Designing HPC Architectures at the Barcelona Supercomputing Center

Miquel Moretó
Senior Researcher

28/Nov/2018
Universidad Complutense de Madrid
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-CNS objectives

Supercomputing services to Spanish and EU researchers

R&D in Computer, Life, Earth and Engineering Sciences

PhD programme, technology transfer, public engagement

BSC-CNS is a consortium that includes

Spanish Government 60%

Catalan Government 30%

Univ. Politècnica de Catalunya (UPC) 10%
People and Resources

Data as of 31st of December 2017

Staff Funding (People): 529

Competitive Funding (362)

Ordinary Budget (92)

Personnel Grants (75)

REVENUE

- Competitive
  - 13.6 M€ European Commission
  - 23.5 M€ State & Autonomous Regional Admin.
  - 5.9 M€ Companies
- Ordinary
  - 23.5 M€
  - 8.5 M€ Investments
  - 7 M€ Spanish Government
  - 4.7 M€ Generalitat de Catalunya

Total 39 M€

EXPENSES

- Payroll costs
- Investments
- Current expenditure

Total 39 M€
Mission of BSC Scientific Departments

**Computer Sciences**
To influence the way machines are built, programmed and used: programming models, performance tools, Big Data, computer architecture, energy efficiency

**Earth Sciences**
To develop and implement global and regional state-of-the-art models for short-term air quality forecast and long-term climate applications

**Life Sciences**
To understand living organisms by means of theoretical and computational methods (molecular modeling, genomics, proteomics)

**CASE**
To develop scientific and engineering software to efficiently exploit super-computing capabilities (biomedical, geophysics, atmospheric, energy, social and economic simulations)
MareNostrum 4

Total peak performance: **13.7 Pflops**

<table>
<thead>
<tr>
<th>Cluster Type</th>
<th>Peak Performance</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Purpose</td>
<td>11.15 Pflops</td>
<td>1.07.2017</td>
</tr>
<tr>
<td>CTE1-P9+Volta</td>
<td>1.57 Pflops</td>
<td>1.03.2018</td>
</tr>
<tr>
<td>CTE2-Arm V8:</td>
<td>0.5 Pflops</td>
<td>?????</td>
</tr>
<tr>
<td>CTE3-KNH?:</td>
<td>0.5 Pflops</td>
<td>?????</td>
</tr>
</tbody>
</table>

MareNostrum 1
2004 – 42.3 Tflops
1st Europe / 4th World
New technologies

MareNostrum 2
2006 – 94.2 Tflops
1st Europe / 5th World
New technologies

MareNostrum 3
2012 – 1.1 Pflops
12th Europe / 36th World
New technologies

MareNostrum 4
2017 – 11.1 Pflops
2nd Europe / 13th World
New technologies
HPC Performance Evolution from the Top500

- 567 PFlop/s
- 93 PFlop/s
- 286 TFlop/s

SUM

N=1

N=500

1 E flop/s

1 Pflop/s

1 T flop/s

1 G flop/s

100 Mflop/s

100 Gflop/s

100 Tflop/s

100 Pflop/s

1000 P flop/s

10000 P flop/s

100000 P flop/s

1000000 P flop/s

10000000 P flop/s

100000000 P flop/s

1E+09 P flop/s

Current HPC systems suffer from significant performance inefficiencies
MN3 (2012) vs MN4 (2017): Are applications really faster with new processors?

<table>
<thead>
<tr>
<th>Application</th>
<th>Cores</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRF</td>
<td>256</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>1.06</td>
</tr>
<tr>
<td>GROMACS</td>
<td>1024</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>NAMD</td>
<td>512</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>728</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>1.17</td>
</tr>
<tr>
<td>VASP</td>
<td>64</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>2.0</td>
</tr>
<tr>
<td>HPL</td>
<td>96</td>
<td>2.24</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>2.21</td>
</tr>
</tbody>
</table>

Performance of new processors marginally increases
The MultiCore Era
Moore’s Law + Memory Wall + Power Wall

Chip MultiProcessors (CMPs)

POWER4 (2001)

Intel Xeon 7100 (2006)

UltraSPARC T2 (2007)
How Multicores Were Designed at the Beginning?

**IBM Power4 (2001)**
- 2 cores, ST
- 0.7 MB/core L2, 16MB/core L3 (off-chip)
- 115W TDP
- 10GB/s mem BW

**IBM Power7 (2010)**
- 8 cores, SMT4
- 256 KB/core L2, 16MB/core L3 (on-chip)
- 170W TDP
- 100GB/s mem BW

**IBM Power8 (2014)**
- 12 cores, SMT8
- 512 KB/core L2, 8MB/core L3 (on-chip)
- 250W TDP
- 410GB/s mem BW
Vision in the Programming Revolution

Need to decouple again

Applications

PM: High-level, clean, abstract interface

Power to the runtime

ISA / API

Application logic
Arch. independent

General purpose
Single address space

The efforts are focused on efficiently using the underlying hardware

The need to decouple again is highlighted, focusing on efficiently using the underlying hardware through high-level, clean, abstract interfaces. Applications remain arch. independent, requiring power to the runtime for effective execution.
Runtime Aware Architectures

- The runtime **drives** the hardware design
  - Tight collaboration between software and hardware layers

**Applications**
- PM: High-level, clean, abstract interface

**Runtime**
- Task based PM annotated by the user
- Data dependencies detected at runtime
- Dynamic scheduling

“Reuse” architectural ideas under new constraints

• Management of hybrid memory hierarchies with scratchpad memories (ISCA’15, PACT’15) and stacked DRAMs (ICS’18)
• Runtime Exploitation of Data Locality (PACT’16, TPDS’18, SC’18)
• Exploiting the Task Dependency Graph (TDG) to reduce data movements (ICS’18)
• Architectural Support for Task-dependence Management (IPDPS’17, HPCA’18)
• Criticality-aware task scheduling (ICS’15) and acceleration (IPDPS’16)
• Approximate Task Memoization (IPDPS’17)
• Vector Extensions to Optimize DBMS (Micro’12, HPCA’15, ISCA’16)
• Dealing with Variation due to Hardware Manufacturing (ICS’16)
The International Exascale Challenge

- Exascale computing: $10^{18}$ FLOPS
  - 150x faster than MareNostrum 4
- Sustained real life application performances, not just Linpack...
- Exascale will not just allow present solutions to run faster, but will enable new solutions not affordable with today HPC technology
- From simulation to high predictability for precise medicine, energy, climate change, autonomous driven vehicles...
- The International context (US, China, Japan and EU...)
- The European HPC programme
- The European Processor Initiative (EPI)
- BSC role
Worldwide HPC roadmaps

From K computer... with domestic technology.
... to Post K with domestic technology.

From Tianhe-2.. with domestic technology.
...to Tianhe-2A with domestic technology.

IPCEI on HPC

From the PPP for HPC... to future PRACE systems...
...with domestic technology?
World Top 20 machines (status November 2017)

EU not in HPC world leaders

Europe has only 4 machines in world top 20
- Italy (CINECA) – Nr 14
- UK (Meteorological office) – Nr 15
- Spain (BSC, Barcelona) – Nr 16
- Germany (HLRS, Stuttgart) – Nr 19
EU HPC Ecosystem

- Specifications of exascale prototypes
- Technological options for future systems
- Collaboration of HPC Supercomputing Centres and application CoEs
- Provision of HPC capabilities and expertise
- Identify applications for co-design of exascale systems
- Innovative methods and algorithms for extreme parallelism of traditional & emerging applications

Centers of Excellence in HPC applications
Mont-Blanc projects in a glance

**Vision:** to leverage the fast growing market of mobile technology for scientific computation, HPC and non-HPC workloads

2012 2013 2014 2015 2016 2017 2018

Mont-Blanc

Mont-Blanc 2

Mont-Blanc 3

Allinea

ARM

AVL

CEA

CINECA

ETH Zürich

HLRS

Inria

Jülich Forschungszentrum

Irizar

University of Bristol

Universidad de Cantabria

UNIGRAZ

Université de Versailles St-Quentin-en-Yvelines

MB2020 started!
BSC contributions in Mont-Blanc

**ARM-based prototypes**
- Mobile technology
- Server technology
- Custom design

**System software**
- Scientific libraries
- Performance analysis tools
- Support for runtimes
- Power monitor

**Scientific applications**
- Porting and benchmarking of mini-apps and full scale applications
- Scalability study on real ARM-based platforms

**Resiliency**
- Application based fault tolerance
- Fault tolerance support in the runtime
- Reliability study of the Mont-Blanc prototype

**Next-generation studies**
- big.LITTLE studies
- Limitation analysis
- Performance projections
Mont-Blanc HPC Stack for Arm

Industrial applications

- Pharmacelera™
- MUREX™
- DASSAULT AVIATION
- AVL
- Thermo Fluids
- Cenaero
- Rolls Royce

Applications

- GÉNIE
- CINECA
- HLR
- Uni Graz
- University of Bristol

System software

- ARM®
- ETH Zürich
- Inria

Hardware

- ARM®
- BSC
- Bull
- CEA
- UC
A big challenge, and a huge opportunity for Europe

- Extend current mobile chips with the needed HPC features
  - Explore the use of vector architectures in mobile accelerators (vector processor ARM-based, 15+ Teraflops chip, 150 watts) ... unique opportunity for Europe
- One design for all market segments: mobile, data centers, supercomputers

Built with the best of the market

- 256 nodes
- 250 GFLOPS
- 1.7 Kwatt

Built with the best that is coming

- 120 TFLOPS
- 80 Kwatt

Integrated ARM + GPU

What is the best that we could do?

- 200 PFLOPS
- ~10 MWatt

GFLOPS / W


250 GFLOPS

PRAECE


2017

2016

2015

2014

2013

2012

2011
Mont-Blanc 3: Dibona test platform

- Power supply units
- 48 computes nodes: 16 blades, 96 CPUs, 3000 cores, 12000 threads
- Hydraulics for Direct Liquid Cooling (ultra-energy efficient with hot water cooling)
- IB EDR switches
- Redundant management server including storage
- Internal Ethernet management network
Dibona Processor: Cavium ThunderX2

1U form factor
Direct liquid cooling
3 compute nodes per blade with:
- 2 Cavium ThunderX2 processors (up to 32 Armv8 cores per CPU, 4 threads per core, up to 2.5GHz)
- 16 DDR4 DIMM slots
- 1 Interconnect mezzanine board (EDR)
Mont-Blanc 2020: Architectural Vision

32 Compute Clusters
Shared Last-Level Cache
Designing Mont-Blanc 2020 Compute Clusters

Mesh router

Cluster Shared L3 (MBs)

System Shared Last Level Cache (slice of MBs)
BSC and the European Commission

Final plenary panel at ICT - Innovate, Connect, Transform conference, 22 October 2015 Lisbon, Portugal.

The transformational impact of excellent science in research and innovation

“Europe can develop an exascale machine with ARM technology. Maybe we need an consortium for HPC and Big Data”.

Seymour Cray Award Ceremony  Nov. 2015
Mateo Valero
The European Commission and HPC

European Commission President Jean-Claude Juncker
"Our ambition is for Europe to become one of the top 3 world leaders in high-performance computing by 2020"

Paris, 27 October 2015

Vice-President Andrus Ansip
"I encourage even more EU countries to engage in this ambitious endeavour"

• Ministers from seven MS (France, Germany, Italy, Luxembourg, Netherlands, Portugal and Spain) sign a declaration to support the next generation of computing and data infrastructures

Digital Day Rome, 23 March 2017
The EuroHPC Declaration

Declaration signed in Rome, March 23rd, 2017 by:

France        Germany        Italy        Luxembourg        Netherlands        Portugal        Spain

Six more countries signed the Declaration:

Belgium        Slovenia        Bulgaria        Switzerland        Greece        Croatia

Agree to work towards the establishment of a cooperation framework - EuroHPC - for acquiring and deploying an integrated exascale supercomputing infrastructure that will be available across the EU for scientific communities as well as public and private partners.
EuroHPC news:

Europa portal: (January 2018)

Commission proposes to invest EUR 1 billion in world-class European supercomputers

Brussels, 11 January 2018

The European Commission unveiled today its plans to invest jointly with the Member States in building a world-class European supercomputers infrastructure.

Supercomputers are needed to process ever larger amounts of data and bring benefits to the society in many areas from health care and renewable energy to car safety and cybersecurity.
European Processor Initiative (EPI)

- Dates: end 2018 – end 2022
- Budget: **120M€** (+ external funding to be defined)
- EPI objectives aligned with EuroHPC initiative
- Consortium:
  - ATOS/Bull (coordinator), BSC, Infineon, BMW, ST, ExToll, E4, SemiDynamics, Juelich, ETHZ, CEA, FORTH, Kalray, ...
- Define, design and manufacture:
  - European **general purpose processor** → probably based on Arm ISA
  - European **accelerator** → based on RISC-V ISA
- Suitable for multiple markets: HPC, Big Data, automotive, Internet-of-Things, etc.
- **Goal**: deploy pre-Exascale (2021) and Exascale systems (2023) based on EPI technology in selected European Institutions
EPI 23 partners, from research to industry
from consortium to EU high tech fabless

Fabless company
Industrial hand of EPI
Incorporated by a
couple EPI members
and external investors

1st EPI production
Three streams

- General purpose and Common Platform
  - Arm SVE or other candidates...
  - Bull: System integrator → chip integrator
  - Mont-Blanc 2020 Architectural Vision

- Accelerator
  - RISC-V
  - EU design: BSC, CEA, Chalmers, ETHZ, EXTOLL, E4, FORTH, Fraunhofer, IST, UNIBO, UNIZG, Semidynamics

- Automotive
  - Infineon, BMW...
EPI ROADMAP

Core Technology
- ARM CPU?
- RISC-V ACCEL.

Key Markets
- HPC
- CARS

2018
SGA 1&2

2019

2020

2021

2022

2023

2024

2025
SGA4

Gen 1
- HPC Chip & system
- Accel. Chip & system

Gen 2
- HPC Chip & system
- Accel. Chip & system

Gen 3
- Crossover chip

HPC System
- PreExascale
- Exascale

Automotive CPU
- Proof of Concept
- Product

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

European Processor Initiative
RISC-V accelerator vision @ EPI

- High throughput devices
  - Long Vectors (a la Cray? A la Cyber205? ...)
    - Decouple Front end - Back end engines
    - Optimize memory throughput ([Command vector, 98])
    - Explicit locality management (long register file)

- ISA is important
  - Decouple/hide again hardware details, reuse SW technologies (compilers, OS,...), Specific instructions?
  - “limited” number of control flows

- Low power: ~ low voltage x ~ low frequency

- Hierarchical Acceleration
  - Nesting

- MPI+OpenMP
  - Task based, throughput oriented programming approach
  - Malleability in application + Dynamic resource (cores, power, BW) management
  - Intelligent runtimes & Runtime Aware Architectures
    - Architectural support for the runtime

- Accelerator for ML
  - Specialized “non Von-Neumann” compute and data motion engines (neural/stencil)
  - Tuned numerical precision
BSC is Hiring

BSC is looking for talented and motivated students and professionals with expertise in the design and verification of IPs to be integrated into top-level HPC SoC designs. The immediate responsibilities of this group will be related to The European Processor Initiative.

PhD students and experienced professionals (Engineers and/or PhD holders) wanted for:

- RTL/Microarchitecture
- Verification
- FPGA Design

Find out more: https://www.bsc.es/join-us/job-opportunities/103csrre
Or contact: rrhh@bsc.es
MareNostrum 5 in 2023...
Conclusions

• BSC is designing the next generation of supercomputers
  • Runtime-Aware Architectures
  • Mont-Blanc European project series
  • European Processor Initiative (EPI)

• Goal: Ensure technological leadership of Europe for the next decade in High Performance Computing (HPC)

• We need help from all of you to achieve such an ambitious goal 😊
THANK YOU!

Please, contact me at:

miquel.moreto@bsc.es

www.bsc.es