MONT-BLANC

http://www.montblanc-project.eu

Pre(-pre)-exascale experiences, contributions and future challenges

Etienne Walter

Project Manager at Bull/ATOS

Coordinator of the Mont-Blanc 3 project

Filippo Mantovani

Senior Researcher at Barcelona Supercomputing Center

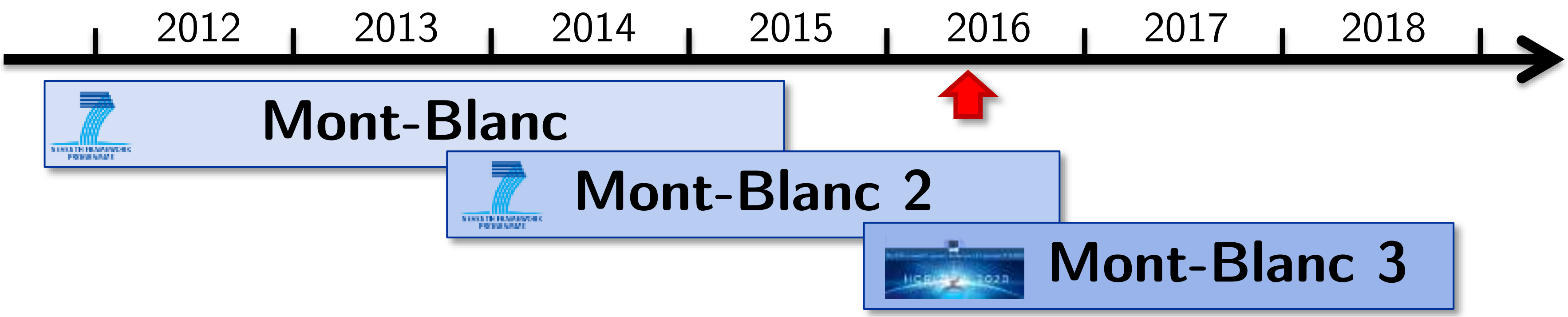
Technical coordinator of the Mont-Blanc 1 and 2 projects

exdci European Extreme Data & Computing Initiative

This project and the research leading to these results has received funding from the European Community's Seventh Framework and H2020 Programmes under grant agreement n° 610402 and 671697.

Mont-Blanc projects in a glance

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



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

allinea

UNI GRAZ

Bull
atos technologies

ARM

JÜLICH
FORSCHUNGSZENTRUM

CINECA

lrz Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

Inria
INVENTEURS DU MONDE NUMÉRIQUE

University of BRISTOL

cea

UC UNIVERSIDAD DE CANTABRIA

AVL

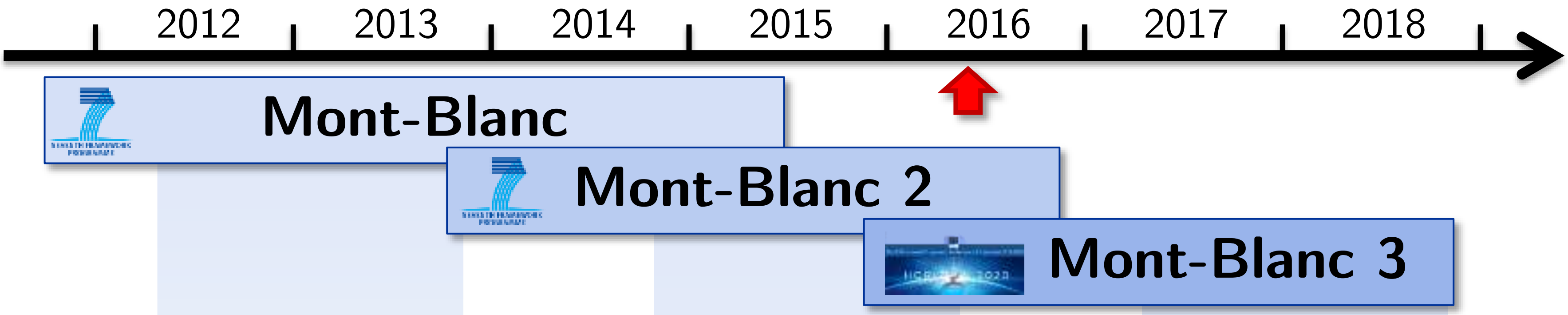
cnrs

HLRS

ETH Zürich

UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

Mont-Blanc objectives



- HPC prototype based on mobile embedded technology
- Port and test real scientific applications
- Learn from the experience, plan for future architecture

Extend

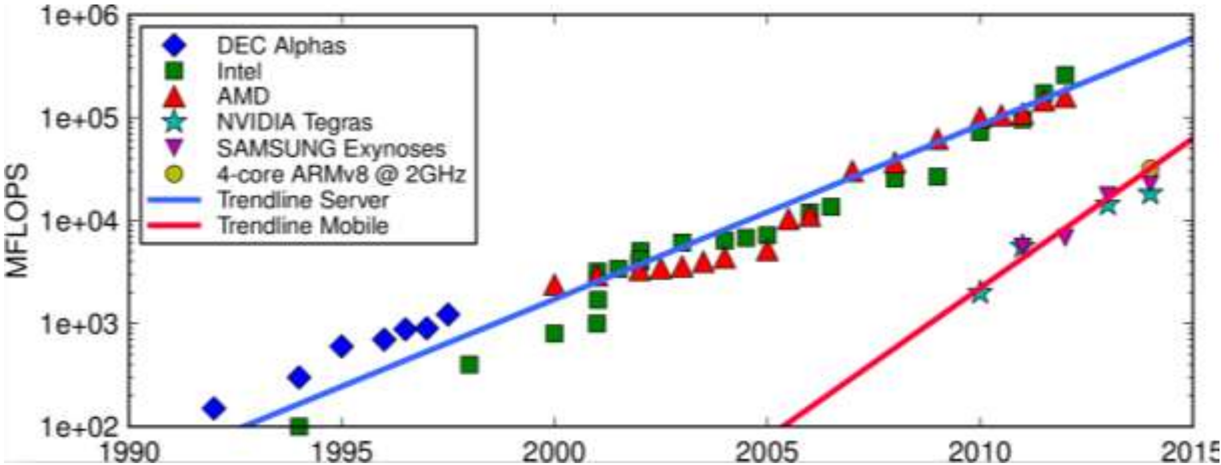
- Support hw and system sw:
  - OmpSs programming model
  - Productivity tools
- New scientific and industrial applications
- Next generation Mont-Blanc architecture

Explore

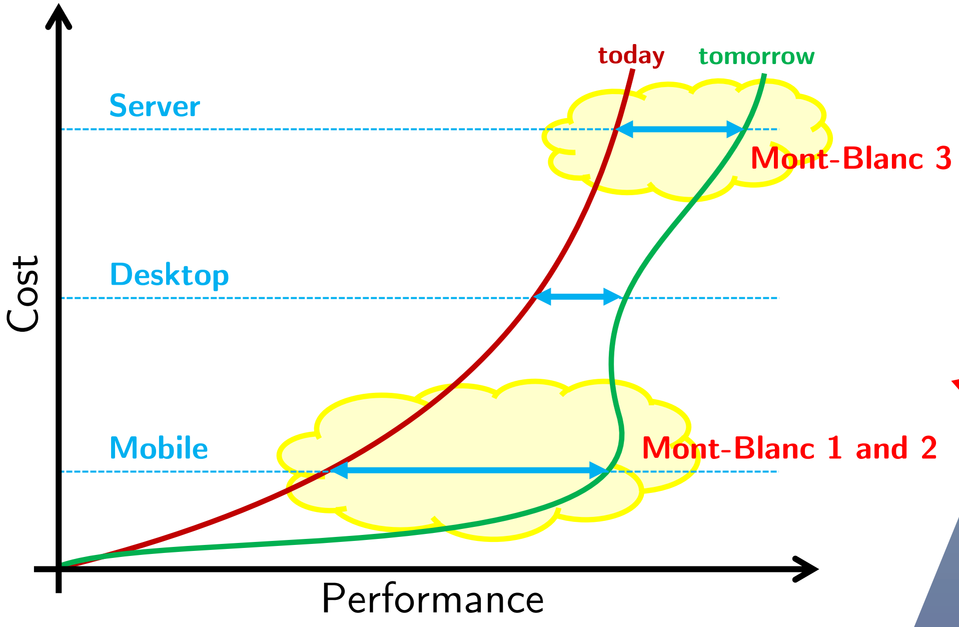
- ARM 64-bit
- Fault tolerance and resiliency
- Market of ARM-based platforms for mini-clusters

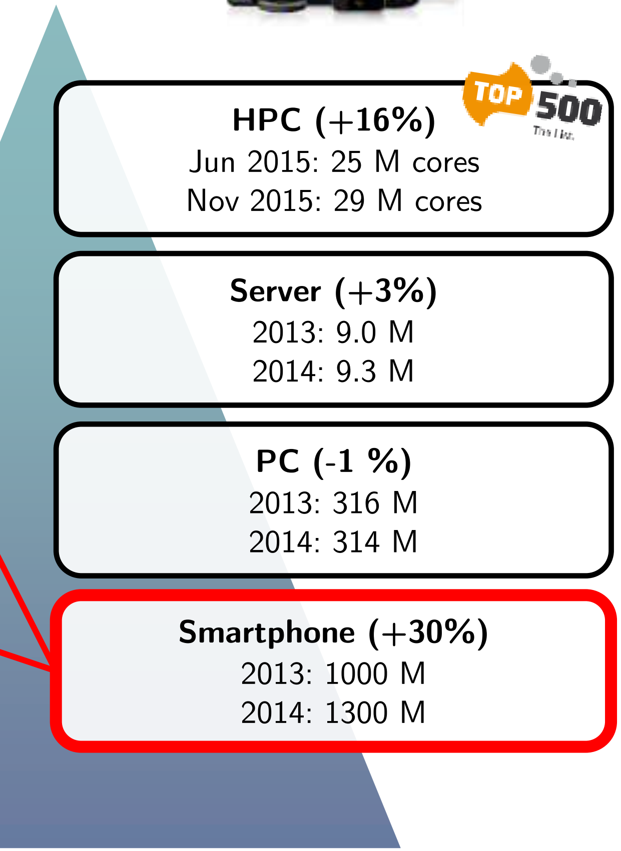
- Balanced ARM-based architecture targeting pre-exascale performance
- Focus on compute efficiency:
  - New high-performance ARM architecture
  - Throughput-oriented compute accelerators
- Co-design approach:
  - Architecture
  - System software
  - Applications

Leveraging a fast-growing market



...and we are still ignoring tablets: >200M





HPC (+16%)
Jun 2015: 25 M cores
Nov 2015: 29 M cores

Server (+3%)
2013: 9.0 M
2014: 9.3 M

PC (-1 %)
2013: 316 M
2014: 314 M

Smartphone (+30%)
2013: 1000 M
2014: 1300 M

Mont-Blanc contributions

- Prototyping
  - Custom integration
  - Solutions on the market
- Enabling co-design
  - System software for ARM
  - Scientific applications at scale
  - Architectural studies for next-generation platforms
- Outreach
  - Industry
  - Research
  - Education

What are you offering to Ecosystem in terms of services/access/cooperation?

The Mont-Blanc prototype ecosystem

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

PRACE prototypes
- Tibidabo
- Carma
- Pedraforca

Mini-clusters
- Arndale
- Odroid XU
- Odroid XU-3
- NVIDIA Jetson

Mont-Blanc prototype

ARM 64-bit mini-clusters
- APM X-GENE2
- Cavium ThunderX
- NVIDIA TX1

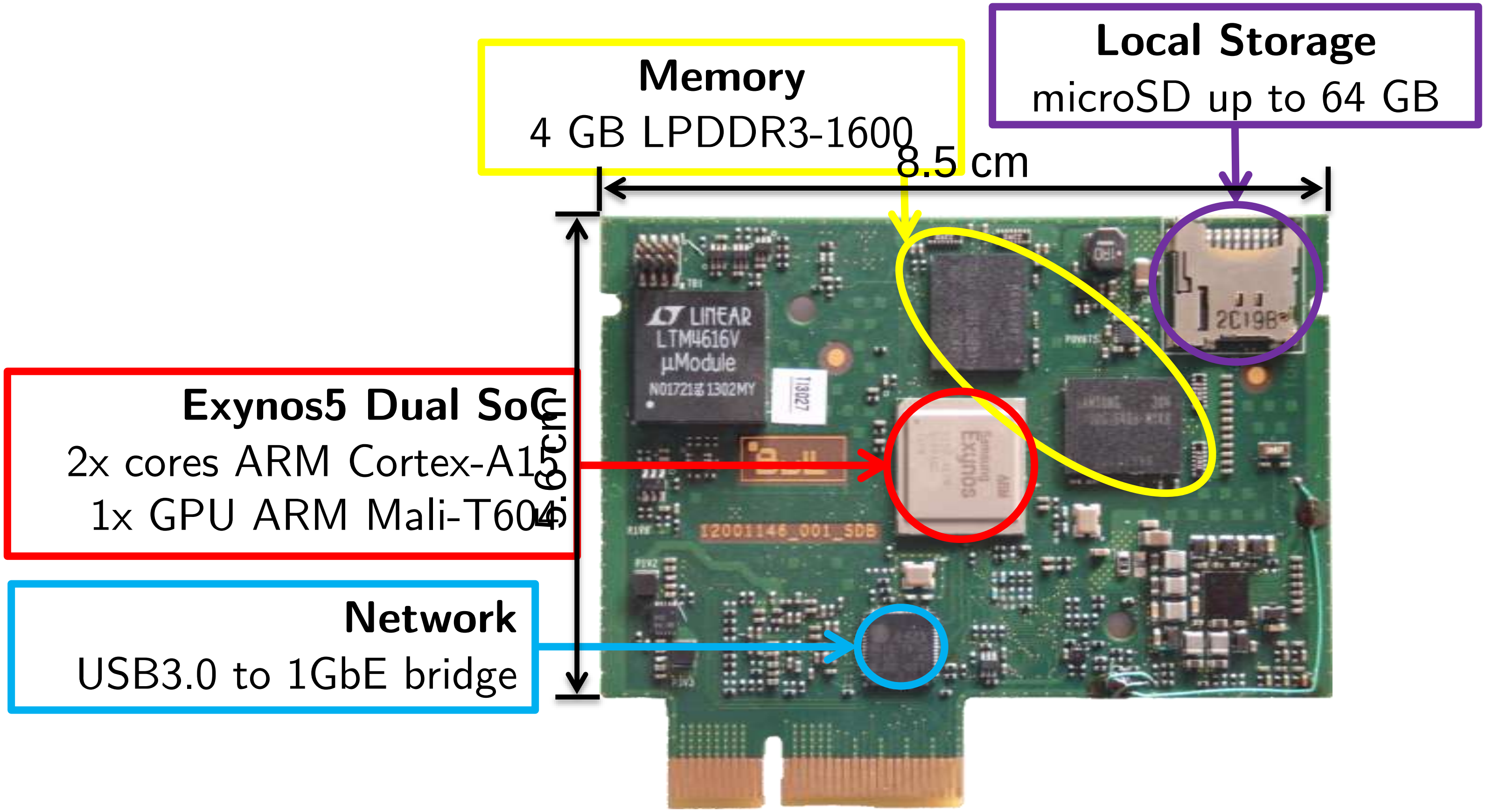
Mont-Blanc 3 demonstrator
- Based on new-generation ARM 64-bit processors
- Targeting HPC market

2012 2013 2014 2015 2016 2017

Mont-Blanc Server-on-Module (SoM)

CPU + GPU + Memory + Local Storage + Network

Form factor: 8.5 x 5.6 cm



Memory
4 GB LPDDR3-1600

Local Storage
microSD up to 64 GB

Exynos5 Dual SoC
2x cores ARM Cortex-A15
1x GPU ARM Mali-T604

Network
USB3.0 to 1GbE bridge

The Mont-Blanc prototype

Exynos 5 compute card
2 x Cortex-A15 @ 1.7GHz
1 x Mali T604 GPU
6.8 + 25.5 GFLOPS
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 Carrier blade
135 x Compute cards
4.3 TFLOPS
2.7 kWatts
1.6 GFLOPS/W

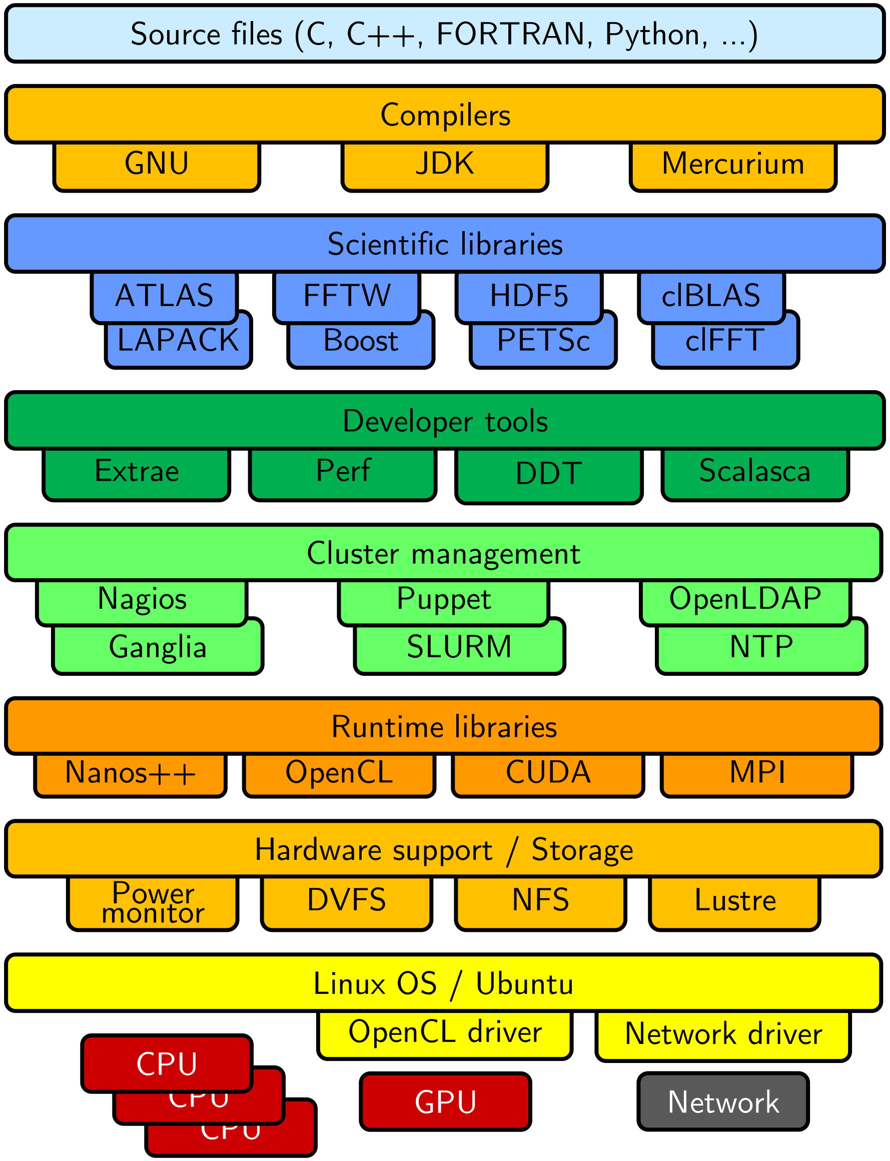
Rack
8 BullX chassis
72 Compute blades
1080 Compute cards
2160 CPUs
1080 GPUs
4.3 TB of DRAM
17.2 TB of Flash

35 TFLOPS
24 kWatt

A perfect playground for addressing important questions for next-generation architectures:

- Can we take real advantage of HMP?
- Can we survive without ECC?
- Can we scale 'something' with one Gigabit Ethernet network?
- Can we learn 'something' from power profiles?

System software stack for ARM



Source files (C, C++, FORTRAN, Python, ...)

Compilers
- GNU
- JDK
- Mercurium

Scientific libraries
- ATLAS
- LAPACK
- FFTW
- Boost
- HDF5
- PETSc
- clBLAS
- clFFT

Developer tools
- Extrae
- Perf
- DDT
- Scalasca

Cluster management
- Nagios
- Ganglia
- Puppet
- SLURM
- OpenLDAP
- NTP

Runtime libraries
- Nanos++
- OpenCL
- CUDA
- MPI

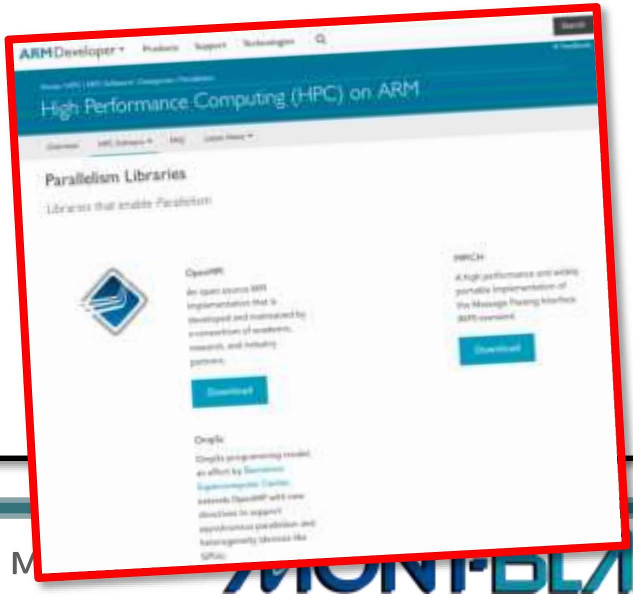
Hardware support / Storage
- Power monitor
- DVFS
- NFS
- Lustre

Linux OS / Ubuntu
- OpenCL driver
- Network driver

1. Based on open-source packages

2. Deployable on any ARM-based platform: a common ground for collaborations with other H2020 projects

3. Impact on the ecosystem



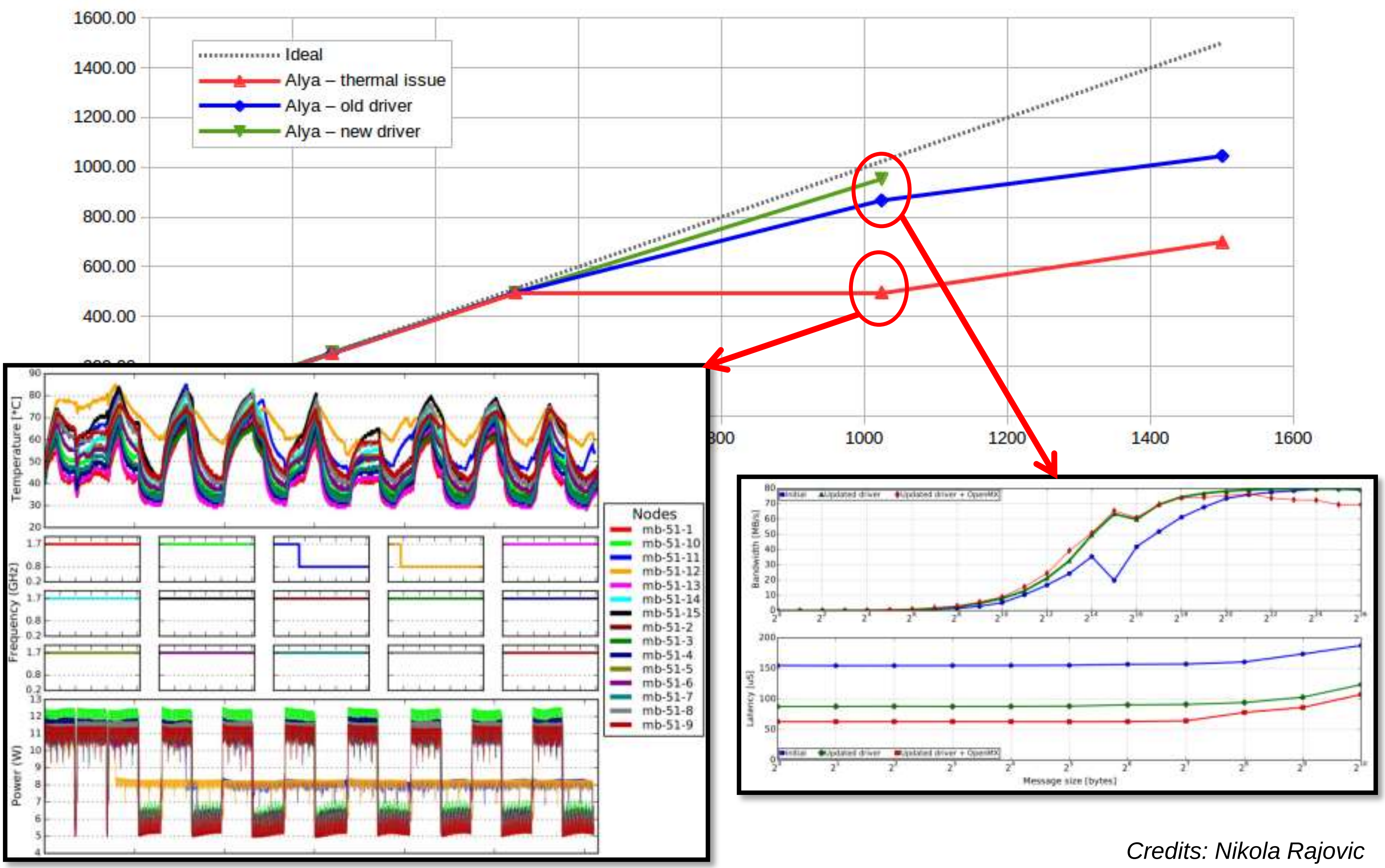
ARM Developer

High Performance Computing (HPC) on ARM

Parallelism Libraries

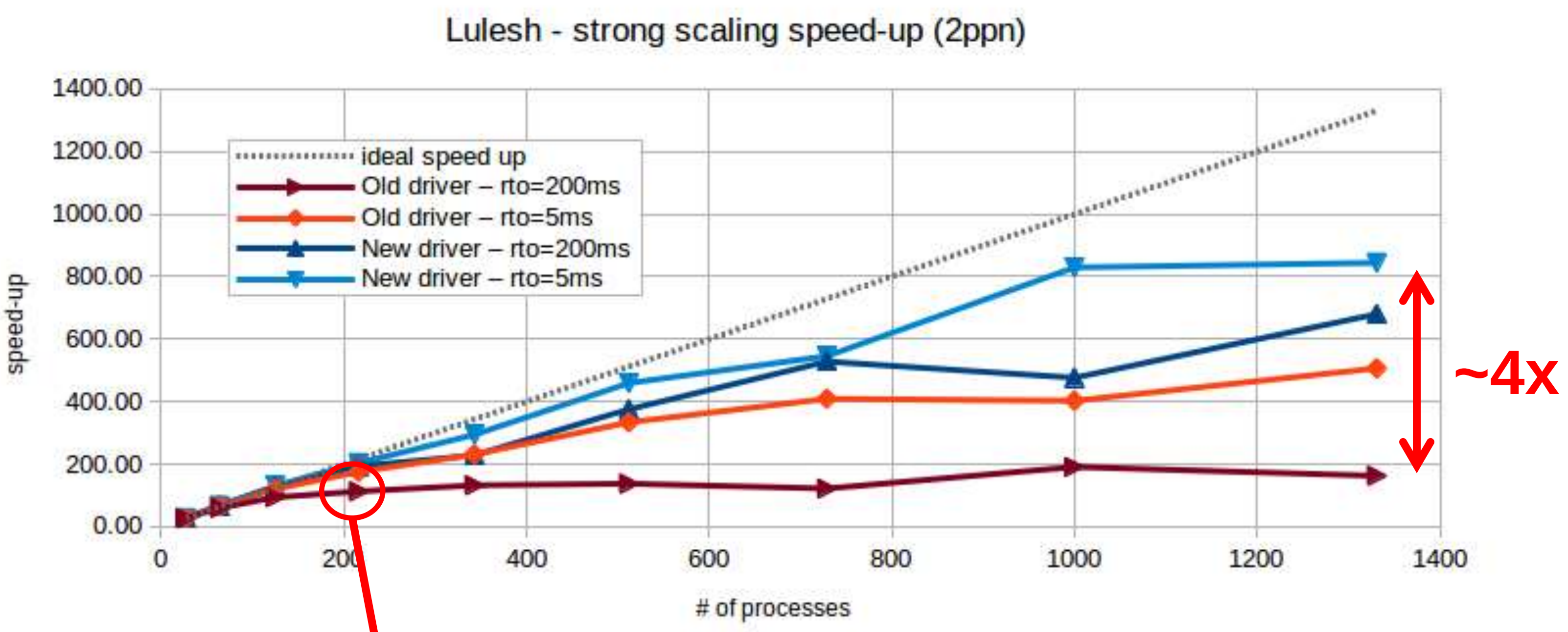
Libraries that enable Parallelism

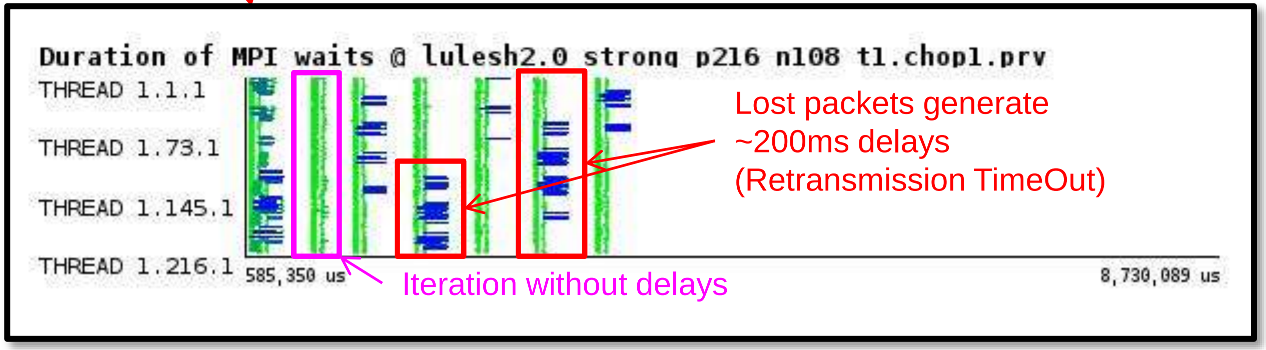
Alya RED on the Mont-Blanc prototype



Credits: Nikola Rajovic

Lulesh on the Mont-Blanc prototype

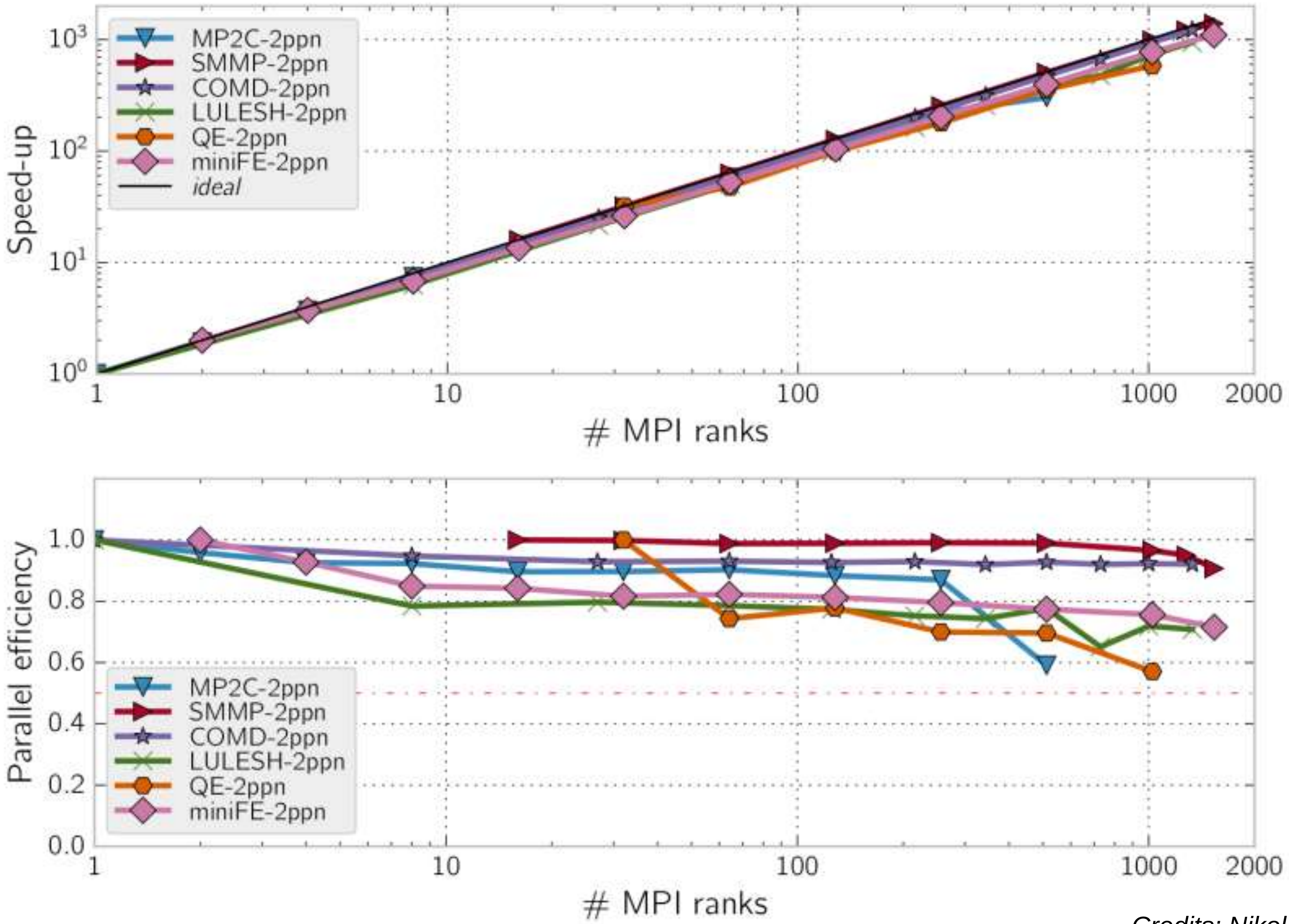



```
Duration of MPI waits @ lulesh2.0 strong p216 n108 t1.chop1.prv
THREAD 1.1.1
THREAD 1.73.1
THREAD 1.145.1
THREAD 1.216.1
585,350 us                                         8,730,089 us
```

Lost packets generate
~200ms delays
(Retransmission TimeOut)

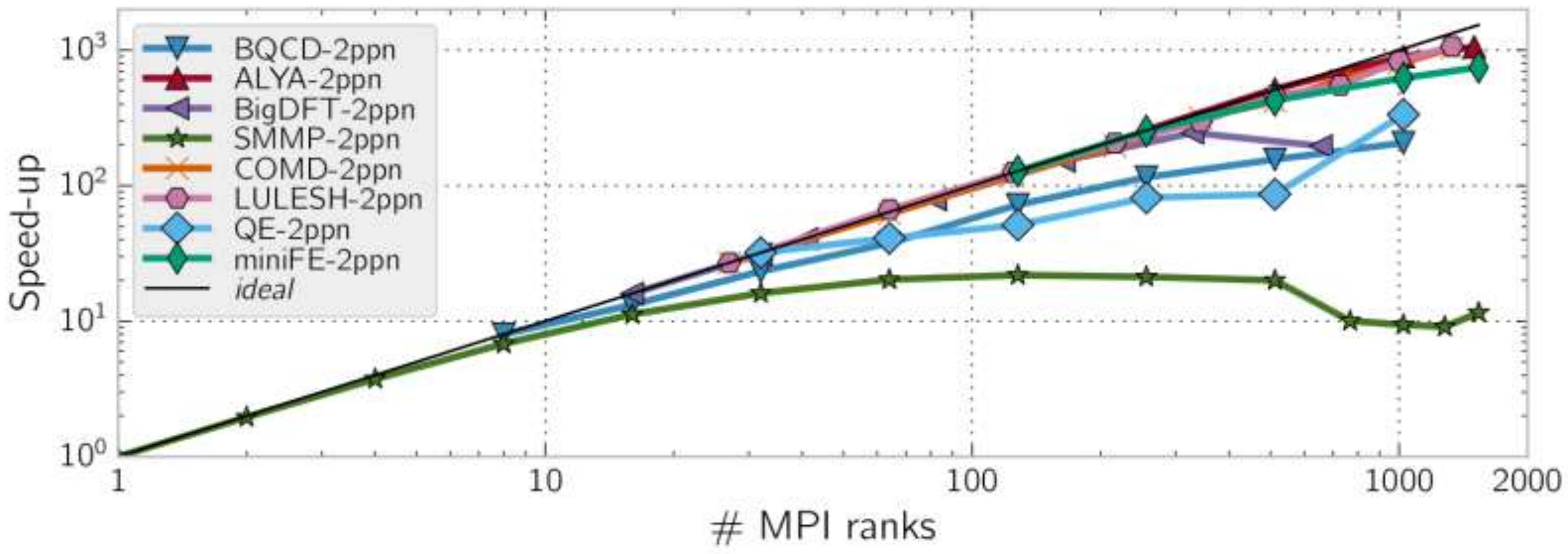
Iteration without delays

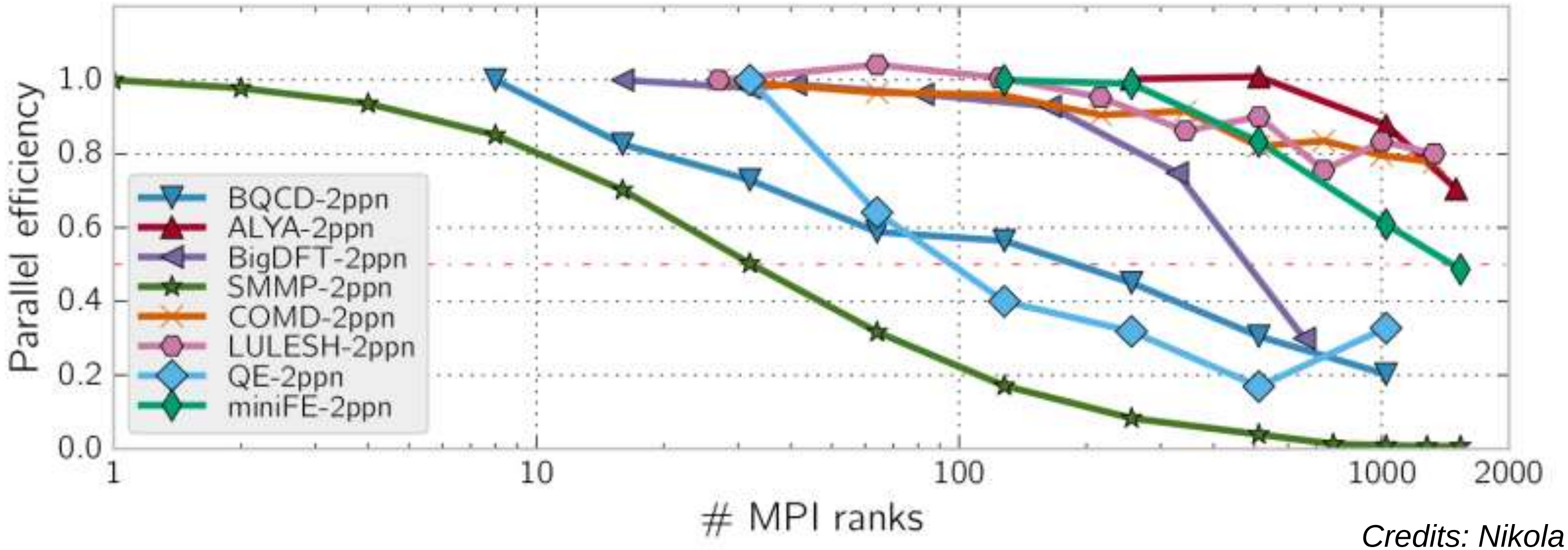
Applications - weak scaling



Credits: Nikola Rajovic

Applications - strong scaling



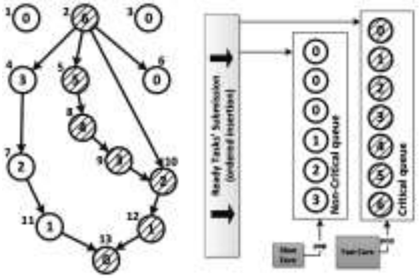


Credits: Nikola Rajovic

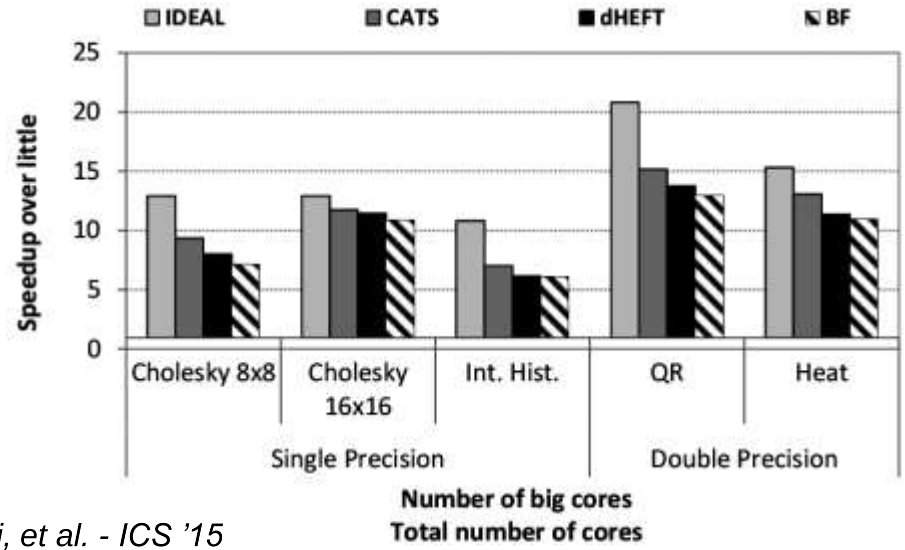
Criticality-Aware Task Scheduler

CATS dynamically assigns critical tasks to fast cores to improve performance in a heterogeneous system, e.g. big.LITTLE

- Scheduling information discoverable at runtime
  - No need of profiling
- Applies to task-based programming models supporting task dependencies
- Evaluation based on Odroid-XU3 + kernels + OmpSs



- Samsung Exynos 5422
- 2GB LPDDR3@933MHz
- 4x Cortex-A15@2.0GHz
- 4x Cortex-A7@1.4GHz

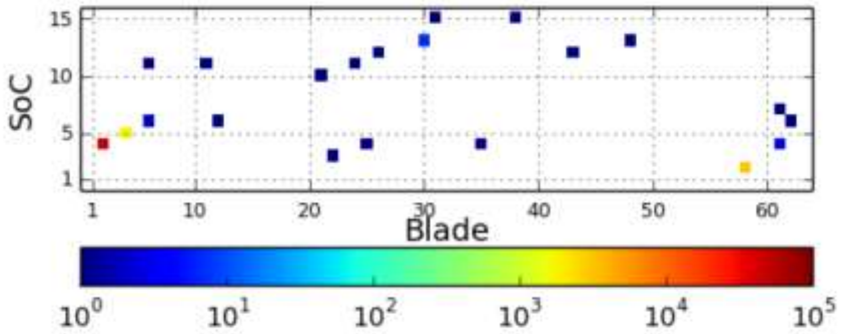


K. Chronaki, et al. - ICS ’15

Memory fault statistics and analysis

Fact: Memory of Mont-Blanc prototype is not ECC protected
Can we survive with this? What do we learn from this?

- Developed a simple in-house memory scanner
- Scanning user-space (~3GB/node) when nodes are idle
- Collected statistics for 13 months / 12 Tbytes/hours
- MTBF per node ~41 hours
- MTBF for the whole prototype ~10 mins



After data analysis, a model of
"quarantine technique"
has been developed.
MTBF from 2.1 hours to 156.9 hours.
Affecting 0.5% of the cluster resources.

99% of the errors are concentrated in 3 nodes

Credits: L. Bautista, F. Zyulkyarov, O. Unsal

Industrial outreach: End-User Group

Commercial ARM based platforms

- Develops a synergy among industry, research centers and partners of the project
- Validates the HPC technologies of the project
- Provides feedback to the project

ROLLS ROYCE

Cenaero
Excellence in Simulation Technologies for Aeronautics

Pharmacelera™

MUREX™

ArcelorMittal

DASSAULT AVIATION

CybeleTech
Technologies numériques pour le monde végétal

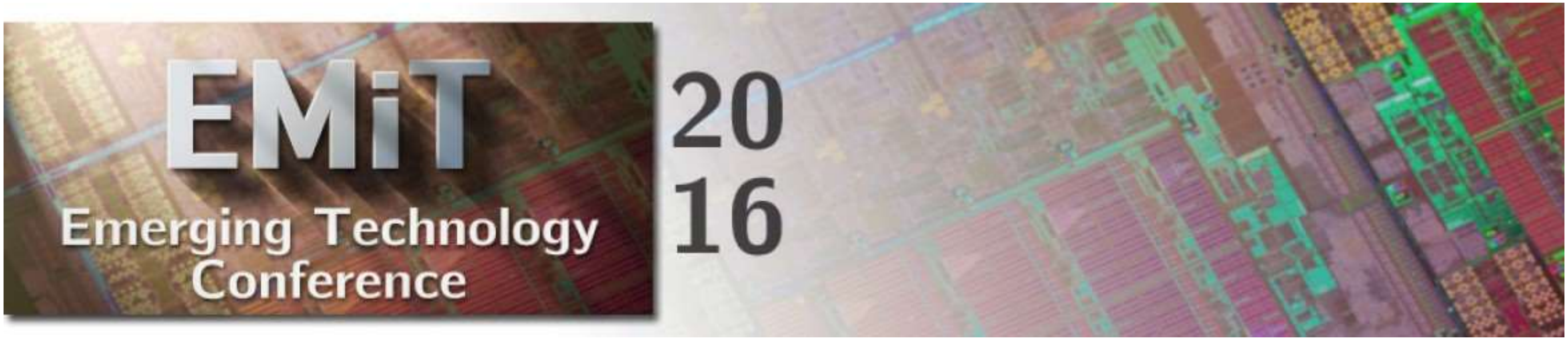
HIGH END COMPUTE

Termo Fluids

Mont-Blanc provides EUG members with:

- Remote access to Mont-Blanc prototype platforms
- Support in platform evaluation and performance analysis
- Invitation to the Mont-Blanc training program

Research outreach: EMiT 2016



Where?	Barcelona, UPC Campus Nord
When?	2nd and 3rd of June 2016
What?	Bring together compute experts that are developing and exploiting emerging technologies

More info @ http://emit.tech/

Educational outreach: Student Cluster Competition

- 11 teams of
  6 undergraduate students
  - From all over the world
  - At the biggest supercomputing conference of Europe
- 3 kW power budget
- 5+ applications
  - Known in advance
  - Some “secret” application
- 3 awards to win
  - Highest HPL
  - 1st, 2nd, 3rd overall places
  - Fan favorite

Team 2015

Team 2016

http://www.montblanc-project.eu

Next step: Mont-Blanc 3 phase

Etienne Walter

Project Manager at Bull/ATOS

Coordinator of the Mont-Blanc 3 project

Filippo Mantovani

Senior Researcher at Barcelona Supercomputing Center

Technical coordinator of the Mont-Blanc 1 and 2 projects

This project and the research leading to these results has received funding from the European Community's Seventh Framework and H2020 Programmes under grant agreement n° 610402 and 671697.

The Mont-Blanc 3 consortium

Short	*Name*	*Type*	*Country*
Bull	Bull	Industry	France
BSC	Barcelona Super Computing	Academic	Spain
ARM	ARM	Industry	UK
ETHZ	ETH Zurich	Academic	Switzerland
LIRMM	CRNS (Centre national de la Recherche Scientifique) / LIRMM	Academic	France
AVL	AVL	Industry (HPC User)	Austria
USTUTT	HLRS, University Stuttgart	Academic (HPC user)	Germany
UNICAN	Universidad de Cantabria (Santander)	Academic	Spain
U Graz	Institute for Scientific Computing of KF Univ. Graz	Academic	Austria
UVSQ	Université de Versailles Saint Quentin	Academic	France

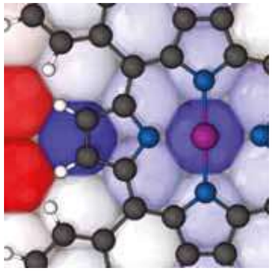
Going forward: Technical challenges (1/2)

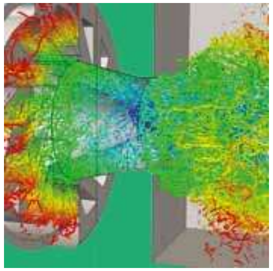
- New challenges with the diminishing technology scaling observed (on core performance, frequency ...)
  - need to leverage new paradigms
- How to optimize the global compute efficiency ?
  - find optimal balance for
    - I/O bandwidth & latency, memory, compute nodes
  - avoid / eliminate unnecessary latencies
    - ➢ from latency-limited to throughput-limited models
  - investigate benefit of heterogeneous architectures

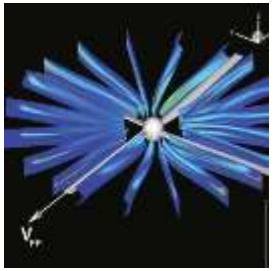
Going forward: Technical challenges (2/2)

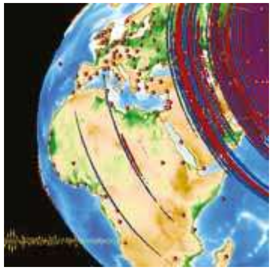
- Assess planned architecture with simulation work
- Enhance the software environment

  run times (OmpSs/OpenMP), tasks operations (MPI), scheduler, compiler & math libraries
- Always keep in mind applications
  - assess solution with real applications
  - enhance performance & energy monitoring tools (MAQAO, system level)
  - hide/minimize the impact of HW architecture



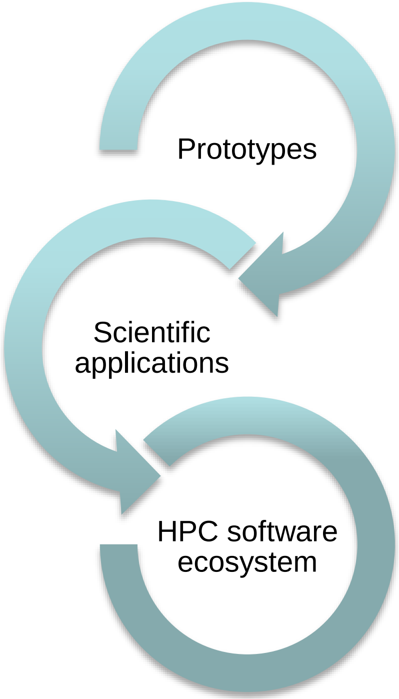




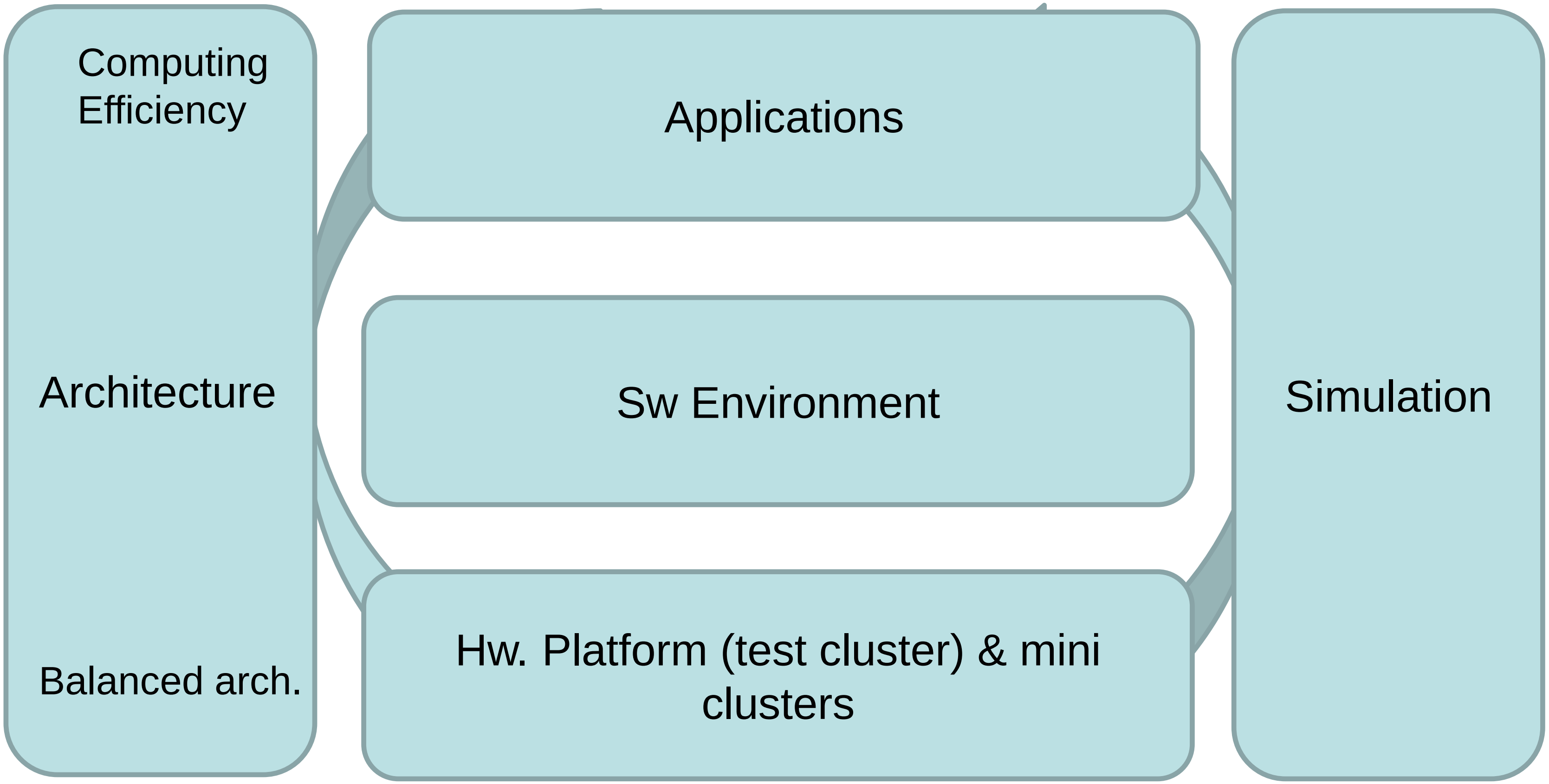


Mont-Blanc 3 project key objectives

- Design a compute node based on ARM architecture for an exascale system
  - Well balanced : Memory, Interconnect, IO
  - Simulation will be used to evaluate the options on applications
- Assess different options for compute efficiency
  - Heterogeneous cores, new option for VPU, high performance core
  - Some assessment with existing solutions will be done using applications
  - Main idea : prepare to transform applications from being latency limited to throughput limited
- Develop the software ecosystem needed for market acceptance of ARM solutions



Project organization



Computing Efficiency

Architecture

Balanced arch.

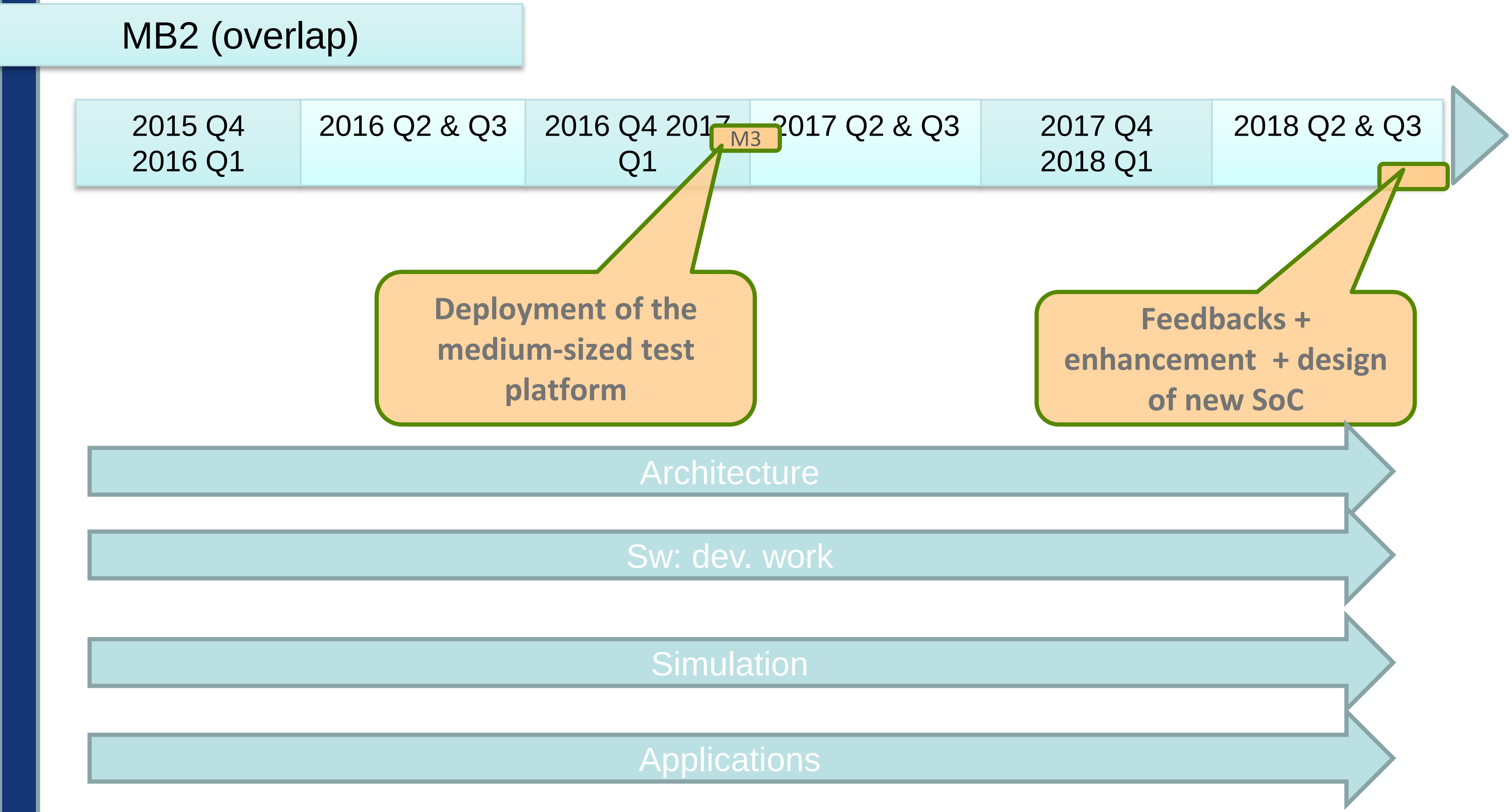
Applications

Sw Environment

Hw. Platform (test cluster) & mini clusters

Simulation

Project general schedule



MB2 (overlap)

2015 Q4 2016 Q1

2016 Q2 & Q3

2016 Q4 2017 Q1

M3

2017 Q2 & Q3

2017 Q4 2018 Q1

2018 Q2 & Q3

Deployment of the medium-sized test platform

Feedbacks + enhancement + design of new SoC

Architecture

Sw: dev. work

Simulation

Applications

EXDCI questions

What are you offering to Ecosystem in terms of services/access/cooperation?

- Cooperation for dissemination
- Sw environment dedicated to ARM based systems
  - shared today
    - on Mont-Blanc project site
    - on https://developer.arm.com/hpc portal
- Access to our test platforms & mini-clusters (to Users member of the End User Group)
  - in return we ask for feedbacks on the realized tests
- Trainings (common trainings for instance – as already done with DEEP/DEPP-ER projects)

EXDCI questions: International cooperation

- Partners individually involved in several international cooperation

- At project level:
    - cooperation initiated with the Argo project (Argonne Nat. Labs, USA) for OS & runtime enhancements

Alignment with the SRA:

- HPC system architecture and components:
  - design a processor improving the energy efficiency in line with the SRA objective
  - work on reducing the data movement inside a compute node to both increase the performance and reduce the power consumption
- System & software management
  - cluster management software: prescriptive maintenance, performance counters
  - resource management & job scheduling (contribution to SLURM)
- Programming environment:
  - research done on runtime, MPI & compiler
- System and environment characteristics:
  - addressing the "energy-efficient design of computer systems" researches : improving the proximity of memory, energy efficient CPU and heterogeneous compute elements
- Balance compute subsystem, IO and storage performance:
  - address the optimization of data transfer through interconnect for both communication between nodes and storage.

Relationship with other FETHPC/CoE :

- open to discussion

- in touch with H2020 projects:
  - DEEP & DEEP-ER (FP7 - *Fault tolerance*)
  - ExaNode (*ARM based compute node*)
  - SAGE (*Storage*)
  - (...)

- CoEs: in touch with several CoEs (PoP, ESiWACE, MAX ...)
  - can potentially propose a compute solution from 2018
  - looking forward feedbacks and further requirements

Brief status (today , May 10th)

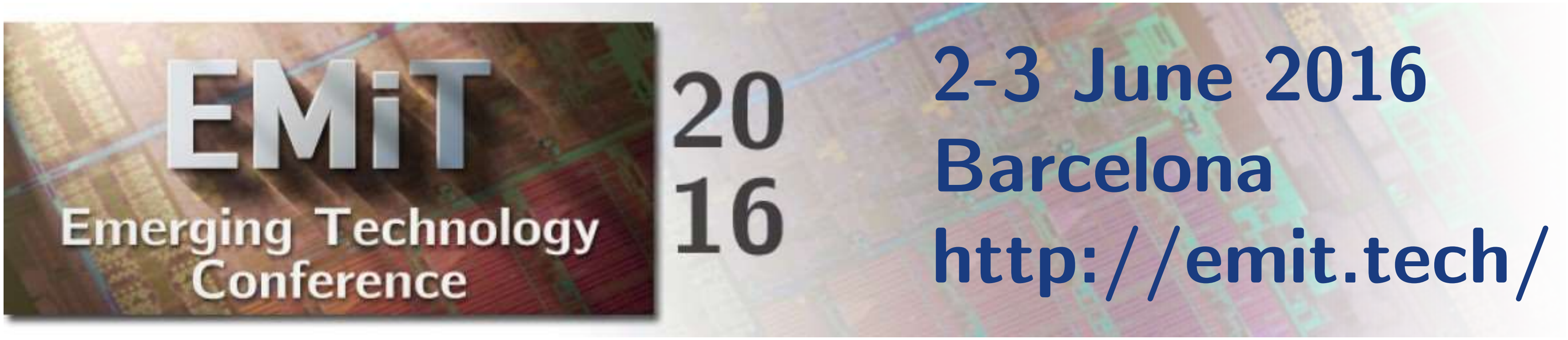
- Work progressing in all domains

- Specific effort done on
    - node architecture (definition of test platform)
    - simulation environment definition
    - software environment preparation

- Specific focus on industrial integration – with a takeover in the project leadership

- But still willing to leverage on ideas generated by scientific research

Mont-Blanc project



2-3 June 2016
Barcelona
http://emit.tech/

filippo.mantovani@bsc.es
etienne.walter@atos.net

montblanc-project.eu

MontBlancEU

@MontBlanc_EU

> “The secret is to win going as slowly as possible.”
> Niki Lauda