# Table of Contents

# Table of Contents

# Question

Question 3. Please describe the storage system (e.g., DPM, dCache, SLAC xrootd, EOS, POSIX-direct access, Object-Store direct access) you use to make the storage available to the users and to grid middleware (in general).Is there any significant or unusual configuration in how the storage middleware is using the underlying storage? Examples of unusual configuration include more robust storage (e.g., storing multiple replicas or using erasure coding).Do you prefer any particular data access protocol? If so, why?

# Answers

## CERN

EOS

EOS system with xrootd interface for data access and gridftp gateway for data transfer.

Ceph

Remote block devices for VMs S3 for applications : Atlas event service, CVMFS, BOINC CMS data bridge, IT infrastructure, Cephfs for HPC storage, Openstack Manila (filer use cases), linuxsoft, Atlas PanDa, BOINC Object interface used by Castor public instance.

Castor

Archival storage - tape with disk pool

CERNBox

Sync n share frontend to EOS.

AFS

We have an AFS installation which is gradually being phased out.

A new tape backend for EOS, named CERN Tape Archive ( CTA ) is under development.

## hephy-Vienna

Legacy Systems DPM and OpenAFS. We move to EOS. We will be at an HPC facility, so we will use also their NFS based storage.

## KI-LT2-QMUL

storage presented as 4.8 PB Lustre file system providing POSIX-direct access using StoRM for WAN access

## UKI-LT2-RHUL

DPM

## RO-13-ISS

xrootd, experiment protocol

## Nebraska

We use XrootD and gridftp to allow access to the storage to the grid. The storage is also available for direct POSIX-like reads to worker nodes via FUSE mounts of the HDFS system and available if a job is aware of our storage topography. We are phasing out gridftp in the near future as x509 is losing favor in the grid world.

Likely DAVS/HTTPS writes via xrootd will take over that role.

## INFN-ROMA1

We use DPM is a standard way, although we are looking forward to interface it with our ceph facility, possibly via the S3 api.

## NDGF-T1

Distributed dCache centrally operated through non-priviledged accounts on servers run by resource providers.

## BEgrid-ULB-VUB

MASS STORAGE: * dCache * no replicas or EC (done on hardware level via raid) * wide variety of protocols thanks to dCache:

- grid: xrootd, http, srm, gridftp
- local: nfs, dcap

## NCG-INGRID-PT

StoRM on top of Lustre. Lustre also provides POSIX direct file access. We also support xrootd.

## IN2P3-IRES

The storage is made available with DPM (SRM, xRootD and HTTPS) and with CEPH (S3)

## LRZ-LMU

dCache

## CA-WATERLOO-T2

Independent dedicated RAID6 servers for dcache pools, admin/head nodes on shared VM structure (with CEPH backend). Support gridftp, dcap, xrootd, webdav.

## CA-VICTORIA-WESTGRID-T2

dCache. We have a few "hot pools" that replicate popular files , more for performance and load balancing than resilience.

## Taiwan_LCG2

1. We use DPM to support WLCG user. 1.2. We use cephFS to offer POSIX-direct access for our local scientifitic applications. 1.3. We also use EOS+CERNBOX for provide dropbox-like service for local user data. 2. We don't use unusual configuration for storage middleware. 3. We would prefer https and posix-direct access, as they are much more common.

## IN2P3-SUBATECH

EOS with dual (master/slave) manager

## asd

## MPPMU

dCache

## INFN-LNL-2

for CMS: dCache supporting dcap, xrootd and gsiftp access protocol for ALICE: native xrootd

## Australia-ATLAS

DPM

dCache on raid boxes with gridftp, xrootd and http doors, directly integrated in NDGF-T1 dCache instance, separate local dCache instance for Belle II. Ceph on separate hardware used for CephFS, ObjectStore only experimental. All input files are cached on CephFS, local scratch is not used for job inputs. ARC-CE + ARC Cache to manage this storage, push mode used for ARC-CE to prefetch inputs in cache. direct xrootd I/O over WAN (NDGF-T1) for analysis jobs.

## KR-KISTI-GSDC-02

dCache

## UKI-LT2-IC-HEP

dCache, gridftp (it works and isn't trendy)

## -UCL

POSIX-direct access

## UKI-SOUTHGRID-BRIS-HEP

DMLite + HDFS plugin, xrootd - local users use HDFS commands for read & write and POSIX mount for read-only

## GR-07-UOI-HEPLAB

DPM

## UKI-SOUTHGRID-CAM-HEP

DPM

## USC-LCG2

DPM

## EELA-UTFSM

dCache

## DESY-ZN

dCache

## PSNC

DPM : SATA HDD , 0.500TiB, Xrootd : SATA HDD, 60TiB, dCache: HDD 1PiB, tape : 10PB

## UAM-LCG2

dCache

## T2_HU_BUDAPEST

DPM

## INFN-Bari

GPFS + StoRM

## IEPSAS-Kosice

dCache 3.2 with Centos 7 , EOS 4.4.46 with Centos 7 , Xrootd 4.9.1 sl 6.10

## IN2P3-CC

Disk infrastructure is splitting on two part

First one is based on dCache available for the grid middleware (usable capacity of 19 326 TiB). We have no preferences concerning the protocol ( xrootd, webdav,  .) Second one is based on Xrootd available for Alice Grid use ( 5 048 TiB usable capacity)

## NONE_DUMMY

blah

## WEIZMANN-LCG2

STORM on Lustre file system

## RU-SPbSU

EOS

## USCMS_FNAL_WC1

Not really understanding the question -- disk storage is dCache, with access exclusively via xrootd, gridFTP. Also deploy EOS for LPC user space, also xrootd, gridFTP, with limited FUSE mount access allowed.

## RRC-KI-T1

EOS for ALICE's disk storage dCache for LHCb and ATLAS disk storages dCache + Enstore for ALICE, ATLAS and LHCb tape-based storage We don't prefer any particular data access protocol.

## vanderbilt

posix mounts for user access, gridftp/xrootd plugins to speak directly to filesystem

## UNIBE-LHEP

DPM for the SE, lustre for the ARC cache / shared scratch for the clusters

## CA-SFU-T2

dCache, plain configuration

## _CSCS-LCG2

dCache on top of Spectrum Scale; we also have a scratch filesystem for jobs

## T2_BR_SPRACE

dCache, we prefer SRM/GridFTP

## T2_BR_UERJ

We use HDFS and XRootD available to the users and to grid middleware

## GSI-LCG2

XRootD, POSIX

## UKI-NORTHGRID-LIV-HEP

Yes, we provide to local users and grid. Nothing unusual.

## CIEMAT-LCG2

We use disk-only (no tape back-end) dCache, accessible via GriDFTP, xrootd, WebDAV and NFS (local users only).

We use RAID6 for our disk servers, but, in general we don't store several replicas or use erasure codes (though we're considering the latter for future deployments with CEPH underlying dCache pools).

We do have multiple replicas (using dCache's migration module) for valuable data that is not replicated in other WLCG sites (basically, local users data, both from CMS and non-WLCG users). We are considering moving from the migration module to dCache Resilience Manager.

Regarding data access protocol, we haven seen certain problems with NFS (though, we are still very interested in its use locally), and it seems that GridFTP is going away. HTTP (WebDav) and xrootd are OK with us, but we would like to have Kerberos and token-based authentication for dCache's xrootd, which is not there yet.

## a

## T2_US_Purdue

In general - all compute nodes access our storage through XRootD. One of the clusters fuse-mounts HDFS. Frontends (interactive login nodes) also fuse-mount HDFS.

## IN2P3-LAPP

High perf. High Avail. Storage Service - used as distributed File System across the cluster - hosting users HOME directories, shared area with the CEs, scientific data storage for non grid communities High capacity Storage - based on DPM - LHC and EGI VOs

## TRIUMF-LCG2

We run dCache on top of xfs systems, underlying is LUNs from different storage systems. Providing gridftp, xrootd, https and dcap protocols.

## KR-KISTI-GSDC-01

We deployed XRootD and EOS for ALICE experiment. We are using a general configuration only difference is that we deployed ALICE token authentication package upon XRootD and EOS. We prefer XRootD protocol since currently we are only supporting ALICE experiment.

## GRIF

DPM, SLAC xrootd(~1PB)

## IN2P3-CPPM

DPM

## IN2P3-LPC

DPM and native xrootd

## IN2P3-LPSC

DPM, no prefered data access protocol (xrootd, htppd, gridftp, rfio)

## ZA-CHPC

EOS, xrootd access for ALICE, gridftp/webdav/xrootd for ATLAS

## JINR-T1

dCache

## praguelcg2

xrootd for ALICE, DPM for others

## UKI-NORTHGRID-LIV-HEP

DPM. Yes, we provide to local users and grid. Nothing unusual.

## INDIACMS-TIFR

DPM

## TR-10-ULAKBIM

DPM

## prague_cesnet_lcg2

DPM DOME flavor

## TR-03-METU

DPM

## aurora-grid.lunarc.lu.se

## SARA-MATRIX_NKHEF-ELPROD__NL-T1_

dCache

## -UNIBA

dpm, xrootd, webdav

## DESY-HH

dCache with partly 2 replicas. Local users see and can access posix like name spaces, grid users use grid protocols.

## T3_PSI_CH

dCache

## SAMPA

XRootD, it don't unusual configuration, i don't prefer any particular protocol

## INFN-T1

StoRM, GridFTP, XrootD, WebDAV over GPFS

## GLOW

HDFS based storage is accessible via SLAC xrootd, Globus gridftp, and FUSE file system. We store 2x replica of all files. Hence the usable capacity is exactly half the raw space.

## UNI-FREIBURG

dcache 2.16 , xrootd access directly to poolserver, gridftp via proxy doors atm. network: IPv4 (IPv6 in prep)

## Ru-Troitsk-INR-LCG2

DPM

## T2_Estonia

Gridftp/xrtood have hdfs module builtin. Compute/user nodes user hdfs over fuse mount or over hdfs command line tool.

## pic

Our storage system is based on dCache (v. 4.2.32). We support protocols XRootD, NFS, HTTPS,GFtp and SRM. There are four instances for different purposes: test , development , pre-production (Middleware Readiness) and prod.

For CMS in Spain we have deployed a regional XRootD re-director, a service that allows jobs from CIEMAT or PIC to read -first- data from the region, then find data outside the region. We enabled this, since we are flocking jobs from PIC to CIEMAT, and vice versa, and we enforce those jobs to read data from the Spanish sites, instead of getting data remotely from distant sites.

At PIC we don t replicate data automatically. We improved the reliability of the disk service by using RAID6 hardware as explained in Q2.

# ifae

The ifae site is hosted at PIC. Our storage system is based on dCache (v. 4.2.32). We support protocols XRootD, NFS, HTTPS,GFtp and SRM. There are four instances for different purposes: test , development , pre-production (Middleware Readiness) and prod.

For CMS in Spain we have deployed a regional XRootD re-director, a service that allows jobs from CIEMAT or PIC to read -first- data from the region, then find data outside the region. We enabled this, since we are flocking jobs from PIC to CIEMAT, and vice versa, and we enforce those jobs to read data from the Spanish sites, instead of getting data remotely from distant sites.

At PIC we don t replicate data automatically. We improved the reliability of the disk service by using RAID6 hardware as explained in Q2.

# NCBJ-CIS

DPM with NFS shares mounted at disk nodes.

# RAL-LCG2

Ceph Object Store backend. XRootD + GridFTP gateways providing frontend. We run XRootD 'gateways' on each WN so they can directly access the backend storage. Also provide cloud APIs S3 and SWIFT. These can be access via Grid tool using DynaFed. Prefer the use of Cloud/DynaFed. GridFTP is being deprecated and XRootD is a HEP specific protocol with a very small number of experts so support is limited.

When using XRootD prefer to copy complete files (to the scratch machine of WN) as opposed to opening files and reading them directly from storage. Backend storage is optimised for high throughput and the number of things that could go wrong with a copy (as opposed to keeping files open for extended periods of time) is greatly reduced.

# T2_IT_Rome

dCache

# BNL-ATLAS

dCache is used to provide the mass disk storage service to users. It supports SRM/Xrootd/webdav protocol. We also provide NFSv4 for local dCache users. We ensure two copies of disk-only files inside dCache. We configure dedicated staging pools for interfacing with tape system. For tape, we use HPSS. For tape access optimization, we have home-developed staging request queuing system called ERADAT, plus monitoring tools for tape usage, operations and error handling.

# FZK-LCG2

dCache for ATLAS,CMS,LHCb. dCache pools are on large GPFS file systems, which makes managing the disk space and performance very comfortable. However, since no dCache pools can share the same files, a dCache pool server outage still makes data unavailable.

SLAC xrootd for ALICE. All servers have access to all files in GPFS so loadbalancing and failover in case of problems on a server a really simple.

WebDAV/xrootd are our preferred protocols. If a very simple protocol for stage requests could be implemented, we would likely consolidate our storage middleware infrastructure.

# INFN-NAPOLI-ATLAS

DPM 1.12. Available protocols srm, xrood, http, gridftp. srm could be dismissed

-- OliverKeeble - 2019-08-22

---

This topic: LCG > QosSurveyAnswersQ3
Topic revision: r3 - 2019-09-20 - OliverKeeble