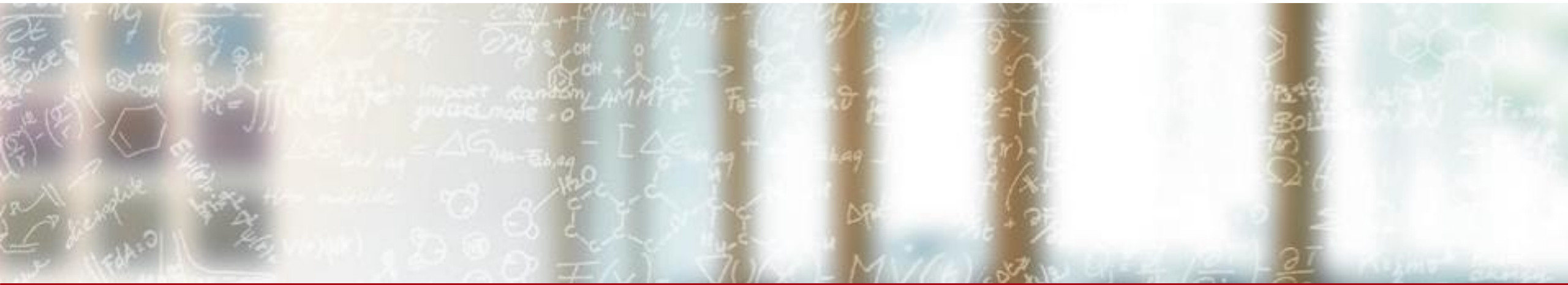




CSCS

Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Spectrum scale at CSCS

HPCXXL – Summer 2018

Giuseppe Lo Re, CSCS

Sep 19, 2018

Outline

- CSCS overview
- Spectrum Scale in Openstack4HPC
- Monitoring
- Conclusion

CSCS overview

- CSCS is the Swiss National Supercomputing Centre
- Unit of the Swiss Federal Institute of Technology in Zurich (ETH Zurich), located in Lugano
- CSCS's resources are open to academia, industry and the business sector
- 2000 m² machine room with no single supporting pillar or any partitioning
- Important collaborations:
 - Human Brain Project
 - Marvel (Material Science)
 - Paul Scherrer Institute
 - CERN
 - ...



CSCS overview

Some operational HPC facilities:

- Piz Daint (Cray XC40/XC50)
- Kesch+Escha (Cray CS-Storm)
- Phoenix (CERN, Grid Cluster)
- Monte Leone (High-memory cluster)
- Gran Tavé (KNL R&D)
- ~~Mönch~~ (NEC Cluster)
- **Mythen (OpenStack for HPC)**
 - Fulen (Virtual HPC cluster)
 - Zupo (Virtual TDS)



CSCS overview

Available data services:

- Posix file systems
 - GPFS
 - Lustre
- Data movers
 - GridFTP
 - Slurm queue
 - ~~Active File Management~~
- TSM Backup/Archive
 - mmbbackup
 - GPFS/HSM
 - Arema (BA API)
- NAS
 - GPFS CES for NFS and SMB
- dCache
- Object store
 - GPFS CES Openstack Swift

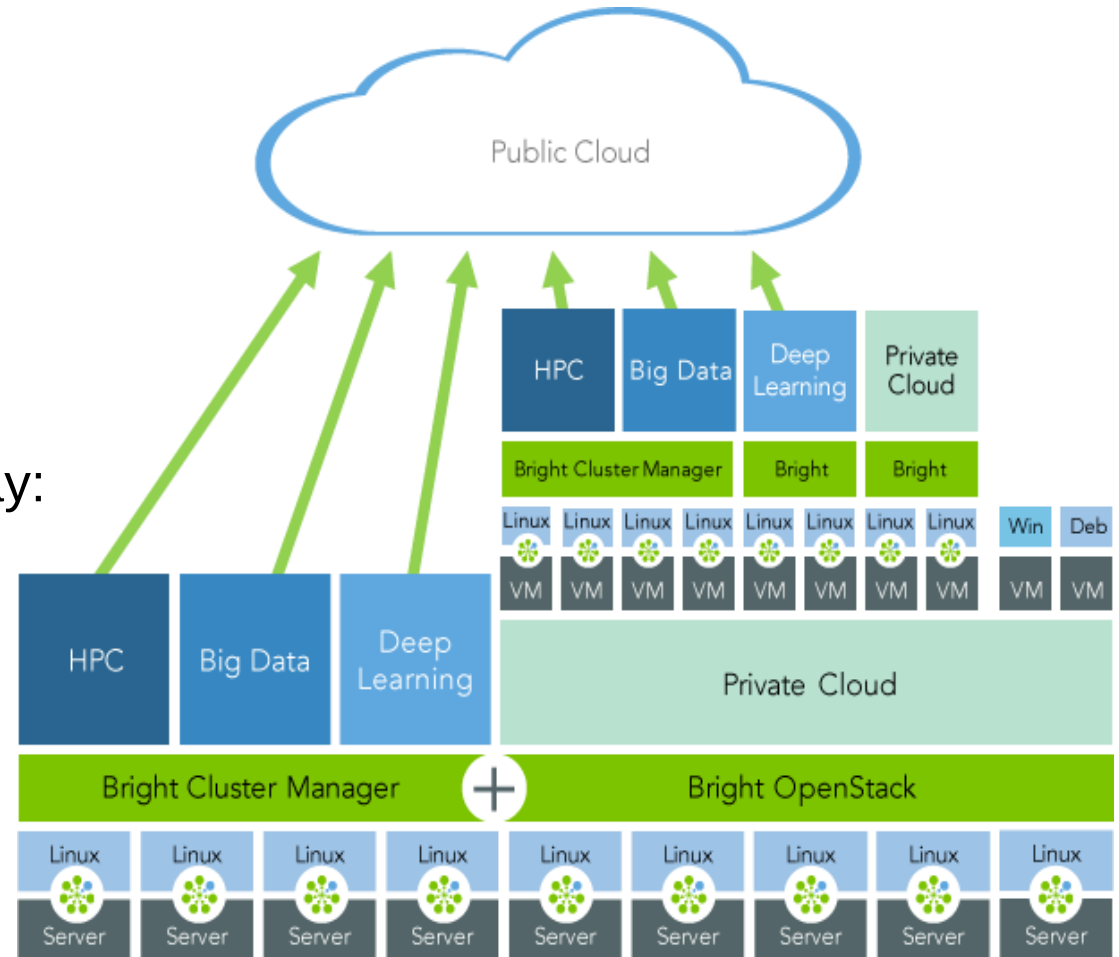
Filesystem0	Size (TiB)	Type
/users	86	GPFS
/apps	58	GPFS
/project	4812	GPFS
/store	3891	GPFS
/scratch/lcg	688	GPFS
/scratch/shared	1434	GPFS
/scratch/snx3000	9011	Sonexion
/scratch/snx2000	904	Sonexion
/scratch/snx1600	2765	Sonexion
Escha /scratch	73	Lustre
Kesch /scratch	73	Lustre
dCache	5386	dCache
NAS	248	GPFS
/object	165	GPFS
/scratch/mythen	546	GPFS
/openstack/mythen	110	GPFS

Spectrum Scale in Openstack4HPC

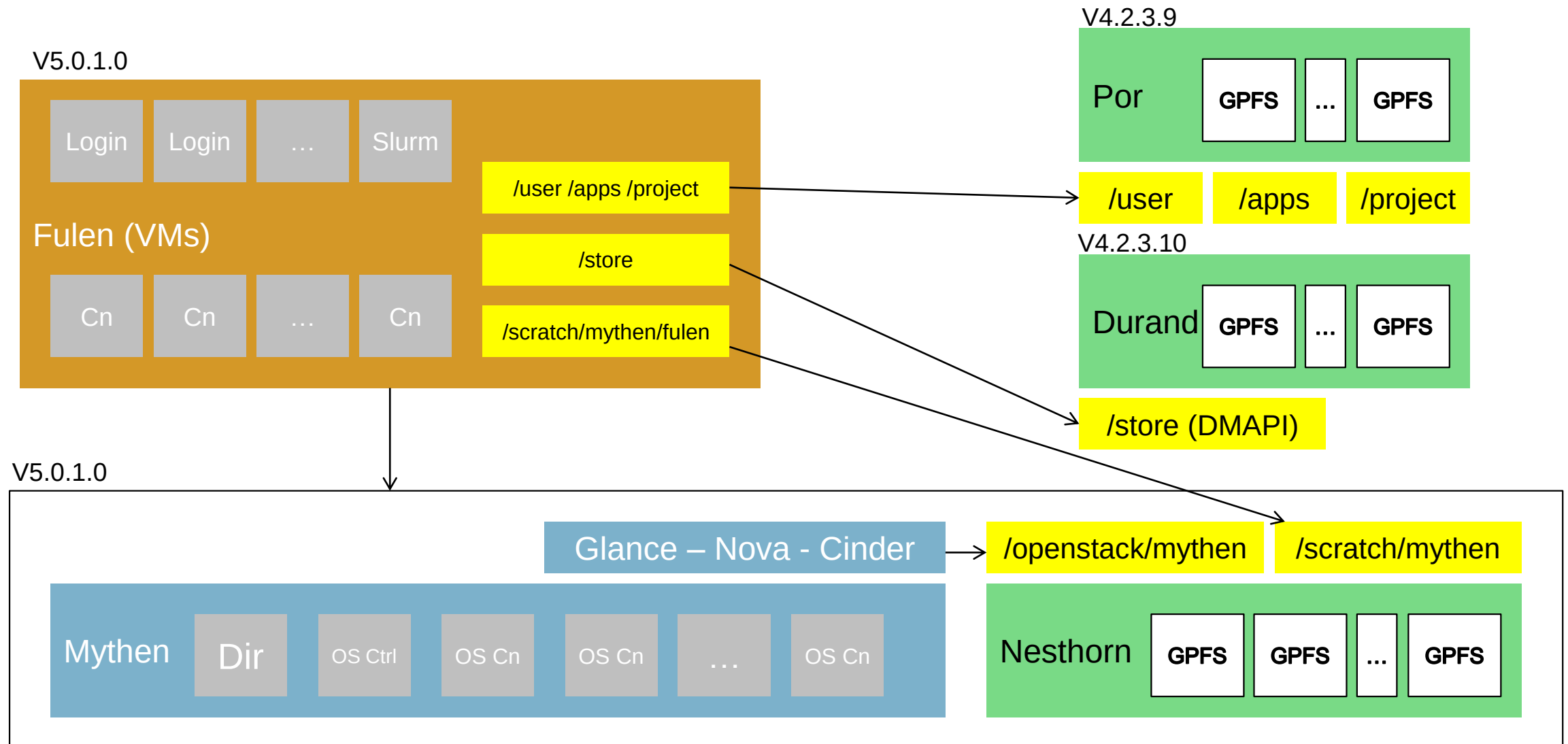
- Continue the infrastructure consolidation process
 - Reduce the number of bare-metal HPC clusters
- Provide HPC optimized clusters with
 - Customer dedicated/isolated environment
 - Flexibility in resource allocation
 - Ability to expand-shrink compute resources
- Reuse in-house OpenStack knowledge developed for the IAAS service
 - Already running OpenStack system for portals, agents, user-to-user services
 - Different scope: want to provide HPC services managed by us
- Availability of commercial support

Spectrum Scale in Openstack4HPC

- RedHat
 - Currently supporting the IAAS system
 - Complex (TrippleO)
 - Support not always satisfactory
- Bright Computing
 - Bright Openstack (OS made easy ...), they say:
 - Installation wizard
 - Start from bare metal
 - Integrated monitoring
 - Easier network setup
 - Good EU support
 - Good feedback from previous projects



Spectrum Scale in Openstack4HPC



Spectrum Scale in Openstack4HPC

- GPFS perf degradation in the Virtual cluster
 - Seq READ bandwidth: 4%
 - Seq WRITE bandwidth: 20%
 - Why 20% drop?
 - Need to test also random IO
- Cinder setup
 - GPFS and Openstack servers in the same GPFS cluster
 - Easier Cinder setup
 - Not comfortable from the serviceability stand point
 - Two logically different clusters
 - Different administrators
 - Pros/Cons of multi-cluster Cinder setup?
- Issues between Virtual cluster and DMAPI fs (Case TS001278091)
 - Mmbbackup failure: fixed upgrading from v4.2.3-9 to v4.2.3-10
 - Expel problem: almost every time we run a ILM scan (mmbbackup or reporting scripts). Fixed by applying kernel/GPFS tunings

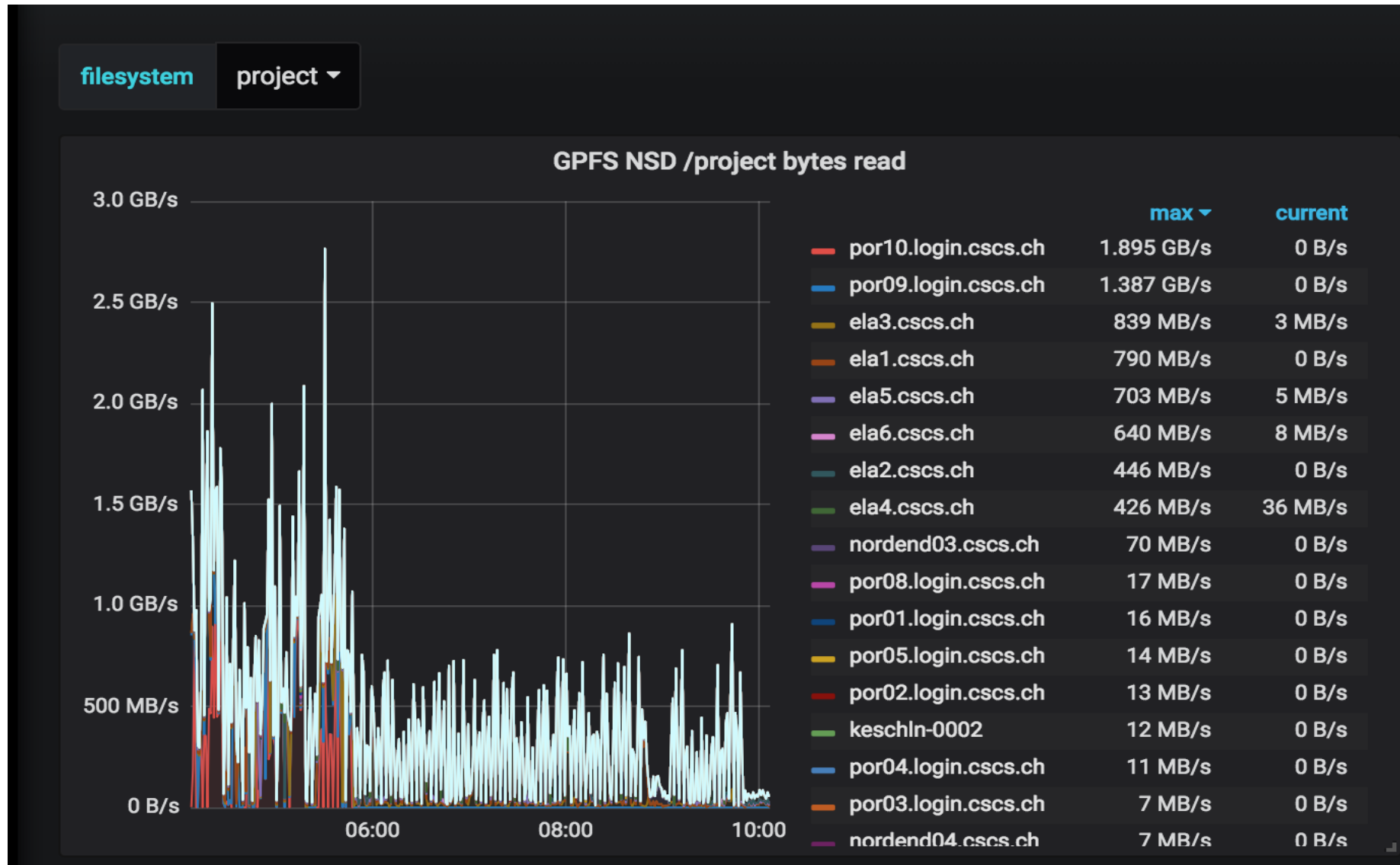
Monitoring

- Dismissed Ganglia/mmpmon monitoring
 - gmond often getting stuck
 - In-house mmpmon wrapper needed additional config steps for sysadmins
 - Limited number of metrics
- Adoption of Zimon/mmperfmon, because
 - Now possible, version $\geq 4.2.3$ everywhere
 - Packaged with GPFS, easier deployment
 - Wide coverage(including metrics for OBJ, NFS, SMB)
 - Data collection is more reliable
 - Gmond stack situations
 - Gmond high CPU usage
- Zimon/mmperfmon issue
 - Work at cluster level -> data disperse on many different servers
 - Monitoring data preservation difficult, backup of tens of separate DBs?
 - We prefer to have a central repository for GPFS monitoring data
 - Grafana bridge doesn't help with that, it leaves data in ZimonDB

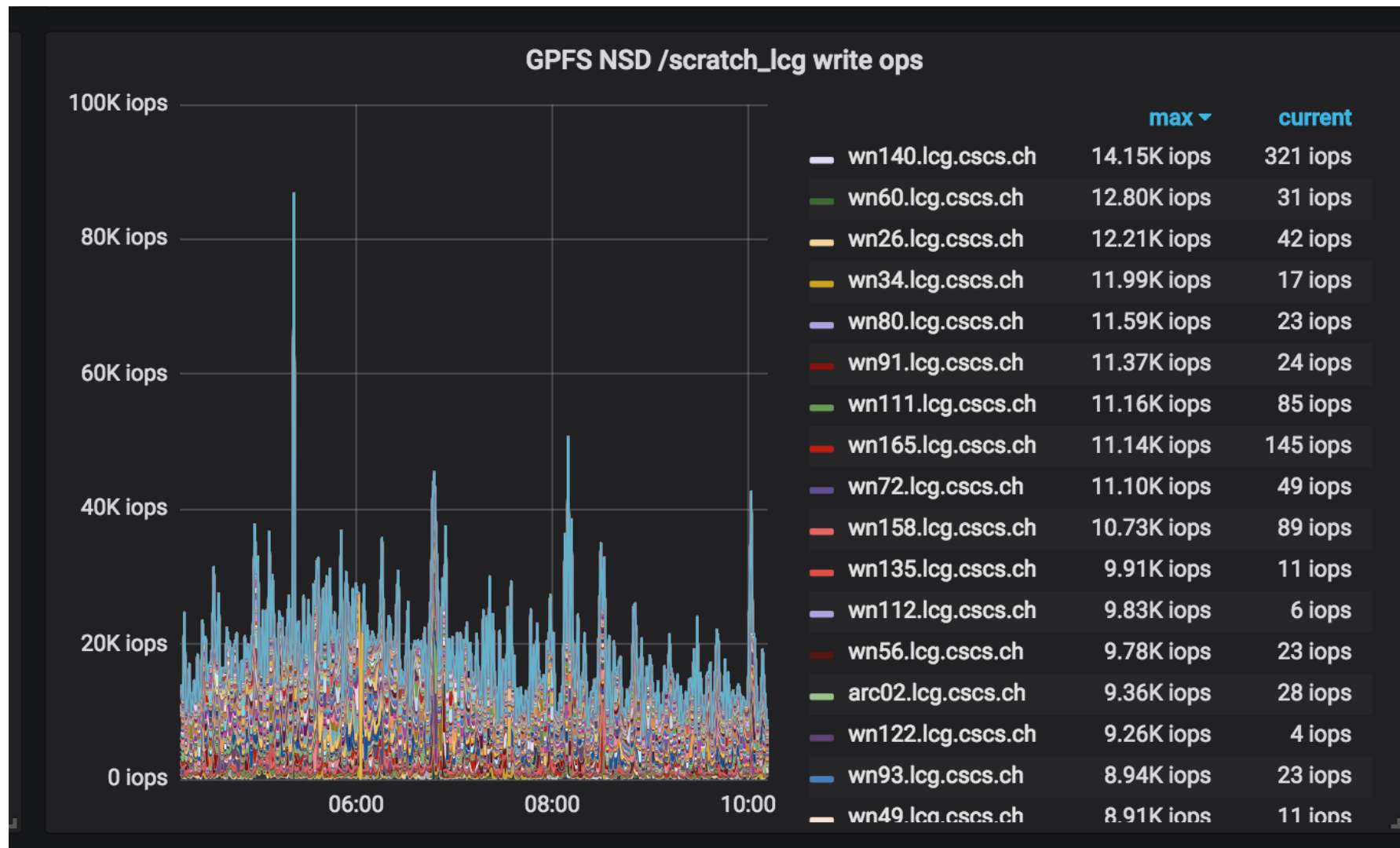
Monitoring

- Implemented a simple collector which
 - Connects to each cluster pmcollector service
 - Gets metric values
 - Sends them to our ElasticSearch service
- Grafana for the visualization
 - one single dashboard for all file systems

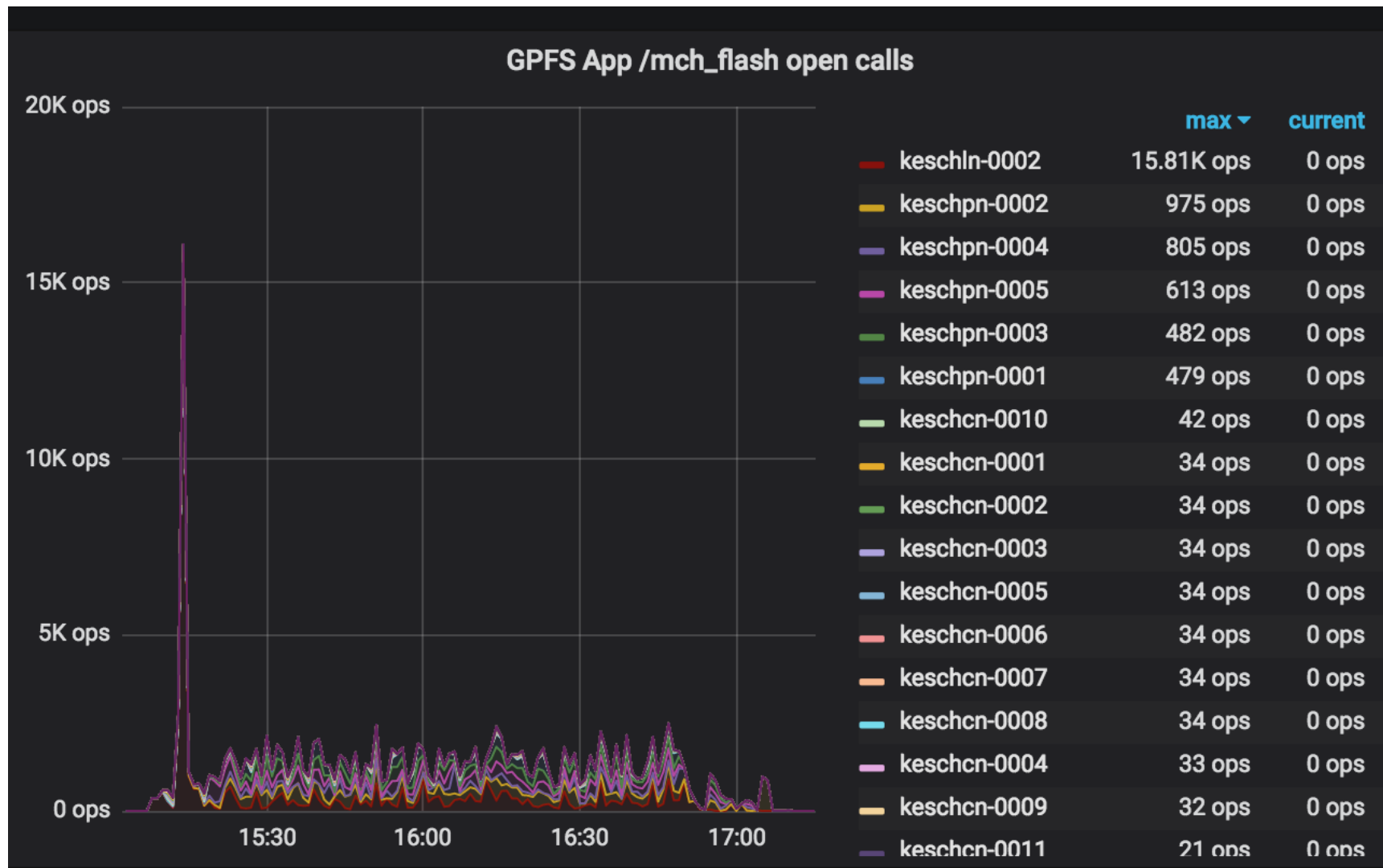
Monitoring



Monitoring



Monitoring



Conclusions

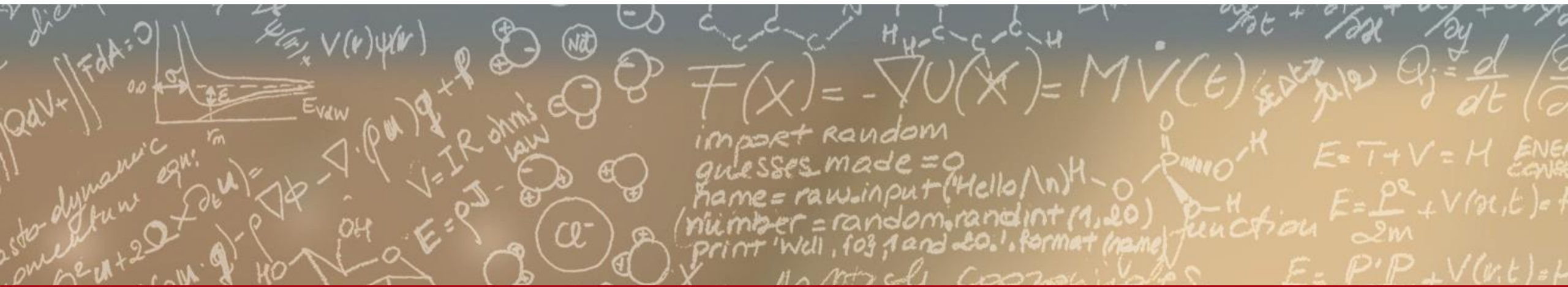
- GPFS in Openstack for HPC env going production
- Open points
 - Performance degradation
 - Cinder/GPFS setup
- Monitoring is now based on Zimon
- ElasticSearch to store monitoring data
- Grafana for data visualization
- Open points
 - Time granularity, 1 min now, smaller interval is reliable?
 - Metrics documentation?



CSCS

Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Thank you for your attention.

CSCS overview

