D7.8 Prototype System Deployed
Version 1.0

Document Information

Contract Number	288777
Project Website	www.montblanc-project.eu
Contractual Deadline	PM42
Dissemination Level	PU
Nature	P
Author	Hervé Gloaguen
Contributors	Said Derradji (BULL SAS)
Reviewer	Filippo Mantovani, Nikola Rajovic (BSC)
Keywords	Prototype

Notices:
The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement n° 288777.

Change Log

Version	Description of Change
V0.0	Initial Draft for internal project review
V0.1	Review by Nikola Rajovic
V0.2	Review by Filippo Mantovani
V0.3	Review by Nikola Rajovic
V1.0	Final version sent to the EC

Table of Contents

Table of Figures

Executive Summary

This deliverable provides the technical description of the final prototype system delivered to BSC for the use within the Mont-Blanc project.

This system consists of 1080 nodes that are deployed in two separate partitions, a small one for test and development and a large one for running applications. The latter is a separate entity and has its own interconnect and storage subsystems.

1 Introduction

One of the main objectives of the Mont-Blanc project is to build a prototype based on mobile embedded technology that can be used by the partners to run their application and do experiments in order to evaluate ARM-based technology under high-performance computing workloads. The Mont-Blanc prototype is made of Bull specific parts and Mont-Blanc specific or commodity parts, these being sourced from AsteelFlash through a call for tender.

2 Technical description of the final prototype

2.1 Compute nodes

The Mont-Blanc compute blade is organized the following way:

- It is the association of multiple boards:
  - The node card, named Samsung Daughter Board (SDB): hosting the ARM-based Samsung Exynos 5250 System-on-Chip and the related components. This board has been specifically manufactured for Mont-Blanc project.
  - CMC2: hosting the Board Management Controller (BMC). This board is a standard Bull board.
  - TSM-C: hosting the controller of the embedded Ethernet switch. This board is a standard Bull board.
  - Ethernet Mother Board (EMB): hosting 15 SDB, 1 CMC2, 1 TSM-C and the related components.
- It is enclosed in a double-width mechanical structure, embedding four fans on the front side and a Mont-Blanc logo on the handle.

The design of an SDB is depicted on Figure 1:

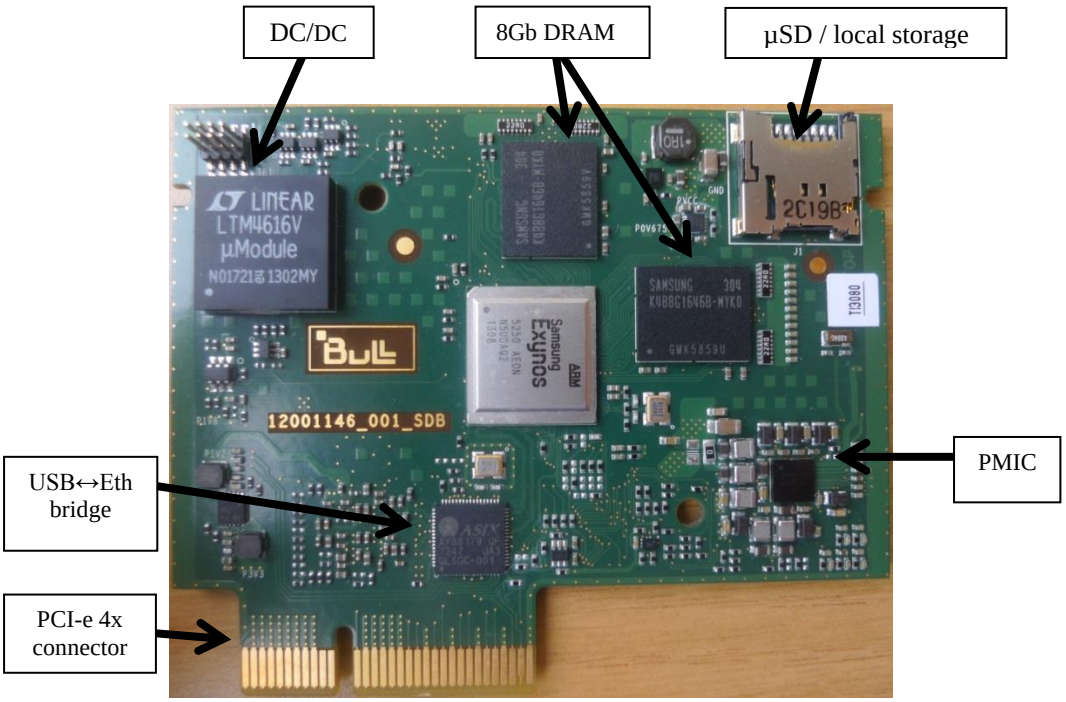

Figure 1: Samsung Daughter Board (SDB) - physical view

The Ethernet Mother Board (EMB) is the board that provides all blade level functionalities. Its main design objectives are:

- Integrate all blade components while optimizing the density
- Fit in a double-width Bullx blade

The EMB, shown in Figure 2, provides the following functionalities:

- Capability of hosting up to 15 SDB
- First level of 1 Gb Ethernet interconnection of the 15 SDB
- Two 10 Gb Ethernet ports (with SFP+ connectors) for connection with other EMBs
- Management and sideband interconnection of the 15 SDB
- Per SDB power consumption monitoring
- Blade-level temperature monitoring
- Power supply for all the components, with blade-level hot-plug

The EMB with its 15 SDBs fits in an adapted Bullx blade, so that these blades fit in a Bullx chassis and can reuse its management, power supplies and cooling (fans).

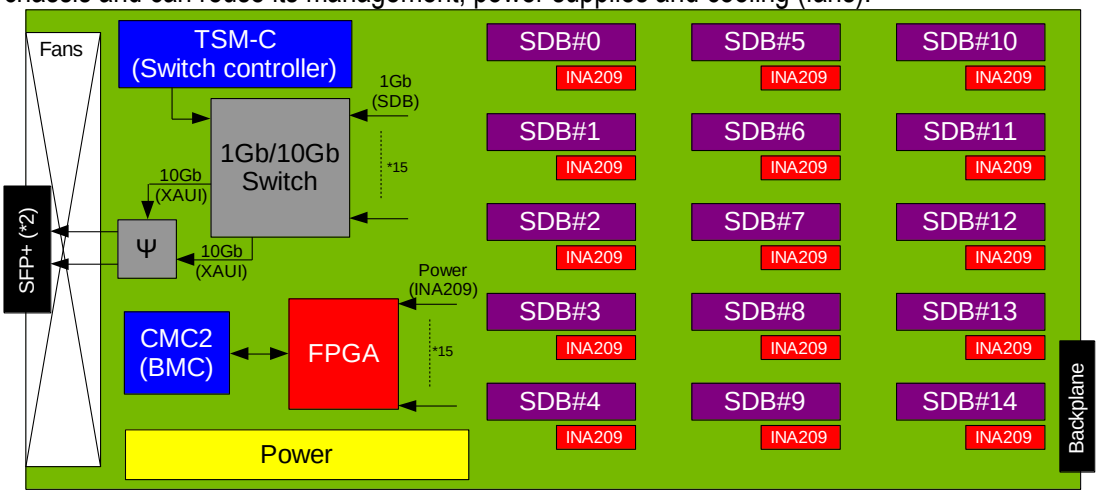

Figure 2: EMB - schematic view

The hardware in charge of compute capabilities is composed of 8 chassis. Each chassis houses 9 EMBs and each EMB contains 15 SDBs. In total 1080 computational nodes have been delivered.

2.2 Storage

The network file system is based on a Storage Bridge Bay (SBB) from Supermicro.
The model 2027B-DE2R24L contains two hot-pluggable systems (nodes) in a 2U form factor.
Each node supports the following:

1. Dual socket B2 (LGA 1356) supports Intel® Xeon® processor E5-2400 v2
2. Up to 192GB DDR3 1600MHz ECC registered DIMM; 6x DIMM sockets (3 per CPU)
3. 3x PCI-E 3.0 slots per node (can be used for host or storage expansion)
4. I/O ports: 2 GbE LAN (1 for IPMI 2.0 w/ Virtual Media/KVM over LAN),SAS 2.0 (6Gbps) x1 JBOD ports
5. Dedicated node to node connectivity featuring high performance PCI-E 3.0 x8 and IPMI for robust node fail-over support

6. 24x Hot-swap 2.5" SAS2 HDD bays Internal 2x SAS/SATA ports for OS load on SLC SATA DOM
7. 920W Redundant Power Supplies Platinum Level Certified

Each node is connected to 2 Ethernet switches with 10 Gb Ethernet copper cables. In order to provide this double 10 Gb Ethernet connection, the node contains one PCIe card with dual-port 10Gb iSCSI Ethernet NIC.

New generation of Western Digital HDD disks are used in the storage system, with capacity of 600GB and sustained read/write bandwidth from 129MB/s to 224MB/s, depending on the location of the data on the disk. The SBB server is equipped with 16 disks, 8 per node in RAID5. The total raw capacity is 9.6TB.

2.3 Interconnect

The first level of switches is embedded in the compute blade (EMB), Figure 3 depicts the topology of the Ethernet network inside the Mont-Blanc chassis, for both cluster interconnect (18x10 Gb Ethernet outputs) and management interconnect (1 Gb Ethernet output).

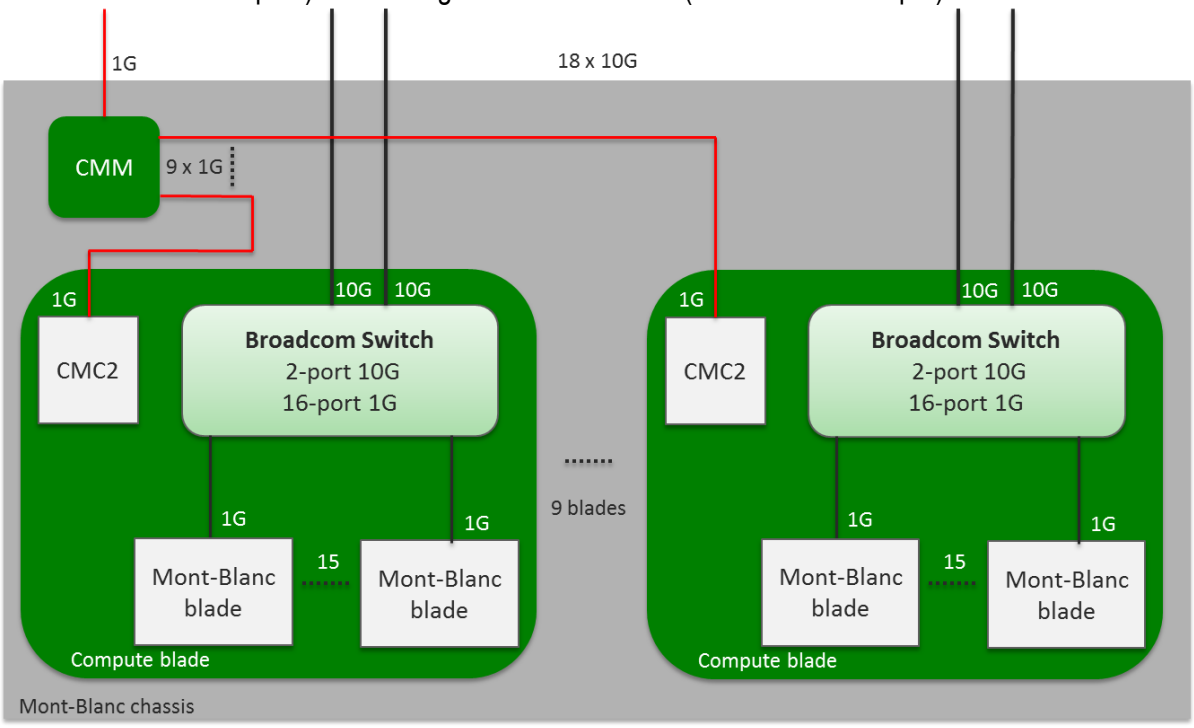

Figure 3 : Chassis Ethernet networks topology

Figure 4 illustrates the topology of the cluster interconnect built to interconnect the Mont-Blanc chassis between each other's and to the storage nodes.

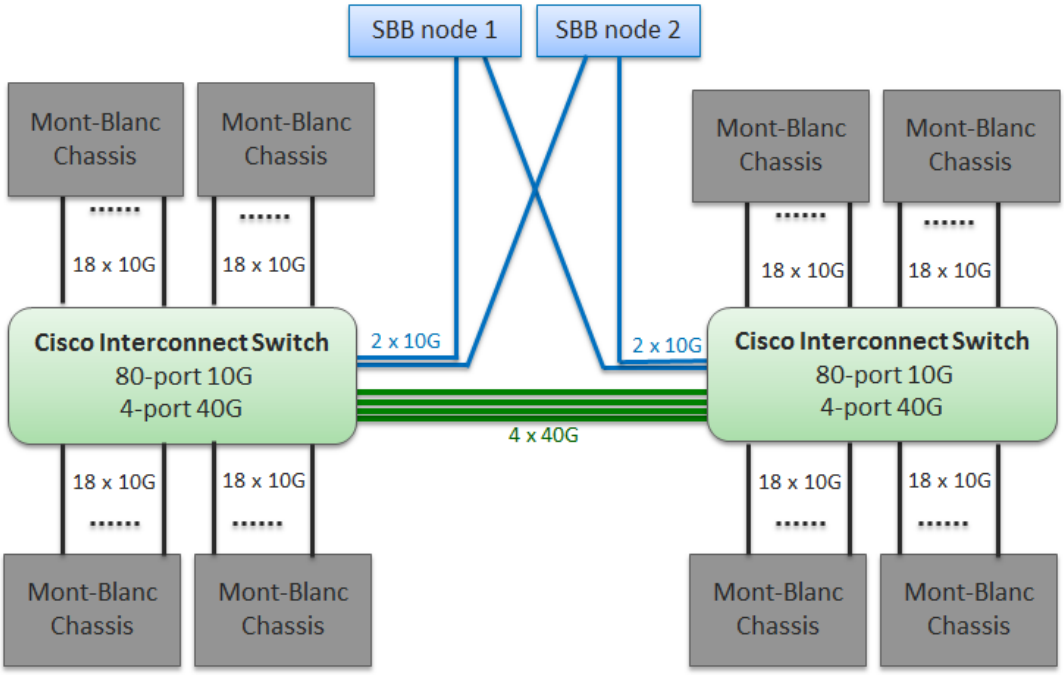

Figure 4 : Cluster interconnect topology

Each Mont-Blanc compute blade provides two 10 Gb Ethernet ports. This implies 18 SFP+ connectors per chassis. A total of (18x8) 144x10 Gb Ethernet ports are required for the blades. Moreover, the storage solution consists in two nodes offering each two 10 Gb Ethernet ports. Therefore additional 4x10 Gb Ethernet ports are required for the storage.
The interconnect network is implemented using two Nexus switches from Cisco 5500 Series. The selected model is the Cisco Nexus 5596UP Switch, a 2RU 10 Gigabit Ethernet switch offering up to 1920 Gb/s of throughput and up to 96 ports.

Each Nexus switch provides 78 connections in Mont-Blanc prototype. The detail of the port's configuration is:

- 72 ports SFP+ 10 Gb Ethernet to local compute blades resident inside the rack
- 2 ports SFP+ 10 Gb Ethernet to the storage nodes
- 4 ports QSFP 40 Gb Ethernet for interconnection between racks.

2.4 Management

For the purpose of cluster management, is a separate Ethernet network is implemented. This management network is based on 1 Gb Ethernet.
The first level of management switches is implemented inside the chassis with a CMM board. The 1 Gb Ethernet switch in the CMM is connected to the management controller CMC2 in each compute blade through the Bullx chassis backplane. CMC2 and CMM have their own IP address.

The second level of management network of the cluster is illustrated in the next Figure.

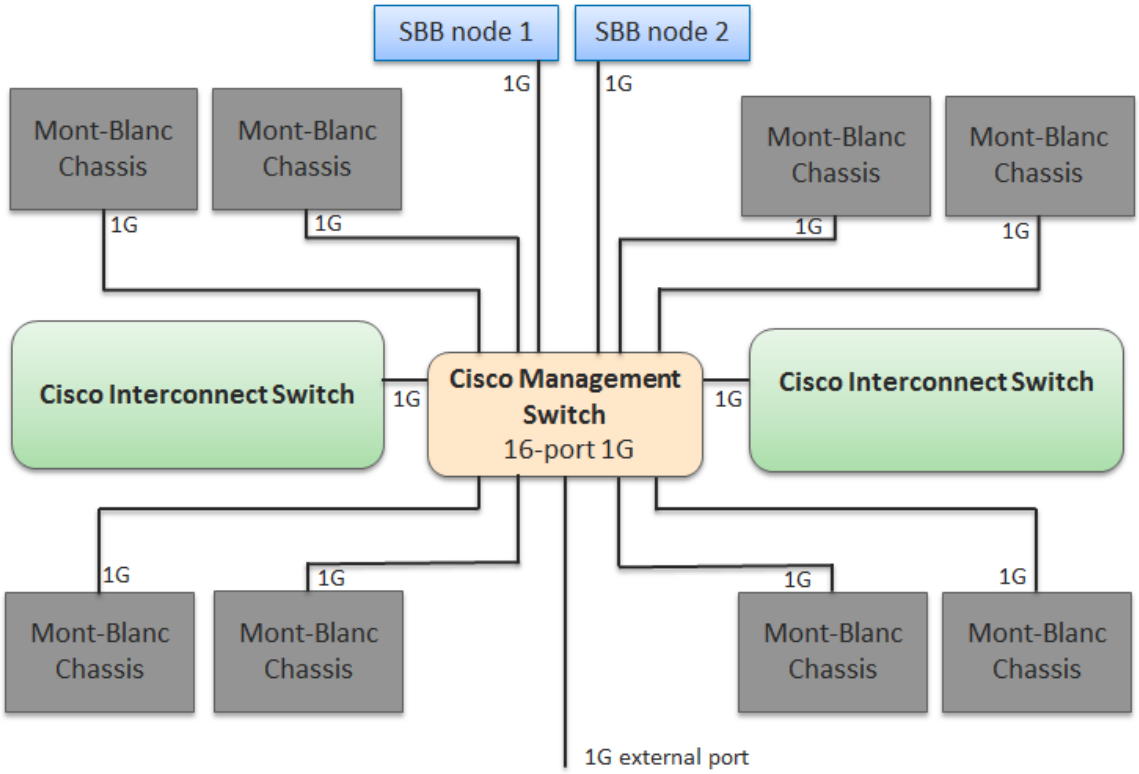

Figure 5 : Management interconnect topology

The 12 components of the cluster (8 chassis, 2 storage nodes and 2 interconnect switches) are connected to a specific Management switch, that connects the management network to the lab backbone with a 1 Gb Ethernet link. For that purpose, a 16-port 1 Gb Ethernet switch is included in the cluster configuration. The switch is the SG300-20 model from Cisco.

3 Testing and Installation

Test of the single components has been described in D7.7. Before delivering the final prototype a comprehensive test of the whole prototype has been performed in order to check the functionalities and the procedures for the final installation.

Installation of the prototype at BSC's facilities has been to split it in 2 partitions:

- The test partition, including one chassis with 135 nodes, is kept in the Mont-Blanc BSC lab for the purpose of software and firmware testing before final deployment in production cluster.
- The production partition, including 7 chassis with 945 compute nodes, is installed in the "Torre Girona Chapel" at the Technical University of Catalunya and is dedicated for running large applications and testing their scalability.

Both partitions are accessible to all partners.
An image of the production partition is presented in Figure 6.

Figure 6: Operational Mont-Blanc production system

4 Schedule

This last section analyses the schedule with an explanation of the delays, as compared to the initial schedule.

Activity	DOW (deliverables)	at P1 review	at P2 review	final
SoC selection	M12	M14	M14	M14
Prototype evaluation board	M18	M21	M23	M23
Prototype system deployed	M27	M29	M36	M40

At P1 review, the main shift was due to the delay on the choice of the SoC, as there were some discussions between the two last candidates (TI and Samsung).

At P2 review, the first round of prototype was done, so the technical risks were covered. But the impact was due to the way the final system can be procured, implying some administrative work that had been initially overlooked.

The last shift between the P2 review and the actual schedule was due to the date when the call for tender was eventually issued by BSC, also due to administrative and legal discussions. Part of this final delay was anyway mitigated by some anticipation in the procurement, as the normal lead time is 22 weeks.