# COMPASS production system: Frontera experience

Artem Petrosyan

GRID'2021, JINR, Dubna
July 5, 2021

Recent COMPASS HPC allocations

- 2016-2019 — Blue Waters, University of Illinois at Urbana Champaign

  - Edge jobs management service: Multi-Job Pilot

- 2019-present — Frontera, Texas Advanced Computing Centre

  - Edge jobs management service: Harvester

- On both machines COMPASS ran data reconstruction tasks

- Calibration database instance ran on each computing node

COMPASS setup on Frontera 1/2
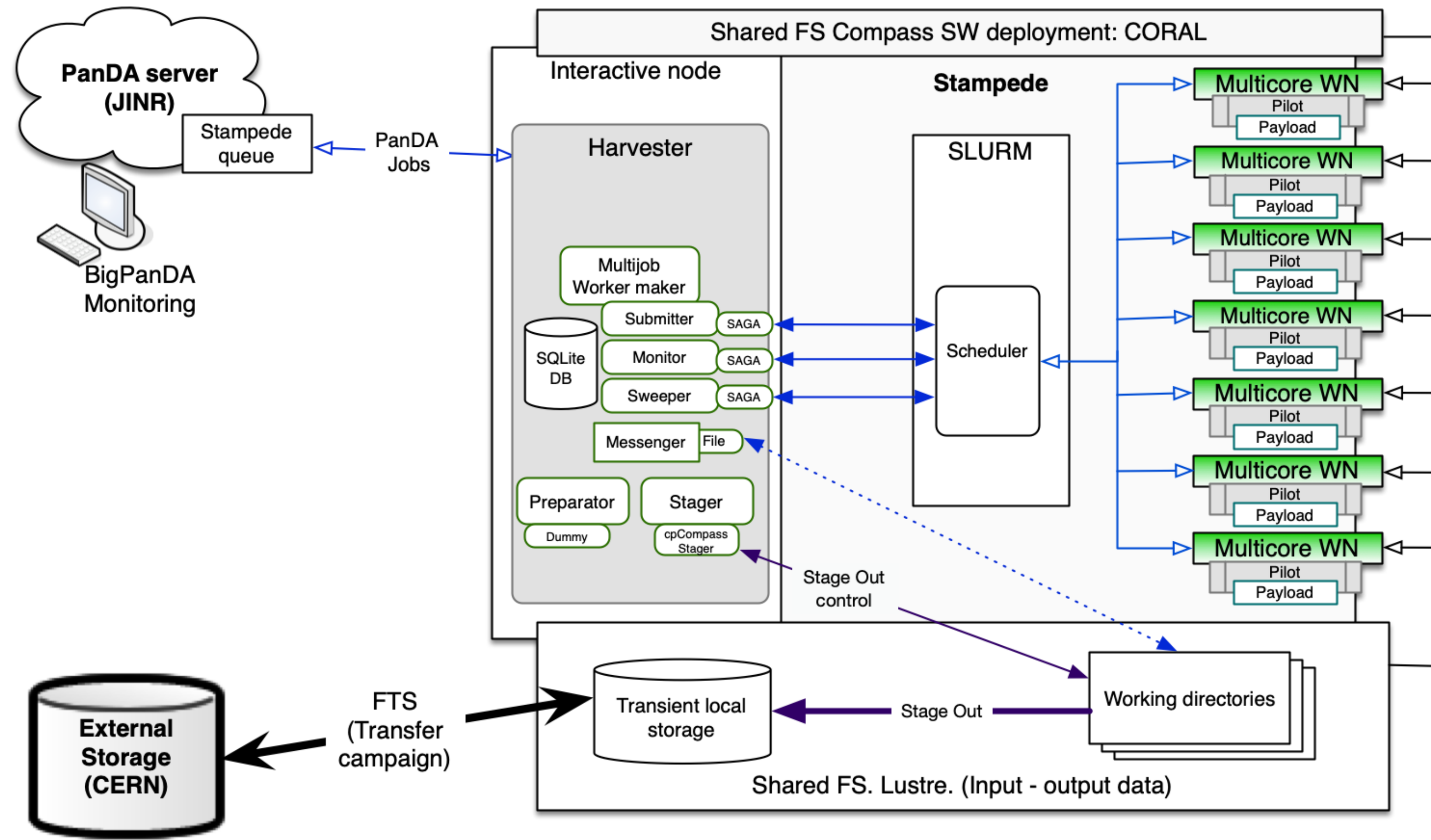
COMPASS setup on Frontera 2/2

- Applied software installed at the local file system

- Calibration db runs at each computing node, i.e. 1 per each 56 jobs, first CPU on the node starts the db instance, others wait

- PanDA Harvester as an edge service is used, extended by COMPASS logic

- 4 Harvester instances deployed in order to reach the expected processing rate

- Submission size: each Harvester configured to run up to 5 submissions of 50 nodes each, 14000 individual jobs in total

- Processing rate: 25-56K individual jobs

Harvester and calibration database

- On HPC expected job execution time is declared before the submission, usually COMPASS data reconstruction job spends up to 8 hours

- Calibration database starts on each node before all payloads

- The db runs as a separate process, and continue to run even when all jobs are finished

- Harvester monitoring checks that all jobs are done and sends signal to kill the submission when all jobs are in finished state

- A special monitoring routine was developed in Harvester monitor module to kill the submission when all jobs are done: it checks that all job are in final status ('failed' or 'finished') and sends signal to Slurm to remove the submission

Harvester and COMPASS data

- Each reconstruction job produces 5 output files: mini data summary tree (mDST), histogram, events dump, stderr and stdout

- Each result file must be transferred to the corresponding directory

- New stager module for Harvester was developed to perform such behaviour
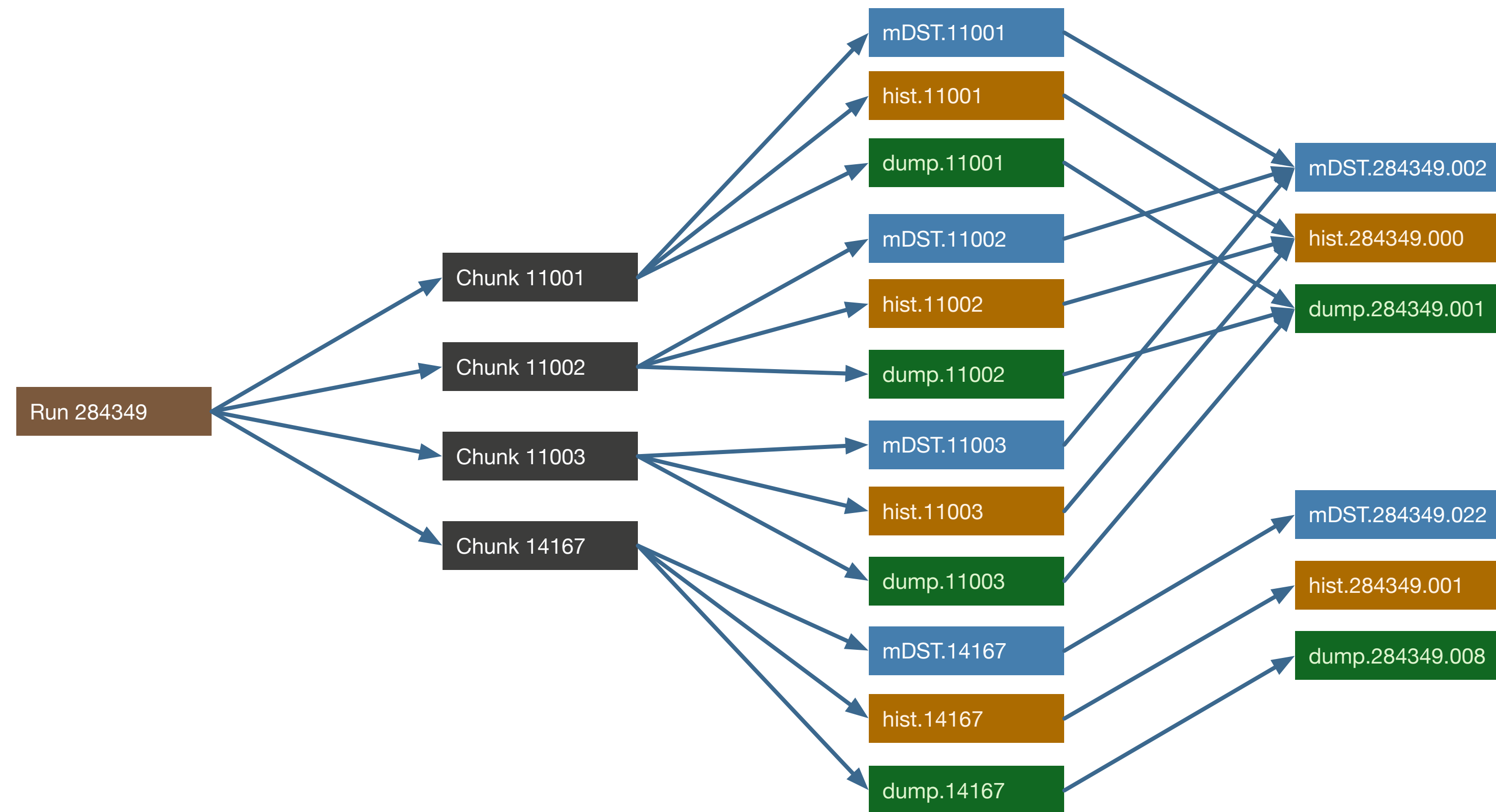
Submission types management

- COMPASS performs full-chain workflows on Frontera, including processing and merging

- Reconstruction job requires up to 8 hours

- Merging job usually spends 1 hour

- Each Harvester instance can manage several logical queues with different configuration

- Two queues were defined to work with Frontera: standard (submission to 50 nodes and 10 hours) and merging (1 node and 2 hours)
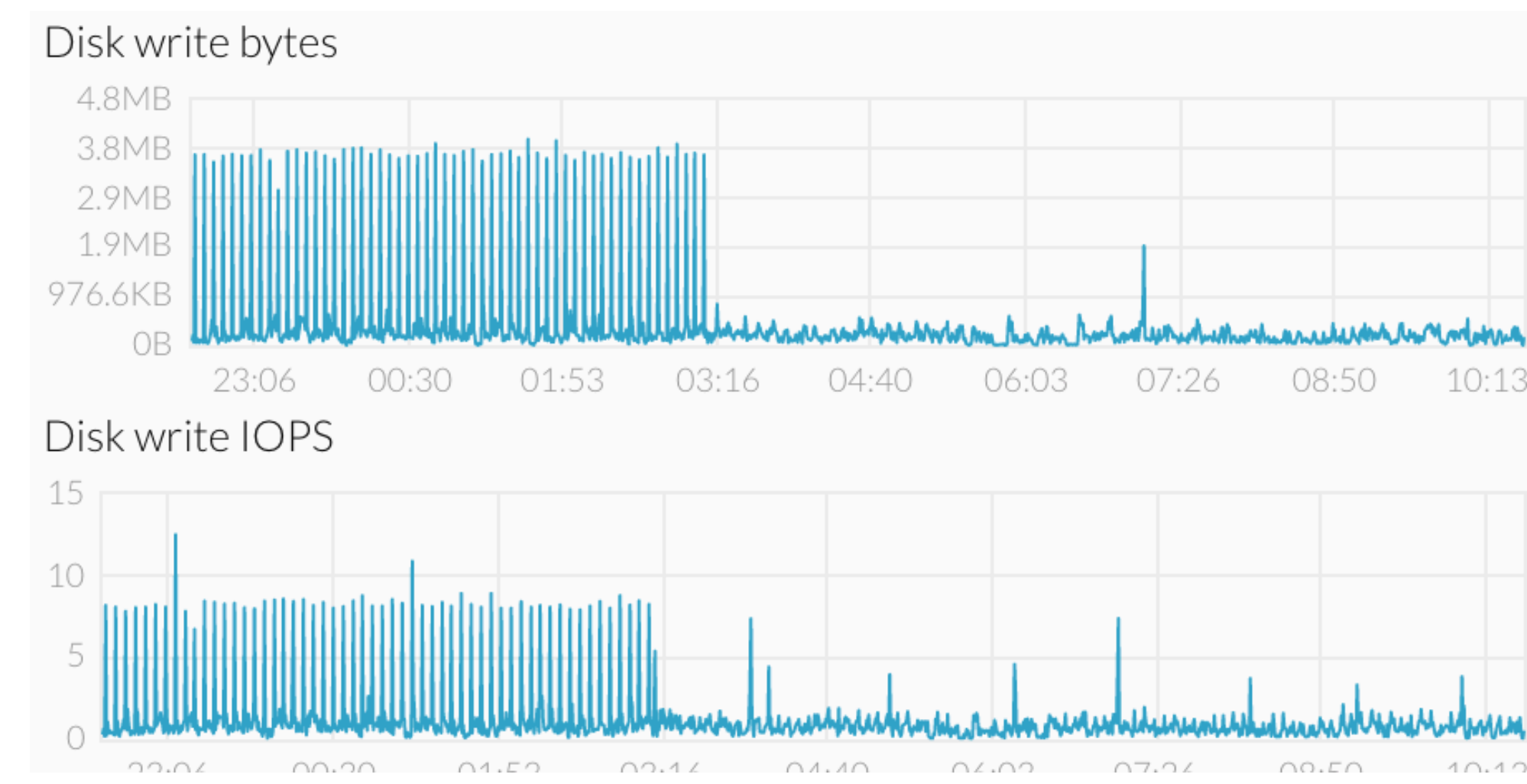
Jobs submission features

- Submission rate at Frontera for COMPASS is limited to 10 jobs per individual user, because more jobs overload the file system

- If submission generates too high load on the file system it will be removed

- A special software module is being used, which redefines all I/O operators and adds ability to work with the file system taking into account its current load

- All payloads are being transferred to the local cache disc of the node

- Applied software is being copied to the local cache disc of the individual node

- Each Harvester instance keeps 5 submissions in the queue, which allows to have 10 submissions in waiting status while 10 are running in case of 4 Harvester instances

PanDA server and long waiting jobs

- Frontera is a large machine and local batch manager sets higher priority for larger jobs in the queue

- Jobs can spend weeks waiting in the queue before being processed

- PanDA server usually restarts each 24 hours to reload CRLs and to release logs

- With high load each such restart leads to loss of jobs

- Fetch-crl was disabled on PanDA and Harvester sides, rotation of logs has been moved to Apache's responsibility, logging level was reduced to CRITICAL

- A dedicated PanDA server was deployed and LIT Cloud to work only with Frontera

9

Disk write bytes

4.8MB
3.8MB
2.9MB
1.9MB
976.6KB
0B
     23:06   00:30   01:53   03:16   04:40   06:03   07:26   08:50   10:13

Disk write IOPS

15

10

5

0

Waiting and running submissions

```
2020-05-04 12:13:39,236 panda.log.monitor: DEBUG     run <id=monitor-login2.frontera.tacc.utexas.edu_260278-2b7be3a52700> done a DB cycle : took 123.118 sec
2020-05-04 12:20:14,206 panda.log.monitor: DEBUG     run <id=monitor-login2.frontera.tacc.utexas.edu_260278-2b7be3c53700> done a DB cycle : took 15.499 sec
2020-05-04 12:20:49,478 panda.log.monitor: DEBUG     run <id=monitor-login2.frontera.tacc.utexas.edu_260278-2b7be3851700> done a DB cycle : took 50.771 sec
2020-05-04 12:26:49,967 panda.log.monitor: DEBUG     run <id=monitor-login2.frontera.tacc.utexas.edu_260278-2b7be3a52700> done a DB cycle : took 118.957 sec
2020-05-04 12:28:57,948 panda.log.monitor: DEBUG     run <id=monitor-login2.frontera.tacc.utexas.edu_260278-2b7be3650700> done a DB cycle : took 79.672 sec
2020-05-04 12:32:04,956 panda.log.monitor: DEBUG     run <id=monitor-login2.frontera.tacc.utexas.edu_260278-2b7be3c53700> done a DB cycle : took 108.188 sec
2020-05-04 12:32:34,189 panda.log.monitor: DEBUG     run <id=monitor-login2.frontera.tacc.utexas.edu_260278-2b7be344f700> done a DB cycle : took 105.660 sec
2020-05-04 12:40:06,085 panda.log.monitor: DEBUG     run <id=monitor-login2.frontera.tacc.utexas.edu_260278-2b7be3851700> done a DB cycle : took 87.977 sec
2020-05-04 12:40:57,136 panda.log.monitor: DEBUG     run <id=monitor-login2.frontera.tacc.utexas.edu_260278-2b7be3a52700> done a DB cycle : took 114.419 sec
[(harvester2) login2.frontera(1283)$ squeue --user=apetr
             JOBID   PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
            791806      normal tmp_Cd5G    apetr PD       0:00      1 (Nodes required for job are DOWN, DRAINED or reserved for jobs in higher priority partitio
ns)
            789361      normal tmp__LO7    apetr  R    4:57:04     50 c120-171,c121-011,c140-[012,111],c141-[001,192,194],c142-174,c146-[064,102,122,151,153,161
],c147-193,c148-[024,143,213],c149-073,c150-133,c151-[044,061,153],c158-214,c159-[032,111,192],c160-[113,152],c162-[001,062,083,193],c163-[004,022,042,062,064,1
11,134,142,144,171,183,191],c164-[053-054],c165-[082,092,104]
            789301      normal tmp_FHhF    apetr  R    6:36:44     50 c118-[141,162],c122-[121,124,161,191],c124-201,c125-[083,104,114,143,173,191,194],c126-[00
4,041,044,094,133],c127-[001,173,203],c128-[012,153,172],c129-[032,142,174],c130-[004,144,152,154,193],c131-[034,052,111,201],c132-[083,101,133,173,192,202],c13
3-[094,112,142,164,204],c134-[042,053]
            789571      normal tmp_pDdY    apetr  R    4:34:55     50 c101-[072-073,131],c102-[003,013,054,174,213],c103-[131,134,203],c104-011,c107-184,c114-[0
31,033-034,113],c116-[142,161],c117-[094,121-122,134,194],c119-032,c138-004,c153-101,c154-[151,163,192,194],c167-161,c168-192,c169-141,c170-[041,083-084,101,104
,111,143-144,211-212],c172-[043,141,153],c173-111,c176-[074,081]
            790133      normal tmp_P2Gn    apetr  R      35:16     50 c153-[084,112,173],c154-[041,043,051,053],c155-[062,071],c156-[043,191],c157-[071,093,112,
134,173,183,194],c160-084,c162-[113,143],c163-[102,153],c164-004,c169-001,c171-[033,072,081],c172-041,c173-[133,141],c174-101,c175-011,c177-032,c178-[004,022],c
179-[044,052,141,164],c180-082,c181-[051,112,194],c182-[063,073,161],c183-122,c184-004,c186-054
            790148      normal tmp_nNHd    apetr  R      34:58     50 c105-092,c110-[172-173],c112-[014,021,213-214],c115-[152-153],c123-[163-164],c135-[181-182
],c136-[043-044,051,123-124],c139-[103-104],c143-[023-024,031-032,174,181-183],c144-[154,161],c152-[122-123,184,191],c185-[142-143],c187-[003-004,141-143,211-21
3],c190-[063-064,071,123-124,131]
            790175      normal tmp_4nS4    apetr  R      34:58     50 c103-[084,091],c104-[174,181],c119-[132-133],c120-[212-213],c122-[033-034,042-043,102-103,
204,211,213-214],c124-[154,161],c128-[163-164],c130-[171-172],c131-[094,101],c134-[064,071],c138-[094,101],c140-[063-064,123-124,201-202],c141-[142-143],c142-[0
13-014,054,061,074,081],c154-[054,061],c155-[063-064,072-073]
            790442      normal tmp_tUB_    apetr  R      34:58     50 c157-[192-193],c159-[182-183],c162-[114,121],c163-[094,101,103-104],c165-[162-163],c169-[0
02-003],c174-[183-184],c175-[003-004],c176-[144,151],c181-[054,061],c183-[001-002,021-022],c186-[003-004,072-073,084,091,114,121,132-133],c187-[034,041],c188-[0
44,051,081-082,092-093],c189-[002-003],c190-[001-002,022-023]
            791660      normal tmp_a9dw    apetr  R      34:29     50 c140-[182,194],c141-141,c142-[021,043,051,053,071,073,114,134],c143-[004,022,184],c144-[06
3,072,122,152],c145-014,c146-[103,121],c147-[082,094,132,192],c148-[003,011,071,083,092,112,122,141,191],c149-[013,024,034,042,063,083,102,163],c150-[022,064,08
3,101,153,204],c151-[012,014]
            791726      normal tmp_Sj9J    apetr  R      20:24     50 c142-[133,142],c144-[043,064,173],c146-212,c147-072,c148-[002,004,101,104],c149-[021,071,1
34,164],c150-[002,123,161],c151-[082,113,122,124],c152-[091,151,181,202],c153-[061,071,073,123,164,172,174,183,203],c154-[034,042,044,052],c155-194,c156-[042,04
4,081],c157-[072,092,094,144,152],c158-171,c159-093
            791753      normal tmp_nVqK    apetr  R      17:23     50 c131-[112,114,133,142],c132-[023,033,054,062,082,132,134,172,191,194],c133-[021,044,053,14
3,174,191],c134-[011,112,144,182],c135-[083,132],c136-054,c137-[144,204],c138-[052,114,124,143],c139-[004,141,151,153,212],c140-[094,162,164,181],c141-134,c142-
[042,044,052,072,113,121],c159-112
            791750      normal tmp_iQz1    apetr  R      58:37      1 c182-012
            791758      normal tmp_ZsmN    apetr  R      58:37      1 c173-063
            791799      normal tmp_Mpck    apetr  R       8:14      1 c109-173
            791800      normal tmp_qpkn    apetr  R       8:14      1 c111-054
(harvester2) login2.frontera(1284)$
```

COMPASS computing infrastructure overview
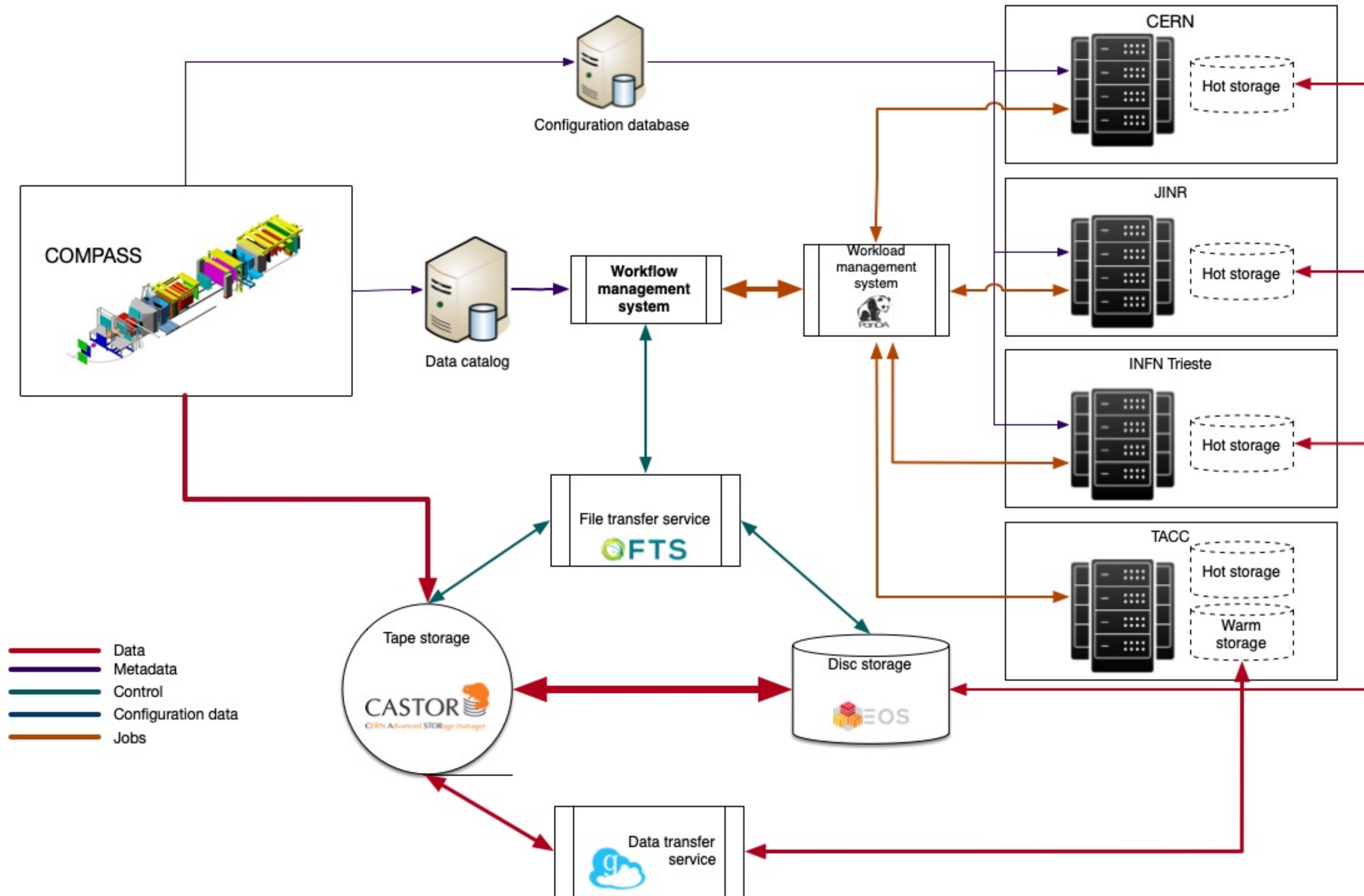
Summary
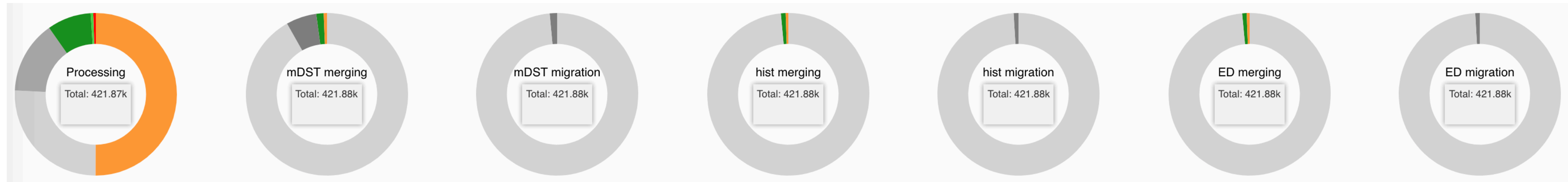
- Since 2019 COMPASS runs data reconstruction tasks on Frontera via the same production system which manages data processing on the Grid

- COMPASS-specific logic was implemented in Harvester: calibration database execution, data handling, monitoring

- Current setup allows to run 56 000 individual jobs, which is 4-5 times more than all other computing resources of the collaboration

- Future plans: start Monte-Carlo processing via the COMPASS production system on Frontera