High Performance
Computing &
Big Data Services

hpc.uni.lu

hpc@uni.lu

@ULHPC

LUXEMBOURG
LET'S MAKE IT HAPPEN

Aggregating and Consolidating two High Performant Network Topologies

The ULHPC Experience

Dr. S. Varrette, H. Cartiaux, T. Valette and A. Olloh

University of Luxembourg (UL), Luxembourg

https://hpc.uni.lu

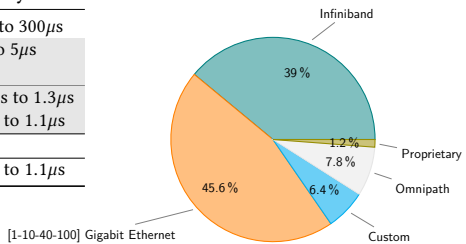
Practice and Experience in Advanced Research Computing (PEARC'22)

July 13th, 2022, Boston, USA

Summary

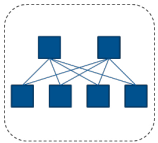
HPC Interconnect Technologies

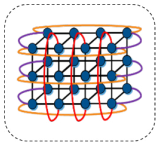
Technology	Interconnect Family	Effective Bandwidth		Latency
Gigabit Ethernet	Ethernet	1 Gb/s	125 MB/s	40μs to 300μs
10 Gigabit Ethernet	Ethernet	10 Gb/s	1.25 GB/s	4μs to 5μs
100 Gigabit Ethernet	Ethernet	100 Gb/s	12.5 GB/s	30μs
Infiniband EDR	Infiniband	100 Gb/s	12.5 GB/s	0.61μs to 1.3μs
Infiniband HDR	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

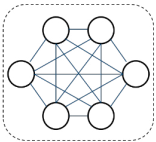


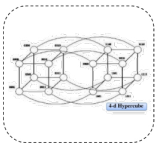
[Source : www.top500.org, Jun 2022]

HPC Interconnect Technologies and Topologies

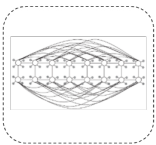

Fat Tree

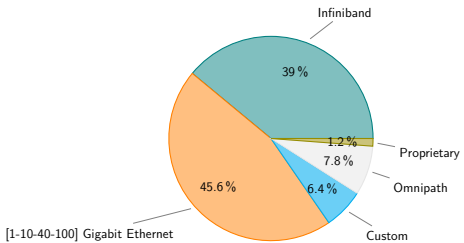

Torus


Dragonfly



Hypercube


HyperX



[Source : www.top500.org, Jun 2022]

- **CLOS Network / Fat-Trees**: versatile, provides high bisection bandwidth
  - ↪ the only topology allowing for a non-blocking network at large-scale

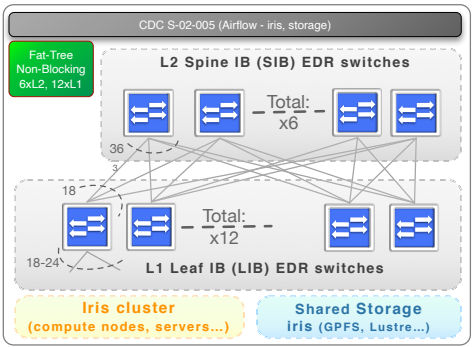
Uni.lu HPC Supercomputers: `iris` cluster

`hpc-docs.uni.lu/systems/iris/`

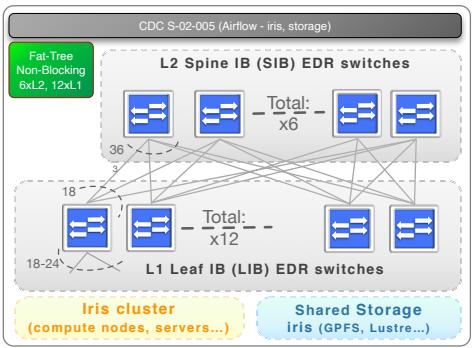
- **Dell/Intel** supercomputer *Air-flow cooling*
  - ↪ 196 compute nodes, **5824 cores**, 52.2 TB RAM
  - ↪ R_{peak}: **1,07 PetaFlop/s**
    - ✓ **regular** nodes (Dual CPU, 128 to 256 GB of RAM)
    - ✓ **GPU** nodes (Dual CPU, 4 NVidia accelerators, 768 GB RAM)
    - ✓ **Large-memory** nodes (Quad-CPU, 3072 GB RAM)
- Stepwise deployment since 2017 two upgrade phases (2018 and 2019)

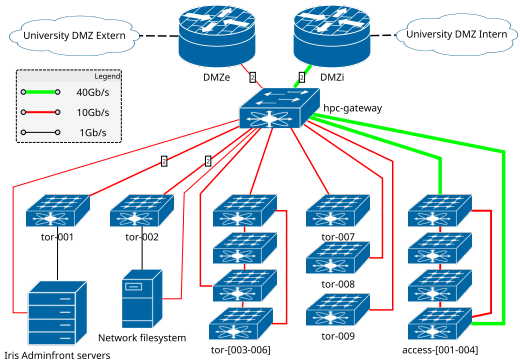
- `iris` **Interconnect Technologies**
  - ↪ **Fast IB EDR network**, **Fat-Tree** Topology
  - ↪ Complementary **Ethernet Network**

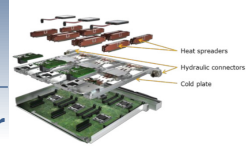
Initial iris IB ... Interconnect



Initial `iris` IB ... and Ethernet Interconnect





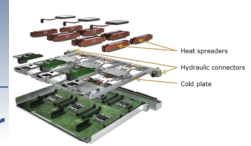


Uni.lu HPC Supercomputers: `aion` cluster

`hpc-docs.uni.lu/systems/aion/`

- Acquisition by European Tender in 2020
  - ↪ **production release** in Oct 2021
- **Atos/AMD** supercomputer, DLC cooling
  - ↪ 4 BullSequana XH2000 adjacent racks
  - ↪ 318 **regular** nodes, **40704 cores**, 81.4 TB RAM
  - ↪ R_{peak}: **1,693 PetaFLOP/s**
- `aion` **Interconnect Technologies**
  - ↪ **Fast IB HDR network**, **Fat-Tree** Topology
  - ↪ Complementary **Ethernet Network**

Uni.lu HPC Supercomputers: `aion` cluster



`hpc-docs.uni.lu/systems/aion/`

- Acquisition by European Tender in 2020
  - ↪ **production release** in Oct 2021
- **Atos/AMD** supercomputer, DLC cooling
  - ↪ 4 BullSequana XH2000 adjacent racks
  - ↪ 318 **regular** nodes, **40704 cores**, 81.4 TB RAM
  - ↪ R_{peak}: **1,693 PetaFLOP/s**
- `aion` **Interconnect Technologies**
  - ↪ **Fast IB HDR network**, **Fat-Tree** Topology
  - ↪ Complementary **Ethernet Network**

- **In this talk**: when **integrating** `aion` into the existing HPC ecosystem:
  - ↪ **Lessons learned from aggregating the IB and Ethernet networks**

Adapting the Fast Local IB Interconnect Network

- **`aion` came with its own internal IB Fat-Tree "*island*"**
  - ↪ 4 spine SIB and 8 LIB **HDR** switches (200 Gb/s)
  - ↪ compute node connected through HDR100 splitter cables (or "*Y-cables*")
    - ✓ permits to drastically reduce the number of installed cables and thus the associated costs
    - ✓ price: **blocking factor 2:1** yet induced bandwidth penalty aligned to `iris` capacities

Adapting the Fast Local IB Interconnect Network

- `aion` **came with its own internal IB Fat-Tree "*island*"**
  - ↪ 4 spine SIB and 8 LIB **HDR** switches (200 Gb/s)
  - ↪ compute node connected through HDR100 splitter cables (or "*Y-cables*")
    - ✓ permits to drastically reduce the number of installed cables and thus the associated costs
    - ✓ price: **blocking factor 2:1** yet induced bandwidth penalty aligned to `iris` capacities

Q: how to merge the two IB islands (`iris` and `aion`) ?

Adapting the Fast Local IB Interconnect Network

- **`aion` came with its own internal IB Fat-Tree "*island*"**
  - ↪ 4 spine SIB and 8 LIB **HDR** switches (200 Gb/s)
  - ↪ compute node connected through HDR100 splitter cables (or "*Y-cables*")
    - ✓ permits to drastically reduce the number of installed cables and thus the associated costs
    - ✓ price: **blocking factor 2:1** yet induced bandwidth penalty aligned to `iris` capacities

Q: how to merge the two IB islands (`iris` and `aion`) ?

- **Approach 1: maintain a non-blocking configuration**
  - ↪ upgraded Fat-tree topology for increased leaf capacity (216 → at least 530)
  - ↪ major recabling on `iris` required!
  - ↪ quickly discarded solution from past experience on cluster moving:
    - ✓ massive re-cabling always prone to errors (network fiber cables remain fragile components)

Adapting the Fast Local IB Interconnect Network

- **Approach 2: allow for a blocking yet balanced configuration**
  - ↪ target low blocking factor with a good bisection bandwidth
  - ↪ minimizing recabling operation

a. **Introduce an additional top level layer (L3)**
   - ↪ several 'super' spine switches enabling to bridge the two IB islands.
   - ↪ would impact latency expected for I/O operations (expecially from `aion`)

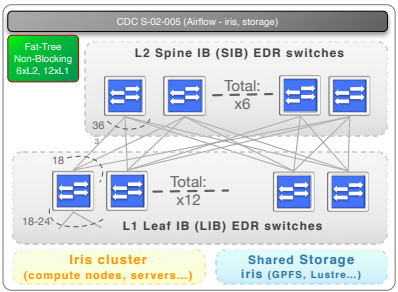
Adapting the Fast Local IB Interconnect Network

- **Approach 2: allow for a blocking yet balanced configuration**
  - ↪ target low blocking factor with a good bisection bandwidth
  - ↪ minimizing recabling operation

a. **Introduce an additional top level layer (L3)**
   - ↪ several 'super' spine switches enabling to bridge the two IB islands.
   - ↪ would impact latency expected for I/O operations (expecially from `aion`)

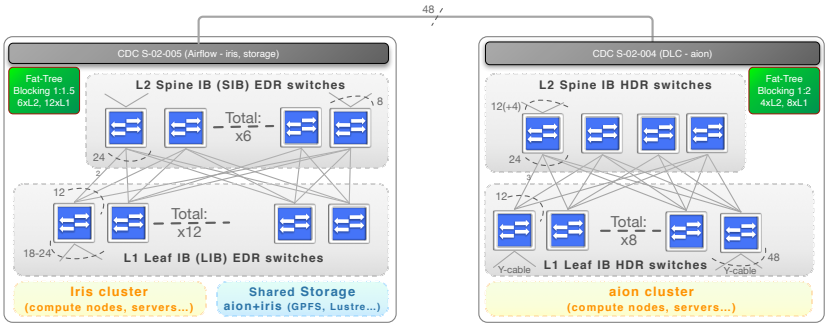
b. (our proposal) **Alternative topology kept on 2 layers only**
   - ↪ DragonFly inspired, maintain Fat-tree height
   - ↪ keep a low blocking factor (different on both cluster)
     - ✓ minimizing congestion and other performance degrading factors.
   - ↪ **Leaf capacity increase**: $216 \rightarrow 12 \times 24 + 8 \times 48 = \mathbf{672}$ end-points (**+311%**)

Adapting the Fast Local IB Interconnect Network



- **before** integration of `aion` (`iris` alone)

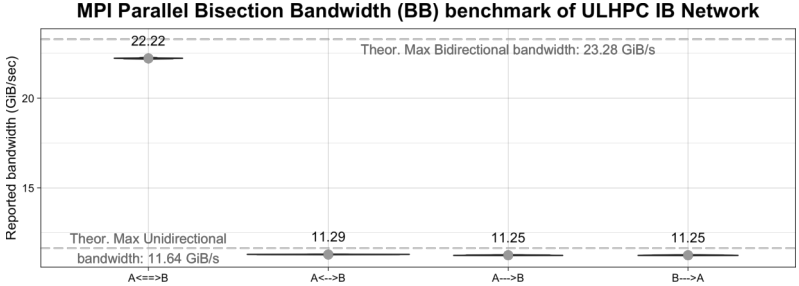
Adapting the Fast Local IB Interconnect Network



- **after** merging `iris` and `aion` IB islands. *In practice*:
  - ↪ 6 LIB ↔ SIB cables removed within `iris` IB island to free 12 ports on each L2 SIB switches
    - ✓ used to connect (2-by-2) 4 Aion L2 SIB switches with the 6 Iris L2 SIB switches
  - ↪ Adaptation of the subnet manager configuration (routing engine, root GUIDs etc.)

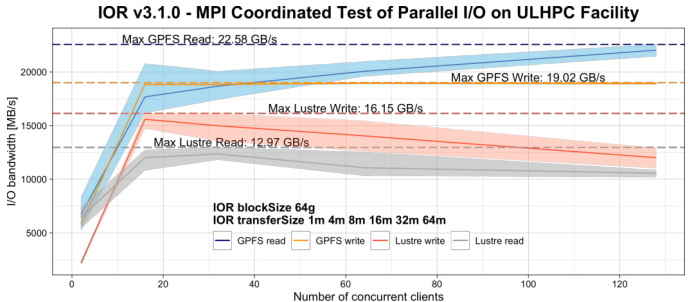
IB Network Aggregation Validation and Impact

- **Network sanity validation** (once link state/speed and SM config carefully validated)
  - ↪ OSU Microbenchmarks (version 5.6.3) for MPI collectives performance evaluation etc.
  - ↪ IB Bisection Bandwidth (**BB**) benchmarks: **96,99% efficiency**



IB Network Aggregation Validation and Impact

- **Network sanity validation** (once link state/speed and SM config carefully validated)
  - ↪ OSU Microbenchmarks (version 5.6.3) for MPI collectives performance evaluation etc.
  - ↪ IB Bisection Bandwidth (**BB**) benchmarks: **96,99% efficiency**
- **Marginal performance penalties**
  - ↪ IOR: **less than 3% (resp. 0.3%) Read (resp. Write) bandwidth degradation**



IB Network Aggregation Validation and Impact

- **Network sanity validation** (once link state/speed and SM config carefully validated)
  - ↪ OSU Microbenchmarks (version 5.6.3) for MPI collectives performance evaluation etc.
  - ↪ IB Bisection Bandwidth (**BB**) benchmarks: **96,99% efficiency**
- **Marginal performance penalties**
  - ↪ IOR: **less than 3% (resp. 0.3%) Read (resp. Write) bandwidth degradation**
  - ↪ cf. also HPL, HPCG, Graph500, Green500, GreenGraph500 performance evaluation [HPCCT22]

	Benchmark	#N	*(Main parameters)*	Best Performance	Efficiency	Improvement*	Equivalent Worldwide Rank	
Aion	**HPL** (Top500)	**318**	*(NB=192,P×Q=48×53)*	$R_{\max}$ = **1255.36 TFlops**	**74.10%**	+1.9%	>500 (Nov 2021)	**#490** (Jun 2020)
	Green500	318		5.19 GFlops/W		+12.83%	**#60** (Nov 2021)	**#56** (Jun 2021)
	HPCG	318		16.842 TFlops		+15.35%	**#144** (Nov 2021)	**#135** (Jun 2021)
	Graph500 BFS	2^8=256	*(Scale: 36,Edge:16)*	975 GTEPS		+64%	**#27** (Nov 2021)	**#23** (Jun 2021)
	GreenGraph500	2^8=256		6.14 MTEPS/W		+180%	**#37** (Nov 2021)	**#36** (Jun 2021)
				**: performance improvement with the minimal acceptance threshold set in the* Aion *tender document*				
	IO500 (isc21 release)	128		**11.345219**			**#42** (Nov 2020 - latest release)	

[HPCCT22] S. Varrette H. Cartiaux, S. Peter, E. Kieffer, T. Valette, and A. Olloh, "*Management of an Academic HPC & Research Computing Facility: The ULHPC Experience 2.0*". In 6th ACM HPC and Cluster Technologies Conference (HPCCT 2022), Fuzhou, China (2022).

Difficulties Met and Lesson Learned

Take Away Messages for PEARC community

- **Align to a *compliant* MOFED version each island before merging**
  - ↪ check for kernel requirements from deployed OS
    - ✓ MUST match deployed GPFS/Lustre expectations (`gplbin`: GPFS portability layer)
  - ↪ heterogeneous HW complexifies the selection (switches models, CX{3,4,6} HCA...)
  - ↪ **MOFED upgrade comes with ALL equipement FW alignment**
    - ✓ Careful with the **upgrade path**

Difficulties Met and Lesson Learned

Take Away Messages for PEARC community

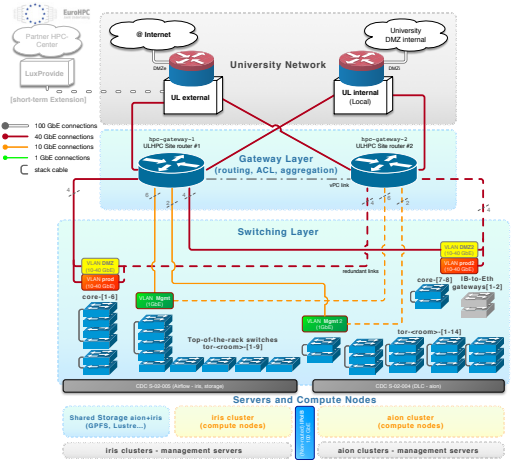
- **Align to a *compliant* MOFED version each island before merging**
  - ↪ check for kernel requirements from deployed OS
    - ✓ MUST match deployed GPFS/Lustre expectations (`gplbin`: GPFS portability layer)
  - ↪ heterogeneous HW complexifies the selection (switches models, CX{3,4,6} HCA. . .)
  - ↪ **MOFED upgrade comes with ALL equipement FW alignment**
    - ✓ Careful with the **upgrade path**
- **Redundant IB Subnet Manager** (OpenSM)
  - ↪ **Routing engine**: ~~`ar_ftree`~~ (proved to be **not** compliant with CX4) → `ftree`
  - ↪ **Careful definition of `root_guid` file!** (all L2 switches GUID)
    - ✓ **Otherwise**: any cable error will lead to revert to `minhop` routing (== bad performances)
  - ↪ plan dedicated and fast path to the IO targets
    - ✓ mitigating the risk of runtime "jitter" for time critical jobs

Difficulties Met and Lesson Learned

Take Away Messages for PEARC community

- **Align to a *compliant* MOFED version each island before merging**
  - ↪ check for kernel requirements from deployed OS
    - ✓ MUST match deployed GPFS/Lustre expectations (`gplbin`: GPFS portability layer)
  - ↪ heterogeneous HW complexifies the selection (switches models, CX{3,4,6} HCA. . .)
  - ↪ **MOFED upgrade comes with ALL equipement FW alignment**
    - ✓ Careful with the **upgrade path**
- **Redundant IB Subnet Manager** (OpenSM)
  - ↪ **Routing engine**: ~~`ar_ftree`~~ (proved to be **not** compliant with CX4) → `ftree`
  - ↪ **Careful definition of `root_guid` file!** (all L2 switches GUID)
    - ✓ **Otherwise**: any cable error will lead to revert to `minhop` routing (== bad performances)
  - ↪ plan dedicated and fast path to the IO targets
    - ✓ mitigating the risk of runtime "jitter" for time critical jobs
- `ibdiagnet` and `ibnetdiscover` are (as always) your friends

Complementary Ethernet Network



`hpc-docs.uni.lu/interconnect/ethernet/`

- Flexibility of Ethernet-based networks still required
- **2-layers** topology
  - ↪ *Upper* level: **Gateway Layer**
    - ✓ routing, switching features, network isolation and filtering (ACL) rules
    - ✓ meant to interconnect only switches.
    - ✓ allows to interface University network (LAN/WAN)
  - ↪ *bottom* level: **Switching Layer**
    - ✓ [stacked or clustered using vPC] core switches
    - ✓ TOR (Top-the-rack) switches
    - ✓ meant to interface HPC servers and compute nodes

Complementary Ethernet Network

- Compared to the precedent setup:
  - ↪ **enhanced service availability** using Fault-Tolerance techniques (redundancy, link aggregation...)
  - ↪ **improved maintainability** Ex: firmware/security updates on switches *without* service interruption
  - ↪ **scalability**: ready for new clusters
- **Strict security policies** enforced and implemented via ACLs on the layer 3

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)

Complementary Ethernet Network

- Compared to the precedent setup:
  - ↪ **enhanced service availability** using Fault-Tolerance techniques (redundancy, link aggregation...)
  - ↪ **improved maintainability** Ex: firmware/security updates on switches *without* service interruption
  - ↪ **scalability**: ready for new clusters
- **Strict security policies** enforced and implemented via ACLs on the layer 3
- Network **validation** *(outside classical sanity checks)* and **performance evaluation**
  - ↪ multithreaded `iperf3` across the network. $\geq$ **94.1% bandwidth efficiency (1-10GbE)**

VLAN	Interconnect Path	Theoretical Bandwidth	Measured Bandwidth mean	Measured Bandwidth 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

Conclusion

- **In this talk**:
  - ↪ **Implemented topology adaptation** when **integrating** a new supercomputer `aion`
  - ↪ **Proposed IB topology** allowed to keep the global Fat-tree height (2 levels)
    - ✓ migration from non-blocking topology to a blocking configuration on `iris`
    - ✓ stable and sustainable bandwidth efficiencies and marginal performance penalities
  - ↪ **Major Ethernet network reorganization** into within a **2-layer topology**
    - ✓ improved robustness, availability, maintainability and scalability
    - ✓ secure and consistent network rules, VLANs etc.
  - ↪ **Successfully deployed and in production for more than 1 year**
    - ✓ applicable to broad range of HPC infrastructures to consolidate their own interconnect stacks

Conclusion

- **In this talk**:
  - ↪ **Implemented topology adaptation** when **integrating** a new supercomputer `aion`
  - ↪ **Proposed IB topology** allowed to keep the global Fat-tree height (2 levels)
    - ✓ migration from non-blocking topology to a blocking configuration on `iris`
    - ✓ stable and sustainable bandwidth efficiencies and marginal performance penalities
  - ↪ **Major Ethernet network reorganization** into within a **2-layer topology**
    - ✓ improved robustness, availability, maintainability and scalability
    - ✓ secure and consistent network rules, VLANs etc.
  - ↪ **Successfully deployed and in production for more than 1 year**
    - ✓ applicable to broad range of HPC infrastructures to consolidate their own interconnect stacks

- **Perspectives and Future directions**
  - ↪ **Smooth integration with Euro-HPC infrastructures**
    - ✓ *transparently* outsource Research Computing/data analytic workflows to Tier-0 systems
  - ↪ **Ready for further HPC capacity expansions over the implemented topologies**
    - ✓ (*normally*) with minimal changes

Thank you for your attention...

Questions?

Sebastien Varrette, Hyacinthe Cartiaux, Teddy Valette & Abatcha Olloh
Aggregating & Consolidating two High Performant Network Topologies: The ULHPC Experience – ACM PEARC'22 - www
University of Luxembourg, Belval Campus
Maison du Nombre, 4th floor
2, avenue de l'Université
L-4365 Esch-sur-Alzette
mail: firsname.lastname@uni.lu

High Performance Computing @ Uni.lu
mail: hpc@uni.lu

1. **Introduction: Context and Motivations**
2. **Proposed IB Topology when Merging the two IB Islands**
3. **Proposed Ethernet Topology**
4. **Conclusion & Perspectives**

High Performance Computing @ Uni.lu

`hpc.uni.lu`

ULHPC Technical Docs

`hpc-docs.uni.lu`