MONT-BLANC

http://www.montblanc-project.eu

European scalable and power efficient HPC platform based on low-power embedded technology

Alex Ramirez

Barcelona Supercomputing Center

Technical Coordinator

This project and the research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement n° 288777.

Project goal

- To develop an **European** exascale approach
- Based on embedded **power-efficient technology**

- Funded under FP7 Objective ICT-2011.9.13 Exa-scale computing, software and simulation
  - 3-year IP Project (October 2011 - September 2014)
  - Total budget: 14.5 M€ (8.1 M€ EC contribution),
    - 1095 Person-Month

Project objectives

- Objective 1: To deploy a **prototype HPC system** based on currently **available energy-efficient embedded technology**
  - Scalable to 50 PFLOPS on 7 MWatt
    - Competitive with Green500 leaders in 2014
  - Deploy a full HPC system software stack

- Objective 2: To design a next-generation HPC system and **new embedded technologies** targeting HPC systems that would overcome most of the limitatior encountered in the prototype system
  - Scalable to 200 PFLOPS on 10 MWatt
    - Competitive with Top500 leaders in 2017

- Objective 3: To port and optimise a small number of **representative exascale applications** capable of

Power defines performance

- Prototype goal: 50 PFLOPS on 7 MWatt
  - 7 GFLOPS / Watt efficiency
- Required improvement on energy efficiency
  - 3.5x over BG/Q
  - 5x over ATI GPU systems
  - 7x over Nvidia GPU systems
  - 8.5x over SPARC64 multi-core
  - 9x over Cell systems

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2097.19	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 2	40.95
2	1684.20	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 1	38.80
3	1375.88	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR	34.24
4	958.35	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	1243.80
5	891.88	CINECA / SCS - SuperComputing Solution	iDataPlex DX360M3, Xeon 2.4, nVidia GPU, Infiniband	160.00
6	824.56	RIKEN Advanced Institute for Computational Science (AICS)	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	9898.56
7	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54

Challenges and Opportunities

- Challenges
  - Exploit massive number of low-power processors
    - Exploit compute accelerators
  - Sustain performance with lower bandwidth components
    - Interconnect
    - Memory
  - Programmability

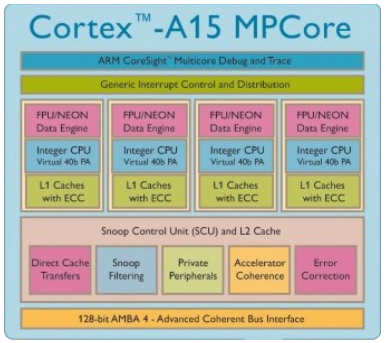
- Why do we think we can make it?
  - Energy-efficient building blocks
  - Hybrid MPI+OmpSs programming model

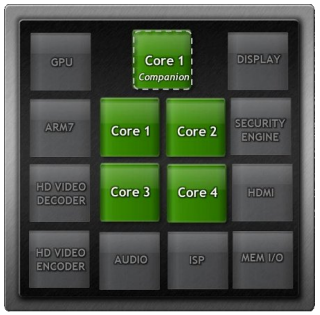
Energy-efficient building blocks

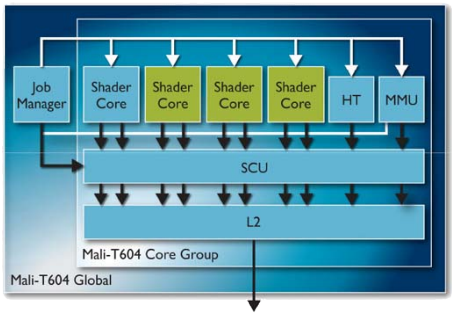
- Integrated system design built from mobile / embedded components

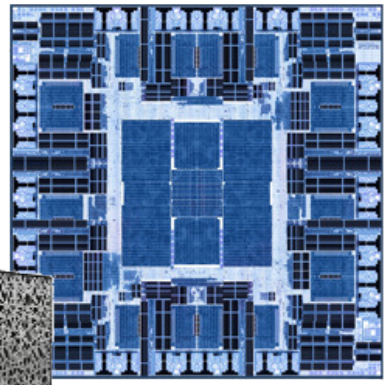
- ARM multicore processors
  - Nvidia Tegra / Denver, Calxeda, Marvell Armada, ST-Ericsson Nova A9600, TI OMAP 5, …
- Mobile accelerators
  - Mobile GPU
    - Nvidia GT 500M, …
  - Embedded GPU
    - Nvidia Tegra, ARM Mali T604
- Low power 10 GbE switches
  - Gnodal GS 256

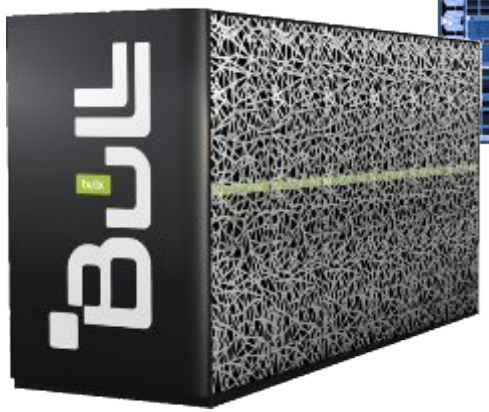
- Tier-0 system integration experience
  - BullX systems in the Top10



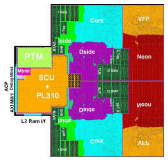






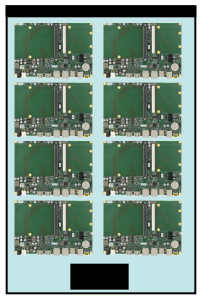


PRACE prototype @ BSC: ARM multicore

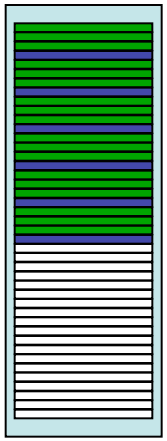


Tegra2 SoC:
2x ARM Corext-A9 Cores
2 GFLOPS
0.5 Watt

Tegra2 Q7 module:
1x Tegra2 SoC
2x ARM Corext-A9 Cores
1 GB DDR2 DRAM
2 GFLOPS
~4 Watt
1 GbE interconnect



1U Multi-board container:
1x Board container
8x Q7 carrier boards
8x Tegra2 SoC
16x ARM Corext-A9 Cores
8 GB DDR2 DRAM
16 GFLOPS
~35 Watt



Rack:
32x Board container
10x 48-port 1GbE switches
256x Q7 carrier boards
256x Tegra2 SoC
512x ARM Corext-A9 Cores
256 GB DDR2 DRAM
512 GFLOPS
~1.7 Kwatt

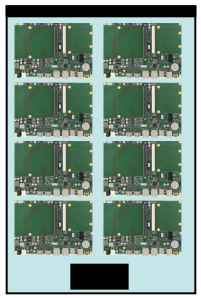
300 MFLOPS / W

- First large-scale ARM cluster prototype
- Proof-of-concept to demonstrate HPC based on low-power components
  - Built entirely from COTS components
  - Mont-Blanc integrated design could improve substantially
- Enabler for early software development and tuning
  - Open-source system software stack
  - Application development and tuning to ARM platform

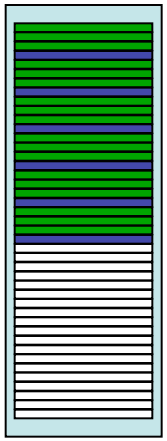
PRACE prototype @ BSC: ARM + mobile GPU

Tegra3 Q7 module:
1x Tegra3 SoC
  4x Corext-A9 @ 1.5 GHz
4 GB DDR3 DRAM
6 GFLOPS
~4 Watt
1 Gbe interconnect

Nvidia GeForce 520MX
48 CUDA cores @ 900 MHz
142 GFLOPS
12 Watts
11.8 GFLOPS / W



1U Multi-board container:
1x Board container
8x Q7 carrier boards
  32x ARM Corext-A9 Cores
  8x GT520MX GPU
32 GB DDR3 DRAM
1.2 TFLOPS
~140 Watt



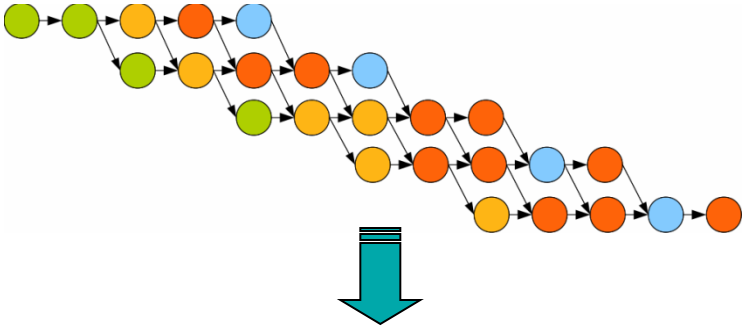
Rack:
32x Board container
10x 48-port 1GbE switches
256x Q7 carrier boards
  256x Tegra3 SoC
  1024x ARM Corext-A9 Cores
  256x GT520MX GPU
  1TB DDR3 DRAM
38 TFLOPS
~5 Kwatt

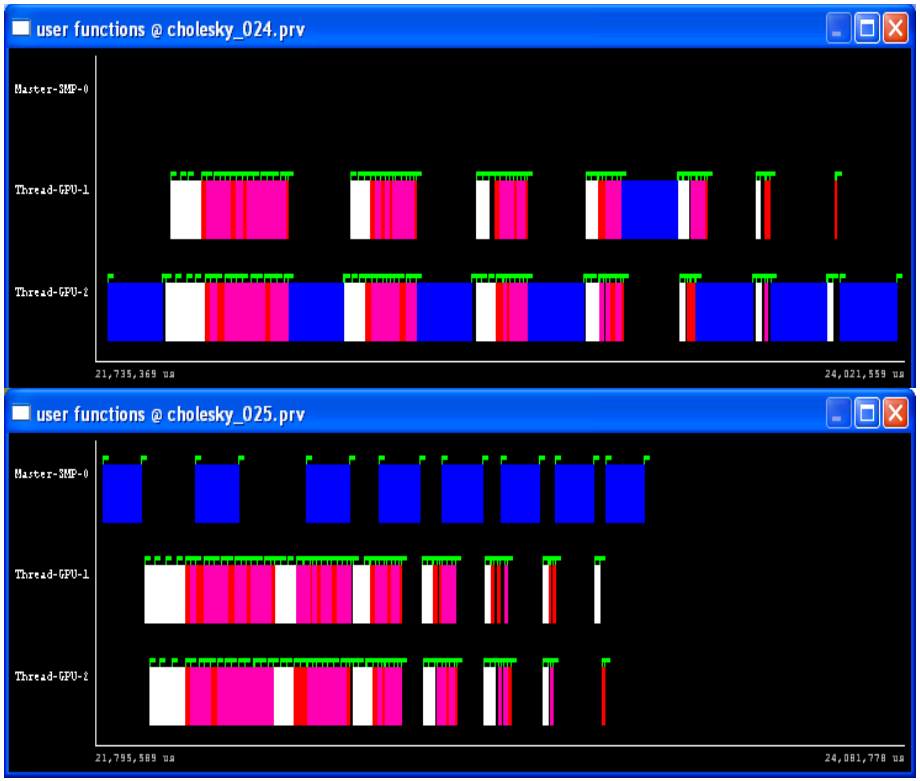
7.5 GFLOPS / W

- Increasing number of Top500 systems use GPU accelerators
- Validate the use of their energy efficient counterparts
  - ARM multicore processors
  - Mobile Nvidia GPU accelerators
- Perform scalability tests to high number of compute nodes
  - Higher core count required when using low-power processors
  - Evaluate impact of limited memory and bandwidth on low-end solutions

Hybrid MPI + OmpSs programming model

- Hide complexity from programmer
- Runtime system maps task graph to architecture
  - Many-core + accelerator exploitation
  - Asynchronous communication
    - Overlap communication + computation
  - Asynchronous data transfers
    - Overlap data transfer + computation
  - Strong scaling
    - Sustain performance with lower memory size per core
  - Locality management
    - Optimize data movement





System software porting + tuning

- Linux OS
- Filesystem
  - NFS, Lustre
- Parallel programming model + Runtime libraries
  - OmpSs, OpenMP, MPI, OpenCL
- Scientific libraries
  - ATLAS, FFTW, HDF5, LAPACK, MAGMA, ...
- Performance tools
  - Hardware performance counters
  - EXTRAE, PARAVER, SCALASCA
- Cluster management
  - Slurm, Ganglia

Target Mont-Blanc applications

- Real applications currently running in PRACE Tier-0 systems or National HPC facilities
  - YALES2 — Fluid Dynamics
  - EUTERPE — Fluid dynamics
  - SPECFEM3D — Seismic wave propagation
  - MP2C — Multi-particle collisions
  - BigDFT — Electronic structure
  - QuantumESPRESSO — Electronic structure
  - PEPC — Coulomb + gravitational forces
  - SMMP — Protein folding
  - ProFASI — Protein folding
  - COSMO — Meteorological modeling
  - BQCD — Quantum ChromoDynamics

Project results

- Prototype HPC system based on European embedded processors
  - Demonstrate potential of embedded technology for HPC
  - Target maximum power efficiency
  - Limited by currently available technology
- Design of a next-generation system
  - Full scale system paving the way towards Exascale computing
  - Proposal and definition of the required technologies to achieve it
- Open source system software stack
  - Operating system, runtime libraries, scientific libraries, performance tools
- Up to 11 full-scale scientific applications
  - Capable of exploiting the benefits of this new class of HPC architectures