I introduce asynchronous communication?

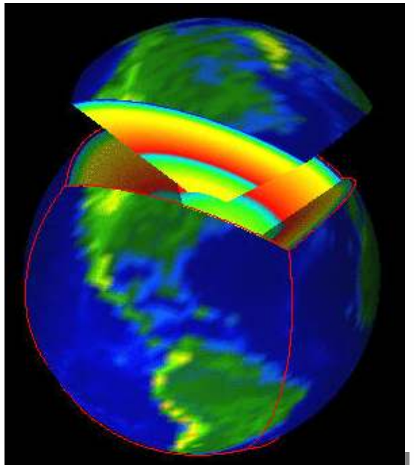


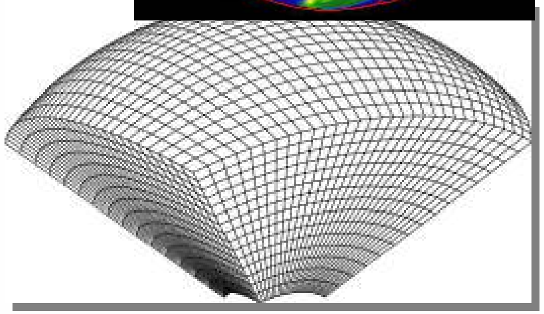
Courtesy Dimitri Komatitsch



Specfem3D

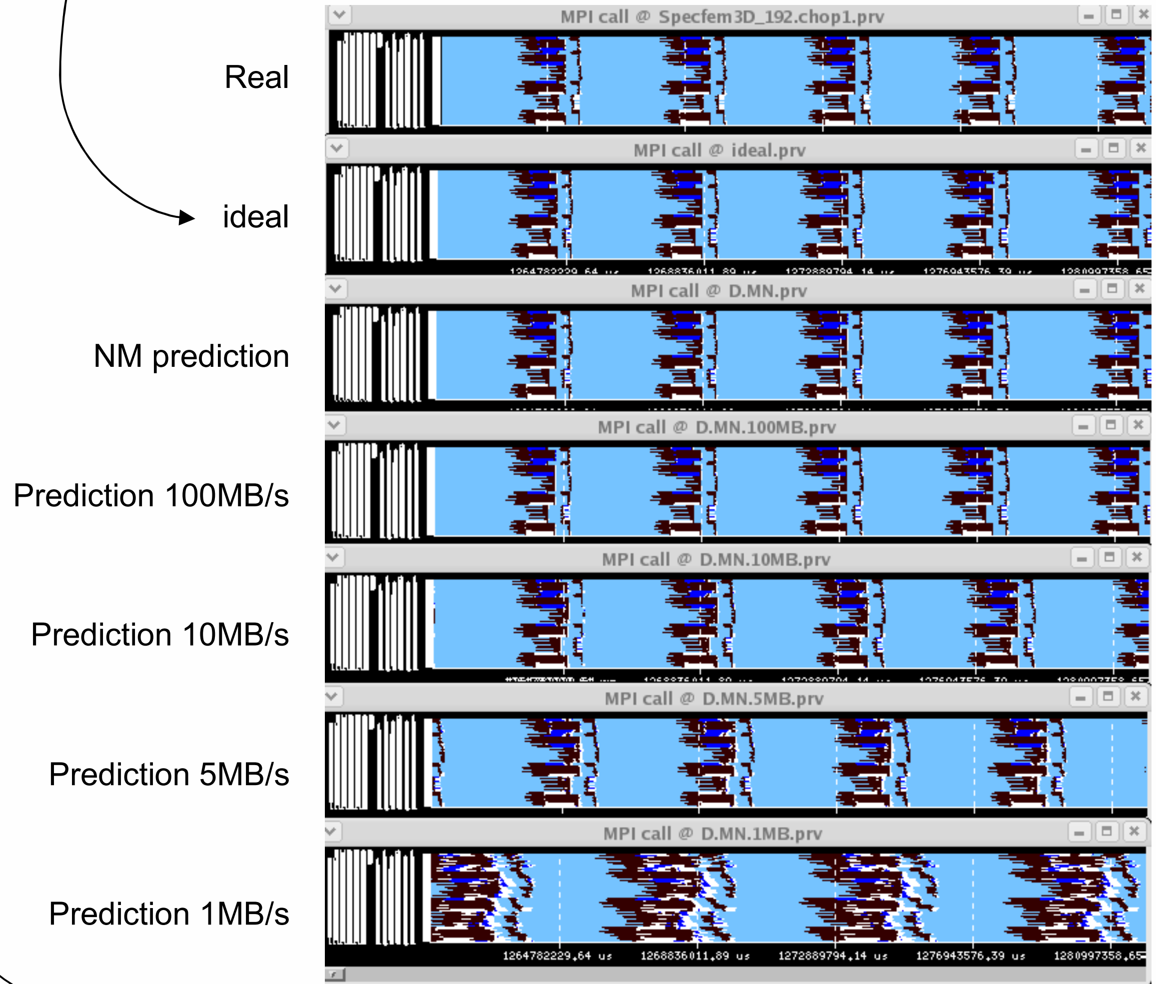
- Should I introduce asynchronous communication?





Courtesy Dimitri Komatitsch

Prediction:
IDEAL communications NO IMPACT

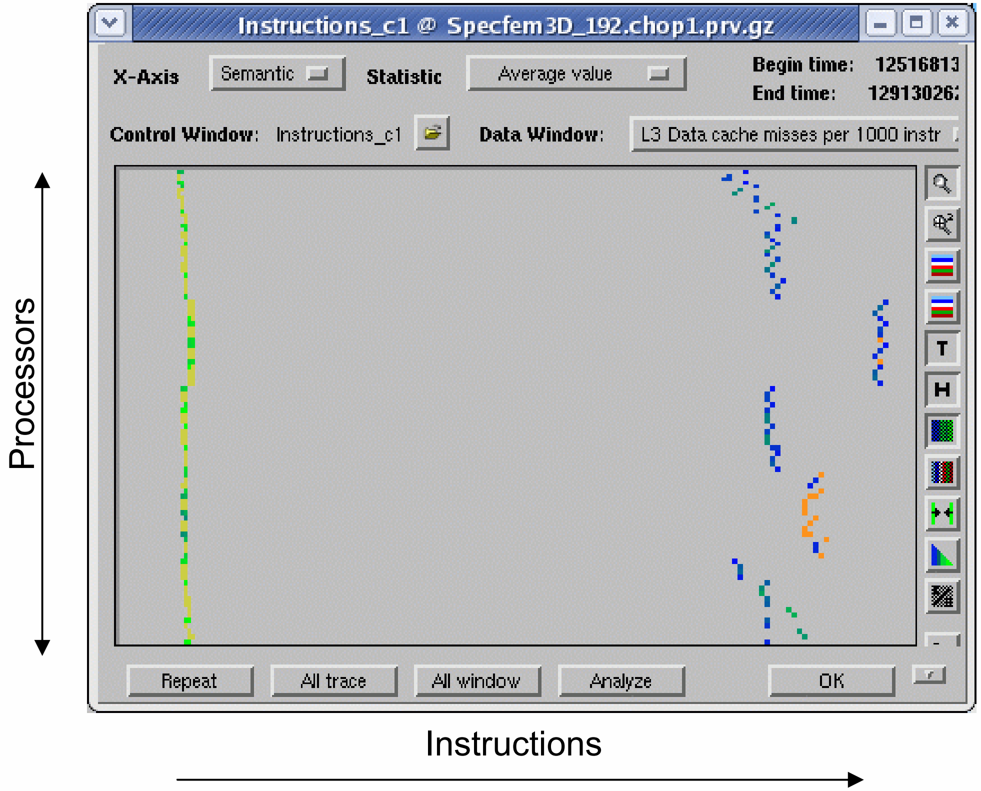


Tolerates very low bandwidths

Specfem3D

- Load Balance? Instructions and cache misses

@ 192 processors



Color: cache misses

Even more

many

Few

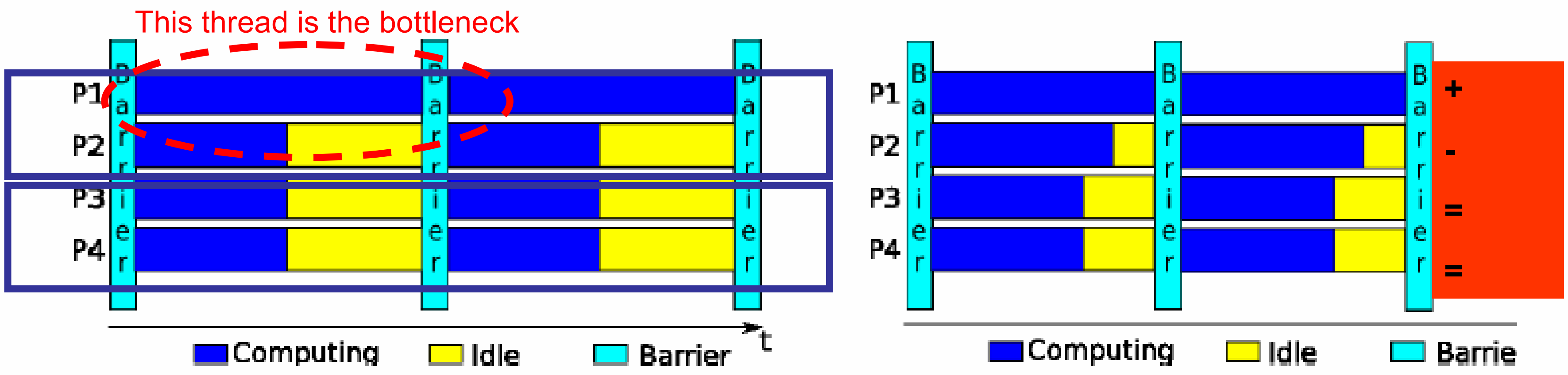
Sources of unbalancing

- Intrinsic to the algorithm:
  - Non-perfect partitioning
  - Some applications need dynamic load balancing: (e.g. molecular dynamics)
  - Several computational phases
  - Data-dependent access pattern

- Caused by resources:
  - Cache misses
  - Processor heterogeneity in a chip/board
  - OS noise/user daemons: in some computing nodes the OS or user daemons could delay the running process
  - Network topology and contention

SMT priorities and load balancing

- Increasing the priority of the threads executing longer[1]
- Assume a 4 process MPI application running on a POWER5
  - Further assume that P1 computes longer than P2, P3, P4
  - P1, P2 run on one core and P3, P4 in the other core
- Increasing throughput is not the solution to unbalance[1]
- By increasing P1's priority the application execution time decreases

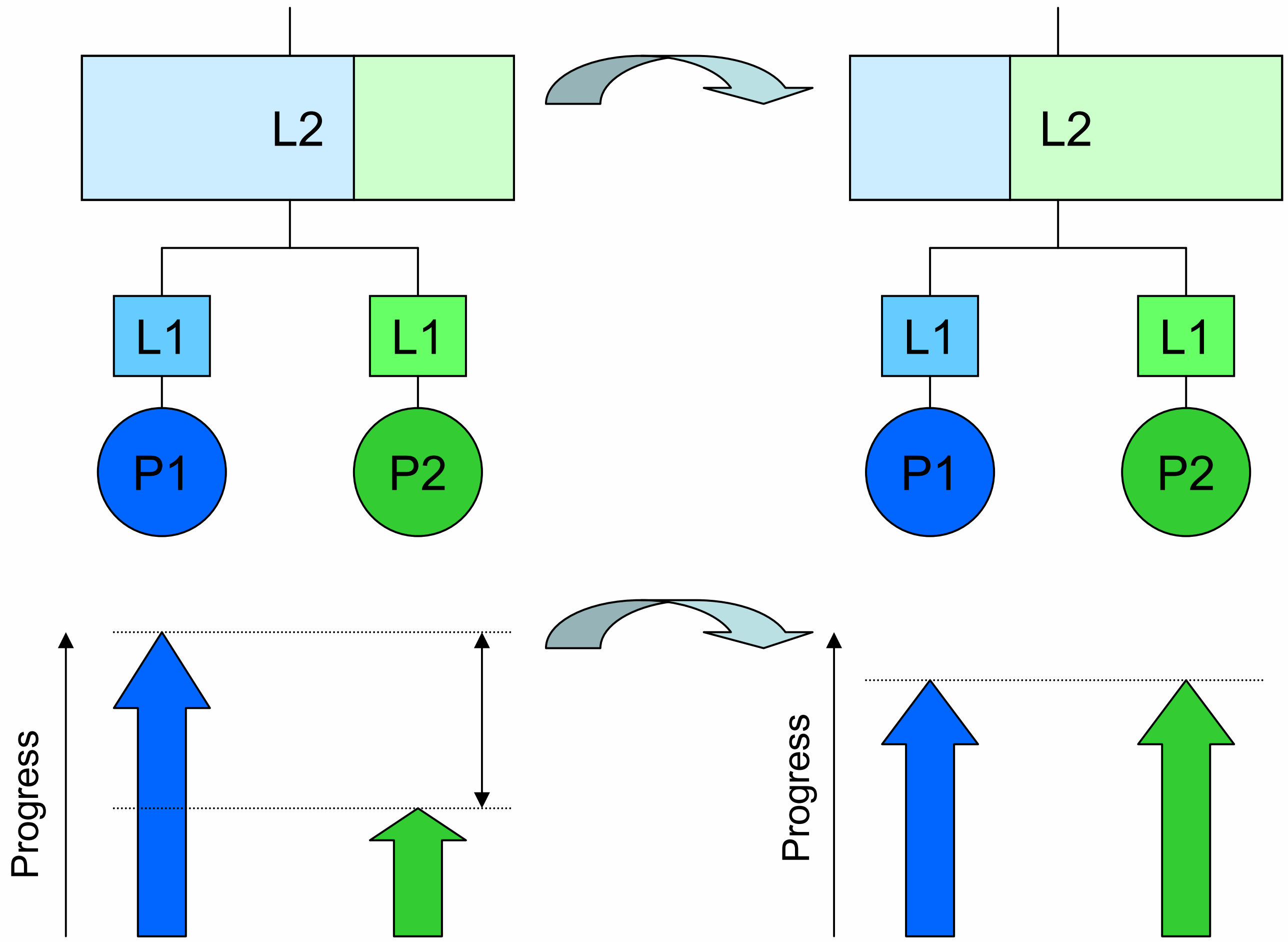


(a) Unbalanced HPC application

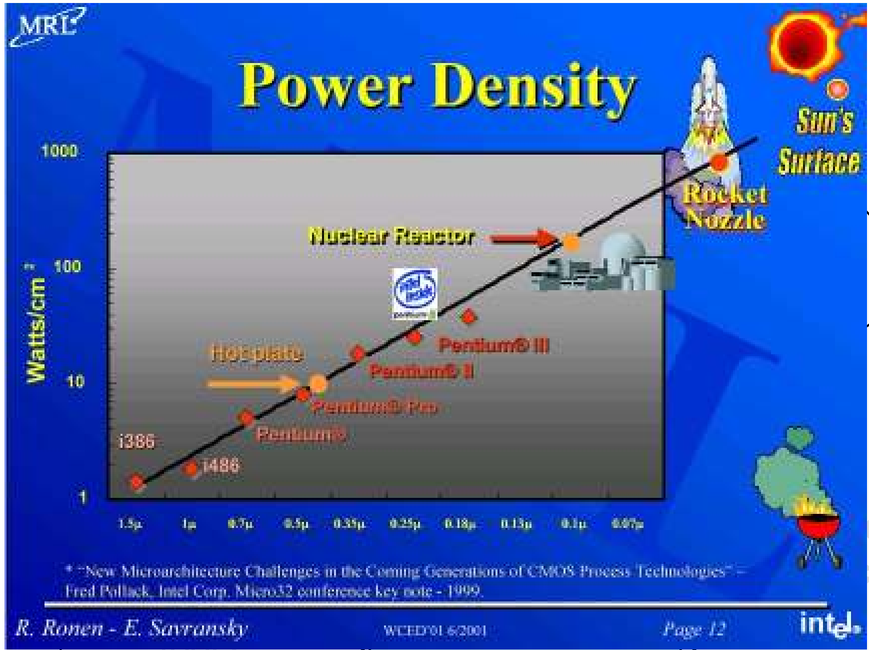
(b) More balanced HPC application

QoS through shared resource management

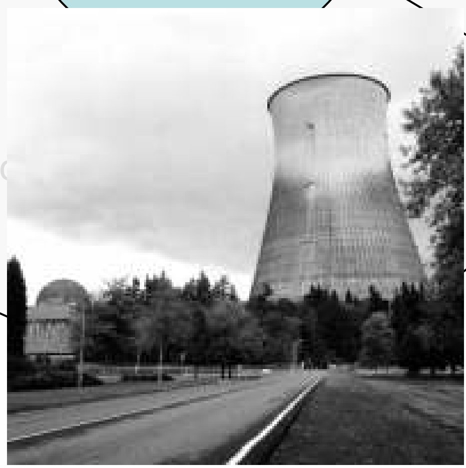
- Balance thread progress by managing the shared L2 cache

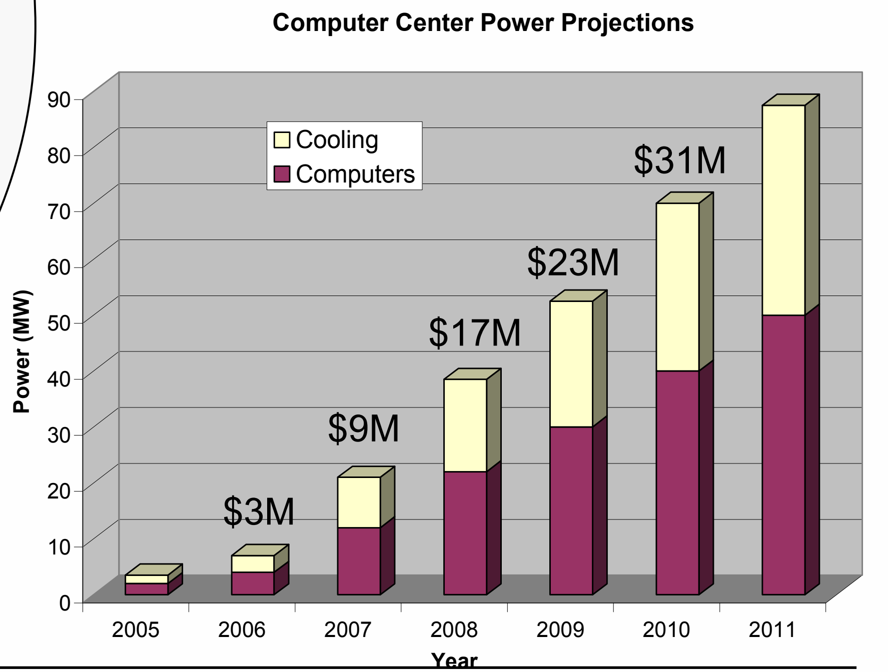


Multidisciplinary top-down approach



Google™





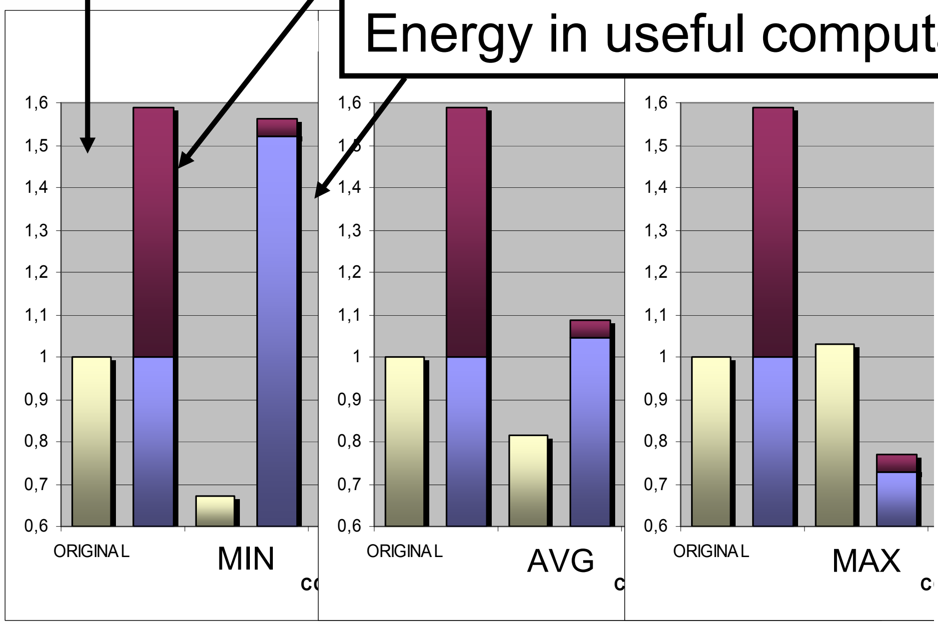
Load balance and power

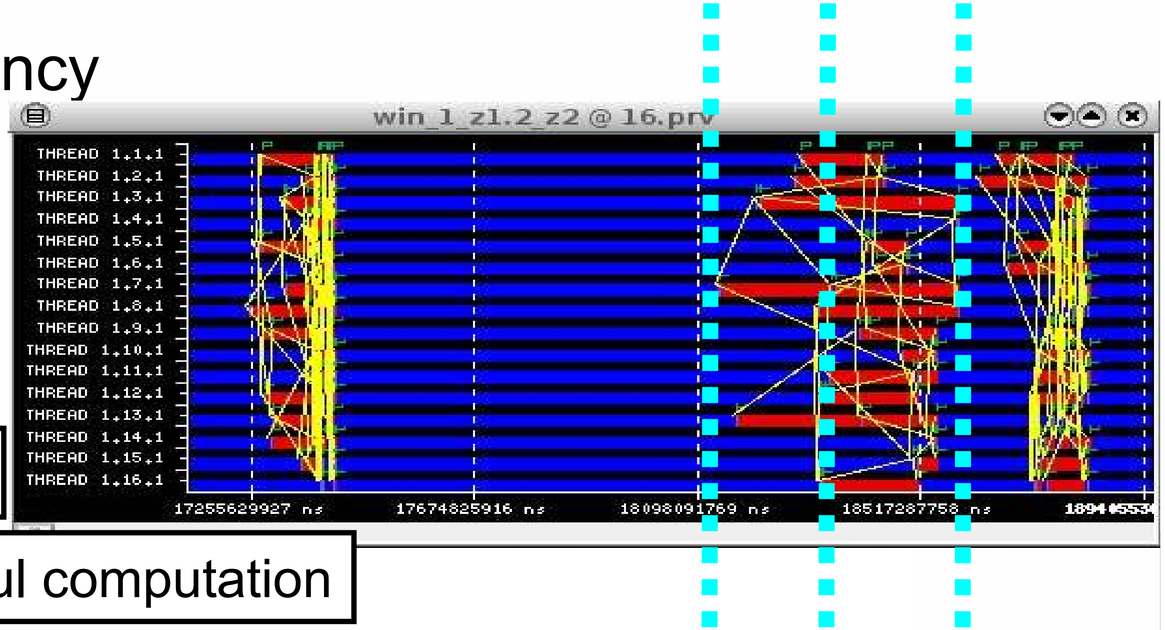
- Changing core frequency

Elapsed time

Energy in MPI

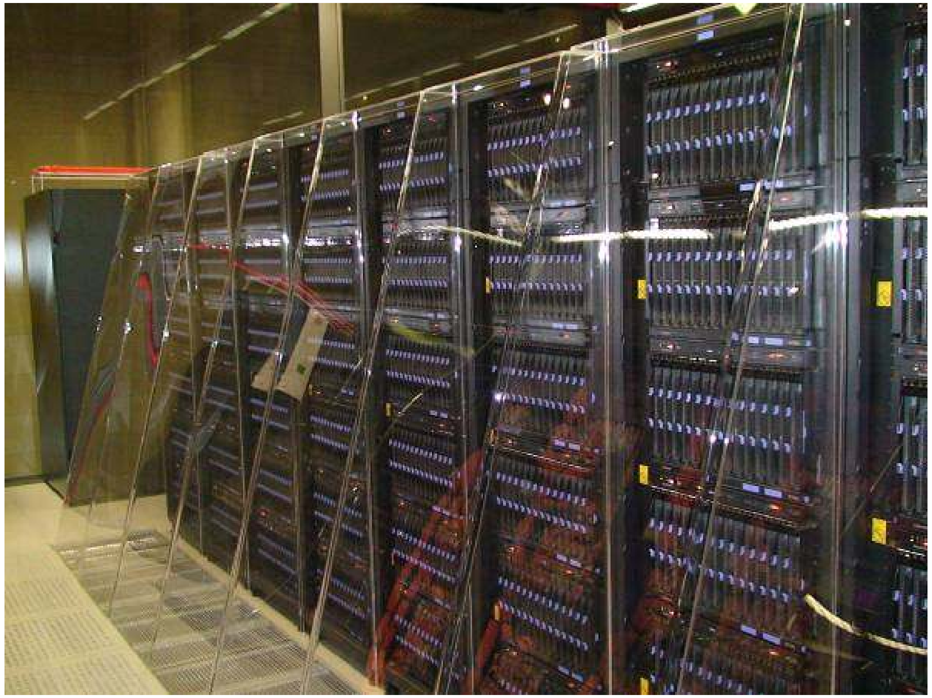
Energy in useful computation

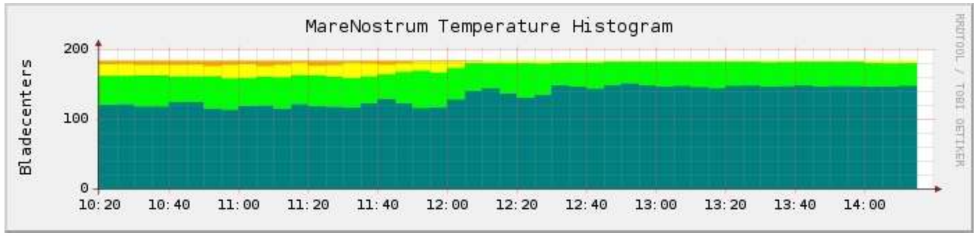


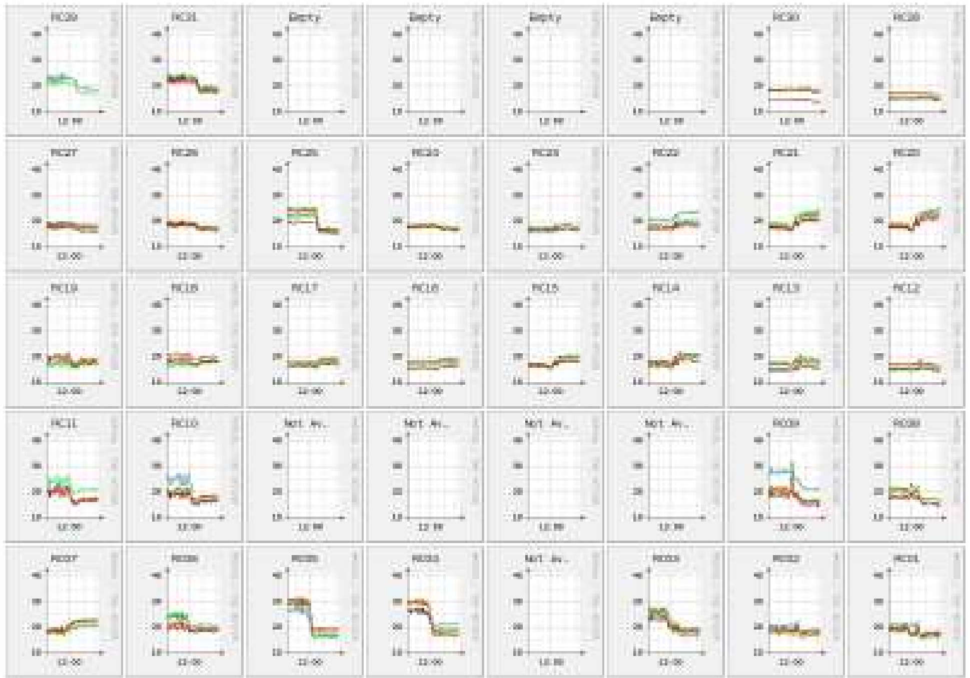


MIN AVG MAX

Power and money







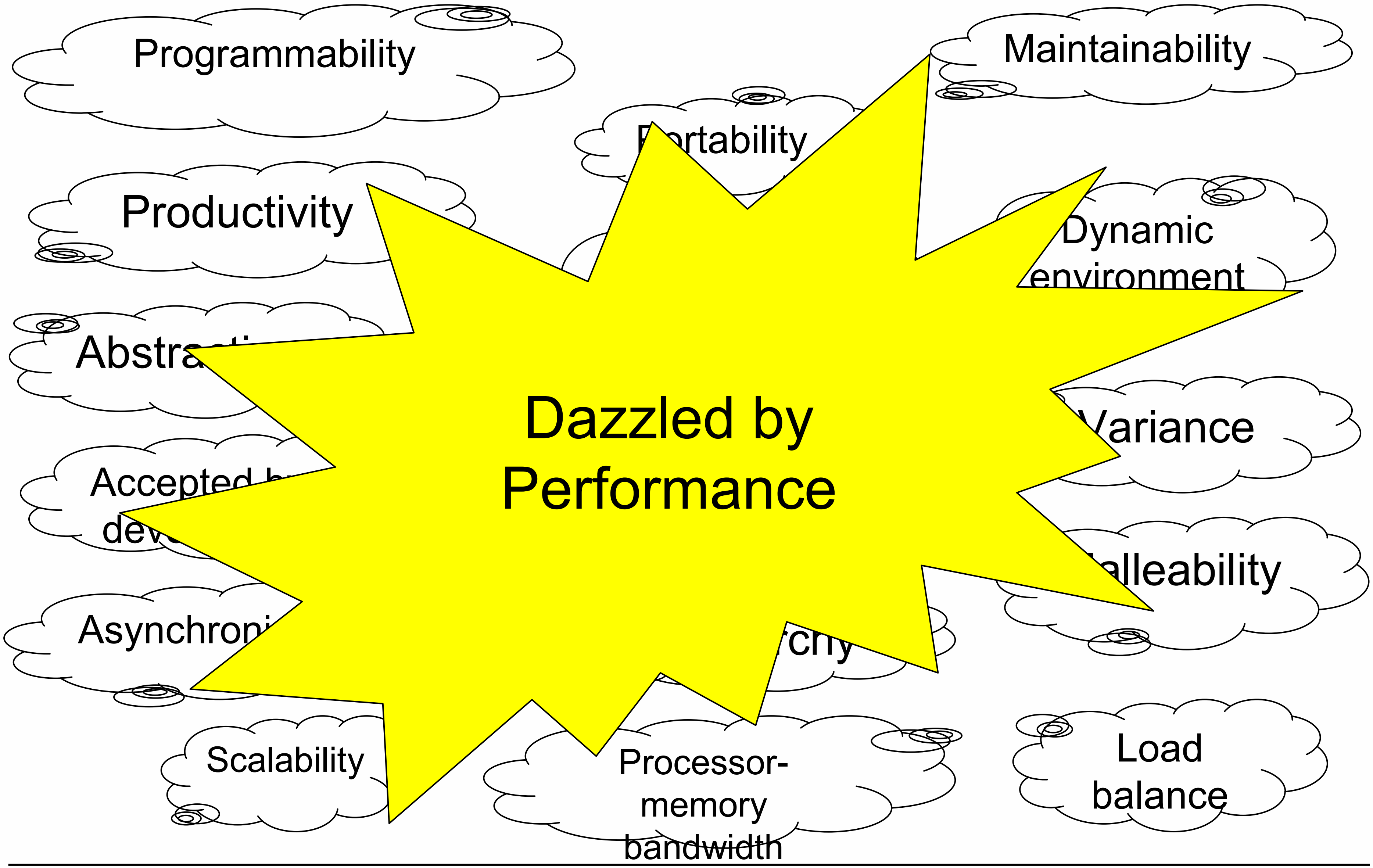
Talk outline

- Supercomputing from the past
  - Architecture evolution
  - Applications and algorithms
- Supercomputing for the future
  - Technology trends
  - Multidisciplinary top-down approach
- **Conclusions**

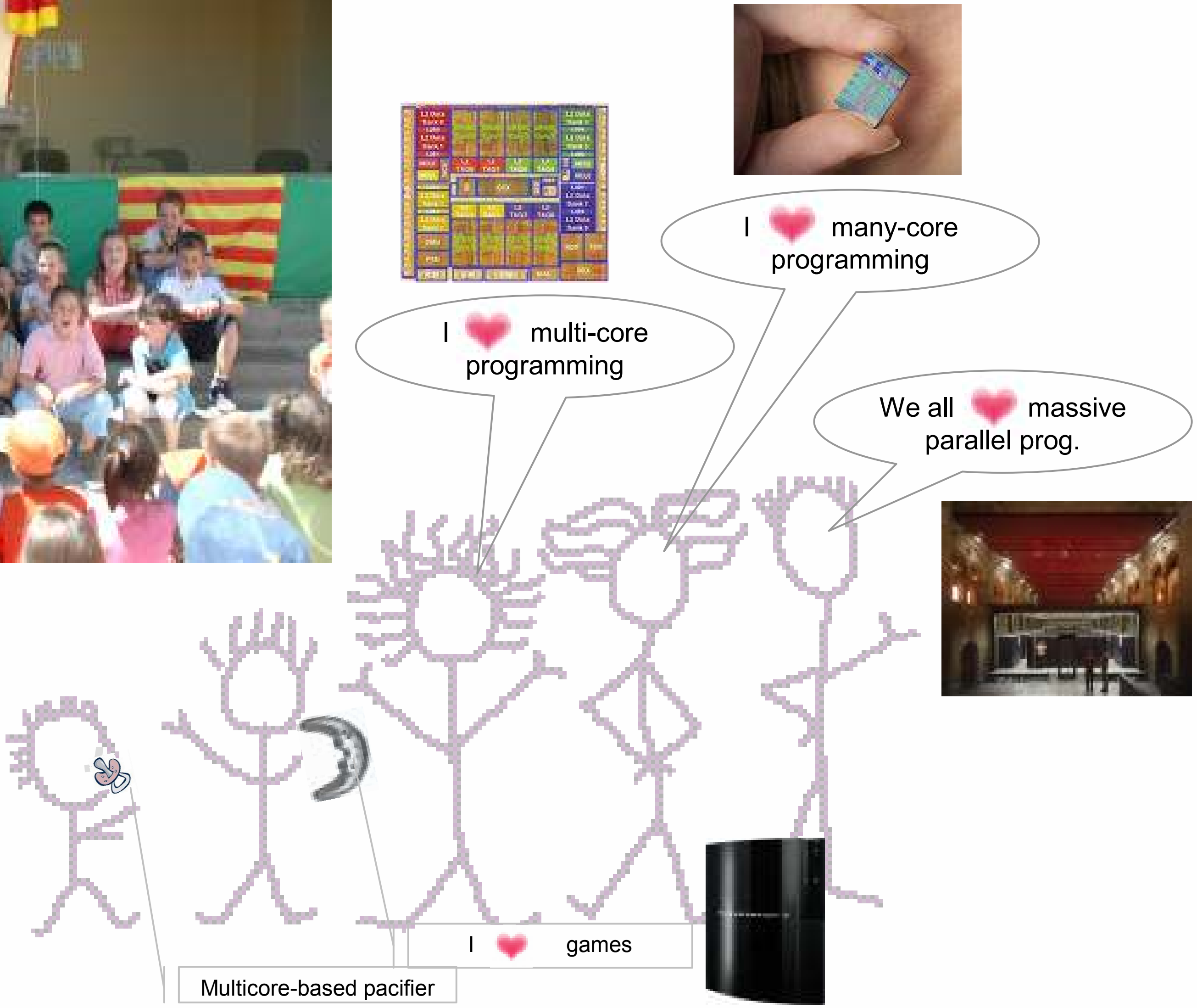
Key issues

- Programmability
- Maintainability
- Portability
- Productivity
- Dynamic environment
- Heterogeneous functionality and performance
- Abstraction
- Variance
- Accepted by developers
- % memory used
- Memory association
- Malleability
- Asynchronism
- Hierarchy
- Scalability
- Processor-memory bandwidth
- Load balance

Key issues


Programmability
Maintainability
ortability
Productivity
Dynamic
environment
Abstra
Dazzled by
Performance
Variance
Accepted b
dev
alleability
Asynchroni
rchy
Scalability
Processor-
memory
bandwidth
Load
balance

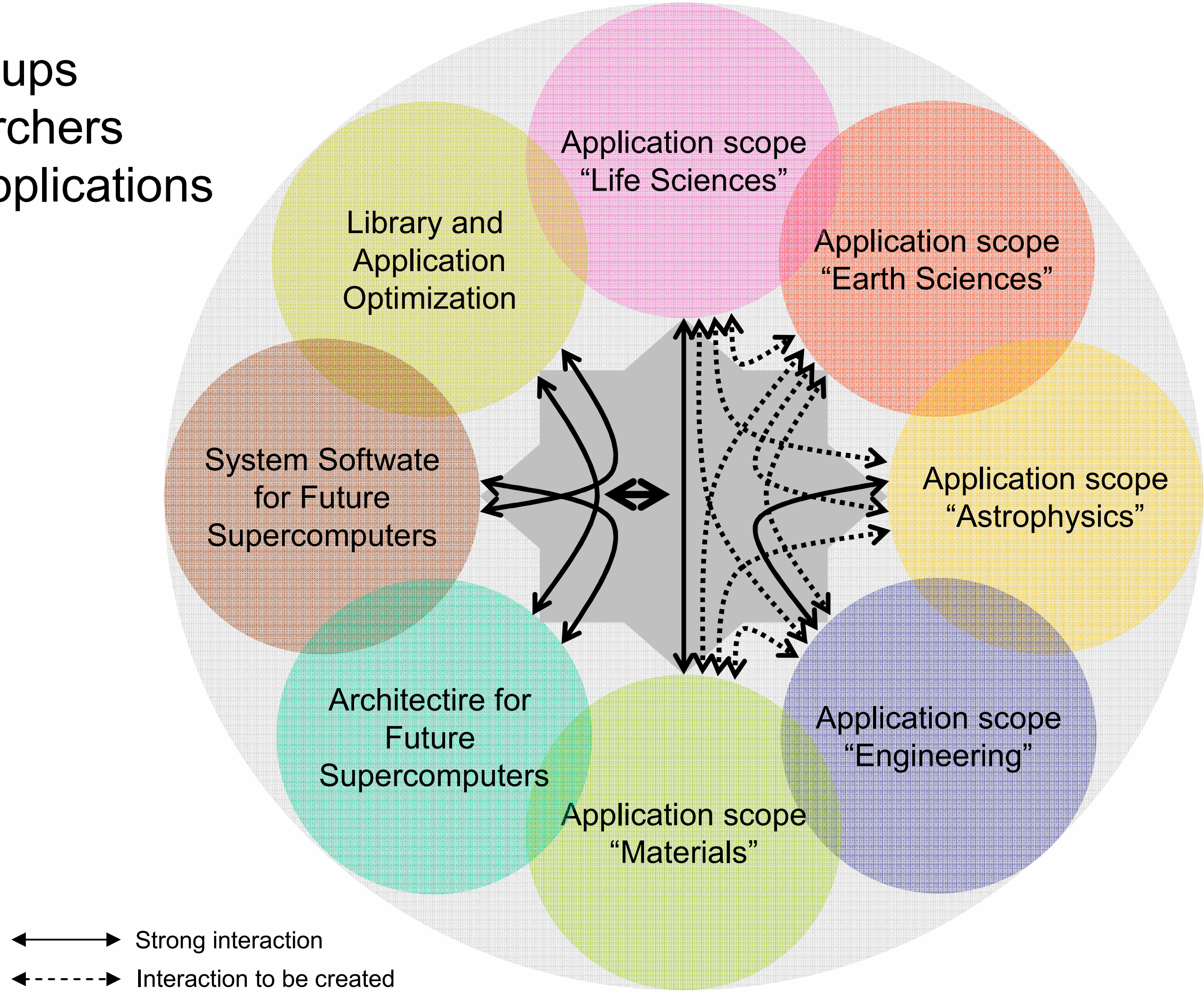
Education for Parallel Programming



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Supercomputing and e-Science Consolider program

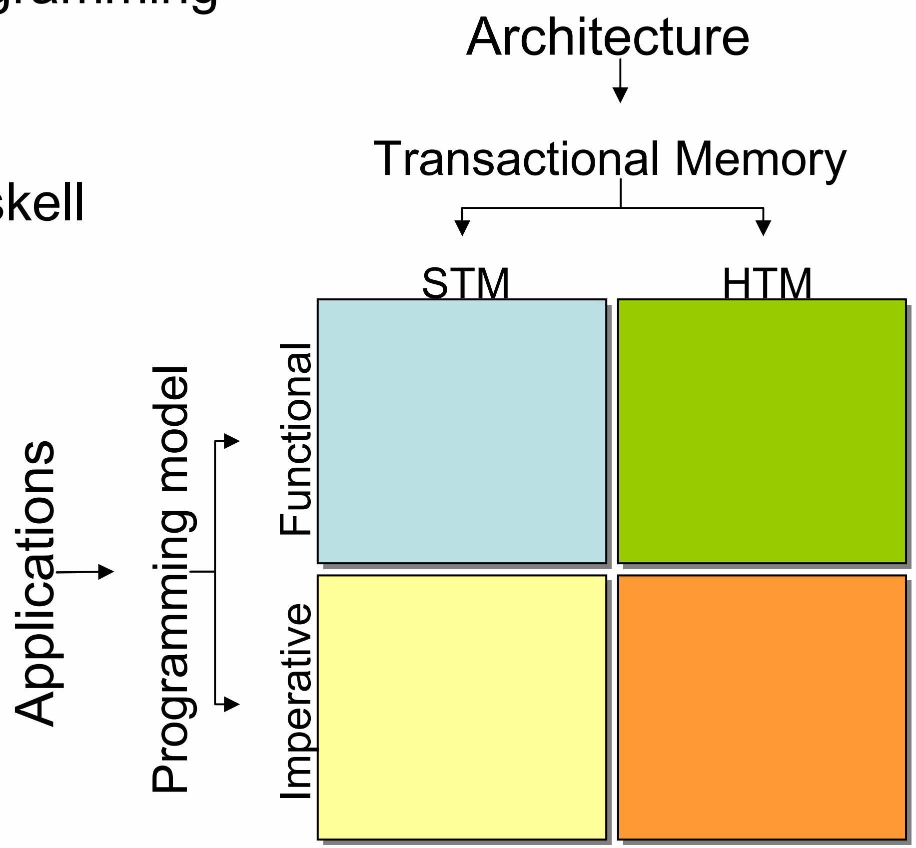
22 Spanish groups
119 senior researchers
5 Grand Challenge applications



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-Microsoft Many-core Project

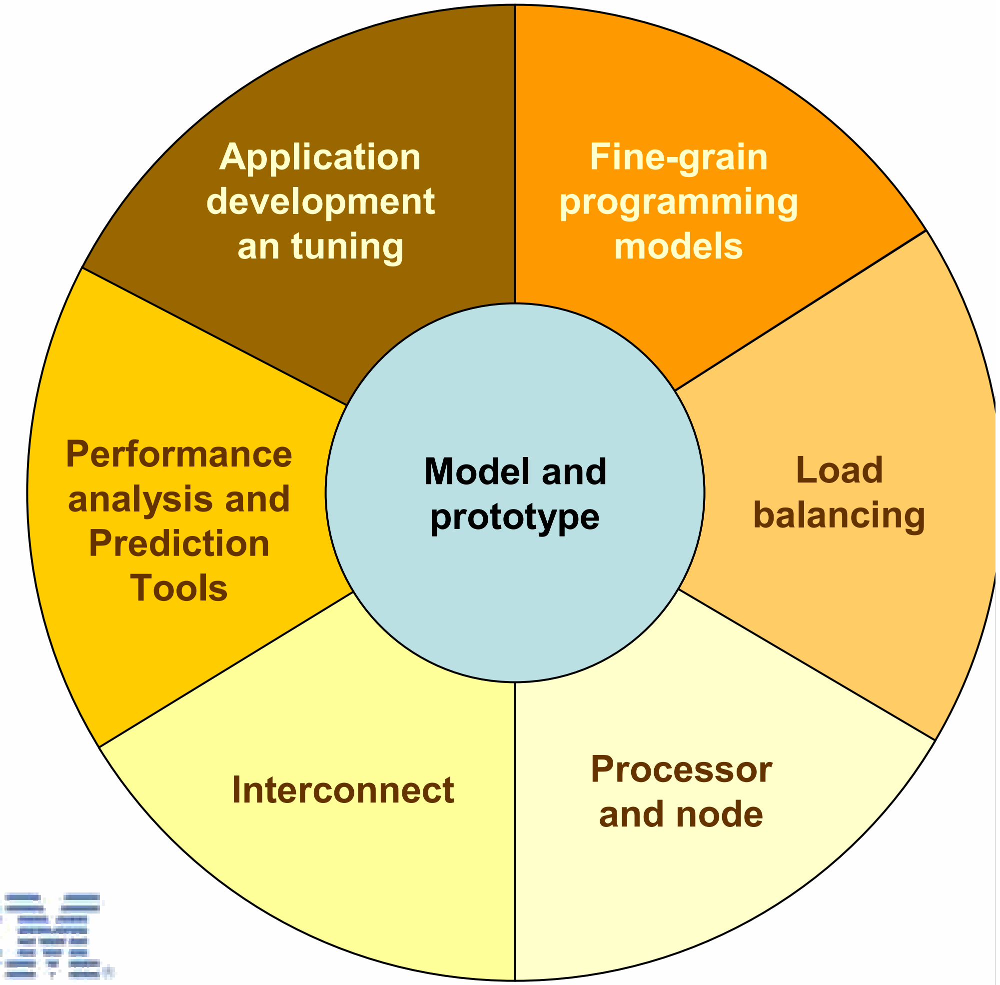
- Programming models for future many-core architectures
- Architectural support to programming models
  - OpenMP+TM
  - HW acceleration for Haskell
- Many-core architecture
- Power-aware



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

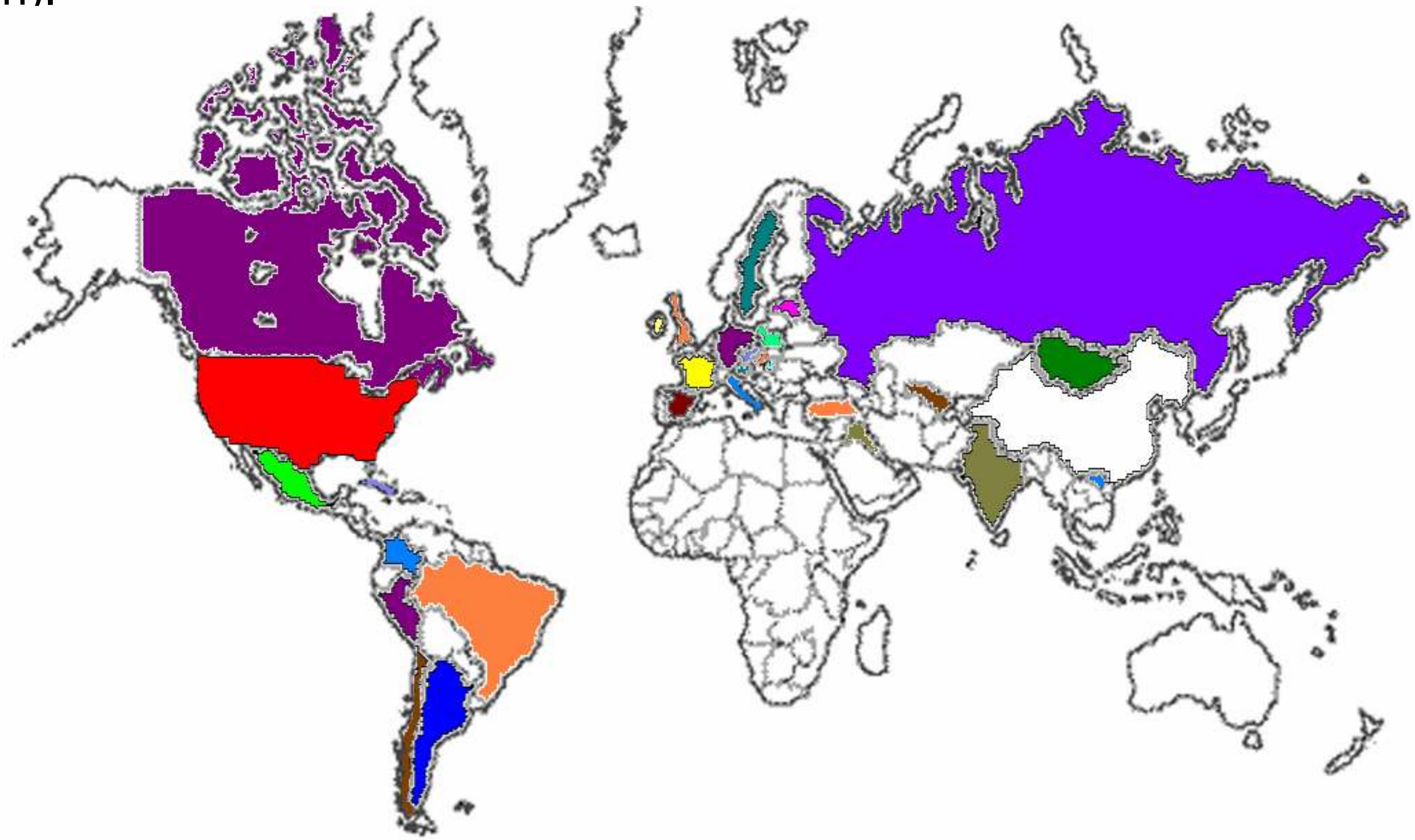
An overall picture of the IBM MareIncognito project

- Our 10-100 Petaflop research project for BSC (2010)
- Port/develop applications to reduce
  time-to-production once installed
- Programming models (MPI, OpenMP, CellSuperScalar)
- Tools for application development
  and to support previous evaluations
- Evaluate node architecture (heavily multicored):
- Evaluate interconnect options



Staff Evolution

BSC-CNS has 195 members at October of 2007 and hailed from 23 different countries (Alemania, Argentina, Belgium, Brazil, Bulgaria, Canada, Colombia, China, Cuba, France, Germany, India, Ireland, Italy, Jordania, Lebanon, Mexico, Poland, Russia, Serbia, Turkey, the United Kingdom, the United States and Spain).



Thank you !