PhEDEx – The Main Component in Data Management System for the CMS Experiment

V. V. Mitsyn, N. N. Voytishin[a]

Joint Institute for Nuclear Research, 6 Joliot-Curie, 141890, Dubna, Russia

E-mail: [a] voitishinn@gmail.com

The CMS experiment along with other major LHC experiments produces enormous amounts of data that need to be managed at the level of storage systems and distributed computing resources. Physics Experiment Data Export (PhEDEx) has been developed as a data management system for the CMS experiment and handled the assigned tasks very well so far. We present a short overview of this system pointing out its structure, features and some case study analysis.

Keywords: CMS, PhEDEx, Tier site, LHC, JINR

Introduction

One of the main functions of the JINR Multifunctional Information and Computing Complex (MICC) is to provide computing and storage resources for the main LHC experiments. The Tier1 [Astakhov, Baginyan, Belov, 2016] and Tier2 sites are particularly used for processing and storage of CMS data. PhEDEx [Barrass, Newbold, Tuura, 2005] is used for these sites as a data-placement management tool. It handles the movement of data within CMS and ensures reliable delivery of the data.

It's main components are:

- an Oracle database, hosted at CERN
- a website and data-service, which users (humans or machine) use to interact with and control PhEDEx
- a set of central agents, hosted at CERN, each taking care of specific tasks such as:
    - routing;
    - history keeping;
    - request execution;
    - et.al.

- a set of site-agents, one set for every site that receives data:
    - file download;
    - file export;
    - file deletion;
    - et. al.

PhEDEx keeps track of all the transfers that were performed. That information is used by PhEDEx central agents for selecting the source for data transferring when a transfer request is made. This way only the name of the dataset and the destination is required from the request submitter. The rest is handled by the agents.

PhEDEx instances

The set of agents of a particular site combined together form an instance. There are currently three instances in use today:

- Production;
- Debug;
- Development.

These instances are independent and there is no interaction or information sharing between each other. Putting it all together we get the operating scheme of PhEDEx, which is illustrated on Fig.1.

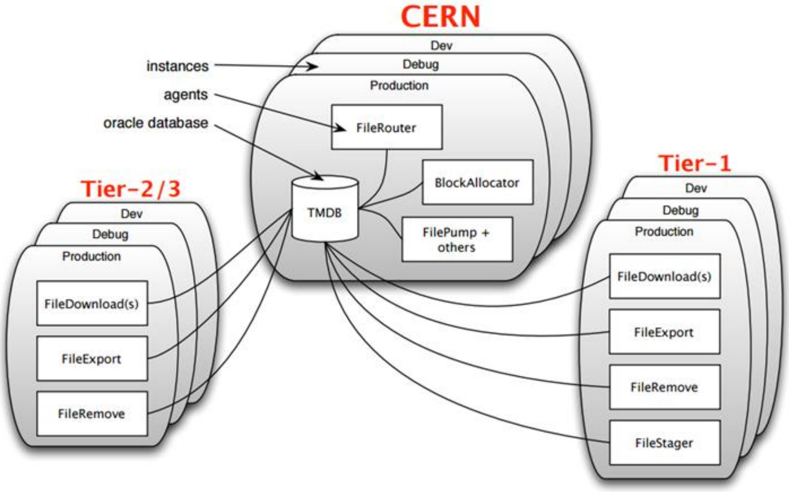

Fig.1 Interaction between sites at PhEDEx level.

The production instance is used for transferring all CMS experimental and simulated data.

The debug is a critical instance used for site commissioning and readiness. The site commissioning procedure, along with agent adjustment, was one of the most difficult challenges that we faced while setting up our sites. This procedure is needed for ensuring that there are valid links between the site that is being commissioned and other existing Tier0/Tier1/Tier2 sites. In order for a site to pass the link commissioning first all links to/from Tier0 and 8 Tier1 sites is performed. After this is done the site needs to have at least 50% of Tier2 sites to have commissioned to/from links. The commissioning of a single link consists of multiple steps. First an injection dataset is created and a transfer of this dataset is initiated to the destination site. If the transfer succeeds, we can proceed to the next steps. If not, we need to debug the transfers and fix the errors. The problem that we faced at this step was that the approval of these requests was moving very slow. This can be explained by the fact that the destination sites' admins are checking the debug instance for pending requests very rarely, because once their site is commissioned the debug instance is not needed so much. So, let's suppose that the transfer got approved and succeeded, our next step is to submit a transfer of a bigger amount of injection files in order to find out the average rate of the transfers for the initiated link. If the rate is sufficient (>20 MB/s for links from Tier1 or >5 MB/s for links from Tier2 sites) the link will become commissioned in the production instance on the next day. There can be commissioned 3-4 links per day in such a way. If there are transfer problems that cannot be resolved a link can be decommissioned in order to get rid of unreliable links.

The development instance is used for testing and validation of the PhEDEx software before its release.

PhEDEx features

The PhEDEx system has two main features. It manages all the data transfers within the CMS experiment and provides divers ways to monitor the transfer state from different perspectives.

The transfers are submitted by user through the web interface where you have to choose the desired dataset and the destination node where you want the data to be transferred. Once the transfer is approved by the destination site administrator, the movement of the data begins. The transfer state can be followed on the subscription page. Besides the subscription page, PhEDEx gives us the possibility to keep track of the transfer rates, link state, recent errors detailed logs and many others. This makes it a very helpful monitoring tool for site administrators as well as for simple users.

PhEDEx controls only the reception of data on the site that it’s installed. The transfers from a particular site are handled by File Transfer System (FTS) once the destination site administrator approves the PhEDEx transfer request. Meanwhile the source site has no possibility to manipulate this transfer through PhEDEx. The only possible tool for this is the stage-in procedure. This is available only for sites with tape-robots and is meant to control the reading of files from tapes to disk buffers.

One more very useful feature implemented through PhEDEx is the file invalidation procedure. Ones the file gets lost or broken on the site's storage element (SE) it can be invalidated. For this purpose a ticked must be submitted to the central admins with the name of the file indicated. If copies of the requested file are present on other sites, this file is invalidated locally, triggering the re-transfer of the file to the requester site. In case the file is registered only at the requester site, it is invalidated globally. This means that the file is lost irretrievably and all its records are deleted from the CMS database. The described procedure is very time consuming, because local site admins have to gather the list of files for invalidation almost manually, searching for these files in the error logs or running scripts over our SEs. Our experts came up with a utility that is meant to make the list gathering easier. This utility scans the error logs by error types and makes the lists of files for each node. This way we can form invalidation lists from time to time in order to get rid of the bad files on our SEs lowering the number of errors at our site.

The entire PhEDEx is written in Perl and gives developers the opportunity to add new plugins that extend the data management capability.

Conclusion

PhEDEx is a continuously evolving system. New data management protocols and utilities are added to optimize and enlarge its functionality. This system proved to be stable and robust through Run 1 of the LHC. It easily deals with the current challenges of about 100 TB/day per site. PhEDEx is not CMS specific. It only requires a hierarchical description of data and it doesn't care about the file type. All that PhEDEx needs for identification and checking is the file size and checksum.

This system flexibility along with the enormous experience in PhEDEx configuration and exploitation acquired over the past years by JINR experts makes PhEDEx a good candidate for the large scale projects at JINR. The best placed candidate is the NICA Project [Bashashin, Kekelidze, Kostromin, Korenkov, Kuniaev, Morozov, Potrebenikov, Trubnikov, Philippov, 2016]. The simulation of NICA computing and dataflow [Korenkov, Nechaevskiy, Ososkov, Pryahina, Trofomov , Uzhinskiy, 2016] are well suited in terms of PhEDEx usage.

References

Astakhov N.S., Baginyan A.S., Belov S.D. et al. JINR Tier-1 centre for the CMS Experiment at LHC. Particles and Nuclei, Letters, v.13,no 5, pp.1103-1107, 2016.

Barrass, Newbold and Tuura, The CMS PhEDEx System: a Novel Approach to Robust Grid Data Distribution, UK e-science Programme All Hands Meeting, Nottingham, UK, 2005

Bashashin M. V., Kekelidze D. V., Kostromin S. A., Korenkov V. V., Kuniaev S. V., Morozov V. V., Potrebenikov Yu. K., Trubnikov G. V., Philippov A. V. NICA project management information system Particles and Nuclei Letters, v.13,no 5, pp.969-973, 2016.

Korenkov V. V., Nechaevskiy A. V., Ososkov G. A., Pryahina D. I., Trofomov V. V., Uzhinskiy A. V. Simulation concept of NICA-MPD-SPD Tier0-Tier1 computing facilities Particles and Nuclei Letters, v.13,no 5, pp.1074-1083, 2016.