Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Technical Requirements for Mont-Blanc

Each entry in the following tables has been classified as:

R – Requirement, which must be satisfied by any proposed solution

D – Desired attribute, which would be valued in any solution

1. Introduction:

This document describes the technical requirements for a prototype energy-efficient cluster, with distributed memory, single-chip integrated multicore CPU, and graphical processor units supporting OpenCL and double precision floating point.

The following describes the various components and internal networks that comprise the cluster:

- The cluster will be built from 1080 node cards, each with two or more ARM processor cores and an integrated GPU that supports OpenCL 1.1 and double precision floating point, 4 GB or more of LPDDR3/DDR3 DRAM, 1 GbE network or better, and a flash card slot for local storage.

- Up to 15 node cards can be configured as login nodes.

- The cluster will be connected to a storage system that can guarantee a minimum of 8 TB disk space and can be accessed with a bandwidth of 2 GB/s. The storage system needs a minimum of two separate nodes.

- The interconnection network must aggregate multiple 1 GbE links into a minimum number of 10 - 40 GbE links, with a worst case bisection bandwidth of 160 Gb/s.

- The 1 GbE network will be also used for cluster management.

- The cluster must integrate a minimum of 540 nodes per rack.

- It must be possible to remotely reset or check the status of single node cards and perform maintenance tasks on faulty nodes without shutting down the entire cluster.

The next section describes the system and each component in more detail.

2. Hardware Description

2.1 Computational Nodes

Ref	Description
R2.1.1	A node card will be the basic computational element of the cluster. An amount of 1080 node cards, with all the necessary infrastructure to build a cluster, will be deployed.
R2.1.2	Each node card will be deployed after being tested and validated.
R2.1.3	Each node card will have the following minimum characteristics: - 2 processor cores with ARM ISA ARMv7 and VFPv4; e.g. Samsung Exynos 5 Dual, with a minimum frequency of 1.7 GHz - On-chip integrated GPU with peak 25 GFLOPS (double-precision) per node, programmable using OpenCL 1.1 - 4 GB main memory per node, 1600 MHz or more, dual-channel, technology LPDDR3/DDR3 - 16 GB local storage in a micro SD card (class 10)
R2.1.4	Each node card must be connected to the rest of the components of the cluster via a single Gigabit Ethernet interface.
R2.1.5	The documentation must contain a block diagram of the node, showing the bandwidths between the components (maximum and used) for the processors, memory, and I/O.
R2.1.6	The cluster must support power consumption monitoring at the level of every single node card, with a frequency of one measurement every second, and an accuracy of 5%.
D2.1.7	Any improvement beyond the above-mentioned minimum values would be preferred, whether in the number of nodes or the specification of each node.
D2.1.8	Any improvement in the capacity or bandwidth of the micro SD card beyond the above minimum values would be preferred.
D2.1.9	The following table (Table 1 – Node Hardware Description) must be completed, so that it contains both the minimum required values given above and the values for the proposed system.
D2.1.10	Any improvement in any entry of Table 1 would be valued.

2.2 Network File System

Ref	Description
R2.2.1	A network storage server with at least 8 TB capacity, with a minimal bandwidth of 2 GB/s to/from the disks, will be provided.
R2.2.2	The systems will include at least 25 commodity SSD disks with a minimal bandwidth of 80 MB/s each.
R2.2.3	The storage system needs a minimum of two nodes in order to support metadata and storage server and disk space for local OS.
R2.2.4	File server used as storage node needs 2x10GbE links. File server used as metadata node needs 1x10GbE link.

Table 1 – Node Hardware Description

Description	Minimum value	Offered value
Characteristics of node card		
Number SoCs per node	1	
SoC model		
Cores per SoC	2	
Frequency of each core	1.7 GHz	
RAM capacity per node	4 GB per node	
Memory technology and access frequency	LPDDR3/DDR3 800MHz	
Internal micro SD capacity		
Internal micro SD drive interface		
Number of Ethernet out-of-band management interfaces	0	
Number of 1 GbE interfaces	1	
Peak performance ARMv7 per core [GFLOPS]	3.4	
Peak performance - GPU per node [GFLOPS]	25	
LINPACK Rmax [GFLOPS]		
LINPACK Rmax / W [GFLOPS]		
Global characteristics of compute cluster		
Total number of computational nodes	1080	
Characteristics of storage server(s)		
Number of nodes	2	
Total capacity	8 TB	
Bandwidth read/write	2 GB/s	
Number of disks	25	
Bandwidth per disk	80 MB/s	
10 GbE links	2+1	

2.3 Networks and Switches

Ref	Description
R2.3.1	The documentation must contain a block diagram showing the connection of each network in the cluster
R2.3.2	The following table (Table 2 – Switch Hardware Description) must be completed, so that it contains both the minimum required values and the values for the proposed system.
D2.3.3	An improvement in any entry, with minimum value or not, will be valued.

Ref	Description
Cluster interconnect	
R2.3.4	All necessary hardware must be included (switches, cables, etc.), in order to build the interconnect with Gigabit Ethernet (GbE).
R2.3.5	Functional requirements: - Level 2 filtering and policies - Access Control Lists - SSHv2 support - Private VLAN support - 802.1Q support - Complete management using SNMPv1, 2 and 3 - Protection against Spanning tree loops - Support for redundant links (multiple links between distinct nodes on the network), using LACP protocol - Filtering of BPDUs from STP
R2.3.6	This interconnect should: - Connect each node using one Gigabit Ethernet interface - Aggregate the maximum 1 GbE links into 10 GbE links without over-subscription - The (worst case) bisection bandwidth must be at least 160 Gb/s - Allow the cluster management software and scheduler to access power counters and other status information available on each node card
D2.3.7	Minimization of the hop count or an increase in the bisection bandwidth would be valued.
D2.3.8	Support for Energy-Efficient Ethernet (802.3az) would be valued. Please specify which components in the network support Energy-Efficient Ethernet.
R2.3.9	At least 5% of the network ports should be free, to allow for future upgrades (the GBIC, where necessary, need not be included)

Ref	Description
Management interconnect	
R2.3.10	All necessary hardware must be included (switches, cables, etc.), in order to build the interconnect with Gigabit Ethernet.
R2.3.11	Functional requirements: - Level 2 filtering and policies - Access Control Lists - SSHv2 support - Private VLAN support - 802.1Q support - Complete management using SNMPv1, 2 and 3 - Protection against Spanning tree loops - Support for redundant links (multiple links between distinct nodes on the network), using LACP protocol - Filtering of BPDUs from STP
R2.3.12	This interconnect should: - Connect each node using one Gigabit Ethernet interface - Aggregate the maximum 1 GbE links into 10 GbE links without over-subscription - The (worst case) bisection bandwidth must be at least 160 Gb/s - Allow the cluster management software and scheduler to access power counters and other status information available on each node card
D2.3.13	Minimization of the hop count or an increase in the bisection bandwidth would be valued.
D2.3.14	Support for Energy-Efficient Ethernet (802.3az) would be valued. Please specify which components in the network support Energy-Efficient Ethernet.
R2.3.15	At least 5% of the network ports should be free, to allow for future upgrades (the GBIC, where necessary, need not be included)

Table 2 – Switch Hardware Description

Description	Minimum value	Offered value
Cluster interconnect		
Level 1		
Number of switches		
Switch manufacturer		
Switch model		
Number of ports per switch		
Number of free ports per switch		
Input link bandwidth	1 Gb/s	
Uplink bandwidth	10 Gb/s	
Level 2		
Number of switches		
Switch manufacturer		
Switch model		
Number of ports per switch		
Number of free ports per switch		
Input link bandwidth	10 Gb/s	
Uplink bandwidth		
Level 3		
Number of switches		
Switch manufacturer		
Switch model		
Number of ports per switch		
Number of free ports per switch		

3. Operational

Ref	Description
Operational requirements	
R3.1	Components must fit in standard 42U - 19” racks and all components within the racks must be supplied.
R3.2	Content of each rack must weigh no more than 900 kg.
D3.3	A lower number of required racks would be preferred.
R3.4	The documentation should contain a diagram showing the contents of each rack, specifying clearly the hardware and number of U’s occupied, at each position in the rack.
R3.5	The total system peak power consumption must be no greater than 35 kW, including all components.
R3.6	The idle power consumption per rack must be no greater than 19 kW, including all components without exception.
D3.7	Lower total system power consumption would be preferred. Indicate the total system power consumption, for the following situations: - Maximum power consumption - Power consumption in idle state - If known, average power consumption when executing Linpack
D3.8	Redundant power supplies would be preferred, provided that they can be switched off in software.
D3.9	Telephone assistance and/or assembly guidelines would be valued. On-site assistance during installation, on delivery in Barcelona, would be highly valued.

4. Software

Ref	Description
R4.1	The operating system must be UNIX-like, and compatible with X/Open Standard POSIX 1003 (IS/IEC 9945). Linux would be preferred.
R4.2	All software required to manage all components of the system, including - cluster management software - switch management software - operating system - OpenCL drivers must be included.
R4.3	Operating system and system software will be deployed under the form of bootable images that can be flashed on the micro SD of the node cards.
R4.4	Standard cluster and switches management software will be deployed under the form of pre-loaded and pre-tested firmware.

5. Maintenance and support

This section describes the minimum requirements for maintenance and support of the whole system. It also describes the improvements in maintenance and support that would be valued.

Ref	Description
R5.1	Two-year guarantee and support on all standard hardware components.
D5.2	A guaranteed response time, during office hours and/or 24-hour, would be valued. A lower guaranteed response time would be preferred.
D5.3	An increase in the length of the guarantee and/or support would be valued.
D5.4	On-site assistance would be valued, comprising cluster configuration, and troubleshooting and resolving incompatibility or other problems that arise when the system is installed.
D5.5	Pro-active support would be valued, comprising notification whenever a software upgrade is available, for any standard component.
R5.6	Documentation, in digital form, should be included with the solution, comprising: - General description of the system components - Connection diagram and IP addresses - Values of configuration parameters - Installation instructions - How to startup the machine, disaster recovery
R4.7	For each standard hardware component, a single point of contact should be provided.