Simplified Multi-Tenancy for Data Driven Personalized Health Research

Diego Moreno

HPC Storage Specialist @ Scientific IT Services, ETH Zürich

hpc-ch Forum on Storage Technologies and Data Management, Lugano

Agenda

- Scientific IT Services
- Personalized Health Research in Switzerland
- Leonhard: A cluster for Personalized Health Research
- Why Lustre?
- Multi-tenancy at ETH Zurich
- Evolution of Leonhard

Who am I

- 10 years of experience in the storage and HPC industry:
  - 5 years of Lustre R&D @ Atos, France
  - 2 years of Storage and Filesystem benchmarking @ Atos, France
  - 3 years of Storage and Filesystem L2 support and consulting @ DDN Storage, Worldwide
- Recently joined the HPC group @ Scientific IT Services
- My favourite topics: Lustre, filesystems, storage hardware and flash
- Now looking at clusters from the other side of the wall is exciting and challenging

Scientific IT Services

- Division of ETH IT Services dedicated to data management, analysis and other services for researchers
- Currently managing 2 centralized clusters for ETH's research community:

Euler

~ 2,270 nodes

1.5 PB DDN Lustre 2.7

1 PB NetApp FAS

General purpose HPC

Leonhard

~ 150 nodes

~ 600 GPUs

1.5 PB DDN GPFS -> Lustre

2 PB DDN Lustre 2.10

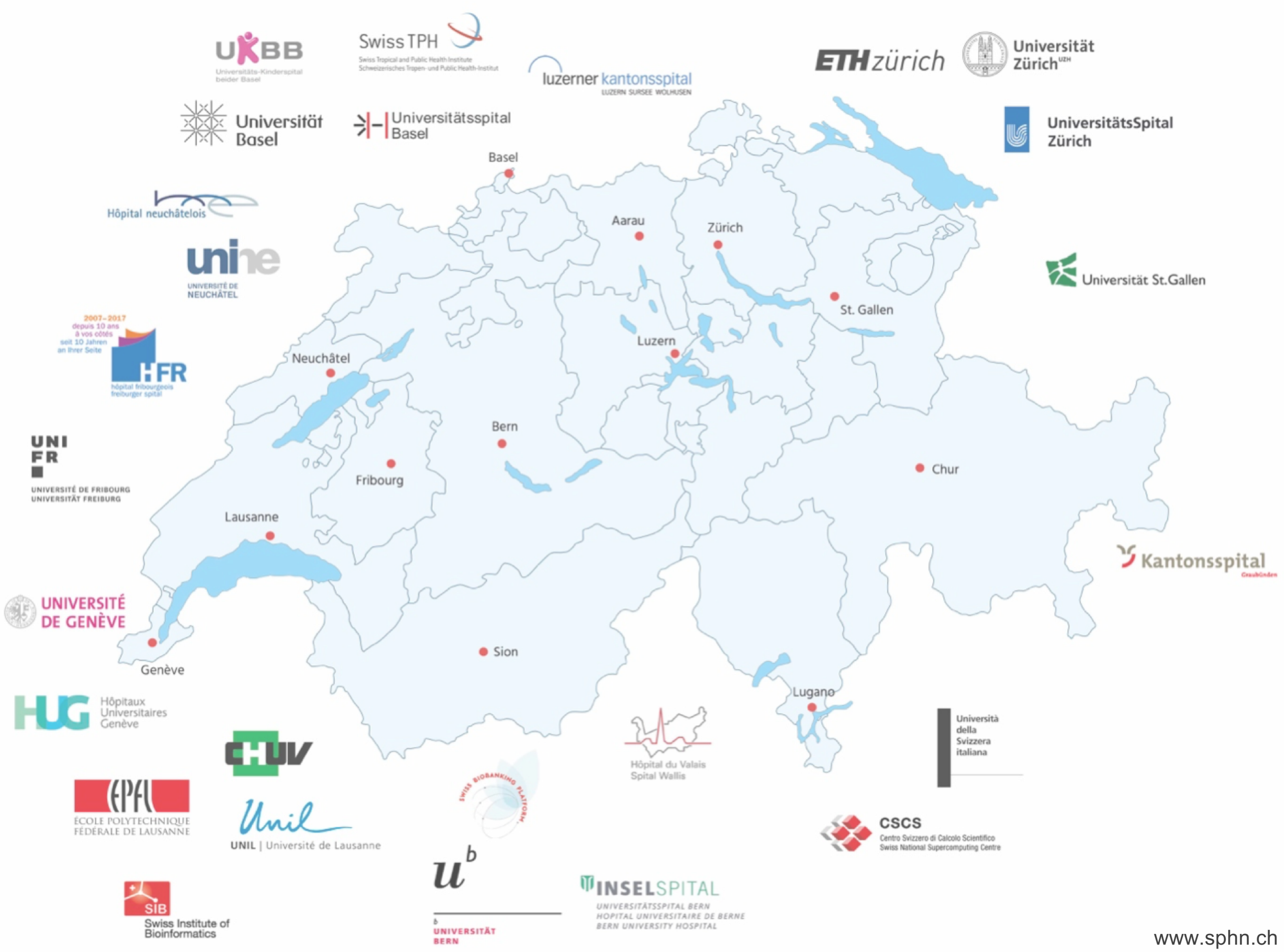
0.5 PB NetApp FAS

Data driven cluster for special projects

Data Driven Personalized Health in Switzerland

SPHN Swiss Personalized Health Network

ETH zürich Strategic Focus Area **Personalized Health and Related Technologies**



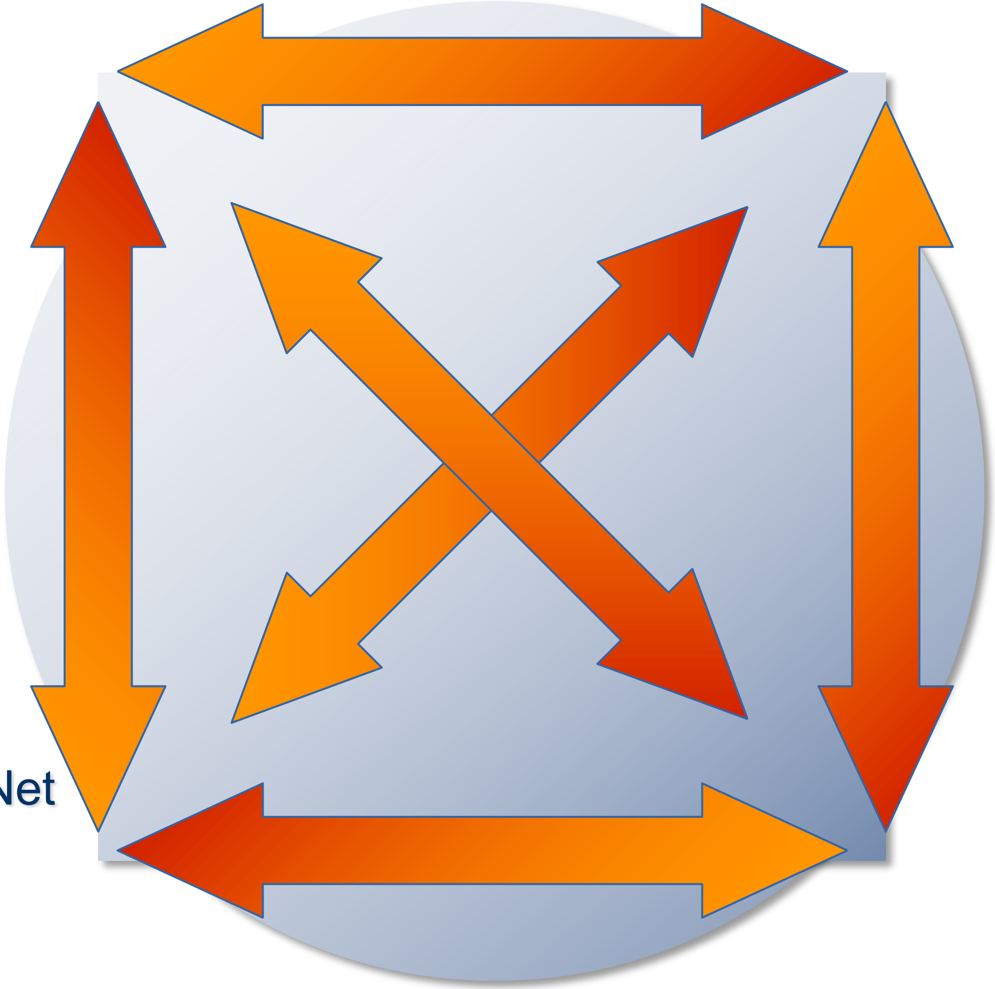
www.sphn.ch

Leonhard: From classic HPC to Health Research Informatics

Personalized Health Research cluster in the heart of Zurich

Leonhard – Challenge

Regulations
- Legal
- Ethical
- Best Practices
- CH, USA, EU



High Performance
- Fast Network
- GPUs
- Parallel Filesystems

Easy to use
- As on the notebook
- No security hassles
- Free access to the Net
- Interactive

Flexible
- Fast changes
- Cutting edge software
- State full nodes
- DB servers

Leonhard – Infrastructure Security

- Physical security
  - Leonhard is located in physically secured room, with access limited to specific persons.
- Network access control
  - Access to Leonhard is only possible through a DMZ, multifactor authentication required.
  - Access from Leonhard to the Internet is strictly controlled – no access to generic websites
- Logging and monitoring
  - Access and exit nodes are audited, to monitor all relevant user action
- Backup
  - Encrypted backup to tape. Data leaves Leonhard encrypted only.
- Multiple projects in parallel

Why Lustre?

Why Lustre?

Well, first it was GPFS…

Why Lustre?

Well, first it was GPFS…

- Choice initially driven by customers asking for GPFS encryption
- Well, they actually did not mean encryption but isolation…
- GPFS limitations on **this** setup (2017)
  - Maximum of 8 encryption keys per filesystem
  - No root squash in the GPFS local cluster
  - VMs: GPFS through NFS gateway vs Native Lustre client
  - Network isolation per tenant/project is hard to achieve
  - Network flexibility
  - Lustre multi-tenancy kicked in

Why Lustre?

Well, first it was GPFS…

- Choice initially driven by customers asking for GPFS encryption
- Well, they actually did not mean encryption but isolation…
- GPFS limitations on **this** setup (2017)
  - Maximum of 8 encryption keys per filesystem
  - No root squash in the GPFS local cluster
  - VMs: GPFS through NFS gateway vs Native Lustre client
  - Network isolation per tenant/project is hard to achieve
  - Network flexibility
  - Lustre multi-tenancy kicked in

Disclaimer: GPFS can be great, but likely not for this setup

Why Lustre?

Network flexibility

Performance

Scalability

Security

Multi-tenancy

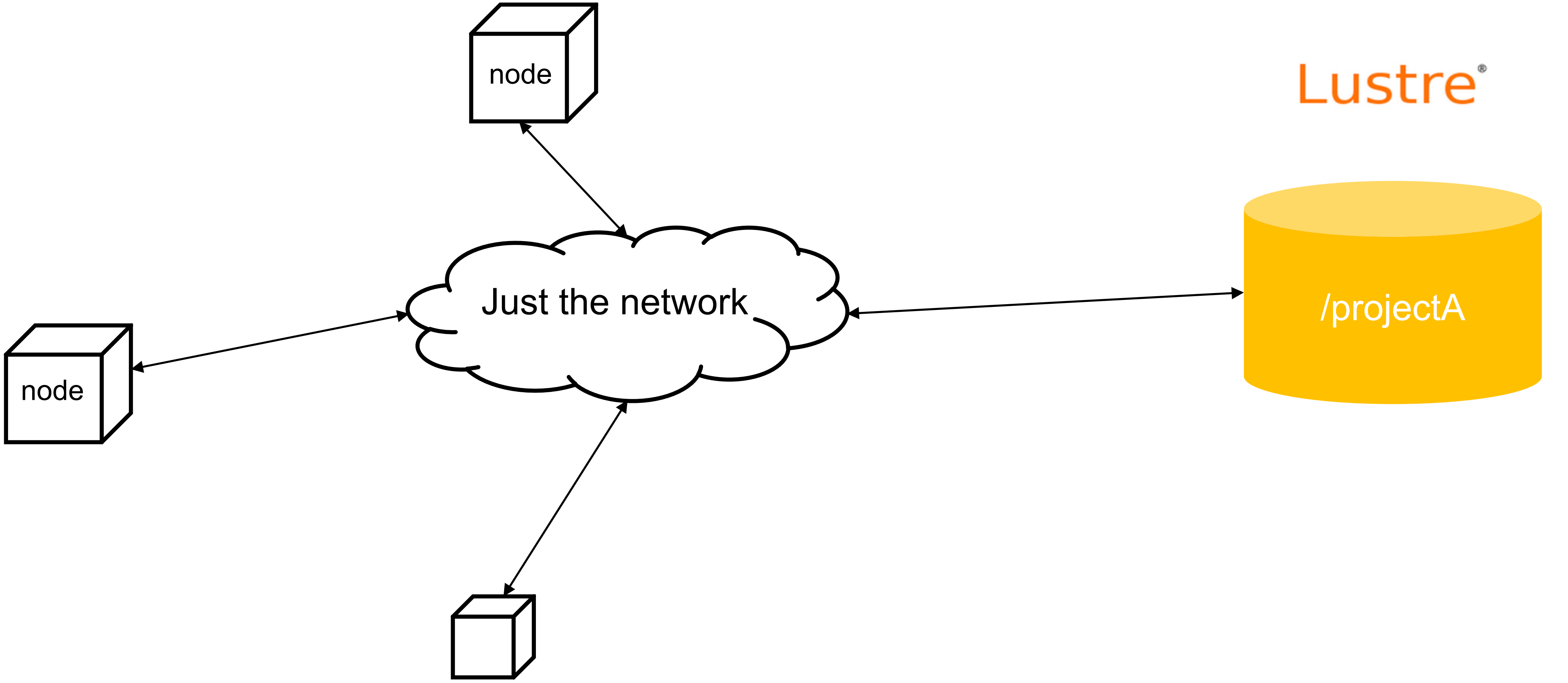
Community experiences

♥ Lustre

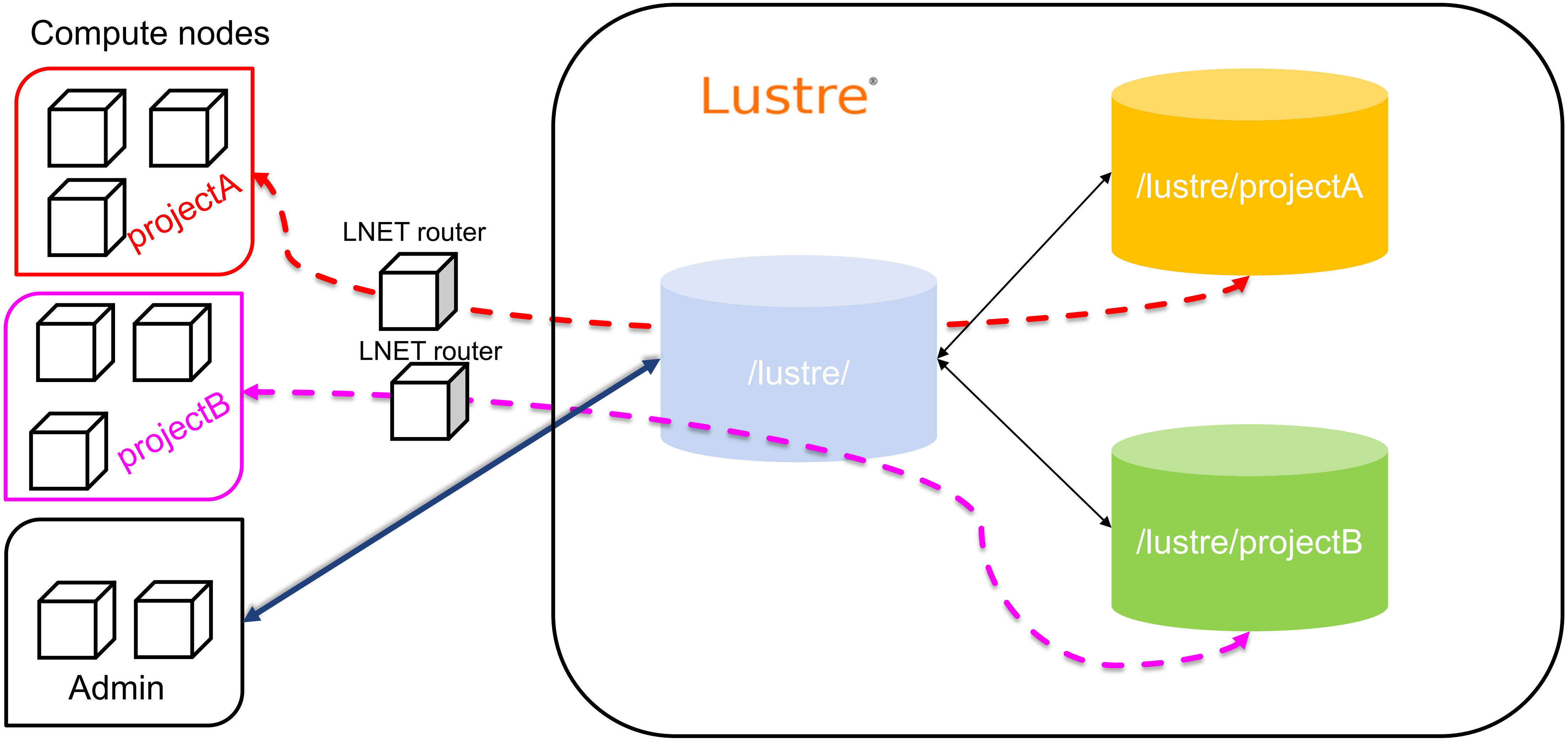
Multi-tenancy in Lustre

- Ensure isolation between tenants/projects: e.g. network and storage
- In reality all tenants are under the same Lustre filesystem and network:
  - Easier for administration: backup, maintenance, etc…
  - Resource sharing made effective
- Specific multi-tenancy for Lustre already discussed in Lustre workshops:
  - Dave Holland (Wellcome Sanger Institute) @ LAD’17 (Paris, France)
  - Sebastien Buisson (DDN presenting Uppsala University, SE) @ LUG’18 (Argonne, US)

Multi-tenancy – The view of a projectA user



Multi-tenancy – The typical sysadmin view



Multi-project vs multi-tenancy at ETH Zurich

- **Often 1 tenant = 1 user**
- **At ETH Zurich we want isolation per project, not per user**
- **So, we prefer to talk about multiple projects instead of multi-tenancy**
- **A project is a group of nodes having common access rights to datasets**

  Each group of nodes lives in one VLAN that can have 1, 2 or more Lustre's LNETs living in it
- **Dataset**

  Data belonging to a project that needs to be independently shared with specific nodes

  E.g.: subdirectory in Lustre containing confidential data linked to a tumor profile project

Multi-project at ETH Zurich

- Use VLANs to isolate projects (no *tenants* but **projects** at ETH Zurich)
  - **Removes LNET router*** overhead **- performance**
  - Provides a good framework for our model of **bare metal provider - adaptability**
  - But **do not exclude LNET routers** in the future if necessary - **flexibility**
  - A compromised node cannot access other projects – **isolation**

*** LNET router: server routing only Lustre packages between networks**

"Simplified" Multi-project at ETH Zurich – The network

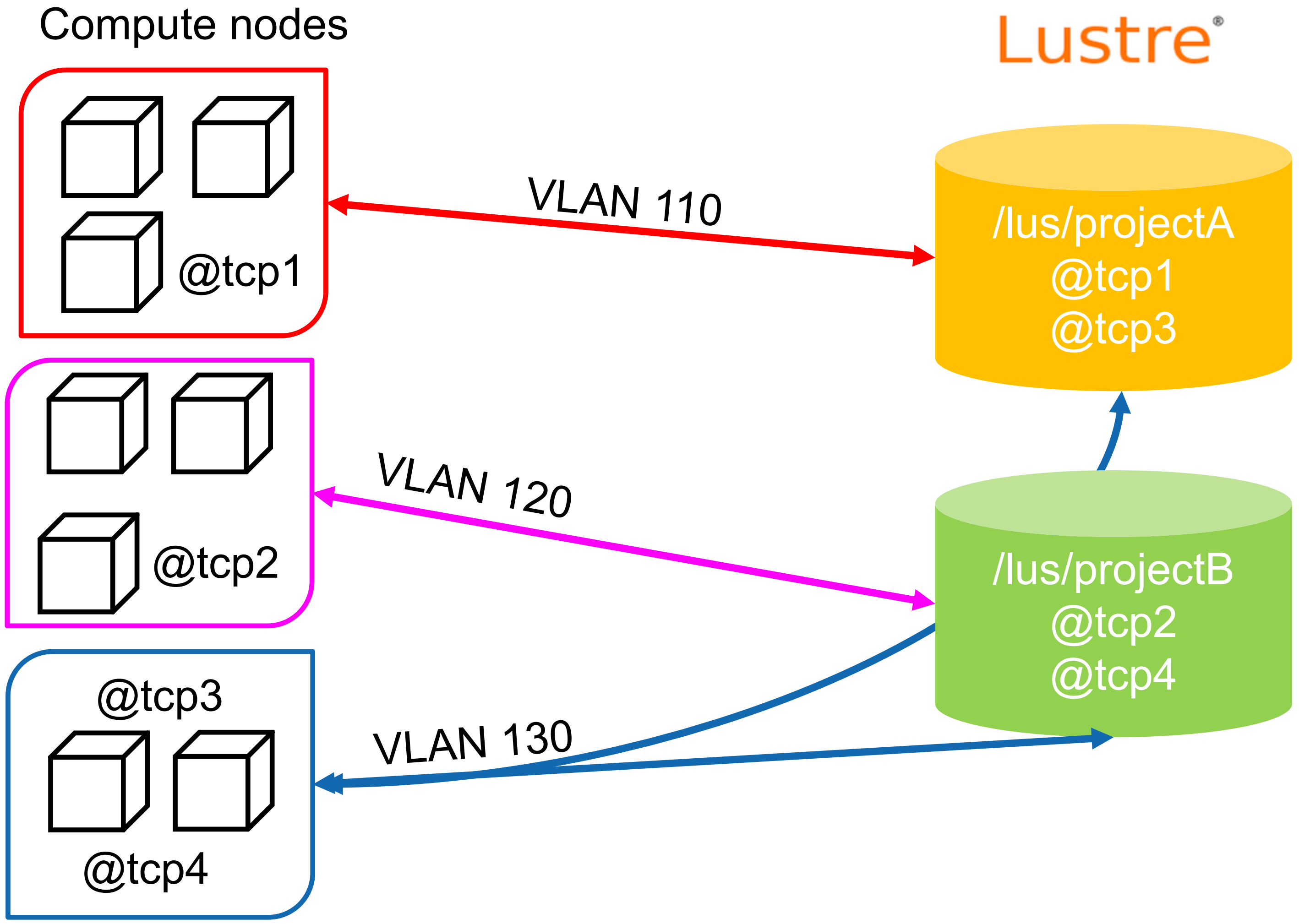
- 10 x Mellanox Ethernet SN-2100 (Cumulus OS):
  - Enforcing VLAN port tagging and switches' ACLs where needed
- On Lustre servers:
  - LNETs and logical interfaces management (1 IP per VLAN)
  - *lctl nodemap* configuration:
    - Assign **subdirectories** as the root filesystem entry point for <u>specific IPs</u>
  - Access control and port management (e.g. ssh only for mgmt. interfaces)

Then simplified becomes a bit more complex…

Shared Multi-project at ETH Zurich

- **Some specific groups can have access granted to 2 or more datasets**
  - Dangerous but possible for specific projects
  - They must not access the root filesystem or other groups of nodes they are not allowed to
  - They must not be accessible by nodes having access to just one of the datasets
  - Needs excellent data management on the user side: “***don’t move data from A to B”***
- **Implementation**
  - 1 LNET per group AND dataset
  - Lustre’s nodemap configuration allows several LNETs for one subdirectory

Shared Multi-project @ ETH



--name pA --range <ip>@tcp1

--name pA --range <ip>@tcp3

nodemap.pA.fileset=/projectA

--name pB --range <ip>@tcp2

--name pB --range <ip>@tcp4

nodemap.pB.fileset=/projectB

Evolution of Lustre’s Leonhard in next months

- Possibility of adding LNET routers later if needed:
  - Cloud computing
  - Cluster with Infiniband or any other interconnect
  - Other clusters on remote sites (with encryption enabled)
- Kerberization of selected projects:
  - Authentication only: authorization to mount the filesystem
  - Partial header encryption (integrity)
  - Full encryption (privacy) for remote projects: with penalty-performance, of course

Evolution of Lustre’s Leonhard in next years

- Some cool features on next Lustre LTS version (2.13?):
  - Data-on-Metadata: up to x KiB the data is stored together with the inode
  - Dynamic File Striping: the layout of the file spreads over storage while the file grows
  - **Audit on Changelogs: which files are accessed, when and who**

Conclusions

- Lustre is a real choice in clusters for personalized health thanks to multiple features
- Exploring security concerns in Lustre is a big topic
- A different implementation of multi-tenancy in Lustre, without LNET routers
- Network design drives the LNET configuration and vice versa: careful decisions
- If you live in Switzerland, well, you might live longer thanks to Lustre ;-)

Thanks!

Allen Neeser
allen.neeser@id.ethz.ch

Christian Bolliger
christian.bolliger@id.ethz.ch

Diego Moreno
diego.moreno@id.ethz.ch

Olivier Byrde
olivier.byrde@id.ethz.ch

Steven Armstrong
steven.armstrong@id.ethz.ch

Eric Muller
eric.mueller@id.ethz.ch

ETH Zurich
Scientific IT Services
High Performance Computing Group
Weinbergstrasse 11
8092 Zürich
https://sis.id.ethz.ch