



 [montblanc-project.eu](https://montblanc-project.eu) | @MontBlanc\_EU

# Energy efficiency @ Mont-Blanc

**Etienne Walter, Bull**  
with contributions from the whole team  
special thanks to Roxana Rositoru, Arm

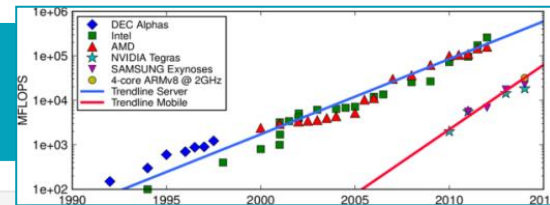
ETP4HPC workshop, Ljubljana



# Mont-Blanc – Origin & context

# The Mont-Blanc pitch

**Vision:** leverage the fast growing market of mobile technology for scientific computation



Design SoC  
Develop IPs

*EPI*

**Mont-Blanc 2020**

Prepare industrial solution  
Test market acceptance

**Mont-Blanc 3**

Extend the concept  
and explore new  
possibilities

**Mont-Blanc 2**

Proof of concept : HPC  
computing based  
on mobile embedded  
technology

**Mont-Blanc**

2012

2013

2014

2015<sup>3</sup>

2016

2017

2018

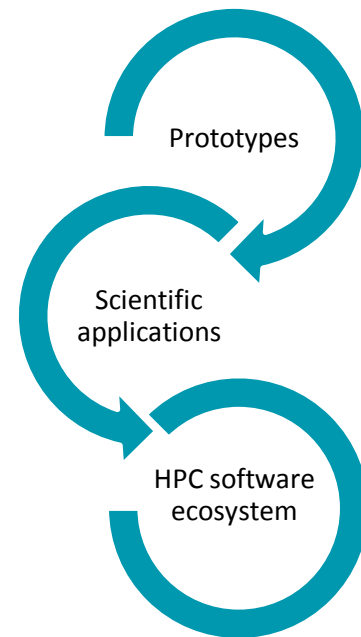
2019

2020

# Mont-Blanc 3 key objectives

- **Design a compute node based on ARM architecture for a pre-exascale system**
  - Well balanced : Memory, Interconnect, IO
  - Use of simulation to evaluate the options on applications
  - Energy efficient
- **Evaluate new high-end ARM core and accelerator, and assess different options for compute efficiency**
  - Heterogeneous cores, vectorization, high performance core
  - Assessment with existing solutions & applications
- **Develop the software ecosystem needed for market acceptance of ARM solutions**

Key ideas: **Well-balanced architecture, (Energy & Compute) Efficiency, Throughput computing, Co-design, SoC/SoP design**



BSC  
Barcelona Supercomputing Center  
Centro Nacional de Supercomputación

CNRS

Bull  
Solutions

ARM

UNIVERSITÉ DE  
VERSAILLES  
ST-GERMAIN EN LAIS

HLRS

AVL

ETH Zürich

UNI  
DUISBURG  
ESSEN

UC  
CLERMONT

**MONT-BLANC**

# A look at last HPCG results (Nov. 17)

Rank	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Fraction of Peak
1	RIKEN Japan	<b>K computer</b> – , SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10.510	10	0.603	5.3%
2	NSCC / Guangzhou China	<b>Tianhe-2 (MilkyWay-2)</b> – TH-IVB-FEP Cluster, Intel Xeon 12C 2.2GHz, TH Express 2, Intel Xeon Phi 31S1P 57-core NUDT	3,120,000	33.863	2	0.580	1.1%
3	DOE/NNSA/LANL/SNL USA	<b>Trinity</b> – Cray XC40, Intel Xeon E5-2698 v3 300160C 2.3GHz, Aries Cray	979,072	14.137	7	0.546	1.8%
4	CSCS Switzerland	<b>Piz Daint</b> – Cray XC50, Intel Xeon E5-2690v3 12C 2.6GHz, Cray Aries, NVIDIA Tesla P100 16GB Cray	361,760	19.590	3	0.486	1.9%
5	NSC Wuxi China	<b>Sunway TaihuLight</b> – Sunway MPP, SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93.015	1	0.481	0.4%
6	Joint Center for Advanced High Performance Computing Japan	<b>Oakforest-PACS</b> – PRIMERGY CX600 M1, Intel Xeon Phi Processor 7250 68C 1.4GHz, Intel Omni-Path Architecture Fujitsu	557,056	13.555	9	0.385	1.5%
7	DOE/SC/LBNL/NERSC USA	<b>Cori</b> – XC40, Intel Xeon Phi 7250 68C 1.4GHz, Cray Aries Cray	632,400	13.832	8	0.355	1.3%
8	DOE/NNSA/LLNL USA	<b>Sequoia</b> – IBM BlueGene/Q, PowerPC A2 1.6 GHz 16-core, 5D Torus IBM	1,572,864	17.173	6	0.330	1.6%
9	DOE/SC/Oak Ridge NL USA	<b>Titan</b> – Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray	560,640	17.590	5	0.322	1.2%
10	GSIC Center, Tokyo Japan	<b>TSUBAME3.0</b> – SGI ICE XA (HPE SGI 8600), IP139-SXM2, Intel Xeon E5-2680 v4 15120C 2.9GHz, Intel Omni-Path Architecture, NVIDIA TESLA P100 SXM2 with NVLink HPE	136,080	8.125	13	0.189	1.6%

Diversity of hw platforms

Only a limited fraction of peak capacity used



Applications

Sw Environment

Hw. platform

# Scientific applications: methodology

## → Applications

- Benchmarks
- Mini-apps
- Production / Industrial codes

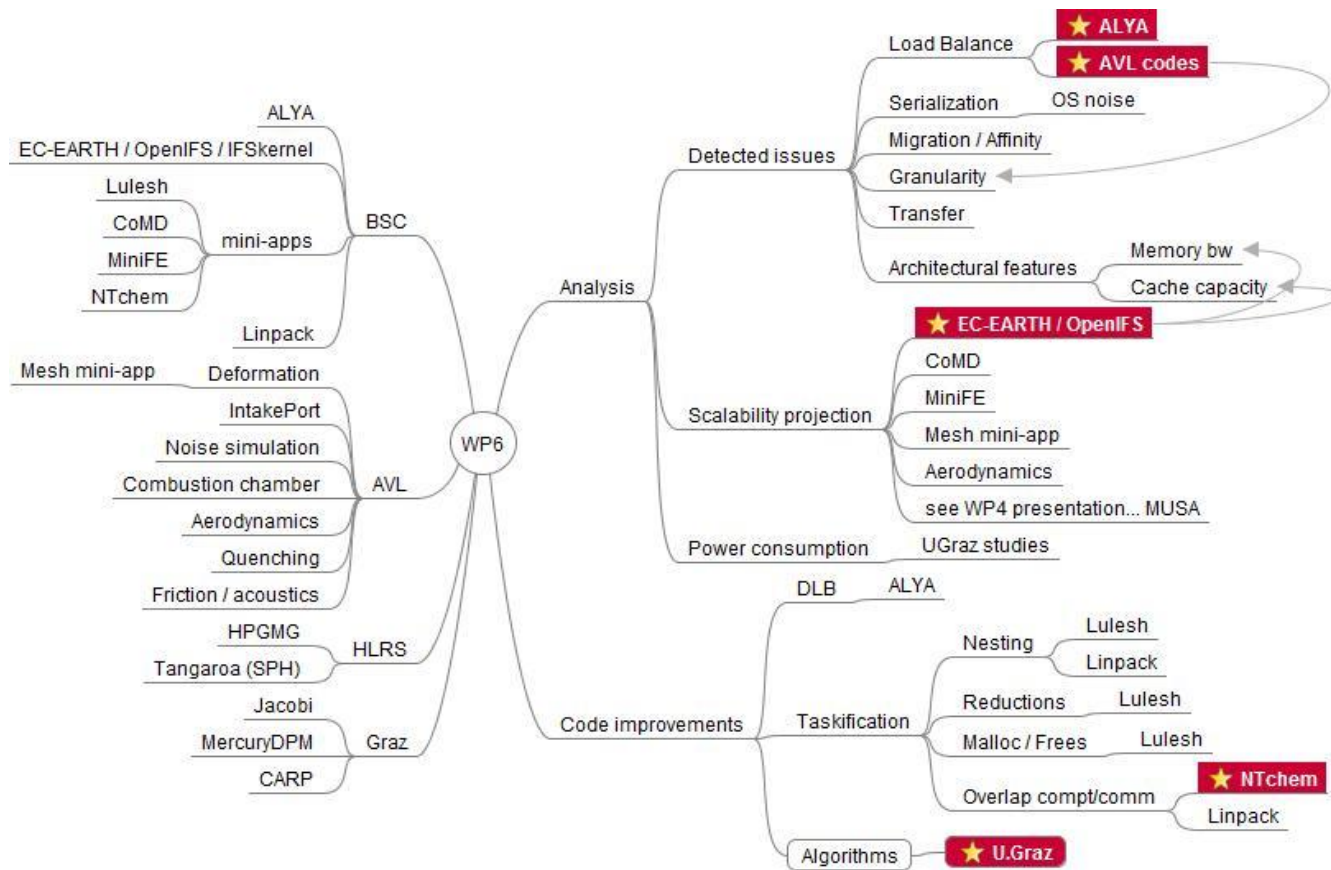
## → Tracing applications with the objective of...

- Test current solutions and provide feedbacks to technology providers
  - Test SoC, e.g. PAPI on Cavium ThunderX
  - Measure power consumption and correlate it with performance
  - Evaluation of HPC Compiler(s)
- Understanding code limitations and helping the developers in restructuring it
  - applying OmpSs/OpenMP4.0 and analyze the effect
  - Benefit of taskification
  - Exploring new techniques, e.g. Dynamic Load Balancing
- Have insights to perform extrapolation studies using next generation machine parameters

Tech report: <http://upcommons.upc.edu/handle/2117/107063>

Poster accepted at SC'17

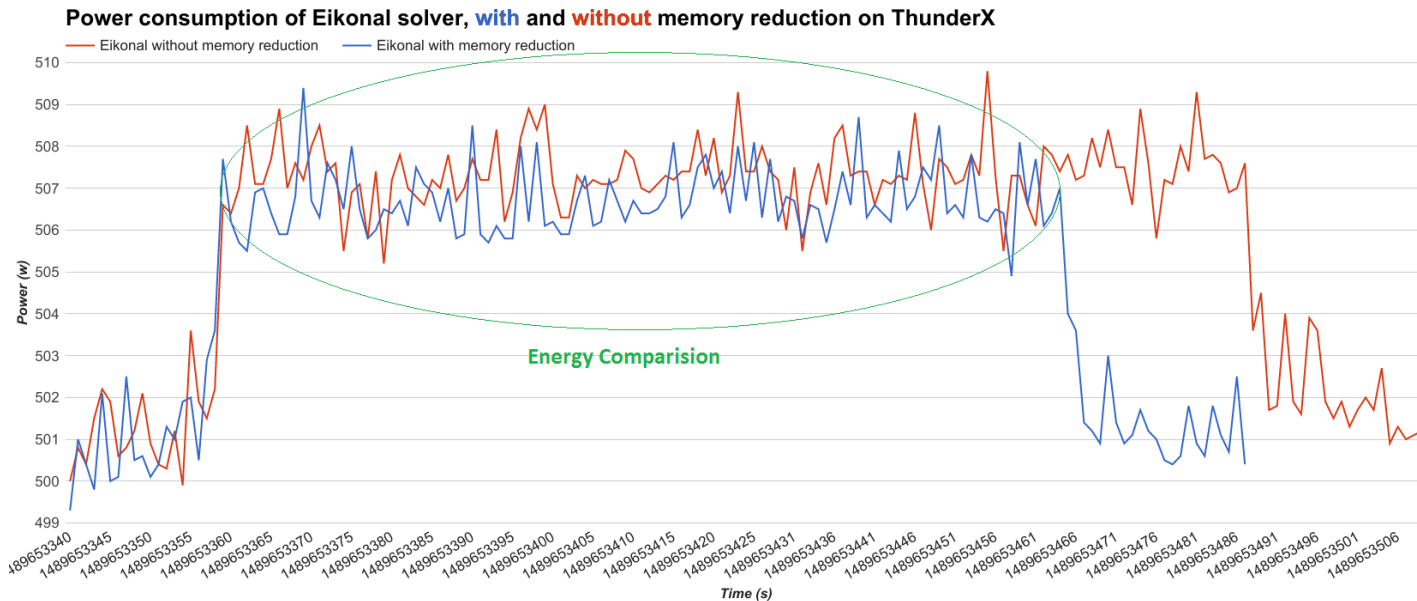
# Applications





## → 13% Energy Efficiency achieved by reducing memory footprint

- Eikonal solver on the ThunderX cluster 20 cores @ 1.8GHz
- Physical measurements: Yokogawa Power Meter - Monitoring for overall cluster.
- **Time-to-solution improvement: 22%**



# Code/algorithms improvements (Alya – BSC)

## → Parallelize Finite Element codes

- Interaction Load balance and IPC
- Reductions with indirect accesses on large arrays

## → Taskification

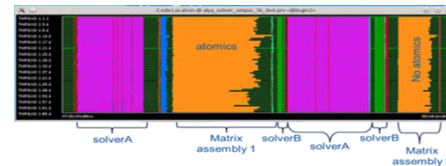
- Towards throughput computing
- Dynamic Load Balancing helps in all cases

## → Coupling codes

- Almost constant performance independent of configuration and kind of coupling

## → Scaling

- Scaling up to 16k cores
- Can manage fine granularities




# Application issues

- Taskification
- Data layout
- Unused hardware resources
- Vectorization
- Memory bandwidth
- Dynamic load balance
- & more...



Courtesy of <https://goo.gl/H56VpH>

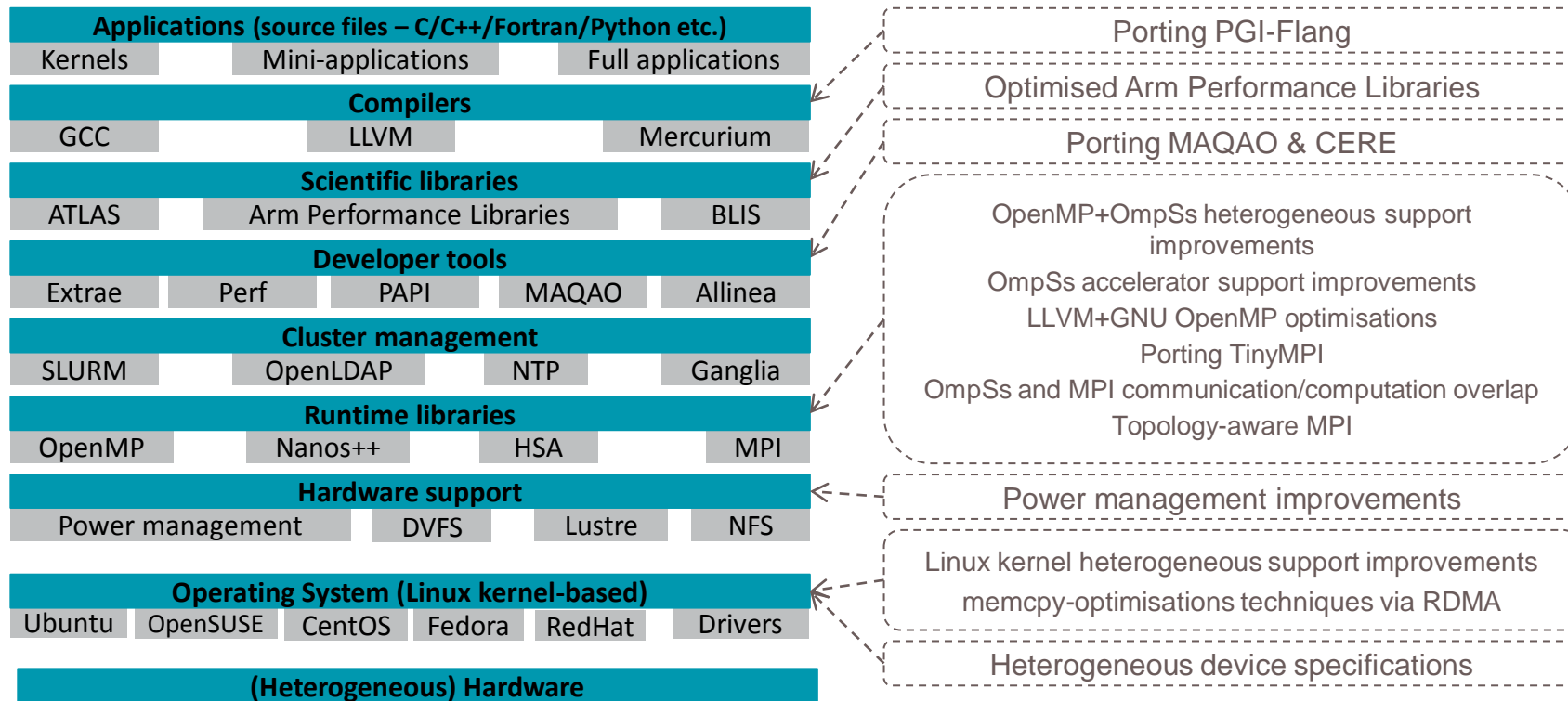


Applications

Sw Environment

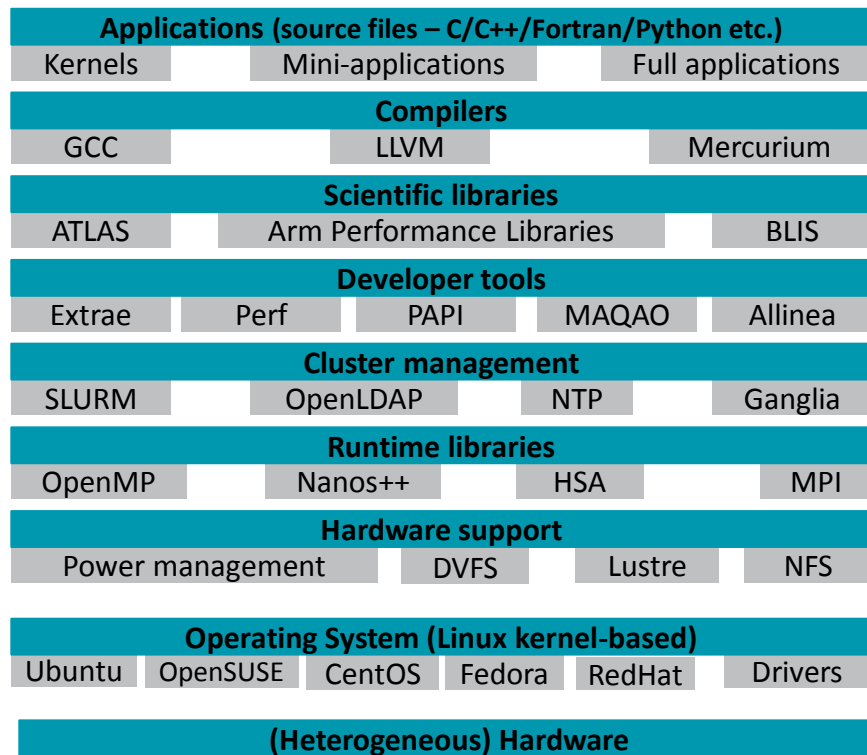
Hw. platform

# HPC Arm Software Stack



+ Contribution to OpenHPC (from 1.2)

# HPC Arm Software Stack



improvements mainly aiming  
(power) efficiency

Porting PGI-Flang

Optimised Arm Performance Libraries

Porting MAQAO & CERE

OpenMP+OmpSs heterogeneous support  
improvements

OmpSs accelerator support improvements

LLVM+GNU OpenMP optimisations

Porting TinyMPI

OmpSs and MPI communication/computation  
overlap

Topology-aware MPI

Power management improvements

Linux kernel heterogeneous support improvements

memcpy-optimisations techniques via RDMA

Heterogeneous device specifications

+ Contribution to OpenHPC (from 1.2)

**MONT-BLANC**

# ARM Performance Libraries

Enable the wide variety of ARM cores available today without adding complexity to the software ecosystem.

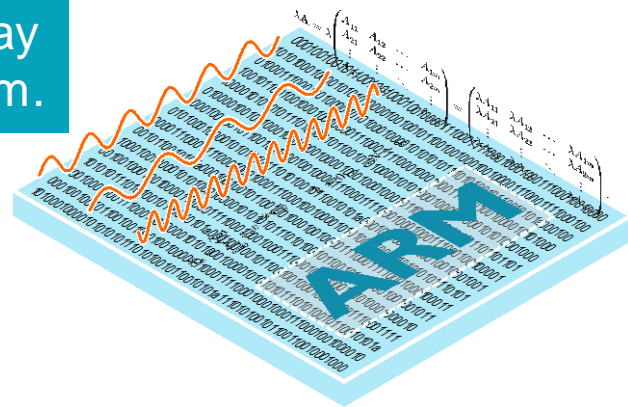
- **Commercially supported 64-bit ARMv8 vendor math libraries for scientific computing.**
- **Built and validated using technology from the Arm Numerical Algorithms Group (NAG).**
- **ARM silicon partners provide tuned kernels.**

## Capabilities:

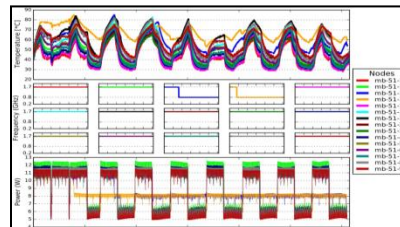
- **BLAS**
- **LAPACK**
- **FFT**

## Tuned for:

- Cortex-A53,A57,A72
- Applied Micro X-Gene®
- Cavium® ThunderX



## → Fine grained power monitoring



- 
- Effective useful RIPS in phase2000 - 2000 @ Alike-150MHz\_per\_merge-pr.v2
- iss - power
- MIPS calls @ Alike-150MHz\_per\_merge-pr.v2
- iss - useful instructions in

- 16





Applications

Sw Environment

Hw. platform

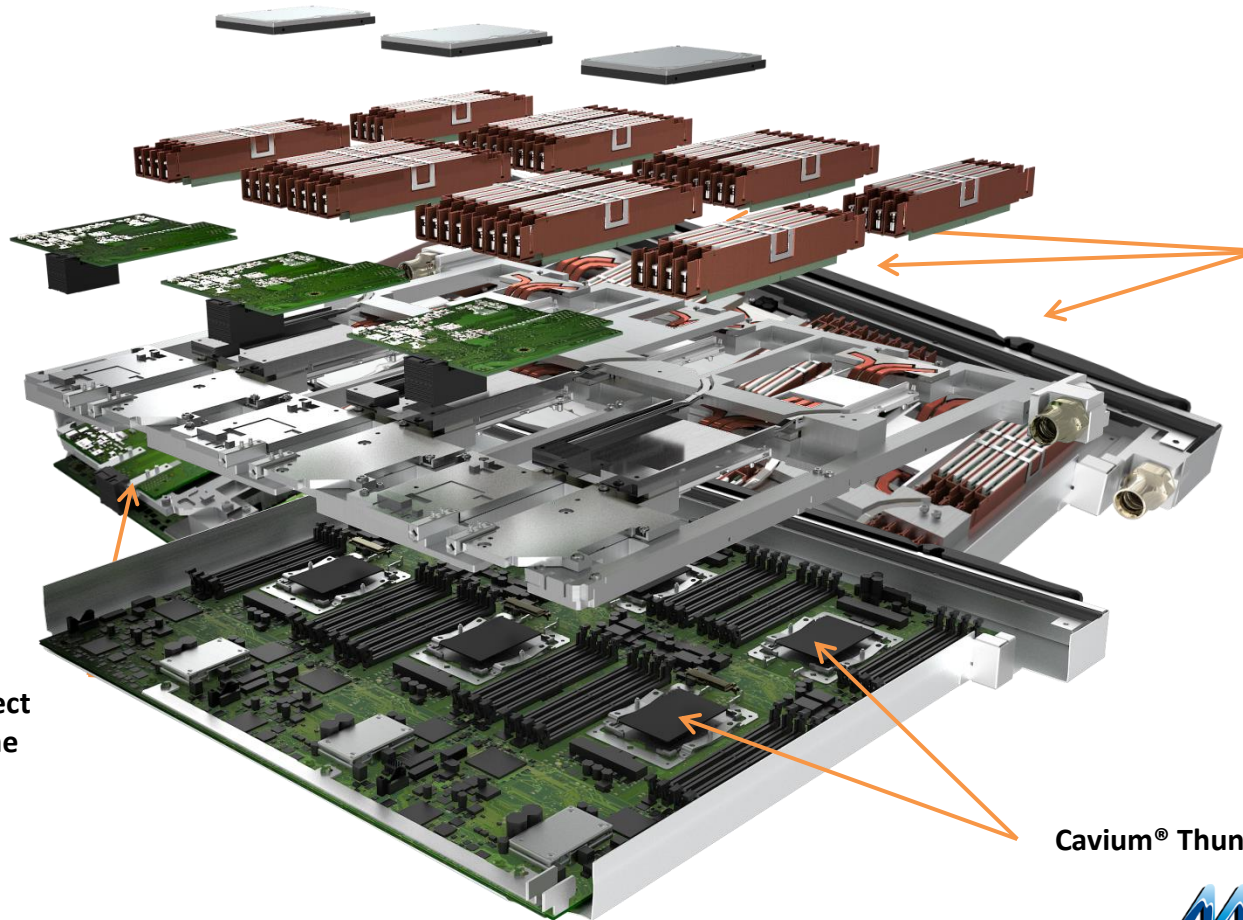
# Dibona: our new test platform



# Dibona (Mont-Blanc 3 test platform)



# Cavium ThunderX2 choice



**1U form factor**

**Direct liquid cooling**

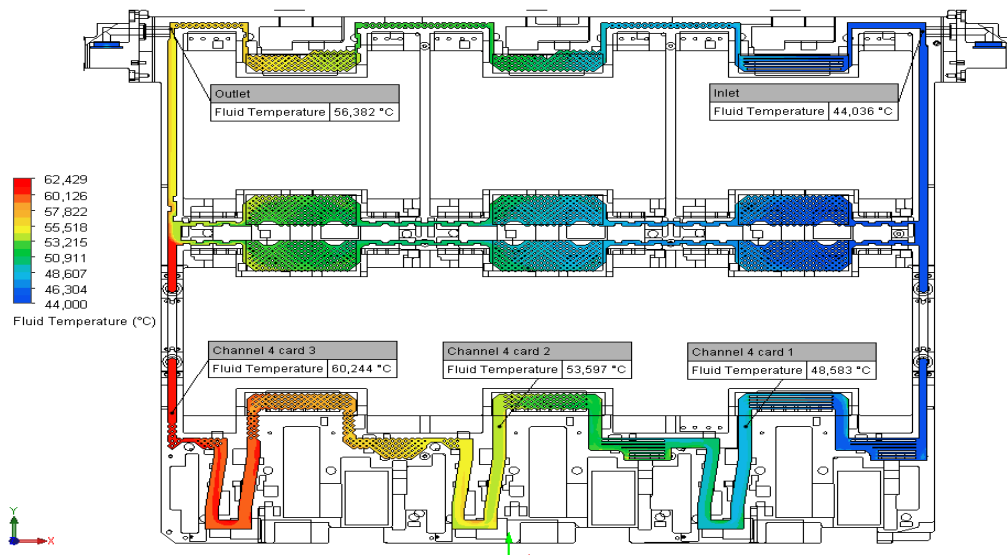
**3 dual-socket compute nodes per blade with:**

- 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)

**Interconnect  
Mezzanine**

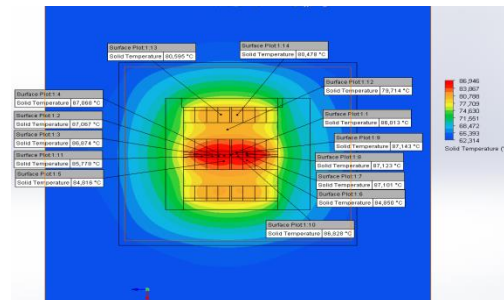
**Cavium® ThunderX2™**

# CVN thermal design & mechanical enclosure

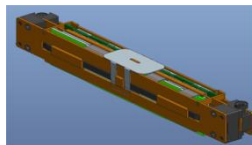


← new coldplate

new CPU heatspreader

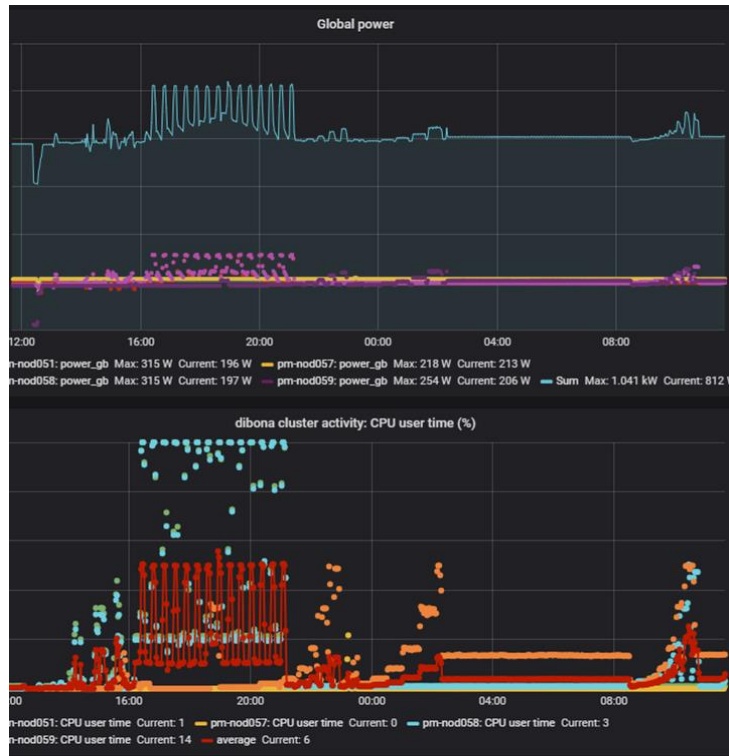


new DDR heatspreader



# HDEEM Energy accounting @ Dibona

- ➔ Original work initiated with TU Desden, and implemented in Mont-Blanc (1) cluster
- ➔ Now implemented in Dibona
- ➔ Energy accounting characteristics:
  - Energy cumulated through time (Blade + VRs + NICs) in Joules
  - High frequency FPGA energy calculation (100Hz for VRs, 1000Hz for Blade)



Applications

Sw Environment

Hw. platform

Simulation

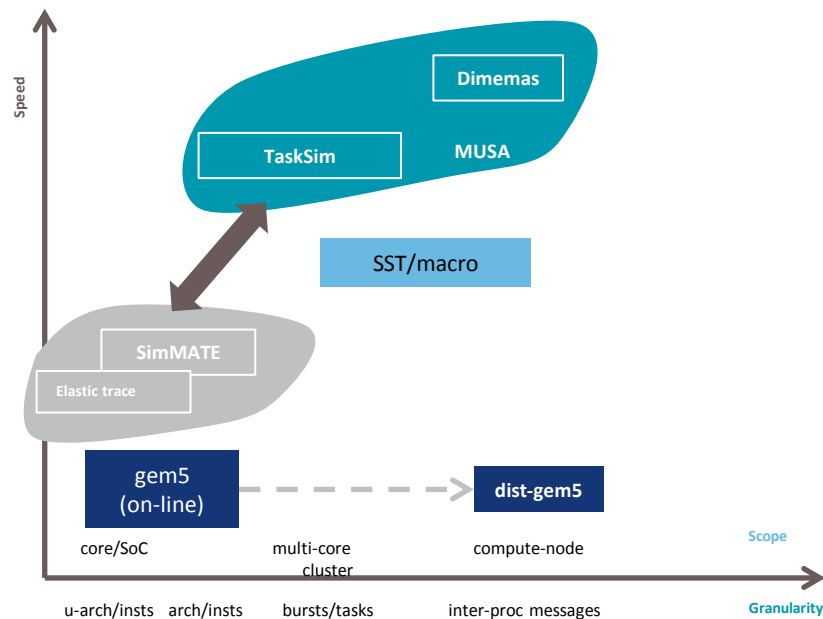
Computing  
Efficiency

Architecture

Balanced  
Architecture

# Simulation Tools

Name	on-line/of fline	scope	granularity	speed(up)
gem5 (classic memory)	on-line	core/SoC	u-arch/insts	~100-200 KIPS
Elastic trace (gem5)	off-line	interconnect/memory system	u-arch/insts	~ 7x (gem5)
SimMATE (gem5)	off-line	memory system	u-arch/insts	~6x – 800x (gem5)
Garnet (gem5)	on-line	interconnect		~0.2x (gem5)
TaskSim	off-line	compute node/task scheduler	arch/tasks	~ 10x native (burst) ~20x gem5 (memory)
Dimemas	off-line	cluster/off-chip network	bursts/messages	“very fast”
SST/macro	on-line	cluster/off-chip network	bursts/messages	~ 0.3-3x native



## Simulation tools map



## Integration in a global framework



Applications

Sw Environment

Hw. platform

Simulation

Computing  
Efficiency

Architecture

Balanced  
Architecture

# Design Space Exploration (DSE)

## → New Experiments

### SIMULATOR PARAMETERS/COMPONENTS

Cores	Issue width / ROB	Frequency	Cache size L3 / L2	Memory	Vector width
1	2 / 40 (lo-end)	1.5	64MB/1MB	4ch DDR4	128
32	4 / 180 (thunX)	2.0	32MB/1MB	8ch DDR4	256
64	6/224 (skylak)	2.5	32MB/256K		512
	8/300 (hi-end)	3.0			

↓ ↓ ↓ ↓ ↓ ↓  
Simulate every possible combination of these

=

865 detailed arch simulations per app

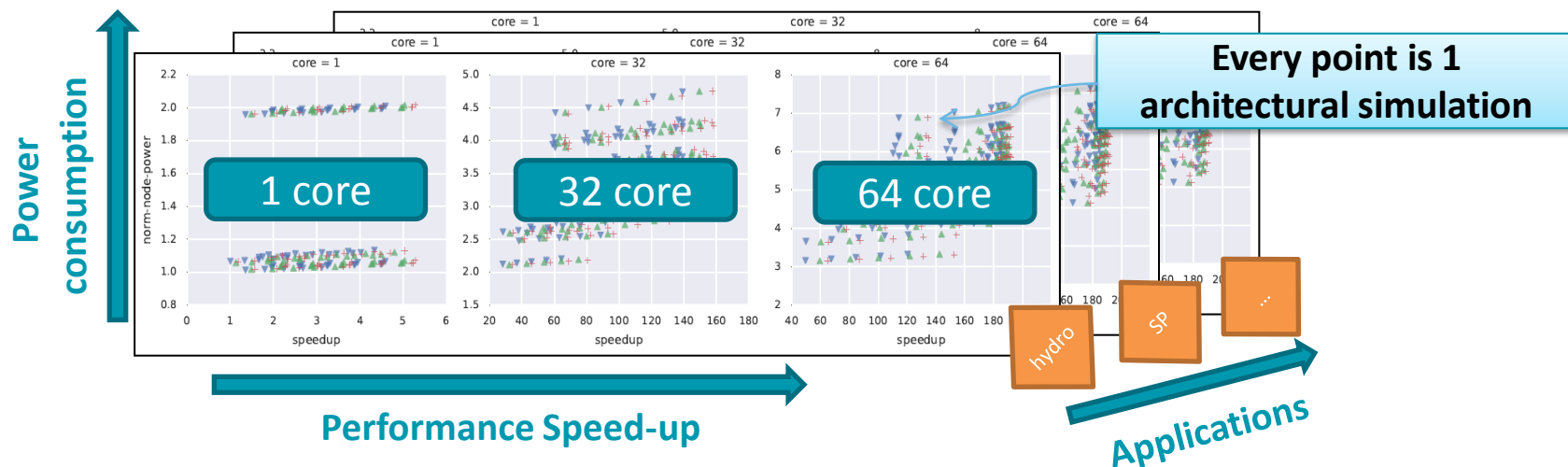
## → New Features

- FP Vector Width Simulation
- Chip and Main Mem. Power measurements

## → Applications

- BT, SP, Hydro, LULESH, Specfem3D

# MUSA DSE Results: (BSC)

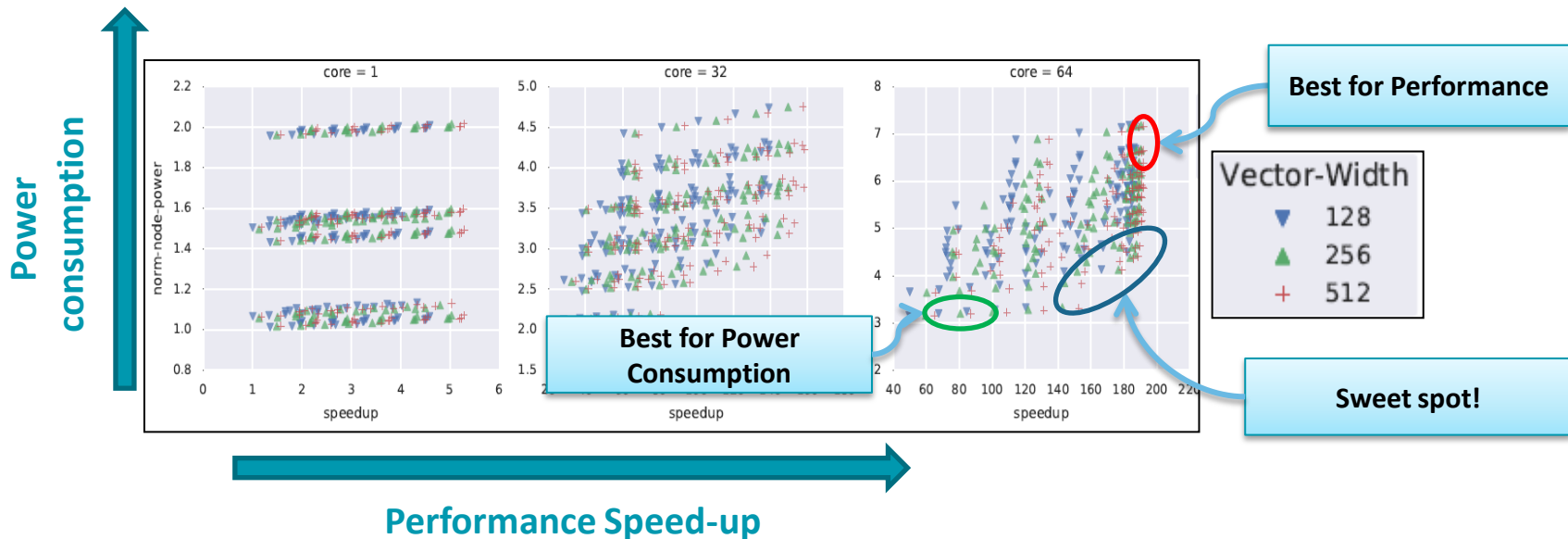


## → Results of this extended space exploration

- Wide study across architectural components
  - Interaction between components
  - Power / Performance tradeoffs
- Identifying trends and optimal cases. Understanding bottlenecks.

# MUSA DSE Results: Exploration (HYDRO)

## → Example: Identifying Optimal Cases



- Hydro simulation: 1 Rank during 1 iteration.
- 400 Architectural configurations per Core
- Legend coded by Vector Width used in the simulation

# Scalable Vector Extension (SVE)

## Key Architectural Features

- **Vector Length (VL)** is a hardware choice, from 128 to 2048 bits, in increments of 128
  - **Vector Length Agnostic (VLA)** programming adjusts dynamically to the available VL
  - **No need to recompile, or to rewrite hand-coded SVE assembler or C intrinsics**
  - **SVE is not an extension of Advanced SIMD**
    - Focus is HPC scientific workloads
  - **SVE also begins to address some of the traditional barriers to auto-vectorization**
- ✓ **Scalable vector length** (up to a current maximum of 2048 bits)
  - ✓ **Vector length Agnosticism**
  - ✓ **Per-lane predication**
  - ✓ **Gather-load and scatter-store**
  - ✓ **Fault-tolerant speculative vectorization**
  - ✓ **Vector partitioning**
  - ✓ **Horizontal and serialized vector operations**

- High performance and high efficiency cache coherent CPU clusters
- Use the right processor at the right time
- Potential for energy savings
- Flexible and transparent to the applications
  - Enabled by the Global Task Scheduling (GTS) or Energy Aware Scheduling (EAS)
- PoC
  - Simulation (on Juno development card)
  - To be emulated on our Dibona cluster (BIOS settings permitting to define core frequency (at core level))

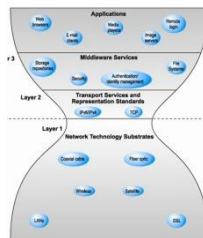


- **Move runtime overhead to an additional thread (task & dependence management)**
- **Implemented on top of OmpSs (a research runtime similar to OpenMP)**
  - Unit of compute: task
- **Functionalities**
  - Task [allocation and] submission within OmpSs
  - Push/pop tasks to/from the task graph
- **Use big or LITTLE cores to execute DAST**
- **Platform:**
  - Odroid-XU3 platform, ARM v7 (4xA7, 4xA15, 2GB RAM), Linux version 3.10.92

# Takeaway

- **No more silver bullet (Moore law is fading)**
- **Energy efficiency is a long term goal – efforts need to be sustained**
  - at all level: holistic approach needed
  - need to create the right tools for such a global approach
    - common abstraction
    - energy monitoring / control tools
- **A kind of hourglass model**
  - Many apps, many hw
  - Interoperability needed however
  - and also in time dimension:
    - Apps & hw : different lifecycles
    - but a need for common objectives

*Applications*



*Hws*



# MONT-BLANC

🌐 [montblanc-project.eu](https://montblanc-project.eu) | @MontBlanc\_EU

## Thank you for your attention

etienne.walter@atos.net



Credits : Mont-Blanc partners (third phase but also first & second phases)

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

