Evaluating Federated Data Infrastructure in Russian Academic Cloud for LHC experiments and Data Intensive Science

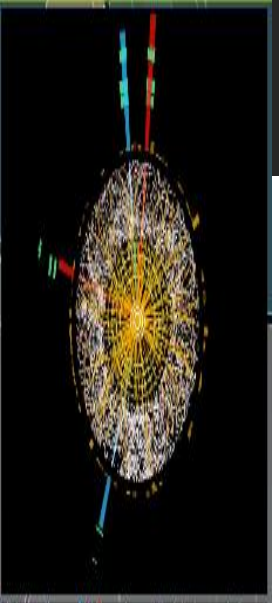
17th International workshop on Advanced Computing and Analysis Techniques in physics research (ACAT)

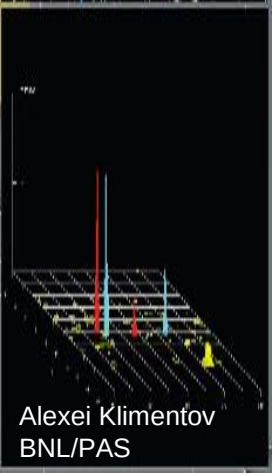
A.Klimentov, A.Kiryanov, D.Krasnopevtsev, A.Zarochentsev and P.Hristov

Brookhaven National Laboratory
National Research Center "Kurchatov Institute"
St.-Petersburg State University
Petersburg Nuclear Physics Institute
CERN

BROOKHAVEN NATIONAL LABORATORY

January 2016, Valparaiso, Chili

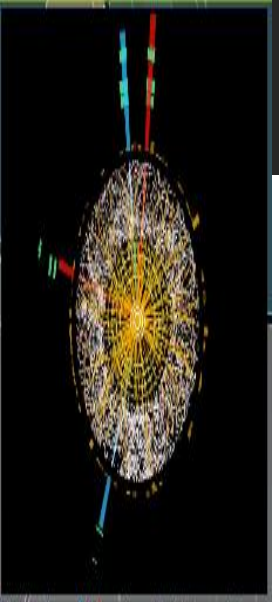


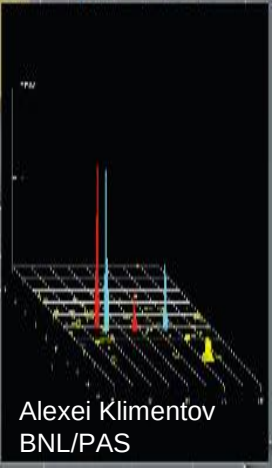


Main Topics

- **Original Motivation**
- **Requirements to Federated Data Storage**
- **Technology Choice**
- **Test infrastructure and methodic**
  - Synthetic tests
  - ATLAS and ALICE test suite
- **Preliminary results and future plans**

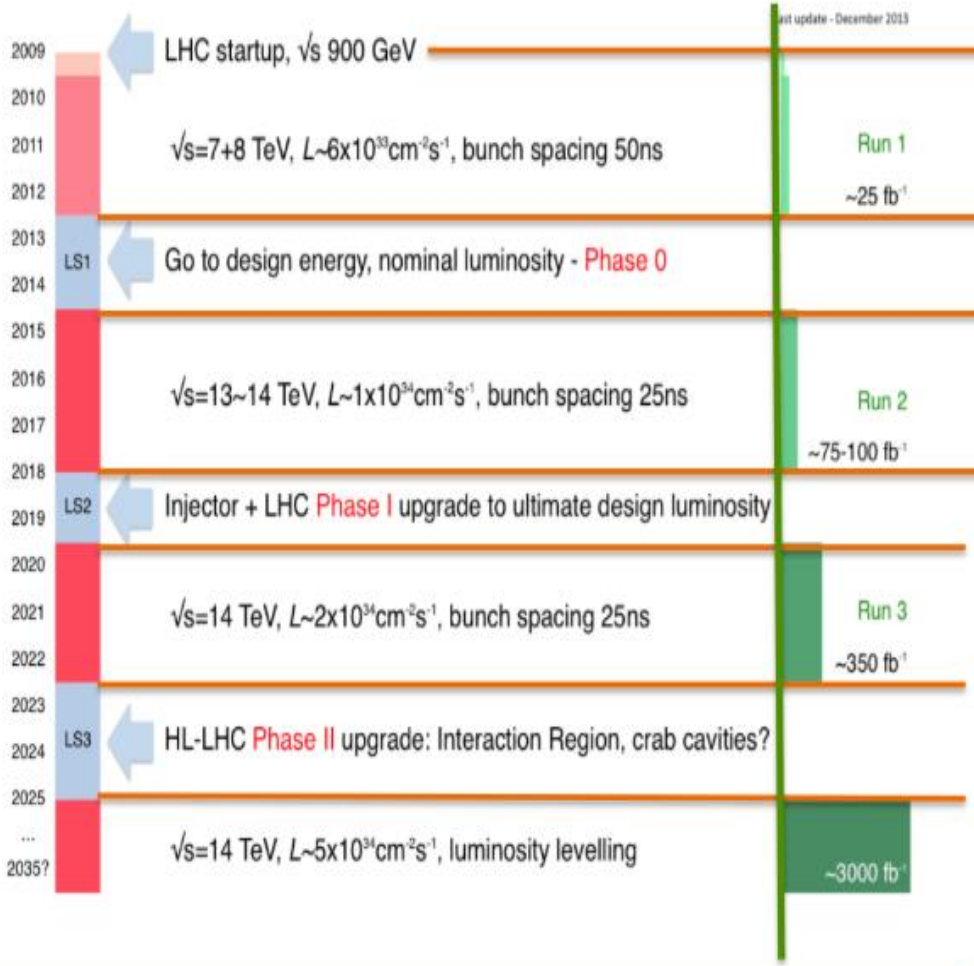
Alexei Klimentov
BNL/PAS





Alexei Klimentov
BNL/PAS

LHC Challenges. The Run2 and beyond



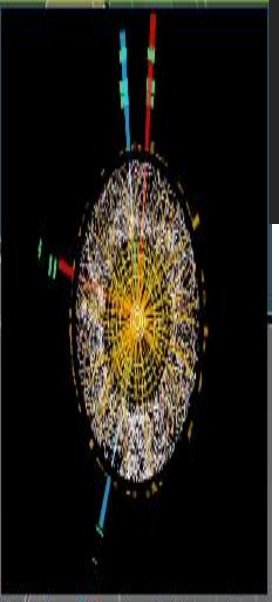
Trigger rate: ~400 Hz
Pile-up: ~20

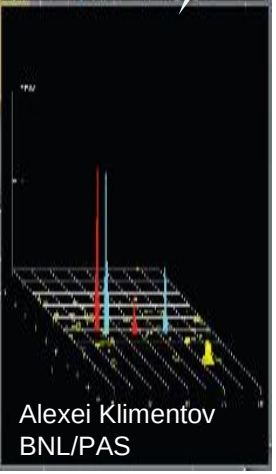
Trigger rate: ~1k Hz
Pile-up: ~40

**Resources constrained by "flat budget"
(no increase in funding for computing)**

A new detector

e.g. tracking, calorimeters





LHC Computing Scale of Needs

CPU needs per event

Run1

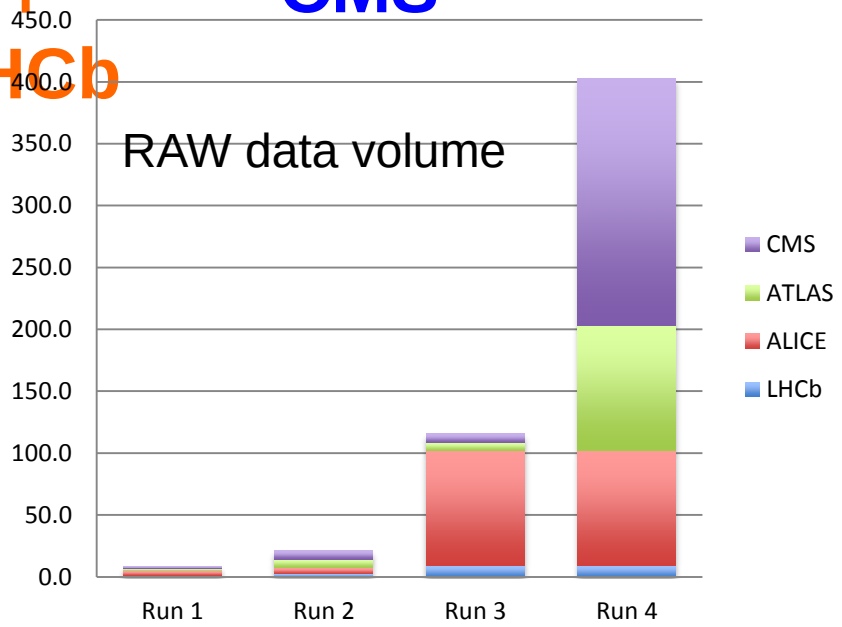
Run2

Run3

ALICE
+
LHCb

Run4

ATLAS
+
CMS



- CPU needs (per event) will grow with track multiplicity (pileup) and energy
- Storage needs are proportional to accumulated luminosity
- Grid resources are limited by funding

Alexei Klimentov
BNL/PAS

LHC Computing Challenges

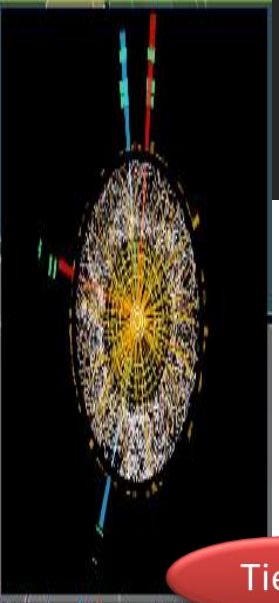
- **A lot of data in a highly distributed environment.**
  - Petabytes of data to be treated and analyzed
    - For example ATLAS managed data volume ~160 PB, distributed world-wide to O(100) computing centers and analyzed by O(1000) physicists
    - ATLAS Detector generates about 1PB of raw data per second
    - More than a hundred of computing centers had to work together
  - Dozens of complex applications
- **Very large international collaboration**
  - Hundreds Institutes and Universities from many countries
  - Thousands of physicists analyze the data

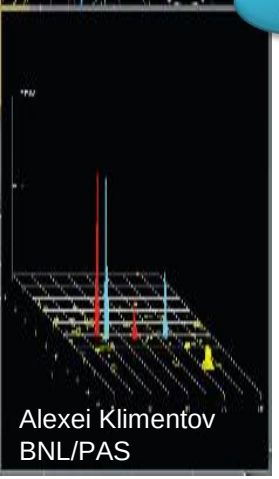
But the landscape has changed
Have we updated the maps ?

LHC Experiments use grid computing paradigm to organize distributed resources;

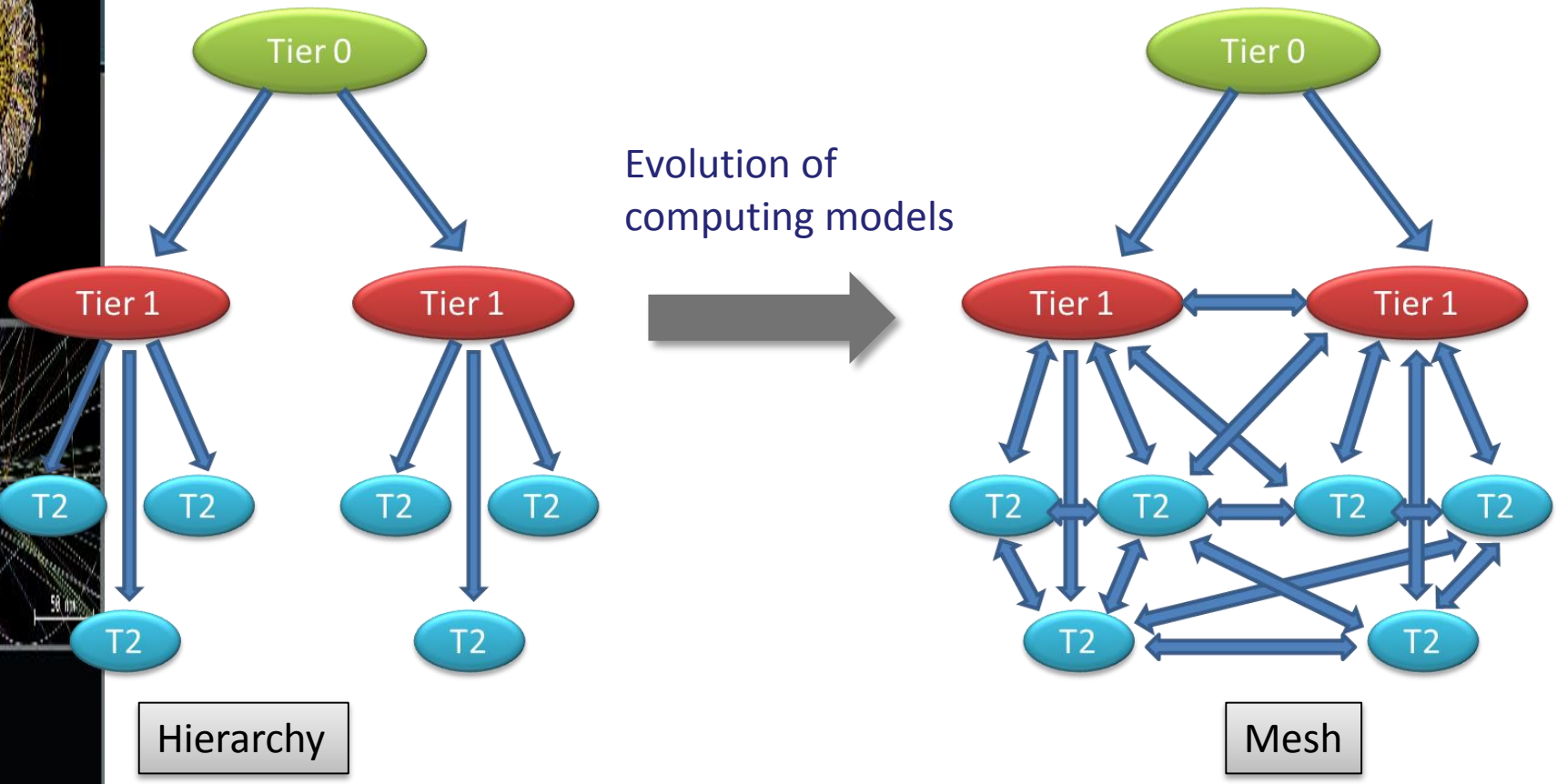
A few years ago Cloud Computing RnD projects were started to explore virtualization and clouds

Now we are evaluating how high-performance and super-computers can be used for data processing and analysis and disks federation and try to reduce complexity.





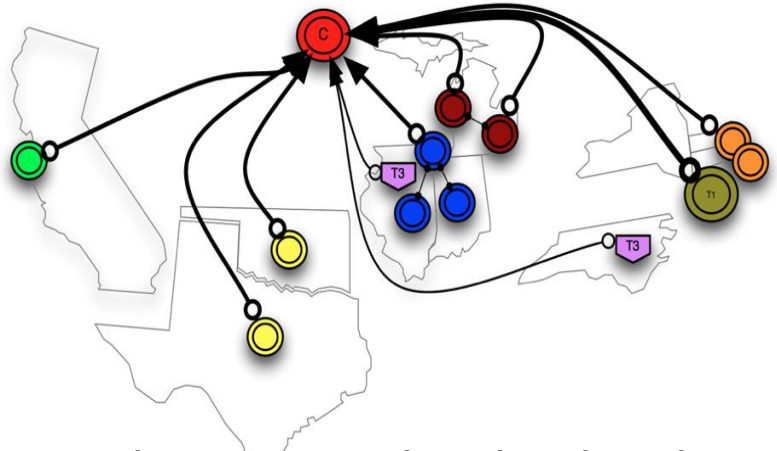
How to Face the Run2+ Challenges. Reducing Complexity.



- Network capabilities and data access technologies have significantly improved our ability to use resources independent of location
- Relaxing hierarchical model : Flat instead of Tiered Grid model

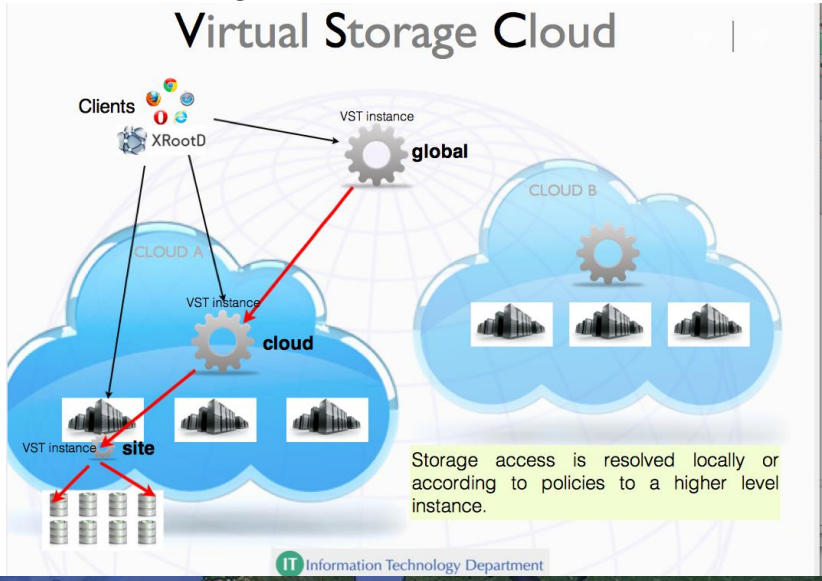
Alexei Klimentov
BNL/PAS

How to Face the Run2+ Challenges. Reducing Complexity. Cont'd.



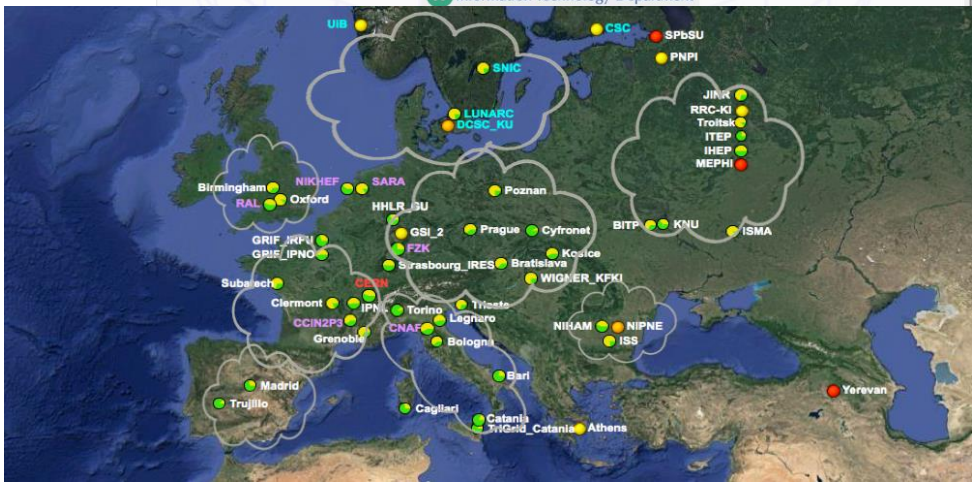
Xrootd ATLAS Federation (FAX)

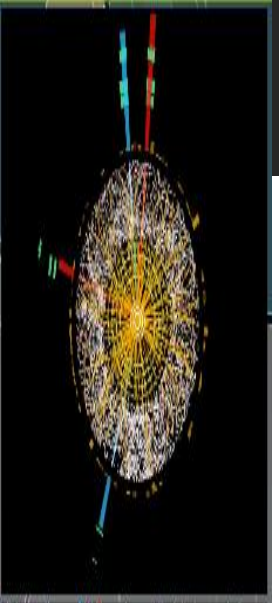
CERN-IT : site or cloud is represented by Virtual Storage Cloud Node

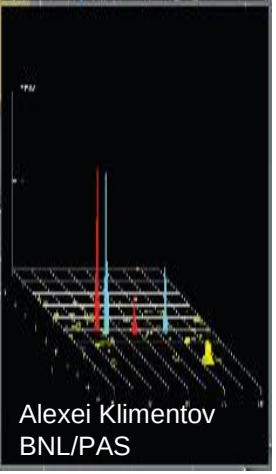


ALICE :

- Virtually joining together the sites based on proximity (latency) and network capacity into Regional Data Clouds
- Each cloud/region provides reliable data management and sufficient processing capability
- Dealing with handful of clouds/regions instead of the individual sites



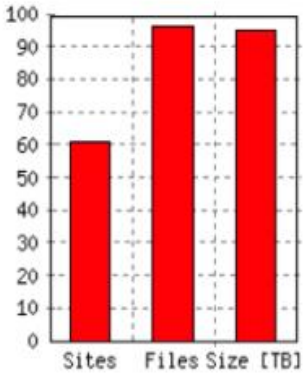




ATLAS Federation (FAX).

Remote data access: the Xrootd ATLAS Federation (FAX)

Goal reached ! ~100% data covered



We deployed a Federate Storage Infrastructure: all data accessible from any location

Increase resiliency against storage failures: FAILOVER

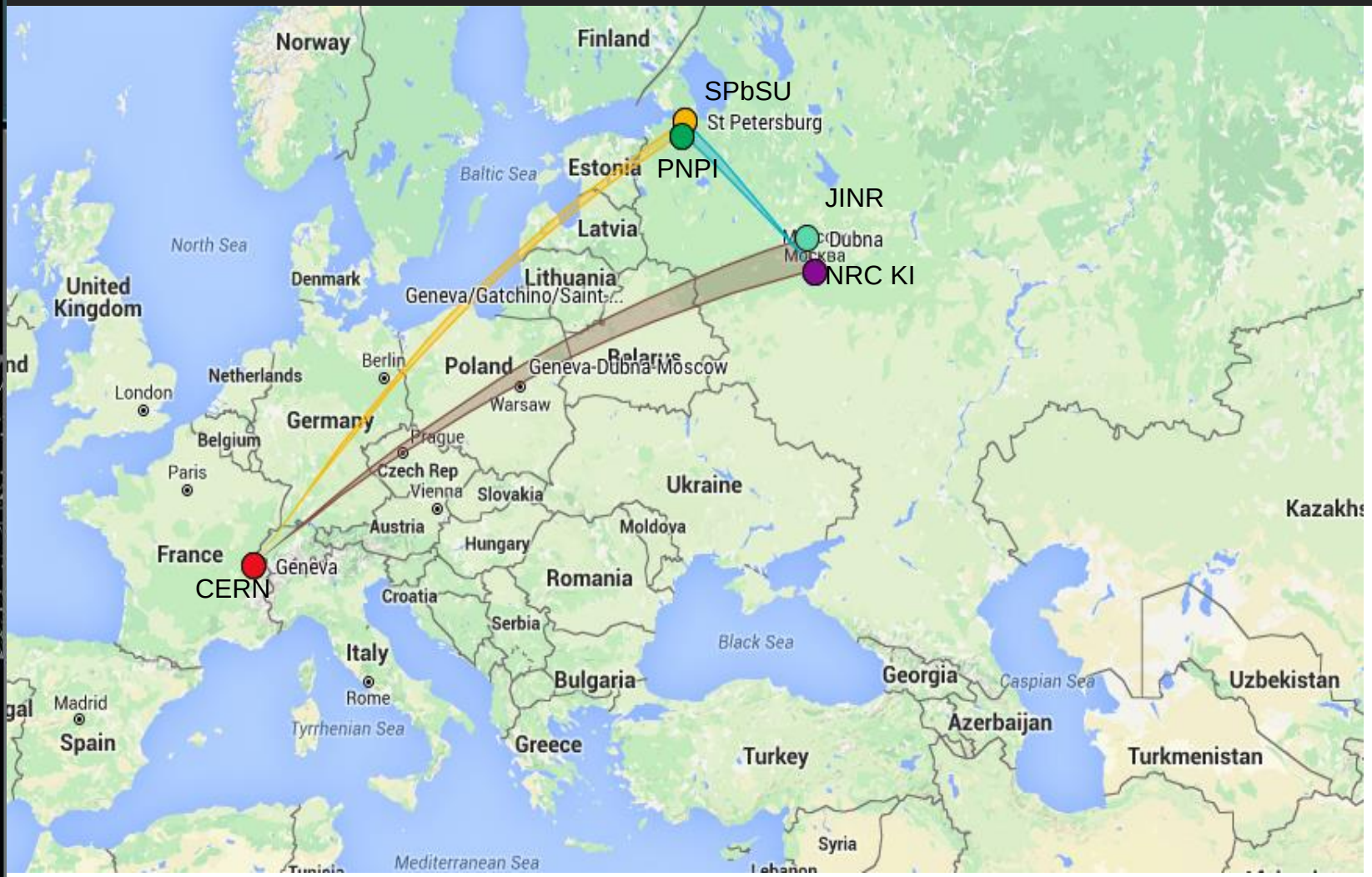
Jobs can run at sites w/o data but with free CPUs: OVERFLOW (up to 10% of jobs)

S.Campana CHEP2015

R&D Project Motivation

- **Computing models for the Run3 and HLLHC 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 fully-fledged software stack. Distributed stuff like FAX, ALICE xrootd, EOS@CERN, AAA / CMS, dCache, etc (mostly) works.**
  - In our project we try to analyze how to set up distributed storage in one region and how it can be used from Grid sites, from HPC, academic and commercial clouds, etc.

Sites and “triangles”.



Tested : T2 (ATLAS, PNPI, Gatchina), T2 (ALICE, SPbSU, Petergof), CERN
Ready for tests : T2 (ATLAS, PNPI, Gatchina), T2 (ALICE, SPbSU, Petergof), T1 (NRC-KI, Moscow)
Planned : T0 (JINR, Dubna, NICA), T0/T1 (NRC KI, Moscow, FAIR), T0 (CERN, LHC)

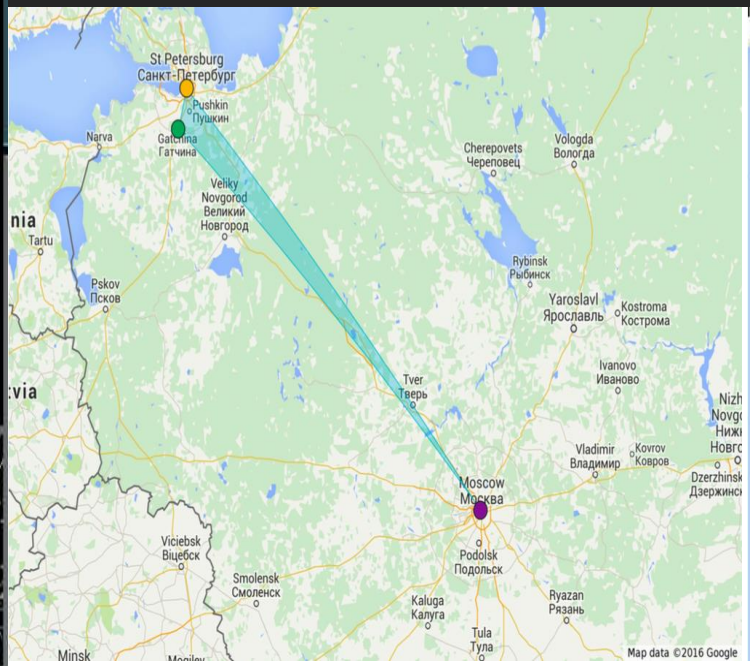
Incomplete List of Requirements

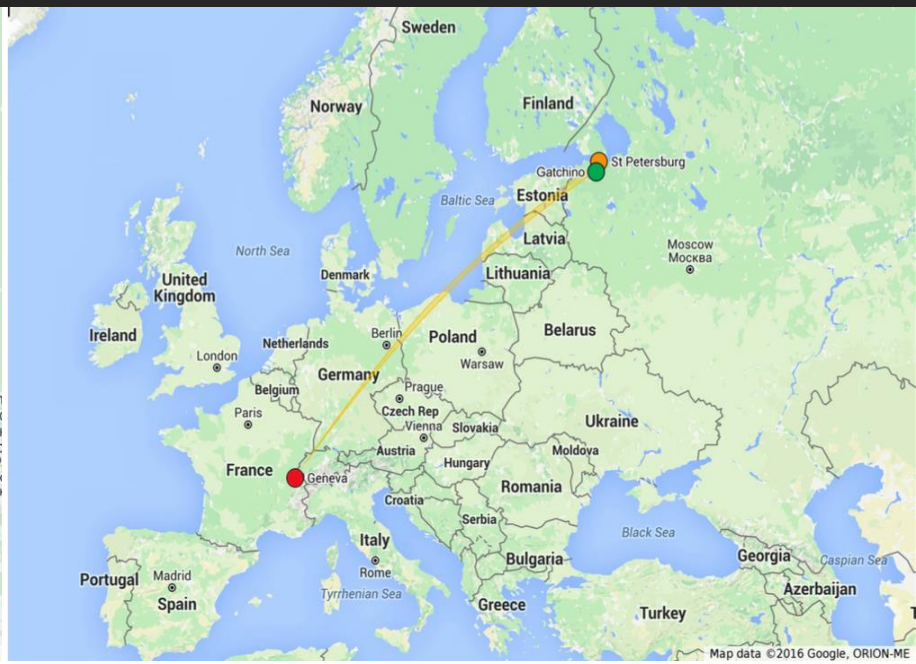
- **Single entry point**
- **Should be usable by major "players" (at least 2-3 LHC experiments)**
  - And it should be interested to the future experiments beyond LHC
- **Scalability and Integrity**
  - it should be easy to add new sites
- **Data transfer optimization: transfers should be routed directly to the disk servers avoiding intermediate gateways and other bottlenecks**
- **Stability. Fault tolerance.**
  - core components redundancy
- **Built-in virtual namespace, no dependency on external catalogues**

Technology choice.

- **We had to find a solution that supports federation of distributed storage resources. This very much depends on a transfer protocol support for redirection. Two protocols that are good at it are xroot and HTTP.**
  - HTTP-based federation is implemented in DynaFed software developed by IT/SDC 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.
  - xroot-based solution is EOS. It's also developed (and supported) at CERN, has characteristics closely matching our requirements.
    - We decided to try it first and then to repeat the same with dCache.

Testbed Infrastructure and Configuration



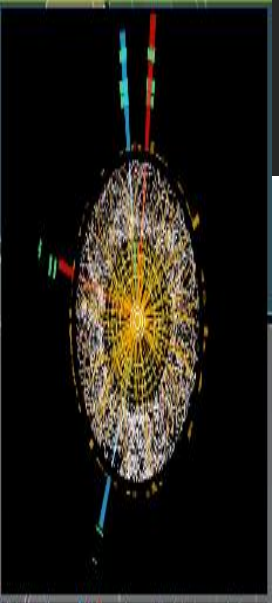


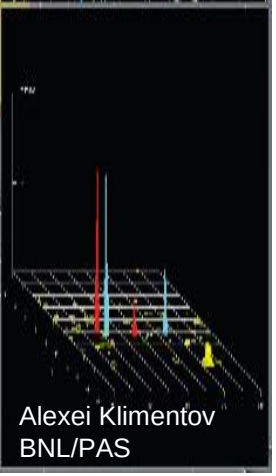
- **CERN :**
  1. MGM (master) + perfSonar + UI
- **NRC-KI**
  - perfSonar
- **SPbSU**
  1. MGM (master) + FST+ perfSonar + UI
  2. MGM (slave) + FST + perfSonar + UI
- **PNPI :**
  1. MGM (master) + FST + perfSonar + UI
  2. MGM (slave) + FST + perfSonar + UI

- **No dedicated network**
- **Commodity servers, disks and switches**
- **Test servers set-up**
  - Base OS: SL6/x64
  - EOS Aquamarine
  - Authentication scheme: GSI
- **Tests**
  - Bonnie++ (file I/O test on FUSE-mounted file system)
  - ATLAS test: standard ATLAS event reconstruction workflow with Athena
  - ALICE test: sequential ROOT event processing
- **PerfSONAR**
  - Network performance monitoring

EOS
- *MGM – management server*
- *FST – file storage server, storing file's by file id*
- *UI : User's I/F to handle system configuration*

perfSonar
- *widely-deployed test and measurement infrastructure that is used by science networks and facilities around the world to monitor and ensure network performance*

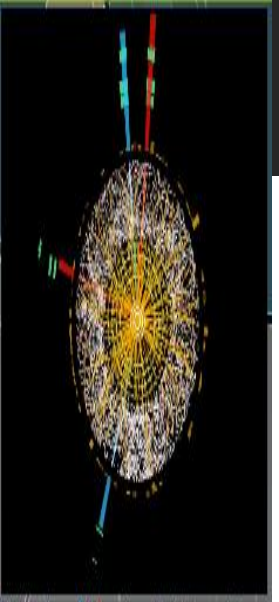


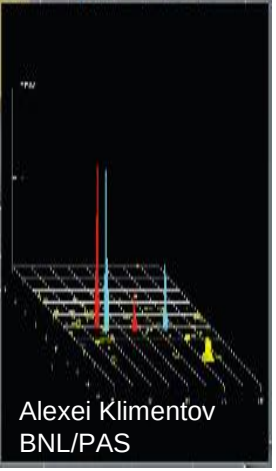


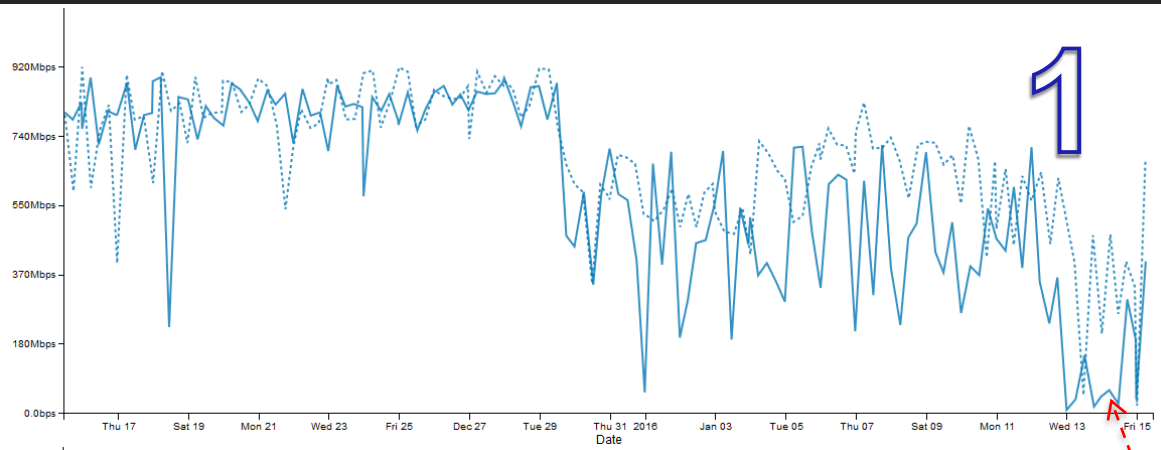
Conducted Tests

- **Synthetic tests :**
  - Bonnie++
    - PNPI FST local
    - SPbSU FST local
    - UI PNPI to ALL
    - UI SPbSU to ALL
    - UI CERN to ALL
  - PerfSonar
    - CERN-PNPI
    - CERN-SPbSU
    - PNPI-SPbSU
- **HENP test suite**
  - ATLAS TRT Reconstruction/Analysis (files distributed between two servers)
    - ATLAS objectives:
      - One of the important studies to understand ATLAS Inner Detector performance at high occupancy conditions is a reconstruction of proton-proton events with large number of interactions in Transition Radiation Tracker. It is actually great challenge for computers since events reconstruction in TRT require reconstruction of each hit on track for events with high number of simultaneous interactions. Important feature of every ATLAS standard reconstruction job is the presence of very detailed log files with information about CPUs used during the operation. Calculations from these log files can be used to study the performance of new federated system. Output files after the reconstruction can be also tested on physics to validate SW on a different clouds (storage parts).:
    - Input RAW data files for reconstruction are located on the different storages:
    - Data accessed locally and remotely, the output is always local
    - UI PNPI to ALL (fuse, xrootd)
    - UI SPbSU to ALL (fuse, xrootd)
    - UI CERN to ALL (fuse, xrootd)
    - Launching of the identical tasks with identical input files but located on alternative storage elements allow us to estimate I/O efficiency and benefits of distributed federated storage vs local (Grid) storage
  - ALICE events selection (files distributed between two servers)
    - ALICE Objectives.
      - Read event by event, select events according to criteria (TPC “pattern”) and store them in output file. Both input and output are ROOT files
    - UI PNPI to ALL (fuse, xrootd)
    - UI SPbSU to ALL (fuse, xrootd)
    - UI CERN to ALL (fuse, xrootd)

Network Monitoring with PerfSonar



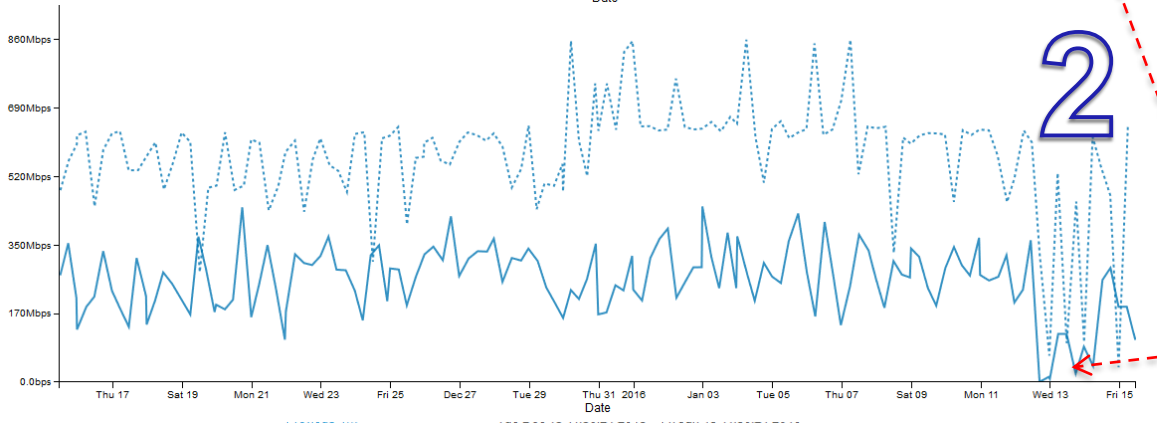




1

Monthly Throughput in Mbit/s (15.12.15-15.01.16)

1. PNPI- SPbSU
2. CERN-SPbSU
3. PNPI-CERN

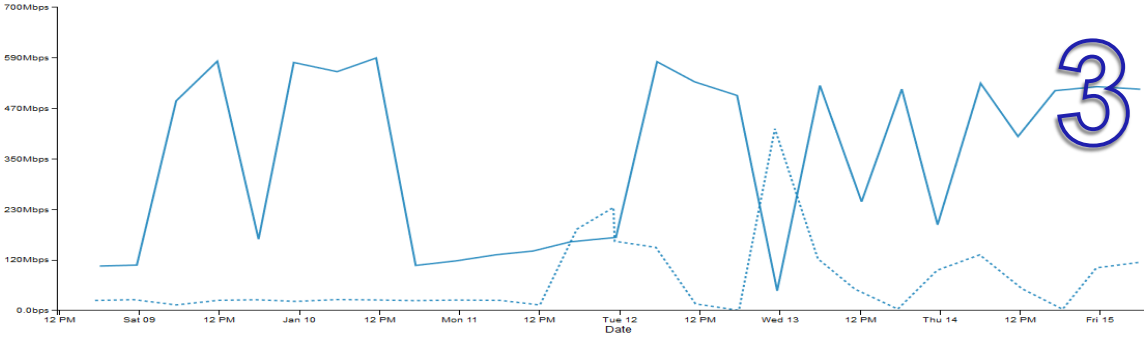


2

Throughput

Reverse Throughput

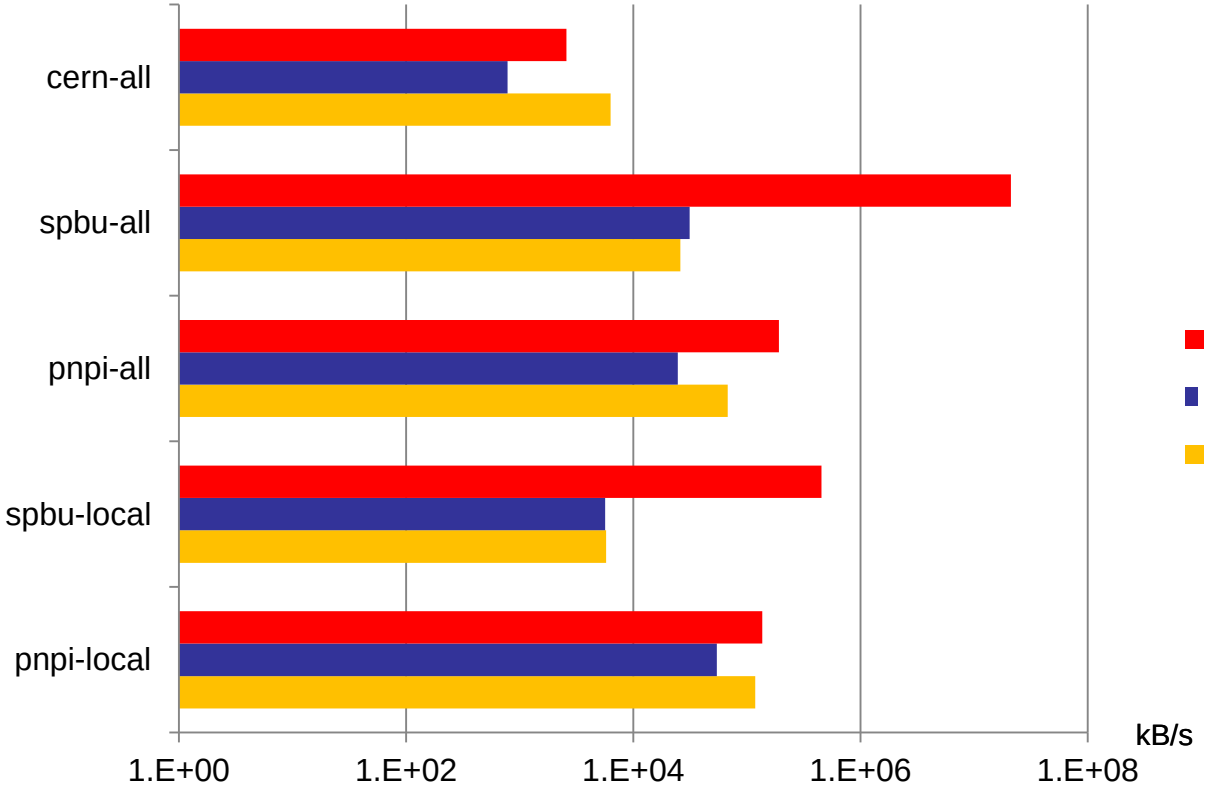
Jan issues with WAN



3

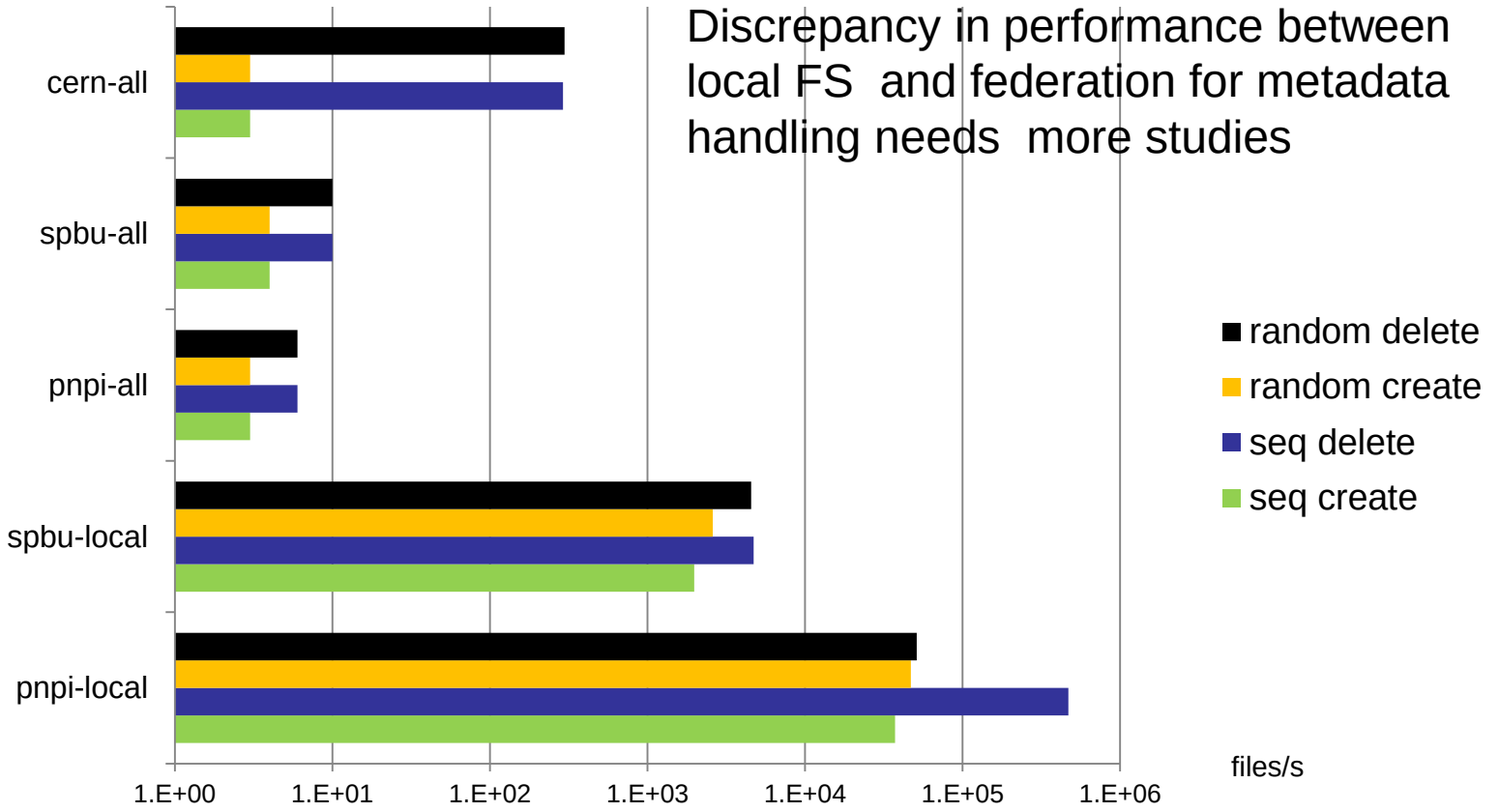
PerfSonar is used to monitor network state Federation sites have no dedicated network

Synthetic Tests I. Bonnie++ Bloc I/O



- **Legend**
  - CERN-ALL : client@CERN, data@PNPI+SPbSU, MGM@CERN
  - SPBU-ALL:client@SPbSU, data@PNPI+SPbSU, MGM@CERN
  - PNPI-ALL: client@PNPI, data@PNPI+SPbSU, MGM@CERN
  - SPBU-LOCAL : client@SPbSU, data@SPbSU, no federation
  - PNPI-LOCAL : client@PNPI, data@PNPI, no fedration

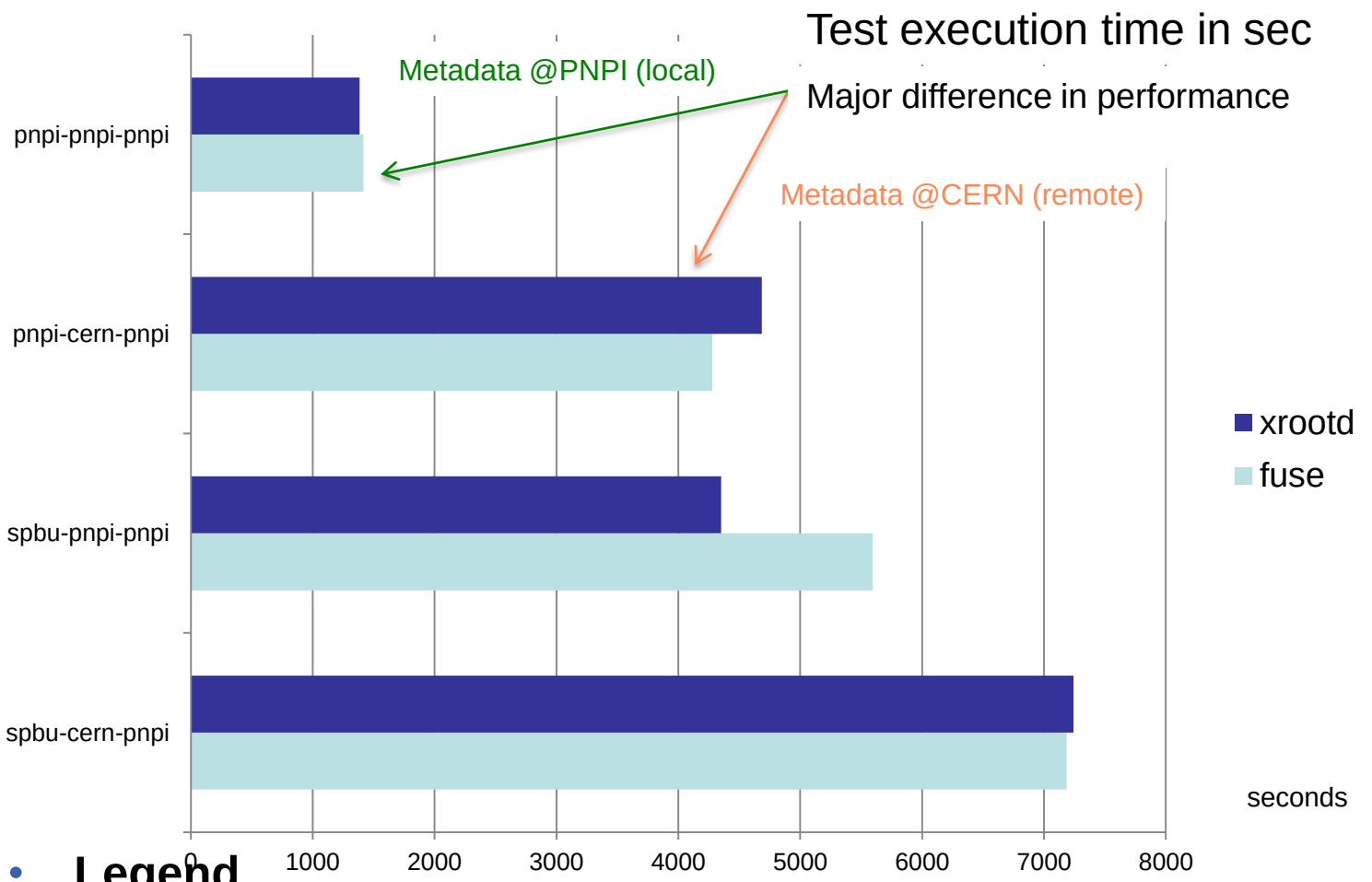
Synthetic Tests II. Bonnie++ File Metadata



Discrepancy in performance between local FS and federation for metadata handling needs more studies

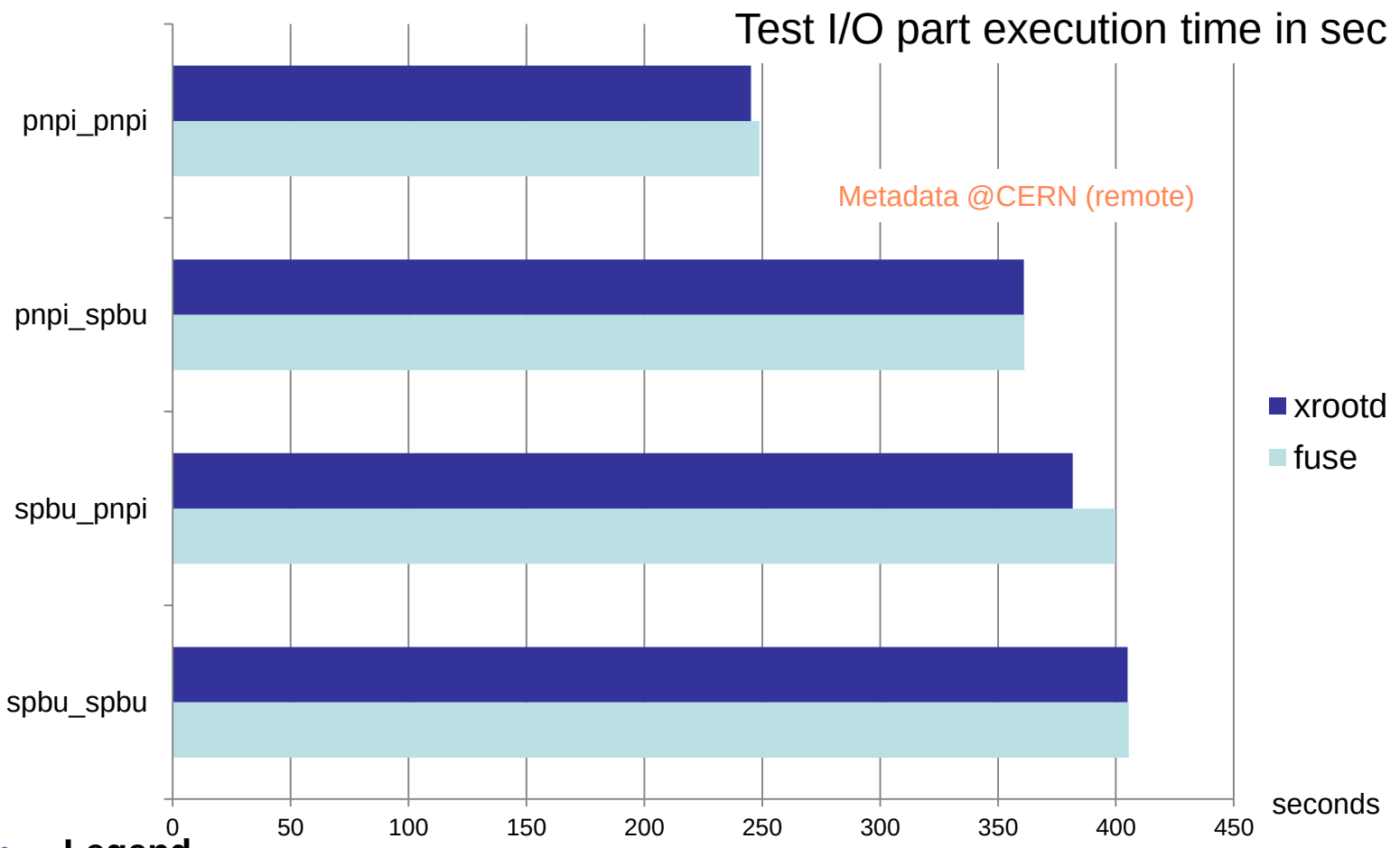
- **Legend**
  - CERN-ALL : client@CERN, data@PNPI+SPbSU, MGM@CERN
  - SPBU-ALL:client@SPbSU, data@PNPI+SPbSU, MGM@CERN
  - PNPI-ALL: client@PNPI, data@PNPI+SPbSU, MGM@CERN
  - SPBU-LOCAL : client@SPbSU, data@SPbSU, no federation
  - PNPI-LOCAL : client@PNPI, data@PNPI, no fedration

HENP Tests I. ALICE



- **Legend**
  - PNPI-PNPI-PNPI : client@PNPI, data@PNPI, MGM@PNPI
  - PNPI-CERN-PNPI: client@PNPI, data@PNPI, MGM@CERN
  - SPBU-PNPI-PNPI: client@SPBU, data@PNPI, MGM@PNPI
  - SPBU-CERN-PNPI : client@SPBU, data@PNPI, MGM@CERN

HENP Tests II. ATLAS



- **Legend**
  - PNPI-PNPI : client@PNPI, data@PNPI, MGM@CERN
  - PNPI-SPBU: client@PNPI, data@SPBU, MGM@CERN
  - SPBU-PNPI: client@SPBU, data@PNPI, MGM@CERN
  - SPBU-PNPI : client@SPBU, data@PNPI, MGM@CERN
  - SPBU-SPBU: client@SPBU, data@SPBU, MGM@CERN

- **Meta-data are always at CERN**
- **FST mounted via xrootd and fuse**

Outline

- **EOS Federation including three sites have been setup**
  - No dedicated network
  - Commodity servers, disks and switches
  - Technical support from EOS core SW team is essential for the success
- **Run ATLAS and ALICE event reconstruction/analysis realistic applications to study federation performance.**
  - No major obstacles to have it as production like federation
    - But it is important to have EOS features in place
  - No major loss in data analysis/reconstruction performance
  - No Federation impact on total execution time for CPU intensive applications
- **JINR (Dubna) is interested to join R&D**
- **The same exercise will be done for more combinations**
  - CERN-JINR-NRC/KI
  - SPbSU-PNPI-NRC/KI
  - SPbSU-PNPI-CERN-JINR
  - SPbSU-PNPI-CERN-NRC/KI
  - …and with access from HPC machines to the Federation

Thanks

- **This talk drew on presentations, discussions, comments, input from many**
- **Thanks to all, including those I've missed**
  - *P.Buncic, S.Campana, K.De, D.Duellmann, P.Fuhrmann, M.Grigorieva, V.Ilyin, O.Keeble, V.Korenkov, M.Lamanna, A.Peters, L.Robertson, E.Ryabinkin, M.Schulz, V.Velikhov …*

- Acknowledgements.
  - This project was funded in part by the Russian Fund of Fundamental Research under contract "**15-29-07942 офи_м"** and U. S. DOE, Office of Science, High Energy Physics and ASCR under Contract No. DE-AC02-98CH10886