Supercomputing based on mobile processors

INFN – CNAF

Bologna, 9 Jan 2014

Filippo Mantovani
filippo.mantovani@bsc.es

Supercomputing based on mobile processors

Is it really possible?!?

INFN – CNAF

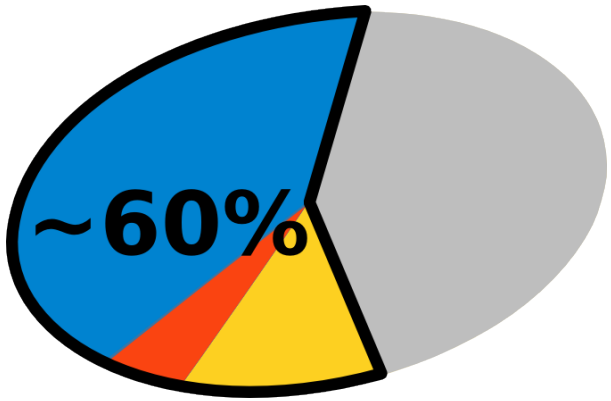
Bologna, 9 Jan 2014

Filippo Mantovani
filippo.mantovani@bsc.es

Once upon a time...

First Top500 list (June 1993) dominated by data-level (vector/SIMD) parallel architectures:

- Cray vector, 41%
- MasPar SIMD, 11%
- Convex/HP vector, 5%
- ...



This trend was pretty stable: Fujitsu Wind Tunnel was #1 in 1993-1995.

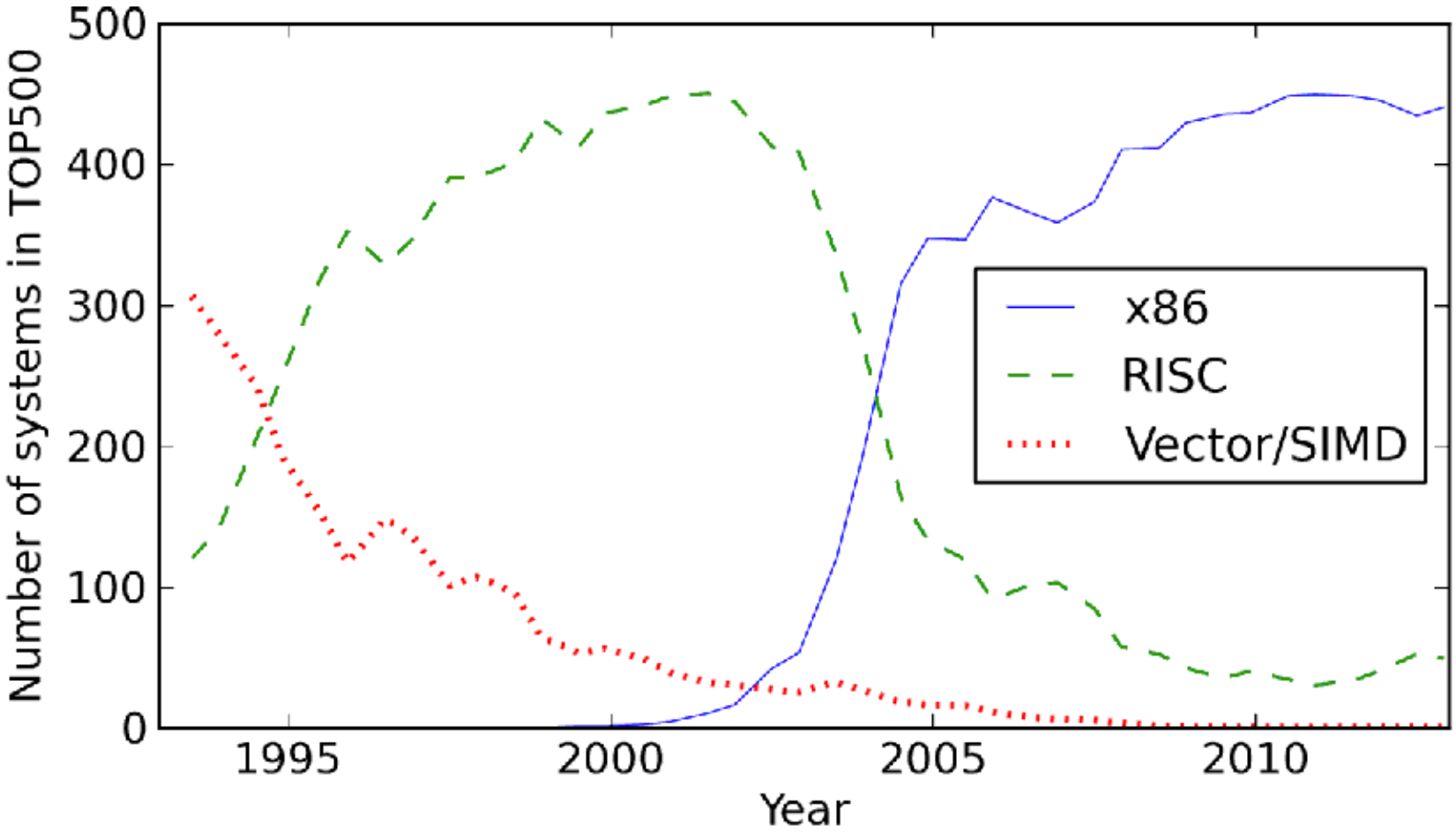
Then, commodity took over special purpose

1997 - ASCI Red, Sandia:
1 TFLOPS, 9298 cores @ 200 MHz Intel Pentium Pro, upgraded to Pentium II Xeon in 1999 (3.1 TFLOPS).

2001 - ASCI White, LLNL:
7.3 TFLOPS, 8192 cores @ 375 Mhz, IBM Power 3.

From **vector parallelism** to **message passing** programming models...

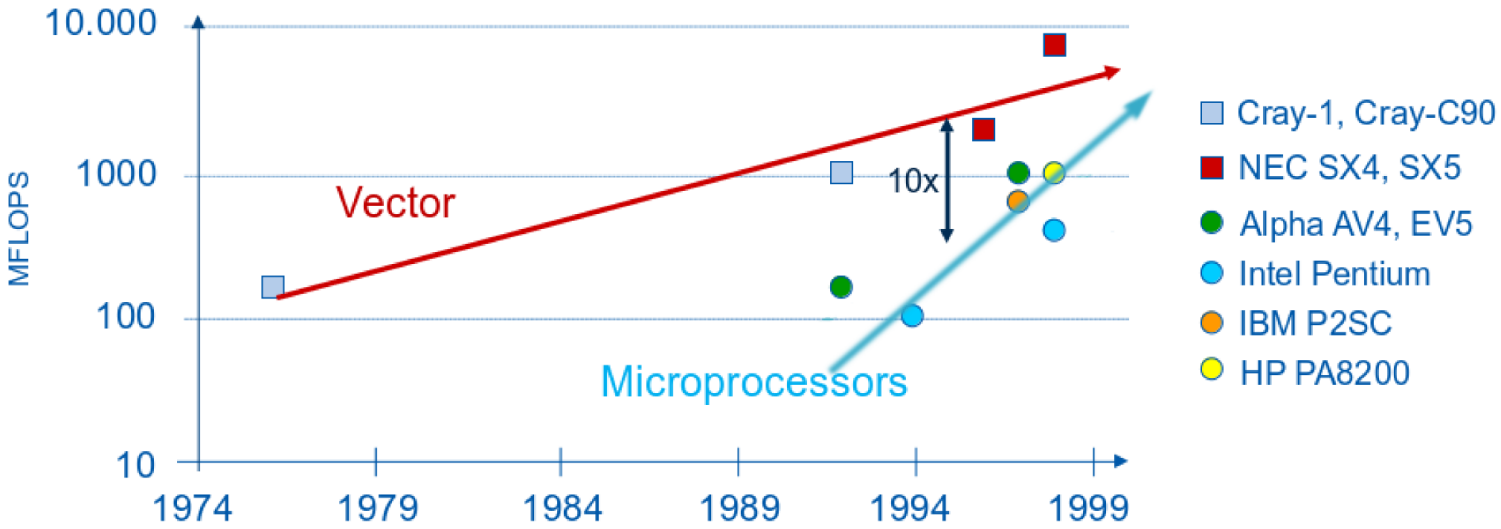
And now commodity components drive HPC



- ➜ RISC processors replaced vectors
- ➜ x86 processors replaced RISC

  Vector processors survive as (widening) SIMD extensions

The killer microprocessors



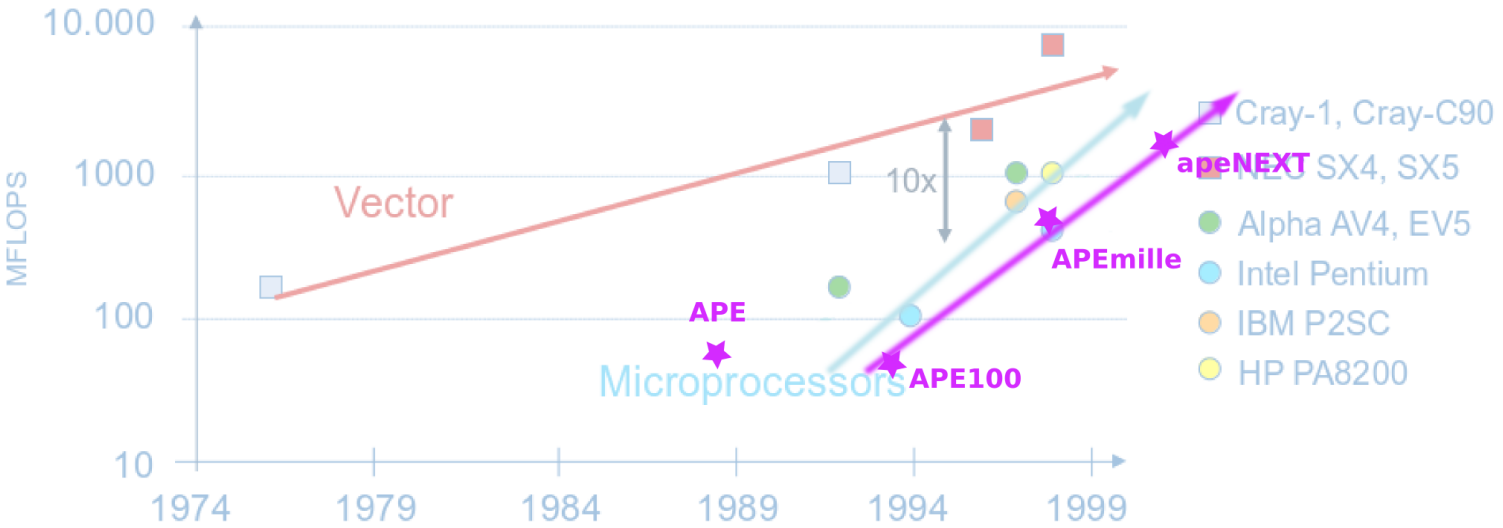
- ➜ Microprocessors killed the vector-based supercomputers.

  They were not faster, but they were significantly cheaper and greener.

- ➜ Need 10 microprocessors to achieve the performance of a single vector CPU.

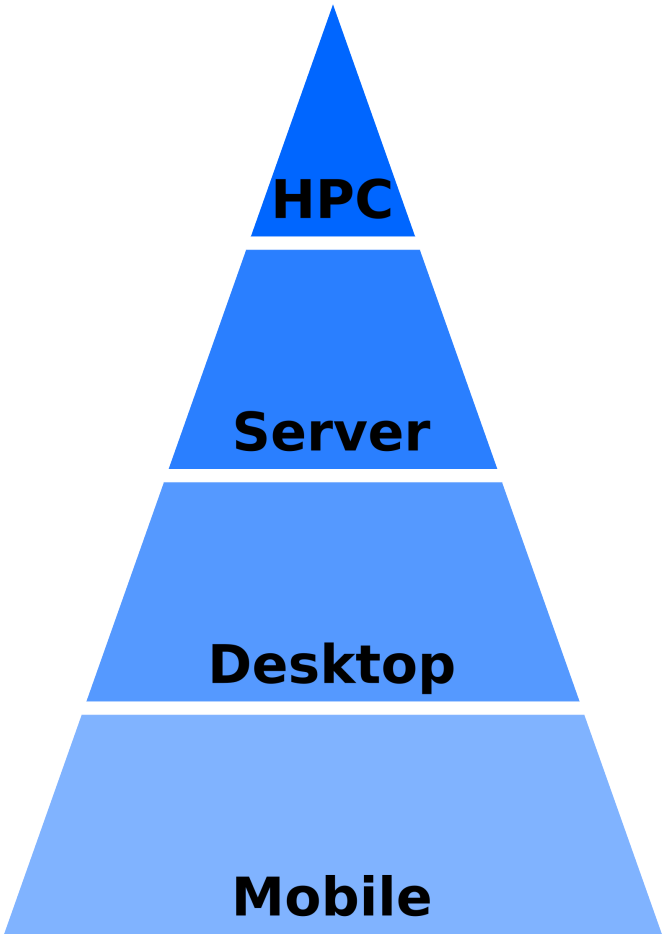
  SIMD vs. MIMD programming paradigms.

Maybe some of you are part of the story...



- ➜ Scaling faster than vector but slower than microprocessors
- ➜ The true lesson: only the fastest processors are good for HPC systems

What's "commodity" nowadays?



➜ ∼21M cores in Nov'13 Top500

Sold in 2012:

➜ >10M servers

➜ >350M PC's

➜ >100M tablets

➜ >700M smartphones

>210M smartphones (1stQ 2013) and counting...

SoCs under study

NVIDIA Tegra 2 – 2011

2 x ARM Cortex-A9 @ 1GHz

1 x 32-bit DDR2-333 channel

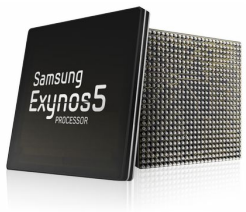
32KB L1 + 1MB L2

NVIDIA Tegra 3 – 2012

4 x ARM Cortex-A9 @ 1.3GHz

1 x 32-bit DDR3-750 channel

32KB L1 + 1MB L2

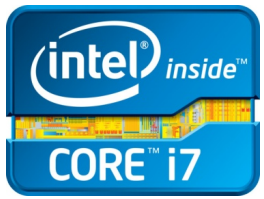


Samsung Exynos 5 – 2012

2 x ARM Cortex-A15 @ 1.7GHz

2 x 32-bit DDR3-800 channels

32KB L1 + 1MB L2



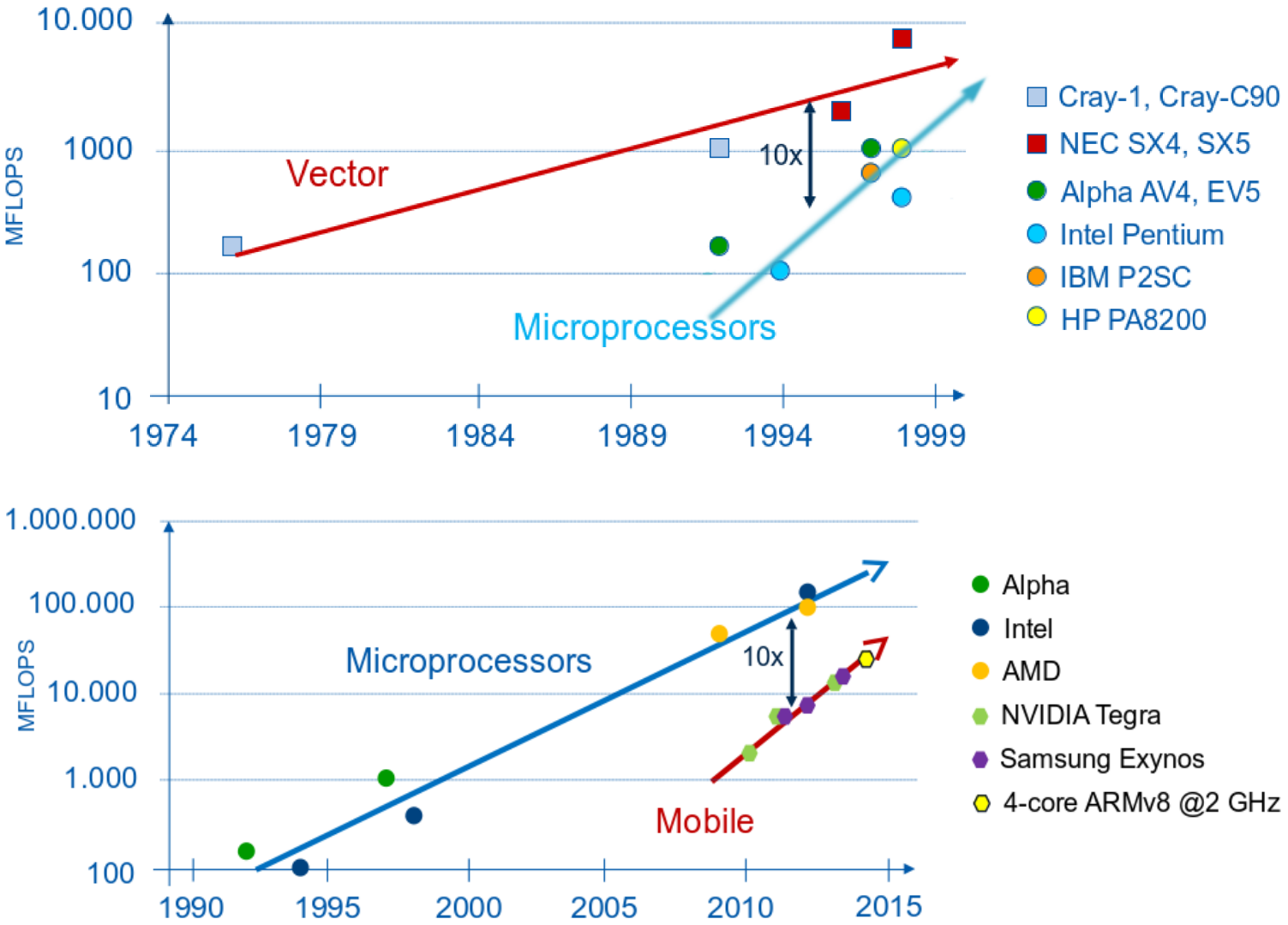
Intel Core i7-2760QM – 2012

4 x Intel SandyBridge @ 2.4GHz

2 x 64-bit DDR3-800 channels

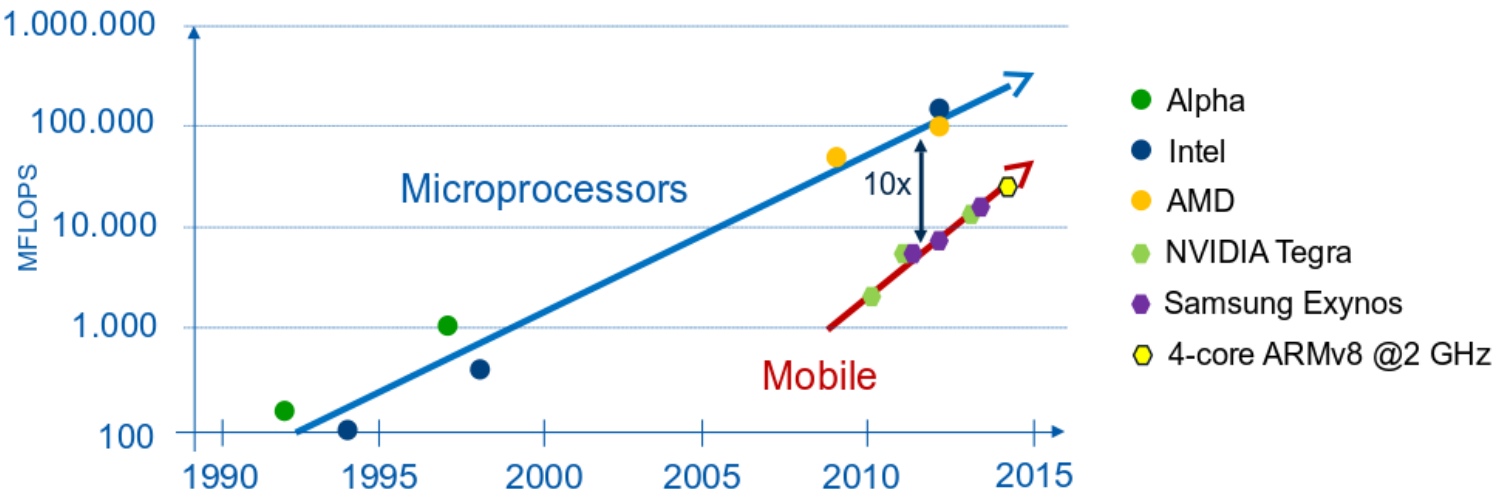
32KB L1 + 1MB L2 + 6MB L3

History may be about to repeat itself



History may be about to repeat itself

- In 2013, mobile SoCs are slower...
  but performance gap seems to close pretty fast.
- SoCs are significantly cheaper due to volume and (maybe) less power greedy.



Mobile SoC vs Server - side by side

153 GFLOPS — 1500$

30x

5.2 GFLOPS — 21$

70x

10x

15.2 GFLOPS — ~20$

70x ?

5x

7+25 GFLOPS — ~20$

Mont-Blanc project overview

- To develop an **European** Exascale approach
- Leverage commodity and **embedded power-efficient** technology


Supported by EU FP7 with 16M€ under two projects:

- **Mont-Blanc:** October 2011 - September 2014
  14.5 M€ budget (8.1 M€ EC contribution), 1095 Person-Month
- **Mont-Blanc 2:** October 2013 - September 2016
  11.3 M€ budget (8.0 M€ EC contribution), 892 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 in 2014
- ➜ Deploy a full HPC system software stack

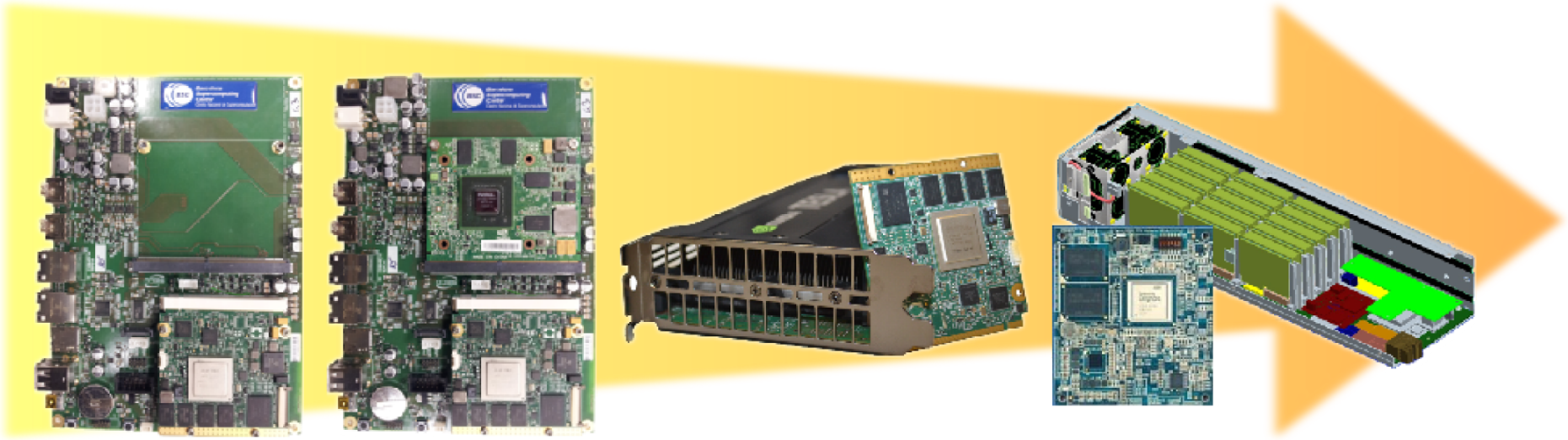
Objective 2: To design a next-generation HPC system and new embedded technologies targeting HPC systems that would overcome most of the limitations encountered in the prototype system

- ➜ 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 exploiting this new generation of HPC systems

- ➜ Up to 11 full-scale applications

ARM-based prototypes at BSC

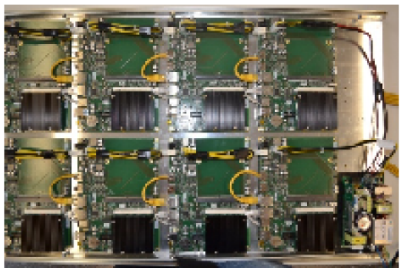


2011	2012	2013	2014
Tibidabo	Kayla	Pedraforca	Mont-Blanc
ARM multicore	ARM + GPU CUDA on ARM	ARM + GPU Inifinband	ARM + GPU (custom board) OpenCL on ARM GPU

Tibidabo: first ARM HPC multicore cluster

Q7 Tegra 2 Module
2 x Cortex-A9 @ 1GHz
2 GFLOPS
5 Watts (?)
0.4 GFLOPS / W

Q7 carrier board
2 x Cortex-A9
2 GFLOPS
1 GbE + 100 MbE
7 Watts
0.3 GFLOPS / W



1U Rackable blade
8 nodes
16 GFLOPS
65 Watts
0.25 GFLOPS / W

2 Racks
32 blade containers
256 nodes
512 cores
10x 48-port 1GbE switch
8x 48-port 100 MbE switch

512 GFLOPS
3.4 Kwatt
0.15 GFLOPS / W

First entry
Green500
Nov'11:
2.0 GFLOPS / W

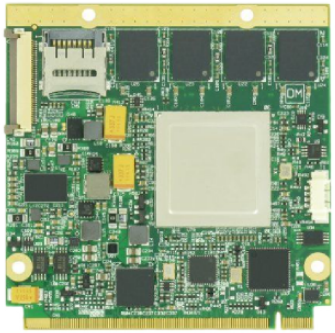
➜ Proof of concept:

‘‘It is possible to deploy a cluster of smartphone processors’’

➜ Enable software stack development.

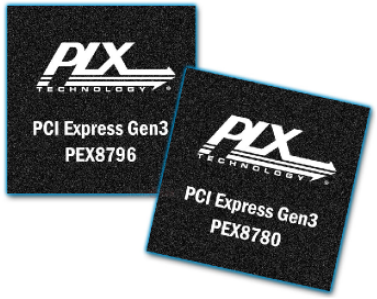
Pedraforca

Tegra 3 Q7 Module
4x ARM Cortex A9 @ 1.3 GHz
2GB DDR3



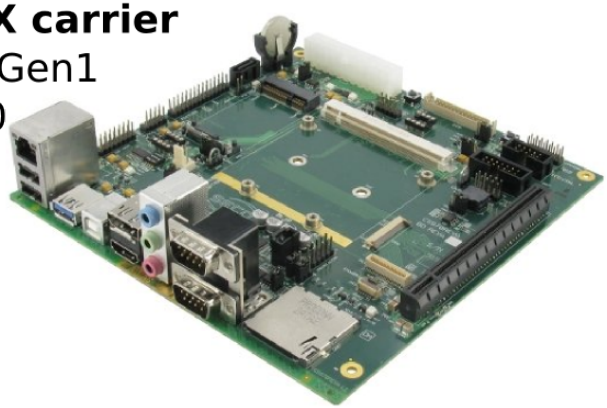
2.5'' SSD
250GB SATA 3

PCIe switch PLX 8632
32 lane Gen2
12 port

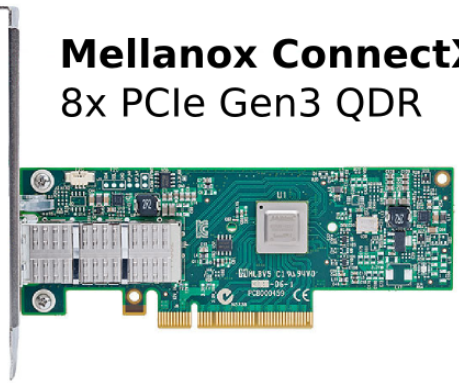


NVIDIA K20
16x PCIe Gen3
1170 GFLOPS
(peak)

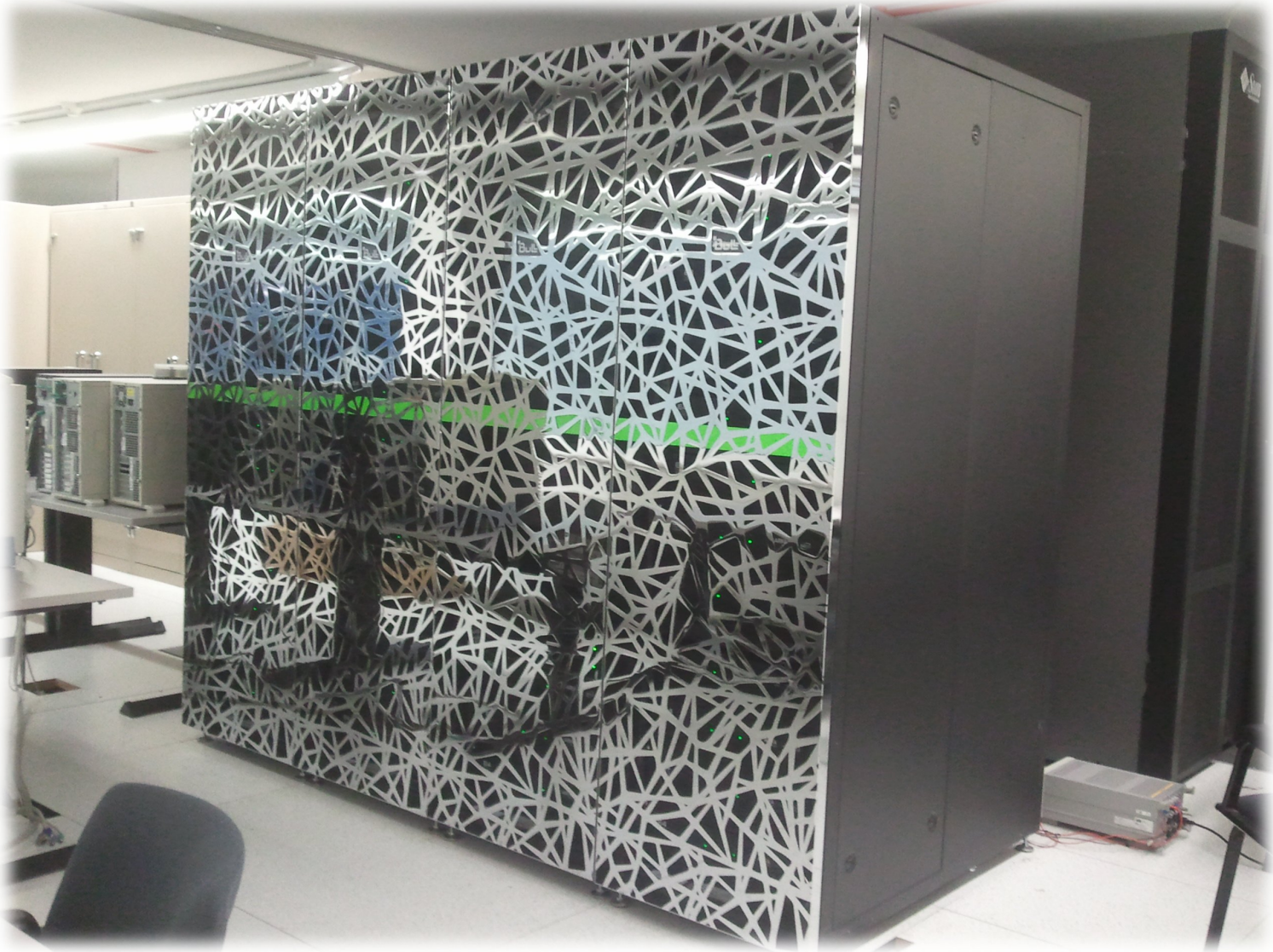
Mini-ITX carrier
4x PCIe Gen1
SATA 2.0
1 GbE



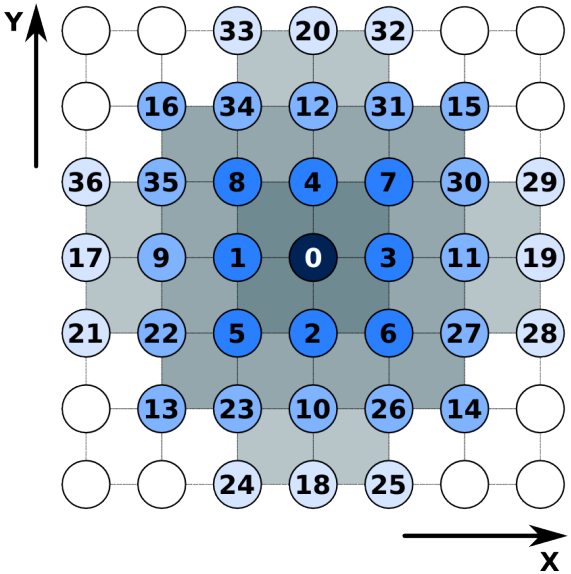
Mellanox ConnectX-3
8x PCIe Gen3 QDR



Pedraforca



Lattice Boltzmann on Pedraforca



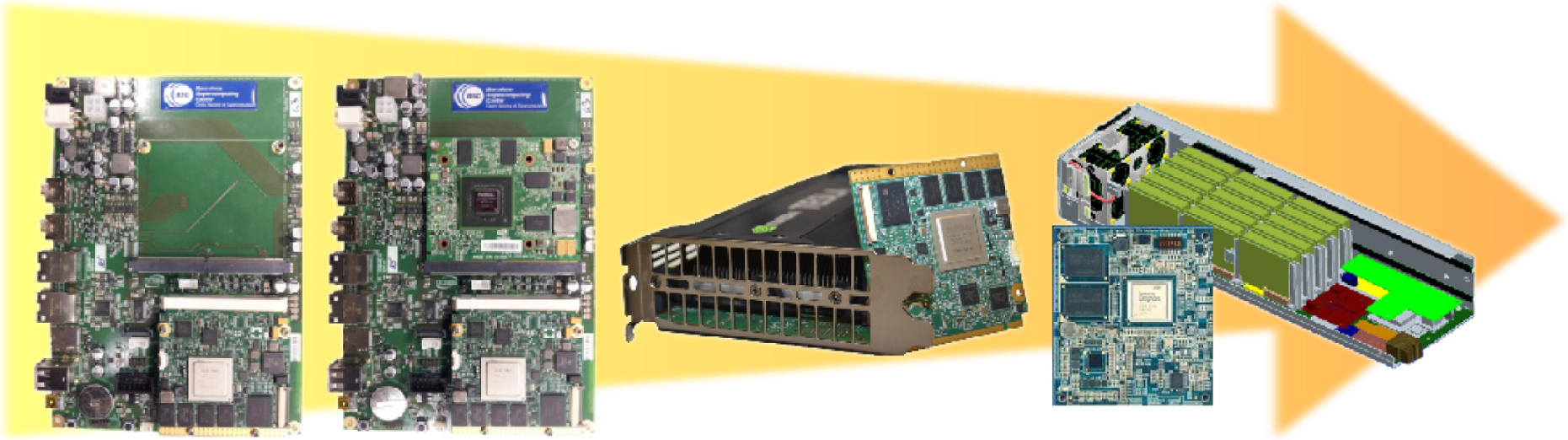
Fluid dynamics simulations evolving a 2D array of particles (double) interacting with their third neighbours.

This translates in a pretty regular pattern of pure floating point computation "**collide**" and memory accesses "**propagate**".

	Propagate				Collide			
Machine	Power [W]	Performance [GB/s]	Perf/Power [GB/J]	Time per iteration [ms]	Power [W]	Performance [GFLOPS]	Perf/Power [GOP/J]	Time per iteration [ms]
Pedraforca	148	129.57	0.88	41.95	187W	383.23	2.05	9.58
Coka	233	128.16	0.55	34.85	300W	461.28	1.54	9.68

Reference: INFN Poster at SC13.

ARM-based prototypes at BSC



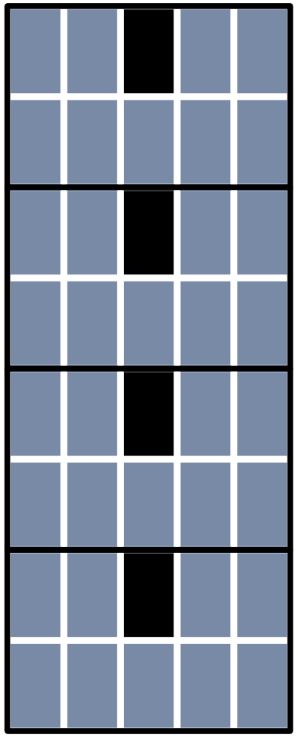
2011	**2012**	**2013**	**2014**
Tibidabo	Kayla	Pedraforca	Mont-Blanc
ARM multicore	ARM + GPU CUDA on ARM	ARM + GPU Inifinband	ARM + GPU (custom board) OpenCL on ARM GPU

Mont-Blanc

Exynos 5 compute card
2 x Cortex-A15 @ 1.7GHz
1 x Mali T604 GPU
6.8 + 25.5 GFLOPS (peak)
15 Watts
2.1 GFLOPS / W

Carrier blade
15 x Compute cards
485 GFLOPS
1 GbE to 10 GbE
300 Watts
1.6 GFLOPS / W

Blade chassis 7U
9 x Carrieir blade
135 x Compute cards
4.3 TFLOPS
2.7 KWatts
1.6 GFLOPS / W

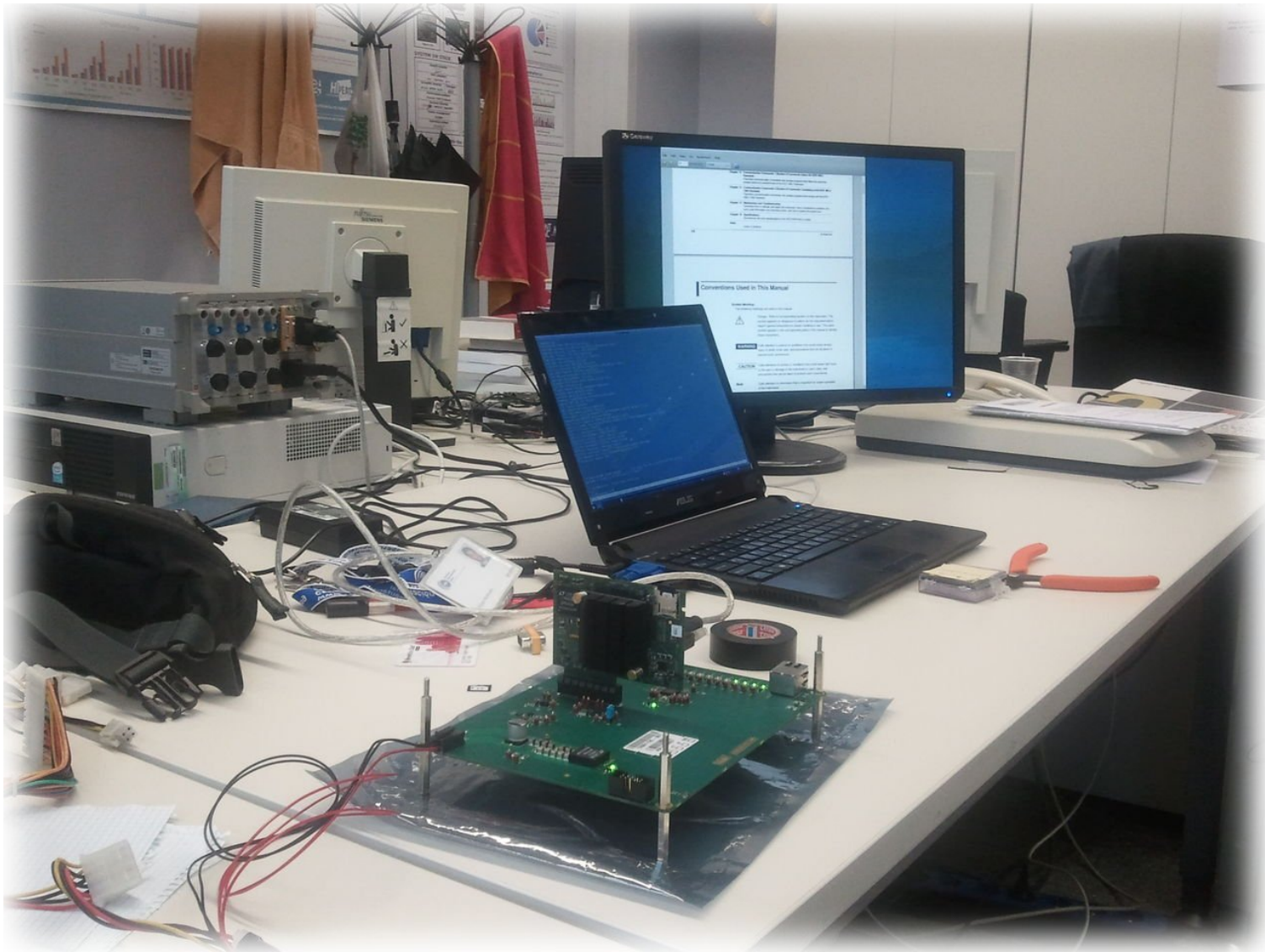


1 Rack
4 x blade chassis
36 carrier blades
540 compute cards
2x 36-port 10GbE switch
8-port 40GbE uplink
17.2 TFLOPS
11 Kwatt
1.5 GFLOPS / W

First entry
Green500
Nov'13:
4.5 GFLOPS / W

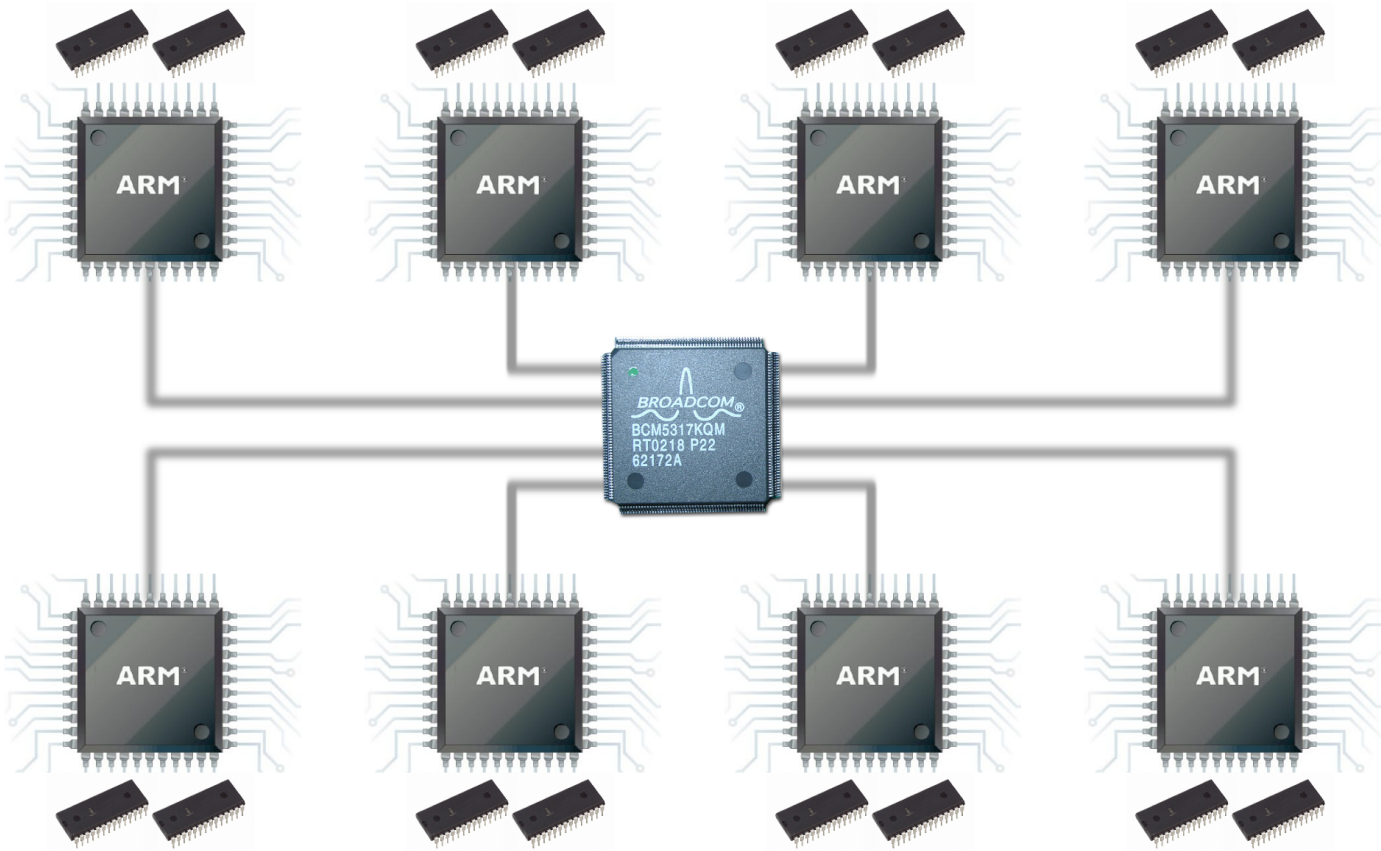
- ➜ Mont-Blanc prototype limited by SoC timing + availability
  Exynos 5 Dual is the 1st ARM Cortex-A15 SoC
- ➜ New mobile SoCs keep appearing in the market
  Exynos 5 Octa, Tegra 4, Snapdragon 800, Tegra 5 “Logan”, ...

Mont-Blanc: Sep. 2013



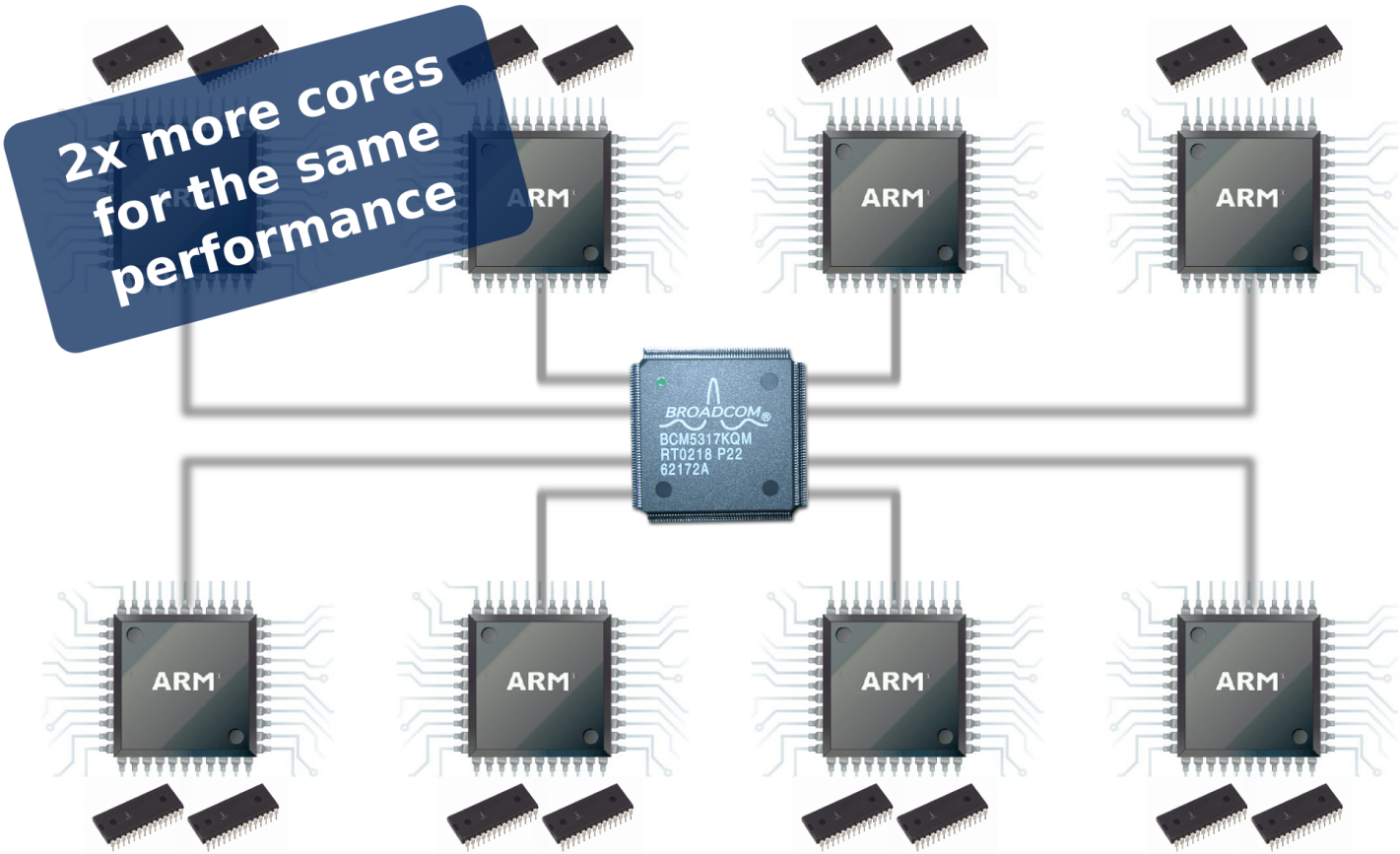
Still challenging...

Compare to a current HPC system:



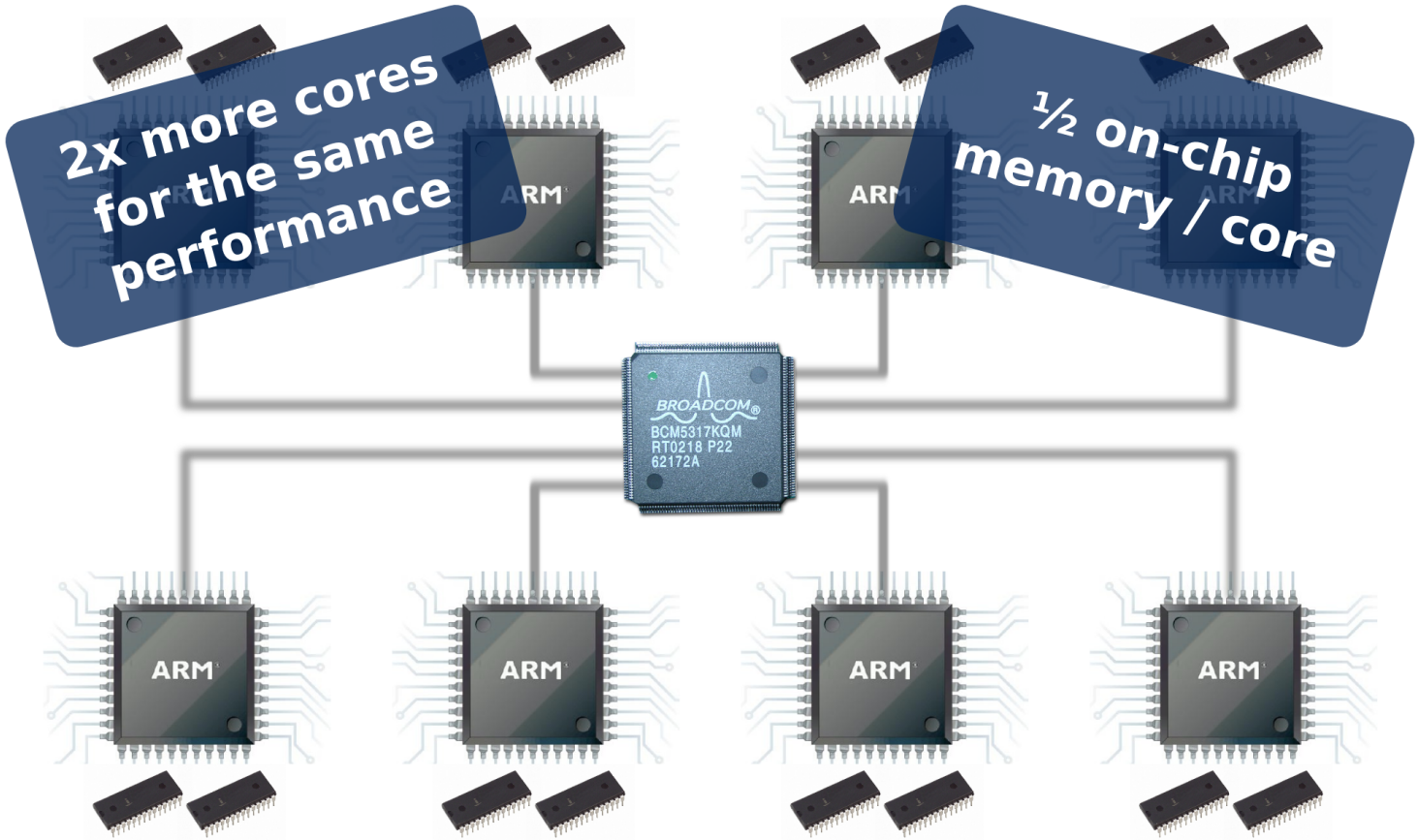
Still challenging...

Compare to a current HPC system:



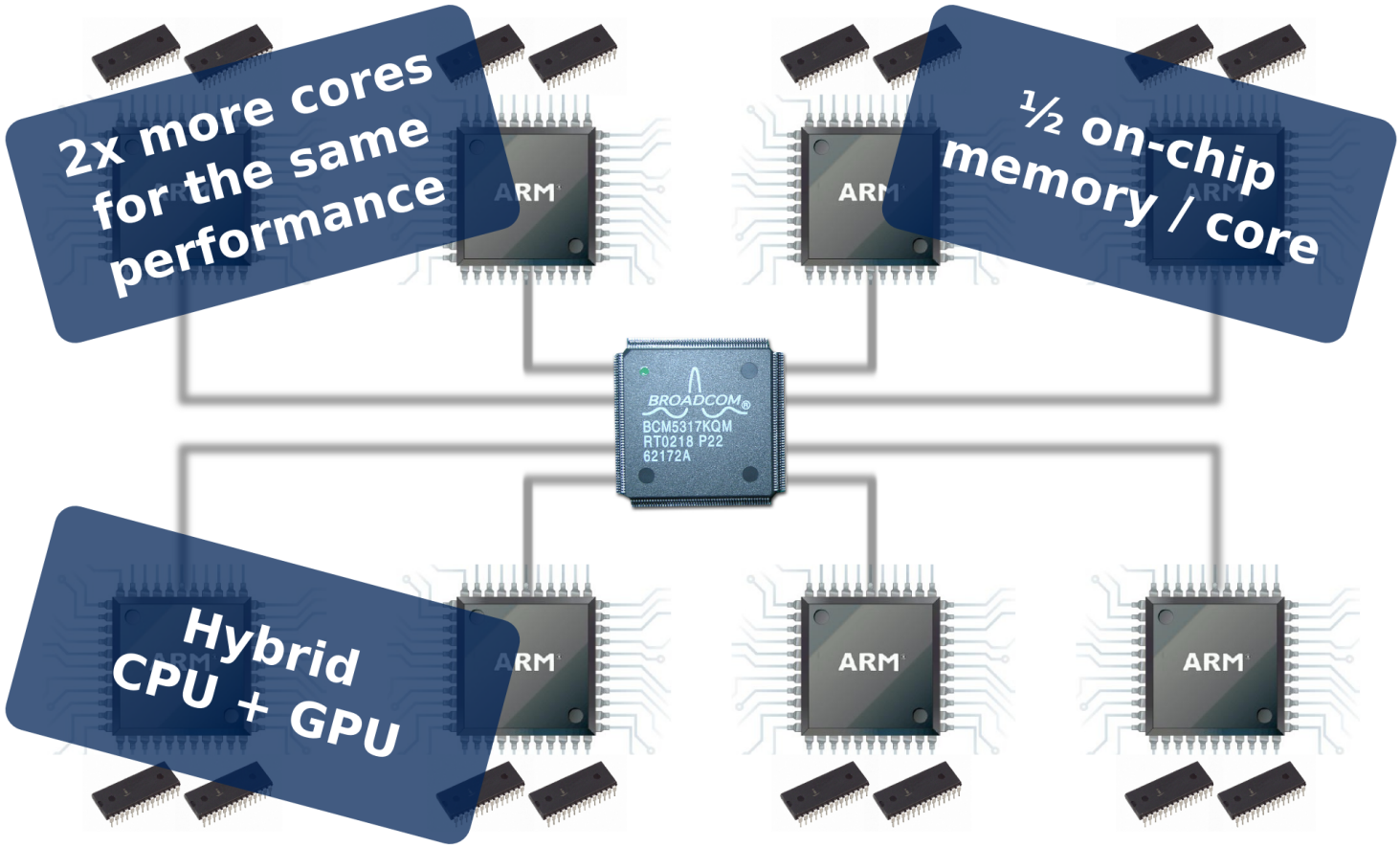
Still challenging...

Compare to a current HPC system:



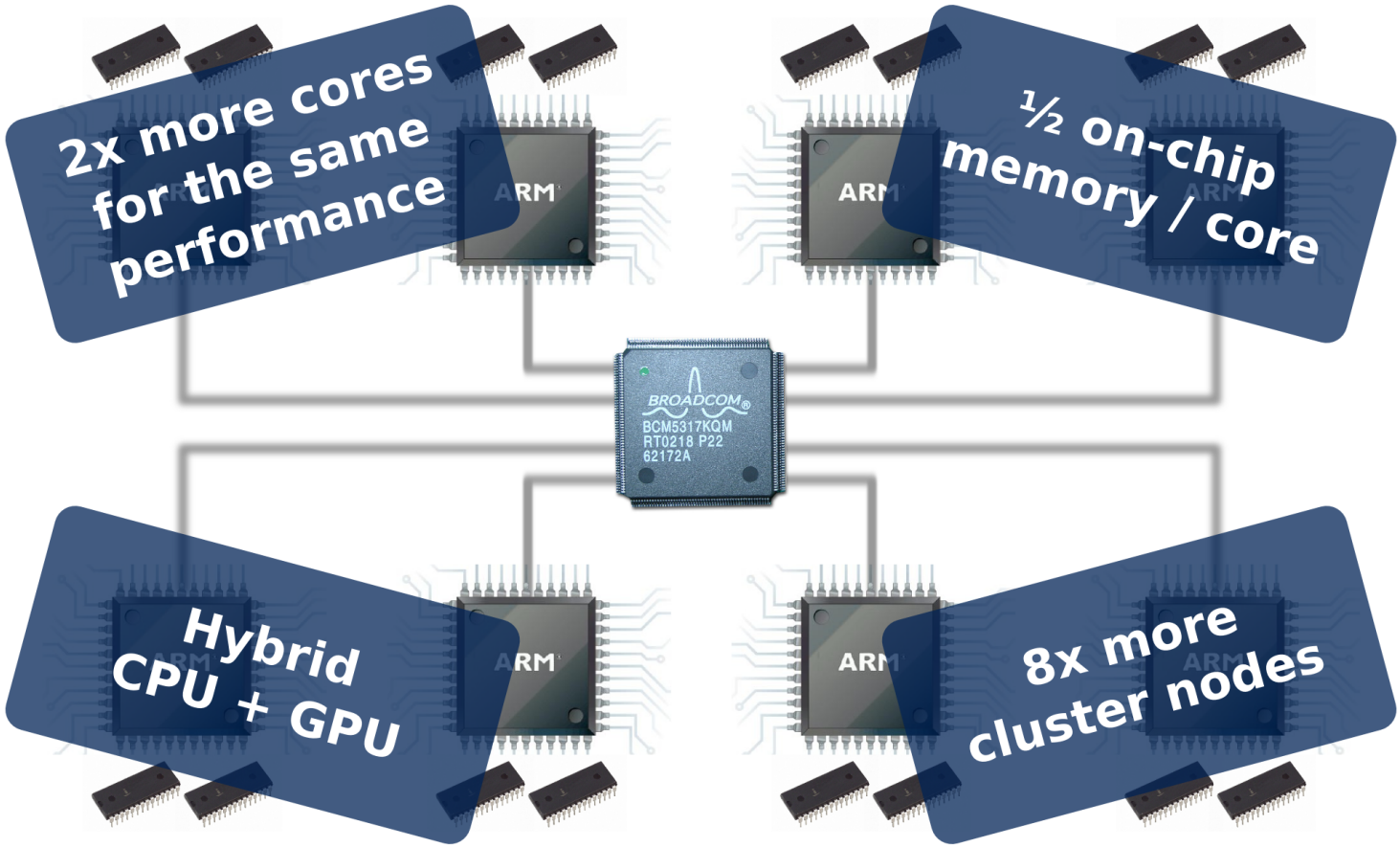
Still challenging...

Compare to a current HPC system:



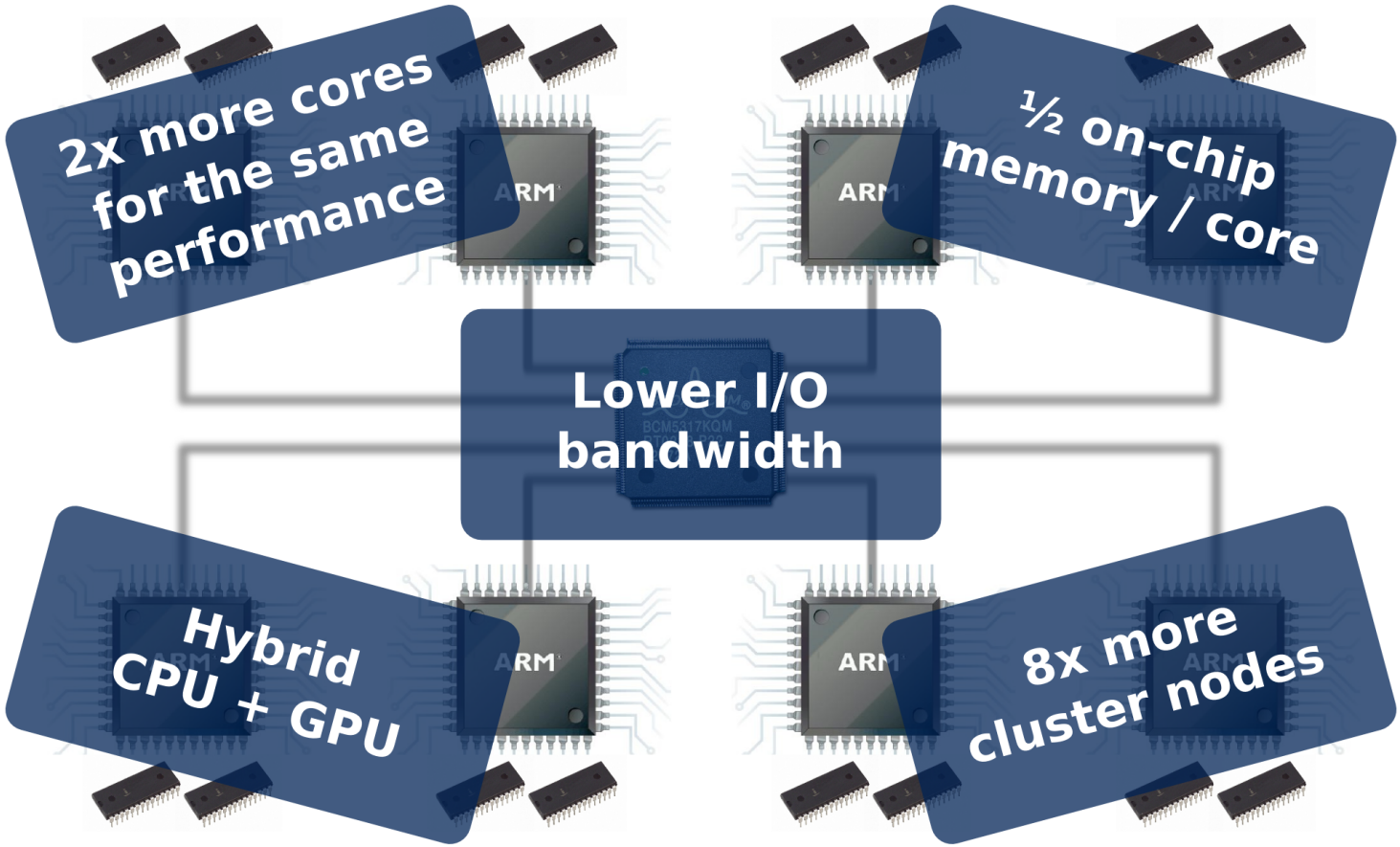
Still challenging...

Compare to a current HPC system:

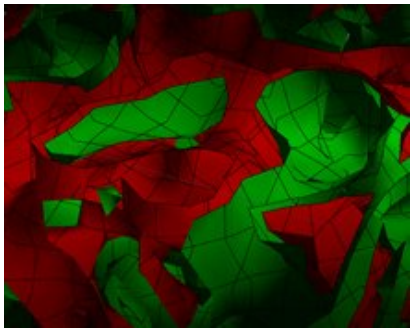


Still challenging...

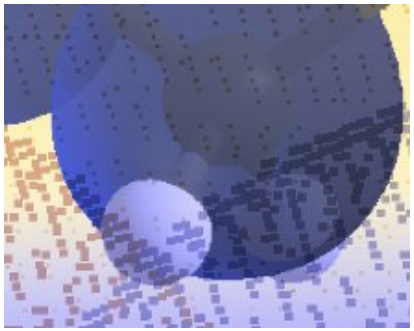
Compare to a current HPC system:



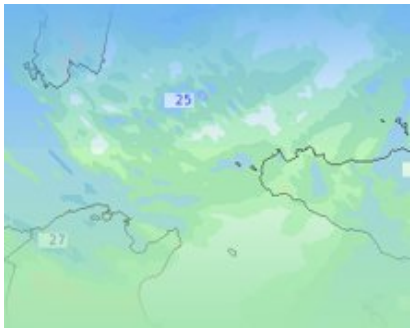
Mont-Blanc: applications



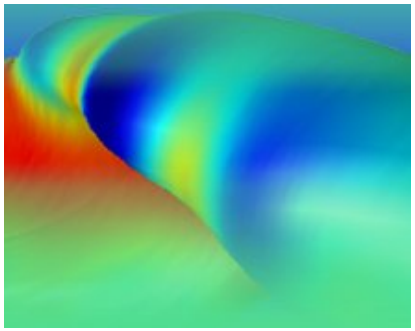
BQCD
Particle physics



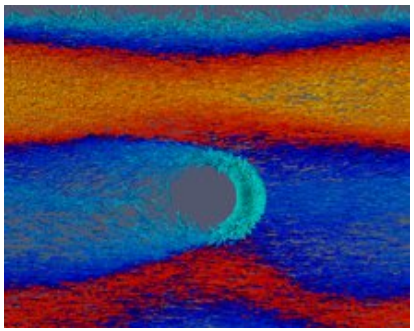
BigDFT
Elect. Structure



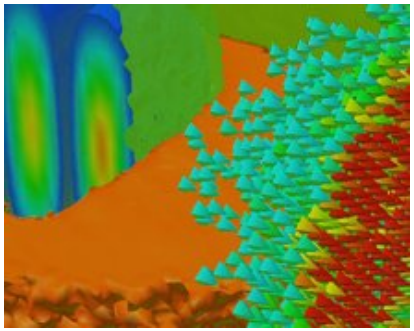
COSMO
Weather forecast



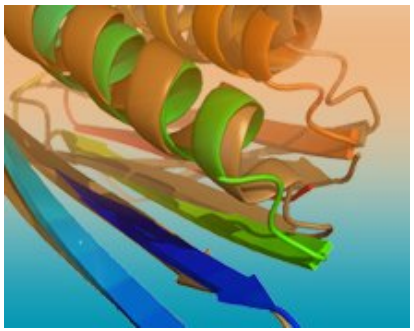
EUTERPE
Fusion



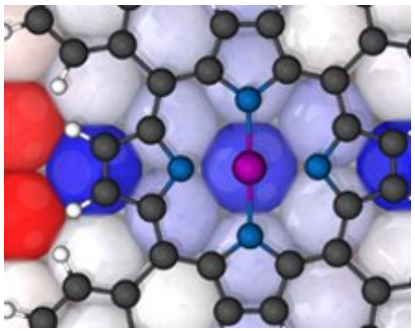
MP2C
Multi-particle collisions



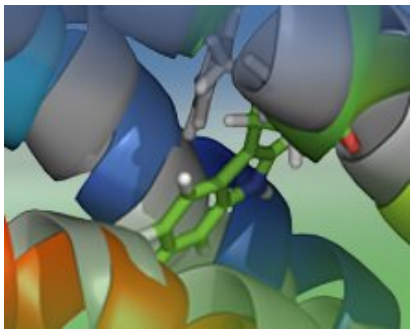
PEPC
Coulomb+Grav. Forces



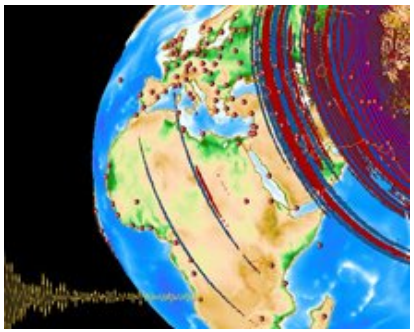
ProFASI
Protein folding



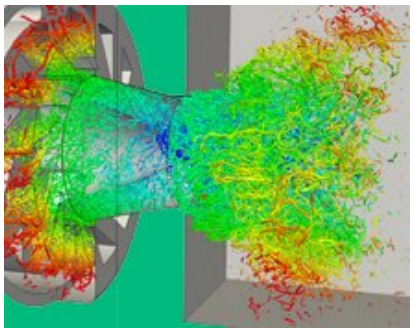
Quantum
ESPRESSO Elect. Structure



SMMP
Protein folding



SPECFEM3D
Wave propagation



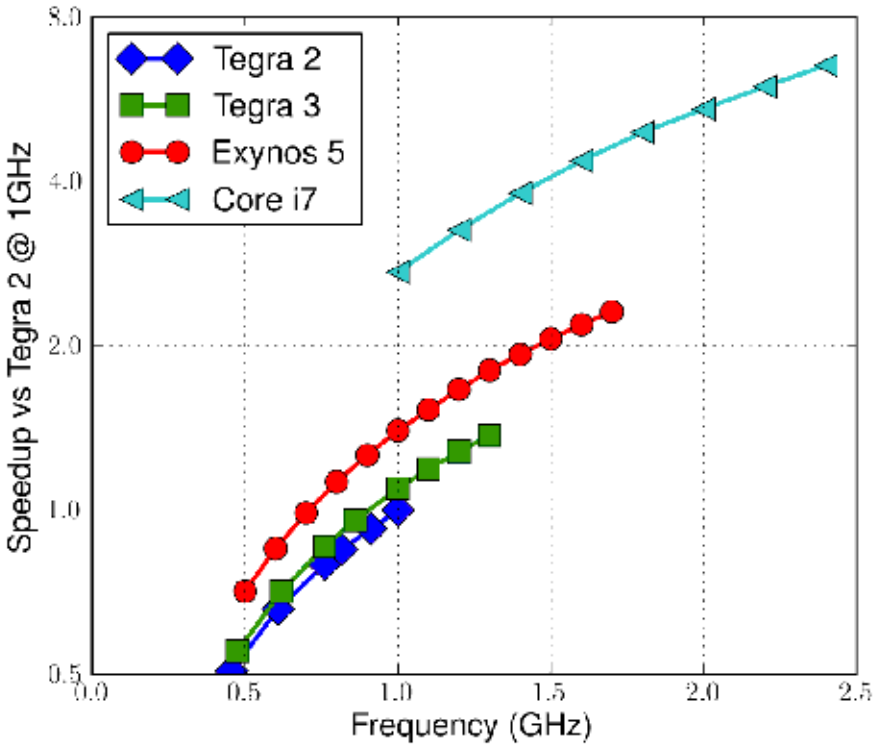
YALES2
Combustion

Mont-Blanc: our toolbox

Set of microbenchmarks coded in C, ported to serial, pthreads, OpenMP, OmpSs, CUDA, OpenCL.

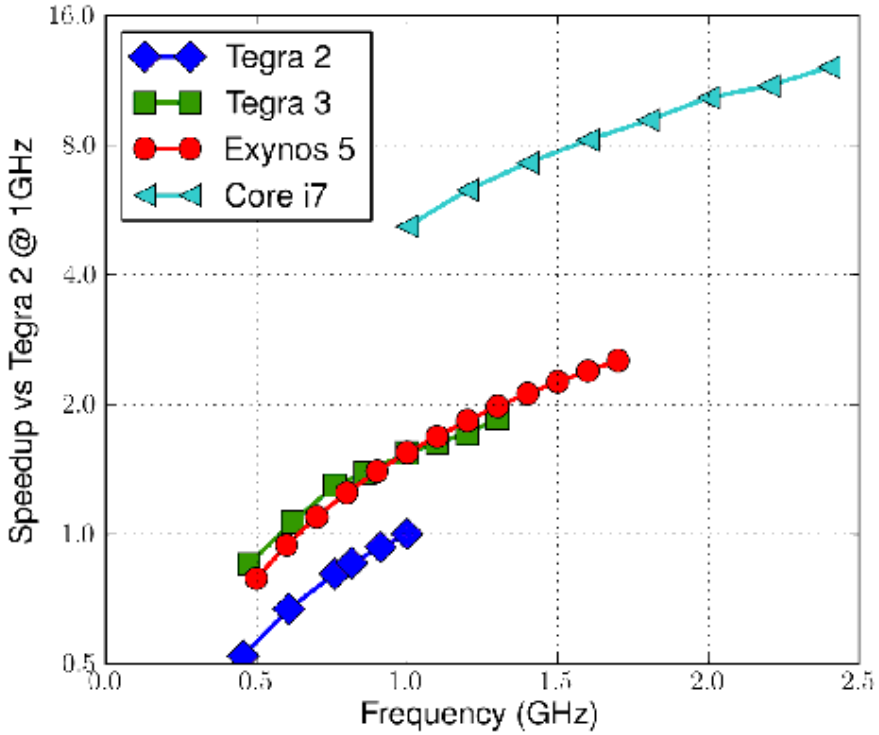
Benchmark	Properties
Vector Operation (vecop)	Common operation in regular codes
Dense Matrix-Matrix Multiplication (dmmm)	Data reuse and compute performance
3D stencil (3dstc)	Strided memory accesses (7-point 3D stencil)
2D Convolution (2dcon)	Spatial locality
Fast Fourier Transform (fft)	Peak floating-point, variable-stride accesses
Reduction (red)	Varying levels of parallelism (Scalar sum)
Histogram (hist)	Local privatisation and reduction
Merge Sort (msort)	Barrier synchronisation
N-Body (nbody)	Irregular memory accesses
Atomic Monte-Carlo Dynamics (amcd)	Embarrassingly parallel: compute performance
Sparse Vector-Matrix Multiplication (spwm)	Load imbalance

Mont-Blanc: benchmark results (single core)



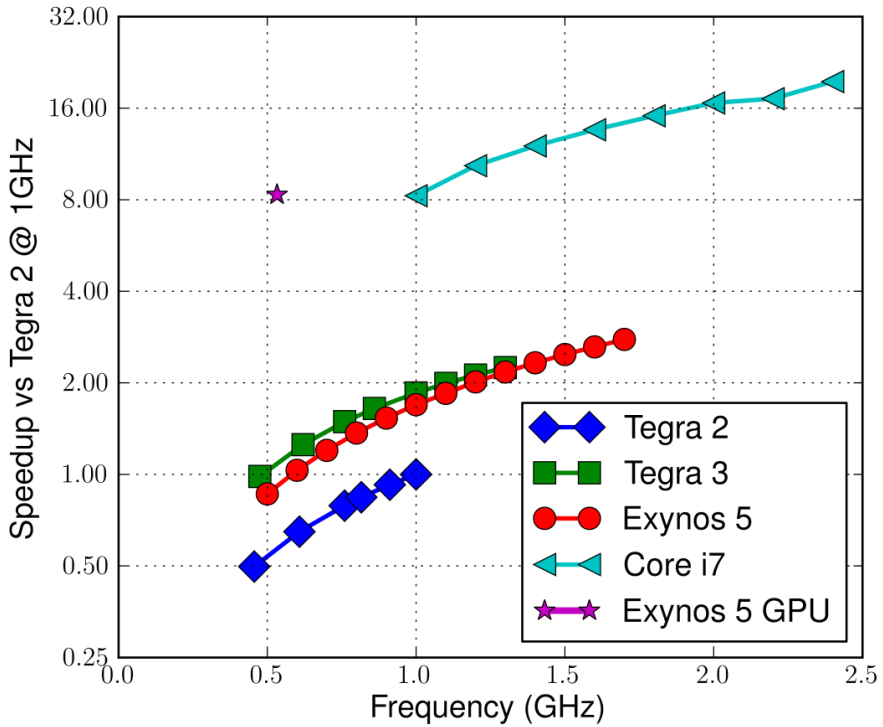
- Cortex-A9 in Tegra3 is 1.4x faster than Tegra2 (higher clock frequency)
- Cortex-A15 in Exynos5 is 1.7x faster than Cortex-A9 in Tegra3
  Higher clock frequency, higher memory bandwidth, and better core microarchitecture
- Core i7 is ∼3x faster than Cortex-A15 in Exynos5 at maximum frequency, 2x faster at the same frequency

Mont-Blanc: benchmark results (multi core)



- Tegra3 platform as fast as Exynos5 platform
  4-core Cortex-A9 vs. 2-core Cortex-A15
- Corei7 is 6× faster than Exynos5 at maximum frequency
- ...and the GPU is still not used in these tests

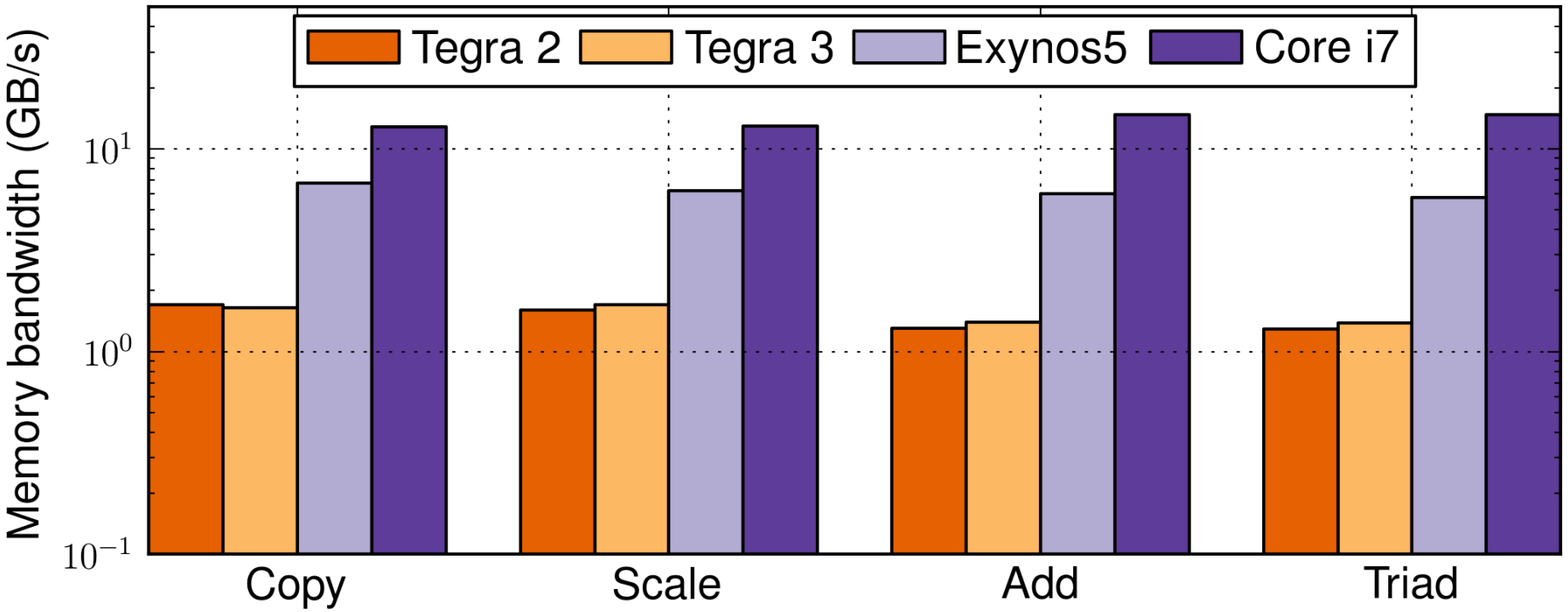
Mont-Blanc: benchmark results (with GPU)



- Exynos 5 also integrates a compute capable Mali-T604 GPU
- Exynos 5 GPU platform as fast as Core i7 platform at 1GHz

  3× faster than ARM Cortex-A15 dual core at max. frequency

Mont-Blanc: STREAM benchmark



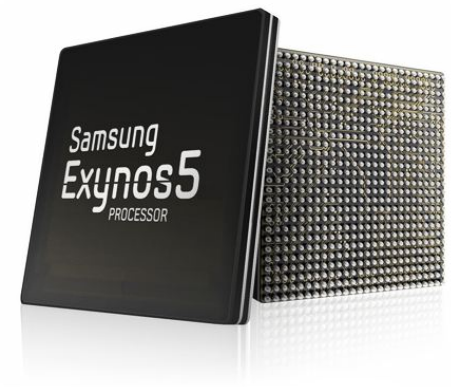
- Exynos 5 improves dramatically over Tegra (4.5x)
  * Dual-channel DDR3
  * ARM Cortex-A15 sustains more in-flight cache misses
- Corei7 provides 2x more memory bandwidth than Exynos5

Interconnect evaluation: SoCs under study

Nvidia Tegra 2

1 GbE (on PCIe)

100 Mbit (on USB 2.0)

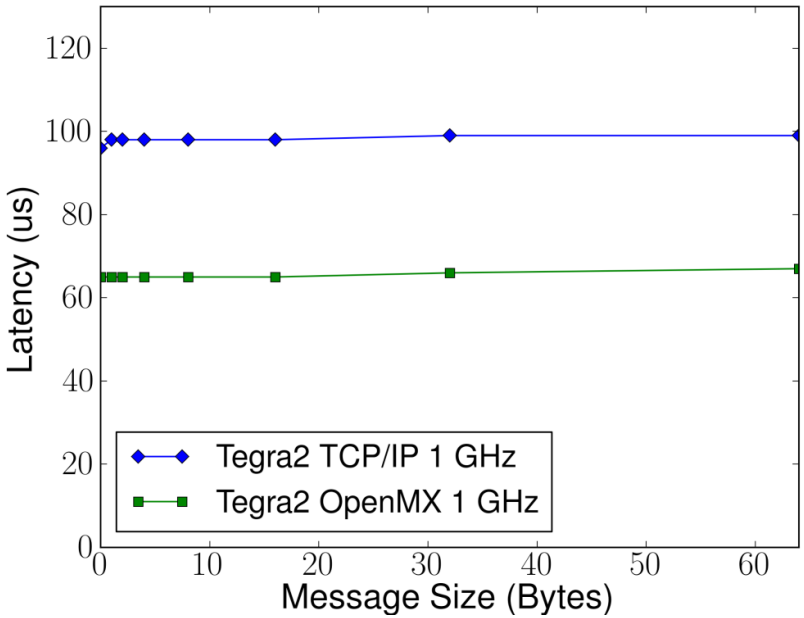


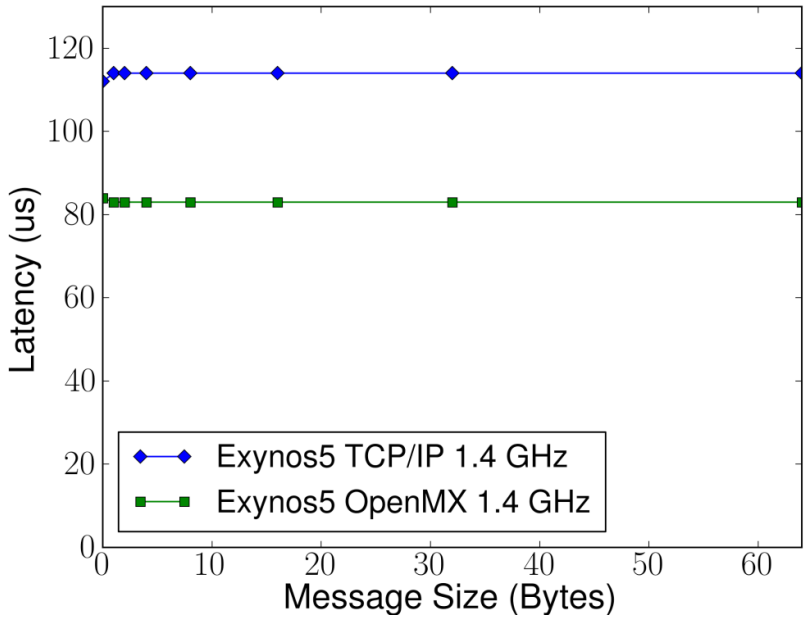
Samsung Exynos 5 Dual

1 GbE (on USB3.0)

100 Mbit (on USB 2.0)

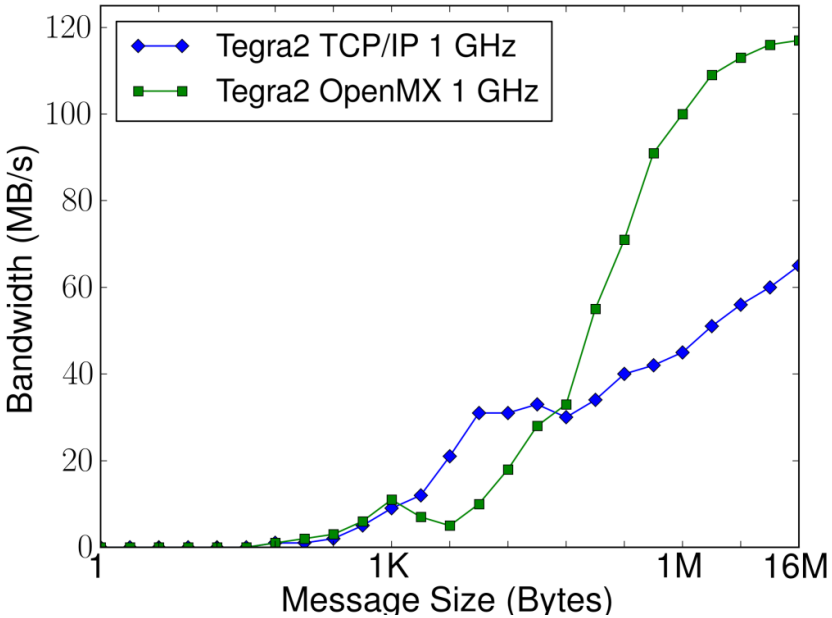
Interconnection Latency

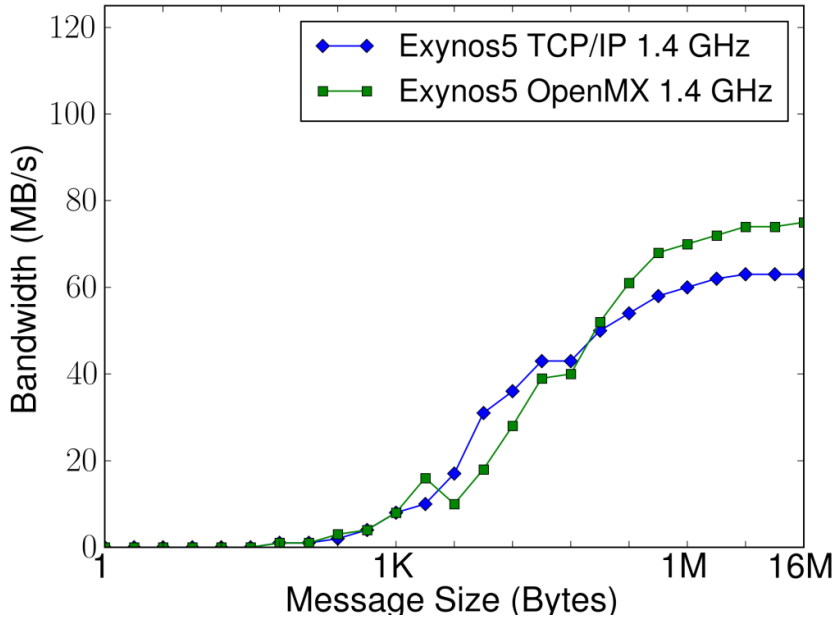




- TCP/IP adds significant CPU overhead
- OpenMX (Myrinet Express) driver interfaces "directly" to the Ethernet NIC
- USB in Exynos5 adds extra latency on top of network stack

Interconnection Bandwidth





- TCP/IP overhead prevents Tegra2 from achieving full bandwidth
  OpenMX does achieve peak bandwidth
- USB overheads prevent Exynos 5 from achieving full bandwidth, even with OpenMX
- The # of FLOP needed to hide a data transfer is $\sim$ comparable with a high end HPC system (Intel+IB)

Mobile SoC limitations for HPC

- 32-bit memory controller
  Even though ARM Cortex-A15 offers 40-bit address space
- No ECC protection in memory
  Limiting factor for scalability after certain number of nodes
- No standard server I/O interfaces
  Provide USB 3.0, SATA and (minimal) PCIe
- No network protocol offload engines
  TCP/IP, OpenMX, USB protocol stacks run on the CPU
- Thermal package not designed for sustained full-power operation

These are only **design decisions**, not really **unsolvable problems**.
ARM server SoCs don't have any of these restrictions!

If vendors decide to include a minimum set of required features...

Conclusions and remarks

1. Mont-Blanc architecture is shaping up:
   - ➜ ARM multicore + integrated OpenCL accelerator
   - ➜ Ethernet NIC
   - ➜ High density packaging
2. A full set of tools and "underworld" for scientific computing on SoCs is developing:
   - ➜ linux distributions
   - ➜ programming model/tools
   - ➜ scientific libraries
   - ➜ applications
   - ➜ micro-benchmarks
3. Interconnection of many SoCs is an issue, as they are built to be self-sufficient

Commodity SoCs still immature, but they are funny and evolve pretty fast and we do not want to be caught imprepared the day that they are eventually ready for HPC!!!

Reference: http://dl.acm.org/citation.cfm?doid=2503210.2503281