





EOS across ~ 1000 km (~ 620 mi)

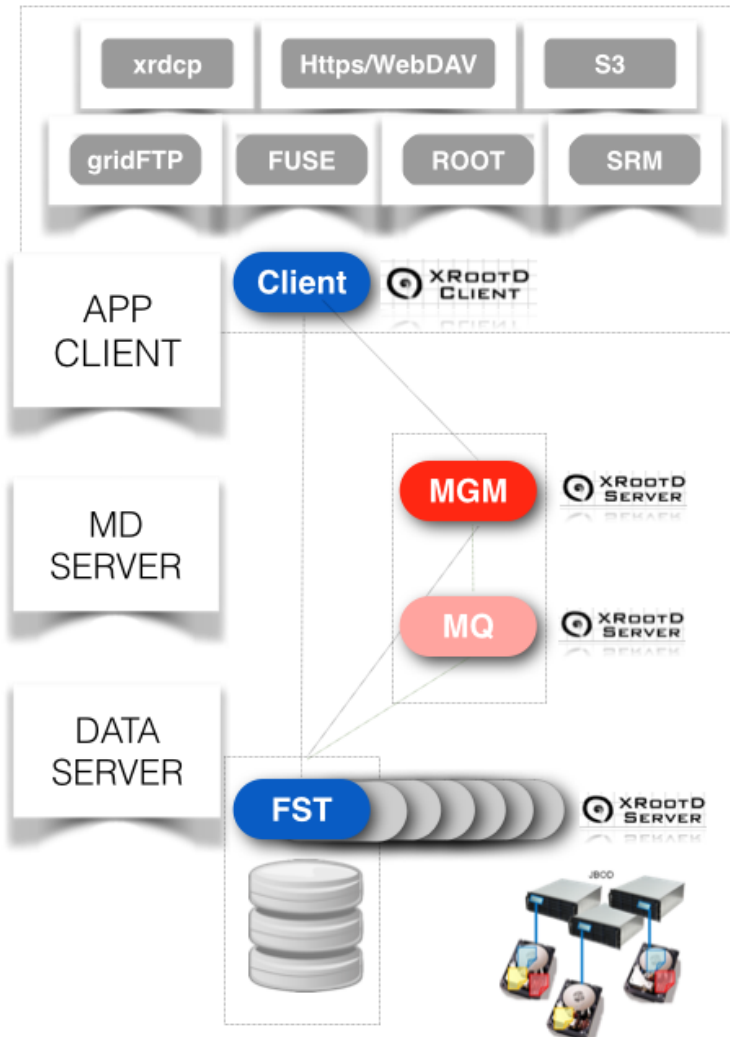
**Luca Mascetti
CERN/IT-DSS**

Outline

- EOS Architecture & New Features
- EOS Evolution
- Wigner Computer Centre
- EOS Deployment
- Geo Scheduling
- EOS Infrastructure Awareness
- Other Functionalities
- Federations
- Conclusion



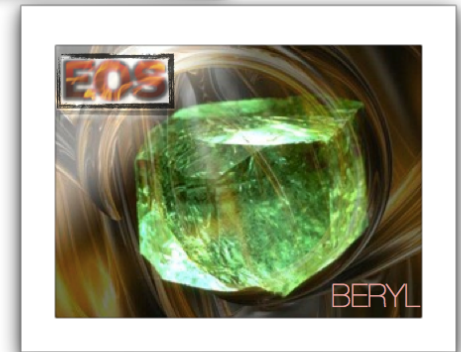
Architecture



- Project started in April 2010
- Production since May 2011
- Simple and scalable solution
- Simple to operate
- In-memory namespace
- Strong authorization
- Quotas
- Network RAID (RAIN)
- Tunable QoS
- Dev&Ops in CERN/IT-DSS

EOS Improvements & New Features

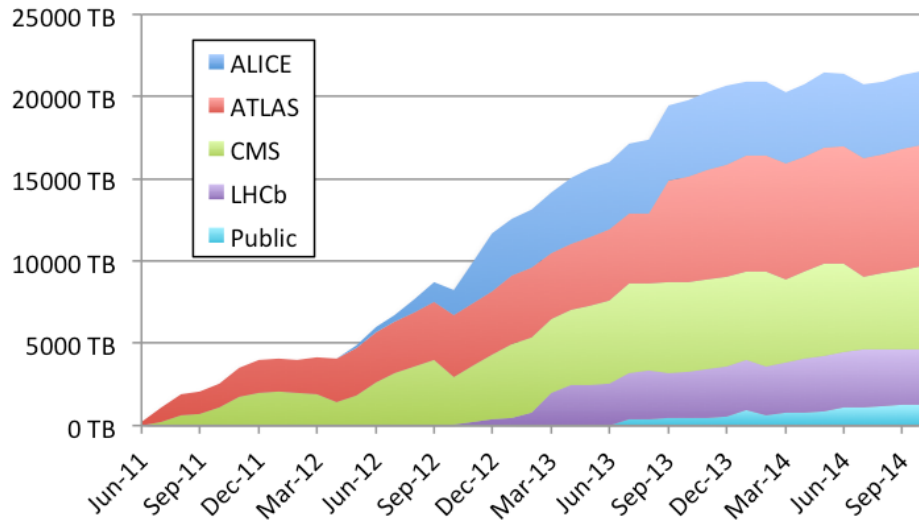
- Improved
 - multi-user FUSE client
- New
 - master/slave(ro) instant failover
 - recycle bin + new ACLs
 - vector reads
 - multiple RAIN layouts
 - geo balancer
- Coming Soon
 - archive functionality
 - IPv6 compliant (XRRootD4)
 - infrastructure awareness
 - authentication delegation



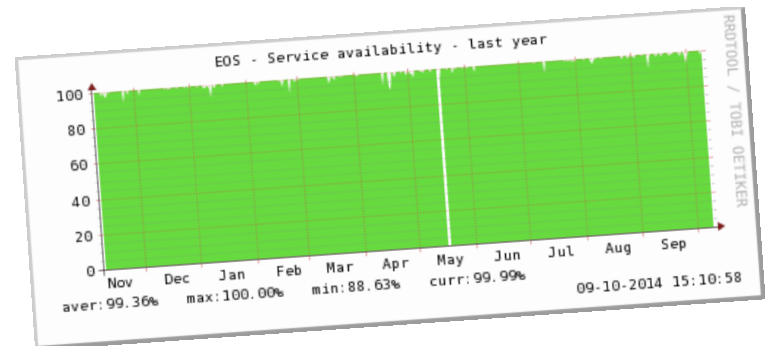
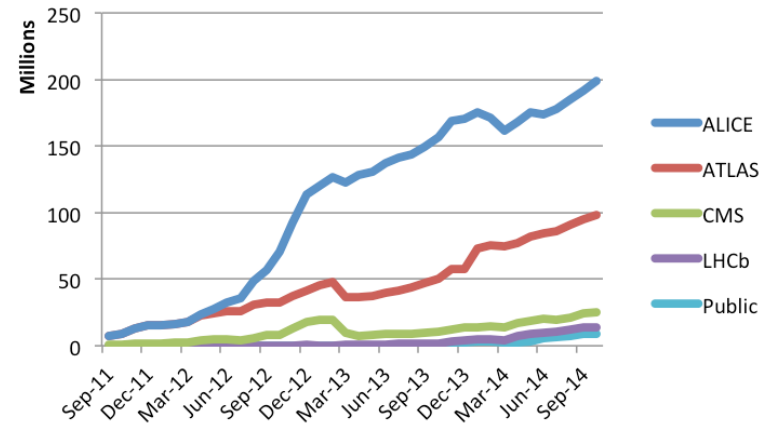
EOS Evolution

21.5 PB and 200 M files used by the experiments in 5 EOS instances

Space Used on EOS (logical)



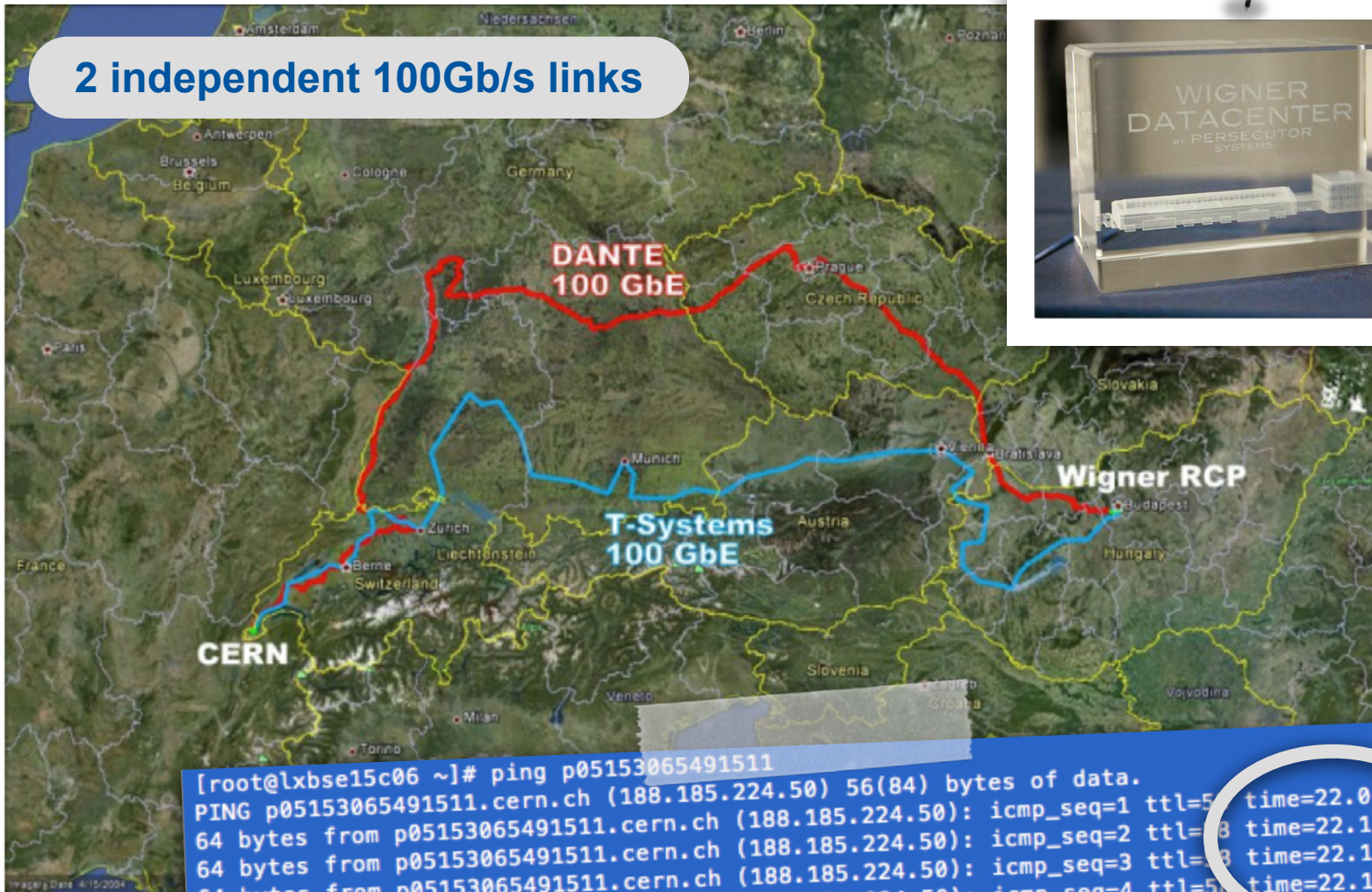
Files Stored on EOS



~99.4% availability for our "picky" SLS

Wigner Computer Centre

2 independent 100Gb/s links

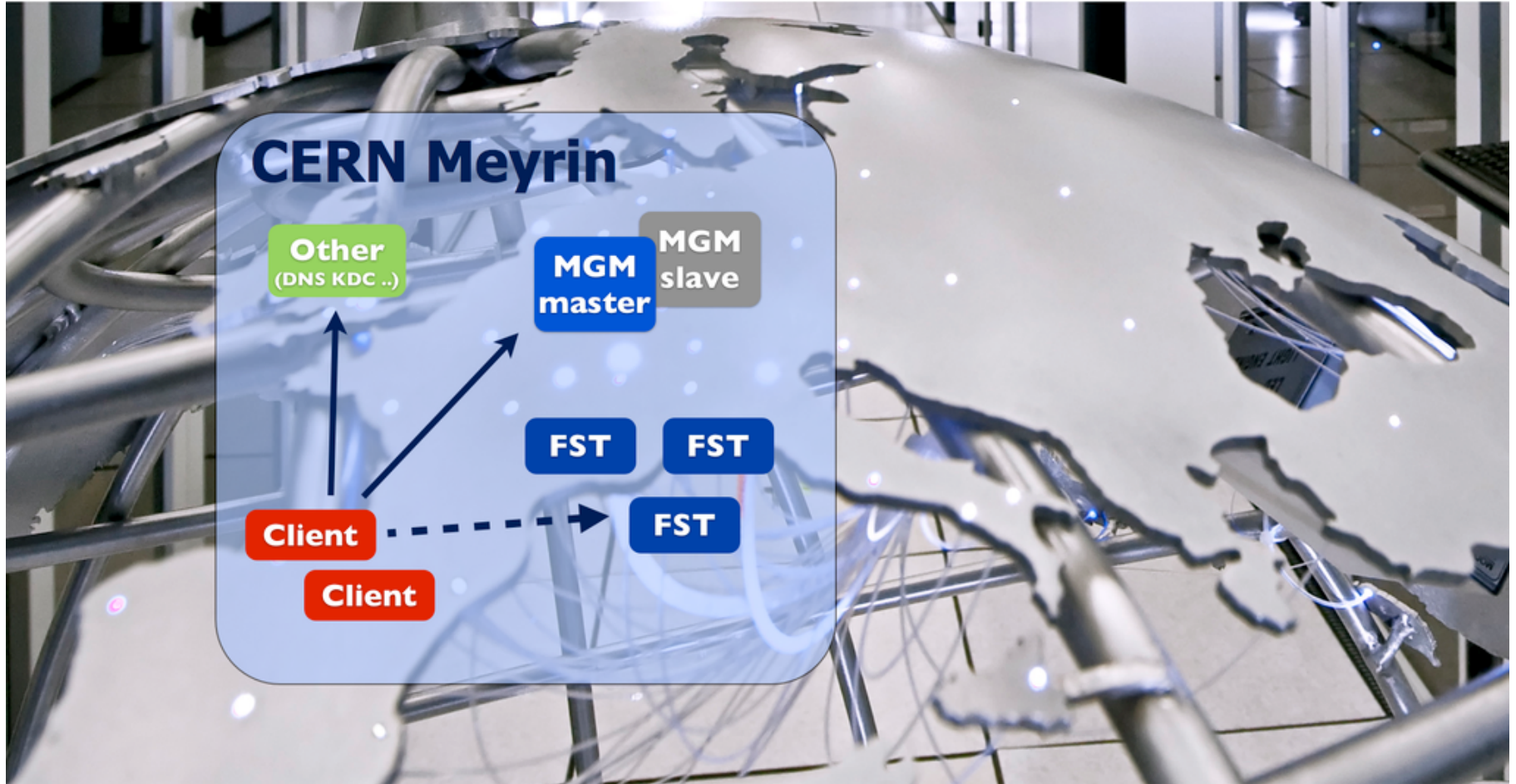


```
[root@lxbse15c06 ~]# ping p05153065491511
PING p05153065491511.cern.ch (188.185.224.50) 56(84) bytes of data.
64 bytes from p05153065491511.cern.ch (188.185.224.50): icmp_seq=1 ttl=58 time=22.0 ms
64 bytes from p05153065491511.cern.ch (188.185.224.50): icmp_seq=2 ttl=58 time=22.1 ms
64 bytes from p05153065491511.cern.ch (188.185.224.50): icmp_seq=3 ttl=58 time=22.1 ms
64 bytes from p05153065491511.cern.ch (188.185.224.50): icmp_seq=4 ttl=58 time=22.1 ms
```

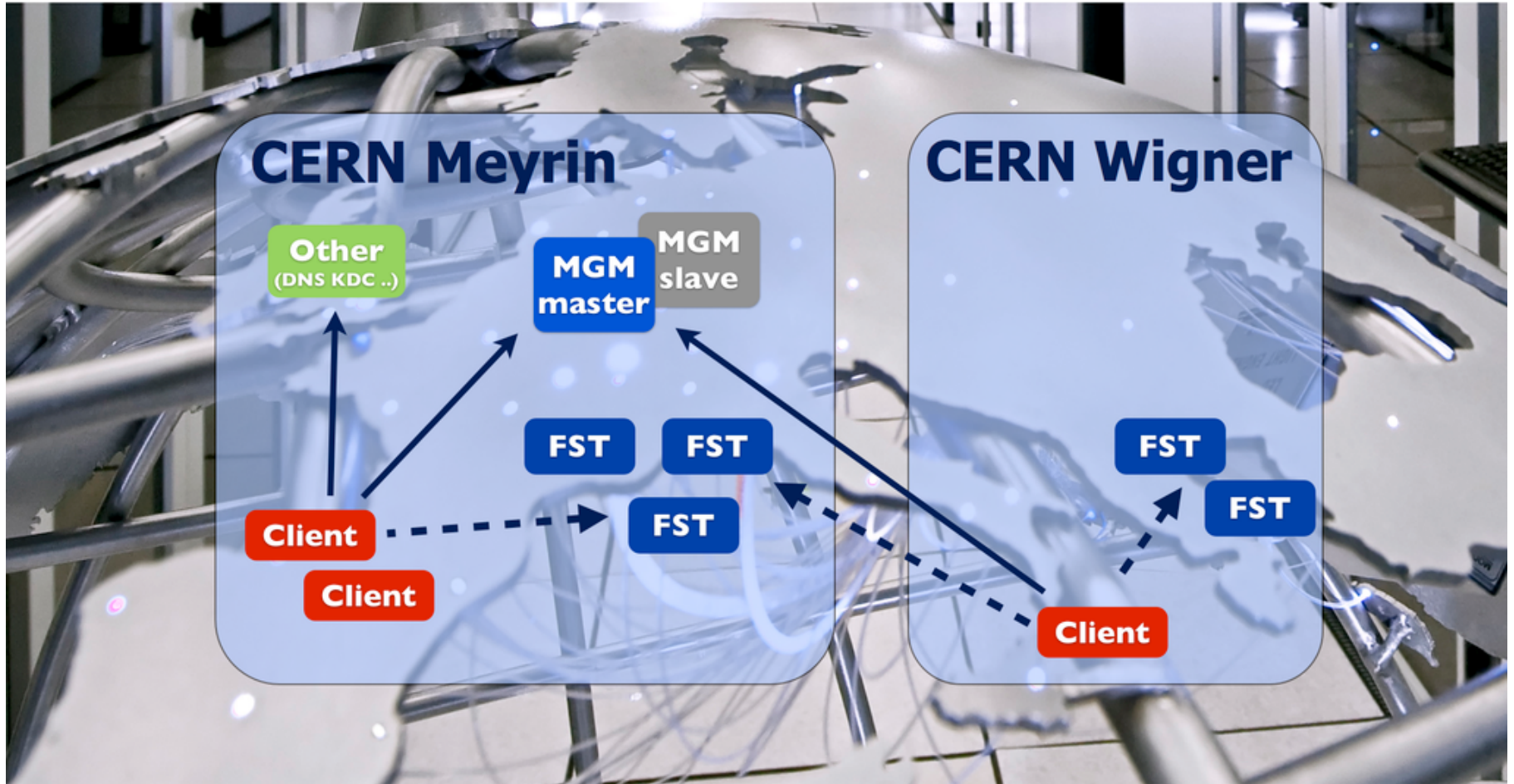
22ms latency 7



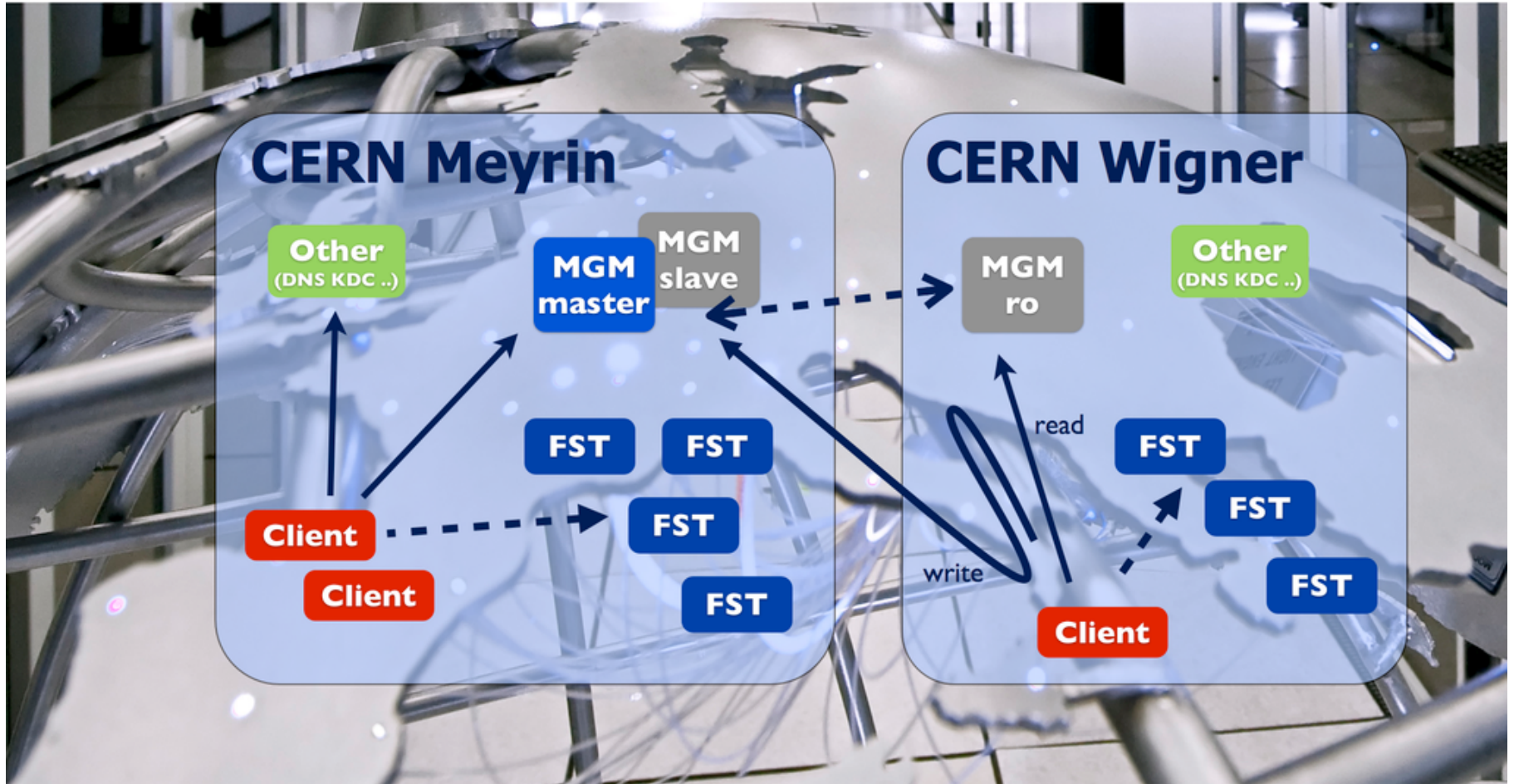
EOS 2013 Deployment



EOS 2014 Deployment

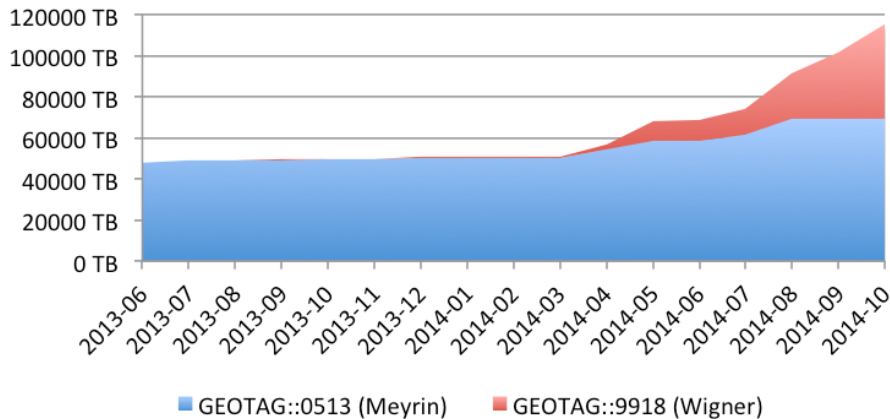


EOS 2015 Deployment



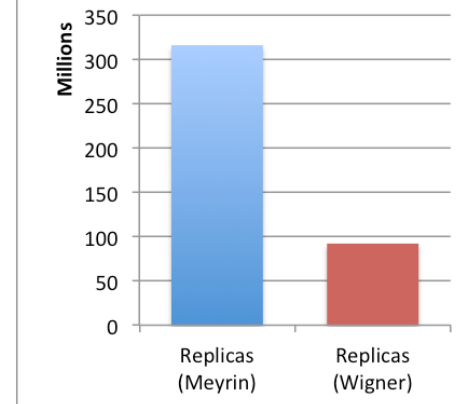
Installed Capacity and Replicas Distribution

EOS Installed Raw Capacity

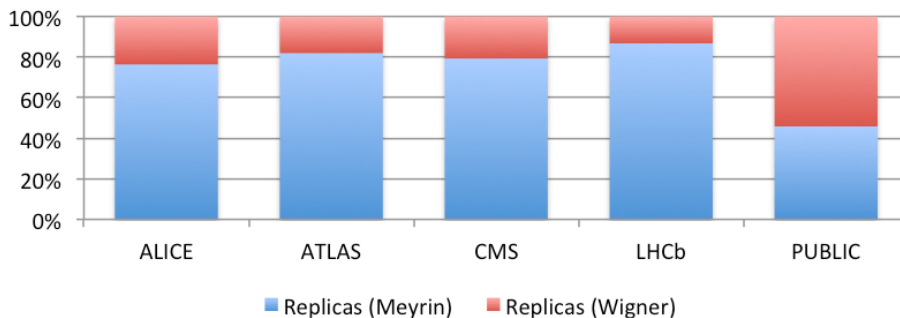


Depending on the next hardware delivery we expect to reach ~50% of storage capacity or stabilizing around the 60% vs. 40% ratio

EOS Replicas Distribution



Replicas Distribution in % per Experiment



Distribution of replicas between the two centres is currently ~ 3:1

Not (yet) all the installed capacity is fully available to the experiments for using it (preparation for 2015)

Puppet Contribute

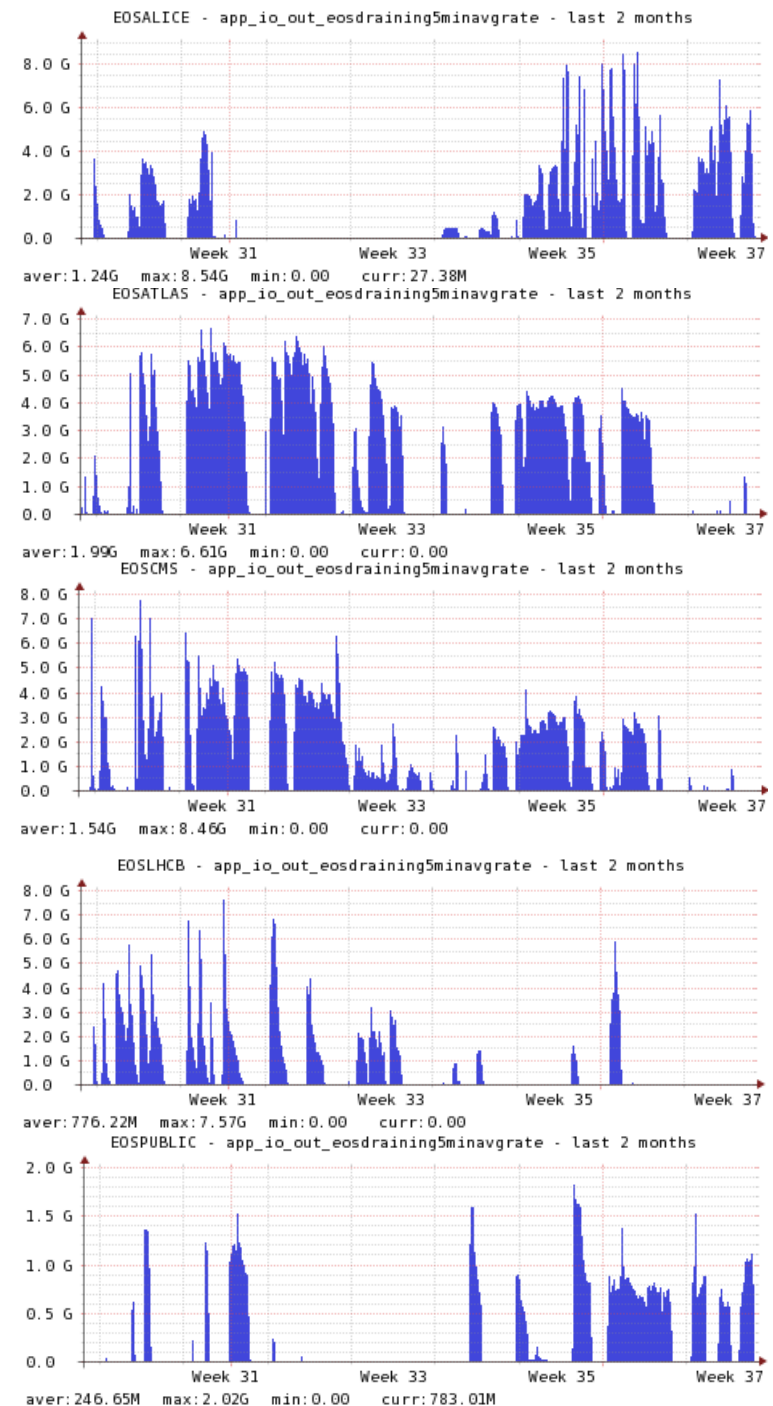
The puppetization of EOS (migration from Quattor to Puppet) helped in moving part of the replicas from Meyrin to Wigner

Disksevers were drained/replicated and then reinstalled from scratch with the new configuration system

The draining/replication activity was done in a controlled way on top of standard user analysis and production activity

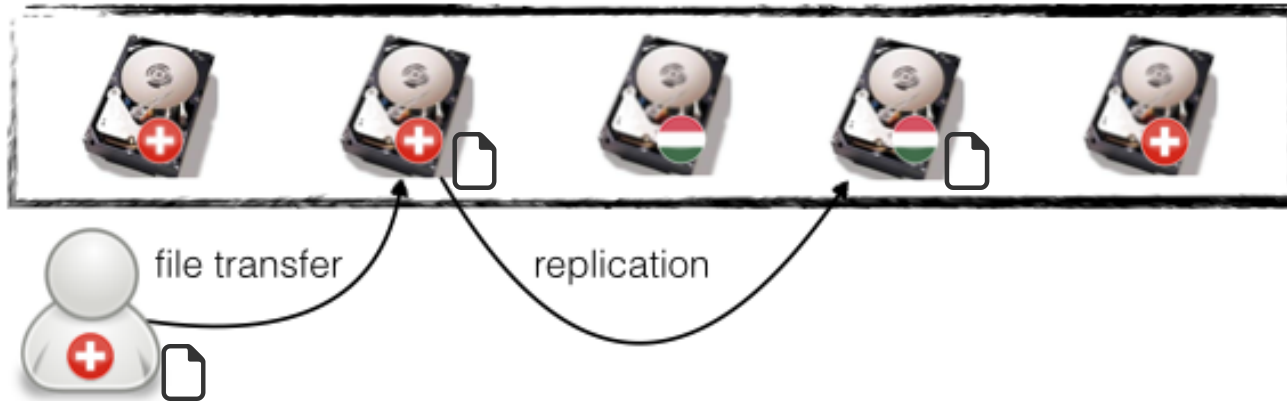
From the plots during Jul/Aug/Sep is possible to see the performance of the EOS draining system with peaks of 8GB/s per instance

Overall during all July and August we moved in this way around 34PB of data



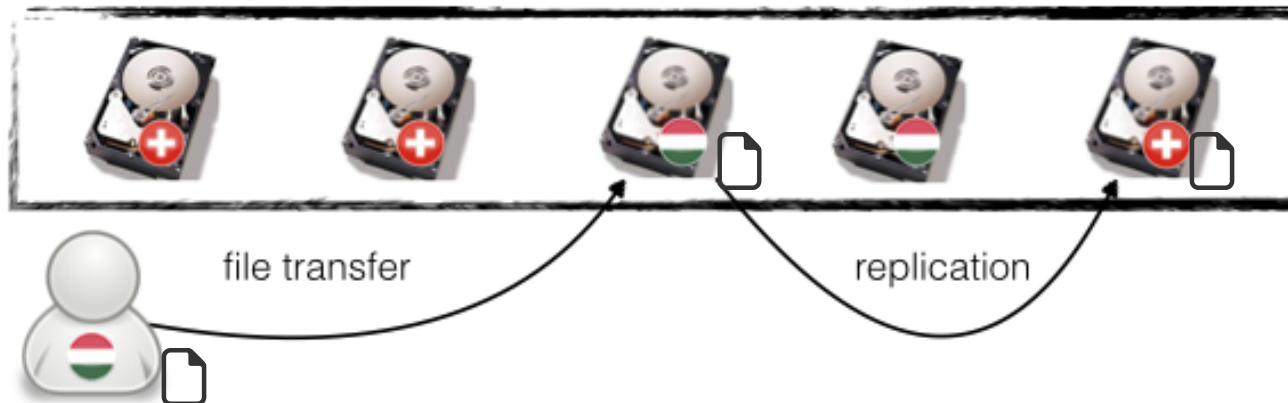
Geo Scheduling (write)

Scheduling Group



By construction EOS place replicas in two different disk servers, moreover if there are no space constraints EOS by default try to place a replica in each of the two sites

Scheduling Group



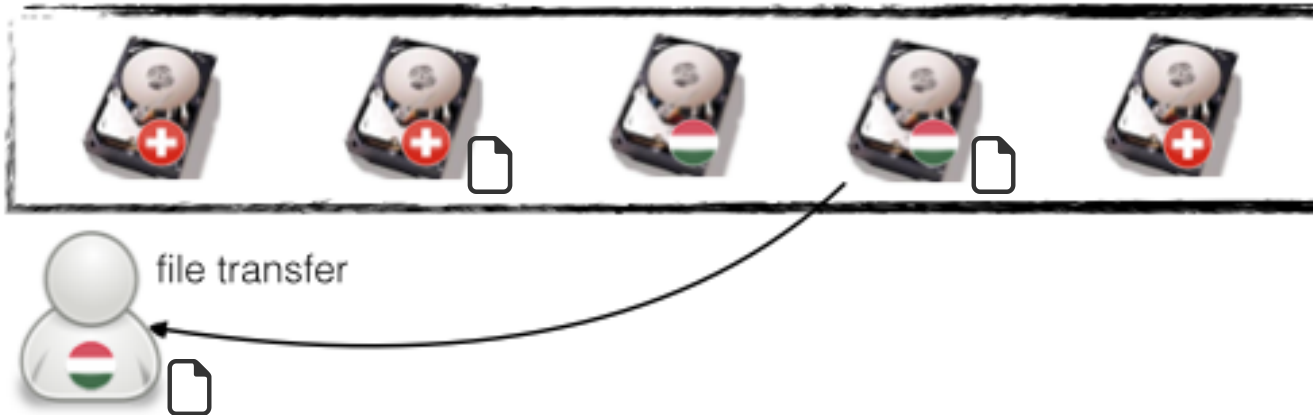
Geo Scheduling (read)

Scheduling Group



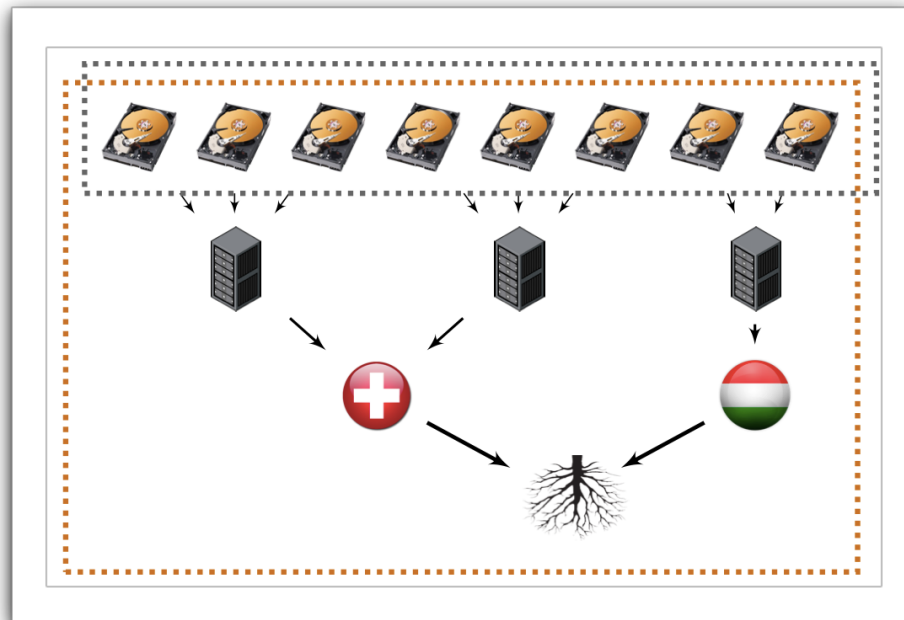
For a read operation EOS by default will select the "closest" replica (if available) with a 95% probability, since the algorithm take into consideration also the load and the availability of the disk servers

Scheduling Group



EOS Infrastructure Awareness

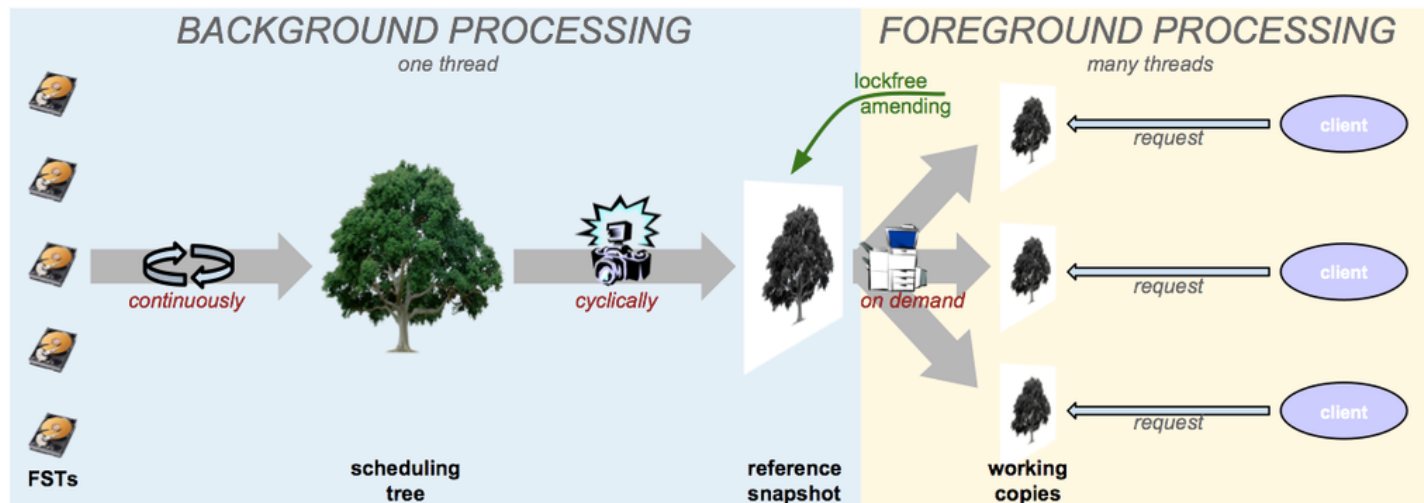
- Flat scheduling (current version)
 - stateless
 - work only with 2 locations
 - single scheduling policies available
 - possible CPOF
- Tree scheduling (next version)
 - stateful (more complex)
 - tree efficient data structures
 - more scheduling policies available
 - avoid CPOF





EOS Infrastructure Awareness

Implementation:

- periodical update of instance state
- estimate delay compensation
- using tree for easy updating and snapshots
- no mutex contention
- low scheduling latency $O(0.1\text{ms})$



Other functionalities (Http/WebDAV)



[drop a file here to upload into the current directory]

refresh my collaborative PAD my ID: URL: Invite to my PAD

Quota Whoami Who Upload Mkdir Rmdir Info Space Nodes Groups Filesystems

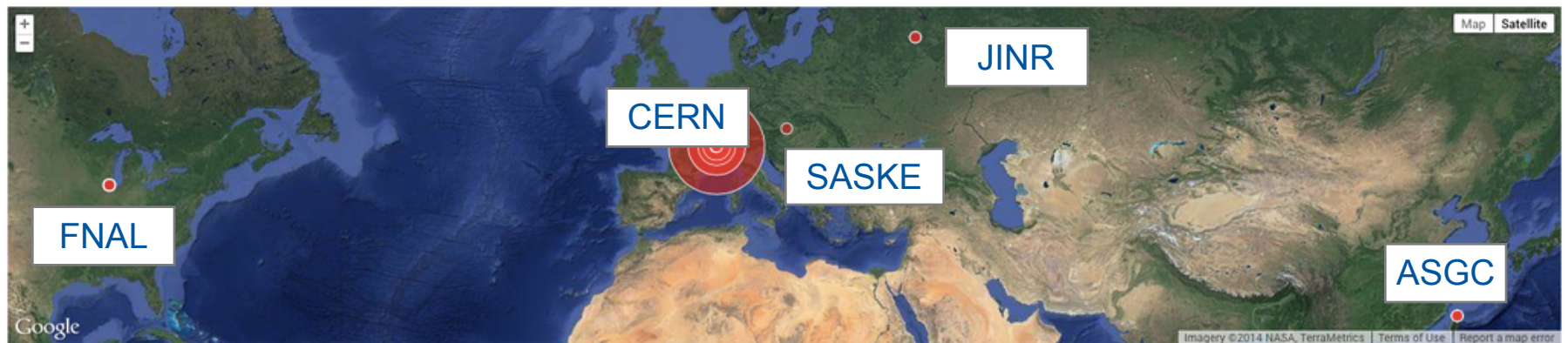
quota	gld	space	usedbytes	usedlogicalbytes	usedfiles	maxbytes	maxlogicalbytes	maxfiles	percentageusedbytes	statusbytes	statusfiles
node	project	/eos/lhcb/	0	0	0	0	0	0	100.00	ignored	ignored

nobody@eospps]: /eos/lhcb

Path	Size	Created	Mode	owner	group	AcI
./		Nov 05 2013 10:08	drwxrwsr-+	Imascett	c3	
../		Apr 26 2011 11:03	drwxrwsr-+	root	root	
grid/		Dec 02 2013 18:09	drwxrwsr-+	Imascett	z5	
opstest/		Nov 05 2013 11:23	drwxrwsr-+	Imascett	c3	
proc/		Nov 05 2013 11:23	drwxrwsr-+	Imascett	c3	
test/		Nov 05 2013 11:23	drwxrwsr-+	Imascett	c3	
testCfg/		Nov 05 2013 11:23	drwxrwsr-+	Imascett	c3	

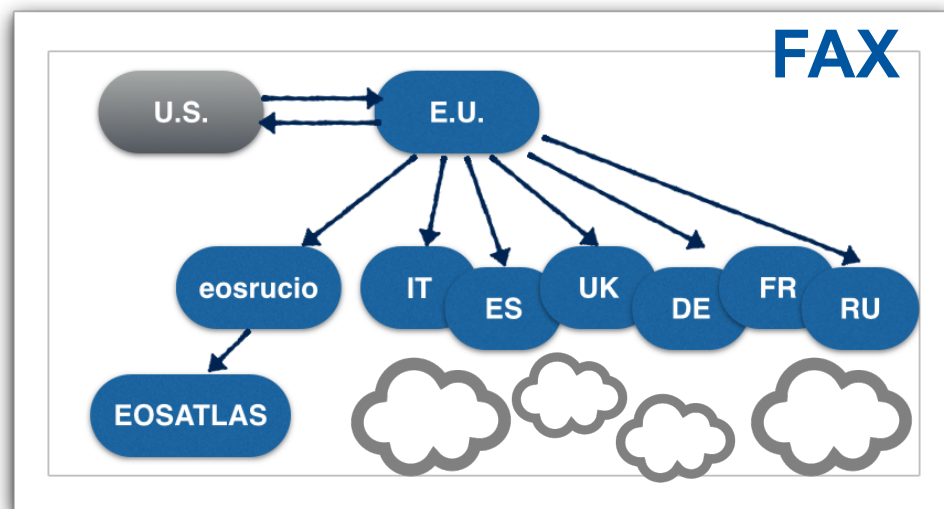
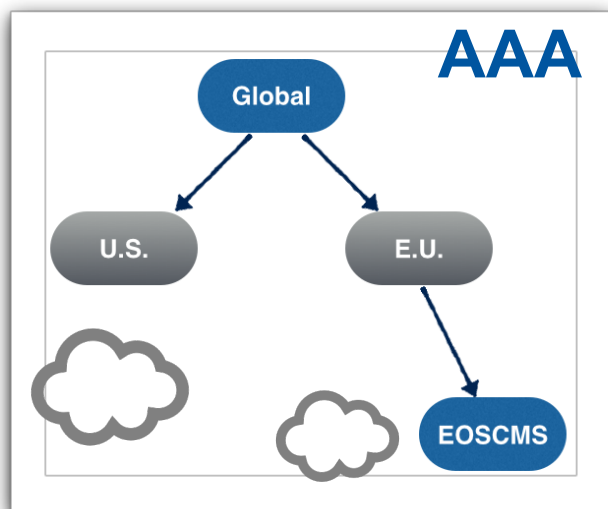
Other functionalities (VST)

- Virtual Storage Interface
 - overview of EOS instances
 - display generic parameters
 - configuration
 - utilization
 - in the future could be used to implement Storage Clouds
 - Virtual Storage Cloud (VSC)
 - Introducing higher level of redirection



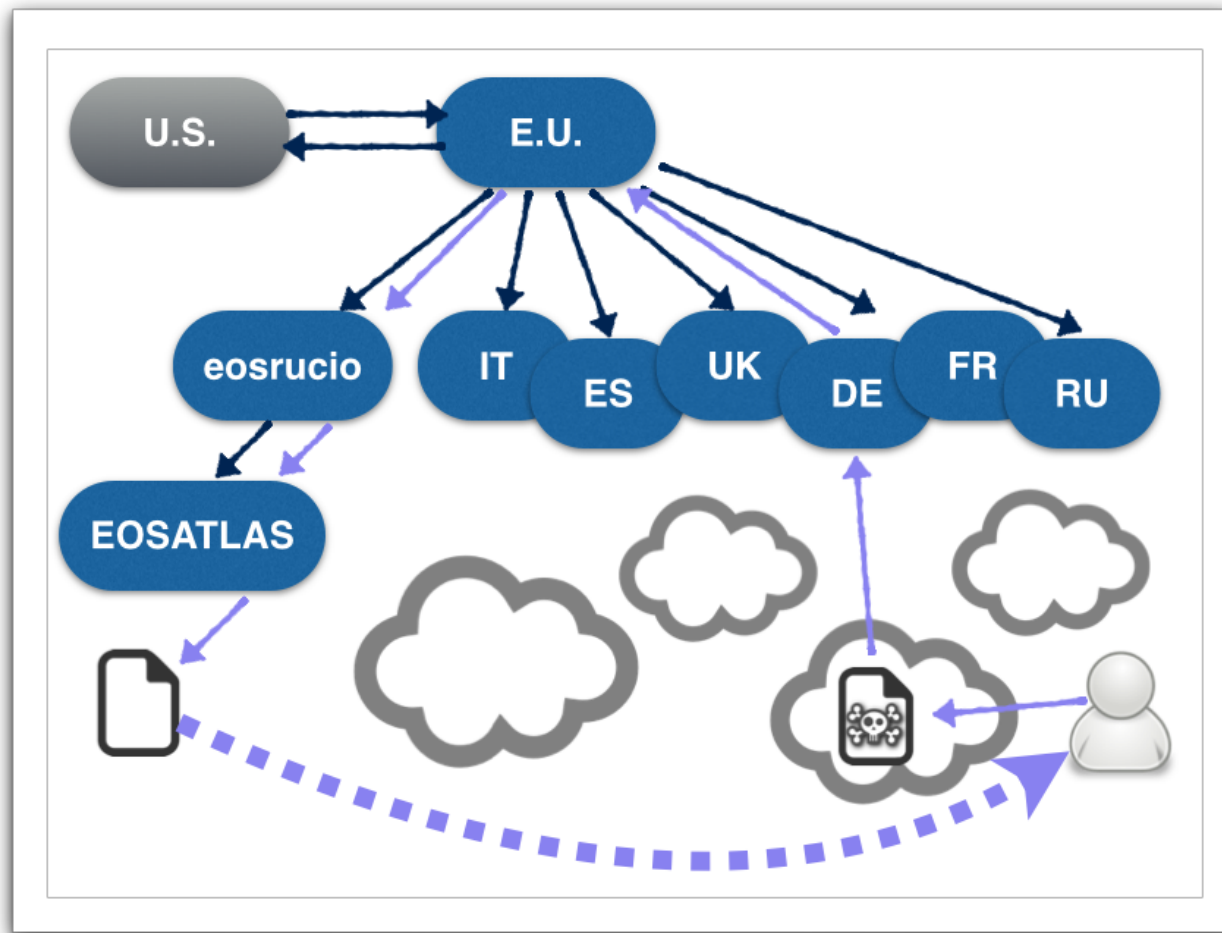
XRootD Federation

- AAA for CMS and FAX for ATLAS
- CERN/IT-DSS operates several redirectors
- cope with multiple sites and a "unified" namespace



XRootD Federation

how it works:



ALICE Federation

ALICE already use a federation model

Storage federated via AliEn

Data files are accessed remotely

EOS take part to this federation with EOSALICE at CERN

Other EOS sites are part of it

Table: Disk storage elements

Summary statistics from the table:

Total	22.02 PB	16 PB	6.084 PB	376,439,034	35.94 PB	25.11 PB	10.83 PB
--------------	-----------------	--------------	-----------------	--------------------	-----------------	-----------------	-----------------

Overlaid logos for EOS sites: CERN, MEPHI, ROC_K1_T1, Subatech, UNAM_T1.



Summary

- EOS usage at CERN continue growing
 - new functionalities available
- EOS currently cope well with 2 sites
 - when possible it try to be smart
 - smooth operation
- next scheduling version even better
 - full infrastructure aware
- looking forward to business continuity
 - completely feasible only with CC size comparable
- New challenges ahead
 - EOS will play a more central role during Run2
 - CERNBox integration



www.cern.ch

Questions ?