montblanc-project.eu | @MontBlanc_EU

Mont-Blanc (Arm for HPC)

Etienne Walter, Bull
with contributions from the whole team

Arm User Group workshop, Dallas

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

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

Prepare industrial solution
Test market acceptance

Mont-Blanc 2020

Extend the concept and explore new possibilities

Mont-Blanc 3

Proof of concept : HPC computing based on mobile embedded technology

Mont-Blanc 2

Mont-Blanc

2012 | *2013* | *2014* | *2015*[3] | *2016* | *2017* | *2018* | *2019* | *2020*

Mont-Blanc & Mont-Blanc 2 key results

- **Showed HPC computing was possible with ARM based hardware**
  - growing interest in using ARM for HPC
- **Showed that application scalability is possible even with very low end component**
- **Need denser compute - more cores**
- **Need a better interconnect (Native Ethernet)**
- **Optimizing energy efficiency at the system-level is hard**

Applications

Sw Environment

Hw. platform

Applications

Sw Environment

Hw. platform

MONT-BLANC

HPC Arm Software Stack - Mont-Blanc 3 contributions

Applications (source files – C/C++/Fortran/Python etc.)
Kernels | Mini-applications | Full applications

Compilers
GCC | LLVM | Mercurium

Scientific libraries
ATLAS | Arm Performance Libraries | BLIS

Developer tools
Extrae | Perf | PAPI | MAQAO | Allinea

Cluster management
SLURM | OpenLDAP | NTP | Ganglia

Runtime libraries
OpenMP | Nanos++ | HSA | MPI

Hardware support
Power management | DVFS | Lustre | NFS

Operating System (Linux kernel-based)
Ubuntu | OpenSUSE | CentOS | Fedora | RedHat | Drivers

(Heterogeneous) Hardware

- 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)

Most elements are available on Dibona platform

MONT-BLANC

Applications

Sw Environment

Hw. platform

Dibona: our new test platform

Dibona (Mont-Blanc 3 test platform)

Power supply units

48 computes nodes:
16 blades
96 CPUs
3072 cores
12 288 threads

Hydraulics for Direct Liquid Cooling (ultra-energy efficient with hot water cooling)

Redundant management server including storage

IB EDR switches
Fat tree topology

Internal Ethernet management network

now operational

ThunderX2 powered blades

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)

Thermal design & mechanical enclosure

new coldplate

new CPU heatspreader

new DDR heatspreader

Power management features (@ Dibona)

- **Energy and Power monitoring**
  - ✓ HDEEM data collection
    - ✓ Energy cumulated through time (Blade + VRs + NICs) in Joules
    - ✓ High frequency FPGA energy calculation (100Hz for VRs, 1000Hz for Blade)
  - ✓ SLURM Power plugin
  - ✓ Grafana visualisation

- **DVFS, Power capping and Turbo support**

Software stack and tools status (@ Dibona)

- RHEL 7.5 linux kernel-4.14.0-22
- Open MPI 2.0.x
- SLURM 17.x
- Mellanox OFED 4.3.x
- PAPI 5.4.x
- GCC 5.3.0, 6.2.1, 7.3.0, 8.20
- ARM HPC Compiler 18.4, 19.0
- HPCToolkit 2017.06
- the Allinea Studio tool :
  - the Arm HPC compiler 18.2. for RH 7.2+
  - the Allinea Forge debugging and profiling tool
  - the Allinea Performance Reports reporting tool
- Patched perf tool (providind 900+ uncore events for L3 partitions and DMCs)
- Patched PAPI (26 available events)

now operational

- **48 bi-socket TX2 nodes**
  - (32 cores B0 155 W SKU)
  - Theoretical peak performance > 49 Tflops
- **Derived into a product: BullSequana X1310**
- **An R&D asset to test sw frameworks & further SoCs**

Applications

Sw Environment

Hw. platform

MONT-BLANC

Applications

HPCG MPI-only architecture comparison (preliminary results)

MN4 → Intel Skylake,
P9 → IBM Power9,
MB3 → Cavium/Marvell ThunderX2 (Dibona)

Credits: Filippo Mantovani, Marta Garcia, Julian Morillo (BSC)

Alya + Multidependences (Preliminary results)

Credits: Marc Josep, Marta Garcia (BSC)

TensorFlow in Dibona

Credits: Guillem Ramirez (BSC)

Preparing the future SoCs

Applications

Sw Environment

Hw. platform

Simulation

Computing Efficiency

Architecture

Balanced Architecture

MONT-BLANC

Simulation Tools

Name	on-line/offline	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

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

MONT-BLANC

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

- Initial Datamining, plotting and analysis of results (Identifying Optimal Cases)

Takeaway

- **The arm ecosystem is maturing :**
  - Real High End solutions are here
- **We are now at the point where users can play seriously**
  - Access and provide further feedbacks
- **Openness (induced by the arm business model) is a key value**
  - We can expect to have more and more technical solutions,
  - while guarantying a good sustainability to users

MONT-BLANC

Thank you for your attention

Follow us:

@SC:

- Booth - #237 European Exascale Projects
- Wednesday, 14 November – 11:30am, room C146 Runtime-Assisted Cache Coherence Deactivation in Task Parallel Programs (BSC)
- Thursday, 15 November – 10.00am – 5.00pm, Aloft Hotel : Arm SVE Hackathon - powered by Dibona prototype

Other events :

- ICT18 in Vienna (6-8 Dec. in Vienna)

etienne.walter@atos.net

Credits : the Mont-Blanc team

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