Commodity embedded technology for future computational platforms

Filippo Mantovani
Barcelona Supercomputing Center
Technical Coordinator
Mont-Blanc projects goals

• To develop an European Exascale approach
• Leverage commodity and embedded power-efficient technology

Supported by EU FP7 with 16M€ under two projects:

• Mont-Blanc: October 2011 – September 2014 + 9 months
  14.5 M€ budget (8.1 M€ EC contribution), 1095 Person-Month
• Mont-Blanc 2: October 2013 – September 2016
  11.3 M€ budget (8.0 M€ EC contribution), 892 Person-Month
Mont-Blanc: Project objectives

- To deploy a prototype based on **currently available** energy-efficient embedded technology
  - Competitive with Green500 leaders in 2014
  - Deploy a full HPC system software stack

- To design a next-generation HPC system and new embedded technologies targeting HPC systems that would **overcome most of the limitations** encountered in the prototype
  - Learn from the experience and prepare for the future

- To port and optimize a small number of **representative scientific applications** capable of exploiting this new generation of systems
  - Up to 10 full-scale scientific applications
  - And not only HPC workload... We are at BDEC!
Mont-Blanc 2: Project objectives

- **Continue support for** the Mont-Blanc consortium
  - Mont-Blanc prototype(s) operation
  - Wider set of applications
  - Increased dissemination effort (End-User Group)

- Complement the effort on the Mont-Blanc *system software stack*
  - Development tools: debugger, performance analysis/prediction
  - OmpSs programming model
  - Resiliency
  - ARMv8 ISA

- Initial definition of future Mont-Blanc *Exascale architectures*
  - Continue tracking and evaluation of ARM-based products
  - Deployment and evaluation of small developer kit clusters
  - Performance & power models for design space exploration
Why are we doing this?

1 teraFLOPS supercomputer
ASCI Red
(Sandia – 1997)
Pentium Pro

1 petaFLOPS supercomputer
Roadrunner
(IBM / Los Alamos NL - 2008)
AMD Opteron + PowerXCell 8i

>10 petaFLOPS supercomputer
Titan
(Cray / Oak Ridge NL - 2012)
AMD Opteron + Nvidia K20
What is commodity nowadays?

~22M cores (June '14)

<table>
<thead>
<tr>
<th></th>
<th>Servers</th>
<th>PC</th>
<th>Smartphones</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>8.7M</td>
<td>350M</td>
<td>725M</td>
</tr>
<tr>
<td>2013</td>
<td>9.0M</td>
<td>315M</td>
<td>1000M</td>
</tr>
</tbody>
</table>

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

Source: International Data Corporation
Prototypes are critical to accelerate software development
System software stack + applications
Mont-Blanc Server-on-Module (SoM)

CPU + GPU + DRAM + storage + network all in a compute card just 8.5 x 5.6 cm

Exynos5 Dual:
2x ARM Cortex-A15
ARM Mali-T604

USB 3.0 to 1GbE bridge

4 GB DDR3-1600

μSD slot up to 64 GB
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

<table>
<thead>
<tr>
<th>Mont-Blanc [GFLOPS/W]</th>
<th>Green500 [GFLOPS/W]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov 2011</td>
<td>0.15</td>
</tr>
<tr>
<td>Nov 2014</td>
<td>1.5</td>
</tr>
</tbody>
</table>

*GPUs ~ 3/4 peak CPU ~ 1/4 peak
Limitation of commodity mobile technology

- 32-bit memory controller
  - Even if ARM Cortex-A15 offers 40-bit address space
- No ECC protection in memory
  - Limited scalability, errors will appear beyond a certain number of nodes
- No standard server I/O interfaces
  - Do NOT provide native Ethernet or PCI Express
  - Provide USB 3.0 and SATA (required for tablets)
- No network protocol off-load engine
  - TCP/IP, OpenMX, USB protocol stacks run on the CPU
- Thermal package not designed for sustained full-power operation

All these are implementation decisions, not unsolvable problems. Only need a business case to justify the cost of including the new features (e.g. the HPC and server markets).
Applications results (preliminary)

COSMO

- Atmospheric prediction model
- Weak scalability test
- Single core

![Graph showing performance results for COSMO](image)
Applications results (preliminary)

QuantumEspresso

- Electronic structure
- Strong scalability test
- Single core

Still working at...
- Analyzing traces
- OpenCL version of ZGEMM
- Understanding effects of MPI configuration
- Overlapping computation with communication

![Graph showing speedup (baseline MPI) vs. # of nodes for different configurations: Linear, MPI, MPI+OmpSs, MPI+OmpSs+OpenCL.](image)
Non-MB application (preliminary)

NMWB – weather forecast (BSC)
Global Run, 1.40625° x 1°, 24h forecast, 1h output, CPU only

<table>
<thead>
<tr>
<th></th>
<th>MareNostrum3 32 cores – 2 nodes</th>
<th>MareNostrum3 32 cores – 16 nodes</th>
<th>Mont-Blanc 32 cores – 16 nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run1</td>
<td>488s</td>
<td>169s</td>
<td>2039s (4x / 12x)</td>
</tr>
<tr>
<td>Run2</td>
<td>487s</td>
<td>171s</td>
<td>1958s (4x / 11x)</td>
</tr>
</tbody>
</table>
NON-HPC workload: Hadoop 2.0

- Datanode stores data as distributed by namenode.
- Nodemanager executes map and reduce java process as guided by resource manager.
Teragen and Terasort

• Terasort is one of the important test for hadoop clusters which benchmarks disk, CPU, memory and network performances.

• Teragen generates data, that can be sorted using terasort map reduce.
  
  • It generates data in form of 100 bytes rows.

  • Each row has the format:
    <10 bytes key><10 bytes rowid><78 bytes filler>

  • The key is the id generated by the map task and rowid are serial numbers.
    e.g. generating 1000 lines using 10 maps, the key range will be 1-10 and rowid as 1-100

• The map tasks in terasort will collect the lines based on the keys and reduce tasks will arrange the lines in serial order.
**Hadoop installation on Mont-Blanc platforms**

**NVIDIA Jetson Mini-cluster**
- 256 GB SSD on jetson 1-8
- Local bw 180 MB/s
- Total storage 1.73 TB

**Mont-Blanc Prototype**
- 11.2 GB uSD cards on each datanode
- Local bw 13 MB/s
- Total storage 100 GB
Teragen on Jetson (8 min 25GB, std.dev 2.26%)

All Jetson slaves have similar load footprint for I/O with the SSDs. Hence generated data is equally distributed with a standard deviation of 2.26%.

Average write bw:
52.1 MB/s
~30% of “peak” bw
Terasort on Jetson (28 min for 25 GB)

Not much network in mapping phase. Most maps run locally. Reduction phase also shuffles data across network.

All Jetson slaves have similar load foot print for maps and reduces.
Teragen on MB_proto (45 min 25GB, std. dev. 20.21%)

- All MB nodes have different load footprint.
- I/O with some SDcards is poor. Hence data is unequally distributed with standard deviation of 20.21%.

Average write bw: 9.3 MB/s
~70% of “peak” bw

CPUs under utilized. Potential reason, I/O with SDcards.
Terasort on MB-proto (38 min 25GB)

Good CPU utilization in terasort as it also includes computation and not just I/O.

All MB nodes have almost similar computational footprint, the differences are due to poor performance of some SD cards.
Terasort and Teragen node loads

Terasort: almost balanced load

Teragen: unbalanced load
Hadoop preliminary observations

• MB-proto can catch up with Jetson on terasort with few more nodes available.

• MB-proto ethernet network looks capable enough for hadoop loads.

• Physical memory is the limiting factor for the parallel maps and reduces on Jetson cluster.

• SD cards on MB-proto are limiting the performance of hadoop setup.
End-User Group

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

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
Conclusions:

- Need sustainable EFLOPS technology
  - min(power + space + cost + … )
  - Energy/cost efficiency
  - Commodity market (both mobile and server)

- Preliminary results show acceptable scalability figures for HPC applications

- Preliminary tests of big data load have been performed
  - Highlighted some limiting factors (RAM on Jetson, local storage + network on MB)

- Still a lot to do... but the MB prototype is behaving “incredibly” well under different kind of workloads:
  - HPC
  - Hadoop
  - Other applications (BSC weather forecast, End-User Group)