COMPASS Grid Production System

Artem Petrosyan
NEC'2017, Budva, Montenegro
September 28, 2017

What is COMPASS

- COmmon Muon Proton Apparatus for Structure and Spectroscopy (COMPASS) is a high-energy physics experiment at a Super Proton Synchrotron (SPS) at CERN

- The purpose of the experiment is the study of hadron structure and hadron spectroscopy with high intensity muon and hadron beams

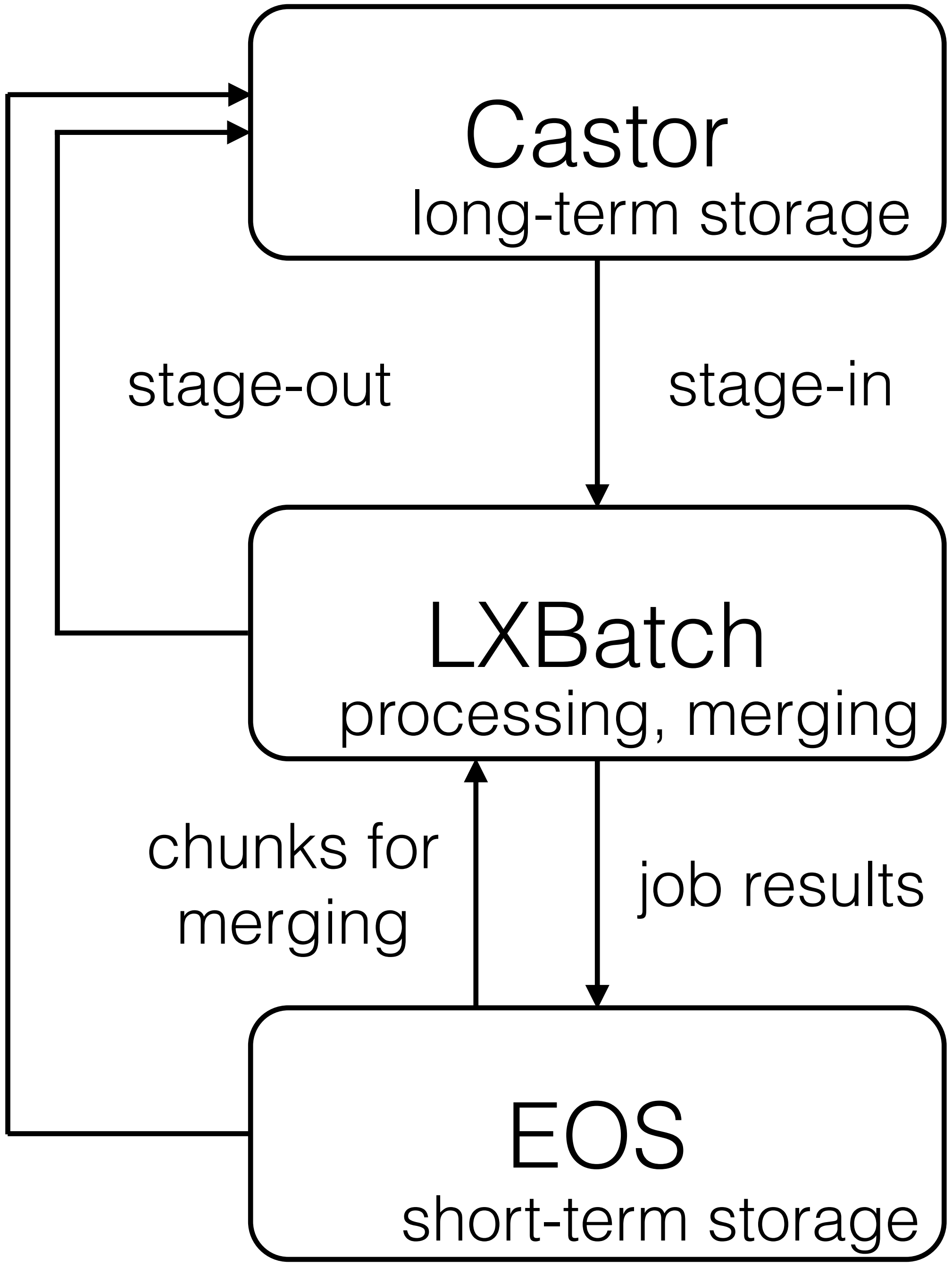
- First data taking run started in summer 2002 and sessions are continue

- Each data taking session containing from 1.5 to 3 PB of data

- More than 200 physicists from 13 countries and 24 institutes are the analysis user community of COMPASS

“Classic” COMPASS production dataflow

- All data stored on Castor
- Data is being requested to be copied from tapes to disks before processing (may take ~6 hours)
- Task moves files directly from Castor to lxbatch for processing, several programs are used for processing
- After processing results are being transferred to EOS for merging or short-term storage or directly to Castor for long-term storage
- Merging
- Results are being copied to Castor for long-term storage



Features of “classic” implementation

- We can run jobs on the only one computing resource and LSF will be decommissioned at the end of 2018

- We strictly connected to AFS, local file system, which will be replaced by EOS

- Strictly connected to CASTOR, which will be replaced by EOS

- User jobs and production jobs are sent directly to computing resources and can not be managed (we can not set priority, quota at user’s level)

Motivation

- Move processing to CERN Condor
  - Even more: get ability to switch computing sites, get more resources, any type, not only LSF
  - Enable processing on Blue Waters HPC
- Get rid of self-written code and start using some “common" solution

Action items to enable processing via PanDA

- PanDA instance installation
- Grid environment setup
- Production jobs execution by PanDA expert
- Physics validation
- Production chain management software preparation
- Production by COMPASS production manager

Grid environment

- AFS COMPASS group
    - Production account
- Local batch queue
- EOS directory
- AFS directory to deploy production software

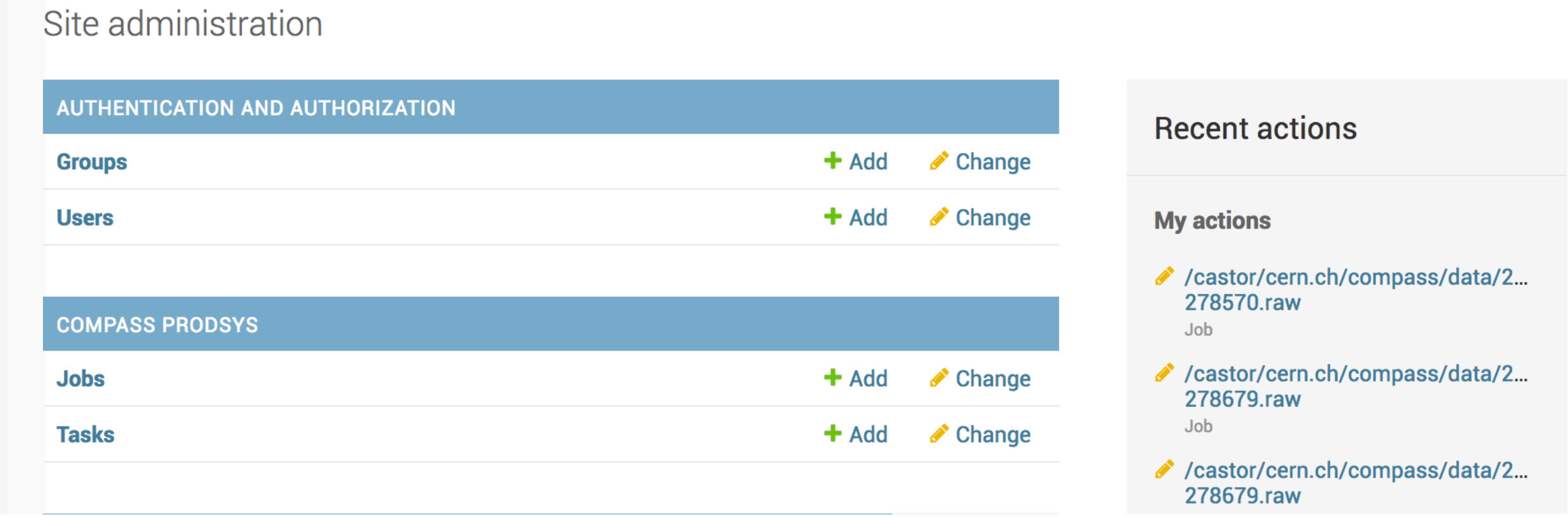
- Virtual organisation
    - Production role
- Computing element
- EOS storage element
- CVMFS

Infrastructure overview

- PanDA server over MySQL, Monitoring, AutoPilotFactory, Production System deployed in Dubna on production area of our cloud service
- ProdSys service deployed on JINR cloud service
- Condor CE at CERN
- EOS SE at CERN
- PBS CE at JINR
- LSF CE at Trieste
- PerfSonar service at JINR cloud network segment to monitor network connectivity between JINR and CERN

ProdSys

- Totally reengineered component
- UI based on Django Admin
- MySQL database backend
- Periodic tasks managed by crontab and Celery
- Communication with PanDA server via PanDA Client
- Manages tasks definition, jobs submission, status tracking, errors handling, retries strategy, merging, cross checking



Site administration

AUTHENTICATION AND AUTHORIZATION		
Groups	+ Add	✎ Change
Users	+ Add	✎ Change

COMPASS PRODSYS		
Jobs	+ Add	✎ Change
Tasks	+ Add	✎ Change

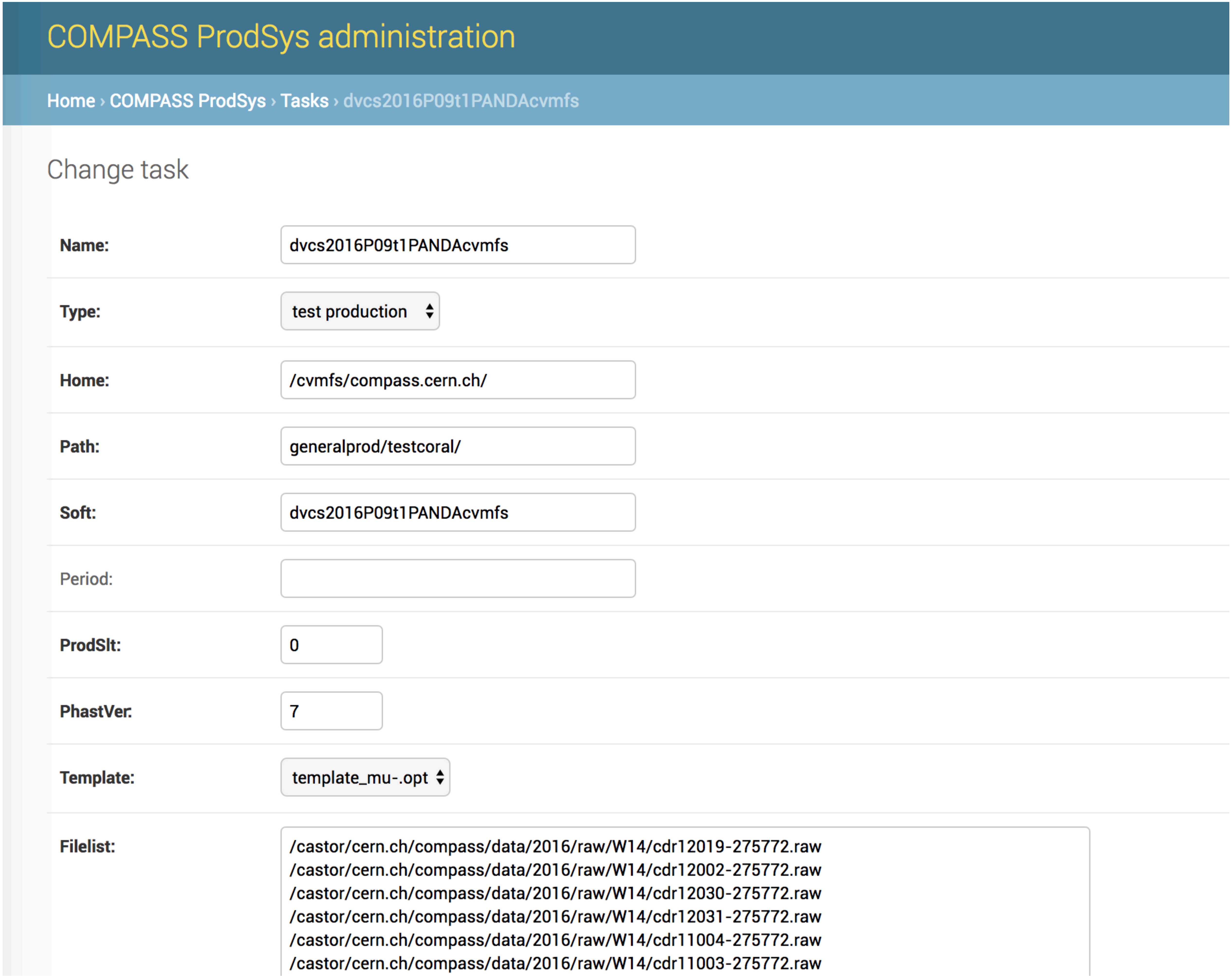
Recent actions

My actions

- ✎ /castor/cern.ch/compass/data/2... 278570.raw
  Job
- ✎ /castor/cern.ch/compass/data/2... 278679.raw
  Job
- ✎ /castor/cern.ch/compass/data/2... 278679.raw

Task statuses and actions

- Draft
    - No automatic actions, production manager defines task, when the task is ready, manager changes status to Ready
- Ready
    - Jobs definitions are generated for task, when it’s done, status changes to Jobs ready
- Jobs ready
    - Production manager check if jobs generated correctly and if yes changes status to Send
- Send
    - Jobs are being sent to PanDA, first running job changes status of task to Running
- Running
    - Jobs are being sent, statuses of jobs gathered automatically, automated resubmitted in case of failure, once all jobs of one run of the task are finished, merging job is prepared and sent, being tracked, and once all merging jobs of the run are finished, cross check job to compare number of events in chunks and merged files is issued, when all cross check jobs return success results task status changes to Done
- Done
- Paused
- Cancelled



COMPASS ProdSys administration

Home › COMPASS ProdSys › Tasks › dvcs2016P09t1PANDAcvmfs

Change task

Name: dvcs2016P09t1PANDAcvmfs

Type: test production

Home: /cvmfs/compass.cern.ch/

Path: generalprod/testcoral/

Soft: dvcs2016P09t1PANDAcvmfs

Period:

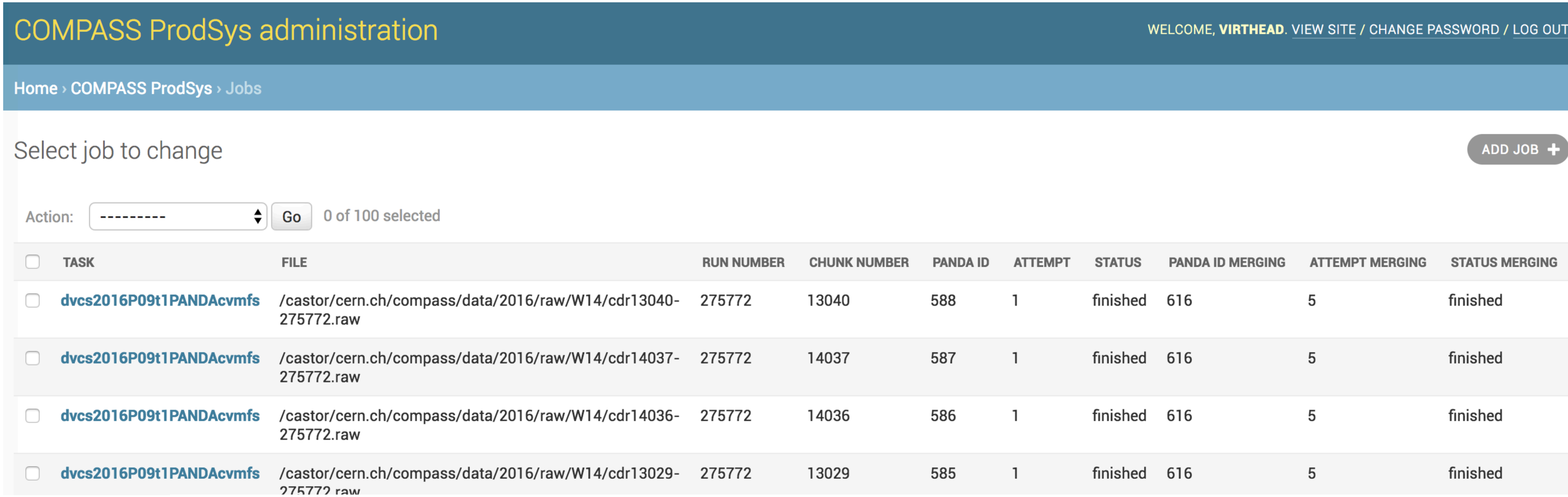
ProdSlt: 0

PhastVer: 7

Template: template_mu-.opt

Filelist:

```
/castor/cern.ch/compass/data/2016/raw/W14/cdr12019-275772.raw
/castor/cern.ch/compass/data/2016/raw/W14/cdr12002-275772.raw
/castor/cern.ch/compass/data/2016/raw/W14/cdr12030-275772.raw
/castor/cern.ch/compass/data/2016/raw/W14/cdr12031-275772.raw
/castor/cern.ch/compass/data/2016/raw/W14/cdr11004-275772.raw
/castor/cern.ch/compass/data/2016/raw/W14/cdr11003-275772.raw
```



COMPASS ProdSys administration

WELCOME, **VIRTHEAD**. VIEW SITE / CHANGE PASSWORD / LOG OUT

Home › COMPASS ProdSys › Jobs

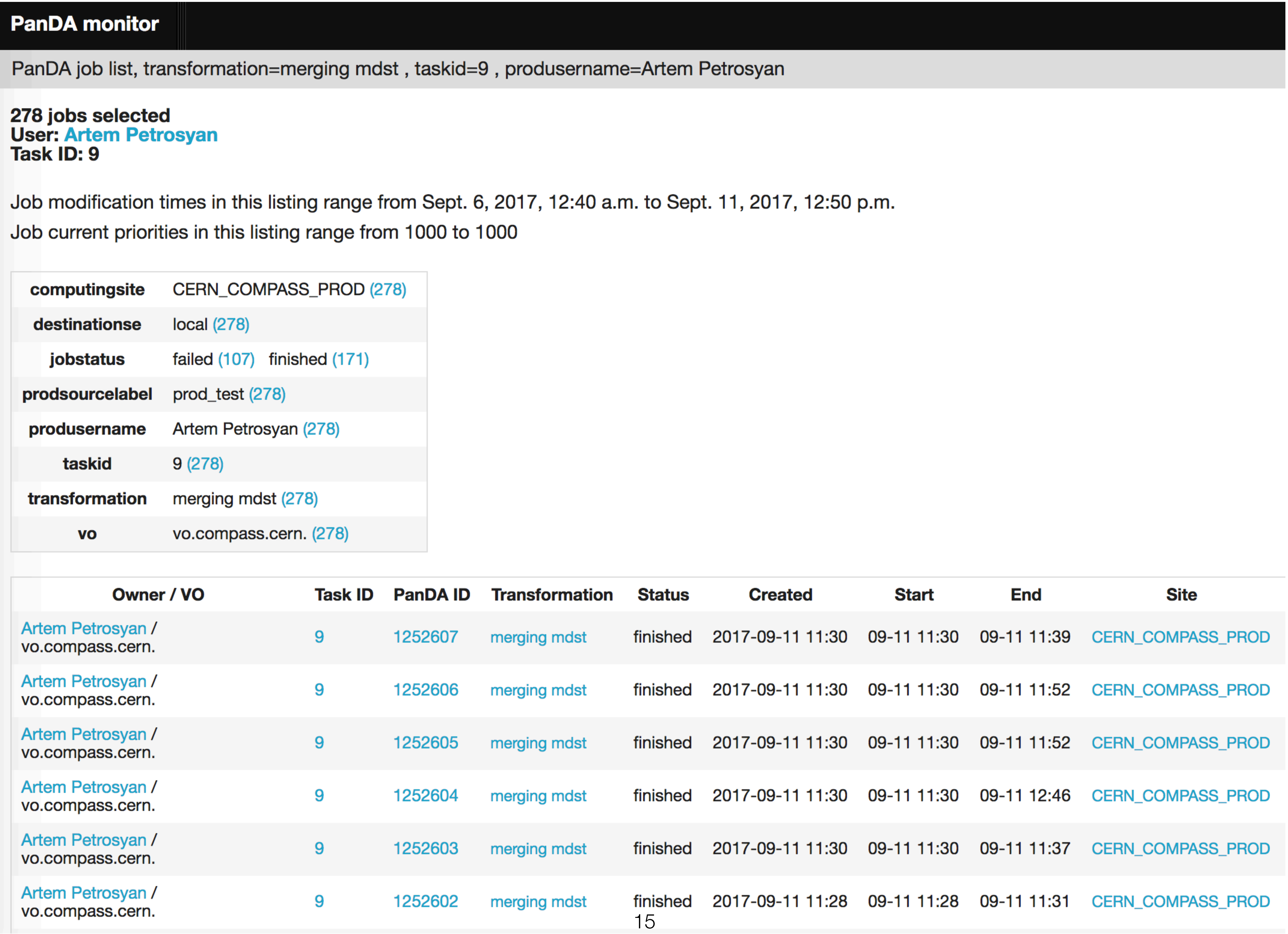
Select job to change

Action: ---------- Go 0 of 100 selected

□	TASK	FILE	RUN NUMBER	CHUNK NUMBER	PANDA ID	ATTEMPT	STATUS	PANDA ID MERGING	ATTEMPT MERGING	STATUS MERGING
□	dvcs2016P09t1PANDAcvmfs	/castor/cern.ch/compass/data/2016/raw/W14/cdr13040-275772.raw	275772	13040	588	1	finished	616	5	finished
□	dvcs2016P09t1PANDAcvmfs	/castor/cern.ch/compass/data/2016/raw/W14/cdr14037-275772.raw	275772	14037	587	1	finished	616	5	finished
□	dvcs2016P09t1PANDAcvmfs	/castor/cern.ch/compass/data/2016/raw/W14/cdr14036-275772.raw	275772	14036	586	1	finished	616	5	finished
□	dvcs2016P09t1PANDAcvmfs	/castor/cern.ch/compass/data/2016/raw/W14/cdr13029-275772.raw	275772	13029	585	1	finished	616	5	finished

Production job types

- Normal
  - File downloads from CASTOR to computing node
  - After processing being transferred to EOS
- Merging
  - Data stages in from EOS
  - Up to 40 results of normal jobs merged into one file with desired filesize (4Gb)
  - After processing result file being transferred to EOS
- Cross check
  - Internal job, uses PanDA job metrics
  - Compares number of events in file chunks and in merged file per run



PanDA job list, transformation=merging mdst , taskid=9 , produsername=Artem Petrosyan

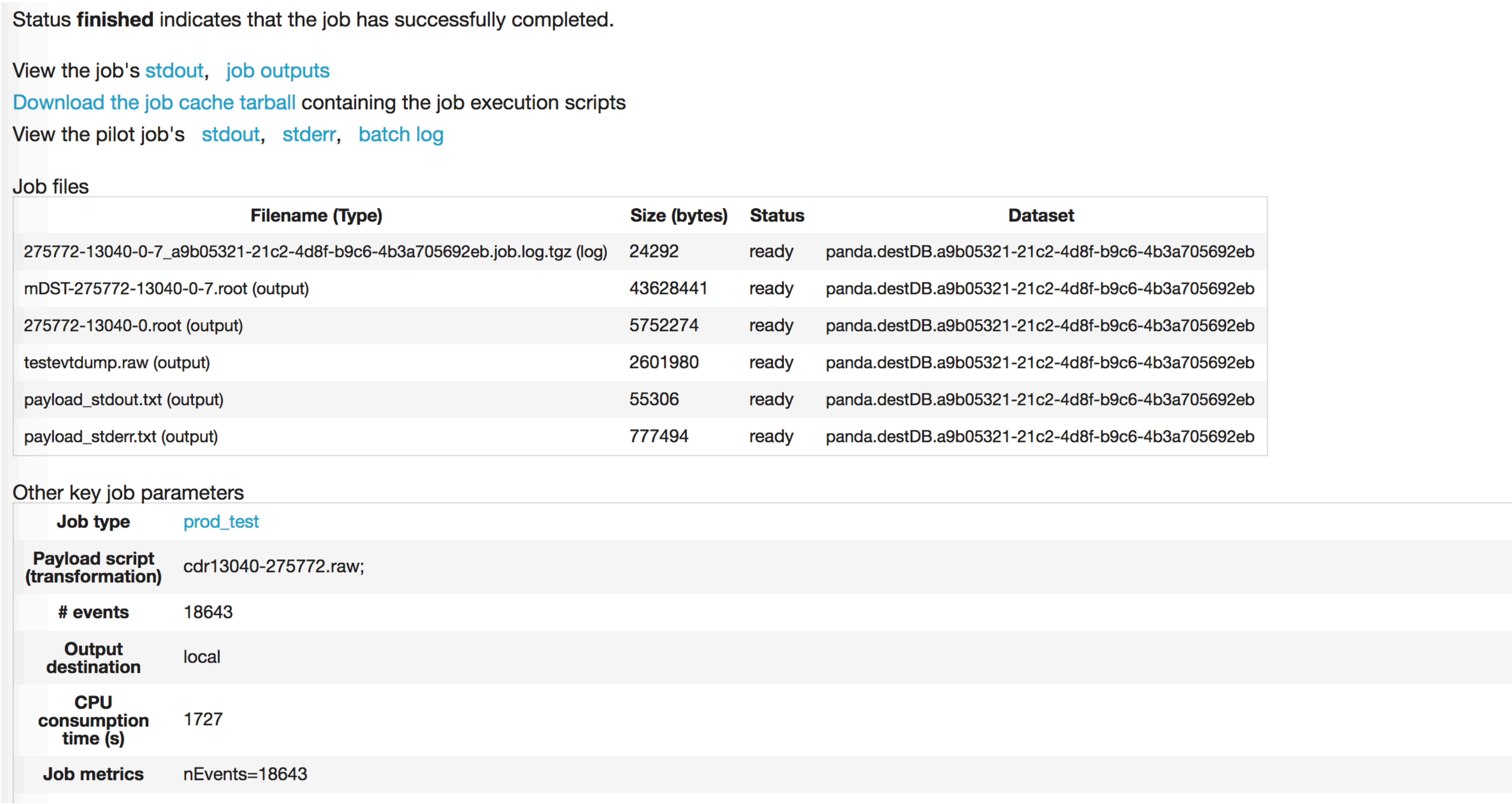
278 jobs selected
User: Artem Petrosyan
Task ID: 9

Job modification times in this listing range from Sept. 6, 2017, 12:40 a.m. to Sept. 11, 2017, 12:50 p.m.

Job current priorities in this listing range from 1000 to 1000

computingsite	CERN_COMPASS_PROD (278)
destinationse	local (278)
jobstatus	failed (107) finished (171)
prodsourcelabel	prod_test (278)
produsername	Artem Petrosyan (278)
taskid	9 (278)
transformation	merging mdst (278)
vo	vo.compass.cern. (278)

Owner / VO	Task ID	PanDA ID	Transformation	Status	Created	Start	End	Site
Artem Petrosyan / vo.compass.cern.	9	1252607	merging mdst	finished	2017-09-11 11:30	09-11 11:30	09-11 11:39	CERN_COMPASS_PROD
Artem Petrosyan / vo.compass.cern.	9	1252606	merging mdst	finished	2017-09-11 11:30	09-11 11:30	09-11 11:52	CERN_COMPASS_PROD
Artem Petrosyan / vo.compass.cern.	9	1252605	merging mdst	finished	2017-09-11 11:30	09-11 11:30	09-11 11:52	CERN_COMPASS_PROD
Artem Petrosyan / vo.compass.cern.	9	1252604	merging mdst	finished	2017-09-11 11:30	09-11 11:30	09-11 12:46	CERN_COMPASS_PROD
Artem Petrosyan / vo.compass.cern.	9	1252603	merging mdst	finished	2017-09-11 11:30	09-11 11:30	09-11 11:37	CERN_COMPASS_PROD
Artem Petrosyan / vo.compass.cern.	9	1252602	merging mdst	finished	2017-09-11 11:28	09-11 11:28	09-11 11:31	CERN_COMPASS_PROD



Status **finished** indicates that the job has successfully completed.

View the job's stdout, job outputs
Download the job cache tarball containing the job execution scripts
View the pilot job's stdout, stderr, batch log

Job files

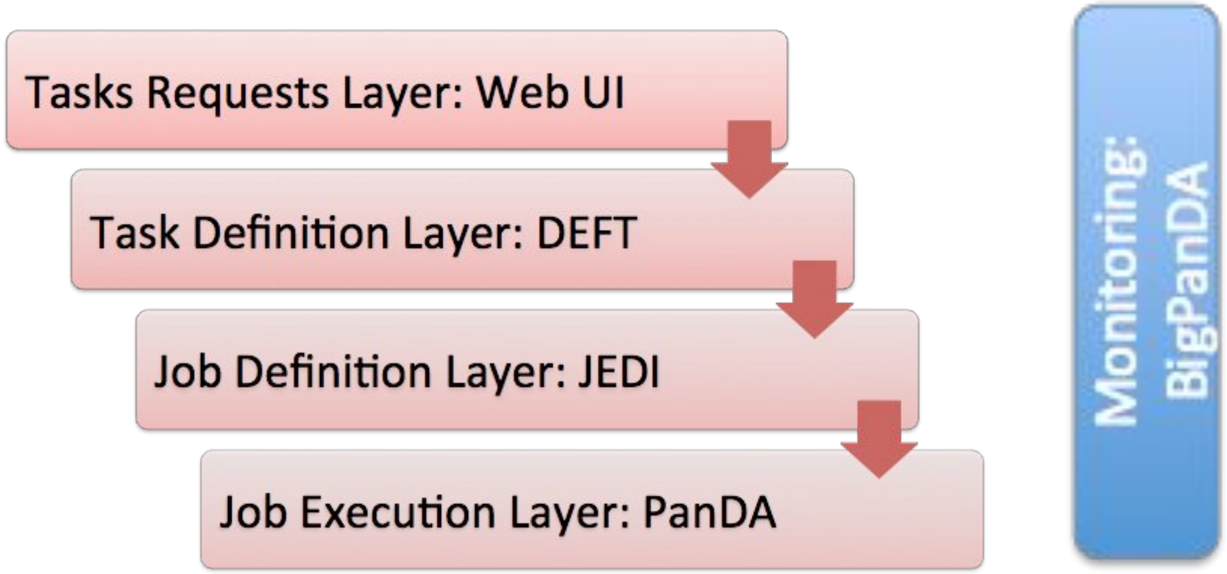
Filename (Type)	Size (bytes)	Status	Dataset
275772-13040-0-7_a9b05321-21c2-4d8f-b9c6-4b3a705692eb.job.log.tgz (log)	24292	ready	panda.destDB.a9b05321-21c2-4d8f-b9c6-4b3a705692eb
mDST-275772-13040-0-7.root (output)	43628441	ready	panda.destDB.a9b05321-21c2-4d8f-b9c6-4b3a705692eb
275772-13040-0.root (output)	5752274	ready	panda.destDB.a9b05321-21c2-4d8f-b9c6-4b3a705692eb
testevtdump.raw (output)	2601980	ready	panda.destDB.a9b05321-21c2-4d8f-b9c6-4b3a705692eb
payload_stdout.txt (output)	55306	ready	panda.destDB.a9b05321-21c2-4d8f-b9c6-4b3a705692eb
payload_stderr.txt (output)	777494	ready	panda.destDB.a9b05321-21c2-4d8f-b9c6-4b3a705692eb

Other key job parameters

Job type	prod_test
Payload script (transformation)	cdr13040-275772.raw;
# events	18643
Output destination	local
CPU consumption time (s)	1727
Job metrics	nEvents=18643

ATLAS production system components

- **Web UI** for Managers and Users provides the interface for task* and production request managing and monitoring at the higher level
- Database Engine for Tasks (**DEFT**): is responsible for formulating the tasks, chains of tasks and also task groups (production request), complete with all necessary parameters
  - ○ It also keeps track of the state of production requests, chains and their constituent tasks



Tasks Requests Layer: Web UI

Task Definition Layer: DEFT

Job Definition Layer: JEDI

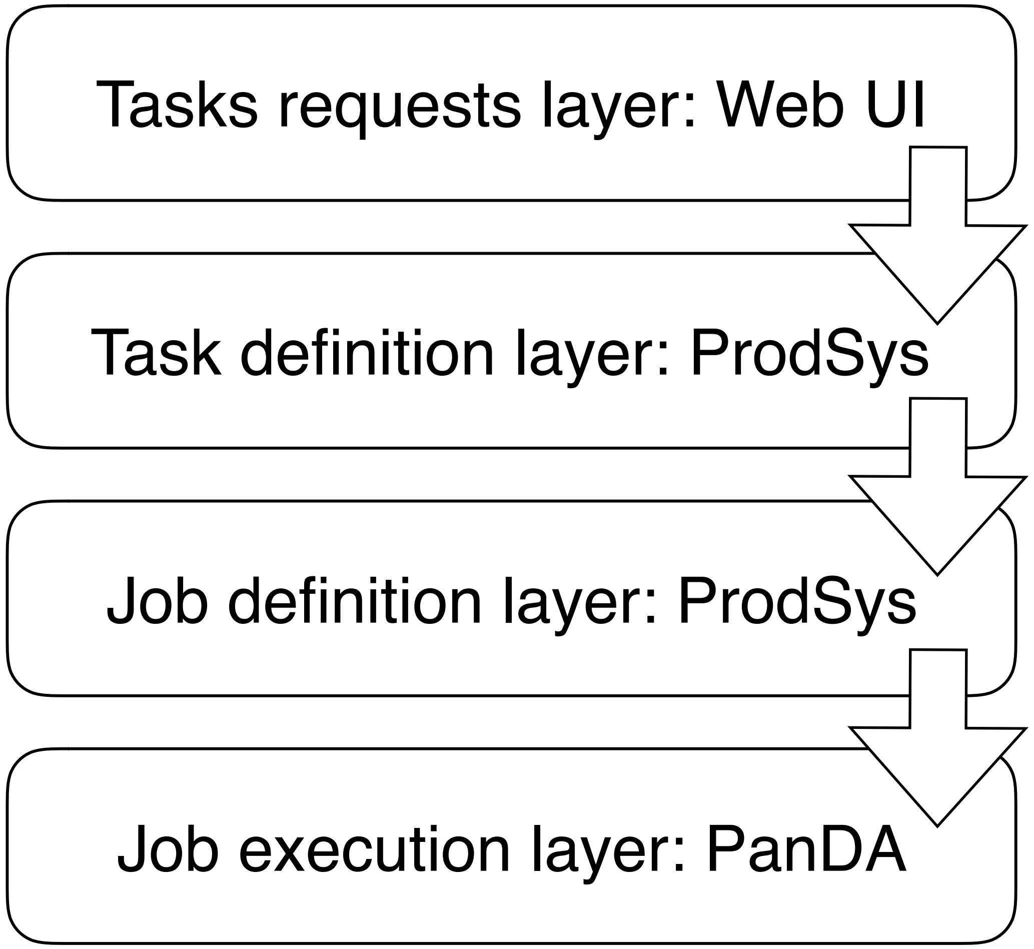
Job Execution Layer: PanDA

Monitoring: BigPanDA

- Job Execution and Definition Interface (**JEDI**): is an intelligent component in the **PanDA** server to have capability for **task-level** workload management.
  - ○ Key part of it is **'Dynamic'** job definition, which highly optimizes resources usage compared to 'Static' model used in ProdSys1.
    - ■ Dynamic job definition in JEDI is also crucial for multi-core, HPCs and other new requirements
- Monitoring (**BigPanDA**): progress, status and error diagnostics for all components.
- The PanDA **pilot** is an execution environment used to prepare the computing element, request the actual payload (a production or user analysis job), execute it, and clean up when the payload has finished. Input and output are transferred from/to storage elements, including object stores.

*Task consists of jobs that all run the same program.

COMPASS ProdSys components



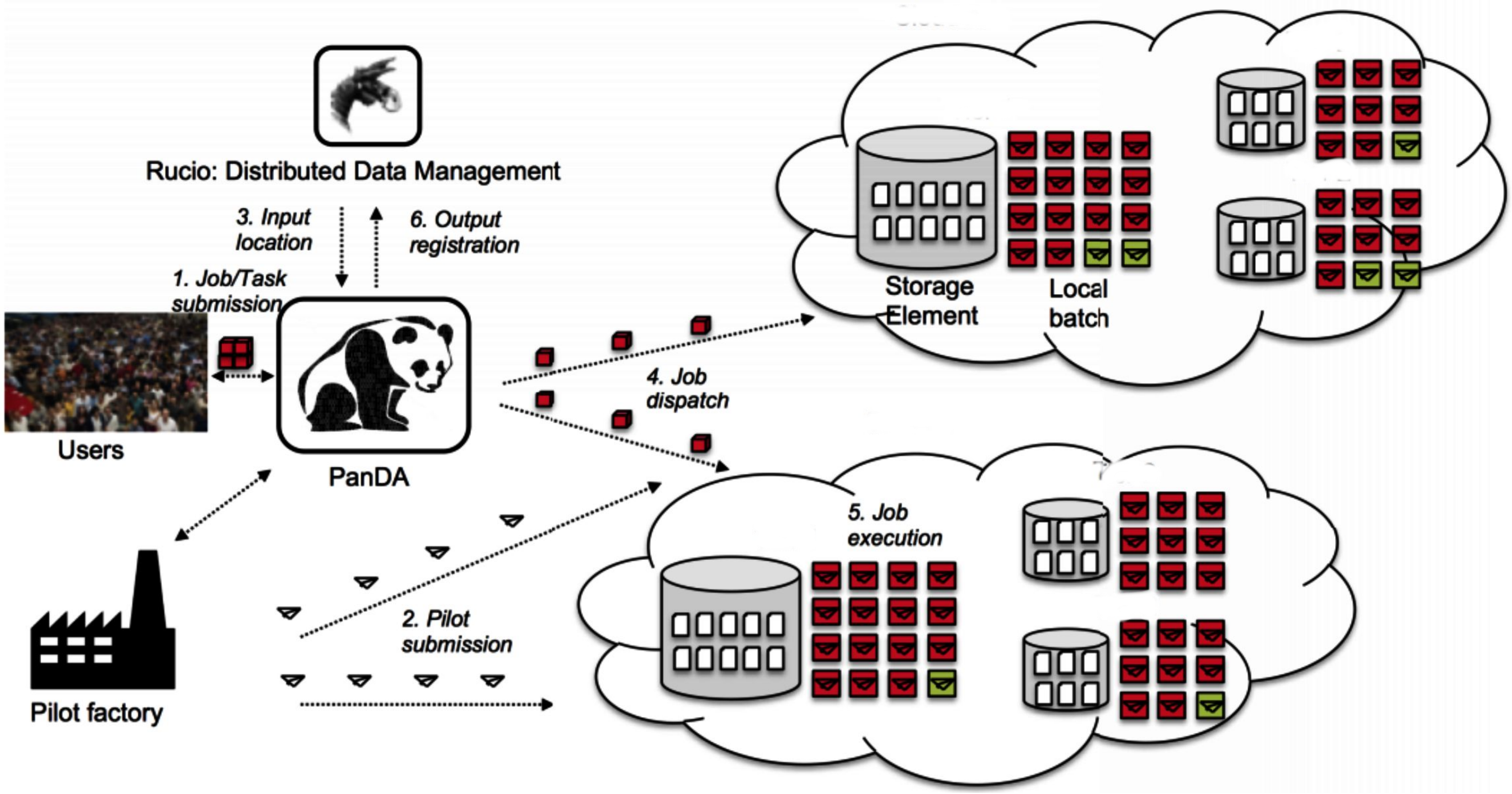
Tasks requests layer: Web UI

Task definition layer: ProdSys

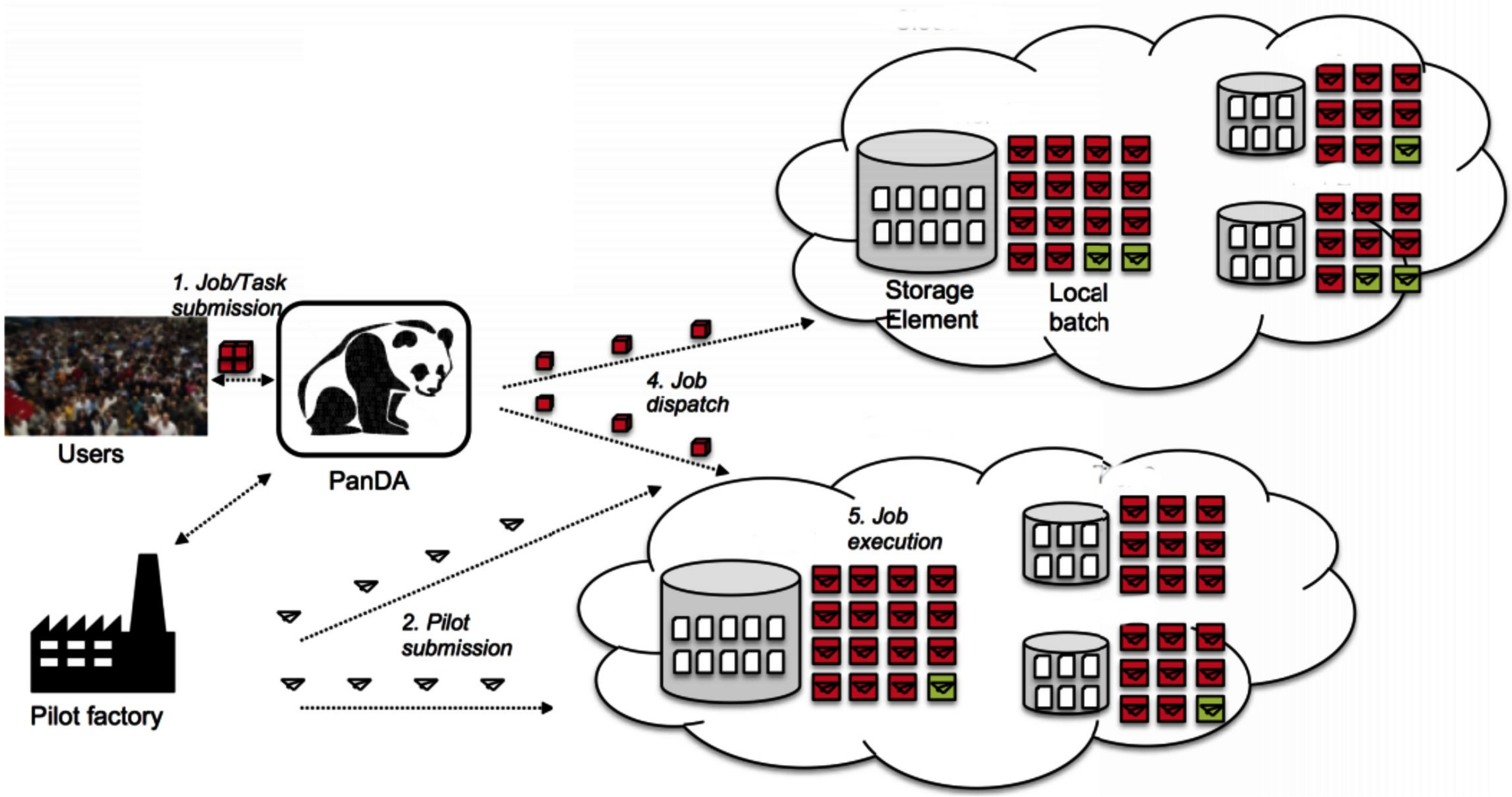
Job definition layer: ProdSys

Job execution layer: PanDA

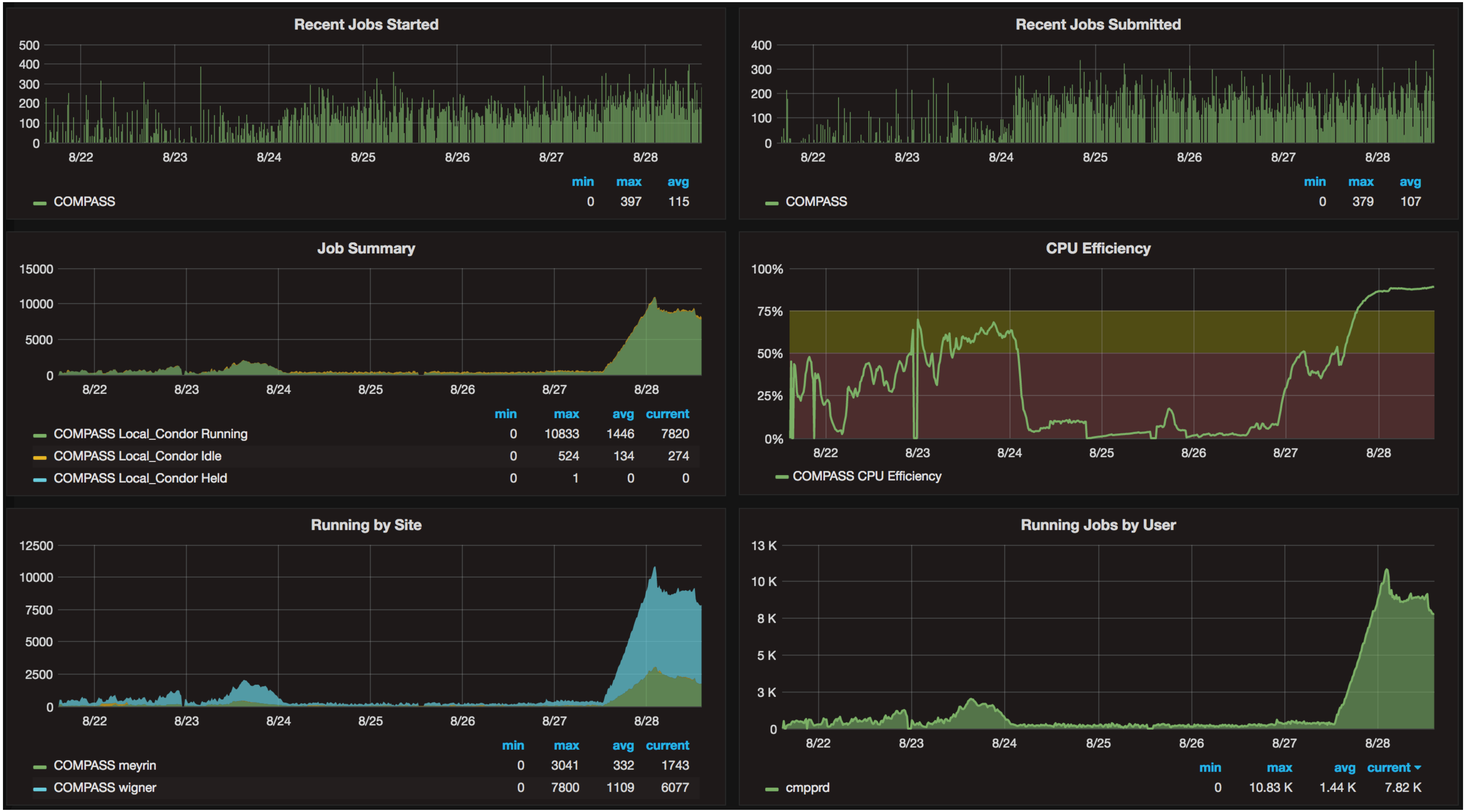
High level overview



High level overview



CERN Condor monitoring



Recent Jobs Started

	min	max	avg
COMPASS	0	397	115

Recent Jobs Submitted

	min	max	avg
COMPASS	0	379	107

Job Summary

	min	max	avg	current
COMPASS Local_Condor Running	0	10833	1446	7820
COMPASS Local_Condor Idle	0	524	134	274
COMPASS Local_Condor Held	0	1	0	0

CPU Efficiency

COMPASS CPU Efficiency

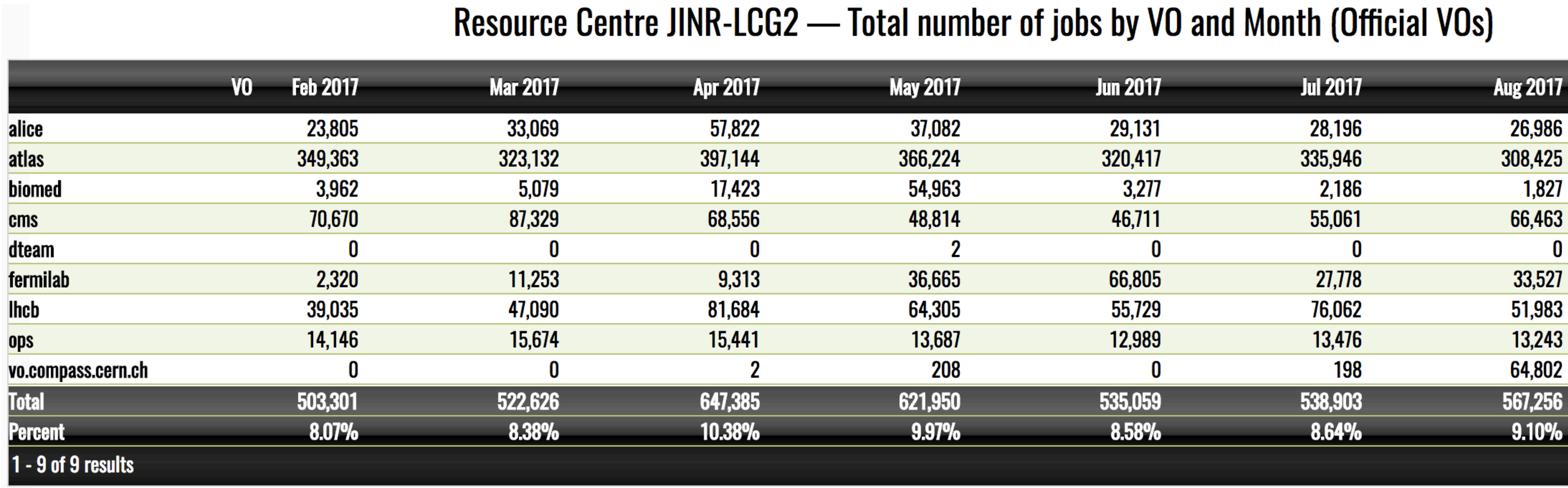
Running by Site

	min	max	avg	current
COMPASS meyrin	0	3041	332	1743
COMPASS wigner	0	7800	1109	6077

Running Jobs by User

	min	max	avg	current ▾
cmpprd	0	10.83 K	1.44 K	7.82 K

JINR T2 jobs per VO stats



Resource Centre JINR-LCG2 — Total number of jobs by VO and Month (Official VOs)

VO	Feb 2017	Mar 2017	Apr 2017	May 2017	Jun 2017	Jul 2017	Aug 2017
alice	23,805	33,069	57,822	37,082	29,131	28,196	26,986
atlas	349,363	323,132	397,144	366,224	320,417	335,946	308,425
biomed	3,962	5,079	17,423	54,963	3,277	2,186	1,827
cms	70,670	87,329	68,556	48,814	46,711	55,061	66,463
dteam	0	0	0	2	0	0	0
fermilab	2,320	11,253	9,313	36,665	66,805	27,778	33,527
lhcb	39,035	47,090	81,684	64,305	55,729	76,062	51,983
ops	14,146	15,674	15,441	13,687	12,989	13,476	13,243
vo.compass.cern.ch	0	0	2	208	0	198	64,802
Total	503,301	522,626	647,385	621,950	535,059	538,903	567,256
Percent	8.07%	8.38%	10.38%	9.97%	8.58%	8.64%	9.10%

1 - 9 of 9 results

Plans

- Migrate all types of data processing to the new system
    - MC
    - Users analysis
- Prepare COMPASS-specific monitoring
    - Tatiana Korchuganova already at CERN
    - Wish list includes COMPASS-specific items to be presented on the monitoring pages, such as data taking periods, years, runs, etc.
- Enable processing on BlueWaters HPC
    - There is a team adopting COMPASS software to run on HPC, looks like pretty soon it will be ready to be run by PanDA
- Discussion about possibility of adding Rucio to organise namespace data catalog is ongoing
- Validation of CRIC information system

Summary

- New system works in production mode
- 3 computing sites, 22 physical queues: CERN (Condor), JINR (PBS), Trieste (LSF) wrapped by one PanDA queue
- 1 storage element: EOS at CERN
- Processing with only one storage allowed to get rid of DDM, files management done by pilot at the stage in and out steps
- Observed maximum so far ~12000 simultaneously running jobs
- ~1 million jobs organised as 40 tasks processed during last two months
- ~100TB of data taken during runs of 2015, 2016 and 2017 processed already
- PanDA server over MySQL, Monitoring, AutoPilotFactory, Production System deployed in Dubna on production area of our cloud service
- Production system is a brand new one, based on Django framework, prepared to manage COMPASS production chain tasks and jobs
- We also created simple Web UI for PanDA configuration in order to manage users, sites, queues, etc.

Growing PanDA Ecosystem

COMPASS
LHC
Google Cloud Platform
NERSC
AMS-02
Alpha Magnetic Spectrometer
Europe • Asia
North America
ATLAS
NICA
nectar
amazon web services
EC2
16.99 [sec]
~4-5 weeks
135 chunks
~4 days
megaPanDA

	Titan System (Cray XK7)		
Peak Performance	27.1 PF 18,688 compute nodes	24.5 PF GPU	2.6 PF CPU
System memory	710 TB total memory		
Interconnect	Gemini High Speed Interconnect	3D Torus	
Storage	Lustre Filesystem	32 PB	
Archive	High-Performance Storage System (HPSS)	29 PB	
I/O Nodes	512 Service and I/O nodes		

time for Mammoths DNA samples from weeks to days.

Thanks!