CERN

Storage strategy

Alberto Pace,

with input from German Cancio, Dirk Duellmann, Massimo Lamanna, Luca Mascetti , Jakub Moscicki, and Dan van der Ster

Current Strategy

- EOS
  - An architecture designed for Exabyte scale. In production
- CERNBOX
  - The sync & share service for offline access to the entire EOS storage
  - Includes web access and an application platform (Gallery, Office, …)
- Online Access
  - Fuse mount from Linux (used by lxplus and lxbatch)
  - Samba access from Windows (planned for Terminal Services)
- Ceph / CephFS / S3 for OpenStack storage and other special / custom cases
  - Used for block devices, local and cluster storage, HPC, databases, object stores, build clusters, filers, …

Current Status

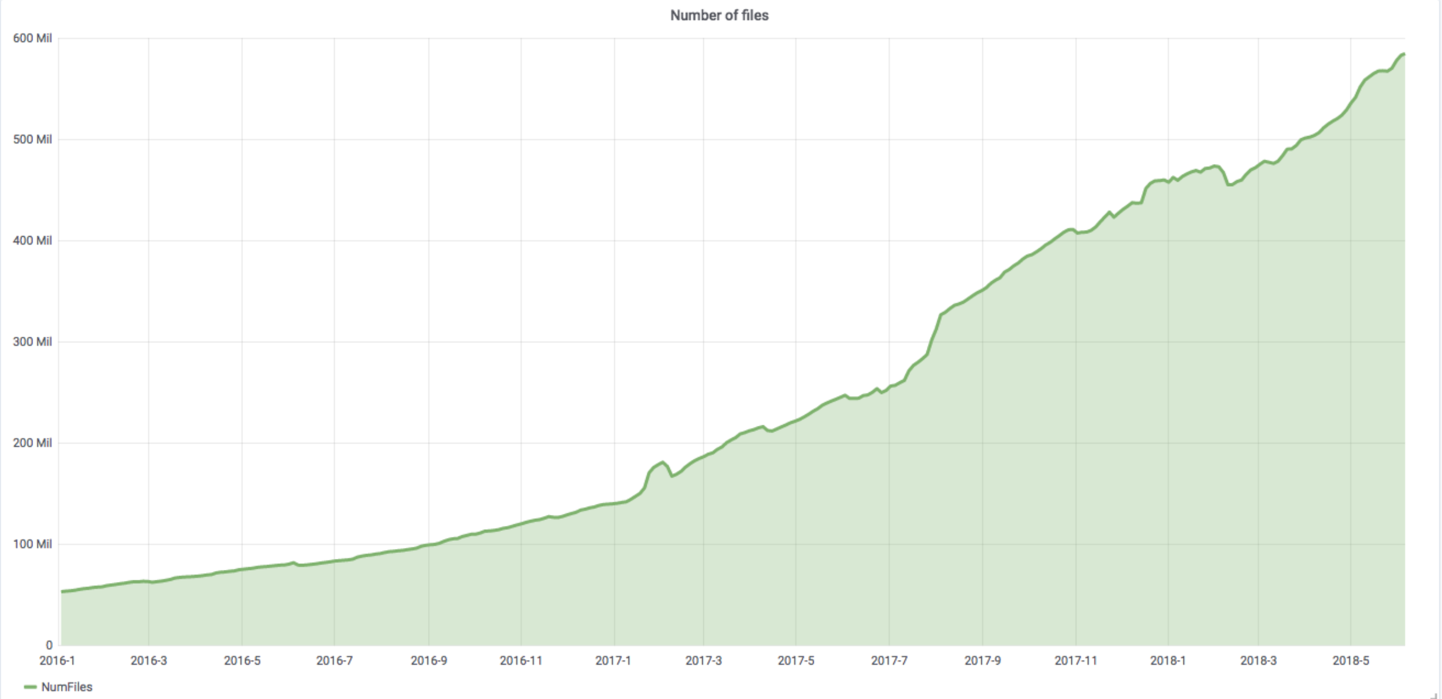
- EOS
  - Ok, instance sizes growing rapidly
  - Need smaller instances for better scale out (see plan on next slides)
- CERNBOX
  - Ok, also growing fast: +240% last 12 months
- Online access using FUSE (old version)
  - Ok, … but access latency not acceptable for some use case
- Ceph
  - Ok, growing fast: +140% last 12 months

Plans

- EOS
  - Needs improvement in isolation and scalability
  - Two plans, both pursued in parallel
    - Split instances into smaller federated ones (eoshome00, eoshome01, …). Unlimited scalability improvements and problem isolation
    - New name server which keeps in RAM only active metadata. Allows 100x scalability increase and 50x reduced reboot time. Hardware split across availability zones (RAFT)
- FUSE - Replace current Fuse with FuseX
    - impressive latency reduction with client side cache
    - direct benefit expected for both Linux and Windows (samba) clients
    - Already deployed on lxplus under /eos/scratch for selected IT users
- Until these improvements are deployed, we have reduced all pressure on end-user to migrate out of AFS

Current figures

- EOSUSER (CERNBOX): 14300 accounts, 584 M files
- Restart time
  - $2\text{x}10^7$ entries per minute today on current EOSUSER
  - less than 1 minute on the new namespace for 10^9 files (100x improvements)
- Scale out instead of scale up required !



EOS detailed plan (with dates)

- May 2018
  - New deployment preparation ✓
  - Batch tests (new FUSE) ✓
  - New MGM stress tests ✓
  - EOSHOME00 up ✓
- June 2018
  - EOSHOME01 up ✓
  - Migration tool to move users from EOSUSER to EOSHOME) ✓
  - Move of IT-ST accounts (EOSUSER aka CERNBox)
- July - September 2018
  - Move of IT accounts
  - New accounts are created on EOSHOME
- Before end of 2018
  - Transparent move of larger groups
  - Finalise the move
    - e.g. Critical account that might have an impact on LHC data taking

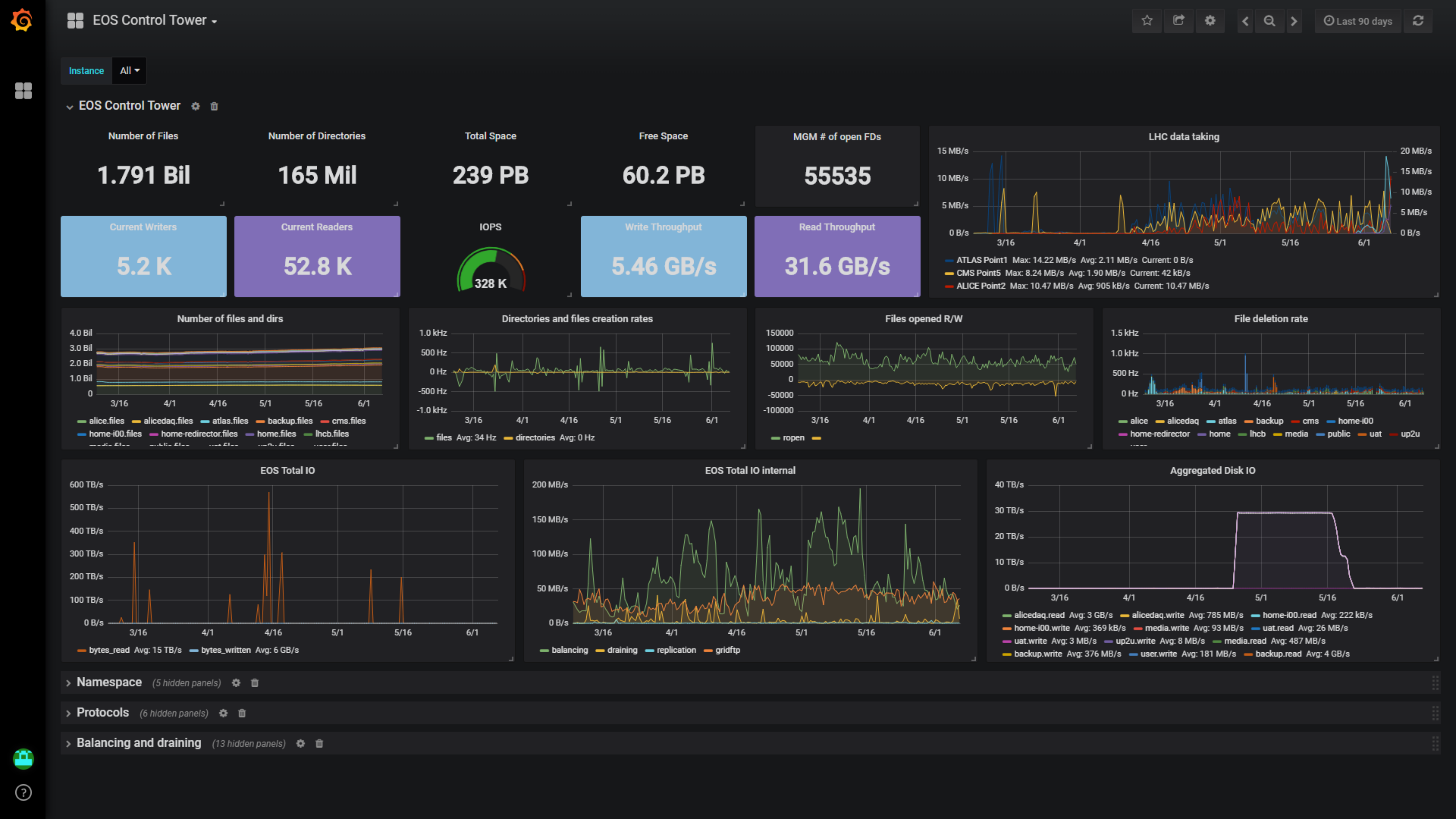
FuseX detailed deployment plan

- May 2018
  - FuseX deployed under /eos/scratch on lxplus + lxbatch
  - Validations and tests form ST, CM, CF, CDA and several other CERN power users
- June 2018
  - Scale test on /eos/scratch: minimum support for 2000 simultaneous clients
- July -August 2018
  - Enable FuseX on EOSHOMExx and on EOSLHCB.
  - FuseX will become the default access software for migrated eos users and for everything under /eos/lhcb
- September - End of 2018
  - Following the LHCB upgrade, transparent move of all other instances

Conclusion

- IT-ST very busy in instances split, namespace and FuseX deployments
- Orders of magnitude improvements expected in problem isolation, service stability, and future service scalability (example: EOSMEDIA)
- We will be addressing all use cases with specific solutions
  - The vast majority of end-user case are or will be covered with EOSHOME and CERNBOX
  - Other requirements (build clusters, databases, object stores, HPC, Filers, …) will continue to require specific solutions that are already addressed using Ceph and will and will be even be better addressed using CephFS

Reserve slides



EOS Control Tower

Number of Files: 1.791 Bil

Number of Directories: 165 Mil

Total Space: 239 PB

Free Space: 60.2 PB

MGM # of open FDs: 55535

Current Writers: 5.2 K

Current Readers: 52.8 K

IOPS: 328 K

Write Throughput: 5.46 GB/s

Read Throughput: 31.6 GB/s

LHC data taking

- ATLAS Point1 Max: 14.22 MB/s Avg: 2.11 MB/s Current: 0 B/s
- CMS Point5 Max: 8.24 MB/s Avg: 1.90 MB/s Current: 42 kB/s
- ALICE Point2 Max: 10.47 MB/s Avg: 905 kB/s Current: 10.47 MB/s

Number of files and dirs

Directories and files creation rates

- files Avg: 34 Hz
- directories Avg: 0 Hz

Files opened R/W

File deletion rate

EOS Total IO

- bytes_read Avg: 15 TB/s
- bytes_written Avg: 6 GB/s

EOS Total IO internal

Aggregated Disk IO

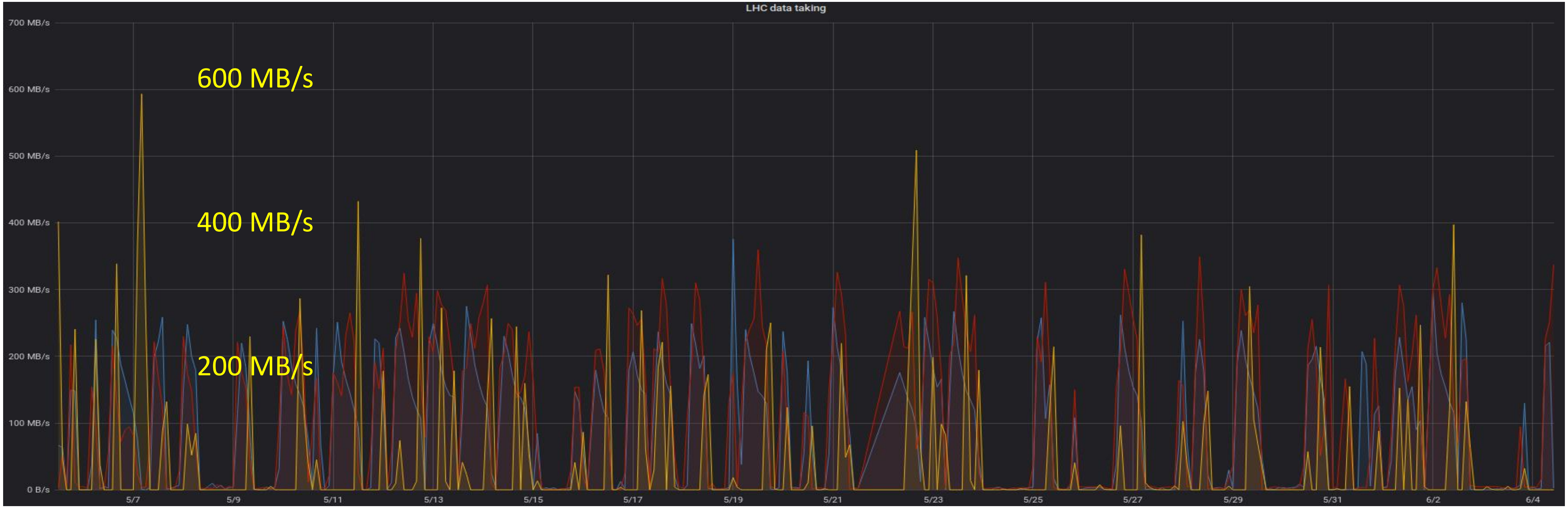
- alicedaq.read Avg: 3 GB/s
- alicedaq.write Avg: 785 MB/s
- home-i00.read Avg: 222 kB/s
- home-i00.write Avg: 369 kB/s
- media.write Avg: 93 MB/s
- uat.read Avg: 26 MB/s
- uat.write Avg: 3 MB/s
- up2u.write Avg: 8 MB/s
- media.read Avg: 487 MB/s
- backup.write Avg: 376 MB/s
- user.write Avg: 181 MB/s
- backup.read Avg: 4 GB/s

> Namespace (5 hidden panels)

> Protocols (6 hidden panels)

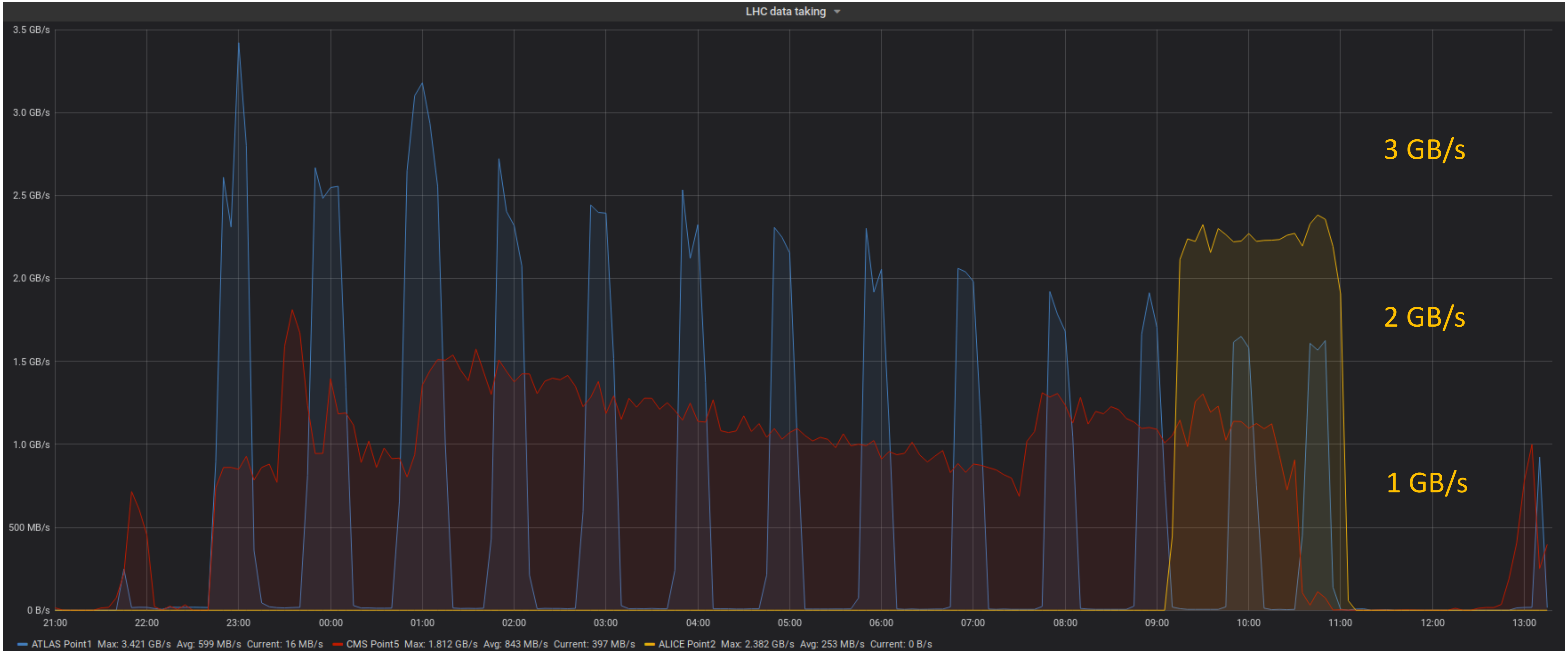
> Balancing and draining (13 hidden panels)

Last 30 days (LHC data acquisition only)



ALICE, ATLAS and CMS write data to EOS directly from the pit. EOS is their source for Repro, Export and Archive

A recent LHC fill

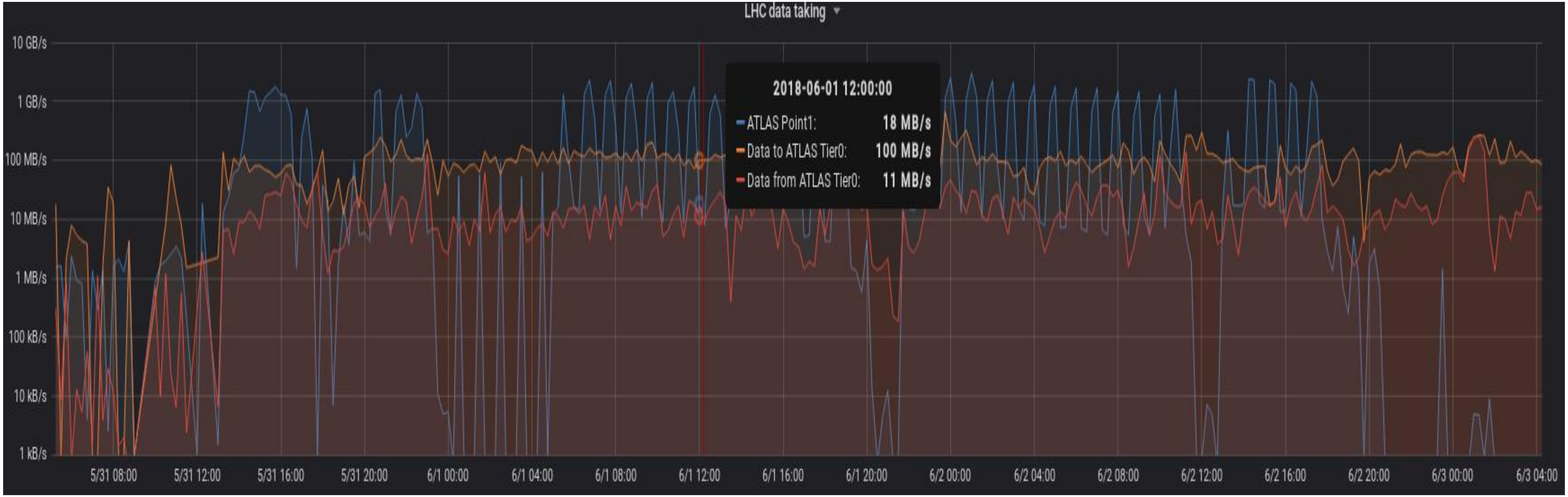


3 GB/s

2 GB/s

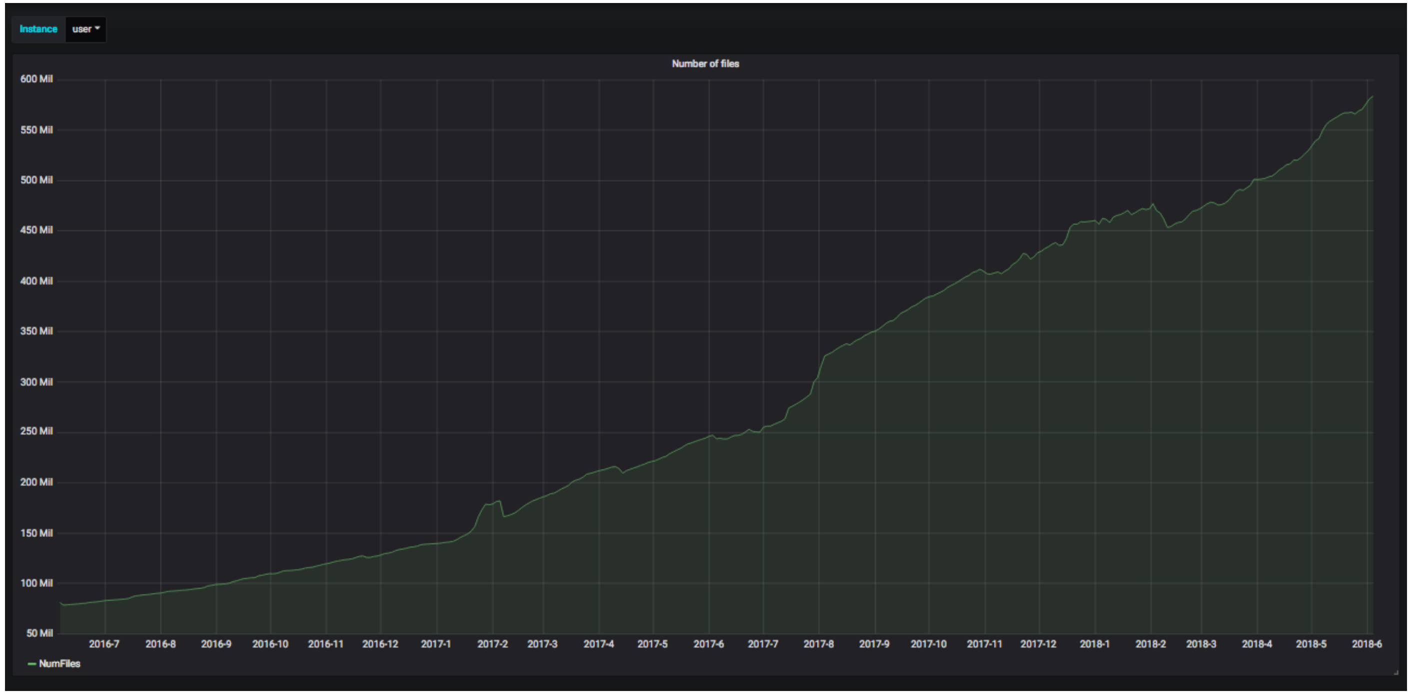
1 GB/s

ATLAS analysis examples



Eosuser 2018 evolution

- Increase of the number of files: ~ x 2.4
  - Now at 584M files
- Increase of disk space: ~ x 2.4
  - Now at 3.27PB
- Increase of the number of users: ~ x 1.4
  - Now at 14300 accounts


Instance
user ▾
Number of files
600 Mil
550 Mil
500 Mil
450 Mil
400 Mil
350 Mil
300 Mil
250 Mil
200 Mil
150 Mil
100 Mil
50 Mil
NumFiles