

MONT-BLANC

<http://www.montblanc-project.eu>

The Mont-Blanc approach towards Exascale

Alex Ramirez

Barcelona Supercomputing Center



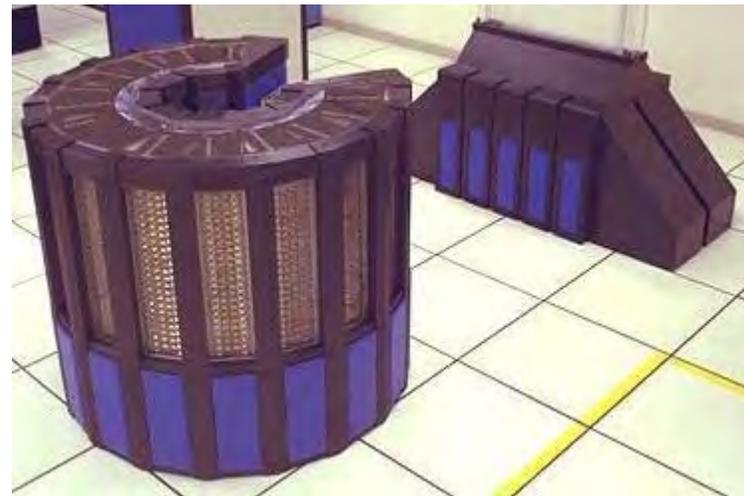
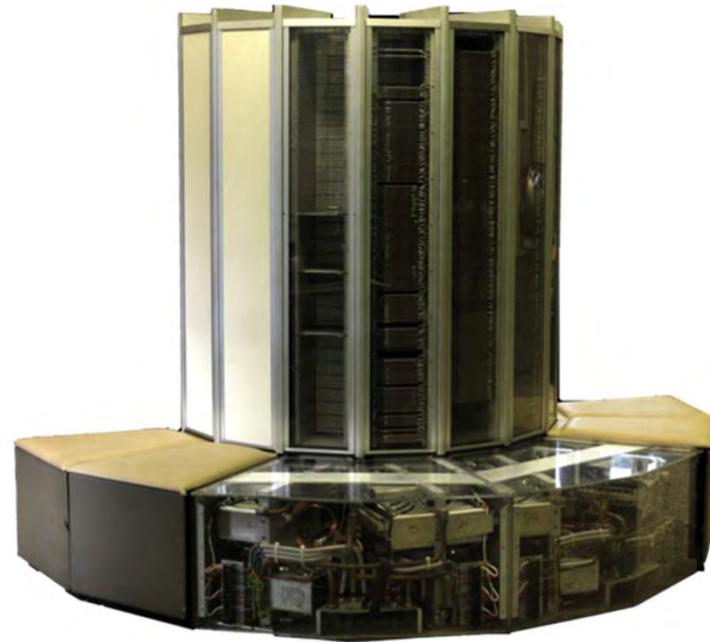
Disclaimer: I only speak for myself, not for the individual members of the consortium. All references to unavailable products are speculative, taken from web sources. There is no commitment from ARM, Samsung, TI, Bull, or others, implied.

Outline

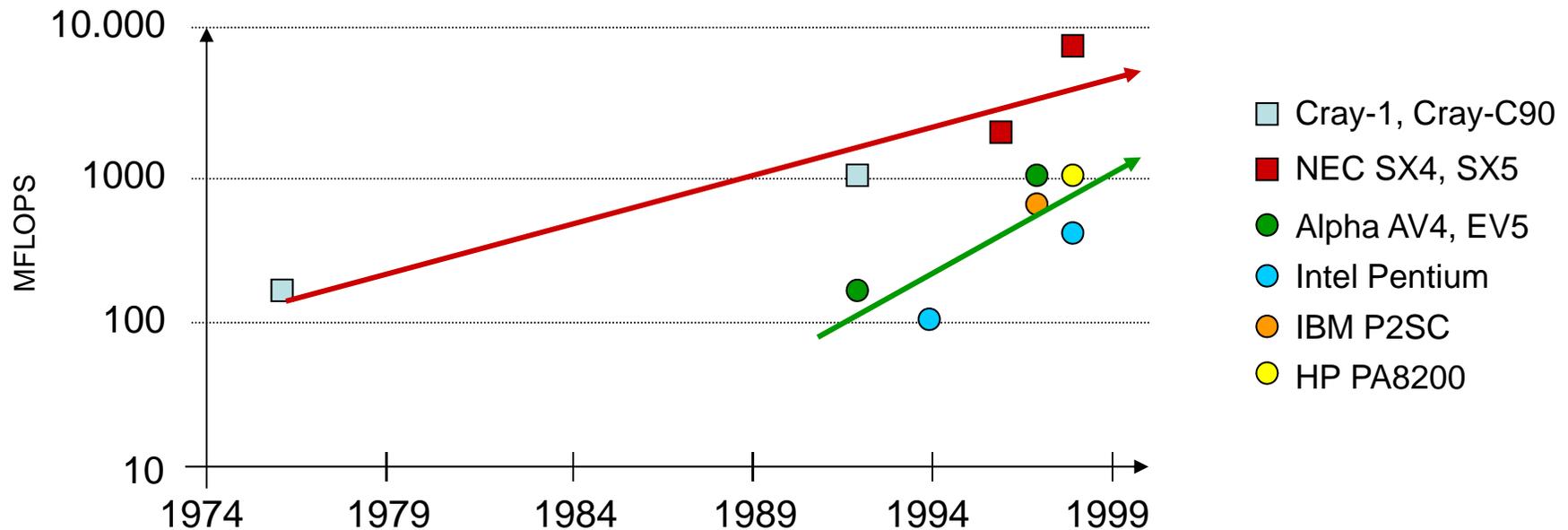
- A bit of history
 - Vector supercomputers
 - Commodity supercomputers
 - The next step in the commodity chain
- Supercomputers from mobile components
 - Killer mobile examples
 - Mont-Blanc architecture strawman
 - Rely on OmpSs to handle the challenges
- BSC prototype roadmap
- Mont-Blanc project goals and milestones

In the beginning ... there were only supercomputers

- Built to order
 - Very few of them
- Special purpose hardware
 - Very expensive
- Control Data, Convex, ...
- Cray-1
 - 1975, 160 MFLOPS
 - 80 units, 5-8 M\$
- Cray X-MP
 - 1982, 800 MFLOPS
- Cray-2
 - 1985, 1.9 GFLOPS
- Cray Y-MP
 - 1988, 2.6 GFLOPS
- Fortran+vectorizing compilers



The Killer Microprocessors



- Microprocessors killed the Vector supercomputers
 - They were not faster ...
 - ... but they were significantly cheaper and greener
- Need 10 micros to achieve the performance of 1 vector CPU
 - SIMD vs. MIMD programming paradigms

Then, commodity took over special purpose



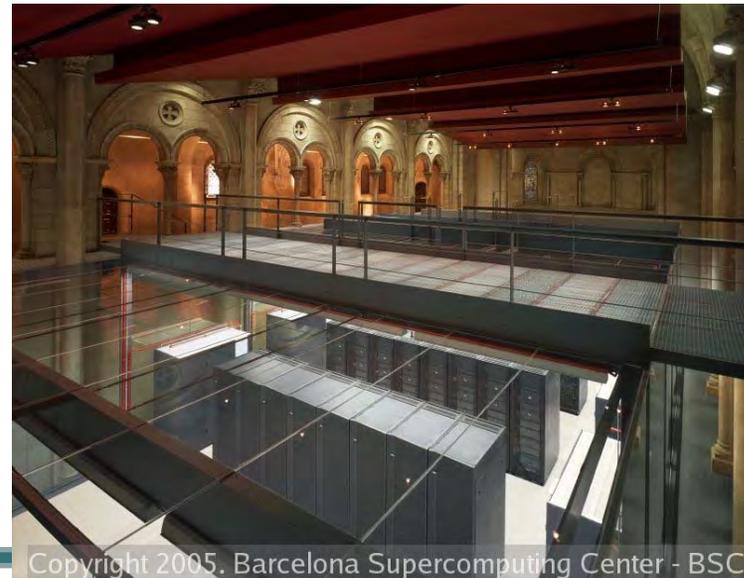
- ASCI Red, Sandia
 - 1997, 1 Tflops (Linpack),
 - 9298 cores @ 200 Mhz
 - 1.2 Tbytes
 - Intel Pentium Pro
 - Upgraded to Pentium II Xeon, 1999, 3.1 Tflops

- ASCI White, LLNL
 - 2001, 7.3 TFLOPS
 - 8192 proc. @ 375 Mhz,
 - 6 Tbytes
 - (3+3) Mwats
 - IBM Power 3

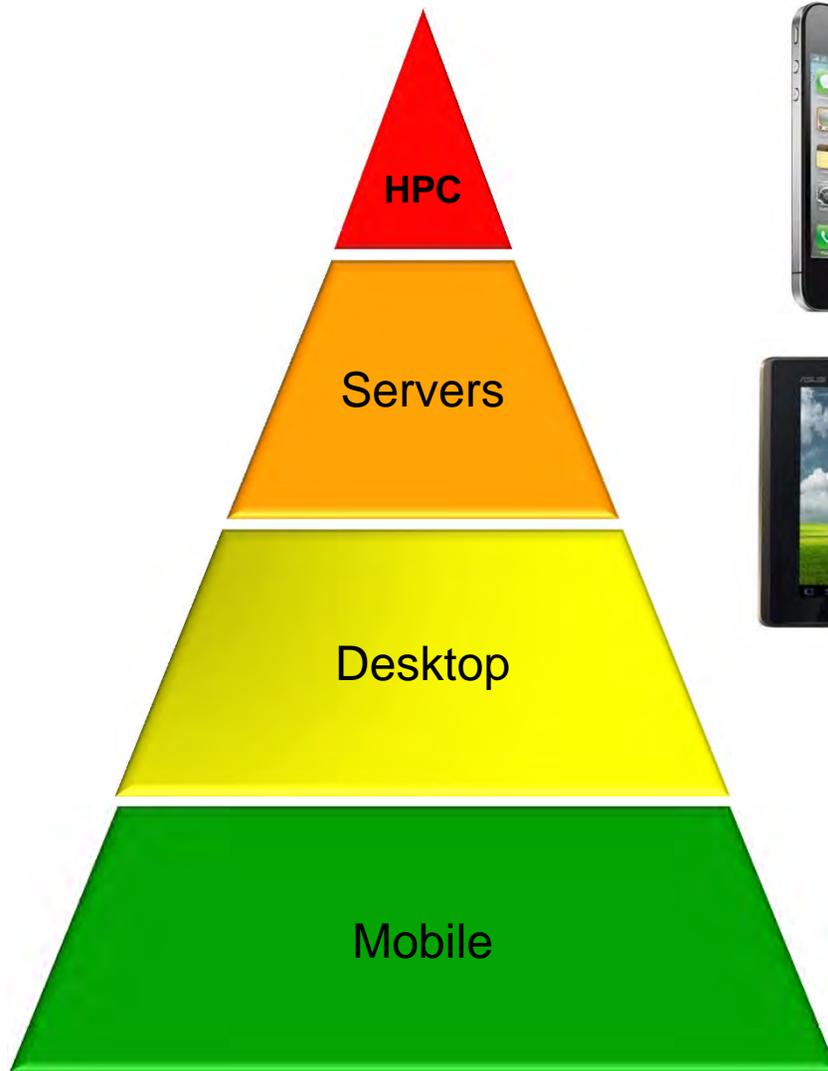
Message-Passing Programming Models

Finally, commodity hardware + commodity software

- MareNostrum
 - Nov 2004, #4 Top500
 - 20 Tflops, Linpack
 - IBM PowerPC 970 FX
 - Blade enclosure
 - Myrinet + 1 GbE network
 - SuSe Linux

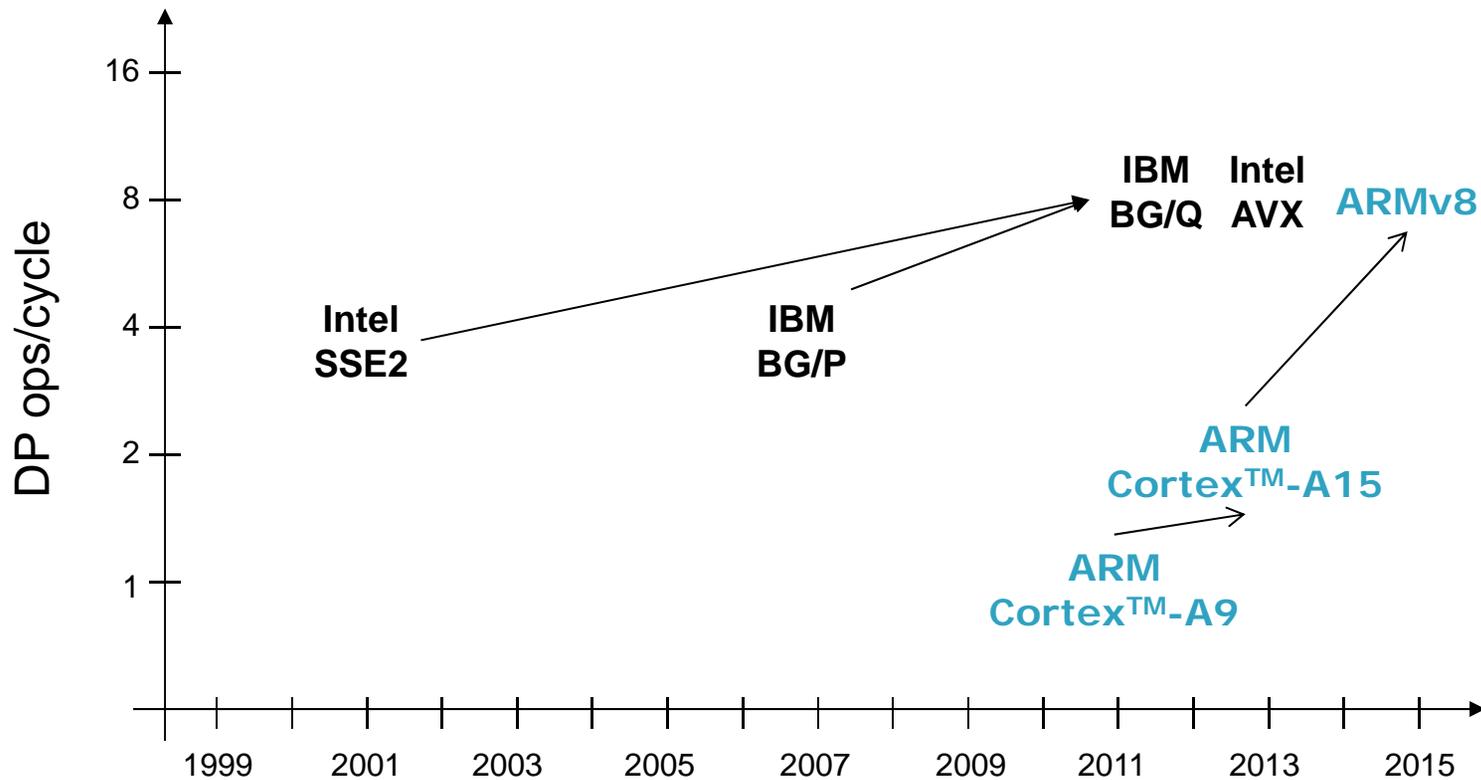


The next step in the commodity chain



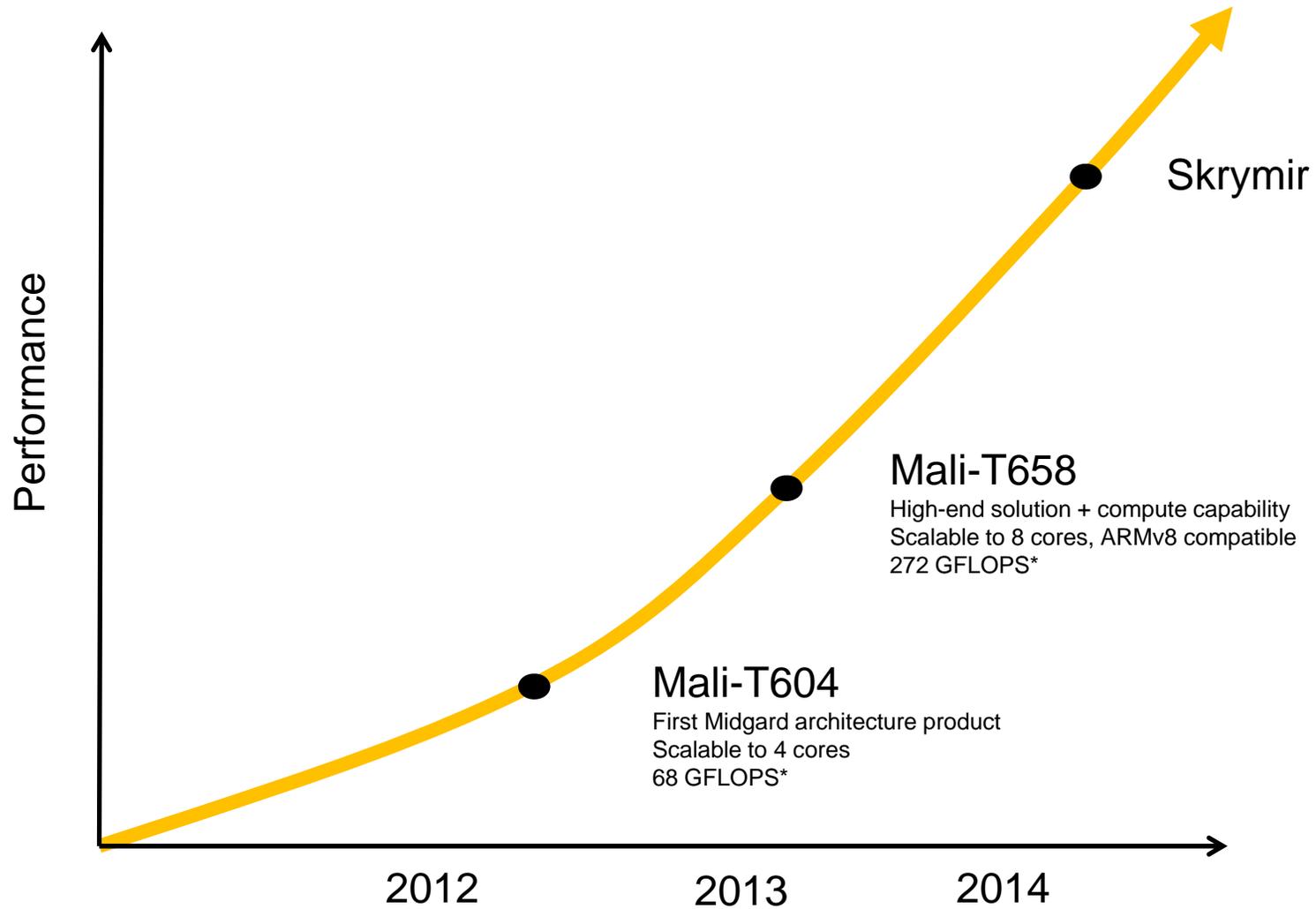
- Total cores in Jun'12 Top500
 - 13.5 Mcores
- Tablets sold in Q4 2011
 - 27 Mtablets
- Smartphones sold Q4 2011
 - > 100 Mphones

ARM Processor improvements in DP FLOPS



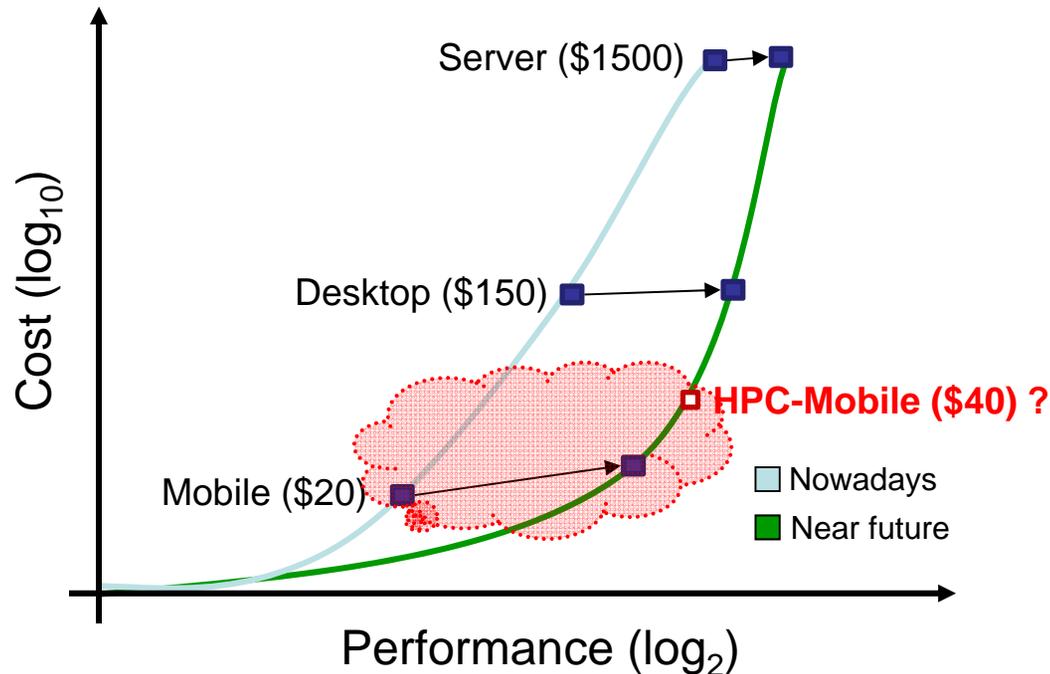
- IBM BG/Q and Intel AVX implement DP in 256-bit SIMD
 - 8 DP ops / cycle
- ARM quickly moved from optional floating-point to state-of-the-art
 - ARMv8 ISA introduces DP in the NEON instruction set (128-bit SIMD)

Integrated ARM GPU performance



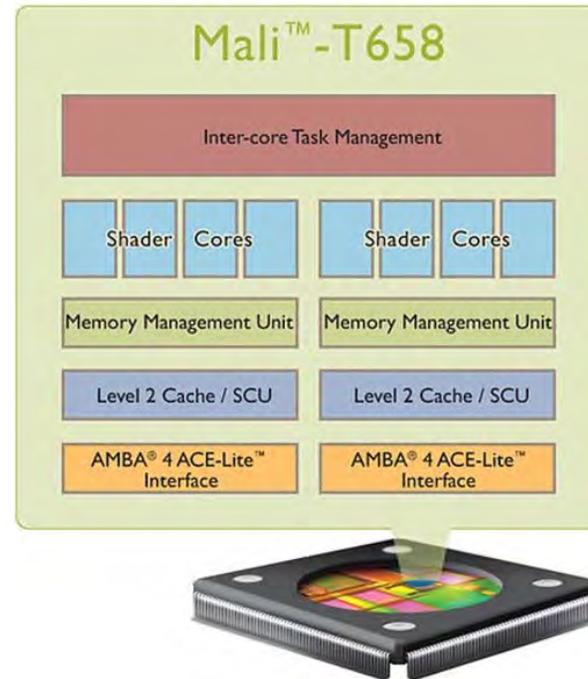
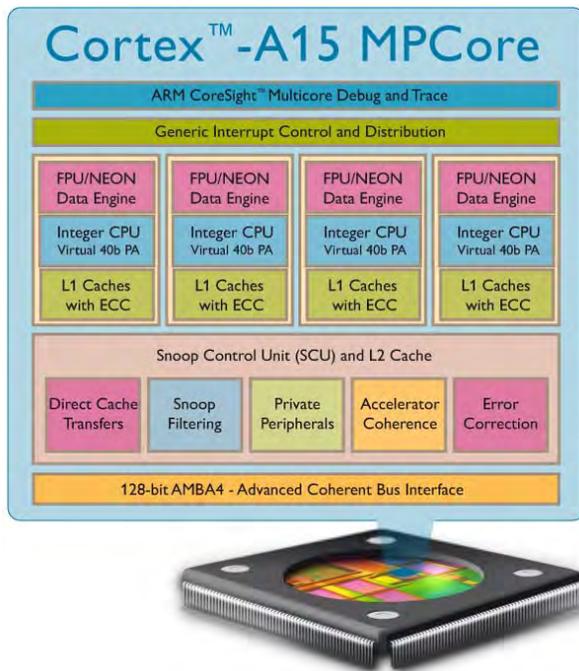
* Data from web sources, not an ARM commitment

Are the “Killer Mobiles™” coming?

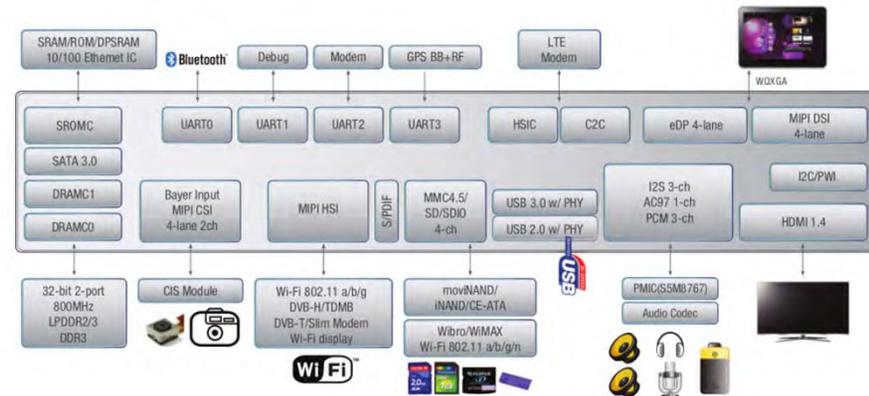


- Where is the sweet spot? Maybe in the low-end ...
 - Today ~ 1:8 ratio in performance, 1:100 ratio in cost
 - Tomorrow ~ 1:2 ratio in performance, still 1:100 in cost ?
- The same reason why microprocessors killed supercomputers
 - Not so much performance ... but much lower cost, and power

Killer mobile™ example: Samsung Exynos 5450 *



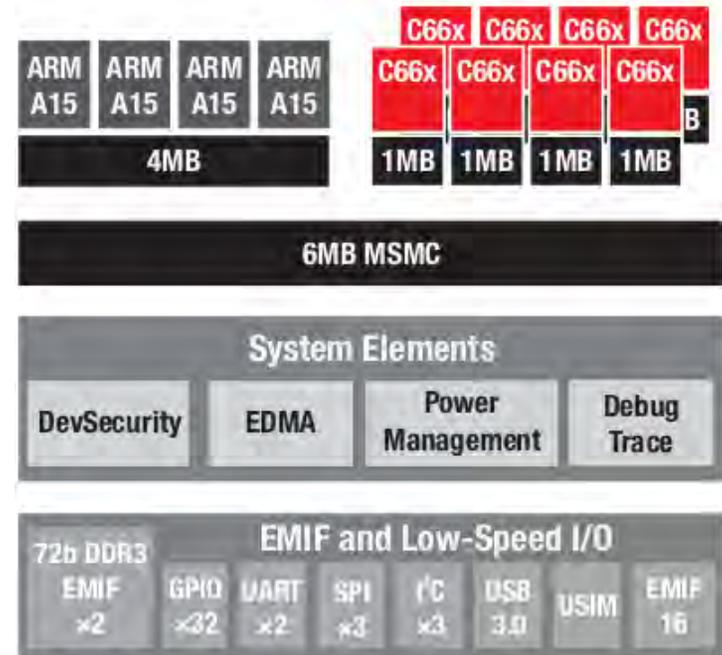
- 4-core ARM Cortex-A15 @ 2 GHz
 - 16 GFLOPS
- 8-core ARM Mali T685
 - 272 GFLOPS*
- Dual channel DDR3 memory controller
- **All in a low-power mobile socket**



* Data from web sources, not an ARM or Samsung commitment

Killer mobile™ example: TI KeyStone II *

- 4-core Cortex-A15 @ 2 GHz
 - 16 GFLOPS
- 8-core C66x DSP
 - 160 SP GFLOPS
 - 60 DP GFLOPS
- Dual channel DDR3 + ECC
- High speed I/O interfaces
- 4-port Gigabit Ethernet switch
- **All in a 10-15W socket***



* Data from web sources, not an ARM or TI commitment

Integrated CPU + GPU

- BSC has low-power prototypes for other architectures ...
 - Homogeneous multicore
 - Tibidabo: Tegra2 cluster (2x ARM Cortex-A9)
 - Heterogeneous multicore + discrete accelerator
 - Pedraforca: Tegra3 + CUDA GPU (4x Cortex A9 + Quadro 1000M)
- If we want to be better, we must be different
$$A > B \Rightarrow A \neq B$$
- Integrated GPU has many advantages
 - Shared memory with CPU
 - Even cache coherent!
 - No power wasted on PCIe bus
 - No power *wasted* on GDDR5 memory
 - Higher energy efficiency + lower cost

Are we building BlueGene again?

- Yes ...
 - Exploit Pollack's Rule in presence of abundant parallelism
 - Many small cores vs. Single fast core
- ... and No
 - Heterogeneous computing
 - On-chip GPU, DSP
 - Commodity vs. Special purpose
 - Higher volume
 - Many vendors
 - Lower cost
 - Lots of room for improvement
 - No SIMD / vectors yet ...
 - Build on Europe's embedded strengths

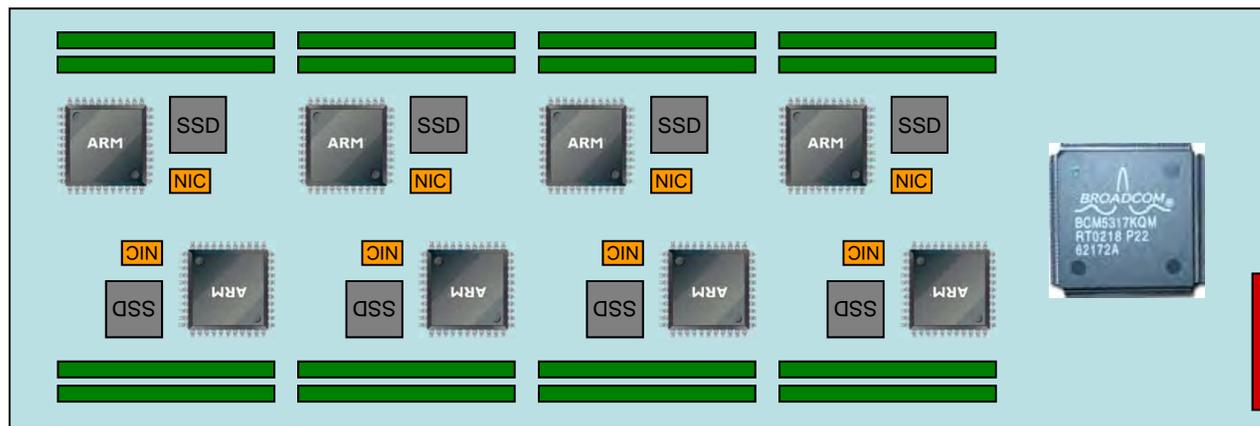


High density packaging architecture

- Standard BullX blade enclosure
- Multiple compute nodes per blade
 - Additional level of interconnect, on-blade network



X86 + Nvidia cluster, Minotauro @ BSC, 1266 MFLOPS / Watt



* Strawman design concept, not the actual Bull implementation

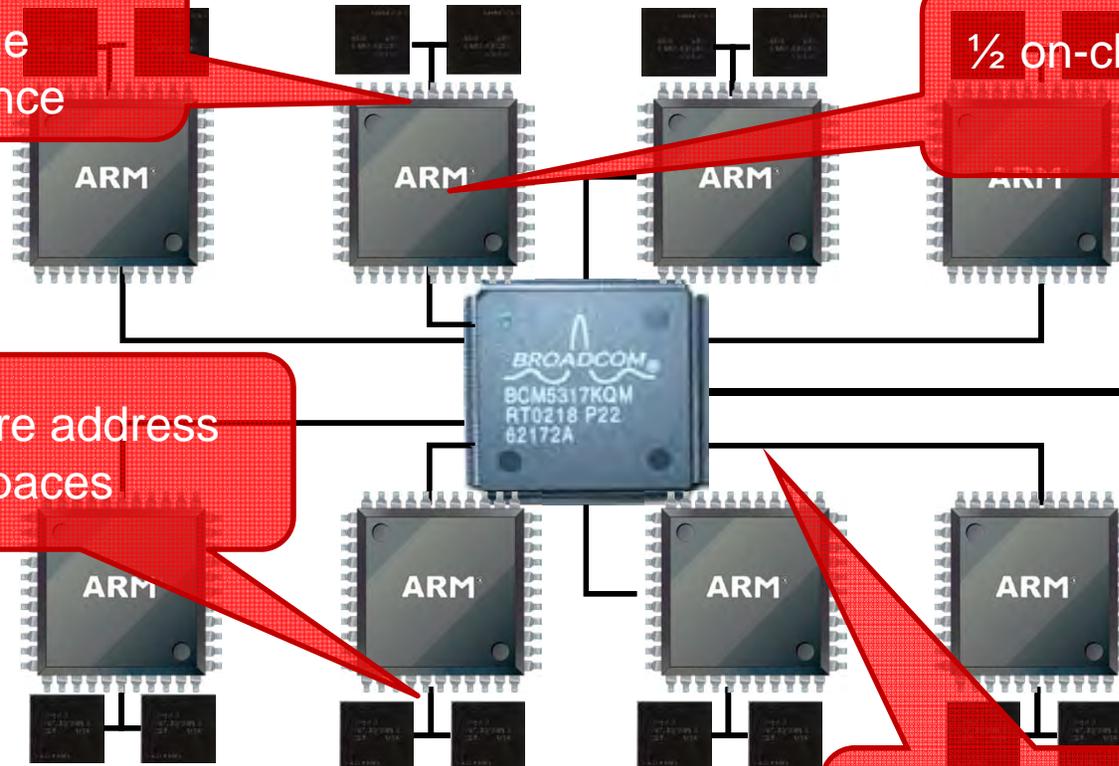
There is no free lunch

2X more cores for the same performance

1/2 on-chip memory / core

8X more address spaces

1 GbE inter-chip communication



Rely on software to handle the challenges

- Programming model and runtime are key components to address the challenges
 - Programming Model: provide mechanisms to
 - Let programmer focus on science, algorithms
 - Provide hints to runtime
 - Runtime: map to resources
 - Most information available on application demands and system state/characteristics
 - Need to put intelligence in it, need to rely on it
- Maybe *macho* programmers can get high performance today...
 - ... but what about the rest? At what cost? How portable?

OmpSs: Generate task graph at run time

```
#pragma omp task in(A, B) out(C)  
void vadd3 (float A[BS], float B[BS],  
           float C[BS]);
```



```
#pragma omp task in(sum, A) out(B)  
void scale_add (float sum, float A[BS],  
               float B[BS]);
```

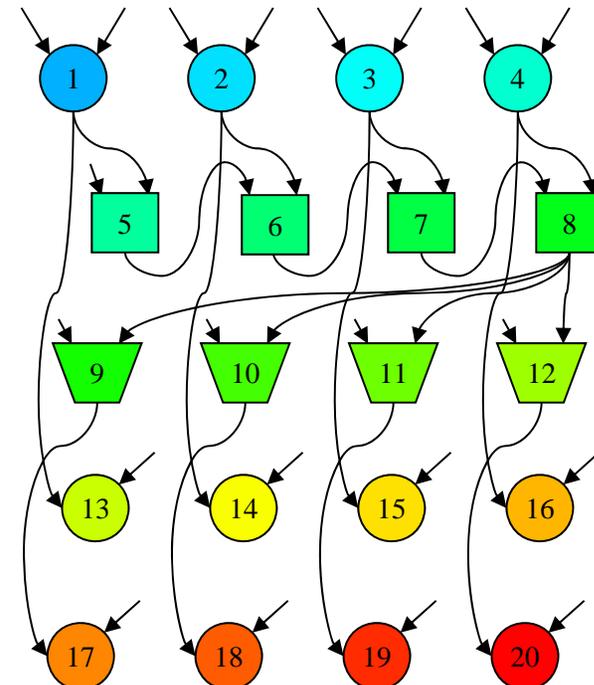


```
#pragma omp task in(A) inout(sum)  
void accum (float A[BS], float *sum);
```



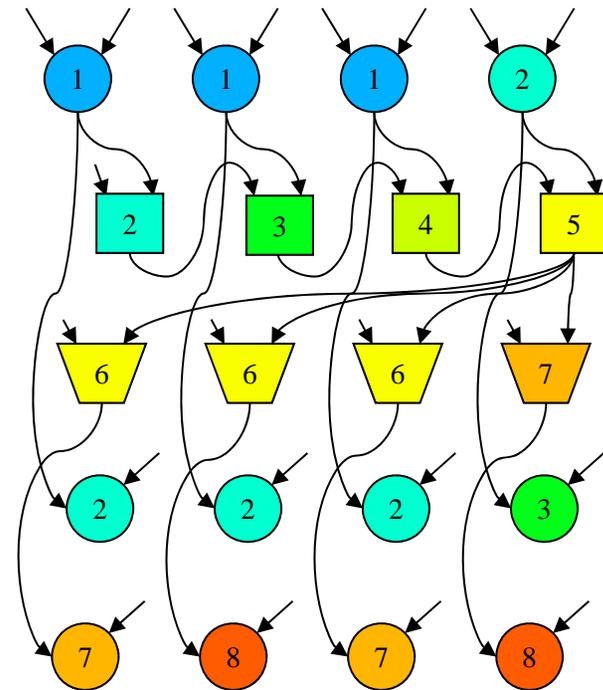
```
for (i=0; i<N; i+=BS) // C=A+B  
    vadd3 (&A[i], &B[i], &C[i]);  
...  
for (i=0; i<N; i+=BS) // sum(C[i])  
    accum (&C[i], &sum);  
...  
for (i=0; i<N; i+=BS) // B=sum*E  
    scale_add (sum, &E[i], &B[i]);  
...  
for (i=0; i<N; i+=BS) // A=C+D  
    vadd3 (&C[i], &D[i], &A[i]);  
...  
for (i=0; i<N; i+=BS) // E=C+F  
    vadd3 (&C[i], &F[i], &E[i]);
```

Simple Program Annotations Task Graph Generation



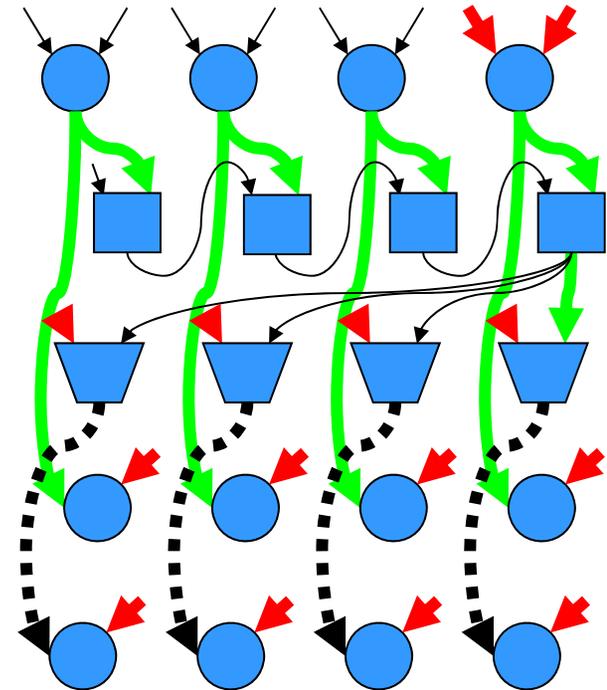
OmpSs & Challenges: 2x more cores

- Flexibility to dynamically generate work and traverse the computation space
 - Asynchronous data flow
 - Overlap
 - Tolerate variability
 - Non structured parallelism
 - Look-ahead
 - Huge task window
 - Do not stall at dependences
 - See what will have to be executed far in advance
 - Nesting
 - Top down
 - All levels contribute
 - Parallelize overheads



OmpSs & Challenges: 1/2 on chip memory

- Potential to automatically implement
 - Prefetch 
 - Reuse 
- Runtime responsibilities
 - Replication management, coherence + consistency
 - Example techniques
 - Minimize reuse distance
 - Lazy write-back
 - Data bypassing

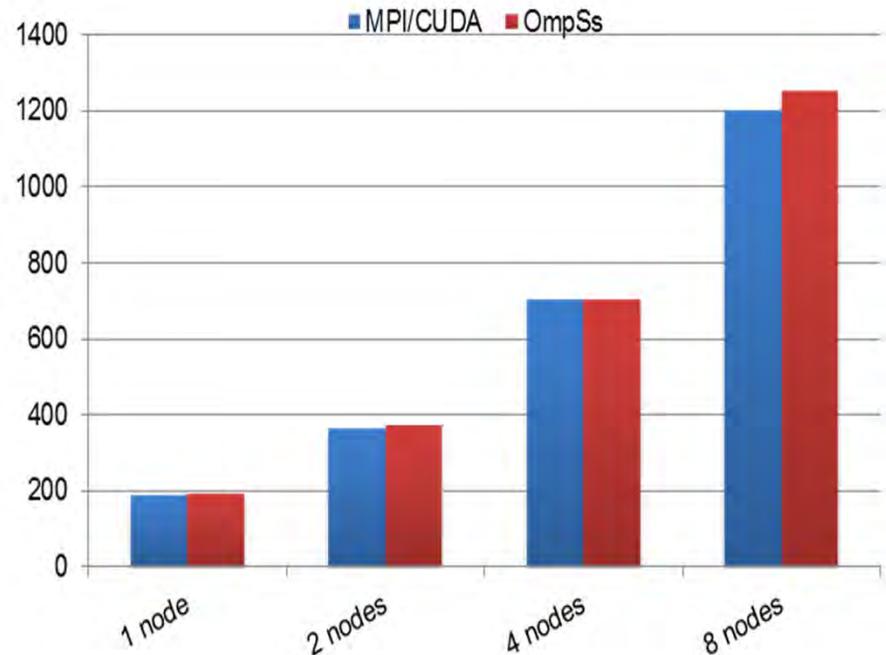


P. Bellens, et al, **CellSs: Scheduling Techniques to Better Exploit Memory Hierarchy**. Sci. Prog. 2009

P. Bellens, J.M. Pérez, R.M. Badia, J. Labarta: **Making the Best of Temporal Locality: Just-in-Time Renaming and Lazy Write-Back on the Cell/B.E.** IJHPCA 25(2): 137-147 (2011)

OmpSs & Challenges: 8x address spaces

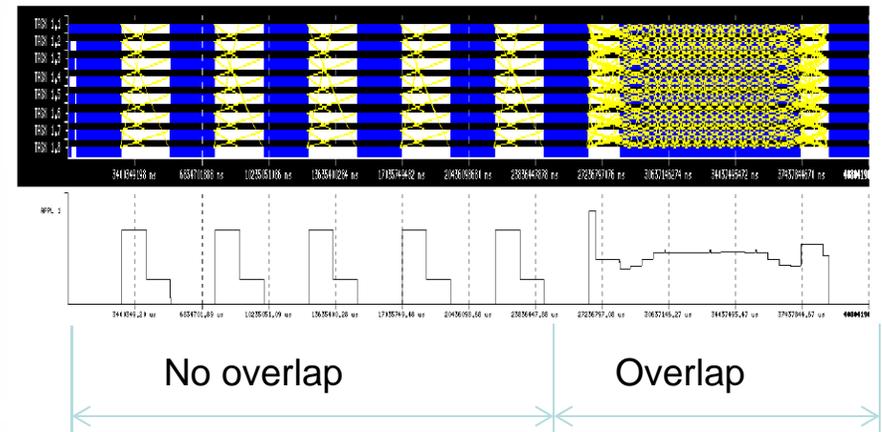
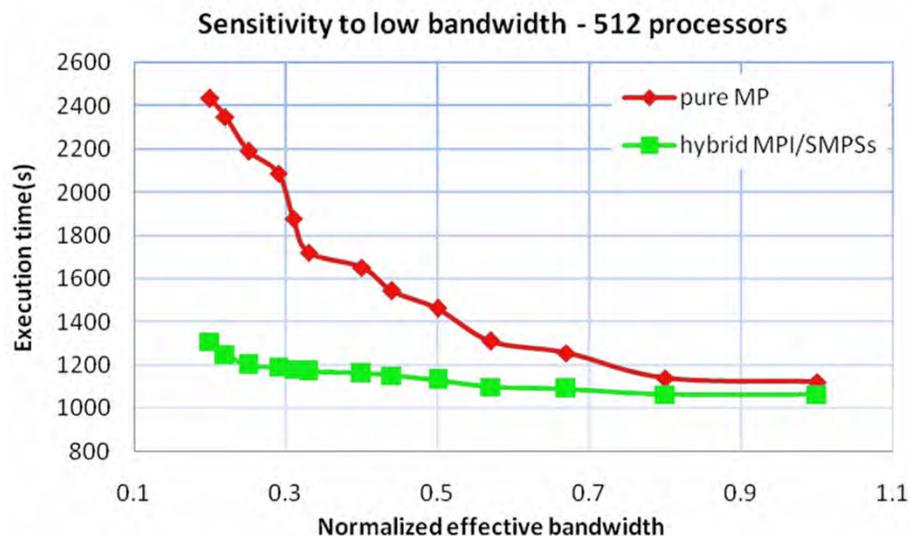
- OmpSs @ Cluster
 - Handles replication and copies
 - Handle coherency and consistency
 - Optimize for locality and reuse
- A single “shared memory” node
 - Built of several separated address spaces
 - Built from heterogeneous nodes
 - CPU + GPU



J. Bueno, A. Duran, R.M. Badia, X. Martorell, E. Ayguade, J. Labarta.
Productive Programming of GPU Clusters with OmpSs. IPDPS'12.

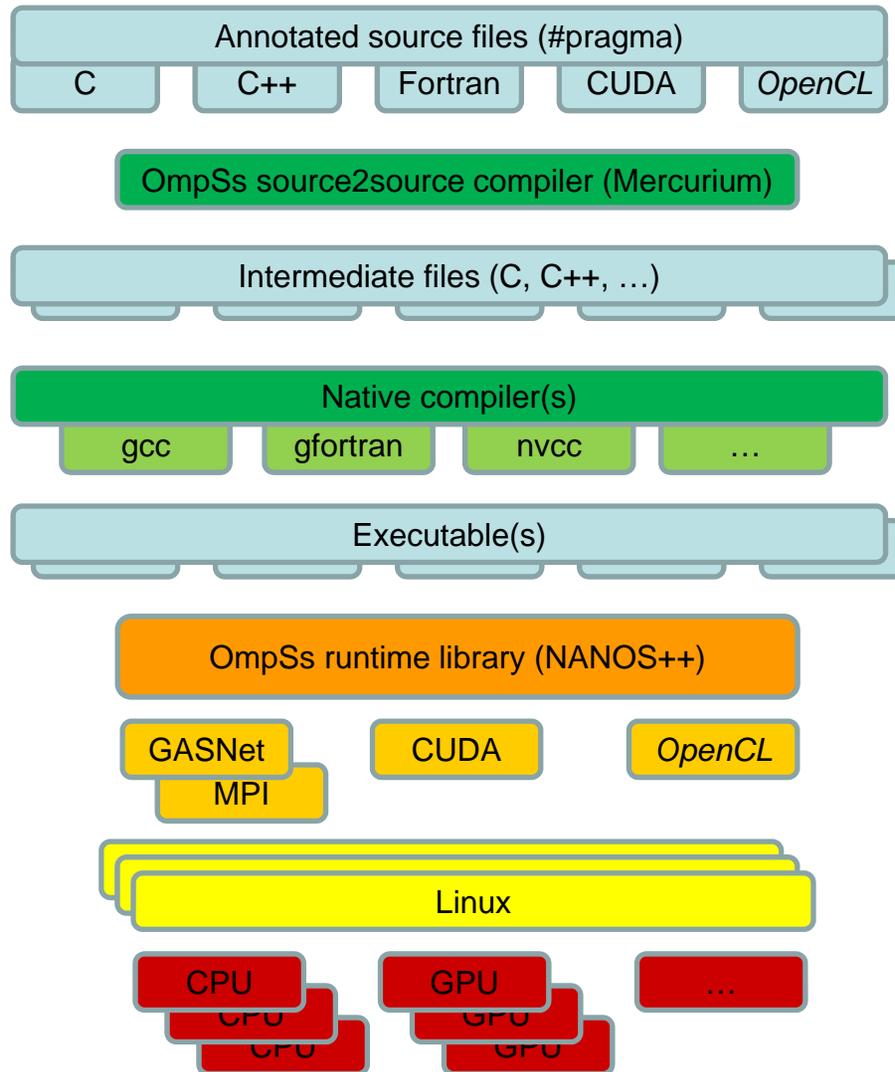
OmpSs & Challenges: Slow interconnect

- Hybrid MPI + OmpSs:
 - Encapsulate MPI messaging into asynchronous tasks
 - Propagate asynchronous behavior to MPI level
 - Overlap communication with computation
 - Hide long network latency and low bandwidth



V. Marjanovic, J. Labarta, E. Ayguadé, M. Valero: **Overlapping communication and computation by using a hybrid MPI/SMPs approach.** ICS 2010: 5-16

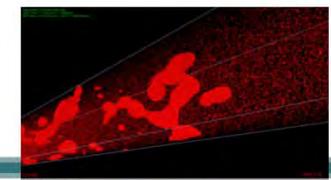
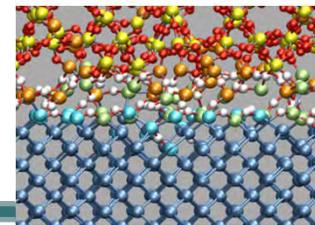
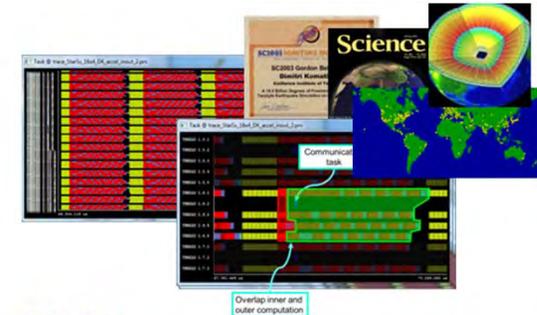
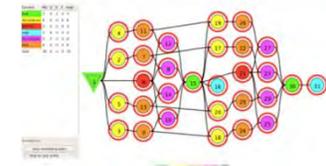
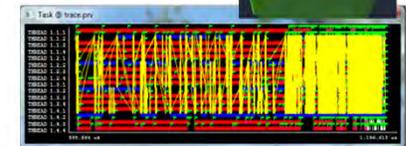
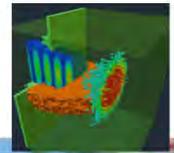
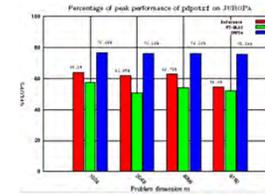
OmpSs runtime layer manages architecture complexity



- Programmer exposed a simple architecture
- Task graph provides lookahead
 - Exploit knowledge about the future
- Automatically handle all of the architecture challenges
 - Strong scalability
 - Multiple address spaces
 - Low cache size
 - Low interconnect bandwidth
- Enjoy the positive aspects
 - Energy efficiency
 - Low cost

Used in projects and applications ...

- Undertaken significant efforts to port real large scale applications:
 - **te~~s~~t**
 - Scalapack, PLASMA, SPECFEM3D, LBC, CPMD PSC, PEPC, LS1 Mardyn, Asynchronous algorithms, Microbenchmarks
 - **MONT-BLANC**
 - YALES2, EUTERPE, SPECFEM3D, MP2C, BigDFT, QuantumESPRESSO, PEPC, SMMP, ProFASI, COSMO, BQCD
 - DEEP
 - NEURON, iPIC3D, ECHAM/MESSy, AVBP, TurboRVB, Seismic
 - G8_ECS
 - CGPOP, NICAM (planned) ...
 - Consolider project (Spanish ministry)
 - MRGENESIS
 - BSC initiatives and collaborations:
 - GROMACS, GADGET, WRF, ...

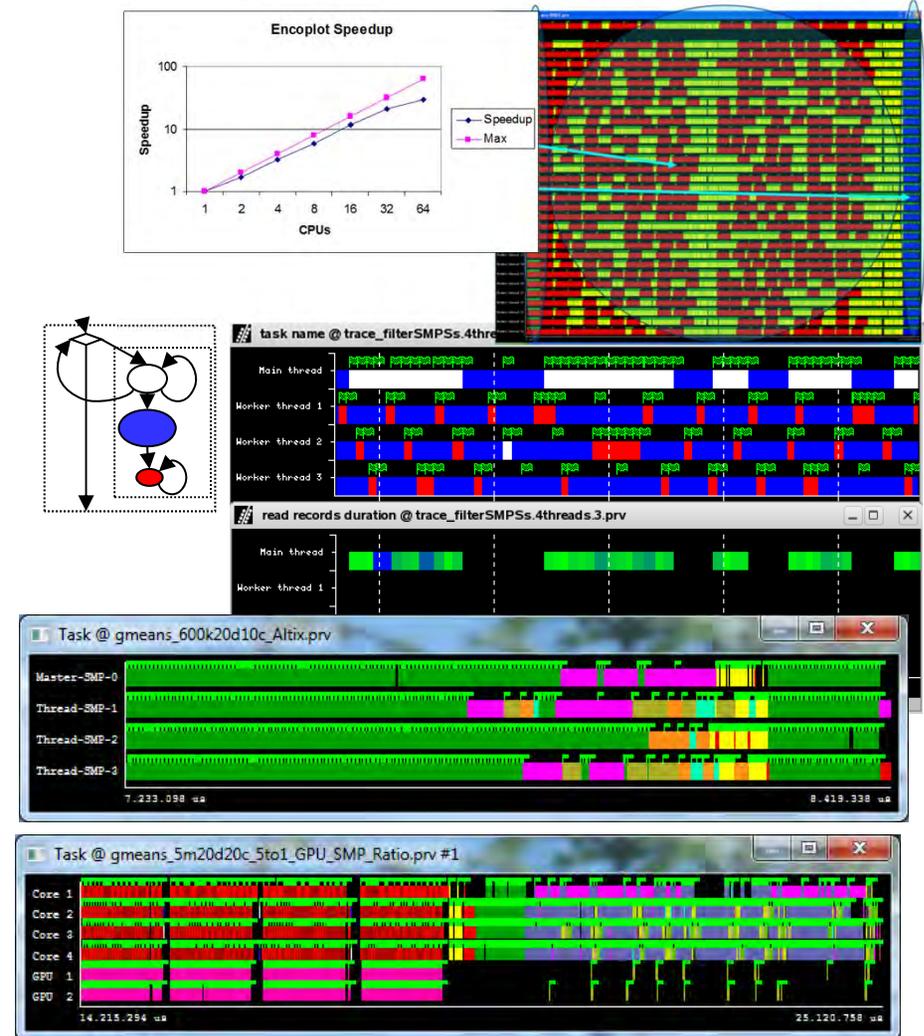


MONT-BLANC

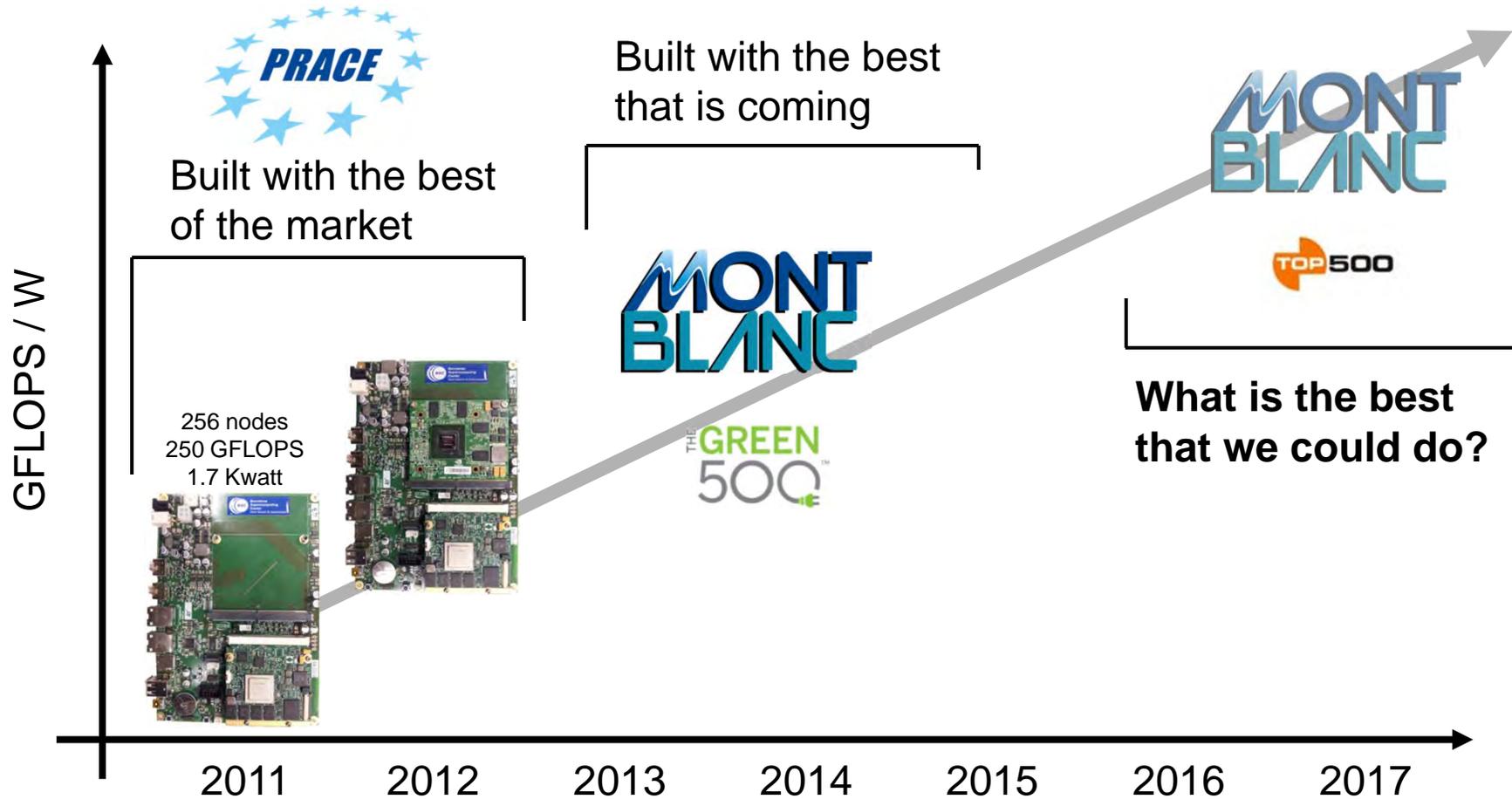
... but NOT only for «scientific computing» ...

- Plagiarism detection
 - Histograms, sorting, ... (FhI FIRST)
- Trace browsing
 - Paraver (BSC)
- Clustering algorithms
 - G-means (BSC)
- Image processing
 - Tracking (USAF)
- Embedded and consumer
 - H.264 (TUBerlin), ...

encore!



A big challenge, and a huge opportunity for Europe



- Prototypes are critical to accelerate software development
 - System software stack + applications

Very high expectations ...

- High media impact of ARM-based HPC
- Scientific, HPC, general press quote Mont-Blanc objectives
 - Highlighted by Eric Schmidt, Google Executive Chairman, at the EC's Innovation Convention

THE WALL STREET JOURNAL

Europe Edition Home | Today's Paper | Video | Blogs | Emails | Journal Community | Mobile | Tablet

World | Europe | U.K. | U.S. | Business | Markets | Market Data | Tech | Life & Style

TOP STORIES IN Technology

U.S. Alleges Collusion On E-Book Prices

Seeking 'Second' Life After Facebook

WSJ BLOGS

Digits
Technology News and Insights

November 14, 2011, 3:35 PM

Barcelona Center Makes Super Bet on Cellphone Chips

Article | Comments (1)

Email | Print | Like | Send | + More | Text

By Don Clark

Supercomputers, once built from handcrafted circuitry, were transformed when companies started assembling them from inexpensive PC-style microprocessors. Researchers in Barcelona are placing an early bet that the next big leap will be cellphone chips.

The [Barcelona Supercomputing Center](#) said Monday it is developing what it believes is the first supercomputer based on the ARM Holdings chip designs used in most cellphones. BSC, as it is called, plans to start with ARM-based chips from Nvidia called Tegra as well as Nvidia graphics processing units, or GPUs—the kind of chips used in videogame systems, which are also shaking up the supercomputer market.

Behind the experiment is a power struggle—that is, a struggle to control the power consumption of supercomputers, which take up huge data centers and draw the



Científicos líderes. Ales Ramirez, jefe de equipo del Barcelona Supercomputing Center, en la UPC. Detrás de él, la joya de la entidad, el Mare Nostrum

El supercibercerebro

Barcelona construye el primer megordenador del mundo basado en teléfonos móviles

OSCAR NUÑOZ

En un pequeño cuarto de uno de los edificios consagrados a la investigación del Campus Nord de la Universidad Politécnica de Catalunya (UPC) hasta ahora lo que para muchos es una sala de aula...

za a la supremacía tecnológica estadounidense y así como en el campo de los ordenadores gigantes. Allí, un equipo de científicos construye el que será el primer superordenador del mundo basado en los teléfonos móviles. La idea es aprovechar la eficiencia en el consumo energético de los

smartphones y de las tabletas que el mayor parte del tiempo no están encendidas a la red eléctrica y funcionan sin sobrecalentarse, para aumentar la capacidad de cálculo sin depender el gasto energético. Todo un reto que debe dar respuesta a las necesidades «crescien» de las empresas

e instituciones, que piden programas, simulaciones cada vez más complicadas. Es uno de los objetivos del jefe Mont-Blanc, liderado por el reto y que da cuenta de la apuesta que la capital catalana debe dar respuesta a las necesidades «crescien» de las empresas

CONTINUA EN LA PAGINA SIGUIENTE

WIRED ENTERPRISE

IT HAPPENS

PREVIOUS POST

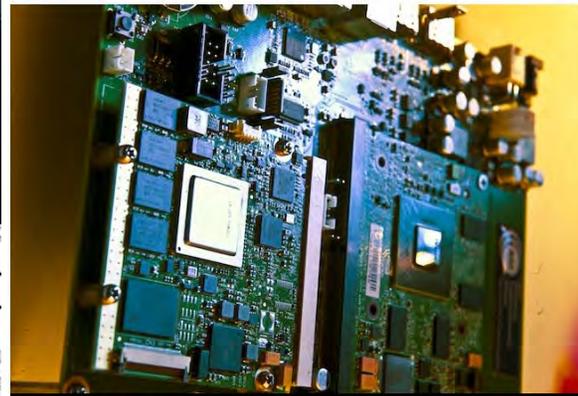
Barcelona Supercomputer ARMed For Assault on World's Fastest Machines

By Robert McMillan | April 3, 2012 | 6:30 am | Categories: Hardware, Microprocessors, Servers

Follow @bobcmillan

Jorge Naranjo and 130 others like this.

210 26 34



A Tegra 2 system with a GPU processor. Is this the future of supercomputing? Photo: Barcelona Supercomputer Center

From **mobile phone** to supercomputer?

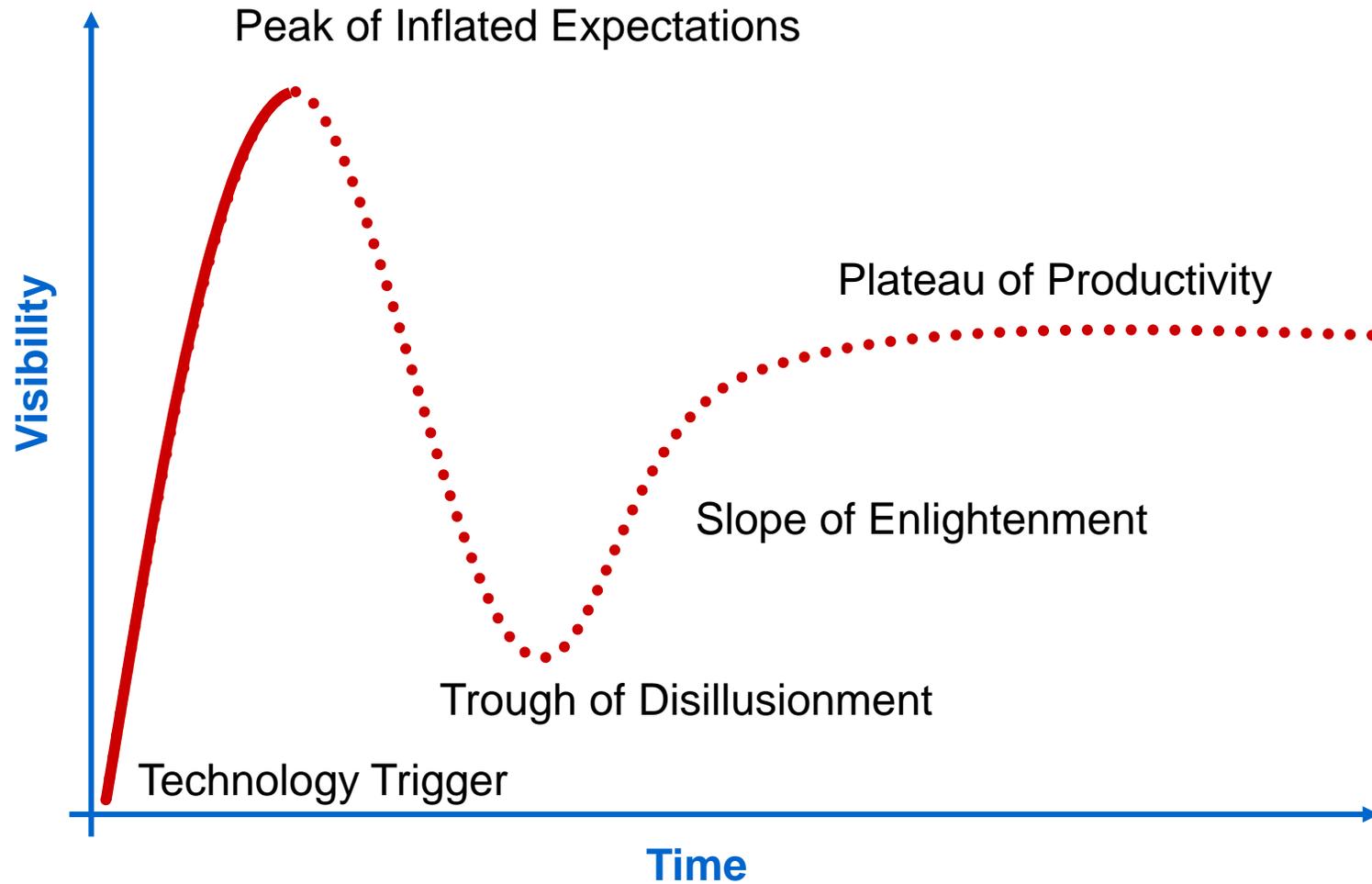
Tom Wilkie looks at the emerging strategies for Exascale computing

machines consume somewhere around 5MW to 10MW of power annually, costing between \$5M and \$10M at current US prices. Exascale machines will run a thousand times faster, so no one can afford simply to scale up existing technology for no-one could bear annual electricity bills of more than \$5 billion.

are currently used in mobile phones and embedded applications because these guys have been facing power-density limitations from the beginning - they work with battery-operated devices where energy consumption was always an issue. But Ramirez does not believe that it will

in early February, just after this issue of

The hype curve



- We'll see how deep it gets on the way down ...

Project goals

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



- Objectives
 - Develop a first prototype system, limited by available technology
 - Design a Next Generation system, to overcome the limitations
 - Develop a set of Exascale applications targeting the new system

Conclusions

- Mont-Blanc architecture is shaping up
 - ARM multicore + integrated accelerator
 - Ethernet NIC
 - High density packaging
- OmpSs programming model to handle hardware challenges
- Many important decisions still pending
 - Contacting providers
 - Comparing alternatives

- Stay tuned!



www.montblanc-project.eu



MontBlancEU



@MontBlanc_EU