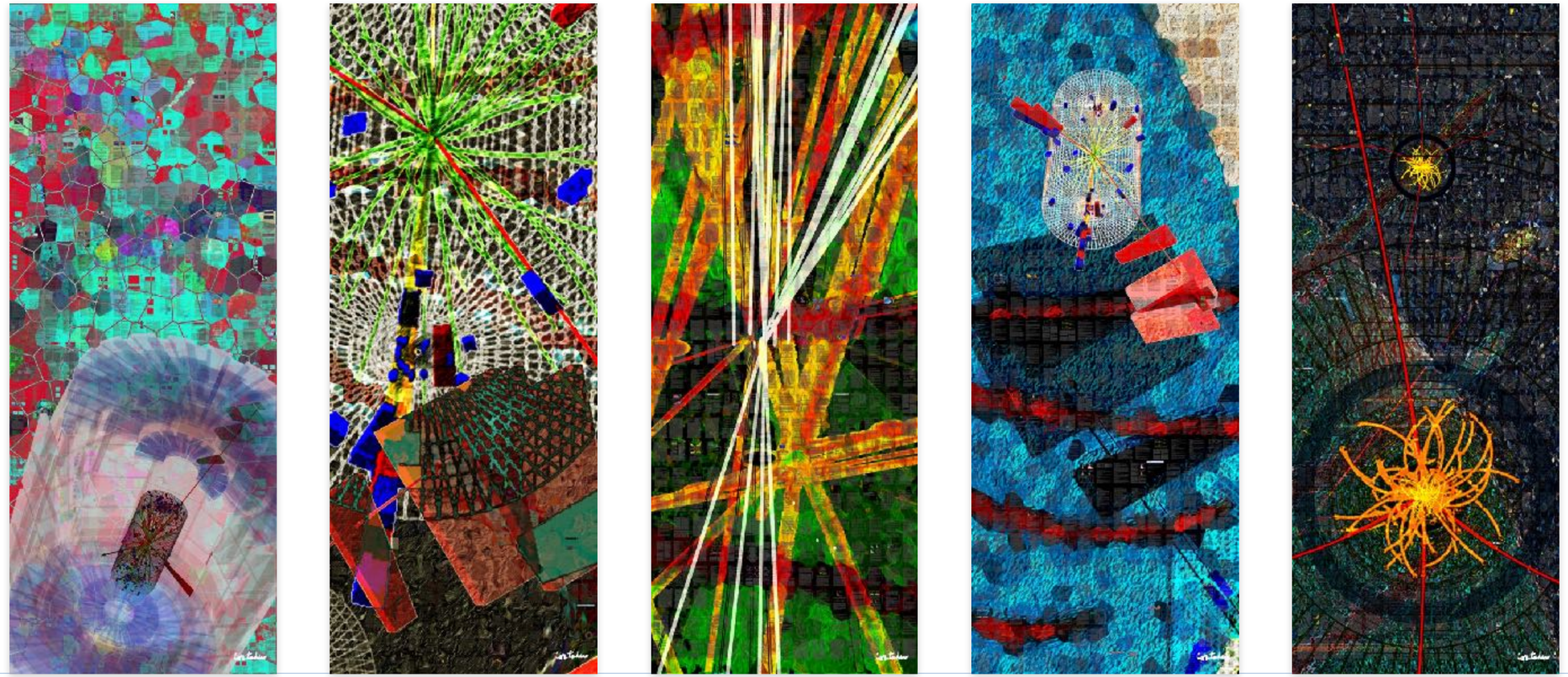
CMS Evaluation of Rucio

Brian Bockelman, Bo Jayatilaka,
Natalia Ratnikova, Eric Vaandering



CMS Data Management Needs

- Our usage currently falls mostly in the “replica management” category
- Current statistics on our data storage and movement
  - Stored on tape O(100 PB)
  - Stored on disk O(50 PB)
  - Production file size O(1 GB), user file size O(100 MB)
  - Per day transfers ~1 PB, 1 M files (combined user, production)
  - 8 sites with tape, O(50-100) with managed disk
- Currently CMS has two DM systems
  - Production data is fully managed
  - Some user data is “lightly managed” (catalogued when produced, not able to be moved)
  - Some user data is completely unmanaged
- Numbers stay more or less constant for next 7-8 years, go up 50x in 2026 and beyond

Current CMS Data Management Situation

- CMS data stored in a three tiered structure:
  - Files - target size 4 GB
  - Blocks - usually about 100 files, designed to be a unit that can be stored and transferred at one site
  - Dataset - some number of blocks, has a physics meaning (often stored all at a site, but no necessarily)
  - All 1:many maps, not many:many (unlike rucio)

- Primary data management is done by PhEDEx
  - Each site hosts a PhEDEx agent to manage its own data. Also manages local tape
  - Maintains a database of the desired states (blocks at sites) and issues FTS commands to achieve it
  - FTS is only one option for moving data, but ~all disk to disk is done with it
- A higher layer, Dynamo, monitors popularity of data and, based on rules, makes subscriptions to dynamically distribute popular data
- DBS (Data Bookkeeping Service) is our meta-data catalog.
  - Shares the same description of Files (including size & checksum), Blocks, Datasets with PhEDEx
  - Also stores physics metadata on the file, block, dataset level

Current CMS User Data Management

- User data is in DBS only, PhEDEx and Dynamo have nothing to do with it
- Produced at Site A, moved to Site B (user has a relationship with Site B)
  - User typically has 2 TB of storage at Site B as part of site pledge to CMS
- Can never be moved to Site C and reflected in DBS
- All done with AsyncStageOut (ASO) which is a thin layer on top of FTS
  - Considering removing even this thin layer

- CMS has largely given up the idea of storage for physics groups

- (I've kind of lied, we have an infrequently used process to turn user data into official data)

- Choices driven by tools, may re-evaluate if we adopt Rucio

CMS Review Process

- Two data management reviews in the last two years
- PhEDEx is aging and we realize its lifetime is limited
  - Now on third generation of developers
  - Overly complex in functionality and operations (effort needed at every Tier2 site)
  - Written in perl and some dependencies are now abandoned
  - Re-tuning as network capacities increase is necessary
- PhEDEx is capable of Run3 data transfers, probably not ideal
  - No confidence it can survive in the HL-LHC era

- Currently exploring two alternatives:
  - In house extension of Dynamo to handle transfers and eventually the catalog
  - Evaluation of Rucio - rest of this talk

- Parallel and related effort by FNAL for current and future experiments - Rob's talk later

CMS Rucio Evaluation

- November: Agreed to do a Rucio evaluation for CMS mostly with Fermilab effort
- Targeted towards July reviews of possible solutions, plan to pick one shortly thereafter
- A few people part time, more recently two collaborators from INFN joined
- Started with trying to install
- Then were offered and used U Chicago server for CMS
  - Many, many thanks to Judith Stephen and Benedikt Riedel!
- Spent some time familiarizing ourselves with subtle differences between ATLAS and CMS models

Evaluation continued

- Spent some time familiarizing ourselves with subtle differences between ATLAS and CMS models
- Remember CMS has Dataset - Block - File
  - Not perfect but fits OK into Rucio model:
    - ★ CMS Dataset - Rucio Container
    - ★ CMS Block - Rucio Dataset
- Also some differences in terms of how storage elements are thought of
- Everyone has also had to get familiar with the rucio CLI tools and concepts
- Given the holidays, progress has been reasonable
- Recently achieved milestones:
  - Inject CMS-like data structures into Rucio (Dataset, Block, File)
  - Upload and transfer files between existing CMS storage end points
- Still to do:
  - Full dataset transfers, rules, and then on to scale testing
  - Tape staging

Interaction with developers

- So far our interactions have been extremely positive. (Almost) all differences between CMS and ATLAS models have been addressed with new development
  - Lots of time spent by both groups understand each other's models
- Recall that CMS has O(200 PB) of existing data. Any transition needs to work with CMS data as is, not physically copying into new system with new names
- Special characters in CMS datasets: / and #
  - Addressed in latest release. Problematic because these characters are special in REST interface as well
- CMS has an existing namespace: /store/dir/subdir/etc/filename.ext (Logical filename)
  - Typically physical filename on a storage device is prefix + LFN
  - Rather different than ATLAS, but also accommodated without too much trouble
- CMS gave up on "scopes" sometime back (e.g. physics groups no longer manage their own data).
  - Not a problem, use a single scope and only use others for limited purposes

Stray questions

- 2e: Workflow management system is WMAgent which is based on GlideinWMS and Condor.
  - Currently its WMAgent which determines where data is located and sends jobs to that data
    - ★ There are other operation modes, but this is the dominant one
  - We have an outer layer which can modify idle jobs should data move
  - Integration between Condor and Rucio to determine the appropriate resources could be very useful

Conclusions

- Rucio meets CMS's immediate scalability needs and is a good enough fit to our existing data model
- Rucio developers have been very accommodating and encouraging
  - It is a concern for CMS if the effort continues to be owned by ATLAS
  - Community project would be ideal
- U Chicago system has been instrumental in helping with a quick start
- We still have lots of milestones to meet to show that CMS could adopt Rucio, but we are all optimistic they can be met before the summer review
  - Transition would take place during 2019-2020 LHC shutdown
  - Still need to map out exactly how this would happen
    - ★ What are the possibilities for running both our systems in parallel for a while