www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The Mont-Blanc approach towards Exascale

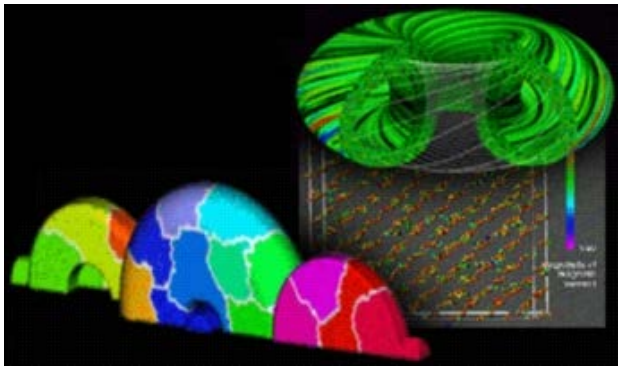
Prof. Mateo Valero
Director

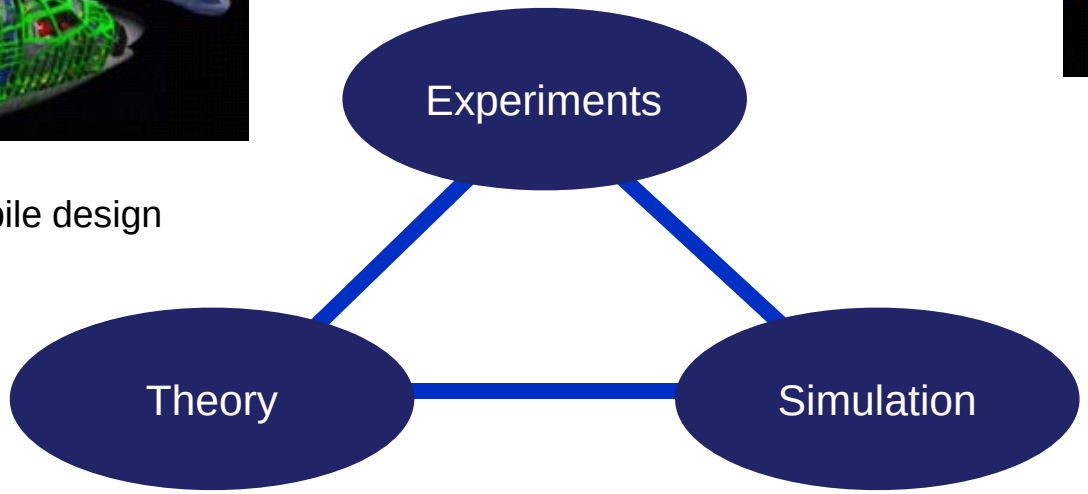
Disclaimer:
All references to unavailable products are speculative, taken from web sources. There is no commitment from ARM, Samsung, Intel, or others implied.

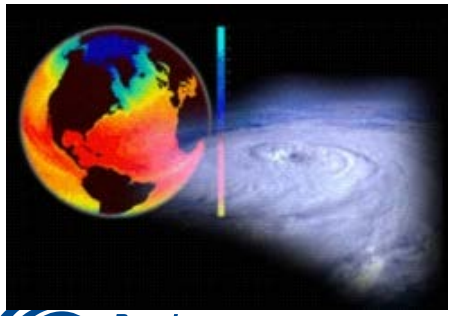
Supercomputers, Theory and Experiments

Aircraft
Automobile design

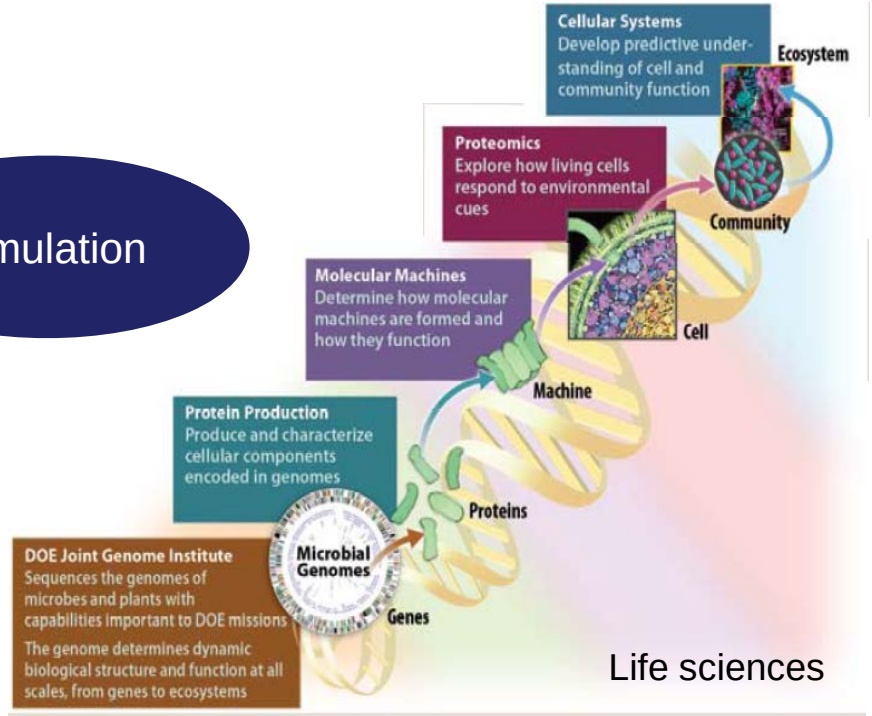
Fusion reactor
Accelerator design
Material sciences
Astrophysics





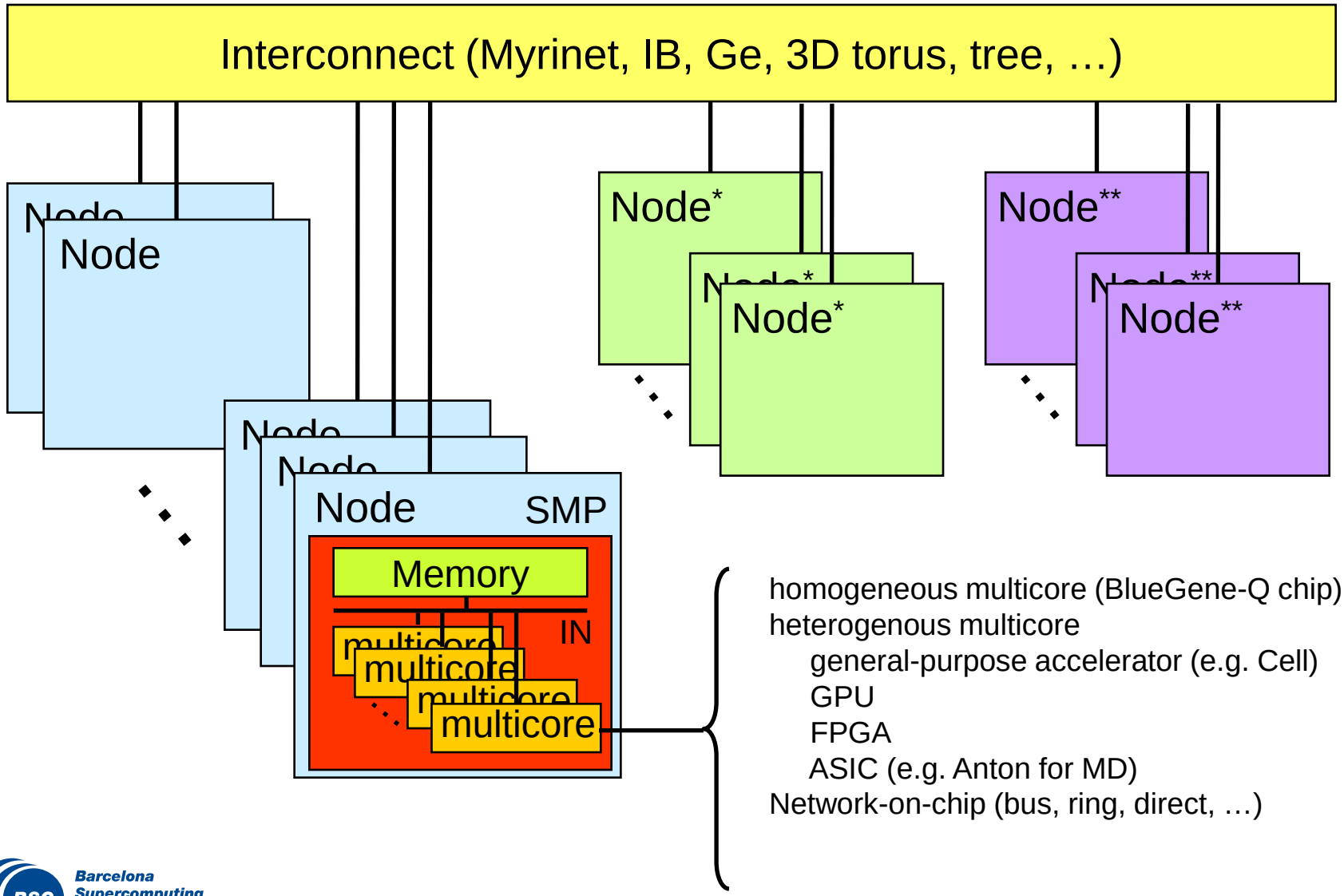


Climate and weather modeling



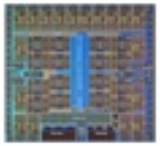
Life sciences

Parallel Systems



Blue Gene/Q packaging hierarchy

1. Chip
16 cores



2. Module
Single Chip

3. Compute Card
One single chip module,
16 GB DDR3 Memory

4. Node Card
32 Compute Cards,
Optical Modules, Link Chips,
Torus

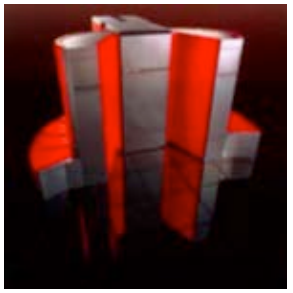
5b. I/O Drawer
8 I/O Cards
8 PCIe Gen2 slots

5a. Midplane
16 Node Cards

6. Rack
2 Midplanes
1, 2 or 4 I/O Drawers

7. System
20PF/s

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 v
    the locally self-consistent multiple scattering met

- 1 PFlop/s; 2008; Cray XT5; 1.5×10^5 Processors
  - Superconductive materials

- 1 EFlop/s; ~2018; ?; 1×10^8 Processors?? (10^9 th

Top10 TOP500 SUPERCOMPUTER SITES June 2012 list

Rank	Site	Computer	Procs	Rmax	Rpeak	Power	GFlops/Watt
1	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	1572864	16,32	20,13	7,89	2,06
2	RIKEN Advanced Institute for Computational Science (AICS)	Fujitsu, K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	705024	10,51	11,28	12,65	0,83
3	DOE/SC/Argonne National Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	786432	8,16	10,06	3,94	2,07
4	Leibniz Rechenzentrum	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050	147456	2,89	3,18	3,51	0,82
5	Tianjin, China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050	186368 100352	2,56	4,70	4,04	0,63
6	DOE/SC/Oak Ridge National Laboratory	Cray XK6, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA 2090	298592	1,94	2,62	5,14	0,37
7	CINECA	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	111104	1,72	2,09	0,82	2,09
8	Forschungszentrum Juelich (FZJ)	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	131072	1,38	1,67	0,65	2,12
9	CEA/TGCC-GENCI	Bullx B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR	77184	1,35	1,66	2,25	0,60
10	Shenzhen, China	XeonX5670+NVIDIA	120640 64960	1,27	2,98	2,58	0,49

Outline

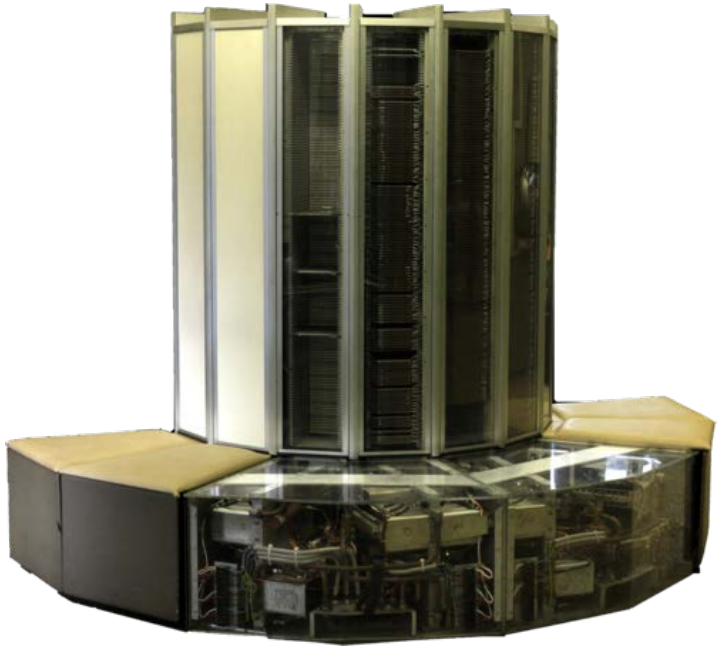
- Brief history of supercomputing
  - Special purpose
  - The “Killer-Micros”
  - The “Killer-Mobiles(tm)"
- Power and cost define performance
- Supercomputers from mobile devices
  - Strawman design
  - Challenges identified
- Addressing the challenges with OmpSs
- The Mont-Blanc roadmap

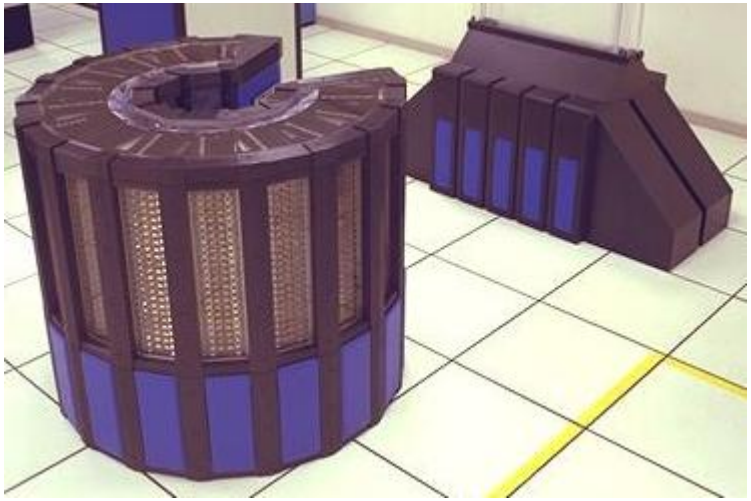
In the beginning ... there were only supercomputers

- Built to order
  - Very few of them
- Special purpose hardware
  - Very expensive

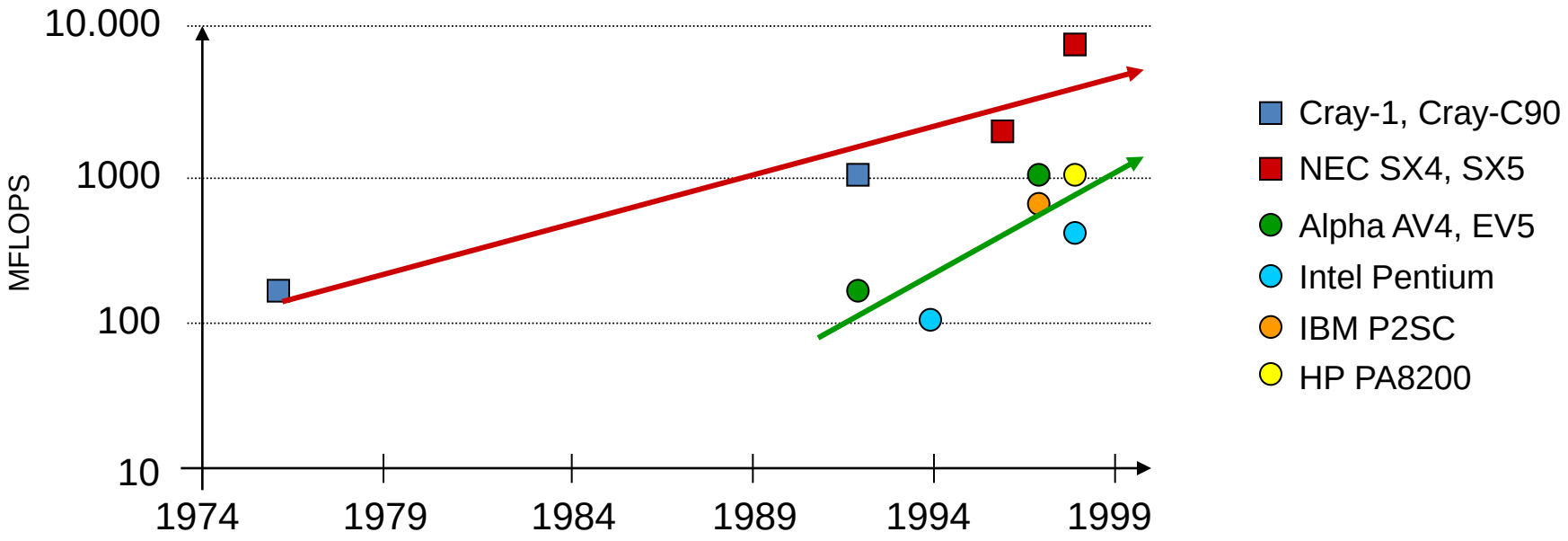
- Control Data
- Cray-1
  - 1975, 160 MFLOPS
    - 80 units, 5-8 M$
- Cray X-MP
  - 1982, 800 MFLOPS
- Cray-2
  - 1985, 1.9 GFLOPS
- Cray Y-MP
  - 1988, 2.6 GFLOPS

- ...Fortran+ Vectorizing Compilers





Killer-Micros



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

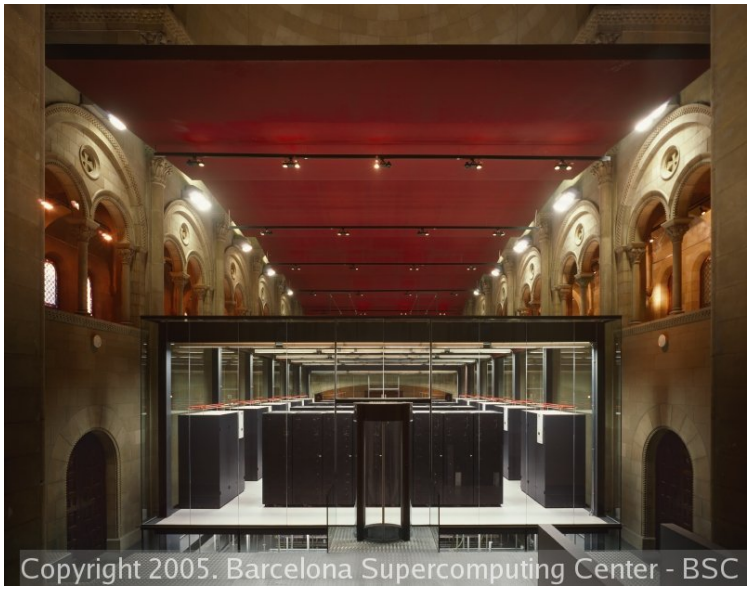
Then, commodity took over special purpose

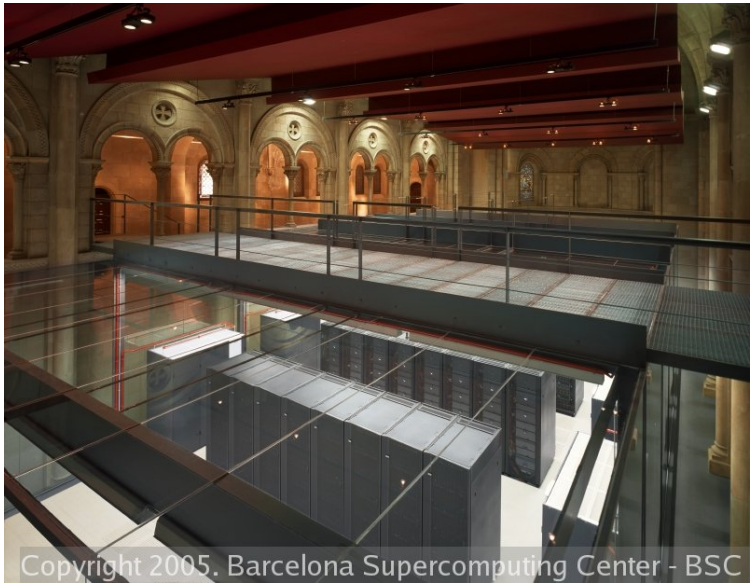
- ASCI Red, Sandia
  - 1997, 1 Tflops (Linpack), 9298 processors at 200 Mhz, 1.2 Tbytes
  - Intel Pentium Pro
    - Upgraded to Pentium II Xeon, 1999, 3.1 Tflops
- ASCI White, Lawrence Livermore Lab.
  - 2001, 7.3 TFLOPS, 8192 proc. RS6000 at 375 Mhz, 6 Terabytes, (3+3) Mwats
  - IBM Power 3

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





Green/Top 500 June 2012

Green500 Rank	Top500 Rank	Mflops/watt	Power	Site	Computer
1	252	2100,88	41,1	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
2	253	2100,88	41,1	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
3	99	2100,86	82,2	DOE/SC/Argonne National Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
4	100	2100,86	82,2	DOE/SC/Argonne National Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
5	102	2100,86	82,2	Rensselaer Polytechnic Institute	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
6	103	2100,86	82,2	University of Rochester	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
7	101	2100,86	82,2	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom
8	20	2099,56	493,1	University of Edinburgh	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
9	13	2099,50	575,3	Science and Technology Facilities Council - Daresbury Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
10	8	2099,46	657,5	Forschungszentrum Juelich (FZJ)	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
11	7	2099,39	821,9	CINECA	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
12	36	2099,14	246,6	High Energy Accelerator Research Organization /KEK	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
13	29	2099,14	328,8	EDF R&D	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
14	30	2099,14	328,8	IDRIS/GENCI	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
15	31	2099,14	328,8	Victorian Life Sciences Computation Initiative	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
16	49	2099,14	164,4	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom
17	50	2099,14	164,4	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom
18	48	2099,14	164,4	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
19	3	2069,04	3945	DOE/SC/Argonne National Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom
20	1	2069,04	7890	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom
21	150	1380,67	72,9	Intel	Intel Cluster, Xeon E5-2670 8C 2.600GHz, Infiniband FDR, Intel MIC
22	456	1379,79	47	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR
23	**177**	**1266,26**	**81,5**	**Barcelona Supercomputing Center**	**Bullx B505, Xeon E5649 6C 2.53GHz, Infiniband QDR, NVIDIA 2090**
24	41	1151,91	366	Center for Computational Sciences, University of Tsukuba	Xtream-X GreenBlade 8204, Xeon E5-2670 8C 2.600GHz, Infiniband QDR, NVIDIA 2090
25	74	1050,26	226,8	Los Alamos National Laboratory	Xtreme-X , Xeon E5-2670 8C 2.600GHz, Infiniband QDR, NVIDIA 2090
36	**358**	**932,19**	**79,8**	**Universidad de Cantabria - SSC**	**iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR**
60	2	830,18	12659,9	RIKEN Advanced Institute for Computational Science (AICS)	SPARC64 VIIIfx 2.0GHz, Tofu interconnect
103	**375**	**467,73**	**154**	**CeSViMa - Centro de Supercomputación y Visualización de Madrid**	**BladeCenter PS702 Express, Power7 3.3GHz, Infiniband**
408	**465**	**93,37**	**683,59**	**Barcelona Supercomputing Center**	**BladeCenter JS21 Cluster, PPC 970, 2.3 GHz, Myrinet**

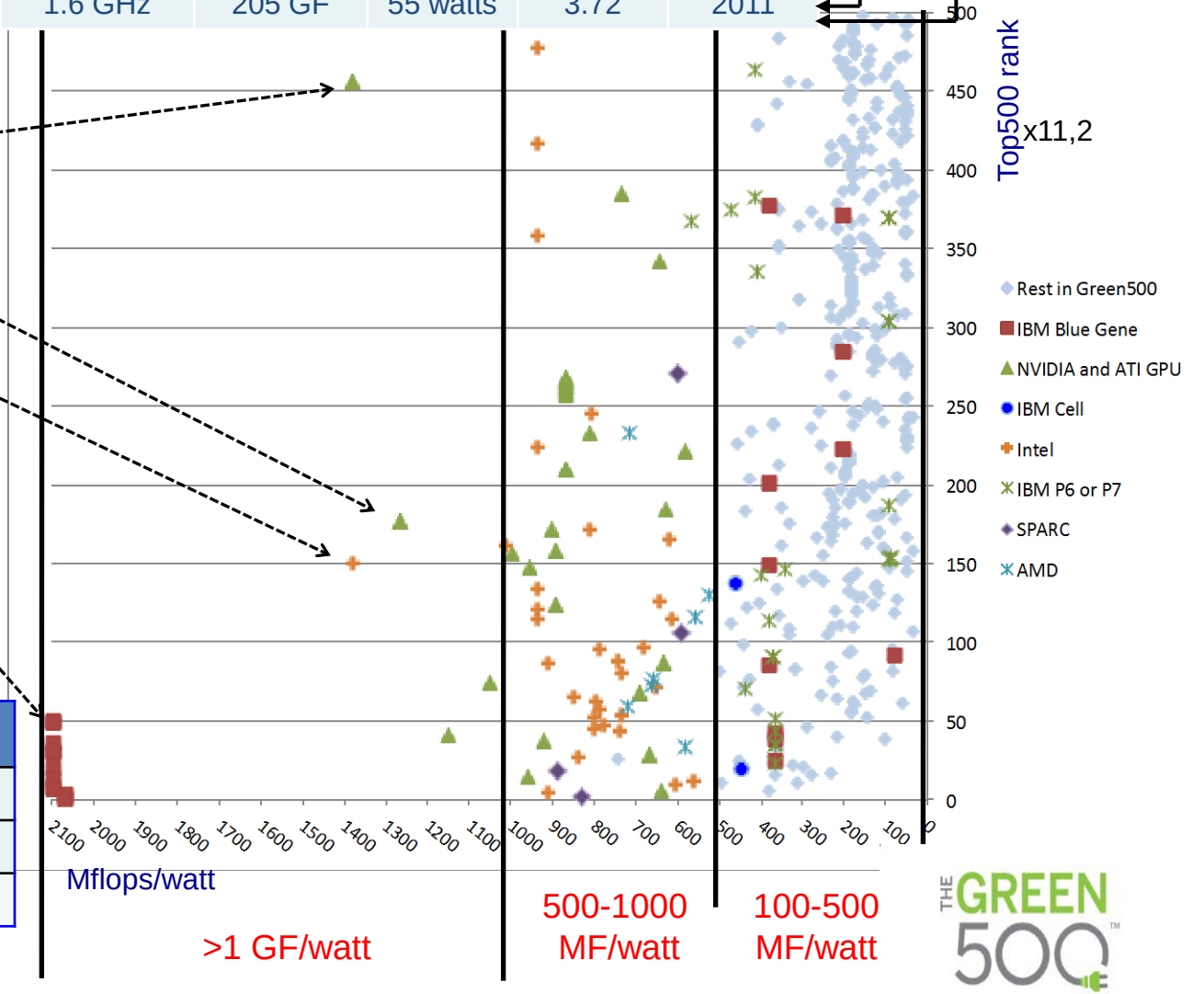
June 2012

Processor	CPU Type	Cores	CPU Speed	Peak FP	Power	Gflops/ wat	Date
BG/L	PPC 440	2	700 MHz	5.6 GF	17 watts	0.33	2004
BG/P	PPC 450	4	850 MHz	13.6 GF	16 watts	0.85	2007
BG/Q	PPC A2	18	1.6 GHz	205 GF	55 watts	3.72	2011

x2,6

x4,4

x11,2

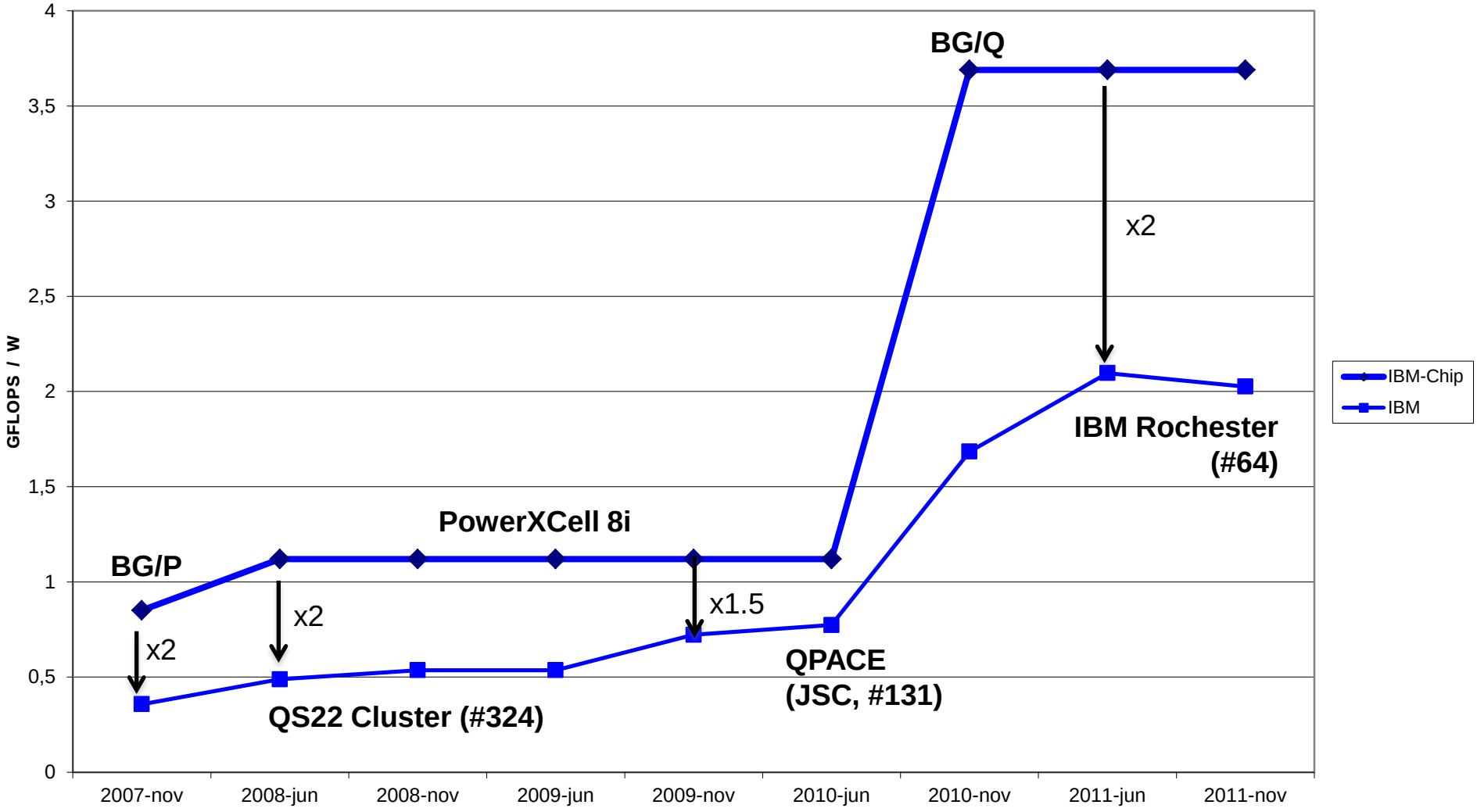


Mflops/watt	Mwatts/Exaflop
2100,86	476
1380,67	724
1379,78	725

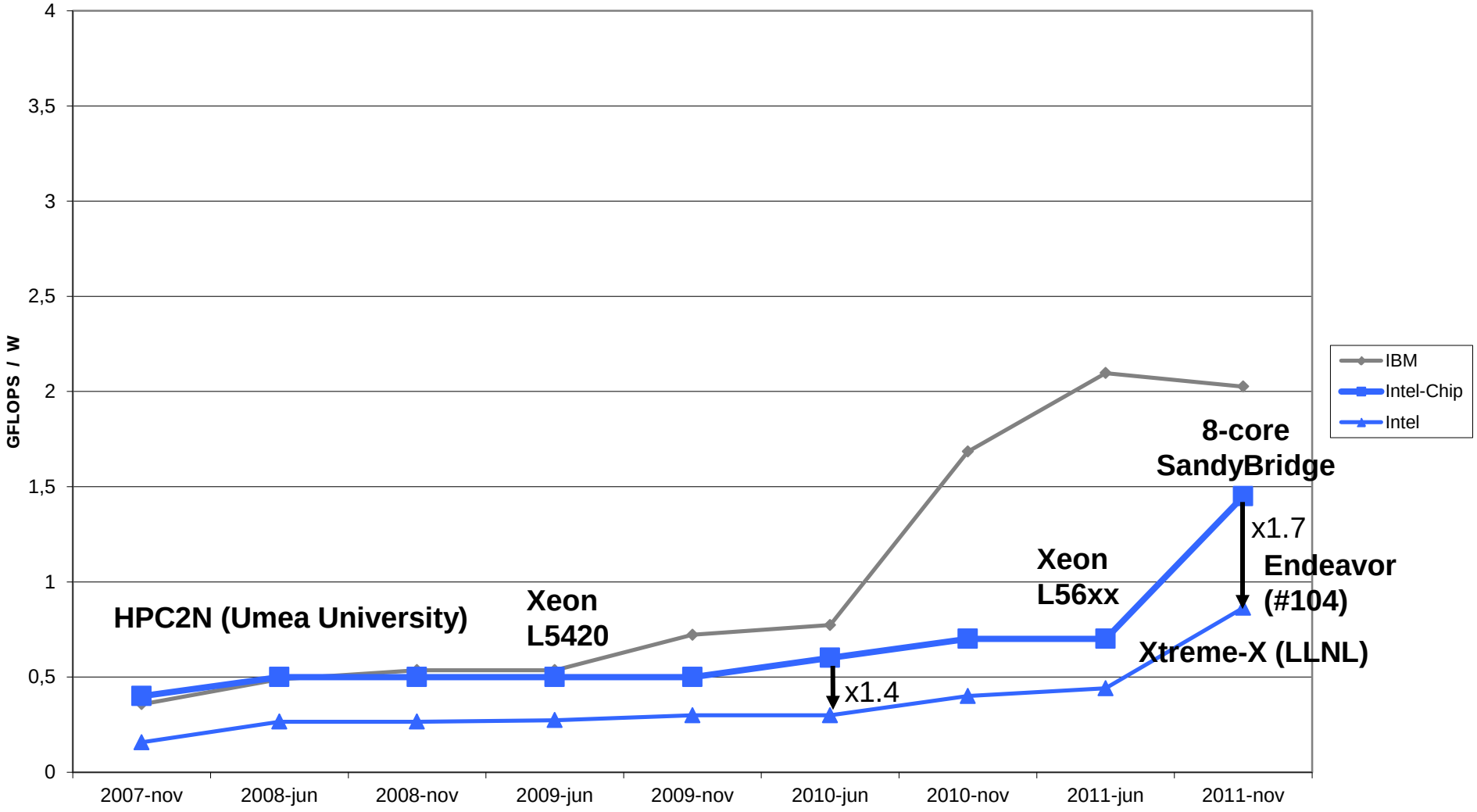
THE GREEN 500

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

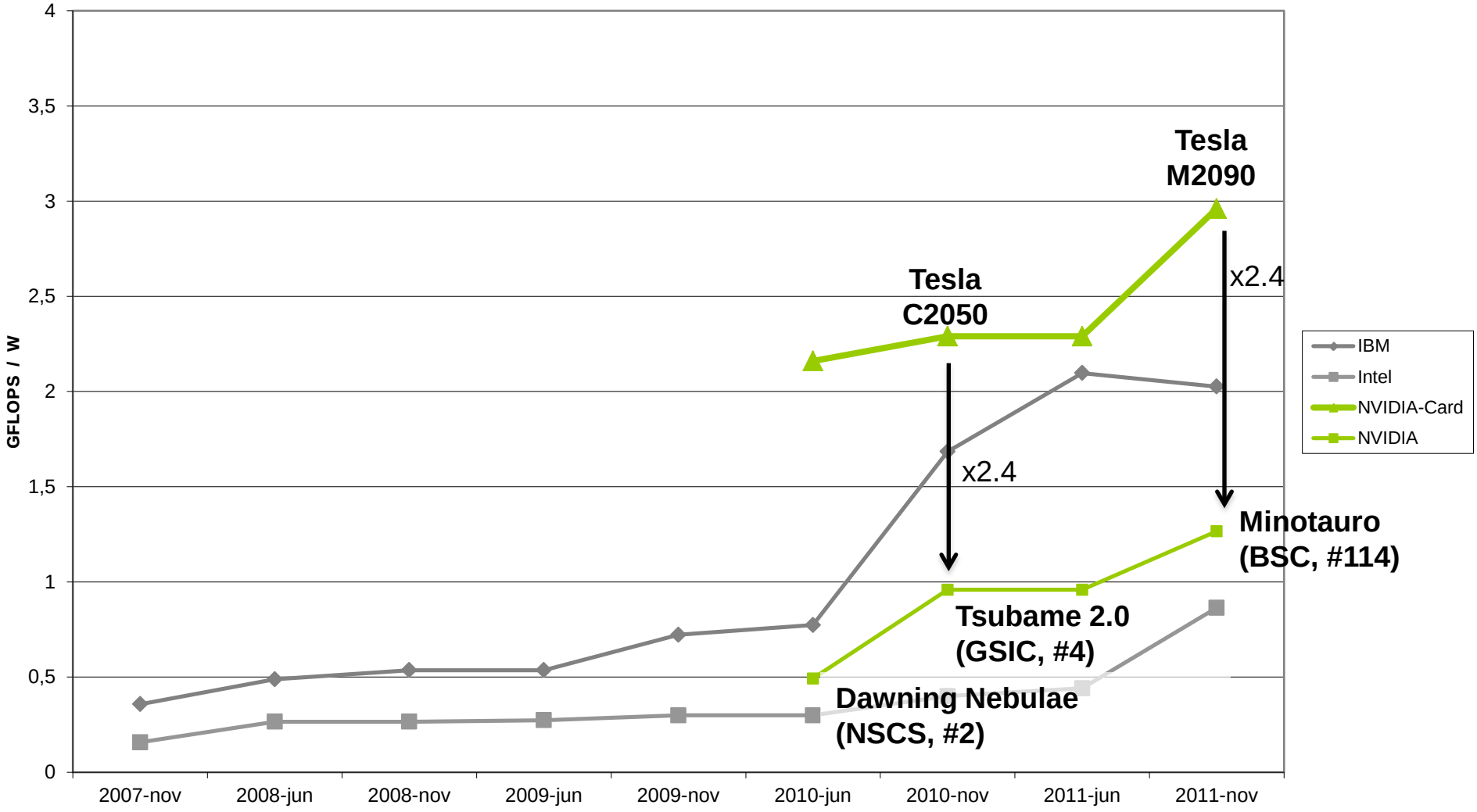
Green500 evolution, chip vs. system: IBM



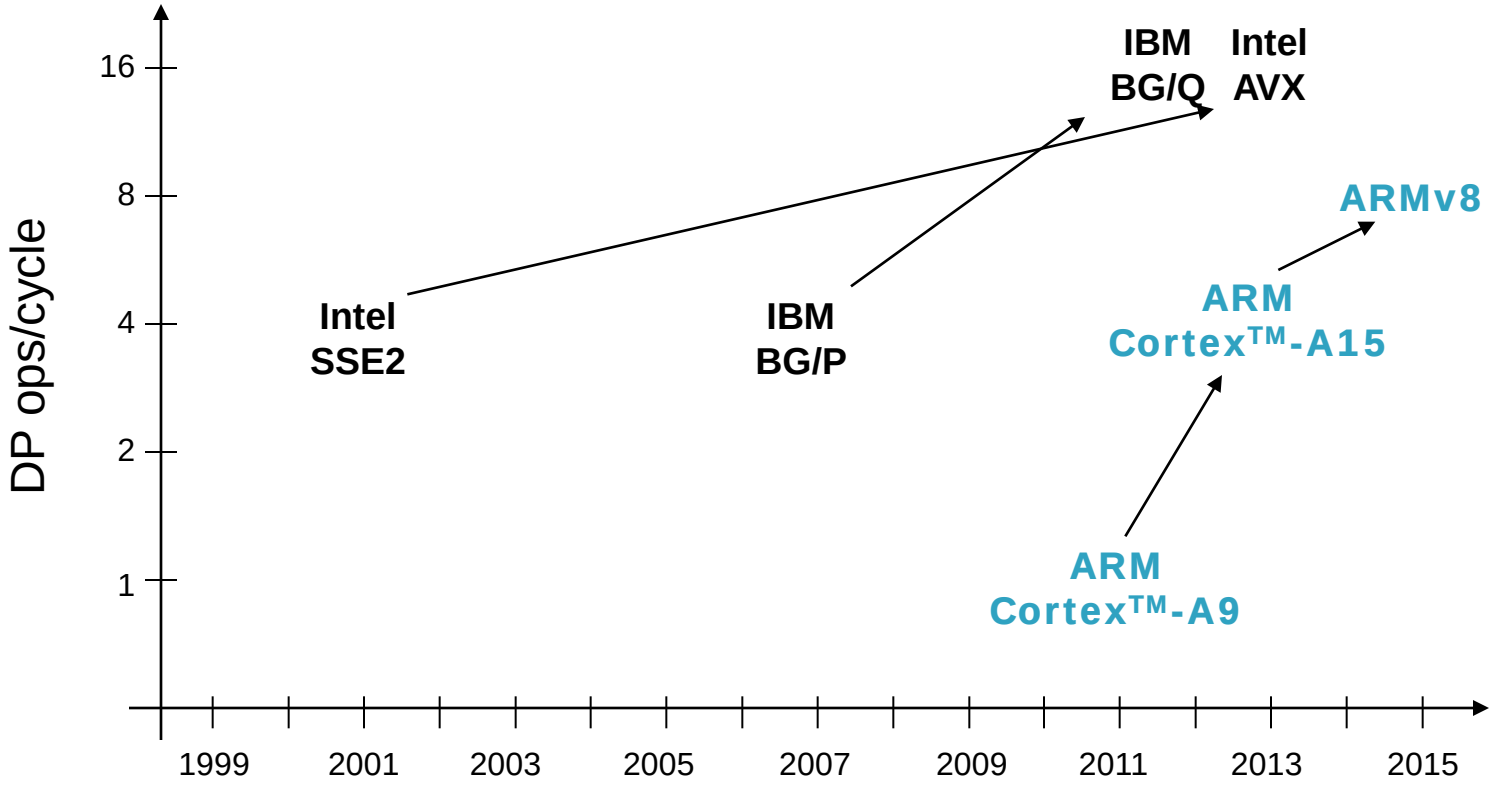
Green500 evolution, chip vs. systems: Intel



Green500 evolution, chip vs. system: Nvidia GPU

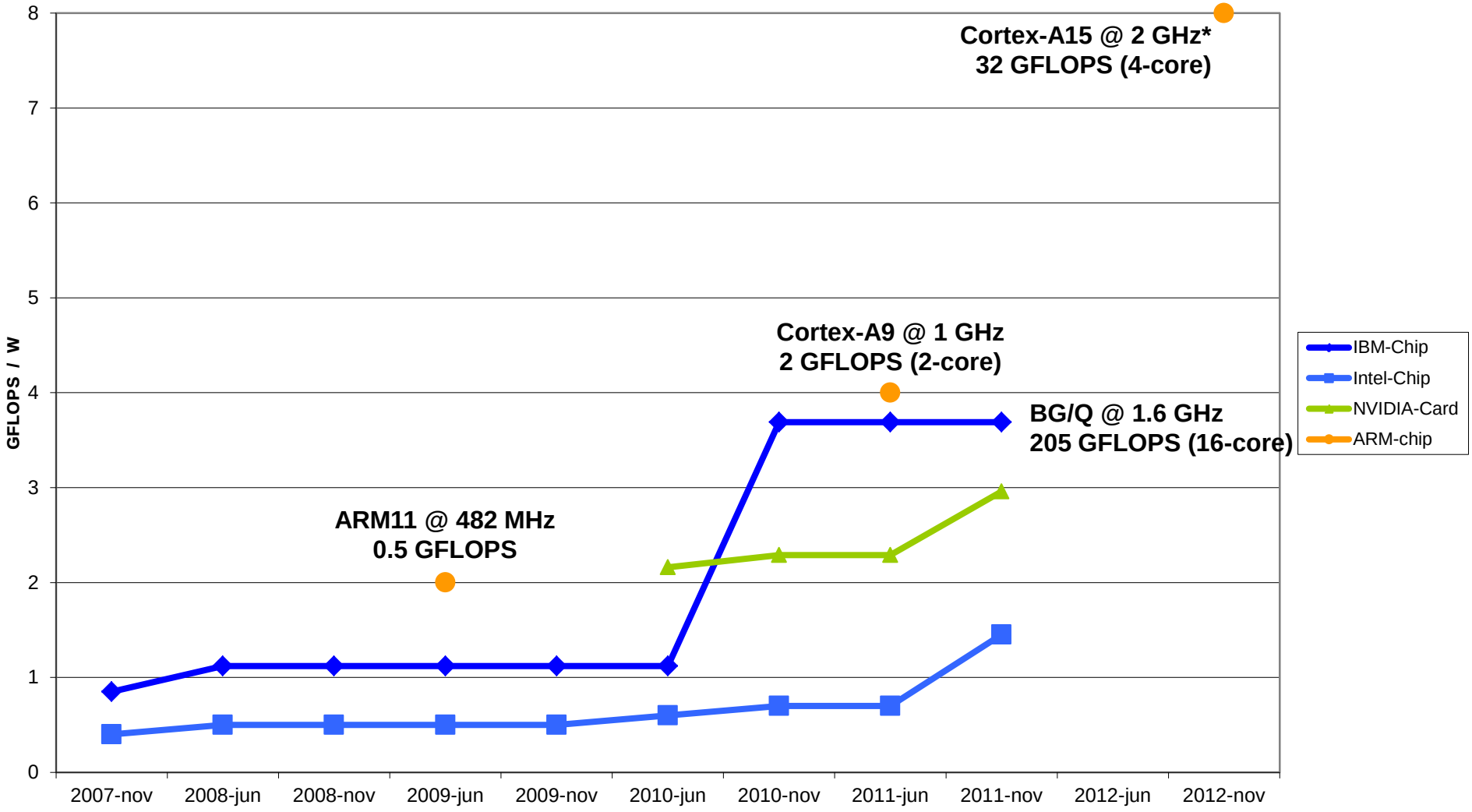


ARM Processor improvements in DP FLOPS



- 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)

ARM processor efficiency vs. IBM / Intel / Nvidia



* Based on ARM Cortex-A9 @ 2GHz power consumption on 45nm, not an ARM comitment

Comparing embedded to HPC cores

CPU	#cores	GFLOPS	On-chip Memory	Off-chip GB/s	Watts	Cost
Intel Sandy Bridge	8	185	22 MB 0.12 B/KFLOP	68 0.37 B/FLOP	135 1.37 GFLOPS/W	~ $1500
AMD Bulldozer	16	133	32 MB 0.24 B/KFLOP	102 0.77 B/FLOP	140 0.95 GFLOPS/W	~ $1500
Fujitsu UltraSPARC VIIIfx	16	128	12 MB 0.09 B/KFLOP	64 0.50 B/FLOP	110 1.16 GFLOPS/W	?
IBM BlueGene/Q	18	205	32 MB 0.15 B/KFLOP	42 0.20 B/FLOP	55 3.72 GFLOPS/W	?
ARM Cortex-A15 @ 2GHz	4	32	2 MB 0.06 B/KFLOP	12.8 0.40 B/FLOP	4* 8 GFLOPS/W	~ $20

* Based on ARM Cortex-A9 @ 2GHz power consumption on 45nm, not an ARM comitment

- Lower per-chip performance
- Lower on-chip cache size

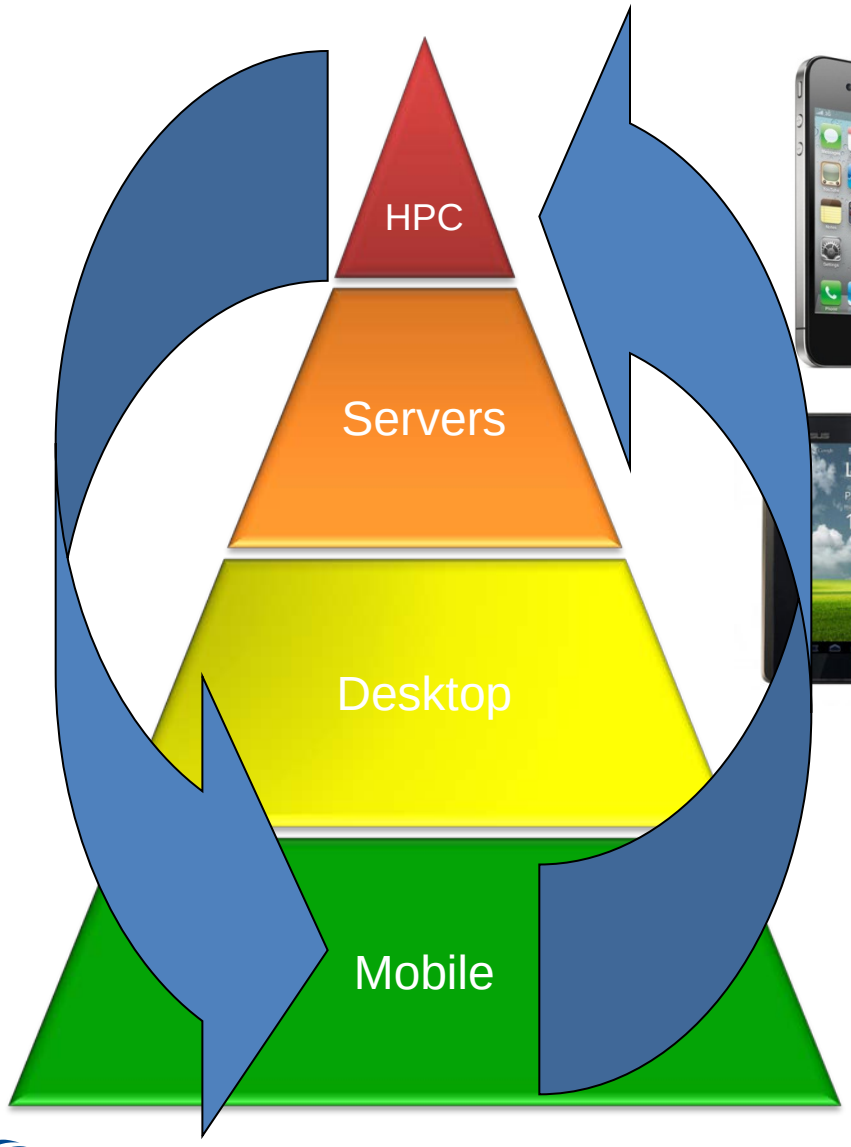
Comparing embedded to HPC cores

CPU	#cores	GFLOPS	On-chip Memory	Off-chip GB/s	Watts	Cost
Intel Sandy Bridge	8	185	22 MB 0.12 B/KFLOP	68 0.37 B/FLOP	135 1.37 GFLOPS/W	~ $1500
AMD Bulldozer	16	133	32 MB 0.24 B/KFLOP	102 0.77 B/FLOP	140 0.95 GFLOPS/W	~ $1500
Fujitsu UltraSPARC VIIIfx	16	128	12 MB 0.09 B/KFLOP	64 0.50 B/FLOP	110 1.16 GFLOPS/W	?
IBM BlueGene/Q	18	205	32 MB 0.15 B/KFLOP	42 0.20 B/FLOP	55 3.72 GFLOPS/W	?
ARM Cortex-A15	4	32	2 MB 0.06 B/KFLOP	12.8 0.40 B/FLOP	4* 8 GFLOPS/W	~ $20

* Based on ARM Cortex-A9 @ 2GHz power consumption on 45nm, not an ARM comitment

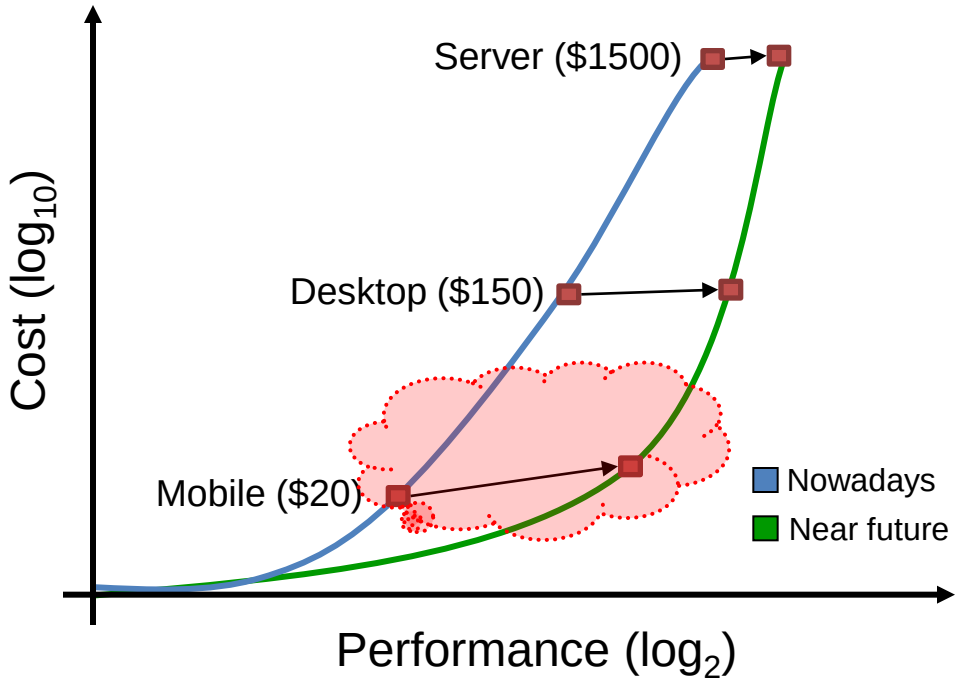
- Lower per-chip performance
- Lower on-chip cache size
- Higher energy efficiency
- Much lower cost

Can we build a supercomputer with mobile chips?



- Total cores in Jun'12 Top500
  - 13.5 Mcores (9.2 in Nov'11)
- Tablets sold in Q4 2011
  - 27 Mtablets
- Smartphones sold Q4 2011
  - > 100 Mphones

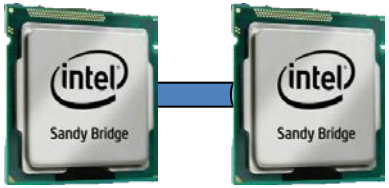
History repeats itself: are the “Killer-Mobiles(tm)" coming?

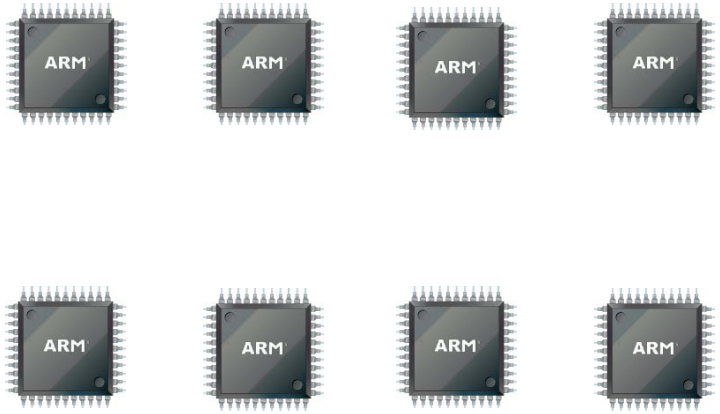


“Live by the sword, die by the sword”
--Mathew 26:52

- Where is the sweet spot? Maybe in the low-end ...
  - Today ~ 1:4 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**

Can we achieve competitive performance?

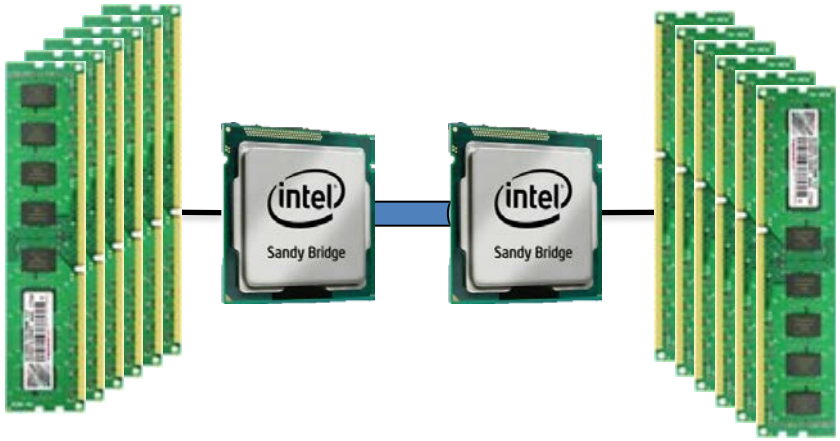


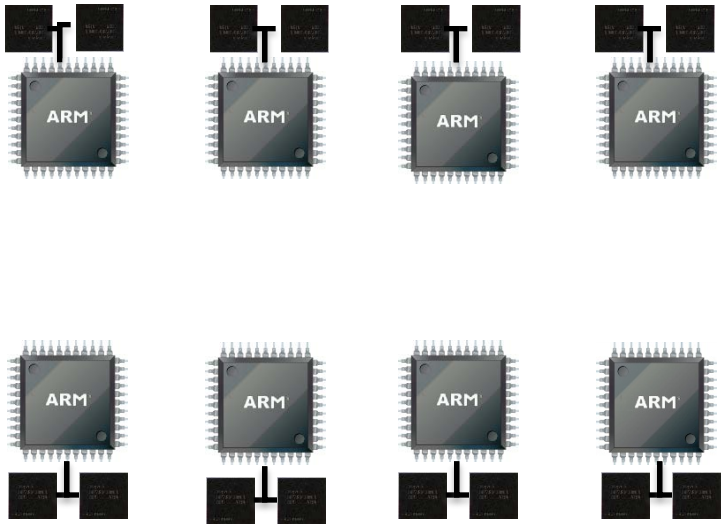


- 2-socket Intel Sandy Bridge
  - 370 GFLOPS
  - 1 address space
- 8-socket ARM Cortex A-15
  - 256 GFLOPS
  - 8 address spaces

Strawman design, not the actual Mont-Blanc architecture

How about cache size? And memory bandwidth?

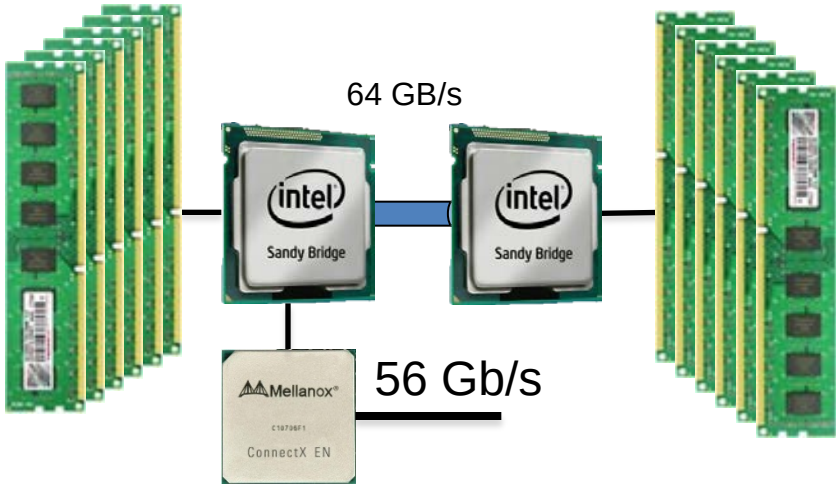


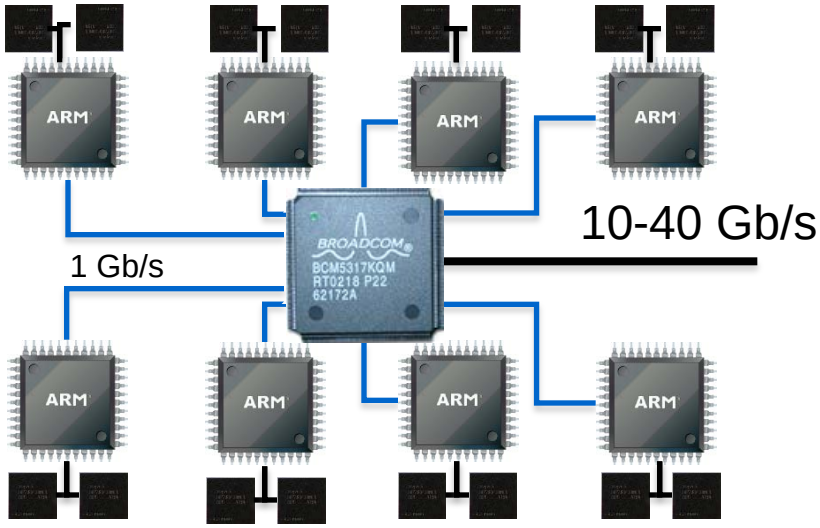


- 2-socket Intel Sandy Bridge
  - 44 MB on-chip memory
    - 0.12 bytes / KFLOP
  - 136 GB/s
    - 0.36 bytes / FLOP
- 8-socket ARM Cortex A-15
  - 16 MB on-chip memory
    - 0.06 bytes / KFLOP
  - 102 GB/s
    - 0.40 bytes / FLOP

Strawman design, not the actual Mont-Blanc architecture

How about interconnect?

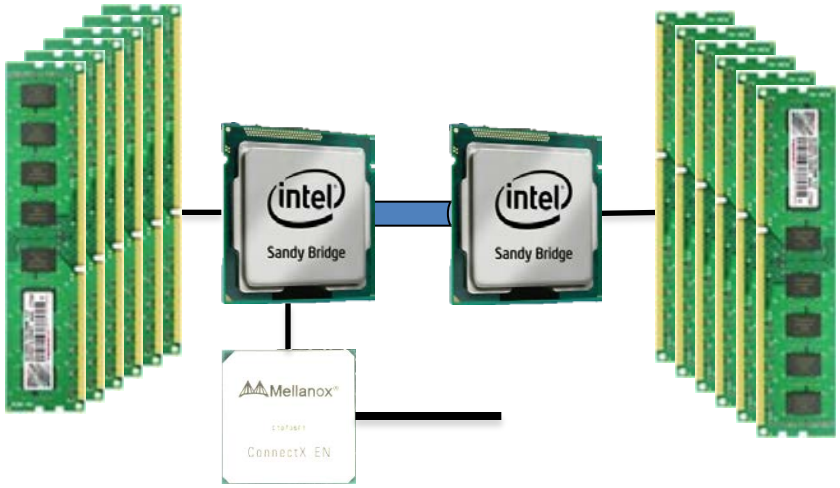


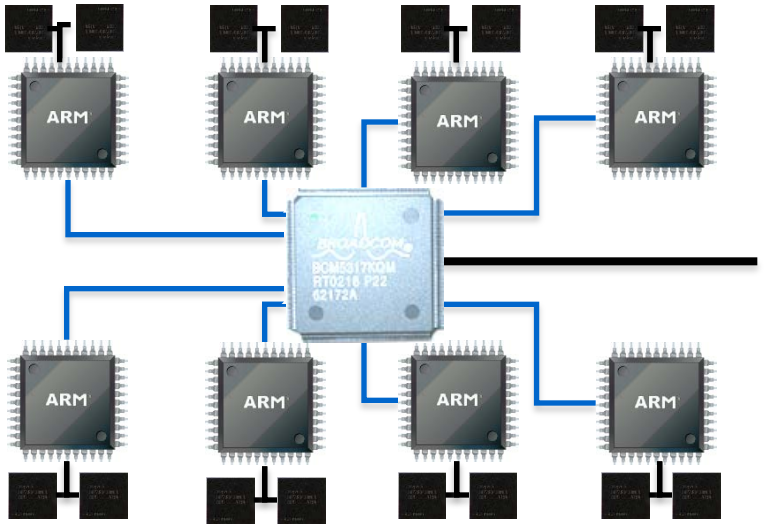


- 2-socket Intel Sandy Bridge
  - 64 GB/s intra-node (2 x QPI)
  - 56 Gb/s inter-node (4x FDR)
    - 32 Gb/s (4x QDR)
- 8-socket ARM Cortex A-15
  - 1 Gb/s intra-node (1 GbE)
  - 10 Gb/s inter-node (10 GbE)
  - 40 Gb/s backbone

Strawman design, not the actual Mont-Blanc architecture

How about power? And cost?



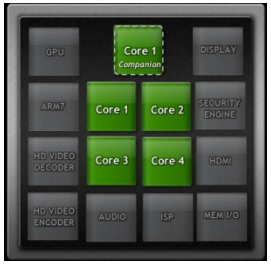


- 2-socket Intel Sandy Bridge
  - > $3.000 (Intel sockets only)
  - > 350W (CPU + DRAM)
- 8-socket ARM Cortex A-15
  - < $200 (ARM sockets only)
  - < 100W (CPU + DRAM)

Strawman design, not the actual Mont-Blanc architecture

Accelerators enter the game

- Application-specific hardware is more energy-efficient
  - Higher performance at lower power
- Several types of accelerators
  - Taken from the commodity (gaming) market
    - Nvidia + ATI (AMD) GPU
  - Developed for HPC
    - Intel MIC
  - Integrated in the mobile SoC
    - IMG PowerVR 6
    - ARM Mali T685
    - Qualcomm Adreno



Comparing embedded GPU to HPC accelerators

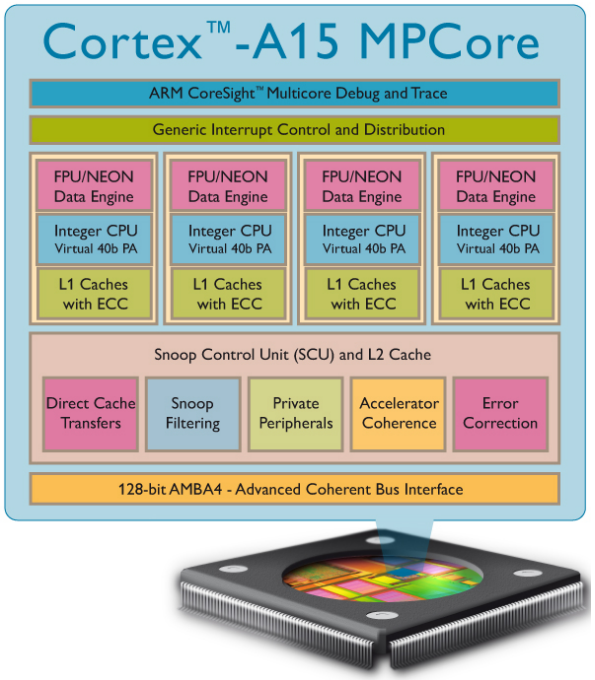
CPU	GFLOPS	CPU-GPU GB/s	On-chip Memory	Off-chip GB/s	Watts	Cost
Intel Knight's Corner	1000	32 ? 0.03 B/FLOP			300 3.33 GFLOPS/W	N/A
ATI Radeon HD 5970	928	16 0.02 B/FLOP		2 x 128 0.27 B/FLOP	294 3 GFLOPS/W	$1000
Nvidia Tesla C2090	665	16 0.02 B/FLOP	15 MB 0.02 B/KFLOP	177 0.19 B/FLOP	225 3 GFLOPS/W	$2000
NVIDIA K20 (Q4 2012)	1500	32 0.02 B/FLOP	16 MB 0.01 B/KFLOP	192 0.13 B/FLOP	140 10 GFLOPS/W	$2000+
ARM Mali T685	168*	12.8 0.08 B/FLOP	256 KB 0.002 B/KFLOP	12.8 0.08 B/FLOP	<5* 35+ GFLOPS/W	free

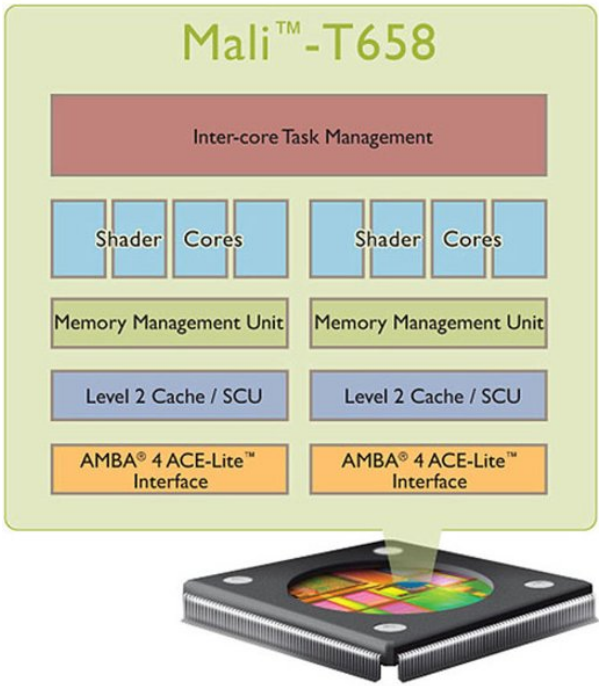
* Based on public information from Internet sources, not an ARM comitment

- Lower per-chip performance
- Lower on-chip cache size
- Lower memory bandwidth

- Higher energy efficiency
- Integrated with the SoC
  - Included in the baseline cost

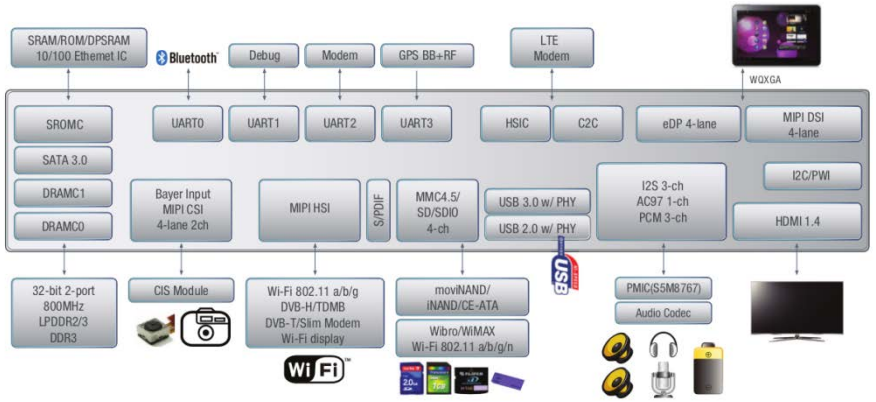
"Mobile-killer(tm)" example: Samsung Exynos 5450



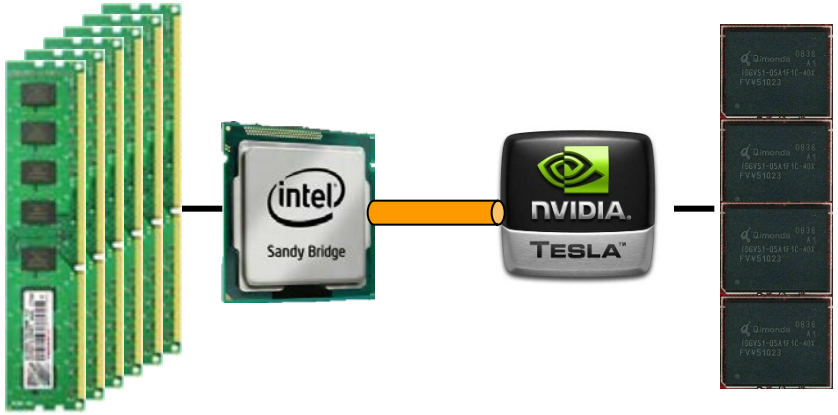


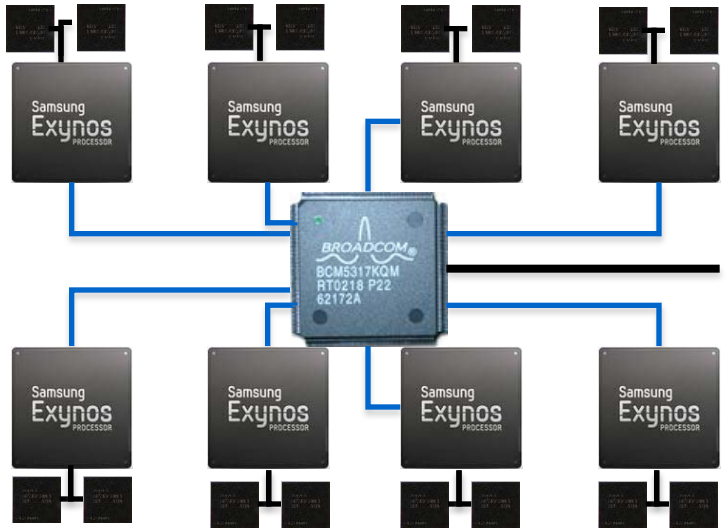
- 4-core ARM Cortex-A15 @ 2 GHz
  - 32 GFLOPS
- 8-core ARM Mali T685
  - 168 GFLOPS
- Dual channel DDR3 memory controller

- **All in a low-power mobile socket**



Can we achieve competitive performance?





- Intel Sandy Bridge + Nvidia K20
  - 1685 GFLOPS
  - 2 address spaces
  - 32 GB/s between CPU-GPU
    - 16x PCIe 3.0
  - 68 + 192 GB/s

- 8-socket Exynos 5
  - 1600 GFLOPS
  - 16 address spaces
  - 12.8 GB/s between CPU-GPU
    - Shared memory
  - 102 GB/s

Strawman design, not the actual Mont-Blanc architecture

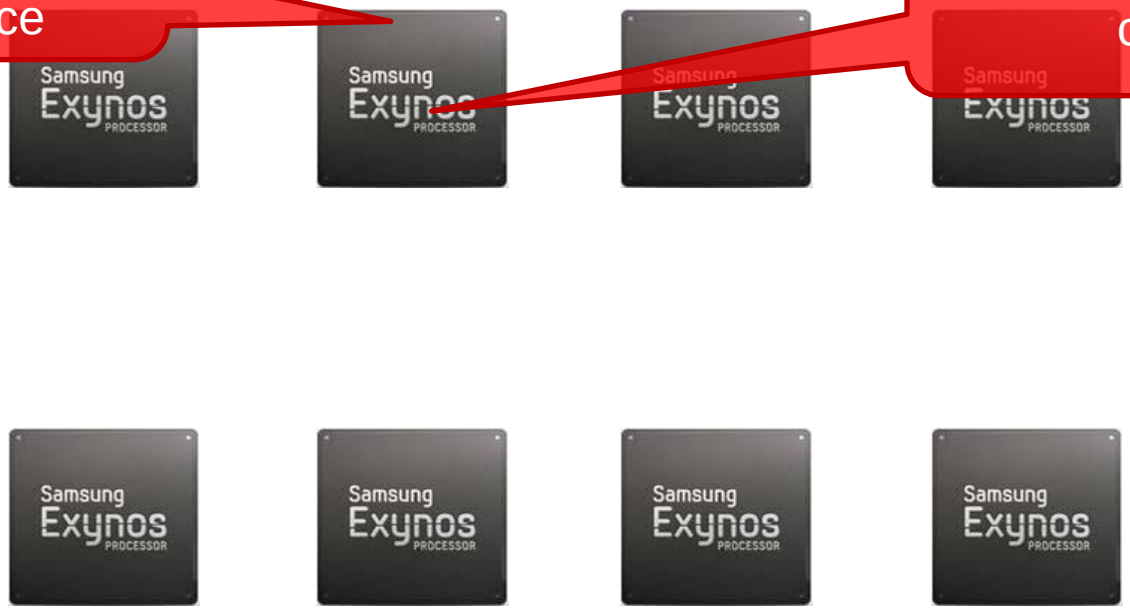
Challenges identified so far

2X more cores for the same performance

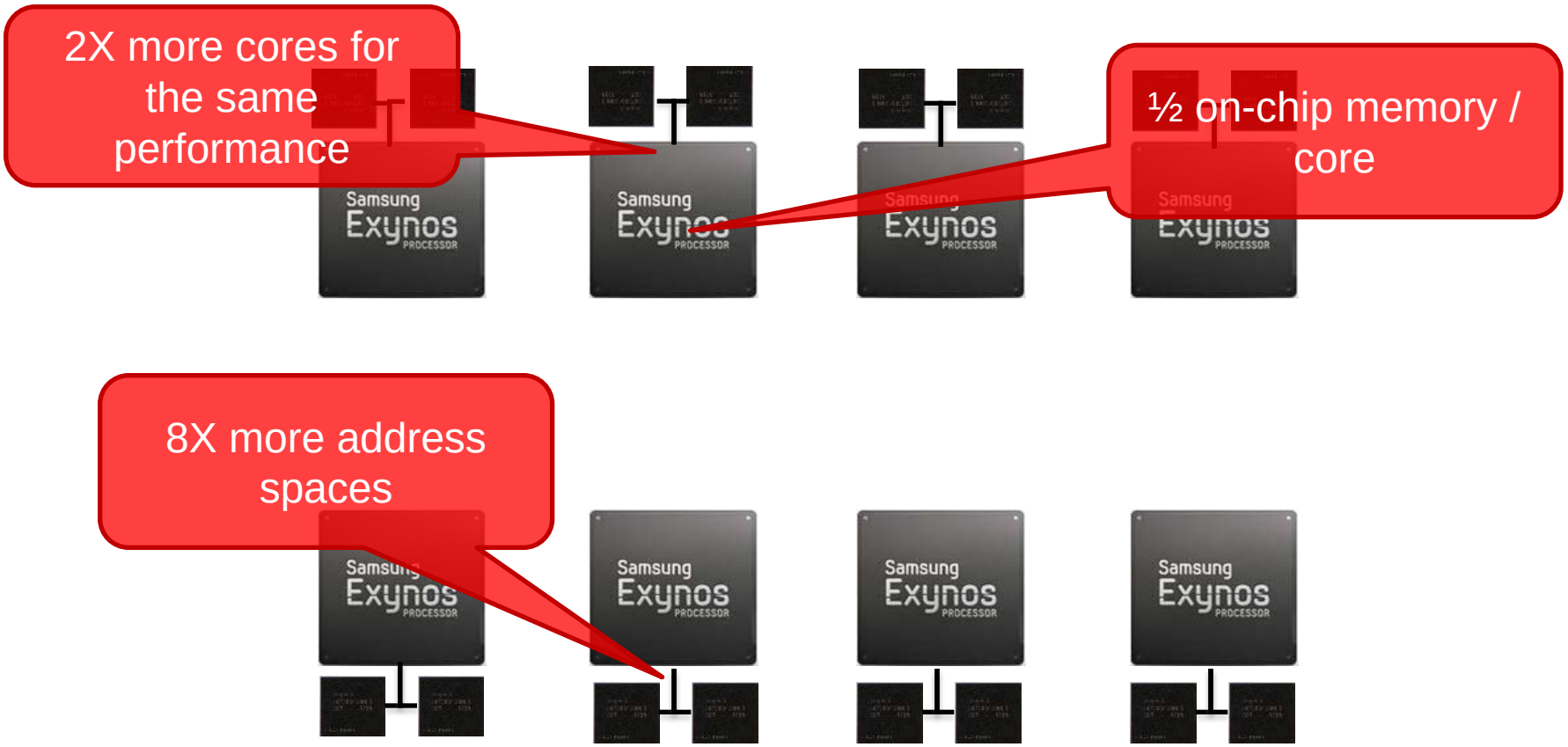
Challenges identified so far

2X more cores for the same performance

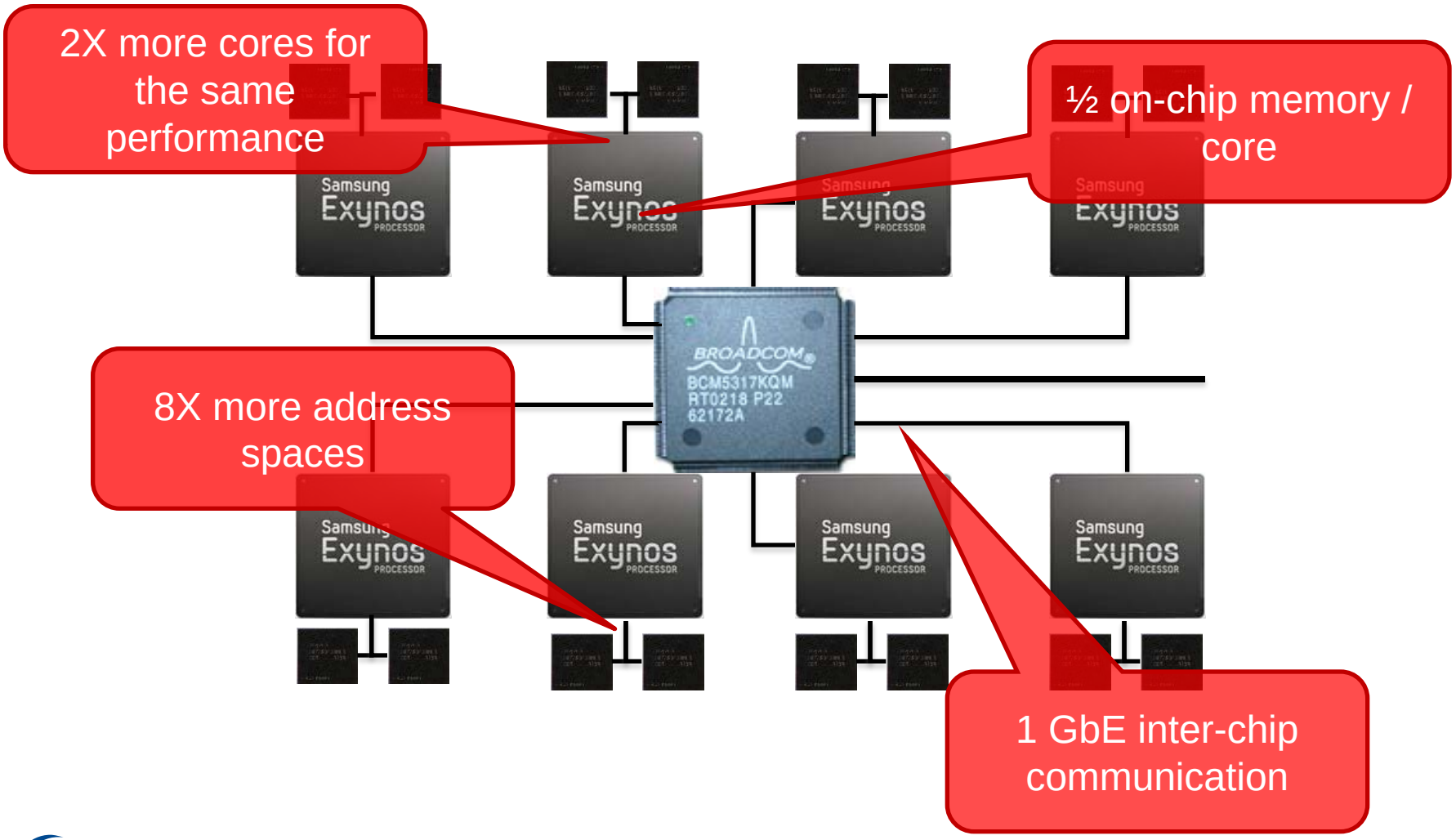
½ on-chip memory / core



Challenges identified so far

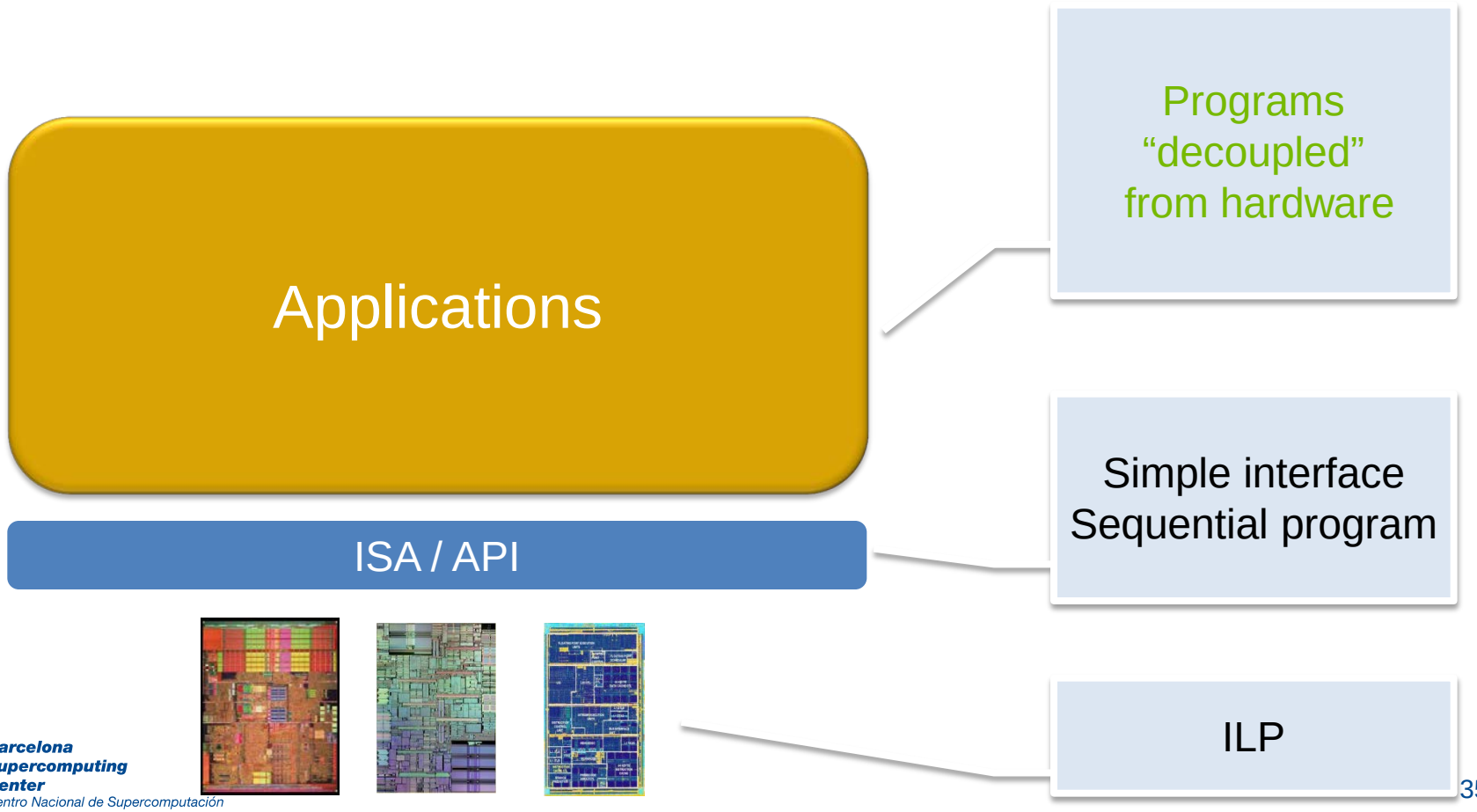


Challenges identified so far



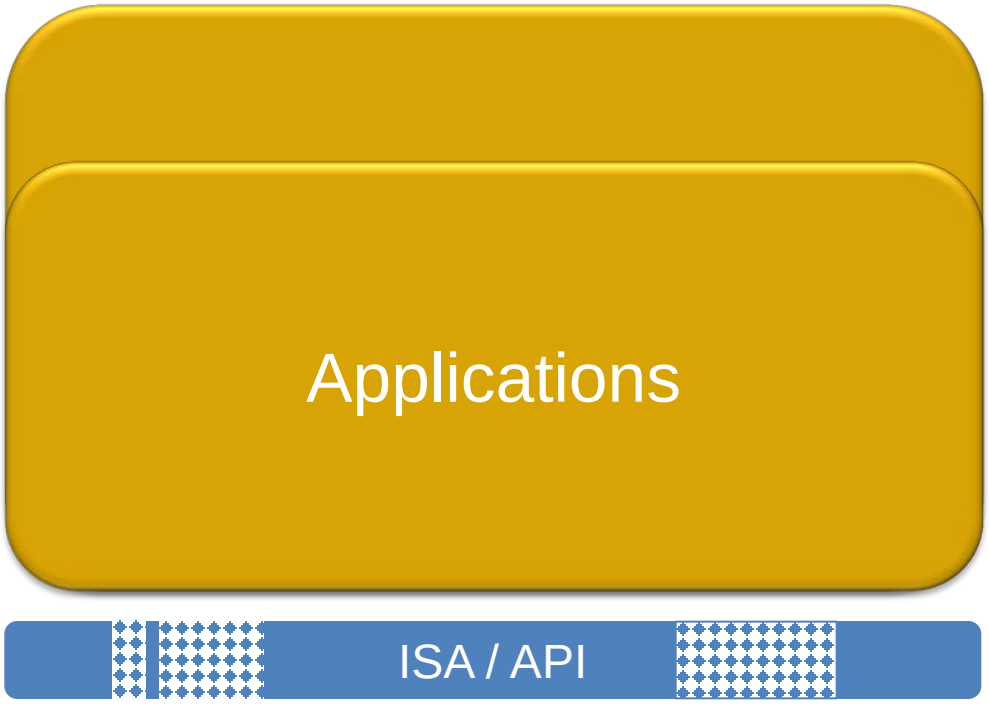
Living in the programming revolution

- At the beginning there was one language



Living in the programming revolution

- Multicores made the interface to leak



Application logic
+
Platform specificites

Address spaces (hierarchy, transfer), control flows, …

BSC Vision in the programming revolution

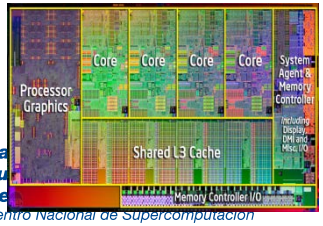
Application logic

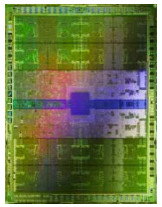
Arch. independent

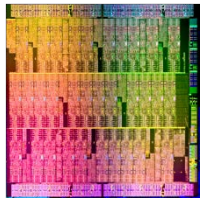
Applications

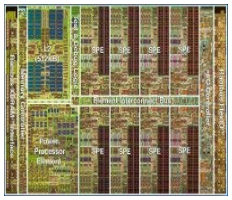
PM: High-level, clean, abstract interface

ISA / API

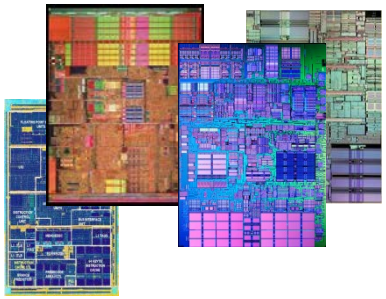






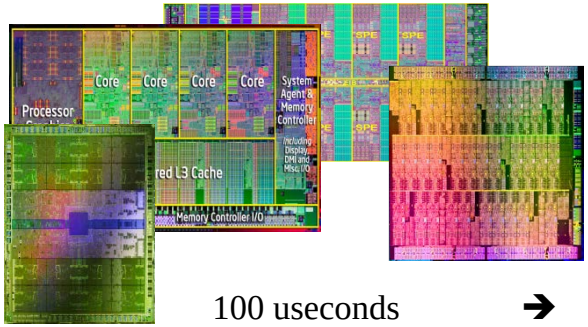


Computing: a matter of perspective



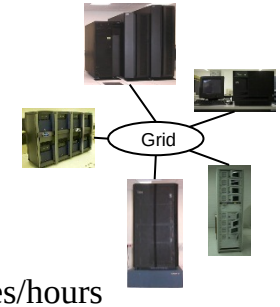
ns ➔

≅



100 useconds ➔

≅



minutes/hours

Mapping of concepts:

Instructions	➔	Block operations	➔	Full binary
Functional units	➔	Cores	➔	machines
Fetch &decode unit	➔	Core	➔	home machine
Registers (name space)	➔	Main memory	➔	Files
Registers (storage)	➔	Local memory	➔	Files

Granularity

Stay sequential

Just look at things from a bit further away

Architects do know how to run parallel

BSC Vision in the programming revolution

Applications

PM: High-level, clean, abstract interface

Power to the runtime

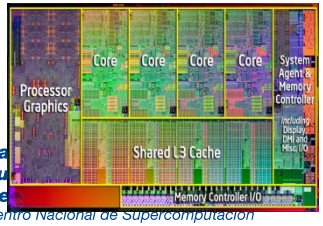
ISA / API

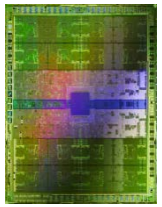
Application logic

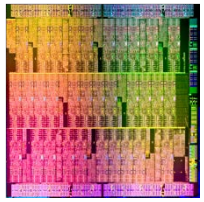
Arch. independent

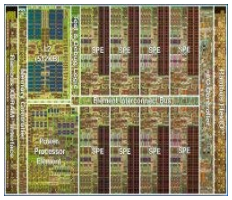
Task based
Single address space

“Reuse” architectural ideas under new constraints









The StarSs family of models: key concept

- Sequential program …
  - Task based program on single address/name space
  - Directionality annotations
    - Used to compute dependences AND to provide data access information
    - Use pattern, NOT resources and forcing actions (copies,…)
  - Order IS defined !!!!

- … happens to execute in parallel
  - Automatic run time computation and honoring of dependencies

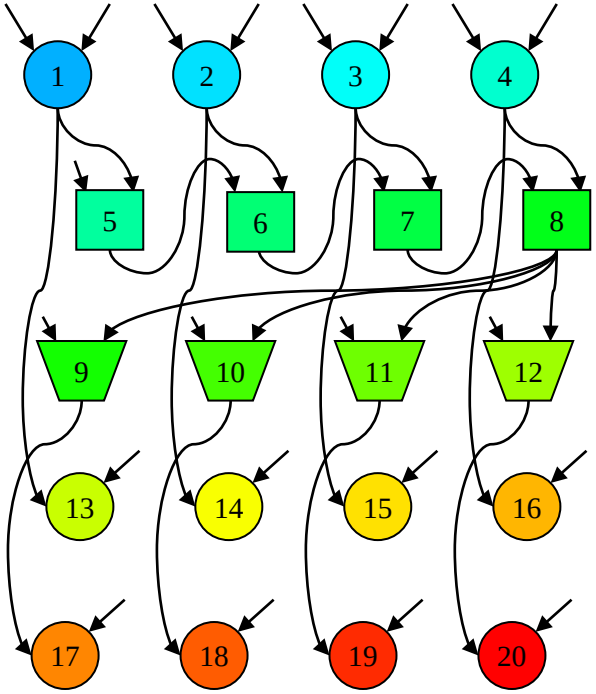
OmpSs: … generates task graph at run time …

```
#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]);
```

Simple Program Annotations

Task Graph Generation

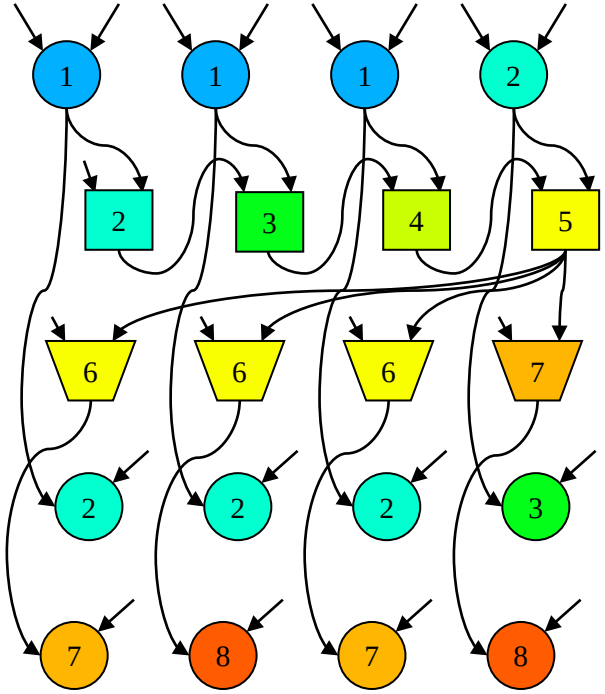


OmpsSs: … and executes as efficient as possible …

```
#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);
```

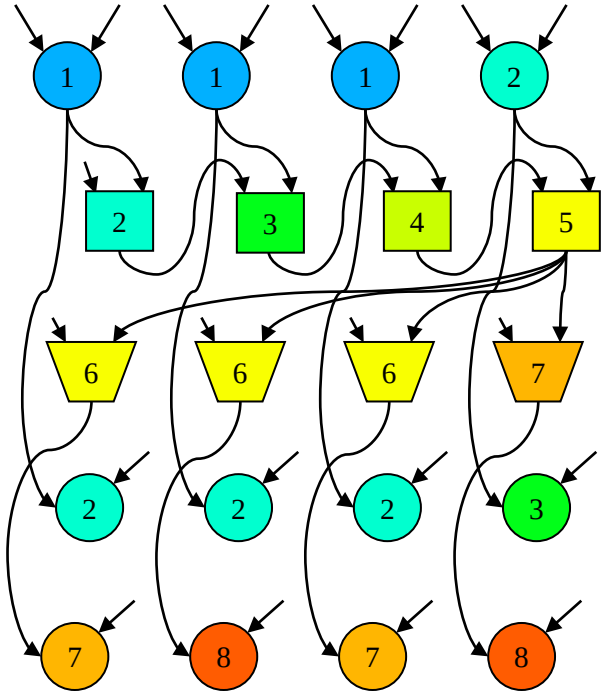
```
for (i=0; i<N; i+=BS)              // C=A+B
   vadd3 ( &A[i], &B[i], &C[i]);
...
for (i=0; i<N; i+=BS)             // sum(C[i])
   accum (&C[i], &sum);
...
for (i=0; i<N; i+=BS)             // B=sum*E
   scale_add (sum, &E[i], &B[i]);
...
for (i=0; i<N; i+=BS)             // A=C+D
   vadd3 (&C[i],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

Task Graph Execution



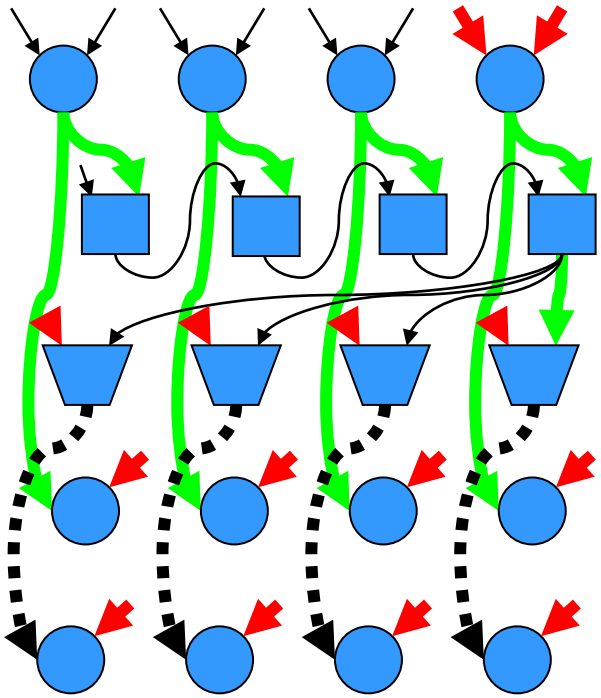
OmpSs & Challenges: 2x more cores

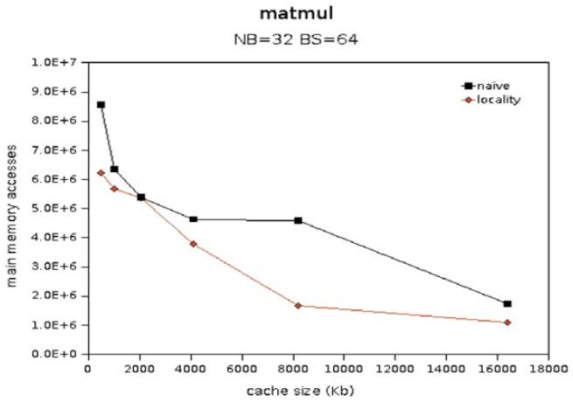
- Flexibility to dynamically generate work and traverse the computation space
  - Asynchronous data flow
    - Overlap
    - Tolerate variability
  - Non structured parallelism
  - Look-ahead
    - Huge task window
    - Do not stall at dependences
    - See what will have to be executed far in advance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



OmpSs & Challenges: ½ on chip memory

- Potential to automatically implement
  - Prefetch →
  - Reuse →
- Runtime responsible
  - Replication management, Coherence/consistency
  - Example techniques:
    - Minimize reuse distance
    - Lazy write-back
    - Data bypassing

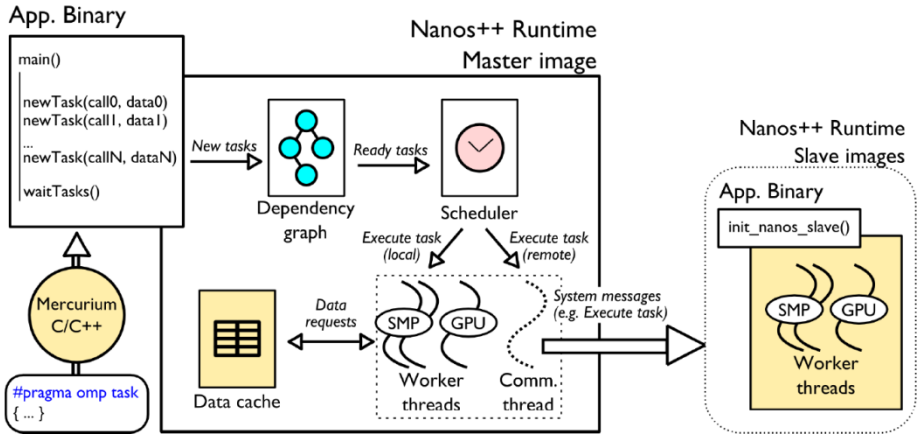


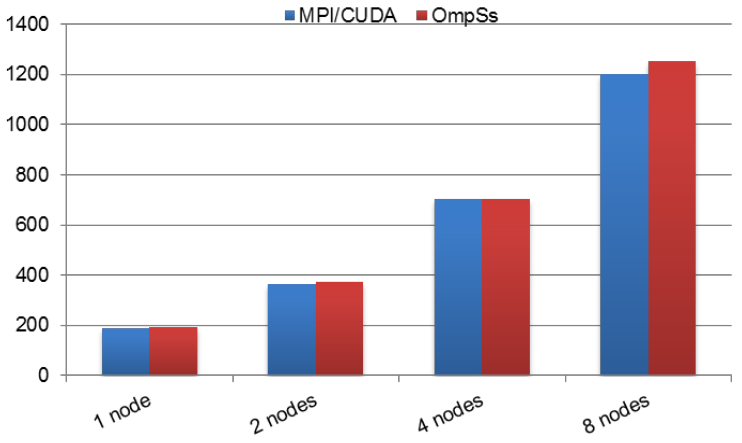


Pieter Bellens, Josep M. Pérez, Rosa M. Badia, Jesús Labarta: **Making the Best of Temporal Locality: Just-in-Time Renaming and Lazy Write-Back on the Cell/B.E**. IJHPCA 25(2): 137-147 (2011)

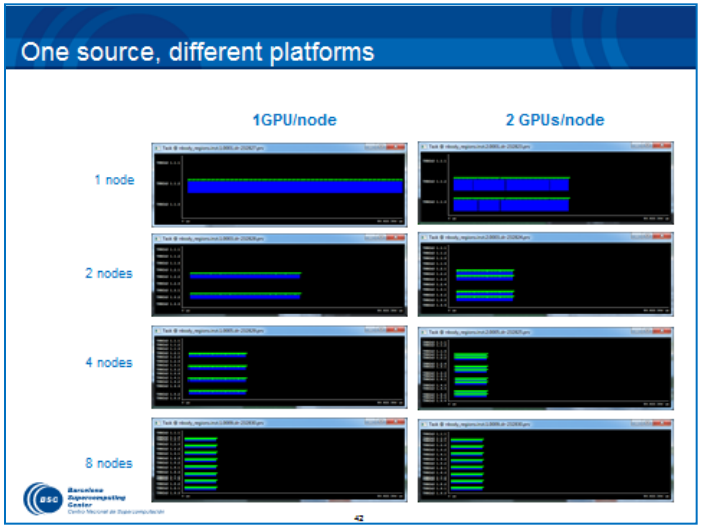
P. Bellens, et al, "CellSs: Scheduling Techniques to Better Exploit Memory Hierarchy" Sci. Prog. 2009

OmpSs & Challenges: 8X address spaces



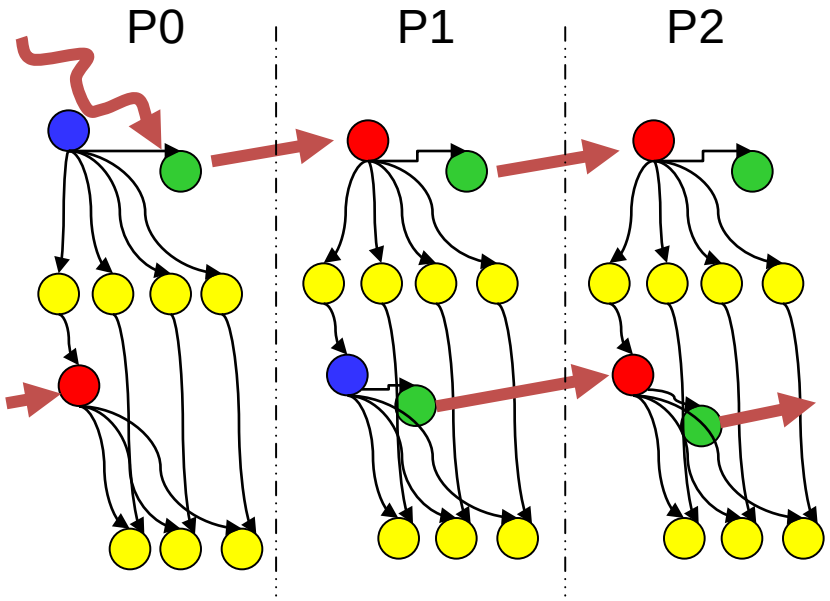


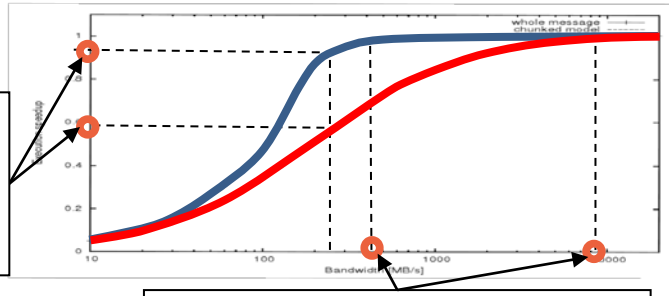
- OmpSs @ Cluster: Implementation that
  - Handles replication and copies
  - Optimize for locality and reuse
  - Maintain coherency and consistency
- A single "shared memory" node built of several separated address spaces



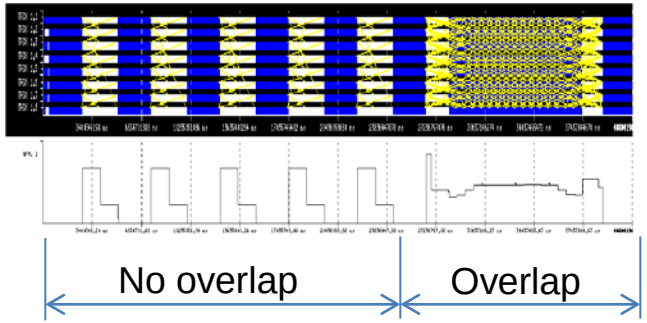
J. Bueno, A. Duran, R.M. Badia, X. Martorell, E. Ayguade, J. Labarta. **Productive Programming of GPU Clusters with OmpSs.** IPDPS'12.

OmpSs & Challenges: slow interconnect



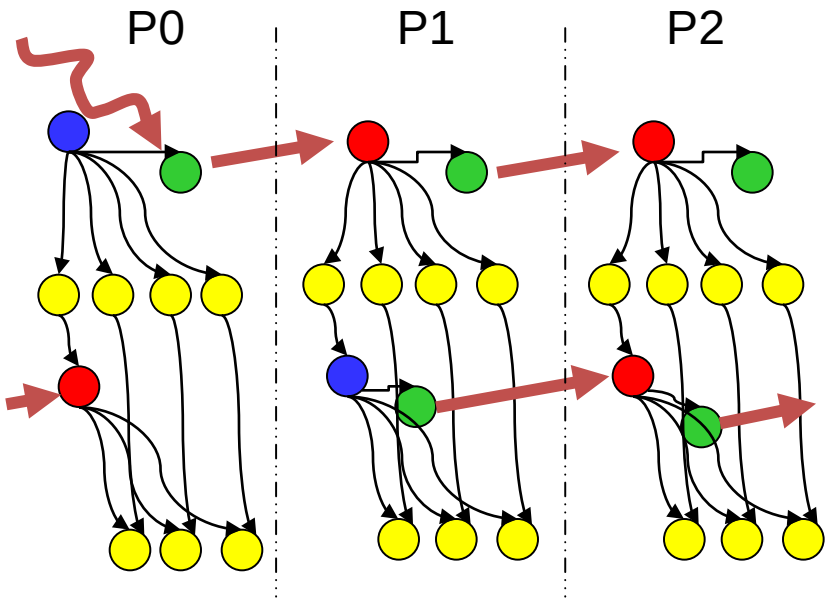


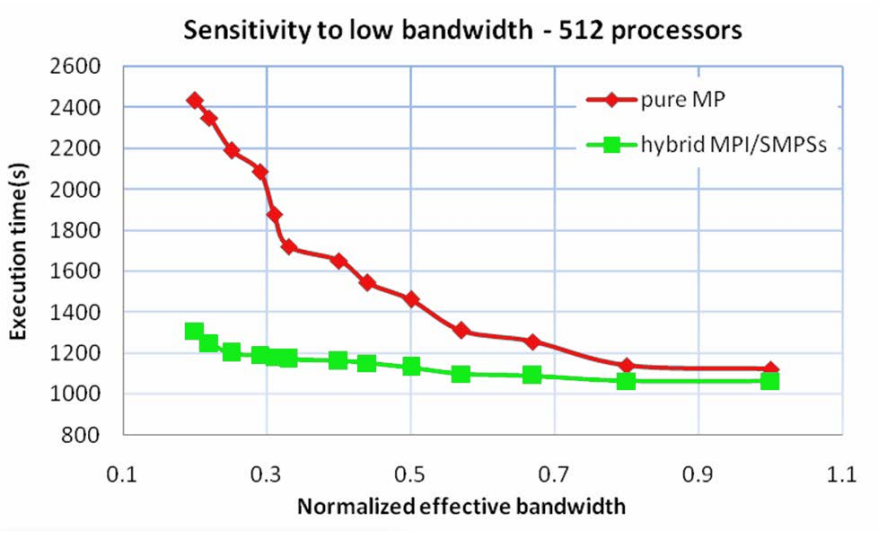
- MPI+OmpSs:
  - Overlap communication with computation
  - Hide long network latency and low bandwidth
  - Propagate asynchronous behavior to MPI level



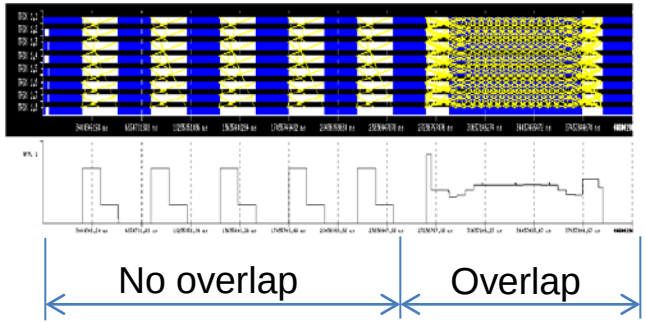
Vladimir Marjanovic, Jesús Labarta, Eduard Ayguadé, Mateo Valero: **Overlapping communication and computation by using a hybrid MPI/SMPSs approach**. ICS 2010: 5-16

OmpSs & Challenges: slow interconnect





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



Vladimir Marjanovic, Jesús Labarta, Eduard Ayguadé, Mateo Valero: **Overlapping communication and computation by using a hybrid MPI/SMPSs approach**. ICS 2010: 5-16

History

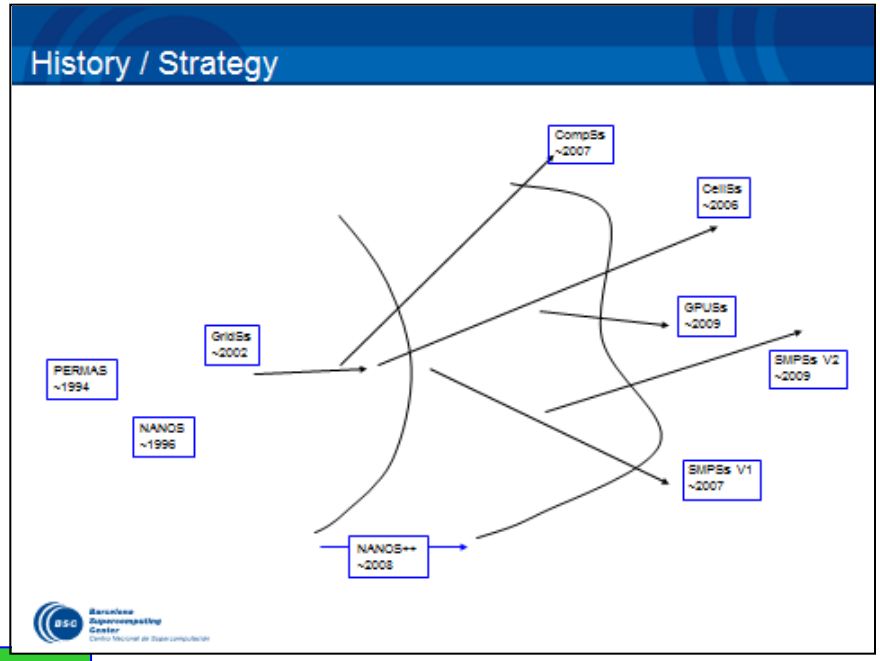
Basic SMPSs

must provide directionality ∀argument
Contiguous, non partially overlapped
Renaming
Several schedulers (priority, locality,…)
No nesting
C/Fortran
MPI/SMPSs optims.

SMPSs regions

C, No Fortran
must provide directionality ∀argument
ovelaping &strided
Reshaping strided accesses
Priority and locality aware scheduling

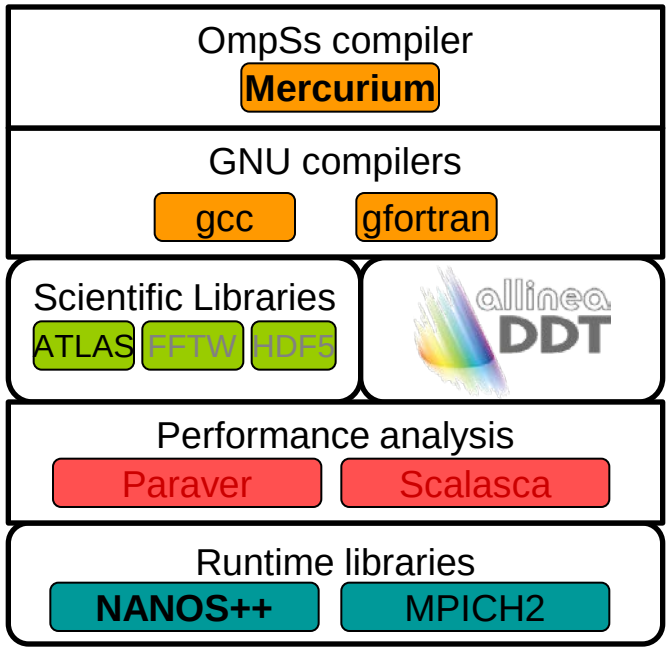
Evolving research since 2005

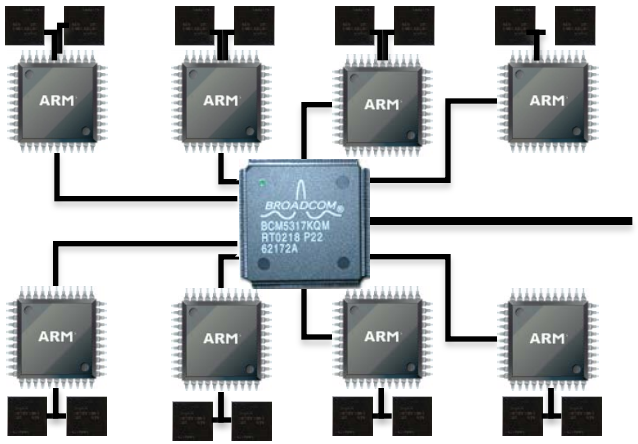


OMPSs

C, C++, Fortran
OpenMP compatibility (~)
Contiguous and strided args.
Separate dependences/transfers
Inlined/outlined pragmas
Nesting
Heterogeneity: SMP/GPU/Cluster
No renaming,
Several schedulers: “Simple” locality aware sched,…

OmpSs runtime layer manages architecture complexity

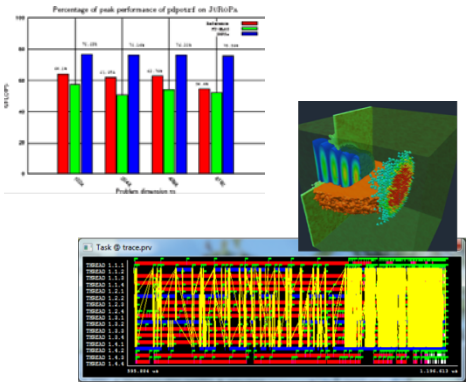
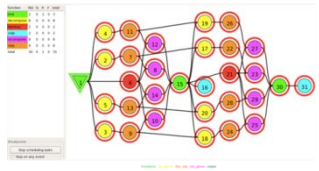
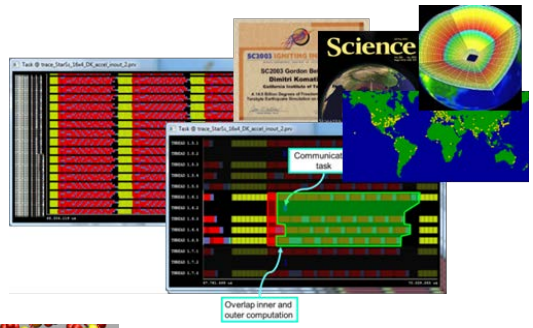


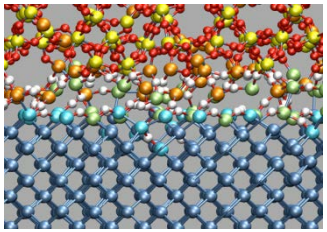
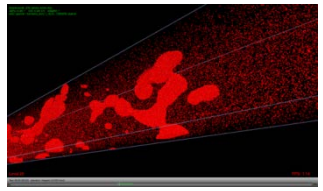


- Programmer exposed a simple architecture
- Task graph provides lookahead
  - Exploit knowledge about the future
- Automatically handle all of the architecture challenges
  - Strong scalability
  - Multiple address spaces
  - Low cache size
  - Low interconnect bandwidth
- Enjoy the positive aspects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

Used in projects and applications …

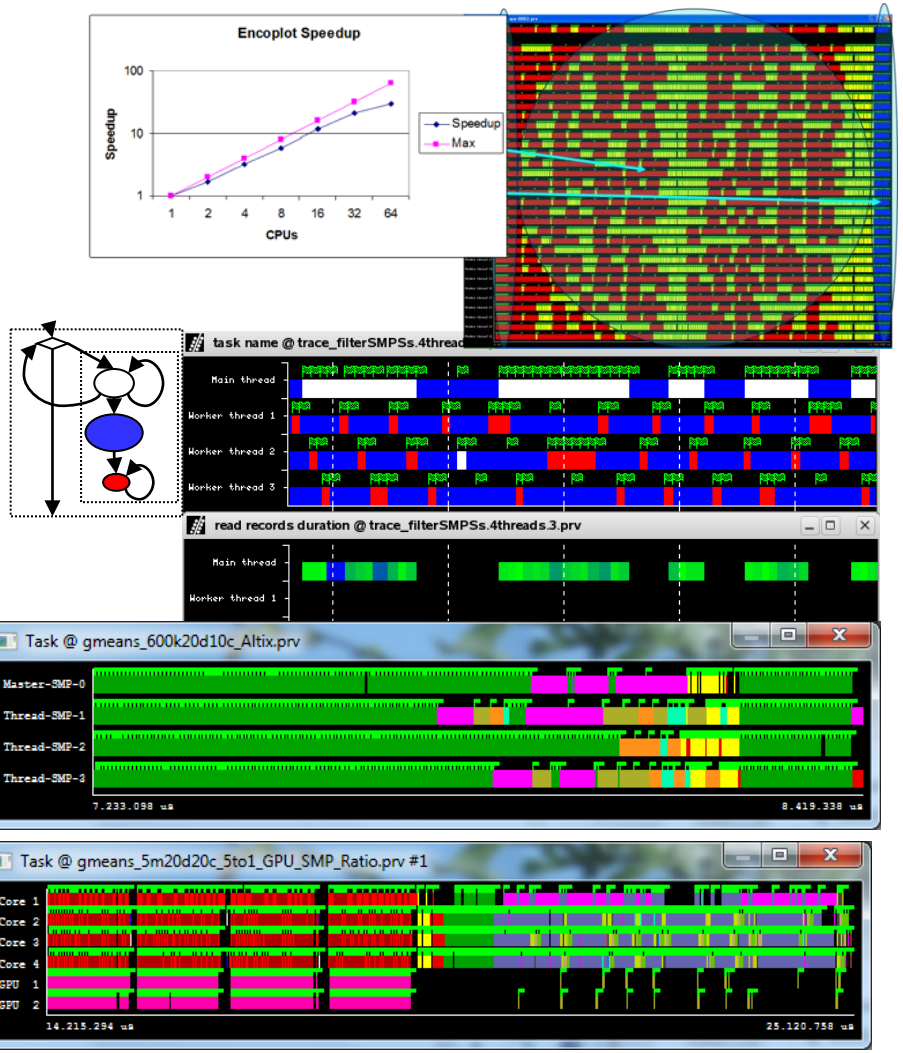
- Undertaken significant efforts to port real large scale applications:
  - text
    - Scalapack, PLASMA, SPECFEM3D, LBC, CPMD PSC, PEPC, LS1 Mardyn, Asynchronous algorithms, Microbenchmarks
  - MONT-BLANC
    - YALES2, EUTERPE, SPECFEM3D, MP2C, BigDFT, QuantumESPRESSO, PEPC, SMMP, ProFASI, COSMO, BQCD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

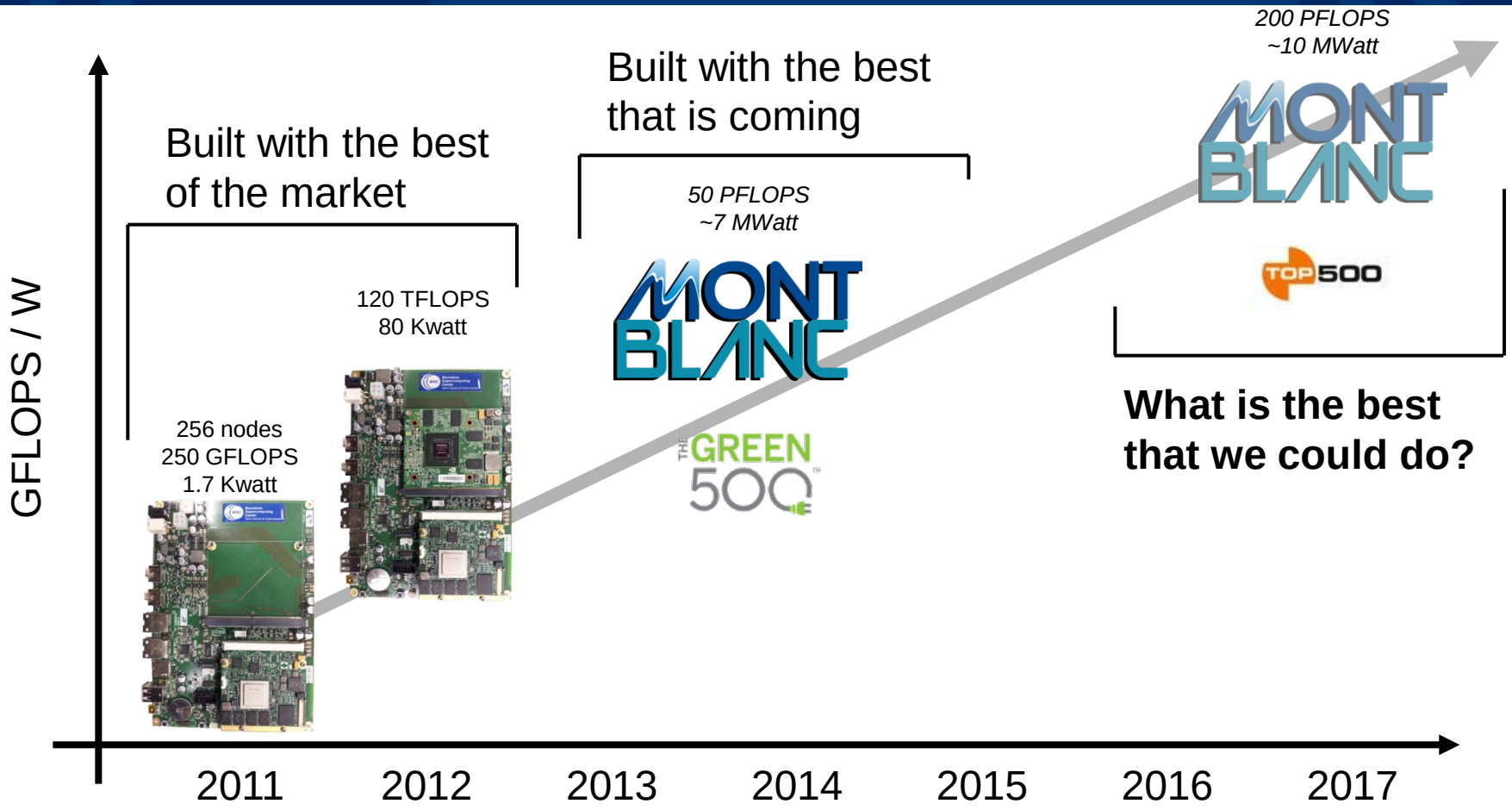
… but NOT only for «scientific computing» …

- 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!

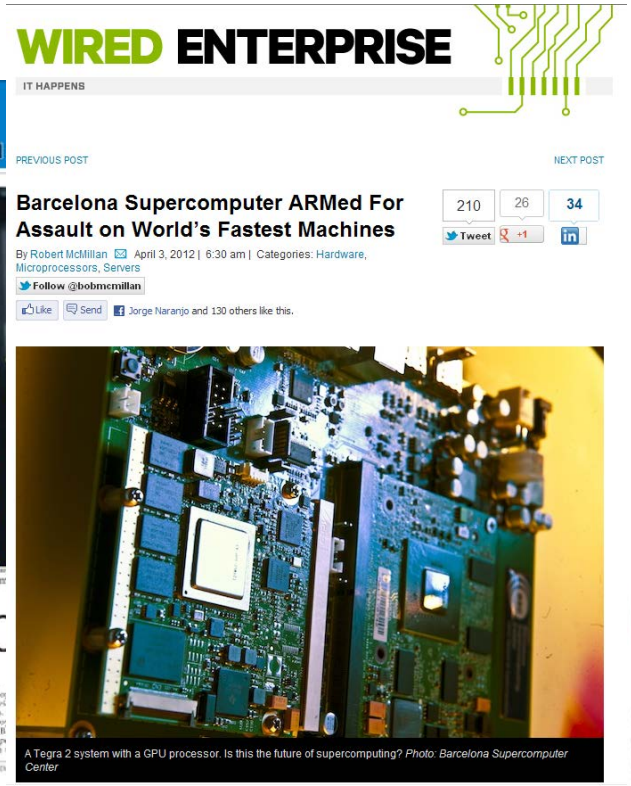
A big challenge, and a huge opportunity for Europe

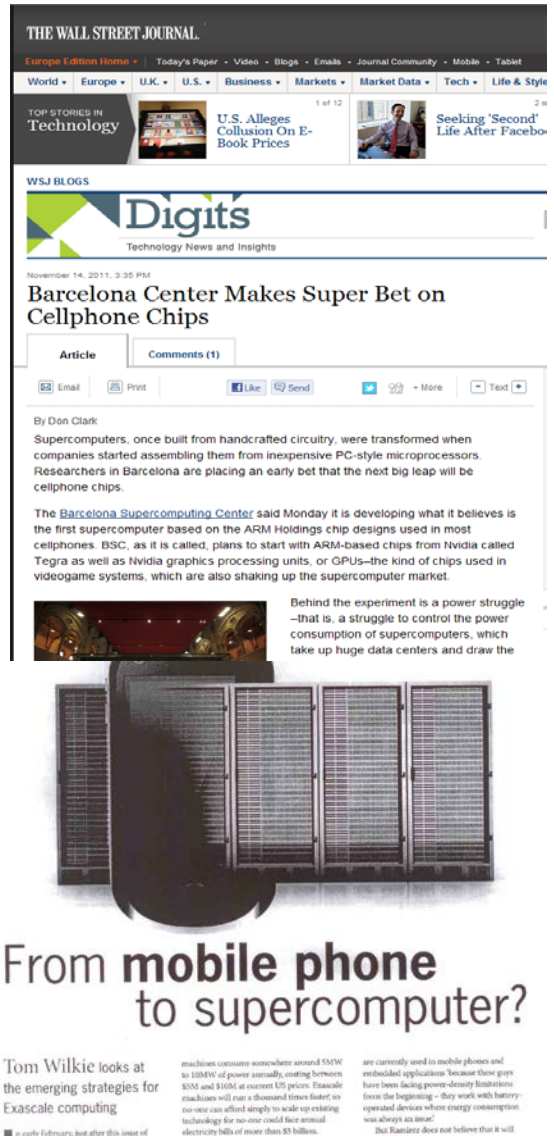


- Extend current mobile chips with the needed HPC features
  - Explore the use **vector architectures** in mobile accelerators
  - One design for all market segments: mobile, data centers, supercomputers

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





Mont-Blanc vision for the future

- A new class of energy-efficient supercomputer
  - Built from mobile + low-cost technology
  - Built on European industry strengths

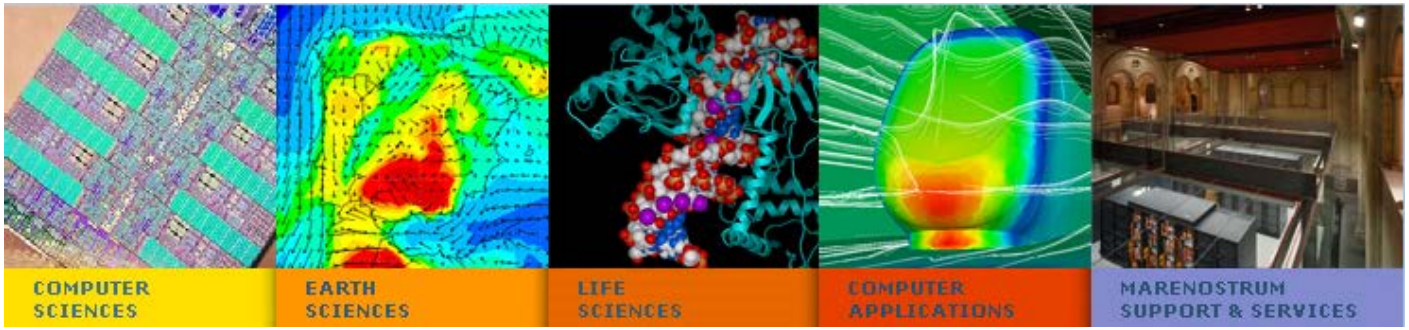
- Mont-Blanc is only the first step
- The potential is enormous, but needs continued investment
- Make the minimal changes to mobile chips to make them HPC-ready
  - Vector accelerator, integrated network, memory ECC, ...

Barcelona Supercomputing Center

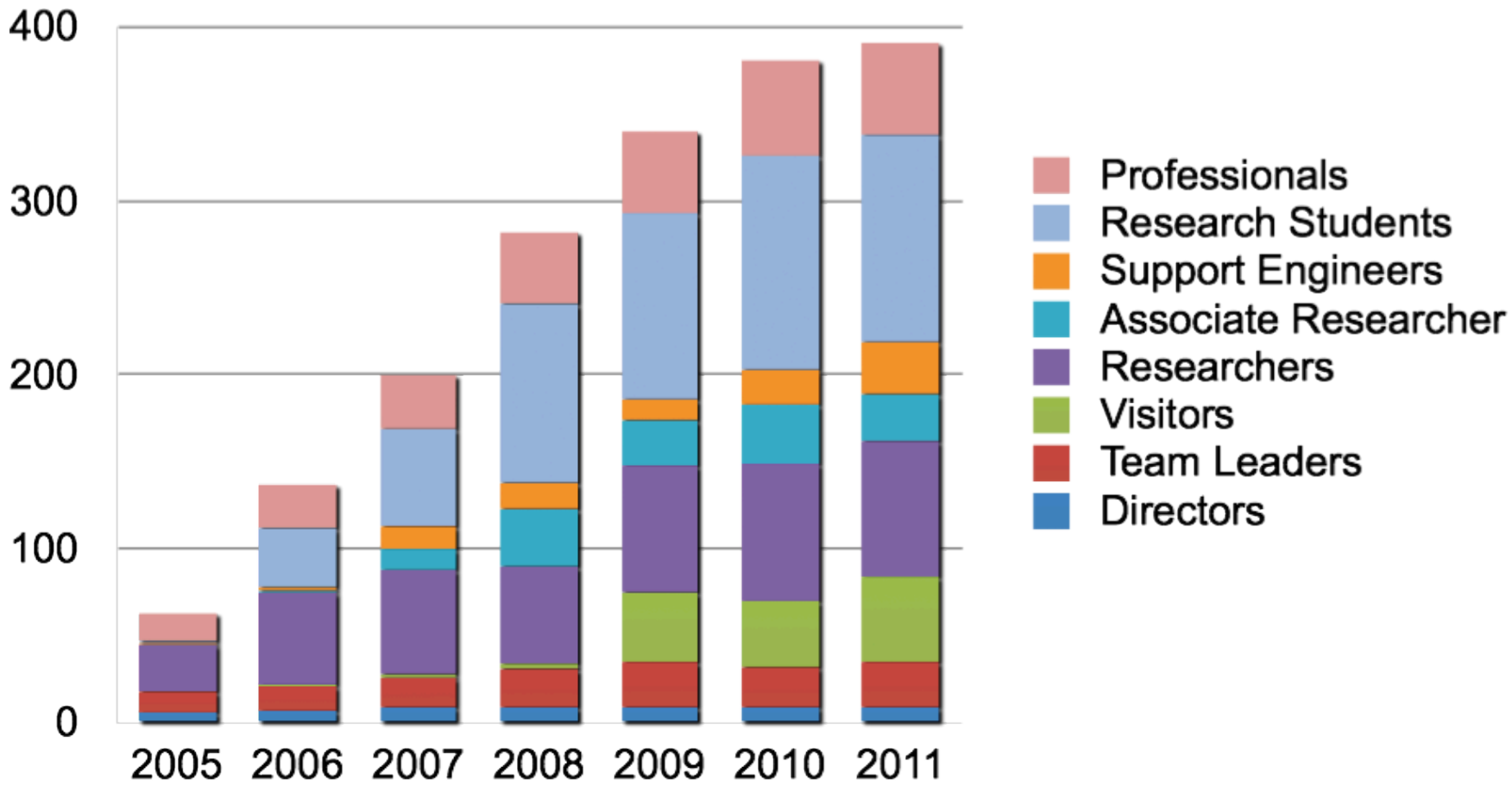
GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Generalitat de Catalunya
Departament d'Economia i Coneixement

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH



BSC Staff


400
300
200
100
0
2005
2006
2007
2008
2009
2010
2011
Professionals
Research Students
Support Engineers
Associate Researcher
Researchers
Visitors
Team Leaders
Directors

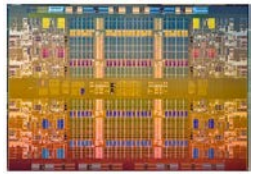
BSC people

- Algeria
- Argentina
- Austria
- Belgium
- Brazil
- Bulgaria
- Canada
- Chile
- China
- Colombia
- Denmark
- Ecuador
- France
- Germany
- Greece
- Hungary
- India
- Iran
- Ireland
- Israel
- Italy
- Japan
- Mexico
- Montenegro
- Pakistan
- Peru
- Poland
- Portugal
- Russia
- Serbia
- Slovakia
- South Africa
- Spain
- Syria
- Sweden
- Thailand
- The Netherlands
- Turkey
- UK
- Ukraine
- USA

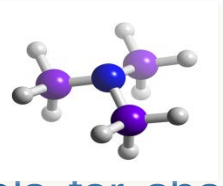
400 people - 40% from foreign countries

Scientific Strategic Issues

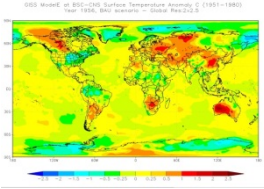
- Influence the way machines are built, programmed and used (Programming models, Performance tools, Computer Architecture, Low Power).



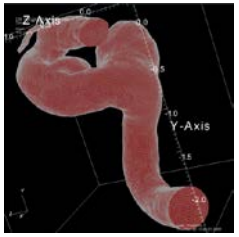
- Understand living organisms by means of theoretical and computational methods (Molecular Modeling, Genomics, Proteomics).



- Develop and implement global and regional state-of-the-art models for short-term air quality forecast and long term climate applications.

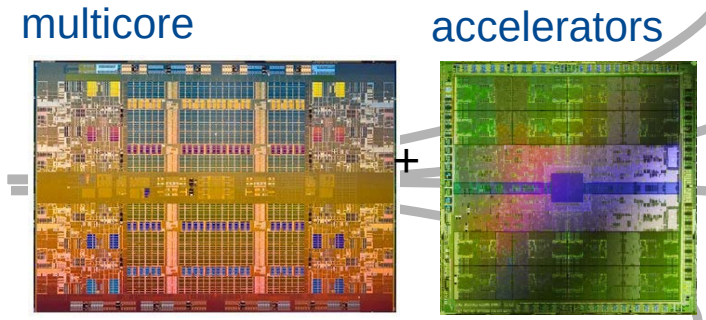


- Develop relevant scientific and engineering software for exploiting efficiently supercomputing capabilities (Biomedical, Geophysics, Atmospheric, Energy, Social and Economic simulations).



Computer Sciences mission

- Hardware and software components for high-performance computing



Mobile multimedia devices

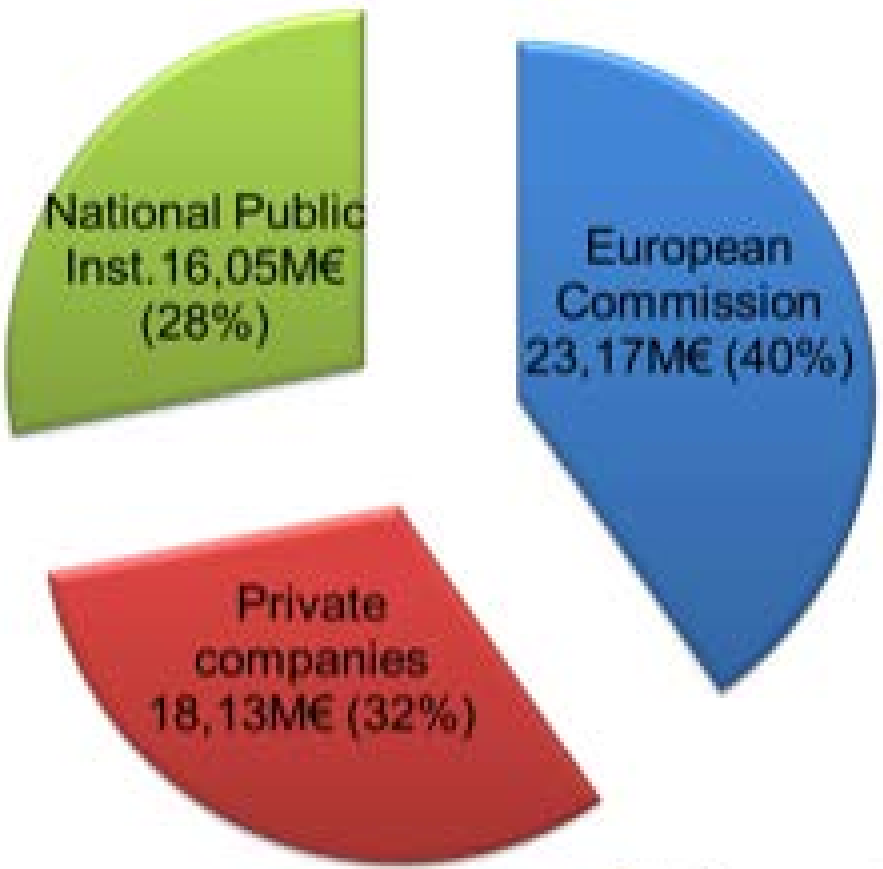
Real-time systems

Personal computers

Computing data centers and Cloud

- Performance, power/energy, reliability and usability

Source Competitive Funding



Total amount (2005/11): 57,35M€

BSC is the 3rd highest public research institution (after CSIC and CNIO) as regards resources from FP7 2007-2010

BSC & Industry

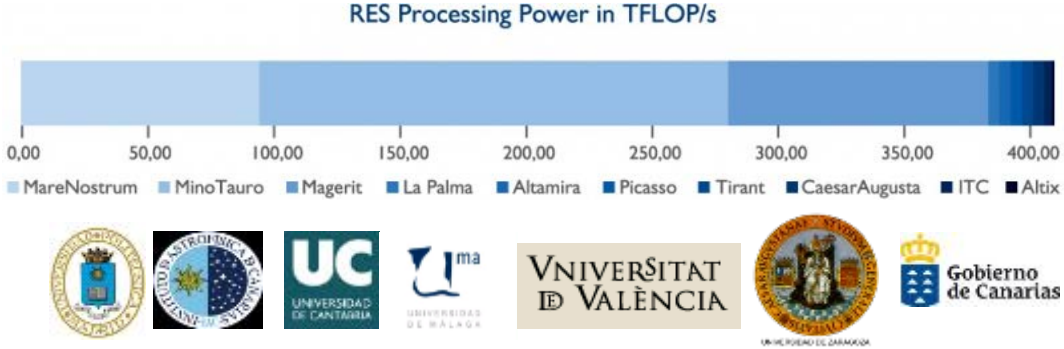
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC & Industry. Long term collaborations

- R&D of High-Tech firms

- Spanish firms

RES members and resources



BSC-CNS (MareNostrum)
Processor: 10240 PowerPC 970 2.3 GHz
Memory: 20 TB
Disk: 280 + 90 TB
Network: Myrinet

BSC-CNS (MinoTauro)
Processor: 128 nodes: 2 Nehalem + 2 M2090
Memory: 3 TB
Network: IB QDR

BSC-CNS (Altix)
Processor: SMP 128 cores
Memory: 1,5 TB

UPM (Magerit II)
Processor: 3.920 (245x16) Power7 3.3 GHz
Memory: 8700 GB
Disk: 190 TB
Network: IB QDR

IAC, UMA, UC, UZ, UV (LaPalma, Picasso, Altamira, Caesaraugusta, Tirant)
Processor: 512 PowerPC 970 2.2 GHz
Memory: 1 TB
Disk: 14 + 10 TB
Network: Myrinet

Gobierno de Islas Canarias - ITC (Atlante)
Processor: 336 PowerPC 970 2.3 GHz
Memory: 672 GB
Disk: 3 TB
Network: Myrinet

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Severo Ochoa: programme

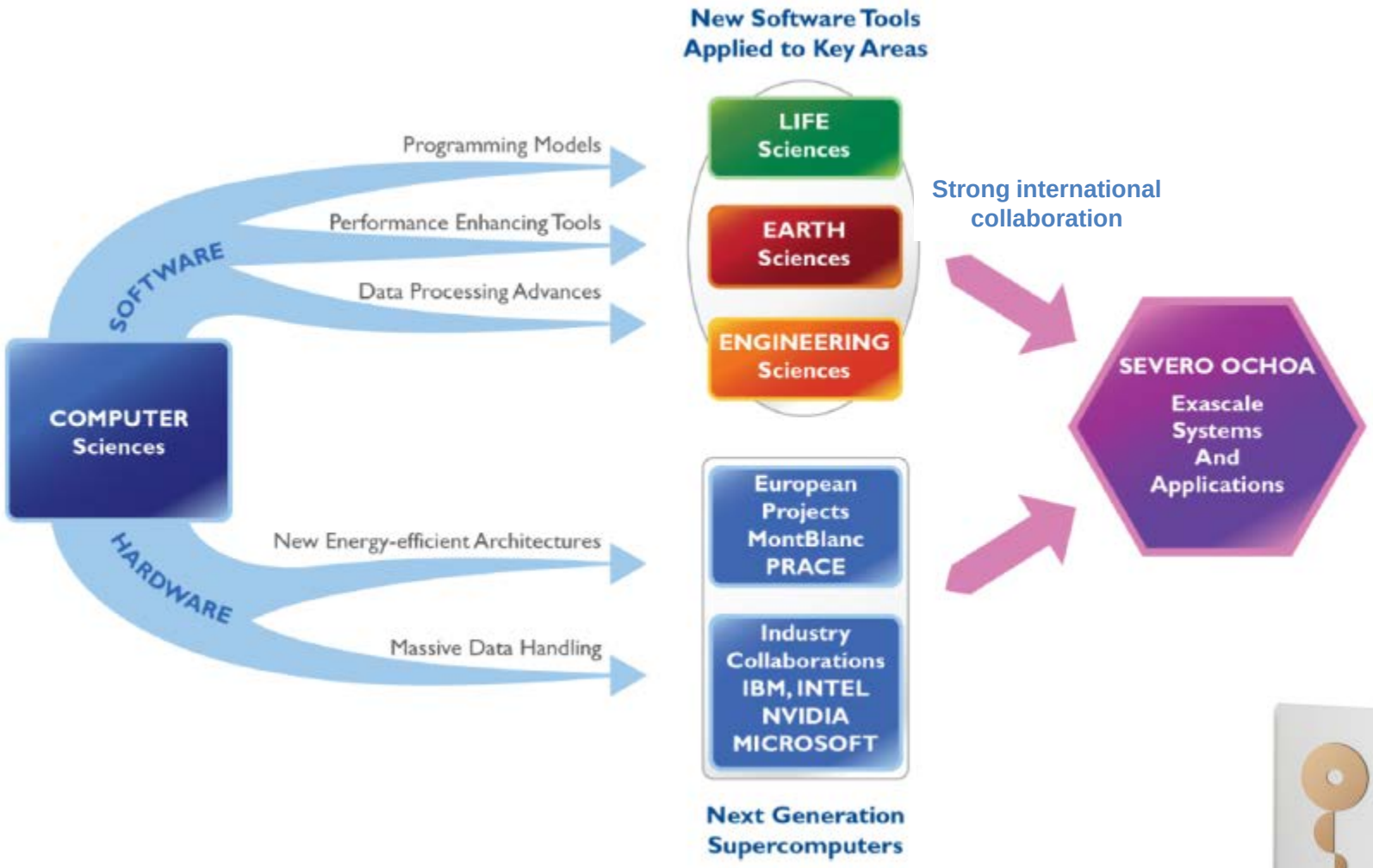
- The BSC is one of only eight Spanish research centres awarded with the prestigious Severo Ochoa grant.

  - The aim of the Severo Ochoa programme is to strengthen the very best Spanish research centres, who are internationally amongst the most competitive in their field.
  - With the Severo Ochoa grant, the BSC-CNS will strengthen its strategic research capacities, human resources, international collaboration and the dissemination of its results to society.

Severo Ochoa: research agenda in a few words

- BSC proposes a multidisciplinary research program to address the complex challenges in the path towards Exascale.
  - proposing high-performance/energy-efficient hardware and software solutions.
  - with challenging applications in:
    - Personalised medicine
    - Modelling of human organs
    - Global models for climate change and air quality prediction
- … leveraging existing national and international collaborations, consolidating BSC-CNS as a world leader in the field.

Severo Ochoa: research agenda in a few words



BSC in the world

- BSC has the will to establish a RIS network (Supercomputing Iberoamerican Network) via CYTED (Iberoamerican Programm for Science and Technology Development)
  - RES plus Iberoamerican centers
  - Including training and research
  - Connected to EU programs

- Countries:
  - Argentina
  - Brazil
  - Belgium
  - Chile
  - Colombia
  - Italy
  - México
  - Portugal
  - Spain

Many thanks …

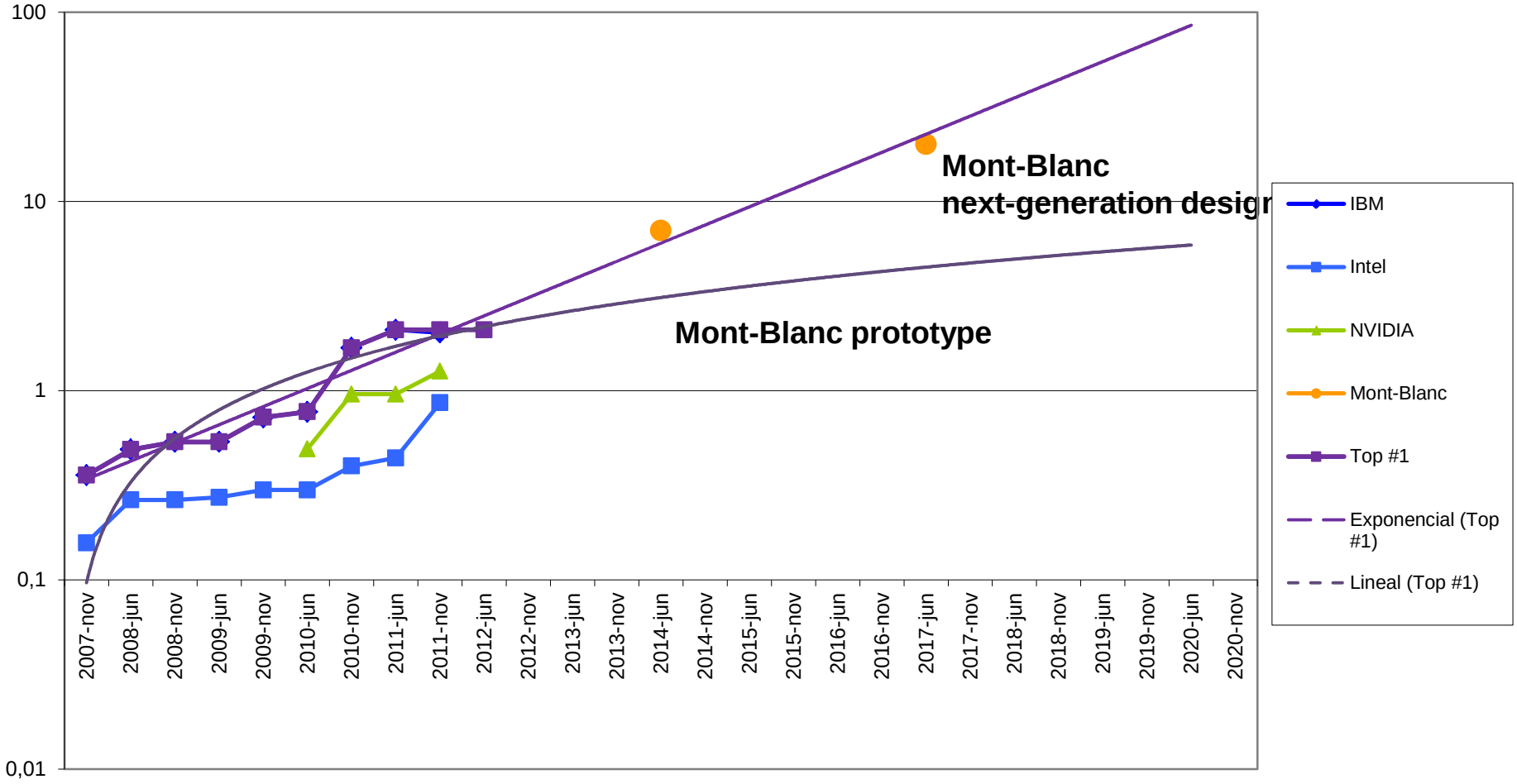
- This presentation would not have been possible without help from a great team
  - Eduard Ayguadé
  - Jesus Labarta
  - Isaac Gelado
  - Nacho Navarro
  - Mario Nemirovsky
  - Alex Ramirez
- And the work of the entire BSC and Mont-Blanc teams!

www.bsc.es

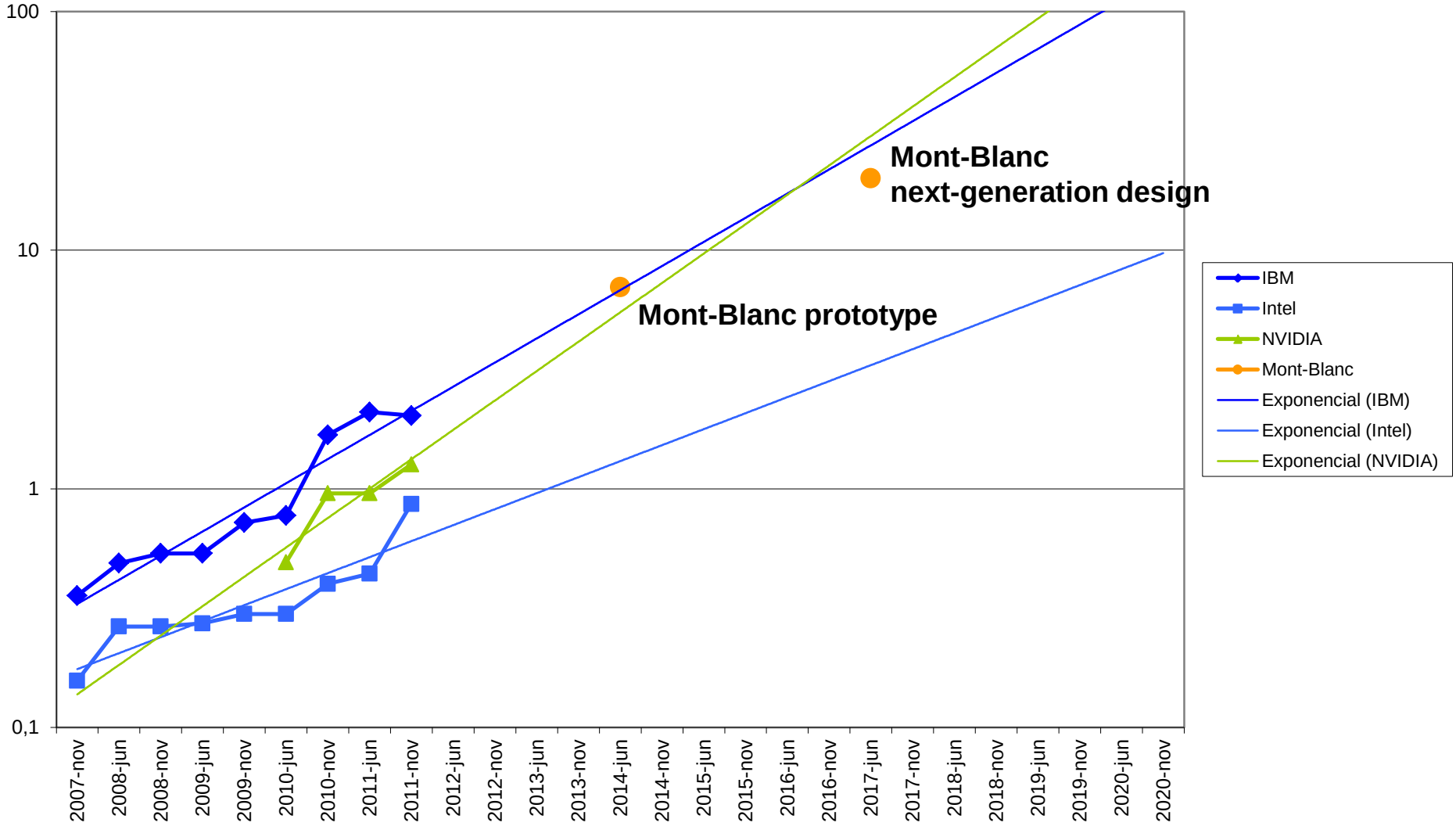
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

extra / backup slides

Projected Green500 + Mont-Blanc objectives



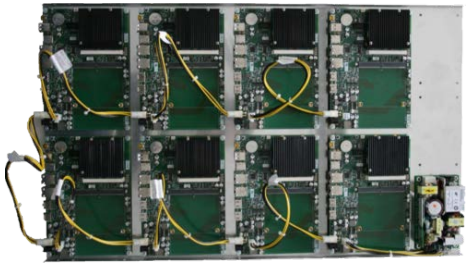
Projected Green500 + Mont-Blanc objectives



Tibidabo: ARM prototype @ BSC

Q7 Tegra 2
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

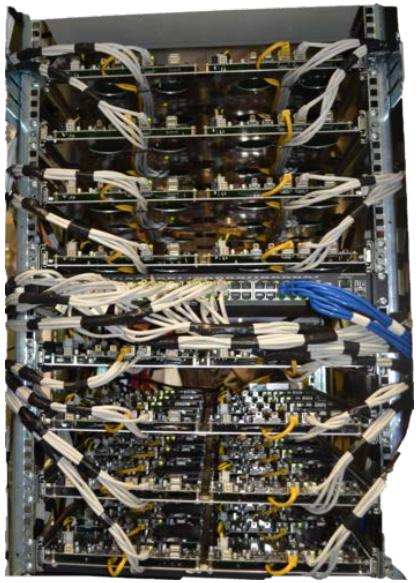


1U Rackable blade
8 nodes
16 GFLOPS
65 Watts
0.25 GFLOPS / W

Rack
16 blade containers
- 128 nodes
- 256 cores

5x 48-port 1GbE switch

256 GFLOPS
1.7 Kwatt
0.15 GFLOPS / W



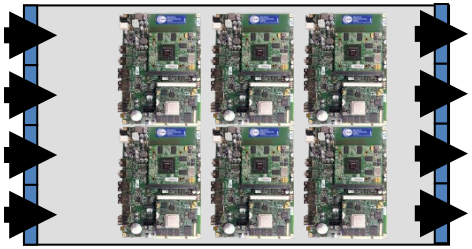
- Proof-of-concept to demonstrate HPC based on low-power components
  - Built entirely from COTS components
- Enable early software development and tuning
  - System software stack
  - Applications

Pedraforca: Tegra3 + Quadro 5010M GPU

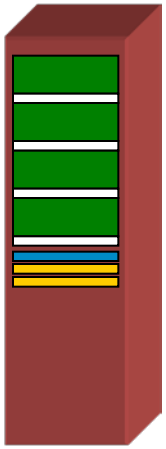
Q7 Tegra 3
4 x Cortex-A9
6 GFLOPS
7 Watts (?)
0.8 GFLOPS / W

Quadro 5010M
384 CUDA cores @ 450 MHz
345 GFLOPS
100 Watts
3.5 GFLOPS / W

Q7 + GPU board
4 x Cortex-A9
1 x GPU
351 GFLOPS
110 Watts
3.5 GFLOPS / W



4U rackable container
6 x (Q7 + GPU) node
Forced airflow (fans)
2.1 TFLOPS
0.6 KWatts
3.3 GFLOPS / W



½ Rack
4 x Blade container
96 cores
24 nodes
1 x 1GbE switch
1-2 PSU

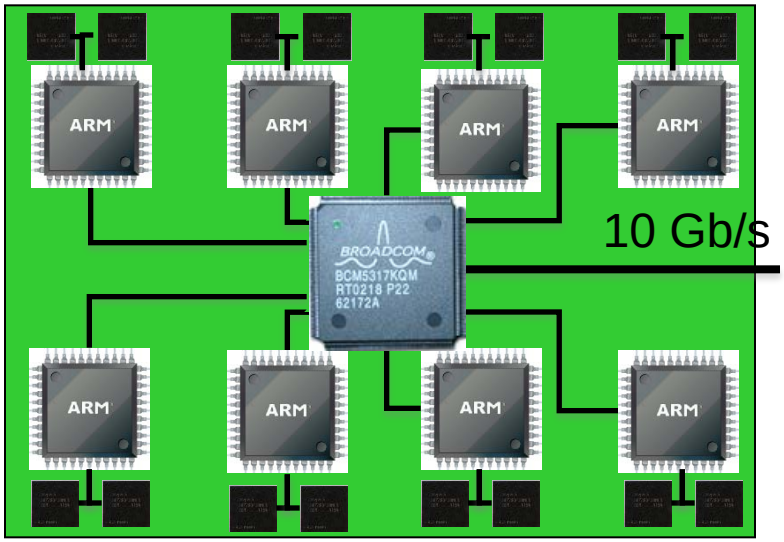
8.4 TFLOPS
<3 Kwatt
3 GFLOPS / W
1.5 GFLOPS/W at 50% efficiency

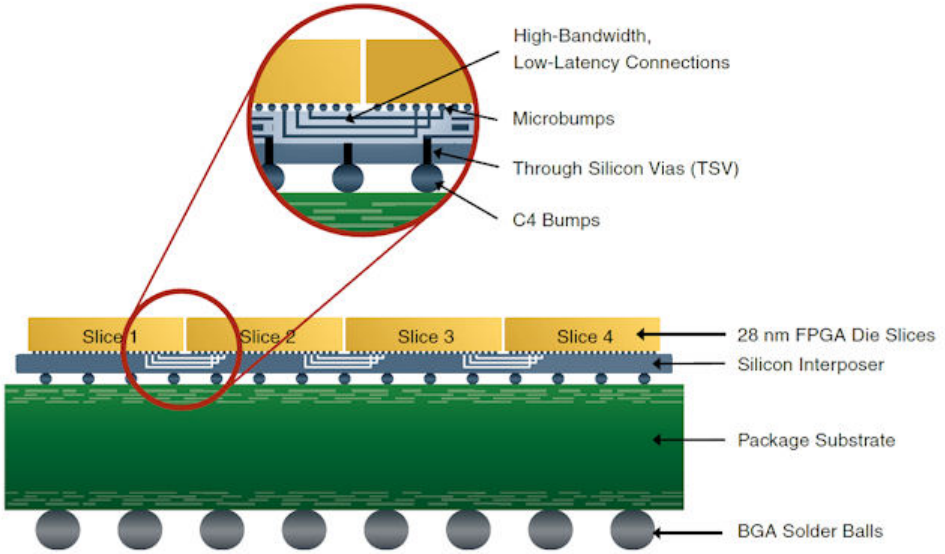
- Increase computation density with an external accelerator
  - GPU now dominates energy efficiency
- Proof of concept
- Software development platform

NVIDIA Kepler

CPU	CPU-GPU GB/s	GFLOPS	GB/s	On-chip Memory	Watts	GFLOPS/ W	Cost
GTX 680	32	125	192	16 MB	195	0.65	$500
GTX 690	32	2 x 125	2 x 192	2 x 16 MB	300	0.83	$1000
K10	32	125	177		225	0.5	
K20 (Q4 2012)	32	1500	192		300	5	

Inter-chip vs. Intra-chip communication



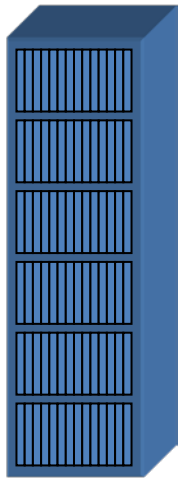


- Inter-chip communication more costly than multi-core
  - High energy, low bandwidth
- Integarte the 8 ARM sockets in the same 3D IC socket
  - Multi-chip package with silicon interposer
- Similar latency, bandwidth, and power to a single chip solution

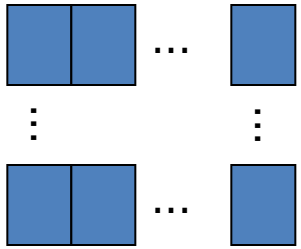
Mont-Blanc architecture (reverse engineering)

- 50 PFLOPS on 7 MWatt
- ARM Cortex-A15 CPU
  - 4 ops/cycle @ 2GHz = 8 GFLOPS
  - 65% efficiency = 5.2 GFLOPS
- Blade-based system design
  - 108 blades / rack
  - 12 sockets / blade
  - 5 Watts / socket
- 50 PFLOPS / 5.2 GFLOPS
  - 10 M cores
- 32% of 7 MWatt = 227 KWatt
  - 227 KWatt / 10 Mcores
  - 0.23 Watts / core
- 5 Watt / socket
  - 22 cores / socket ...
  - 4 cores + GPU / socket
  - 112 GFLOPS / socket

Multi-core chip:
112 GFLOPS
5.2 GFLOPS / core
68 GFLOPS / GPU
4 cores + GPU / chip
5 Watts / socket

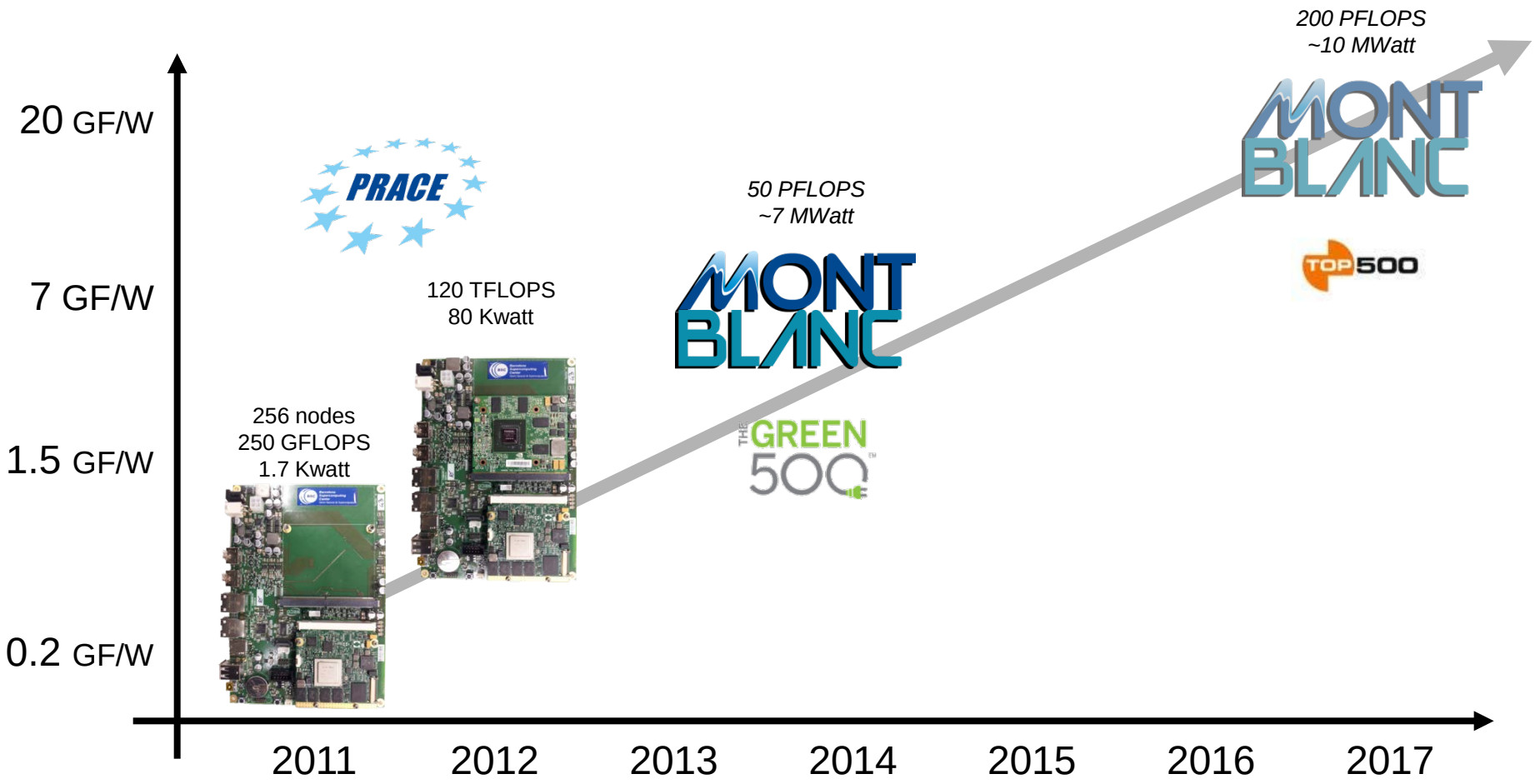


Rack:
108 compute nodes
1.296 chips
18 Kcores
140 TFLOPS
18-20 Kwatts



Full system:
352 racks
19 K blades
10 M cores
50 PFLOPS
7 MWatts

Energy-efficient prototype series @ BSC



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

Rely on software to handle the challenges

- Programming model and runtime are key component 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?