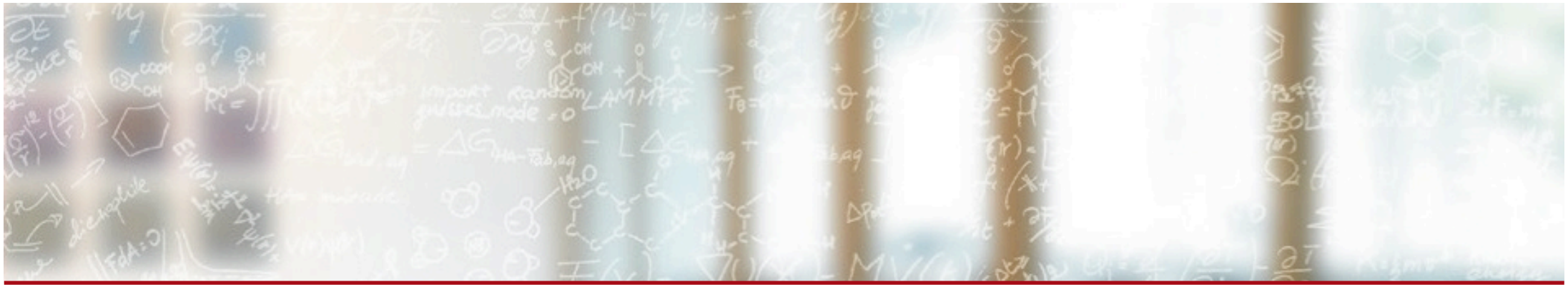
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



HPC storage @ CSCS

Stefano Gorini and Colin McMurtrie, CSCS

June 23rd, 2016

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

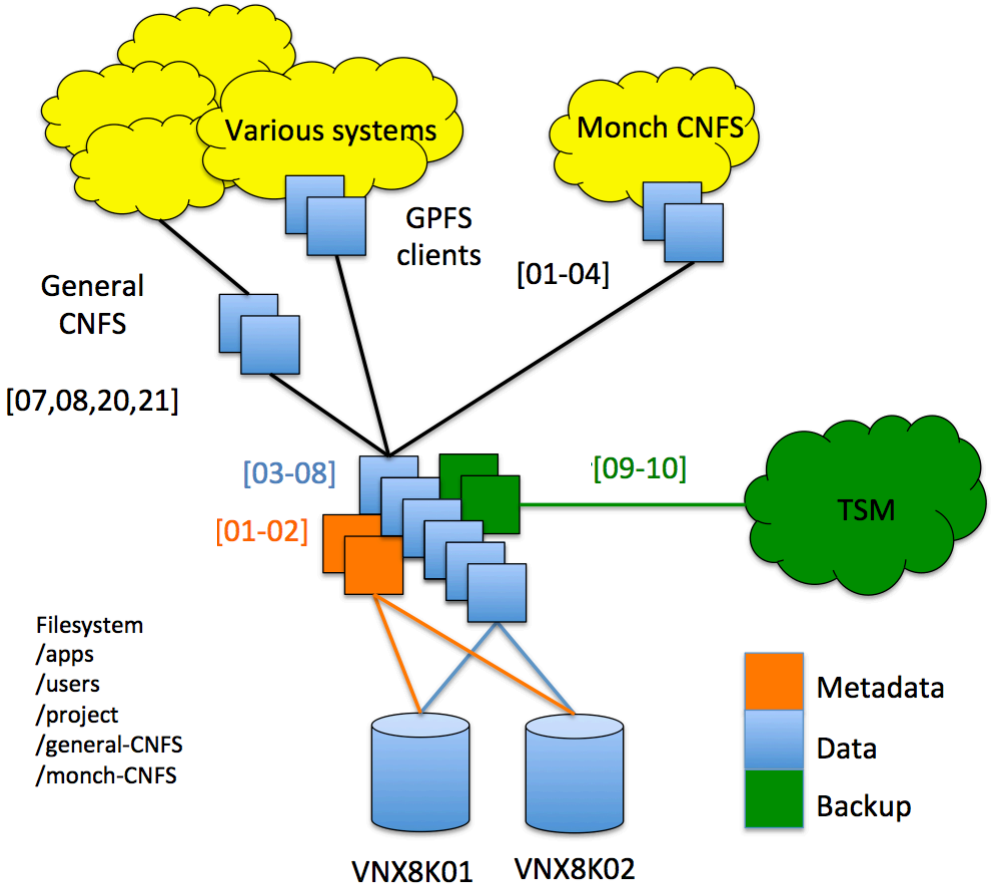
ETH zürich

Online Filesystems

Mission Critical Filesystems

Filesystem	Size
/users	86 TiB
/apps	58 TiB

- optimized for small files
- GPFS 3.5
- blocksize: 256 KiB
- metadata on SSD in double copy
- files < 128 KiB on inodes

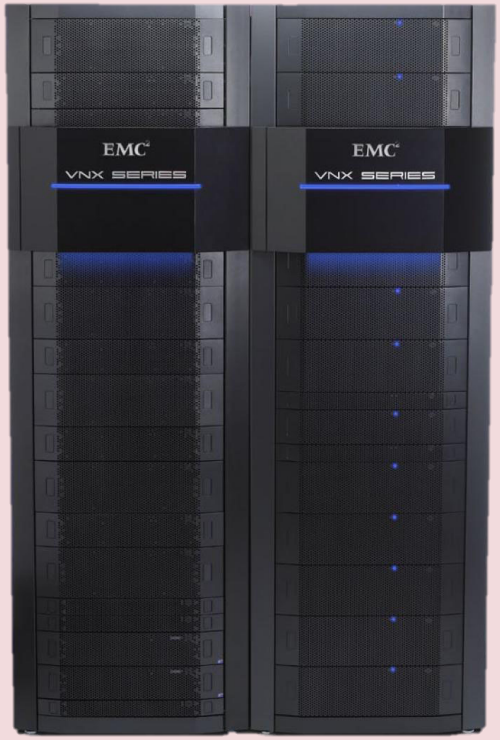


Net: 144 TiB | Raw: 200 TB | 50

Online Filesystems with Backup - /project

Filesystem	Size
/project	5.8 PiB

- optimized for big files
- GPFS 3.5
- blocksize: 1 MiB
- metadata on SSD in double copy
- files < 256 KiB on inodes
- NEW 2015: + 2 PiB
- Quota based on research proposals



- 2 EMC VNX8000
- 37 disk enc. (60 slot each)
- RAID6 + Hot spares
- 4 TB NL-SAS disks
- SSDs for Metadata

Net: 5.94 PiB | Raw: 8.65 PB | 2163

Online Filesystems with Backup - /store

NEW

Filesystem	Size
/store	4.4 PiB

- GPFS 4.1
- 2 Tiers
- Policy based on access time
- Quota based on contracts
- Previous size: 2.6 PiB

- 1 EMC VNX8000
- 8 disk enc. (60 slot each)
- RAID6 + Hot spares
- 4 TB NL-SAS disks
- SSDs for Metadata

- 3 NetApp E5600
- 18 disk enc. (60 slot each)
- parity-declustering RAID
- 6 TB NL-SAS disks

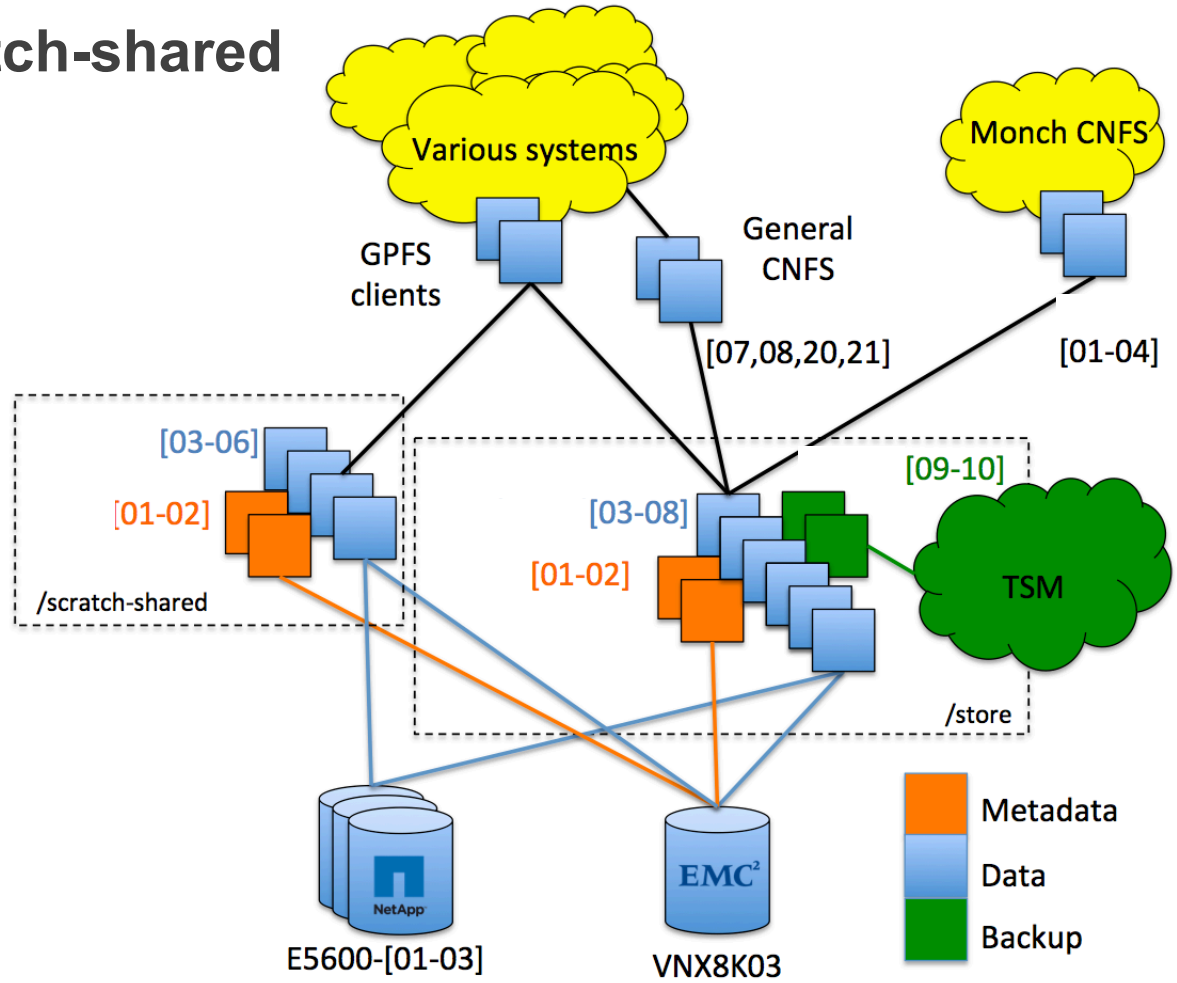
Net: 10.34 PiB | Raw: 15.21 PB | 3412

Online Filesystems - /scratch-shared

NEW

Filesystem	Size
/scratch-shared	1.2 PiB

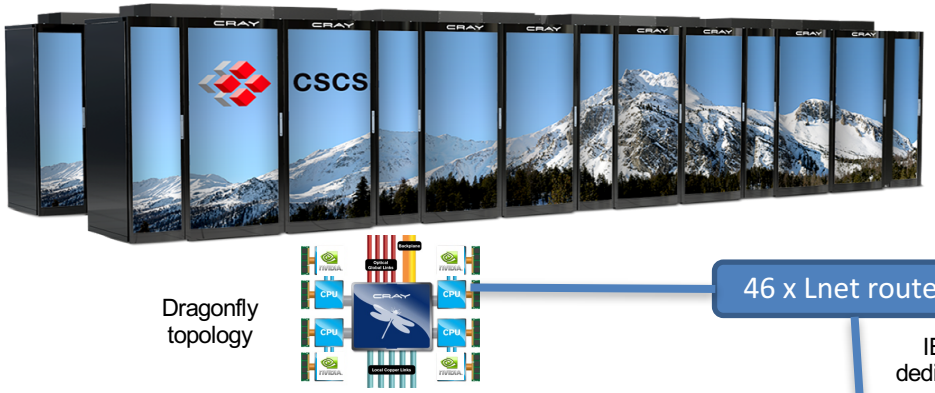
- GPFS 4.1
- No backup
- Prev. size: 642 TiB



Net: 11.54 PiB | Raw: 17.01 PB | 3712

Online Filesystems - /scratch/daint

Filesystem	Size
/scratch/daint	2.7 PiB



- optimized for very big files
- optimized for writes
- 116 GiB/s as peak performance
- Lustre 2.1
- 6582 client nodes (dora+daint)
- Robinhood for cleaning policies

Net: 14.24 PiB | Raw: 20.95 PB | 5680

Other Storage Systems

Filesystem	Size
/scratch/santis	167 TiB
/scratch/dora	904 TiB
/scratch/brisi	226 TiB

- Test and Development Systems for Cray Sonexion 1600

- Cray Sonexion 2000 for Dora and its TDS systems
  - Lustre 2.1
  - Declustered RAID (GridRAID)
  - New Expansion Storage Units
  - 4 OSSs with 2 OSTs each one
  - 41 disks (113 TiB) per OST
  - stripe_count=1

NEW

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

Net: 15.51 PiB | Raw: 22.84 PB | 6172

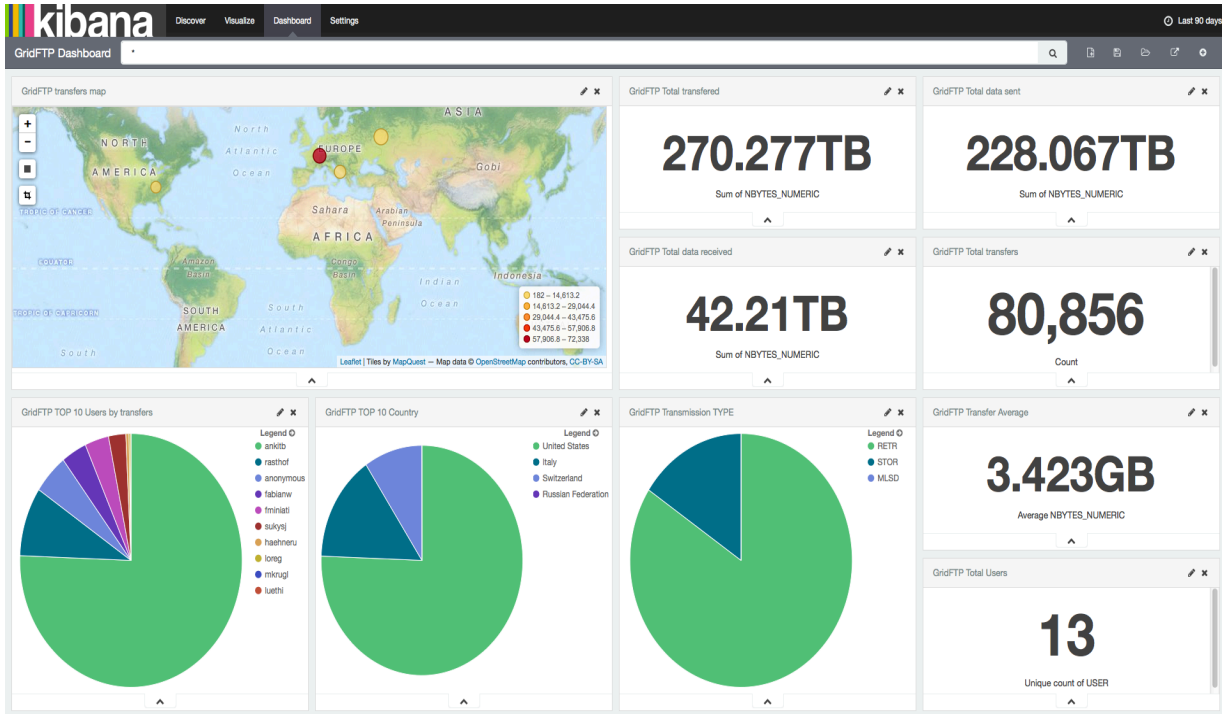
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

Data Services

Data Transfer Service

- Data Movers Services
  - NEW GridFTP (4 nodes cluster)
  - GPFS AFM for HBP
    - Between CSCS and EPFL
    - To be extended to Juelich, Cineca and BSC.



Backup/Archive Service

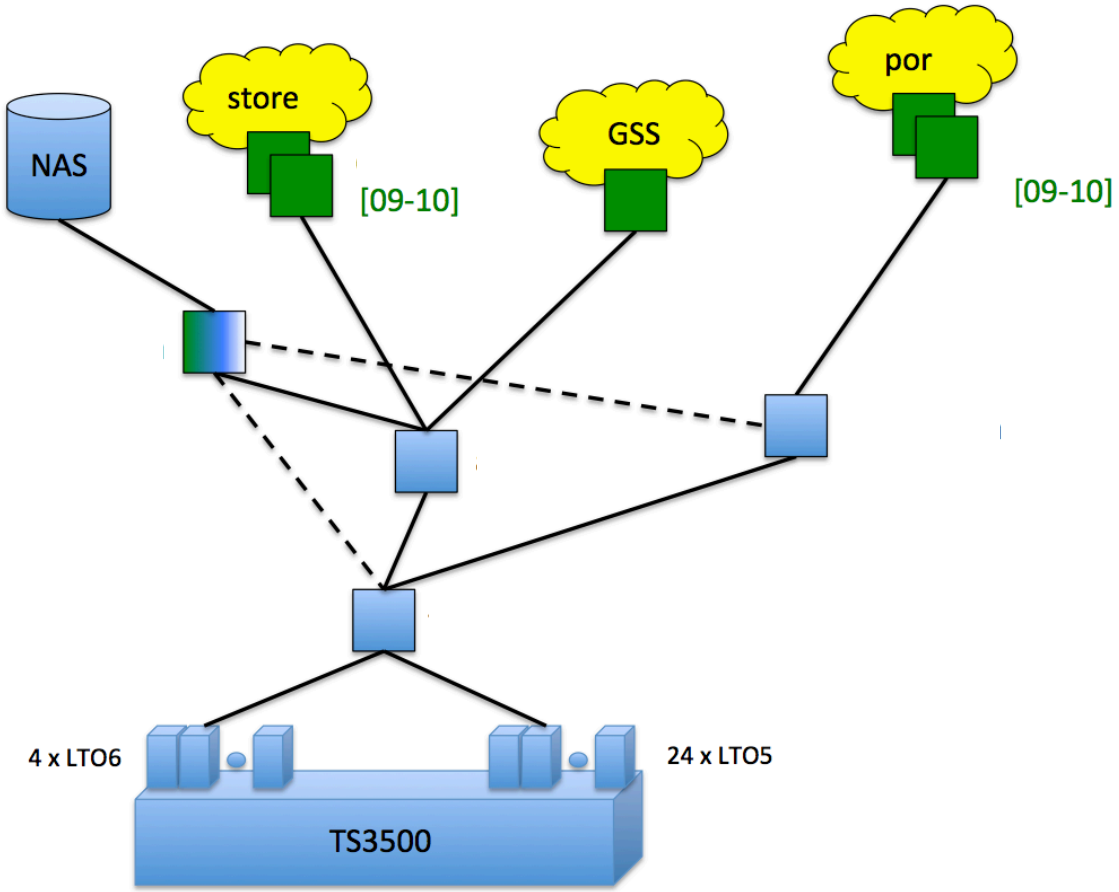
- 3 TSM Servers + 1 spare

- IBM TS3500 Tape Library (18257 slots)
- 28 drives (24 LTO5 + 4 LTO6)
- 12510 LTO5 + 100 LTO6 cartridges

- Mainly used with mmbackup for GPFS
- 5 Storage Agents
- Big DB2 databases (~ 400GB) to keep metadata infos

Raw: 19.01 PB | 12610

Backup/Archive Service



CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

Customized Solutions

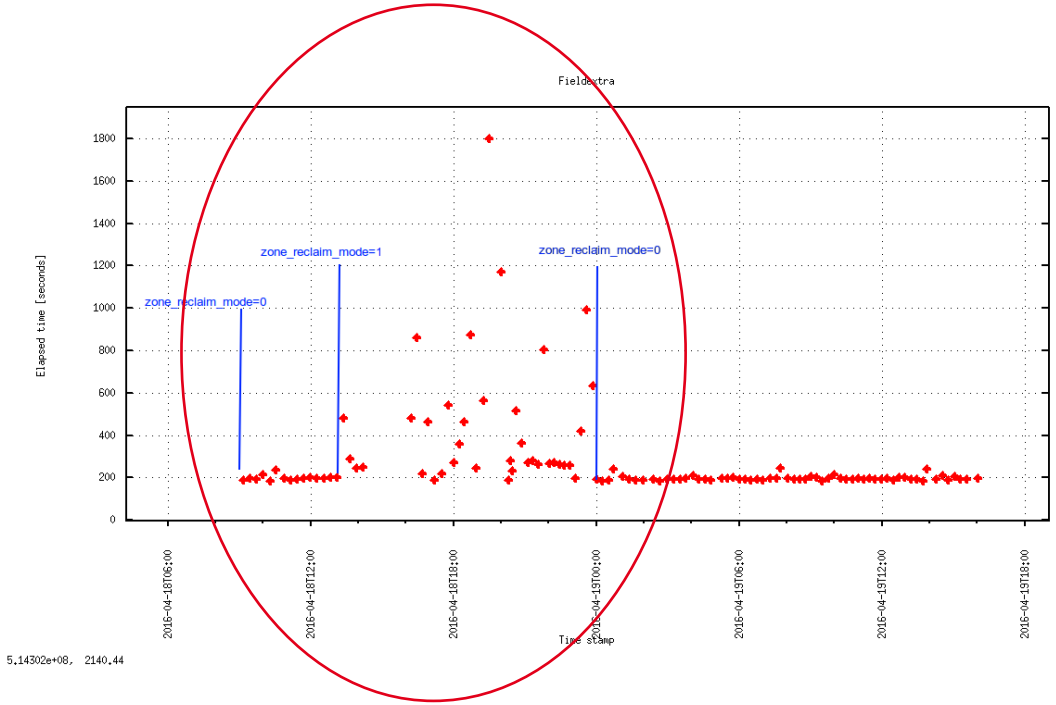
MeteoCH

Filesystem	Size
/workspace	223 TiB
Albis /opr	18 TiB
Lema /opr	18 TiB
Escha /scratch	73 TiB
Kesch /scratch	73 TiB

- Cray Sonexion 1300 for old /workspace
- Cray Lustre for old /opr and new /scratch
- built on NetApp hardware

Description of the Problem

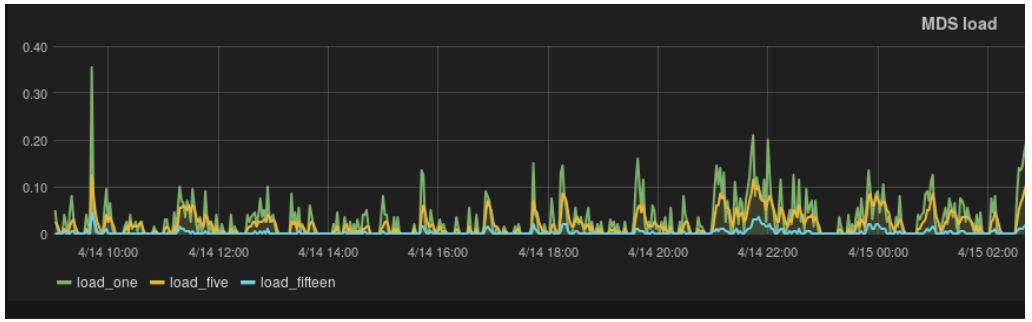
- FIELDEXTRA (pre/post processing Fortran tool) slowdown

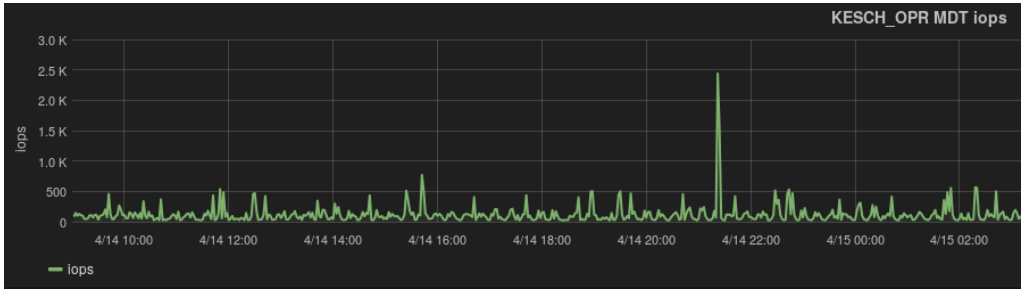


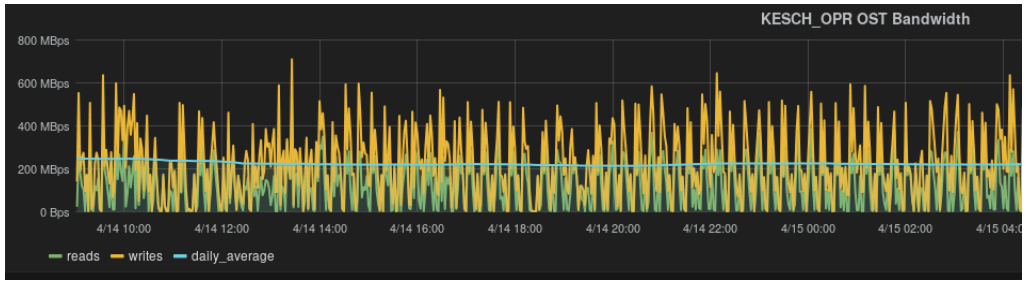
Condition	zone_reclaim_mode	Number of Runs	Average [s]	Standard dev [s]
2	0	**15**	198.533	12.928
3	**1**	**38**	**440.921**	**337.741**
4	**0**	**62**	**193.677**	**27.617**
5	0	**161**	499.379	1133.936
6	**0**	**173**	**199.08**	**11.316**

Is it the FS?

- Lets try GPFS…..
  - No Variation FIELDEXTRA 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 Fieldextra process with ***perf***, in case of slowdown, Fieldextra 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% Fieldextra [kernel.kallsyms] [k] **clear_page_c_e**
  7.58% fieldextra fieldextra_12.2.0_gnu4.9.3_opt_omp [.] spumb_c_
  7.35% fieldextra [kernel.kallsyms] [k] compaction_alloc

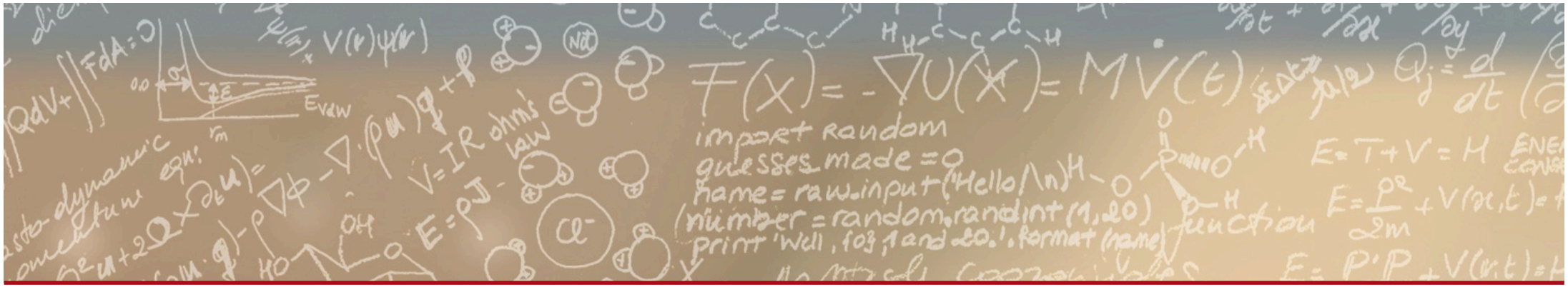
Solution

- MCH redesigned the initialization of data arrays (~40 GB on disk) by doing this initialization stepwise
- With this new version of fieldextra no significant performance fluctuation has been seen → More testing is underway to confirm these results
- Running the test case during more than 12 hours without cache cleaning on all nodes (“condition 5”)
- The new initialization even improves the performance on top of that:

“The test case ~30% faster than the fastest runtime with the current operational executable”

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Q & A

sgorini@cscs.ch - colin@cscs.ch