

MONT-BLANC

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

Scientific computing on ARM-based platforms: evaluation and perspectives

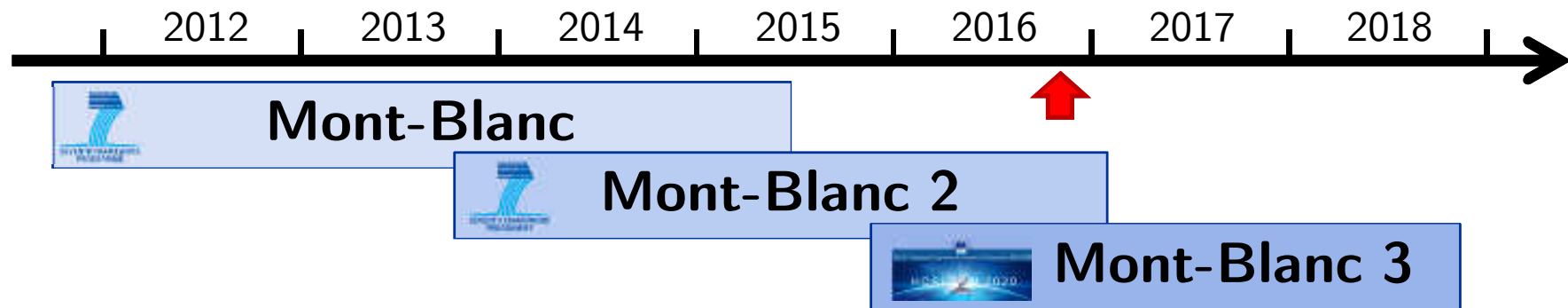
Filippo Mantovani

Senior Researcher at Barcelona Supercomputing Center
Technical coordinator of the Mont-Blanc 1 and 2 projects



Mont-Blanc projects in a glance

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



allinea



HLRIS



University of BRISTOL



Bull
atos technologies

ARM®



Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities



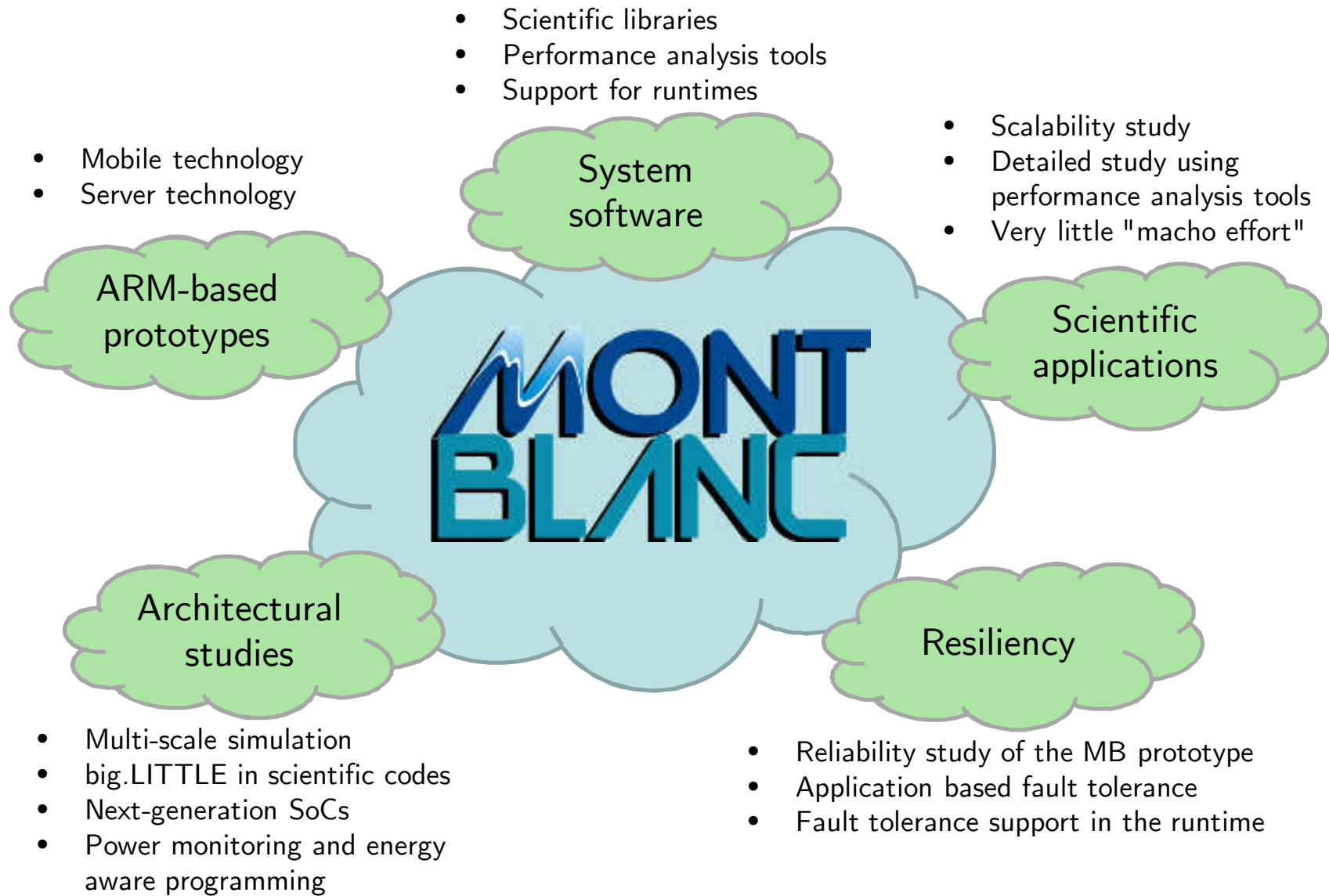
JÜLICH
FORSCHUNGSZENTRUM



UNIVERSITÉ DE
VERSAILLES
ST-QUENTIN-EN-YVELINES

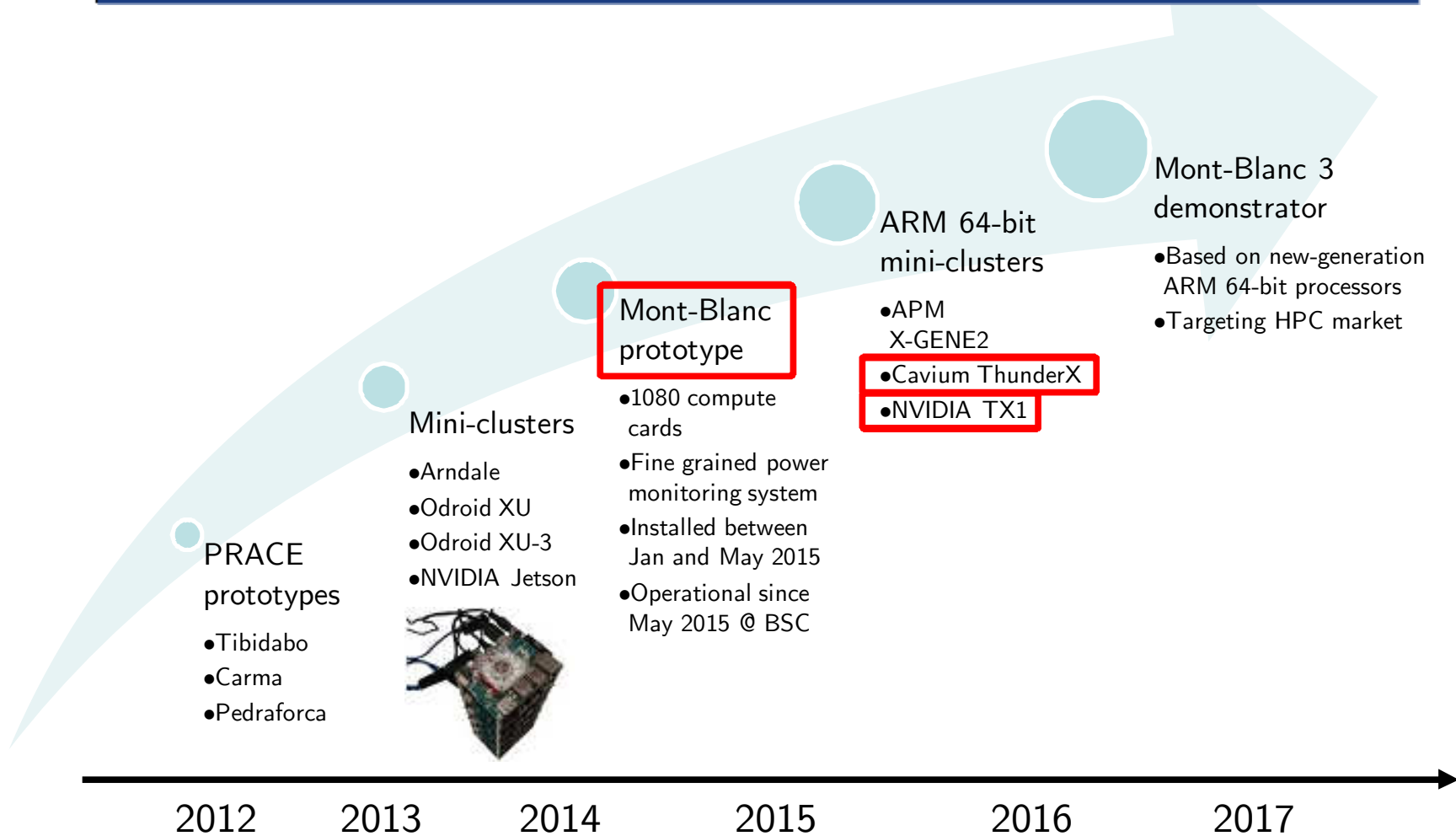


Areas of contribution

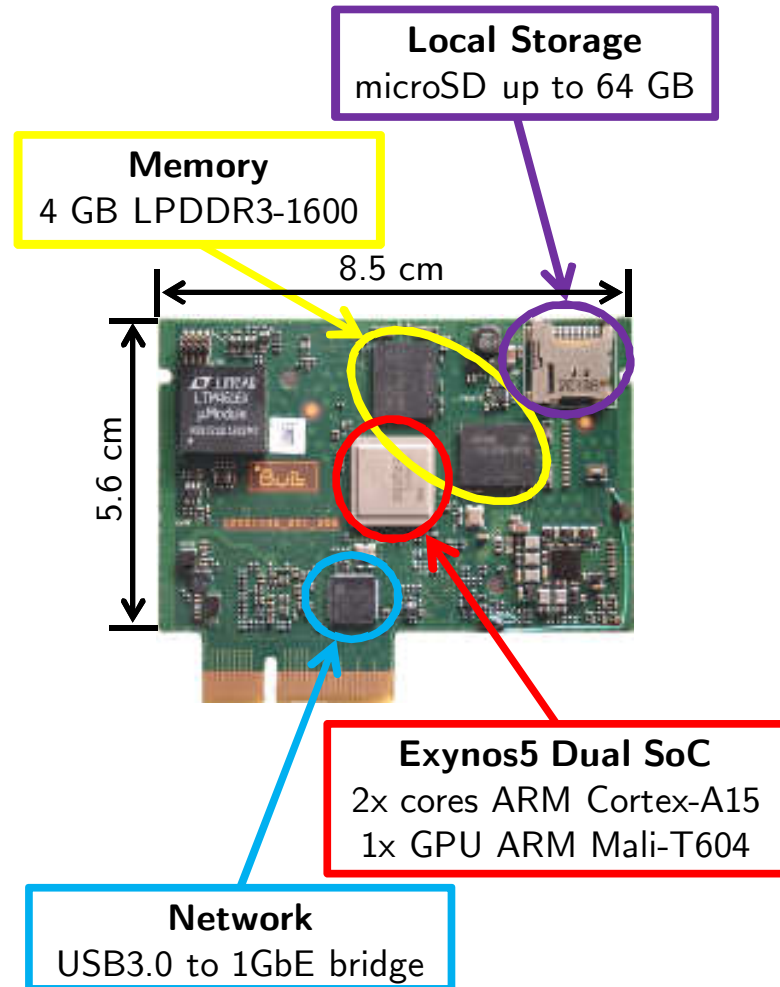


The Mont-Blanc prototype ecosystem

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



Mont-Blanc prototype



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

Operational since May 2015 @ BSC

Jetson TX1 cluster

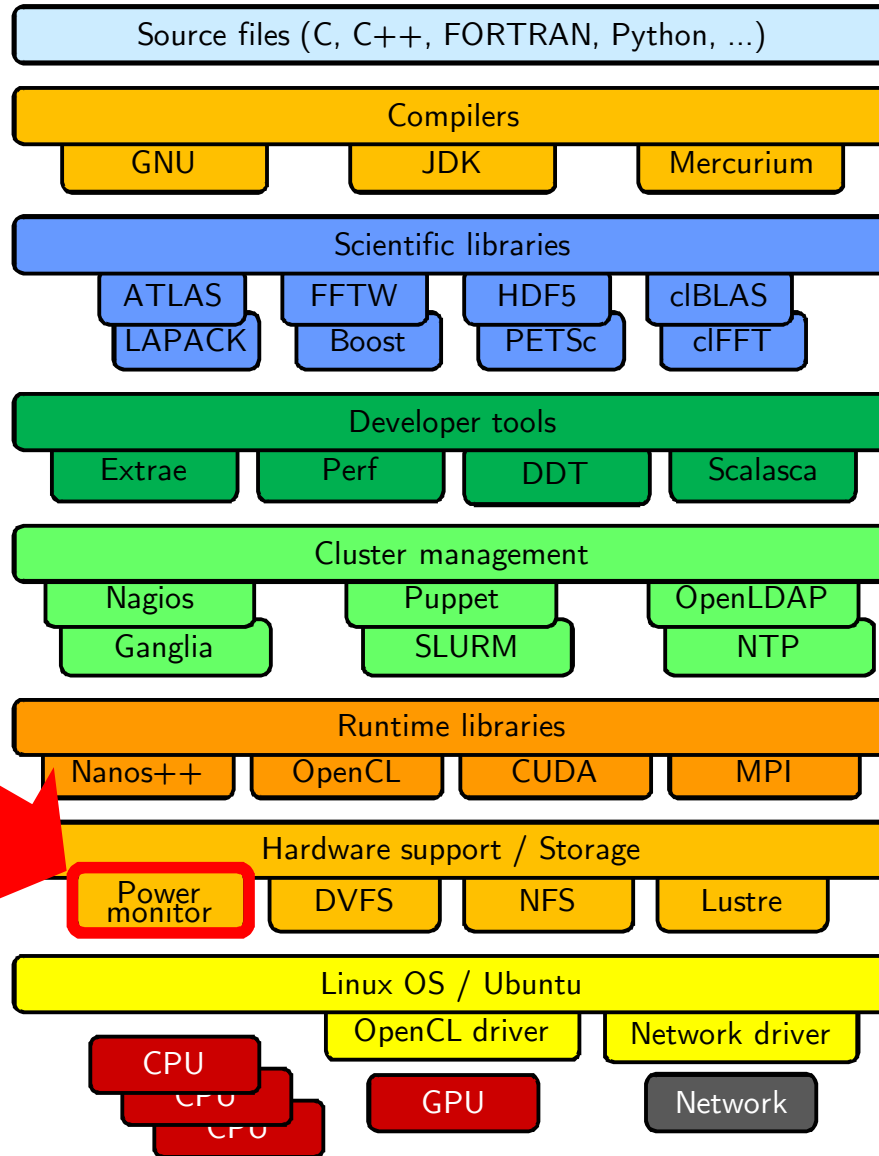
- Same SoC of NVIDIA Shield console
- 1x NVIDIA Tegra X1
 - 4x Cortex-A57 @ 1.73GHz
 - 1x Cortex-A53 (not usable)
- 1x NVIDIA Maxwell GPU
 - 256 CUDA cores
- 4 GB LPDDR4
- 1GbE Network
- Cluster deployed at BSC facilities
 - 16x NVIDIA Jetson TX1 boards
 - Mont-Blanc software stack available



Provided by:



System software stack for ARM



1

Based on open-source packages

2

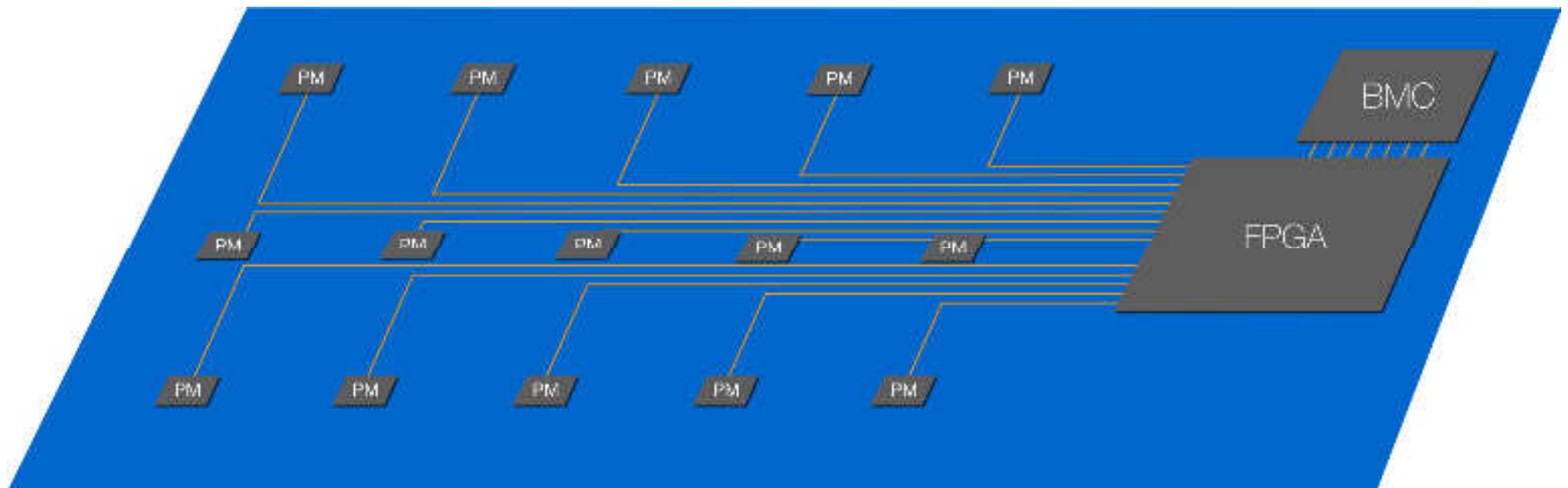
Tested on several ARM-based platform

3



- More than 10 prototypes
- More than 5 years
- More than 4 different ways of measuring the power...
...and still no standards!

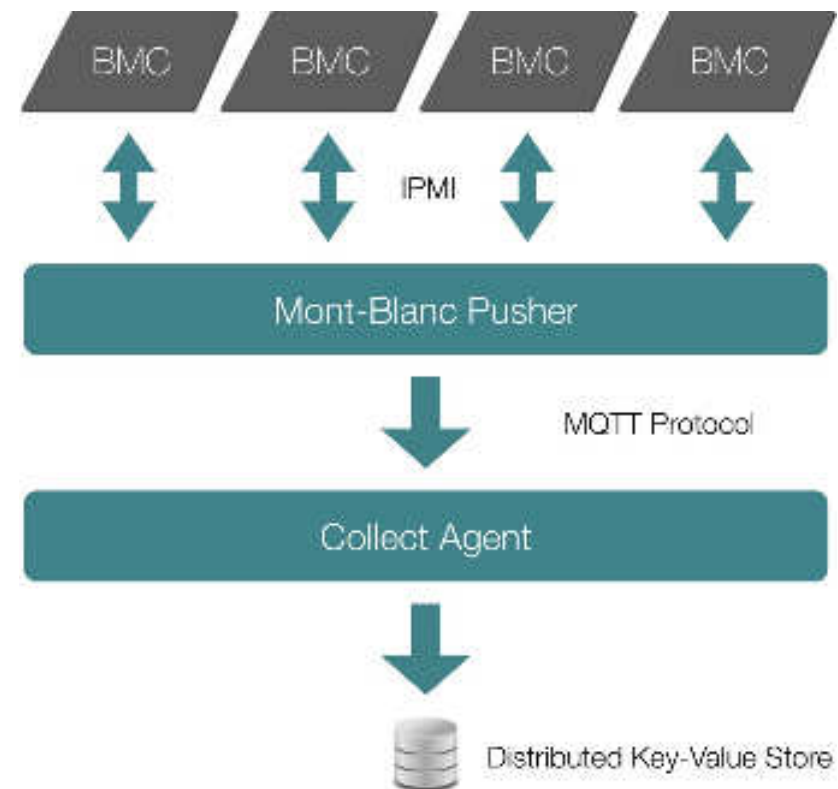
MB-proto: Power monitor – HW infrastructure



Credits: Axel Auweter, Daniele Tafani (LRZ)

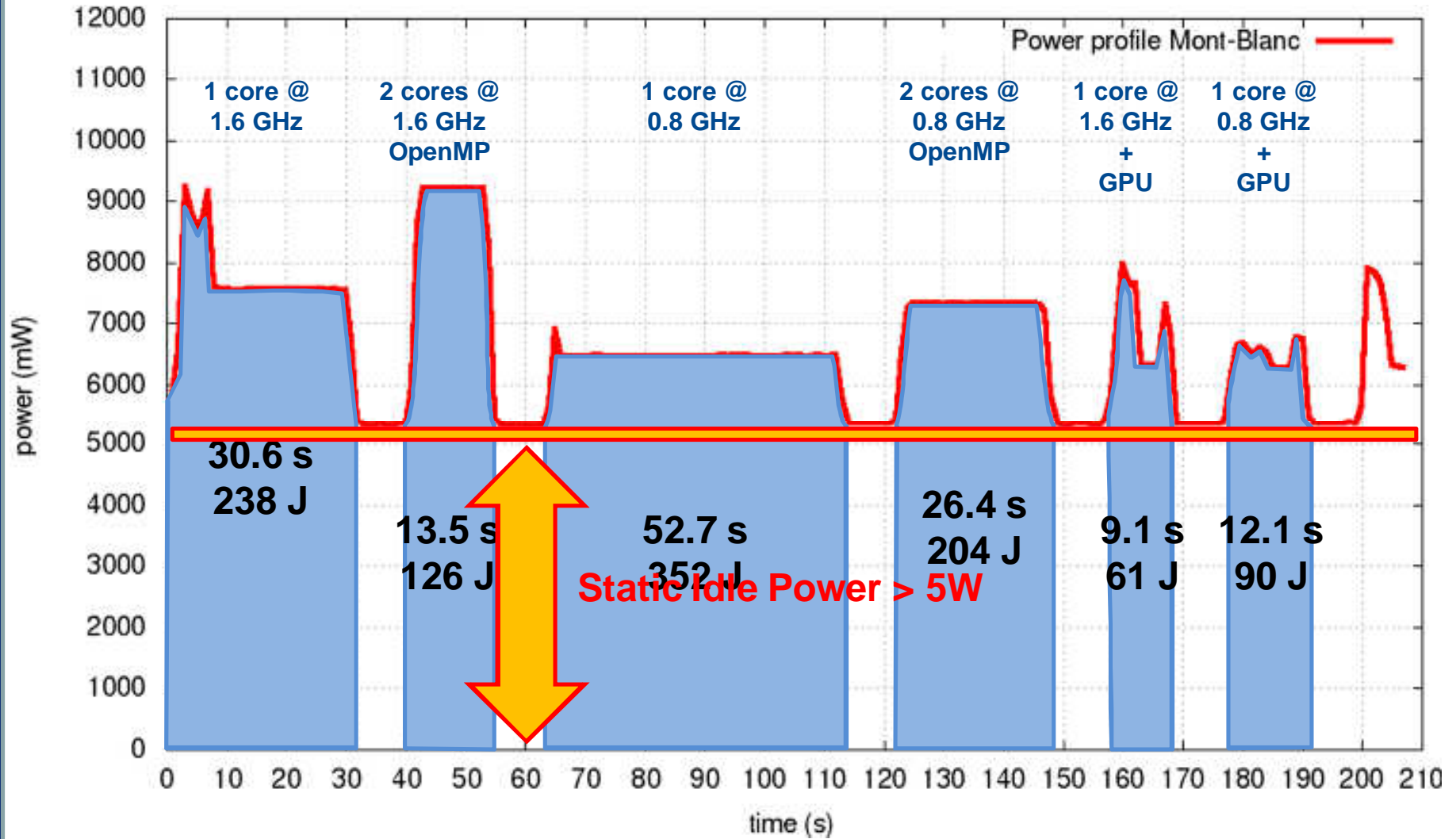
MB-proto: Power monitor – HW / SW interface

- Field Programmable Gate Array (FPGA)
 - Collects power consumption data from all 15 power measurement / sample interval: 70ms
- Board Management Controller (BMC)
 - Collects 1s averaged data from FPGA
 - Stores measurement samples in FIFO
- Mont-Blanc Pusher
 - Collects measurement data from multiple BMCs using custom IPMI commands
 - Forwards data using MQTT protocol through Collect Agent into key-value store

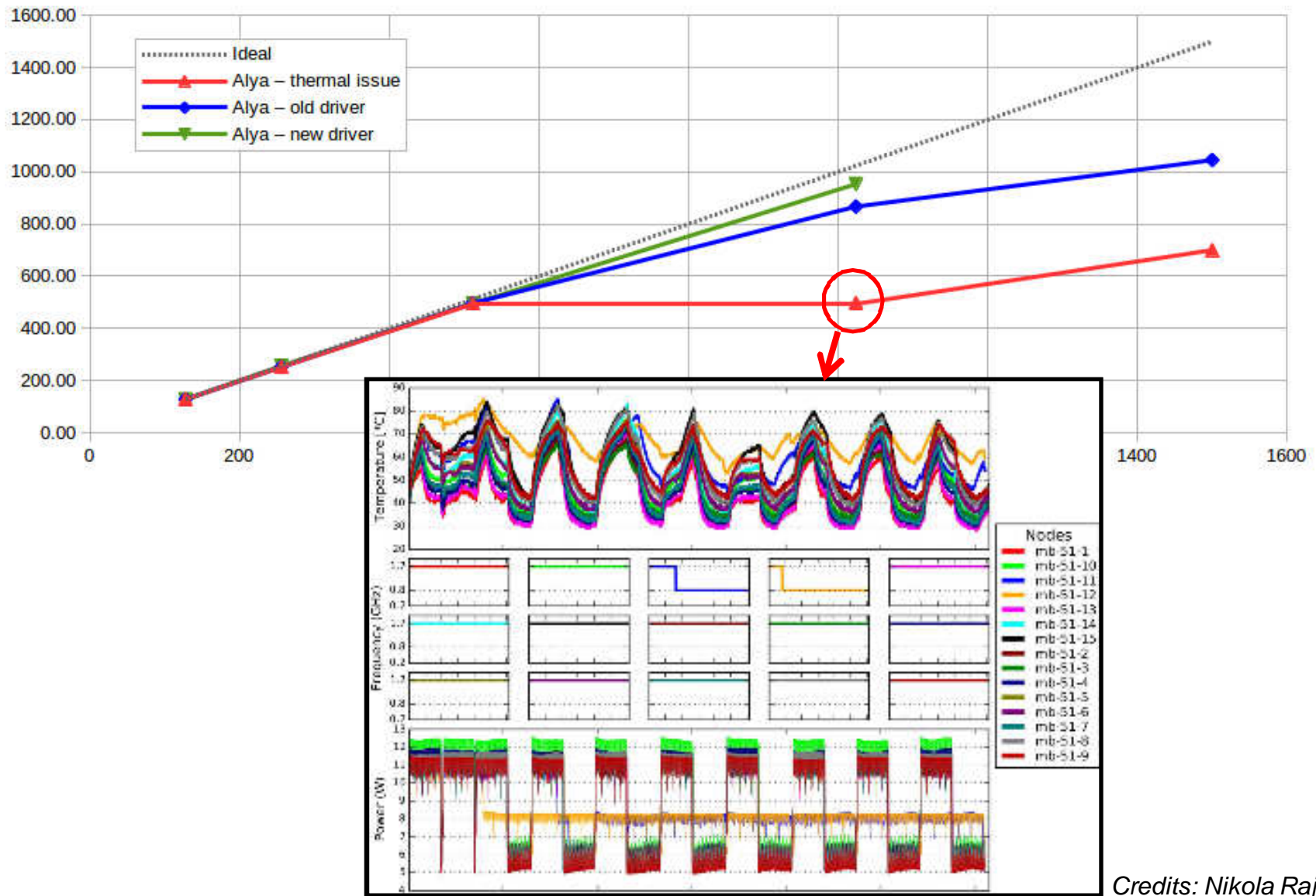


Credits: Axel Auweter, Daniele Tafani (LRZ)

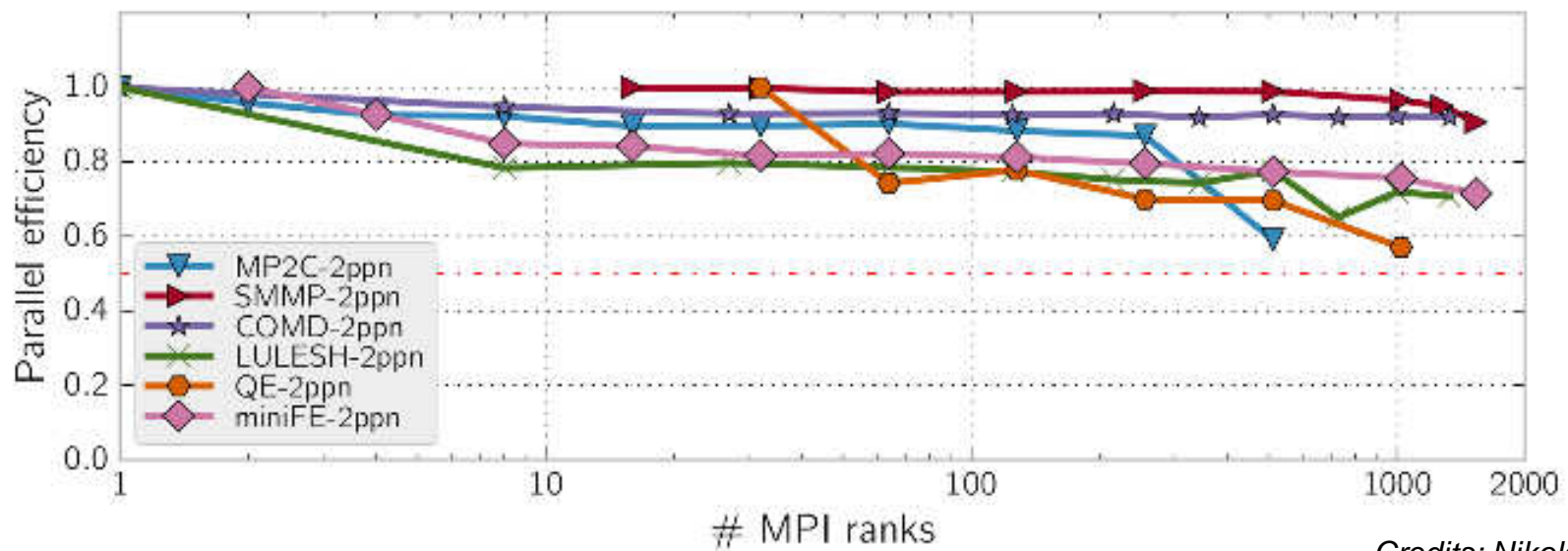
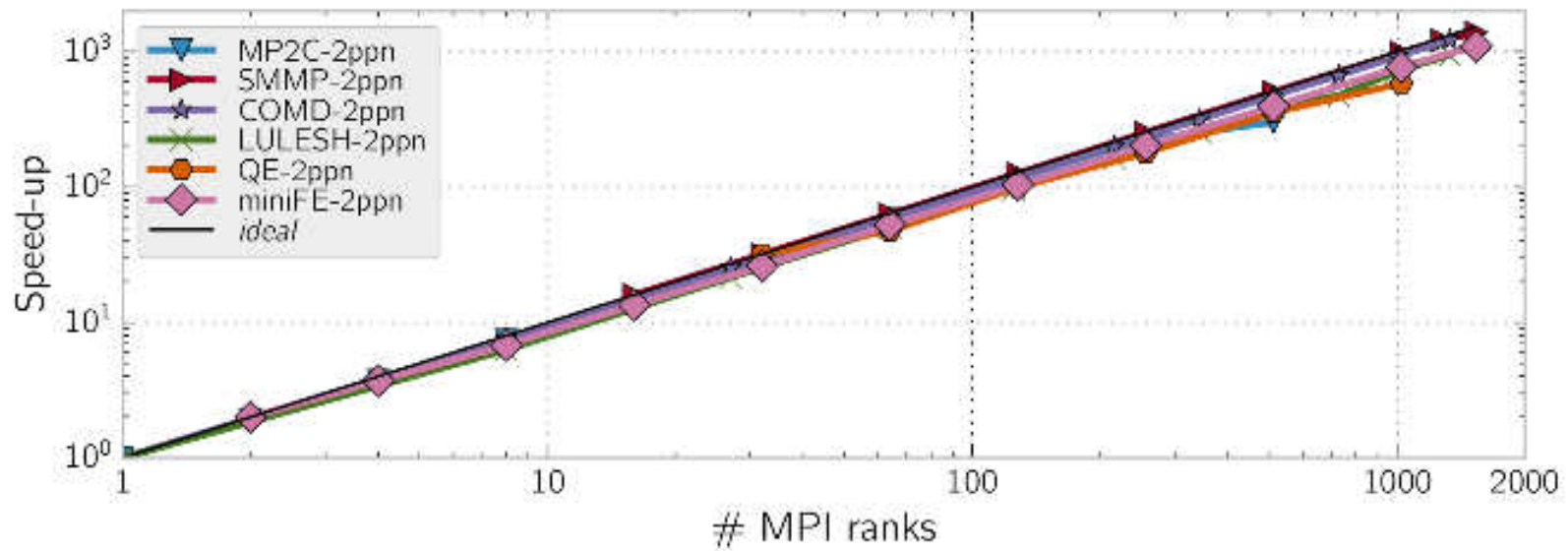
On Mont-Blanc prototype



Alya RED on the Mont-Blanc prototype

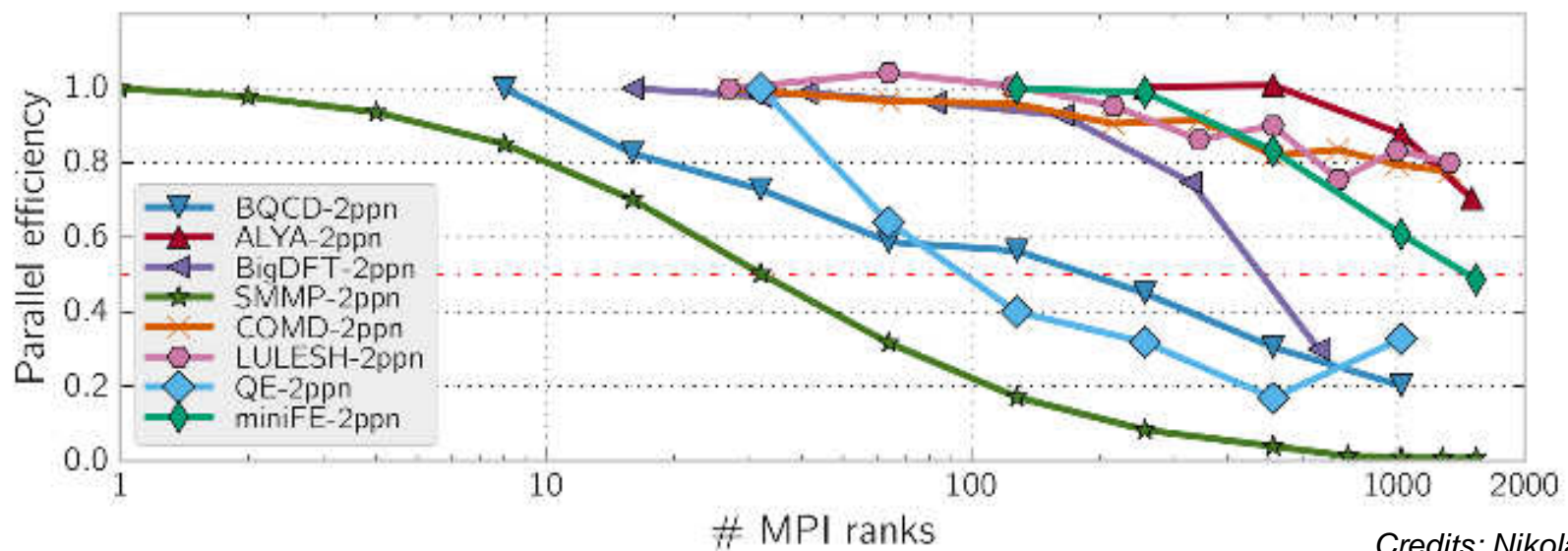
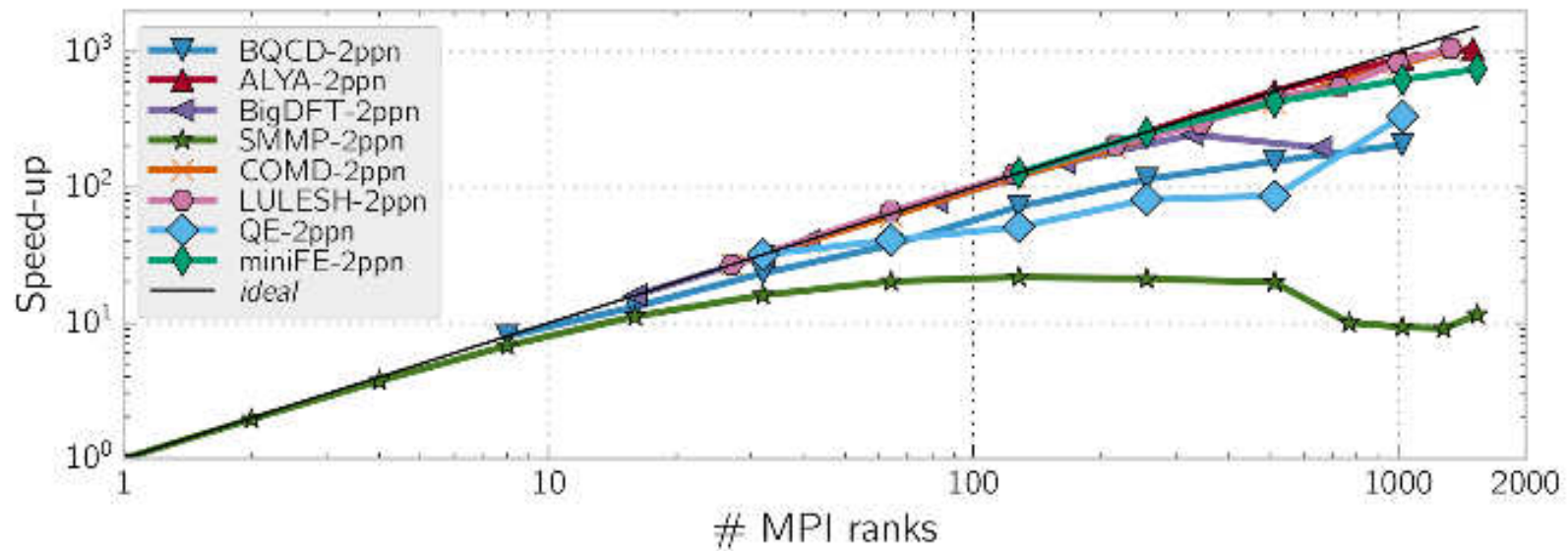


Applications - weak scaling



Credits: Nikola Rajovic

Applications - strong scaling



Credits: Nikola Rajovic

Experimental setup with external power monitor

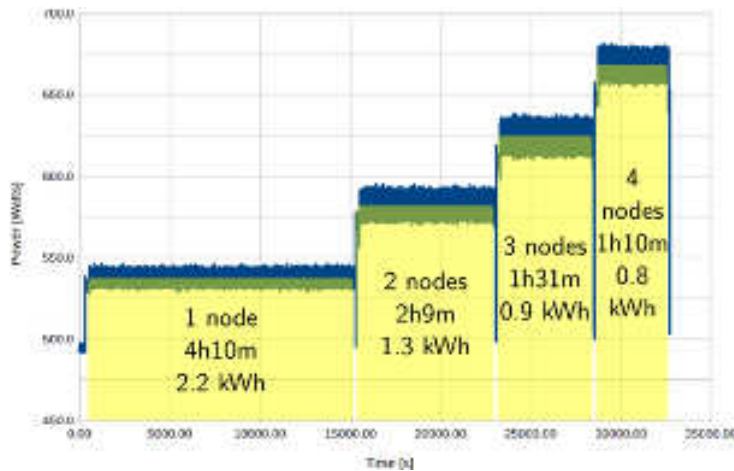


Cluster

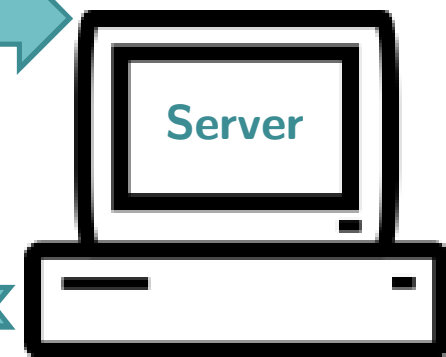
System power plug



- Not “platform specific”
 - Cavium ThunderX
- Full node measurements
 - Including PSU losses



Serial Interface
3 sample/sec



Server



Linux Kernel support for the ThunderX PMU

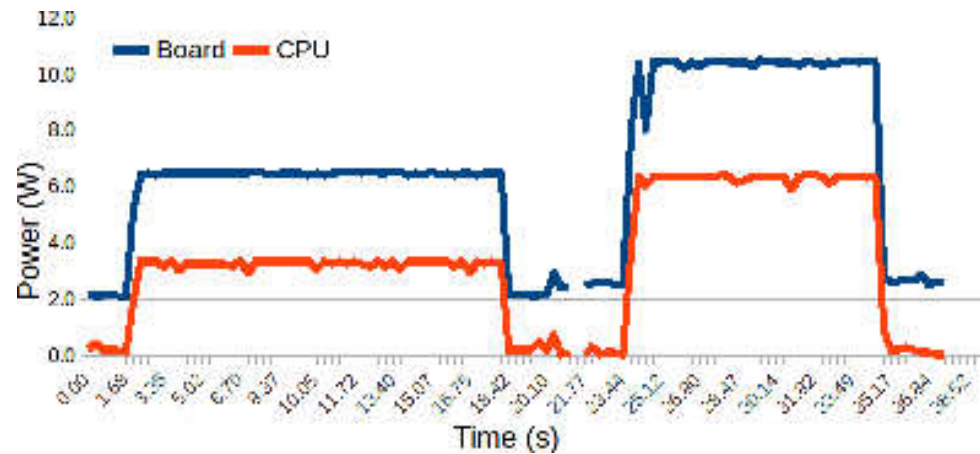
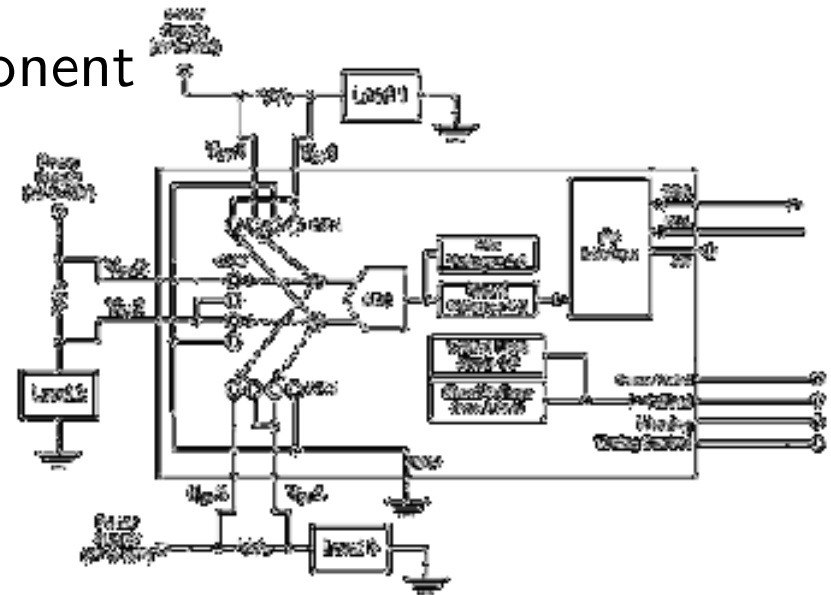
INFO: On ARM CPU, hardware performance counters are handled by a component called Performance Monitor Unit (PMU), part of the SoC:

- Access to PMU is SoC-dependent
- In order to access the PMU
 - Device Tree must include PMU definition
 - Linux Kernel must implement a way to access PMU
- **In our case**
 - Device Tree does not include PMU definition
 - PMU access not supported by the kernel
- **Result**
 - Hardware counters can be accessed with the “perf” command
 - Patch available to the rest of the world

Developed the PAPI extension for accessing hardware counters via PAPI (including preset and native events)

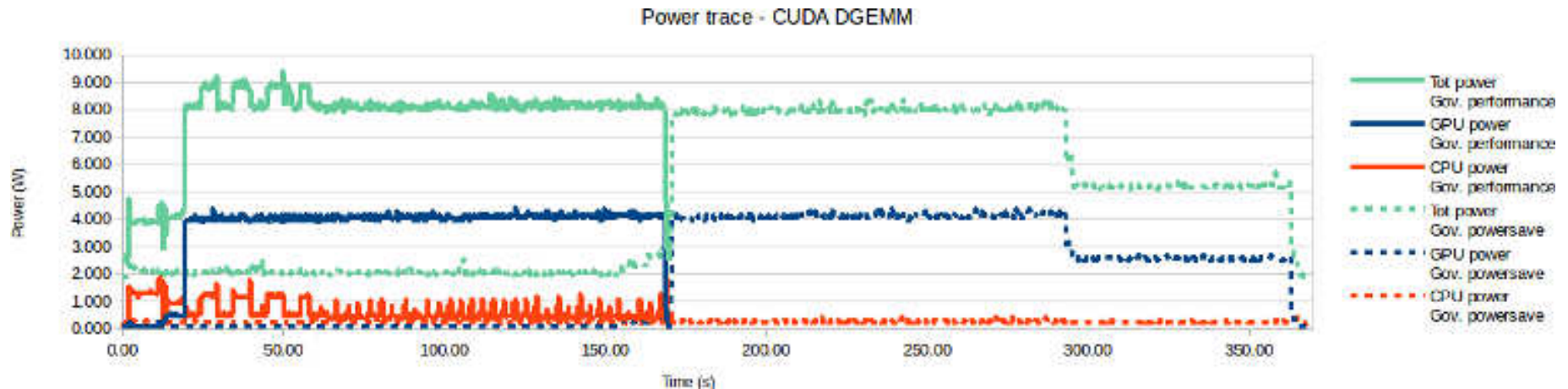
Jetson TX1: “old school” hacking...

- Voltage monitor on-board component
 - Texas Instruments INA3221
 - Connected via I2C
 - No support provided by NVIDIA
 - Hand-written support...
- Measurements validated with external setup
- So we are now able to get power traces on Jetson TX1!

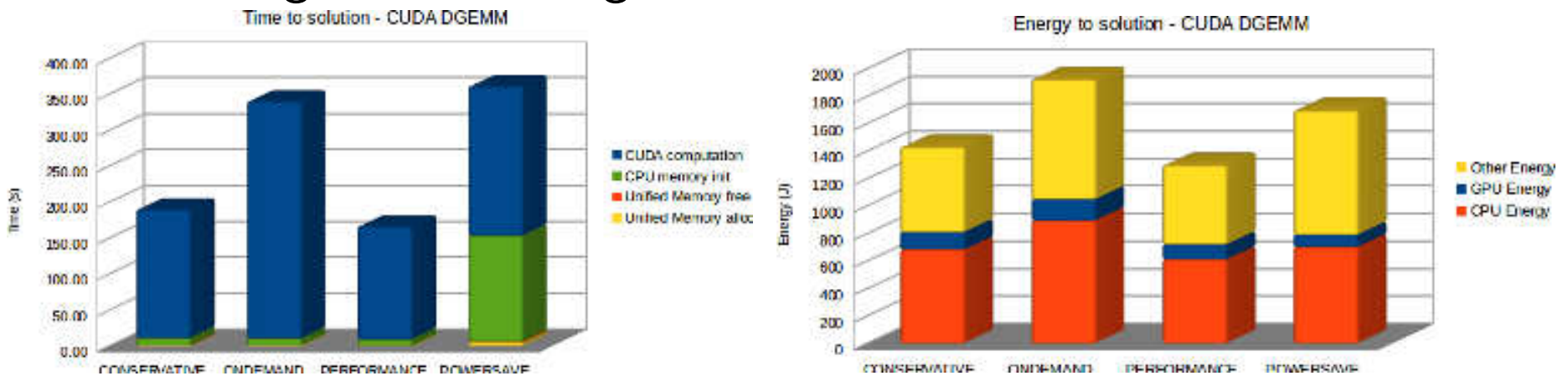


DGEMM with CUDA

- Test of a simple CUDA code performing a DGEMM

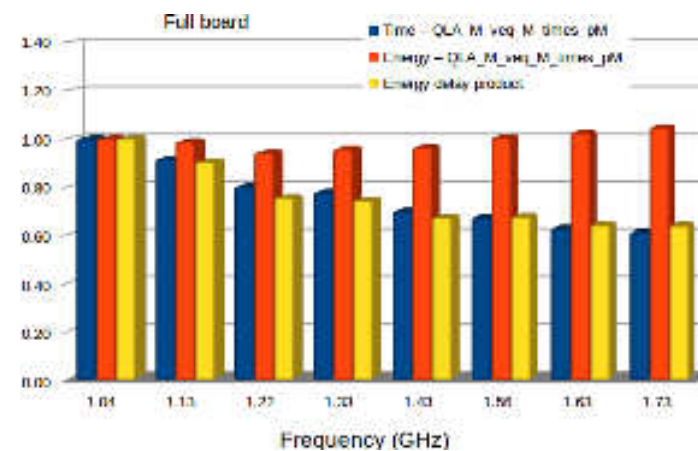
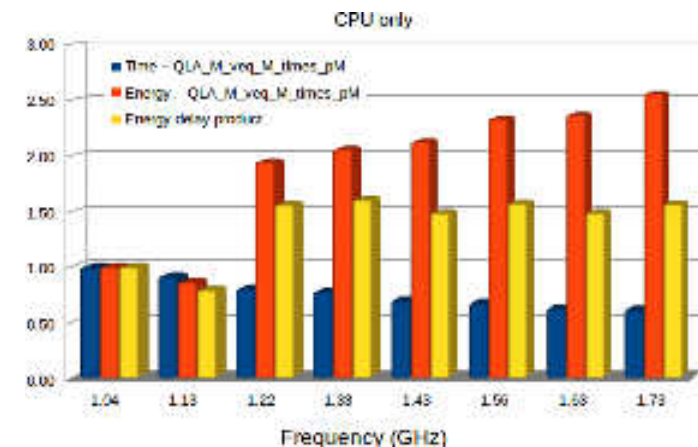
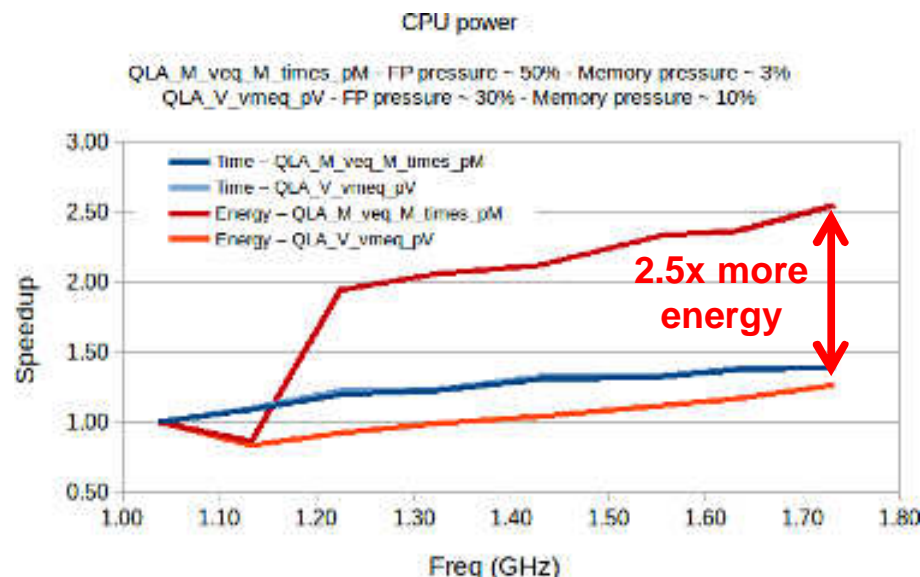


- Basic time to solution / energy to solution study with default governor configurations



MILCmk power analysis (preliminary)

- 7 representative microkernels for the MIMD Lattice Computation (MILC) collaboration code
- CORAL benchmark, C code, OpenMP, double precision
- Study of the computational features using hw counters



Next steps

Short term:

- Power profile complex CUDA codes
- Deeper understanding of governors

Ideally targeting three levels of power information:

- From the application
 - Access to an energy register, PAPI style
 - Possibility of easily powering on-off cores
 - Change of frequency
- From the runtime
 - Direct access to the power registers
 - Possibility of easily powering on-off cores (without kernel support)
- From the outside
 - Gather power data of larger systems “a la Mont-Blanc”
 - Power aware job scheduling

Conclusions

- Highlight of Mont-Blanc activities have been presented
 - Even with low-end hardware components it is possible to achieve decent performance in parallel computation
 - Main-line of Mont-Blanc 3 activity is targeting high-end server market
 - Still researching in cost-efficient platforms
- 3 ARM-based platforms for scientific computing have been introduced
 - With focus on power monitoring
 - There is still a long way for real power aware programming...

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

Niki Lauda



montblanc-project.eu



MontBlancEU



@MontBlanc_EU