ETH zürich

The filesystem as a living creature

Diego Moreno

Scientific IT Services - Informatikdienste

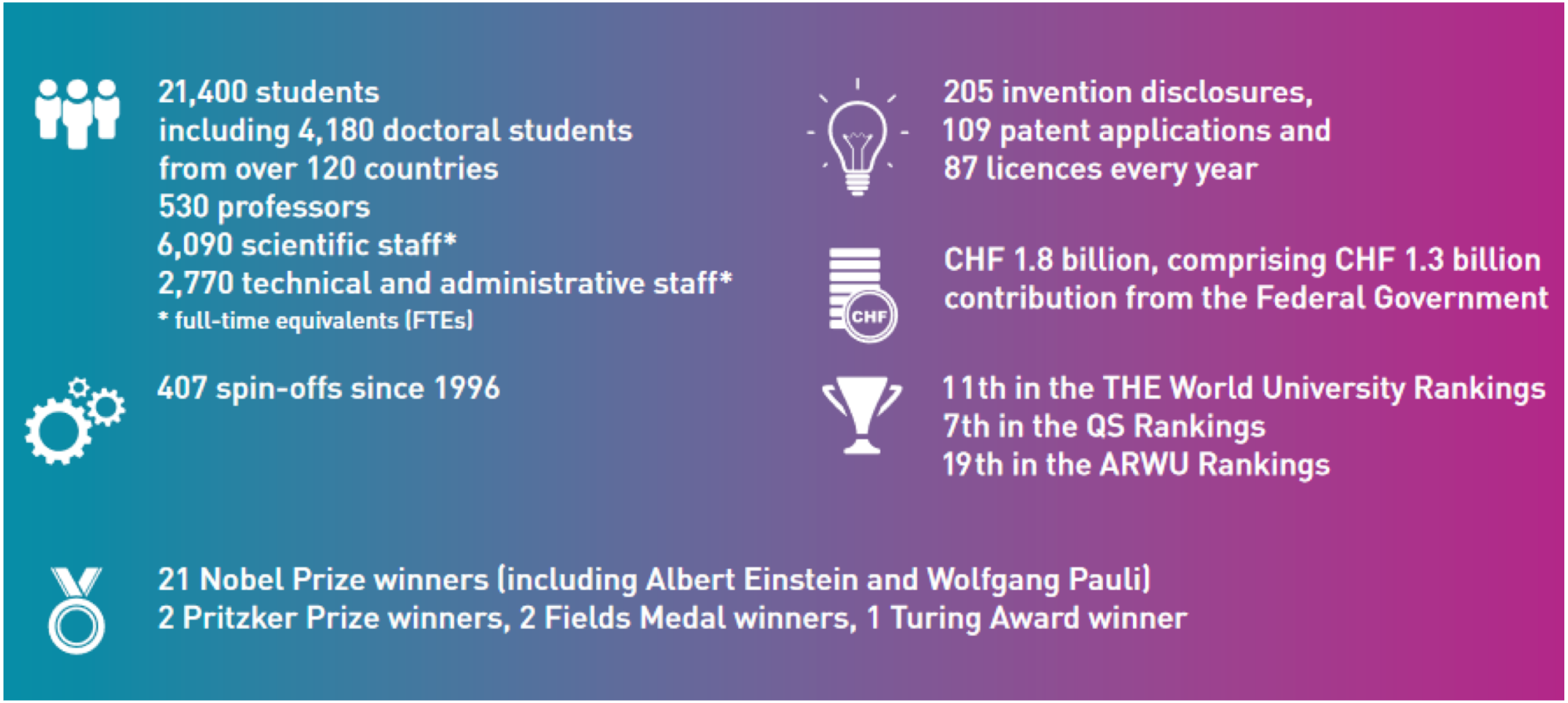
LAD 2019 @ Paris

September, 24th

ETH zürich

Switzerland ➡ Zürich ➡ ETH ➡ IT Services ➡ SIS (Scientific IT Services)

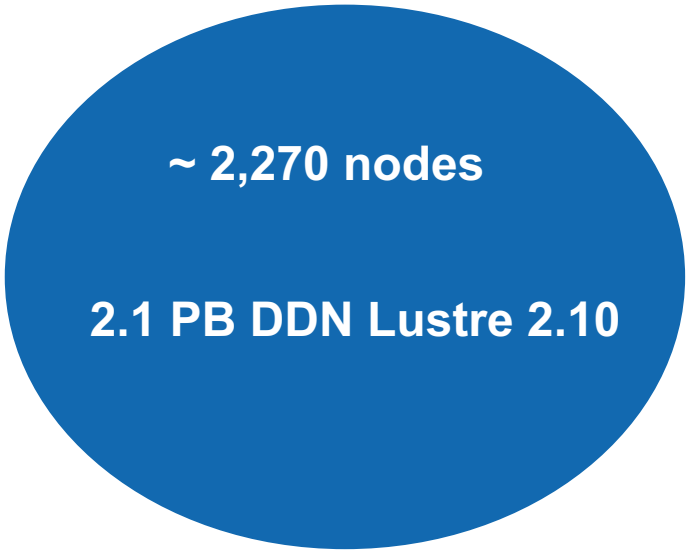
ETH Zurich at a glance



The clusters

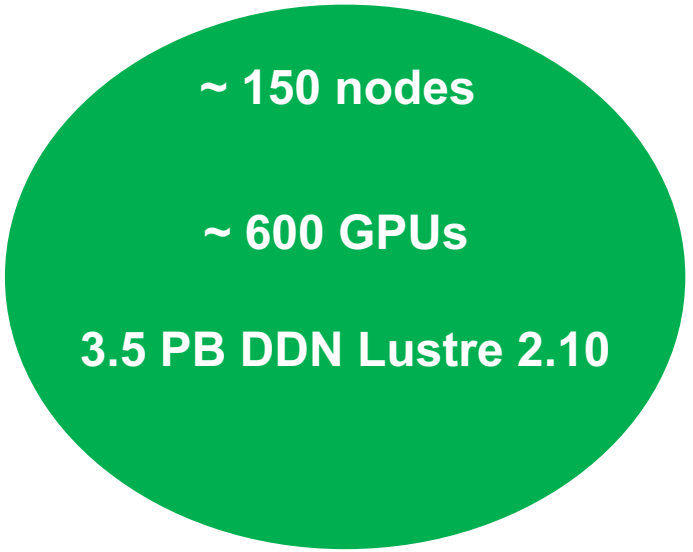
- Currently managing 2 centralized clusters for the ETH's research community:

Euler



General purpose HPC

Leonhard

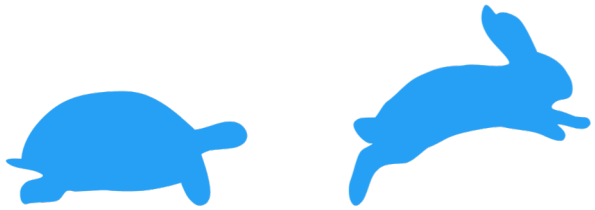


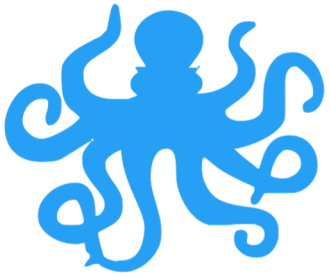
Data driven cluster for special projects

Filesystems as living creatures?

(Disclaimer: this is more a site update than a biology lecture)

- A feature driven comparison in the animal world:





15 years ago: performance vs features

Now: features AND performance AND tiering AND…

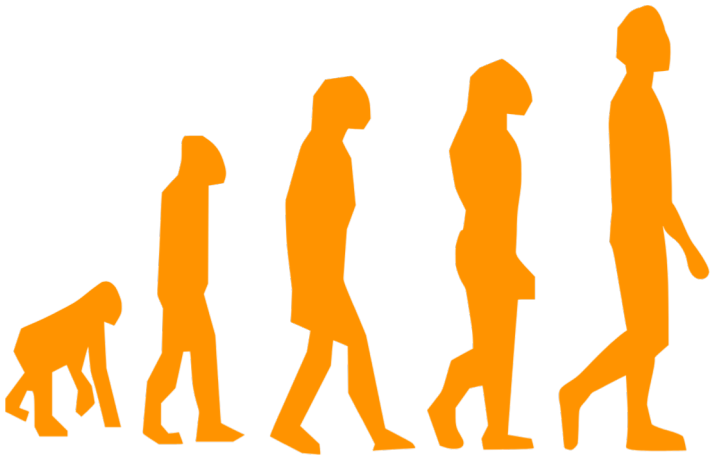
- A lifecycle comparison in the human context:

Basic deployment

Add features

Exascale's era filesystem

Or yet another comparison:



Why the evolution?

- Workloads evolving
- Filesystem usage shifting
- Data capacity
- Flash (r)evolution
- Security requirements increasing exponentially
- Every cluster has its own network architecture

The current Lustre "business as usual" at ETH (2018-2019)

- **Adding online 8 OSTs with 2,500 clients**
- **Adding online 4 OSS (6 OSTs/OSS) and 2 MDS (1 MDT/MDS)**
- Enabling Lustre QoS (TBF per UID)
- Transparent upgrade from 2.7 to 2.10
- Deploying and configuring **multi-rail** LNET routers
- **Shifting from lustre.conf to lnet.conf**
- Enabling Progressive File Layouts

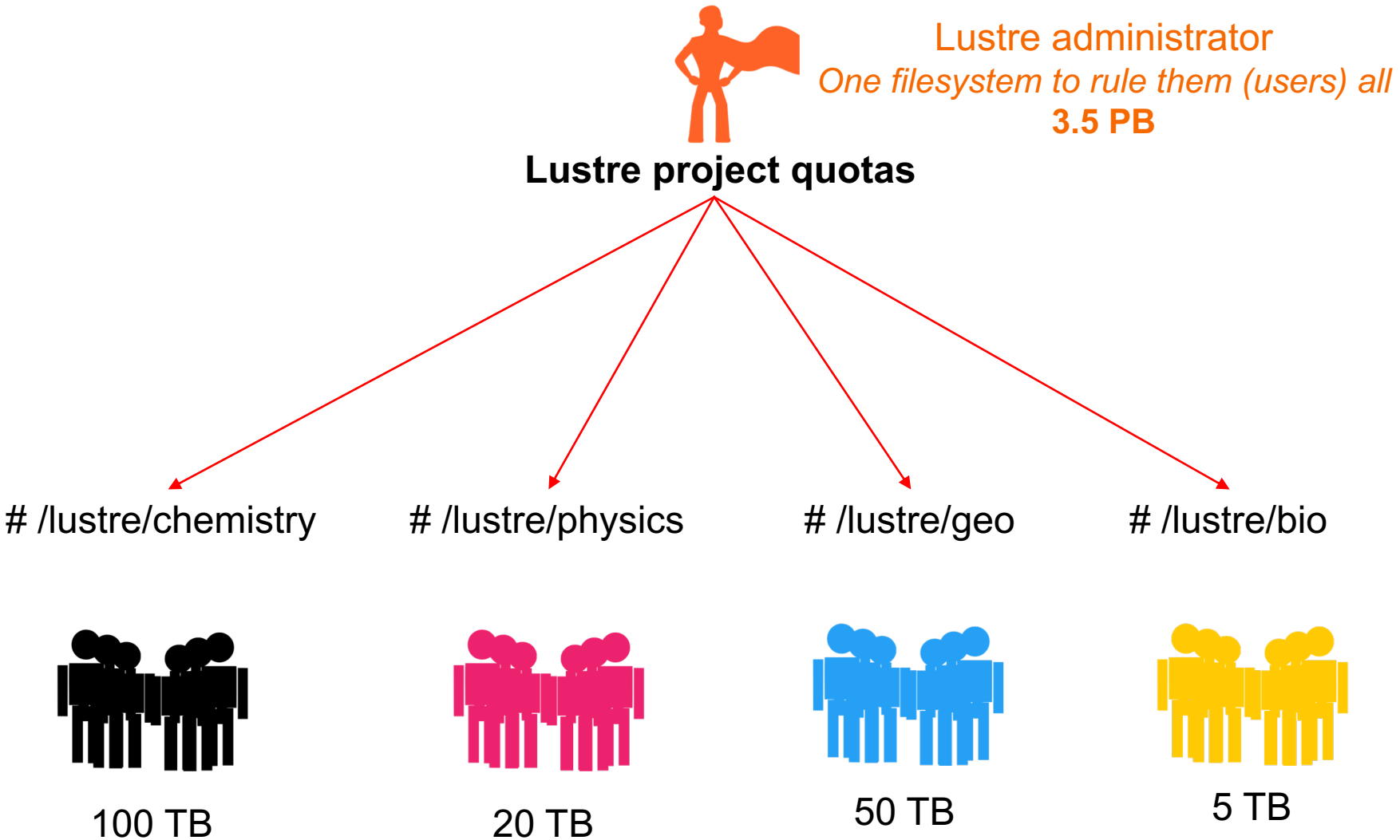
The current Lustre "business as usual" at ETH (2018-2019)

- **Adding online 8 OSTs with 2,500 clients**
- **Adding online 4 OSS (6 OSTs/OSS) and 2 MDS (1 MDT/MDS)**
- Enabling Lustre QoS (TBF per UID)
- Transparent upgrade from 2.7 to 2.10
- Deploying and configuring **multi-rail** LNET routers
- **Shifting from lustre.conf to lnet.conf**
- Enabling Progressive File Layouts

From a classic filesystem to the project model

(Childhood)

The “project” model



Applying the volume/tenant model on a daily basis

- Requires planning and extra tools:
  - Project quotas: Which id? Parameters? Quota?
  - A dedicated tool is a must:

```
lus_vol_create --user diego --group ID-SIS-HPC --perm 3770 --tb 20 --dir /work/sis_hpc
```

  - Wishlist: user quotas per project. Requires external tool (mindblowing to develop in ldiskfs)

Applying the volume/tenant model on a daily basis

- Obtaining the project quota might not be obvious for users
- Extra tools for project quota information:

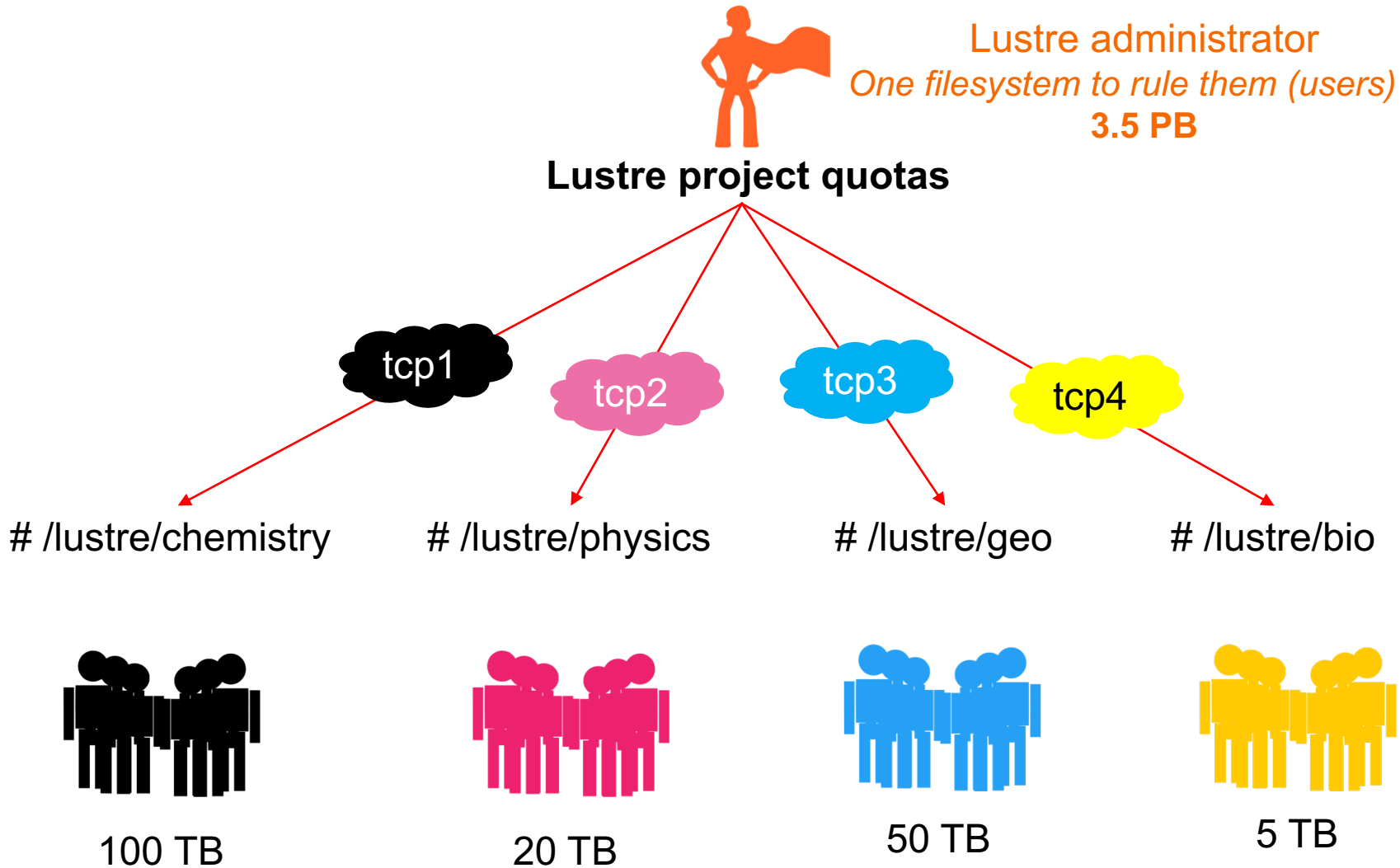
```
[diego@euler ~]$ lquota /cluster/work/sis/
+--------------------------+------------+-----------------+-----------------+-----------------+
| Storage location:        | Quota type: | Used:           | Soft quota:     | Hard quota:     |
+--------------------------+------------+-----------------+-----------------+-----------------+
| /cluster/work/sis/       | space      |        19.73 TB |        30.00 TB |        32.99 TB |
| /cluster/work/sis/       | files      |          177004 |        10000000 |        11000000 |
+--------------------------+------------+-----------------+-----------------+-----------------+
```

- Wrapper for NFS (*quota*) and Lustre (*lfs project –d /dir* + *lfs quota –p id /lustre*)
- Unify units (TiB vs TB)

From the project model to multi-tenancy and isolation

(Adolescence)

The “tenant” model with network isolation



Applying the volume/tenant model on a daily basis

- On Lustre: which NIDs? Fileset? Routing?
  - Tools on the MGS for nodemap creation:

```
create_nodemap --name chemistry --range 10.0.[1-4].[0-255]@tcp11 --admin yes --dir /chem
```

  - Tools to consistently modify while online LNET on servers AND routers:

```
create_lnet --lnet tcp11 --router lnet-router-[1-4] --ips 10.0.1.[10-14] --server o2ib0
```

  - Virtualized LNET routers? Virtualization management
  - Non-virtualized LNET routers? Config management: cdist

Adding admins to “multi-tenancy”

(The teenager has weird friends now)

The volume/tenant model, now with tenant admins

Lustre administrator

One filesystem to rule them (users) all

3.5 PB

Tenant administrator (customer)

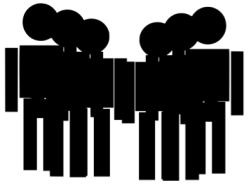
/lustre/chemistry

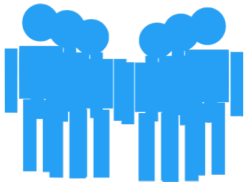
/lustre/physics

/lustre/geo

/lustre/bio

Users





100 TB

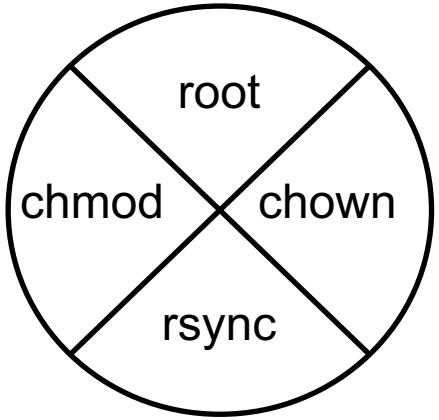
20 TB

50 TB

5 TB

Privileged operations for tenant administrators

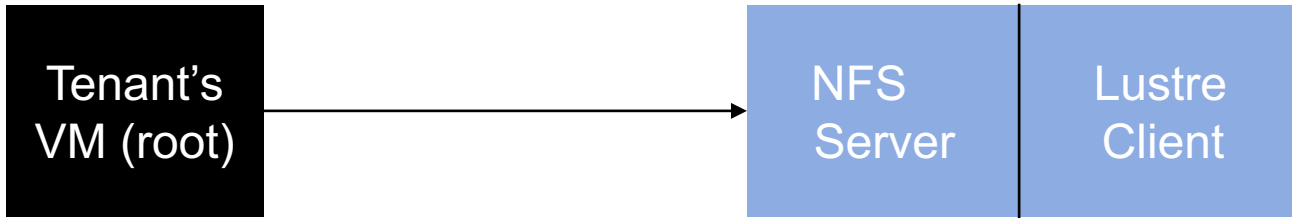
- Tenants' administrators should be allowed to be root on their directories
- First thoughts:
  - Play with POSIX groups? *Not quite convenient*
  - Create a Lustre nodemap with admin flag? *Probably not:*
    - Root on client can change quota settings
    - lfs mkdir and other privileged lfs operations



lu-2299 speaks about this…

Solution: NFS re-exporter for root tenants

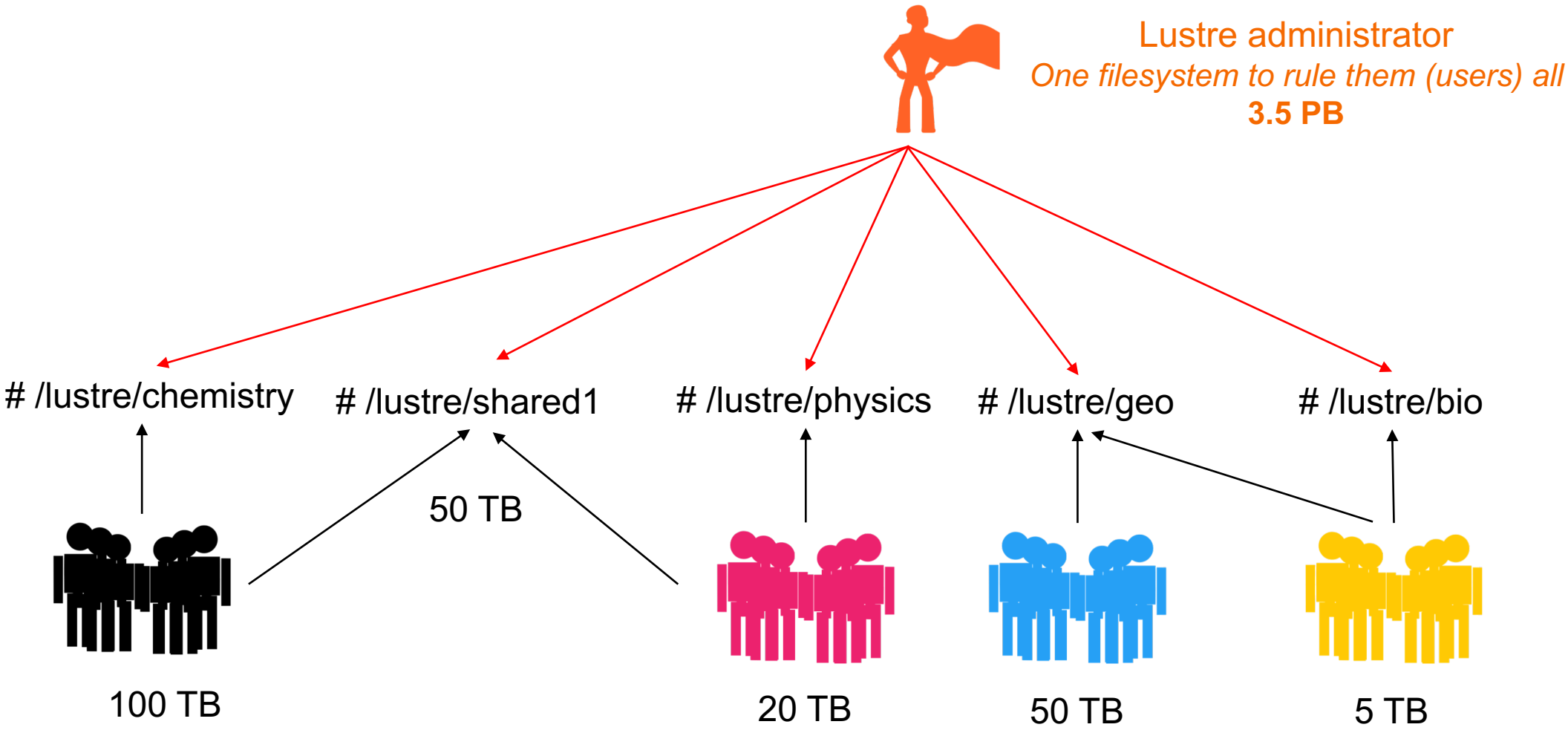
- Classic and convenient solution for this problem: **no_root_squash**
- Apply good practices over NFS: *export to specific IPs, firewall rules, export only the necessary, subtree_check*



Adding multiple filesets to a tenant and sharing them

(Adulthood: life gets complex)

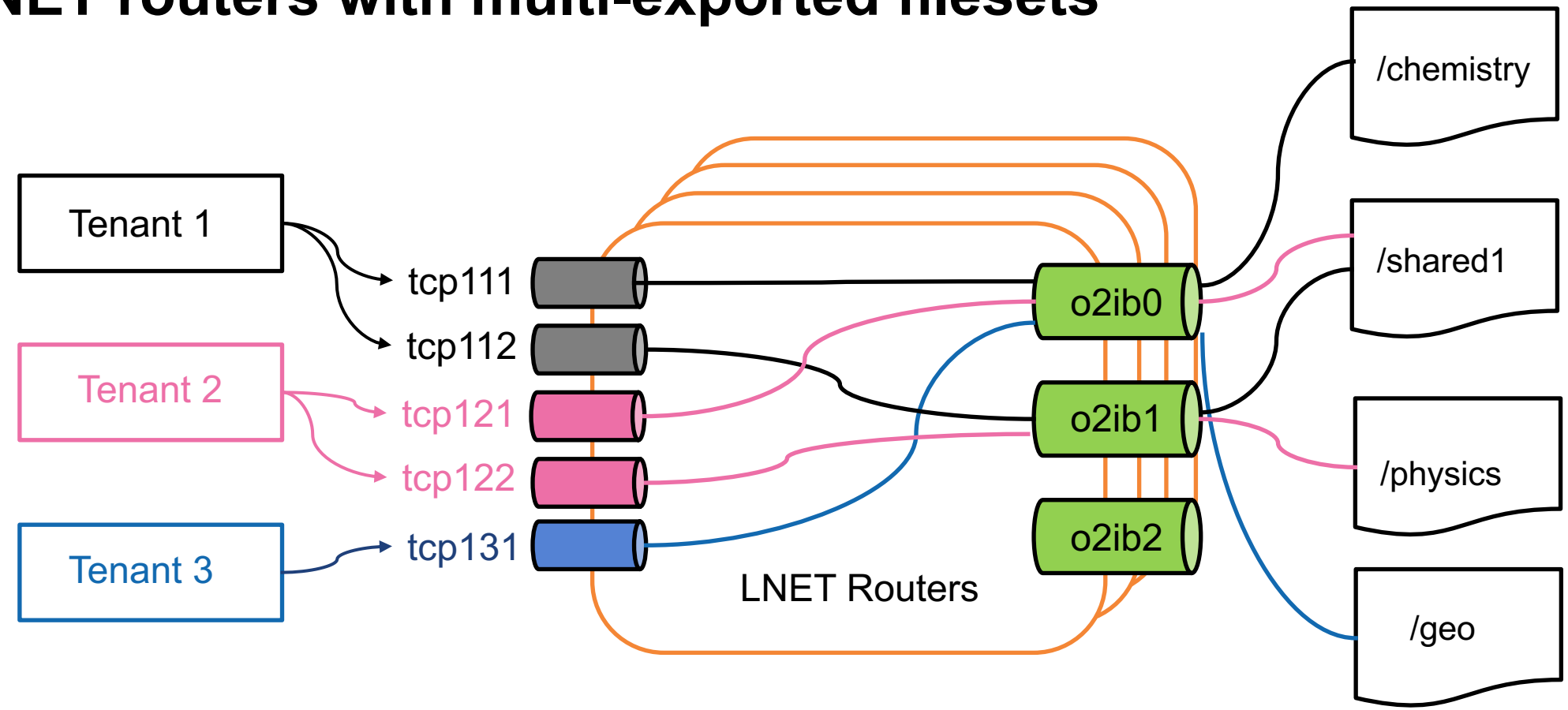
The volume/tenant model, now with multi-exported filesets



Once again: LNET routers vs no LNET routers

- We spoke on LAD'18 about the possibility of getting rid of LNET routers under some conditions
- Well… things happened:
  - Limit of 36 NIDs per device (actually 18 with HA) – so, max 18 filesets without routers…
  - Bug on MGS reconfiguration with writeconf not adding some NIDs
  - Dynamic peer discovery (2.12) not available in Lustre 2.10
- *The man who never alters his opinions is like standing water, and breeds reptiles of the mind (W. Blake)* => Deploy LNET routers and enjoy the advantages

LNET routers with multi-exported filesets




```
mount –o network=tcp121 mgs@o2ib0:/leefs1 /mnt/shared
mount –o network=tcp122 mgs@o2ib1:/leefs1 /mnt/physics
```

* LNET routes: tcp121 <-> o2ib0, tcp122 <-> o2ib1

Monitoring

(Watch the pups)

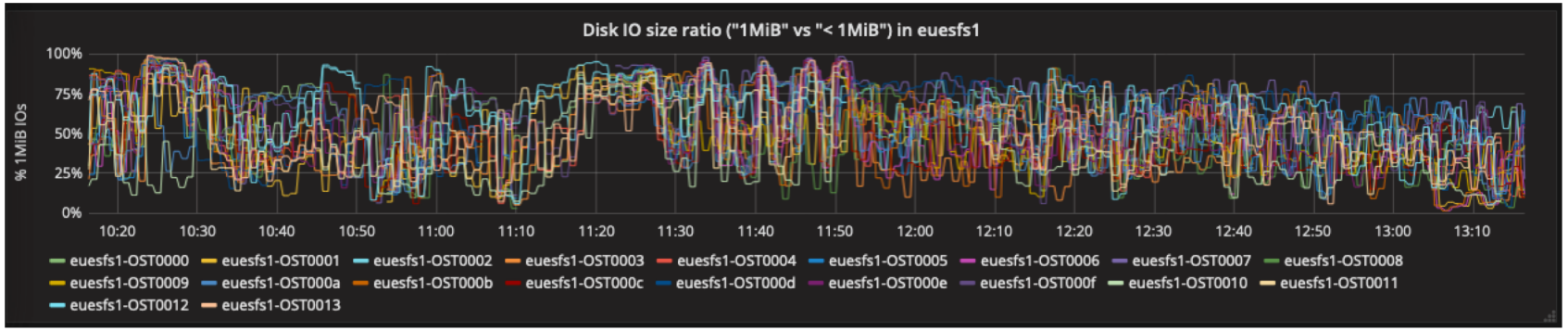
Prometheus + Grafana

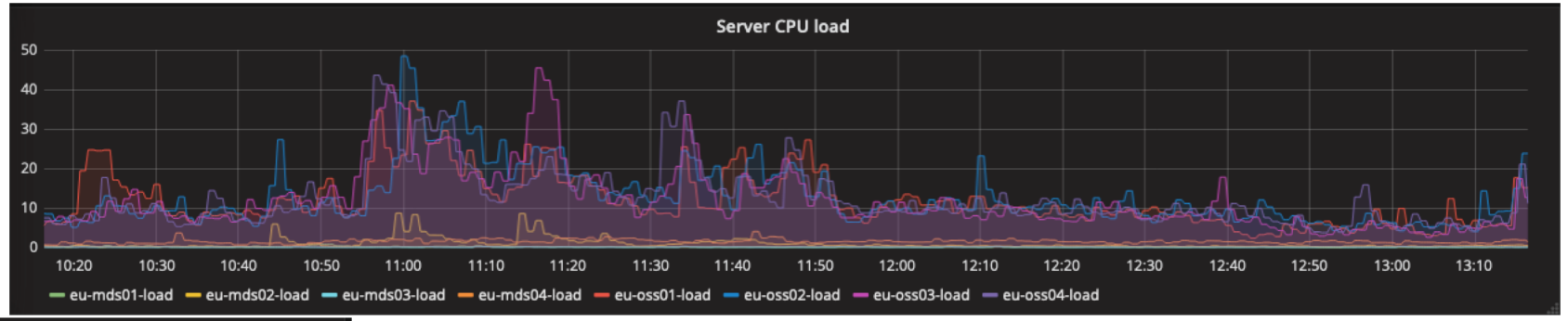
- Prometheus node_exporter + lustre_exporter on servers and routers:

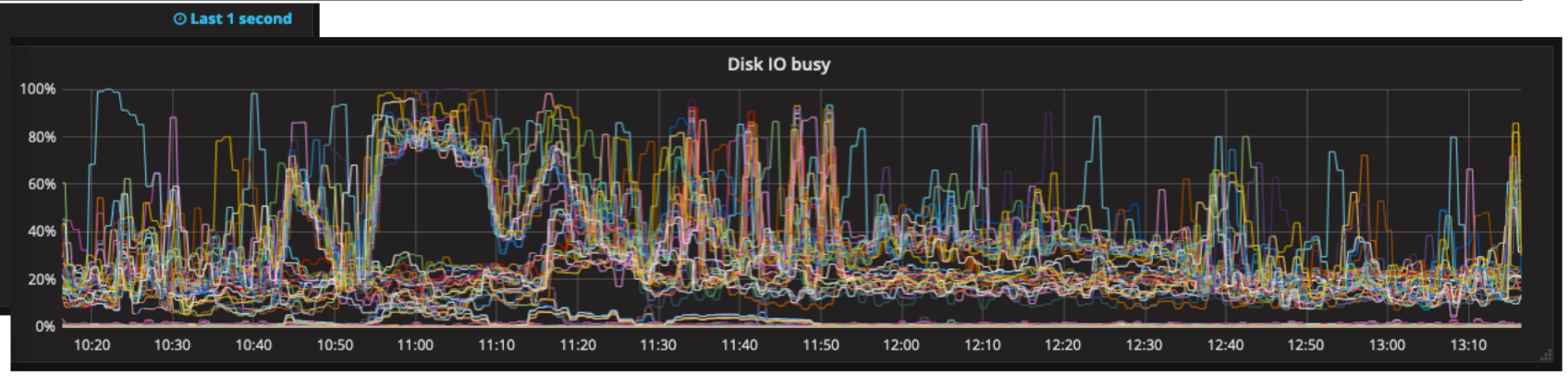
  Thanks to HPE's exporter: https://github.com/HewlettPackard/lustre_exporter

- ETH work on Grafana dashboards published on grafana repos:
  - Lustre Overview, Detailed, General Jobs Stats and Specific Job Stats
  - Support of multiple filesystems
  - Support of stats per device, per server, per filesystem

E.g: Lustre detailed







The creature keeps growing - Future

(Continuous education)

It just keeps moving…

- Prometheus alerting on IOPS, devices and servers' load
- Lustre persistent cache on client
- Dynamic LNET discovery
- Flash pools with pool quotas

Credits

hpc-group
ETH Zurich
Scientific IT Services
IT Services
Weinbergstrasse 13
Zurich

https://sis.id.ethz.ch/

Publisher: IT Services of ETH Zurich
Images: ETH Zürich (diagrams), SVG Image & Icon (Pictures) (Licensed under CC), Cdist (logo)