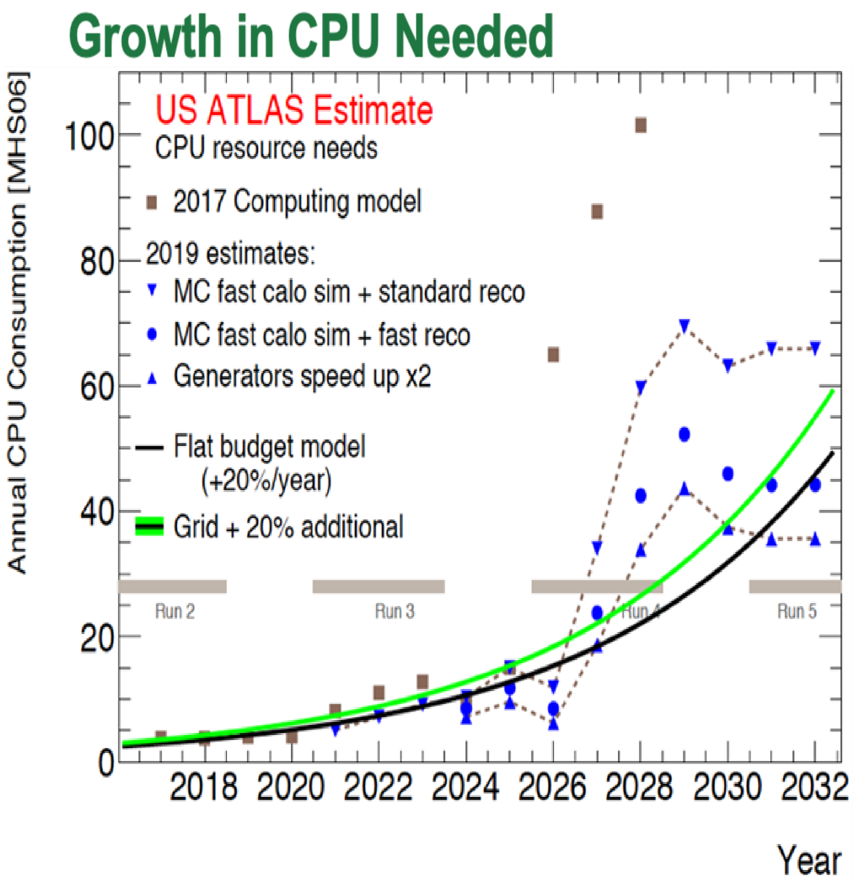
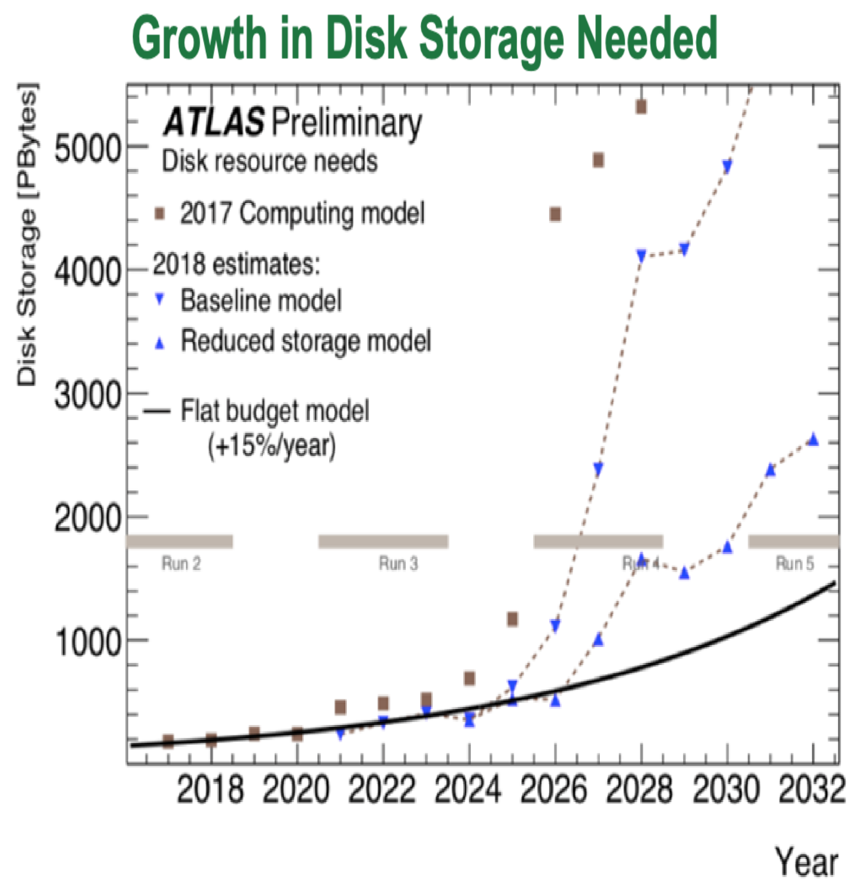
Data Lake: configuration and testing of a distributed data storage system

Aleksandr Alekseev, Xavier Espinal, Stephane Jezequel,
Andrey Kiryanov, Alexei Klimentov, Valery Mitsyn,
Andrey Zarochentsev

October 18, 2019

The High Luminosity LHC Challenge



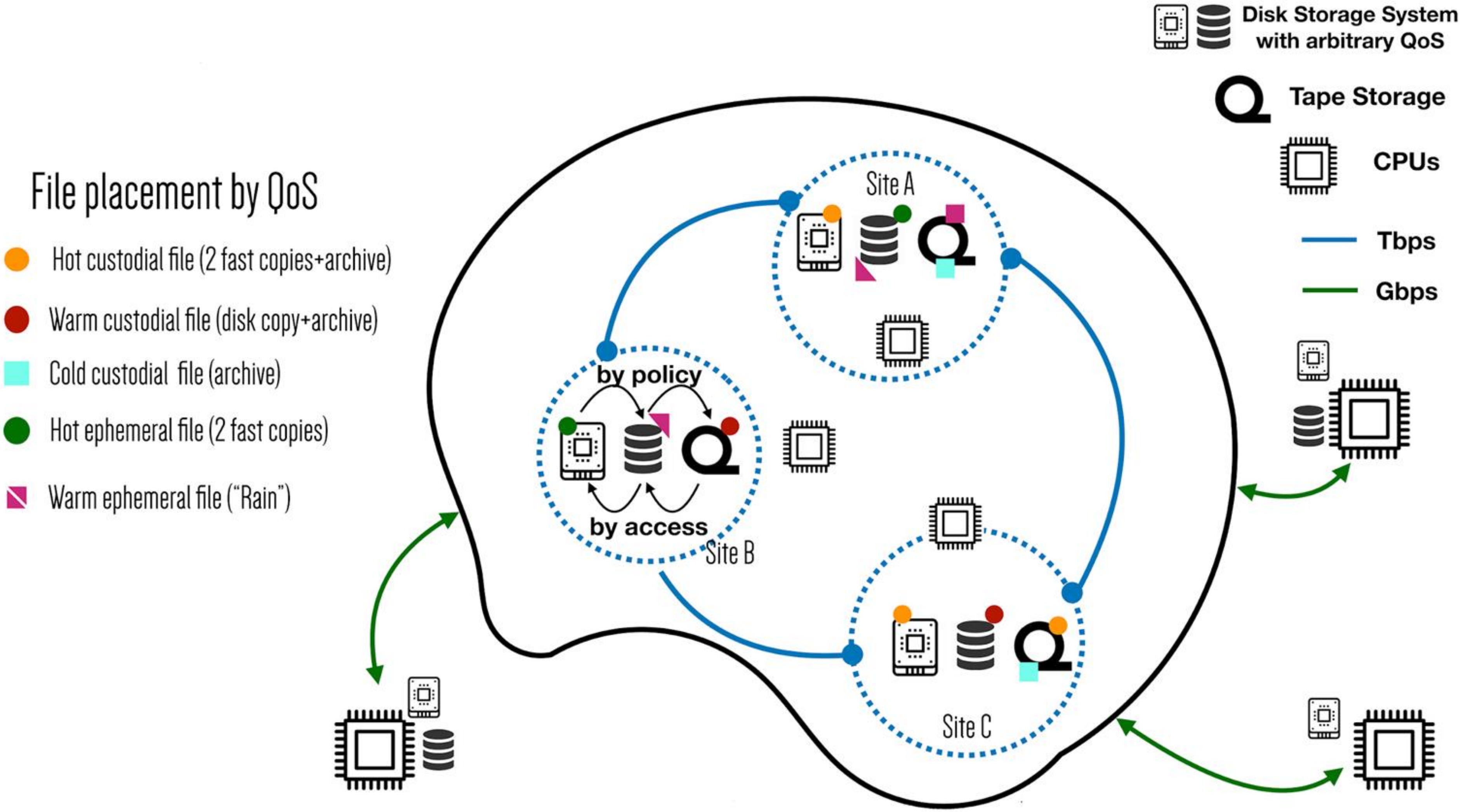


- High Luminosity LHC will be a multi-exabyte challenge where the envisaged Storage and Compute needs are a factor 10 to 100 above the expected technology evolution.
- LHC experiments have successfully integrated HPC facilities into its distributed computing system. “Opportunistic storage” basically does not exist for LHC experiments.
- The HEP community needs to evolve current computing and data organization models in order to introduce changes in the way it uses and manages the infrastructure, focused on optimizations to bring performance and efficiency not forgetting simplification of operations.

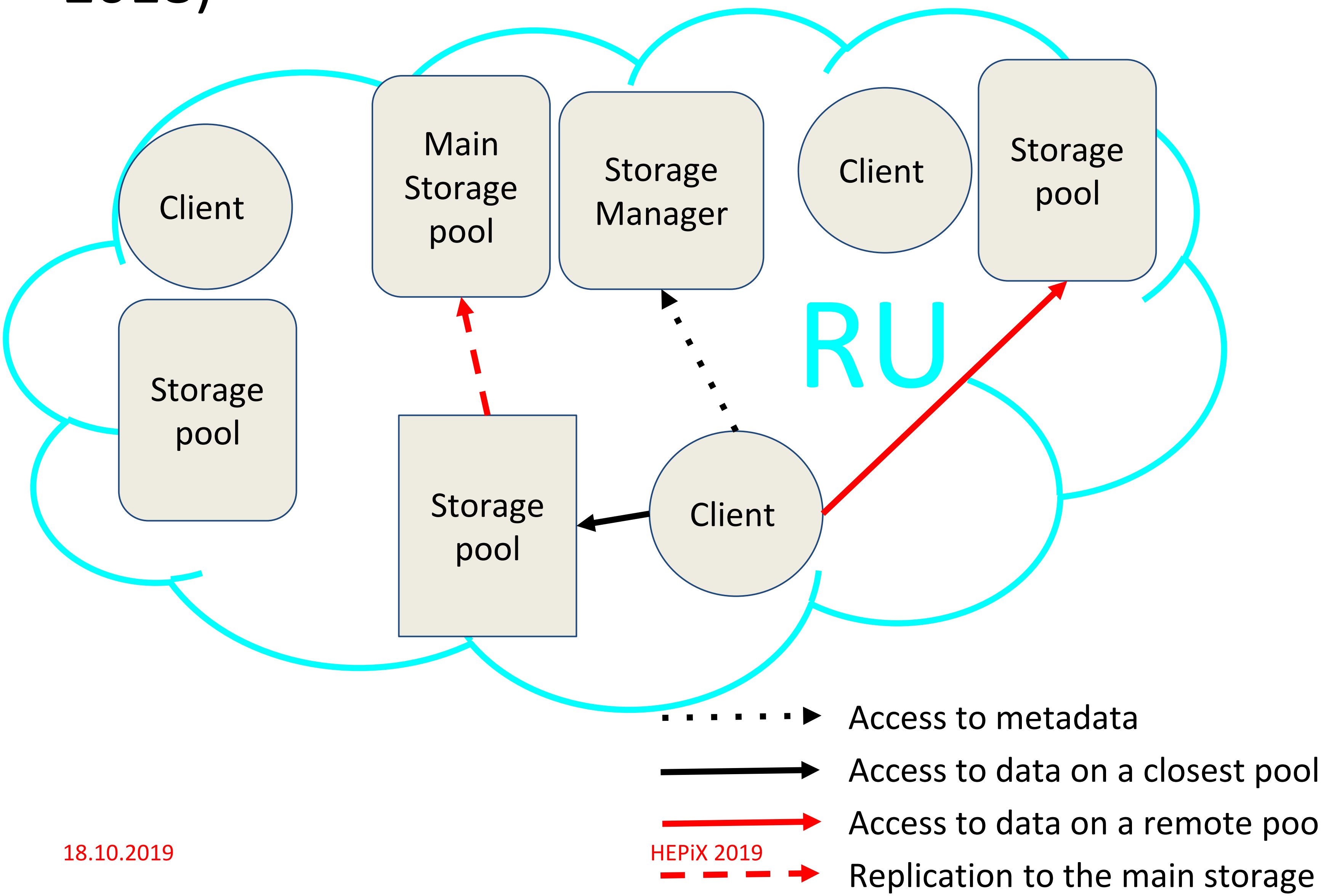
WLCG DOMA Project

- HL-LHC will be a (multi) Exabyte challenge.
- The WLCG community needs to evaluate LHC computing model to store and manage data efficiently.
- The technologies that will address the HL-LHC computing challenges may be applicable for other communities to manage large-scale data volume (SKA, DUNE, CTA, LSST, BELLE-II, JUNO, NICA, etc).
- WLCG has launched Data Organization Management and Access (DOMA) project to address HL-LHC data challenges.
  - the Data Lake R&D is a part of DOMA. The aim is to consolidate geographically distributed data storage systems connected by fast network with low latency.
  - we see the Data Lake model as an evolution of the current infrastructure bringing reduction of the storage and operational costs

Data Lake R&D


File placement by QoS
Hot custodial file (2 fast copies+archive)
Warm custodial file (disk copy+archive)
Cold custodial file (archive)
Hot ephemeral file (2 fast copies)
Warm ephemeral file ("Rain")
Disk Storage System with arbitrary QoS
Tape Storage
CPUs
Tbps
Gbps
Site A
by policy
by access
Site B
Site C

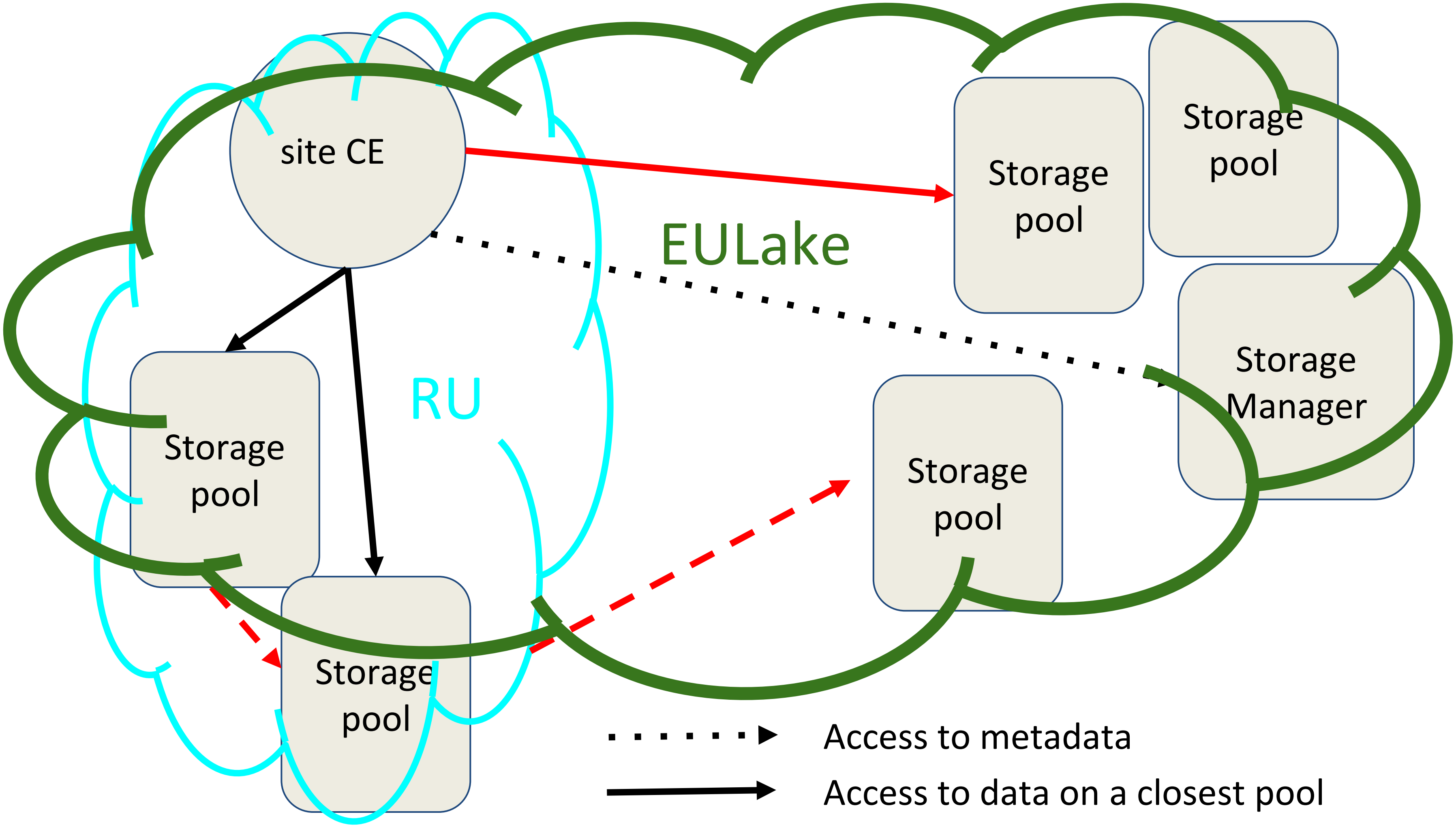
Russian Federated Storage Project (2015 – 2018)



Russian Federated Storage Project (2015 – 2018)

- First distributed infrastructure of such scale, EOS & dCache (including interoperability tests), network monitoring, etc.
- Synthetic tests:
  - Bonnie++: file and metadata I/O test for mounted file systems (FUSE only)
  - xrdstress: EOS-bundled file I/O stress test for xrootd protocol
  - Simple xrdcp: copy files filled with random data
- Experiment-specific tests:
  - ATLAS test: standard ATLAS TRT reconstruction workflow with Athena
  - ALICE test: sequential ROOT event processing
- Experience gained during this project was later used on a EULake prototype (EOS-based Data Lake test with HQ @ CERN)

Russian part of EULake (2018 – 2019)



Federated Storage and EULake test results

1. Basic stuff works as expected (primary result ☺)
2. We have performance gain with copy to the closest pool (but we lose this gain with replication)
3. We have performance gain with copy from the closest pool if we have replica there

2 –> We need “smart” data management

3 –> To ensure availability of local replicas we can use caching

Report on dCache workshop 2017 in Umeo:

https://www.dcache.org/manuals/workshop-2017-05-29-Umea/000-Final/Andrey-dCache-Federated-Storage-V1.pdf

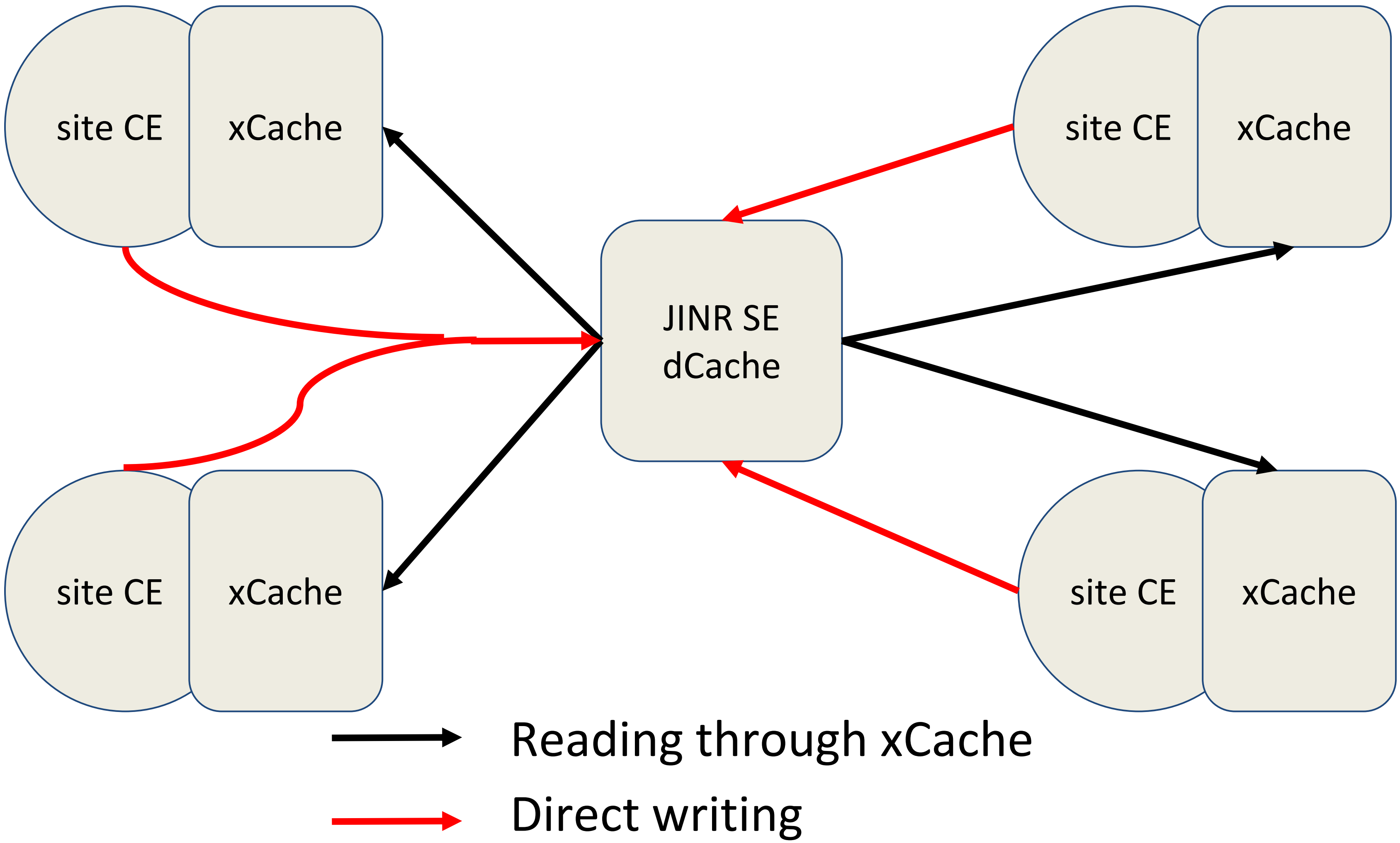
Report on ACAT 2019 in Saas-Fee:

https://indico.cern.ch/event/708041/contributions/3276346/attachments/1809212/2955264/DataLake-021.pdf

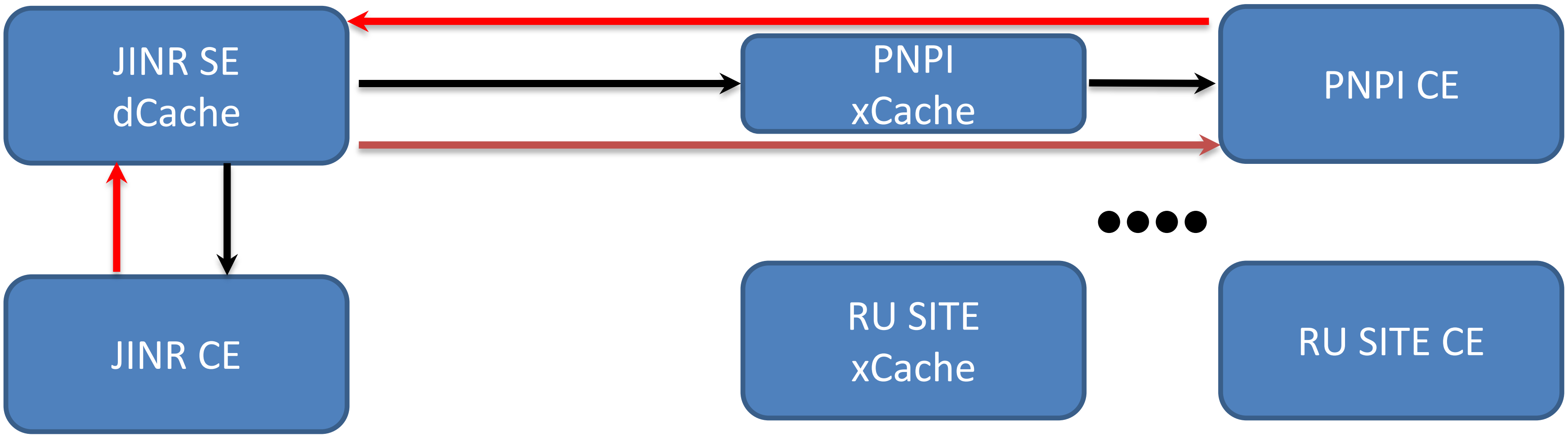
Russian Data Lake for HEP R&D Project

- The project has been launched in 2019
- The work is supported by the Russian Science Foundation award
- It will be 5 years project. Many Russian WLCG sites are involved : JINR, REU, SPbSU, PNPI, MEPhI,...
- We will work in very close collaboration with DOMA

Russian Data Lake Phase 1 (2019 Prototype)

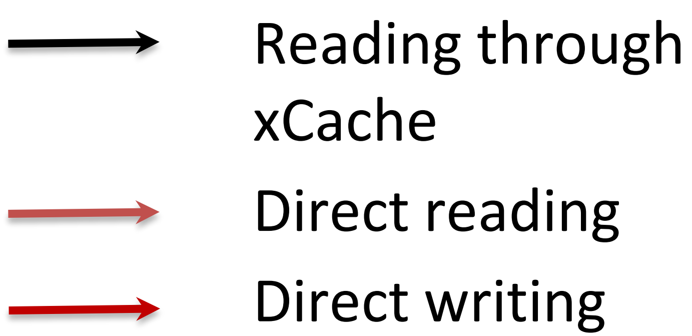


Plan of tests



Submitted tests:

1. Synthetic tests from Worker Nodes by hand and through Cream-CE
2. Two types of standard ATLAS tests through HammerCloud:
   a. Copy2Scratch
   b. Directaccess

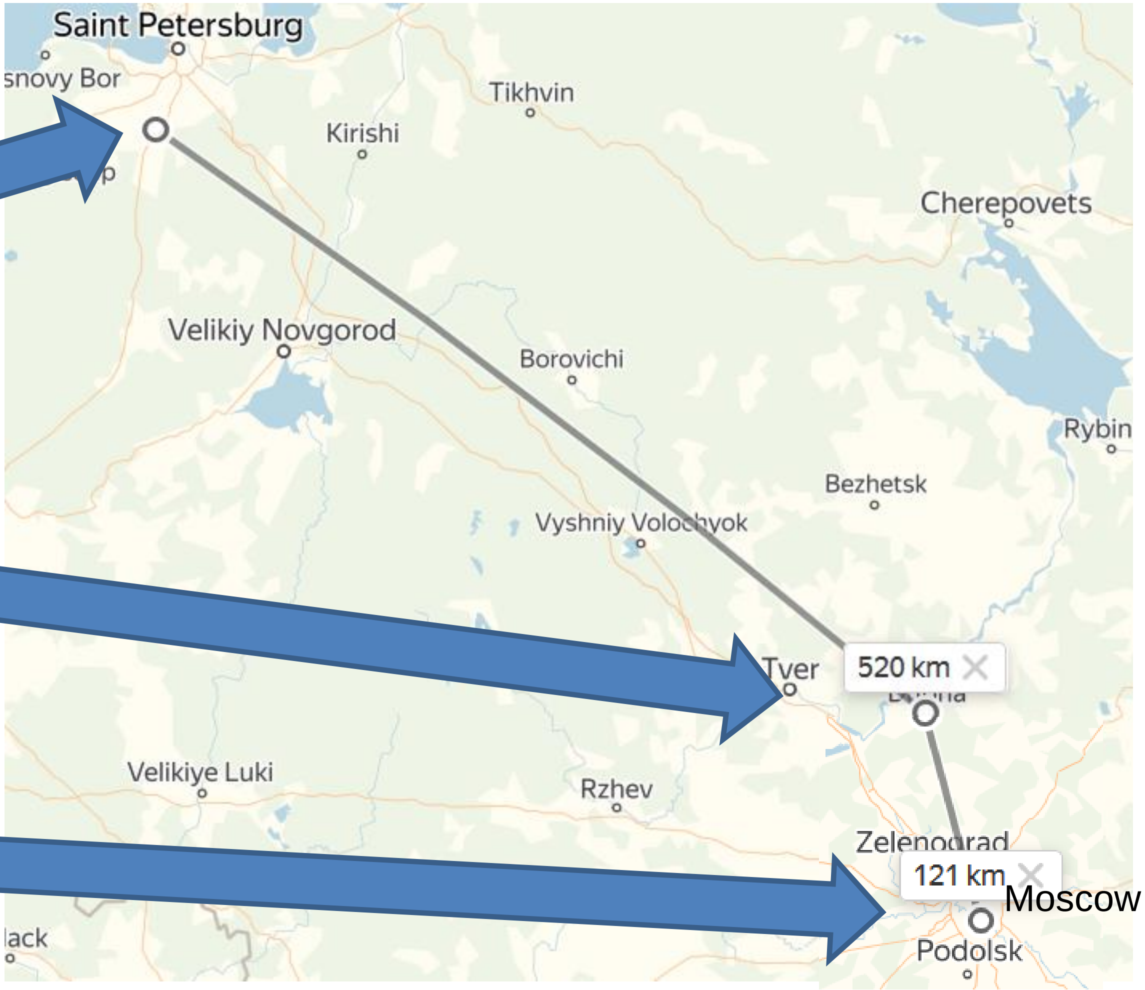


Participating Sites

Petersburg Nuclear Physics Institute

JINR

MEPhI



Authorization

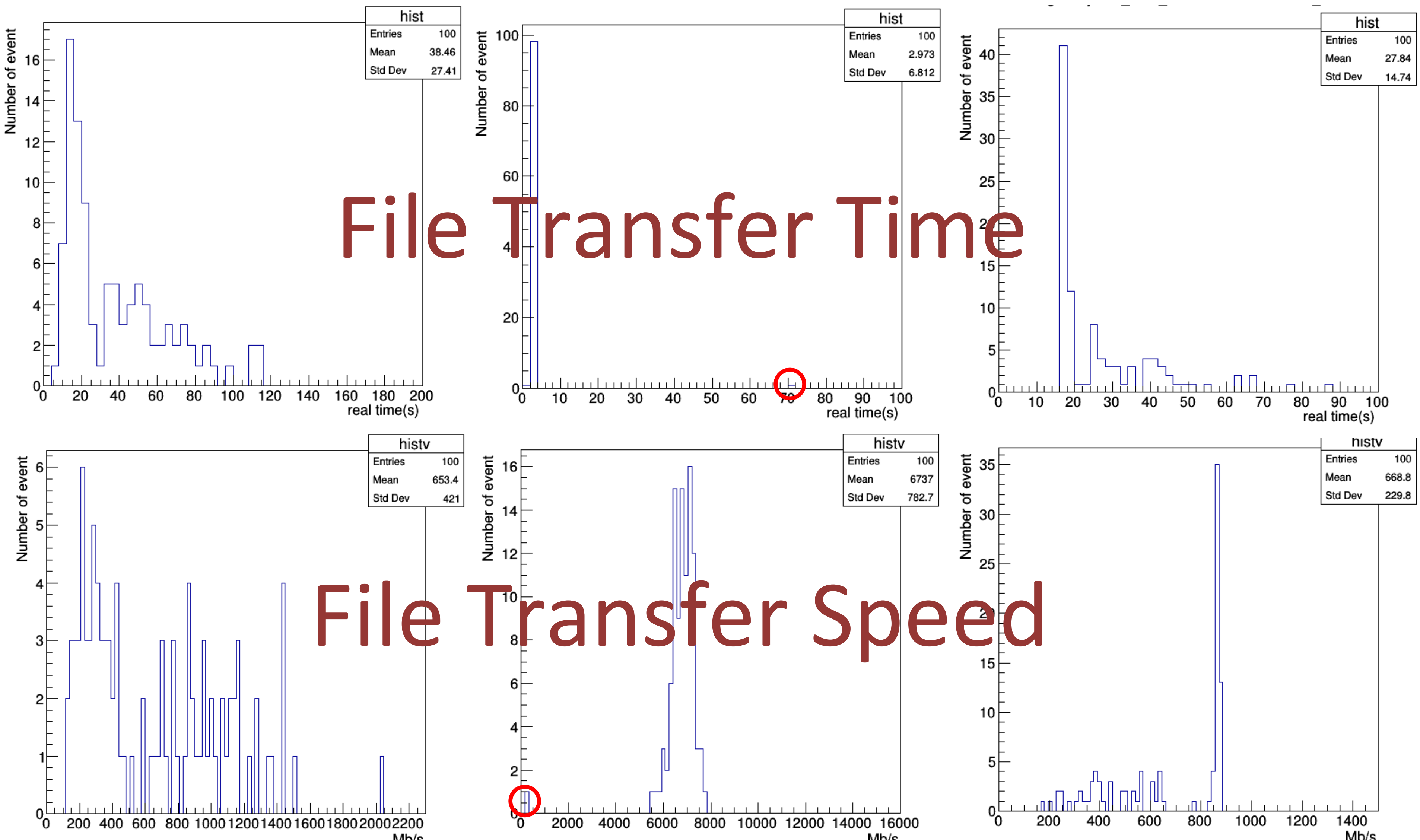
- PNPI xCache ➜ JINR SE: GSI authorization by local gridmapfile on JINR SE
- PNPI WN ➜ PNPI xCache: GSI authorization by VOMS (ALICE & ATLAS)
- PNPI UI ➜ JINR CE, PNPI CE (for local tests): GSI authorization by VOMS (ALICE & ATLAS)
- Hammer Cloud ➜ ALL: GSI authorization by VOMS (ATLAS)
- An external library for VOMS authorization in xCache: https://github.com/opensciencegrid/xrootd-lcmaps
- xCache (and probably xrootd in general) does not actually switch UNIX users, so we use *nobody* user as a stub.
  - "/atlas/Role=production" nobody

Technical specifications

- Worker Node @ JINR: 8 cores, Xeon E5420, 16GB RAM, 8.74 HEP-SPEC06 per Core
- Worker Node @ PNPI: 8 cores, Xeon E5-2680, 32GB RAM (VM), ~11 HEP-SPEC06 per Core
- Local network @ JINR (SE<->CE) 1Gb/s
- Local network @ PNPI (SE<->CE) 10Gb/s
- Network IPv4,6 JINR ➜ PNPI: Latency ~5ms
- Network IPv4,6 PNPI ➜ JINR: Latency ~10ms
- Network IPv4,6 JINR ➜ PNPI: Throughput ~1Gb/s
- Network IPv4,6 PNPI ➜ JINR: Throughput ~1,5Gb/s

Local test results: copy from JINR-SE 1.9 GB root file (100 iter.)

JINR-SE->PNPI-CE JINR-SE->xCache->PNPI-CE JINR-SE->JINR-CE



Local test results: copy from JINR-SE 1.9 GB root file (100 iter.)

- Mean FTS PNPI – Direct-SE: 650±40 Mb/s
  - < 1Gb/s
  - Time 38s
- Mean FTS PNPI – xCache-SE: 6700±700 Mb/s
  - One hit on 219 Mb/s, other hits with minimal deviation
  - Time 2s – We have 95% gain in time
- Mean FTS JINR – SE: 660±220 Mb/s
  - < 1Gb/s, large deviation

HammerCloud tests

state	id	host	clouds	start time (CET)	end time (CET)	total jobs
completed	20146370	hammercloud-ai-11	RU_PROD	24/9/2019 15:00	26/9/2019 15:00	411

Site	S	R	C	F	Eff	T
PNPI_XCACHE-TEST	4	4	204	0	1.00	212
PNPI-TEST	5	1	186	2	0.99	194
JINR_UCORE-TEST	1	0	4	0	1.00	5
Site	S	R	C	F	Eff	T

- Test number 20146370 from Template 1099 (copy2scratch)

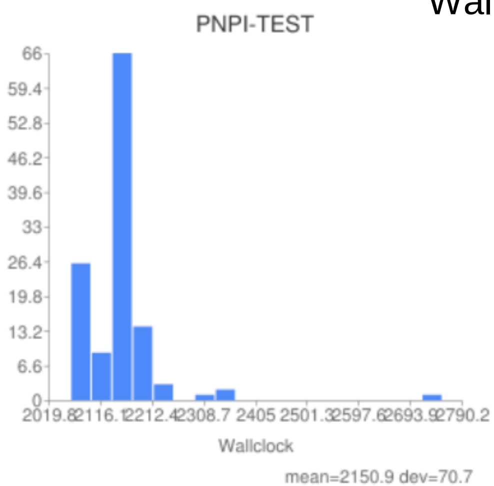
state	id	host	clouds	start time (CET)	end time (CET)	total jobs
completed	20146182	hammercloud-ai-11	RU_PROD	19/9/2019 12:00	21/9/2019 12:00	254

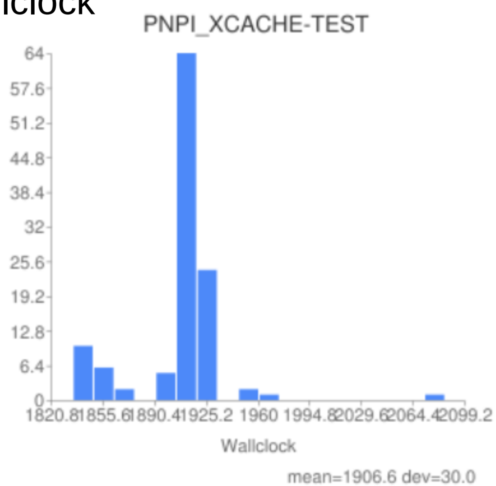
Site	S	R	C	F	Eff	T
PNPI_XCACHE-TEST	5	0	115	0	1.00	120
PNPI-TEST	5	0	122	2	0.98	129
JINR_UCORE-TEST	0	0	4	1	0.80	5
Site	S	R	C	F	Eff	T

- Test number 20146182 from Template 1100 (direct access)
- Weak statistics from JINR-CE for both tests (local problem with JINR-TEST-CE)

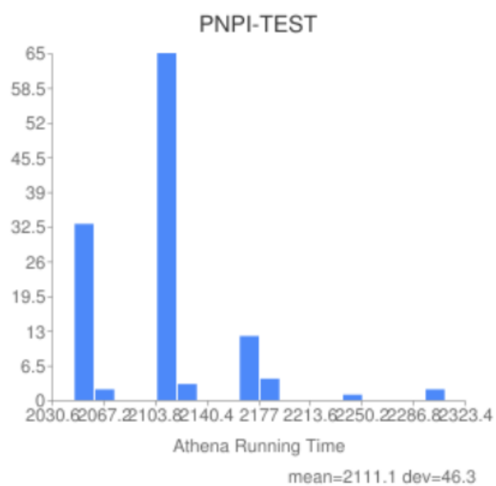
HammerCloud test results - N20146182 from Template 1100 (direct access)

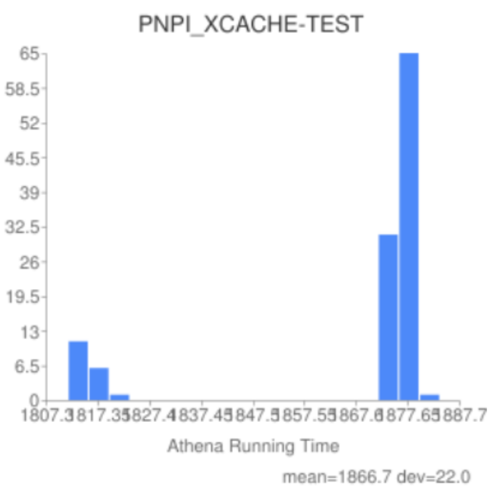
Wallclock





Athena Running Time





Wallclock:

Direct mean time = 2150s ± 70s

xCache mean time = 1906s ± 30s

Difference ~ 250s, ~12%

Download of input files time:

Direct mean time = 12s

xCache mean time = 13s

Athena Run Time:

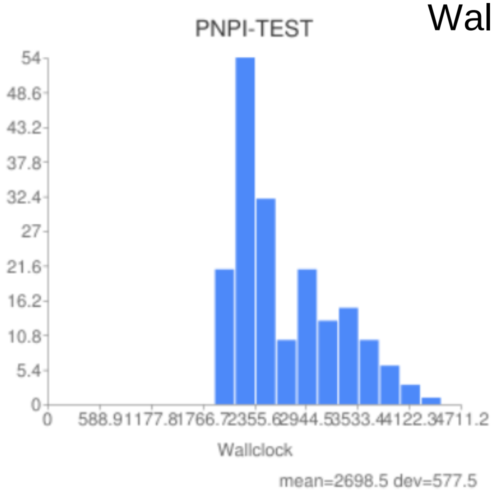
Direct mean time = 2111s ± 46s

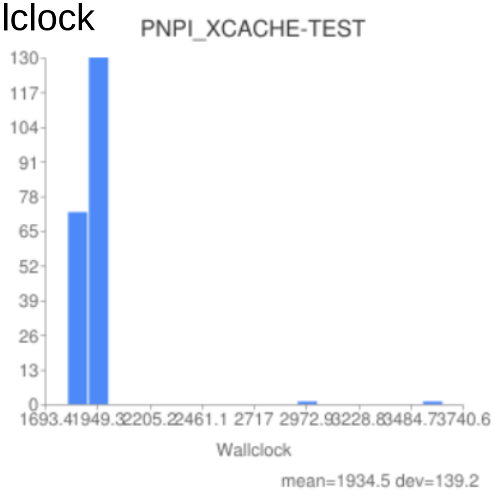
xCache mean time = 1856s ± 22s

Difference ~ 255s, ~12%

HammerCloud test results - N20146370 from Template 1099 (copy2scratch)

Wallclock



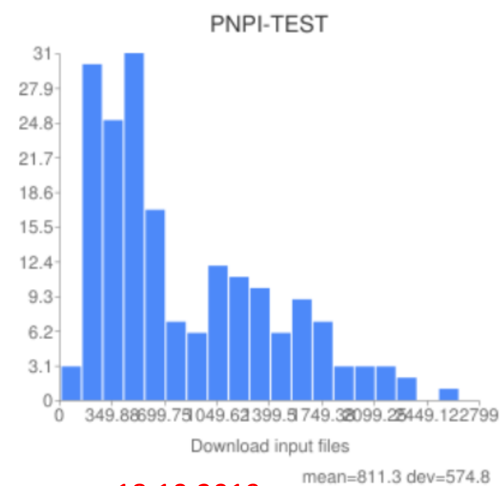


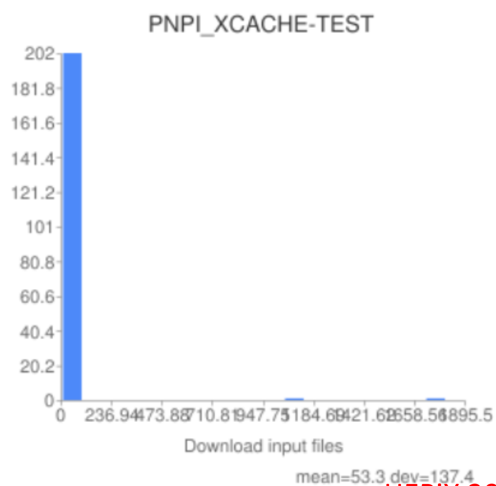
Wallclock:
Direct mean time = 2698s ± 577s
xCache mean time = 1934s ± 139s
Difference ~ 770s, ~30%

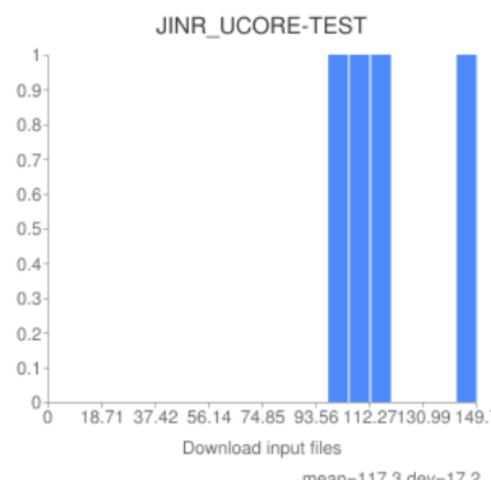
Download input files time:
Direct mean time = 811s ± 574s
xCache mean time = 53s ± 137s
Difference ~ 770s, ~95%

Local (JINR) = 117s ± 17s

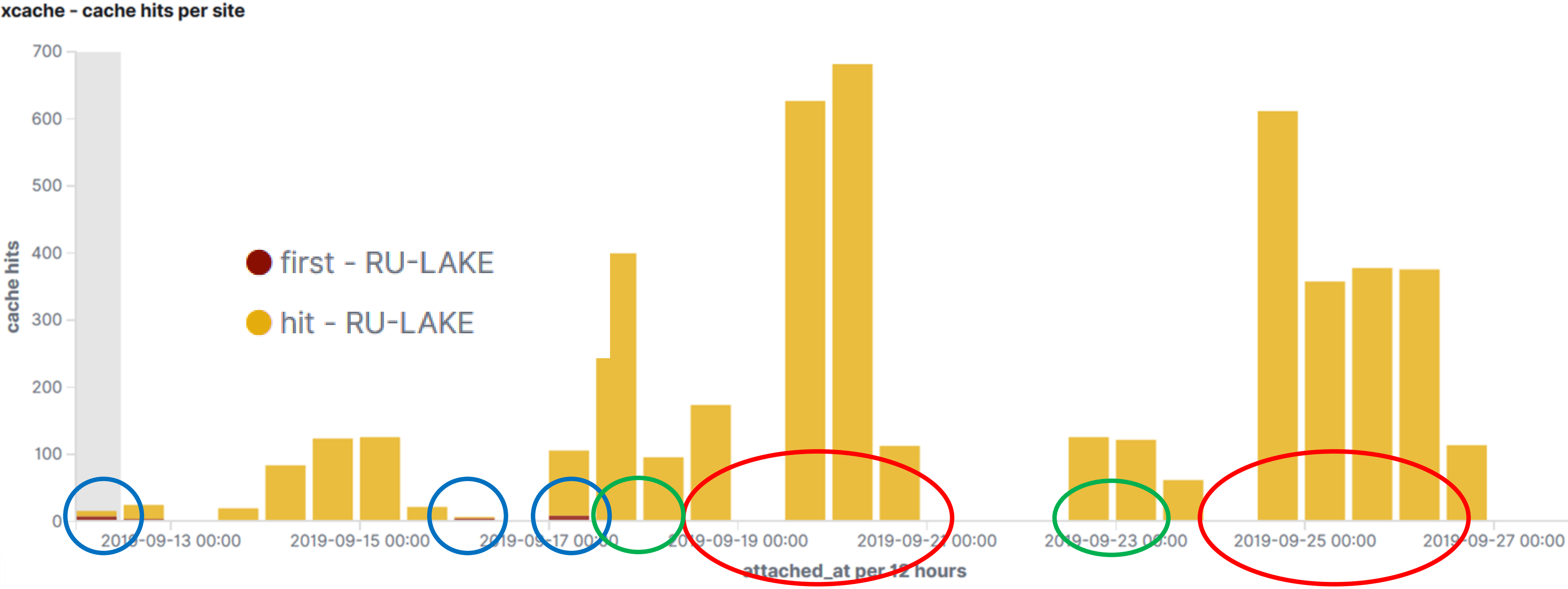
Download input file







xCache Monitoring in Kibana

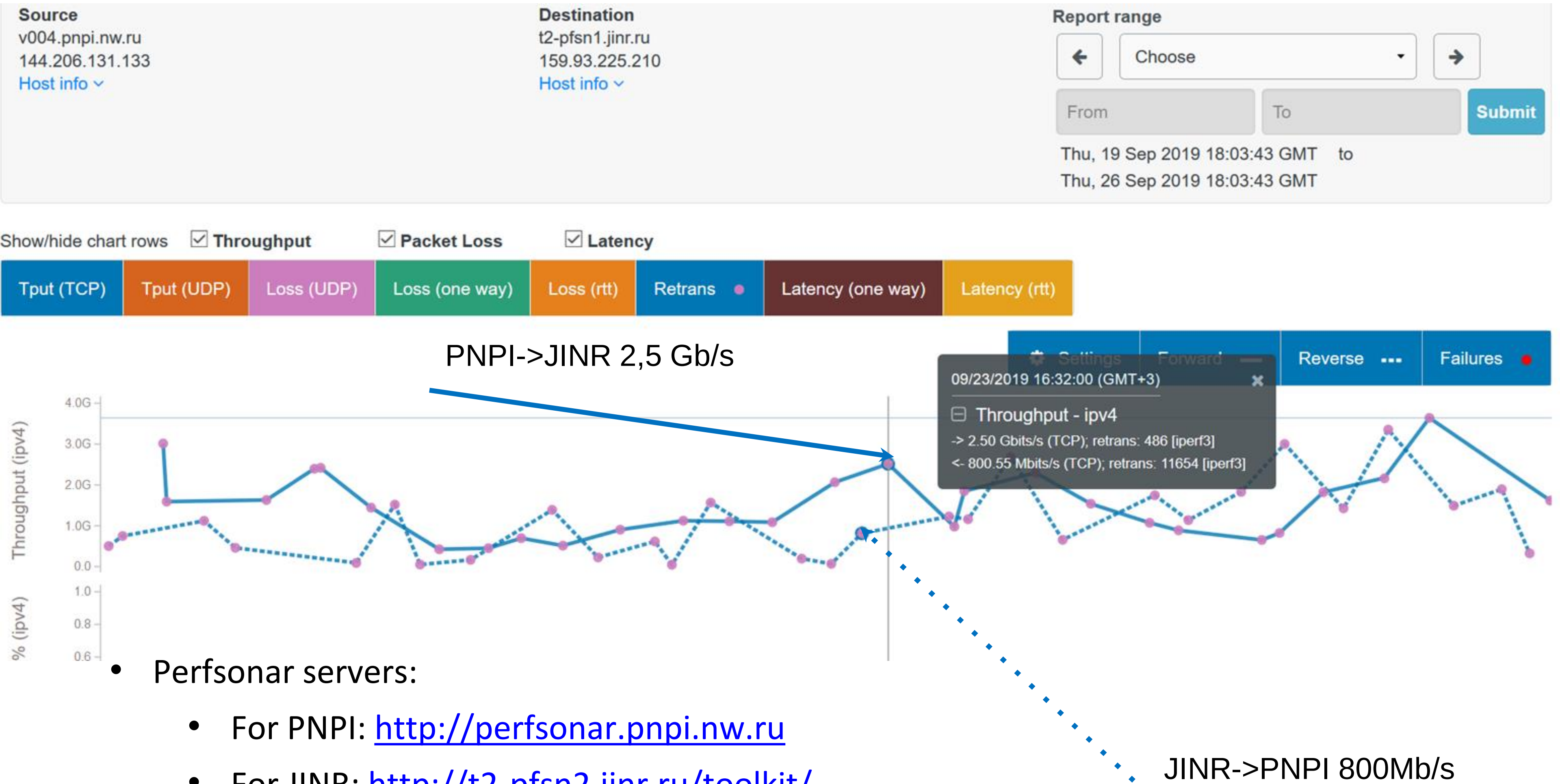


First access to files

Activity of HC tests

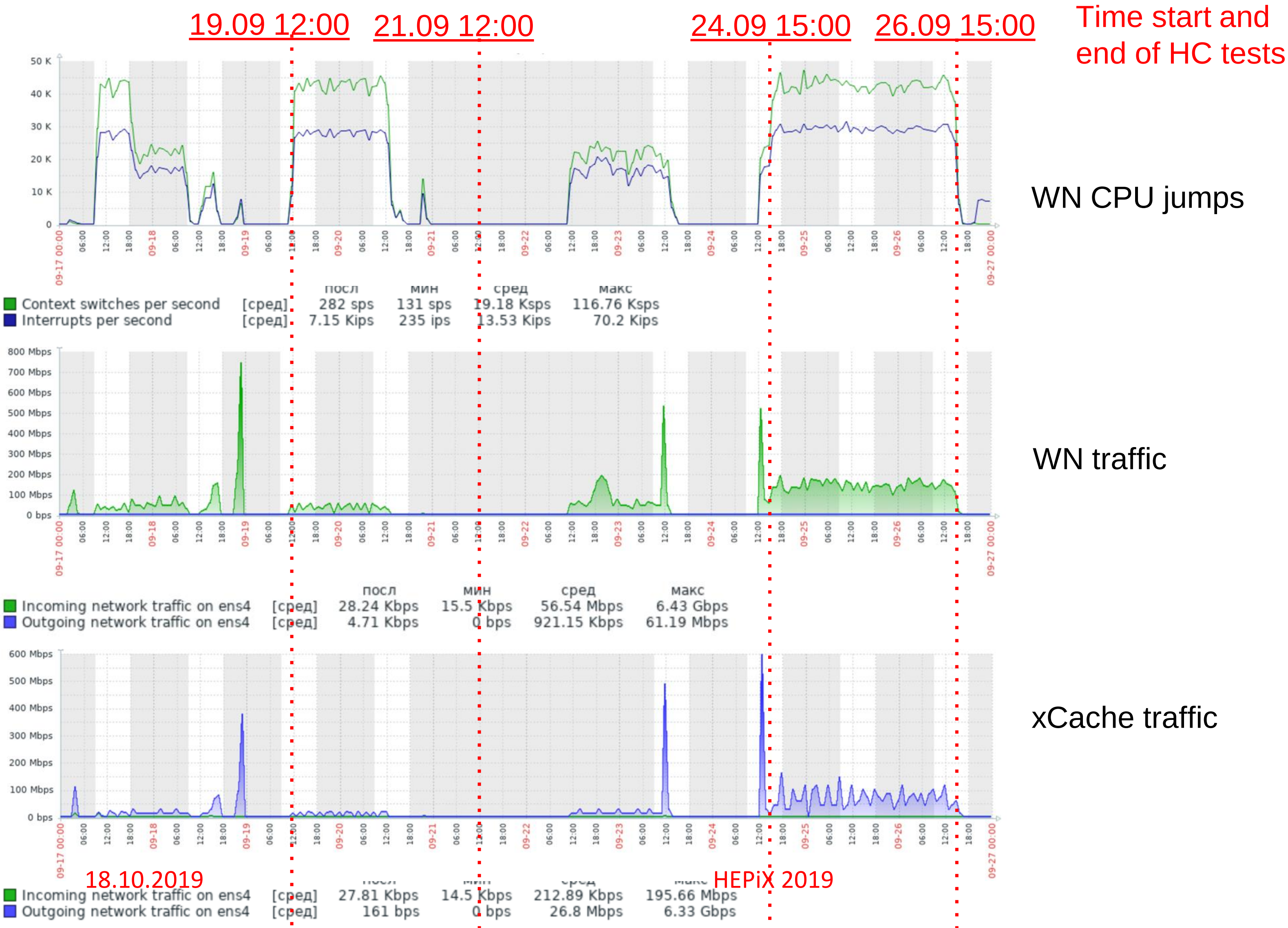
Activity of synthetic tests

PerfSonar 19.09-25.09



- Perfsonar servers:
  - For PNPI: http://perfsonar.pnpi.nw.ru
  - For JINR: http://t2-pfsn2.jinr.ru/toolkit/
    http://t2-pfsn1.jinr.ru/toolkit/

Zabbix monitoring at PNPI

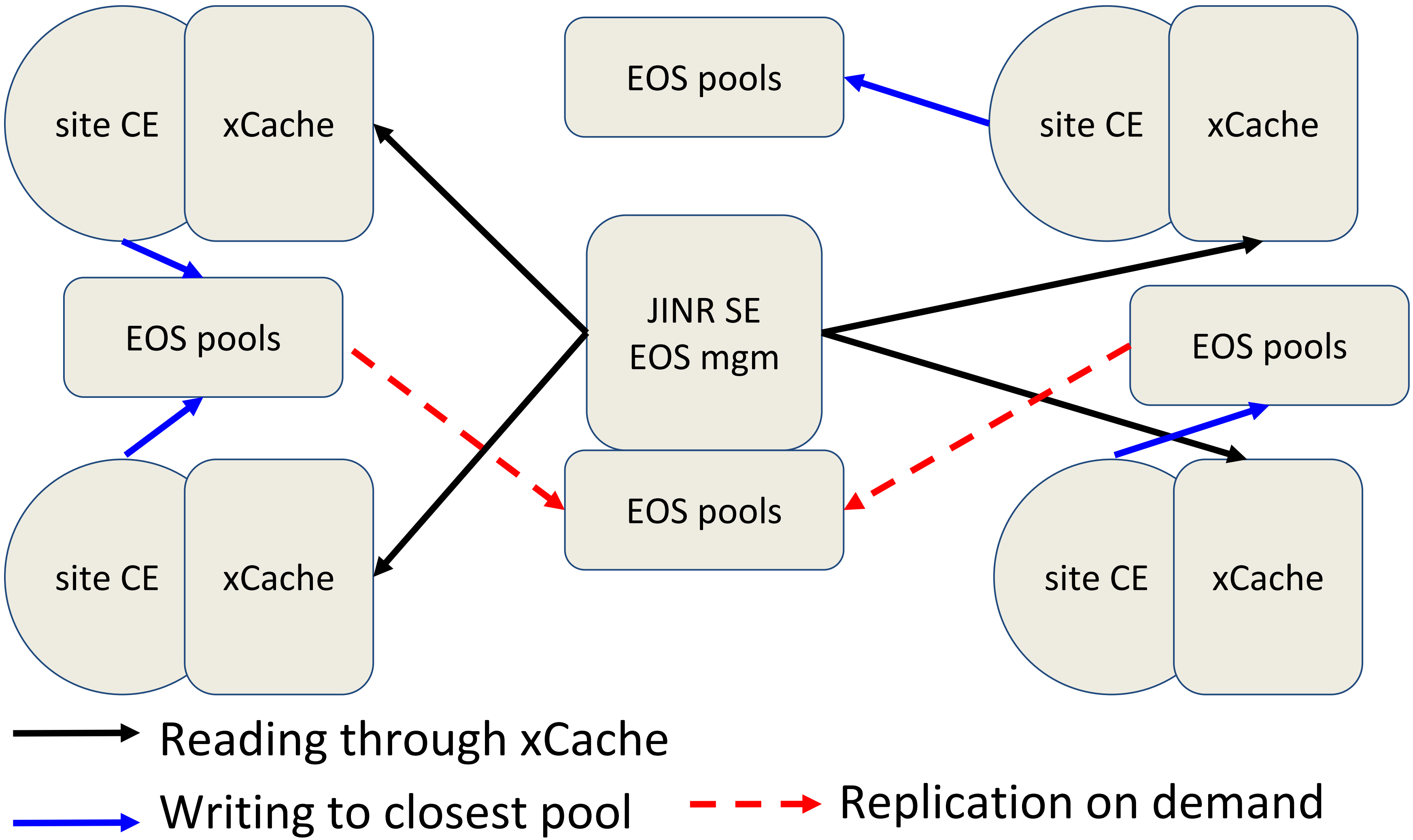


Results of xCache tests

1. Basic stuff works as expected (as usual ☺)
2. Result of synthetic tests demonstrate up to 95% gain in time for a file copy if it is done repeatedly
3. Results of HC tests demonstrate 30% gain in time for “copy2scratch” and 12% gain in time for “Direct access”

We can see correlation of monitoring data from different sources, but we need unified monitoring covering Perfsonar, Kibana, BigPanda monitoring, etc.

Russian Data Lake Phase 2 (2020 – 2021)



Summary

Data Lake R&D project was launched in Russia as a continuation of the successful Federated Data Storage project. Production-grade computing resources and 10–100 Gbps network connectivity with low latency will be used to prototype data lake.

We have a feasible plan for the first two phases of the project:

1) Consolidation of monitoring, better understanding of xCache control, e.t.c.
2) Expansion to other Russian sites with production resources and scattered storages. Continuation of monitoring and testing of the infrastructure.

Acknowledgements

Thank you to all who contributed materials, discussions, use cases and so on.

This work was funded in part by the Russian Science Foundation under contract No. 19-71-30008 (research is conducted in the Plekhanov Russian University of Economics)