Energy efficient computing on Embedded and Mobile devices

Nikola Rajovic, Nikola Puzovic, Lluis Vilanova,
Carlos Villavieja, **Alex Ramirez**

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

PRACE

MONT BLANC

SEVENTH FRAMEWORK PROGRAMME

A brief look at the (outdated) Top500 list

Rank	Site	Computer
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu
2	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C NUDT
3	DOE/SC/Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz Cray Inc.
4	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU Dawning
5	GSIC Center, Tokyo Institute of Technology Japan	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows NEC/HP
6	DOE/NNSA/LANL/SNL United States	Cielo - Cray XE6 8-core 2.4 GHz Cray Inc.
7	NASA/Ames Research Center/NAS United States	Pleiades - SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband SGI
8	DOE/SC/LBNL/NERSC United States	Hopper - Cray XE6 12-core 2.1 GHz Cray Inc.
9	Commissariat a l'Energie Atomique (CEA) France	Tera-100 - Bull bullx super-node S6010/S6030 Bull SA
10	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband IBM

- Most systems are built on general purpose multicore chips
  - Backwards compatibility
  - Programmer productivity

A brief look at the (soon to be outdated) Green500 list

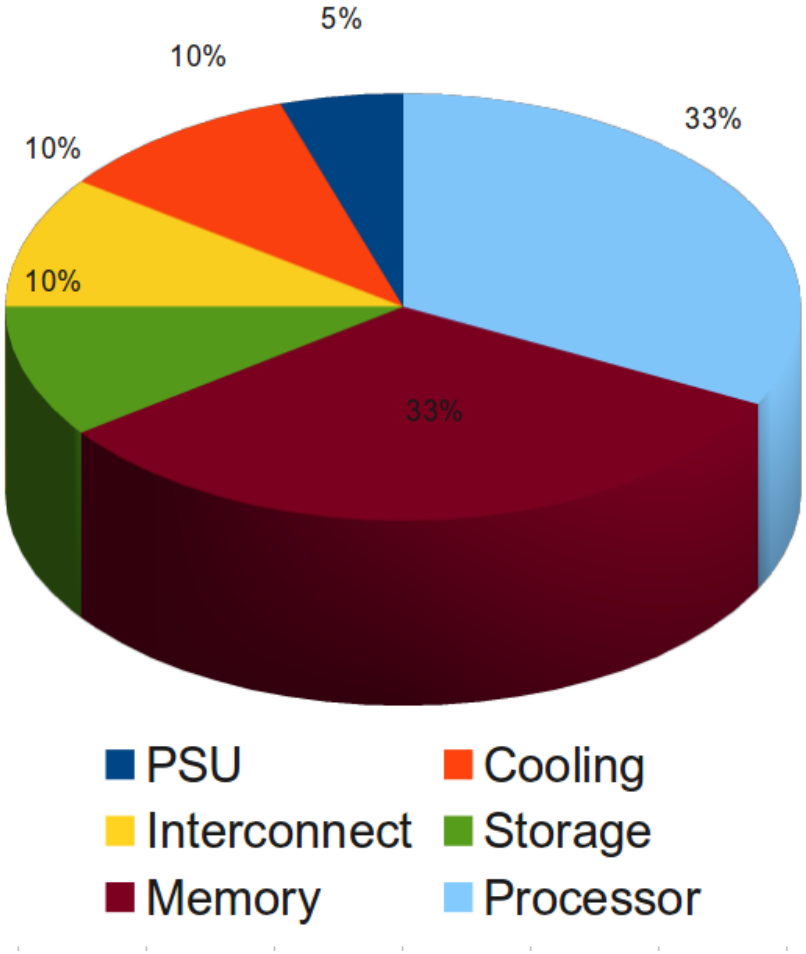
Green500 Rank	MFLOPS/W	Site*	Computer*
1	2097.19	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 2
2	1684.20	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 1
3	1375.88	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR
4	958.35	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows
5	891.88	CINECA / SCS - SuperComputing Solution	iDataPlex DX360M3, Xeon 2.4, nVidia GPU, Infiniband
6	824.56	RIKEN Advanced Institute for Computational Science (AICS)	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect
7	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus
8	773.38	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus
9	773.38	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus
10	718.13	Universitaet Frankfurt	Supermicro Cluster, QC Opteron 2.1 GHz, ATI Radeon GPU, Infiniband

- Most of the Top10 systems rely on accelerators for energy efficiency
  - ATI GPU
  - Nvidia GPU
  - IBM PowerXCell 8i

Some initial assertions

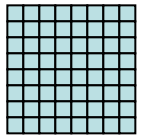
- You may disagree, but bear with me …

- Power distribution
  - 5% Power Supply
  - 10% Cooling
    - Direct water
  - 10% Interconnect
    - Not always active
  - 10% Storage
    - Not always active
  - 32.5% Memory
  - **32.5% Processor**

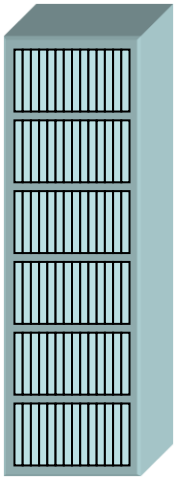


Now, some arithmetic (and some assumptions)

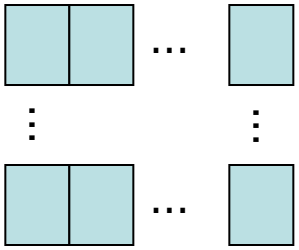
- Objective: 1 EFLOPS on 20 MWatt
- Blade-based multicore system design
  - 100 blades / rack
  - 2 sockets / blade
  - 150 Watts / socket
- CPU
  - 8 ops/cycle @ 2GHz = 16 GFLOPS
- 1 EFLOPS / 16 GFLOPS
  - **62.5 M cores**
- 32% of 20 MWatt = 6.4 MWatt
  - 6.4 MWatt / 62.5 M cores
  - **0.10 Watts / core**
- 150 Watt / socket
  - **1500 cores / socket**
  - 24 TFLOPS / socket



Multi-core chip:
150 Watts
24 TFLOPS
16 GFLOPS / core
1500 cores / chip
0.10 Watts / core



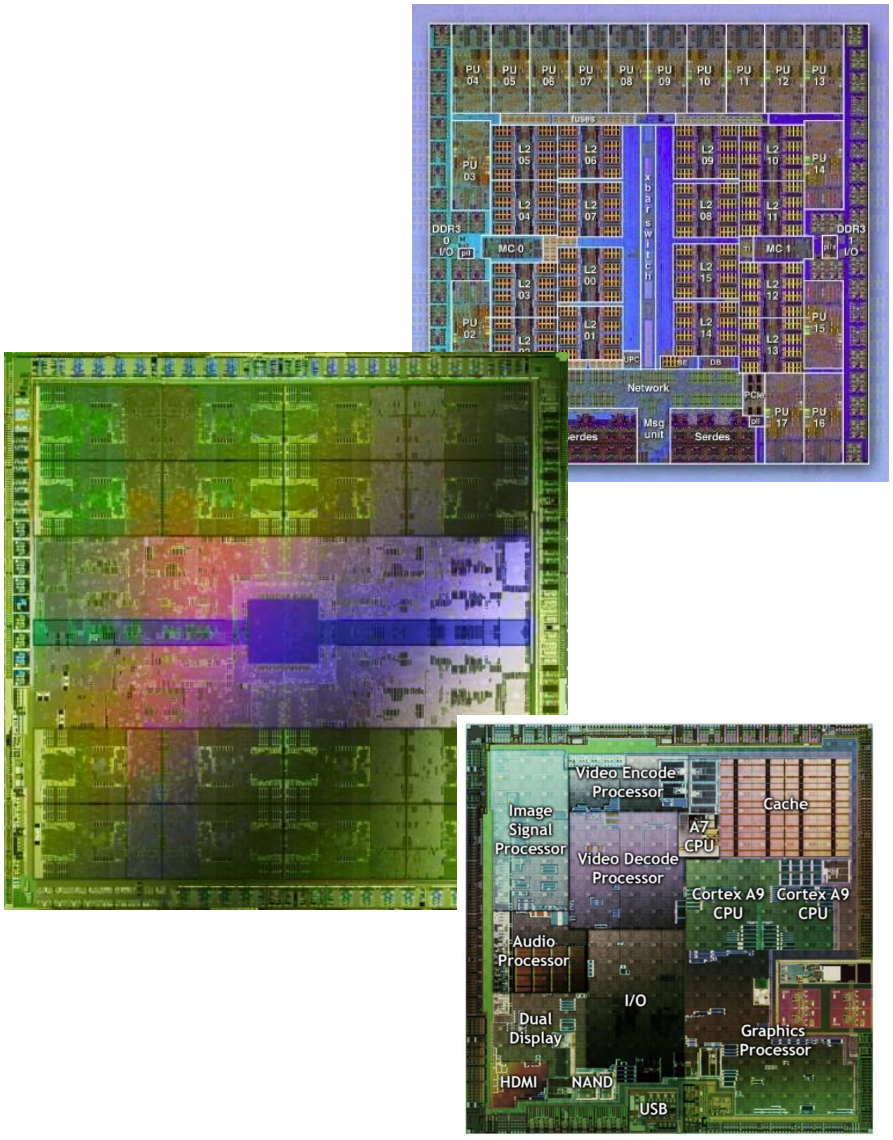
Rack:
100 compute nodes
200 chips
300.000 cores
4.8 PFLOPS
72 Kwatts / rack



Exaflop system:
210 racks
21.00 nodes
62.5 M cores
1.000 PFLOPS
20 MWatts

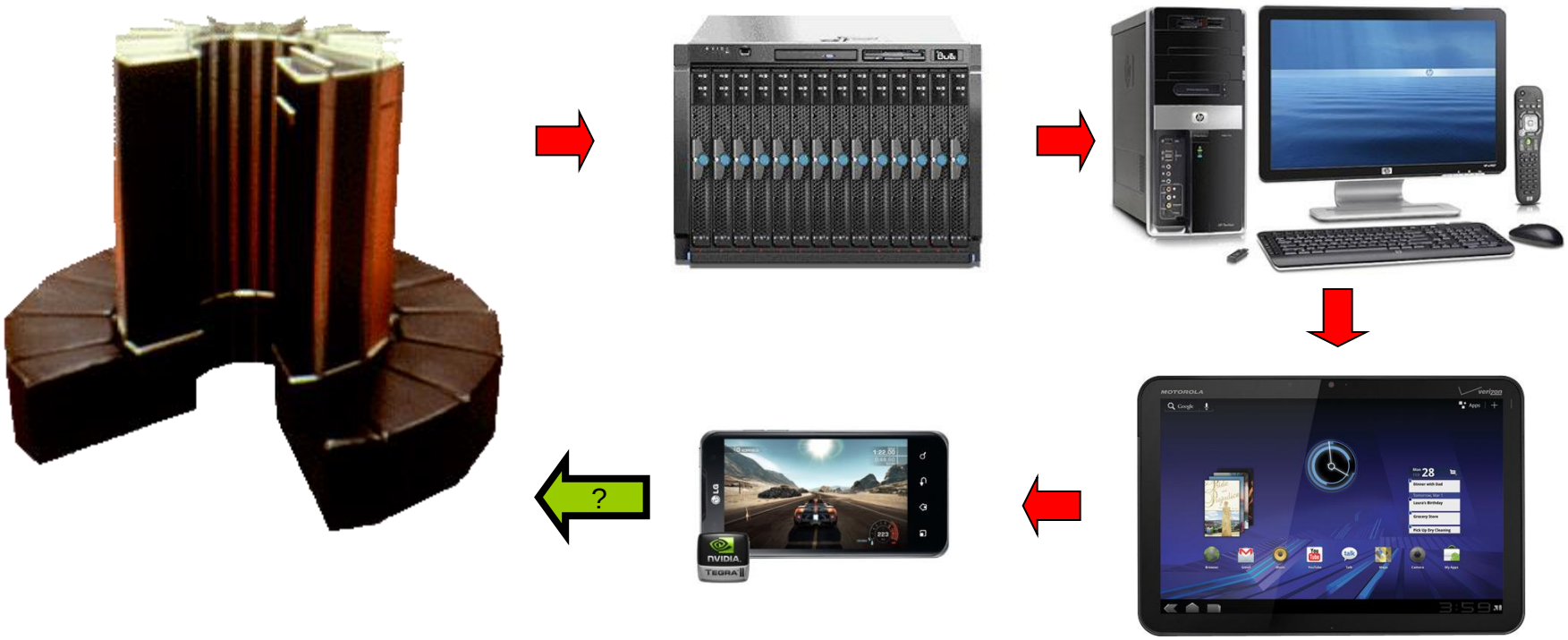
Where are we today?

- IBM BG/Q
  - 8 ops/cycle @ 1.6 GHz
  - 16 cores / chip
    - 16K cores / rack
  - ~2.5 Watt / core
- Fujitsu Ultrasparc VIIIfx
  - **8 ops / cycle @ 2GHz**
  - 8 cores / chip
  - 12 Watts / core
- Nvidia Tesla C2050-2070
  - **448 CUDA cores**
- ARM Cortex-A9
  - 1 ops / cycle @ 800 MHz - **2 GHz**
  - **0.25** - 1 Watt
- ARM Cortex-A15
  - *4 ops / cycle @ 1 - 2.5 GHz**
  - *0.35 Watt**
- All is there … but not together?



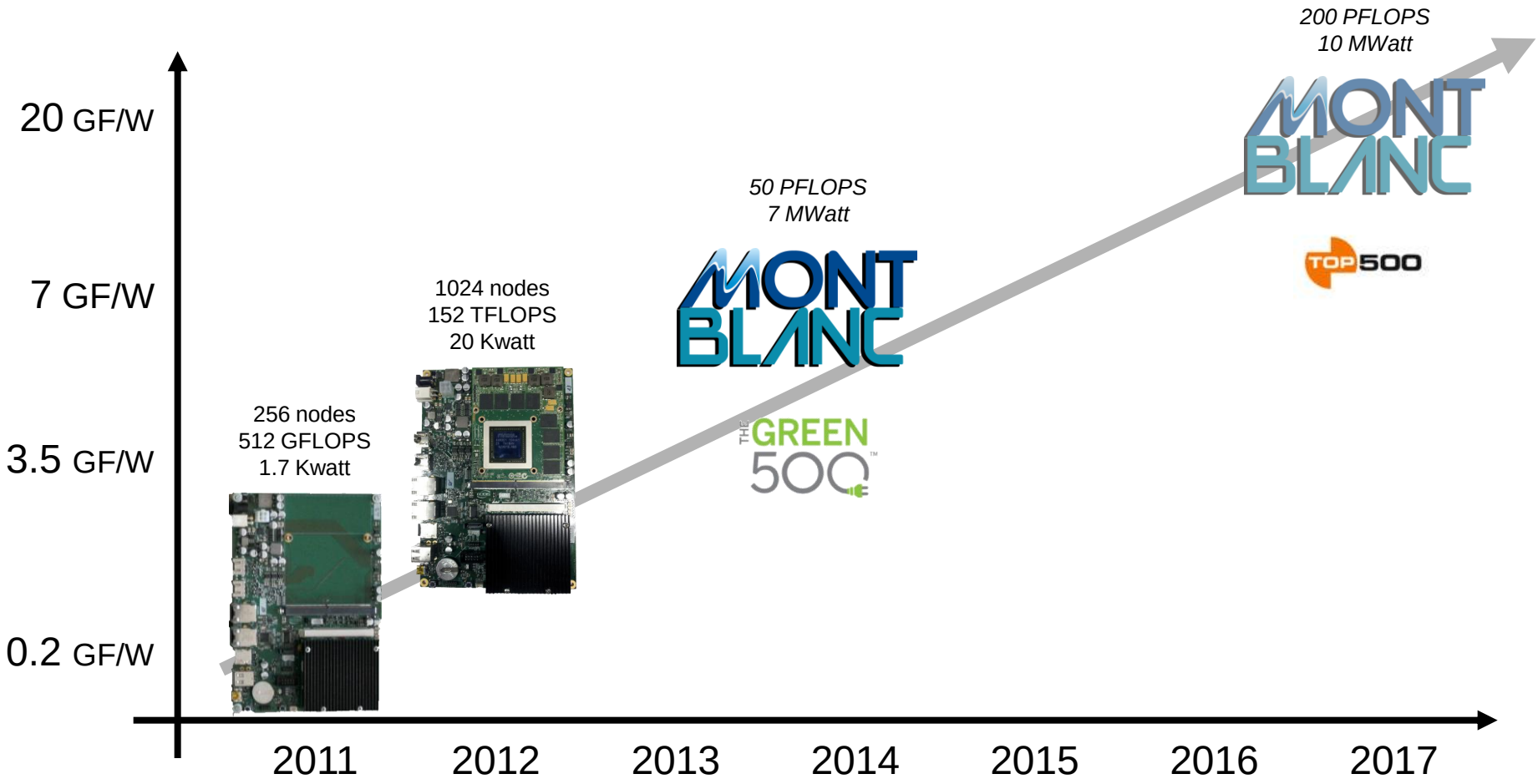
* Estimated from web sources, not an ARM Commitment

Can we build supercomputers from embedded technology?



- HPC used to be the edge of technology
  - First developed and deployed for HPC
  - Then used in servers and workstations
  - Then on PCs
  - Then on mobile and embedded devices
- **Can we close the loop?**

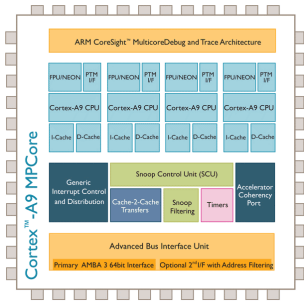
Energy-efficient prototype series @ BSC

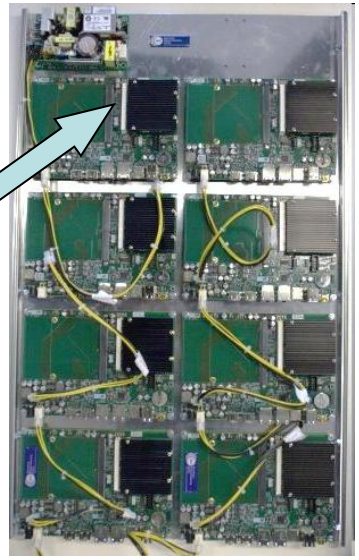


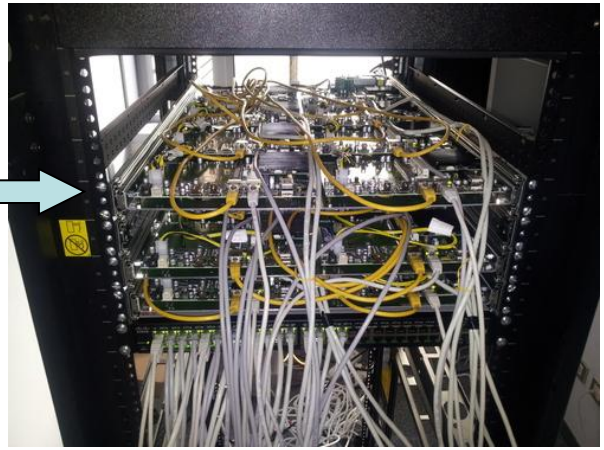
- Start from COTS components
- Move on to integrated systems and custom HPC technology

Tegra2 prototype @ BSC

- Deploy the first large-scale ARM cluster prototype
  - Built entirely from COTS components
- Exploratory system to demonstrate
  - Capability to deploy HPC platform based on low-power components
  - Open-source system software stack
- **Enable early software development and tuning on ARM platform**

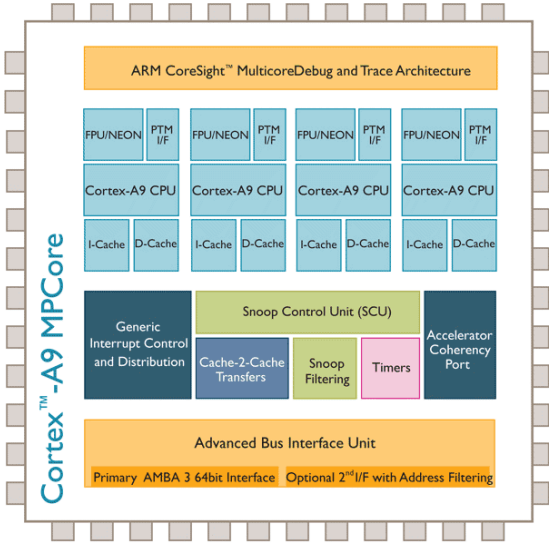






ARM Cortex-A9 multiprocessor

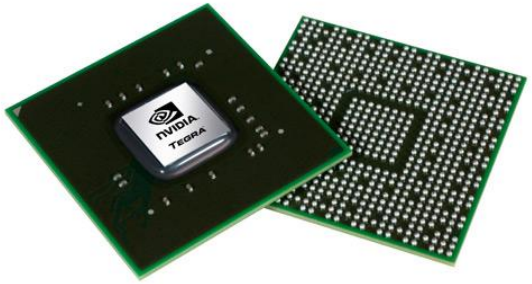
- Energy-efficient application processor
- Up to 4-core SMT
  - Full cache coherency
- VFP Double-Precision FP
  - 1 ops / cycle

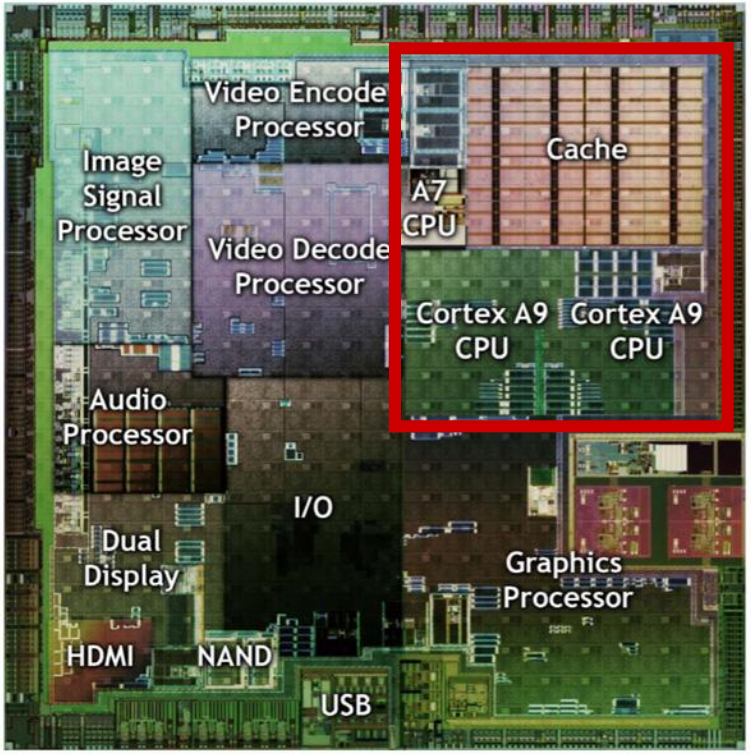


ARM Cortex-A9 Performance Power & Area			
	Cortex-A9 Single Core Soft Macro Trial Implementation	Cortex-A9 Dual Core Hard Macro Implementations	
Process	TSMC 65G	TSMC 40G	
Optimization method	Performance Optimized	Performance Optimized	Power Optimized
Standard Cell Library	ARM SC12	ARM SC12 + High Performance Kit	ARM SC12 + High Performance Kit
Performance (Total DMIPS)	2,075 DMIPS	10,000 DMIPS	4,000 DMIPS
Frequency	830 MHz	2000 MHz (typical)	800 MHz (wc/ss)
Energy Efficiency (DMIPS/mW)	5.2	5.26	8.0
Total power at target frequency	0.4 W	1.9 W	0.5 W
Silicon Area	1.5 mm^2 (excludes caches)	6.7 mm^2 (including L1 parity and all DFT/DFM)	4.6 mm^2 (including all DFT/DFM)

Nvidia Tegra2 SoC

- Dual-core Cortex-A9 @ 1GHz
  - VFP for DP (no NEON)
    - 2 GFLOPS (1 FMA / 2 cycles)
- Low-power Nvidia GPU
  - OpenGL only, CUDA not supported
- Several accelerators
  - Video encoder-decored
  - Audio processor
  - Image processor
- ARM7 core for power management
- 2 GFLOPS ~ 0.5 Watt

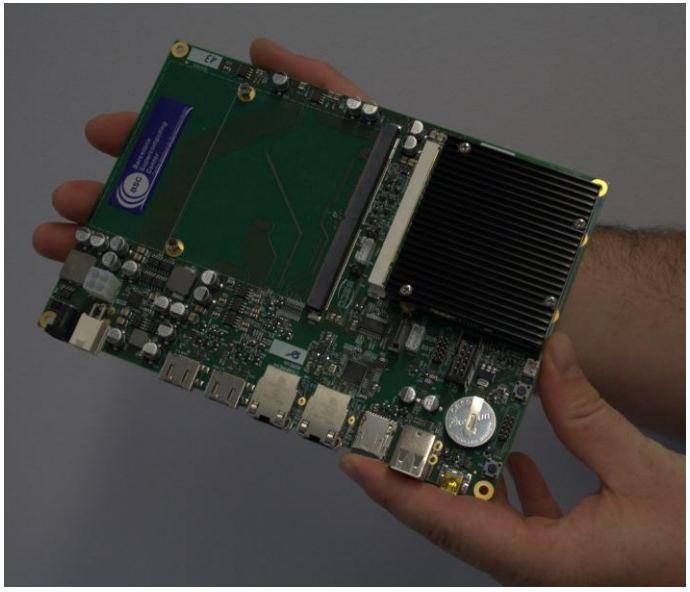




SECO Q7 Tegra2 + Carrier board

- Q7 Module
  - 1x Tegra2 SoC
    - 2x ARM Cortex-A9, 1 GHz
  - 1 GB DDR2 DRAM
  - 100 Mbit Ethernet
  - PCIe
    - 1 GbE
    - MXM connector for mobile GPU
  - 4" x 4"

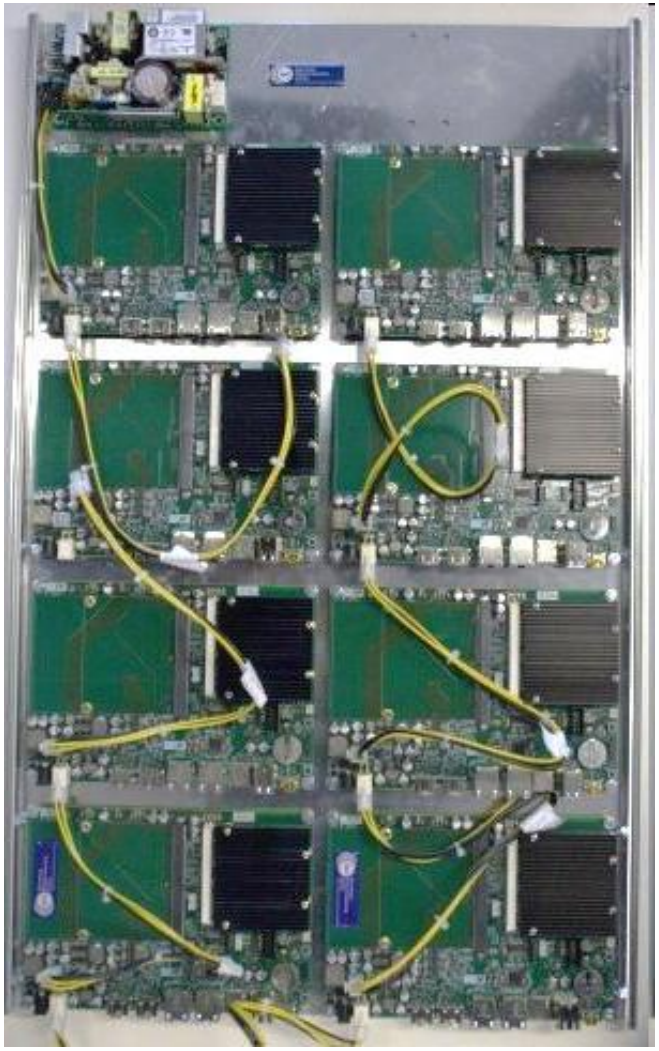
- Q7 carrier board
  - 2 USB ports
  - 2 HDMI
    - 1 from Tegra
    - 1 from GPU
  - uSD slot
  - 8" x 5.6"



- 2 GFLOPS ~ 4 Watt

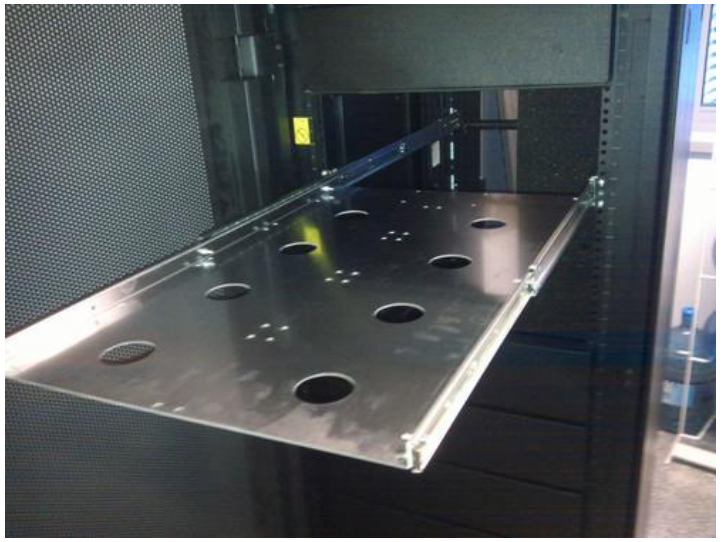
1U multi-board container

- Standard 19" rack dimensions
  - 1.75" (1U) x 19" x 32" deep
- 8x Q7-MXM Carrier boards
  - 8x Tegra2 SoC
    - 16x ARM Cortex-A9
  - 8 GB DRAM
- 1 Power Supply Unit (PSU)
  - Daisy-chaining of boards
  - ~ 7 Watts PSU waste
- 16 GFLOPS ~ 40 Watts

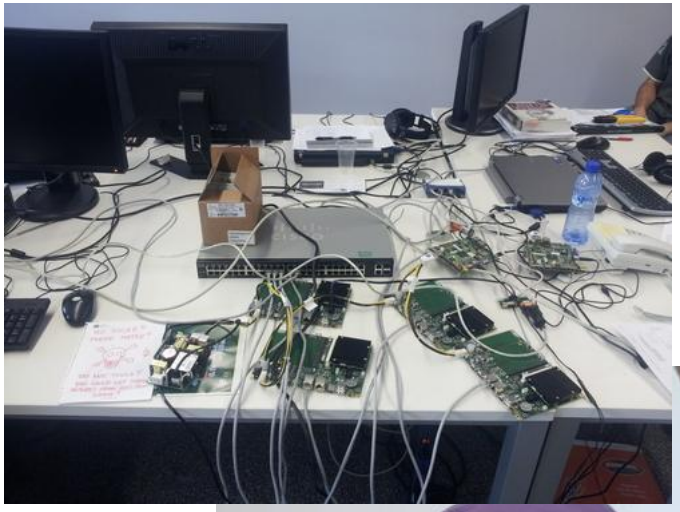


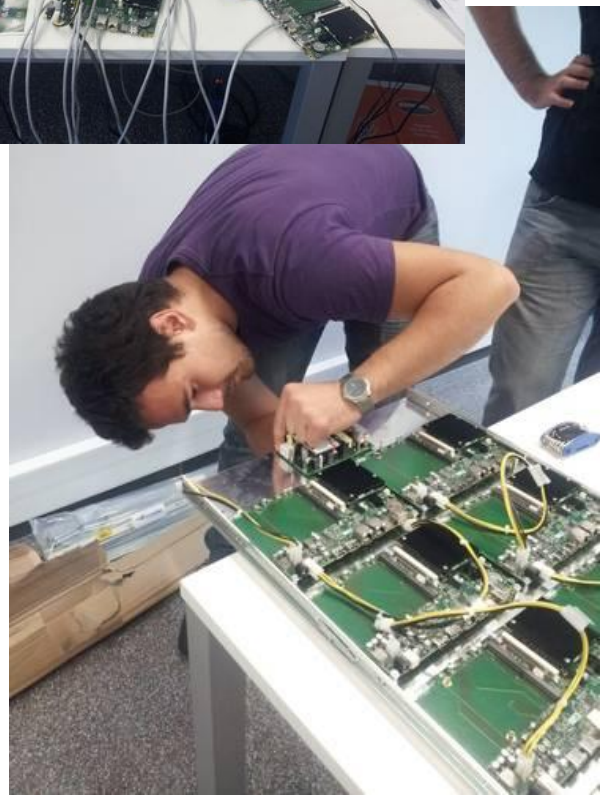
Prototype rack

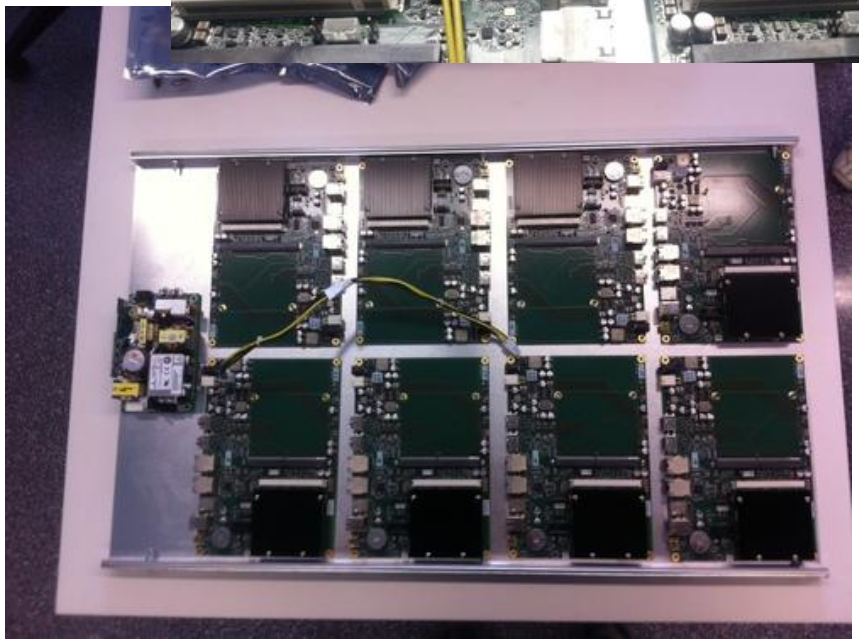
- Stack of 8 x 5U modules
  - 4 Compute nodes
  - 1 Ethernet switch
- Passive cooling
  - Passive heatsink on Q7
- Provide power consumption measurements
  - Per unit
    - Compute nodes
    - Ethernet switches
  - Per container
  - Per 5U
- 512 GFLOPS ~ 1.700 Watt
  - 300 MFLOPS / W
  - 60% efficiency ~ 180 MFLOPS / W



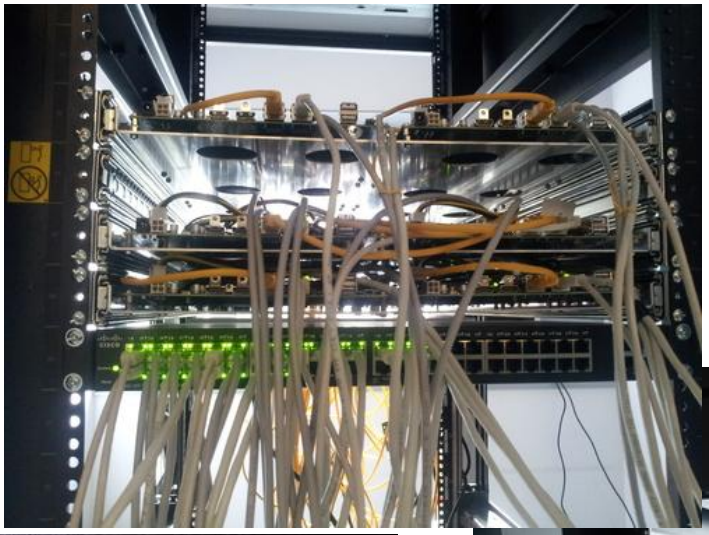
Manual assembly of board container

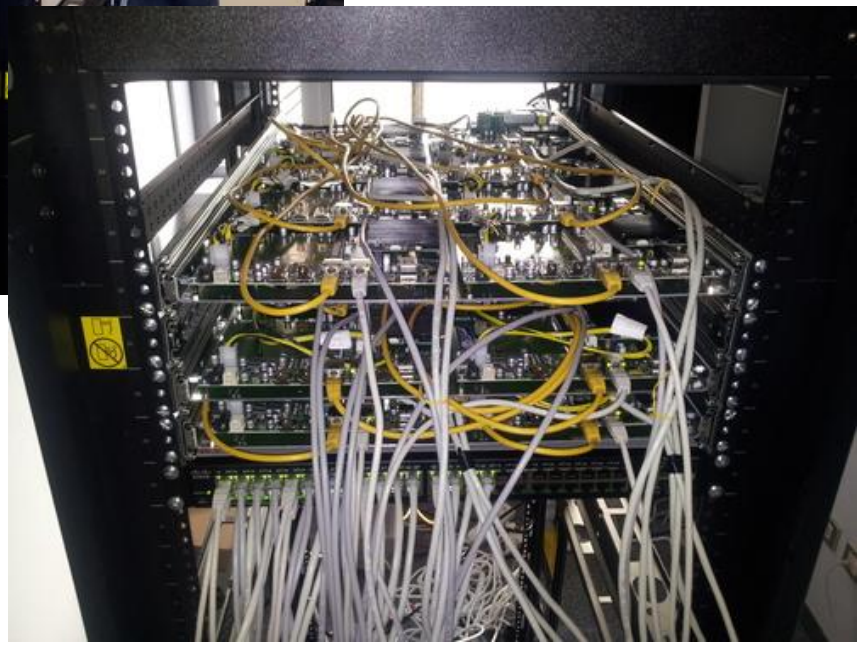


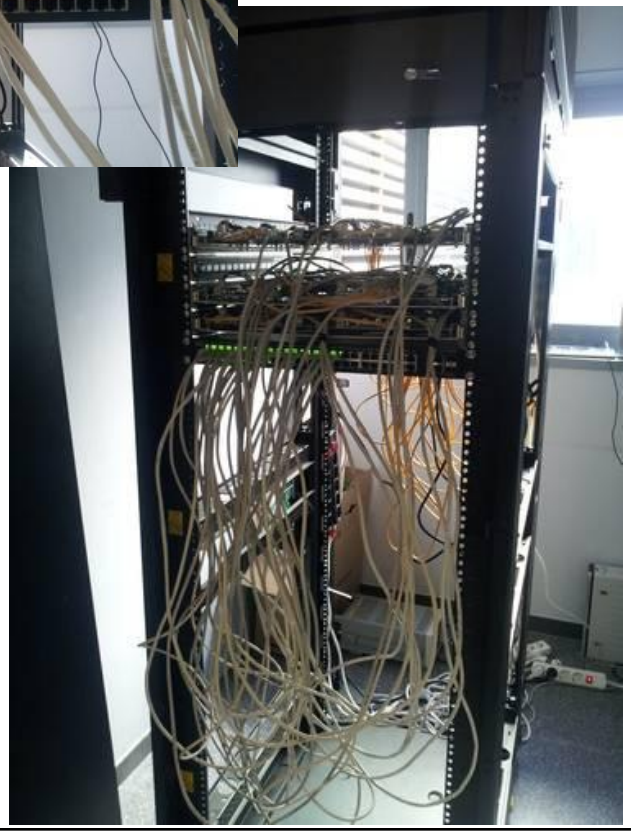




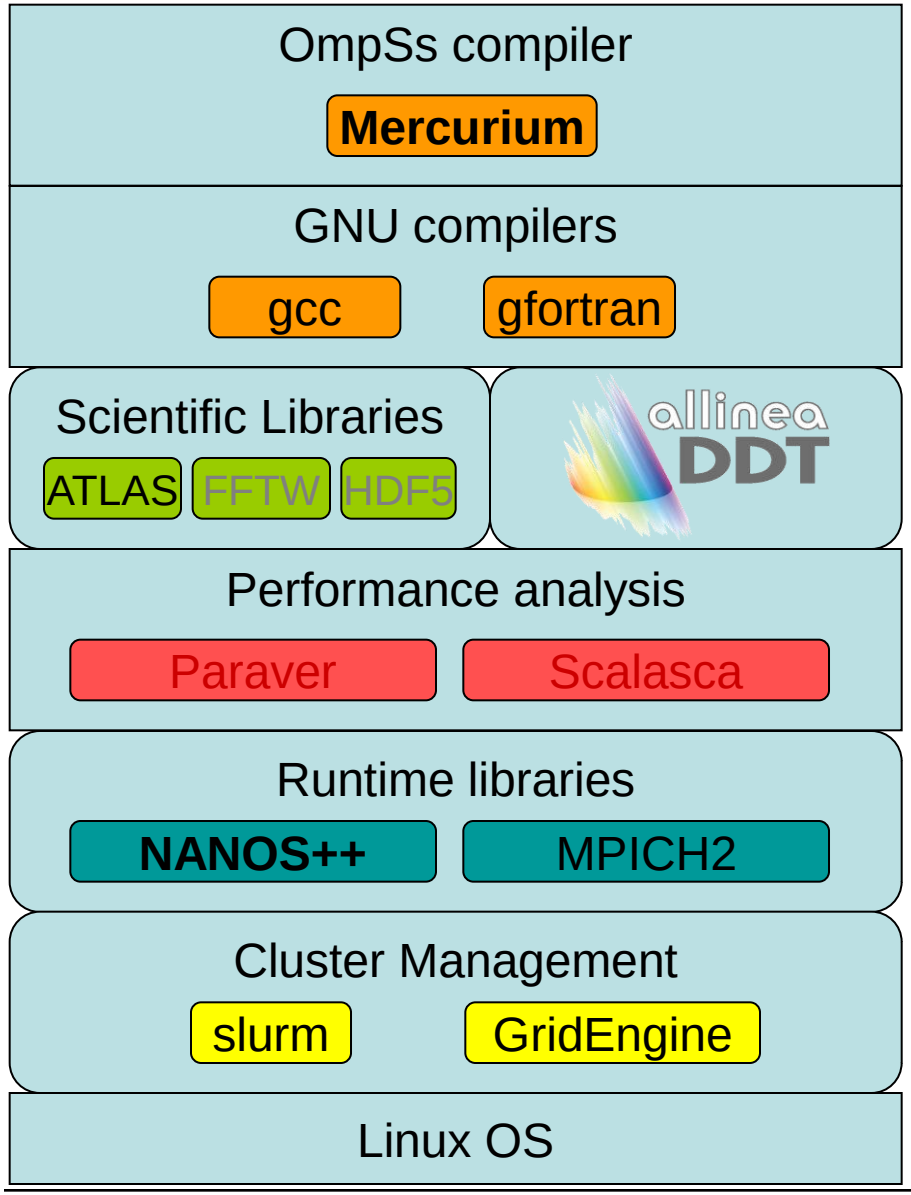
Manual assembly of containers in the rack + interconnect wiring





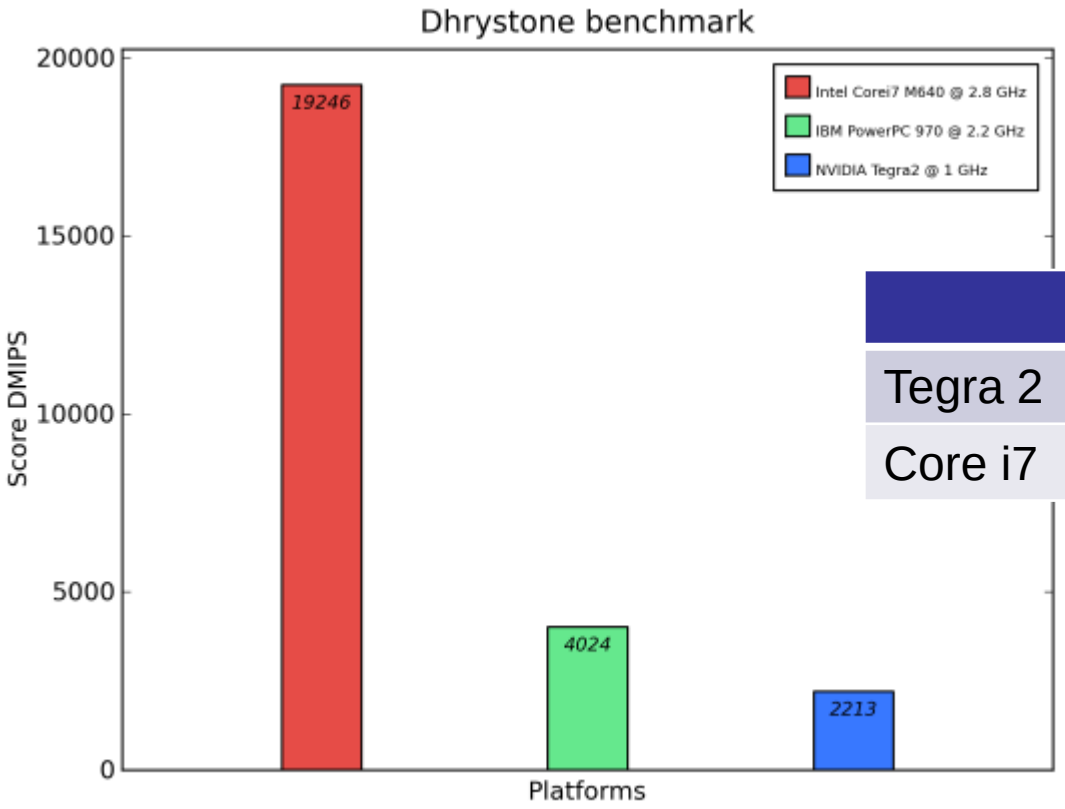


System software stack



- Open source system software stack
  - Linux OS
  - GNU compiler
    - gcc 2.4.6
    - gfortran
  - Scientific libraries
    - ATLAS, FFTW, HDF5
  - Cluster management
- Runtime libraries
  - MPICH2, OpenMPI
  - **OmpSs toolchain**
- Performance analysis tools
  - Paraver, Scalasca
- Allinea DDT 3.1 debugger

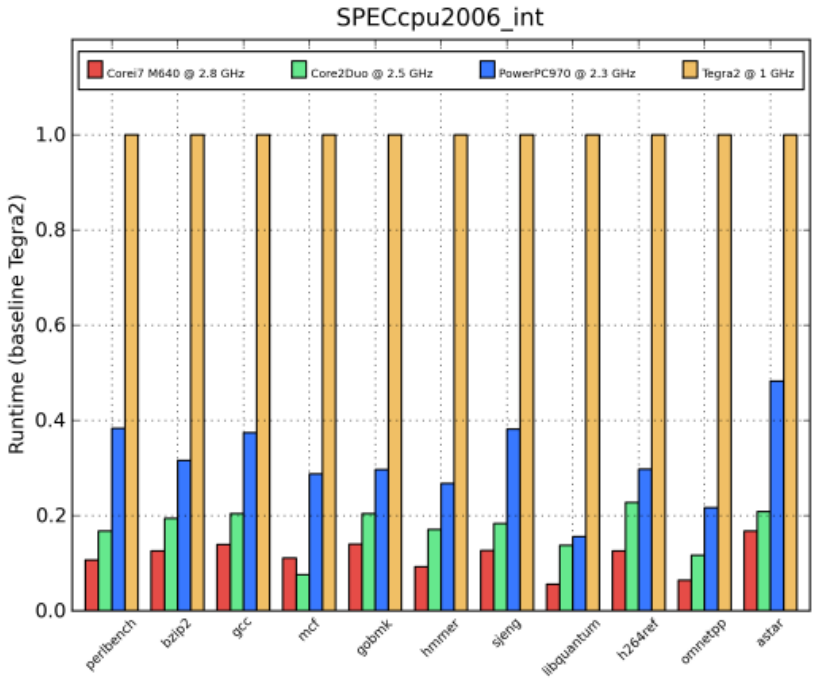
Processor performance: Dhrystone

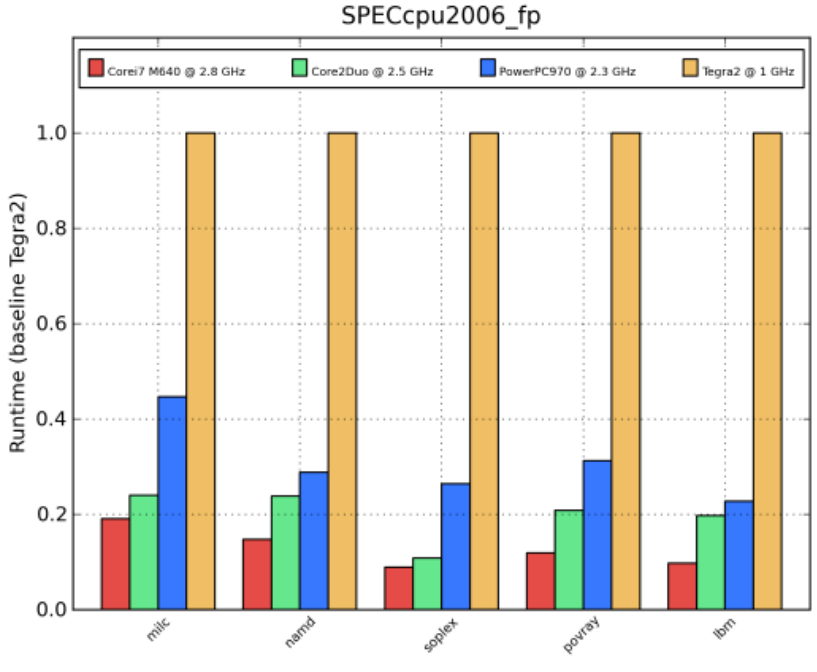


	Energy (J)	Normalized
Tegra 2	110.6	1.0
Core i7	116.8	1.056

- Validate if Cortex-A9 achieves the ARM advertised Dhrystone performance
  - 2.500 DMIPS / GHz
- Compare to PowerPC 970MP (JS21, MareNostrum) and Core i7 (laptop)
  - ~ 2x slower than ppc970
  - ~ 9x slower than i7

Processor performance: SPEC CPU 2006

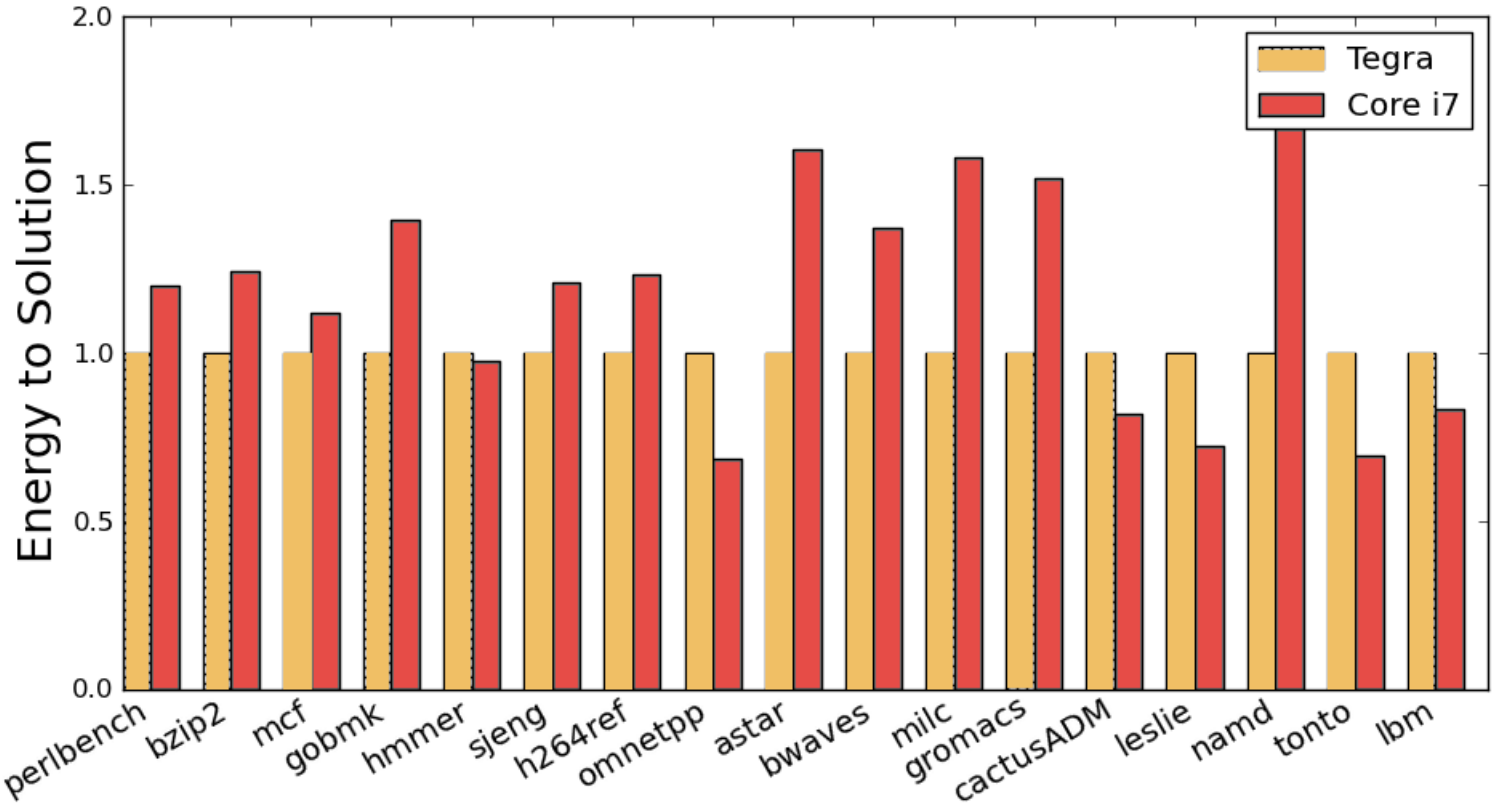




- Compare Cortex-A9 @ 1 GHz CPU performance with 3 platforms
  - ppc970 @ 2.3 GHz ~ 2-3x slower (= if we factor in freq.)
  - Core2 @ 2.5 GHz ~ 5-6x slower
  - Core i7 @ 2.8 GHz ~ 6-10x slower (2-4x slower if we factor freq.)

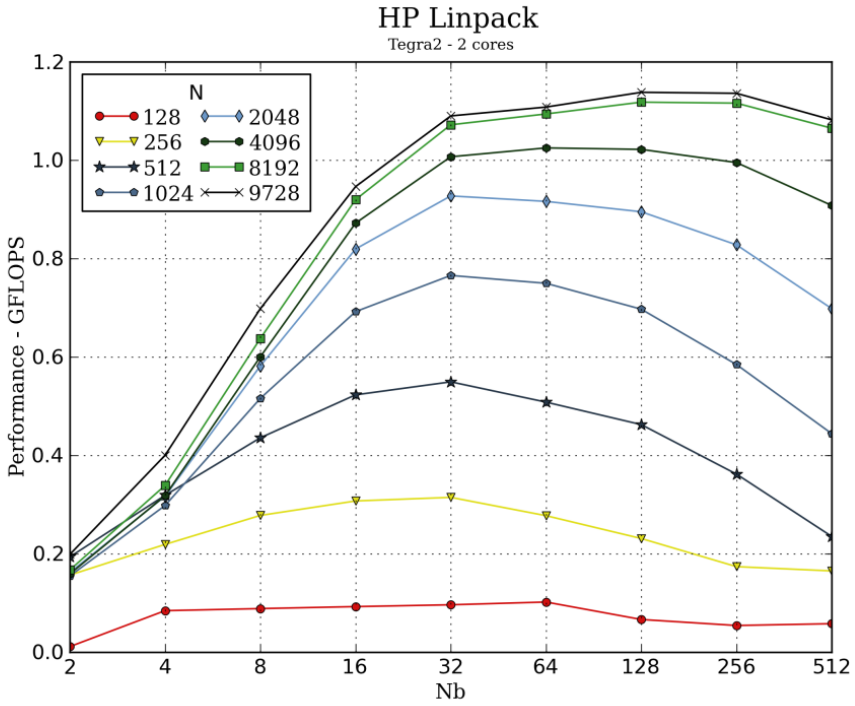
- Is it more power efficient?

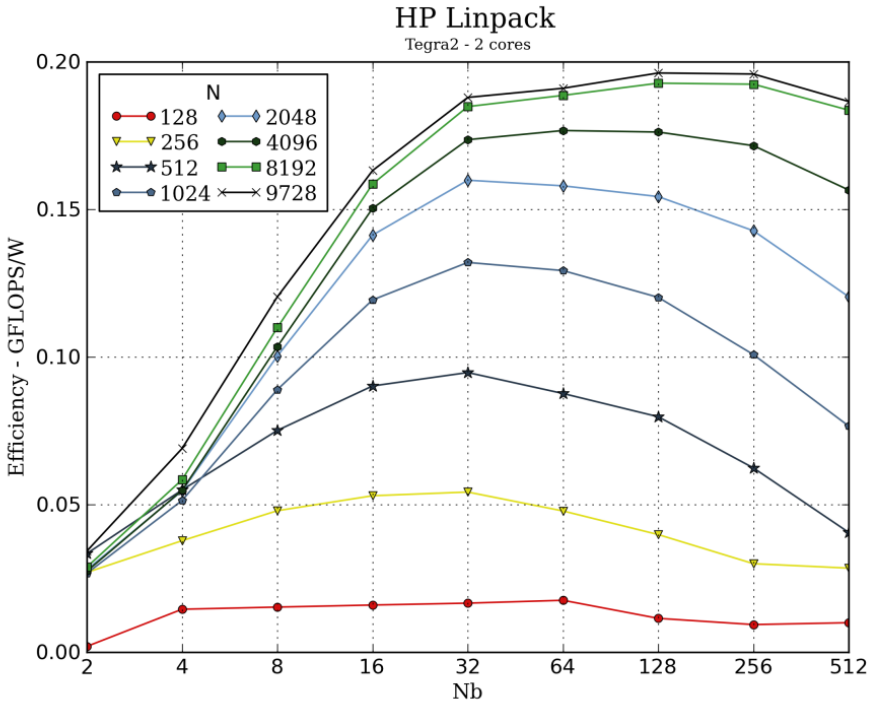
Energy to solution: SPEC CPU 2006



- Tegra2 not always more power-efficient than Core i7
  - i7 efficiency is better for benchmarks where it outperforms A9 by 10x

Node performance: Linpack



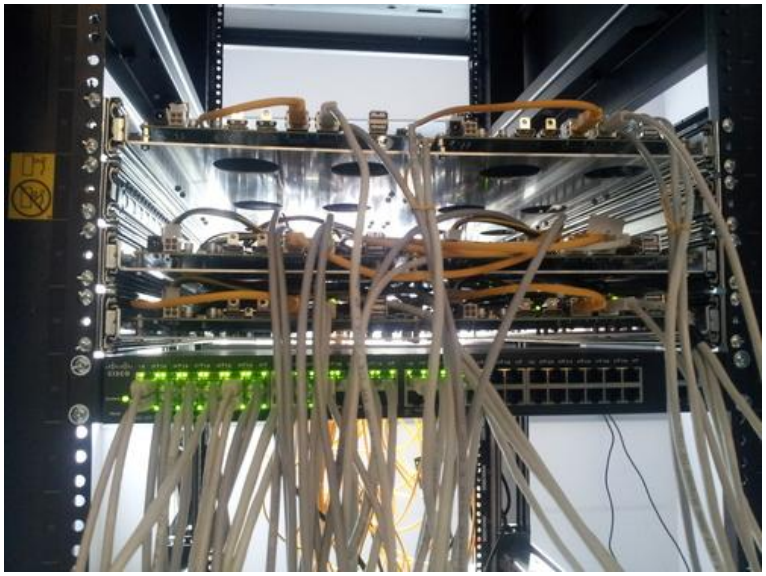


- Standard HPL, using ATLAS library
  - ATLAS microkernels also achieve 1 GFLOPS peak performance
- 1.15 GFLOPS ~ 57% efficiency vs. peak performance
- ~200 MFLOPS / Watt
  - In line with original predictions

Cluster performance: Linpack

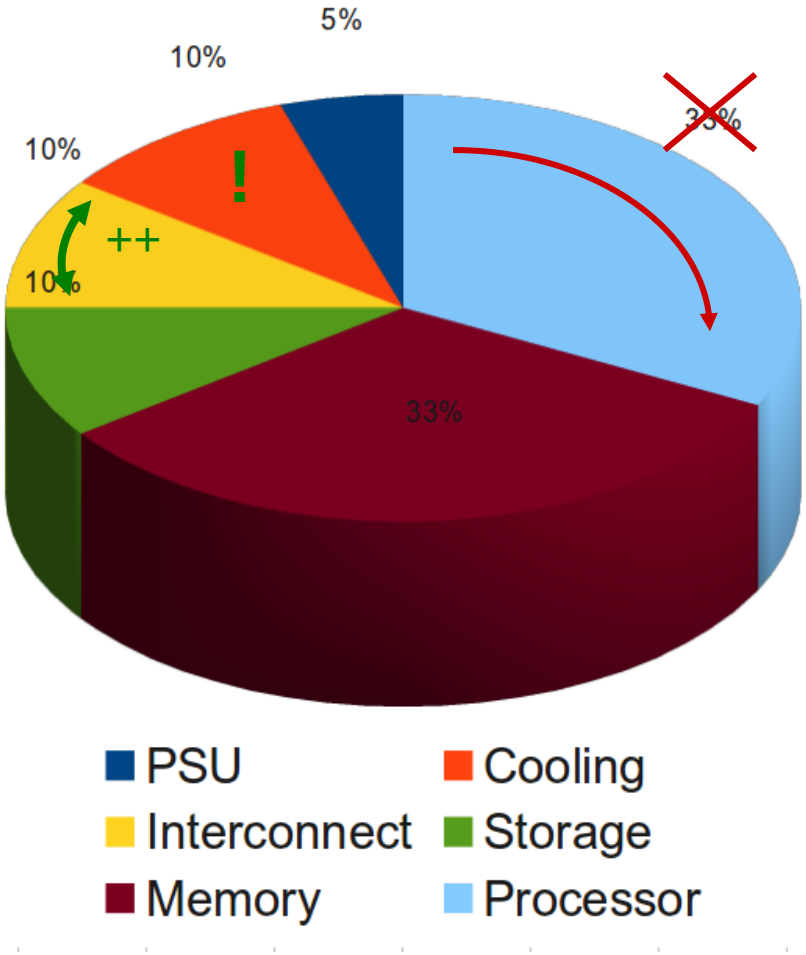
- 24 nodes
  - 3 x 8 boards (48 GFLOPS peak)
  - 1 GbE switch
- 27.25 GFLOPS on 272 Watts
  - 57% efficiency vs. peak
  - 100 GFLOPS / Watt
- Small problem size (N)
  - 280 MB / node
- Power dominated by GbE switch
  - 40 W when idle, 100-150 W active

- 32 nodes
  - 4 x 8 boards (64 GFLOPS peak)
  - 1 GbE switch
- … runs don't complete due to boards overheating
  - Boards too close together
  - No space for airflow



Lessons learned so far

- Memory + interconnect dominates power consumption
  - Need a balanced system design
- Tuning scientific libraries takes time + effort
  - Compiling ATLAS on ARM Cortex-A9 took 1 month
- Linux on ARM needs tuning for HPC
  - CFS scheduler
  - softfp vs. hardfp
- DIY assembly of prototypes is harder than expected
  - 2 Person-Month just to press screws
- Even low-power devices need cooling
  - It's the density that matters



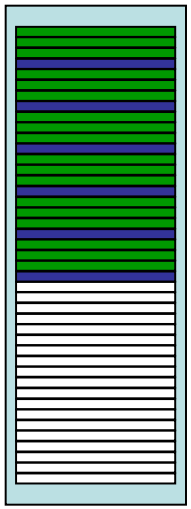
ARM + mobile GPU prototype @ BSC

Tegra3 + GeForce 520MX:

4x Corext-A9 @ 1.5 GHz
48 CUDA cores @ 900 MHz

148 GFLOPS ~ 18 Watts

~ 8 GFLOPS / W



Rack:
32x Board container
256x Q7 carrier boards
  1024x ARM Corext-A9 Cores
  256x GT520MX GPU
8x 48-port 1GbE LBA switches

38 TFLOPS ~ 5 Kwatt
7.5 GFLOPS / W

50% efficiency
3.7 GFLOPS / W

- Validate the use of energy efficient CPU + compute accelerators
  - ARM multicore processors
  - Mobile Nvidia GPU accelerators
- Perform scalability tests to high number of compute nodes
  - Higher core count required when using low-power components
  - Evaluate impact of limited memory and bandwidth on low-end solutions
- Enable early application and runtime system development on ARM + GPU

What comes next?

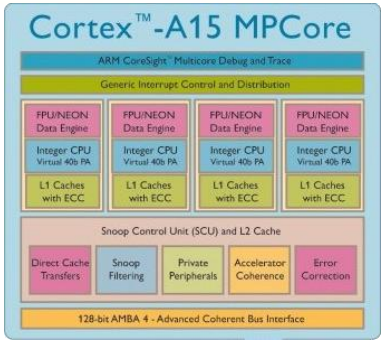
http://www.montblanc-project.eu

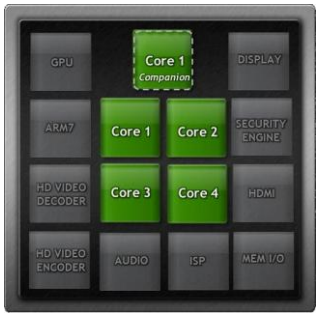
European Exascale approach
using embedded **power-efficient technology**

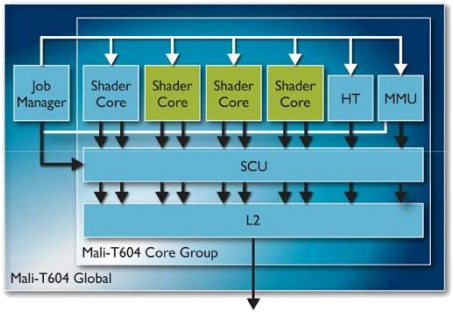
1. To deploy a **prototype HPC system** based on **currently available** energy-efficient embedded technology
2. To design a next-generation HPC system and **new embedded technologies targeting HPC systems** that would overcome most of the limitations encountered in the prototype system
3. To port and optimise a small number of **representative exascale applications** capable of exploiting this new generation of HPC systems

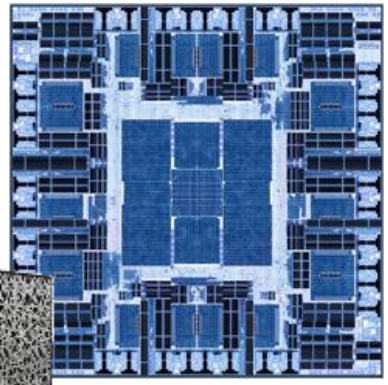
Integrate energy-efficient building blocks

- Integrated system design built from mobile / embedded components
- ARM multicore processors
  - Nvidia Tegra / Denver, Calxeda, Marvell Armada, ST-Ericsson Nova A9600, TI OMAP 5, …
- Low-power memories
- 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



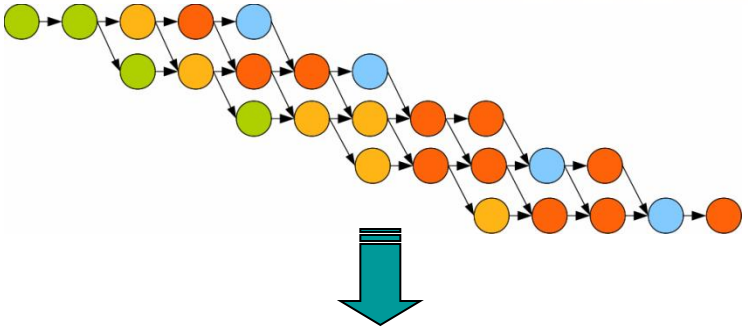


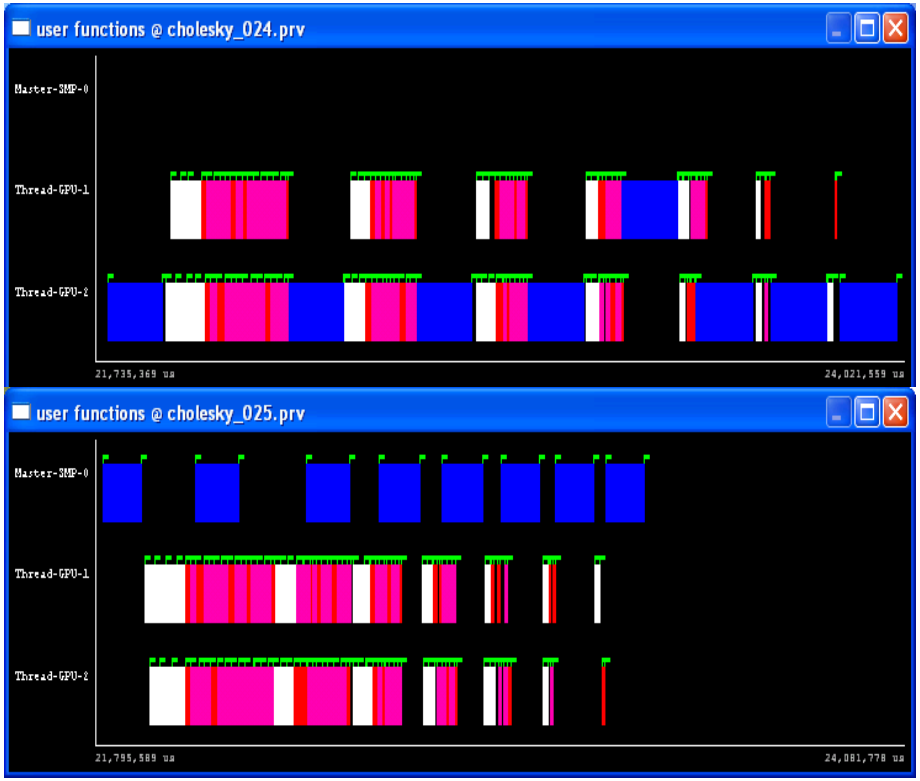




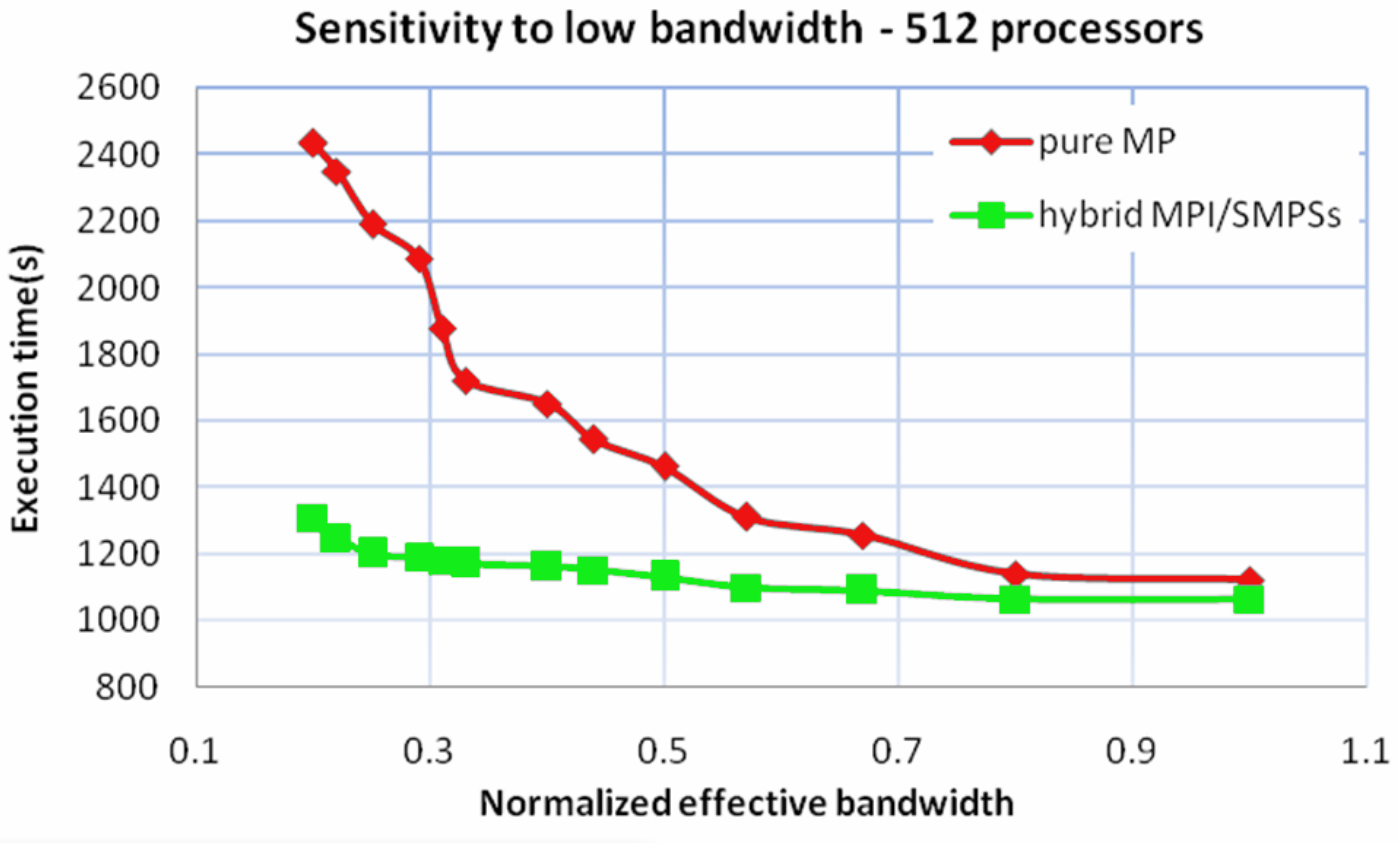
Hybrid MPI + OmpSs programming model

- Hide complexity from programmer
- Runtime system maps task graph to architecture
- Automatically performs optimizations
  - 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



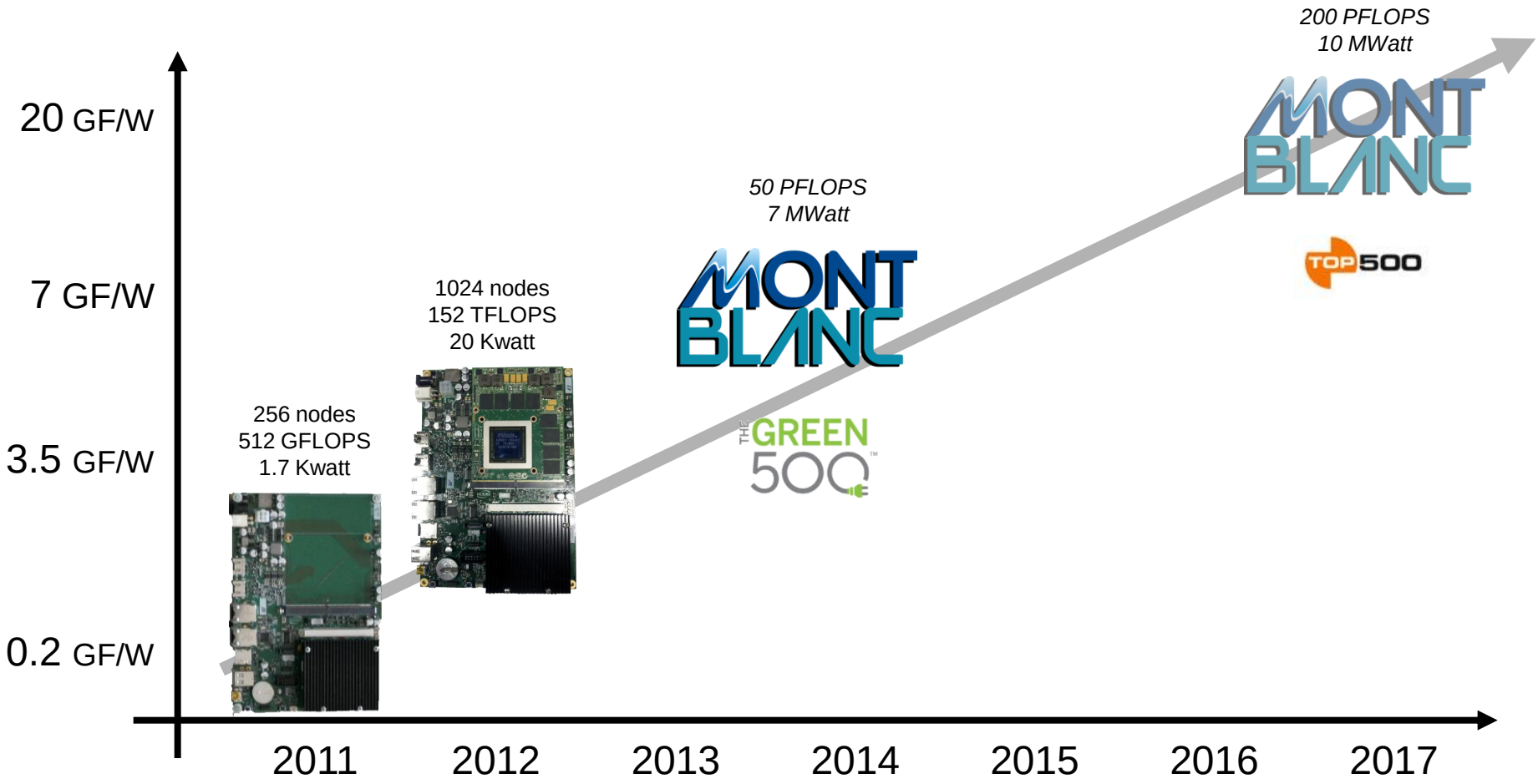


Trade off bandwidth for power in the interconnect



- Hybrid MPI + SMPSs Linpack on 512 processors
- 1/5th the interconnect bandwidth, only 10% performance impact
- Rely on slower, but more efficient network?

Energy-efficient prototype series @ BSC



- A very exciting roadmap ahead
- Lots of challenges, both hardware and software!