Distributed cross-site storage with single EOS end-point

D.Krasnopevtsev, A.Kiryanov,
A.Klimentov, A.Petrosyan,
E.Ryabinkin, A. Zarochentsev

Motivation

1. There's an increasing demand in storage resources for LHC experiments, especially in view of HL LHC
2. Operation of a large Tier-2 site is quite complicated and requires unique expertise of attending personnel
3. Smaller Tier-3 sites may help, but it's not easy to centrally coordinate their activity

Our solution to aforementioned problems is a federation of small sites that looks like a large site ("cloud") from outside

We evaluate federated storage as a first step towards this idea

Requirements for a federated storage

- Single entry point
- Universality: should be usable by at least four major LHC experiments
- Scalability: it should be easy to add new resources
- Data transfer optimality: transfers should be routed directly to the optimal disk servers avoiding intermediate gateways and other bottlenecks
- Fault tolerance: replication/redundancy of core components
- Built-in virtual namespace, no dependency on external catalogues

Finding possible solutions

We had to find a software solution that supports federation of distributed storage resources. This very much depends on a transfer protocol support for redirection. We have found three storage systems whose access protocols allow this:

- DynaFed is an HTTP-based federator developed by IT-ST group at CERN. This software is highly modular and only provides a federation frontend while the storage backend(s) have to be chosen separately. It would be interesting to try it out but we were looking for more all-in-one solution.
- EOS is an xroot-based solution. It is also developed at CERN (we knew where to ask for help), has characteristics closely matching our requirements, and is already used by all major LHC experiments. We decided to give it a try.
- dCache is a storage system developed at DESY which uses dCap protocol. Depending on the Persistency Model, dCache provides methods for exchanging data with backend (tertiary) Storage Systems as well as space management, pool attraction, dataset replication, hot spot determination and recovery from disk or node failures.

Software alternatives

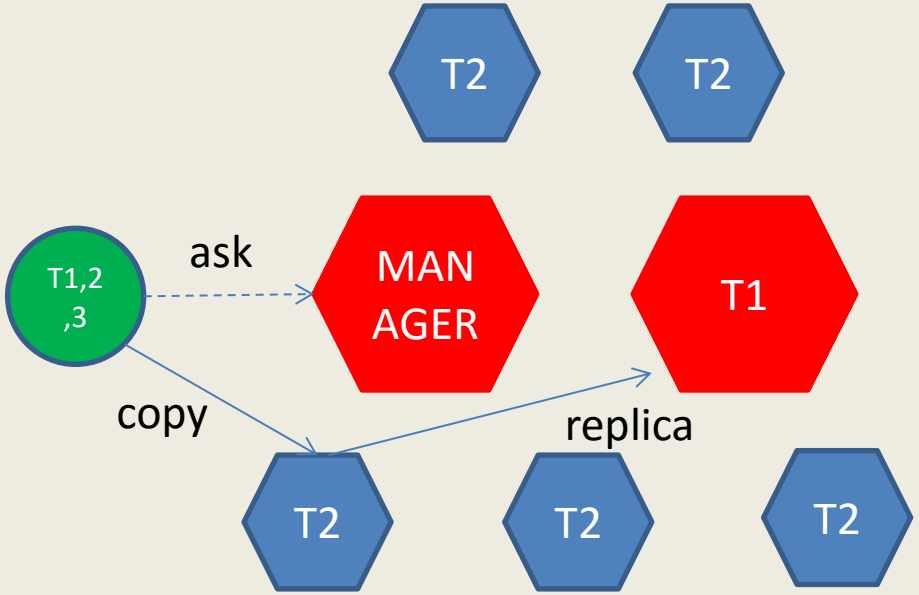
- EOS
  - Our current infrastructure and tests use EOS. We have already performed an extensive testing with this software, some results will be shown later on.
- dCache
  - We plan to deploy dCache on our infrastructure in the near future and repeat all tests we made with EOS.
- DynaFed
  - This software is currently under active development. We will give it a try when EOS and dCache testing will be finished.

Generic testbed structure

- T1 sites as end-points and main storage servers.
- T2 sites – secondary storage servers (for local replicas)
- T1,2,3 – clients .

Write process: client asks manager (end-point) about nearest SE, copies to nearest SE , copies replica from nearest SE to another SE on demand.

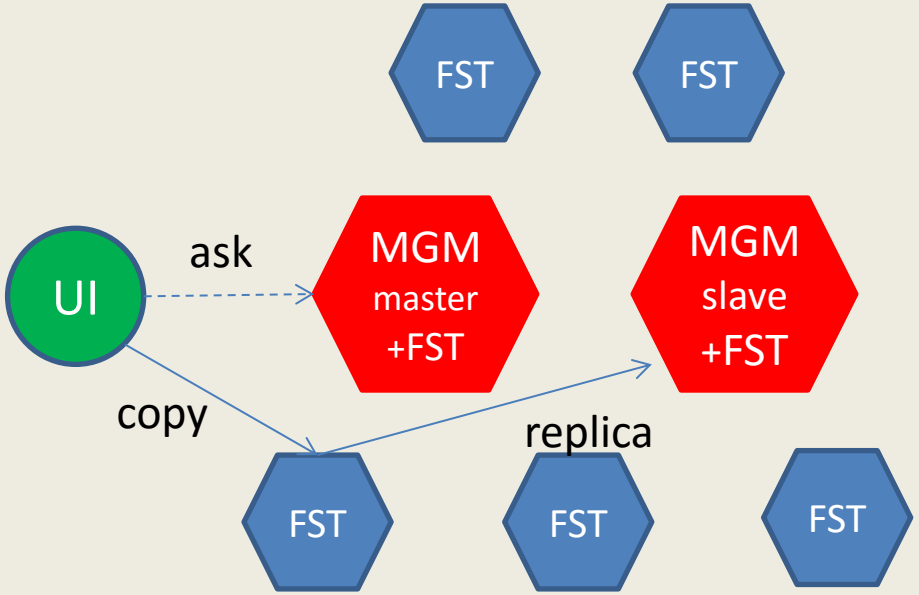
Read process: Client asks manager for a nearest SE with required file and copies that file locally.



Testbed structure for EOS

- T1 – MGM master and slave + FST servers.
- T2 sites – FST servers
- T1,2,3 – UI

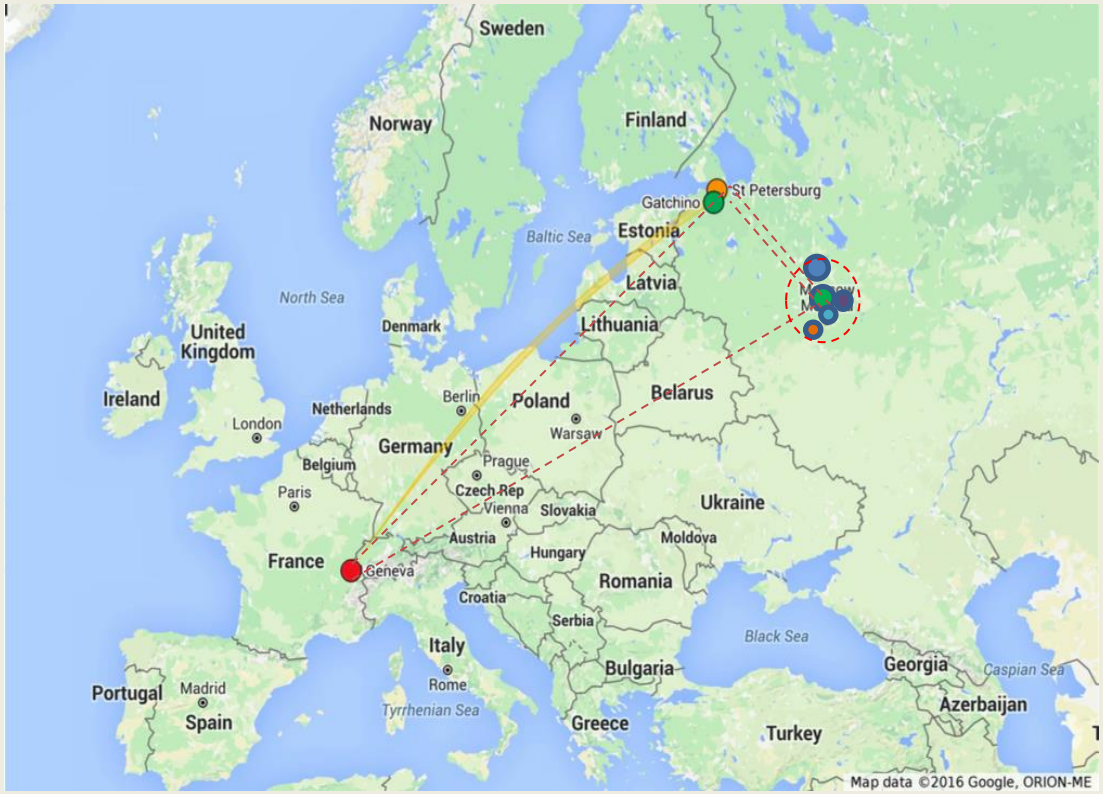
MGM as “Manager” on T1



Required tests

- **Proof-of-concept test:** install and configure distributed EOS, hook up GSI authentication, test basic functionality (file/directory create/delete, FUSE mount, access permissions)
  - **Reliability test:** MGM master-slave migration
  - **Performance tests:** file and metadata I/O, real-life experiment software, network
  - **Redirection impact test:** check if there's performance degradation with remote "head" node
  - **Data locality test:** evaluate EOS geo-tags role in data distribution

Participating sites



- CERN -MGM,UI
- KI – MGM, FST, UI (T1)
- JINR - MGM, FST, UI (T1)
- PNPI – MGM,FST,UI
- SPbSU – MGM,FST,UI
- SINP – FST, UI
- MEPHI - FST, UI
- ITEP - FST, UI

1st testbed - SPbSU ,PNPI ,CERN

2nd testbed - all

Tests

- **Proof-of-concept test**: ok on 2 testbeds: SPbSU+PNPI+CERN, KI + all.
- **Reliability test**: master slave migration support 2 MGMs only (1 master+ 1 slave), migration process is unstable.
- **Performance tests**: sinthetic test Bonnie+ and experiments tests – see later
- **Redirection impact test**: see “Reliability test” and “Performance tests”.
- **Data locality test**: see later (Geotag tests)

Geotag tests

Tests plan:

1)Write test: write N files, check by fileinfo
2)Read test: create folder with replica, write N files, read N files , check by logs.
3)Start bonnie tests without replicas
4)Start bonnie tests with replicas
5)Experiment test (with and without replicas)

Write check

```
~]$ eos attr ls eos/spbcloud/test/zar/tmp/rep
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="4k"
sys.forced.checksum="adler"
sys.forced.layout="replica"
sys.forced.nstripes="2"
sys.forced.space="default"
~]$ for i in {1..10}; do xrdcp /etc/passwd root://alice01.spbu.ru//eos/spbcloud/test/zar/tmp/rep/pass_petergoff.$i; done
[xrootd] Total 0.00 MB  |====================|  100.00 % [inf MB/s]
[xrootd] Total 0.00 MB  |====================|  100.00 % [inf MB/s]
……..
 ~]$ for i in {1..10}; do eos -b fileinfo eos/spbcloud/test/zar/tmp/rep/pass_petergoff.$i; done | grep default
  0    1      alice20.spbu.ru        default        /pool0   booted         rw     nodrain   online            petergoff
  1    3       eos.pnpi.nw.ru        default         /data   booted         rw     nodrain   online            gatchina
……
```

```
~]$ eos attr ls eos/spbcloud/test/zar/tmp/out0
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="4k"
sys.forced.checksum="adler"
sys.forced.layout="replica"
sys.forced.nstripes="1“
sys.forced.space="default"
~]$ for i in {1..10}; do xrdcp /etc/passwd root://alice01.spbu.ru//eos/spbcloud/test/zar/tmp/out0/pass_petergoff.$i; done
[xrootd] Total 0.00 MB  |====================|  100.00 % [inf MB/s]
[xrootd] Total 0.00 MB  |====================|  100.00 % [inf MB/s]
……..
~]$ for i in {1..10}; do eos -b fileinfo eos/spbcloud/test/zar/tmp/out0/pass_petergoff.$i; done | grep default
  0    3       eos.pnpi.nw.ru        default         /data   booted         rw     nodrain   online            gatchina
  0    1      alice20.spbu.ru        default        /pool0   booted         rw     nodrain   online            petergoff
  0    3       eos.pnpi.nw.ru        default         /data   booted         rw     nodrain   online            gatchina
……
```

Geotag test results

1) Write test: - ~~works for 2 and more replicas, does not work for 1 replica~~. - OK

space.geo.access.policy.write.exact=on

2) Read test – all ok (checked by logs and tcpdump)

3) ,4).. – ~~We're waiting for a fix for 1)~~ – next step

Performance tests

Network test

PerfSONAR suite was used for an independent network performance measurements. Two significant metrics that were measured between each pair of resource centers are bandwidth and latency. These metrics allow to understand the impact of network parameters and topology on performance test results.

Bonnie++ test

In order to test file I/O performance we used a synthetic Bonnie++ test, which is capable of measuring file and metadata access rate on a filesystem. EOS supports mounting as a virtual filesystem via Linux FUSE mechanism, which makes it possible for Bonnie++ to test both file read-write speeds (FST access) and metadata transaction rates (MGM access) independently.

Experiments tests (Atlas)

ATLAS test application performs a proton-proton event reconstruction with additional requirement on signal recovery from all Transition Radiation Tracker (TRT) detector drift tubes using so-called “raw” data as an input. Reconstruction of a signal from each of the proportional drift tubes in TRT is one of the most challenging and highly CPU-bound task, especially in high-luminosity conditions.

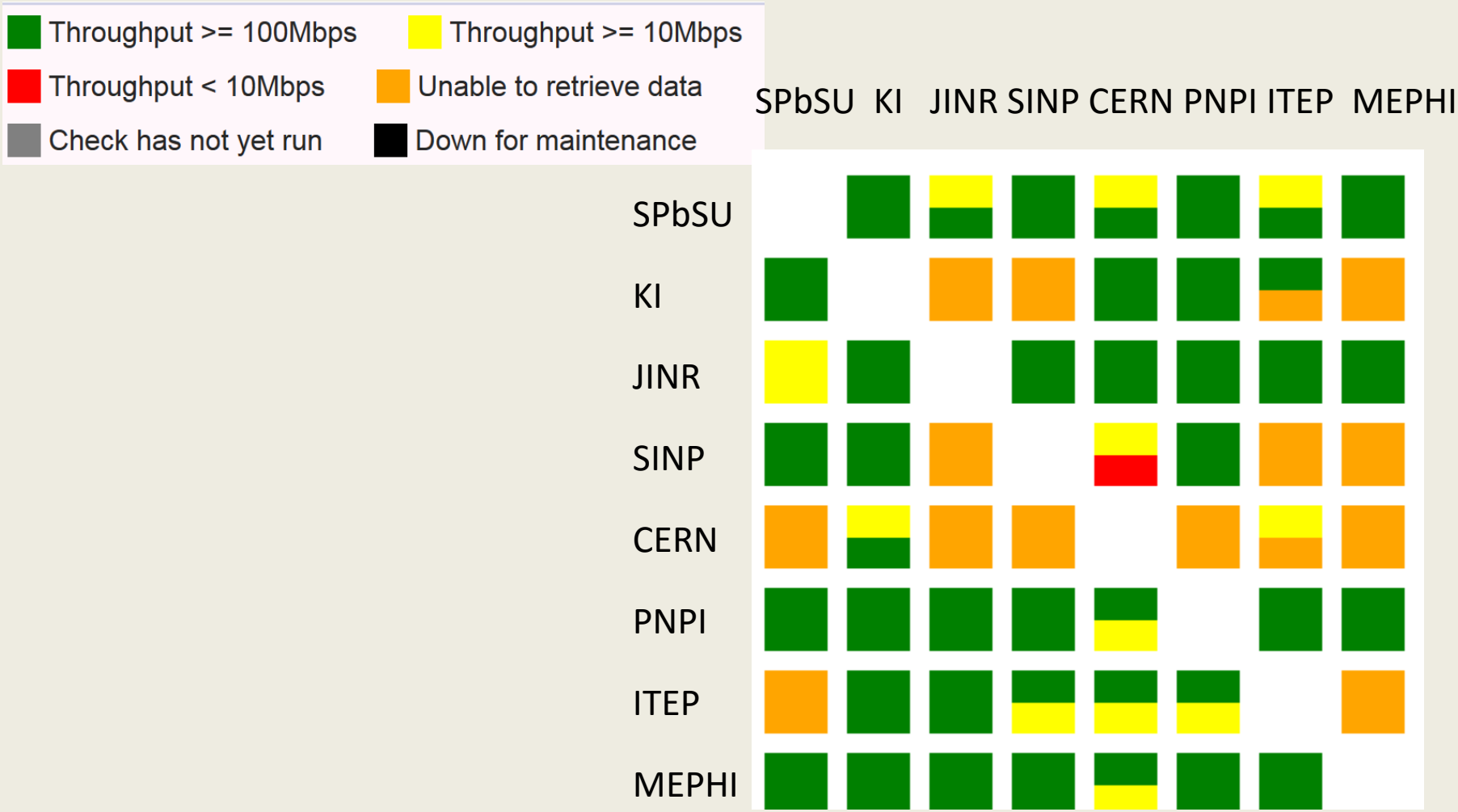
Input and output data, combined in a single file or a dataset, may be accessed both locally, on the filesystem, and remotely via xroot protocol.

Upon completion test application produces a comprehensive log file containing information about consumed computing resources on three key stages: environment initialization, event reconstruction event by event and finalization. In addition, during event processing, application records in the log file information about resources consumed during processing of each individual event.

Experiments tests (Alice)

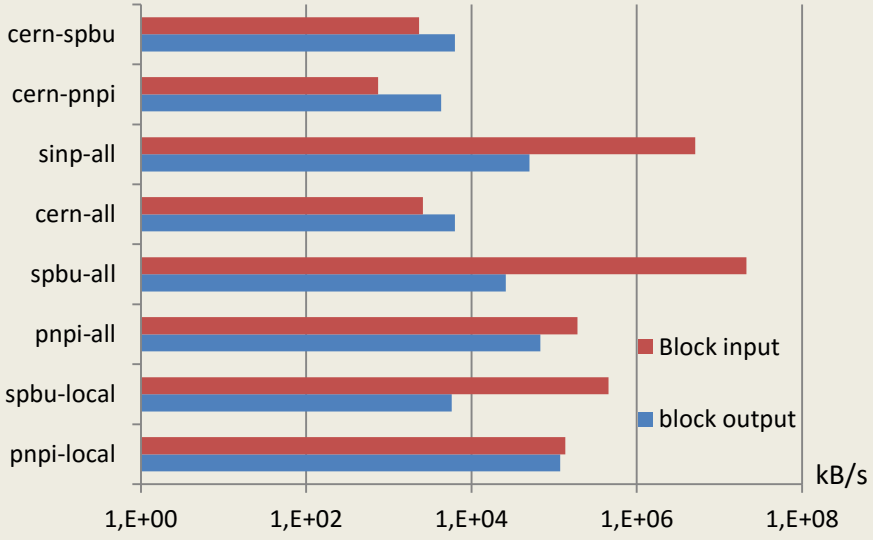
ALICE test application sequentially reads events from a file, analyzes them and produces information about the most "interesting" event according to the specified selection parameters. This application was specifically invented to evaluate the performance of the storage system. Just like before, input dataset can be accessed both locally and remotely via xroot protocol.

Network test (maddash webui)

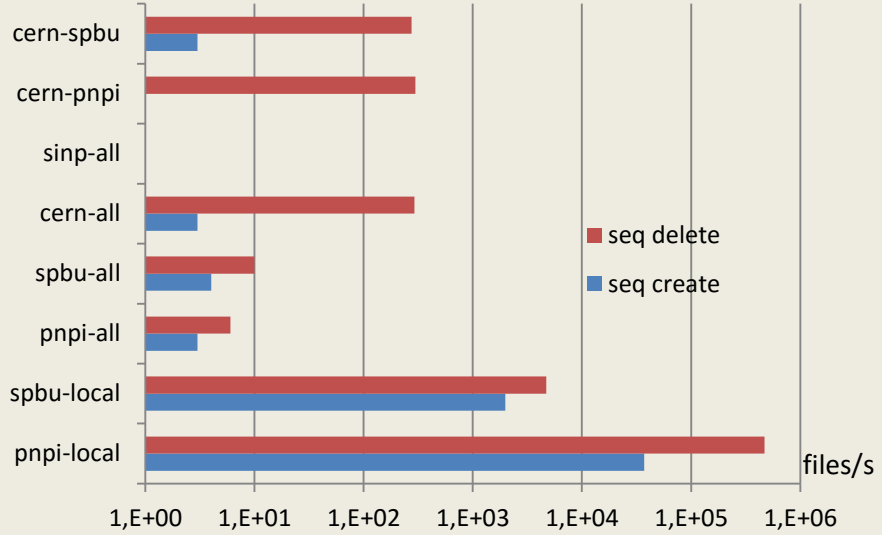


Bonnie test for 1st testbed – MGM at CERN, FST at SPbSU and PNPI

Data read-write



Metadata read-write



pnpi-local - local test on PNPI SE

spbu-local - local test on SPbSU SE

pnpi-all – UI-PNPI,MGM –CERN,SE -Federation

spbu-all – UI-SPbSU,MGM –CERN,SE –Federation

cern-all – UI-CERN,MGM –CERN,SE -Federation

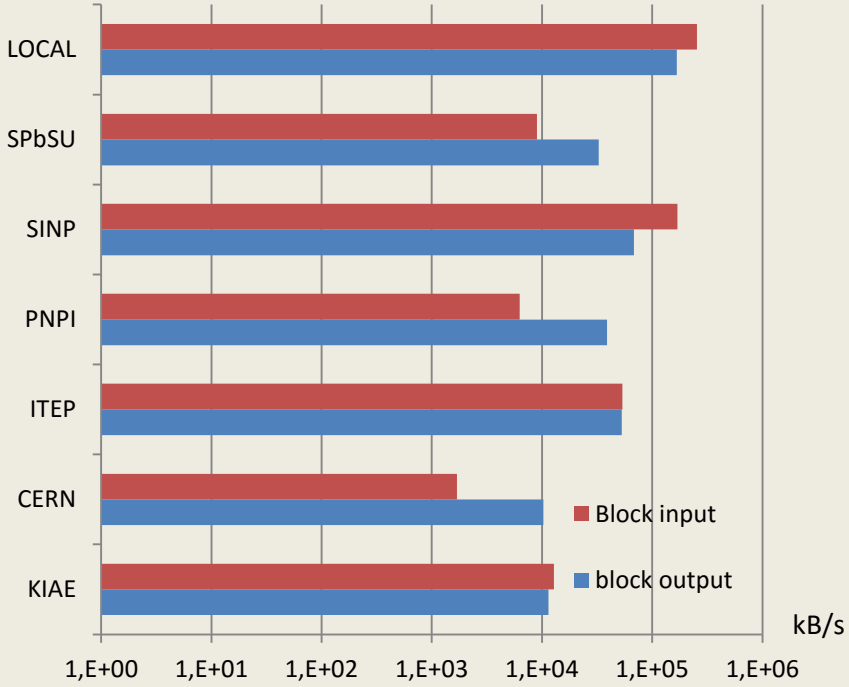
sinp-all – UI-SINP,MGM –CERN,SE -Federation

cern-pnpi – UI-CERN,MGM –CERN,SE -PNPI

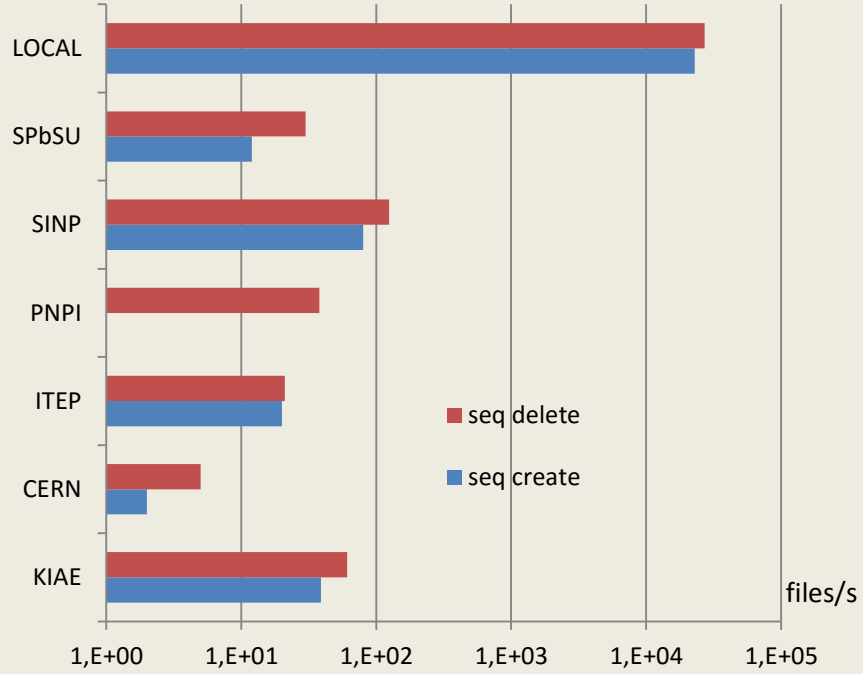
cern-spbu – UI-CERN,MGM –CERN,SE -SPbSU

Bonnie test for 2nd testbed – MGM and FST at KI

Data read-write



Metadata read-write



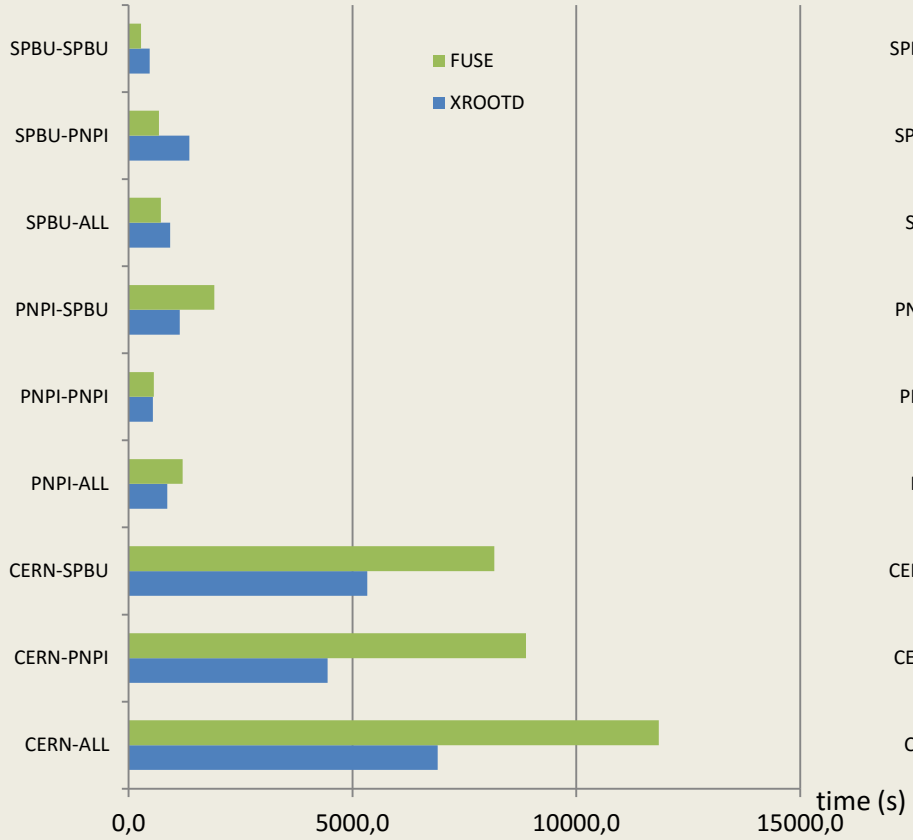
ALICE test

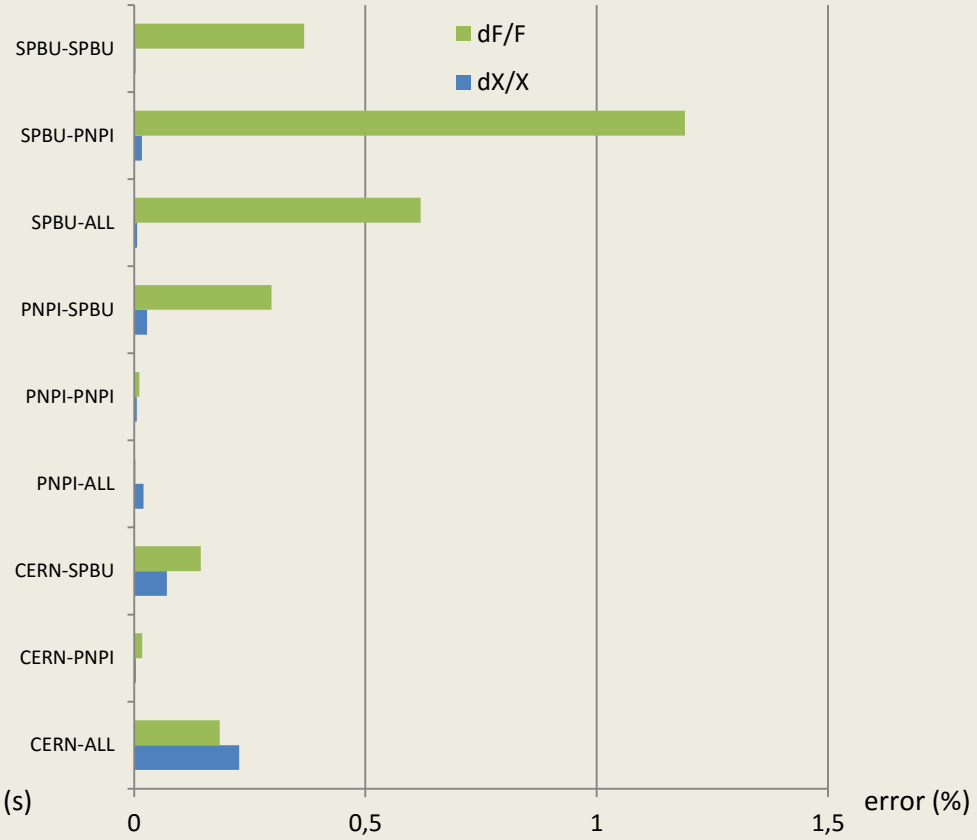
```
[zar@eostest scripts]$ cat ~/list_spbu_5_xrootd
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808029.18.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808030.11.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808030.18.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808031.13.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808032.13.root
[zar@eostest scripts]$ cat ~/list_pnpi_5_fuse
~/eos/spbcloud/test/alice_pnpi/11000167808029.18.root
~/eos/spbcloud/test/alice_pnpi/11000167808030.11.root
~/eos/spbcloud/test/alice_pnpi/11000167808030.18.root
~/eos/spbcloud/test/alice_pnpi/11000167808031.13.root
~/eos/spbcloud/test/alice_pnpi/11000167808032.13.root
[zar@eostest scripts]$ cat ~/list_10_xrootd
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808029.18.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_pnpi/11000167808029.18.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808030.11.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_pnpi/11000167808030.11.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808030.18.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_pnpi/11000167808030.18.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808031.13.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_pnpi/11000167808031.13.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_data/11000167808032.13.root
root://rueosfed.cern.ch//eos/spbcloud/test/alice_pnpi/11000167808032.13.root
```

Number of files = 10
Summary size = 12.68 GB

```
[zar@eostest scripts]$ cat alicefulltest.sh
#!/bin/bash
fileres=~/file.`date +%s`.alice.out
hostname >> $fileres

flist="../list_spbu_5_fuse ../list_spbu_5_xrootd
../list_pnpi_5_fuse  ../list_pnpi_5_xrootd
../list_10_xrootd"

time for i in {1..4}; do
      echo "********************  $i
************************";
      for l in $flist; do
            echo "================ $l
====================";
            date;
            echo ./alicetest.sh $l ;
            ls -la $l;
            ./alicetest.sh $l ;
      done
done | tee   $fileres

echo $fileres
```

Alice test out





spbu-spbu - UI-SPbSU,MGM –CERN,SE – SPbSU
spbu-pnpi - UI-SPbSU,MGM –CERN,SE - PNPI
spbu-all – UI-SPbSU,MGM –CERN,SE -Federation
pnpi-spbu – UI-PNPI,MGM –CERN,SE -SPbSU

pnpi-pnpi – UI-PNPI,MGM –CERN,SE –PNPI
pnpi-all– UI-PNPI,MGM –CERN,SE - Federation
cern-pnpi – UI-PNPI,MGM –CERN,SE -PNPI
cern-spbu – UI-PNPI,MGM –CERN,SE -SPbSU
cern-all – UI-CERN,MGM –CERN,SE -Federation

Alice out (in second)

CERN-ALL-F	CERN-ALL-X	CERN-PNPI-F	CERN-PNPI-X	CERN-SPBSU-F	CERN-SPBSU-X	PNPI-ALL-F	PNPI-ALL-X	PNPI-PNPI-F
16476,8	9468,0	4384,5	2066,7	5088,22	2828,39	1178,23	862,23	288,13
13093,9	5949,8	4376,5	2064,6	3770,94	3048,37	1178,28	860,89	281,46
10005,3	5930,5	4557,6	2045,6	3753,99	2512,4	1170,54	821,55	280,07
7790,2	6266,9	4441,6	2710,6	3729,95	2276,68	1297,72	920,98	282,89

PNPI-PNPI-X	PNPI-SPBSU-F	PNPI-SPBSU-X	SPBU-ALL-F	SPBU-ALL-X	SPBU-PNPI-F	SPBU-PNPI-X	SPBU-SPBU-F	SPBU-SPBU-X
274,14	1444,32	603,11	1 199,01	933,41	1019,38	679,95	109,65	234,25
271	783,86	556,1	1 257,80	943,81	109,86	709,85	109,55	234,48
270,59	823,69	580,81	219,78	948,61	109,62	688,38	223,92	235,95
271,77	783,54	540,18	219,78	892,76	115,11	632,46	110,02	234,39

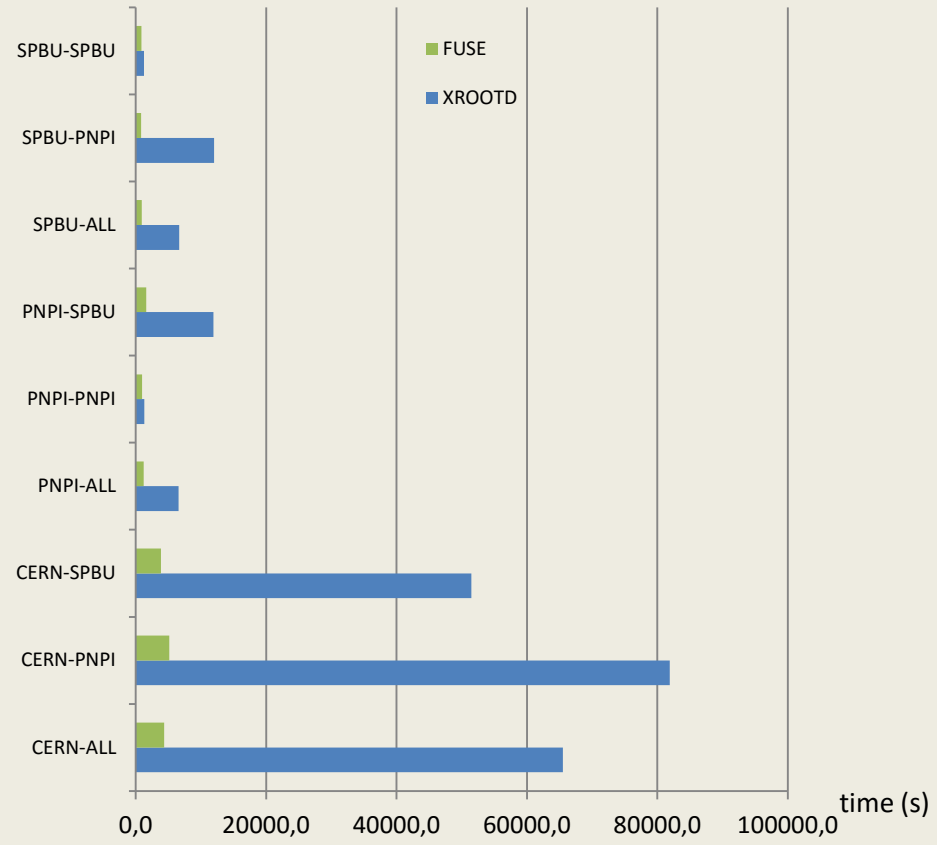
Atlas test

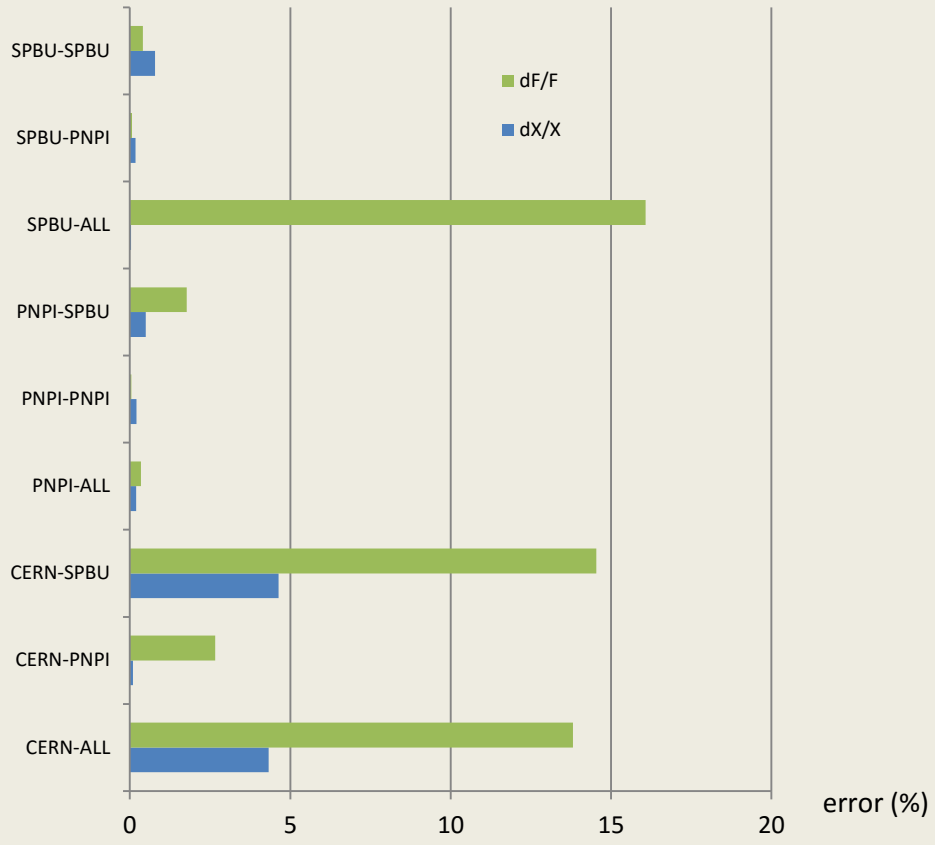
```
[zar@fedcloudui scripts]$ ls ~/list_atlas_
list_atlas_f_all   list_atlas_f_pnpi  list_atlas_f_spbu  list_atlas_x_all   list_atlas_x_pnpi  list_atlas_x_spbu
[zar@fedcloudui scripts]$ ls ~/list_atlas_f_pnpi
/home/zar/list_atlas_f_pnpi
[zar@fedcloudui scripts]$ cat ~/list_atlas_f_pnpi
/home/zar/eos/spbcloud/test/atlas_pnpi/NTUP_SMWZ.01281424._000061.root.1,/home/zar/eos/spbcloud/test/atlas_pnpi/NTUP_SMWZ.01281424._000021.root.1,
[zar@fedcloudui scripts]$ cat ~/list_atlas_x_all
root://rueosfed.cern.ch//eos/spbcloud/test/atlas_spbu/NTUP_SMWZ.01281424._000061.root.1,root://rueosfed.cern.ch//eos/spbcloud/test/atlas_spbu/NTUP_SMWZ.01281424._000021.root.1,root://rueosfed.cern.ch//eos/spbcloud/test/atlas_pnpi/NTUP_SMWZ.01281424._000061.root.1,root://rueosfed.cern.ch//eos/spbcloud/test/atlas_pnpi/NTUP_SMWZ.01281424._000021.root.1,
[zar@fedcloudui scripts]$ cat atlastesfull.sh
#!/bin/bash
fileres=~/file.`date +%s`.atlas.out
hostname >> $fileres

filelist="/home/zar/list_atlas_f_pnpi  /home/zar/list_atlas_f_spbu  /home/zar/list_atlas_x_pnpi  /home/zar/list_atlas_x_spbu /home/zar/list_atlas_x_all /home/zar/list_atlas_f_all"

cd ~/scripts/atlas_test2/

#source ~/eos/spbcloud/test/scripts/atlasinit_spbu.sh

for i in {1..4}; do
    echo "**********  $i **********";
    for file in $filelist; do
        ls -la $file
        echo root -b -q run.C\(\"$file\"\);
        (time -p root -b -q run.C\(\"$file\"\)) 2>&1
    done
done| tee   $fileres

echo $fileres
```

Number of files = 4

Summary size = 4.7 GB

Atlas test





spbu-spbu - UI-SPbSU,MGM –CERN,SE – SPbSU
spbu-pnpi - UI-SPbSU,MGM –CERN,SE - PNPI
spbu-all – UI-SPbSU,MGM –CERN,SE -Federation
pnpi-spbu – UI-PNPI,MGM –CERN,SE -SPbSU

pnpi-pnpi – UI-PNPI,MGM –CERN,SE –PNPI
pnpi-all– UI-PNPI,MGM –CERN,SE - Federation
cern-pnpi – UI-PNPI,MGM –CERN,SE -PNPI
cern-spbu – UI-PNPI,MGM –CERN,SE -SPbSU
cern-all – UI-CERN,MGM –CERN,SE -Federation

Atlas out (in second)

PNPI-ALL-F	PNPI-ALL-X	PNPI-PNPI-F	PNPI-PNPI-X	PNPI-SPBSU-F	PNPI-SPBSU-X
1224,43	6559,78	489,64	662,58	733,91	5889,06
1225,1	6581,38	489,98	665,57	743,65	5912,59
1234,48	6597,55	490,37	665,92	770,74	5967,66
1283,78	6612	504,58	674,09	954,85	6082,16

SPBU-ALL-F	SPBU-ALL-X	SPBU-PNPI-F	SPBU-PNPI-X	SPBU-SPBU-F	SPBU-SPBU-X
1174,15	6637,99	413,99	5997,79	418,4	624,31
839,26	6642,73	414,5	6019,03	421,38	626,32
840,33	6645,44	418,78	6024,67	423,22	636,49
843,03	6661,21	419,37	6036,71	442,74	637,16

Plans

- Expand 2nd testbed by adding slave MGM from JINR and FSTs from SINP and MEPhI.
- Submit experimental tests on 2nd testbed with geotags with 1 and 2 replicas.
- Exploit a dCache-based federation on the same resources with the same tests.
- Run the same tests on existing WLCG resources (with the consent from the experiments).

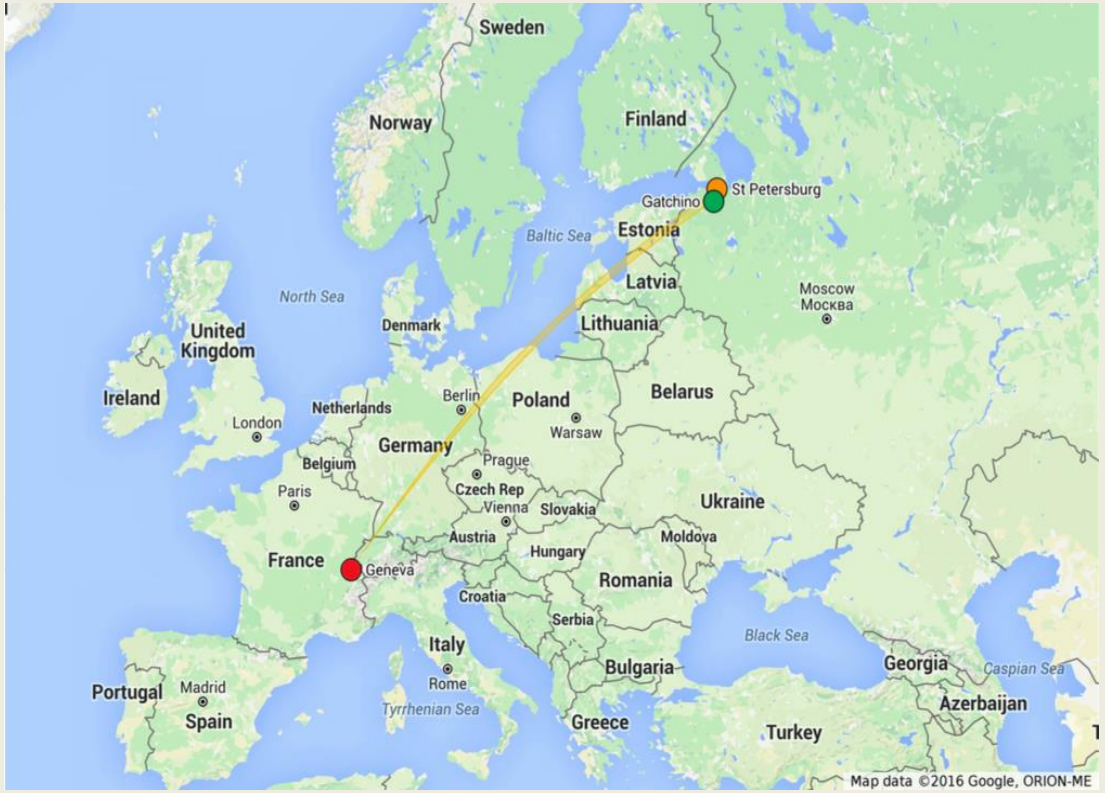
Conclusions

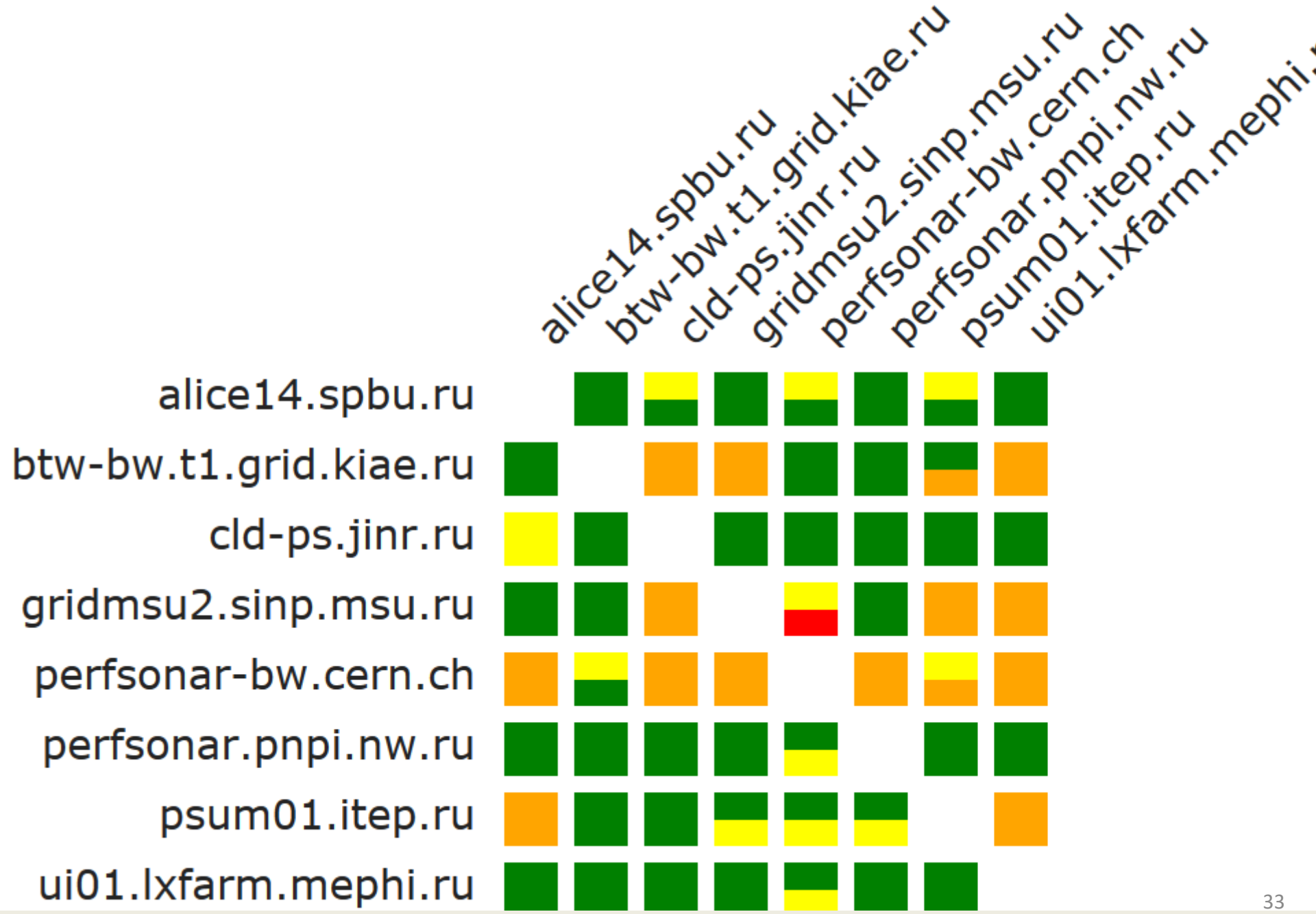
- We have established a vast, geographically distributed testbed for federated storage.
- We have a full mesh of perSONARs measuring all-to-all connectivity.
- We have performed an extensive testing of EOS-based federation with both synthetic tests and real-life workloads.
- There’s still much to be done, we have a dense roadmap in our hands.

Thank you for your attention!

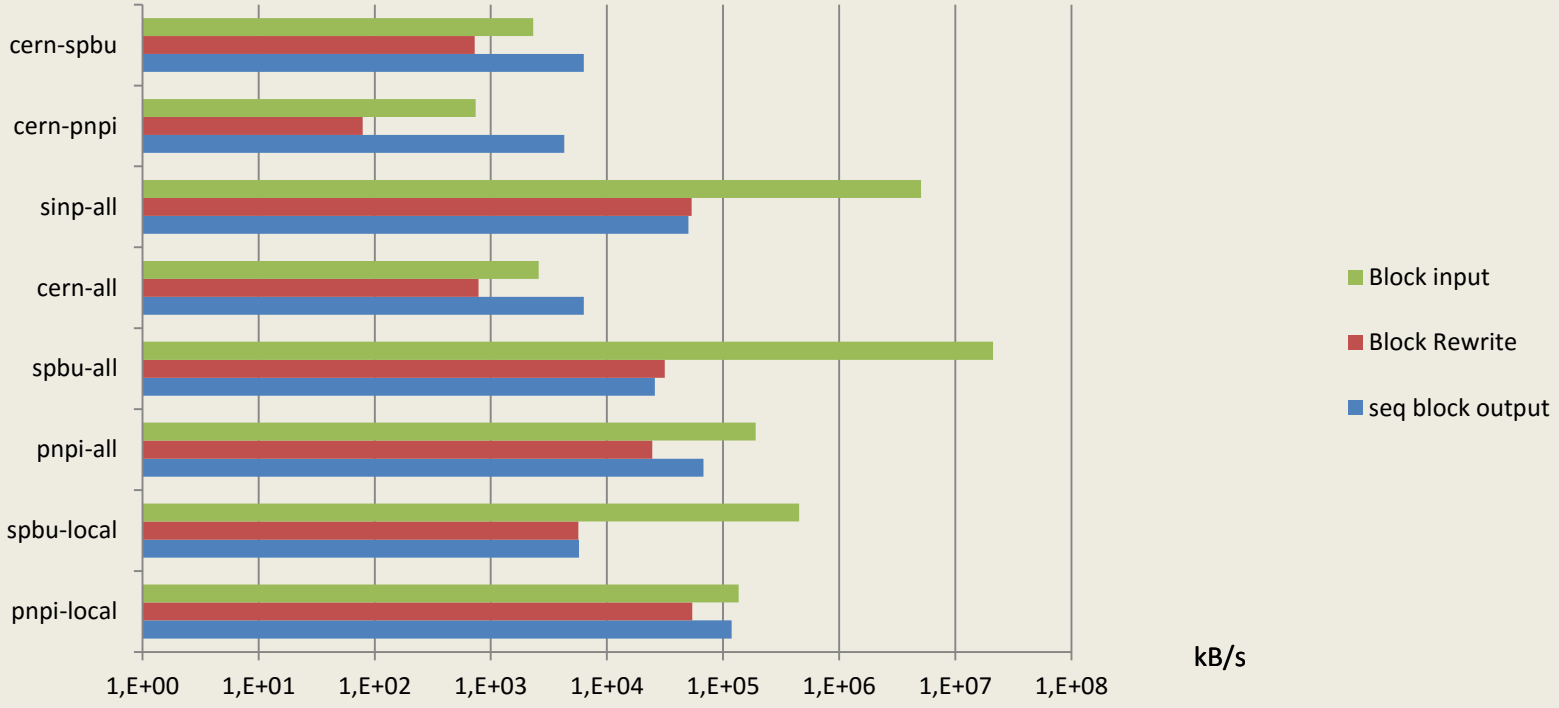
BACKUP

Participants and site structure




alice14.spbu.ru
btw-bw.t1.grid.kiae.ru
cld-ps.jinr.ru
gridmsu2.sinp.msu.ru
perfsonar-bw.cern.ch
perfsonar.pnpi.nw.ru
psum01.itep.ru
ui01.lxfarm.mephi.ru
alice14.spbu.ru
btw-bw.t1.grid.kiae.ru
cld-ps.jinr.ru
gridmsu2.sinp.msu.ru
perfsonar-bw.cern.ch
perfsonar.pnpi.nw.ru
psum01.itep.ru
ui01.lxfarm.mephi.ru

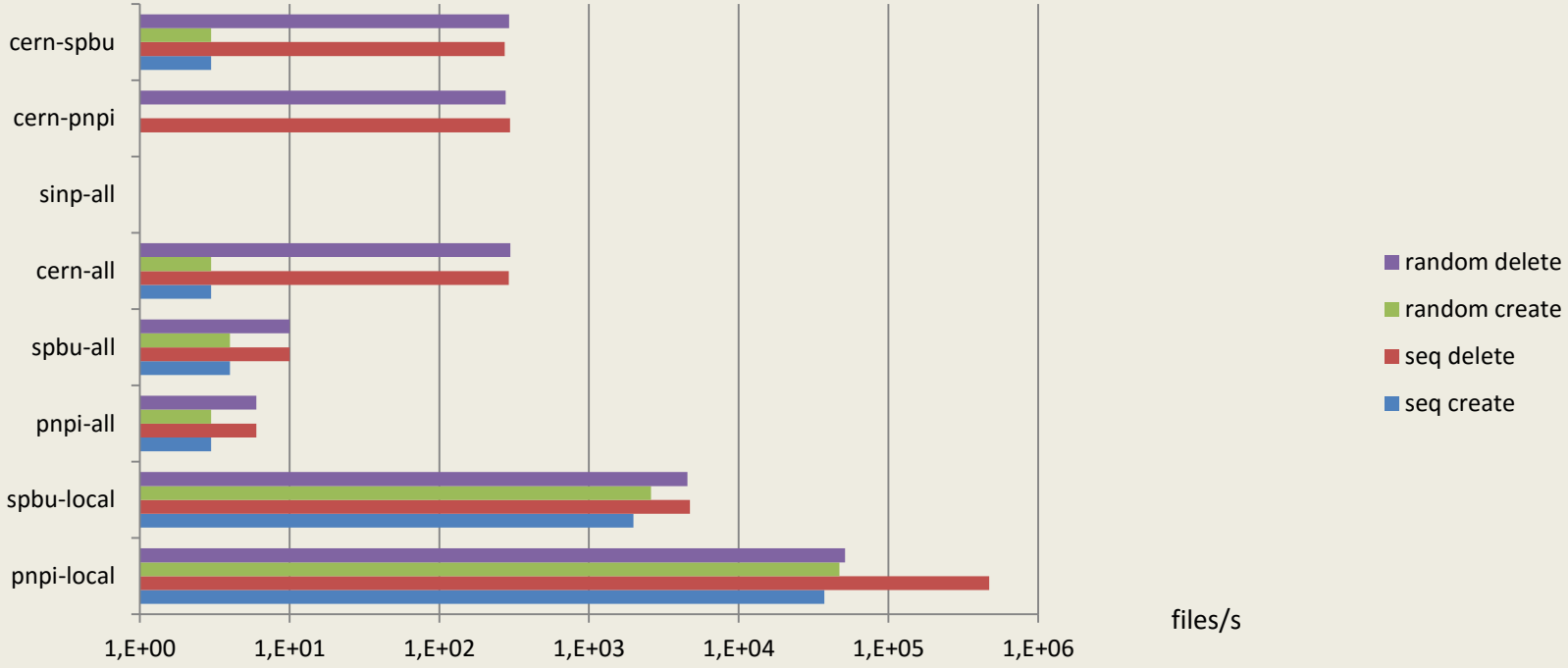
Bonnie tests on data read-write



pnpi-local - local test on PNPI SE
spbu-local - local test on SPbSU SE
pnpi-all – UI-PNPI,MGM –CERN,SE -Federation
spbu-all – UI-SPbSU,MGM –CERN,SE –Federation

cern-all – UI-CERN,MGM –CERN,SE -Federation
sinp-all – UI-SINP,MGM –CERN,SE -Federation
cern-pnpi – UI-CERN,MGM –CERN,SE -PNPI
cern-spbu – UI-CERN,MGM –CERN,SE -SPbSU

Bonnie tests on metadata read-write



pnpi-local - local test on PNPI SE
spbu-local - local test on SPbSU SE
pnpi-all – UI-PNPI,MGM –CERN,SE -Federation
spbu-all – UI-SPbSU,MGM –CERN,SE –Federation

cern-all – UI-CERN,MGM –CERN,SE -Federation
sinp-all – UI-SINP,MGM –CERN,SE -Federation
cern-pnpi – UI-CERN,MGM –CERN,SE -PNPI
cern-spbu – UI-CERN,MGM –CERN,SE -SPbSU