D2.4

ExaNoDe Infrastructure Requirements

Workpackage:	2	Co-Design for Exa-scale HPC System
Author(s):	Elyès Zekri	BULL-ATOS
	Dirk Pleiter	JUELICH
Authorized by	Elyès Zekri	BULL-ATOS
Reviewer	Emre Ozer	ARM
Reviewer	Petar Radojkovic	BSC
Reviewer	Guillaume Colin de Verdière	CEA
Dissemination Level	Public	

Date	Author	Comments	Version	Status
2017-07-28	E. Zekri	Initial draft	V0.0	Draft
2017-09-07	E. Zekri D. Pleiter	Added content for section2	V0.1	Draft
2017-09-13	E. Zekri D. Pleiter	Added content for section3	V0.2	Draft
2017-09-15	E. Zekri D. Pleiter	Document finalised and submitted to review	V1.0	Ready for QA
2017-09-27	E. Zekri D. Pleiter	Modifications following review process	V1.1	Ready for QA (2nd interation)

Executive Summary

In the context of ExaNoDe Task 2.4, we are aiming to describe and analyse datacentre infrastructure requirements for an Exascale High Performance Computing (HPC) system using ExaNoDe compute solution.

Therefore, we present general HPC infrastructure components and their related requirements regarding power, cooling, storage, interconnect network and datacentre equipment monitoring, then we describe our long-term vision for integrating ExaNoDe compute system in a HPC infrastructure.

The main outcomes of this analysis are related to these topics:

- analysis of HPC power efficient infrastructure requirements and the description of enhancement possibilities regarding power efficiency and storage,
- description of a compute blade for ExaNoDe which is compliant with modern HPC infrastructure requirements,
- suggestion of an interconnect network solution based on 3D-torus taking into consideration Unimem hardware constraints,
- presentation of a set of requirements, for connecting ExaNoDe nodes to a storage subsystem, in terms of bandwidth, network, and software.

Table of Contents

1 Introduction 1
2 HPC infrastructure general requirements 1
2.1 HPC compute system packaging 1
2.2 Power system 2
2.2.1 Power distribution 2
2.2.2 Uninterruptible power 2
2.2.3 Power distribution enhancement 2
2.3 Cooling system 3
2.3.1 Power efficiency metrics 3
2.3.2 Boards cooling solutions 4
2.4 Interconnect network 5
2.4.1 Network topologies 5
2.4.2 Network packaging 7
2.5 Storage system 8
2.5.1 Background on future storage architectures 8
2.5.2 Exascale requirements 9
2.6 Visualization subsystem 10
2.7 Infrastructure management and monitoring 10
2.7.1 Key infrastructure equipment monitoring 10
2.7.2 Rack management 11
3 Proposed rack-scale solution for current ExaNoDe compute system 12
3.1 Compute blade 12
3.1.1 Compute blade description 12
3.1.2 Compute blade internal links 13
3.1.3 Compute blade external links 14
3.1.4 Compute blade cooling 14
3.2 Interconnect network topology 15
3.3 Rack management network 16
3.3.1 Management network components 16
3.3.2 Ethernet network 16
3.4 Storage subsystem 17
3.5 Maximum power consumption estimation 17
3.6 Rack composition 18
3.7 Datacentre operating conditions 19
4 Conclusions 20
5 References 20

Table of Figures

Figure 1. Network types and topologies 6
Figure 2. Decreasing the longest cable length in a torus (a) by “folding” it (b). 7
Figure 3. Rack management architecture 11
Figure 4. MCM Compute Board (CB) 13
Figure 5. MCM Compute Blade 13
Figure 6. Compute Blade 14
Figure 7. Cooling components of Bull Sequana X1210 compute blade 15
Figure 8. Rack-scale interconnect network topology (3D-Torus, N=288) 16
Figure 9. Ethernet Management Network 17
Figure 10. Rack composition 19

Table of Tables

Table 1. Liquid cooling solutions classification 5
Table 2. Network diameter, nodal degree and bisection bandwidth (n is dimension, N is total number of nodes) 7
Table 3. Two-tier storage subsystem parameters 8
Table 4. Compared performance numbers between NERSC-9 and targeted exaflopic system 9
Table 5. Key equipment measurements 11
Table 6. Maximum power consumption estimation 18

1 Introduction

The main constraint of Exascale is the power required to achieve Exascale computing. At present, a single computing system can take up to 13 MW from the power source and the Exascale target systems will draw about 20 to 25 MW [1]. This expected huge increase in power consumption, makes it necessary to introduce more power efficient solutions regarding datacenter infrastructure: compute components, power distribution, storage and cooling systems design.

In the context of ExaNoDe Task 2.4 we are aiming to describe datacentre infrastructure requirements for an Exascale High Performance Computing (HPC) system using ExaNoDe compute solution. ExaNoDe solution is based on Multi Chip Module (MCM) and Unimem memory system and aims to deliver a prototype-level system demonstrating that those technologies are promising candidates towards the definition of a compute node for the Exascale computing [2].

This deliverable is organized in two main sections:

- Section 2 of this document presents general HPC infrastructure components and their related requirements regarding power, cooling, storage, interconnect network and datacentre equipment monitoring,
- Section 3 describes a long-term vision for ExaNoDe compute solution integration in a HPC infrastructure inspired by Bull-Atos Sequana Platform [3].

2 HPC infrastructure general requirements

This section is dedicated to the description of HPC datacentre infrastructure general requirement with a focus on power efficiency and performance constraints which are the main constraints of an Exascale system.

2.1 HPC compute system packaging

Datacentre compute system is characterized by a set of components that is enclosed within a packaging where compute modules containing the processing nodes are the core components of this packaging. To work properly, compute modules need a set of other associated components like:

- Power Supply Units (PSU)[1] in charge of delivering Direct Current (DC) power to compute modules,
- Routing modules containing switch chips ,
- Backplane high-speed cables to provide compute modules interconnection,
- Fans or liquid cooling modules (depending on the implemented cooling solution).

Compute modules are contained within a rack (or cabinet) enclosures. The system consists of one or more rack enclosures with the necessary cables connecting the router ports according to the network topology.

Moreover, the network cables may aggregate multiple network links into a single cable to reduce both cost and cable bulk.

Furthermore, datacentre component packaging determines the notion of density which is often measured in kW per rack. Currently, a high density datacentre is one where each rack consumes more than 10 kW. An alternative measure of datacentre density is the amount of power consumed per square foot of floor space, which is typically expressed in Watt per

[1] PSU can be embedded in each compute module or centralized by providing power to several modules. Centralized PSU solution is generally preferred because of its relatively lower cost.

square foot. In general, a density of more than 150 W per square foot (about 1.5kW per m²) is considered high density[2].

Datacentre density depends, most of the time, on client constraints regarding available floor surface. It is also closely related to chosen solutions for power delivery and cooling system.

2.2 Power system

The power distribution and cooling accounts for about 25% of the total datacentre cost [4].

They must be designed to accommodate the worst-case power consumption at 100% utilization (i.e. running the Linpack benchmark). In practice, however, a large cluster system rarely operates at full utilization (50-60% on average).

Furthermore, efficient packaging, power, and cooling have a large impact on both the capital and operating cost of the cluster.

2.2.1 Power distribution

The important increase in power density, resulting from expected evolution to Exascale computing (20-25 MW), requires to provide enhancements to power distribution network with the aim to potentially decrease network losses, maintenance costs and improve the efficiency of energy utilization (see §2.3.1).

A utility typically delivers power across transmission High Voltage (HV) lines using 110kV (or above) to reduce energy loss across long distances. The incoming transmission lines are stepped down at datacenter level to Medium Voltage (MV) power lines usually ranging from 6 to 20 kV. Distribution transformers are generally used to step down the voltage from MV to Low Voltage (LV: 230V Alternating Current (AC) one phase or 400V AC three phases) voltages which are brought into Uninterruptible Power Supplies (UPS) (see §2.2.2) whose output circuits are connected to a distribution board in the server room. The energy from the server room distribution boards is then supplied to the server racks, equipped with their own distribution equipment called Power Distribution Unit (PDU). The receiving device, such as data processing equipment, data storage systems and network devices are powered directly from the PDU.

2.2.2 Uninterruptible power

Uninterruptible power must be supplied for some components in the datacentre to ensure their continuous operation or to reduce risks of damages to fragile ones. Uninterrupted operation of critical services, such as network and storage devices, critical data processing systems or a set of computing machines, is one of the prerequisites to ensure the quality of services provided by these objects. Achieving a reliable and uninterrupted power supply for critical facilities can be accomplished by making the power supplies or the power sources redundant or by installing Uninterruptible Power Supplies (UPS) which can take the form of batteries permitting to resist to more or less long power outages (depending on batteries capacity).

2.2.3 Power distribution enhancement

Within the described power network, the first element that generates considerable power losses in the distribution network is the Medium Voltage to Low Voltage (MV/LV) transformer. This equipment generates no-load losses caused by constant magnetization and demagnetization of the core and additional losses when the transformer operates (load losses), caused by current flow through the windings. Reducing the losses generated by the distribution transformers can be achieved by the use of better materials in the creation of a transformer core and for the copper windings.

[2] This takes into account that empty floor space in the front and the back of a rack is needed.

Other devices that generate high losses in the datacentre power distribution network are static UPS devices which performs double energy conversion: Alternating Current (AC) is straightened by the rectifier and passed to the Direct Current (DC) internal power rail. The battery (or capacitor) is connected to the DC power rail and forms an energy storage. The energy from the DC internal rail is then converted back to AC using an inverter, which provides voltage matching the one required by the devices in the computing centre.

Each of these conversions (AC to DC and DC to AC) causes power losses, which in large Static UPS devices reaches about 10 percent. Such loss can be reduced by using UPS only for critical components.

One other possible optimization of the power supply distribution consists in using an HVDC (High Voltage Direct Current bus), in place of the traditional AC line feed. The principle is to replace the standard AC input (either 230V AC one phase, or 400V AC three phases) by a direct current input (230-400V DC). The expected benefits are mainly regarding the overall efficiency of the power distribution (including distribution losses in cables, power supply efficiency), ease of adding a UPS, and the possibility to connect to an auxiliary source (like a local wind turbine or solar cell supply) made simpler.

Using ultra-capacitors power supplies at rack level instead of UPS to handle micro power outages is an interesting alternative in terms of cost. However they should be used when input power is considered as sufficiently stable with outages lasting less than 800ms.

2.3 Cooling system

Cooling systems must evacuate the heat generated by the processor sockets, DRAM, and networking equipment.

Heat removal can be done via convection (blowing air across the hot components). Fans in each rack are used to blow air across the component in combination with a heat sink (or heat spreader) to increase the surface area of the component, thereby improving its cooling efficiency.

However, the liquid cooling method is more and more used in HPC datacentre. In fact, this cooling method is less influenced by the ambient temperature and it permits to remove fan consumption and noise. A water-cooled system uses pipes, pumps, solenoid valves, etc in the rack to circulate coolant through the system. Water is a common coolant used for such applications. But other more efficient coolants are commonly used like polyethylene glycol or fluorinert[3] [5].

2.3.1 Power efficiency metrics

One of the most commonly used power metrics is the Power Usage Efficiency (PUE) of a datacentre. It is defined by the Green Grid consortium [6] as the ratio of total amount of energy used by a computer datacentre facility to the energy delivered to computing equipment. The closer a PUE is to 1, the more the datacentre is power efficient.

PUE can be calculated from:

$$PUE = \frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy} = 1 + \frac{Non\ IT\ Facility\ Energy}{IT\ Equipment\ Energy}$$

In the case of liquid-cooled HPC infrastructure, one way to decrease PUE is by increasing the energy dissipated by water with respect to the total rack dissipation. This could be achieved by adding cold plates or heat exchangers wherever it is necessary (CPU, memory, switches, PSU).

[3] non-flammable and with less environmental impacts than polyethylene glycol

First liquid cooling implementations used using mechanical water chillers to generate cold liquid (T~6°C) leading to a relatively high PUE (higher than 1.3).

Within more optimized cooling implementations, PUE could be improved by using warm liquid (T~40-45°C as input temperature) jointly with a free cooling approach.

Free cooling approach consists in lowering the liquid temperature in the datacentre by using ambient outdoor cool air (cooling towers) instead of mechanical refrigeration. With free cooling, PUE could be significantly decreased.

When input temperature is higher than 45°C, PUE could also be improved by using adsorption chiller (instead of mechanical chiller) which can produce cold water from the heat present in the warm water loop (this kind of cooling method is still under investigation and needs improvement).

This cold water may be reused in liquid-cooled doors, added in air-cooled systems in order to reduce heat propagation in the data centre and air conditioning cost. Indeed investment in liquid cooling can be done in compute part of the cluster, but not in the storage and administration part.

Within implementation of cooling solution reusing the heat produced by datacentre, it is important to take into account the positive effect of the energy reuse term on the overall energy budget of a site (laboratory, campus, research centre…) which accommodates the datacentre. Thus, other metrics have been developed like the ERE (Energy Reuse Efficiency) by the Green Grid consortium [7]. ERE is defined so that it is lower than the PUE when there is a benefit from the heat reuse for the global site energy consumption:

$$ERE = \frac{(Cooling + Power + Lighting + Misc. + IT - Reuse)}{IT}$$

The ERE can be lower than 1.0 if the Reuse term is sufficiently large.

2.3.2 Boards cooling solutions

To cool the CPU board with the coolant, it is necessary to build a mechanical heat transfer chain between these two elements. In order to have an efficient thermal chain, it is necessary to create a strong proximity between the coolant and the CPUs.

Since it is necessary to be able to remove the electronic board from the rack, it is necessary to have a system which makes possible to simply connect and disconnect the thermal chain.

The following table (Table 1) classifies three possible compute blades cooling solutions according to CPU-coolant proximity level.

Solution	CPU-Coolant proximity (performance)	Description	Pros	Cons
Immersion	++++	The boards are immersed in a diphasic liquid bath. This liquid is cooled by the customer network via an exchanger	*Optimal performance *No mechanical constraints for integration (any board could be used) *Reduced noise	*Density: difficulty to stack bathes *Installation constraints: bath filling, bath weight

Solution	CPU-Coolant proximity (performance)	Description	Pros	Cons
Soft cold plate	+++	The heat transfer occurs between CPU and a cold plate (containing the coolant) directly, without heat spreader	*Performance quite similar to "Immersion" solution	*Complex packaging with multi-socket boards (cold plate should be modular) * non suitable for DIMM cooling
Rigid cold plate	++	Cooling board is ensured by a cold plate which is crossed by the coolant. Unlike the "soft cold plate" solution, a heat spreader connects (mechanically and thermally) the CPU and the cold plate.	*Good performance *Provide good quality-price ratio *Easy blade disassembly and component replacement	* Potential risk of leakage due to the use of rapid coupler for hydraulic connections

Table 1. Liquid cooling solutions classification

2.4 Interconnect network

The interconnect network topology plays a central role in both the performance and cost of the network. It also determines some of the packaging and cabling requirements as well as fault resilience. In the following paragraphs we try to highlight the key points related to interconnect networks.

2.4.1 Network topologies

Network topologies can be divided into two different types: direct and indirect [8].
A direct network has processing nodes attached directly to the switching fabric; that is, the switching fabric is distributed among the processing nodes. An indirect network has the endpoint network independent of the endpoints themselves (dedicated switch nodes exist and packets are forwarded indirectly through these switch nodes) [9].
Examples of direct network include mesh, torus, and hypercubes as well as high-radix topologies such as the flattened butterfly. Indirect networks include conventional butterfly topology and fat-tree topologies (Figure 1).

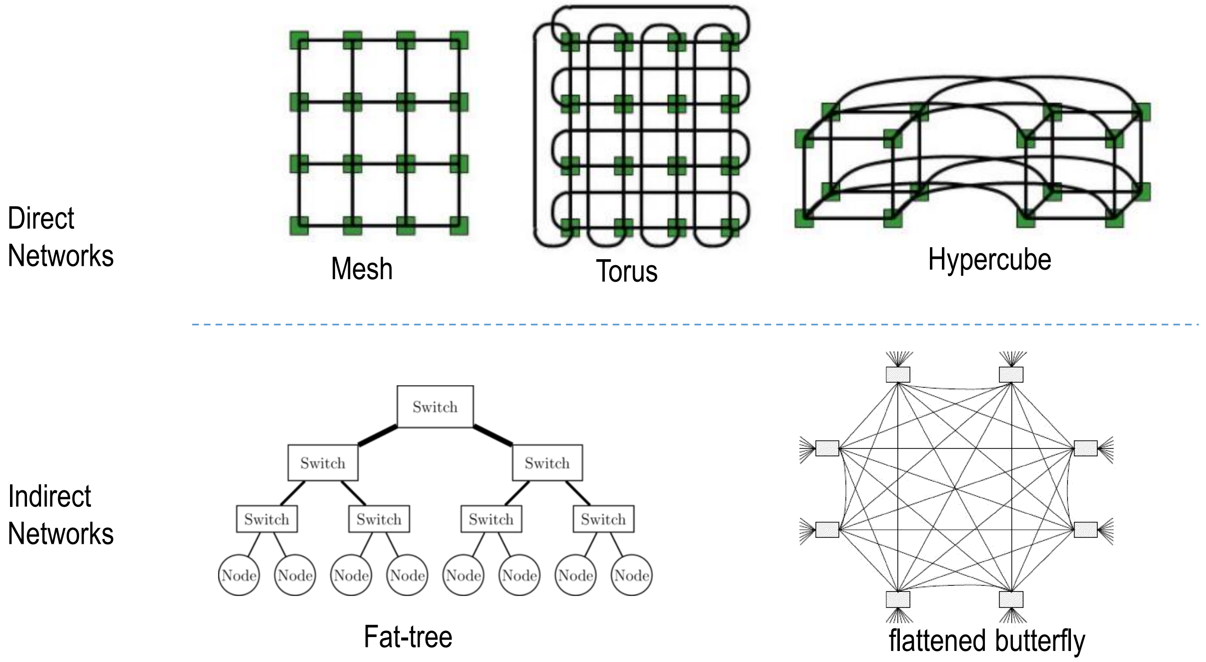


Figure 1. Network types and topologies

Choosing a network topology type depends essentially on performance (latency and bandwidth) and on power consumption.

Generally, direct networks are more efficient than indirect networks in terms of power consumption because they don’t use switches, however they are less efficient regarding latency and bandwidth (see Table 2).

In the context of this deliverable we will focus on direct topologies and particularly mesh and torus network topologies which don’t need dedicated switch nodes. Thus, these topologies should be the most suitable to use with Unimem [10]interconnect solution.

These topologies are often referred to as k-ary n-mesh or k-ary n-cube. The scalability of the network is largely determined by the radix (number of nodes within a dimension) k, and number of dimensions n, with N = kn total endpoints in the network. In practice, the radix of the network is not necessarily the same for every dimension (irregular mesh or torus). Therefore, a more general way to express the total number of endpoints is given by Equation:

$$N = \prod_{i=0}^{n-1} k_i$$

The *worst-case* distance (measured in hops) that a packet must traverse between any source and any destination is called the *diameter* of the network. The network diameter is an important metric as it bounds the worst-case latency in the network. Since each hop entails an arbitration stage to choose the appropriate output port, reducing the network diameter will, in general, reduce the *variance* in observed packet latency. The network diameter is independent of traffic pattern, and is entirely a function of the topology, as shown in Table 2.

Moreover, another important metrics representing the network characteristics are:

- The bisection bandwidth which is the smallest bandwidth between half of the node to another half of the nodes,
- The nodal degree which is the number of connections the node has to other nodes.

Network family	**Network type**	**Diameter (hops)**	**Nodal degree**	**Bisection bandwidth (in units of link bandwidth)**
Direct network	nD-Mesh	$n(N^{1/n} - 1)$	2n	$N^{(n-1)/n}$
	nD-Torus	$nN^{1/n} / 2$	2n	$2N^{(n-1)/n}$
Indirect network	Flattened butterfly (k-by-k switches)	$\log_k N$	k	N/2
	k-ary tree	$2\log_k N$	k+1	1

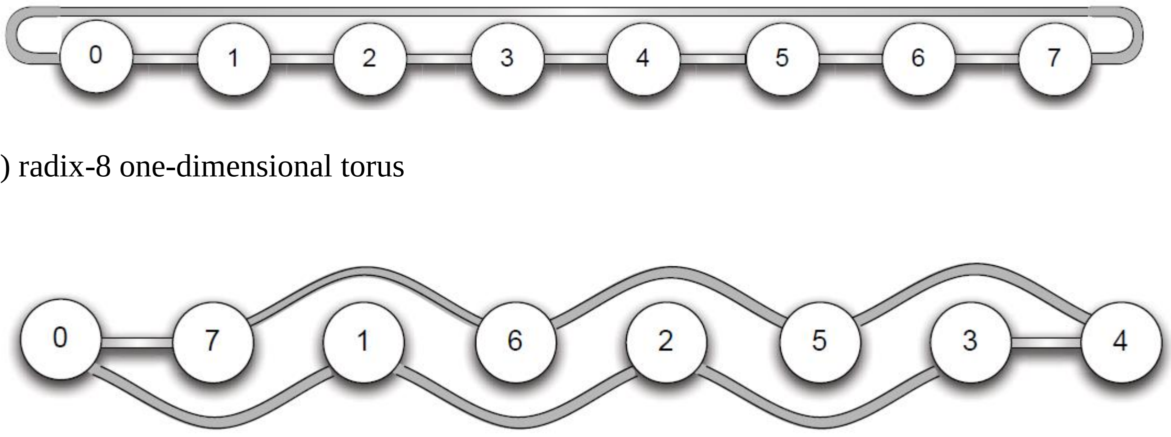
Table 2. Network diameter, nodal degree and bisection bandwidth (n is dimension, N is total number of nodes)

2.4.2 Network packaging

One often overlooked property of a network is how a given topology maps to physical packaging.

For example, a torus or mesh network which connect to their neighboring nodes makes most links very short (see Figure 2(a)). The wraparound links in a torus can be made shorter by cabling the system as a folded torus as shown in (see Figure 2 (b)). Mesh and torus networks have several packaging advantages [9]:

- a portion of one dimension can be implemented on the printed circuit board (PCB) by connecting the adjacent nodes on the same board with PCB trace,
- a portion of one dimension can be implemented within a PCB or cable backplane to connect adjacent nodes within the same rack enclosure,
- cabling the mesh or torus is very regular,
- it requires relatively short cables which can operate at high signal rates and generally have a cost advantage over longer cables, and
- it requires only a small number of different cable lengths.



(a) radix-8 one-dimensional torus

(b) Folded torus implementation

Figure 2. Decreasing the longest cable length in a torus (a) by "folding" it (b).

2.5 Storage system

2.5.1 Background on future storage architectures

Attaching a high-performance storage subsystem will be a major challenge for future exascale systems. On the one hand, the performance of the storage systems have to improve significantly as the compute performance increases and emerging data-intensive applications result in more stringent I/O performance requirements. On the other hand, the number of compute nodes with clients accessing the storage subsystem is expected to increase significantly and the compute architecture is likely to change, e.g. due to the use of a large number of relatively weak CPU cores.

In this section, the assumption is made that future online storage subsystems need to be organised in a hierarchical manner in order to cope both with capacity as well as performance requirements.[4] The reasons for this development are the roadmaps of storage technologies. Today's online storage systems are mainly based on HDDs. This technology continues to improve significantly in terms of capacity, however improvements in terms of bandwidth are moderate and in terms of I/O access rates negligible. Increasing bandwidth thus requires increasing the number of disks up to the point that capacity exceeds the needs. The emerging alternative of using storage technologies based on non-volatile memory technologies, e.g. SSDs, is affected by the limited endurance of the currently predominant NAND Flash technology. Only high quality devices meet the HPC requirements resulting in a high cost versus capacity ratio. These devices do allow to realise storage tiers featuring high bandwidth and high access rates, but costs limit the affordable capacity.

While future storage architectures might comprise multiple storage tiers[5], we consider in the following two tiers:

- A *Large Capacity Storage Tier* (LCST) is a storage tier that is optimised for primarily for capacity;
- A *High Performance Storage Tier* (HPST) is a storage tier that is optimised for performance (both bandwidth as well as high access rates).[6]

Such architecture can be characterised by the following parameters (Table 3).

B_{HPST}	Bi-section bandwidth between compute system and the High Performance Storage Tier
C_{HPST}	Capacity of the High Performance Storage Tier
B_{LCST}	Bi-section bandwidth between compute system and the Large Capacity Storage Tier
C_{LCST}	Capacity of the Large Capacity Storage Tier

Table 3. Two-tier storage subsystem parameters

In future, non-volatile memory will likely be integrated also in the compute system. This is, in particular, of interest for facilitating check-pointing. This part is not considered here.

[4] A tape band storage backend, which can be considered as an additional, but offline storage tier, is not considered here.

[5] For instance, the SAGE project is working towards storage architectures with a larger number of tiers.

[6] Currently the term "Burst Buffer" is very popular when referring to such a fast storage tier. However, the concept of a burst buffer is narrower than that of a HPST.

2.5.2 Exascale requirements

Exascale I/O requirements depend both on the need for reading and writing data as well as writing check-point dumps. For this reason we consider for the HPST a capacity that is a small multiple of the computing systems main memory capacity. This would allow to hold at least 2 full memory check-points and have still space available for staging data for reading and buffer data for writing.

For these parameters we make in the following an attempt to perform an extrapolation towards exascale. For the HPST we assume that the capacity should be a small multiple of the computing systems main memory capacity. Assuming a factor[7] of 4 and a moderate memory capacity for an exascale system of CMEM = 10 PByte the capacity of the HPST would be about C_{HPST} = 40 PBytes. Allowing for a full dump of the main memory to take 4-5 minutes results in the following estimate of the bandwidth: B_{HPST} = 40 TBytes/s.

The capacity of the storage systems attached to today's PRACE Tier-0 systems providing a computing capability of several PFlop/s is about 20 PByte. With performance capabilities of the computing system increasing roughly by a factor 200, the capacity of the storage system should increase by at least a factor 20, i.e. the target LCST capacity is C_{HPST} = 400 PByte. We assume that the bandwidth towards the LCST could be an order of magnitude smaller compared to the HPST, i.e. we expect B_{LCST} = 4 TByte/s.

The following table (Table 4) compares the performance numbers assumed above with the target performance number of the planned pre-exascale system NERSC-9, which should be realised towards the end of 2020 [11]:

	NERSC-9	**This work**
B_{fp}	150-300 PFlop/s	1 EFlop/s
C_{mem}	>3 PByte	10 PByte
B_{HPST}	>5 TByte/s	40 TByte/s
C_{HPST}	>90 PByte	40 PByte
B_{LCST}	1 TByte/s	4 TByte/s
C_{LCST}	50 PByte	400 PByte

Table 4. Compared performance numbers between NERSC-9 and targeted exaflopic system

Today different approaches are used for realising the interconnect between a compute and a storage system. As the costs for a dedicated storage network is often discarded for costs reason, typically the compute nodes access the storage subsystem through the network used for interconnecting the compute nodes. Storage nodes are either directly attached to this network or accessible through gateway nodes or switches. The latter has the disadvantage that such an architecture typically do not allow for RDMA-based communication. The latter is important for maximising performance, in particular in terms of throughput of I/O operations. Today's storage subsystems are attached using Infiniband or Ethernet.

They are significant uncertainties concerning the software stack used for accessing this storage. On today's supercomputers external storage resources are accessed through a parallel file system through a POSIX compliant interface. While this meets the requirements of many

[7] When using HPST for check-point, at least 2x the memory capacity is needed to overwriting a previous memory dump. Another factor 2 is to allow for having a I/O read and write buffer of the same size of the memory

applications or higher-level I/O middleware components (e.g., MPI-IO), there are known scalability issues. This problem can be mitigated by shipping I/O calls to a smaller set of I/O nodes. However, this does not address another limitation of POSIX compliant interfaces: These typically do not allow to fully exploit the performance of modern storage devices based on non-volatile memory technologies. At exascale the role of today's dominating technologies like Lustre and GPFS might change.

Check-pointing speed may be even more improved by making non-volatile memory available at node level, which would allow for higher aggregate I/O bandwidth. Solutions like the Scalable Checkpoint / Restart (SCR) Library can take benefit of such memory as they support multi-level check-pointing [12].

2.6 Visualisation subsystem

Integration of visualisation is becoming increasingly important due to the demands for visualising larger and larger data-sets as well as to monitor large-scale simulations.

The classical approach where data is first written to disk and later processed on a dedicated visualisation facility does often not meet the requirements anymore. Alternative approaches could consist of:

- In-situ visualisation: a significant part of the visualisation pipeline is co-located on the compute node. In this case, part of the available resources has to be reserved for this purpose, which typically leads to higher node-level memory capacity requirements. Furthermore, typically local rendering capabilities are required, which results in the need of relevant software stacks like OpenGL to be supported (with or without hardware support).
- In-transit visualisation: data is moved through the interconnect network from the compute nodes to dedicated visualisation nodes (with GPU for example) without hitting the external storage system. Non-volatile memory integrated into the node as well as an HPST can serve as fast intermediate buffers of the data. This approach leads to high network performance requirements.

Such visualisation sub-systems do not add specific requirements to the compute infrastructure. As a consequence, there is no specific study with respect to ExaNoDe infrastructure in the framework of this deliverable.

2.7 Infrastructure management and monitoring

2.7.1 Key infrastructure equipment monitoring

High performance computing (HPC) facilities contain a large number and various types of equipment. With each type of equipment, there is a different need when it comes to metering. The table below (Table 5) shows a generic breakdown of the various systems found within an HPC facility with liquid cooling system [13]. The table also lists the potential components within each system and the key measurements to be obtained. The measurements for each system can be used to calculate various performance metrics for the system and the facility.

System	Components	Measurements
General measurements		Indoor/Outdoor temperature, Indoor/Outdoor relative humidity
Server/Storage/Networking	Internal fans	Current, Voltage, Power, Temperature
UPS/PDU		Current, Voltage, Power
Transformers		Current, Voltage

Datacentre air conditioners	Compressors, fans, pumps, …	Temperature, Flow rate, Current, Voltage, Power
Chillers	Compressors, heat exchangers	Temperature, Flow rate, Current, Voltage, Power
Cooling towers	Fans, pumps	Temperature, Flow rate, Current, Voltage, Power, Pressure
Heat exchangers		Temperature, Flow rate
Lighting		Current, Voltage

Table 5. Key equipment measurements

2.7.2 Rack management

HPC datacentre management and monitoring system permits various equipment controlling and monitoring using dedicated management nodes usually organized in a hierarchical network to provide highly scalable management architecture.
At rack-level, management is performed thanks to several management controllers located in various spots of the rack and controlling power, cooling, switches and compute nodes.
These management controllers are connected through an embedded Ethernet network and could be additionally connected by an alternative sideband network dedicated to management controller maintenance.

The following figure (Figure 3) describes an example of a 2-layer management network controlling and monitoring datacentre racks.

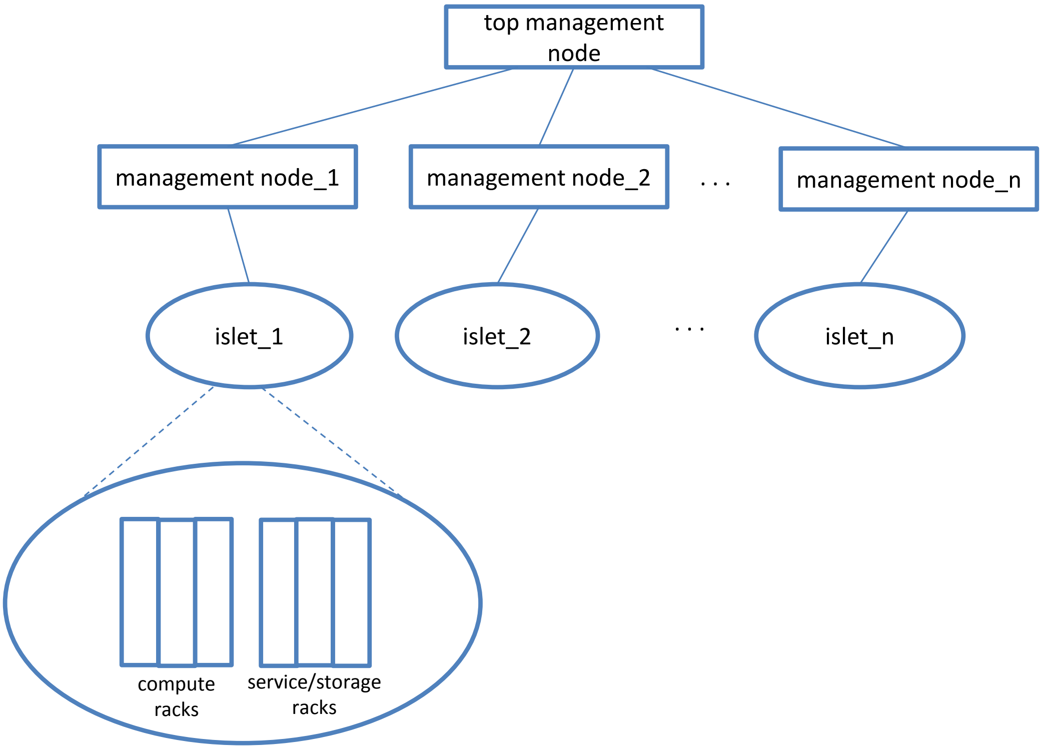


Figure 3. Rack management architecture

Management functions deal with:

- Hardware control: power on/off, emergency actions,…
- Data collection: data collected from sensors of temperature, voltage, power consumption,…
- Data processing and analysis: data aggregation, monitoring, reporting,…
- Application management and scheduling: automated installation and configuration of the nodes.

3 Proposed rack-scale solution for current ExaNoDe compute system

In this section we try to provide an example of implementation describing a rack-scale solution integrating ExaNoDe Multi Chip Module (MCM) in accordance with its current architectural specifications [2].

This proposed example is a long-term vision that may have to evolve according to modifications that could subsequently impact MCM daughter board, and relatively to the results of the future Unimem system evaluation.

In addition, the main objective of this approach is to demonstrate the possibility to package ExaNoDe compute system in a rack designed for large-scale HPC infrastructure and based on Bull Sequana solution [3].

3.1 Compute blade

As currently proposed ExaNoDe daughter board [2] is designed to be integrated with ExaNeSt cabinet [14], this board is not compliant with Bull Sequana cooling solution.

In fact, Bull Sequana compute blade is a 1U-blade cooled with a rigid cold plate witch is a board cooling solution providing good quality-price ratio and permitting easy blade disassembly and component replacement (see §2.3.2). Therefore, we propose an alternative integration solution which described below.

3.1.1 Compute blade description

We suggest in the framework of this example, an alternative physical design for daughter board. This suggested alternative design keeps the same current logical daughter board design while taking into account the compute blade dimensions (600x500mm) and the blade cooling method (cold plate).

The proposed modifications (Figure 4) are summarized as following :

- 4x2 MCMs are integrated in one Compute Board (CB), where each 2 MCM pattern is logically equivalent to current ExaNoDe daughter board,
- Instead of using SODIMMs, we propose to use regular DIMMs which are more compliant with rigid liquid-cooled cold plate used in Bull Sequana,
- M.2 SDD are inserted horizontally instead of vertically (for thinner blades),
- Current High-Speed (HS) connector is removed and GTH links are directly routed toward compute blade connector in order to improve signal attenuation and integrity.

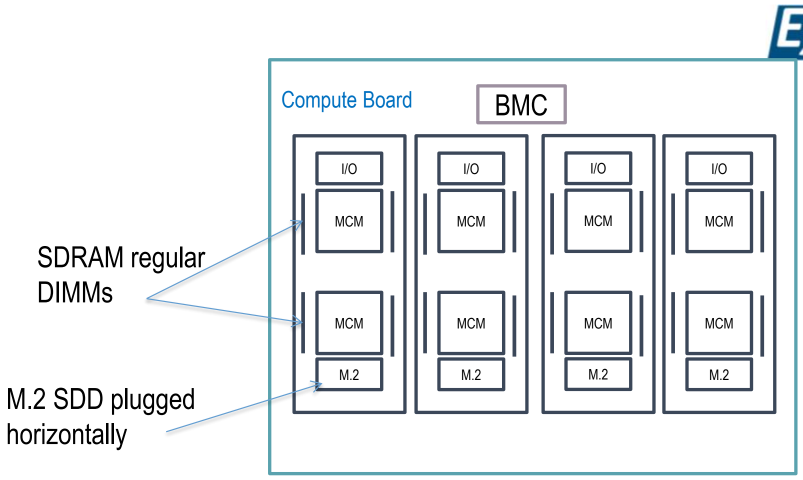


Figure 4. MCM Compute Board (CB)

Each compute board is managed by a Base Board Management Controller (BMC) controlling a set of 8 MCM and their associated I/O FPGAs, SDRAM Modules and M.2 SDD.
In addition, proposed compute blade will contain two compute boards (Figure 5).

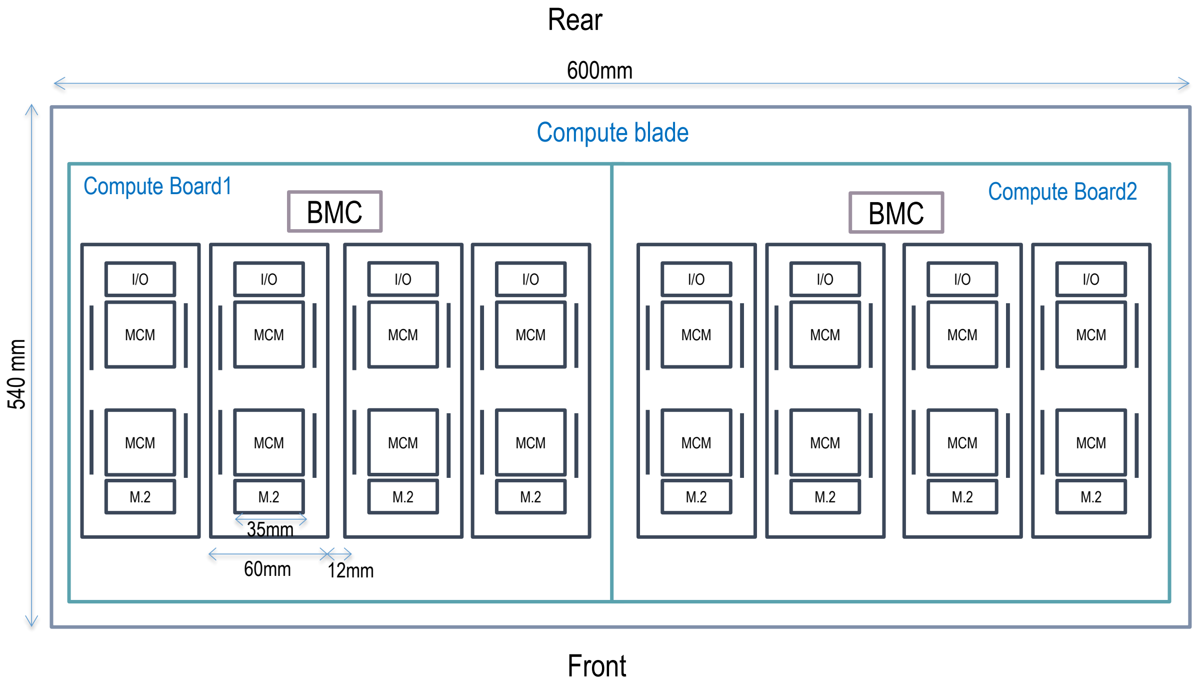


Figure 5. MCM Compute Blade

3.1.2 Compute blade internal links

MCMs inside a Compute blade are interconnected with High-Speed (HS) (10~16Gbps) links. The interconnection ring consists in a pattern of one dimension of 3D-torus interconnect topology described in §2.4.1 (Figure 6).
Moreover, this ring is designed according to the folded torus implementation (as presented in §2.4.2).

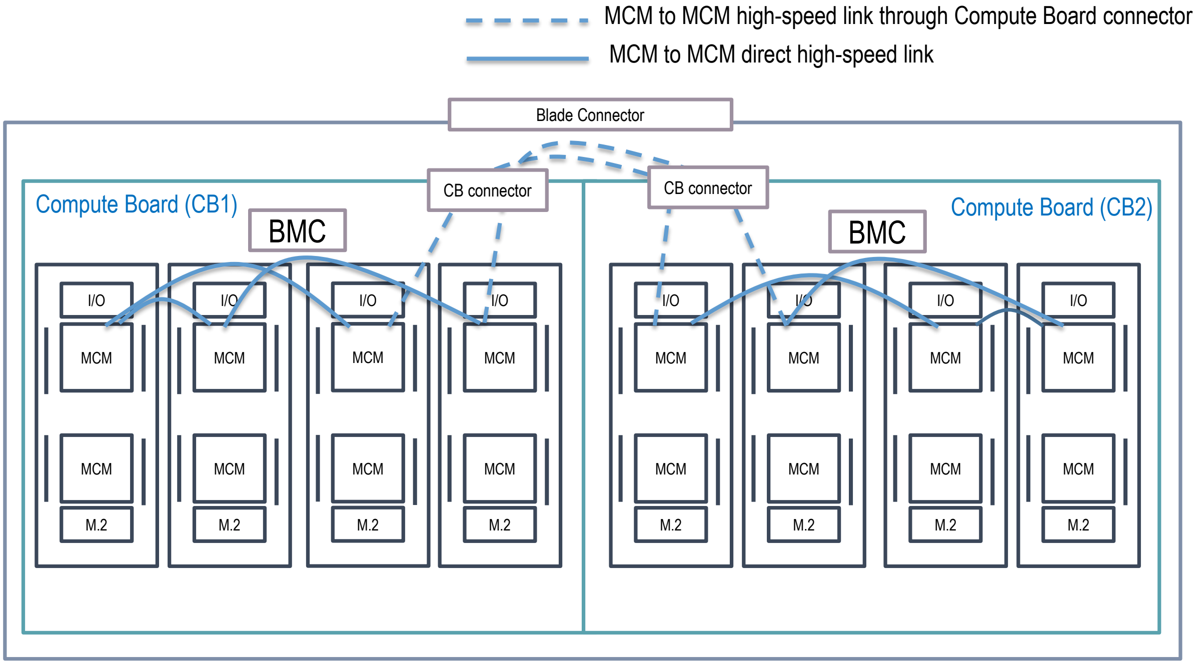


Figure 6. Compute Blade

3.1.3 Compute blade external links

Proposed compute blade exposes through its connector (Figure 6) 64 bidirectional high-speed links[8]. Given that each high-speed link bandwidth is ranging from 10 to 16 Gbps, total compute blade bandwidth will range from 640 to 1024 Gpbs.

Compute blade connector will also provide 10 Ethernet management links: 8 links coming from I/O FPGAs and 1 link from each compute board BMC.

Finally, 2 alternative management links (Sideband signals) dedicated to BMC powering and Ethernet connectivity fixing, are provided.

3.1.4 Compute blade cooling

Rack cooling system is performed by hydraulic modules located in the bottom of the rack and containing as main components: a heat exchanger, a pump and regulation valves. Hydraulic modules main function is to maintain rack internal hydraulic circuit temperature to a fixed regulation temperature.

Cooling is based on Bull Sequana compute blade cooling by direct contact with Direct Liquid Cooling (DLC) cold plate (Figure 7).

Heat spreaders are used to dissipate heat from CPU sockets and hydraulic connectors connects the cold plate to the rack internal hydraulic circuit.

[8] Each MCM has 10 GTH links where 2 links are used in the internal ring connecting MCMs inside compute blade

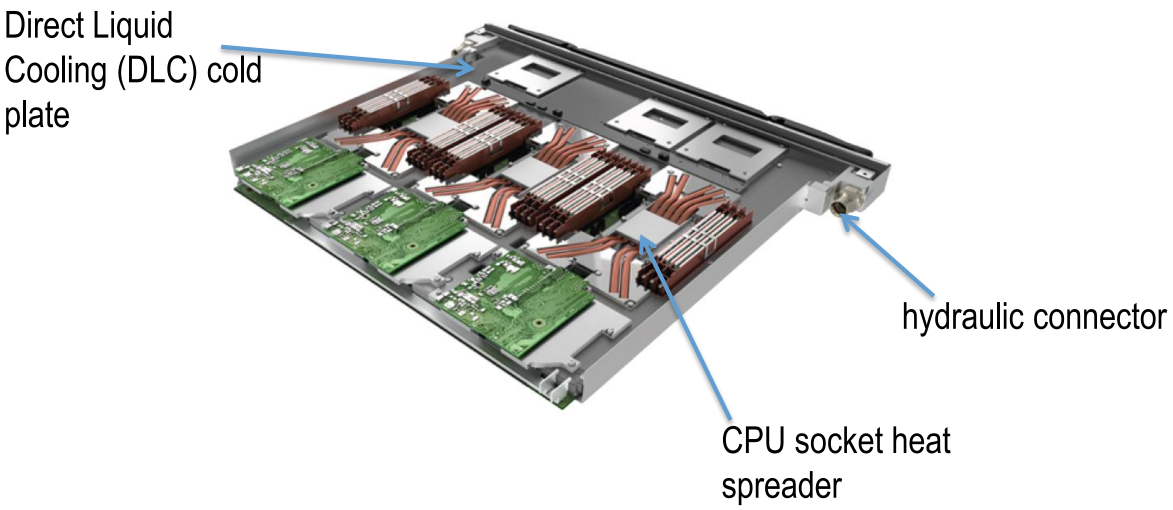


Figure 7. Cooling components of Bull Sequana X1210 compute blade

3.2 Interconnect network topology

The suggested interconnect network topology at rack level takes into consideration the proposed packaging for the compute blade (refer to §3.1.1), as well as the number of rack slots dedicated to compute blades (a maximum of 18 compute blade slots at each rack side).

Thereby, the proposed interconnect network topology is a direct network based on an irregular 3D-torus containing N=8x6x6=288 endpoints packaged in 36 compute blades.

Each network endpoint is a pattern composed of 2 MCMs and their associated I/O FPGA, SDRAM Modules and M.2 SDD (Figure 8).

The first dimension of the 3D-torus network contains 8 endpoints and corresponds to a compute blade.

Rack interconnections are not considered in the scope of this proposal which focuses on rack-scale solution, however 4 free high-speed links per network node could possibly be used[9] within a 4th torus dimension dedicated to rack interconnections.

[9] Only 6 GTH links among 10 available links per network node are used in the proposed 3D-torus network

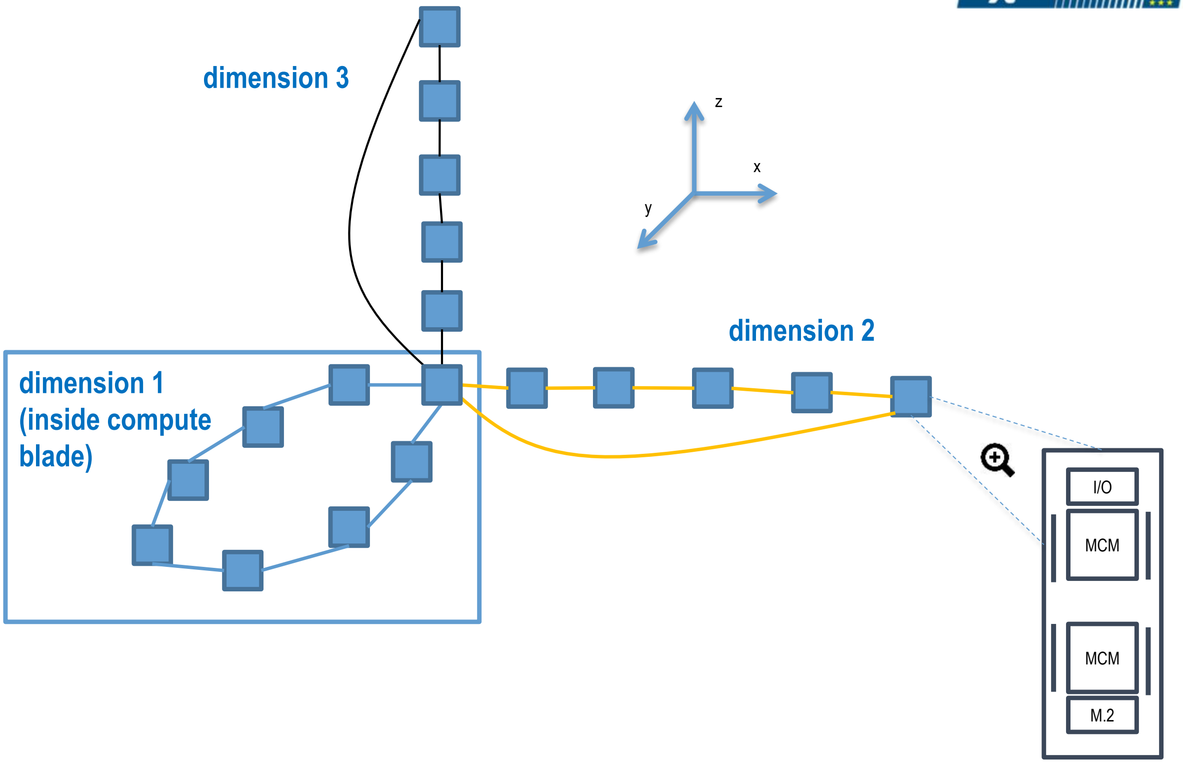


Figure 8. Rack-scale interconnect network topology (3D-Torus, N=288)

3.3 Rack management network

3.3.1 Management network components

Rack management network is composed of 2 types of Ethernet switches (Figure 9).

- 18 Leaf Ethernet switches (LSW): 9 at each rack side, each LSW is connected to 2 compute blades (4 BMC)
- 2 Top Ethernet switches (TSW): one at each rack side. TSW#1 is connected to 6 LSW and TSW#2 is connected to 12 LSW

Management network also contains a set of management controllers:

- 2 Top switch Management Controllers (TMC), each one managing each 1 TSW
- 6 Leaf switch Management Controllers (LMC), each one managing 3 LSW
- 1 Power Management Controller (PMC) managing Power Supply Unit (PSU) modules
- 3 Cooling Management Controller (CMC), each one controlling 3 hydraulic modules
- 72 Baseboard Management Controller (BMC), each one managing a compute board

3.3.2 Ethernet network

Rack management network is a tree based on 2 layers of Ethernet switches composed, for the first layer, of Top Ethernet switches (TSW) and, for the second layer, of Leaf Ethernet switches (LSW).

- 18 Leaf Ethernet switches (LSW) : 9 at each rack side, each LSW is connected to 2 compute blades (4 BMC)
- 2 Top Ethernet switches (TSW) :
  - TSW#1 (at rear side) is connected to 6 LSW, 2 LMC, 1TMC, 1PMC, 3CMC
  - TSW#2 (at front side) is connected to 12 LSW, 4 LMC and 1 TMC

LSW and TSW are embedded in switch blades. LSW blade contains 3 LSW and 1 LMC, where TSW blade contains 1 TSW and 1 TMC (Figure 9).

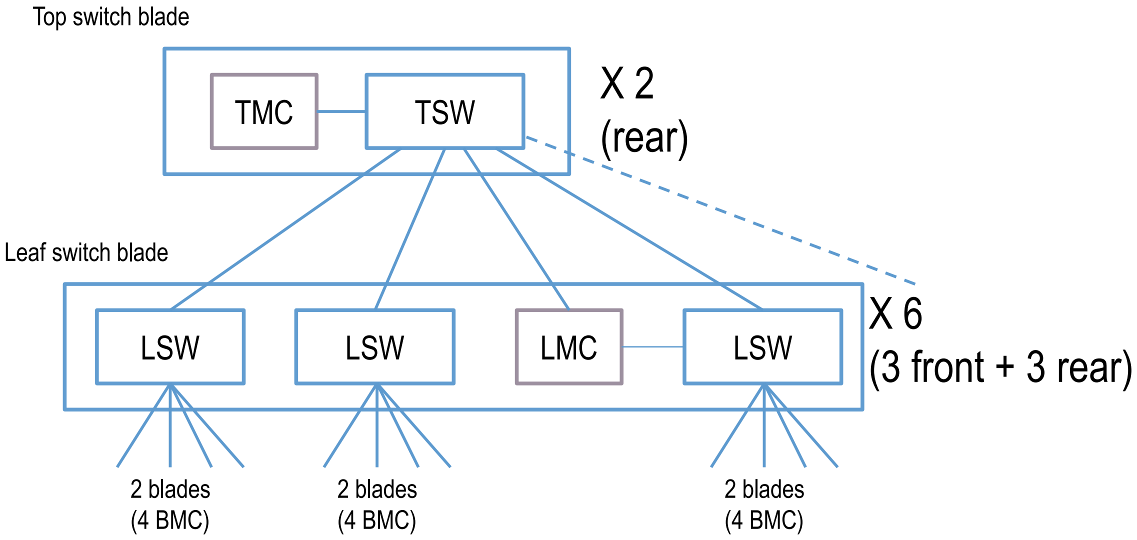


Figure 9. Ethernet Management Network

3.4 Storage subsystem

Based on the analysis described in §2.5.2 we derive the following requirements for connecting ExaNoDe nodes to the storage subsystem:

- To reach the target aggregate I/O bandwidth, the per node I/O bandwidth towards the HPST, b_{HPST}, must be larger than (b_{fp} / B_{fp}) BHPST, where b_{fp} is the compute performance per node. To allow for a smaller number of nodes to saturate this bandwidth, b_{fp} should be an order of magnitude larger. Assuming b_{fp} = 2.4 TFlop/s we thus arrive at the following requirement: $b_{\mathrm{HPST}} \geq 1$ GByte/s.
- To facilitate high throughput of I/O operations and efficient exploitation of the storage network, communication between the compute node and the HPST should be realised using a network technology supporting RDMA.
- Relevant I/O middleware software stacks must be supported by the ExaNoDe architecture. For the time-being this includes support for clients of relevant parallel file systems like GPFS, Lustre and BeeGFS.

3.5 Maximum power consumption estimation

An estimation of component power consumption is detailed in Table 6. It is based on the estimation of ExaNoDe daughter board maximum power consumption (about 150W). This estimation is used to dimension the power provision (peak consumption), not the energy efficiency.

Components	Max. Power consumption
36 x Compute blades (1 Compute blade = 16xMCM + 8xI/O FPGA + 8xM.2 SSD + 2xBMC)	~41kW
6 x Leaf switch blades (1 Leaf switch blade = 3xSwitch boards + 1xManagement Controller)	~800W

Components	Max. Power consumption
2 x Top switch blades (1 Top switch blade = 1xSwitch board + 1xManagement Controller)	~120W
Other Management controllers (1xPMC + 3xCMC)	~80W
Total	**~42kW**

Table 6. Maximum power consumption estimation

Theoretically, 3 liquid-cooled power shelves, each one providing 15kW, are sufficient to provide total rack maximum power capacity (42kW).

For ExaNoDe, we propose to support 2N power shelves redundancy (potentially with 2 independent power sources) by integrating 6 liquid-cooled power shelves in the rack.

Liquid cooling supporting rack outlet temperatures in the range of 40°C and 44°C is proposed for Exanode in order to allow using free cooling method (see $2.3) in most areas of Europe throughout the year.

This cooling will be ensured by 3 hydraulic modules (2 + 1 redundant) where each module could dissipate until 35kW.

Moreover, short power outage handling is ensured by a set of 6 ultracapacitor modules (15 kW capacity for each one).

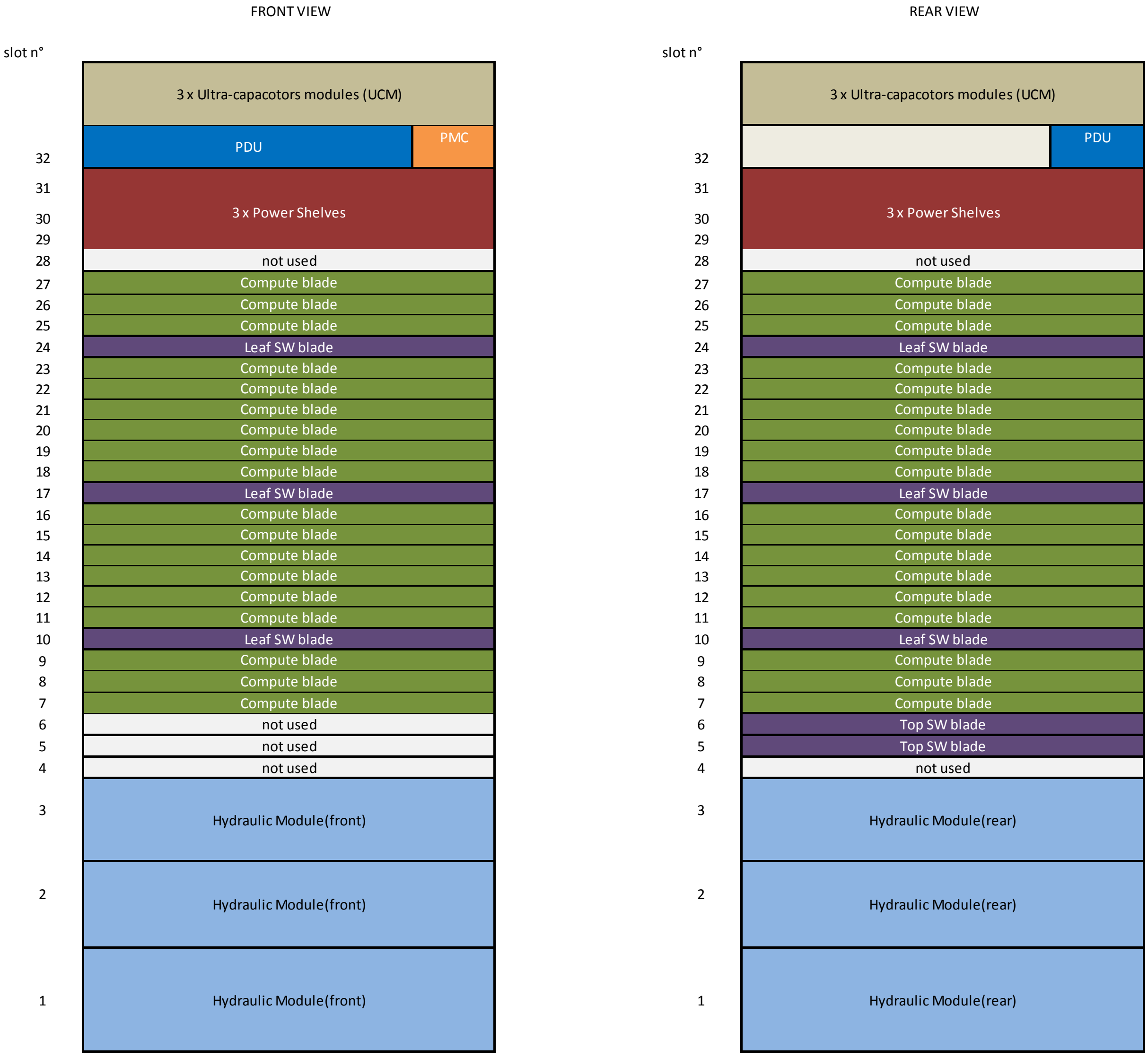
Note that power consumption of each pump contained in a hydraulic module is estimated to 1kW. However this pump power is not delivered by rack PSUs, it is provided by another datacentre power source.

3.6 Rack composition

According to the previously presented elements on compute blades packaging, interconnect network, cooling and management network, an example of rack composition is presented in Figure 10.

The rack is composed of:

- 6 Ultracapacitor modules: 3 modules located on top of the rack at each rack side
- 6 Liquid-cooled power shelves: 3 power shelves at each rack side
- 32 Compute blades: 16 blades at each rack side
- 6 Leaf switch blades: 3 blades at each rack side
- 2 Top switch blades at rear side
- 3 Hydraulic modules
- 1 rack Power Distribution Unit (PDU)
- 1 PMC module containing power management controller



slot n°	FRONT VIEW	REAR VIEW
	3 x Ultra-capacotors modules (UCM)	3 x Ultra-capacotors modules (UCM)
32	PDU PMC	PDU
31	3 x Power Shelves	3 x Power Shelves
30		
29		
28	not used	not used
27	Compute blade	Compute blade
26	Compute blade	Compute blade
25	Compute blade	Compute blade
24	Leaf SW blade	Leaf SW blade
23	Compute blade	Compute blade
22	Compute blade	Compute blade
21	Compute blade	Compute blade
20	Compute blade	Compute blade
19	Compute blade	Compute blade
18	Compute blade	Compute blade
17	Leaf SW blade	Leaf SW blade
16	Compute blade	Compute blade
15	Compute blade	Compute blade
14	Compute blade	Compute blade
13	Compute blade	Compute blade
12	Compute blade	Compute blade
11	Compute blade	Compute blade
10	Leaf SW blade	Leaf SW blade
9	Compute blade	Compute blade
8	Compute blade	Compute blade
7	Compute blade	Compute blade
6	not used	Top SW blade
5	not used	Top SW blade
4	not used	not used
3	Hydraulic Module(front)	Hydraulic Module(rear)
2	Hydraulic Module(front)	Hydraulic Module(rear)
1	Hydraulic Module(front)	Hydraulic Module(rear)

Figure 10. Rack composition

3.7 Datacentre operating conditions

The operating conditions (air in the vicinity of the rack) are compliant with ASHRAE A2 class (version 2011) [15]:

- ambient air temperature between 10°C and 35°C,
- relative humidity between 20% and 80%,
- dew-point up to 21°C.

The temperature of all the rack walls must be at least at the ambient temperature.

In case the internal liquid temperature has gone down below 17°C after the rack has been powered off, a condensation risk may occur on the rack parts in contact with the liquid that can be below 17°C. Even if this is a seldom condition, it should be taken into account by the temperature regulation mechanism which can activate a preheating phase of the liquid if necessary, during the rack powering on procedure.

4 Conclusions

In this deliverable, we describe and analyse general datacentre infrastructure requirements for high performance computing, large-scale, high-density and power efficient systems with a focus on ExaNoDe compute system.

Several infrastructure aspects have been analysed: power, cooling, network, storage and management.

The main outcomes of our analysis are:

- The characteristics of a HPC power efficient infrastructure are described and some enhancement possibilities regarding cooling and storage are presented,
- A solution is proposed for the integration of ExaNoDe Multi Chip Modules (MCM) in a HPC infrastructure based on Bull Sequana design. An alternative physical design for daughter board was suggested. This design keeps the same current logical daughter board design while taking into account Bull Sequana compute blade specificities.
  Thus, a compute blade for ExaNoDe which is compliant with HPC infrastructure requirements is described and its internal and external interfaces were detailed according to the interconnect network topology.
- A recommended interconnect network solution based on 3D-torus is presented. It takes into consideration Unimem hardware constraints (switch-less).
- Requirement for connecting ExaNoDe nodes to a storage subsystem are described in terms of bandwidth, network, and software.

5 References

[1] M. Pospieszny et al. Electricity in HPC Centres. Partnership for Advanced Computing in Europe (PRACE). Available online at www.prace-ri.eu

[2] C. Pinto et al. “Paving the way towards a high energy-effcient and highly integrated compute node for the Exascale revolution: the ExaNoDe way”. 2016

[3] Bull Sequana X1000 Series commercial whitepaper. https://bull.com/wp-content/uploads/2016/08/f-sequana-en1_web_11.pdf

[4] Albert Greenberg, James Hamilton, David A. Maltz, and Parveen Patel. The cost of a cloud: research problems in data center networks. SIGCOMM Comput. Commun. Rev., 39(1):68–73, 2009.

[5] 3MCorporation. http://www.3m.com/product/information/Fluorinert-Electronic-Liquid.

[6] The Green Grid consortium. https://www.thegreengrid.org

[7] Energy Reuse Efficiency (ERE). https://eehpcwg.llnl.gov/documents/infra/06_energyreuseefficiencymetric.pdf

[8] W. J. Dally and B. Towles. Principles and Practices of Interconnection Networks. 2004.

[9] D. Abts and J. Kim. High Performance Datacenter Networks Architectures, Algorithms, and Opportunities. Synthesis Lectures On Computer Architecture #14. 2011.

[10] Euroserver project. http://www.euroserver-project.eu/

[11] E.C. Joseph et al. (IDC), “Analysis of the Characteristics and Development Trends of the Next-Generation of Supercomputers in Foreign Countries”, December 2016.

[12] A. Moody et al. “Design, Modeling, and Evaluation of a Scalable Multi-level Checkpointing System”. 2010.

[13] T. Wenning et al. “High Performance Computing Data Center Metering Protocol”. 2010

[14] ExaNeSt project. http://www.exanest.eu/

[15] ASHRAE TC 9.9. “Thermal Guidelines for Data Processing Environments”. 2011. www.tc99.ashraetcs.org.