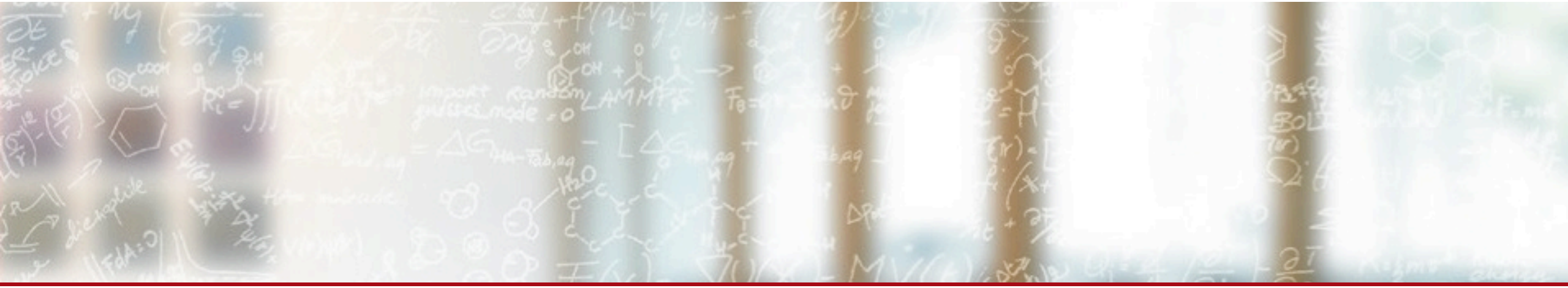
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

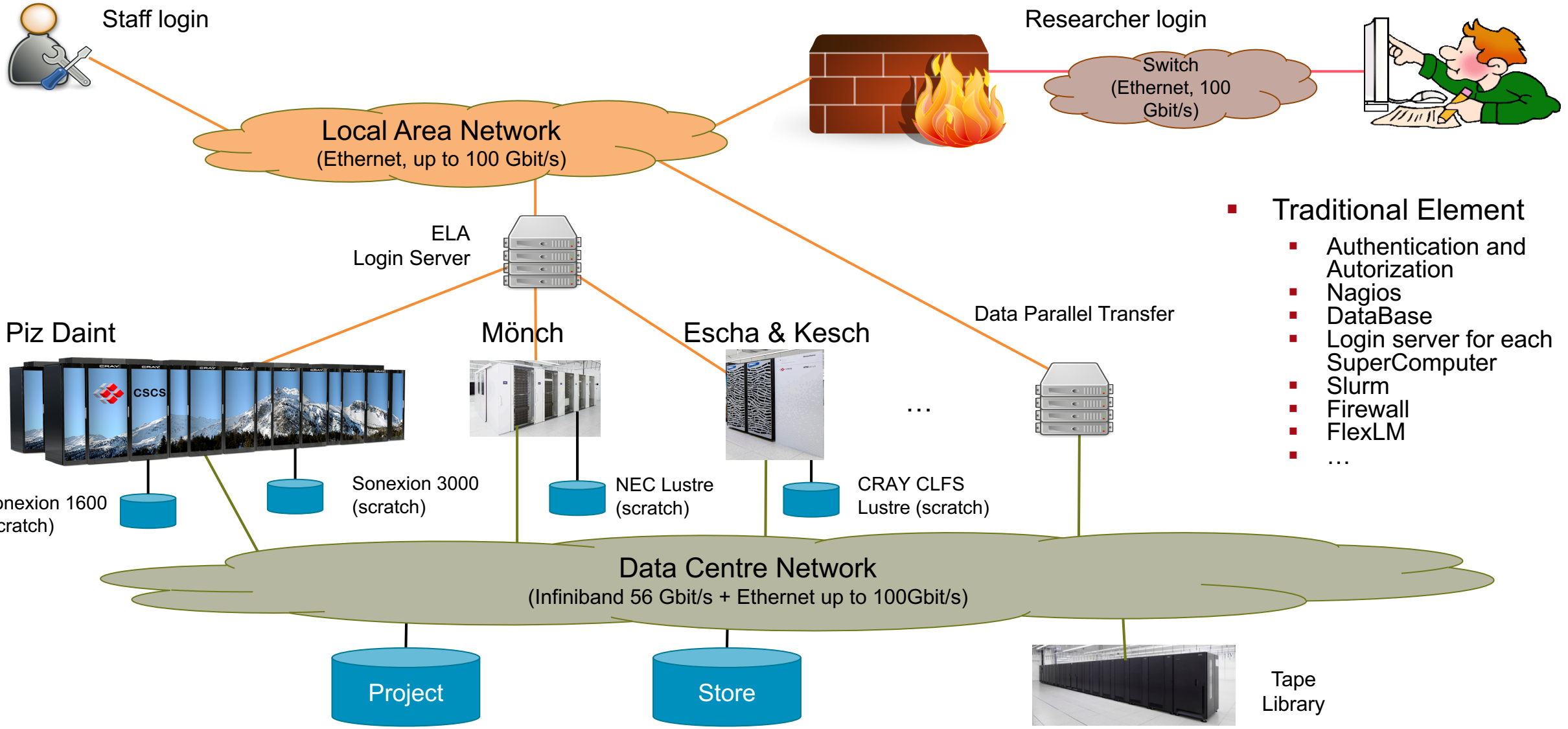


CSCS Site Update

Stefano Gorini, Matteo Chesi, Carmelo Ponti, CSCS

May 31st, 2017

CSCS IT Architecture



- Traditional Element
  - Authentication and Autorization
  - Nagios
  - DataBase
  - Login server for each SuperComputer
  - Slurm
  - Firewall
  - FlexLM
  - …

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

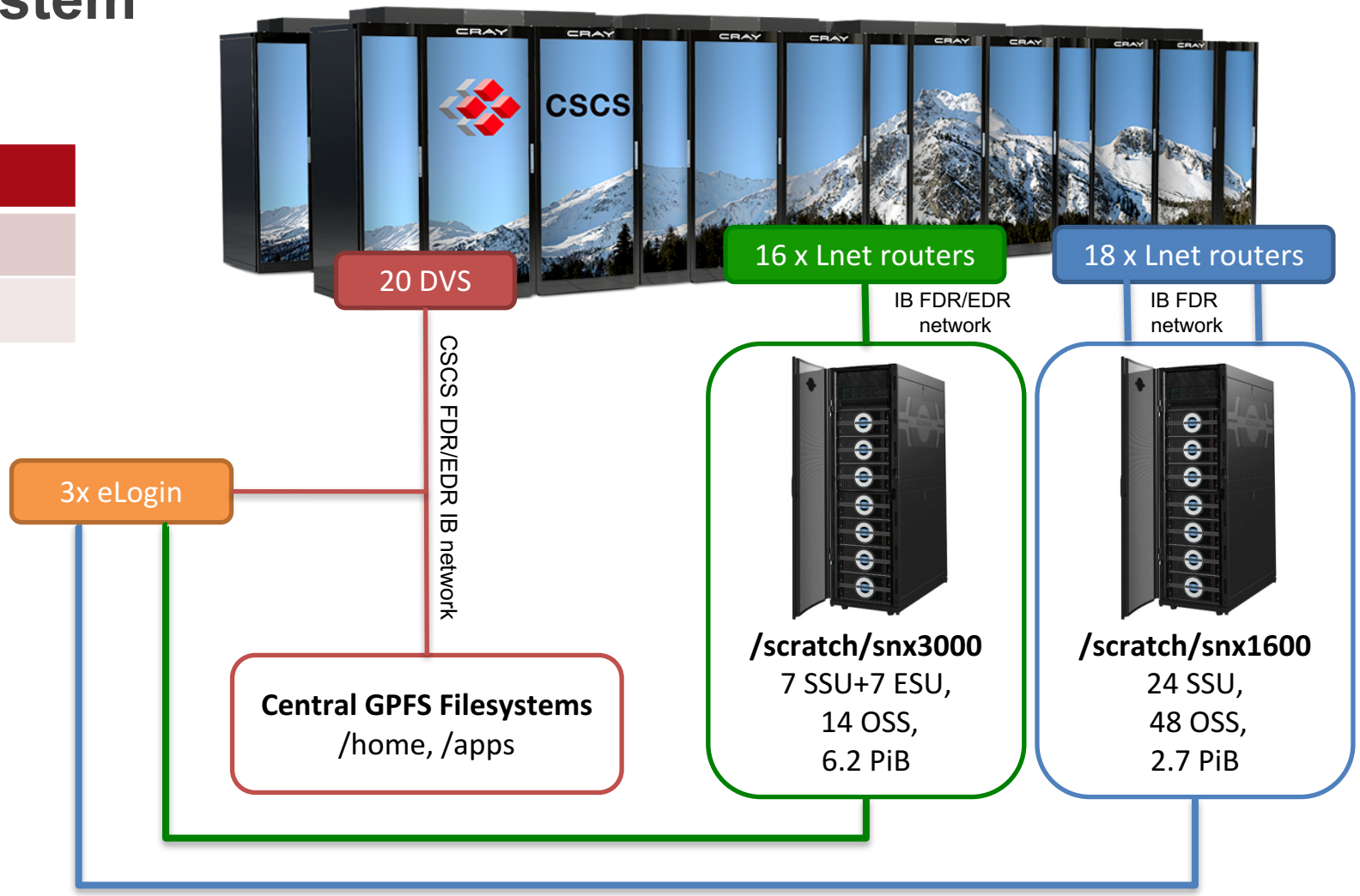
ETH zürich

Lustre Filesystems @ CSCS

Lustre for the Flagship System

Filesystem	Size	GB/s
/scratch/snx1600	2.7 PiB	120
/scratch/snx3000	6.4 PiB	80

- optimized for very big files
- optimized for writes
- Lustre 2.5
- ~6K client nodes
- Robinhood for cleaning policies



Lustre for TDS and R&D System

- Test and Development Systems: Cray Sonexion 1600 & 2000

- Cray Sonexion 2000 for R&D systems
  - Lustre 2.5
  - Declustered RAID (GridRAID)
  - New Expansion Storage Units
  - 4 OSSs with 2 OSTs each one
  - 41 disks (113 TiB) per OST
  - stripe_count=1

- Management Infrastructure (Nagios, Ganglia, Puppet, Greylog, custom solutions…)

MCH System – CRAY CLFS Lustre

Filesystem	Size
Escha /scratch	73 TiB
Kesch /scratch	73 TiB

NetApp 2760 (2TB drives, NL-SAS)

2 CLFS Servers (OSS, MDS, MGS)

Server:
CentOS release 6.4 (Final)
Lustre: 2.5.0

Client:
Red Hat Enterprise Linux Server release 6.7 (Santiago)
Lustre: 2.5.4

Monch – NEC Lustre

Filesystem	Size
Monch /scratch	350 TiB

Server:
CentOS release 6.4 (Final)
Lustre: 2.1.6

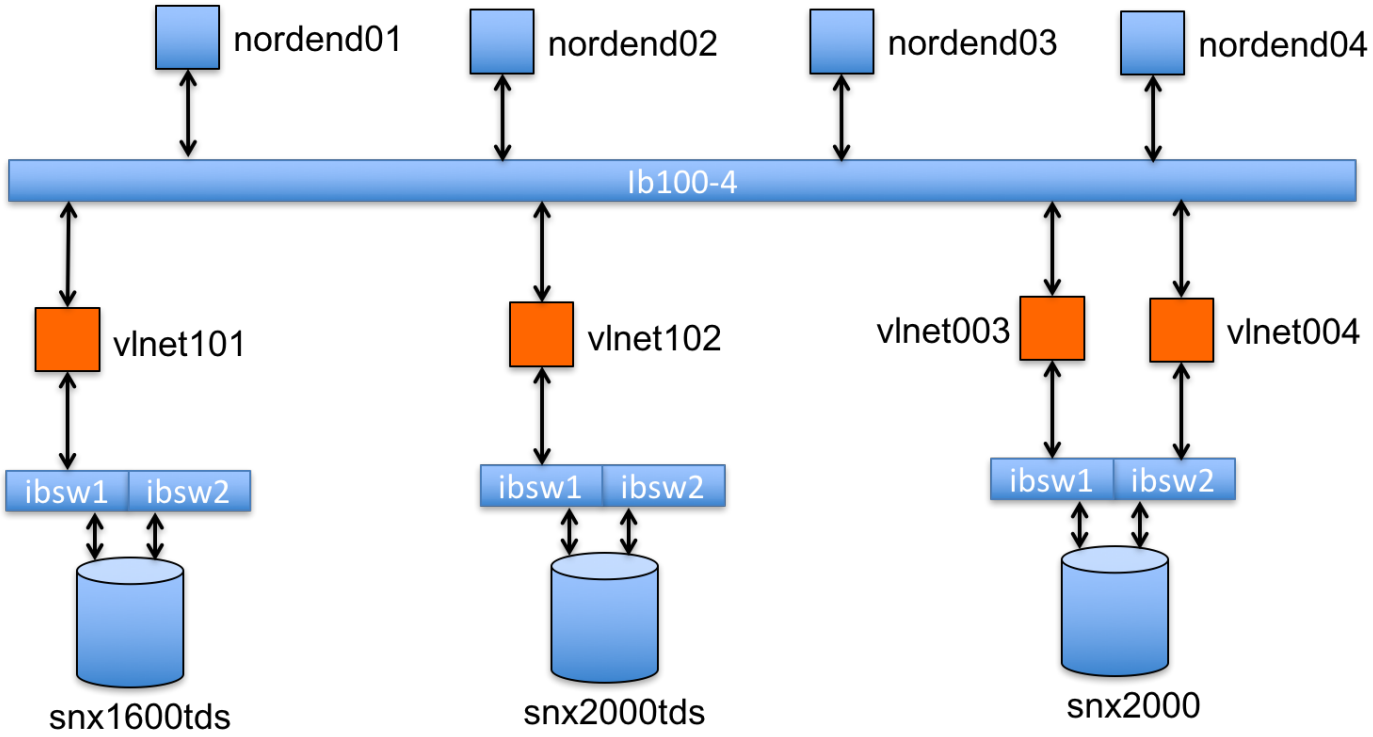
Client:
CentOS release 6.7 (Final)
Lustre: 1.8.9

Data Movers

Data Transfer Service

The Data Mover nodes are managed via SLURM in order to create dependency and a clear workflow with HPC Jobs and data movement via specific tools:

- GRIDFTP
- move
- Cp
- Rsync
- …



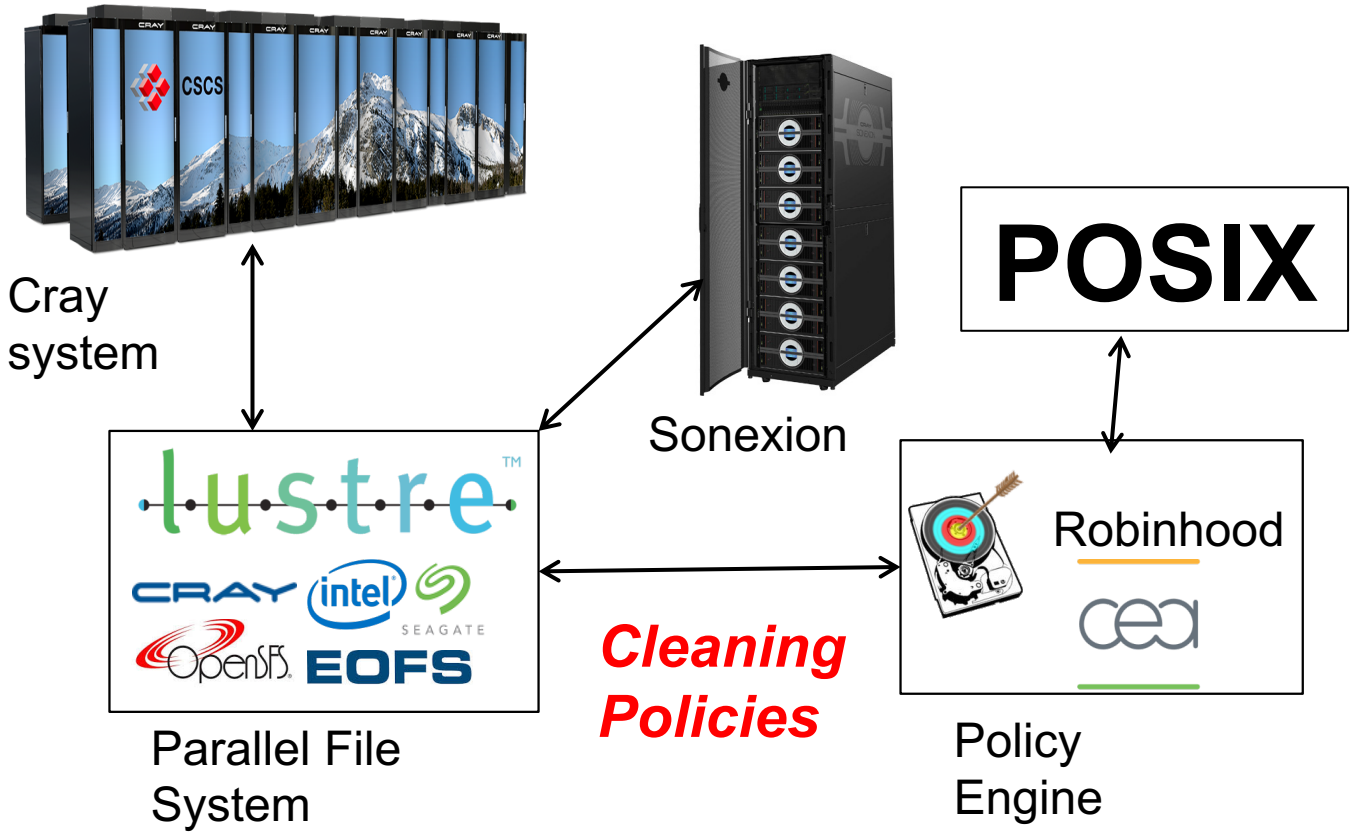
RobinHood

RobinHood

Each Lustre at CSCS as a dedicate RobinHood Server to perform the proper cleaning policy

Unfortunately on the main HPC System we are not able to do a real time check of the file system:

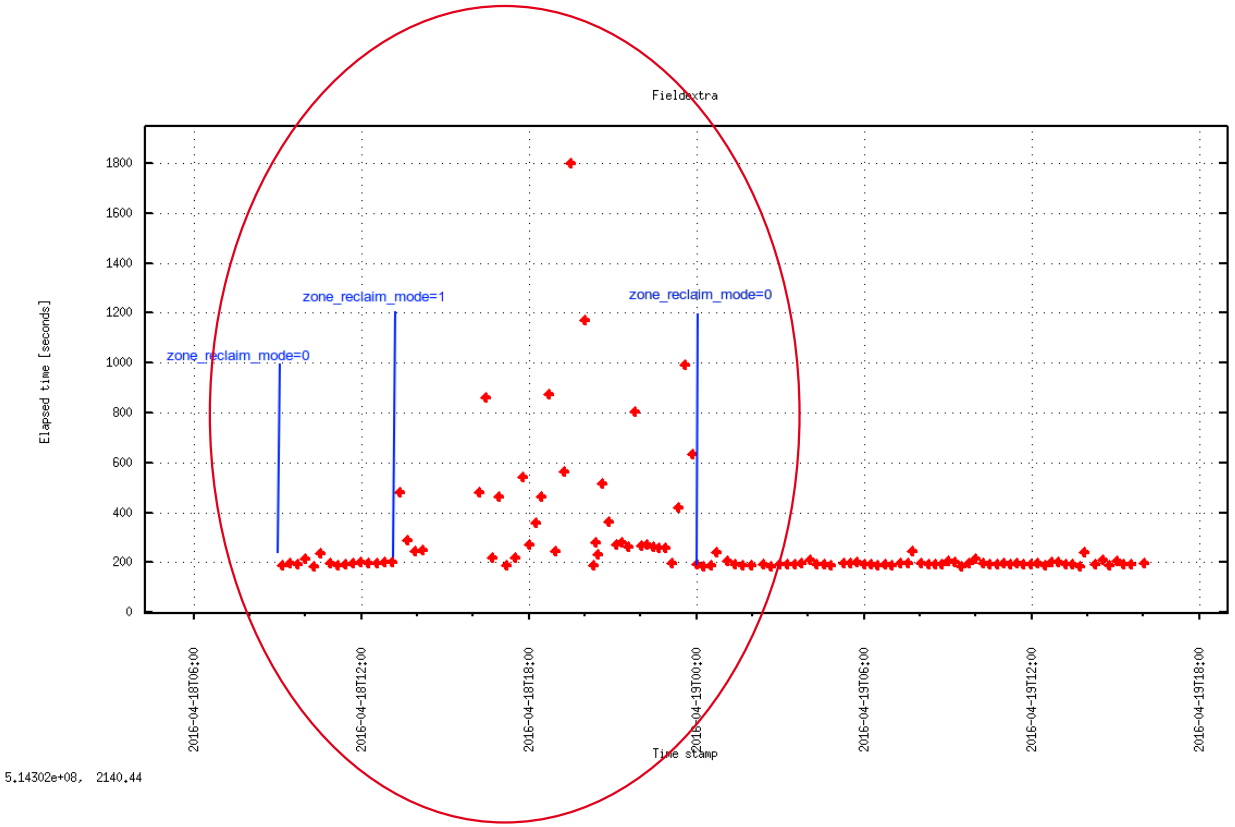
- Changelog is too slow in respect of the change rate we have on the FS
- Cleaning Policy is running base on a 24h FS Scan



Application Vs Lustre

Description of the Problem

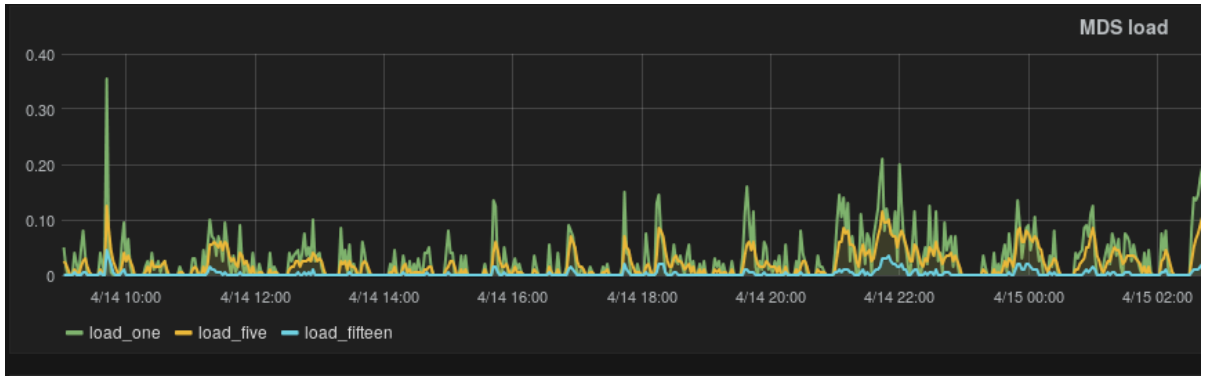
- Application (pre/post processing Fortran tool) slowdown

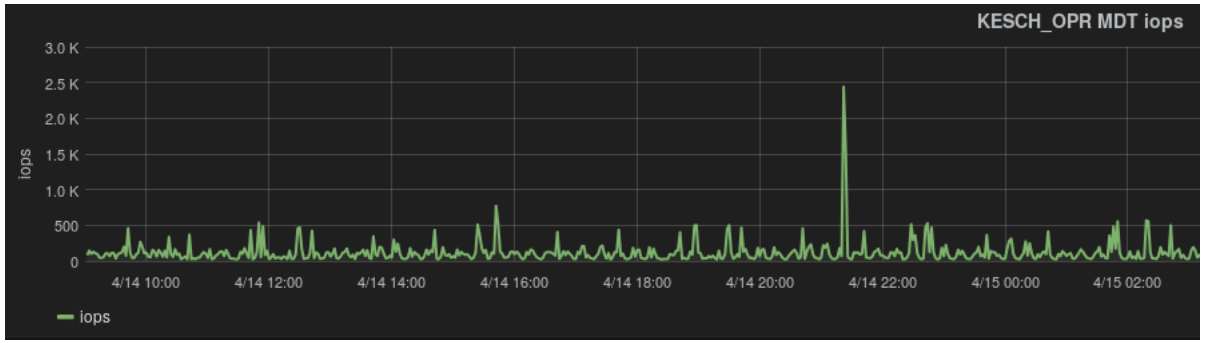


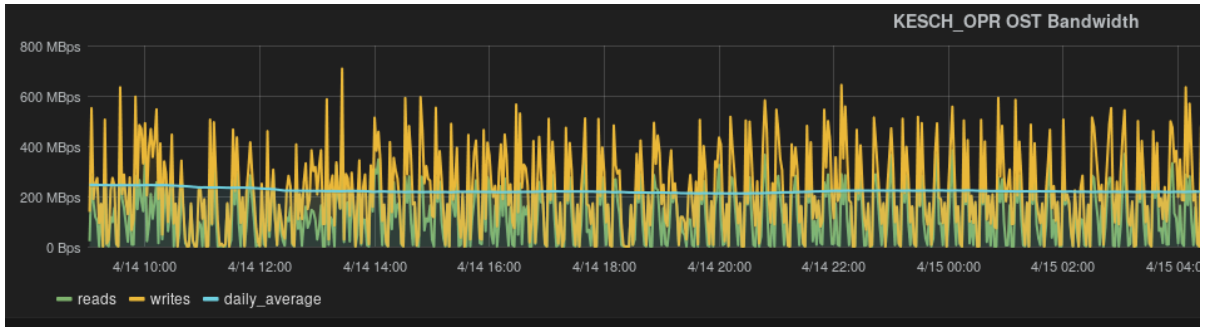
Condition	zone_reclaim_mode	Number of Runs	Average [s]	Standard dev [s]
2	0	**15**	198.533	12.928
3	**1**	**38**	**440.921**	**337.741**
4	**0**	**62**	**193.677**	**27.617**
5	0	**161**	499.379	1133.936
6	**0**	**173**	**199.08**	**11.316**

Is it the FS?

- Lets try GPFS…..
  - No Variation the application always perform the same
- So is it Lustre FS storage HW? ….No







Dedicated Test and Analysis Session

- All the problems are not related to an high load on the Lustre file system
- The kernel parameter reclaim `vm.zone_reclaim_mode` has a significant effect on the slowdown (“condition 5”)
- Running the suite on the same node mitigates the slowdown

- Important Remark:

  During the analysis of the application process with ***perf***, in case of slowdown, The application was spending a lot of time with the kernel function **clear_page_c_e**:

  Samples: 1M of event 'cycles', Event count (approx.): 854374192198
  13.12% [kernel.kallsyms] [k] **clear_page_c_e**
  7.58% application_12.2.0_gnu4.9.3_opt_omp [.] spumb_c_
  7.35% [kernel.kallsyms] [k] compaction_alloc

Solution

- The customer redesigned the initialization of data arrays (~40 GB on disk) by doing this initialization stepwise.
- With this new version of the library no significant performance fluctuation has been seen.
- Running the test case during more than 12 hours without cache cleaning on all nodes (“condition 5”)
- The new initialization even improves the performance
- BUT the problem is still there:

“Lustre 2.9: is it fixed in this version?”

Next Challenges

What next ?

CRAY system

Sonexion

CRAY TIERED ADAPTIVE STORAGE

HPSS
High Performance Storage System

POSIX

lustre™
CRAY intel SEAGATE
OpenSFS EOFS

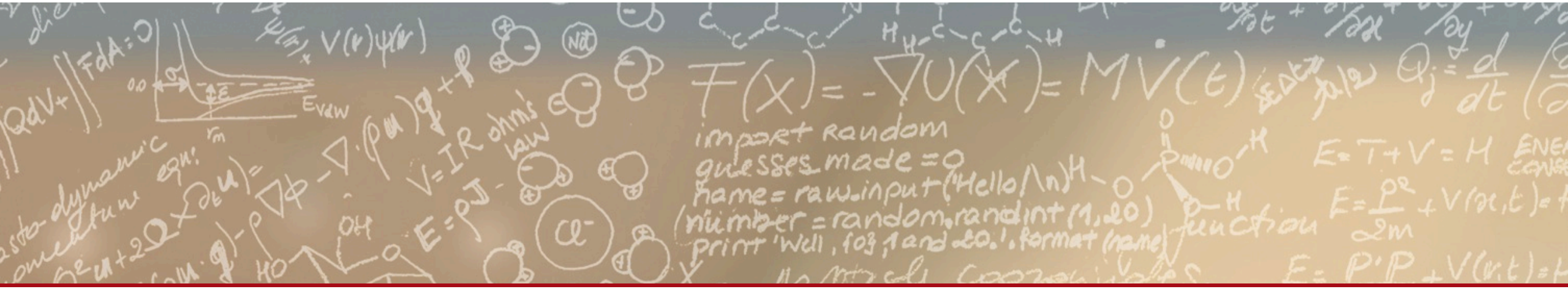
Parallel File System

Robinhood
cea

Policy Engine

openstack
CLOUD SOFTWARE

Swift


import random
guesses_made = 0
name = raw_input('Hello! What is your name?\n')
number = random.randint(1, 20)
print 'Well, ' + name + ', I am thinking of a number between 1 and 20.'

Q & A