montblanc-project.eu | @MontBlanc_EU

Butterfly effect of porting scientific applications to ARM-based platforms

Filippo Mantovani

June 22nd, 2017

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement n° 671697

Mont-Blanc projects in a glance

Vision: to leverage the fast growing market of mobile technology for scientific computation, HPC and data centers.

2012 2013 2014 2015 2016 2017 2018

Mont-Blanc

- **Hardware**
  - Mobile technology (ARMv7)
- **Software**
  - Enabling ARM based clusters
- **Next generation studies**
  - Gathering info from first prototypes and extrapolate

Mont-Blanc 2

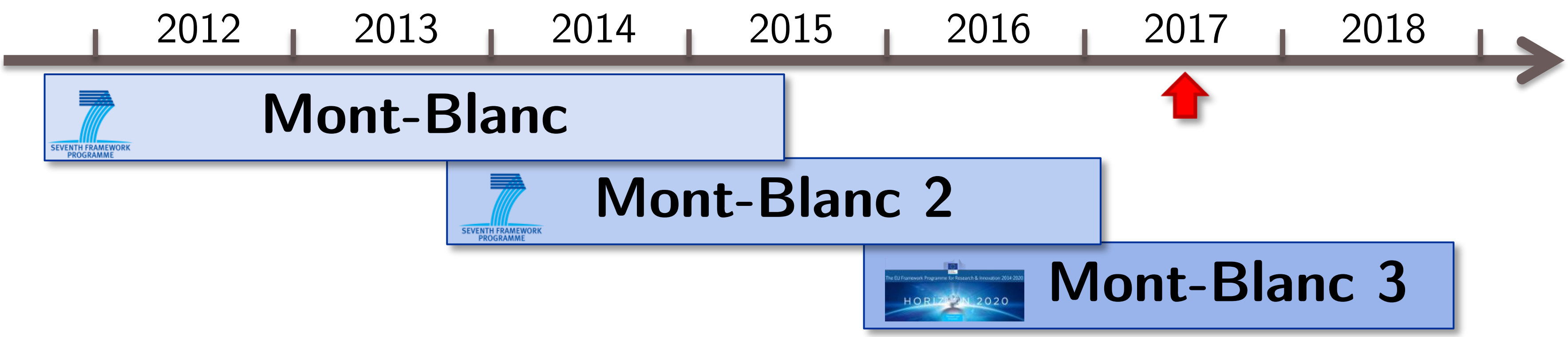
- **Hardware**
  - Mobile and server technology (ARMv8)
- **Software**
  - Massive improvement of the ecosystem (development tools, resiliency)
- **Next generation studies**
  - Pre-exascale HPC compute node

Mont-Blanc 3

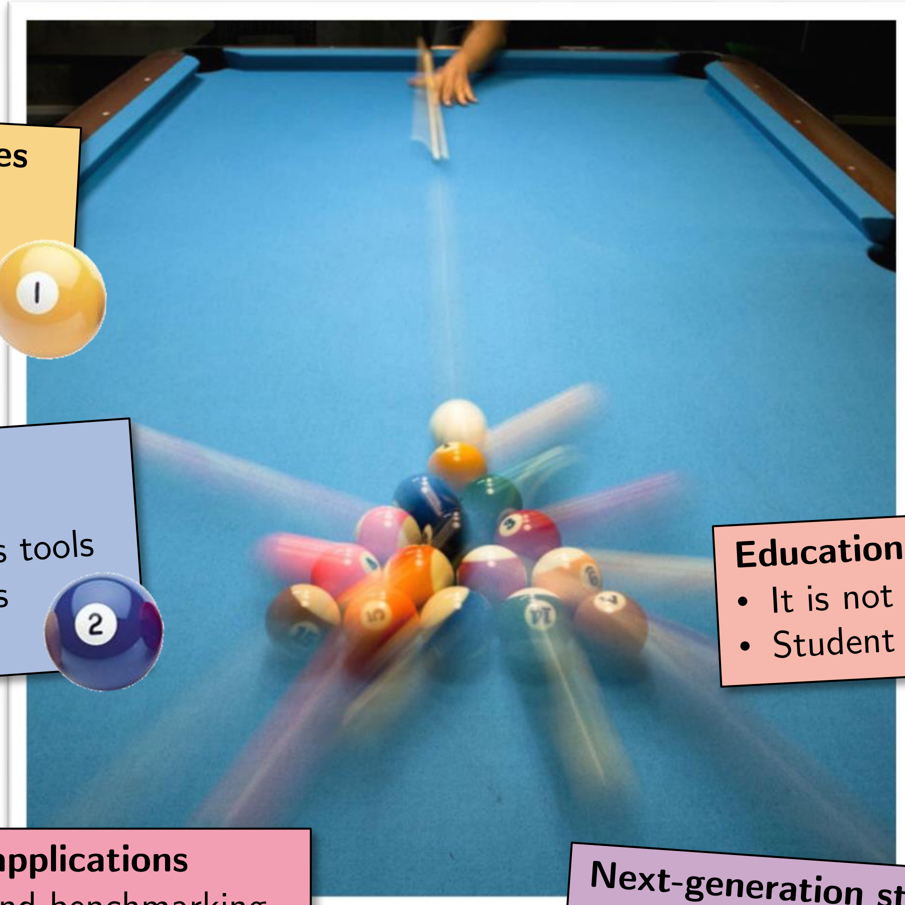
- **Hardware**
  - Server technology (ARMv8)
- **Software**
  - Focusing on study of mini-apps and real industrial applications
- **Next generation studies**
  - Focusing on an HPC SoC
  - Get a market-ready system!

Mont-Blanc projects in a glance

Vision: to leverage the fast growing market of mobile technology for scientific computation, HPC and data centers.



Mont-Blanc contributions



ARM-based prototypes
- Mobile technology
- Server technology
- System integration

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

Next-generation studies
- Performance projections
- From SoC to full system simulation: Multi-scale simulation infrastructure

Educational challenge
- It is not only a business problem…
- Student Cluster Competition

The Mont-Blanc prototype ecosystem

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

2012 – PRACE prototypes
- Tibidabo
- Carma
- Pedraforca

2013 – Mini-clusters
- Arndale
- Odroid XU
- Odroid XU-3
- NVIDIA Jetson

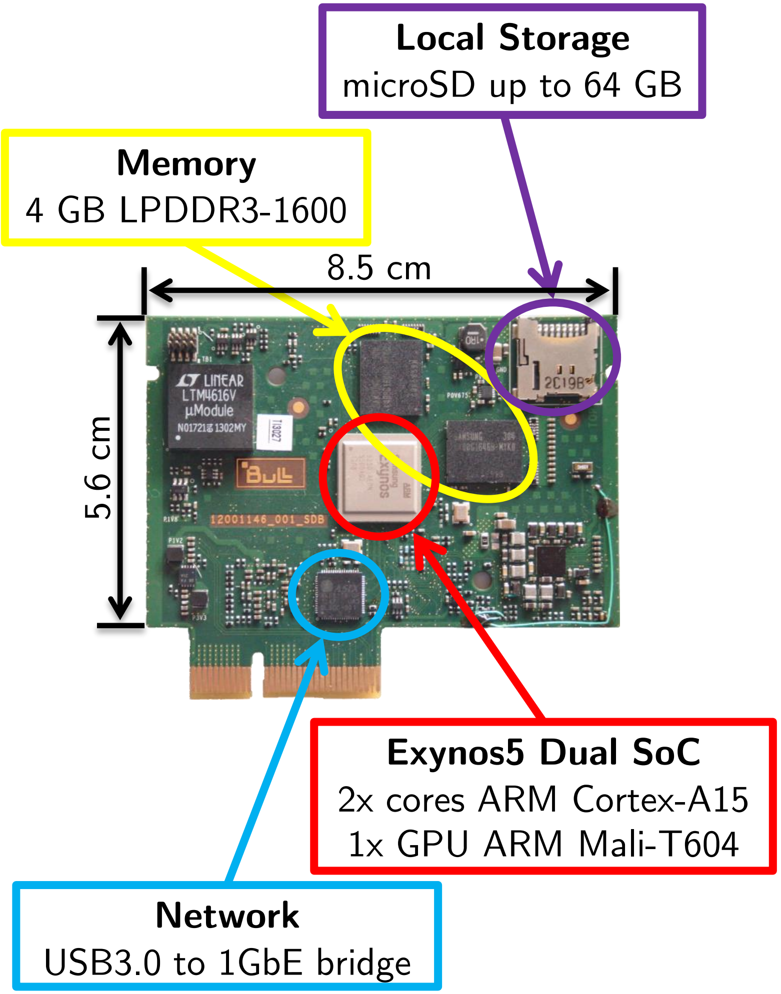
2014 – Mont-Blanc prototype
- 1080 compute cards
- Fine grained power monitoring system
- Installed between Jan and May 2015
- Operational since May 2015 @ BSC

2015 – ARM 64-bit mini-clusters
- APM X-GENE2
- Cavium ThunderX
- NVIDIA TX1

2016/2017 – Mont-Blanc 3 demonstrator
- Based on new-generation ARM 64-bit processors Cavium ThunderX2 SoC
- Targeting HPC market

2012 2013 2014 2015 2016 2017

The first 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

The Mont-Blanc 3 demonstrator

➔ **Codename: "Dibona"**

Power supply units

48 compute nodes
96 Cavium ThunderX2 CPUs
3000 cores
12000 threads

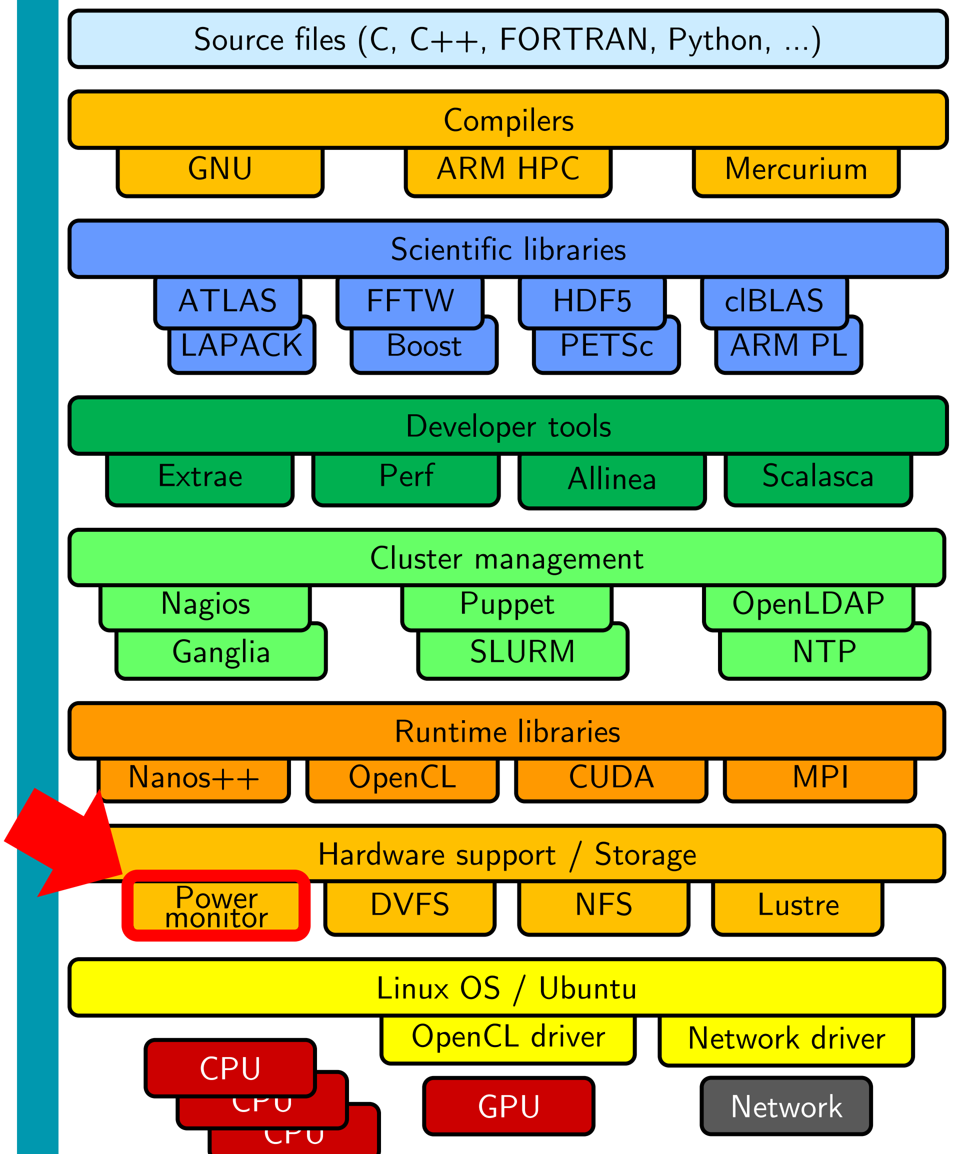
Direct Liquid Cooling

Redundant management server and storage

Infiniband EDR switches

Management network

System software stack for ARM



1. Developed since 2011! Today in collaboration with all major OpenHPC partners
2. Tested on several ARM-based platform
3. Mostly based on open-source packages
4. Effort in standardize power measurements formats for efficient use of existing systems

MONT-BLANC

Scientific applications: methodology

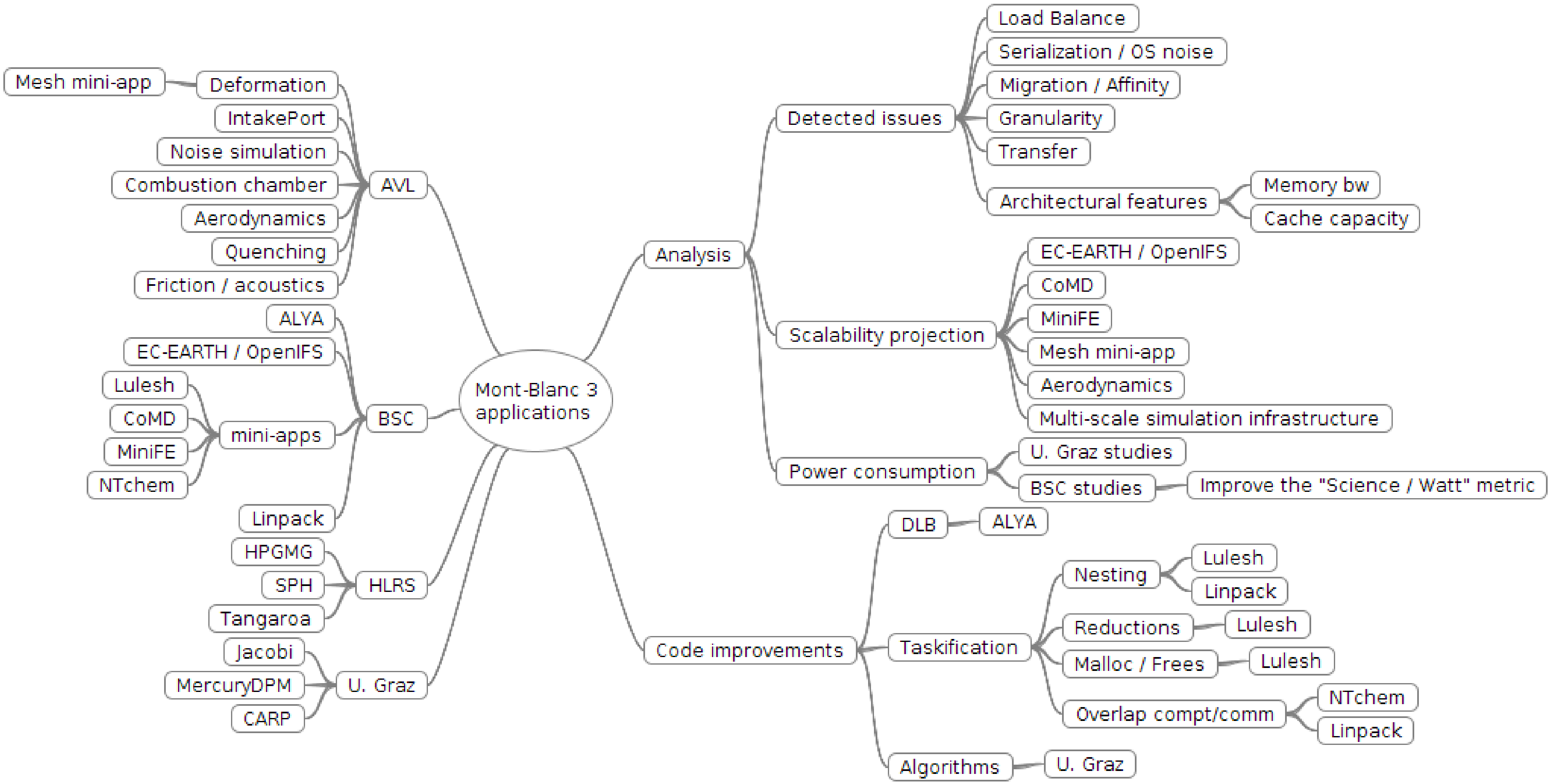
Applications

- Benchmarks
- Mini-apps
- Production / Industrial codes

Tracing applications with the objective of...

- Fixing features, aka limitations, of current systems implementation
  - e.g. memory affinity on Cavium ThunderX
- Applying OmpSs/OpenMP4.0 and analyze the effect
  - Benefit of taskification
  - Exploring new techniques, e.g. Dynamic Load Balancing
- Understanding code limitations and helping the developers in restructuring it
- Performing extrapolation studies using next generation machine parameters
  - MUSA

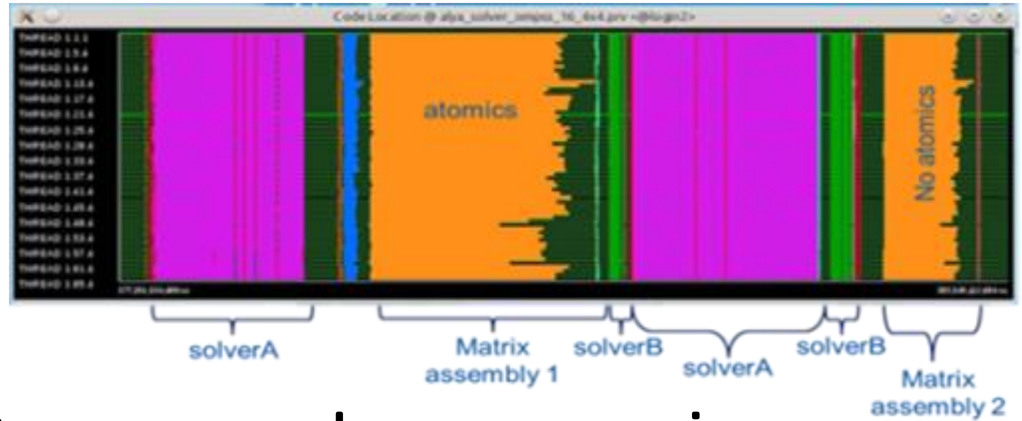
Scientific applications: methodology



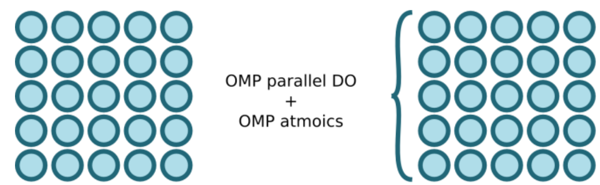
Alya: BSC code for multi-physics problems

Parallelization of finite elements code

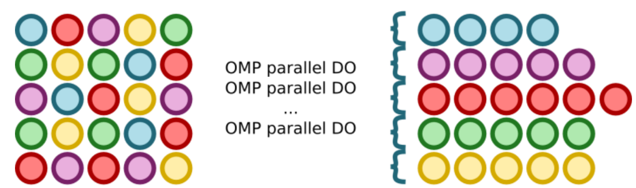
- **Analysis with Paraver:**



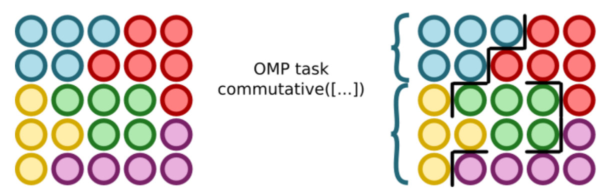
- **Reductions with indirect accesses on large arrays using**
  - No coloring



  - Coloring

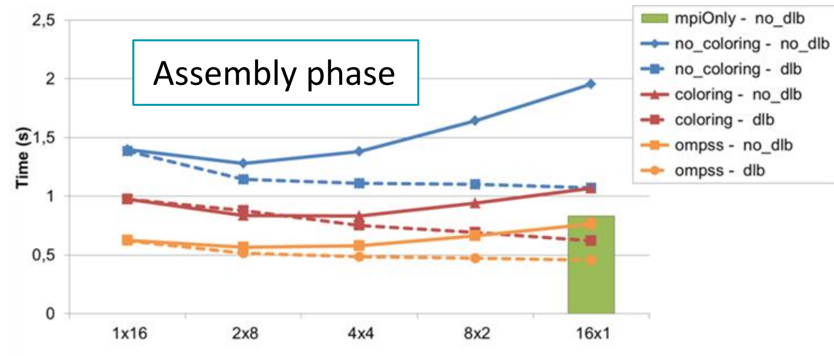


  - Commutative Multidependences (OmpSs feature to be hopefully included in OpenMP)



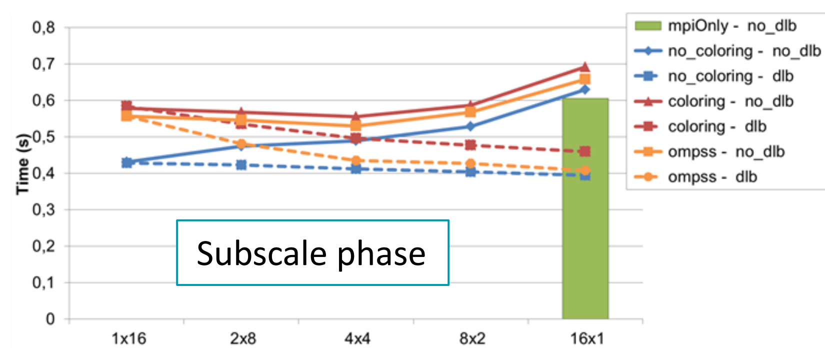
Alya: taskification and dynamic load balancing

- **Towards throughput computing**
  - Tasks + DLB → dotted lines



- **DLB helps in all cases**
  - Even more in the bad ones

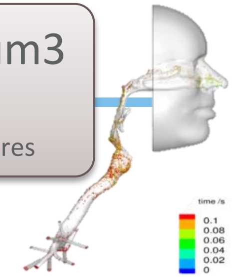
16 nodes x P processes/node x T threads/process



- **Side effects**
  - Hybrid MPI+OmpSs Nx1 can perform better than pure MPI!
  - Nx1 + DLB: hope for lazy programmers

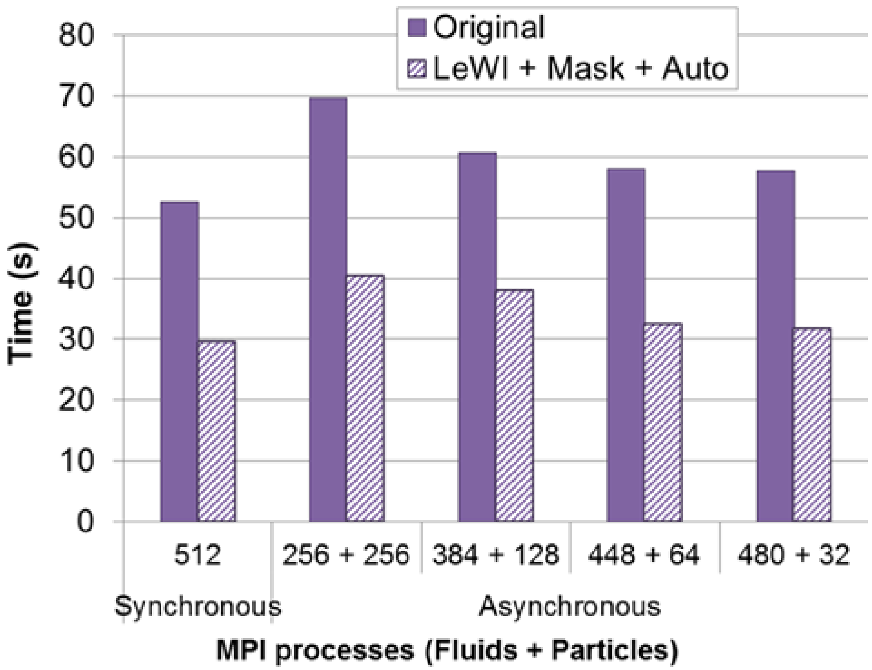
Alya: coupled codes

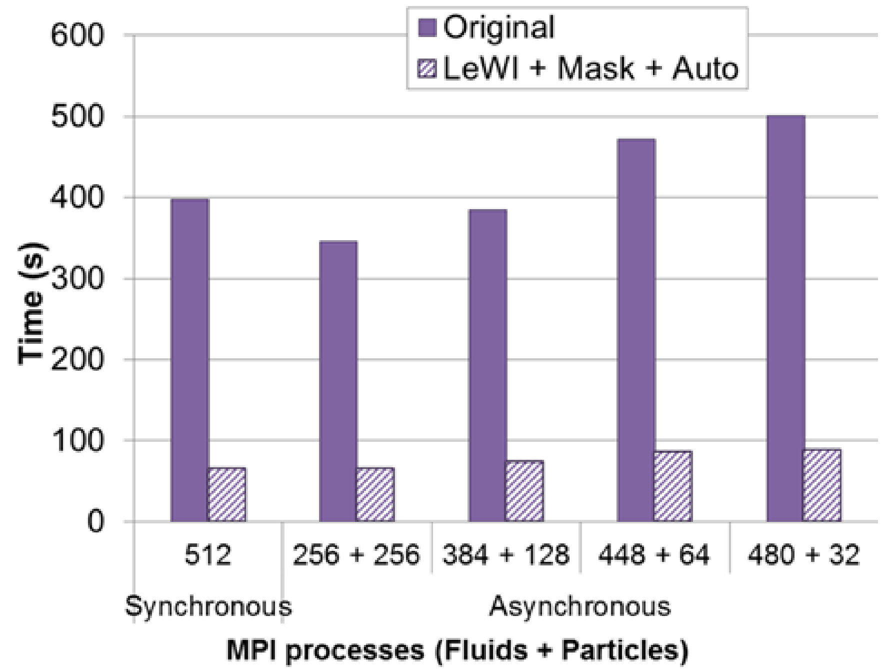
- Marenostrum3
  - 32 nodes
    - 512 cores



➔ Observations

- Big impact configuration and kind of coupling in original version
- Important improvement with DLB-Lend-When-Idle in all cases
- Almost constant performance independent of configuration and kind of coupling





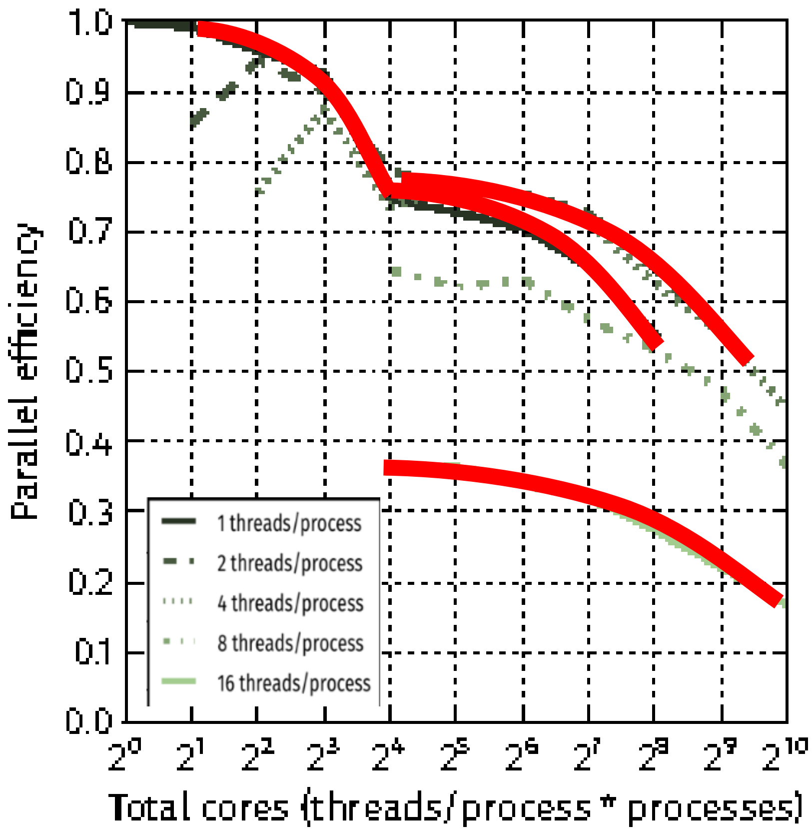
Fluid dominated

Fluid
Particle

Particle dominated

Credits: J. Labarta, M. Garcia

NTChem: hybrid MPI + OpenMP

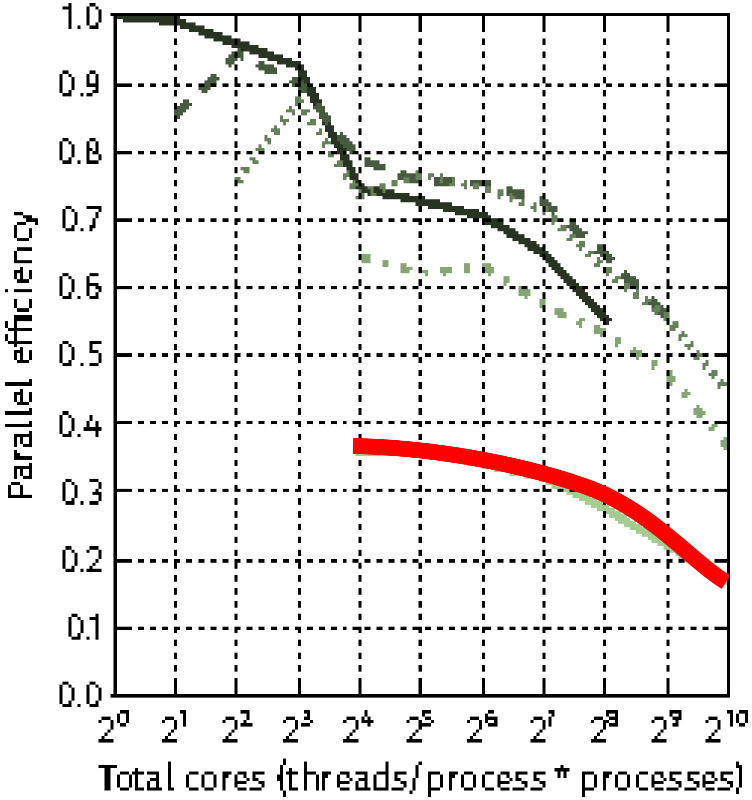


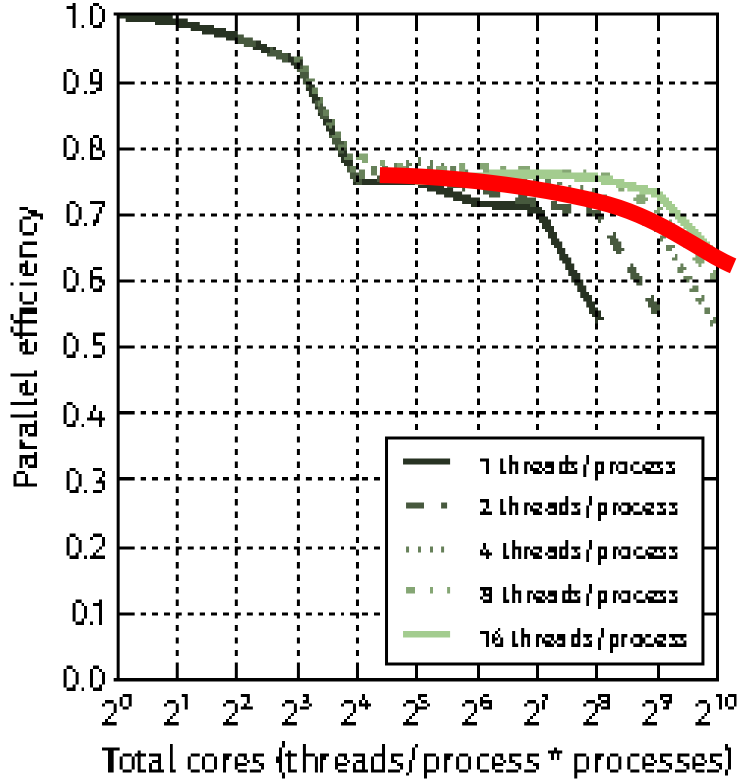
- **As the node is not fully populated, the system activates TurboBoost**
  - Increasing frequency from 2.6 to 3.1GHz
- **Load imbalance**
  - Global serialization
  - Noise
- **Some gain when using low number of threads**
- **High overhead, fine granularity at large thread count**

Data from MareNostrum3

NTChem: OmpSs taskification

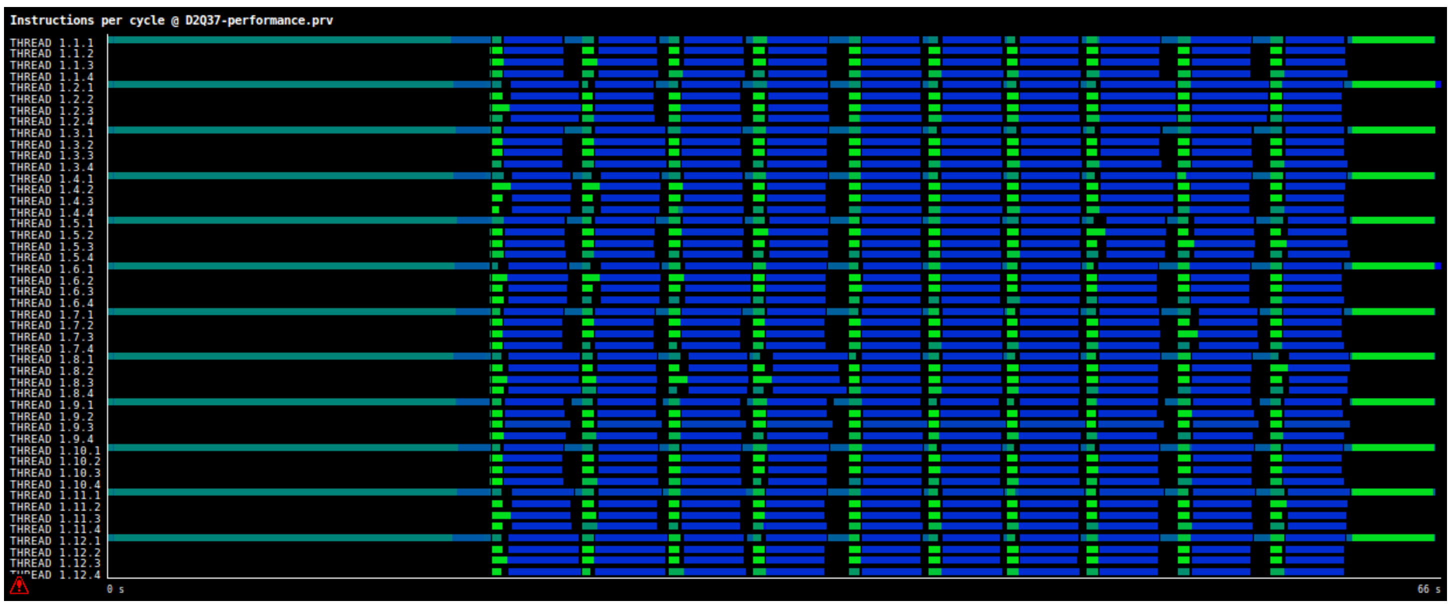
- **Simpler code, better performance!**
  - Sufficient task granularity
  - Communication computation overlap
  - DLB needed at very large scale (TBC)

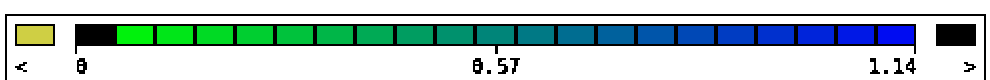




Lattice Boltzmann D2Q37

- Fluid dynamic code MPI+OpenMP for simulation of e.g. mixing layer evolution of fluids ad different temperature/density
- Simple structure
  - Serial initialization + closing
  - Propagate (memory bound)
  - Collide (compute bound)

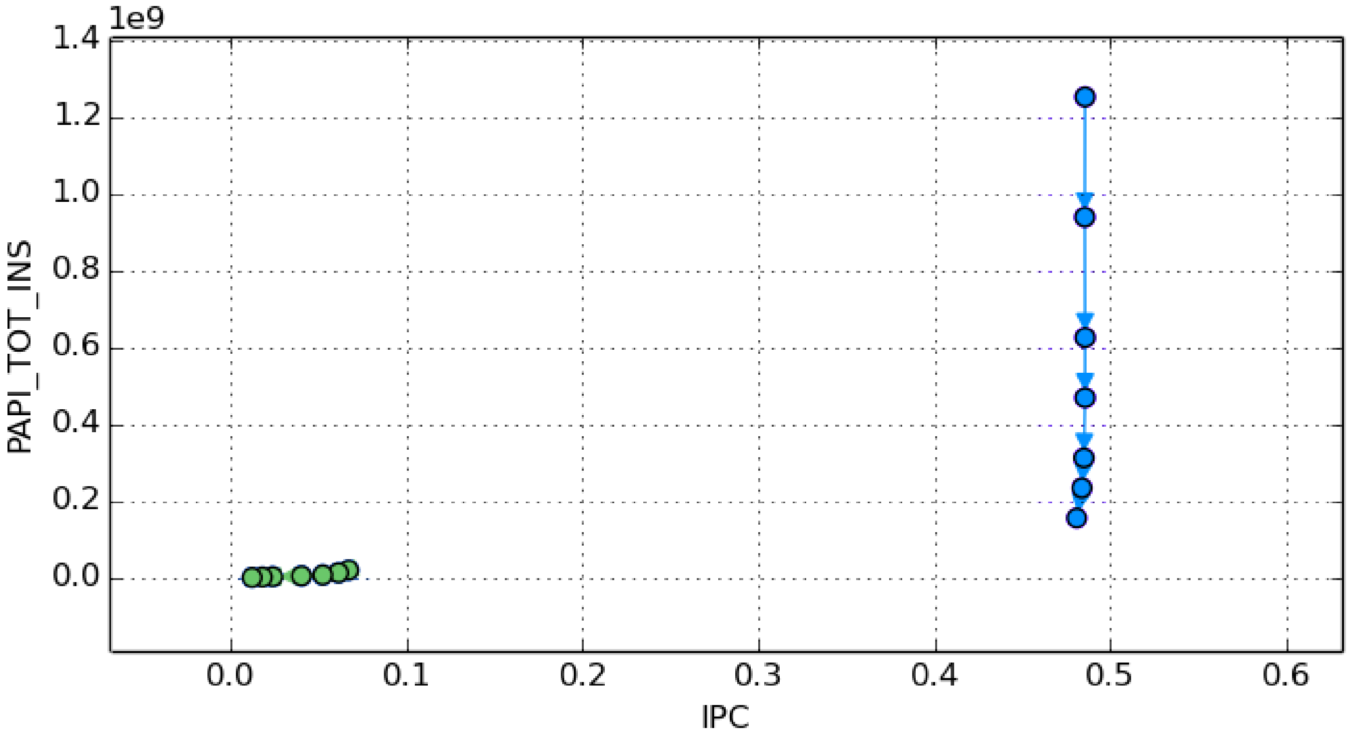




Credits: E. Calore

D2Q37 Clustering analysis (on ThunderX)

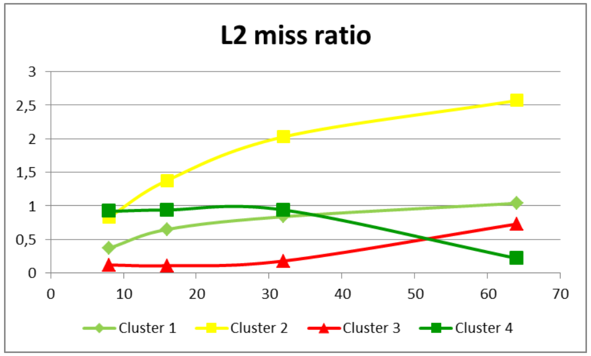
- **The different runs were performed over the same lattice size with a varying number of threads: 6, 8, 12, 16, 24, 32, 48.**
  - Collide function is scaling almost perfectly up to 48 threads
  - For the Propagate, increasing the number of threads makes threads competing for the same resource (memory)

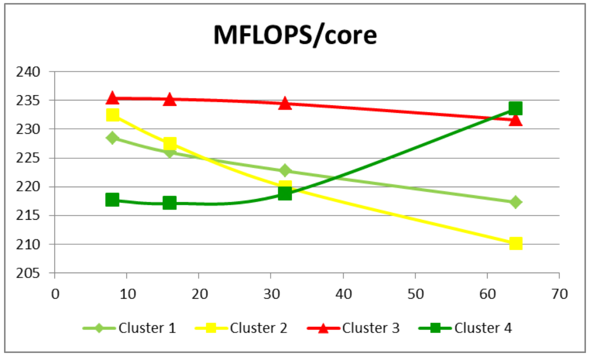


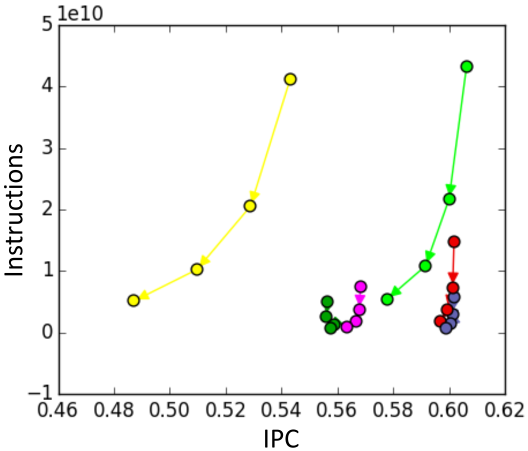
Clustering of OpenIFS in ThunderX

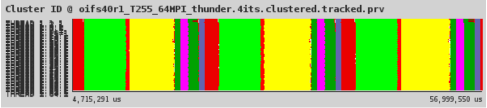
Strong scaling tracking of IPC

- Shared L2 effect







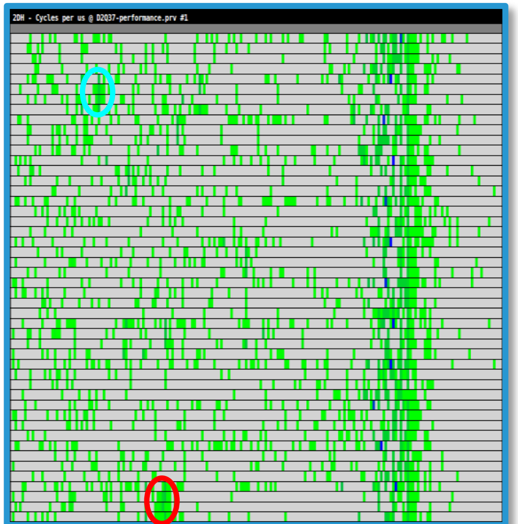


- Core Cost Model
  - Ideal CPI (1.6) , L2 hit cost (62.2) , L2 miss cost (129,4)
  - "Reverse" engineering form real production code runs

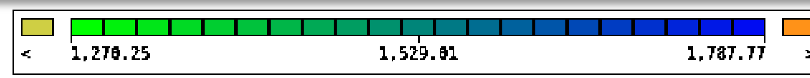
- Socket IPC: ~ 26 (!!!)

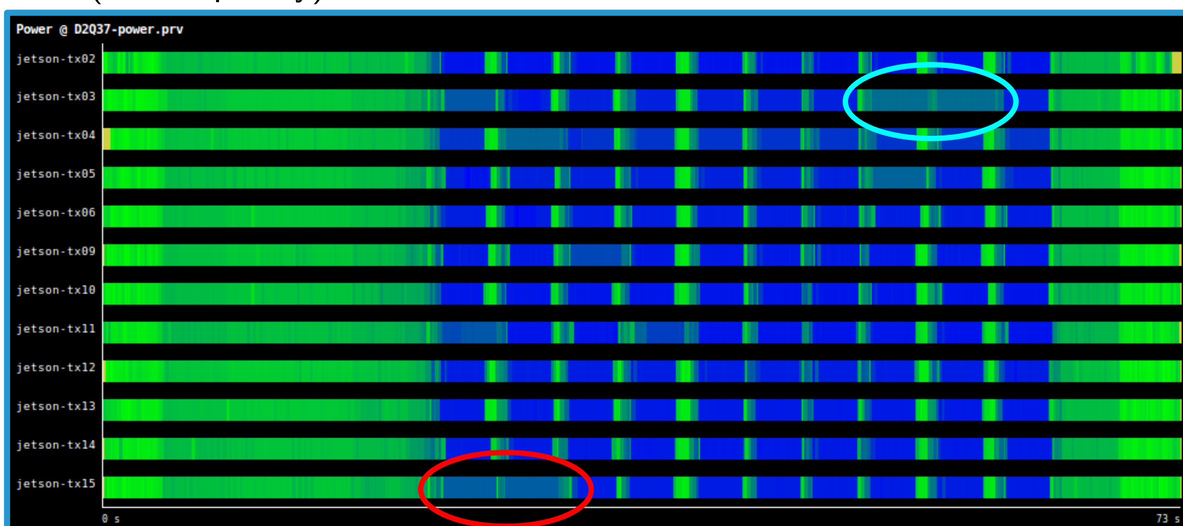
MONT-BLANC

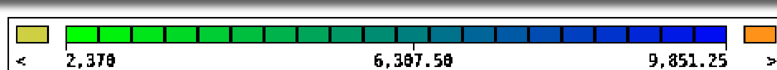
D2Q37 Correlating performance, power, energy



Histogram of cycles per us (i.e. frequency)







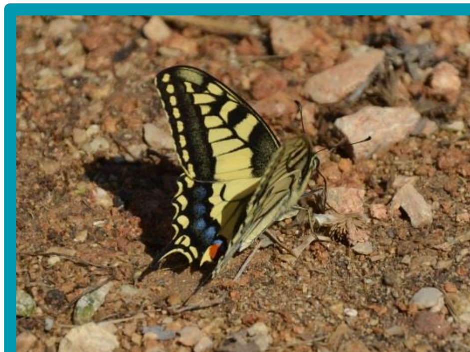
Energy to solution @ D2Q37-powe

	[2,370.00..10,638.75]
jetson-tx02	469,427.16
jetson-tx03	439,638.69
jetson-tx04	430,995.02
jetson-tx05	453,455.81
jetson-tx06	458,372.55
jetson-tx09	457,328.90
jetson-tx10	469,052.97
jetson-tx11	469,194.46
jetson-tx12	456,111.53
jetson-tx13	459,625.07
jetson-tx14	469,675.26
jetson-tx15	450,010.75
Total	5,482,888.18
Average	456,907.35
Maximum	469,675.26
Minimum	430,995.02
StDev	11,753.48
Avg/Max	0.97

Credits: Enrico Calore

Conclusions / Vision

- **Most hardware limitations will evolve, eventually**
  - In the original market of the devices
  - When extending to the server market
  - Pushed by other markets (e.g. automotive)
- **Programming model and runtime will help "overcome"**
  - Asynchrony and overlap
  - Resilience
  - Variability / Load balancing
- **Tools can help understand the real problems and suggest/evaluate alternatives**
  - e.g. correlating performance and power
- **The work performed on Mont-Blanc 3 applications is not strictly ARM specific**
  - Benefits for any new platform!!!



Student Cluster Competition

➔ Rules

- 12 teams of 6 undergraduate students
- 1 cluster operating within 3 kW power budget
- 3 HPC applications + 2 benchmarks

➔ One team from University Politecnica of Catalunya (UPC-Spain)

- Participating with Mont-Blanc technology

➔ 3 awards to win

- Best HPL
- 1st, 2nd, 3rd overall places
- Fan favorite

For more information

montblanc-project.eu

@MontBlanc_EU

filippo.mantovani@bsc.es