LUMI Introduced: Opportunities and limitations

Kurt Lust
LUMI User Support Team (LUST)
University of Antwerp

October 2023

Contents

LUMI Consortium

- LUMI consortium
- LUMI hardware
- LUMI software environment
- Requesting access to LUMI
- LUMI support and training

LUMI Consortium

- EuroHPC project
  50% European funding, 50% countries
- Unique consortium of 10 countries with strong national HPC centres (soon 11 countries)
- The resources of LUMI are allocated per the investments
- The share of the EuroHPC JU (50%) is allocated by a peer-review process (cf. PRACE Tier-0 access) and available for all European researchers
- Belgian share is 7.4% for all current resource categories

LUMI, the Queen of the North

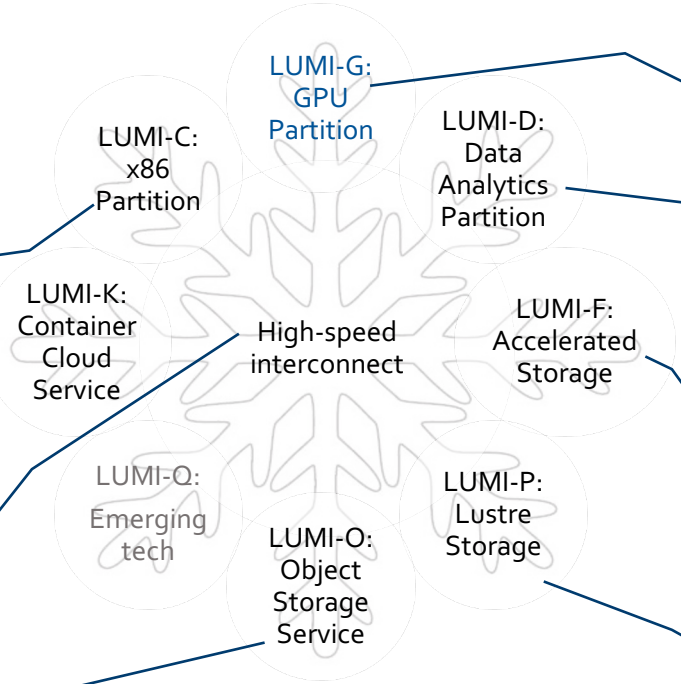
LUMI
hardware

LUMI is a Tier-0 ***GPU-accelerated supercomputer*** *that enables the convergence of* ***high-performance computing***, ***artificial intelligence***, *and* ***high-performance data analytics.***

- Supplementary CPU partition
- ~260,000 AMD EPYC CPU cores

Possibility for combining different resources within a single run. HPE Slingshot technology.

30 PB encrypted object storage (Ceph) for storing, sharing and staging data



Tier-0 GPU partition: over ~~550~~ 490 Pflop/s FP64 vector performance powered by AMD Instinct GPUs

Interactive partition with 4 TB CPU-only and 2 TB NVIDIA GPU nodes for data analytics and visualization

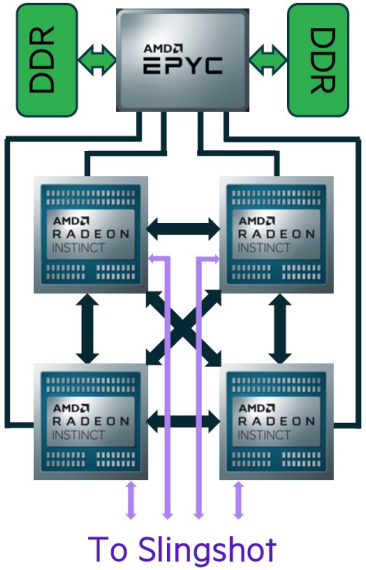
8.5 PiB Flash-based storage layer with extreme I/O bandwidth of 2 TB/s and IOPS capability. Cray ClusterStor E1000.

4 x 18 PiB parallel file system. Cray CusterStor E1000

LUMI compute node configurations

LUMI hardware

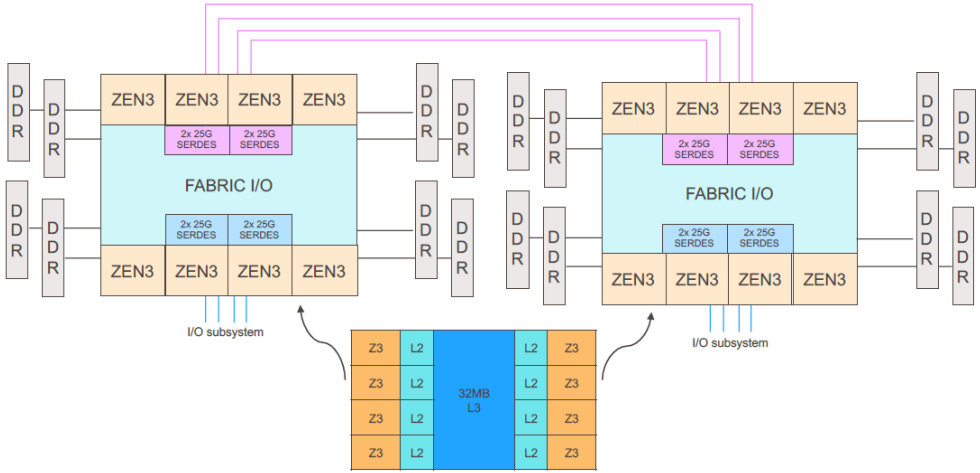
LUMI-G (marketing view)



2978 nodes with 1 64-core AMD Trento CPU, 4 MI250x and 512+4x128 GB memory

- But 8 cores and 32 GB reserved
- Compute GPU, not render GPU

LUMI-C



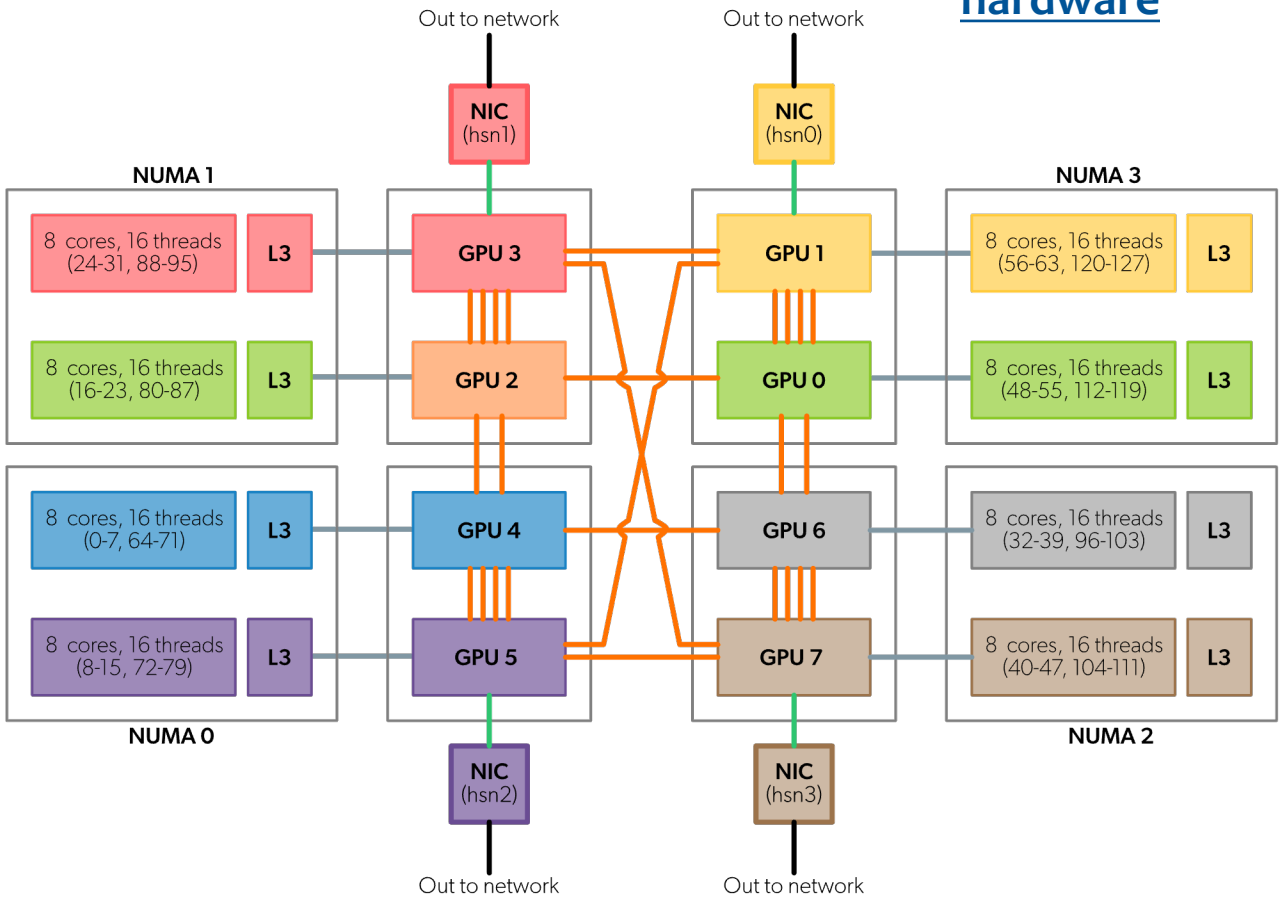
2x 64-core AMD Milan processors per node
1888 nodes with 256 GB, 128 with 512 GB and 32 with 1 TB

- But 32 GB reserved

Real LUMI-G node

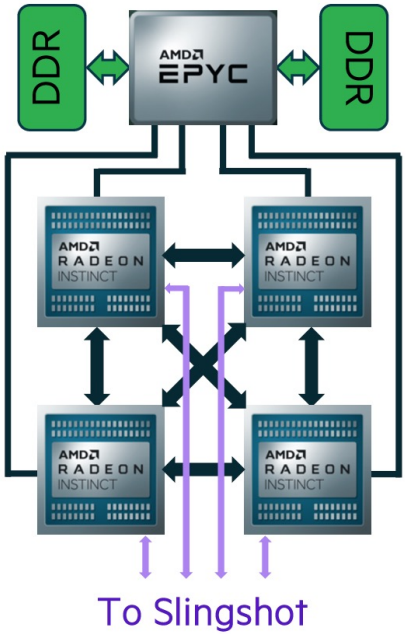
LUMI hardware

- 4 GPUs behave as 8 with 64 GB each
- Bandwidth between the GPU dies is low
- Binding to the CCDs is important for performance: Each GPU die closely associated to an L3 cache region
- Read the docs and/or take a course!

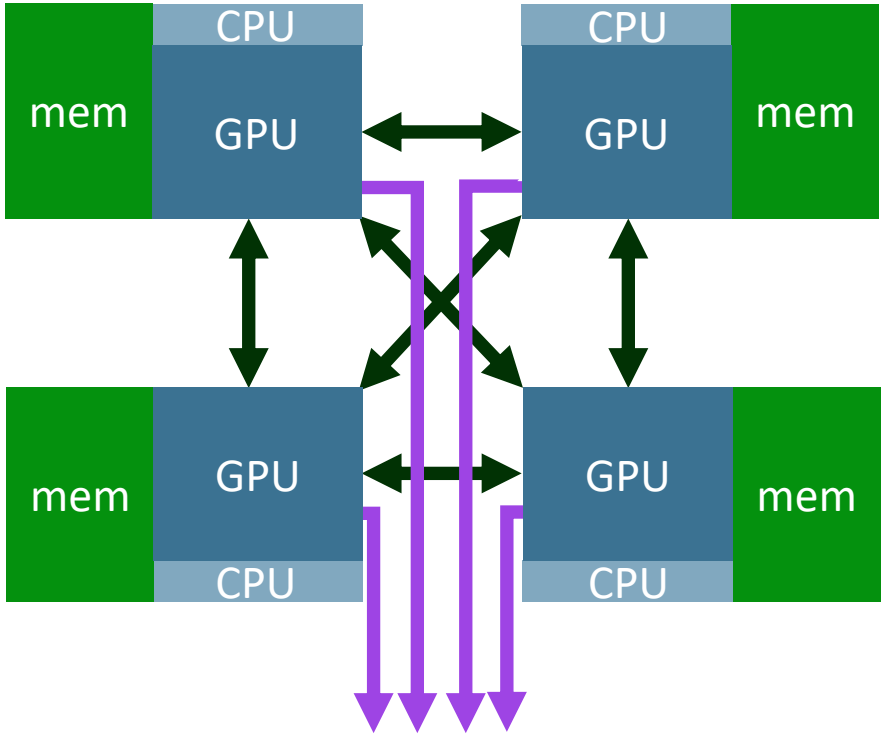


The future according to AMD

LUMI hardware



- LUMI (marketing view)



- MI300A APU for El Capitan (marketing view???)

LUMI file systems

LUMI
hardware

- LUMI-P and LUMI-F are based on Lustre
  - 32 OSTs on LUMI-P, 72 OSTs on LUMI-F
  - Needs user help for good performance or you'll be using just a fraction of the bandwidth => Important topic in our courses!
- LUMI-P definitely doesn't like lots of small files
- LUMI-F still open for arbitration
  - Claims high bandwidth and high IOPS
  - But somewhere the restrictions of a parallel file system will show up
- LUMI-O object storage
  - Good as an intermediate station for large data transfers
  - Direct use from a job currently less practical due to sometimes long wait times in the queue and short expiration time of keys

What do we work with software-wise?

LUMI
software

- LUMI is built for extreme scalability in the first place
- Compute nodes run Cray OS (COS), derived from SUSE 15 SP4
  - Reducing OS jitter is a central element (and introduces restrictions)
  - Yet needed to reserve some CPU cores for the OS on the GPU nodes
- Cray PE, but not the same components as on Lucia
  - Compilers: Cray, GNU, AMD AOCC + ROCm. No Intel.
    - ROCm version tends to lag, follow the LTS system software distribution from HPE
  - HPE Cray LibSci + accelerator version + FFTW + netCDF/HDF5
  - Cray MPICH, using libfabric, with GPU-aware MPI. No UCX!
  - Cray performance analysis and debugging tools
- Additional tools from AMD, Linaro, …
- Slurm scheduler

Software stack design considerations

LUMI
software

- Very leading edge and inhomogeneous machine (new interconnect, new GPU architecture with an immature software ecosystem, some NVIDIA GPUs for visualisation, a mix of zen2 and zen3)
  - Slingshot interconnect and AMD GPUs are both challenging
  - Need to remain agile
- Users that come via 11 different channels (+subchannels), with different expectations
- Small central support team considering the expected number of projects and users and the tasks the support team has
- Cray Programming Environment is a key part of our system, need to integrate with that
- Need for customised setups
  - Everybody wants a central stack as long as their software is in there but not much more
  - Large and diverse user community, cannot settle on a single version of everything
  - Look at the success of conda, Python virtual environments, containers, …

The LUMI solution

LUMI
software

- Software organised in extensible software stacks based on a particular release of the PE
  - Many base libraries and some packages already pre-installed
  - Easy way to install additional packages in project space
- Modules managed by Lmod
  - More powerful than the (old) Modules Environment
  - Powerful features to search for modules
- EasyBuild is our primary tool for software installations
  - But uses HPE Cray specific toolchains
  - Offer a library of installation recipes
  - User installations integrate seamlessly with the central stack
  - We do have a Spack setup but don't do development in Spack ourselves

Policies

LUMI software

- Bring-your-own-license except for a selection of tools that are useful to a larger community
  - There is even no proper mechanism to control access to packages globally due to the distributed user management system
  - Even for software on the system, users remain responsible for checking the license!
- LUST tries to help with installations of recent software, but porting or bug fixing is not their work
  - Not all Linux or even supercomputer software will work on LUMI
  - LUST is too small a team to do all software installations, so don't count on them to do all the work
- Conda, (large) Python installations need to go in containers
  - LUMI offers lumi-container-wrapper and cotainr to do that, and some pre-built AI containers

Getting access for open research

LUMI access

- For Belgian users via LUMI-BE and EuroHPC
  - LUMI-BE tries to sit between the Tier-1 systems and EuroHPC allocations
  - Repeat users are expected to try the EuroHPC channel also
- Separate channels for
  - Benchmarking to prepare a regular project application (4 months in most cases)
  - Software development (1 year)
  - Regular projects (1 year)
  - EuroHPC extreme scale projects (1 year) aimed primarily at projects with very scalable codes and very large simulations
- CPU-only resources are limited compared to Tier-1 in Flanders and Wallonia
  - Belgian share basically the equivalent of a 150 node cluster

Getting access for open research (2)

LUMI access

- Access via project applications
  - Scientific and technical review for EuroHPC regular and extreme scale projects
  - Technical review only for LUMI-BE projects if there is a corresponding reviewed science project already
  - Research not open: see next slide
  - *Participation in the promotion expected as we need to show funding agencies that access to large machines is important*

Become a company user

LUMI

- Private–Public collaboration project
  - Project in cooperation with Belgian university or research institution (academic partner)
  - Free of charge if results are open and published. If the results are owned by the company, the use will be charged according to the pricelist of LUMI computing services.
- Pay per use model
  - Company signs a contract and pays for the resources according to the pricelist. The results of projects are owned by the company and they don´t need to be public.

Become a company user (2)

LUMI

- Test before use (“Try&Buy”) (cf. “VSC’s Exploratory access”)
  - A company can get familiar with and test the suitability of LUMI computing services for the intended purpose free-of-charge through a “Try&Buy” -project.
  - The company project is created with two user accounts.
  - Project will have limited CPU-, GPU-, and data storage resources for testing purposes equivalent to a preparatory project.
  - Expert support will be available to get started with the LUMI usage, provided there is some effort from the company itself as well to share knowledge. (Same model as the former PRACE SHAPE program.)
  - The testing project is available for a limited time.

Become a company user (3)

LUMI

- Caveat
  - Meant for research and development, not for production runs.
  - Not all (commercial) software might run due to the specific hardware and software environment.
  - LUMI is a technology development platform and does not have the same level of stability and quality-of-service as clusters based on established technology.
  - No data backup and not the level of redundancy that commercial providers offer.
  - System software on LUMI follows the "long-term service releases" from HPE which implies that the latest and greatest may not always be present.
  - Only basic support included (getting started, issues while running jobs, …).
  - All jobs independent of the access channel enter in a batch queue. No special priority for paid use.

Billing

LUMI access

- Basic idea: You are billed for all resources that someone else cannot use.
- CPU billing units (in core·hours) for the CPU-only nodes
  - Billing based on cores reserved AND memory consumption in slices of 2 GB.
- GPU billing units (in GPU·hours) for the GPU nodes
  - GPU hour definition based on the 4 MI250X GPUs per node rather than the 8 seen in Slurm
  - Billing based on number of GPUs as seen by Slurm, number of cores (groups of 7) and CPU memory consumption (slices of 60 GB)
- Storage use is also billed
  - Idea is that someone may need a large quota but can clean up after a run
    - Quotas limit maximum use, storage billing limits average use
  - In TB·hours: Using 1 TB for one hour costs 1 TB·hour on LUMI-P storage
    - x 10 for flash storage on LUMI-F (so 10 TB·hours if you use 1 TB for 1 hour)
    - x 0.5 for object storage on LUMI-O
  - Don't be too greedy: We can use 3.5-4 PB on average on LUMI-P and 0.4 PB on LUMI-F

Billing – Paid use

LUMI access

Service	Price
LUMI computing project base package	1,000.00 €
LUMI-C – computing nodes with CPU (AMD Milan) *1 CPU-node·hour (node·h) equals to 128 CPU-core·hours billing units*	0.57 € / CPU-node·h
LUMI-G – computing nodes with GPU (AMD MI250x) *Same definition as on the previous slide*	0.535 € / GPU·h
LUMI-P – Disk-based Lustre parallel file system	0.005 € / TiB·h
LUMI-F – Flash-based Lustre parallel file system	0.05 € / TiB·h
LUMI-O – CEPH object storage	0.0025 € / TiB·h

- Caveat:
  - CPU/GPU billing based on core-, memory- and GPU use, whichever is proportionally the largest
  - All storage use is reported in storage billing units which are TiB·h for LUMI-P

Pay attention to…

LUMI access

- Strict file quota on LUMI (block quota are flexible)
  - Not the machine to dump your big data set as millions of small files
- Strict limits on number of jobs in Slurm
  - and array job counts as many jobs,
  - so capacity computing users will need an additional level to manage their jobs.
- Software that comes as binaries does not always work on LUMI
  - MPI usually the culprit, GPU driver could be a problem also
- Only Cray MPICH, no Open MPI
  - And process starter is srun, no mpirun/mpiexec
- Cray CPE uses universal compiler wrappers for each language, no mpicc etc.
- VSC users: No Torque wrappers for Slurm, no vsc-mympirun
- No ssh to compute nodes
- Limits are the same for company use also as they are the result of limitations of technology

Pay attention to… (2)

LUMI access

- GUI applications can be problematic (due to COS constraints)
- GPU programming
  - HIP and OpenMP offload are the prefered models
  - OpenACC only in Cray Fortran
  - OpenCL support unclear
  - Try to support AdaptiveCpp (formerly hipSYCL) and looking at Intel DPC++ also (as we can get support for that)
  - **NO CUDA!**
- Check software compatibility before you apply
  - LUMI-BE and LUST are there to help advise you

LUMI support

LUMI support

- Distributed support system
- Allocations: Support comes from the organisation that granted the allocation
  - Belgian allocation: lumi-be-support@enccb.be
- LUMI User Support Team offers L1 and L2 support (but cannot help with allocation problems)
  - Via web forms on lumi-supercomputer.eu/user-support/need-help
- Users in Flanders: VSC has a Tier-0 support project
  - Not yet fully staffed
  - Also via lumi-be-support@enccb.be
- Users in Wallonia: Some limited support via Orian Louant

LUMI support (2)

LUMI
support

- EuroHPC has granted the EPICURE project to set up a network for advanced L2 and L3 support across EuroHPC centres
  - Belgium participates as a partner in the LUMI consortium
- In principle the EuroHPC Centres of Excellence should play a role in porting of specific applications

Training

LUMI support

- System-specific trainings
  - LUST with HPE and AMD organise trainings
    - 1-day introductory
    - 4-day comprehensive with more attention on how to run efficiently and to development and profiling tools
    - Announcements on lumi-supercomputer.eu/events and various mailing lists and NCC web site
    - Training materials on lumi-supercomputer.github.io/LUMI-training-materials
  - Working on a BE version of the 1-day training
- Application-specific trainings
  - Should come from the EuroHPC centres of excellence and other organisations
- Application-on-system trainings
  - Small target group, but may have some about AI on LUMI in the future
- Awaiting a new EuroHPC training initiative to succeed PRACE…

Questions?

LUMI