Aggregating and Consolidating two High Performant Network Topologies: The ULHPC Experience

Sebastien Varrette
Hyacinthe Cartiaux
Teddy Valette
Abatcha Olloh
sebastien.varrette@uni.lu
hyacinthe.cartiaux@uni.lu
teddy.valette@uni.lu
abatcha.olloh@uni.lu
Faculty of Science, Technology and Medicine (FSTM),
University of Luxembourg
Esch-sur-Alzette, Luxembourg

ABSTRACT

High Performance Computing (HPC) encompasses advanced computation over parallel processing. The execution time of a given simulation depends upon many factors, such as the number of CPU/GPU cores, their utilisation factor and, of course, the interconnect performance, efficiency, and scalability. In practice, this last component and the associated topology remains the most significant differentiators between HPC systems and lesser performant systems. The University of Luxembourg operates since 2007 a large academic HPC facility which remains one of the reference implementation within the country and offers a cutting-edge research infrastructure to Luxembourg public research. The main high-bandwidth low-latency network of the operated facility relies on the dominant interconnect technology in the HPC market *i.e., Infiniband* (IB) over a Fat-tree topology. It is complemented by an Ethernet-based network defined for management tasks, external access and interactions with user's applications that do not support Infiniband natively. The recent acquisition of a new cutting-edge supercomputer *Aion* which was federated with the previous flagship cluster *Iris* was the occasion to aggregate and consolidate the two types of networks. This article depicts the architecture and the solutions designed to expand and consolidate the existing networks beyond their seminal capacity limits while keeping at best their Bisection bandwidth. At the IB level, and despite moving from a non-blocking configuration, the proposed approach defines a blocking topology maintaining the previous Fat-Tree height. The leaf connection capacity is more than tripled (moving from 216 to 672 end-points) while exhibiting very marginal penalties, *i.e.* less than 3% (resp. 0.3%) Read (resp. Write) bandwidth degradation against reference parallel I/O benchmarks, or a stable and sustainable point-to-point bandwidth efficiency among all possible pairs of nodes (measured above 95.45% for bi-directional streams). With regards the Ethernet network, a novel 2-layer topology aiming for improving the availability, maintainability and scalability of the interconnect is described. It was deployed together with consistent network VLANs and subnets enforcing strict security policies via ACLs defined on the layer 3, offering isolated and secure network environments. The implemented approaches are applicable to a broad range of HPC infrastructures and thus may help other HPC centres to consolidate their own interconnect stacks when designing or expanding their network infrastructures.

KEYWORDS

HPC Management, Network, Performance Evaluation, Ethernet, Infiniband

ACM Reference Format:
Sebastien Varrette, Hyacinthe Cartiaux, Teddy Valette, and Abatcha Olloh. 2022. Aggregating and Consolidating two High Performant Network Topologies: The ULHPC Experience. In *Practice and Experience in Advanced Research Computing (PEARC '22), July 10–14, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3491418.3535159

1 INTRODUCTION

HPC is crucial in academic environments to achieve high-quality results in all application areas. All world-class universities require this type of facility to accelerate the conducted research and ensure cutting-edge and timely results. The University of Luxembourg (UL) operates since 2007 a large research computing facility referred to hereafter as *ULHPC*, which remains a reference implementation within the country. It offers a cutting-edge research infrastructure to Luxembourg public research and serves as edge access to PRACE and EuroHPC supercomputers. Installed in the premises of the University's Centre de Calcul (CDC), the ULHPC facilities provides as of 2022 a total computing capacity of 2.76 PetaFlops and a shared storage capacity of 13.6 PetaBytes. A central component of the operated infrastructure which actually differentiates HPC systems from over distributed computing platforms remains the interconnect technology and topology. While *Ethernet* is established for decades as the dominant interconnect standard for mainstream commercial computing requirements, the underlying protocol has

PEARC '22, July 10–14, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9161-0/22/07.
https://doi.org/10.1145/3491418.3535159

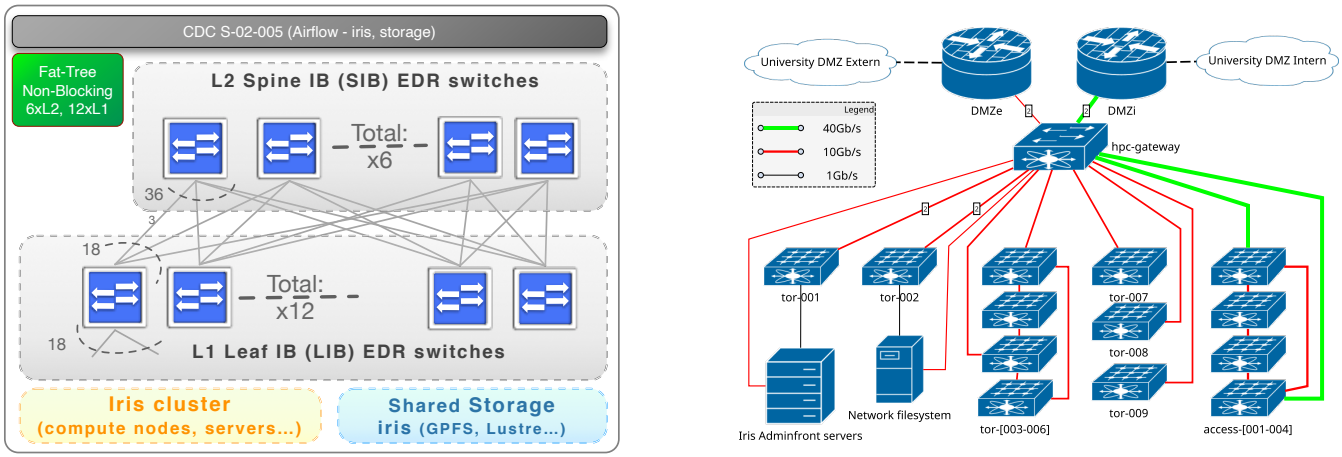

Figure 1: Overview of the high-performant IB (left) or Ethernet (right) network topologies before the integration of the *Aion* supercomputer.

inherent limitations preventing low-latency deployments expected in real HPC environment. When in need of high-bandwidth and low-latency communications, better options have emerged and are considered such as *InfiniBand* (IB), an industry standard defined by the InfiniBand Trade Association which originated in 1999 to specifically address workload requirements that were not adequately addressed by Ethernet. Designed for scalability using a switched fabric network topology together with Remote Direct Memory Access (RDMA) to reduce CPU overhead, IB is for several years the dominant interconnect technology in the HPC market; Alternatively, *vendor specific interconnects* can be considered. Nowadays, this mainly corresponds to the technology provided by three HPC vendors: Cray/HPE Slingshot, Intel's Omni-Path Architecture (OPA) or, to a lesser extent, Atos/Bull BXI. Table 1 depicts the theoretical performance characteristics of the different HPC interconnect technologies. When looking at the interconnect family distribution within the current generation of HPC systems as reported by the latest Top500 list [3], 35.6% of the listed systems rely on the IB network technology, a proportion increasing to 60% when restricting to the Top 50 systems. Within the ULHPC facility, the main high-bandwidth low-latency network relies on IB interconnect technology, more specifically in the latest HDR (High Data Rate – 200Gbps) and EDR (Enhanced Data Rate – 100Gbps) flavors. While different topologies are commonly deployed in large-scale HPC deployments *i.e.*, Fat tree, Hypercube, Torus or Dragonfly [5], Fat-tree was always promoted on all ULHPC clusters due to its versatility, high bisection bandwidth and well understood routing which remains very efficient at avoiding superposition of routes on the same link for all to all or many to many communication patterns. It is also the only topology allowing for a non-blocking network at large-scale.

For this reason, the seminal setup over the flagship cluster *Iris* (in production since 2017 and totaling 196 compute nodes) was relying on a non-blocking 1:1 Fat-Tree topology saturating the leaf connections, where all links have a bandwidth capacity of 100Gb/s due to the used Mellanox EDR InfiniBand technology. The corresponding topology is illustrated in the left Figure 1 and involves 6 spine switches (Level 2 in the Fat Tree, thus labelled "L2 SIB") and 12 leaf (Level 1/L1 LIB) switches allowing to connect a maximum of 216 leaf connections. Each LIB switch is connected with 3 links to each SIB switch to generate the 1:1 Fat-tree topology. In addition, the ULHPC facilities exploit a complementary Ethernet-based network defined for management tasks, external access and interactions with user's applications that do not support Infiniband natively. The initial configuration is reported in the right Figure 1 and used to comprise a core gateway with links to the Internet and the University's Intranet, Top-of-the-Rack (TOR) switches for out of band management, and access switches for WAN/LAN connections of servers, services, compute and storage equipment. More specifically, TOR switches (at least one per rack) are used for 1Gb Ethernet management network (RJ45 connections on all equipment). The access switches are used to connect the compute nodes of *Iris* within a 10Gb Ethernet network (with SFP+ connections), linking them to the WAN and LAN's services hosted on our administrative

Technology	Interconnect Family	Effective Bandwidth		Latency
Gigabit Ethernet	Ethernet	1 Gb/s	125 MB/s	40μs to 300μs
10 Gigabit Ethernet	Ethernet	10 Gb/s	1.25 GB/s	4μs to 5μs
100 Gigabit Ethernet	Ethernet	100 Gb/s	12.5 GB/s	30μs
Infiniband EDR	Infiniband	100 Gb/s	12.5 GB/s	0.61μs to 1.3μs
Infiniband HDR	Infiniband	200 Gb/s	25 GB/s	0.5μs to 1.1μs
Intel Omnipath	OmniPath	100 Gb/s	12.5 GB/s	0.9μs
Cray Slingshot	Proprietary Network	200 Gb/s	12.5 GB/s	0.3μs to 1.1μs

Table 1: Characteristics of the main HPC interconnect technologies.

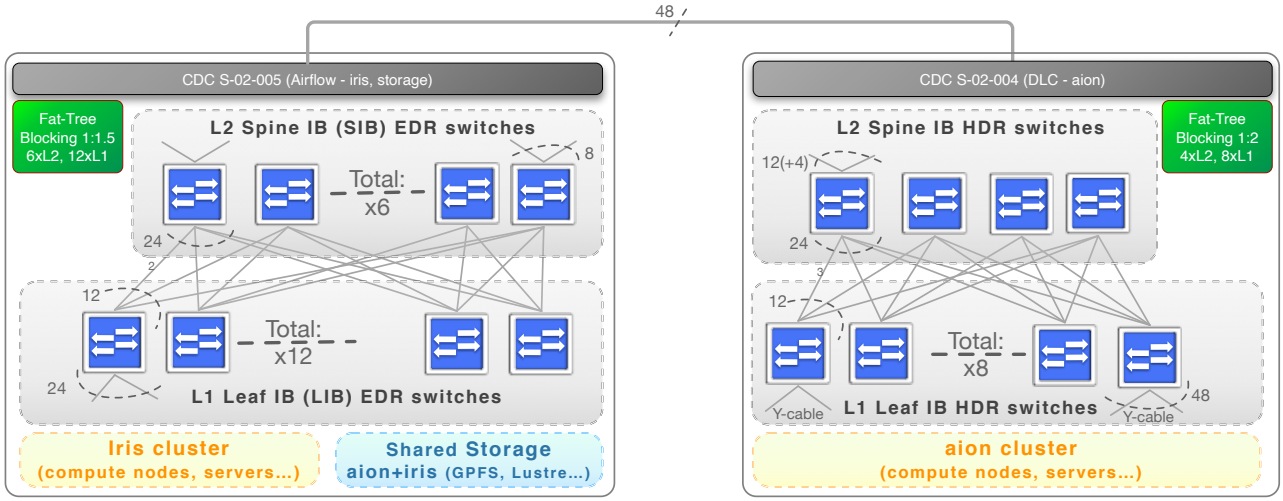

Figure 2: Overview of the current high-bandwidth low-latency IB network topology.

servers managing the HPC services. They are stacked together with redundant 40Gb Ethernet up-links to the HPC Gateway. The recent acquisition of a new cutting-edge supercomputer *Aion* (released in October 2021 and consisting of 318 compute nodes and several management servers) required to aggregate and consolidate the interconnect topologies in place. This article depicts the architecture and the solutions designed to expand the existing networks beyond their seminal capacity limits while keeping at best the Bisection bandwidth of the aggregated networks and minimizing the re-cabling operations.

2 PROPOSED IB TOPOLOGY WHEN MERGING THE TWO IB ISLANDS

The new supercomputer *Aion* came with its own internal IB "*island*" based on a Fat-tree topology composed by 4 spine and 8 leaf HDR switches. *Aion*'s compute node are connected to the leaf switches through HDR100 splitter cables (also called "Y-cables") which permits to drastically reduce the number of installed cables and thus the associated costs at the price of a blocking factor 2:1. The induced bandwidth penalty (*i.e.* 100 Gb/s instead of 200 Gb/s, thus aligned to *Iris* capacities relying on IB EDR technologies) was considered affordable as nowadays, very few applications are finally really able to fully exploit 200 Gb/s networks. Then, the main challenge was to adapt and extend the existing Fat-tree topology within the *Iris* IB island (which was set with a non-blocking configuration) to allow the integration of the new system while ensuring its "transparent" access over the IB network to the shared storage facilities. The only way to maintain a non-blocking configuration would mean a complete recabling of the existing solution over an upgraded Fat-tree topology designed with increased leaf capacity so as to sustain the connection of all ULHPC end-points, *i.e.*, *Iris* **and** *Aion* compute nodes (thus totalling 514 compute nodes), and all the storage and management interfaces. From our past experience of the complete moving of the *Iris* cluster hardware components from one server room to another, this solution was quicky discarded as it would induce a non-negligible cost overhead while putting the existing infrastructure at high risk. Indeed, a massive re-cabling is always prone to errors as the network fiber cables remain fragile components. For this reason, the proposed approach aimed at reaching a blocking yet balanced configuration (with a low blocking factor) with a good bisection bandwidth while minimizing recabling operations. This could have been done by introducing an additional top level layer (a third level) with several 'super' spine switches enabling to bridge the two IB islands. Yet this topology would impact the latency expected for I/O operations as it would enforce to cross three level in the Fat-Tree hierarchy for operations performed from the *Aion* compute nodes.

This article reports an alternative topology kept on 2 layers only (thus maintaining the Fat-tree height) that permits to keep a low blocking factor (different on both cluster), thus minimizing congestion and other performance degrading factors. The proposed solution, inspired by the DragonFly topology, is depicted in the Figure 2. In practice, we removed 6 cables on each of the L1 LIB switches within the *Iris* IB island (*i.e.*, 1 connection to each of the 6 L2 SIB switches) which results in freeing up 12 ports on each of the L2 SIB switches. Then from *each* of the 4 *Aion* L2 SIB switches, 12 interlink cables were used and distributed 2-by-2 over the 6 L2 SIB switches from the existing *Iris* cluster, bringing a total of $4 \times 12 = 48$ interlinks within the global topology. Overall, this approach allowed to increase the leaf connection capacity from 216 to $12 \times 24 + 8 \times 48 = 672$ end-points (**+311%**). This changed the blocking factor for *Iris* from full non-blocking to 1:1.5. In return, the proposed topology update prooved to induce very marginal performance penalties. For instance, **less than 3% (resp. 0.3%) Read (resp. Write) bandwidth degradation** were observed when evaluating the impact on the parallel I/O performance of the shared storage infrastructure (either SpectrumScale/GPFS or Lustre) through IOR[1] [1, 7]. The resulting IB network configuration was also validated with the MPI Bisectional Bandwidth (BB) benchmark widely used to provide an evaluation of a topology's performance [6]. This measures the IB

[1]This reference parallel IO benchmark is used to measure I/O throughput using various interfaces and access patterns subjected to a synthetic workload.

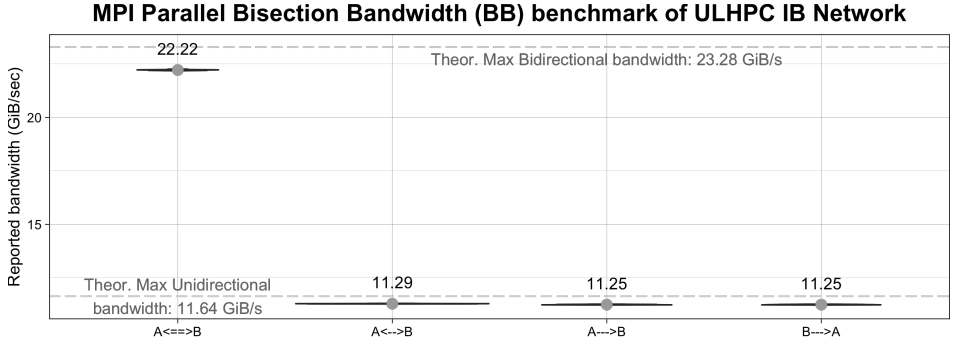

Figure 3: MPI Bisectional Bandwidth (BB) IB performance between ULHPC compute nodes.

bandwidth between pairs of nodes. Considering the theoretical effective throughput of the implemented network at the compute nodes level (100 Gb/s), unidirectional (resp. bidirectional) point-to-point bandwidth evaluations are expected to reach 11,64 GiB/s (resp. 23,28 GiB/s). The measured performances are reported in the Figure 3, demonstrating stable and sustainable performance results for all possible pairs of nodes, *i.e.*, **96.65% (resp. 95.45%) unidirectional (resp. bidirectional) point-to-point bandwidth efficiency** when compared to these theoretical maximum attainable performances. Finally, it is worth to report that the selected routing engine (`ftree`) enabled on a redundant set of IB subnet managers was configured with dedicated and fast path to the IO targets, avoiding congestion on the high-speed network and mitigating the risk of runtime "jitter" for time critical jobs [4].

3 PROPOSED ETHERNET TOPOLOGY

Having a single high-bandwidth and low-latency network to support efficient HPC and Big Data workloads would not provide the necessary flexibility brought by the Ethernet protocol. For this reason, an additional Ethernet-based network is defined for management tasks, external access and non-IB compliant user's applications inside the research computing system. The different flows and streams are separated inside dedicated Virtual Local Area Network (VLAN) (see table 2) as detailed in the sequel. Compared to the seminal configuration displayed in the Figure 1 which exhibits as evident single point-of-failure the HPC gateway switch, the Ethernet network has been heavily reorganized as a novel 2-layer topology summarized in Figure 4. The upper level (*Gateway Layer*) is dedicated for routing, switching features, network isolation and filtering (ACL) rules and is meant to aggregate *only* switches unlike the initial setup. This layer is handled by a now *redundant* set of site routers (HPC gateways), featuring many 40GbE and 10GbE ports. It allows to interface the University network for both internal (LAN) and external (WAN) communications.

The bottom level (*Switching Layer*) is composed by *core* switches as well as the *TOR* network equipment, meant to interface the HPC servers and compute nodes. As before, the TOR switches are typically 1GbE switches with redundant 10GbE uplinks, possibly stacked (a configuration enforced on *Aion*), and connecting all out-of-band interfaces for hardware management. The core switches (previously called access switches) are 10GbE switches with redundant 40GbE uplinks, consistently stacked or clustered using Cisco vPC technology (Virtual Port Channel). This new topology aimed at circumventing the limitations met with the precedent setup. Concretely, it provides the following features:

(1) enhanced service *availability* using fault tolerance techniques: critical network equipment are fully redundant; critical servers are connected using link aggregations etc.;
(2) improved *maintainability*. For instance, it is easy to apply firmware and security updates on the switches, without requiring a service interruption or a maintenance window;
(3) *scalability*: additional clusters or racks of computing equipment can be added in the coming years, without requiring any major topology change or physical cabling.

In addition, we have reworked a consistent set of network rules to be followed for a sane network infrastructure and an easier global administration of the HPC infrastructure. The separation of the main Ethernet networks into VLANs and subnets, with strict security policies enforced and implemented via ACLs on the layer 3, offers an isolated environment from the UL internal network. More

VLAN	Typical capacity	Description
`Interco`	40-100 GbE	Interconnection with the University network.
`DMZ*`	10-40 GbE	Demilitarized zone (DMZ) network for services *i.e.*, user-accessible entry point.
`prod*`	10-40 GbE	User-level data transfer (excluding very-high-bandwidth, low-latency transfers as well as I/O) and Internet access, in-band management
`mgmt*`	1 GbE	Management network containing all management card (BMC) for all installed equipment (server, racks, censors etc.)
`IPoIB`	100 GbE	*Non routed* network for IP over InfiniBand (IB)

Table 2: Overview of the configured VLANs.

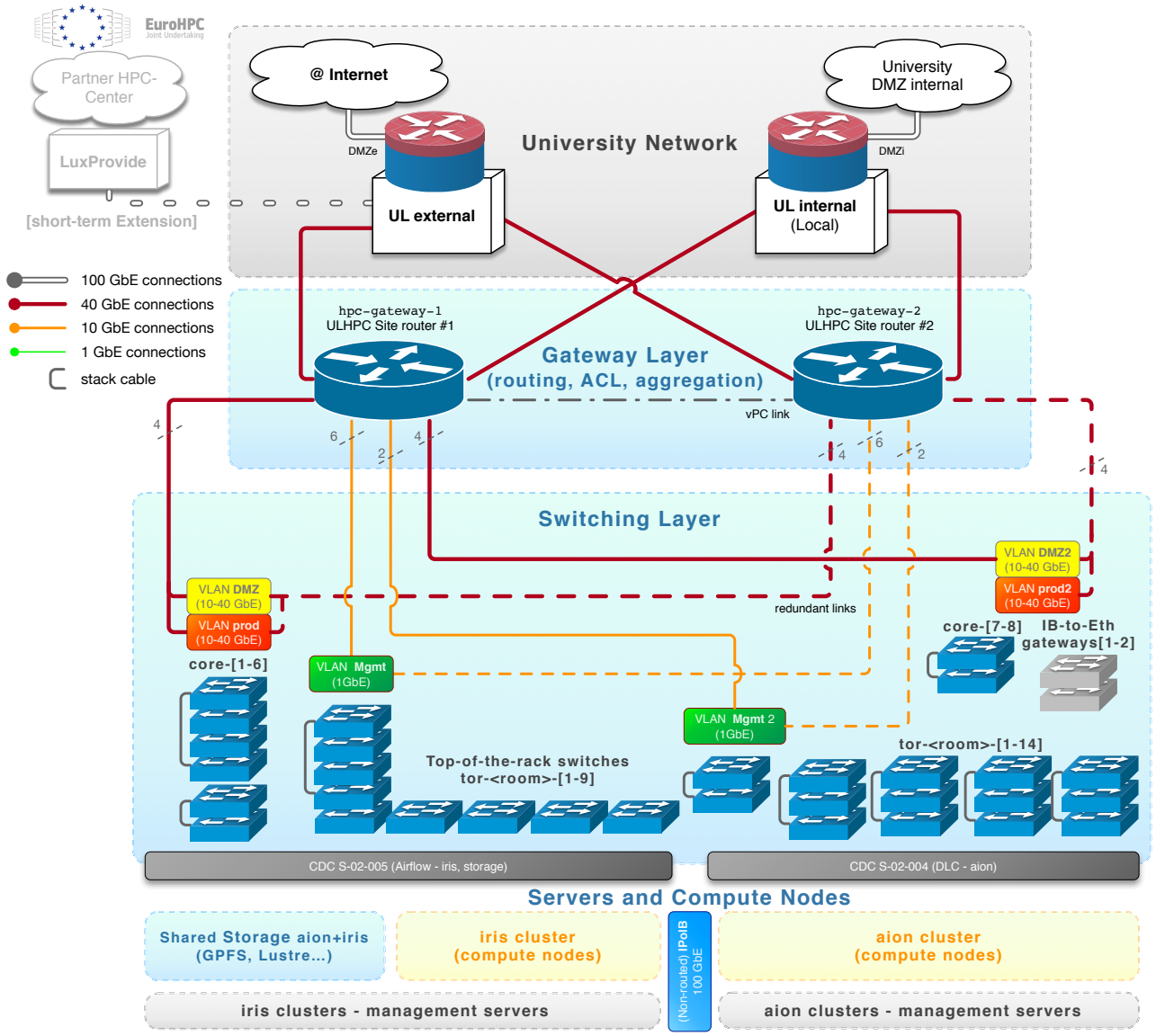

Figure 4: Overview of the current Ethernet network topology.

specifically, the VLANs are defined by cluster and by type of equipment, with the objective of isolating the compute nodes and any user-facing resources inside a private network unreachable from the outside, and to confine the broadcast traffic and network noise within its own source cluster network. Furthermore, network ACLs are defined by flow (fully allowed or not), rather than by specific hosts. Finally, systematic logical rules for IP addressing is enforced, taking into account the type of the equipment (network, storage, server, compute nodes), in order to prevent overlapping with any other internally reserved range. In addition, systematic network cabling guidelines are promoted with identical and unique labels on both ends of each cable, thus facilitating the work of the HPC operators when interviening in the data centre premisces. Exploiting the `iperf3` tool [2], an exhaustive performance evaluation was performed over the proposed Ethernet topology and is reported in the Table 3. It permitted to validate the theoretical point-to-point link capacities – above 94.1% bandwidth efficiency – when evaluating different representative network paths within the network. Starting 40GbE network capacities, the MTU parameter (*Maximum Transmission Unit*) must be tuned to reach such an efficiency. Since this is a non-standard state of practice, we reported the results obtained with the default MTU settings (1500 bytes), which allowed to reach in this case a 74.4% bandwidth efficiency.

4 CONCLUSION AND PERSPECTIVES

This article reports on the implemented topology changes which were introduced within the ULHPC facility upon the release of a newly acquired supercomputer *Aion*. The objective was to expand and consolidate the existing networks beyond theirs seminal

VLAN	Interconnect Path	Theoretical Bandwidth	Measured Bandwidth mean	sd
Interco	UL internal network ⇔ HPC gateway	40000 Mb/s	29757 Mb/s*	1060
prod*	*Iris* access frontend ⇔ *Iris* compute node	10000 Mb/s	9411 Mb/s	11.4
mgmt*	*Aion* deployment server ⇔ *Aion* BMC compute node	1000 Mb/s	942 Mb/s	0.496

*: default MTU parameter

Table 3: Summary of the main `iperf3` multithreaded performance evaluation results.

capabilities and capabilities. It used to affect both the main high-bandwidth low-latency network which relies on the dominant interconnect technology in the HPC market *i.e.,* IB EDR and HDR, as well as the complementary Ethernet-based network defined for management tasks, external access and user's applications that do not support Infiniband natively. With regards the IB network, the proposed approach enforces the migration from a non-blocking topology to a blocking configuration where the associated blocking factor differs between the two supercomputers. In return, following the described strategy allowed for the leaf connection capacity to be more than tripled without increasing the number of levels in the Fat-Tree hierarchy. The performance impact evaluated through several reference benchmarks demonstrates marginal penalties, for instance over parallel I/O throughput (below 3%). Furthermore, stable and sustainable unidirectional and bidirectional point-to-point bandwidth efficiencies (above 95.45%) was observed when measured across all possible pairs of compute nodes. The Ethernet topology on the other side was the subject of a major reorganization within a 2-layer topology aiming for improving the robustness, availability, maintainability and scalability of the corresponding interconnect network. The robustness and high-availability of the implemented network configuration was tested at several occasion and proved to occur without noticeable disruption on the network traffic. This happened during planned maintenance sessions upon firmware upgrades operations over network hardware equipment, or (once) while in production from an unexpected switch shutdown within the implemented gateway layer. As perspectives tied to the presented work, the smooth integration with collaborating EuroHPC infrastructures is expected, as well as further HPC capacity expansions planned starting 2023. Both setups will be performed within the described topologies with minimal changes. In all cases, the reported IB and Ethernet interconnect architectures have been successfully deployed and are in production within the ULHPC facility. They are applicable to a broad range of HPC infrastructures and thus may help other HPC centres to consolidate their own interconnect stacks.

ACKNOWLEDGMENTS

The experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg [7] – See `hpc.uni.lu`.

REFERENCES

[1] [n.d.]. IOR: HPC I/O Benchmark. [online]. https://ior.readthedocs.io/.
[2] [n.d.]. iperf3: A TCP, UDP, and SCTP network bandwidth measurement tool. https://software.es.net/iperf/.
[3] [n.d.]. The Top 500 List. https://top500.org/.
[4] Maciej Besta, Jens Domke, Marcel Schneider, Marek Konieczny, Salvatore Di Girolamo, Timo Schneider, Ankit Singla, and Torsten Hoefler. 2021. High-Performance Routing With Multipathing and Path Diversity in Ethernet and HPC Networks. *IEEE Transactions on Parallel and Distributed Systems* 32, 4 (2021), 943–959.
[5] A. Bhatele, N. Jain, M. Mubarak, and T. Gamblin. 2019. Analyzing Cost-Performance Tradeoffs of HPC Network Designs under Different Constraints Using Simulations. In *Proc. of the ACM SIGSIM Conf. on Principles of Advanced Discrete Simulation (SIGSIM-PADS'19)* (Chicago, IL, USA) *(SIGSIM-PADS '19)*. ACM, New York, NY, USA, 1–12.
[6] Sangeetha Abdu Jyothi, Ankit Singla, P. Brighten Godfrey, and Alexandra Kolla. 2016. Measuring and Understanding Throughput of Network Topologies. In *SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 761–772. https://doi.org/10.1109/SC.2016.64
[7] S. Varrette, H. Cartiaux, S. Peter, E. Kieffer, T. Valette, and A. Olloh. 2022. Management of an Academic HPC & Research Computing Facility: The ULHPC Experience 2.0. In *Proc. of the 6th ACM High Performance Computing and Cluster Technologies Conf. (HPCCT 2022)*. Association for Computing Machinery (ACM), Fuzhou, China.