NEC'2019



XXVII International Symposium on Nuclear Electronics & Computing

Montenegro, Budva, Becici, 30 September - 4 October 2019

Towards Russian National Data Lake Prototype

A. Alekseev, S. Campana, X. Espinal, *A. Kiryanov*, A. Klimentov, V. Mitsyn, A. Zarochentsev

Petersburg Nuclear Physics Institute

NATIONAL RESEARCH CENTER "KURCHATOV INSTITUTE"

St Petersburg University

CERN

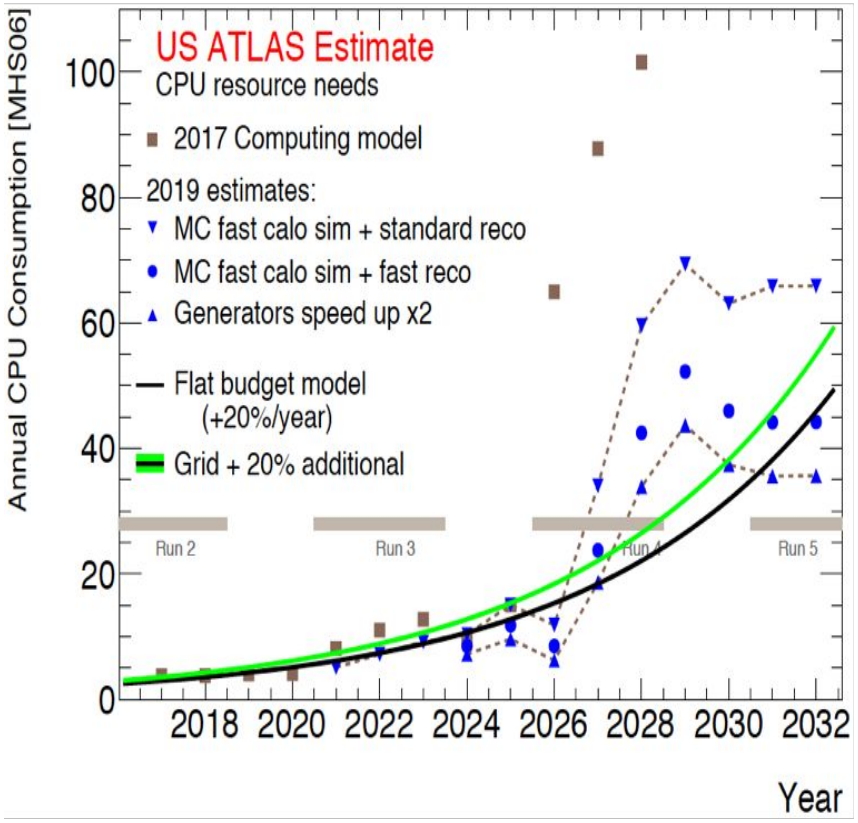
JINR

Outline

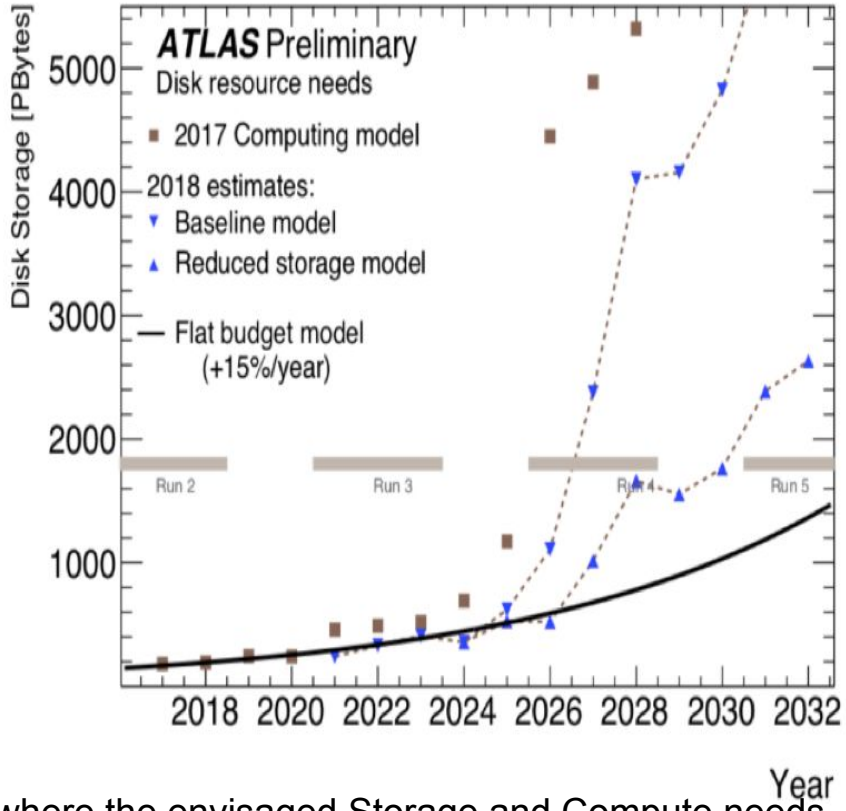
- HL-LHC Computing Challenges
- DOMA Project and Data Lake R&D
- Requirements to Data Storage Infrastructure
- Russian Data Lake R&D Project
- Summary and plans

The High Luminosity LHC Challenge

Growth in CPU Needed



Growth in Disk Storage Needed



- 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

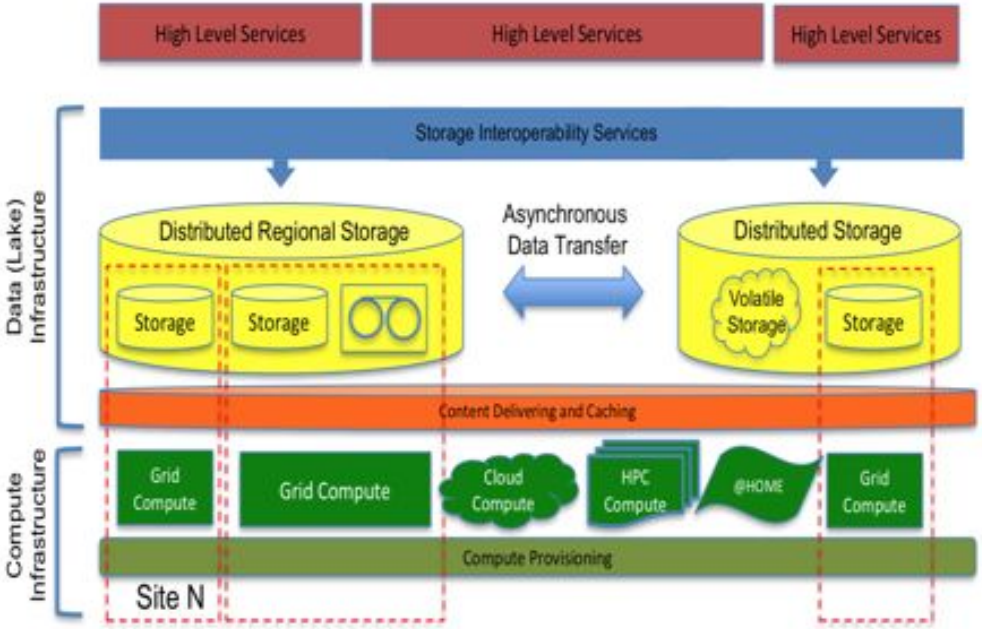
Requirements for a future data storage infrastructure

WLCG has defined the following implementation requirements:

- Common namespace and interoperability
- Coexistence of different QoS
- Geo-awareness
- File transitioning based on namespace rules
- File layout flexibility
- Distributed redundancy
- Fast access to data, latency compensation
  - File R/W cache
  - Namespace cache

Data Lake Concept

- Our sites are linked with (ever higher) high-bandwidth networking
  - We can expect **~100x bandwidth growth** by HL-LHC time
- **Data lakes**: integrated consolidation of distributed storage (and compute) facilities, leveraging high-bandwidth networks



- Data lake encompasses facilities with several levels of storage
  - **Tape**, at a relatively limited number of sites
  - **Standard disk**, at large storage repositories and smaller caches
  - Fast SSD '**edge cache**' for the hottest data
  - Should be able to **place data optimally** based on (dynamic) need

Federated Storage Milestones

2008: Distributed dCache for NDGF

2010: AAA – CMS Federated Storge

2010: FAX – ATLAS Federated Storge

2013: CERN distributed T0 – 2 computing centres 1200 km apart with 50:50 distribution of EOS-managed disk resources

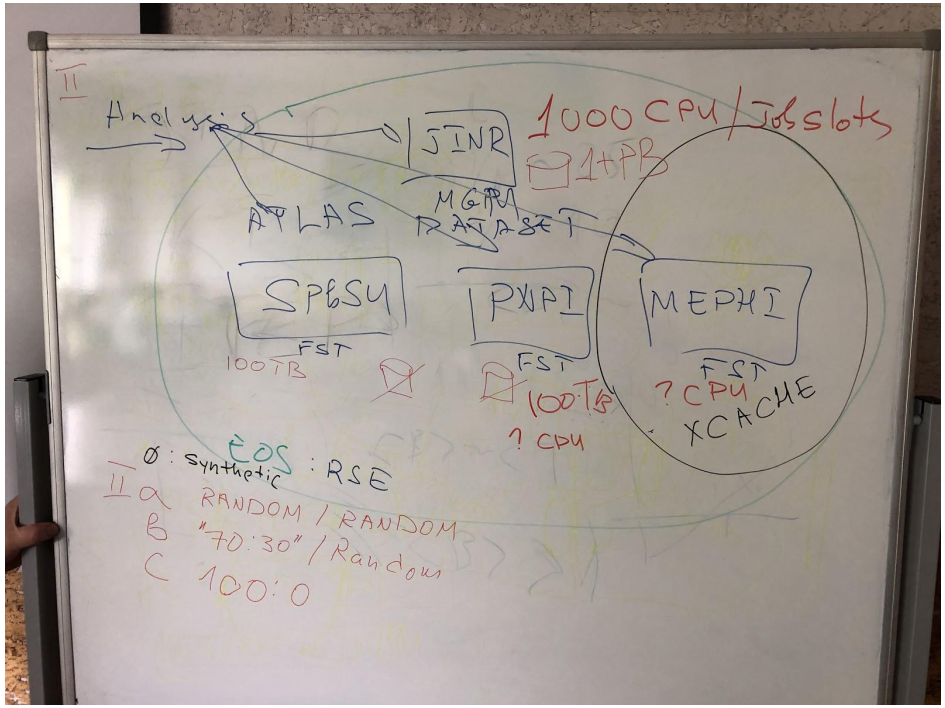
2015: Russian Federated Data Storage prototype – 8 centres, 2 major locations (SPb and Moscow), EOS & dCache

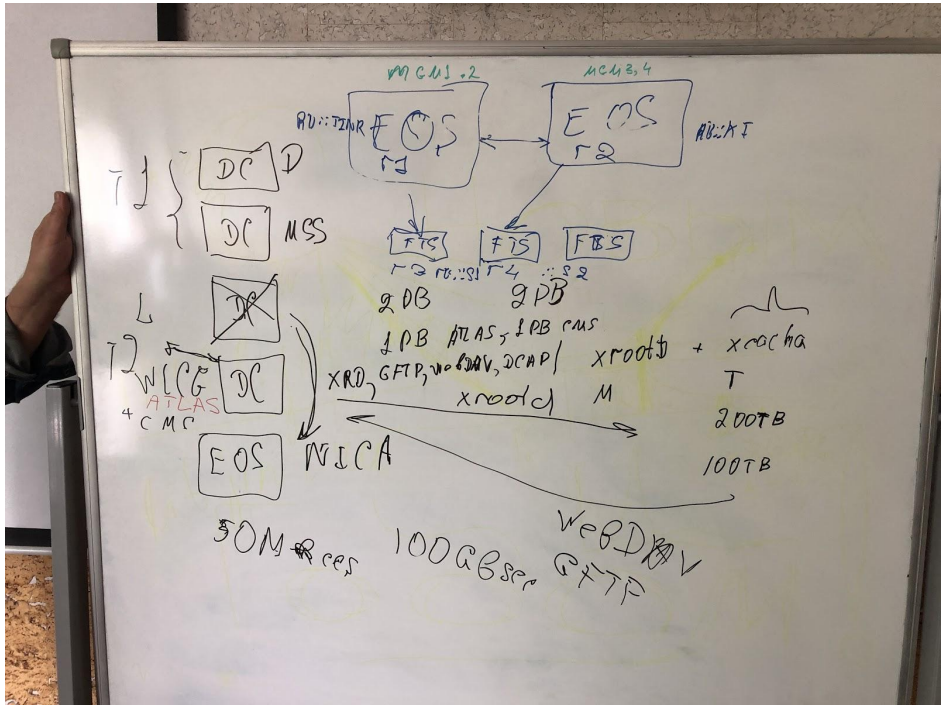
2018: ATLAS/Google “Data Ocean” project – cloud computing can offer attractive solutions and we can learn from industry leaders

2018: EULake – many centres at different locations (CERN, Russia, Spain, Netherlands, Australia, UK)

2019: Russian DataLake

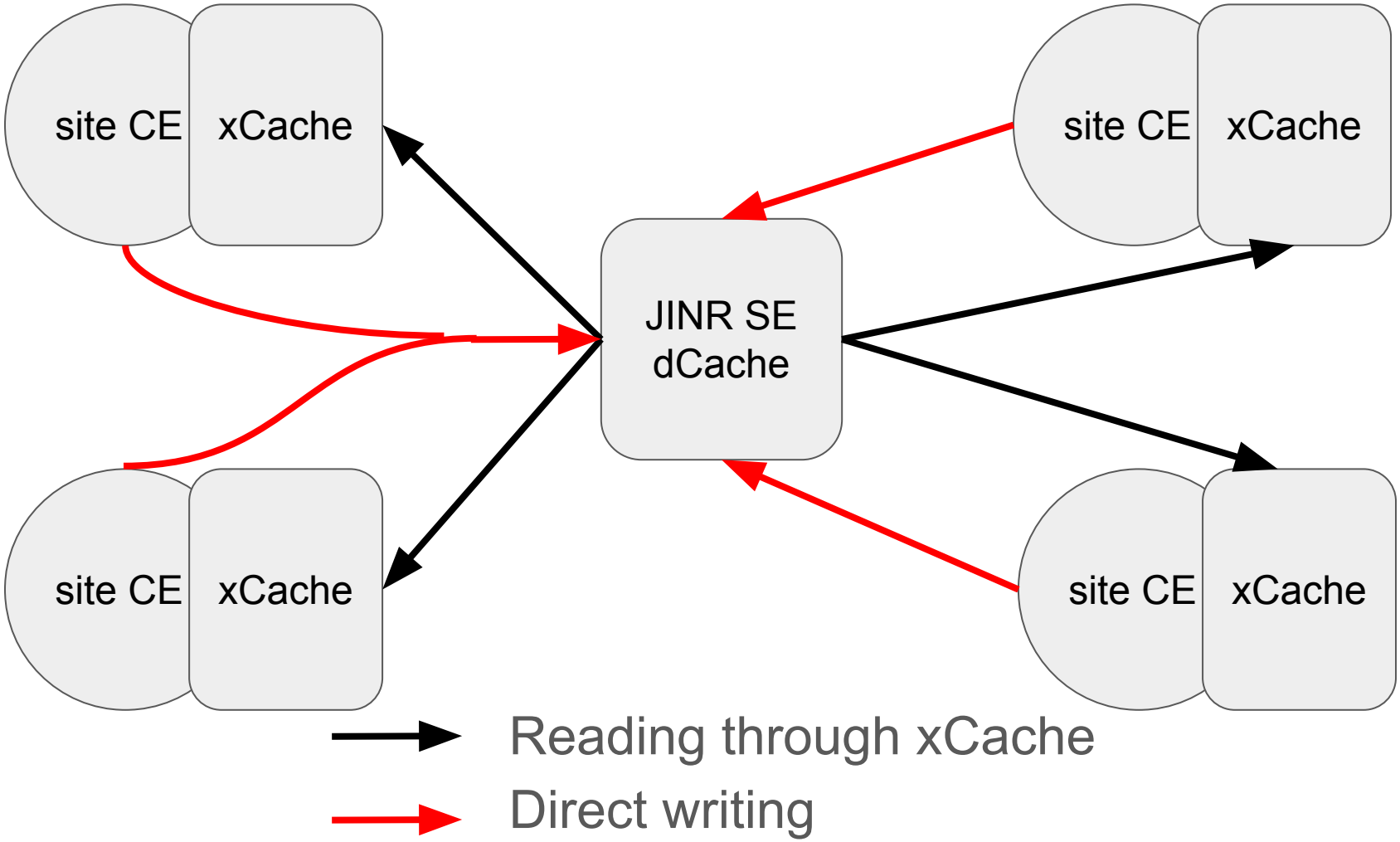
Russian DataLake R&D



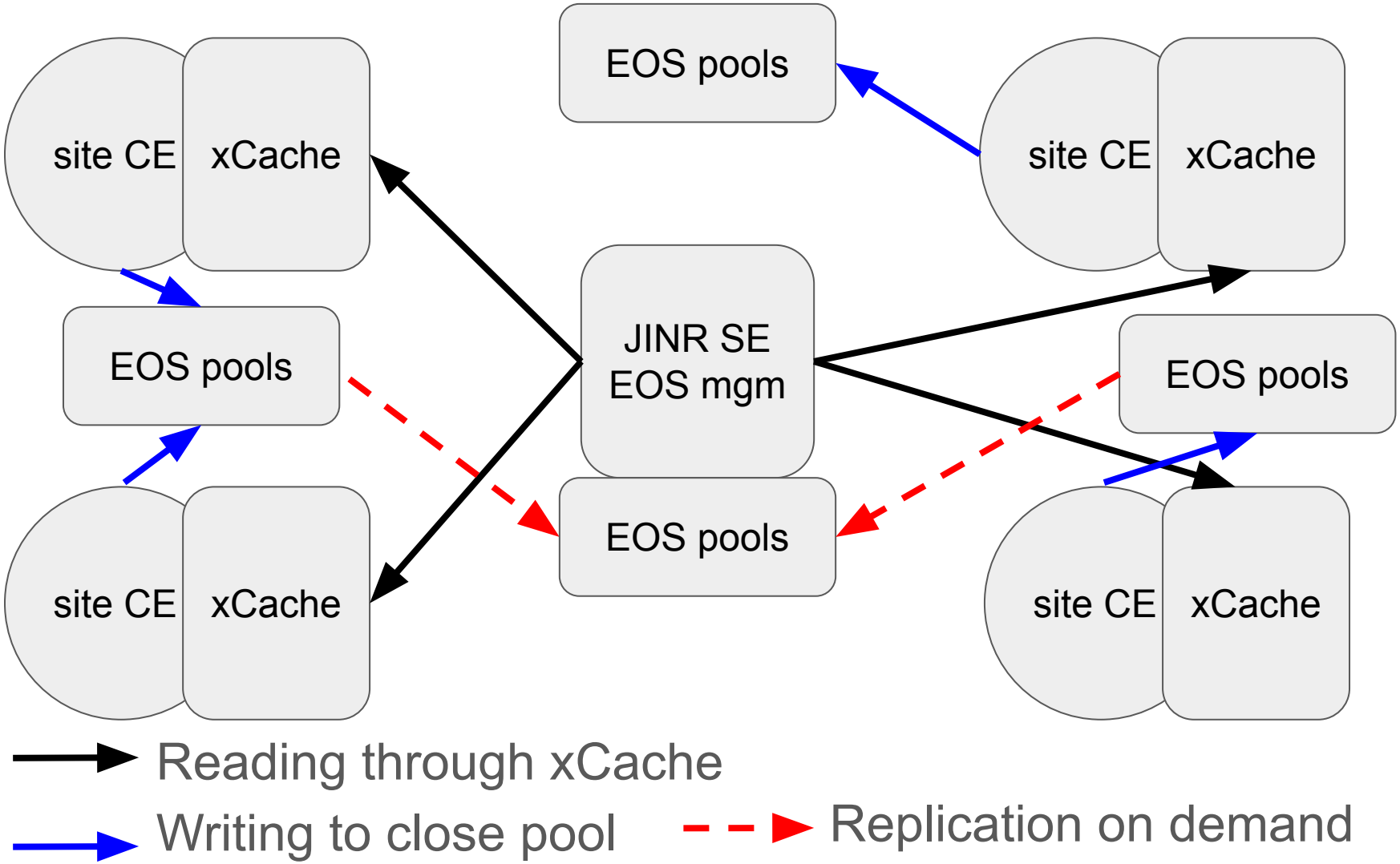


One cannot properly conduct an R&D without a messy whiteboard

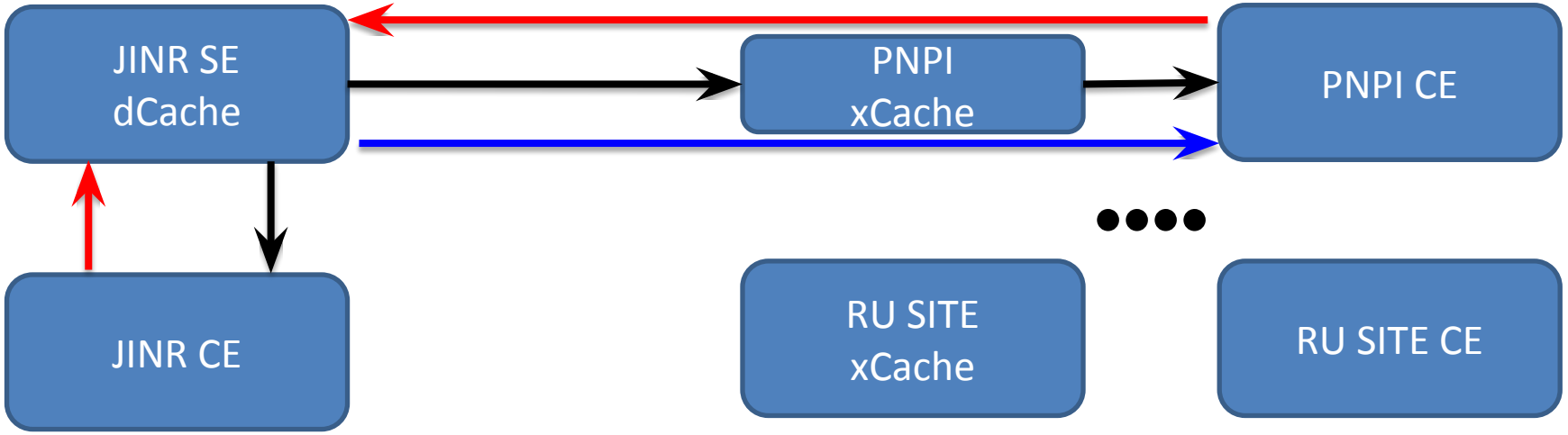
Russian DataLake Phase 1 (2019 Prototype)



Russian DataLake Phase 2 (2020-2021)



Planning Russian DataLake Phase I Prototype Tests



- Atlas standard tests through HammerCloud
- Synthetic tests from Worker Nodes
  - Manually
  - Through Cream-CE

Reading through xCache

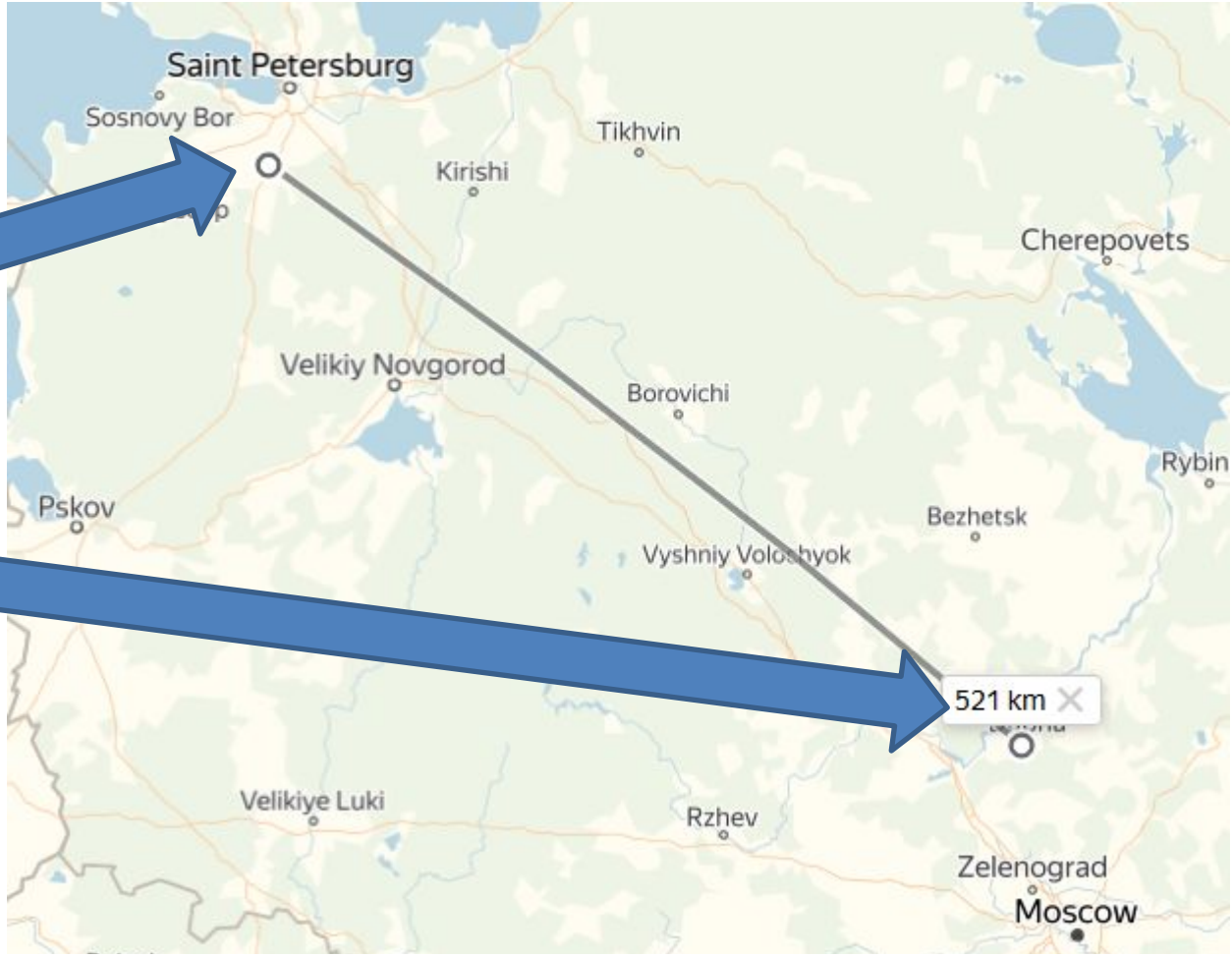
Direct reading

Direct writing

Participating Sites

Petersburg Nuclear Physics Institute

JINR



Authorization

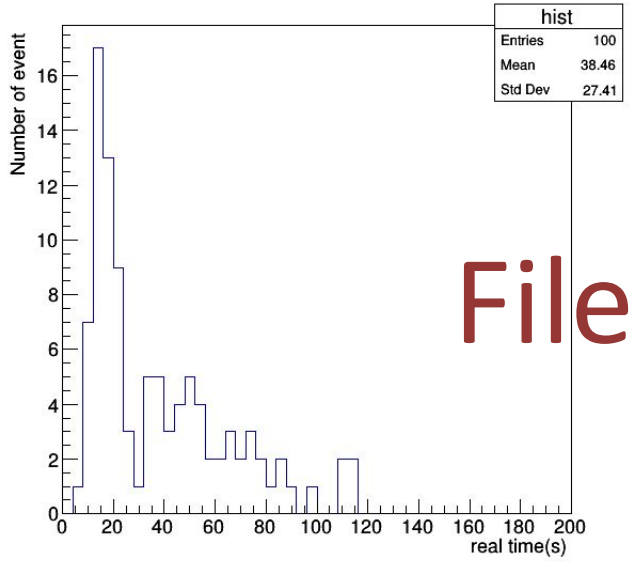
- PNPI xCache ➜ JINR SE: GSI authorization by local gridmapfile on JINR SE
- PNPI WN ➜ PNPI xCache: GSI authorization by VOMS (ATLAS)
- PNPI UI ➜ JINR CE, PNPI CE (for local tests): GSI authorization by VOMS (ALICE & ATLAS)
- Hammer Cloud ➜ ALL: GSI authorization by VOMS (ATLAS)
- An external library 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

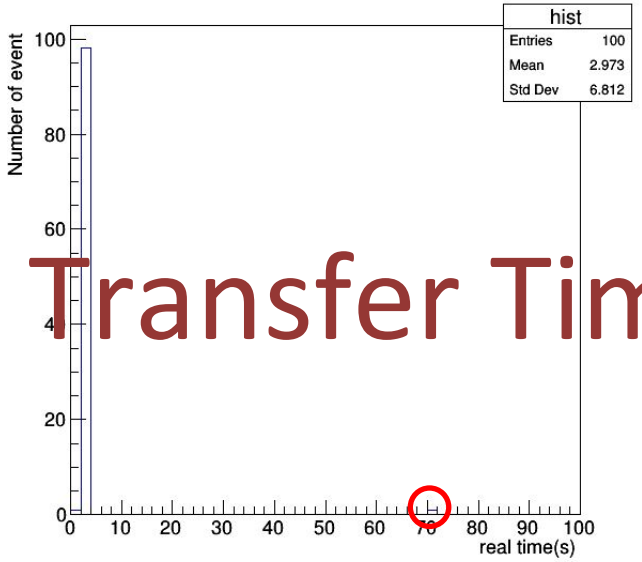
- Worker Node @ JINR: 8 cores, Xeon E5420, 16GB RAM, 8.74 HEP-SPEC06 per Core
- Worker Node @ PNPI: 8 cores, Xeon E5-2680, 32GB RAM (VM), ~11 HEP-SPEC06 per Core
- Local network @ JINR (SE<->CE) 1Gb/s
- Local network @ PNPI (SE<->CE) 10Gb/s
- Network IPv4,6 JINR ➜ PNPI: Latency ~5ms
- Network IPv4,6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

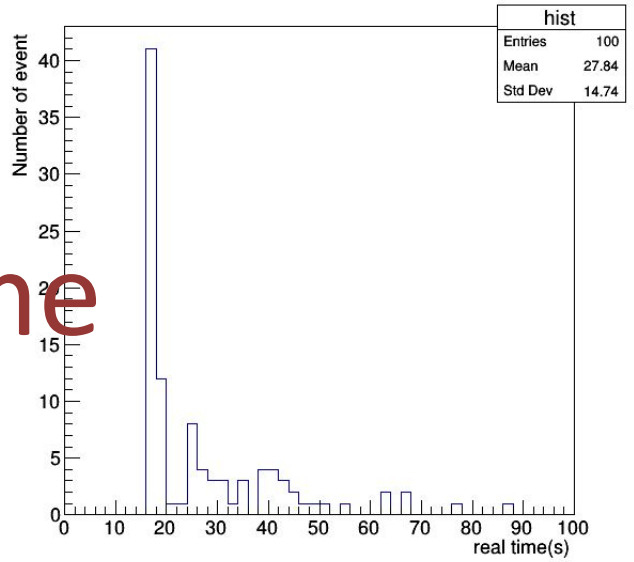
JINR-SE->PNPI-CE



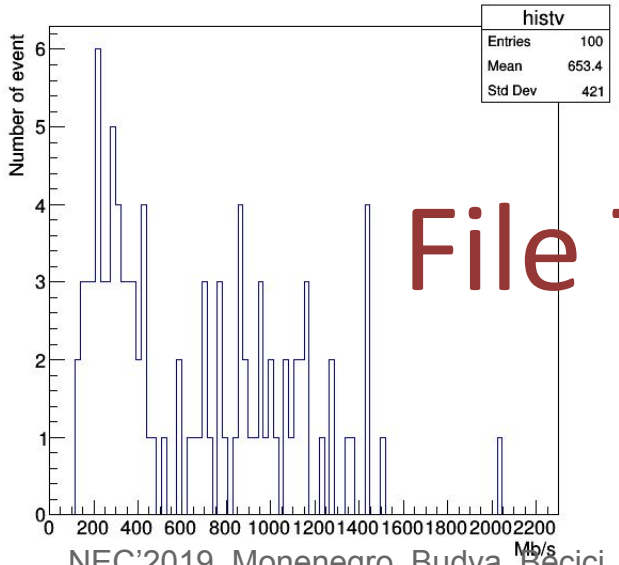
JINR-SE->xCache->PNPI-CE

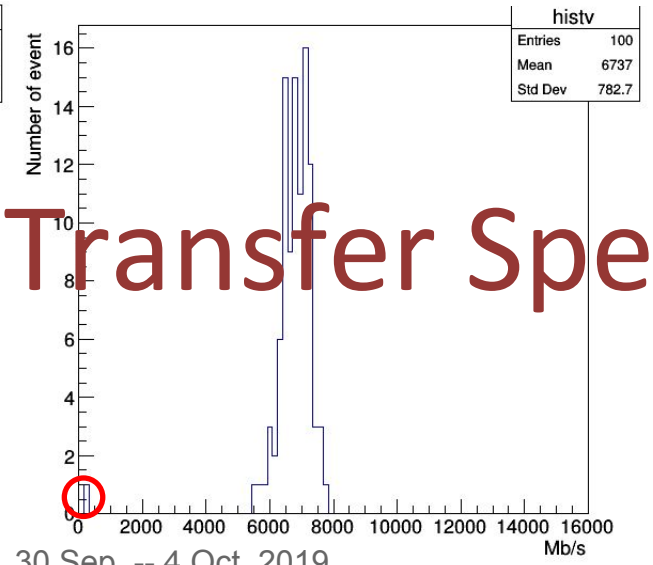


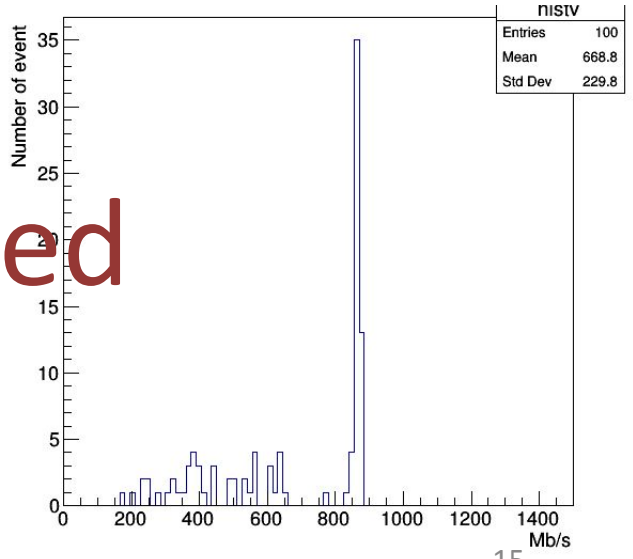
JINR-SE->JINR-CE



File Transfer Time







File Transfer Speed

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**

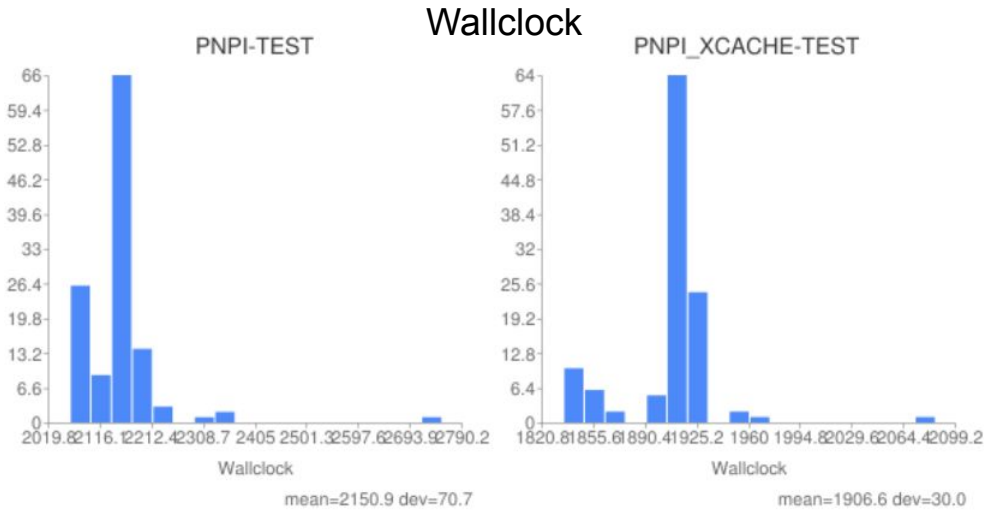
- Test number 20146370 from Template 1099 (copy2scratch)

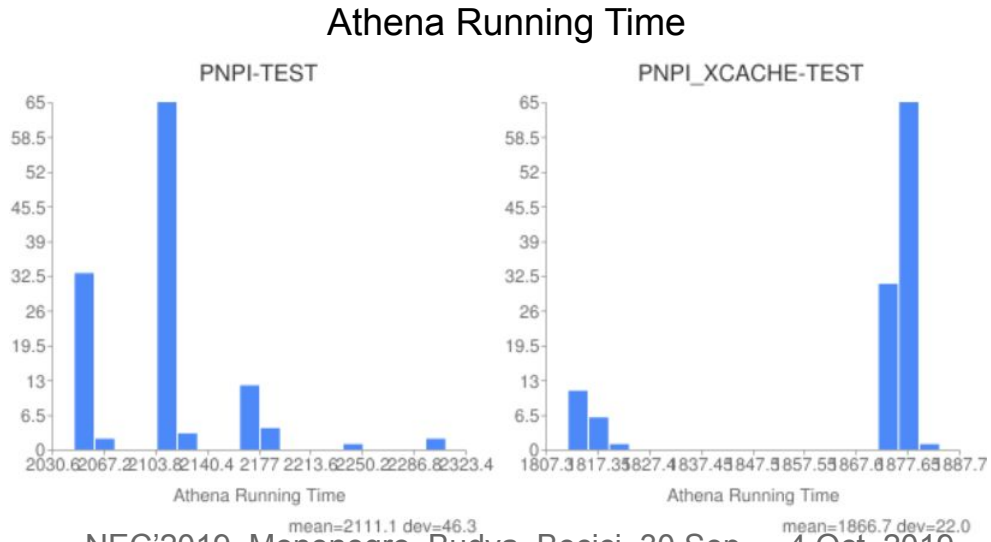
state	id	host	clouds	start time (CET)	end time (CET)	total jobs
completed	20146182	hammercloud-ai-11	RU_PROD	19/9/2019 12:00	21/9/2019 12:00	254

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)





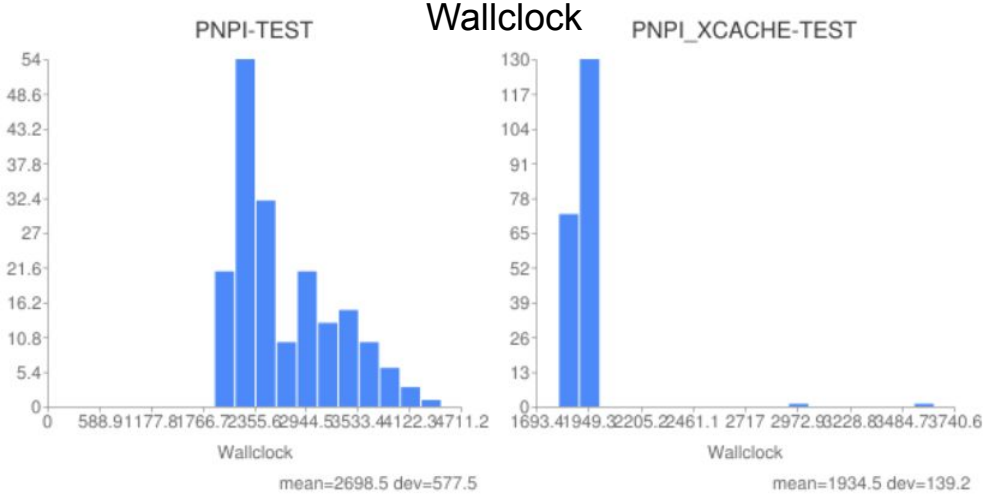
Wallclock:
Direct mean time = 2150s ± 70s
xCache mean time = 1906s ± 30s
Difference ~ 250s, ~12%

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

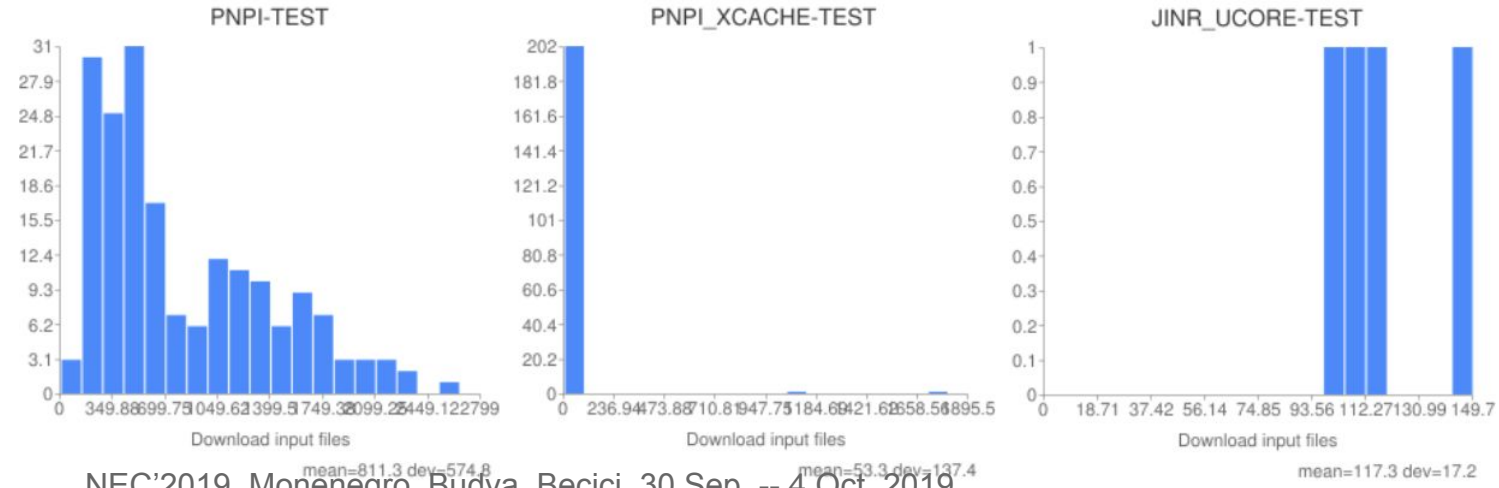


Wallclock:
Direct mean time = 2698s ± 577s
xCache mean time = 1934s ± 139s
Difference ~ 770s, ~30%

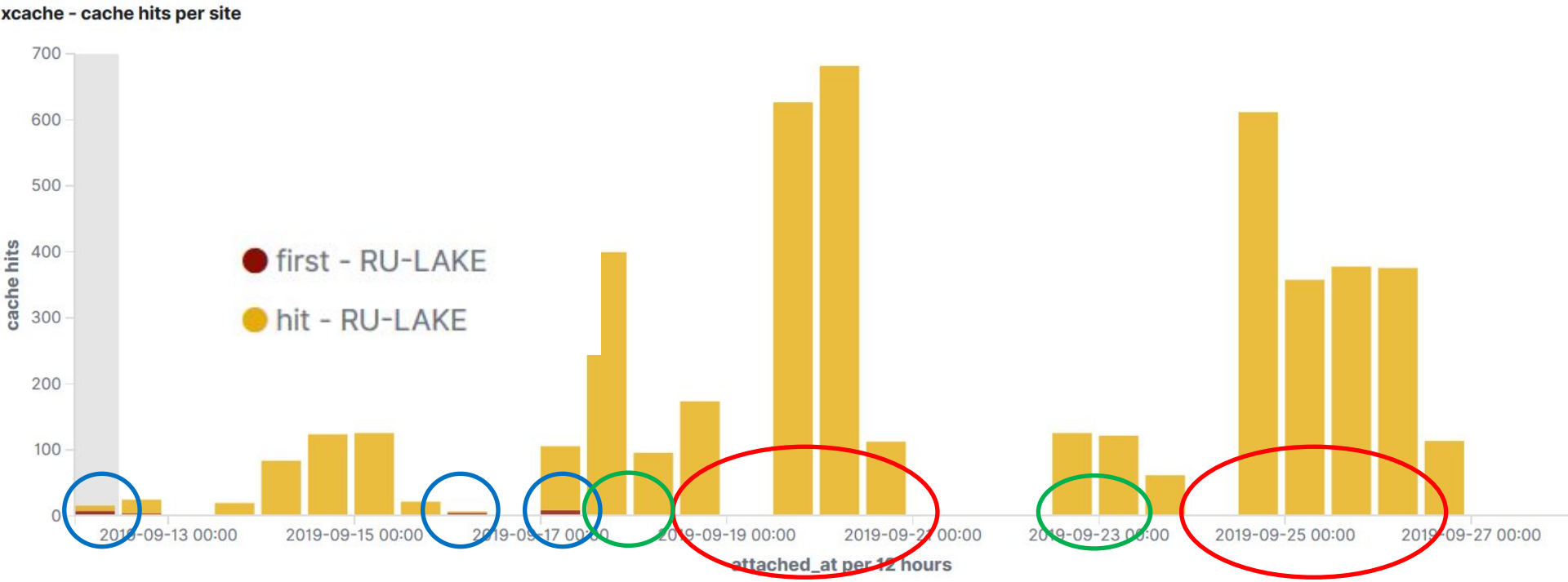
Download input files time:
Direct mean time = 811s ± 574s
xCache mean time = 53s ± 137s
Difference ~ 770s, ~95%

Local (JINR) = 117s ± 17s

Download input file

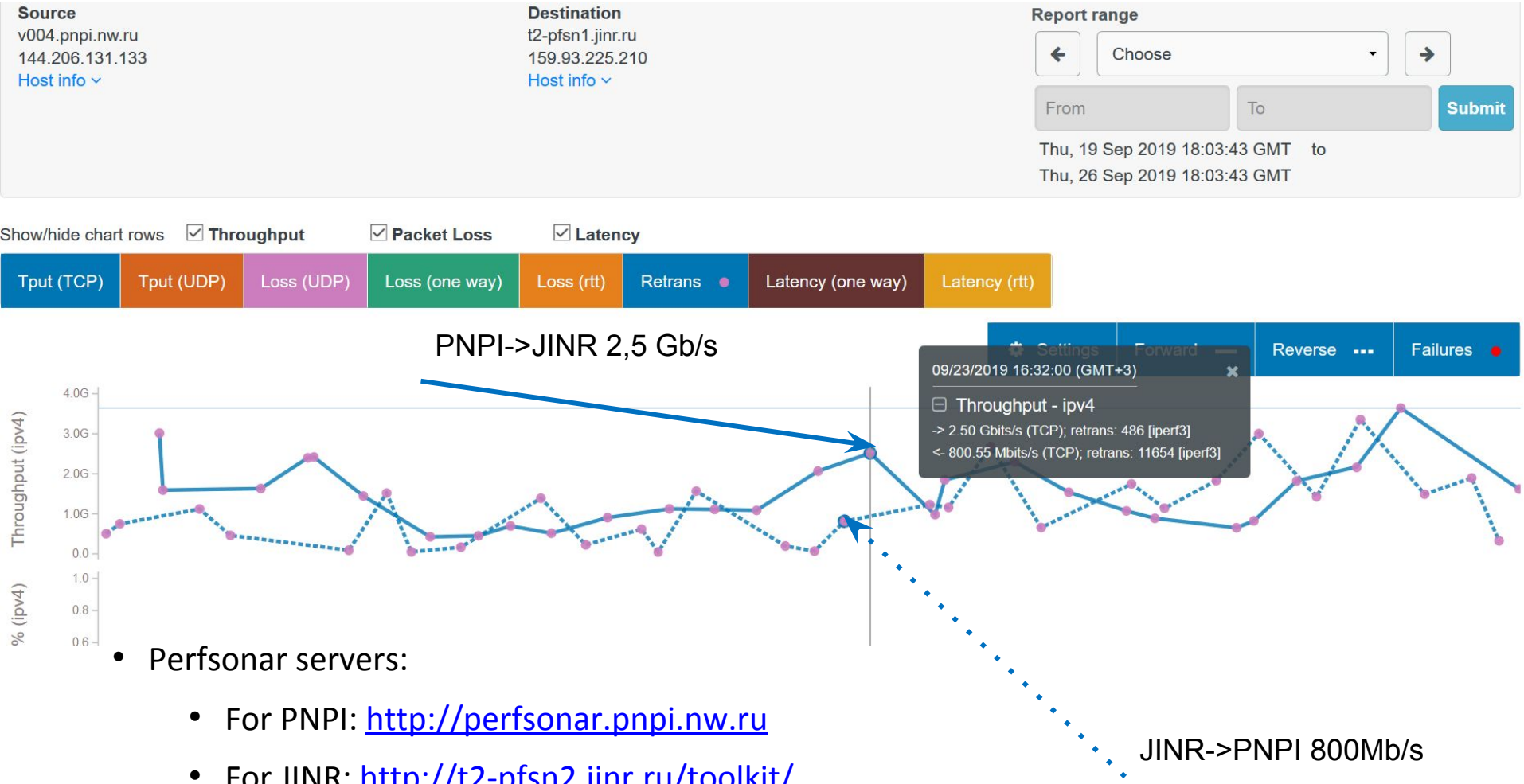


xCache Monitoring in Kibana



- 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/

Summary I

The first phase highlights :

- xCache is configured and works well
- PNPI tests are informative for the time being
- Result of synthetic tests demonstrate 95% gain in time for 100 file copy (if it is done repeatedly)
- Results of HC tests demonstrate 30% gain in time for “copy2scratch” and 12% gain in time for “Direct access”
- xCache and PerfSonar monitoring is available

Summary II

It is planned :

- To conduct scaling tests to/on other Russian sites
- To understand better how to control xCache (cleaning, etc)
- To have an unified monitoring from all sources - Perfsonar, kibana, BigPanda monitoring, etc.

Thanks!

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)

This work was supported by the NRC "Kurchatov Institute" (order No. 1571)