Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

# Space Usage Monitoring
## for distributed heterogeneous data storage

Natalia Ratnikova

HEPiX workshop at BNL

15 October 2015

# New to HEPiX. Many years in HEP software and computing.

**1997** Offline software librarian for HERA-B, DESY Hamburg

**2000** Fermilab computing division, CMS group

- Software release integration, packaging and distribution
- US CMS GRID testbed integration
- Users support, LPC center
- CMS software development tools group

**2008** Scientist at IEKP, KIT, Germany

- CMS Tier-1 site support, data management, site monitoring
- Built CMS remote control center, trained CMS computing shifters
- PhEDEx development, data consistency tools, SpaceMon

**2012** Fermilab CMS-T1 facilities team

- Data storage administration (dCache)
- Data transfers (PhEDEx, xRootd)
- Distributed computing services and operations
- CMS Space Monitoring project lead

# Outline

- CMS data storage overview
- Motivation for Space Monitoring
- System architecture and components
- Deployment campaign
- Conclusions

# CMS storage resources by the beginning of LHC Run 2

- CMS Tier 1 and 2 storage space requirements* :

| Year | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|
| Tier 1 Disk | 26,000 | 26,000 | 26,000 | 33,000 |
| Tier 1 Tape | 50,000 | 55,000 | 74,000 | 100,000 |
| Tier 2 Disk | 26,000 | 27,000 | 29,000 | 38,000 |

- Increased pileup, higher HLT rate, data parking and scouting
- Volume will grow proportionally to LHC life time
- Phase 2 detector upgrade studies

➡ CMS expects severe resource constraints

* *Values are given in Tbytes according to WLCG-rebus pledges summary*

# Evolution of the computing model

- Changed patterns in organized data processing
- Tier 1 disk and tape separation
- AAA xrootd driven data federations
- Dynamic data management
- New data types:
  - MiniAOD
  - phase 2 detector studies
  - parked data
- Diverse user analysis patterns
- Increased share of storage space for users and groups

Multiple data placement processes not necessarily aware of each other sharing the same storage resources

# Space monitoring for distributed storage

- CMS data live in a **global name space**, addressed by a logical file name (LFN), e.g.:

  /store/data, /store/mc, /store/user, /store/group, …

- Data are accessed by physical file names (PFNs) according to the LFN to PFN translation rules specified in the trivial file catalogs provided by the sites
- Space monitoring allows to track the space occupied by each level under /store across the sites.
- CMS central Transfer Management Database keeps track of data maintained by PhEDEx.
- Information on other files, users data, temporary production and test data, is only available from the direct storage dumps.

# Space monitoring for distributed storage

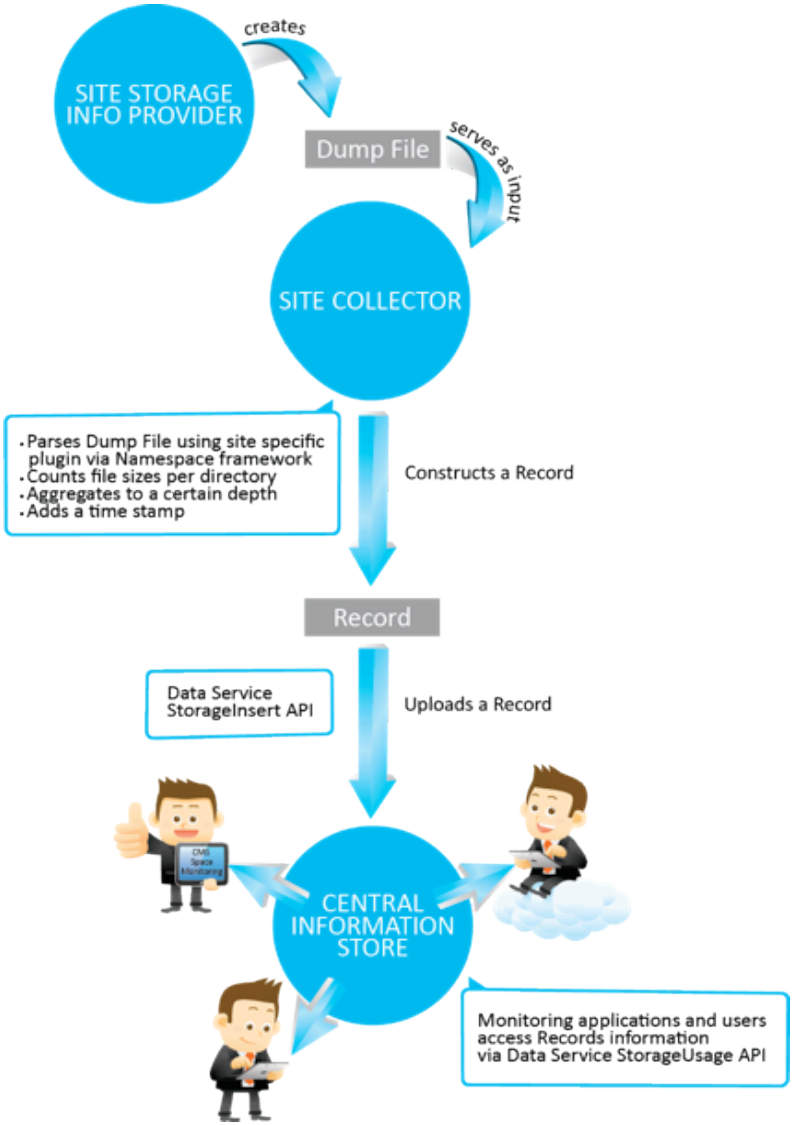
- CMS data live in a **global name space**, addressed by a logical file name (LFN), e.g.:

  /store/data, /store/mc, /store/user, /store/group, …
- Data are accessed by physical file names (PFNs) according to the LFN to PFN translation rules specified in the trivial file catalogs provided by the sites
- Space monitoring allows to track the space occupied by each level under /store across the sites.
- **Main use cases**:
  - Efficient space utilization
  - Fair share between users and groups
  - Resource planning

## Related initiatives

- **CMS monitoring task** force identified a gap in storage accounting (2008)
- The idea for solution came naturally during **validating data consistency** of the new CMS dCache instance at GridKa separated from ATLAS and LHCb
- **Syncat format** presented by Paul Millar at CHEP'09 – demonstrated abstraction from the storage technology
- Proposed the idea as a "**demonstrator**" at WLCG storage jamboree in Amsterdam 2009, supported by Tony & Daniele
- Elisa Lanciotti (WLCG ES group) work on identifying common formats and **tools for storage dumps** at Tier-1 sites
- **CMS Tier 2 data consistency campaign** launched in 2012
- Joint CMS, Atlas and LHCb presentation at **CHEP'12**

# General architecture and workflow






Fermilab

# Components

- Site information providers ( storage dump tools)
  - Storage technologies: Castor, dCache, DPM, EOS, Hadoop, LStore, Lustre, StoRM.
- Site collector (client tool)
  - Aggregates information and uploads to DMWMMON oracle database at CERN
- Central information store: web based data service
  - provides interfaces to upload and retrieve space usage records
  - APIs for authentication, list of sites, and troubleshooting access

# Deployment campaign

Issues encountered during this first stage of this deployment can be categorized into three groups:

1. Questions from sites about why they need to provide storage usage information and at what level of detail
2. Authentication problems uploading the information to the central data service
3. The long time it takes to take a dump for some storage systems.

Also some privacy and security concerns were raised by the sites.

# Conclusions

- CMS has developed Space Monitoring system based on storage dumps and successfully deployed it at across Tier 1 and majority of Tier 2 sites
- Many improvements were done and some still due based on feedback on deployment at the sites. Also CMS specific requirements have been refined.
- More experience, improvements and debugging work are in progress
- The system provides a solution to a general problem without special assumptions about data management model, the interpretation and presentation of the results will reflect the particular model.