New XRootD Monitoring implementation

Borja Garrido, Alessandra Forti, Derek Weitzel, Julia Andreeva, Shawn McKee

Overview

- Motivation
- New implementation
- Current status

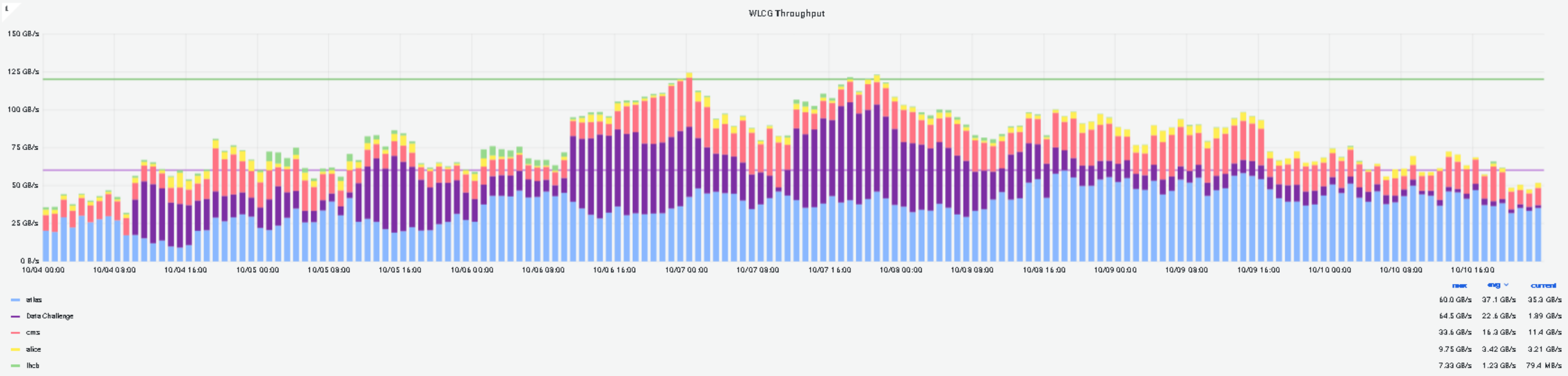
Why is it needed?

Introduction

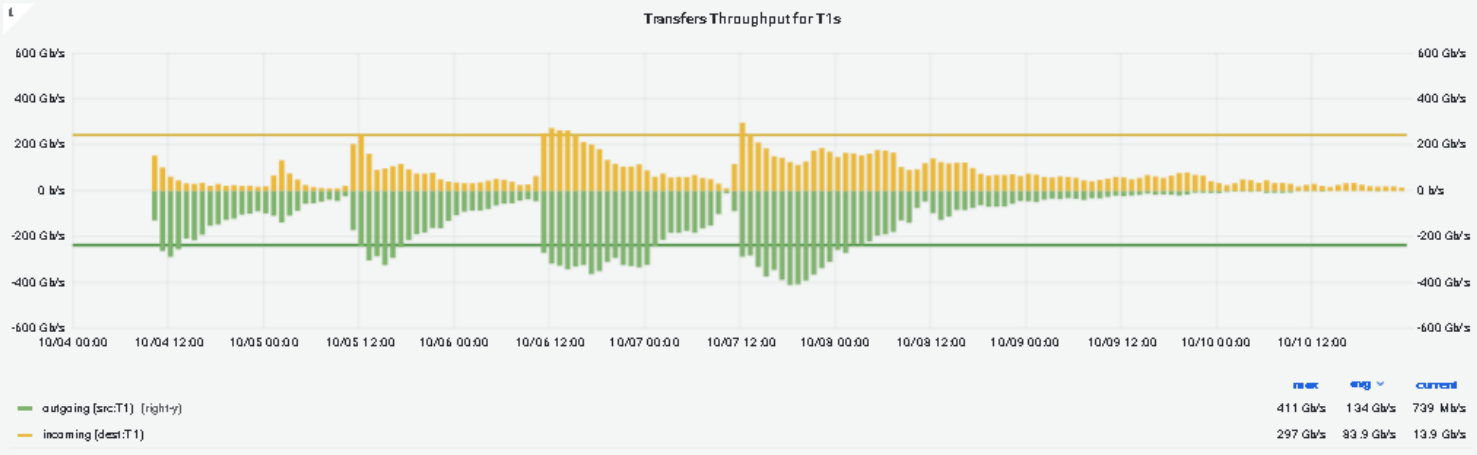
Motivation

Complete and reliable monitoring of the WLCG data transfers is an important condition for effective computing operations of the LHC experiments

WLCG data challenges highlighted the need for improvements in the monitoring of data traffic on the WLCG infrastructure, in particular remote data access via XRootD protocol



T1 Throughput (Beta)



Incoming (dest:T1)

Destination T1	Max Throughput
BNL-ATLAS	59.3 Gb/
IN2P3-CC	57.0 Gb/
FZK-LCG2	51.9 Gb/
SARA-MATRIX	41.6 Gb/
TRIUMF-LCG2	39.8 Gb/
NIKHEF-ELPROD	37.1 Gb/
INFN-T1	29.1 Gb/
JINR-T1	27.9 Gb/
NDGF-T1	21.0 Gb/
RRC-KI-T1	16.1 Gb/
pic	12.3 Gb/

Outgoing (src:T1)

Source T1	Max Throughput
SARA-MATRIX	77.6 Gb/
BNL-ATLAS	75.4 Gb/
FZK-LCG2	66.5 Gb/
IN2P3-CC	65.0 Gb/
INFN-T1	61.7 Gb/
NDGF-T1	59.7 Gb/
NIKHEF-ELPROD	50.3 Gb/
TRIUMF-LCG2	45.5 Gb/
USCMS-FNAL-WC1	35.5 Gb/
RRC-KI-T1	31.1 Gb/
JINR-T1	25.0 Gb/

WLCG Monitoring TaskForce

- **WLCG Monitoring TaskForce was presented in December 2021**
  - During WLCG Operations Coordination meeting
  - Real activities started January 2022: meetings, JIRA project…
- **Core team of ~6 people working in "best effort"**
  - Alessandra Forti, Borja Garrido, Derek Weitzel, Julia Andreeva, Shawn McKee
  - Meeting every 2 weeks for checkpointing and planification
  - Special thanks to Katy Ellis and Robert Currie for contributing on testing the new XRootD flow
- **Focused on three main areas**
  - WLCG transfers harmonization
  - **XRootD monitoring improvements**
  - Site network monitoring integration (Poster available at CHEP)

Main Goal

To provide reliable and consistent monitoring for data transfer/access by the exercise of 2024 Data Challenge

Current State → Desired State

USE CASE	STATUS
FTS	Reliable/Consistent
GLED XRootD	Not Reliable
ALICE XRootD (Monalisa)	Reliable/ Isolated
dCache + XRootD	Not monitored
xCache	Not monitored

USE CASE	STATUS
FTS	Reliable/Consistent
New XRootD - GLED XRootD - ALICE XRoottD - xCache	
dCache + XRootD	

What is needed?

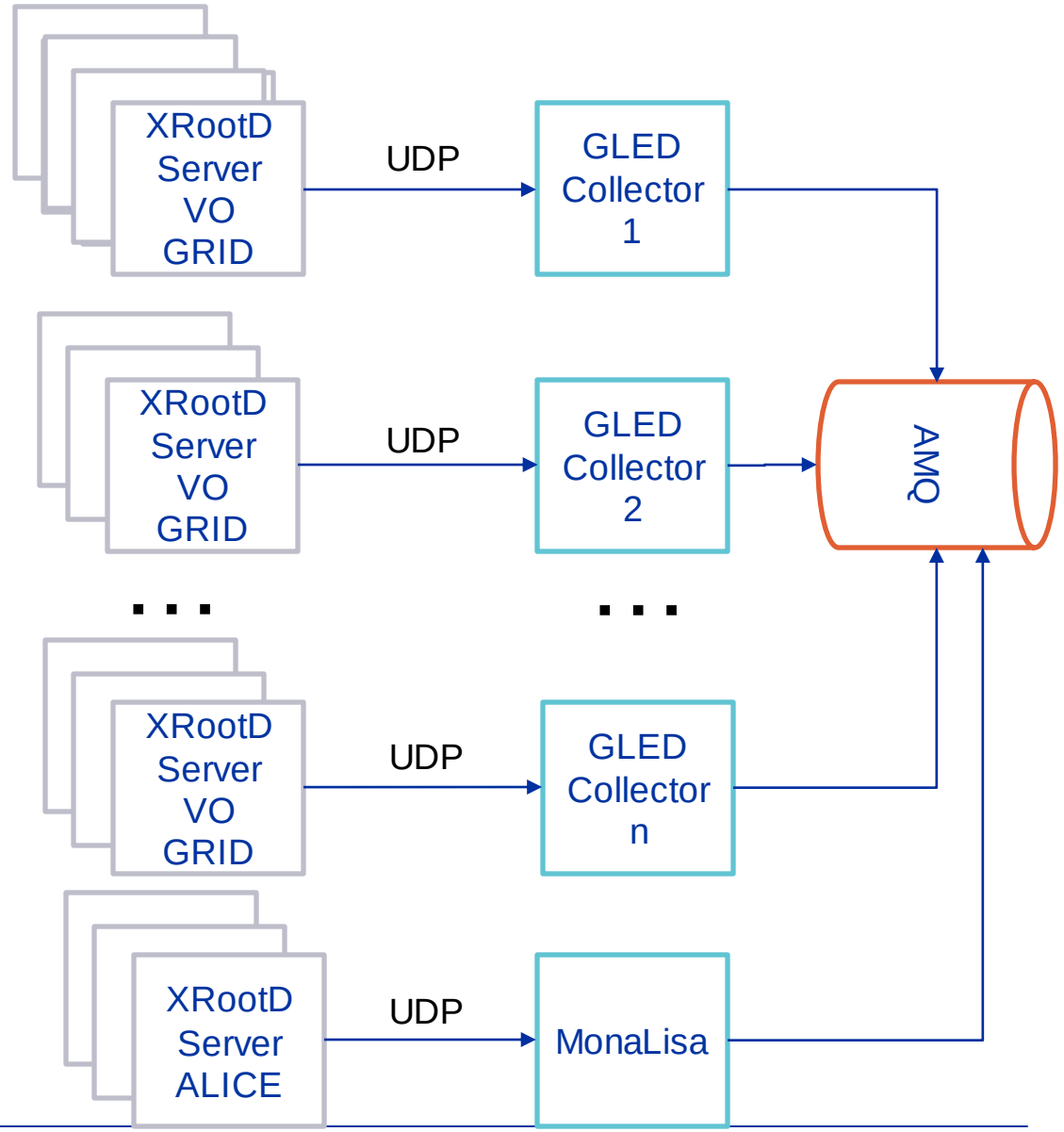
New implementation

Main Directions of Work

- **FTS monitoring**
  - Considered reliable and consistent
  - Won't be touched as part of this iteration
- **XRootD Monitoring**
  - Replace flow based fully on UDP protocol with one relying on a message bus
  - Replace ALICE specific flow with the new one common to all experiments
  - Integrate xCache
  - Integrate dCache + XRootD

Current Architecture

- Based on a **GLED** central collector that receives and aggregates streams into a “transfer document”
- XRootD servers are configured to send **UDP** monitoring streams to the central collector



Current Architecture Issues

- **OSG investigated possible cause of the issues**
  - More information in this **presentation** by **Derek Weitzel & Diego Davila**
  - Produced validation and scale validation reports

- **Main issues identified:**
  - UDP fragmentation: ~40% of streams not delivered successfully
  - Limits in the collector parallel processing: 100 streams/second
  - XRootD stateful streams
    - One transfer document composed of multiple streams
    - Single stream missing could cause full transfer to be wrong

Monitoring Packet Flow

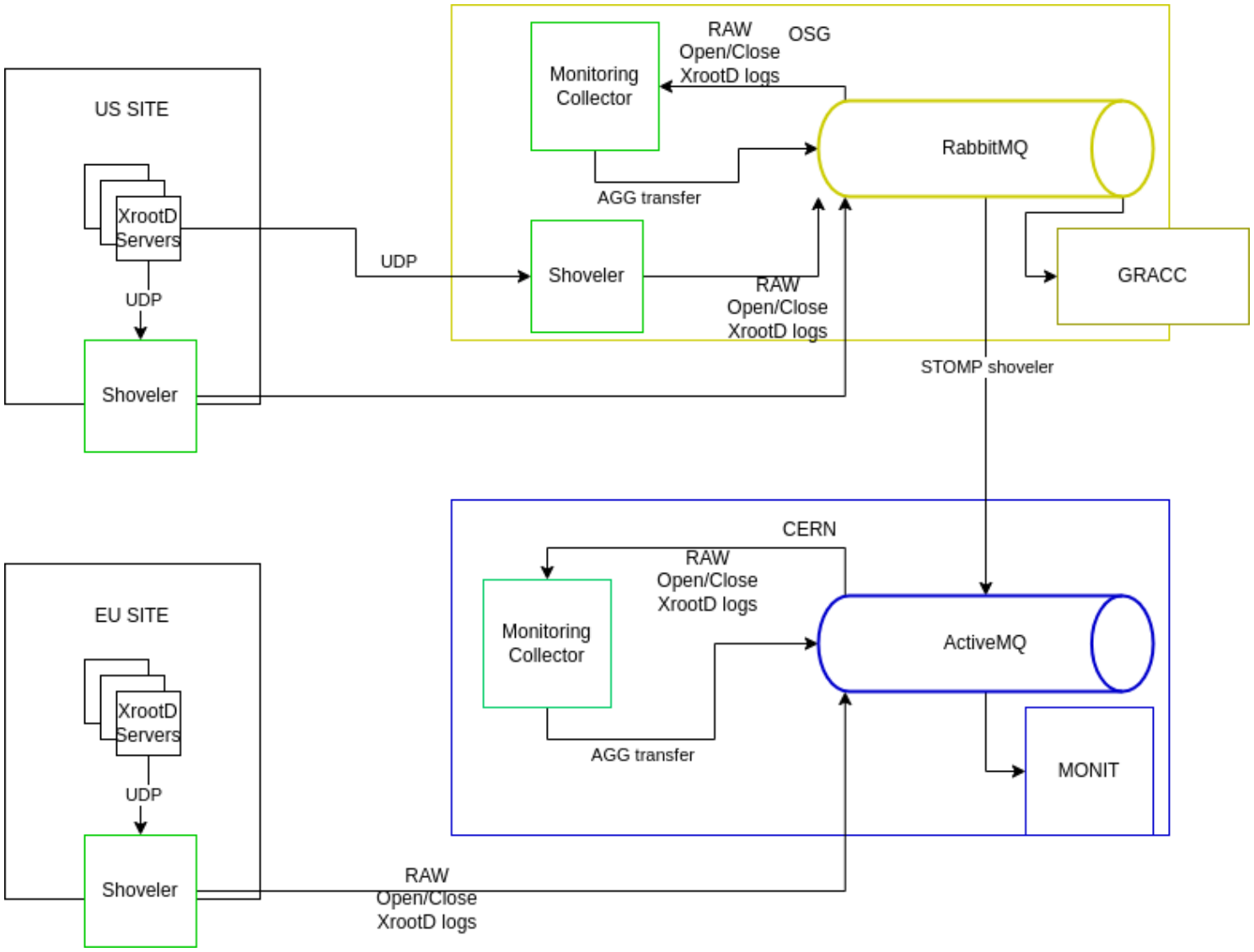
Event	Information
Client Connect	- Cert information - Client IP - Protocol - ClientID
File Open	- File Name - FileID - ClientID
Reads…	Periodic Updates - FileID - Amount Read/Write
File Close	- FileID - Total Read / Write - Total Operations

New Architecture

- **Based in two components**
  - **Shoveler**
    - Runs at sites
    - Collects monitoring UDP streams from XRootD
    - Persists them to a reliable message bus
  - **Collector**
    - Runs centrally
    - Parses monitoring messages
    - Keeps state
    - Processes streams to extract information (VO, type of transfer…)

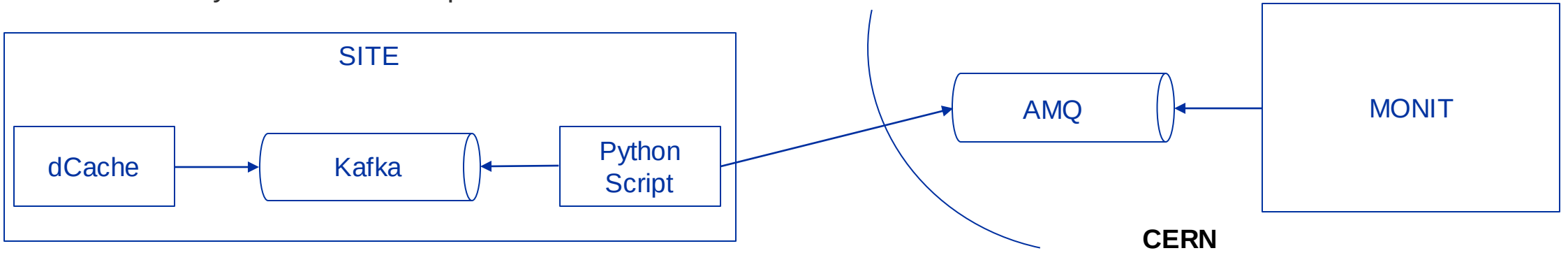
OSG

WLCG


US SITE
XrootD Servers
UDP
Shoveler
UDP
Monitoring Collector
RAW Open/Close XrootD logs
OSG
RabbitMQ
AGG transfer
Shoveler
RAW Open/Close XrootD logs
GRACC
STOMP shoveler
CERN
RAW Open/Close XrootD logs
Monitoring Collector
ActiveMQ
AGG transfer
MONIT
EU SITE
XrootD Servers
UDP
Shoveler
RAW Open/Close XrootD logs

More than XRootD servers (I)

- **dCache + XRootD (for data access)**
  - dCache data transfers monitoring is already covered by FTS
  - Especially important for CMS as pileup sitting in FNAL being accessed constantly by CMS jobs
    - Working in close collaboration with FNAL to enable new workflow there first
  - dCache monitoring flow is completely different from XRootD streams
    - Produces monitoring messages to a Kafka cluster where they can be consumed later
  - Had several meetings with dCache developers
    - Agreed on a new flow to send data to MONIT
    - Schema will be mapped to the expected one provided by the Task Force
    - Currently some issues to provide the correct destination/source IP fields



More than XRootD servers (II)

- **ALICE Monalisa**
  - XRootD servers will report in parallel to Monalisa and new shovelers
  - WLCG Monitoring information will be based on the shovelers flow
- **xCache**
  - OSG already monitors their XCache instances with this new flow
    - The same will be applied for WLCG

New Architecture: WLCG

- **New components already developed and deployed for OSG when WLCG work started**
- **XRootD Shoveler**
  - Implement communication over “STOMP” protocol to communicate with Messaging service at CERN
  - Implement certificate-based authentication to enable TLS to Messaging service
  - Deploy a test battery at CERN for integration tests
- **XRootD Collector**
  - Implement communication over “STOMP” protocol to communicate with CERN messaging
  - Deploy and run centrally for WLCG

Current Status

- **Test bed deployment running on a Kubernetes cluster**
- **Testing of the new flow with pioneer sites (Thanks!)**
  - T1 (CMS, ATLAS) T2 (Edinburgh, Manchester)
- **Troubleshooting**
  - Lack of VO information
  - Wrong "accounting" of fast transfers
  - Missing streams

Summary

- **New implementation should be ready and deployed for the Data Challenges 2024, Improved monitoring is required to perform DC24 exercise correctly**
  - Happening around Q1 2024
- **Network monitoring will allow to complement and validate information received from data servers**
- **Main effort is currently focused on the following tasks:**
  - Fixing the found issues regarding the new flow
  - Integration of dCache monitoring
  - After that we will gradually integrate more sites
- **The outcome of the WLCG Monitoring Task Force work will ensure our capability to reliably monitor all data traffic on the WLCG infrastructure**

Questions & Answers

Contact: wlcgmon-tf@cern.ch

Acknowledgements

Derek’s and Diego’s work plus Derek’s continuing effort in the task force is sponsored by the NFS Grants IRIS-HEP, Grant ID OAC-1836650

home.cern