Simplified Multi-Tenancy for Data Driven Personalized Health Research

Diego Moreno
HPC Storage Specialist @ Scientific IT Services, ETH Zürich

LAD 2018, Paris
Agenda

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

ETH Zurich and Scientific IT Services
ETH Zurich at a glance

- 20,600 students, including 4,100 doctoral students, from over 120 countries
- 500 professors
- 10th in THE ranking
- 7th in QS ranking
- 19th in ARWU ranking
- 380 spin-offs since 1996
- 21 Nobel Prize winners, including Albert Einstein and Wolfgang Pauli
- 1 Fields Medal winner
- 2 Pritzker Prize winners
- 90 patent applications and 200 invention reports every year
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

**Leonhard**
- ~ 150 nodes
- ~ 600 GPUs
- 1.5 PB DDN GPFS -> Lustre
- 2 PB DDN Lustre 2.10
- 0.5 PB NetApp FAS

General purpose HPC

Data driven cluster for special projects
The goal is to provide the **right treatment**, at the **right moment** to the **right patients** (precision medicine) and in the same time to ensure as many people as possible **stay healthy** (prevention; personalized health).
Data Driven Personalized Health in Switzerland
Data Driven Personalized Health in Switzerland

The BioMedIT network

sciCORE Med

Leonhard Med

Vital-IT Med

ETH

USB

USZ
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.

- Multitenancy
Why Lustre?

Well, first it was GPFS… (cough cough)
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 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 is hard to achieve
  - Network flexibility
  - Lustre multi-tenancy kicked in

Disclaimer: GPFS can be great, but not for this setup and this workshop
Why Lustre?

- Performance
- Network flexibility
- Scalability
- Security
- Multi-tenancy
- Community experiences
- Lustre
A reminder on multi-tenancy in Lustre

- Ensure isolation between tenants: e.g. network and storage

- In reality all tenants are under the same filesystem:
  - Easier for administration: backup, maintenance, etc…
  - Resource sharing made effective

- Well covered topic:
  - LAD’17: Dave Holland (Welcome Trust Institute)
  - LUG’18: Sebastien Buisson (DDN)
  - The Lustre Operations Manual 😊
“Simplified” Multi-tenancy at ETH Zurich – The network

- Use **VLANs** to isolate projects
  - 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 tenants - **isolation**
“Simplified” Multi-tenancy at ETH Zurich – The network

- 10 x Mellanox Ethernet SN-200 (Cumulus OS):
  - Enforcing VLAN port tagging and switches’ ACLs where needed

- On Lustre servers:
  - LNETs and logical interfaces management
  - `lctl nodemap` configuration
  - Access control and port management (e.g. ssh only for mgmt. interfaces)
“Simplified” Multi-tenancy at ETH Zurich – The “tenants”

- Group of nodes having common access rights to datasets
  
  *Each group of nodes lives in one VLAN that can have 1, 2 or more LNETs living in it*

- Dataset
  
  *Data belonging to a project that needs to be independently shared with specific nodes*
  
  *E.g.: subdirectory in Lustre*

- Then simplified becomes a bit more complex…
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-tenancy @ ETH

Compute nodes

- VLAN 110
  - @tcp1
  - /lus/projectA
    - tcp1
    - tcp3
  - nodemap.pA.fileset=/projectA
  - --name pA --range <ip>@tcp1
  - --name pA --range <ip>@tcp3

- VLAN 120
  - @tcp2
  - /lus/projectB
    - tcp2
    - tcp4
  - nodemap.pB.fileset=/projectB
  - --name pB --range <ip>@tcp2
  - --name pB --range <ip>@tcp4

- VLAN 130
  - @tcp3
  - @tcp4
LNET routed vs non-routed configuration

- **With LNET Routers**
  - LNET routing between independent clusters with different interconnection networks
  - Additional level of isolation between clients and servers: only LNET traffic is routed to servers
  - Servers in a fixed LNET/networking configuration
  - Ideal on virtualized environments
  - Routing overhead
  - Additional hardware needed
  - Router configuration needed

- **Without LNET Routers**
  - No routing overhead, no extra hardware
  - (Maybe) Easier configuration (add one LNET on cluster vs add one router for each tenant)
  - Isolation provided by network infrastructure: VLANs, partitions, etc..
  - Ideal for bare metal services
  - Compute nodes have direct access to servers
  - Servers and storage devices need to configure one interface/LNET per tenant/group of nodes
  - Switch configuration needed

⚠️ This can be a long discussion...
Evolution of Lustre’s Leonhard in next months

- Possibility of adding LNET routers later if needed:
  - Cloud computing
  - IB cluster
  - Other clusters on remote sites (with encryption enabled)

- Kerberization of selected tenants:
  - Authentication only
  - Partial header encryption (integrity)
  - Full encryption (privacy) for remote tenants
**Evolution of Lustre’s Leonhard in next years**

- All these cool features in next LTS versions:
  - Data-on-Metadata
  - Dynamic File Striping
  - Audit on Changelogs
Lustre is a big actor in clusters for personalized health thanks to multiple features.

Exploring security concerns in Lustre is a big topic.

Yet another example of the possibilities of multi-tenancy in Lustre.

Network design drives the LNET configuration and vice versa: be careful.

If you live in Switzerland, well, you might live longer thanks to Lustre ;-)