PAUL SCHERRER INSTITUT

PSI

Leonardo Sala :: AWI :: Paul Scherrer Institut

DARI, SLS and RA

AWI Department meeting 2022.12.12 / PSI

Outline

- Who we are / What we do
- RA news and highlights
- SLS news and highlights
- Next major projects

Who / What

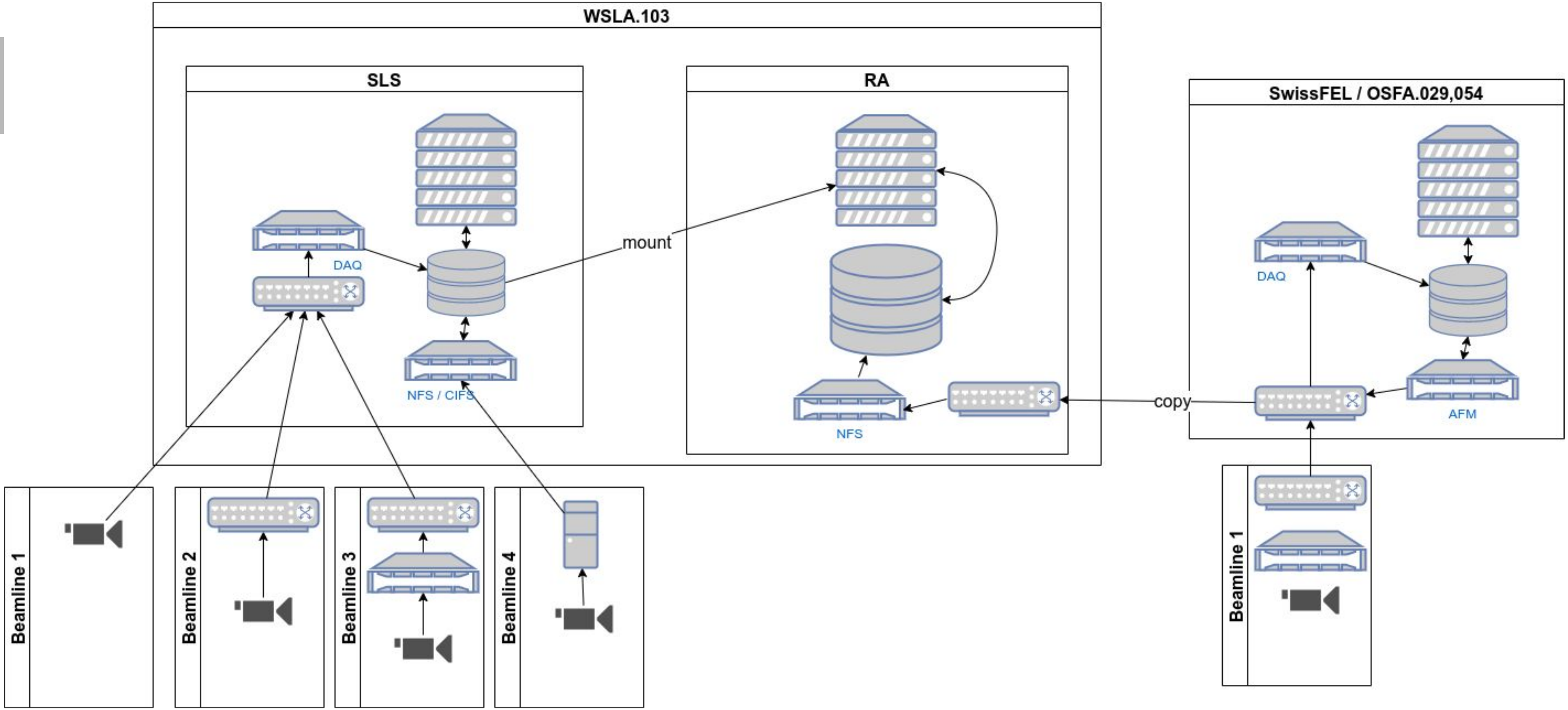
Who we are:

- **Ivano Talamo:** RA responsible and main admin
- **Alvise Dorigo:** SwissFEL responsible and main admin
- **Joshua Taylor:** SLS responsible and main admin
- **Krisztian Pozsa:** Expands / EOSCFuture projects (project)
- **Borys Sharapov:** SLS 2 DevOps / Sysadmin (project)
- **Leonardo Sala:** Group Lead

What we do

- Manage **IT infrastructure** for Photon experiments (storage, compute, network, …)
- Manage special **services** (archiving, opendcim, jupyterhub, …)
- Enable **DevOps** and best practices to enable scientists / staff to do their job

Big picture


WSLA.103
SLS
DAQ
NFS / CIFS
mount
RA
NFS
SwissFEL / OSFA.029,054
DAQ
AFM
copy
Beamline 1
Beamline 2
Beamline 3
Beamline 4
Beamline 1

Some numbers

- **~9000 cores, ~20 PiB, ~70 TB ram, ~30 managed switches (Infiniband, Ethernet)**
- managed by **Puppet and ansible**
  - infrastructure as code backed up by Gitlab
  - Puppet for basic / standard OS installation
  - Ansible for special setups, pipelines and operations
- monitored by **Icinga and InfluxDB / Grafana**
  - for more details see Alvise's talk
  - looking into AIT ELK
- We even have a test Openshift k8s cluster
  - used for gitlab runners and tests

Highlight: server installation

Our server installation is mostly automated:

- physical server installation
- register default admin credentials and required variables
- run playbook that:
  - configure BIOS based on profiles
  - register system in linux inventory
  - configure RAID, boot device, …
  - boot up server
- based in industry-standard Redfish API

Next steps:

- automatic filling of our Data Center management system (opendcim)
- this is possible now as we recently installed a version with a RESTful API

RA updates

Slurm resources management:

- migrated away from full node allocation towards resources allocation
- this allows us to:
  - efficiently manage CPU and GPU resources
  - limit resources usage with cgroups -> less interference
  - have similar setup to Merlin -> can bother Marc even more :D

New Jupyterhub interface

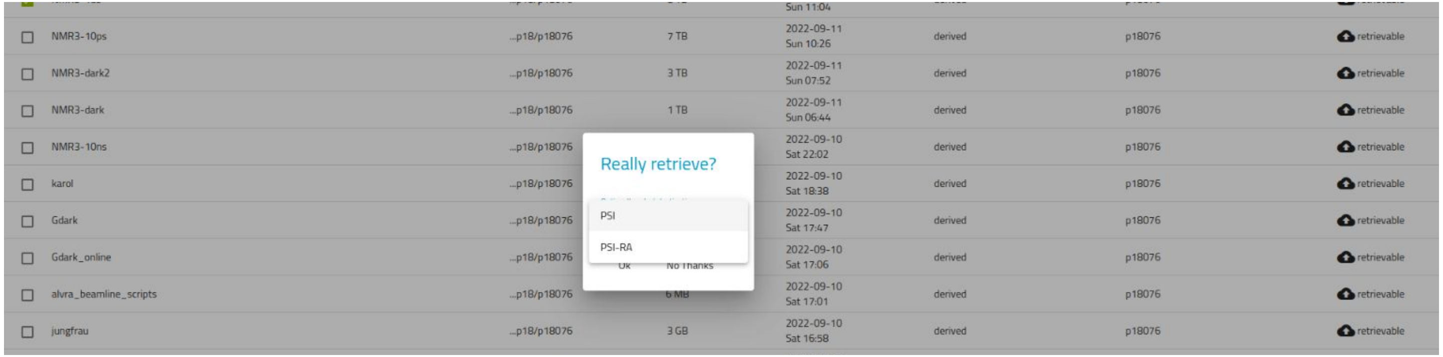
- more dynamic (python + javascript) - depending on queue, shows different options
- more improvements to come, like run time checks

New storage

- project-funded storage is more than 5 years old now
- replacing 2 x 2.2 PiB systems with 1 x 6 PiB system
- delivery this week (Xmas present)

Highlight: tape retrieve

Thanks to a joint effort by AWI (Ivano Talamo, Krisztian Pozsa, Stephan Egli, Carlo Minotti) and AIT (Peter Huesser, Michael Kallmeier), **simple one-click tape retrieve from CSCS Petabyte Archive to RA storage is available now**. Fixing now some bugs



☑				Sun 11:04			
□	NMR3-10ps	...p18/p18076	7 TB	2022-09-11 Sun 10:26	derived	p18076	retrievable
□	NMR3-dark2	...p18/p18076	3 TB	2022-09-11 Sun 07:52	derived	p18076	retrievable
□	NMR3-dark	...p18/p18076	1 TB	2022-09-11 Sun 06:44	derived	p18076	retrievable
□	NMR3-10ns	...p18/p18076		2022-09-10 Sat 22:02	derived	p18076	retrievable
□	karol	...p18/p18076		2022-09-10 Sat 18:38	derived	p18076	retrievable
□	Gdark	...p18/p18076		2022-09-10 Sat 17:47	derived	p18076	retrievable
□	Gdark_online	...p18/p18076		2022-09-10 Sat 17:06	derived	p18076	retrievable
□	alvra_beamline_scripts	...p18/p18076	6 MB	2022-09-10 Sat 17:01	derived	p18076	retrievable
□	jungfrau	...p18/p18076	3 GB	2022-09-10 Sat 16:58	derived	p18076	retrievable

A similar mechanism will allow retrieval from tape to CSCS Object Storage (possibly early 2023)

SLS updates

Automatic quota warning system

- overcomes some icinga limitations
- automatic warning emails sent to beamline scientists
- scientists can self manage thresholds and address list

ACLs

- most data writers run as root -> not good
- explore ACLs usage to write data without root privileges
- prototype with MX successful, plan to propagate to SLS and SwissFEL together with the developers
- A tool to verify ACL policies is being implemented

DAQ support

- support MX Jungfraujoch efforts
- migration away from IBM Power architecture (Filip Leonarski)

SLS updates / II

Migration from Samba wide-links

- current data access over samba mounts e-account home directory, access data directory over symlink
- this is not supported anymore due to security concerns
  - it also creates quota issues when mounted over Windows
- new mountpoints will be created during the long Shutdown

Plan:

- Enable new mountpoints in our prod_2 cluster
- Migrate beamlines from prod_1 to prod_2
- Keep prod_1 running as-is in case of issues
- Reinstall prod_1 and upgrade Spectrum Scale versions

Highlight: services deployment

Quite some work has been put in the past to make services deployment to DAQ nodes reproducible and automatic -> this is the way SLS DAQ nodes are mostly managed since some time

New effort to improve the system and support beamlines-managed services, e.g. MX analysis pipeline

Requirements:

- separate code from data (config files)
- restrict control over code and pipeline definition
- have a simple and intuitive interface

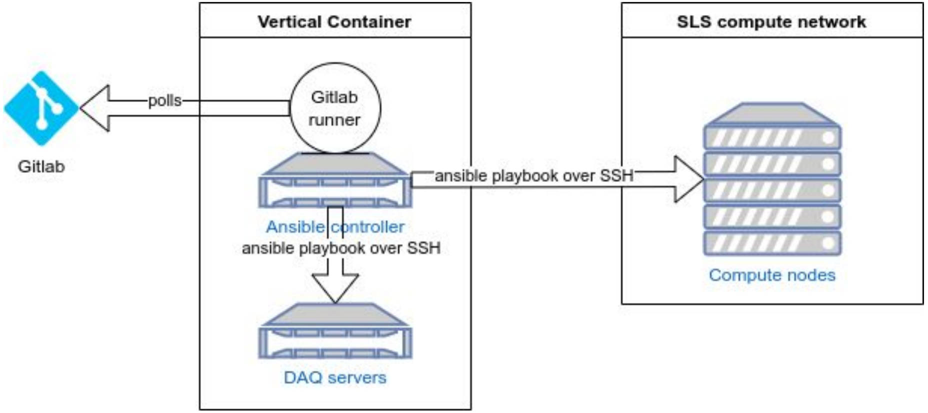
Gitlab pipelines - architecture

Solution:

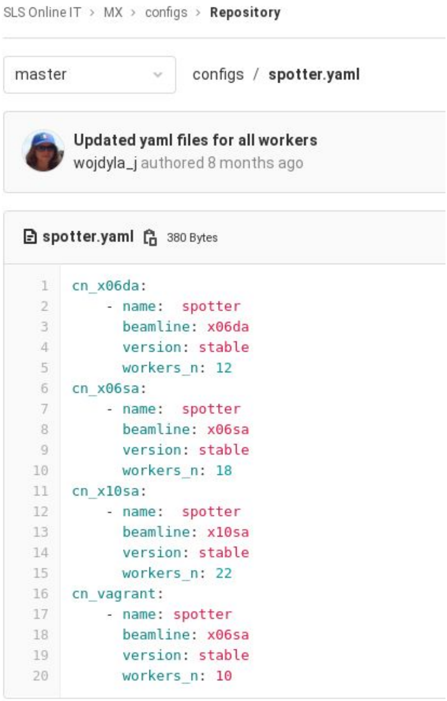
- IaaC with Ansible playbooks
- Gitlab as interface, deployment jobs through pipelines
- access control based on repositories

Advantages:

- fine grained control
- reproducible and versionable
- web interface to output



Pipelines



SLS Online IT > MX > configs > **Repository**

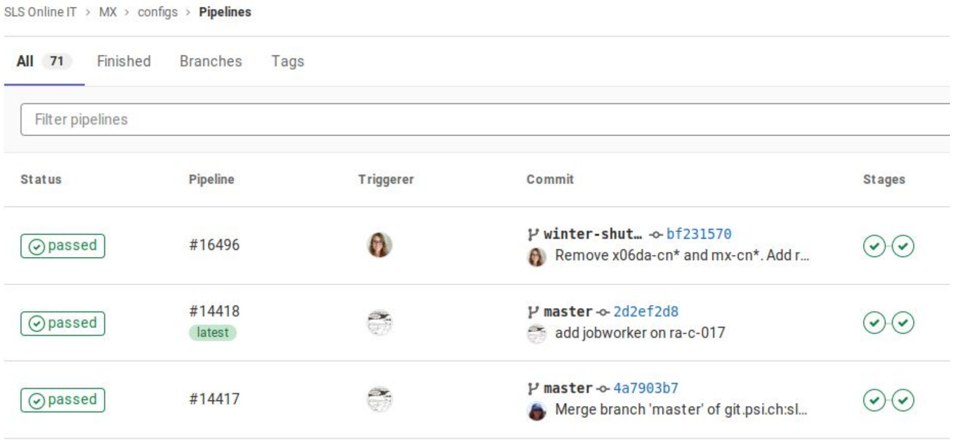
master

configs / **spotter.yaml**

Updated yaml files for all workers
wojdyla_j authored 8 months ago

spotter.yaml 380 Bytes

```
cn_x06da:
    - name:  spotter
      beamline: x06da
      version: stable
      workers_n: 12
cn_x06sa:
    - name:  spotter
      beamline: x06sa
      version: stable
      workers_n: 18
cn_x10sa:
    - name:  spotter
      beamline: x10sa
      version: stable
      workers_n: 22
cn_vagrant:
    - name: spotter
      beamline: x06sa
      version: stable
      workers_n: 10
```

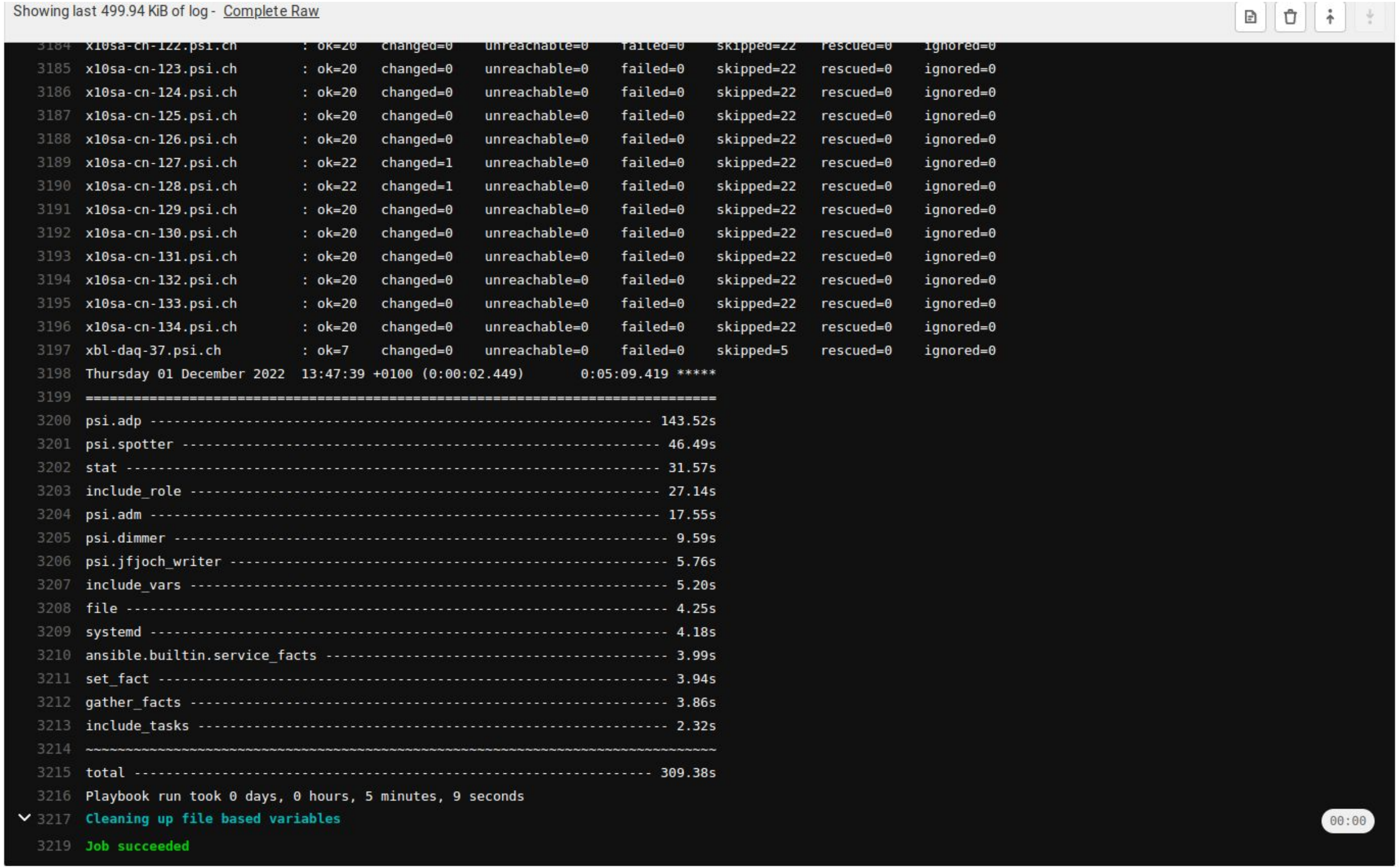


SLS Online IT > MX > configs > **Pipelines**

All 71 Finished Branches Tags

Filter pipelines

Status	Pipeline	Triggerer	Commit	Stages
passed	#16496		winter-shut… bf231570 Remove x06da-cn* and mx-cn*. Add r...	✓ ✓
passed	#14418 latest		master 2d2ef2d8 add jobworker on ra-c-017	✓ ✓
passed	#14417		master 4a7903b7 Merge branch 'master' of git.psi.ch:sl...	✓ ✓



Showing last 499.94 KiB of log - Complete Raw

```
x10sa-cn-122.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-123.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-124.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-125.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-126.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-127.psi.ch        : ok=22   changed=1    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-128.psi.ch        : ok=22   changed=1    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-129.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-130.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-131.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-132.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-133.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
x10sa-cn-134.psi.ch        : ok=20   changed=0    unreachable=0    failed=0    skipped=22   rescued=0    ignored=0
xbl-daq-37.psi.ch          : ok=7    changed=0    unreachable=0    failed=0    skipped=5    rescued=0    ignored=0
Thursday 01 December 2022  13:47:39 +0100 (0:00:02.449)       0:05:09.419 *****
===============================================================================
psi.adp --------------------------------------------------------------- 143.52s
psi.spotter ------------------------------------------------------------ 46.49s
stat ------------------------------------------------------------------- 31.57s
include_role ----------------------------------------------------------- 27.14s
psi.adm ---------------------------------------------------------------- 17.55s
psi.dimmer -------------------------------------------------------------- 9.59s
psi.jfjoch_writer ------------------------------------------------------- 5.76s
include_vars ------------------------------------------------------------ 5.20s
file -------------------------------------------------------------------- 4.25s
systemd ----------------------------------------------------------------- 4.18s
ansible.builtin.service_facts ------------------------------------------- 3.99s
set_fact ---------------------------------------------------------------- 3.94s
gather_facts ------------------------------------------------------------ 3.86s
include_tasks ----------------------------------------------------------- 2.32s
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
total ----------------------------------------------------------------- 309.38s
Playbook run took 0 days, 0 hours, 5 minutes, 9 seconds
Cleaning up file based variables
Job succeeded
```

PAUL SCHERRER INSTITUT PSI

Next big projects

- **Compute node merge**
  - manage SLS and RA compute nodes as a unique pool of resources
  - prototype for feasible switching between different configurations
- **Storage WTO** - we need a new one to replace existing systems and procure SLS2 resources
- **Resources API**
  - Allowing users/beamline to self manage resources with no or minimal intervention from admins e.g. compute node reservation, quota extensions, …
    - possible by exposing resources operations and workflows via APIs
  - early discussions about DUO integration
- **Data lifecycle**: write - read - archive - delete - retrieve
  - streamline policies and workflows
  - including paid storage for projects / grants
- **SLS2**
- **Documentation and dashboards**

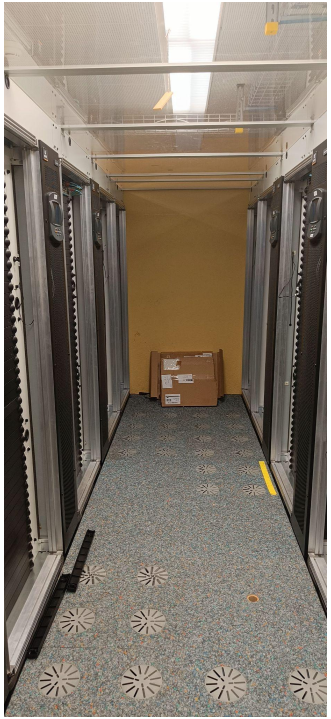
Highlight: WHGA server room

Migration away from SLS server room during dark period.

Started using WHGA server room for RA:

- 4 compute nodes
- Infiniband switch (2x100G, additional links to be added in the next weeks)
- 6 PB Storage in January (Lenovo DSS-G260)

Plan to gradually migrate compute nodes, phase out storage based on lifecycle. No downtimes.



PAUL SCHERRER INSTITUT PSI

Questions?