Updates on Data Carousel

Xin Zhao, Alexei Klimentov (BNL)
ATLAS Site Jamboree, CERN, March 8, 2019

Outline

- Big picture overview
- What’s happening with WFMS/DDM ?
- What’s happening with sites ?
- Next steps

** Team effort --- WFMS team, DDM team, Ops team, Alessandro Di Girolamo, Johannes Elmsheuser and ADC experts, all T0 and T1 storage and tape experts*

Big Picture Overview (1/2)

- We are at phase II of the Data Carousel project
  - From the ADC data carousel (live google doc) :

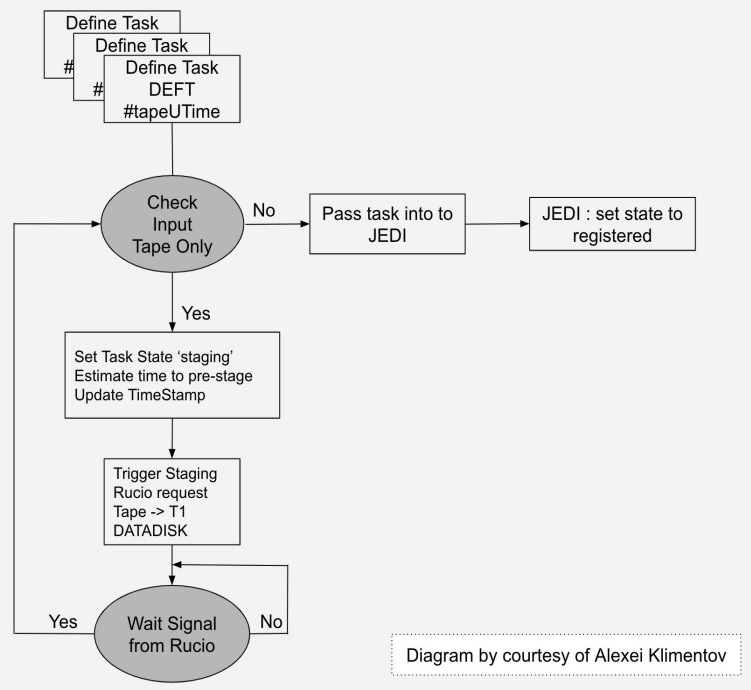
    “... In the second phase we will address the issue of data retrieval/exchange between tape/disk with our data management and workflow management systems: this second phase of R&D will require a deeper integration between the two main distributed computing components... ”

Big Picture Overview (2/2)

- What's happening in Phase II
  - WFMS/DDM (google doc to follow up with progress)
    - Algorithms for intelligent data pre-staging
    - Provide monitoring to sites and users
    - Intelligent writing (including bigger files to tape)
  - Sites
    - tape systems continue to evolve
    - WLCG archival storage WG activities

WFMS/DDM: intelligent data pre-staging (1/5)

- Pre-stage step being added to task definition in ProdSys2, for data from tape
  - Each production request will be assigned a unique (new) hashtag #tapeUTime, for scheduling and tracking staging requests
  - Final task brokerage and possibly datasets redistribution will be done by JEDI
  - JEDI may check for all #tapeUTime tasks
  - JEDI may delete datasets disk replica at T1s after brokerage is done



WFMS/DDM: intelligent data pre-staging (2/5)

- Pre-stage policy : ensure bulk mode
  - Don’t start pre-stage if minimum limit is not reached or grace period is not passed;
  - Don’t start pre-stage if maximum limit is reached
  - Estimate time to finish pre-stage ?
- Use metrics obtained from phase 0, to define site storage characteristics: min/max I/O limit, average throughput. (see table to the right, more details in this google doc)

site	[1][2] Bulk size (max) #nfiles	[3] Bulk size (min) #files	[4] Tape throughput (upper) MB/s	[5] Tape throughput (average)) MB/s
BNL	200000	5000	850	550
FZK	2000		300	300
INFN			300	250
PIC	10000		400	400
TRIUMF			1000	700
CCIN2P3	3500		3000	2000
SARA-NIKHEF	3500		600	600
RRC-KI	50000	>5000	1400	150
RAL	100000		2000	1500
NDGF	200000		500	300

WFMS/DDM: intelligent data pre-staging (3/5)

- In the long run, “intelligent staging” will respect priorities, shares, computing and storage availability

Near-term Plans

PS2/DEFT :
implement stagein messaging
implement task staging status/pass info to JEDI
implement info/warn/alarm for tasks in ‘staging’
implement tapeUTime

Monitoring :
implement tasks grouping in staging state
implement tasks grouping by #tag : tapeUTime

Rucio :
implement staging message
implement #tag meta-info
implement the “grouped FIFO” model in staging requests

Long-term Plans

PS2/DEFT :
implement priority mechanism; implement “staging policy” on the level of request
implement request cancellation
implement pre-stage lifetime (~2 weeks)
implement automatic stage out for datasets

JEDI :
implement staging to remote site
implement replicas deletion after dataset redistribution

Rucio
implement priority mechanism
implement watermark check (automatic redistribution ?)
implement intelligent staging

WFMS/DDM: intelligent data pre-staging (4/5)

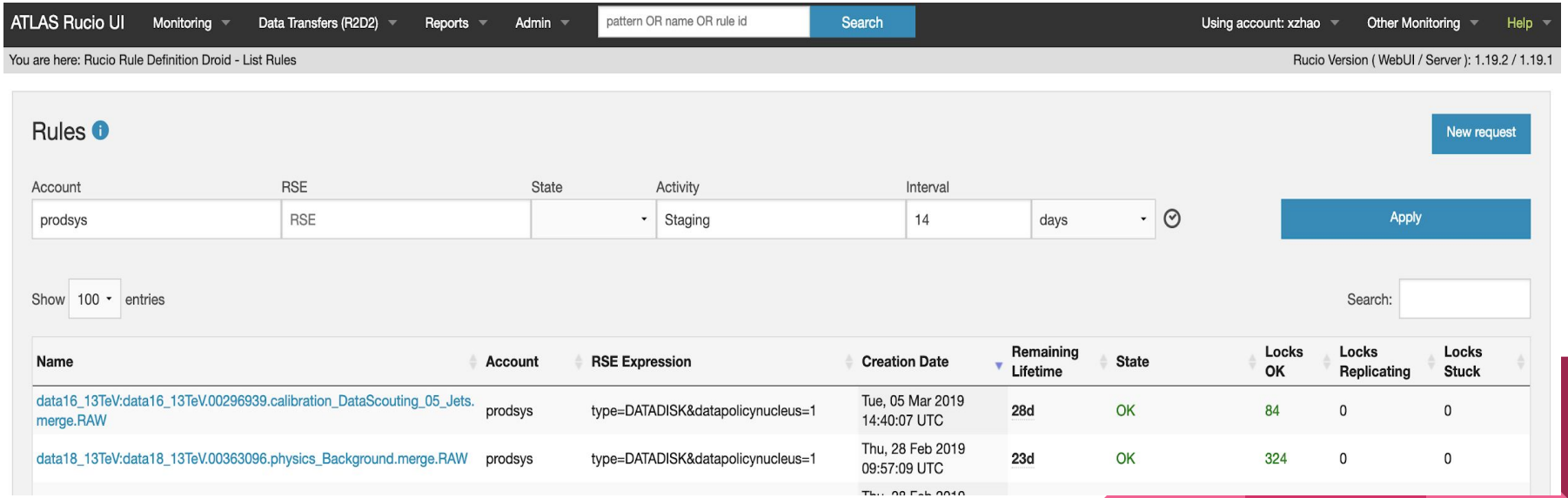
- A recent mini-exercise (Misha, Rod)
  - Several real production requests were submitted with pre-stage option
    - Mostly reprocessing tasks
    - All inputs are on tape
    - Data distribution among T1s
      - 9 out of 10 T1s were used.
      - Distribution (in terms of number of datasets) roughly matches pledge
    - staging requests were submitted to Rucio as they come, no special grouping for bulk mode
    - The staging process ran from 01/16 to 02/22
      - Since March pre-stage option is the default one

WFMS/DDM: intelligent data pre-staging (5/5)

- A recent mini-exercise (cont...)
  - Some characteristics of these data sample:
    - small files:
      - Total datasets: 2945, average files per dataset : 900, average file size: 680MB
    - Wide spread among many tapes (at least on some sites)
  - Results
    - Average time to stage a dataset: (varies with sites) 17~108 hours
    - Overall average time to stage a dataset: 65 hours
    - How about throughput ? DDM dashboard, TBD …

WFMS/DDM: monitoring (1/3)

- What's available now?
  - From R2D2 (snapshot below)



WFMS/DDM: monitoring (2/3)

- New monitoring
  - Integrated monitoring, from both WFMS and DDM data(base) sources
  - display only information essential for data carousel
  - Will be used by sites, production managers and ADC operations
- What to monitor ?
  - Staging progress: per task/dataset, per source site, per destination site
  - Staging throughput and data volume out of tape sites
  - Logging of decision making history from WFMS and DDM
  - ……

WFMS/DDM: monitoring (3/3)

- Sketch from WFMS monitoring team (a snippet below, PanDA monitor)

Tasks list

Campaign	ReqID	taskID	Status	nTotal	nQueued	nActive	nCompleted	nCompleted / nTotal (%)	Source	Destination	Time passed since the start time	Last staged file timestamp
...		$task_1$	staging									
...		...										
...		$task_N$	staging									

- **Inputs from sites and ops are vital !**

WFMS/DDM: intelligent writing

- Increasing file size to tape
  - test ongoing in ADC
    - Rucio supports archival files
    - zip jobs (just like a regular PanDA job) being tested at small scale
  - Target:
    - rough number 10GB
    - different types of data may vary in target size, details under discussion…
- Group writing by dataset ? by container ? by tag ?
  - DDM/Rucio work with FTS and dCache teams, to pass metadata (e.g. dataset name) along with staging requests, to sites
    - Done. Rucio can pass metadata to FTS now
  - Discussion within WLCG archival storage WG (more on later slides)

Sites: a recent survey of tape sites (1/2)

- On tape system (and frontend) evolution ...
  - New purchases of tape libraries and drives at various sites
  - Migration to new tape system, e.g. castor -> CTA; move to new castor instance; TSM -> HPSS
  - Expansion of tape disk space in the frontend, e.g. dCache tape read/write pools grow to O(PB)
  - Tape system upgrade, with some new features (e.g. TOR@HPSS)
  - Performing stress test just between dCache and tape, without rucio/FTS/SRM
  - “... working on increasing the throughput at least a factor of two …”

Sites: a recent survey of tape sites (2/2)

- Feedback to ADC on better usage of tape systems …
  - Bulk request for staging
  - Bigger file size
  - Anyway to be notified in advance of an upcoming massive recall campaign ? How long ahead can it be ?
    - Hours? Monitoring will show how many staging requests are queued
  - Limit as much as possible writing data that will be removed, since intense repacking is a resource-consuming activity
  - Site would like to have a deeper understanding of how a campaign and various workflows are handled, in terms of I/O operations with disk and tape, in the current model and data carousel model, to better appreciate the similarities and differences.
    - Will follow up with ADC/DDM experts on it

Sites: WLCG Archival Storage WG activities

- Current hot topic: how to better organize writing to tape
  - A recent observation from a tape site:
    - “... >75000 files being queued for recall. While it only represents ~45 TB of data, it is actually spread across >1600 tapes. …”
  - CERN proposal on improving writing to tape
    - Discussion ongoing between experiments and tape system experts
    - One issue is how to co-locate data while still guarantee writing throughput in the tape system
  - One good example from an ATLAS T1: TRIUMF
    - They group writing by datasets
    - Presentation given at a recent TCB meeting

Next steps

- Exercise of running derivation from tape
  - After near term plan are implemented in WFM and DDM systems
- Pragmatic and iterative process ...
  - Exercise after every important checkpoint, as data carousel evolves

- Ultimate goal -- to have data carousel in production before Run 3