



Leonardo Sala :: Science IT :: Paul Scherrer Institut

PSI site update

HPCXXL meeting - ORNL, 2018-09-19

“The Paul Scherrer Institute, PSI, is the largest research institute for natural and engineering sciences within Switzerland. We perform world-class research in three main subject areas: Matter and Material; Energy and the Environment; and Human Health. By conducting fundamental and applied research, we work on long-term solutions for major challenges facing society, industry and science.”

PSI operates various Large Scale Facilities:

- SLS: Swiss Light Source synchrotron
- SINQ: Spallation Neutron Source
- S μ S: Swiss Muon Source
- SwissFEL: Swiss Free Electron Laser

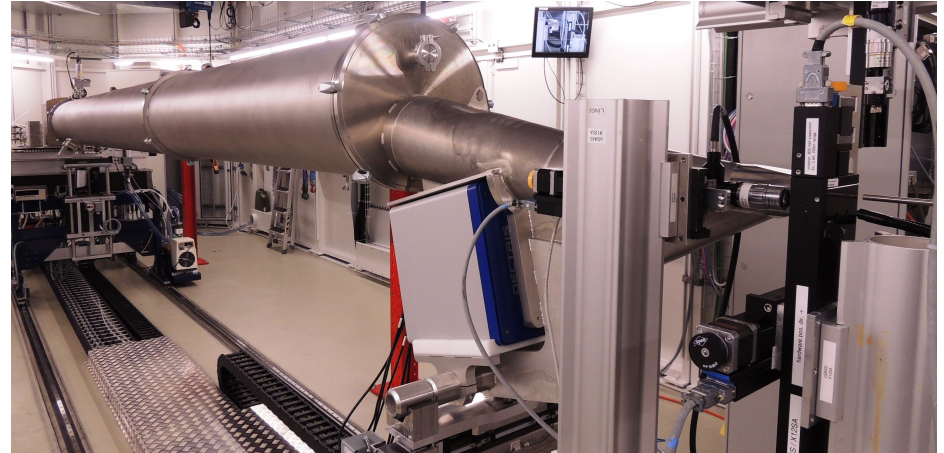
PSI - aerial view



SwissFEL - aerial view



SLS - few pics

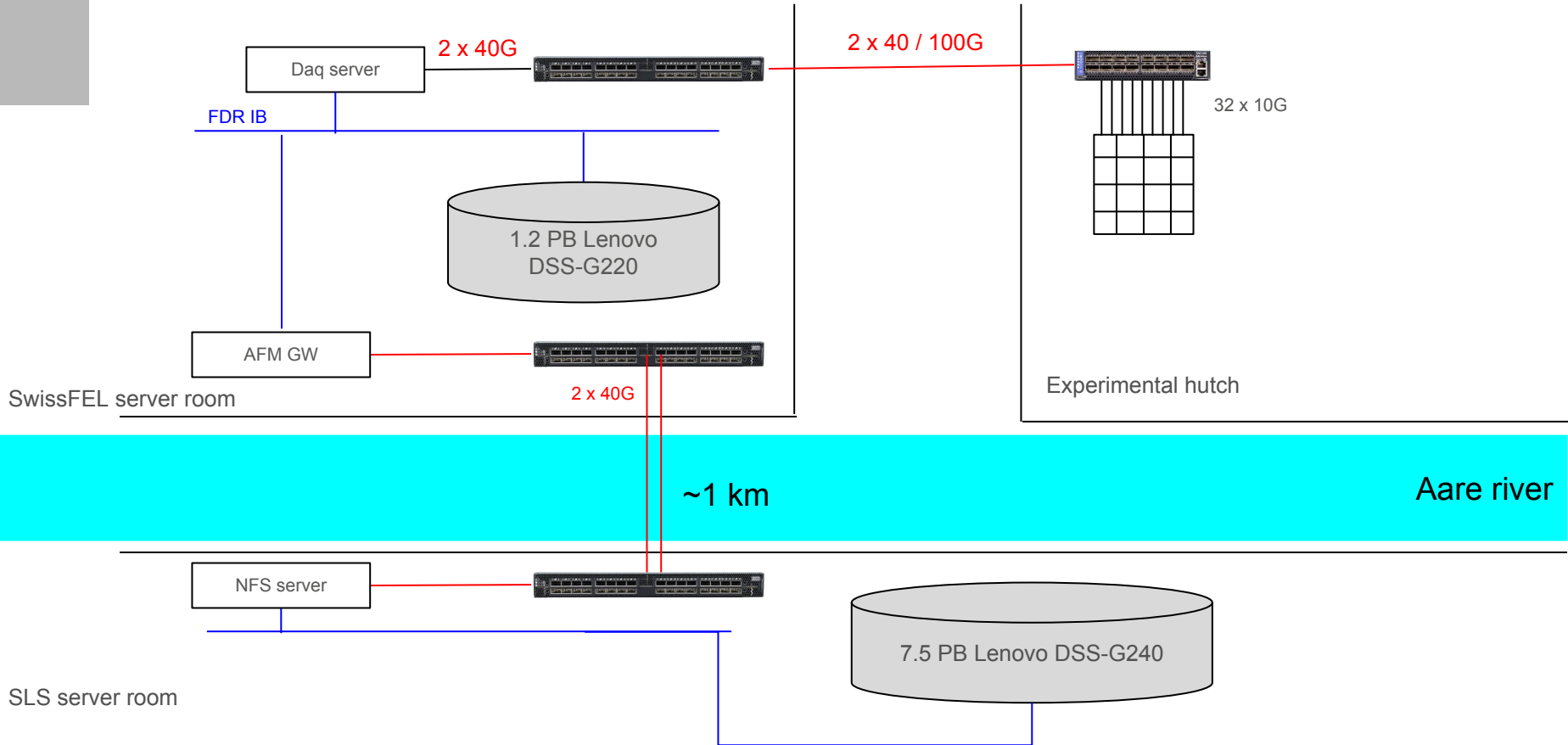


Multiple activities ongoing (short overview)

- SwissFEL new experiments
- CES implementation
- *AFM*
- *Monitoring*
- *Hardware upgrade*

I will focus on the last three items

SwissFEL - Current setup



New experiment in August, 25% production capacity (25 Hz vs 100 Hz):
50T / d, 250 T of RAW data in total (+500 T derived data)

Due to some interactions between HDF5 / GPFS (under investigation), fast feedback analysis moved to offline cluster (RA)

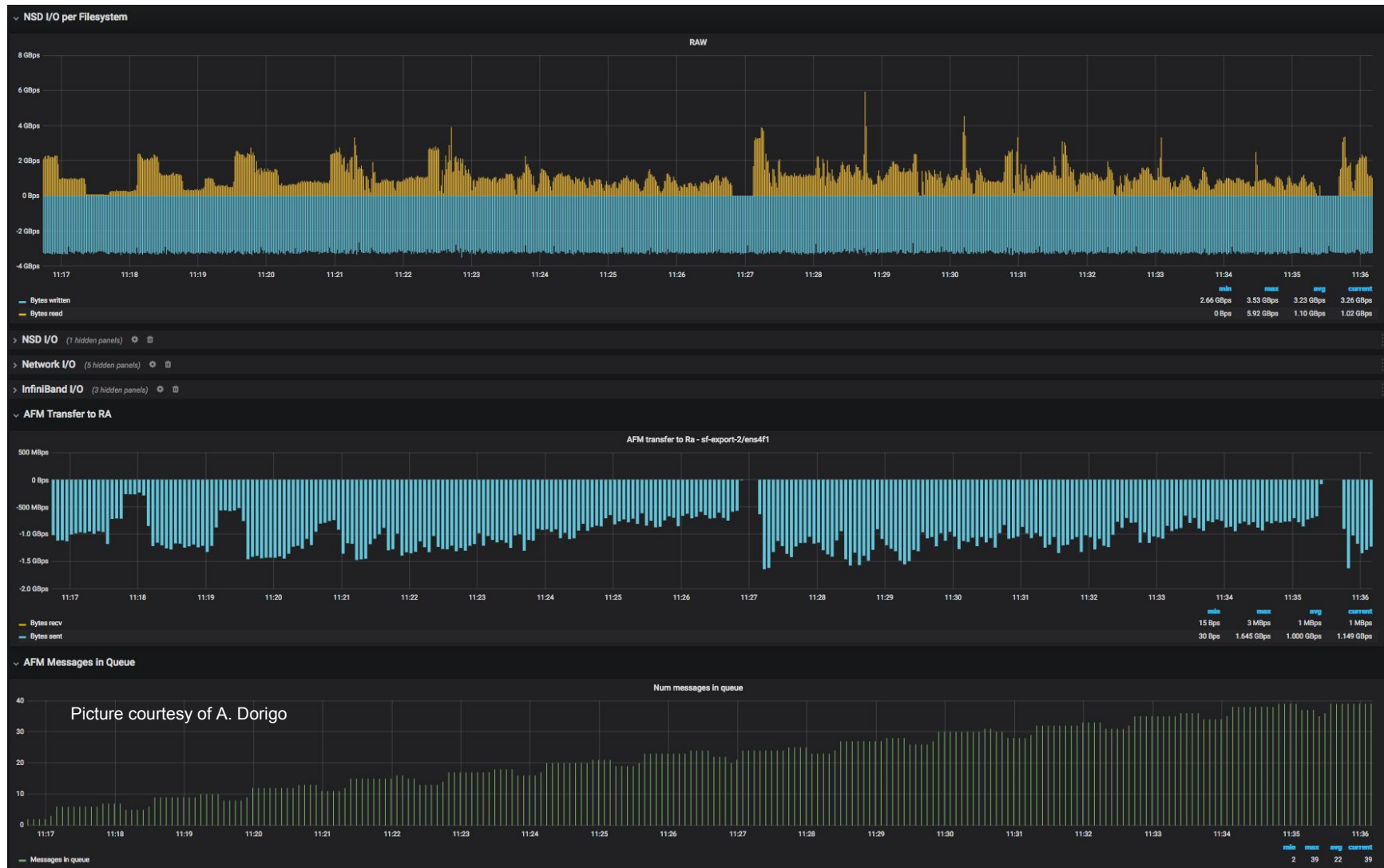
The good:

- AFM kept up nicely with data rates (sustained 800 MB/s write)
- Users never noticed any delay (compared to “online” access)

The bad

- Despite fine tuning with with developer, cannot even get up to NFS limits (~1.6 GB/s)
- “Transfer holes” with full queue

Next: parallel AFM, multiple NFS servers?



Monitoring

We were very interested in Spectrum Scale's PMCollector etc, hoped to have the tools to do trending / etc out of the box

Current setup

- Federated PMcollectors (1 per cluster, 1 central)
- Grafana bridge
- InfluxDB for other metrics (e.g. storage temperature)

We started noticing issues, interested in your experience

Resampling

When period $\neq 1s$, rates are aggregated with SUM. For short term monitoring keep 1s (expensive), but then trending becomes cumbersome, as we found no way to have Grafana deal with that

The Good:

- We can show management plots with 50 GB/s

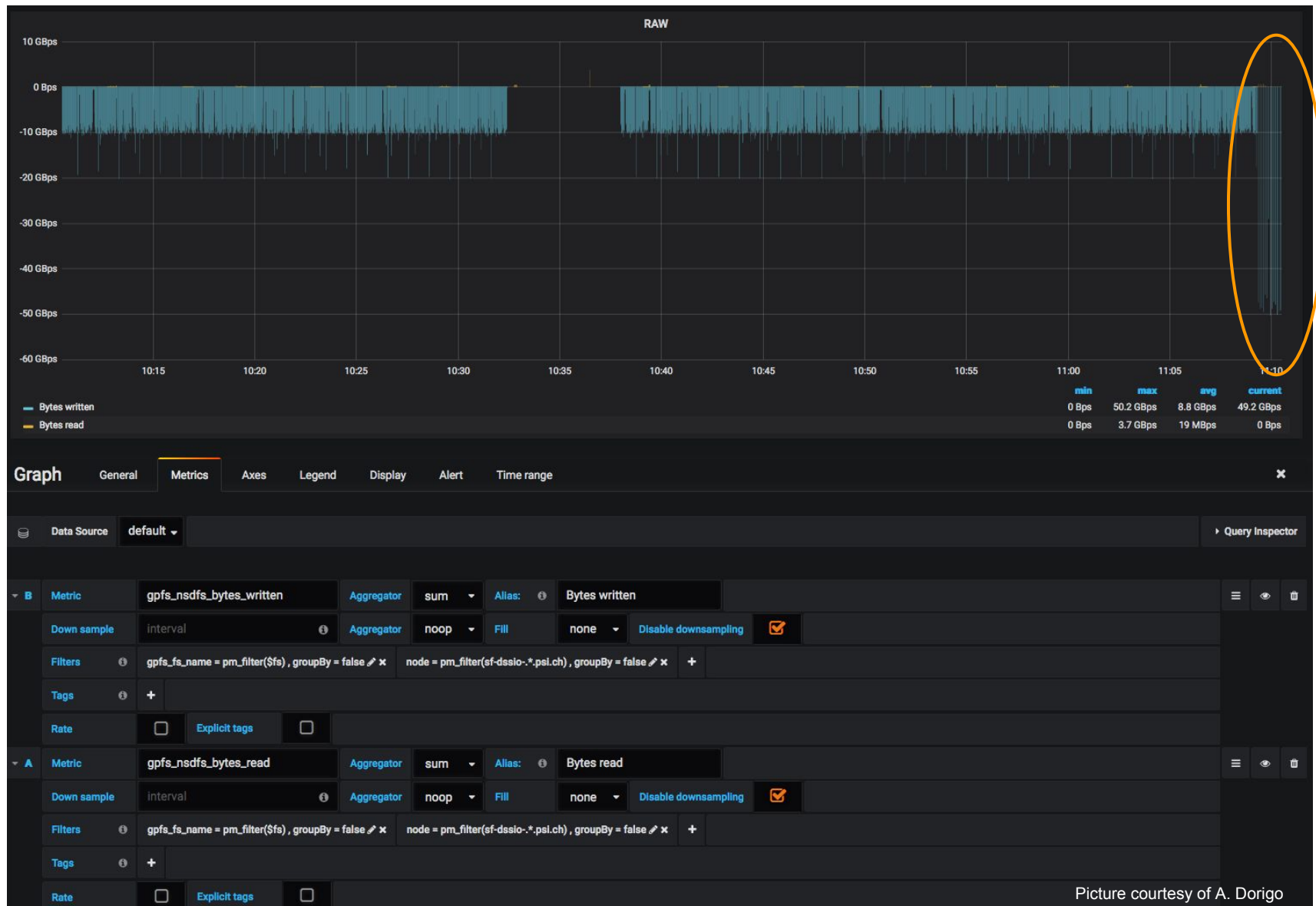
Bad:

- We have plots with 50 GB/s

Next:

- Analysis with external tools (e.g. pandas), but only averages. Not very significant
- Dump PMCollector data in InfluxDB
- Put Prometheus or the likes in the equation

From 10 to 50 GB/s



Picture courtesy of A. Dorigo

Hardware Upgrades

Photonics

We are hungry for storage. At the moment we have ~7.5PB in offline, +3 PB upgrade this year, maybe another 3 PB next year, then we see...

More storage will also lead to more compute in the near future

Exploration of NVMe / SSD burst buffers (current dilemma: our bursts are big and long... are they bursts anymore?)

Stretched cluster with long-haul Infiniband?

Local HPC

Complete renewal of the Merlin5 cluster (2500 cores, 1 PB storage)

Acknowledgements

Derek Feichtinger
(Head Science IT)

The HPC team:

- H. Billich (SLS)
- M. Caubet (local HPC)
- A. Dorigo (SwissFEL)
- N. Loktionova (CMS T3)
- I. Talamo (Analysis Cluster RA)

