Microsoft® Research
Faculty Summit 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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer</th>
<th>Procs</th>
<th>Rmax</th>
<th>Rpeak</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tianjin, China</td>
<td>XeonX5670+NVIDIA</td>
<td>186368</td>
<td>2566000</td>
<td>4701000</td>
</tr>
<tr>
<td>2</td>
<td>Oak Ridge Nat. Lab.</td>
<td>Cray XT5, 6 cores</td>
<td>224162</td>
<td>1759000</td>
<td>2331000</td>
</tr>
<tr>
<td>3</td>
<td>Shenzhen, China</td>
<td>XeonX5670+NVIDIA</td>
<td>120640</td>
<td>1271000</td>
<td>2984300</td>
</tr>
<tr>
<td>4</td>
<td>GSIC Center, Tokyo</td>
<td>XeonX5670+NVIDIA</td>
<td>73278</td>
<td>1192000</td>
<td>2287630</td>
</tr>
<tr>
<td>5</td>
<td>DOE/SC/LBNL/NERSC</td>
<td>Cray XE6 12 cores</td>
<td>153408</td>
<td>1054000</td>
<td>1288630</td>
</tr>
<tr>
<td>6</td>
<td>Commissariat a l'Energie Atomique (CEA)</td>
<td>Bull bulx super-node S6010/S6030</td>
<td>138368</td>
<td>1050000</td>
<td>1254550</td>
</tr>
<tr>
<td>7</td>
<td>DOE/NNSA/LANL</td>
<td>QS22/LS21 Cluster, PowerXCell 8i / Opteron Infiniband</td>
<td>122400</td>
<td>1042000</td>
<td>1375780</td>
</tr>
<tr>
<td>8</td>
<td>National Institute for Computational Sciences/University of Tennessee</td>
<td>Cray XT5-HE 6 cores</td>
<td>98928</td>
<td>831700</td>
<td>1028850</td>
</tr>
<tr>
<td>9</td>
<td>Forschungszentrum Juelich (FZJ)</td>
<td>Blue Gene/P Solution</td>
<td>294912</td>
<td>825500</td>
<td>825500</td>
</tr>
<tr>
<td>10</td>
<td>DOE/NNSA/LANL/SNL</td>
<td>Cray XE6 8-core</td>
<td>107152</td>
<td>816600</td>
<td>1028660</td>
</tr>
</tbody>
</table>
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; 1.5x10^5 Processors
  - Superconductive materials
- 1 EFlop/s; ~2018; ?; 1x10^8 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?

Current #1 (6/2010)
- 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

**Cartagena, Colombia, May 18-20**
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 $\to$ 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

Cartagena, Colombia, May 18-20
# Systems Scaling Projections

<table>
<thead>
<tr>
<th>Design Parameters</th>
<th>BG/L</th>
<th>BG/P</th>
<th>25PF</th>
<th>300PF</th>
<th>1200PF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores / Node</td>
<td>2</td>
<td>4</td>
<td>8-24</td>
<td>32-64</td>
<td>96-128</td>
</tr>
<tr>
<td>Clock Speed (GHz)</td>
<td>0.7</td>
<td>0.86</td>
<td>1.6-4.1</td>
<td>2.3-4.8</td>
<td>2.8-6.0</td>
</tr>
<tr>
<td>Flops / Clock / Core</td>
<td>4</td>
<td>4</td>
<td>8-32</td>
<td>8-32</td>
<td>16-64</td>
</tr>
<tr>
<td>Nodes / Rack</td>
<td>1024</td>
<td>1024</td>
<td>100-1024</td>
<td>256-1024</td>
<td>256-1024</td>
</tr>
<tr>
<td>Racks / Full System Config</td>
<td>64</td>
<td>72</td>
<td>128-350</td>
<td>128-400</td>
<td>256-400</td>
</tr>
<tr>
<td>MB RAM/core</td>
<td>256</td>
<td>512</td>
<td>1024-4096</td>
<td>1024-4096</td>
<td>1024-4096</td>
</tr>
<tr>
<td>Total Power</td>
<td>2.5MW</td>
<td>4.8MW</td>
<td>8MW-20MW</td>
<td>20MW-50MW</td>
<td>30MW-80MW</td>
</tr>
<tr>
<td>Flops / Node (GF)</td>
<td>5.6</td>
<td>14</td>
<td>128-640</td>
<td>640-2000</td>
<td>2000-6000</td>
</tr>
<tr>
<td>Flops / Rack (TF)</td>
<td>5.7</td>
<td>14</td>
<td>200-400</td>
<td>400-1200</td>
<td>1600-4800</td>
</tr>
<tr>
<td>LB Concurrency</td>
<td>5E+05</td>
<td>1E+06</td>
<td>1E6-64E6</td>
<td>100E6-1E9</td>
<td>1E9-10E9</td>
</tr>
</tbody>
</table>

## 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
Cartagena, Colombia, May 18-20

<table>
<thead>
<tr>
<th>Applications</th>
<th>Job Scheduling</th>
<th>Programming Model</th>
<th>Run time</th>
<th>Interconnection</th>
<th>Processor/node architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Applications**
- **Job Scheduling**
- **Programming Model**
- **Run time**
- **Interconnection**
- **Processor/node architecture**

- **Load Balancing**
- **Concurrent load scheduling**
- **Moldability**
- **Resource awareness**
- **User satisfaction**
- **Address space**
- **Work generation**
- **Locality optimization**
- **Topology and routing**
- **External contention**
- **NIC design**
- **Run time support**
- **Memory subsystem**

- **Comput. Complexity**
- **Async. Algs.**
- **Malleability**
- **Overhead**
- **Power efficiency**
- **Concurrency extraction**
- **Concurrency extraction**
- **Hw counters**
- **Core Structure**

**Towards exaflop**

**Holistic approach**...
BSC-CNS: International Initiatives (IESP)

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

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

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
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](http://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
The EU, Latin America and the Caribbean region must enhance their cooperation to face common challenges ahead, from climate change to taking 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 cooperation 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

**Applications**
- 4 relevant apps:
  - Materials: SIESTA
  - Geophysics imaging: RTM
  - Comp. Mechanics: ALYA
  - Plasma: EUTERPE
- General kernels

**Performace analysis tools**
- Automatic analysis
- Coarse/true grain prediction
- Sampling
- Clustering
- Integration with Peekperf

**Models and prototype**
- Contention, Collectives
- Overlap computation/communication
- Slimmed Networks
- Direct versus indirect networks

**Programming models**
- StarSs: CellSs, SMPSs
- OpenMP@Cell
- OpenMP++
- MPI + OpenMP/StarSs

**Load balancing**
- Coordinated scheduling:
  - Run time, Process, Job
  - Power efficiency

**Processor and node**
- Contribution to new Cell design
- Support for programming model
- Support for load balancing
- Support for performance tools
- Issues for future processors

Cartagena, Colombia, May 18-20
Kaleidoscope Project

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

- 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

Cartagena, Colombia, May 18-20
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,...

Cartagena, Colombia, May 18-20
The holistic approach ...

<table>
<thead>
<tr>
<th>Applications</th>
<th>Performance Tools</th>
<th>Programming Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Merge nicely with node level</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Asynchrony/dependences/dataflow</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abstract/simple memory model</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fine grain DLP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Basic transformations (unroll,…)</td>
</tr>
</tbody>
</table>

### Towards exaflop

**Load Balancing**
- Monitor/set tunable and scheduling variables

**Interconnection**
- Detect data production to enable overlapped transfer

**Processor/node architecture**
- Efficient support for basic mechanisms, data transfers, thread creation, dependence handling

---

Cartagena, Colombia, May 18-20
StarSs: ... generates task graph at run time...

```c
#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

Cartagena, Colombia, May 18-20
Cartagena, Colombia, May 18-20

StarSs: ... and executes as efficient as possible ...

```c
#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, handwritten scheduling


Cartagena, Colombia, May 18-20
Hybrid MPI/SMPSs: Linpack example

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

```c
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


Cartagena, Colombia, May 18-20
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

Cartagena, Colombia, May 18-20

Open Source
http://www.bsc.es/smpsuperscalar
http://nanos.ac.upc.edu/
Related FP7 projects

- Dataflow at all levels
- Data flow and transactional memory

- Chip and memory architecture for 256+ cores
- Focused on task based programming models
The TEXT project

• Towards EXaflop applications

• Demonstrate that Hybrid MPI/SMPSs addresses the Exascale challenges in a 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 GPUs)

• Support other users
  • Identify users of TEXT applications
  • Identify and support interested application developers

• Contribute to Standards (OpenMP ARB, PERI-XML)
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

Cartagena, Colombia, May 18-20
CEPBA-Tools environment

Analysis of a run
Predictions/expectations
System design studies

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

Machine description
Time analysis, filters
Contention analysis environment
Instr. Level Simulators

Stats Gen
PeakPerf
Data Display Tools

XML control
MRNET

FSIM
IBM-Zurich

Controversial
Multidisciplinary top-down approach

- Application and algorithms
- Programming models
- Performance analysis and prediction tools
- Interconnect
- Processor and node
- Power
- Load balancing

- 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)

Cartagena, Colombia, May 18-20
NVIDIA Fermi Architecture

16 Streaming-Multiprocessors (512 cores) execute Thread Blocks

**620 Gigaflops**

Unified 768KB L2 cache serves all threads

GigaThread hardware scheduler assigns Thread Blocks to SMs

Wide DRAM interface provides 12 GB/s bandwidth

Cartagena, Colombia, May 18-20
The holistic approach ... Processor Architecture

Applications
Performance Tools
Programming Model
Load Balancing
Interconnection

Processor/node architecture

Applications
Performance Tools
Programming Model
Load Balancing
Interconnection

Latency
Bandwidth
Parallelism
Scheduling
Power

Towards exaflop

Cartagena, Colombia, May 18-20
Multidisciplinary top-down approach

Power Density

Watts/cm²

Nuclear Reactor

Rocket Nozzle

Hot plate

Pentium® II

Pentium® III

Intel

Computer Center Power Projections

$3M

$9M

$17M

$23M

$31M

2005 2006 2007 2008 2009 2010 2011

Year

Power (MW)

R. Ronen - E. Savransky

WCED'01 6/2001

Page 12

Cartagena, Colombia, May 18-20
<table>
<thead>
<tr>
<th>Green500 rank</th>
<th>Top500 rank</th>
<th>Mflops per watt</th>
<th>Total power</th>
<th>Site</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>115</td>
<td>1684.20</td>
<td>38.8</td>
<td>IBM Thomas J. Watson Research Center</td>
<td>NNSA/SC BlueGene/Q Prototype</td>
</tr>
<tr>
<td>2+</td>
<td>353</td>
<td>1448.03</td>
<td>24.58512</td>
<td>National Astronomical Observatory of Japan</td>
<td>GRAPE-DR accelerator Cluster, Infiniband</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>958.35</td>
<td>1.243,8</td>
<td>GSIC Center, Tokyo Institute of Technology</td>
<td>HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows</td>
</tr>
<tr>
<td>3</td>
<td>404</td>
<td>933.06</td>
<td>36.</td>
<td>NCSA</td>
<td>Hybrid Cluster Core i3 2.93Ghz Dual Core, NVIDIA C2050, Infiniband</td>
</tr>
<tr>
<td>4</td>
<td>170</td>
<td>828.67</td>
<td>57.96</td>
<td>RIKEN Advanced Institute for Computational Sci.</td>
<td>K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect</td>
</tr>
<tr>
<td>5</td>
<td>209</td>
<td>773.38</td>
<td>57.54</td>
<td>Universitaet Wuppertal</td>
<td>QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus</td>
</tr>
<tr>
<td>5</td>
<td>208</td>
<td>773.38</td>
<td>57.54</td>
<td>Universitaet Regensburg</td>
<td>QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus</td>
</tr>
<tr>
<td>5</td>
<td>207</td>
<td>773.38</td>
<td>57.54</td>
<td>Forschungszentrum Juelich (FZJ)</td>
<td>QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>740.78</td>
<td>385.</td>
<td>Universitaet Frankfurt</td>
<td>Supermicro Cluster, QC Opteron 2.1 GHz, ATI Radeon GPU, Infiniband</td>
</tr>
<tr>
<td>9</td>
<td>117</td>
<td>677.12</td>
<td>94.4</td>
<td>Georgia Institute of Technology</td>
<td>HP ProLiant SL390s G7 Xeon 6C X5660 2.8Ghz, nVidia Fermi, Infiniband QDR</td>
</tr>
<tr>
<td>10</td>
<td>102</td>
<td>636.36</td>
<td>117.15</td>
<td>National Institute for Environmental Studies</td>
<td>GOSAT Research Computation Facility, nvidia</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>635.15</td>
<td>4.040</td>
<td>National Supercomputing Center in Tianjin</td>
<td>NUDT YH Cluster, X5670 2.93GHz 6C, NVIDIA GPU, FT-1000 8C</td>
</tr>
<tr>
<td>12</td>
<td>72</td>
<td>628.13</td>
<td>160</td>
<td>Lawrence Livermore National Laboratory</td>
<td>Appro GreenBlade Cluster Xeon X5660 2.8Ghz, nVIDIA M2050, Infiniband</td>
</tr>
<tr>
<td>13</td>
<td>145</td>
<td>555.50</td>
<td>94.6</td>
<td>CSIRO</td>
<td>Supermicro Xeon Cluster, E5462 2.8 Ghz, Nvidia Tesla s2050 GPU, Infiniband</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>492.64</td>
<td>2.580</td>
<td>National Supercomputing Centre in Shenzhen</td>
<td>Dawning TC3600 Blade, Intel X5650, Nvidia Tesla C2050 GPU</td>
</tr>
<tr>
<td>15</td>
<td>120</td>
<td>458.33</td>
<td>138.</td>
<td>IBM Poughkeepsie Benchmarking Center</td>
<td>BladeCenter QS22/LS21, PowerXCell 8i 3.2 Gz / Opteron DC 1.8 Ghz</td>
</tr>
<tr>
<td>15</td>
<td>49</td>
<td>458.33</td>
<td>276.</td>
<td>DOE/NNSA/LANL</td>
<td>BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Gz / Opteron DC 1.8 Ghz</td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>444.25</td>
<td>2.345,5</td>
<td>DOE/NNSA/LANL</td>
<td>BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Gz / Opteron DC 1.8 Ghz</td>
</tr>
<tr>
<td>18</td>
<td>28</td>
<td>431.88</td>
<td>480.</td>
<td>IInst Process Eng., Chinese Academy of Sciences</td>
<td>Mole-8.5 Cluster Xeon L5520 2.26 Ghz , nvidia Tesla, Infiniband</td>
</tr>
<tr>
<td>19</td>
<td>129</td>
<td>400.64</td>
<td>141.75</td>
<td>Banking (M)</td>
<td>iDataPlex, Xeon X56xx 6C 2.66 Ghz, Infiniband</td>
</tr>
<tr>
<td>20</td>
<td>178</td>
<td>400.62</td>
<td>118.13</td>
<td>University of British Columbia - Cancer Center,</td>
<td>iDataPlex, Xeon X56xx 6C 2.66 Ghz, Infiniband</td>
</tr>
<tr>
<td>21</td>
<td>176</td>
<td>378.77</td>
<td>126.</td>
<td>Max-Planck-Gesellschaft MPI/IPP</td>
<td>Blue Gene/P Solution</td>
</tr>
<tr>
<td>21</td>
<td>175</td>
<td>378.77</td>
<td>126.</td>
<td>IBM Thomas J. Watson Research Center</td>
<td>Blue Gene/P Solution</td>
</tr>
<tr>
<td>21</td>
<td>174</td>
<td>378.77</td>
<td>126.</td>
<td>IBM - Rochester</td>
<td>Blue Gene/P Solution</td>
</tr>
<tr>
<td>21</td>
<td>173</td>
<td>378.77</td>
<td>126.</td>
<td>Ecole Polytechnique Federale de Lausanne</td>
<td>Blue Gene/P Solution</td>
</tr>
<tr>
<td>21</td>
<td>76</td>
<td>378.77</td>
<td>252.</td>
<td>EDF R&amp;D</td>
<td>Blue Gene/P Solution</td>
</tr>
<tr>
<td>21</td>
<td>34</td>
<td>378.77</td>
<td>504.</td>
<td>King Abdullah Univ. of Science and Technology</td>
<td>Blue Gene/P Solution</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Mflops/watt</th>
<th>Mwatts/Exaflop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1684</td>
<td>594</td>
</tr>
<tr>
<td>1448</td>
<td>691</td>
</tr>
</tbody>
</table>

>1 GF/watt  
500-1000 MF/watt  
100-500 MF/watt  
20-100 MF/watt
Node MCM – 20TF + 256GB

GPU Chip
20TF DP
256MB

1.4TB/s
DRAM BW

150GB/s
Network BW

DRAM Stack

DRAM Stack

DRAM Stack

NV Memory

Cartagena, Colombia, May 18-20
System – to ExaScale and Beyond

Dragonfly Interconnect
400 Cabinets is ~1EF and ~15MW
- **Power5 (Server)**
  - 389 mm²
  - 120 W @ 1,900 MHz
- **Intel Core2 sc (Laptop)**
  - 130 mm²
  - 15 W @ 1,000 MHz
- **ARM Cortex A8 (Automobiles)**
  - 5 mm²
  - 0.8 W @ 800 MHz
- **Tensilica DP (Cell Phones/Printers)**
  - 0.8 mm²
  - 0.09 W @ 600 MHz
- **Tensilica Xtensa (Cisco Router)**
  - 0.32 mm² for 3I
  - 0.05 W @ 600 MHz
Montblanc: Architecture requirements for a 200 PF machine on 10 MW

- 200 PF on 10 MW require a power efficiency of 20 GFLOPS / Watt
  - BG/Q, the current Green500 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**

Cartagena, Colombia, May 18-20
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
Education for Parallel Programming

Cartagena, Colombia, May 18-20
Barcelona Supports

Cartagena, Colombia, May 18-20
Thank you
Are we planning to upgrade?

- Negotiating our next site ;)

Cartagena, Colombia, May 18-20