Federated data storage system prototype for LHC experiments and data intensive science

Andrey Kiryanov, Alexei Klimentov, Dimitrii Krasnopevtsev,
Artem Petrosyan, Eygene Ryabinkin, Andrey Zarochentsev

GRID 2016, JINR LIT, Dubna, 4-9 July

Project motivation

- Computing models for the LHC Run3 and High Luminosity era anticipate a growth of storage needs of at least two orders of magnitude;
- The reliable operation of large scale data facilities need a clear economy of scale;
- A distributed heterogeneous system of independent storage systems is difficult to be used efficiently by user communities and couples the application level software stacks with the provisioning technology at sites;
- Federating the data centers provides a logical homogeneous and consistent reliable resource for the end users;
- Small institutions have no enough people to support a fully-fledged software stack. Distributed stuff like ATLAS FAX, ALICE xrootd, EOS@CERN, CMS AAA, dCache, etc (mostly) works;
- In our R&D project we try to analyze how to set up a distributed storage within national region and how it can be used from Grid sites, from HPC, academic and commercial clouds, etc.
    - Also part of WLCG Federated storage demonstrator

Basic Requirements for a Federated Storage

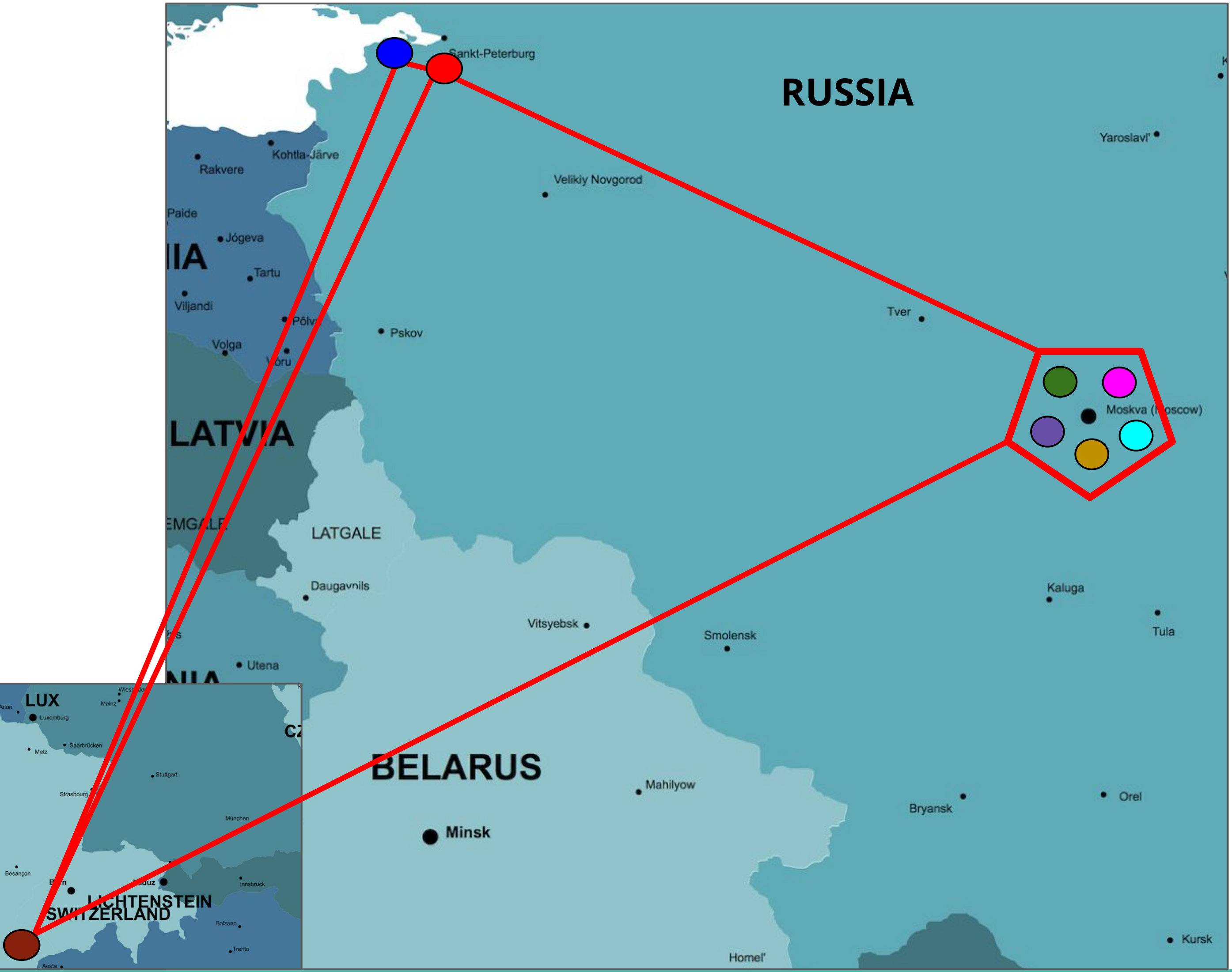
- Single entry point;
- Should be usable by at least two major LHC experiments;
- Scalability and integrity: it should be easy to add new resources;
- Data transfer optimization: transfers should be routed directly to the disk servers avoiding intermediate gateways and other bottlenecks;
- Stability and fault tolerance: redundancy of core components;
- Built-in virtual namespace, no dependency on external catalogues.

Technology choice

We had to find a software solution that is capable of federating distributed storage resources. This very much depends on a transfer protocol support for redirection. Three protocols that are capable of it are xroot, HTTP and pNFS. We have looked through various storage solutions used in WLCG and selected three of them for thorough testing:

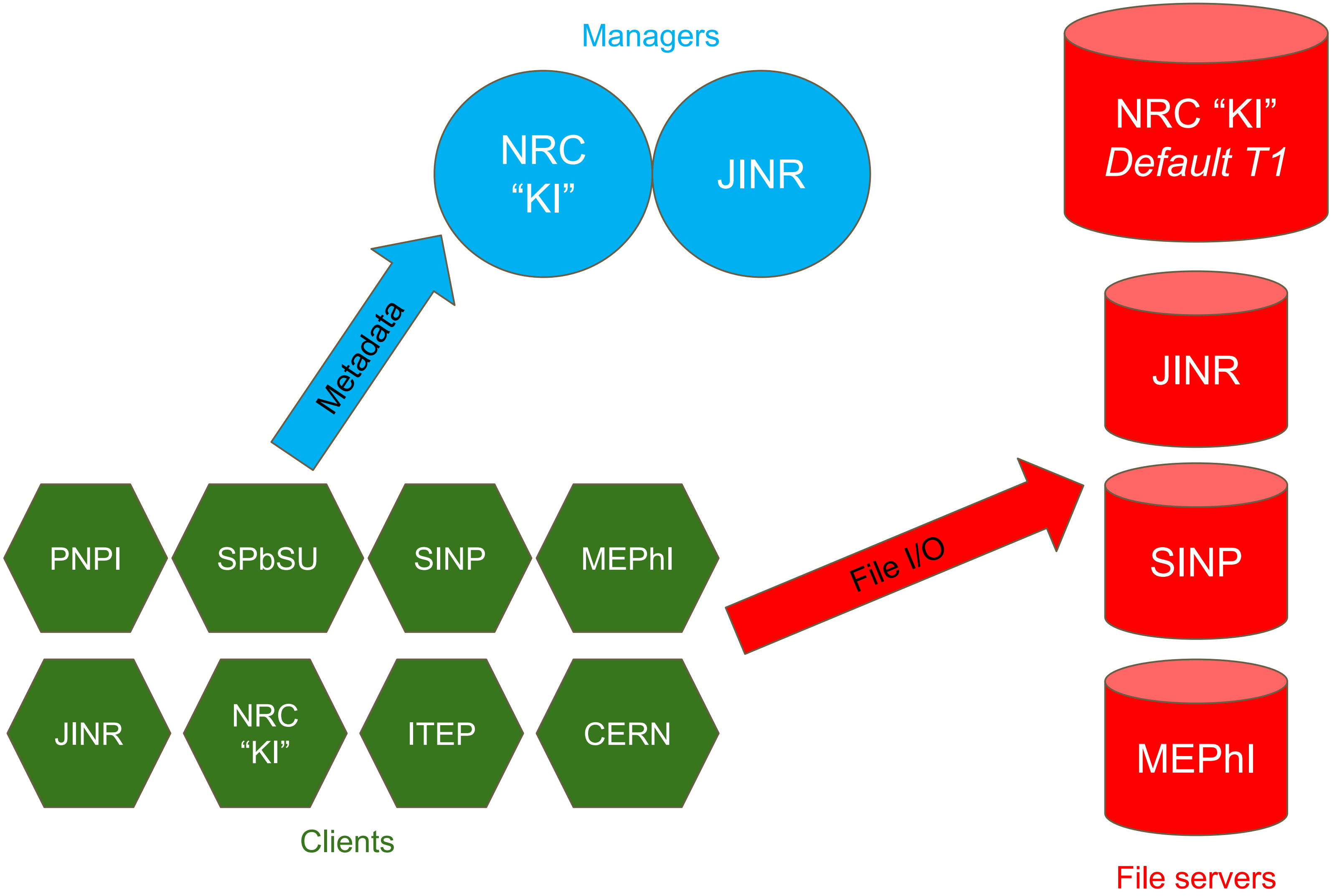
- EOS: xroot-based solution that is developed at CERN (we know where to ask for help), has characteristics closely matching our requirements, and is already used by all major LHC experiments;
- dCache: dCap/pNFS-based storage system developed at DESY. Depending on the Persistency Model, dCache provides 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;
- DynaFed: HTTP-based federator developed at CERN. This software is highly modular but only provides a federation frontend while storage backend(s) have to be chosen separately. While we were looking for more all-in-one solution it would be interesting to try it out eventually.

Participating sites

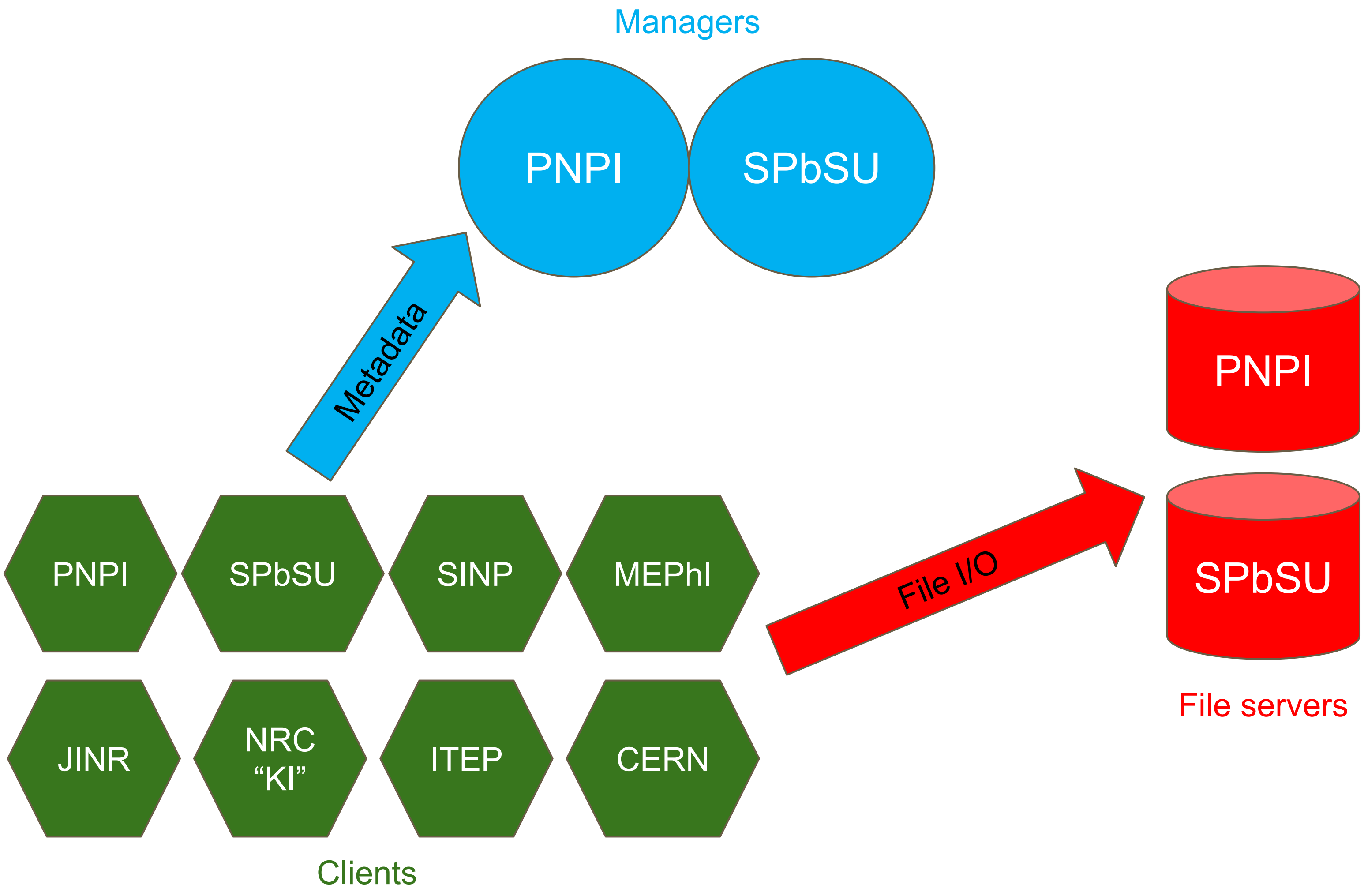


- ● SPbSU
- ● PNPI
- ● JINR
- ● NRC “KI”
- ● MEPhI
- ● SINP
- ● ITEP
- ● CERN

Main testbed



Secondary testbed



Testing plan

- One-shot 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);
  - Redirection impact test: check if there’s performance degradation with remote “head” node;
  - Reliability test: MGM master-slave migration.
- Continuous tests
  - Performance tests: file and metadata I/O, network;
  - Data locality test: evaluate EOS geo-tags role in data distribution;
  - Real-life tests using experiment software and real data.

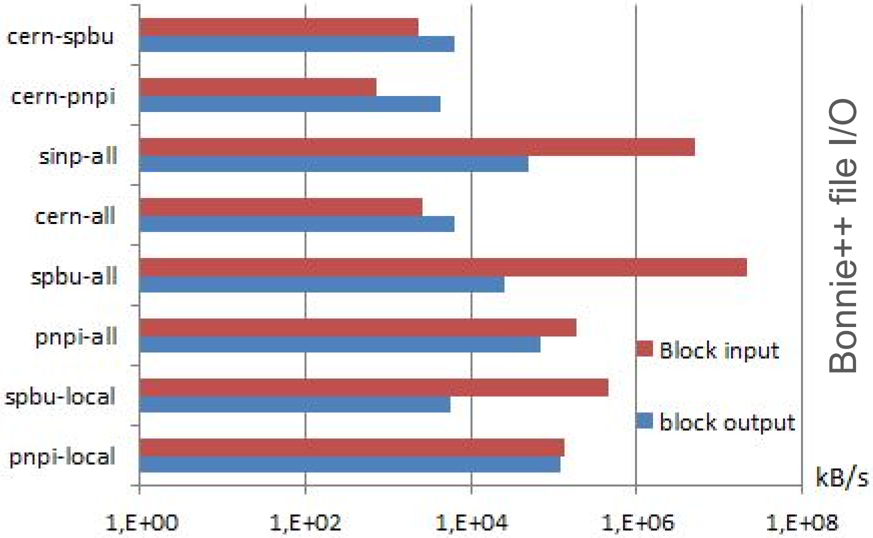
Software and tests

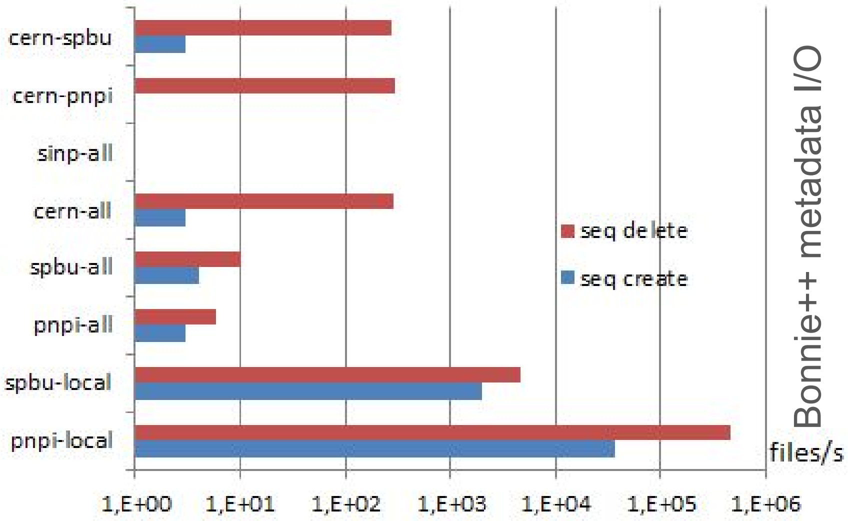
- Base OS: SL6 64bit
- Storage system: EOS Aquamarine
- Authentication scheme: GSI
- Network monitoring: perfSONAR
- Synthetic tests
  - Bonnie++: file and metadata I/O test for mounted file systems (FUSE)
  - xrdstress: EOS file I/O stress test via xroot protocol
- Real-life experiment tests:
  - ATLAS test: standard ATLAS TRT reconstruction workflow with Athena
  - ALICE test: sequential ROOT event processing (thanks to Peter Hristov)

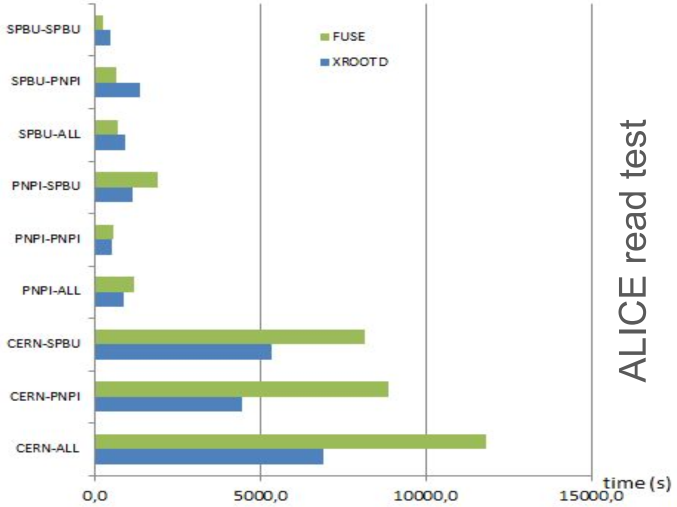
Replication policies

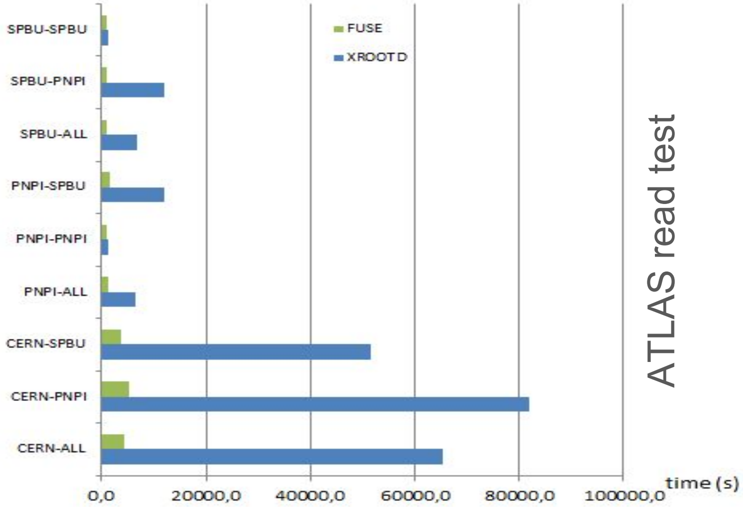
- Desired policy:
  - Read the closest replica
  - Write two replicas: first one on a closest T2, second one on a configured T1
- EOS implementation:
  - *Hybrid* placement policy allows first replica to be placed on a closest FST and the second one “scattered” to a random FST
  - EOS developers think that our desired policy makes sense and may be implemented in the future
- Currently we have three placement policies:
  1. Single replica without geotags: random placement;
  2. Single replica with geotags: write to FST with geotag matching the UI, if there’s no match default FST is used (KIAE);
  3. Two replicas with geotags: 1st replica on a closest FST, 2nd on a random FST.

The first results with EOS. PNPI and SPbSU.


Bonnie++ file I/O


Bonnie++ metadata I/O


ALICE read test


ATLAS read test

First EOS experience and intermediate conclusion

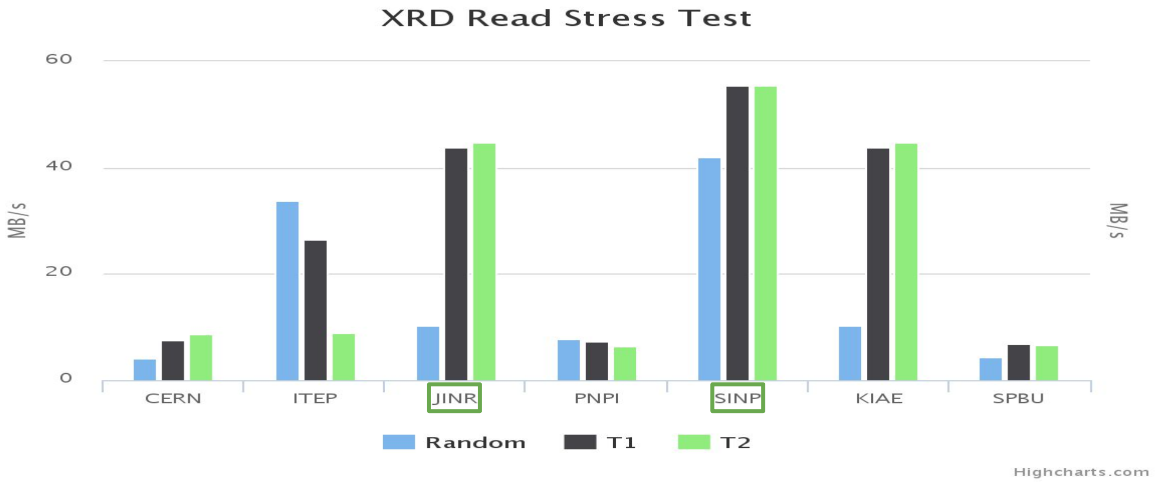
1) Basic stuff works as expected;
2) Problems with master-slave migration were reported to the developers;
3) As expected, metadata I/O performance depends only on Manager location while file I/O performance depends only on File server location;
4) Experiment tests behave differently with different data sets (many small files vs. few large files) and different protocols (pure xroot vs. FUSE-mounted filesystem);
5) Testbed used network connection that was shared with production sites, which had lead to a situation when different test runs produced different results.
   ⇒ we needed to take into account network measurements

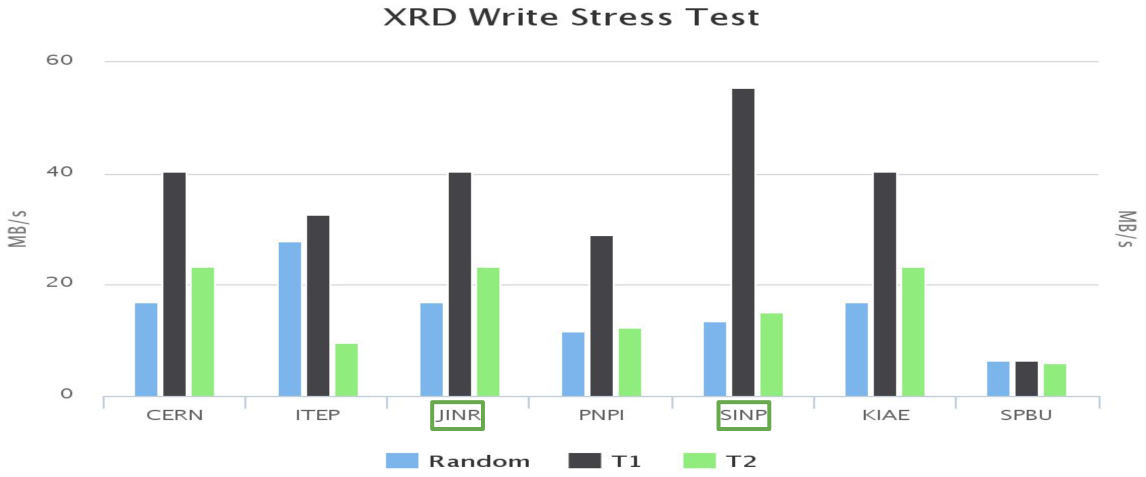
Network performance measurements

- 🟩 Throughput >= 100Mbps
- 🟨 Throughput >= 10Mbps
- 🟥 Throughput < 10Mbps
- 🟧 Unable to retrieve data

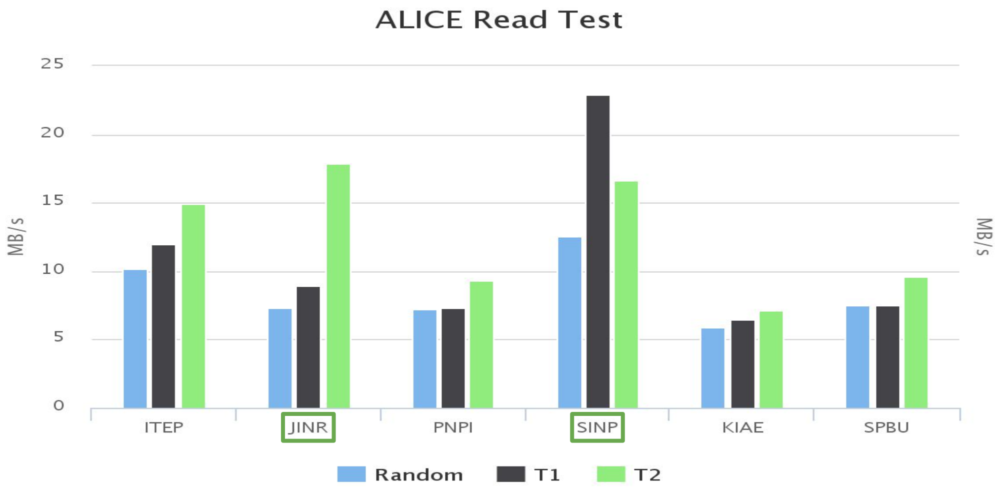
	SPbSU	KI	JINR	SINP	CERN	PNPI	ITEP	MEPHI
SPbSU		🟩	🟨 🟩	🟩	🟨 🟩	🟩	🟨 🟩	🟩
KI	🟩		🟧	🟧	🟩	🟩	🟩 🟧	🟧
JINR	🟨	🟩		🟩	🟩	🟩	🟩	🟩
SINP	🟩	🟩	🟧		🟨 🟥	🟩	🟧	🟧
CERN	🟧	🟨 🟩	🟧	🟧		🟧	🟨 🟧	🟧
PNPI	🟩	🟩	🟩	🟩	🟩 🟨		🟩	🟩
ITEP	🟧	🟩	🟩	🟩 🟨	🟩 🟨	🟩 🟨		🟧
MEPHI	🟩	🟩	🟩	🟩	🟩 🟨	🟩	🟩	

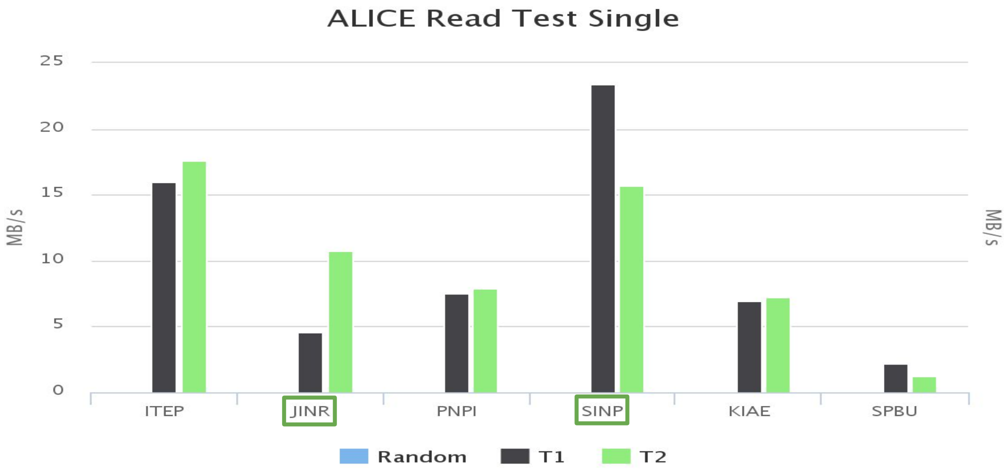
Synthetic tests w/geo-tags on the Main testbed



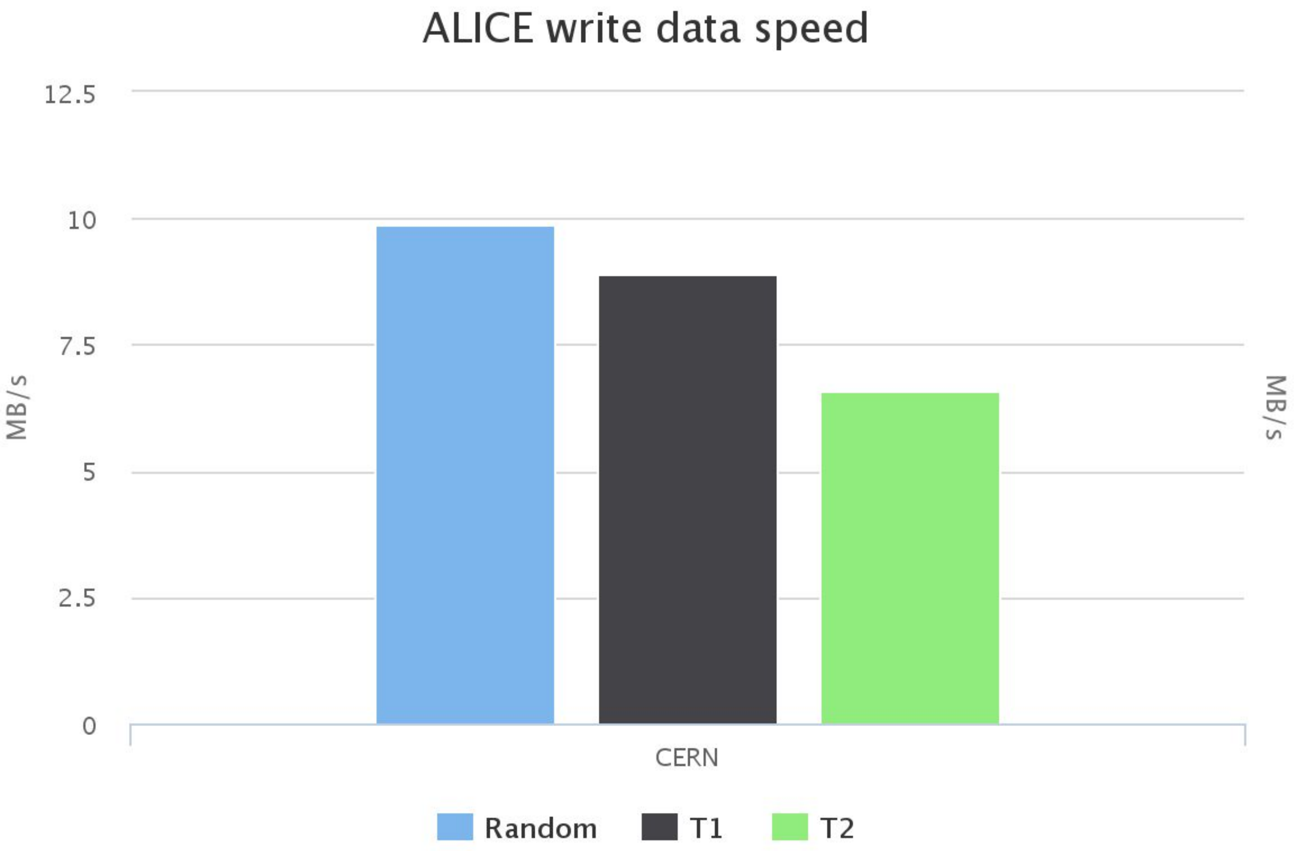


ALICE read test w/geo-tags on the Main testbed





ALICE write test w/geo-tags from CERN



Conclusions

- We have created a working prototype of distributed federated storage system joining seven major Russian scientific organizations, which is accessible from virtually everywhere;
- We have carried out an extensive testing of our infrastructure using various tools. What is even more important, we have understood the results;
- We have exploited EOS as our first technological platform and we have enough confidence to say that it behaves well and has all the features we need;
- The story is not over though, we have other software solutions (dCache and DynaFed) to test.

Next steps

1. Put in order all test results that we’ve got so far. We’re going to need them later for comparison;
2. Get familiar with the next storage system (dCache) and deploy it on our infrastructure;
3. Repeat all of our tests. We may need different tools for new protocols;
4. goto 1

Acknowledgements

This talk drew on presentations, discussions, comments and input from many. Thanks to all, including those we’ve missed

D. Duelmann, P. Fuhrnmann, A. Kryukov, M. Lamanna, E. Lyublev, T. Mkrtchan, A. Peters, S. Smirnov, V. Velikhov

This work was funded in part by the Russian Ministry of Science and Education under Contract No 14.Z50.31.0024 and by the Russian Fund of Fundamental Research under contract “15-29-0794_2 офи_м”

Authors express appreciation to SPbSU Computing Center, PNPI ITAD Computing Center, NRC “KI” Computing Center, JINR Cloud Services, ITEP, SINP and MEPhI for provided resources.

Thank you!

Backup slides

Full test plan

Bonnie++

a. PNPI FST local

b. SPbSU FST local

c. UI PNPI to ALL

d. UI SPbSU to ALL

e. UI CERN to ALL

ATLAS (2 files on two FST)

-----------------------****

a. UI PNPI to ALL (fuse, xrootd)

b. UI SPbSU to ALL (fuse, xrootd)

c. UI CERN to ALL (fuse, xrootd)

ALICE (322 files on two FST)

a. UI PNPI to ALL (fuse, xrootd)

b. UI SPbSU to ALL (fuse, xrootd)

c. UI CERN to ALL (fuse, xrootd)

PerfSonar

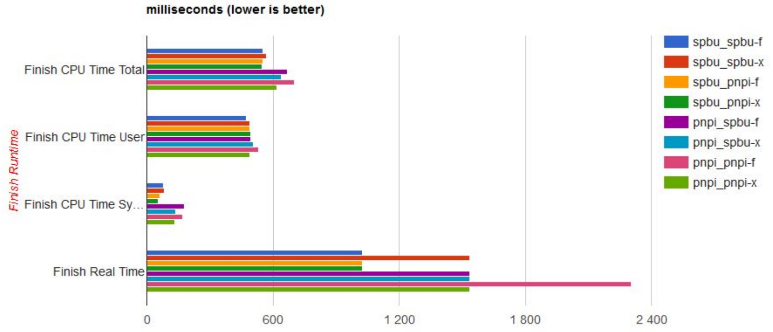
a. CERN-PNPI

b. CERN-SPbSU

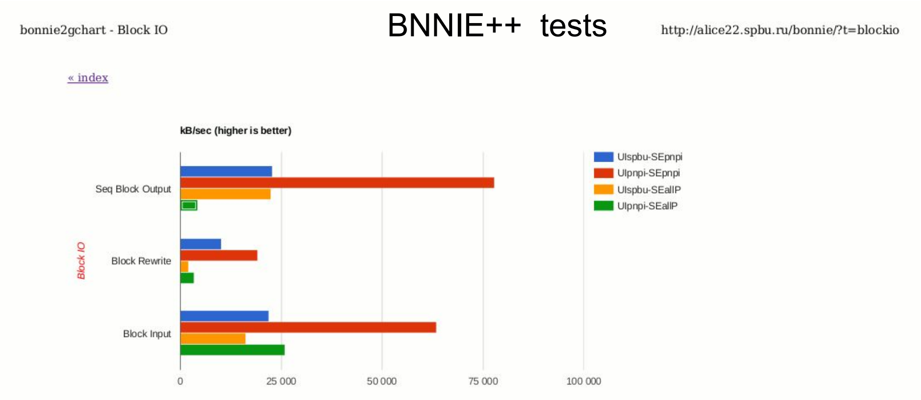
c. PNPI-SPbSU

First results without conclusions

ATLAS tests



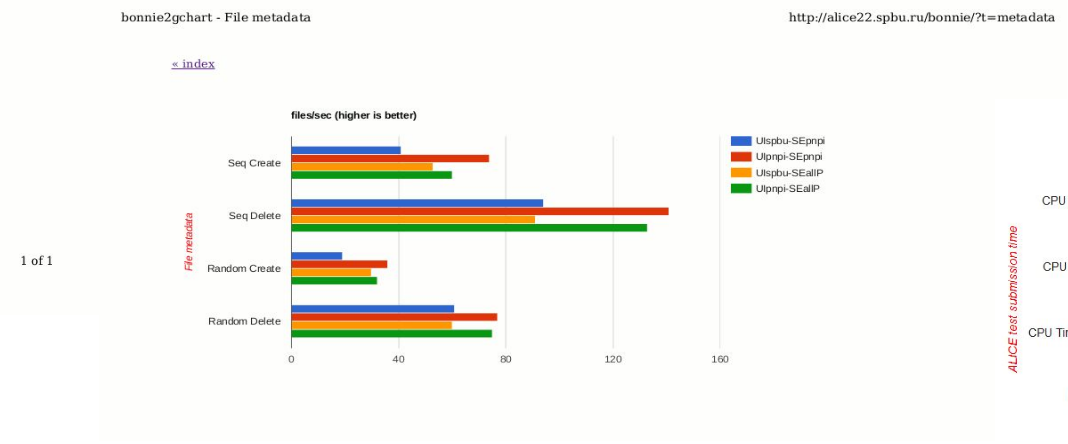
BNNIE++ tests



bonnie2gchart - Block IO

http://alice22.spbu.ru/bonnie/?t=blockio

« index



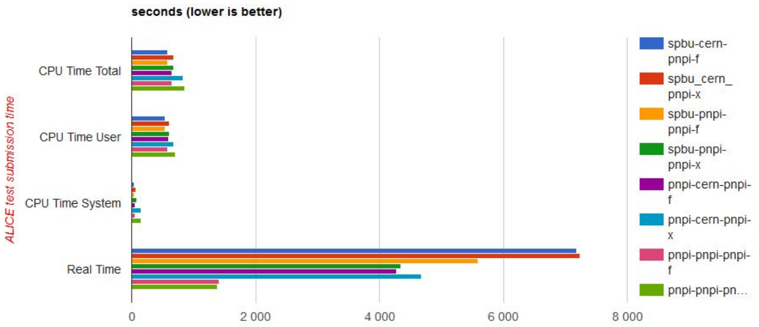
bonnie2gchart - File metadata

http://alice22.spbu.ru/bonnie/?t=metadata

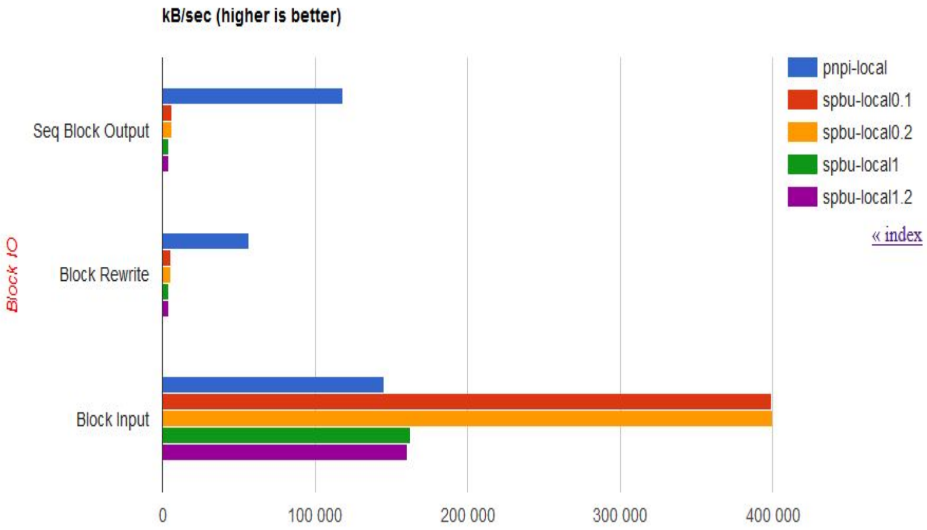
« index

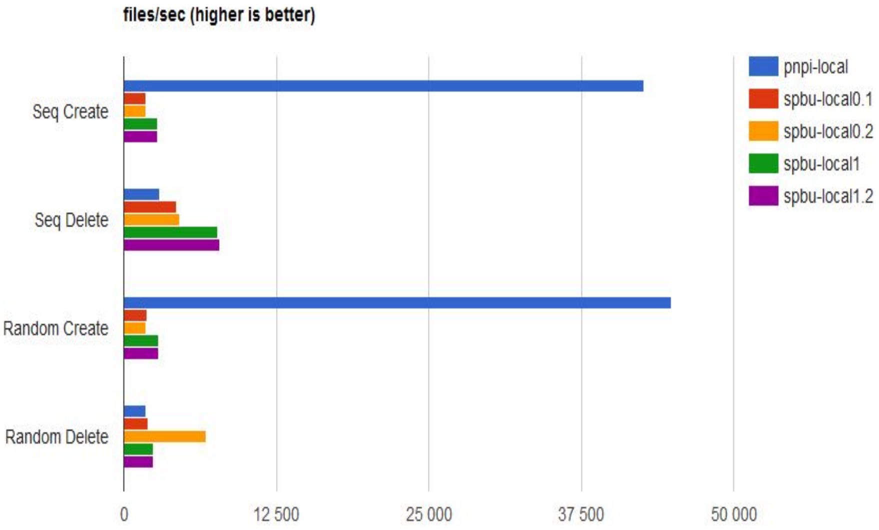
1 of 1

ALICE tests



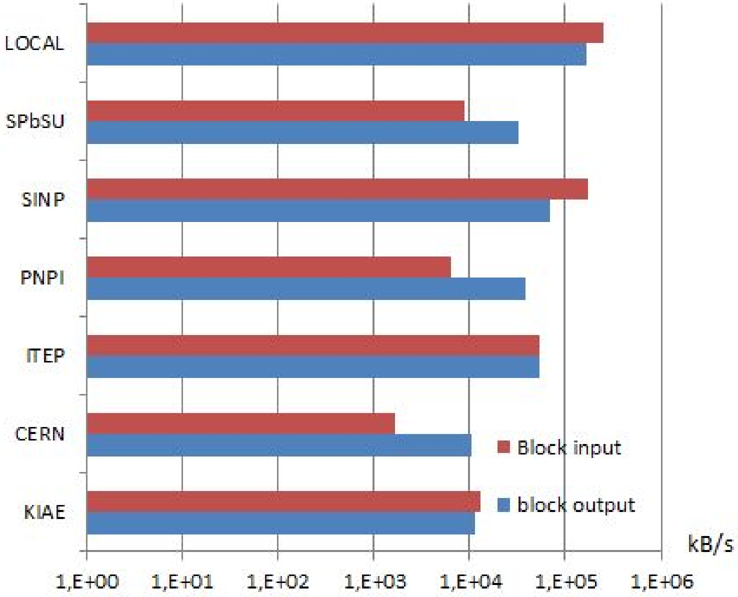
Bonnie local tests - different pools on SPbSU



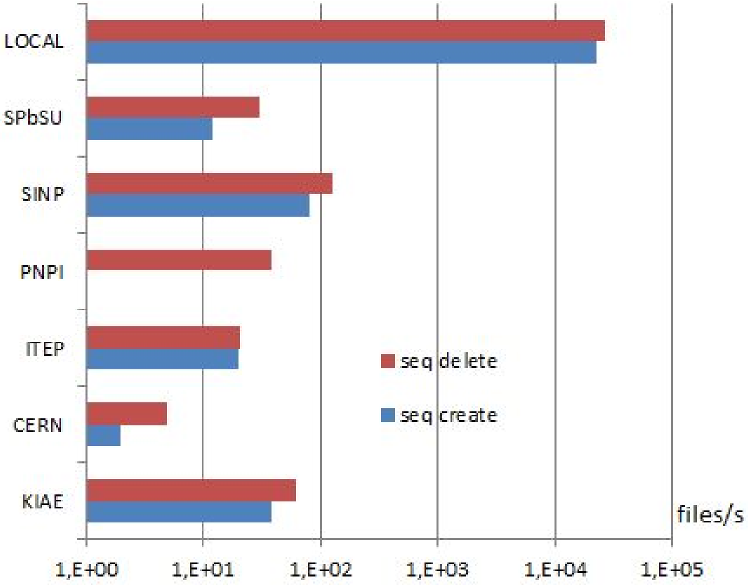


Bonnie test for 2nd testbed – MGM and FST at KI

Data read-write



Metadata read-write


LOCAL
SPbSU
SINP
PNPI
ITEP
CERN
KIAE
🟥 seq delete
🟦 seq create
files/s
1,E+00
1,E+01
1,E+02
1,E+03
1,E+04
1,E+05