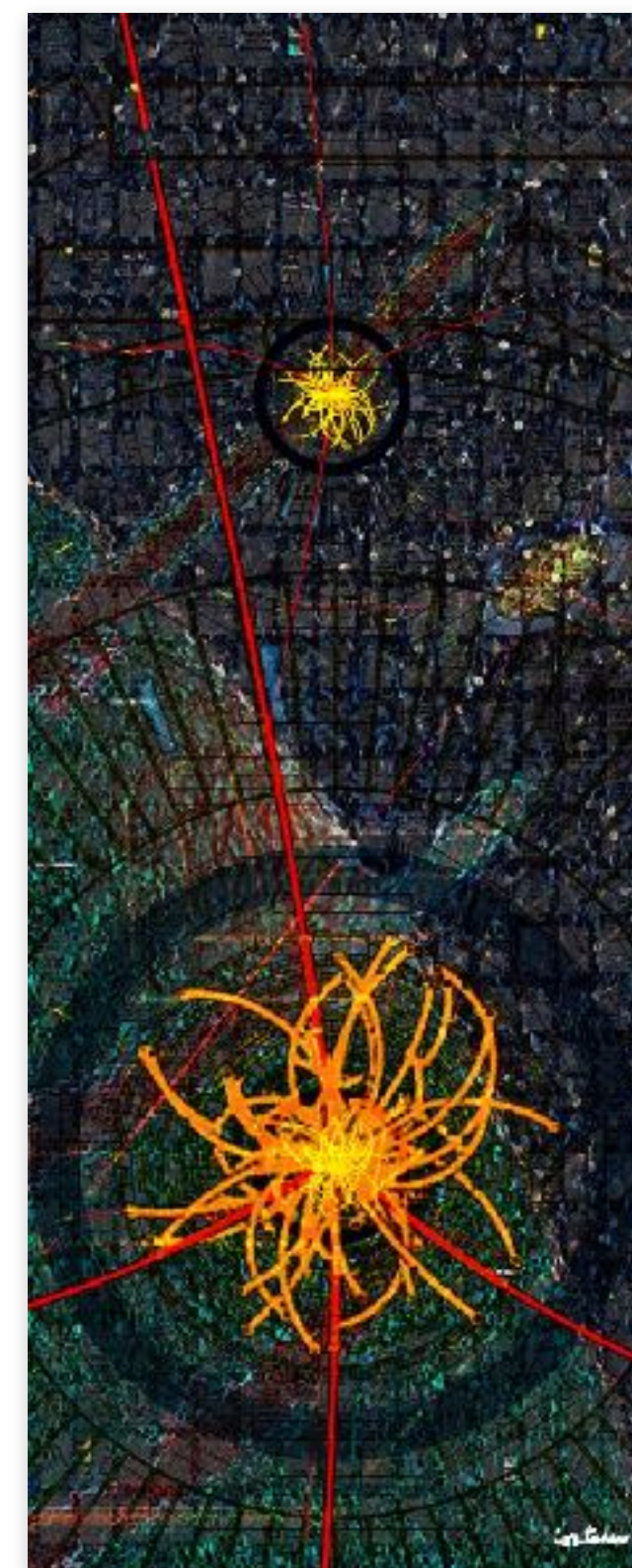
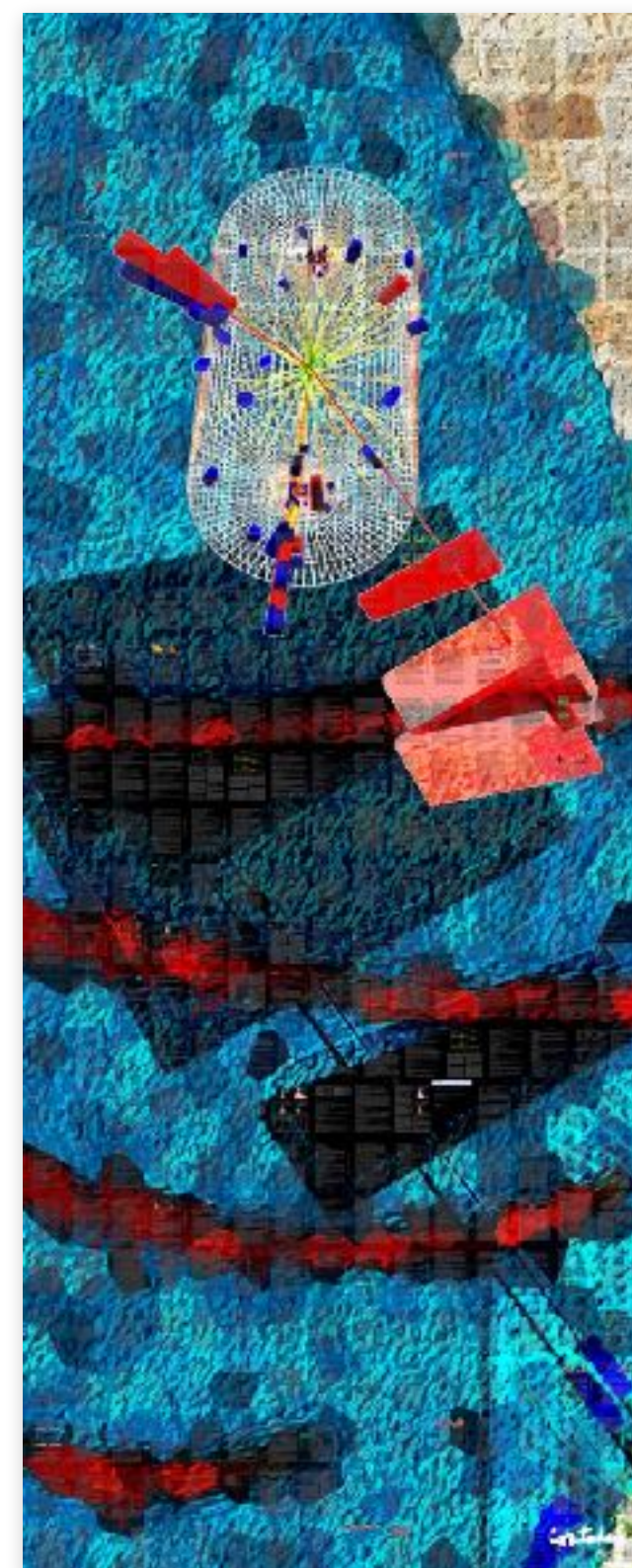
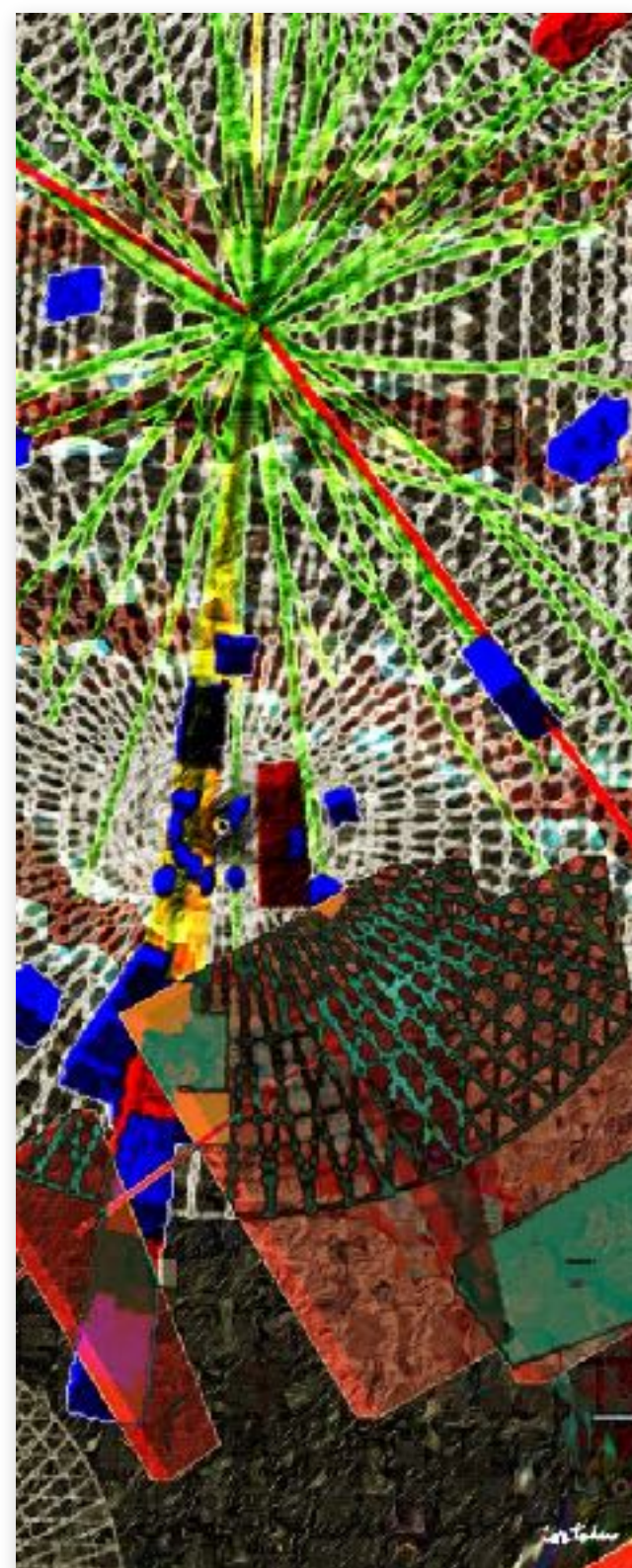
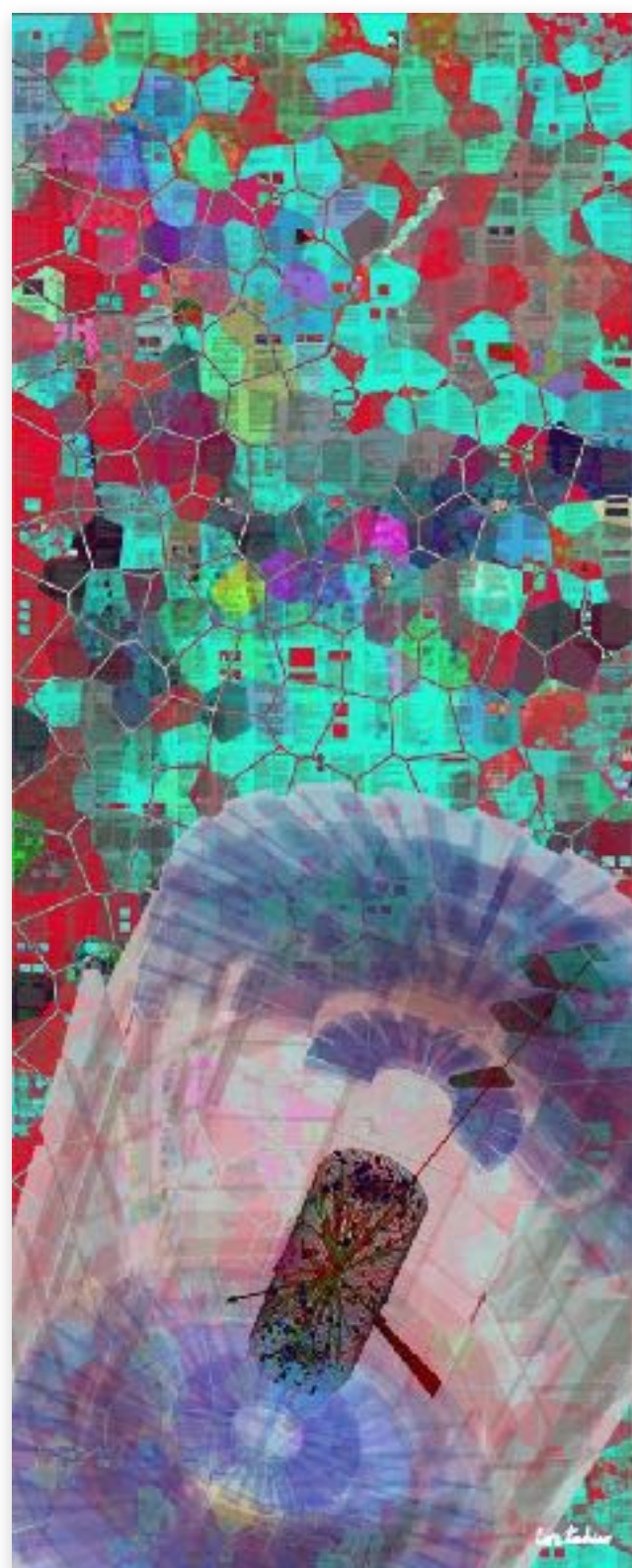


# CMS Evaluation of Rucio

Brian Bockelman, Diego Ciangottini, Bo Jayatilaka,  
Natalia Ratnikova, Andrea Sartirana, [Eric Vaandering](#)



# CMS Data Management Needs

- **Current statistics on our data storage and movement**
  - Stored on tape  $O(100 \text{ PB})$
  - Stored on disk  $O(50 \text{ PB})$
  - Production file size  $O(1 \text{ GB})$ , user file size  $O(100 \text{ MB})$
  - Per day transfers  $\sim 1 \text{ PB}$ , 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**

- 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 typically 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, cleanup unpopular
- 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
- All done with AsyncStageOut (ASO) which is a thin layer on top of FTS
  - Considering removing even this thin layer
- Can never be moved to Site C and have that reflected in DBS
  - I've kind of lied, we have an infrequently used process to turn user data into official data
- CMS has largely given up the idea of storage for physics groups
- Some choices driven by tools, may re-evaluate if we adopt Rucio

- 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 capabilities increase is necessary
- No confidence PhEDEx can survive in the HL-LHC era
  - We have agreed we need to replace PhEDEx by the end of LS2 (before Run3)
- 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

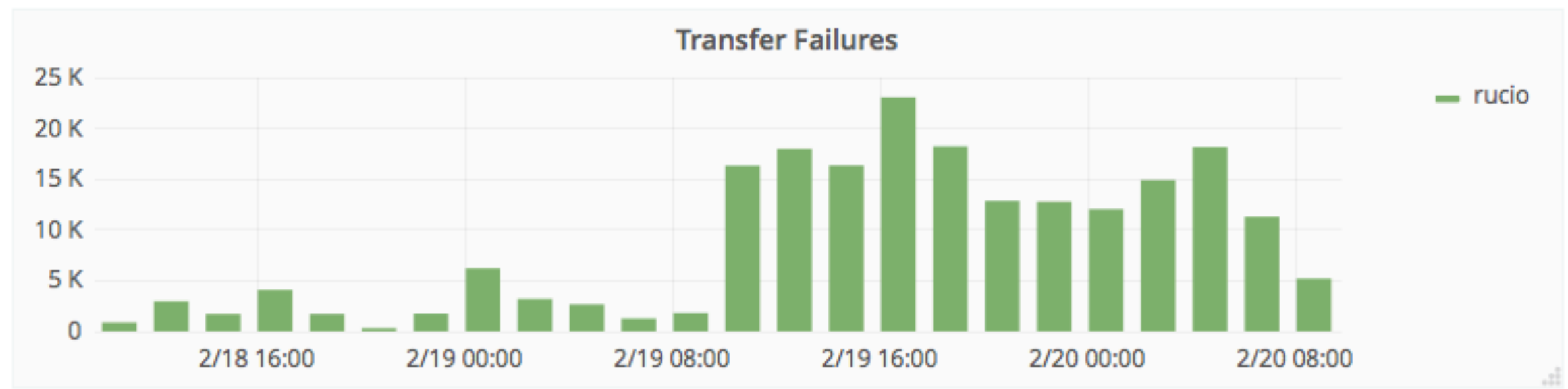
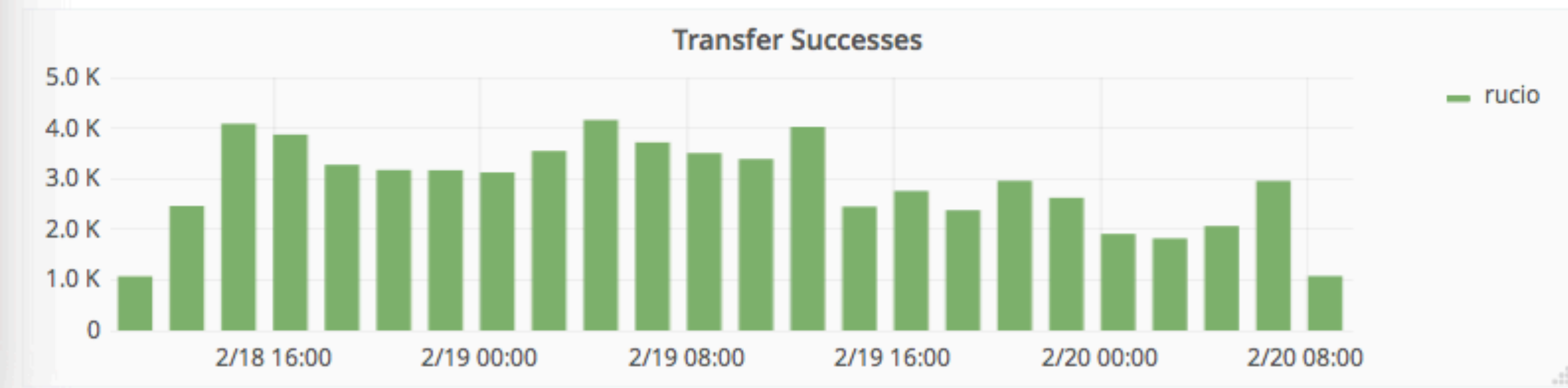
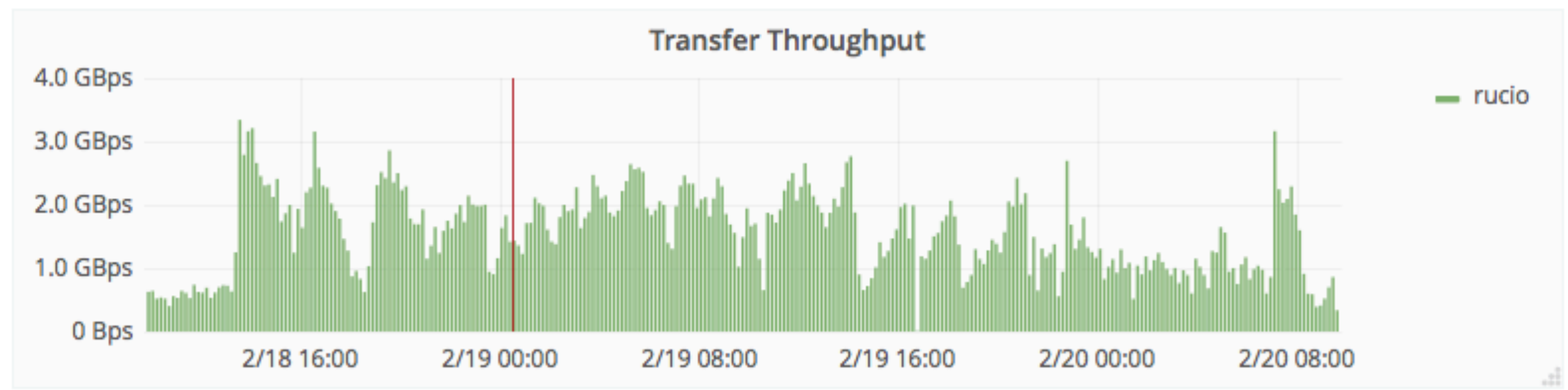
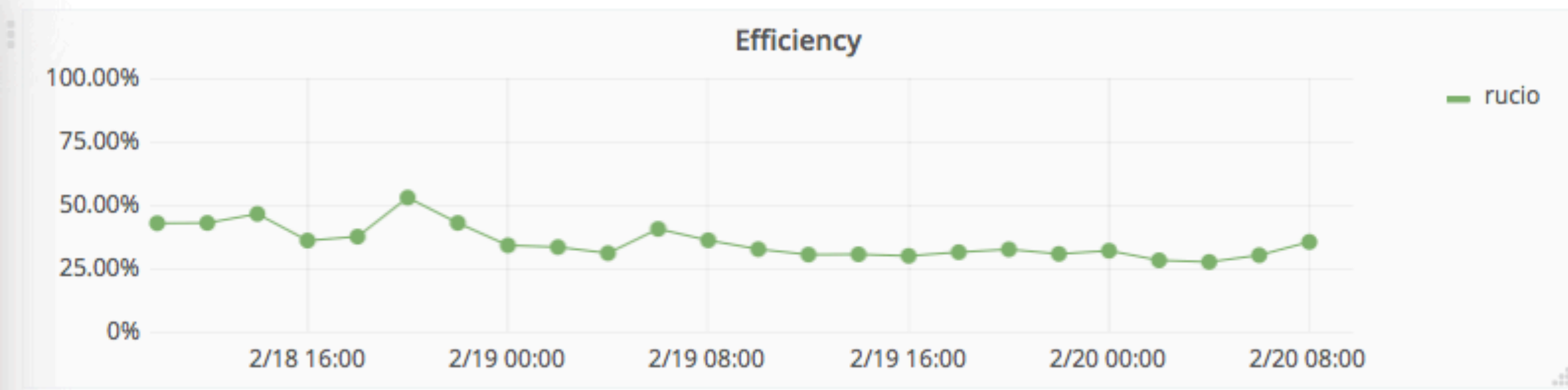
- November 2017: Agreed to do a Rucio evaluation for CMS mostly with Fermilab effort
- Targeted towards July 2018 reviews of possible solutions, plan to pick one shortly thereafter
- A few people part time
- Spent some time familiarizing ourselves with subtle differences between ATLAS and CMS models
- Currently using a server graciously provided by U Chicago
  - Working part time on setting up our own to familiarize

- 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
- Progress has been reasonable
- Recently achieved milestones:
  - Set up two r/o RSEs that are populated with CMS data at existing sites, about a dozen test r/w RSEs
    - ★ Total system size: ~25 RSEs, 3.7 PB tracked, approaching 1 PB transferred
  - Small dataset transfers and distribution using rules (small scale test)
  - Transfer of 600 TB underlying event sample to NERSC HPC site
  - Exercised deletions on rule expiry
  - Initial investigations replacing ASO (user data) with rucio scopes
- Still to do:
  - Tape staging
  - Larger multi-site scale test

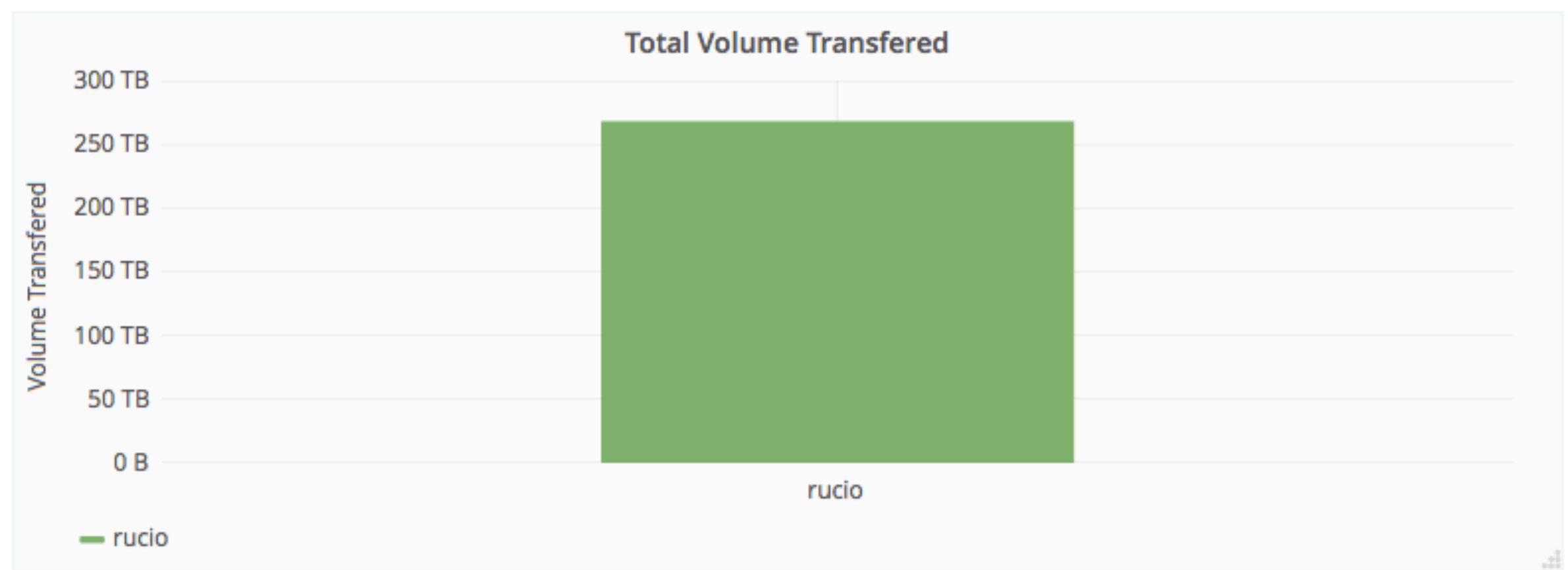
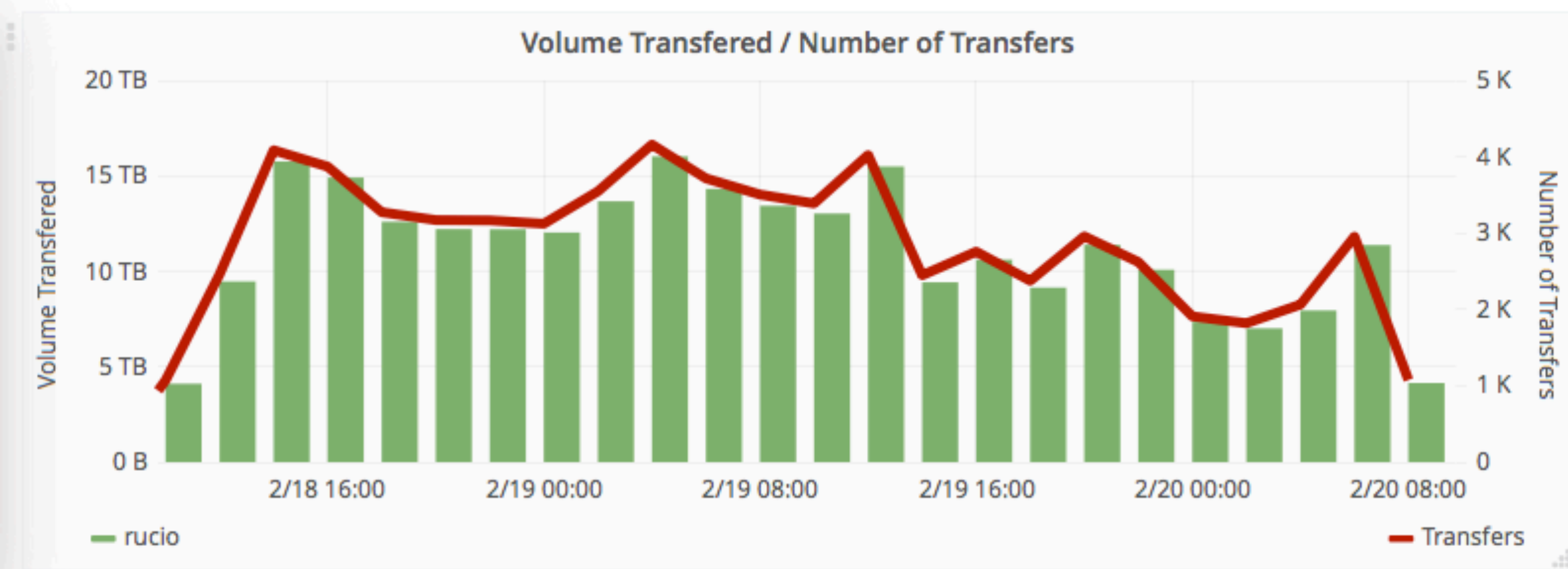


# Transferring 600 TB dataset

FTS Transfers (LAST 30 DAYS)



Volume Statistics





# 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  $O(200 \text{ PB})$  of existing data. Any transition needs to work with CMS data as is, not physically copying into new system with new names
- We now have submitted multiple pull requests to Rucio and had them accepted quickly
  - Special characters in CMS datasets: / and # (special in REST interface)
  - Fine grained control over physical and logical file names
  - CMS does not use space tokens
- Still to come
  - WebUI
  - Model for operations, including debugging
  - Model for interfacing popularity-generated rules
  - LFN to PFN mapping function that replicates the real situation
    - ★ Some sites are simple, some complex
  - 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

- 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 some milestones to meet to show that CMS should adopt Rucio, but we are all optimistic they can be met before the summer review
  - We now know we *could* adopt rucio
  - Transition would take place during 2019-2020 LHC shutdown
  - Still need to map out exactly how this would happen
    - ★ Need to explore how to run both in parallel