A prototype for the ALICE Analysis Facility at GSI

CHEP 2018
Sofia, Bulgaria
July 2018

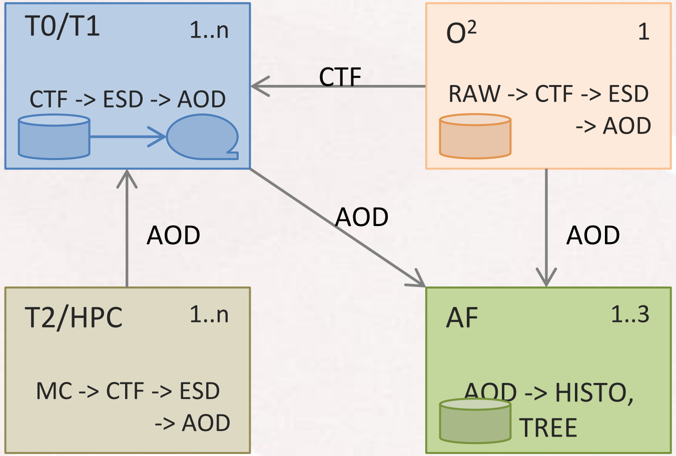
Kilian Schwarz

Sören Fleischer, Raffaele Grosso, Jan Knedlik, Thorsten Kollegger

GSI Helmholtzzentrum für Schwerionenforschung GmbH

for the ALICE Collaboration

ALICE Analysis Facility Prototype Motivation: Run 3 Computing Model



Grid Tiers mostly specialized for given role

- O2 facility (2/3 of reconstruction and calibration), T1s (1/3 of reconstruction and calibration, archiving to tape), T2s (simulation)
- All AODs will be collected on the specialized Analysis Facilities (AF) capable of processing ~5 PB of data within ½ day timescale (a throughput of about 115 GB/s)

The goal is to minimize data movement and optimize processing efficiency

(P. Buncic, ALICE T1/T2 Workshop, Strasbourg, May 2017)

ALICE Analysis Facility Prototype:

current setup

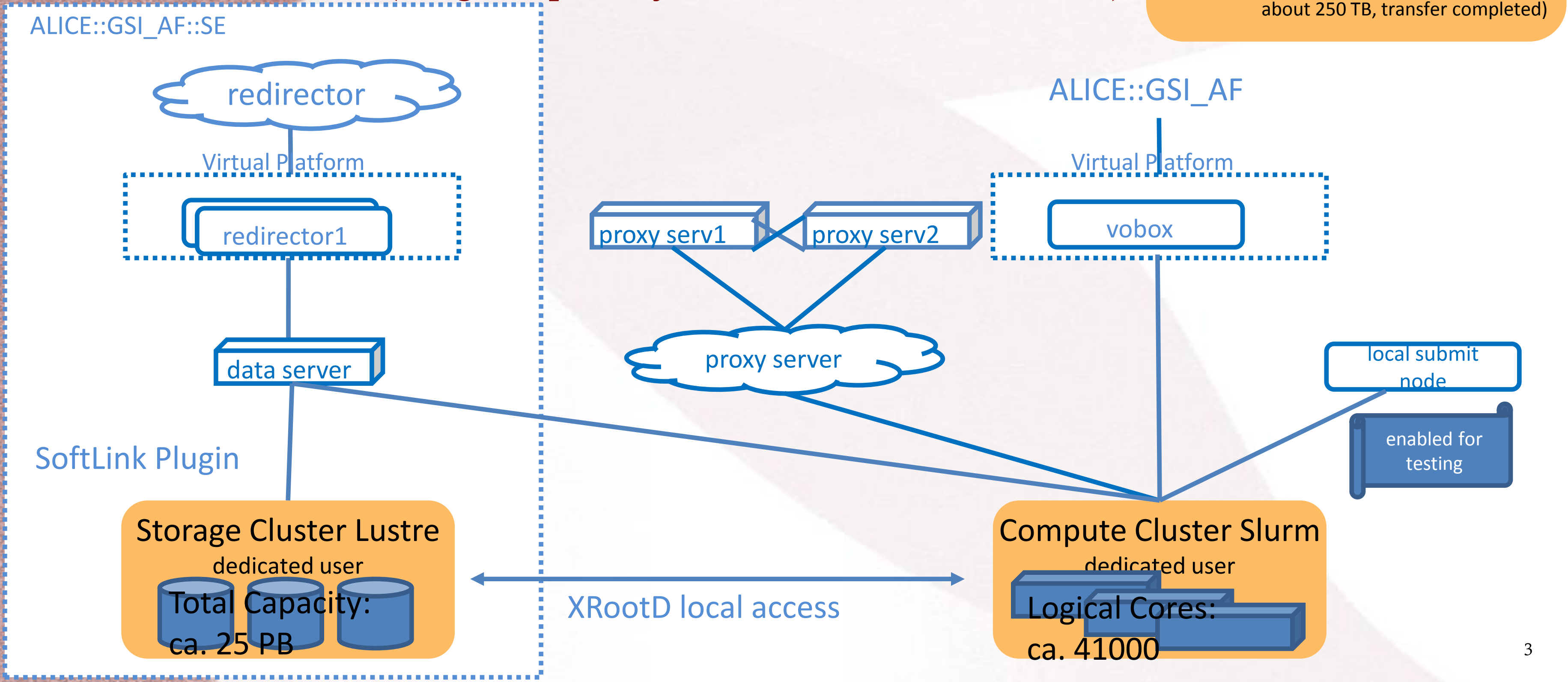
(design inspired by current ALICE Tier2 centre)

Initial Resources

- 0.6 PB disk space
- job slots taken from Tier 2 allocation

Initial dataset

- Full AOD set 2015 Pb-Pb (LHC15o, about 250 TB, transfer completed)



ALICE::GSI_AF::SE

redirector

Virtual Platform

redirector1

data server

SoftLink Plugin

Storage Cluster Lustre
dedicated user
Total Capacity:
ca. 25 PB

XRootD local access

proxy serv1

proxy serv2

proxy server

ALICE::GSI_AF

Virtual Platform

vobox

local submit node

enabled for testing

Compute Cluster Slurm
dedicated user
Logical Cores:
ca. 41000

ALICE Analysis Facility Prototype:

planned improvements

(design inspired by current ALICE Tier2 centre)

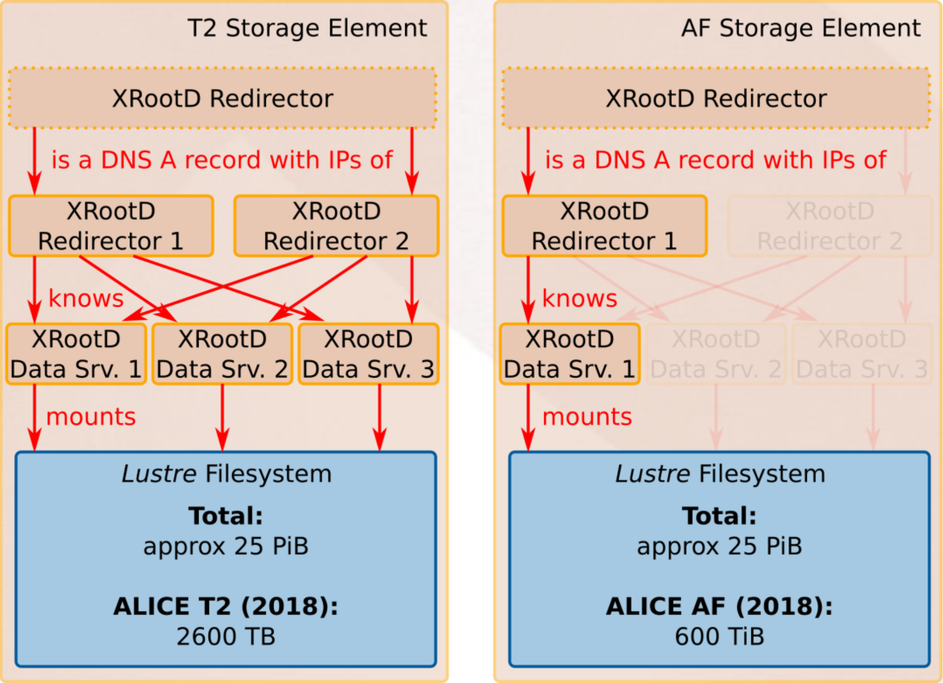
Initial Resources

- 0.6 PB disk space
- job slots taken from Tier 2 allocation

Initial dataset

- Full AOD set 2015 Pb-Pb (LHC15o, about 250 TB, transfer completed)

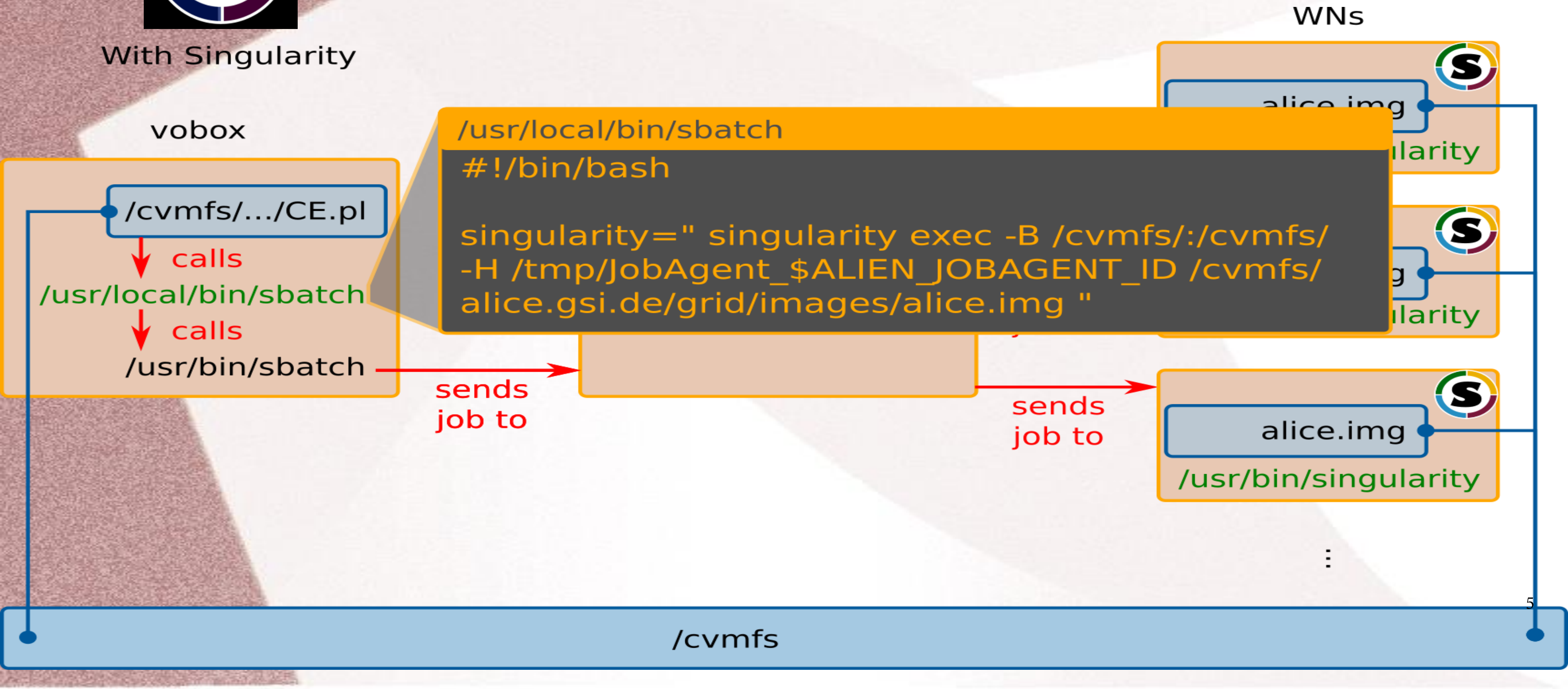
Redundant XRootD Redirectors and Data Servers



GSI ALICE AF Prototype – *Singularity*

Scientific Linux environment provided by Singularity containers on Debian-based HPC cluster -- in production at GSI ALICE Tier2 centre since 2015

With Singularity



/usr/local/bin/sbatch

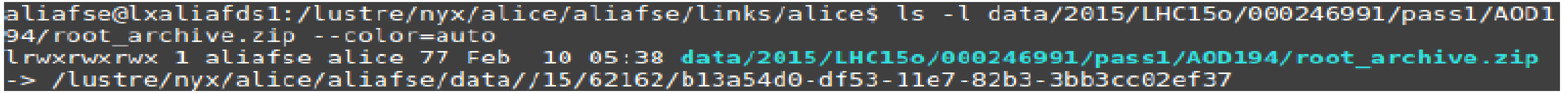
```
#!/bin/bash

singularity=" singularity exec -B /cvmfs/:/cvmfs/
-H /tmp/JobAgent_$ALIEN_JOBAGENT_ID /cvmfs/
alice.gsi.de/grid/images/alice.img "
```

GSI ALICE AF Prototype – XRootD Plugins
a) Symlink Plugin

Feature: Files written to the xrootd data server via AliEn have hash-based filenames. Physicists prefer naturally speaking names.

Solution: Introduce xrootd plugin that creates symbolic links in a different directory that map the AliEn filename (LFN) to the physical filename on storage.



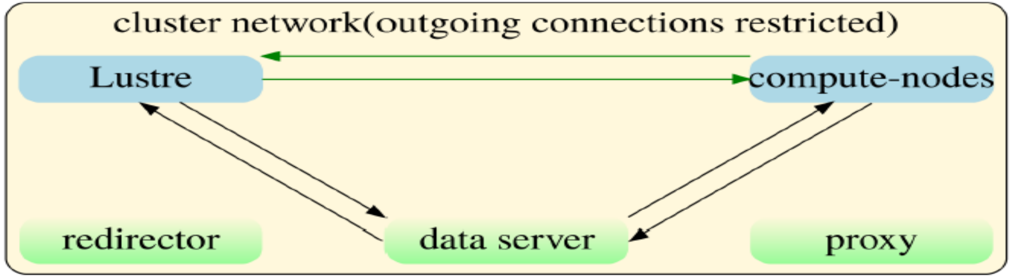
```
aliafse@lxaliafds1:/lustre/nyx/alice/aliafse/links/alice$ ls -l data/2015/LHC15o/000246991/pass1/AOD1
94/root_archive.zip --color=auto
lrwxrwxrwx 1 aliafse alice 77 Feb  10 05:38 data/2015/LHC15o/000246991/pass1/AOD194/root_archive.zip
-> /lustre/nyx/alice/aliafse/data//15/62162/b13a54d0-df53-11e7-82b3-3bb3cc02ef37
```

symlinks are created in XrdAliceTokenAcc during file access authorisation
  checks envelope for read/write access
symlinks are removed in corresponding symlink removal functionality
  checks envelope for delete access

GSI ALICE AF Prototype – XRootD Plugins b) RedirPlugin

XRootD Redir Plug-in :
Reading via XRootD data servers doubles the network traffic inside the infiniband network. This is, especially with a limited number of XRootD servers, a bottleneck in CPU & bandwidth to our setup.
➔ Clients at GSI should read the file directly from Lustre, circumventing XRootD data servers.

see poster presentation J. Knedlik
327. XRootD plug-in based solutions for site specific requirements Track 4- Data Handling



Server (Redirector) Plugin (cms.ofslib).
v4 Client API (XrdCl) needs to be used
Needed Client code in XRootD base starting with version 4.8
(see https://github.com/xrootd/xrootd)

issue:
TkAuthorisation on Xrd data servers is bypassed ➔ complex interplay between local and XRootD file rights, currently redirectLocal only in read mode

GSI ALICE AF Prototype
Local redirections & ROOT (TAlienFile)

- Redirection with ROOT client works as long as TNetXNGFile(new XrdCl) is used and ROOT is compiled against XRootD > 4.8.0
- TAlienFile:
  - Let TAlienFile be derived from TNetXNGFile(new XrdCl)
  - Add an LFN parameter to TNetXNGFile ctor to pass LFN to TFile/TArchiveFile (like in TXNetFile)
  - in case of local file host is set to localhost which prevents that XRootD data servers are being queried
  - Status: Working (including archive files)

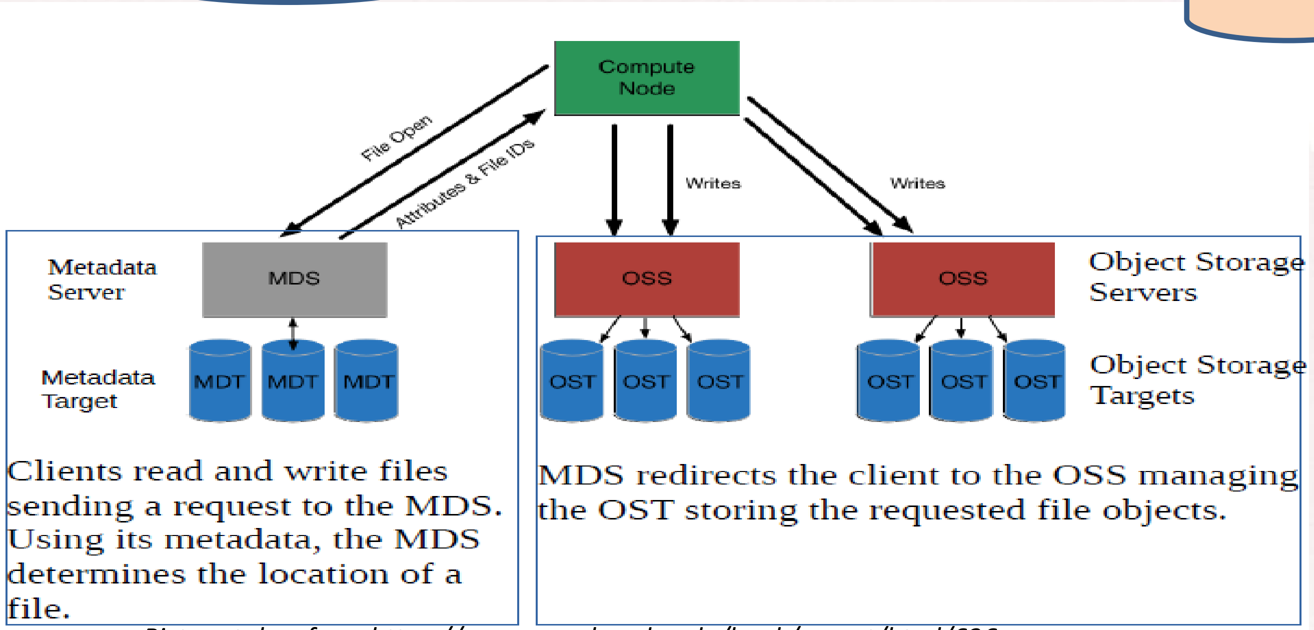
- status TJAlienFile: test environment being prepared

see poster presentation J. Knedlik
327. XRootD plug-in based solutions for site specific requirements Track 4- Data Handling

GSI ALICE AF Prototype more about Lustre

HEBE test Cluster:
Lustre v 2.10
8.2 PB, 30 OSS in total
7 OST for 1 OSS

Production System
Nyx: 17 PB



Clients read and write files sending a request to the MDS. Using its metadata, the MDS determines the location of a file.

MDS redirects the client to the OSS managing the OST storing the requested file objects.

Picture taken from https://www.rc.colorado.edu/book/export/html/626

GSI ALICE AF Prototype
Lustre I/O scaling tests

Method:

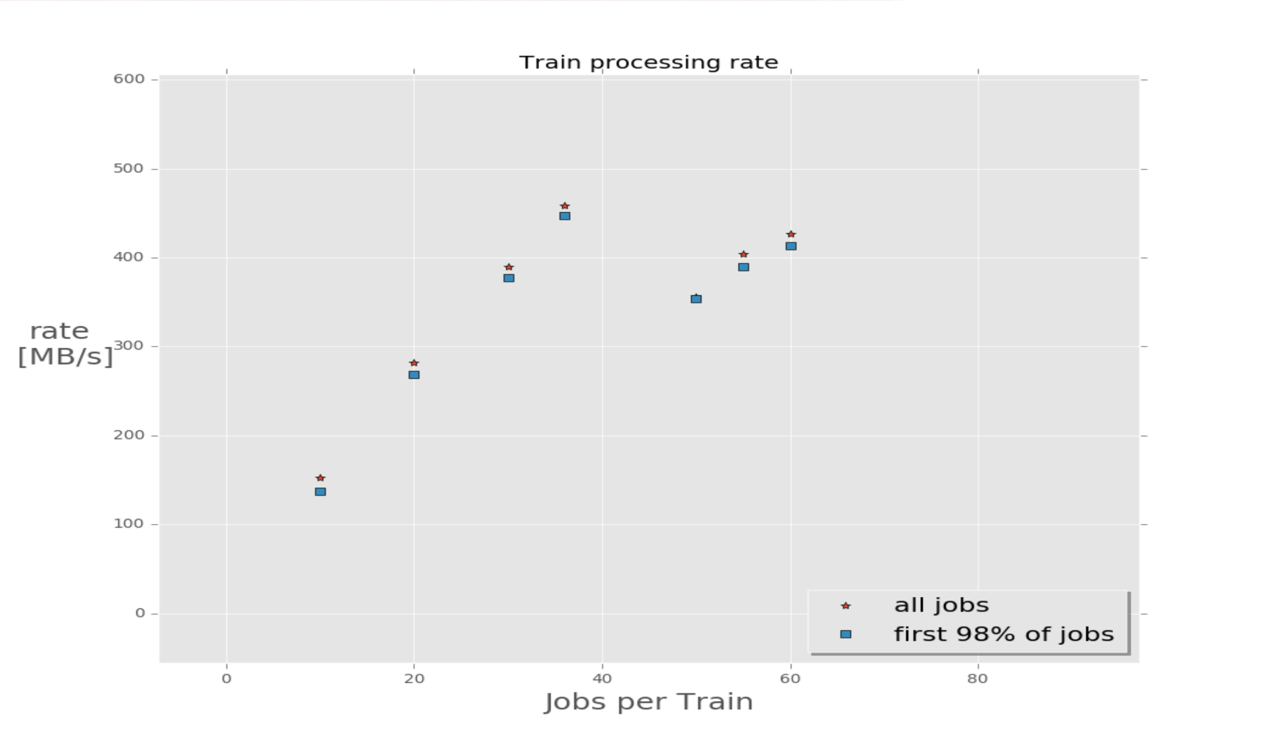
- local ALICE analysis train with simple analysis task submitted to GSI batch farm with varying number of jobs
- Partition has been reserved for exclusive use
- in order to achieve consistency only jobs which started within the first minute are being considered. Plots are generated when 98% of these jobs are finished.
- Software used: patched versions of ROOT v5-34-30-alice & AliRoot v5-09-32
- OST test: only data from single OST have been read
- OSS test: only data from single OSS have been read
- Hebe test:
    - the Hebe test cluster consists of 30 OSS, each OSS manages 7 OSTs
    - Data reading is equally distributed among the OSS
    - Scaling was limited due to size of testing partition (max. of 2500 concurrent jobs)

GSI ALICE AF Prototype
I/O scaling tests OST

OST Test:

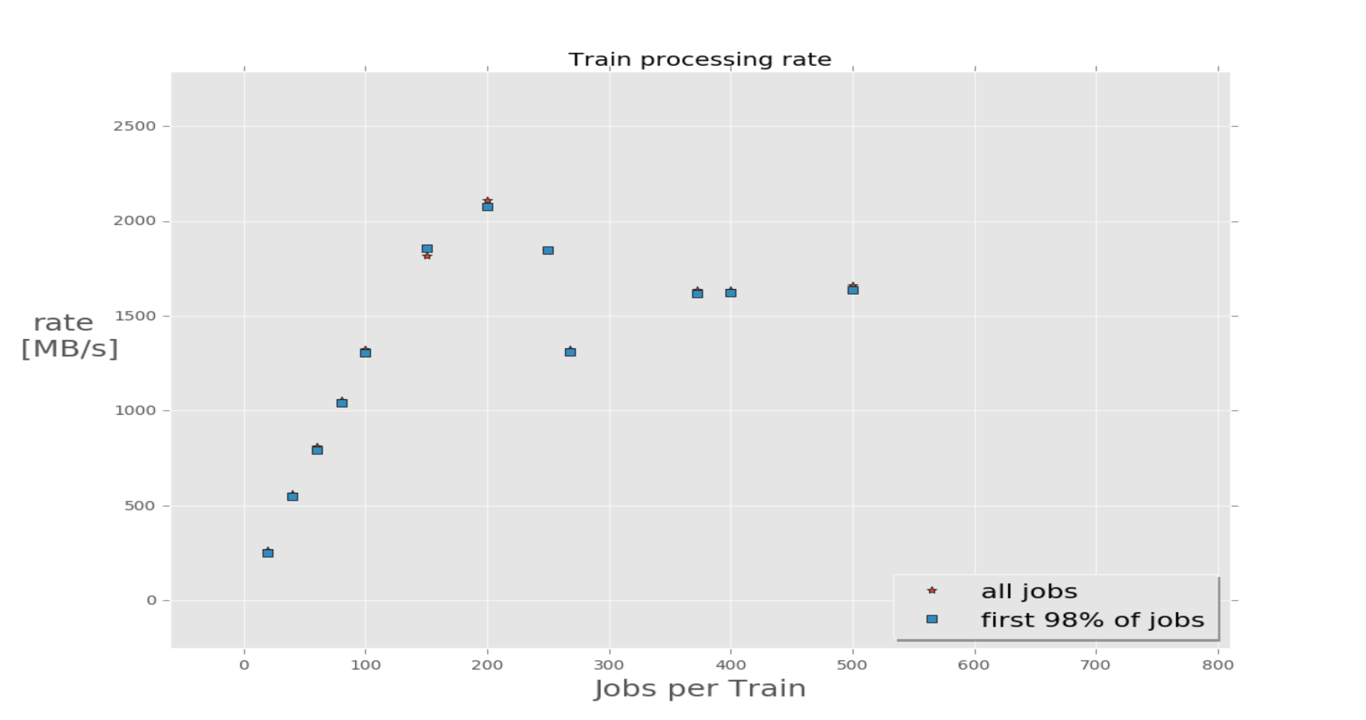
reading with increasing number of jobs from single OST.

Maximum rate: 440 MB/s



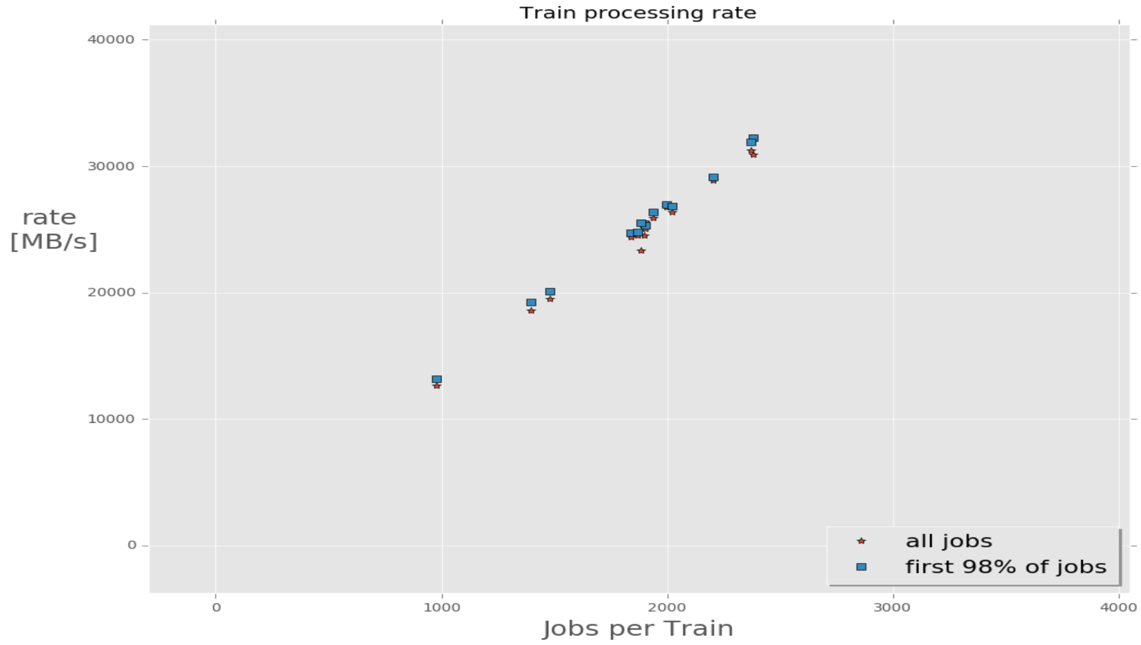
GSI ALICE AF Prototype
I/O scaling tests OSS

OSS Test:
reading with increasing number of jobs from single OSS.
Maximum rate 2100 MB/s



GSI ALICE AF Prototype I/O scaling tests Hebe

Hebe Test:
reading with increasing number of jobs from Hebe Cluster (30 OSS).
Maximum rate: 32 GB/s due to limitation to 2500 concurrent jobs
Desired target rate should be achievable by scaling number of jobs and OSS accordingly.



GSI ALICE AF Prototype
Summary and conclusion

- A prototype of an ALICE Analysis Facility has been set up at GSI
- Key solutions have been implemented using XRootD Plug-Ins
- Performance tests suggest that the target throughput rate of 10 PB/day can be achieved
- Further improvements of the current set up including scaling to production size are on the way