Microsoft® Research

FacultySummit 2011

Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS

Microsoft® Research

FacultySummit 2011

Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS

Towards Exaflop Supercomputers

Prof. Mateo Valero
Director of BSC, Barcelona

National U. of Defense Technology (NUDT) Tianhe-1A

- **Hybrid architecture**:
  - Main node:
    - Two Intel Xeon X5670 2.93 GHz 6-core Westmere, 12 MB cache
    - One Nvidia Tesla M2050 448-ALU (16 SIMD units) 1150 MHz Fermi GPU:
    - 32 GB memory per node
  - 2048 Galaxy "FT-1000" 1 GHz 8-core processors
- Number of nodes and cores:
  - **7168 main nodes** * (2 sockets * 6 CPU cores + 14 SIMD units) = **186368 cores** (not including 16384 Galaxy cores)
- Peak performance (DP):
  - 7168 nodes * (11.72 GFLOPS per core * 6 CPU cores * 2 sockets + 36.8 GFLOPS per SIMD unit * 14 SIMD units per GPU) = **4701.61 TFLOPS**
- Linpack performance: 2.507 PF → **53% efficiency**
- Power consumption **4.04 MWatt**

Source http://blog.zorinaq.com/?e=36

Top10

November 2010 list

Rank	Site	Computer	Procs	Rmax	Rpeak
1	Tianjin, China	XeonX5670+NVIDIA	186368	2566000	4701000
2	Oak Ridge Nat. Lab.	Cray XT5,6 cores	224162	1759000	2331000
3	Shenzhen, China	XeonX5670+NVIDIA	120640	1271000	2984300
4	GSIC Center, Tokyo	XeonX5670+NVIDIA	73278	1192000	2287630
5	DOE/SC/LBNL/NERSC	Cray XE6 12 cores	153408	1054000	1288630
6	Commissariat a l'Energie Atomique (CEA)	Bull bullx super-node S6010/S6030	138368	1050000	1254550
7	DOE/NNSA/LANL	QS22/LS21 Cluster, PowerXCell 8i / Opteron Infiniband	122400	1042000	1375780
8	National Institute for Computational Sciences/University of Tennessee	Cray XT5-HE 6 cores	98928	831700	1028850
9	Forschungszentrum Juelich (FZJ)	Blue Gene/P Solution	294912	825500	825500
10	DOE/NNSA/LANL/SNL	Cray XE6 8-core	107152	816600	1028660

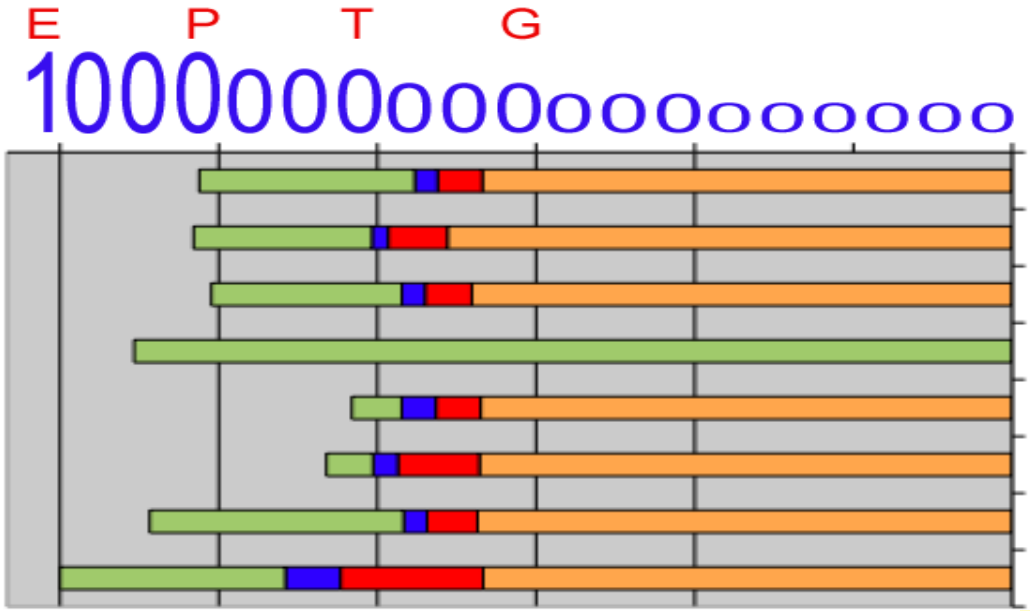
Looking at the Gordon Bell Prize

- 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; ?; 1×10^8 Processors?? (10^9 threads)

Jack Dongarra

Cartagena, Colombia, May 18-20

Evolution towards Exaflop supercomputers



Current #1 (6/2010)

Current #2 (Fermi) (6/2010)

Current #3 (Cell) (6/2010)

All in Top500 (6/2010)

Personal supercomputer (CPU)

Personal supercomputer (CPU/GPU)

Sequoia LANL (announced)

Possible exaflop ?

- 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

	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

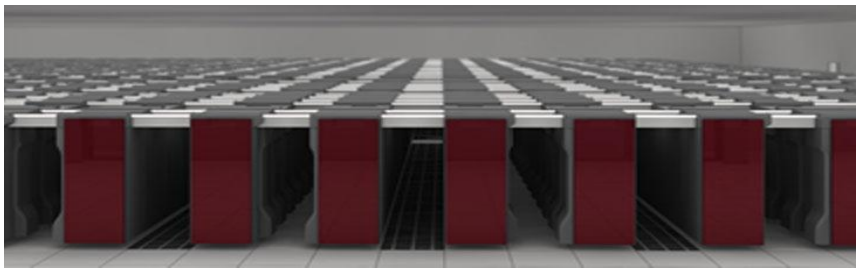
10+ Pflop/s systems planned

- IBM **Blue Waters at Illinois**
  - 40,000 8-core Power7, 1 PB memory, 18 PB disk, 500 PB archival storage, 10 Pflop/s, 2012, $200 million

- IBM Blue Gene/Q systems:
  - **Mira to DOE, Argonne National Lab** with 49,000 nodes, 16-core Power A2 processor (1.6-3 GHz), 750 K cores, 750 TB memory, 70 PB disk, 5D torus, 10 Pflop/s
  - **Sequoia to Lawrence Livermore National Lab** with 98304 nodes (96 racks), 16-core A2 processor, 1.6 M cores (1 GB/core), 1.6 Petabytes memory, 6 Mwatt, 3 Gflops/watt, 20 Pflop/s, 2012

10+ Pflop/s systems planned

- **Fujitsu Kei**
  - 80,000 8-core Sparc64 VIIIfx processors 2 GHz, (16 Gflops/core, 58 watts → 2.2 Gflops/watt), 16 GB/node, 1 PB memory, 6D mesh-torus, 10 Pflops



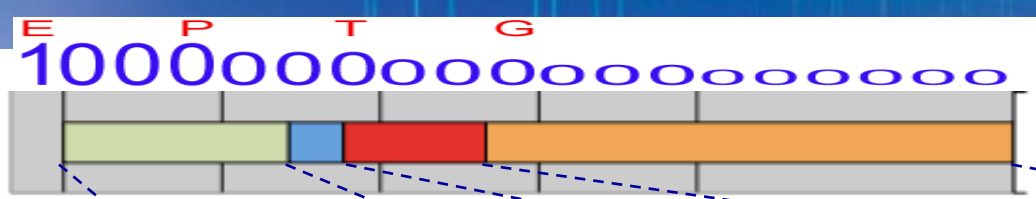
- **Cray's Titan at DOE, Oak Ridge National Laboratory**
  - Hybrid system with Nvidia GPUs, 1 Pflop/s in 2011, 20 Pflop/s in 2012, late 2011 prototype
  - $100 million

Systems Scaling Projections

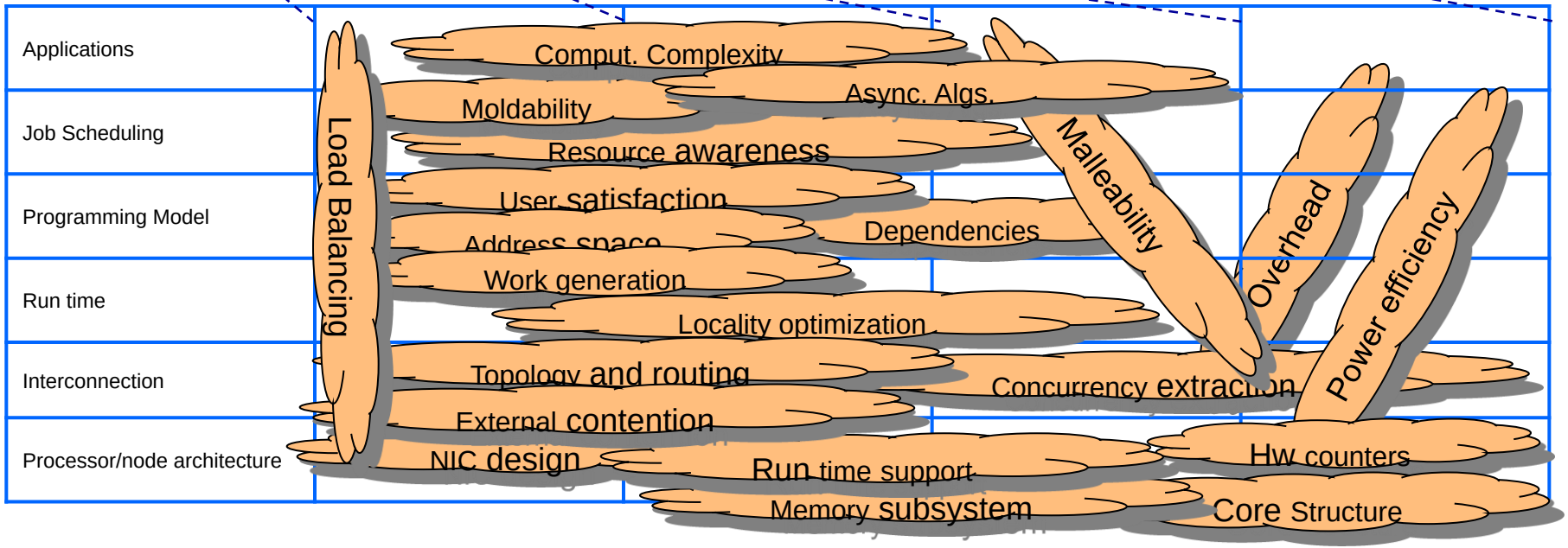
Begin Full System Delivery (Yr)	2004	2007	2012	2015	2019
Design Parameters	**BG/L**	**BG/P**	**25PF**	**300PF**	**1200PF**
Cores / Node	2	4	8-24	32-64-128	96-128-500
Clock Speed (GHz)	0.7	0.85	1.6-4.1	2.3-4.8	2.8-6.0
Flops / Clock / Core	4	4	8-32	8-32	16-64
Nodes / Rack	1024	1024	100-1024	256-1024	256-1024
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**

IESP team

Holistic approach ...



Towards exaflop



BSC-CNS: International Initiatives (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

Cartagena, Colombia, May 18-20

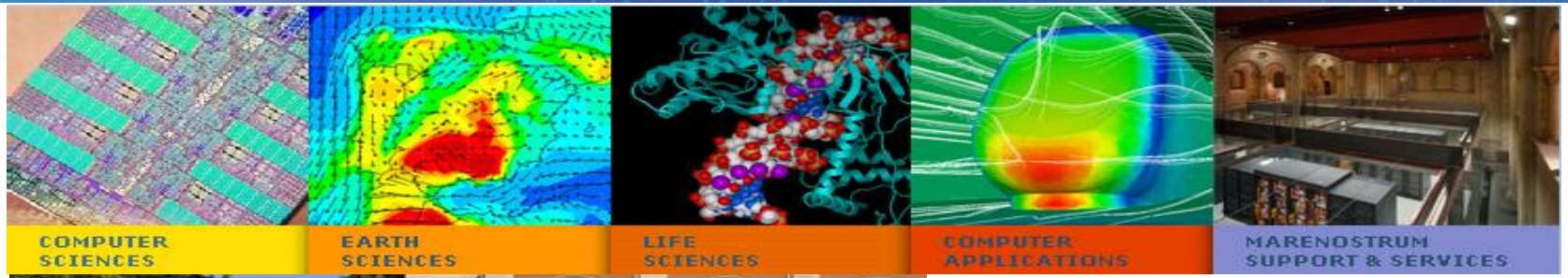
BSC-CNS Introduction

- 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.

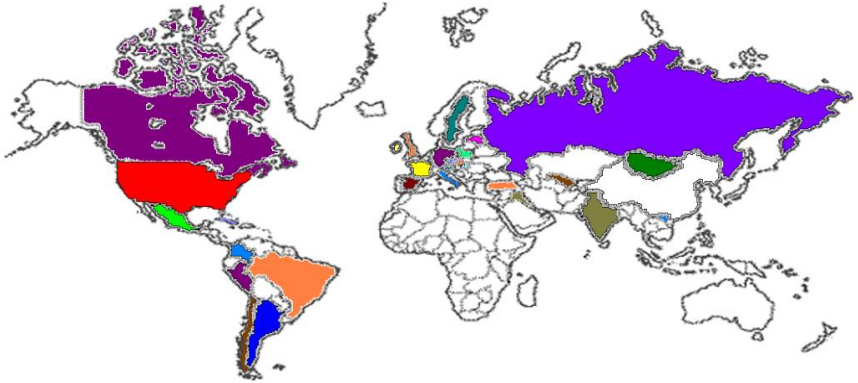
- 5 Scientific/ Technical Departments

BSC: Spanish National Center



Cartagena, Colombia, May 18-20

More than 325 people from 27 different countries (Argentina, Belgium, Brazil, Bulgaria, Canada, Colombia, Cuba, China, Cuba, Dominicana, France, Germany, India, Iran, Ireland, Italy, Jordania, Lebanon, Mexico, Pakistan, Poland, Russia, Serbia, Spain, Turkey, UK,



BSCMSR Centre: general overview

- Microsoft-BSC joint project
  - Combining research expertise
    - BSC Computer Architectures for Parallel Paradigms: computer architecture
    - Microsoft Research Cambridge: programming systems
  - Kickoff meeting in April 2006
  - BSC Total Effort:
    - Very young team! 2 Senior BSC researchers, 15 PhD + 5 MS students
- BSCMSRC inaugurated January 2008
- Open model
  - No patents, public IP, papers and open source main focus
  - Similar research agenda with parallel computing centers in US
    - Berkeley, Illinois, Rice, Stanford

Sample Research Topic: Transactional Memory

- Major focus on developing TM applications, tools, scalable HTM implementations

- Applications:
  - RMSTM, Atomic Quake, QuakeTM, HaskellSTMbench, Wormbench
  - C# versions of STAMP TM applications (using Bartok compiler)
  - Released publicly through www.bscmsrc.eu
  - Published in ICPE11, ICS09, ACM CF, PPoPP09
  - Lessons learned: TM not *yet* easier to program, lacks tools
  - RMSTM best paper award in ICPE11 from 110 submissions
- Tools:
  - TM debugger and bottleneck analysis best paper award in PACT from 240 submissions
- HTM Implementations:
  - EazyHTM: Eager conflict detection, lazy resolution -> fast commit and aborts. Published in Micro-42
  - D1 Data cache for TM: Best paper award in GLSVLSI 2011 Conference
  - Filtering: Eliminate TM overheads by filtering thread-local data out of the read/write sets. Best paper award in HPCC09
- FPGA Emulator
  - Using BEE3 board: 4 Xilinx Virtex5-155T FPGAs, MIPS compatible cores
  - Added TLB, MMU support, cache coherence, multiprocessing facilities, implemented double ring interconnect
  - HTM implementation 16 cores per FPGA, in FCCM11

Sample Research Topic: StarSs and Barrelfish

- StarSs: BSC developed task-based programming model
  - Runtime dynamically detecting inter-task dependencies
  - Provides dataflow execution model

- Barrelfish: Microsoft/ETH developed operating system
  - Message passing based on low-level
  - Can run on shared or distributed memory
  - Designed for heterogeneous systems

- StarSs on Barrelfish
  - Leverages and combines the most attractive aspects of both: heterogeneity, message-passing, dataflow

- Will be developed on the Intel 48-core Cloud Computing Chip (SCC)
  - The SCC is also message-passing based

RISC @ FP7

The EU, Latin America and the Caribbean region must enhance their cooperation to face common challenges ahead, from climate change to tacking full advantage of globalization and economic growth for the benefit of a majority of our citizens.

We are determined to support the efforts of our partners in fighting poverty and strengthening democracy and social cohesion.

President Jose Manuel Durão Barroso

RISC@F7 Partnership

- Barcelona Supercomputer Center
- Universidad de Buenos Aires
- Universidad de Chile
- Universidad Politécnica de Madrid
- Universidade Federal do Rio de Janeiro
- Universidade de Coimbra
- CINECA
- Universidad Veracruzana
- Universidad Autónoma de Manizales
- Menon

RISC@FP7 Mission is to:

I. Survey and assess in detail the potential for HPC R&D co-operation between the EU and Latin America in respect to the challenges above

II. Propose collaborative structures that would substantially increase the number of HPC and ICT R&D collaboration between EU and Latin America R&D organisations and key industries.

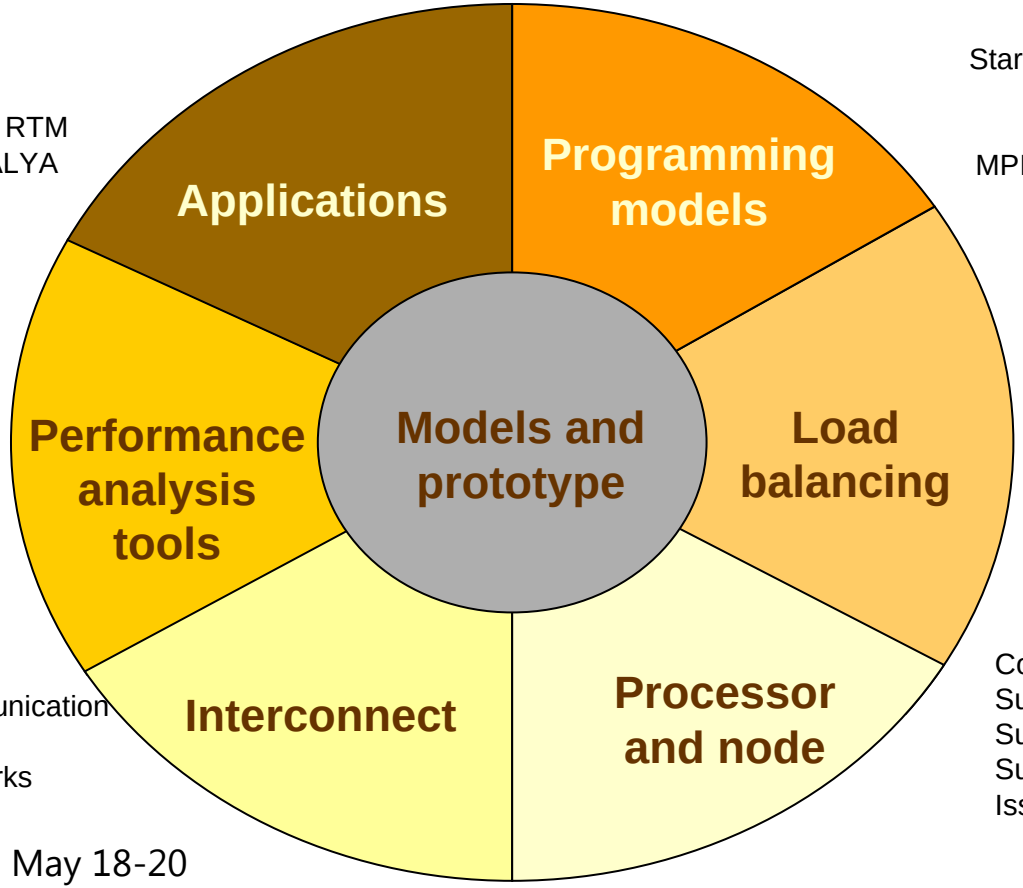
III. Identify major research areas and research clusters which are major drivers of the EU – LA research collaboration.

IV. Facilitate HPC and ICT R&D policy dialogues between the EU Latin America in respect to the above global challenges.

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

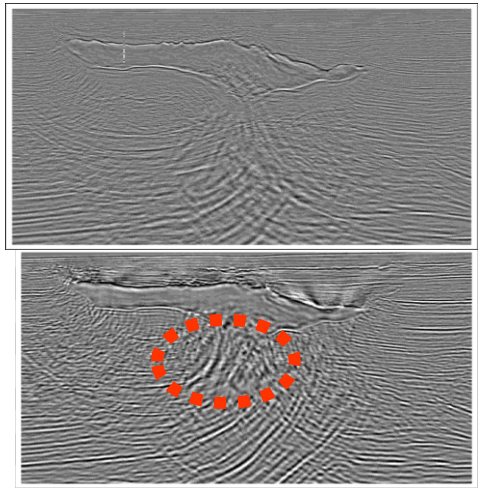
Coordinated scheduling:
Run time,
Process,
Job
Power efficiency

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

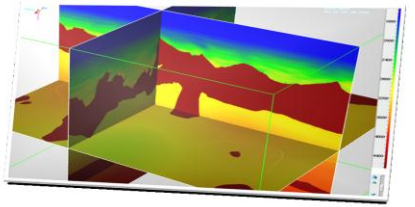
Contribution to new Cell design
Support for programming model
Support for load balancing
Support for performance tools
Issues for future processors

Kaleidoscope Project

Platform	Gflops	Seep-up	Power (W)	Gflops/W
JS21	8,3	1	267	0,03
QS22	116,6	14	370	0,32
2 TESLA 1060	350	42	90+368,8	0,76



- The work of 3 months is now done in
  - 1 week (speed-up 14)
  - 2 days (speed-up 42)

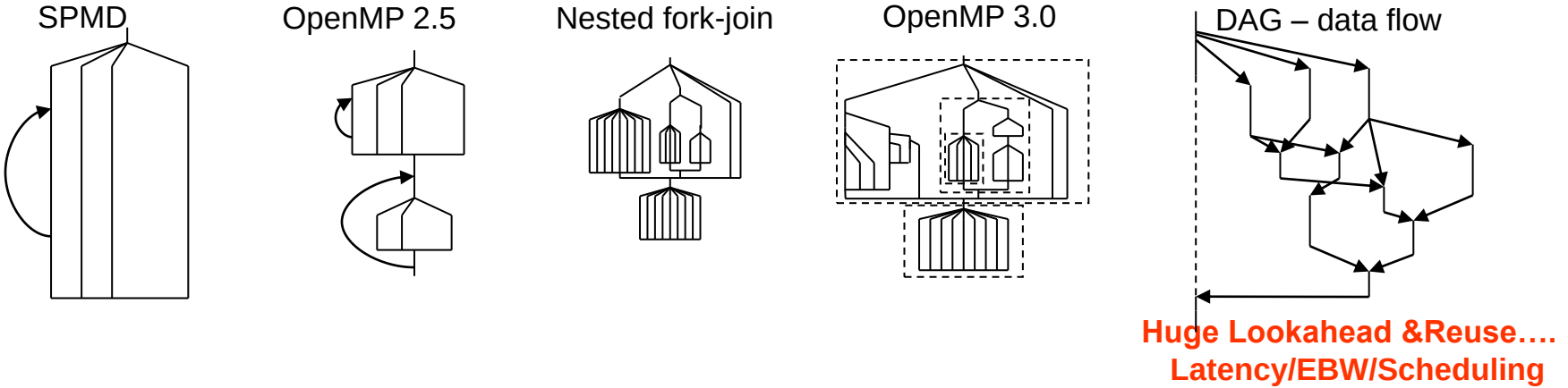


- On the Cell 23.5 GB/s of memory BW used from 25.6 GB/s max BW
- On TESLA the I/O is now the real bottleneck

- Awarded by "IEEE Spectrum" as one of the 2008 top 5 innovative technologies
- Platt's award to the commercial technology of the year 2009

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 …

Towards exaflop

Applications	
Performance Tools	
Programming Model	
Load Balancing	
Interconnection	
Processor/node architecture	

Latency

Bandwidth

Parallelism

Scheduling

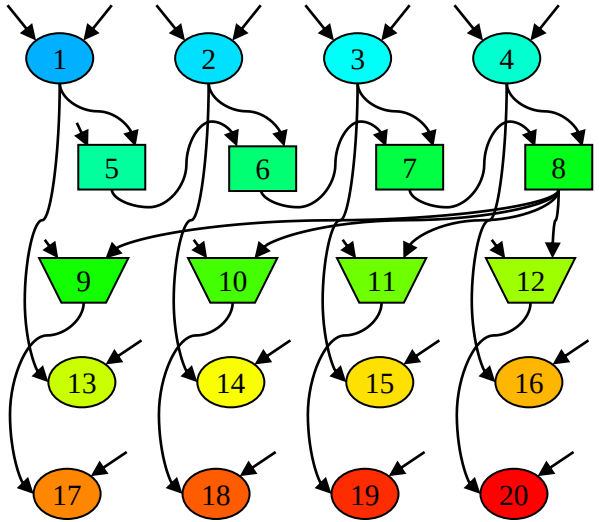
Power

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

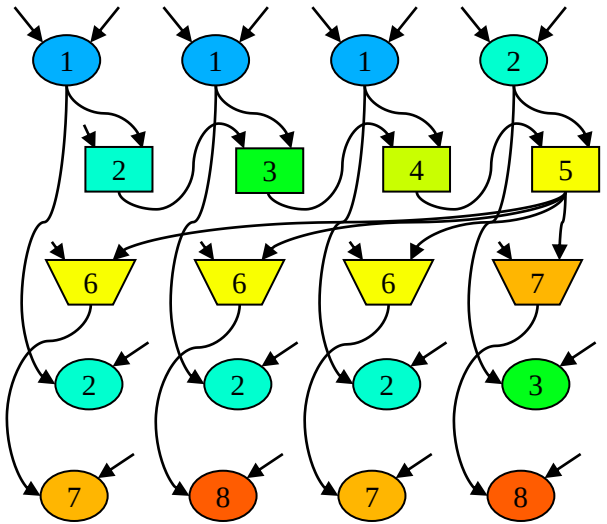


StarSs: … and executes as efficient as possible …

```
#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]);
```

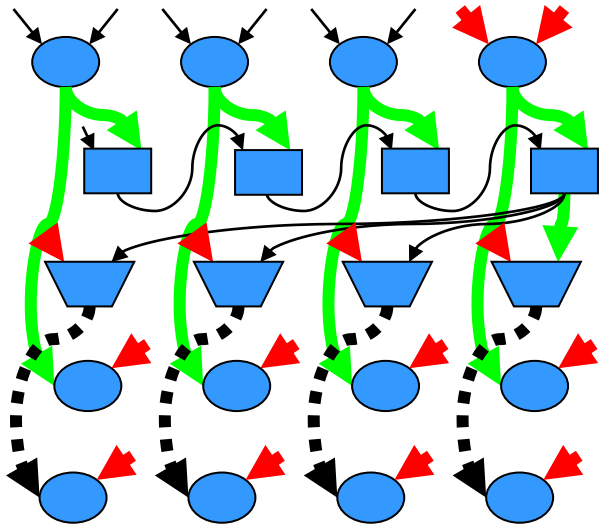
Task Graph Execution



StarSs: … benefiting from data access information

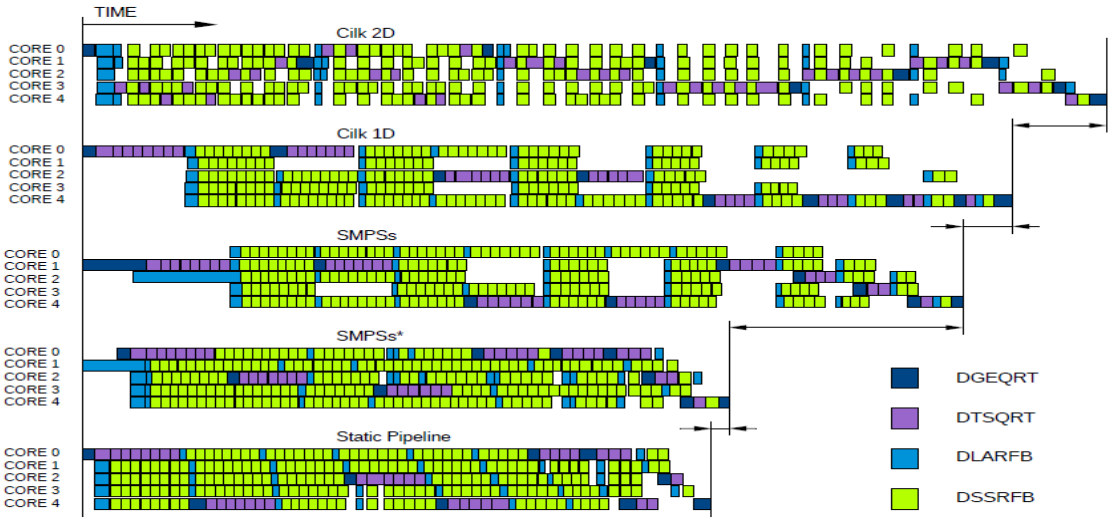
- **Flat global address space seen by programmer**
- Flexibility to dynamically traverse dataflow graph "optimizing"
  - Concurrency. Critical path
  - Memory access

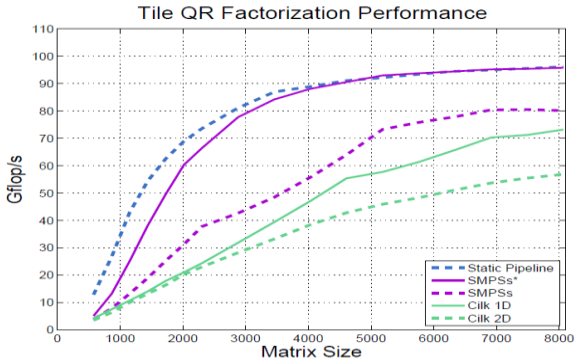
- Opportunities for
  - Prefetch
  - Reuse
  - Eliminate antidependences (rename)
  - Replication management



SMPSs: e.g. QR factorization

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

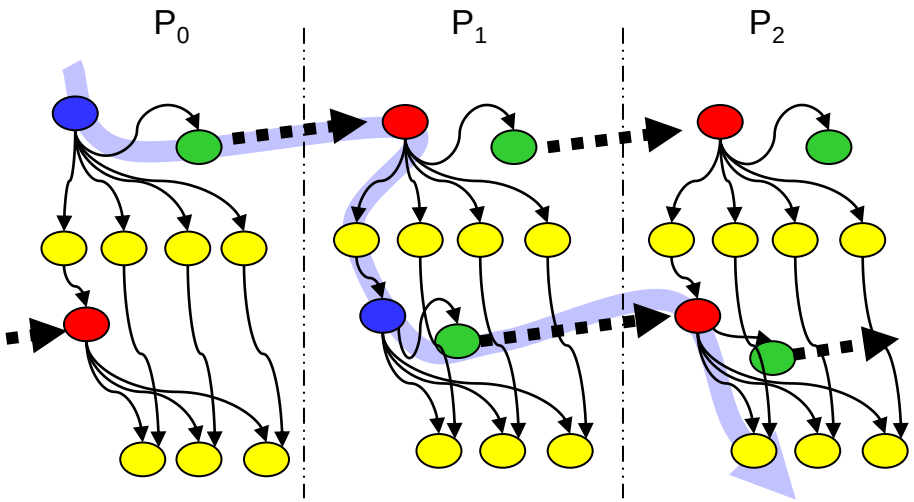




* Kurzak et al, “Scheduling Linear Algebra Operations on Multicore Processors”, LAPACK Working Note 213

Hybrid MPI/SMPSs: Linpack example

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

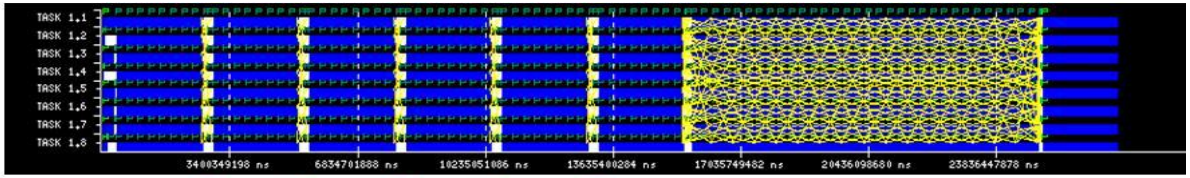



```
…
for (k=0; k<N; k++) {
   if (mine) {
      Factor_panel(A[k]);
      send (A[k])
   } else {
      receive (A[k]);
      if (necessary) resend (A[k]);
   }
   for (j=k+1; j<N; j++)
      update (A[k], A[j]);
```

```
#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);
```

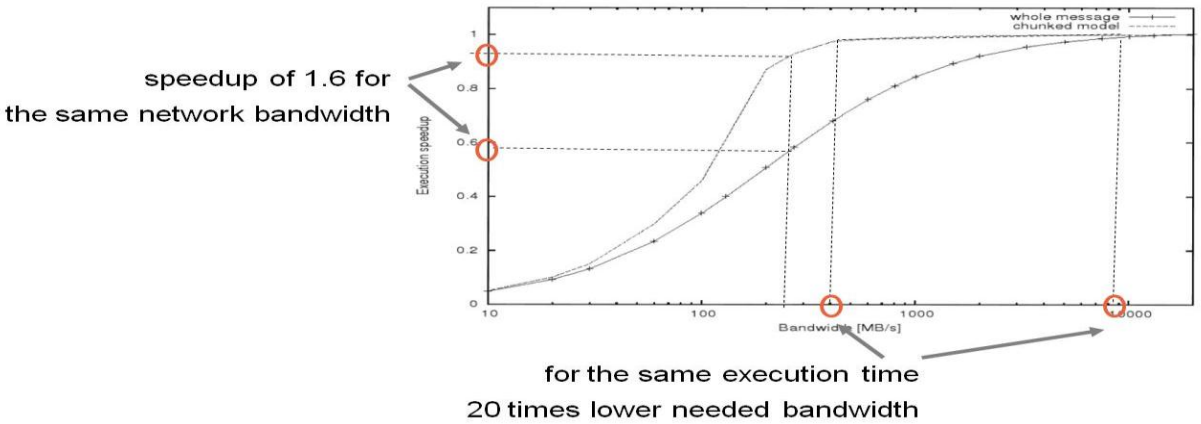
```
#pragma css task input(A[SIZE])
void send(float *A);
#pragma css task output(A[SIZE])
void receive(float *A);
#pragma css task input(A[SIZE])
void resend(float *A);
```

Effects on bandwidth



flattening
communication pattern

thus



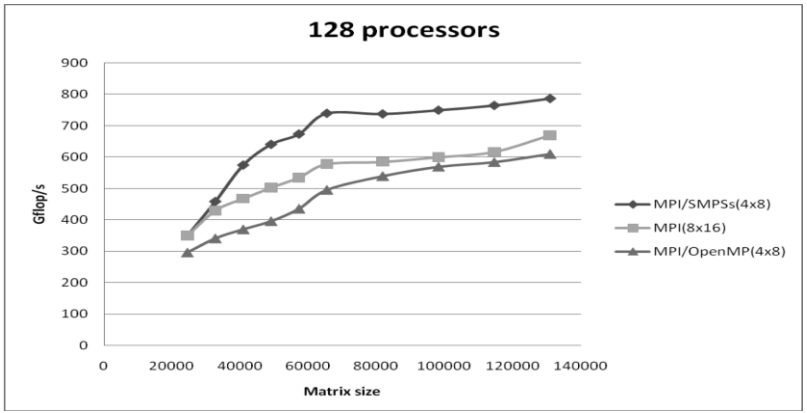
reducing
bandwidth requirements

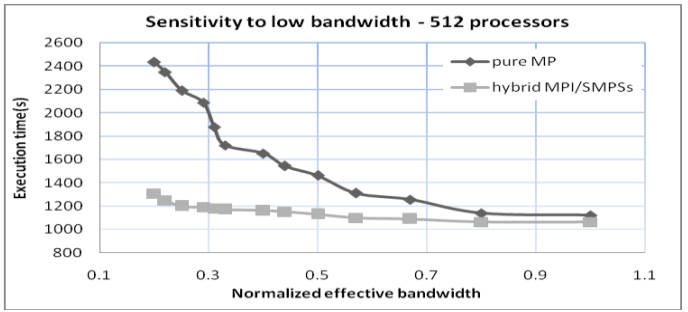
*simulation on application with
ring communication pattern

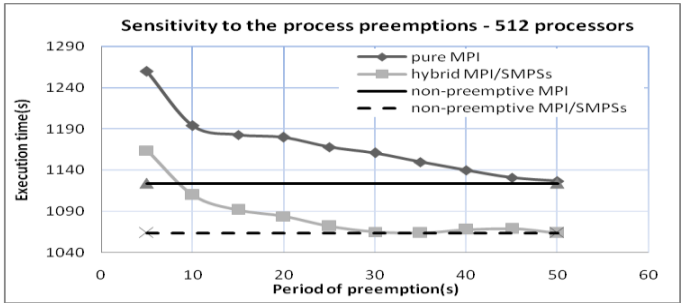
V. Subotic et al. "Overlapping communication and computation by enforcing speculative data-flow", January 2008, HiPEAC

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







A "unified" model

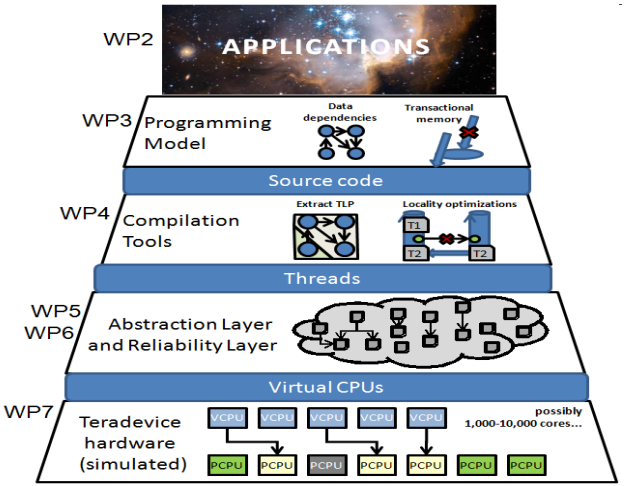
- StarSs
  - A "node" level programming model
  - C/Fortran + directives
  - Nicely integrates in hybrid MPI/StarSs
  - Natural support for heterogeneity
- Programmability
  - Incremental parallelization/restructure
  - Abstract/separate algorithmic issues from resources
  - Disciplined programming
- Portability
  - "Same" source code runs on "any" machine
    - Optimized task implementations will result in better performance.
  - "Single source" for maintained version of a application
- Performance
  - Asynchronous (data-flow) execution and locality awareness
  - Intelligent Runtime: specific for each type of target platform.
    - Automatically extracts and exploits parallelism
    - Matches computations to resources

Open Source **http://www.bsc.es/smpsuperscalar**
http://nanos.ac.upc.edu/

Related FP7 projects

TERAFLUX

- Dataflow at all levels
- Data flow and transactional memory



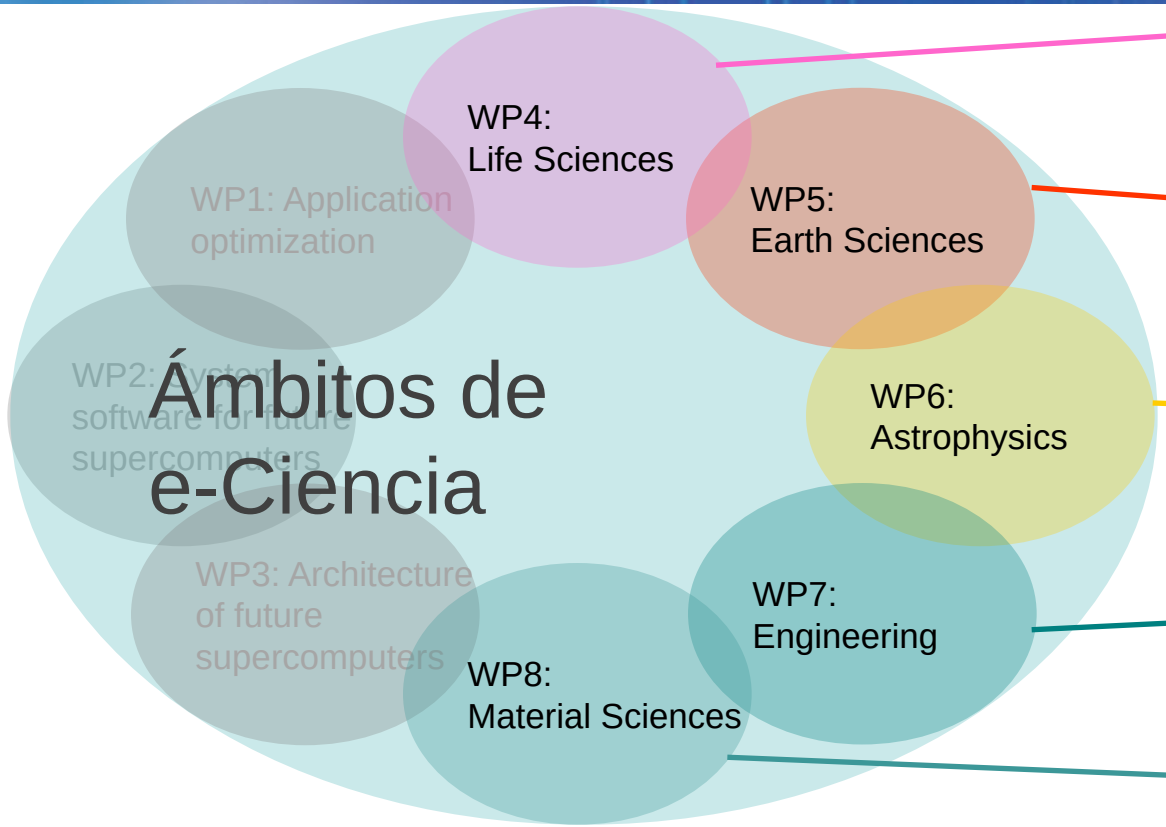
encore

- Chip and memory architecture for 256+ cores
- Focused on task based programming mdels

The TEXT project

- Towards EXaflop applicaTions
- Demonstrate that **Hybrid MPI/SMPSs** addresses the Exascale challenges in a an productive and efficient way.
  - Deploy at supercomputing centers: Julich, EPCC, HLRS, BSC
  - Port Applications (HLA, SPECFEM3D, PEPC, PSC, BEST, CPMD, LS1 MarDyn) and develop algorithms.
  - Develop additional environment capabilities
    - tools (debug, performance)
    - improvements in runtime systems (load balance and GPUSs)
  - Support other users
    - Identify users of TEXT applications
    - Identify and support interested application developers
  - Contribute to Standards (OpenMP ARB, PERI-XML)

Retos científicos y aplicaciones



- Estructuras macromoleculares
  - NAMD, GROMACS, DISCRETE, FTDOCK
- Splicing alternativo y epigenética
  - T-coffee, FLUX, GEM
- Claves genéticas de la leucemia
  - BLAS, HMMER

- WRF, ALYA ,HERMES, EMP, CHIMERE, CMAQ
- Modelo acoplado atmosférico-polvo mineral
- Escalabilidad modelo circulación global
  - WACCM, NEMO, IFS, OASIS

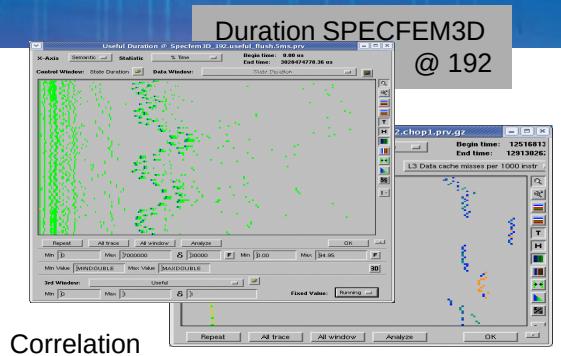
- MHD Relativista
  - GENESIS, RATPENAT
- Cosmología
  - pDEVA, GADGET
- Misión GAIA
  - GASS, GOG

- Simulación DNS de Turbulencia
  - LISO
- Simulación DNS Turbulencia + partículas
  - DNS2

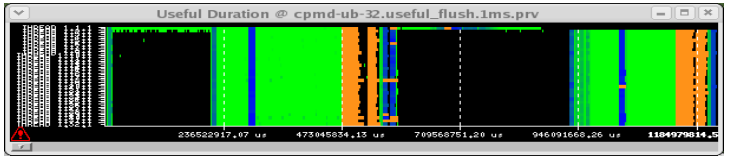
- Escalabilidad de SIESTA
- Daño por radiación
  - SIESTA

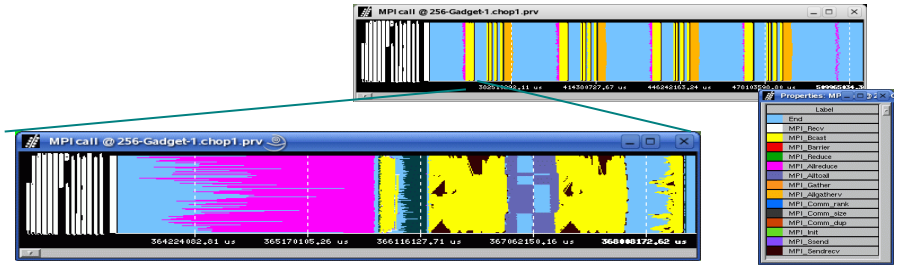
The need of performance analysis tools

- 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



Correlation
load imbalance - L2 misses

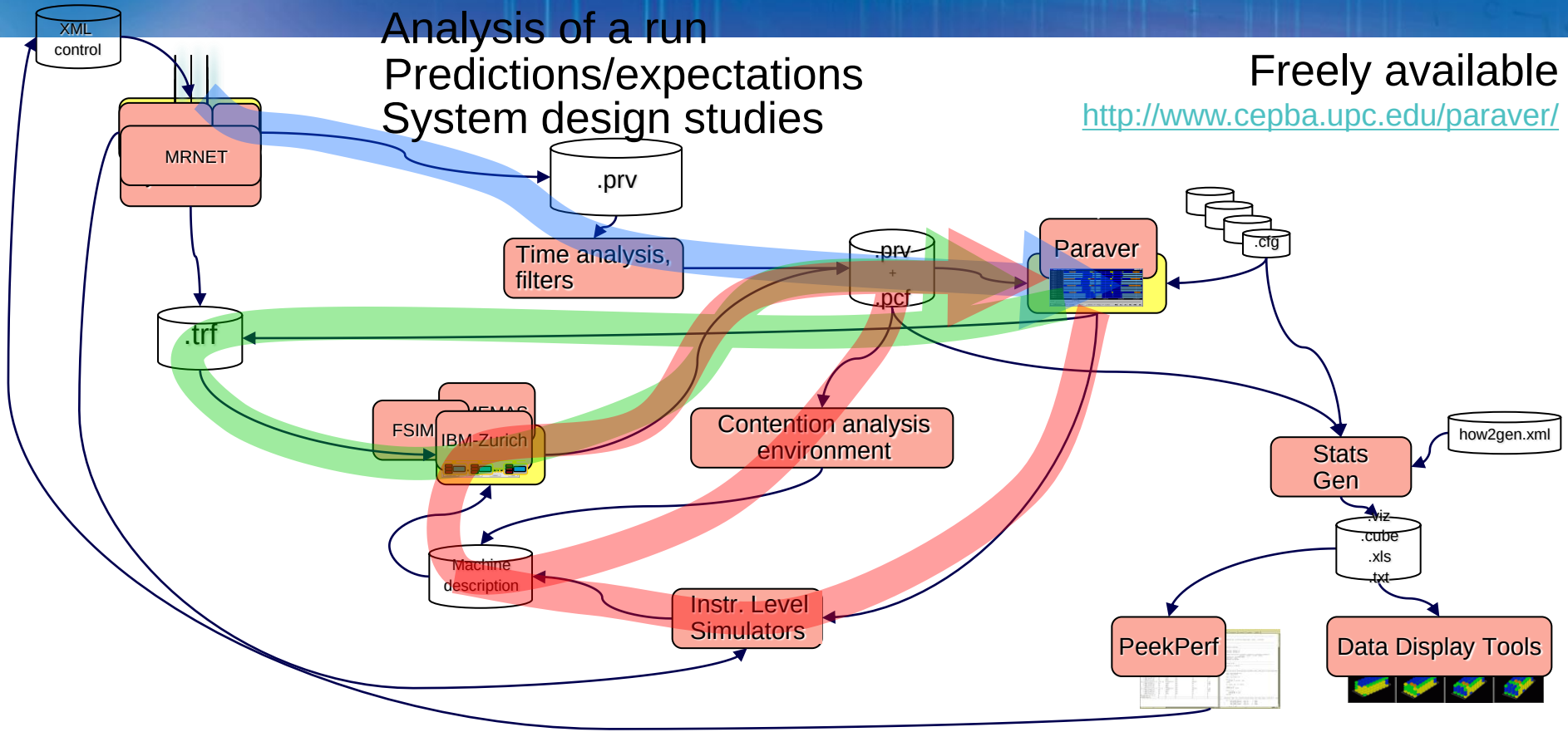




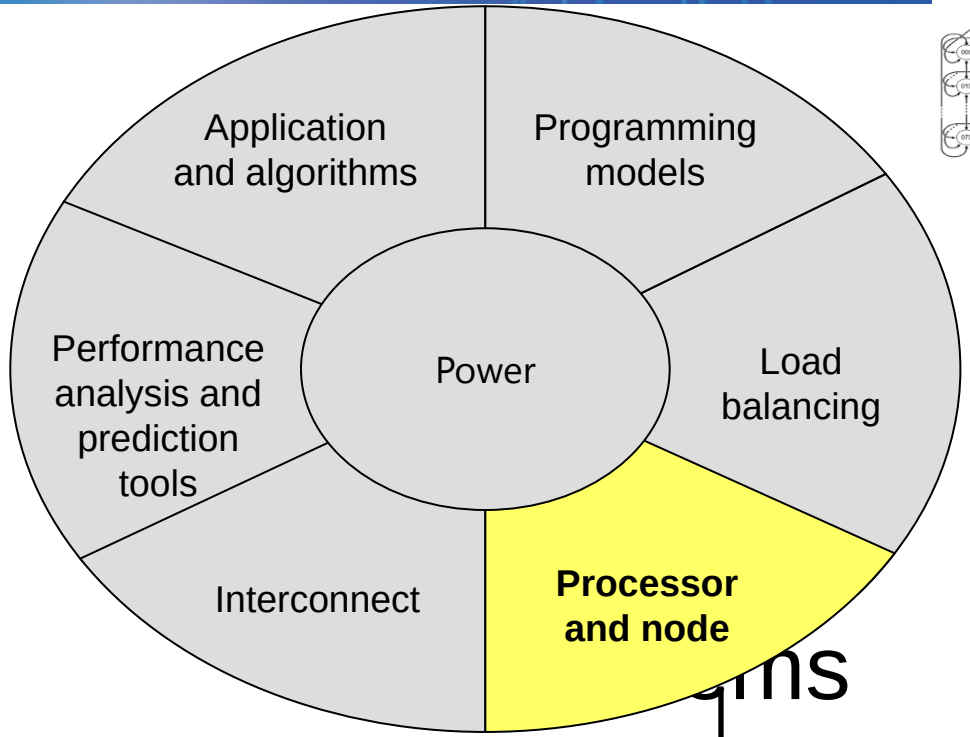
CEPBA-Tools environment

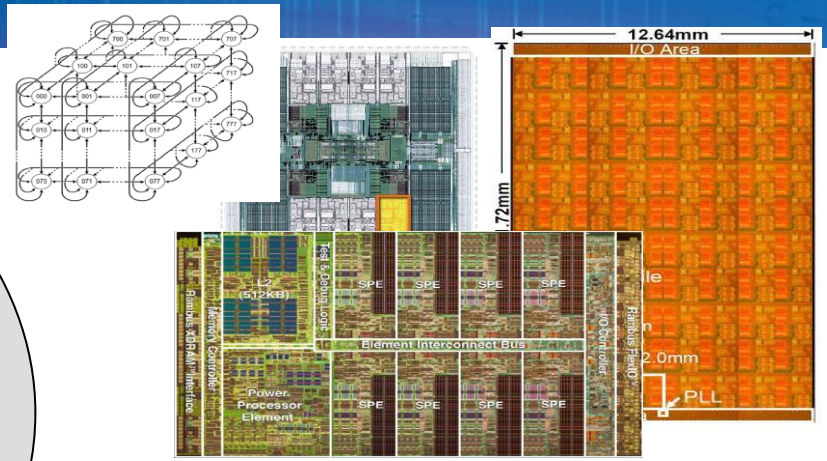
Analysis of a run
Predictions/expectations
System design studies

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



Multidisciplinary top-down approach



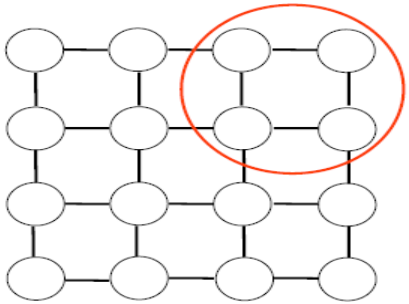


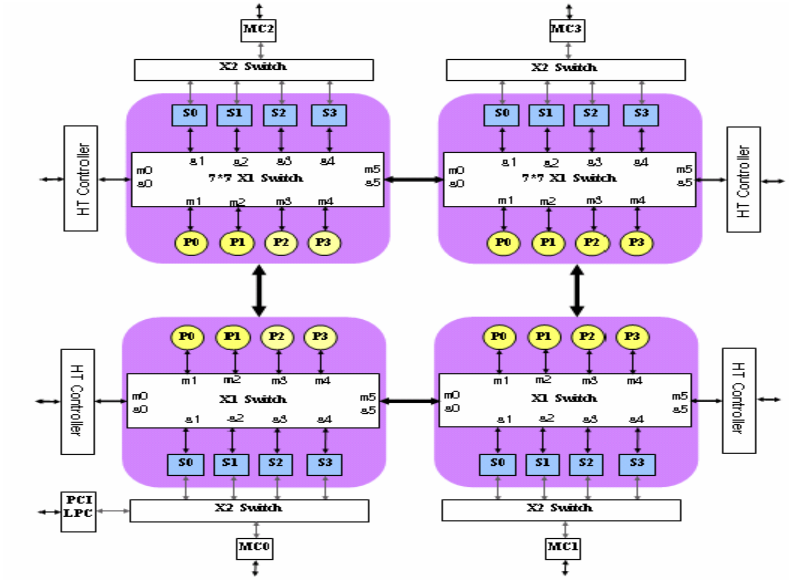
- ASIC
- Homogeneous multicore
- Accelerators
  - GPGPUs
  - FPGAs
  - Cell/B.E.

Dawning 6000, late 2011, 10000 Godson chips, 1 Petaflop

16-core Godson-3C

16 four-issue 64-bit Core
2*256-bit Vector Ext. per core
1.5GHz@28nm
384GFLOPS@15W
4 DDR3, 4 HT Controllers
To be taped out 2011



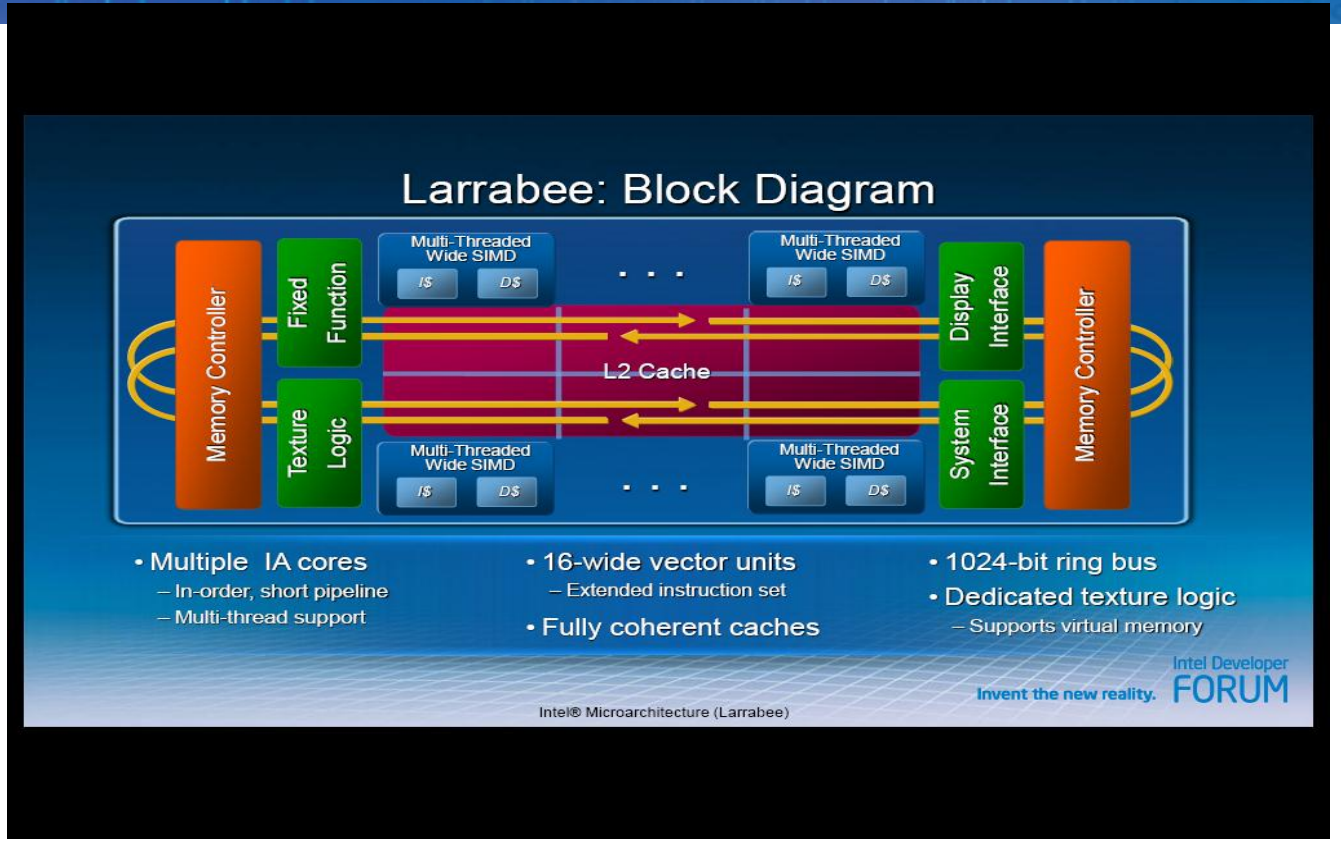


Larrabee

Since 2002 (Roger Espasa, Toni Juan)

40 People

Microprocessor Development (Larrabee x86 many core)



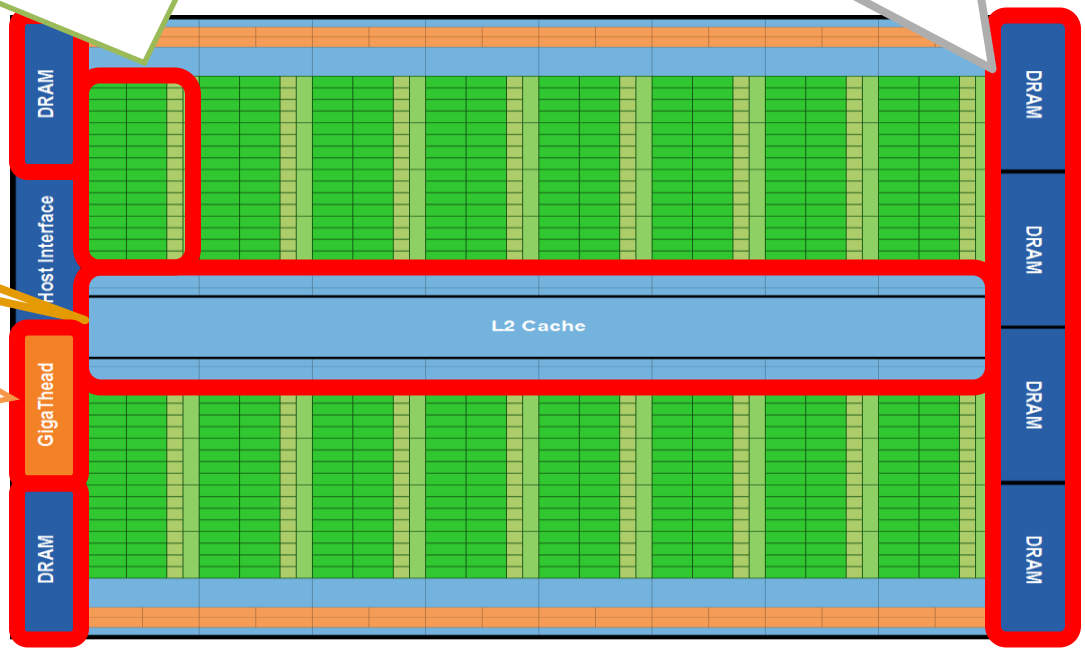
NVIDIA Fermi Architecture

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



The holistic approach …Processor Architecture

Towards exaflop

Applications
Performance Tools
Programming Model
Load Balancing
Interconnection
Processor/node architecture

Latency

Bandwidth

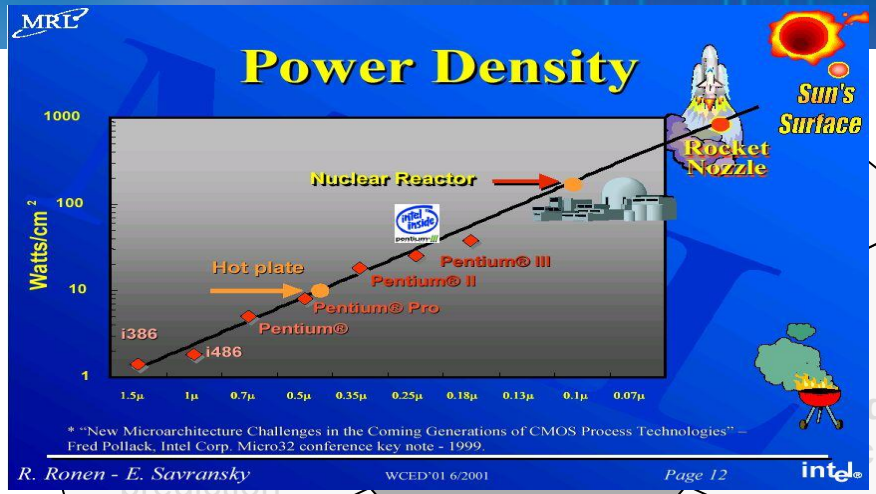
Parallelism

Scheduling

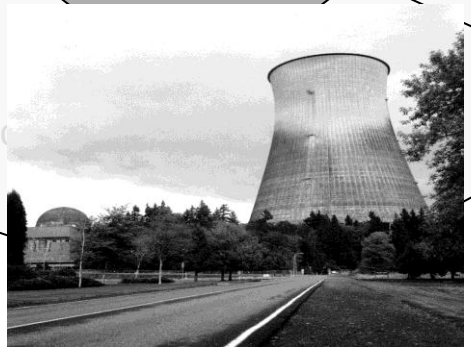
Power

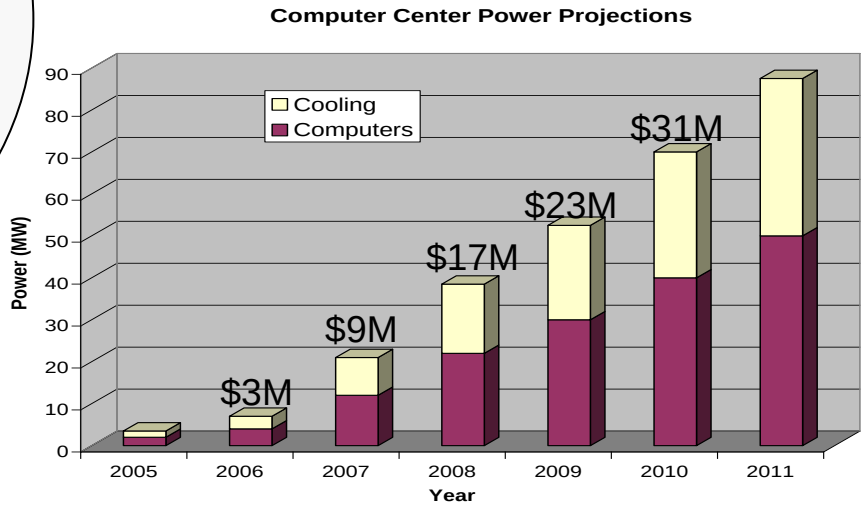
Cartagena, Colombia, May 18-20

Multidisciplinary top-down approach



Google





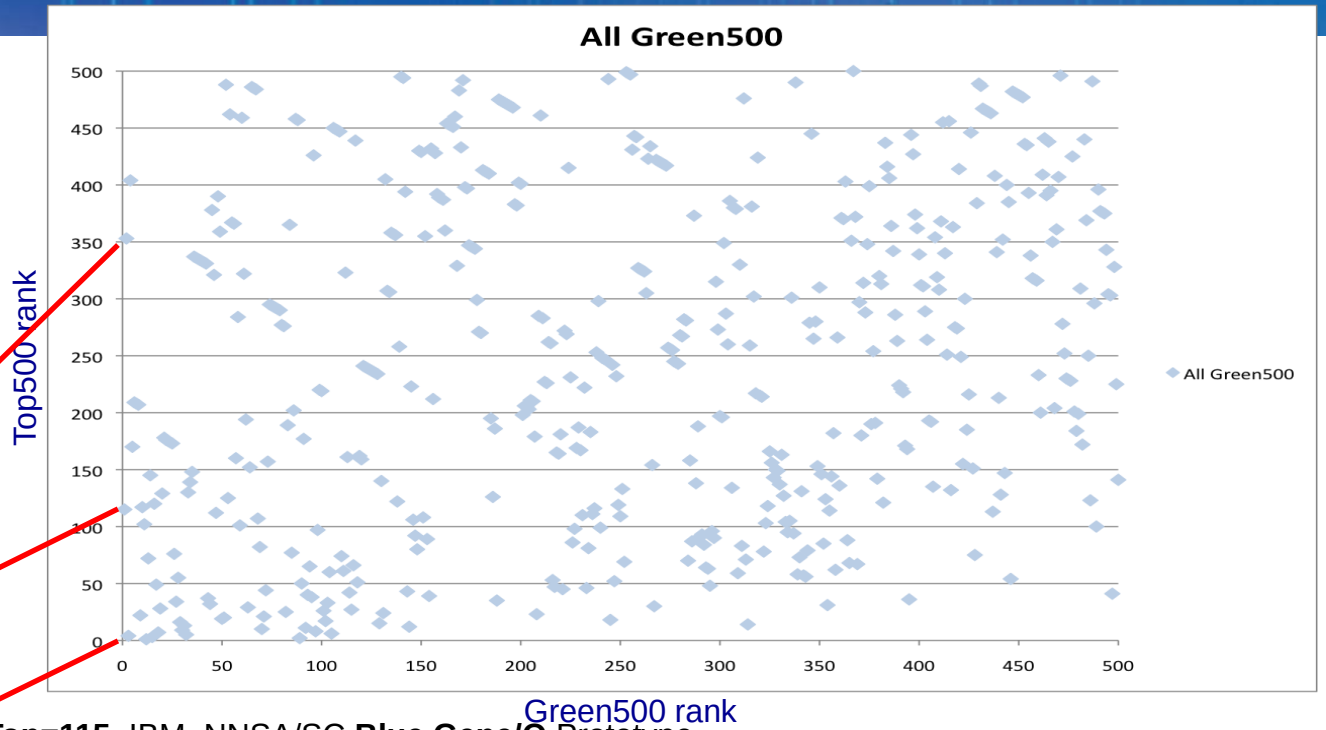
Cartagena, Colombia, May 18-20

Green/Top 500 November 2010

Green500 rank	Top500 rank	Mflops per watt	Total power	Site	Computer
1	115	1684,20	38,8	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype
2+	353	1448,03	24,58512	National Astronomical Observatory of Japan	GRAPE-DR accelerator Cluster, Infiniband
2	4	958,35	1.243,8	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows
3	404	933,06	36,	NCSA	Hybrid Cluster Core i3 2.93Ghz Dual Core, NVIDIA C2050, Infiniband
4	170	828,67	57,96	RIKEN Advanced Institute for Computational Sci.	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect
5	209	773,38	57,54	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus
5	208	773,38	57,54	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus
5	207	773,38	57,54	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus
8	22	740,78	385,	Universitaet Frankfurt	Supermicro Cluster, QC Opteron 2.1 GHz, ATI Radeon GPU, Infiniband
9	117	677,12	94,4	Georgia Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5660 2.8Ghz, nVidia Fermi, Infiniband QDR
10	102	636,36	117,15	National Institute for Environmental Studies	GOSAT Research Computation Facility, nvidia
11	1	635,15	4.040,	National Supercomputing Center in Tianjin	NUDT YH Cluster, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C
12	72	628,13	160	Lawrence Livermore National Laboratory	Appro GreenBlade Cluster Xeon X5660 2.8Ghz, nVIDIA M2050, Infiniband
13	145	555,50	94,6	CSIRO	Supermicro Xeon Cluster, E5462 2.8 Ghz, Nvidia Tesla s2050 GPU, Infiniband
14	3	492,64	2.580,	National Supercomputing Centre in Shenzhen	Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU
15	120	458,33	138,	IBM Poughkeepsie Benchmarking Center	BladeCenter QS22/LS21, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz,
15	49	458,33	276,	DOE/NNSA/LANL	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz,
17	7	444,25	2.345,5	DOE/NNSA/LANL	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz,
18	28	431,88	480,	IInst Process Eng., Chinese Academy of Sciences	Mole-8.5 Cluster Xeon L5520 2.26 Ghz, nVidia Tesla, Infiniband
19	129	400,64	141,75	Banking (M)	iDataPlex, Xeon X56xx 6C 2.66 GHz, Infiniband
20	178	400,62	118,13	University of British Columbia - Cancer Center,	iDataPlex, Xeon X56xx 6C 2.66 GHz, Infiniband
21	176	378,77	126,	Max-Planck-Gesellschaft MPI/IPP	Blue Gene/P Solution
21	175	378,77	126,	IBM Thomas J. Watson Research Center	Blue Gene/P Solution
21	174	378,77	126,	IBM - Rochester	Blue Gene/P Solution
21	173	378,77	126,	Ecole Polytechnique Federale de Lausanne	Blue Gene/P Solution
21	76	378,77	252,	EDF R&D	Blue Gene/P Solution
21	34	378,77	504,	King Abdullah Univ. of Science and Technology	Blue Gene/P Solution

Cartagena, Colombia, May 18-20

Green/Top 500 November 2010



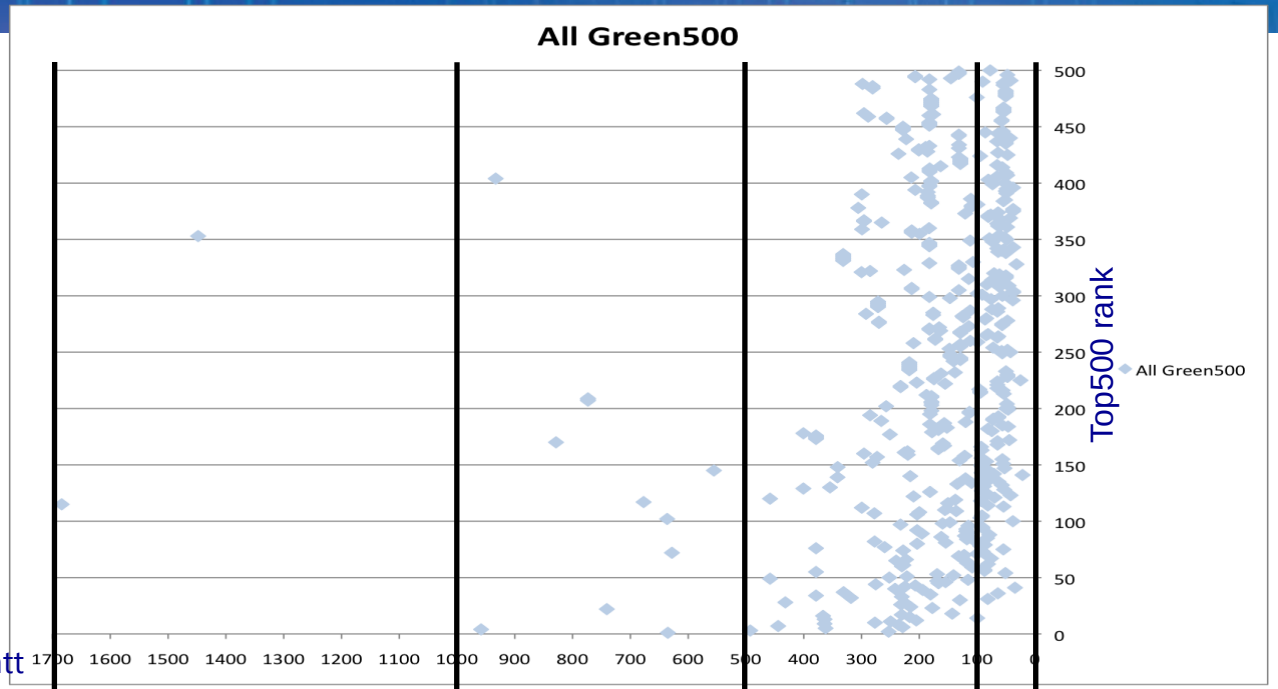
Green=1 (1684 Mflops/watt); Top=115. IBM. NNSA/SC **Blue Gene/Q** Prototype

Green=2+ (1448 Mflops/watt); Top=353. National Astronomical Observatory of Japan **GRAPE-DR accelerator**

Green=2 (958Mflops/watt); Top= 4. GSIC Center, Tokyo Inst. of Technology. HP ProLiant Xeon 6C X5670, **Nvidia M2050**

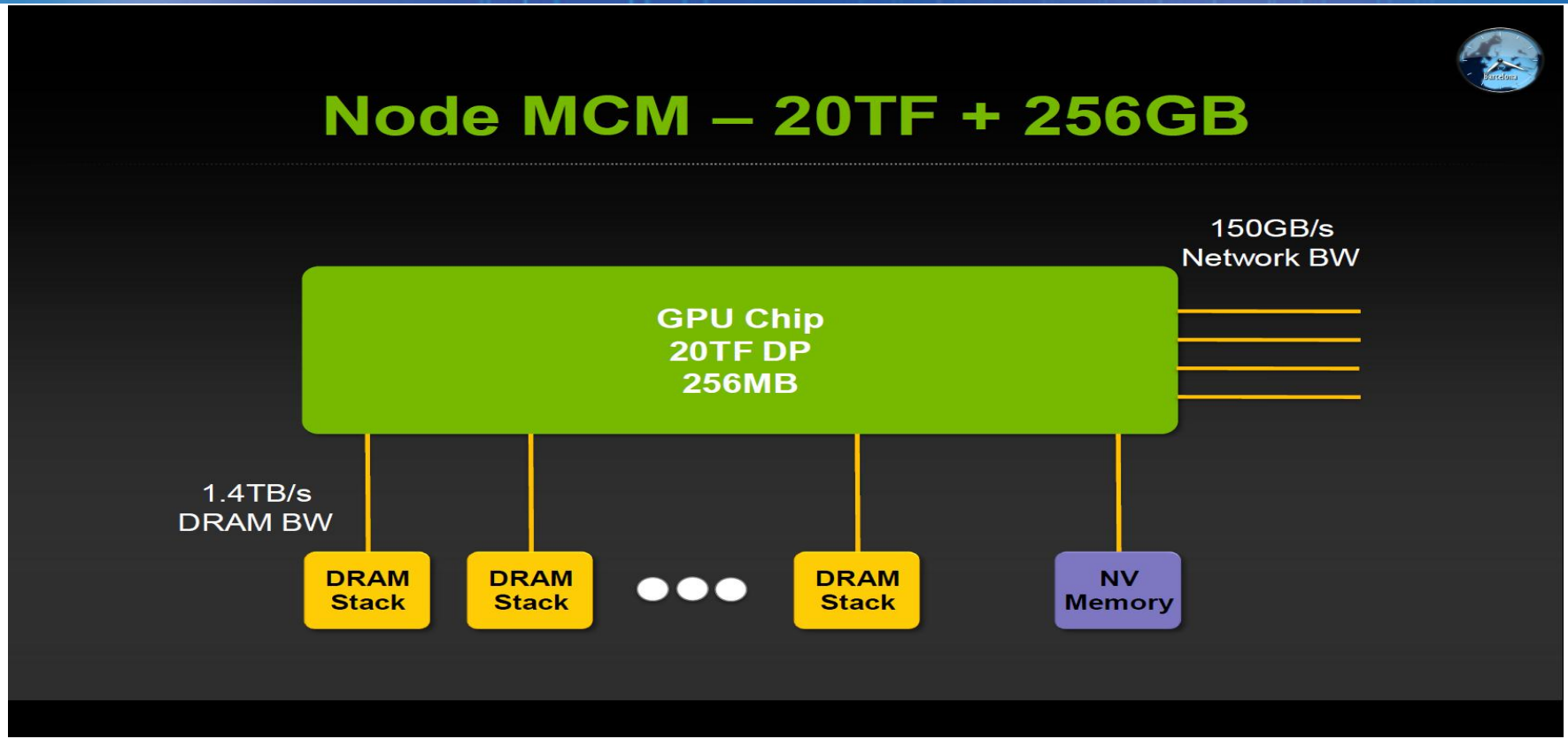
Cartagena, Colombia, May 18-20

Green/Top 500 November 2010



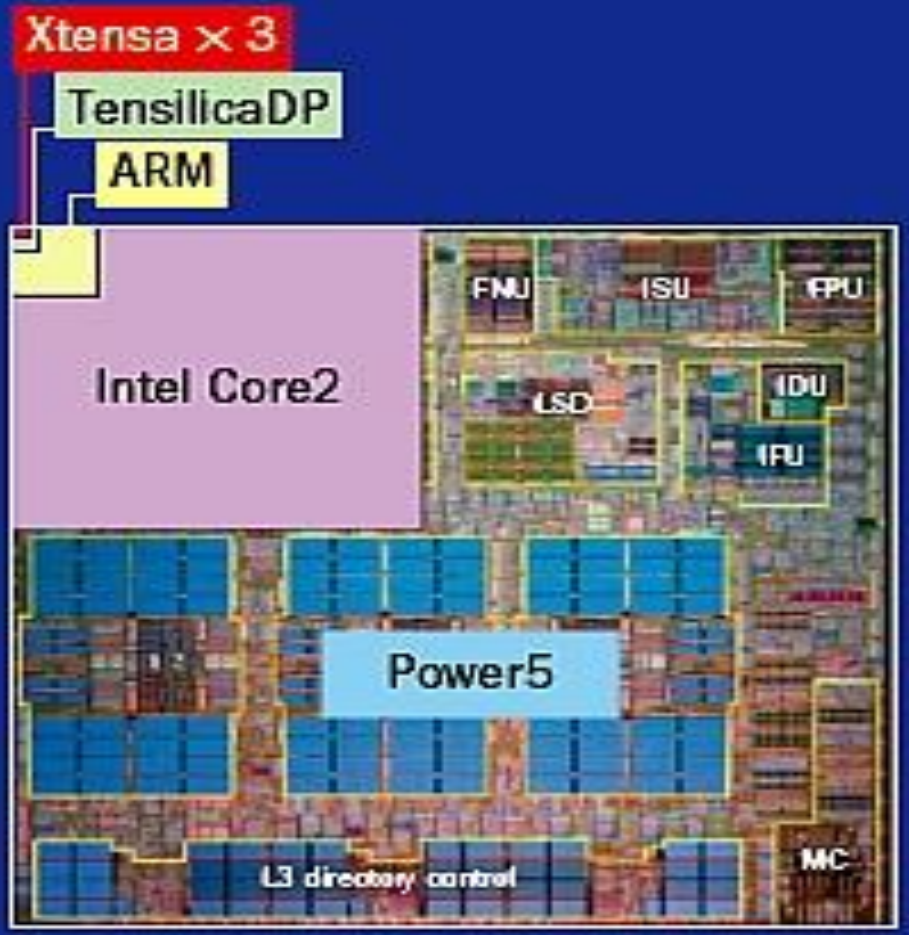
Mflops/watt	Mwatts/Exaflop
1684	594
1448	691

Cartagena, Colombia, May 18-20


Node MCM – 20TF + 256GB
150GB/s
Network BW
GPU Chip
20TF DP
256MB
1.4TB/s
DRAM BW
DRAM
Stack
DRAM
Stack
DRAM
Stack
NV
Memory

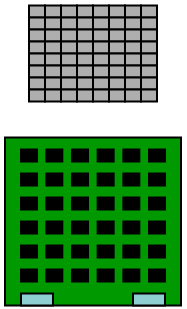
System – to ExaScale and Beyond

Dragonfly Interconnect
400 Cabinets is ~1EF and ~15MW

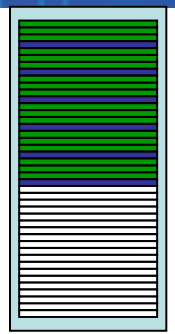


- **Power5 (Server)**
  - 389 mm^2
  - 120 W @ 1,900 MHz
- **Intel Core2 sc (Laptop)**
  - 130 mm^2
  - 15 W @ 1,000 MHz
- **ARM Cortex A8 (Automobiles)**
  - 5 mm^2
  - 0.8 W @ 800 MHz
- **Tensilica DP (Cell Phones/Printers)**
  - 0.8 mm^2
  - 0.09 W @ 600 MHz
- **Tensilica Xtensa (Cisco Router)**
  - 0.32 mm^2 for 3I
  - 0.05 W @ 600 MHz

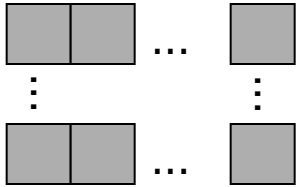
Montblanc: Architecture requirements for a 200 PF machine on 10 MW



Multi-core chip:
60 GFLOPS /W
10 Watts
600 GFLOPS
8 GFLOPS / core
75 cores / chip
0.15 Watts / core
Compute node:
36 chips
2.700 cores
22 TFLOPS
1.000 Watts / node



Rack:
42 compute nodes
1.512 chips
86.400 cores
0.9 PFLOPS
50 Kwatts / rack



Exascale system:
225 racks
16.800 nodes
604.800 chips
4.5 M cores
200 PFLOPS
10 MWatts

- 200 PF on 10 MW require a power efficiency of 20 GFLOPS / Watt
  - BG/Q, the current Geen500 leader only achieves 1.7 GFLOPS / Watt ...
- Only 35% of the total energy is spent on the processors
  - 35% goes to memories, 20% to storage and network, 10% to cooling
- The processor needs to achieve 60 GFLOPS / Watt
  - 600 GFLOPS on a low-power 10 Watt chip
  - 75 cores / chip (assuming 8 GFLOPS / core)
  - 0.15 Watts / core
    - **Current ARM Cortex-A9 @ 800 MHz requires only 0.25 Watts**

B C W

PUMPS Summer School

Barcelona Computing Week
July 18-22, 2011*

Barcelona Supercomputing Center
Universitat Politècnica de Catalunya

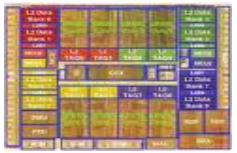
Programming and Tuning Massively Parallel Systems

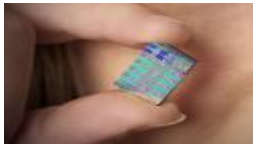
- Invited instructors:
  - Wen-mei Hwu, University of Illinois
  - David B. Kirk, NVIDIA Corporation
- Audience:
  - Parallel tracks specially designed for beginners, advanced and developers
- Programming Languages:
  - CUDA, OpenCL, OpenMP, StarSs, MPI
- Hands-on Labs:
  - Afternoon labs with Teaching Assistants for each audience/level
- Case studies and algorithmic techniques:
  - Graph, tiling, grid, stencil, reductions, sorting and binning, sparse matrices…
  - Molecular dynamics, medical imaging, computer vision, dense linear systems…

More information: http://bcw.ac.upc.edu

Cartagena, Colombia, May 18-20

Education for Parallel Programming

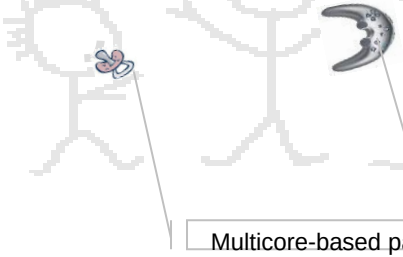




I ❤ many-core programming

I ❤ multi-core programming

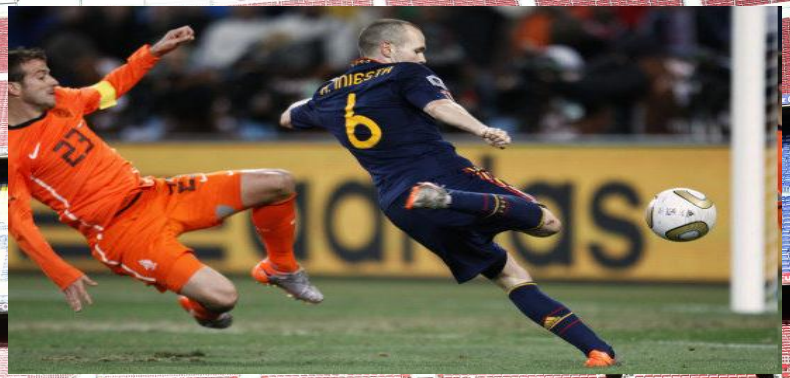
We all ❤ massive parallel prog



I ❤ games

Multicore-based pacifier

Barcelona Supports C



Cartagena, Colombia, May 18-20

Thank you

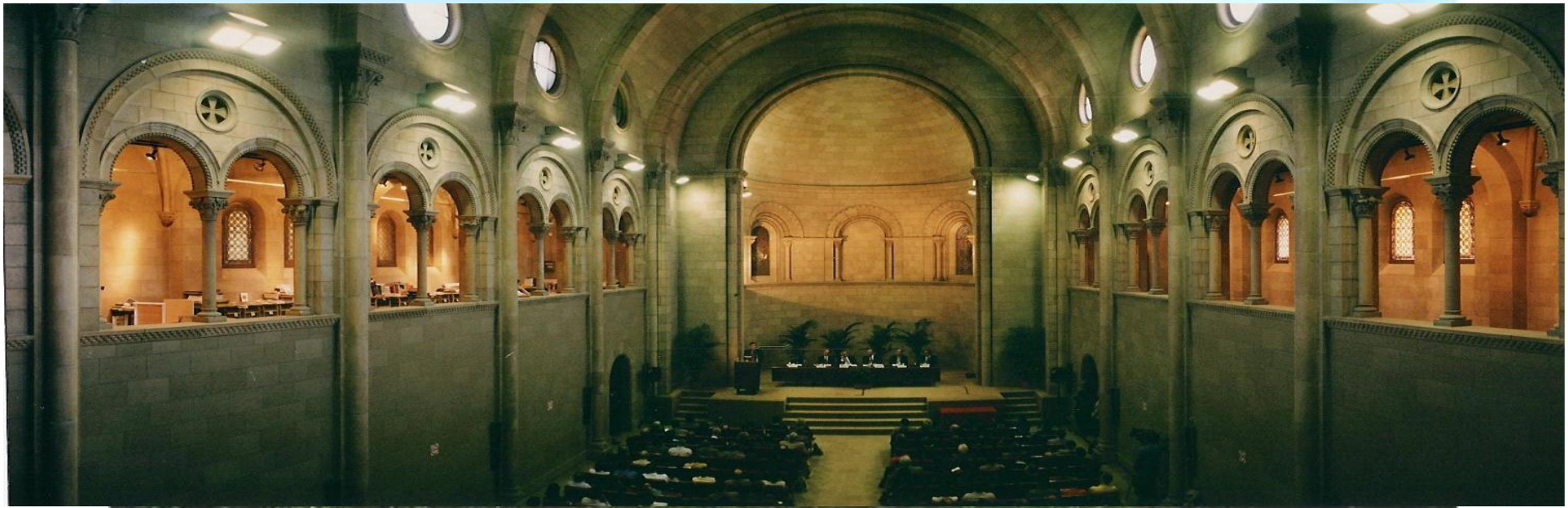
Cartagena, Colombia, May 18-20

Are we planning to upgrade?

- Negotiating our next site ;)

Cartagena, Colombia, May 18-20

Location



Cartagena, Colombia, May 18-20