



Management of an Academic HPC & Research Computing Facility: The ULHPC Experience 2.0

Dr. S. Varrette, H. Cartiaux, S. Peter, Dr. E. Kieffer, T. Valette, A. Olloh

University of Luxembourg (UL), Luxembourg

<https://hpc.uni.lu>

6th HPC and Cluster Technologies Conference (HPCCT 2022)



July 10th, 2022, Fuzhou, China



Summary

1 Overview of the Managed Facility

- Network Organisation
- Tiered Shared Storage infrastructure
- Computing Performance Evaluation and Acceptance Tests

2 User Software Environment

3 User Job Management and the Slurm infrastructure

4 Conclusion and Perspectives

Summary

1 Overview of the Managed Facility

Network Organisation

Tiered Shared Storage infrastructure

Computing Performance Evaluation and Acceptance Tests

2 User Software Environment

3 User Job Management and the Slurm infrastructure

4 Conclusion and Perspectives

University of Luxembourg

www.uni.lu



- **Created in 2003**, moved to Belval (South of the country) in 2015
- Among the top 250 universities in the Times Higher Education (THE) Rankings 2021
 - N°1 worldwide in the THE "international outlook" Rankings
 - N°20 worldwide in the THE Young University Rankings 2021.
 - ✓ N°4 (out of 64) in the THE Millennials Rankings 2021.

Uni.lu HPC (UL HPC) Facility

- Managed and operated since 2007 (Dr. S. Varrette & Co.)
→ 2nd Largest HPC facility in Luxembourg after EuroHPC MeluXina

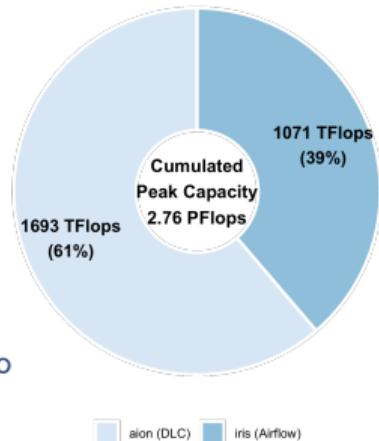


hpc.uni.lu

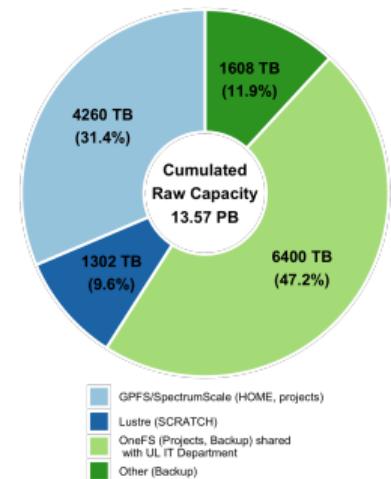
Technical Docs:
hpc-docs.uni.lu

ULHPC Tutorials:
ulhpc-tutorials.rtfd.io

UL HPC Supercomputers (2022)



UL HPC Storage FileSystems (2022)



High Performance Computing & Big Data Services

-  hpc.uni.lu
-  hpc@uni.lu
-  [@ULHPC](#)

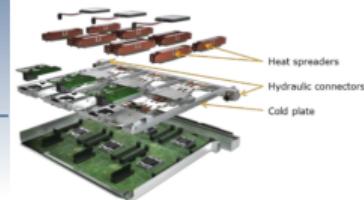
LUXEMBOURG
LET'S MAKE IT HAPPEN

UL 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)
- Fast InfiniBand (IB) EDR network
 - ↪ **Fat-Tree Topology** blocking factor 1:1.5
- Stepwise deployment since 2017
 - ↪ two major upgrade phases (2018 and 2019)

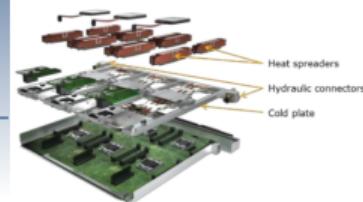


UL HPC Supercomputers: aion cluster

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

- 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
- Fast InfiniBand (IB) HDR network
 - Fat-Tree Topology blocking factor 1:2
- Acquisition by European Tender in 2020
 - production release in Oct 2021 (delayed by COVID)



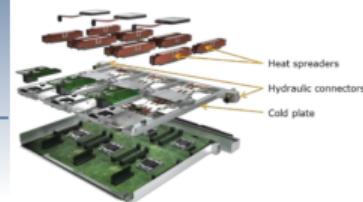


UL HPC Supercomputers: aion cluster

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

- Atos/AMD supercomputer, DLC cooling (**EOY update**)
 - ↪ 4 BullSequana XH2000 adjacent racks
 - ↪ **354 regular nodes, 45312 cores**, 90.6 TB RAM
 - ↪ R_{peak} : **1,885 PetaFLOP/s**
- Fast InfiniBand (IB) HDR network
 - ↪ **Fat-Tree Topology** blocking factor 1:2
- Acquisition by European Tender in 2020
 - ↪ **production release in Oct 2021** (delayed by COVID)
 - ↪ **First upgrade EOY 2022** +36 **regular nodes**





UL HPC Supercomputers: aion cluster

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

- Atos/AMD supercomputer, DLC cooling
 - 4 BullSequana XH2000 adjacent racks
 - 354 regular nodes, 45312 cores, 90.6 TB RAM
 - R_{peak} : 1,885 PetaFLOP/s
- Fast InfiniBand (IB) HDR network
 - Fat-Tree Topology blocking factor 1:2
- Acquisition by European Tender in 2020
 - production release in Oct 2021 (delayed by COVID)
 - First upgrade EOY 2022 +36 regular nodes

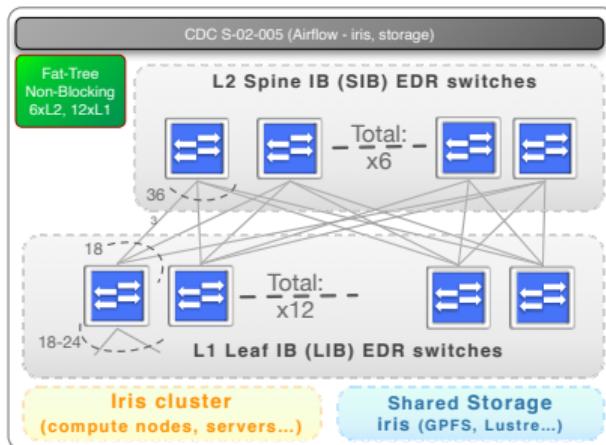


- In this talk:
 - design choices & config. changes when integrating aion, with performance evaluation

Fast Local Infiniband (IB) Interconnect Network

- before integration of aion (iris alone)

hpc-docs.uni.lu/interconnect/ib/

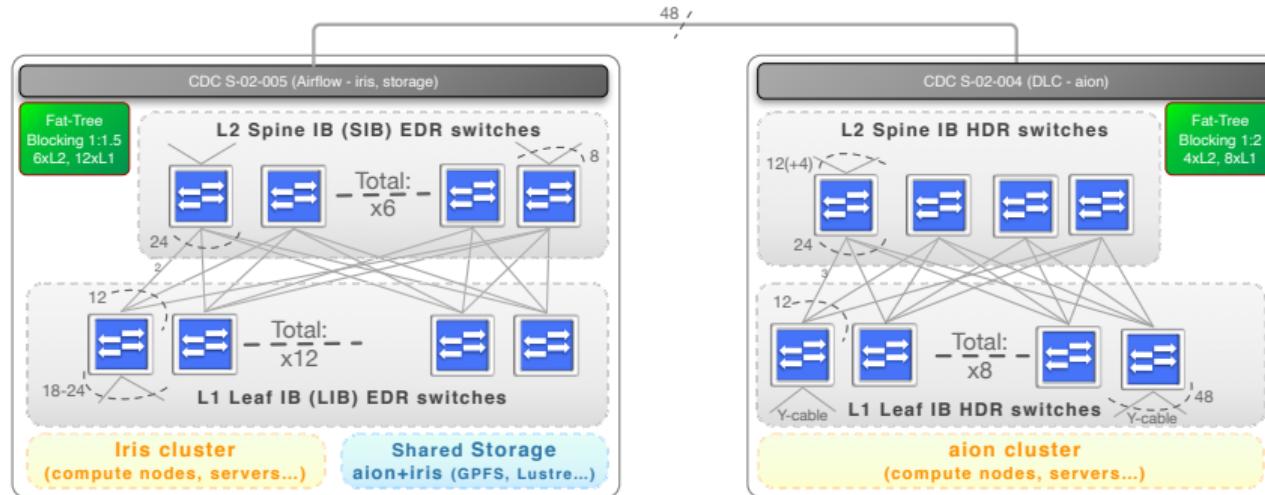


[PEARC22] S. Varrette, H. Cartiaux, T. Valette and A. Ollo, "Aggregating and Consolidating two High Performant Network Topologies: The ULHPC Experience" in ACM Practice and Experience in Advanced Research Computing (PEARC'22), Boston, USA, 2022.

Fast Local Infiniband (IB) Interconnect Network

- after merging iris and aion IB islands

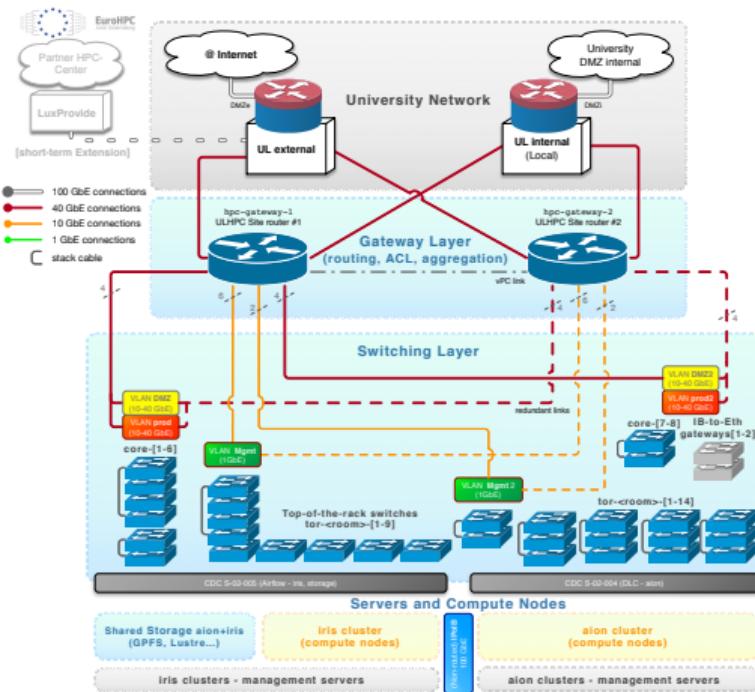
hpc-docs.uni.lu/interconnect/ib/



[PEARC22] S. Varrette, H. Cartiaux, T. Valette and A. Olloh, "Aggregating and Consolidating two High Performant Network Topologies: The ULHPC Experience" in ACM Practice and Experience in Advanced Research Computing (PEARC'22), Boston, USA, 2022.

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] core switches
 - ✓ TOR (Top-the-rack) switches
 - ✓ meant to interface HPC servers and compute nodes

[PEARC22] S. Varrette, H. Cartiaux, T. Valette and A. Ollo, "Aggregating and Consolidating two High Performant Network Topologies: The ULHPC Experience" in ACM Practice and Experience in Advanced Research Computing (PEARC'22)

UL HPC Storage Systems

- Two types of **distributed & parallel FS**
 - ↪ **IBM Spectrum Scale** (formelly GPFS)
 - ↪ **Lustre \$SCRATCH** storage
- Complementary storage infrastructure
 - ↪ **OneFS** (Dell/EMC Isilon)
 - ✓ project data, backup & archival

hpc-docs.uni.lu/filesystems/

File System	Vendor	#Disks	Raw/Effective capacity
GPFS (2017-)	DDN	710 HDDs + 38 SSDs	4260 / 3408 TB
Lustre (2018-)	DDN	Object Storage Targets: 167 HDDs Meta-Data Targets: 19 SSDs	1300 / 920 TB
OneFS (2014-)	Dell/EMC	n/a (NDA)	7100 / 6400 TB

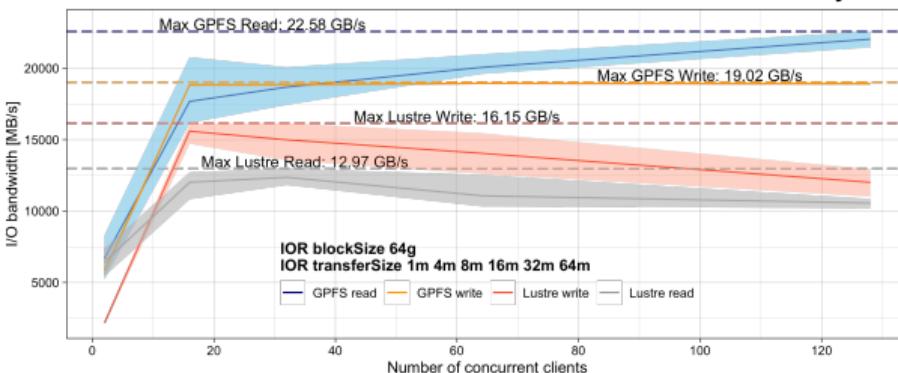
UL HPC Storage Systems

- Two types of **distributed & parallel FS**
 - IBM Spectrum Scale (formally GPFS)
 - Lustre \$SCRATCH storage
- Complementary storage infrastructure
 - OneFS** (Dell/EMC Isilon)
 - ✓ project data, backup & archival

hpc-docs.uni.lu/filesystems/

File System	Vendor	#Disks	Raw/Effective capacity
GPFS (2017-)	DDN	710 HDDs + 38 SSDs	4260 / 3408 TB
Lustre (2018-)	DDN	Object Storage Targets: 167 HDDs Meta-Data Targets: 19 SSDs	1300 / 920 TB
OneFS (2014-)	Dell/EMC	n/a (NDA)	7100 / 6400 TB

IOR v3.1.0 - MPI Coordinated Test of Parallel I/O on ULHPC Facility



UL HPC Storage Systems

- Two types of **distributed & parallel FS**
 - ↪ **IBM Spectrum Scale** (formelly GPFS)
 - ↪ **Lustre \$SCRATCH** storage
- Complementary storage infrastructure
 - ↪ **OneFS** (Dell/EMC Isilon)
 - ✓ project data, backup & archival
- EU's **GDPR** (*General Data Protection Regulation*) and Open Science compliance [APF21]

hpc-docs.uni.lu/filesystems/

[APF21] L. Paseri, S. Varrette, "Protection of Personal Data in HPC Platform for Scientific Research Purposes", in Proc. of the EU Annual Privacy Forum (APF) 2021, LNCS vol. 12703, pp. 123–142.

UL HPC Storage Systems

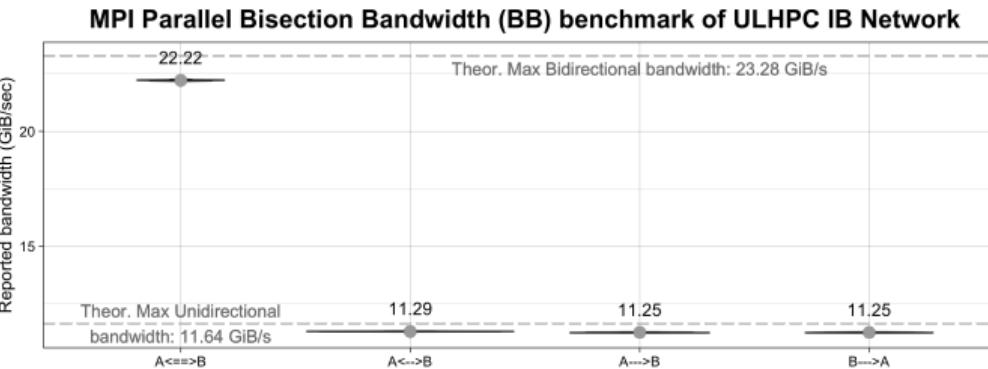
- Two types of **distributed & parallel FS**
 - ↪ **IBM Spectrum Scale** (formelly GPFS)
 - ↪ **Lustre \$SCRATCH storage**
- Complementary storage infrastructure
 - ↪ **OneFS** (Dell/EMC Isilon)
 - ✓ project data, backup & archival
- EU's **GDPR** (*General Data Protection Regulation*) and Open Science compliance [APF21]
- Specific **quota and purging policy** depending on usage pattern/sustaining FS

hpc-docs.uni.lu/filesystems/

	Directory	File System	Backup	Default Quota	Default Inode quota	Purging time
\$HOME	/home/users/<login>	GPFS/Spectrumscale	yes (daily)	500 GB	1 M	-
	/work/projects/<name>		yes (daily)			
\$SCRATCH	/scratch/users/<login>	Lustre	no	10 TB	1 M	60 days
	/mnt/isilon/projects/<name>		yes (snapshot, weekly)			
				1.14 PB globally	-	-

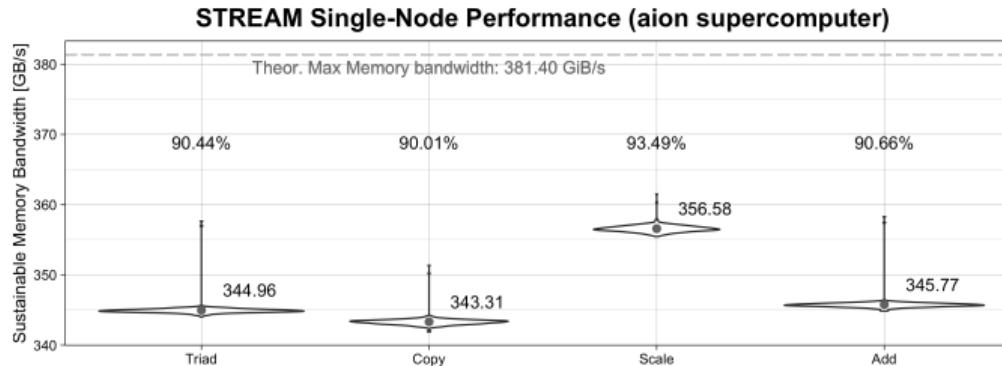
UL HPC Performance Evaluations [selected benchs]

- Bisection Bandwidth (**BB**) benchmarks: **96,99% efficiency**



UL HPC Performance Evaluations [selected benchmarks]

- Bisection Bandwidth (**BB**) benchmarks: **96,99% efficiency**
- **STREAM** sustainable **Memory Bandwidth** performance
 - ↪ above **90,01%** efficiency for 4 highly-intensive memory access pattern



UL HPC Performance Evaluations [selected benchs]

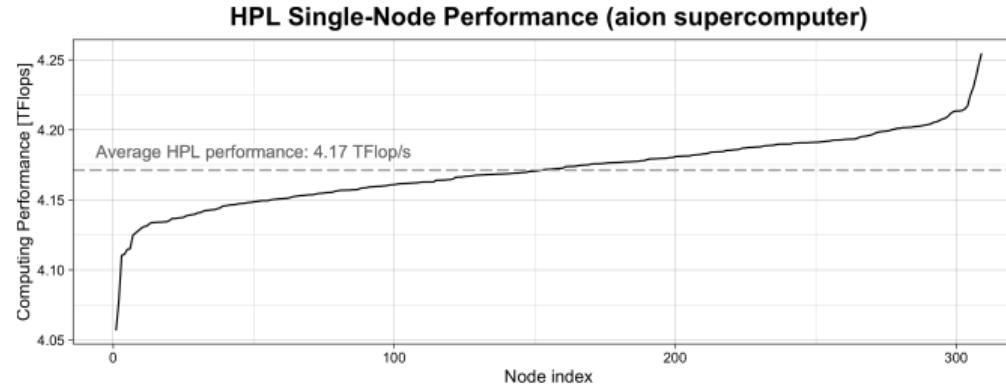
- Bisection Bandwidth (**BB**) benchmarks: **96,99% efficiency**
- **STREAM** sustainable **Memory Bandwidth** performance
 - ↪ **above 90,01% efficiency** for 4 highly-intensive memory access pattern
- **Single-node HPL** performance

Processor/GPU Model	#Cores	Freq.	R _{peak}	Avg. R _{max}
AMD ROME 7H12 (<i>epyc</i>)	64	2.6 GHz	2.66 TFlops	2.09 TFlops
Intel Xeon E5-2680v4 (<i>broadwell</i>)	14	2.4 GHz	0.54 TFlops	0.46 TFlops
Intel Xeon Gold 6132 (<i>skylake</i>)	14	2.3 GHz*	1.03 TFlops	0.94 TFlops
Intel Xeon Platinum 8180M (<i>skylake</i>)	28	2.3 GHz*	2.06 TFlops	1.76 TFlops
NVidia Tesla V100-SXM2	5120+640	1.3 GHz	7.80 TFlops	5.59 TFlops

*: AVX-512 Turbo Frequency

UL HPC Performance Evaluations [selected benchmarks]

- Bisection Bandwidth (**BB**) benchmarks: **96,99% efficiency**
- **STREAM** sustainable **Memory Bandwidth** performance
 - ↪ **above 90,01% efficiency** for 4 highly-intensive memory access pattern
- **Single-node HPL** performance (sorted distribution within aion nodes)



UL HPC Performance Evaluations [selected benchmarks]

	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%	#71 (Jun 2022) #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%	#31 (Jun 2022) #23 (Jun 2021)
	GreenGraph500	2 ⁸ =256		6.14 MTEPS/W		+180%	#43 (Jun 2022) #36 (Jun 2021)
*: performance improvement with the minimal acceptance threshold set in the Aion tender document							
Iris	IO500 (isc21 release)	128		11.345219			#42 (Nov 2020 - latest release)
	HPL (CPU/broadwell)	108		84.75 TFlops	72.98%		
	HPL (GPU/V100 16G)	72	(NB=320, P×Q=12×6)	283.6 TFlops	52.87%		
	HPCG (GPU/V100 16G)	72		8.74 TFlops			
	HPL (GPU/V100 32G)	24	(NB=288, P×Q=6×4)	135.2 TFlops	75.61%		
	HPCG (GPU/V100 32G)	24		2.90 TFlops			

- Reference benchmarks: **HPL, HPCG, Graph500, Green500, GreenGraph500**
 - I/O specific: **IO500, IOR**
- (not presented) Unified European Application Benchmark Suite (**UEABS**)

Summary

1 Overview of the Managed Facility

Network Organisation

Tiered Shared Storage infrastructure

Computing Performance Evaluation and Acceptance Tests

2 User Software Environment

3 User Job Management and the Slurm infrastructure

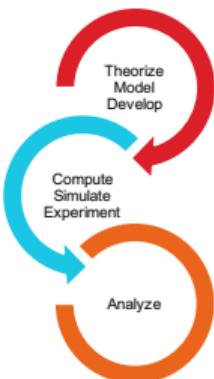
4 Conclusion and Perspectives

Accelerating Research - User Software Sets

- Over 280 software packages available for researchers

- software environment generated using **RESIF 3.0 framework** [PEARC21] over Easybuild
 - ✓ optimized builds organized by architecture, exposed through Environment Modules/Lmod
 - ✓ Categorized Naming Scheme

<category>/<name>/<version>-<toolchain><versionsuffix>



Component	Software set release <version>		
	2019b legacy	2020b prod	2021b devel
binutils	2.32	2.35	2.37
GCCCore	8.3.0	10.2.0	11.2.0
foss	2019b	2020b	2021b
- OpenMPI	3.1.4	4.0.5	4.1.2
intel	2019b	2020b	2021a
- Compilers/MKL	2019.5.281	2020.1.217	2021.4.0
- Intel MPI	2018.5.288	2019.7.217	2021.4.0
Python	3.7.4	3.8.6	3.9.6
RESIF version	3.0	3.0	3.1
#Software Modules	<arch>: 269 gpu: 135	<arch>: 274 gpu: 151	<arch>: 282 gpu: 157

[PEARC21] S. Varrette, E. Kieffer, F. Pinel, E. Krishnasamy, S. Peter, H. Cartiaux, and X. Besseron. "RESIF 3.0: Toward a Flexible & Automated Management of User Software Environment on HPC facility". In ACM Practice & Experience in Advanced Research Computing (PEARC'21) [pdf](#) – [code](#)

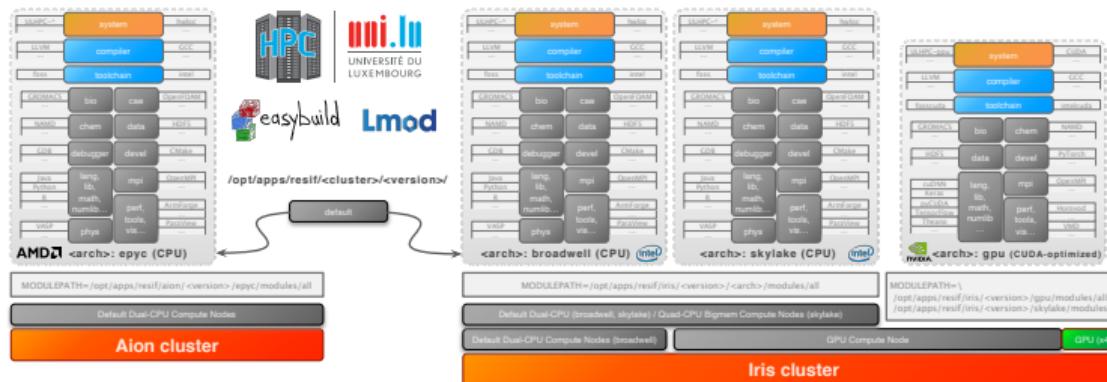
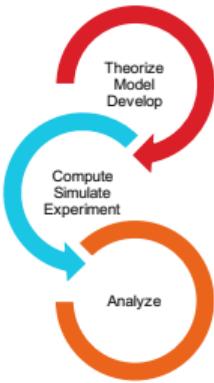
S. Varrette & al. (Univ. of Luxembourg)

Management of an Academic HPC & Research Computing Facility: The ULHPC Experience 2.0

Accelerating Research - User Software Sets

- Over 280 software packages available for researchers

- software environment generated using **RESIF 3.0 framework** [PEARC21] over Easybuild
 - optimized builds organized by architecture, exposed through Environment Modules/Lmod
 - Categorized Naming Scheme



[PEARC21] S. Varrette, E. Kieffer, F. Pinel, E. Krishnasamy, S. Peter, H. Cartiaux, and X. Besseron. "RESIF 3.0: Toward a Flexible & Automated Management of User Software Environment on HPC facility". In ACM Practice & Experience in Advanced Research Computing (PEARC'21) [pdf](#) – [code](#)

S. Varrette & al. (Univ. of Luxembourg)

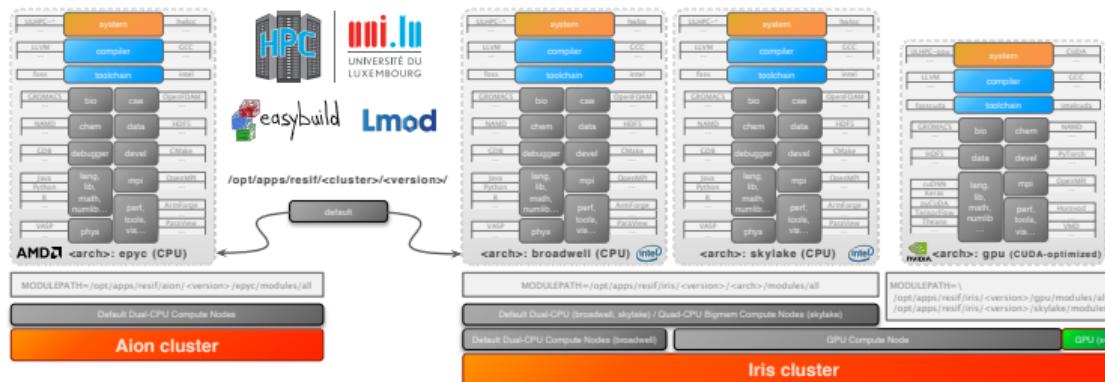
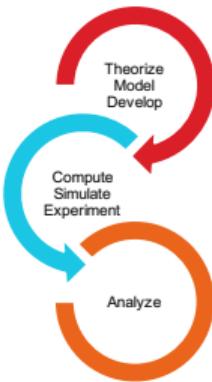
Management of an Academic HPC & Research Computing Facility: The ULHPC Experience 2.0

Accelerating Research - User Software Sets

- Over 280 software packages available for researchers

- software environment generated using **RESIF 3.0 framework** [PEARC21] over Easybuild
 - optimized builds organized by architecture, exposed through Environment Modules/Lmod
 - Categorized Naming Scheme

<category>/<name>/<version>-<toolchain><versionsuffix>



- containerized applications delivered with Singularity system
- user web/application portal (outside regular SSH access): Open OnDemand

Summary

1 Overview of the Managed Facility

Network Organisation

Tiered Shared Storage infrastructure

Computing Performance Evaluation and Acceptance Tests

2 User Software Environment

3 User Job Management and the Slurm infrastructure

4 Conclusion and Perspectives

ULHPC Slurm Partitions and QOS 2.0

AION Partition	Type	#Node	PriorityTier	DefaultTime	MaxTime	MaxNodes
interactive	floating	318	100	30min	2h	2
	batch	318	1	2h	48h	64
IRIS Partition	Type	#Node	PriorityTier	DefaultTime	MaxTime	MaxNodes
interactive	floating	196	100	30min	2h	2
	batch	168	1	2h	48h	64
gpu		24	1	2h	48h	4
	bigmem	4	1	2h	48h	1

[ISPDC22] S. Varrette, E. Kieffer and F. Pinel, "Optimizing the Resource and Job Management System of an Academic HPC and Research Computing Facility" in 21st IEEE Intl. Symp. on Parallel and Distributed Computing (ISPDC'22)

ULHPC Slurm Partitions and QOS 2.0

AION Partition	Type	#Node	PriorityTier	DefaultTime	MaxTime	MaxNodes	
interactive	floating	318	100	30min	2h	2	
batch		318	1	2h	48h	64	
IRIS Partition	Type	#Node	PriorityTier	DefaultTime	MaxTime	MaxNodes	
interactive	floating	196	100	30min	2h	2	
batch		168	1	2h	48h	64	
gpu		24	1	2h	48h	4	
bigmem		4	1	2h	48h	1	
QOS	Partition	Allowed [L1] Account	Prio	GrpTRES	MaxTresPJ	MaxJobPU	Flags
besteffort	*	ALL	1		100		NoReserve
low	*	ALL (default for CRP/externals)	10		2		DenyOnLimit
normal	*	Default (UL,Projects,...)	100		50		DenyOnLimit
long	*	UL,Projects,etc.	100	node=12	node=2	4	DenyOnLimit,PartitionTimeLimit
debug	interactive	ALL	150	node=8		2	DenyOnLimit
high	*	(restricted) UL,Projects,Industry	200			10	DenyOnLimit
urgent	*	(restricted) UL,Projects,Industry	1000			100	DenyOnLimit

[ISPDC22] S. Varrette, E. Kieffer and F. Pinel, "Optimizing the Resource and Job Management System of an Academic HPC and Research Computing Facility" in 21st IEEE Intl. Symp. on Parallel and Distributed Computing (ISPDC'22)

Fairsharing and Accounting 2.0

- New configuration with Multifactor Priority Plugin and Fair tree algorithm
 - migration from Depth-Oblivious Fair-share (*initial setup*)
 - new jobs are immediately assigned a priority fairshare levels

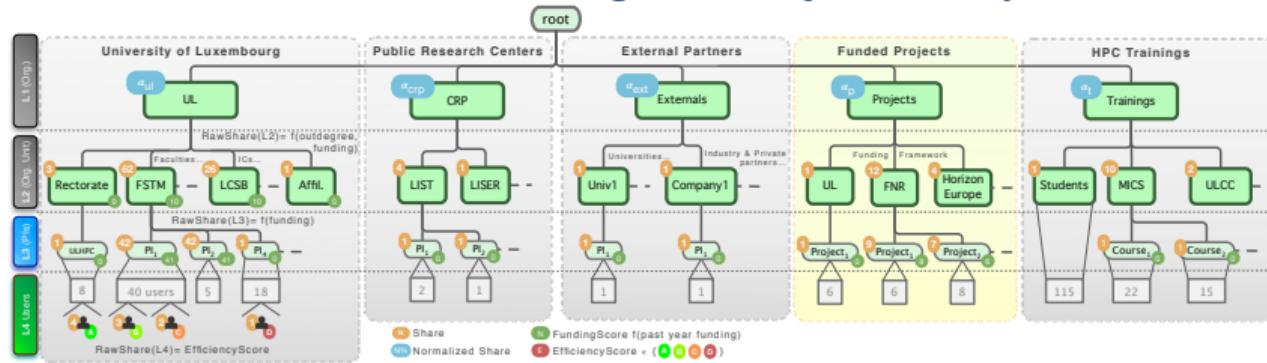
[ISPDC22] S. Varrette, E. Kieffer and F. Pinel, "Optimizing the Resource and Job Management System of an Academic HPC and Research Computing Facility" in 21st IEEE Intl. Symp. on Parallel and Distributed Computing (ISPDC'22)

S. Varrette & al. (Univ. of Luxembourg)

Management of an Academic HPC & Research Computing Facility: The ULHPC Experience 2.0

Fairsharing and Accounting 2.0

- New configuration with Multifactor Priority Plugin and Fair tree algorithm
 - ↪ migration from Depth-Oblivious Fair-share (*initial setup*)
 - ↪ new jobs are immediately assigned a priority fairshare levels
- Accounting records re-organized as a hierarchical tree (3 layers $L_{1,2,3}$ + leafs)
 - ↪ raw share attribution based on funding score and job efficiency



[ISPDC22] S. Varrette, E. Kieffer and F. Pinel, "Optimizing the Resource and Job Management System of an Academic HPC and Research Computing Facility" in 21st IEEE Intl. Symp. on Parallel and Distributed Computing (ISPDC'22)

Fairsharing and Accounting 2.0

- New configuration with Multifactor Priority Plugin and Fair tree algorithm
 - migration from Depth-Oblivious Fair-share (*initial setup*)
 - new jobs are immediately assigned a priority fairshare levels
- Accounting records re-organized as a hierarchical tree (3 layers $L_{1,2,3}$ + leafs)
 - raw share attribution based on funding score and job efficiency

Impact of the new Slurm configuration

- Daily utilization increased by 12.64% to reach 81.56% of available resources
 - measures from workload traces over several months of uninterrupted HPC services
- Wall-time Request Accuracy (WRA) of processed jobs increased by 110,81%
 - evaluation covering 1 year period before and after configuration change
- UL HPC budget incomes increased in 2021 by 10%

[ISPDC22] S. Varrette, E. Kieffer and F. Pinel, "Optimizing the Resource and Job Management System of an Academic HPC and Research Computing Facility" in 21st IEEE Intl. Symp. on Parallel and Distributed Computing (ISPDC'22)

Summary

1 Overview of the Managed Facility

Network Organisation

Tiered Shared Storage infrastructure

Computing Performance Evaluation and Acceptance Tests

2 User Software Environment

3 User Job Management and the Slurm infrastructure

4 Conclusion and Perspectives

Conclusion

- **In this talk:**

- **Design choices** when **acquiring & integrating** a new supercomputer aion
 - ✓ smooth integration within the existing HPC ecosystem
- **Overview of the managed HPC facility**
 - ✓ supercomputer architectures, network organization, tiered shared storage infrastructure
 - ✓ HPC performance evaluation
- **User software environment** & Resource and Job Management System (**RJMS**) adaptation

Conclusion

- **In this talk:**
 - ↪ Design choices when **acquiring & integrating** a new supercomputer aion
 - ✓ smooth integration within the existing HPC ecosystem
 - ↪ **Overview of the managed HPC facility**
 - ✓ supercomputer architectures, network organization, tiered shared storage infrastructure
 - ✓ HPC performance evaluation
 - ↪ **User software environment & Resource and Job Management System (RJMS) adaptation**
- *Not covered here:*
 - ↪ Data center design and characteristics
 - ↪ DevOps Software stack for research computing services management
 - ✓ based on **Puppet** and **Ansible** (**Bluebanquise stack**)

Conclusion

- **In this talk:**

- Design choices when **acquiring & integrating** a new supercomputer aion
 - ✓ smooth integration within the existing HPC ecosystem
- **Overview of the managed HPC facility**
 - ✓ supercomputer architectures, network organization, tiered shared storage infrastructure
 - ✓ HPC performance evaluation
- **User software environment** & Resource and Job Management System (**RJMS**) adaptation

- **Not covered here:**

- Data center design and characteristics
- DevOps Software stack for research computing services management
 - ✓ based on **Puppet** and **Ansible** (**Bluebanquise stack**)

- **Perspectives and Future directions**

- **smooth integration with Euro-HPC infrastructures**
 - ✓ transparently outsource Research Computing/data analytic workflows to Tier-0 systems
- **automatically offload of less-demanding jobs onto virtual cloud resources**

Thank you for your attention...



Questions?

**Sebastien Varrette, Hyacinthe Cartiaux, Sarah Peter, Emmanuel Kieffer,
Teddy Valette and Abatcha Olloh**

Management of an Academic HPC & Research Computing Facility: The ULHPC Experience 2.0 – ACM HPCCT 2022

University of Luxembourg, Belval Campus
Maison du Nombre, 4th floor
2, avenue de l'Université
L-4365 Esch-sur-Alzette
mail: firname.lastname@uni.lu

High Performance Computing @ Uni.lu

mail: hpc@uni.lu

1 Overview of the Managed Facility

Network Organisation
Tiered Shared Storage infrastructure
Computing Performance Evaluation and Acceptance Tests

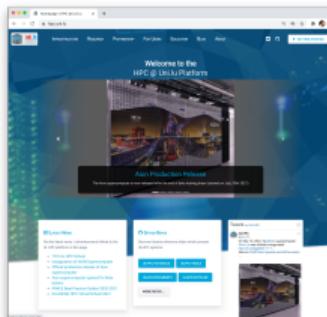
2 User Software Environment

3 User Job Management and the Slurm infrastructure

4 Conclusion and Perspectives

High Performance Computing @ Uni.lu

hpc.uni.lu



ULHPC Technical Docs

hpc-docs.uni.lu