COMPASS Grid Production System

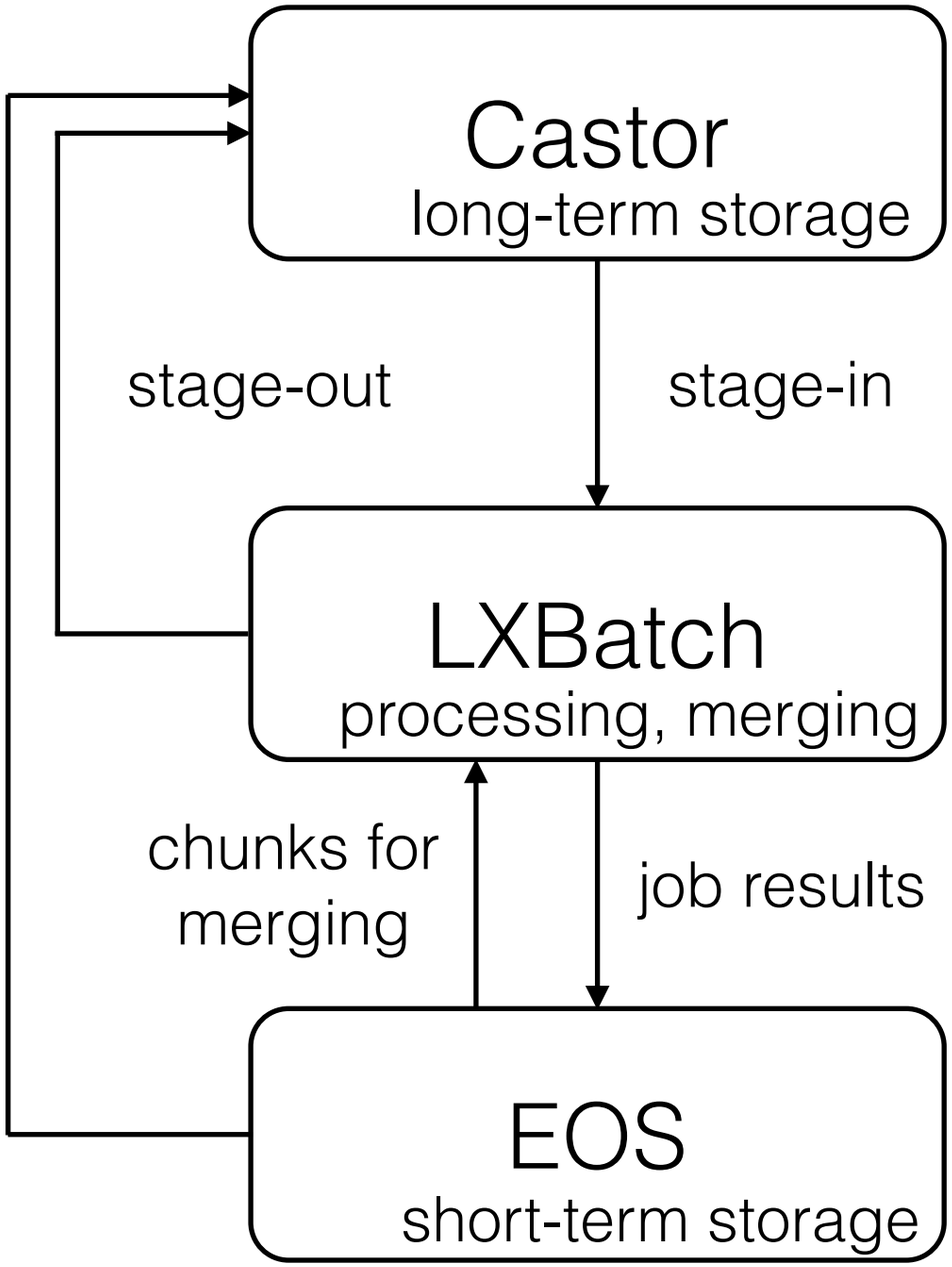
Artem Petrosyan
December 6, 2017

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

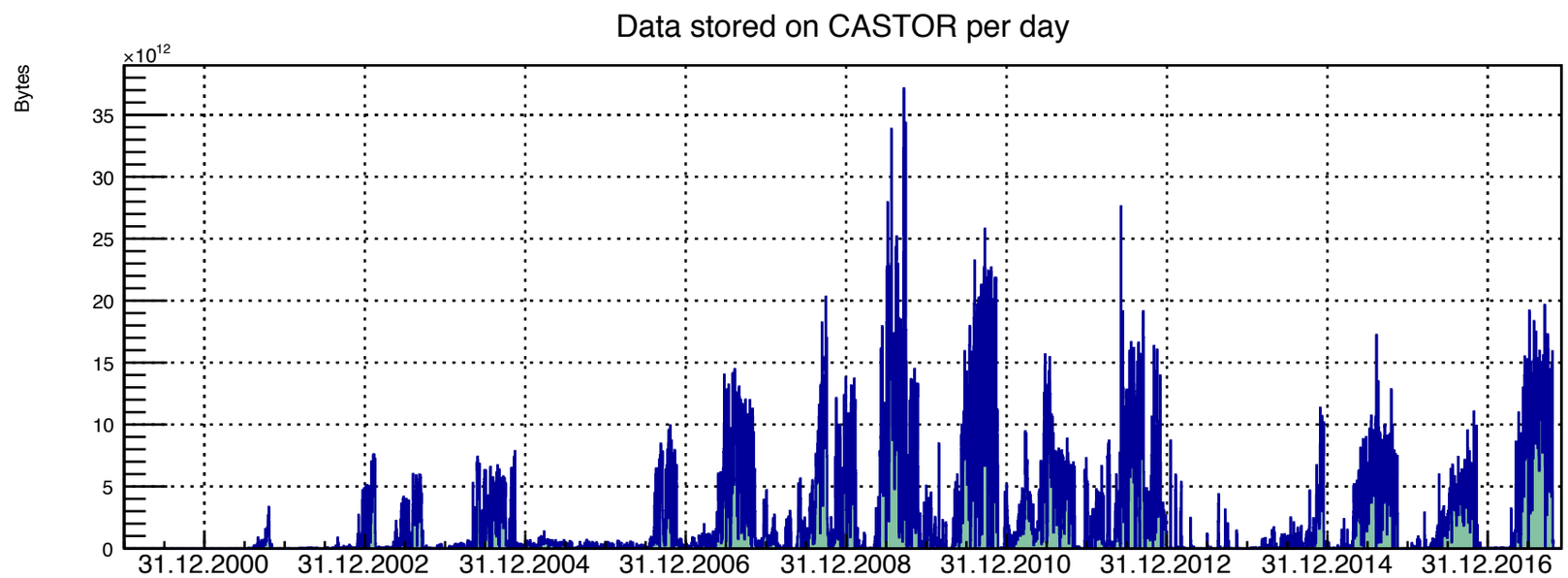
Production work flow

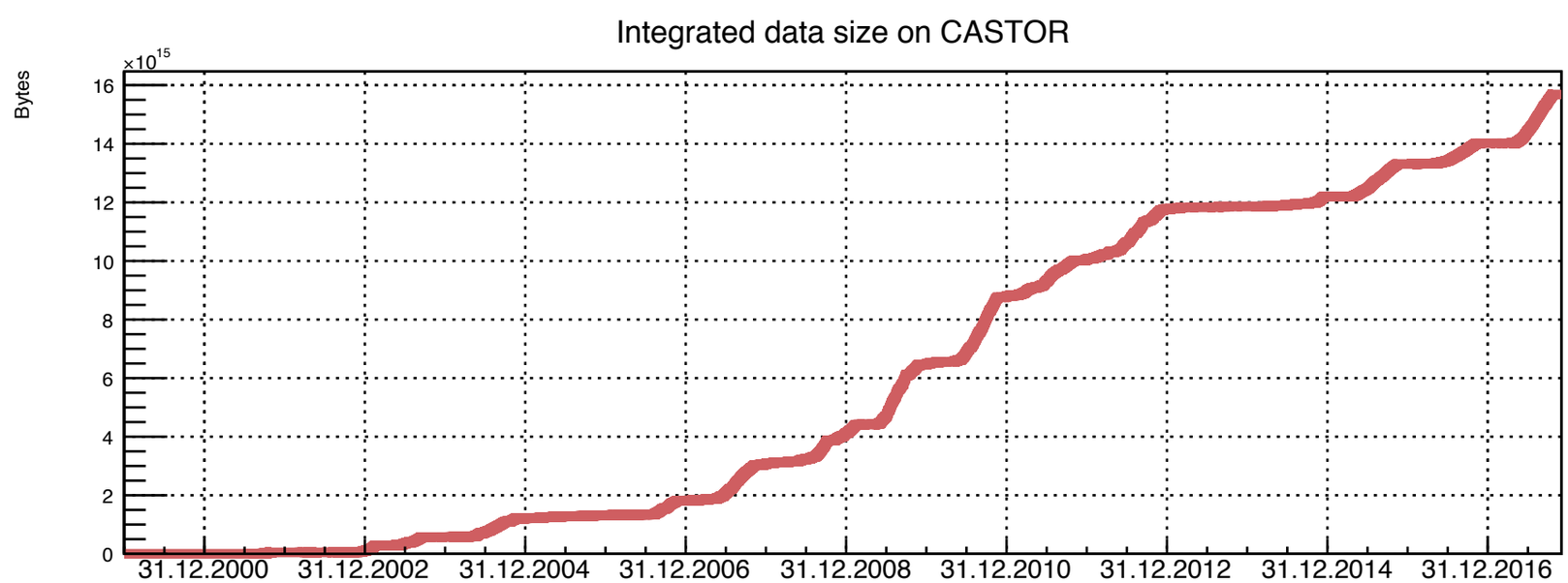
- All data stored on Castor
- Data is being requested to be copied from tapes to disks before processing (may take ~6 hours)
- Task moves files directly from Castor to lxbatch for processing, several programs are used for processing
- After processing results are being transferred to EOS for merging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



Data taken and stored

2001 - 13 TB
2002 - 196
2003 - 230
2004 - 496
2006 - 390
2007 - 912
2008 - 523
2009 - 1223
2010 - 1740
2011 - 518
2012 - 878
2015 - 801
2016 - 571
2017 - 1391





Motivation

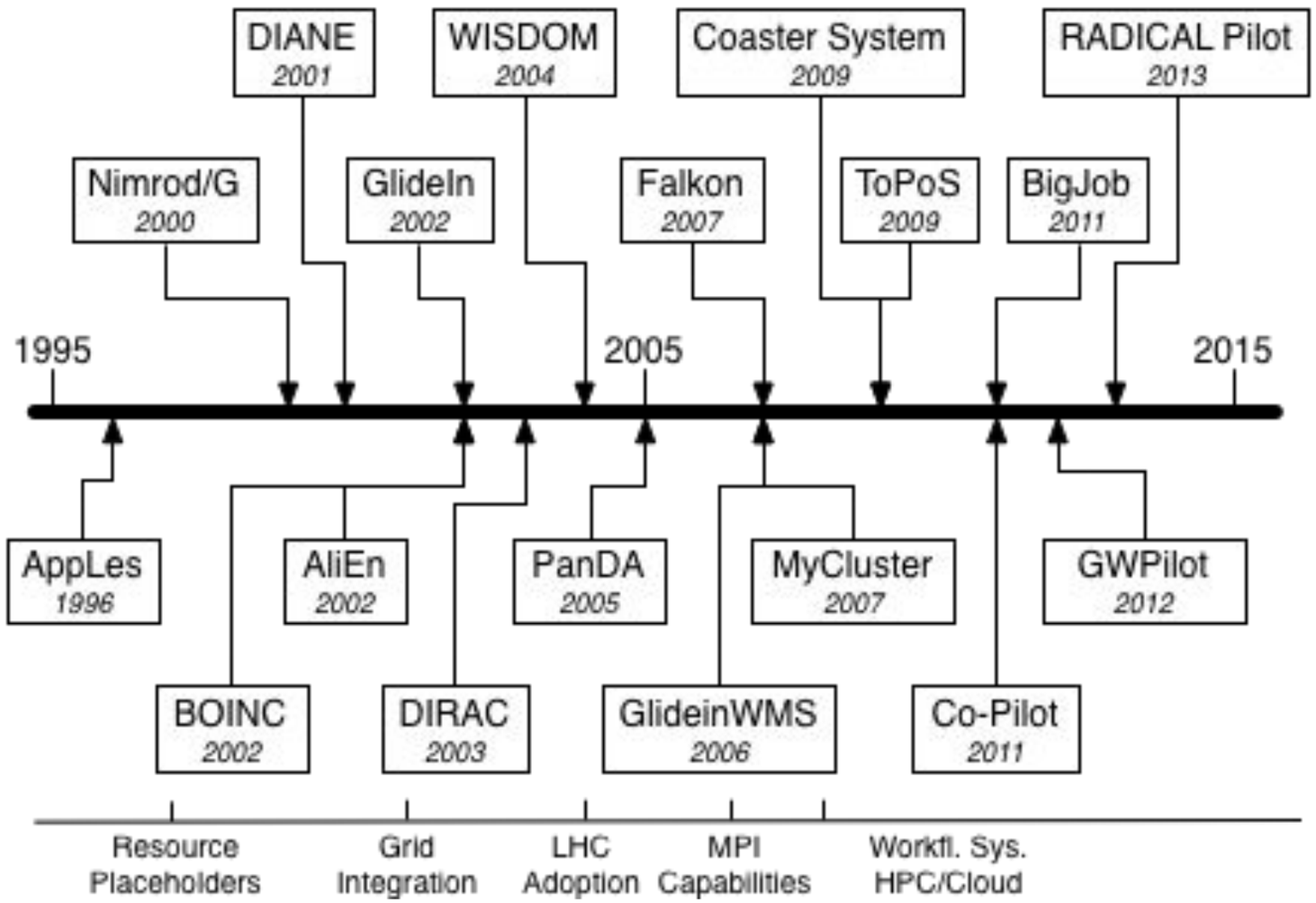
- Replace computing site from LSF, which will be decommissioned by the end of 2018, to Condor
  - Even more: get ability to switch computing sites, get more resources, any type, not only LSF
  - Even more: get machinery which is able to send jobs to some HPC
- Remove strict connectivity to AFS, which will be replaced by EOS FUSE
- Ease connection to CASTOR, which will be replaced by EOS

We need a WMS!

What is WMS?

- WMS — workload management system
- Providing a central queue for all users, makes hundreds of distributed sites appear as local
- Hides middleware while supporting diversity and evolution
  - WMS interacts with middleware, users see only high level workflow
  - Automation engines built in WMS, not exposed to users
- Hides variations in infrastructure
  - WMS presents uniform ‘job’ slots to user
  - Easy to integrate grid sites, clouds, HPC sites
- Uses the same system for simulation, data processing and users analysis
- Similar ideas have been implemented in several independent systems developed by LHC experiments: AliEn, Dirac, PanDA

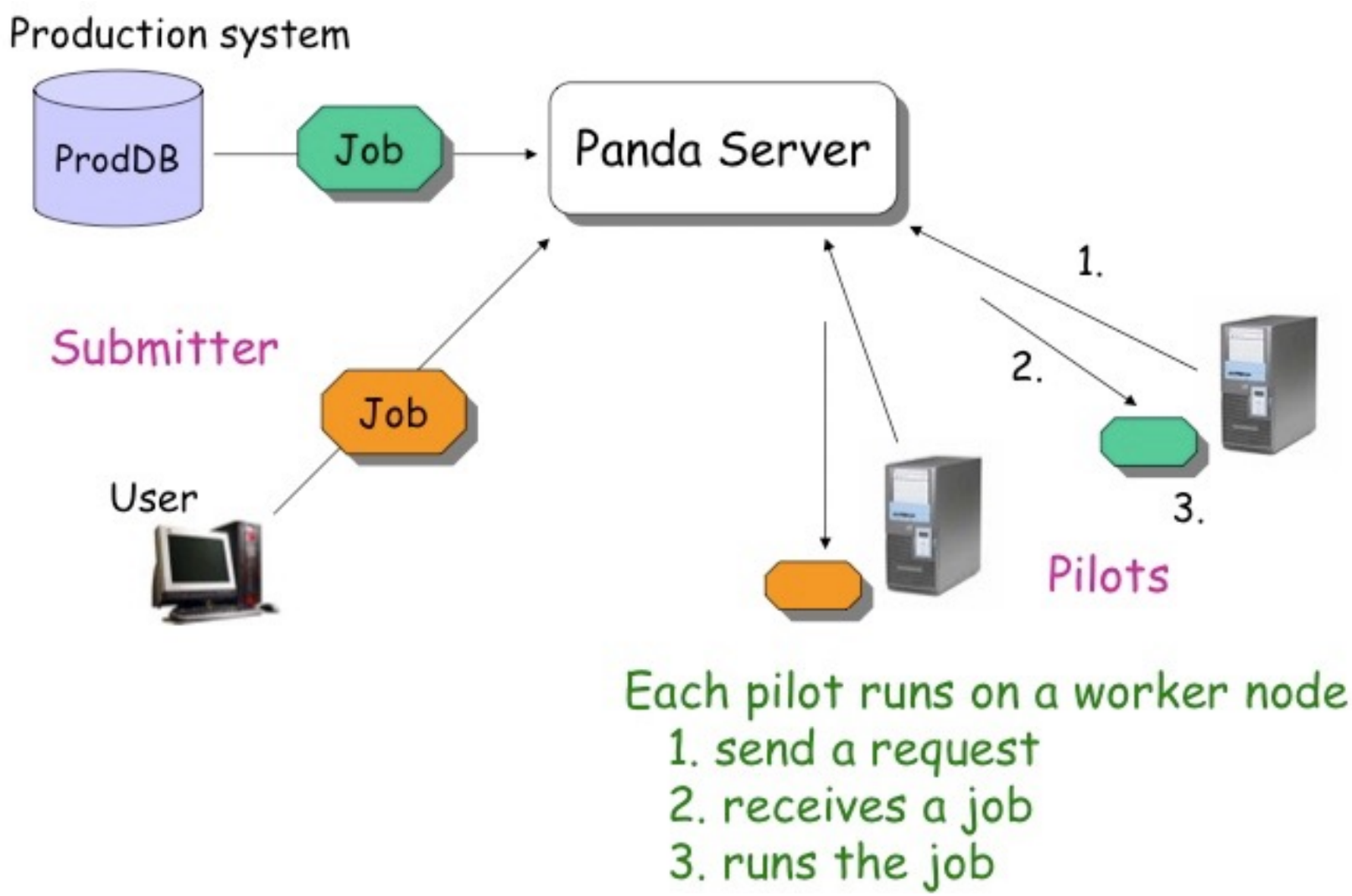
WMS evolution



What is PanDA?

- The PanDA Production and Distributed Analysis System has been developed by ATLAS to meet requirements of data-driven workload management system for production and distributed analysis processing capable at LHC data processing scale
- PanDA manages both user analysis and production jobs via same interface
- PanDA processing rate is 250-300K jobs on ~170 sites every day
- The PanDA ATLAS analysis user community numbers over 1400
- Supports classic Grid computing resources, clouds, HPCs

PanDA job workflow



Steps to enable distributed processing

- WMS instance installation, COMPASS logic implementation in Pilot code
- Production chain workflow and data flow management software preparation
- Grid environment setup
- PanDA monitoring adaptation to COMPASS

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

New ProdSys Components

1. Task requests layer: Web UI over Django framework
2. Job definition layer: automatic, python script
3. Job execution layer: PanDA
4. Workflow management: python scripts
5. Data management: automatic, python scripts
6. Monitoring: PanDA monitoring, adapted to COMPASS

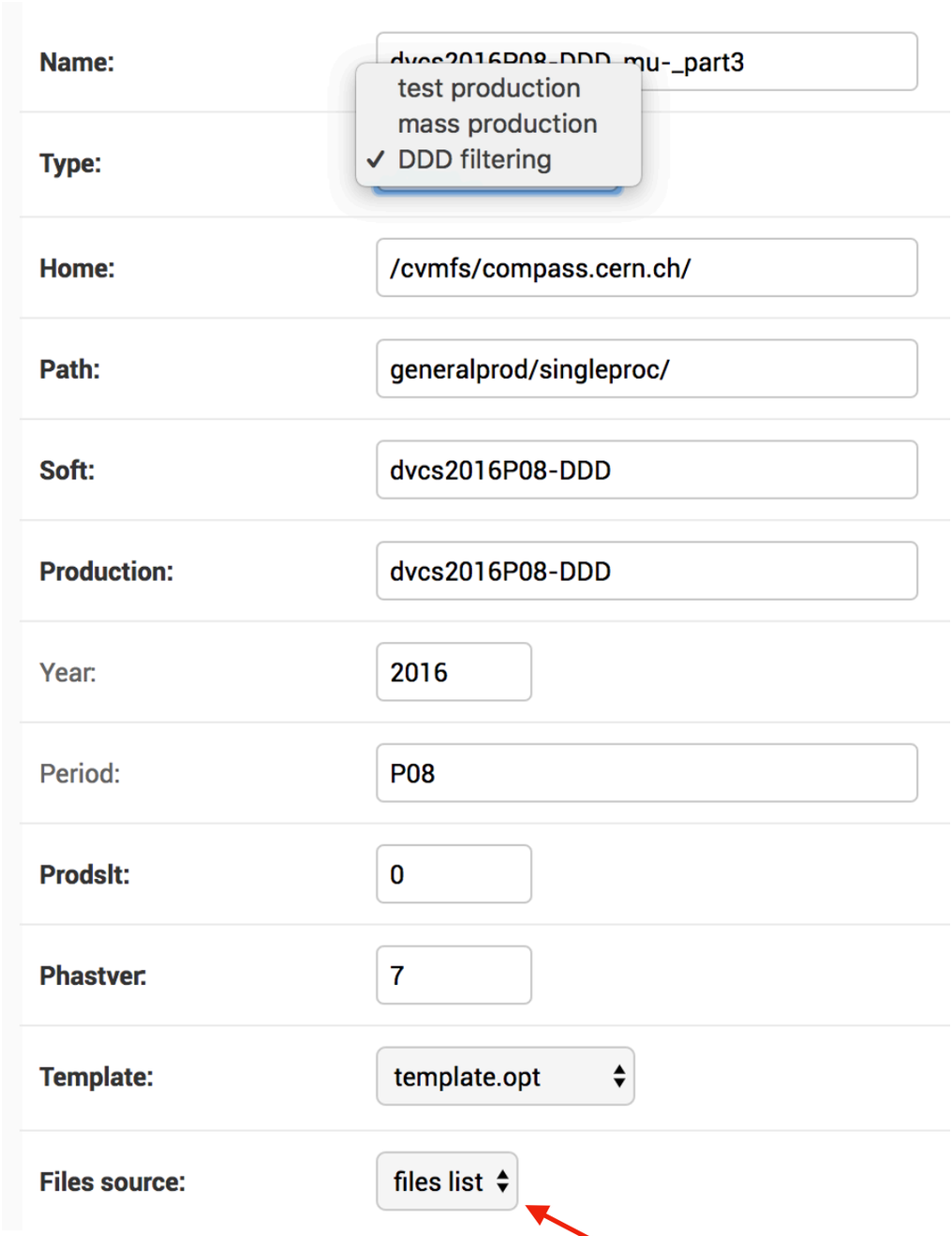
Infrastructure overview

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

1. Task requests layer

Web UI:

- execution parameters
- paths
- version of software
- list of chunks or runs
- etc



May be list of runs as well

2. Job definition layer

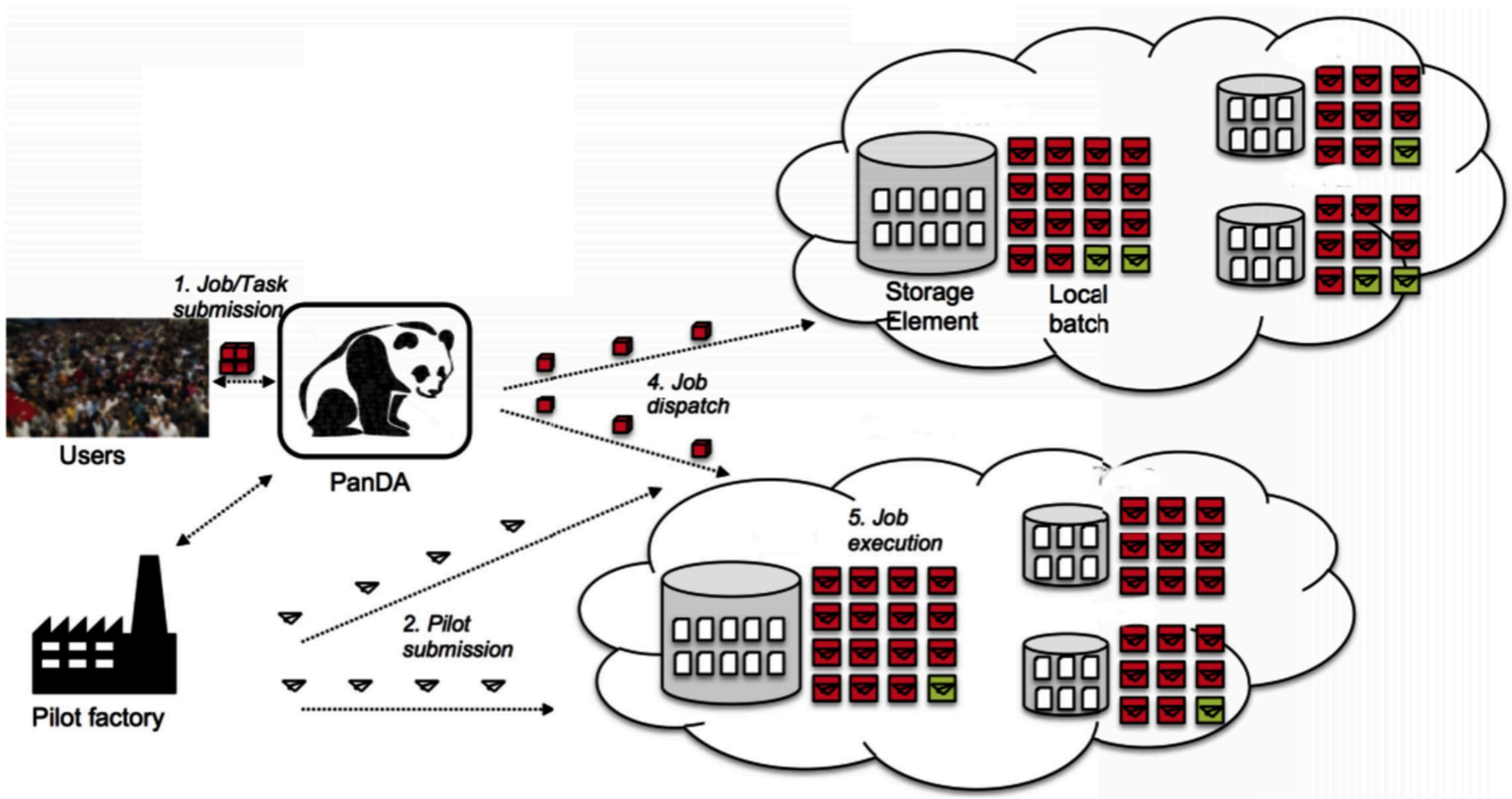
Automatically generates list of jobs for task basing on parameters

Job actions allow to manage any set of selected chunks

Action: ✓ ---------
Delete selected jobs
Resend selected jobs
Resend merging mdst of selected jobs
Resend merging hist of selected jobs
Resend x-check of selected jobs
Resend merging eventdump of selected jobs
Go 0 of 100 selected

TA		RUN NUMBER	CHUNK NUMBER	PANDA ID	ATTEMPT	STATUS
dv	/2016/raw/W14/cdr11091-	275678	11091	2182400	1	finished
dv	/2016/raw/W14/cdr11082- 275678.raw	275678	11082	2182399	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11080- 275678.raw	275678	11080	2182398	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11089- 275678.raw	275678	11089	2182397	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11086- 275678.raw	275678	11086	2182396	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11063- 275678.raw	275678	11063	2182395	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11049- 275678.raw	275678	11049	2182394	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11016- 275678.raw	275678	11016	2182393	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11094- 275678.raw	275678	11094	2182392	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11092- 275678.raw	275678	11092	2182391	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11088- 275678.raw	275678	11088	2182390	1	finished
dvcs2016P09t2r13_mu+	/castor/cern.ch/compass/data/2016/raw/W14/cdr11076- 275678.raw	275678	11076	2182389	1	finished

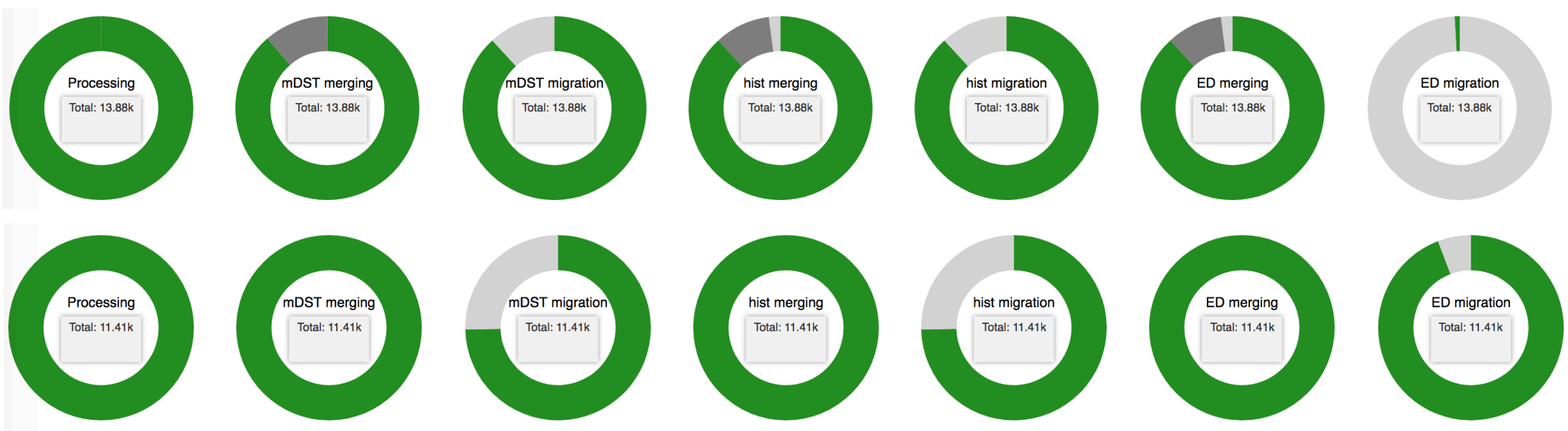
3. Job execution layer: PanDA



4: Workflow management

Decision making mechanisms, guide task from the definition till archive

Each step of each task is managed independently



5: Data management

Automatic stage-in from Castor and stage-out to Castor

```
[2017-11-09 18:40:01,633] INFO [prepare_on_castor:55] Getting jobs with status defined for task dvcs2016P09t2ST03OP_mu+
[2017-11-09 18:40:01,634] INFO [prepare_on_castor:57] Got list of 405 jobs
[2017-11-09 18:40:01,674] INFO [prepare_on_castor:68] Going to request state of file /castor/cern.ch/compass/data/2016/raw/W14/cdr12043-275518.raw
[2017-11-09 18:40:01,674] INFO [prepare_on_castor:70] stager_qry -M /castor/cern.ch/compass/data/2016/raw/W14/cdr12043-275518.raw -S compasscdr
[2017-11-09 18:40:04,746] INFO [prepare_on_castor:73] /castor/cern.ch/compass/data/2016/raw/W14/cdr12043-275518.raw 1568539545@castorns STAGED
[2017-11-09 18:40:04,747] INFO [prepare_on_castor:86] File /castor/cern.ch/compass/data/2016/raw/W14/cdr12043-275518.raw staged, going to update status
[2017-11-09 18:40:04,751] INFO [prepare_on_castor:93] Job 600021 updated at 2017-11-09 18:40:01.427251
[2017-11-09 18:40:04,751] INFO [prepare_on_castor:68] Going to request state of file /castor/cern.ch/compass/data/2016/raw/W14/cdr12044-275518.raw
[2017-11-09 18:40:04,751] INFO [prepare_on_castor:70] stager_qry -M /castor/cern.ch/compass/data/2016/raw/W14/cdr12044-275518.raw -S compasscdr
[2017-11-09 18:40:05,545] INFO [prepare_on_castor:73] /castor/cern.ch/compass/data/2016/raw/W14/cdr12044-275518.raw 1568539546@castorns STAGED
[2017-11-09 18:40:05,545] INFO [prepare_on_castor:86] File /castor/cern.ch/compass/data/2016/raw/W14/cdr12044-275518.raw staged, going to update status
[2017-11-09 18:40:05,548] INFO [prepare_on_castor:93] Job 600022 updated at 2017-11-09 18:40:01.427251
[2017-11-09 18:40:05,549] INFO [prepare_on_castor:68] Going to request state of file /castor/cern.ch/compass/data/2016/raw/W14/cdr12045-275518.raw
[2017-11-09 18:40:05,549] INFO [prepare_on_castor:70] stager_qry -M /castor/cern.ch/compass/data/2016/raw/W14/cdr12045-275518.raw -S compasscdr
[2017-11-09 18:40:06,184] INFO [prepare_on_castor:73] /castor/cern.ch/compass/data/2016/raw/W14/cdr12045-275518.raw 1568539548@castorns STAGED
[2017-11-09 18:40:06,184] INFO [prepare_on_castor:86] File /castor/cern.ch/compass/data/2016/raw/W14/cdr12045-275518.raw staged, going to update status
[2017-11-09 18:40:06,187] INFO [prepare_on_castor:93] Job 600023 updated at 2017-11-09 18:40:01.427251
[2017-11-09 18:40:06,187] INFO [prepare_on_castor:68] Going to request state of file /castor/cern.ch/compass/data/2016/raw/W14/cdr12046-275518.raw
[2017-11-09 18:40:06,187] INFO [prepare_on_castor:70] stager_qry -M /castor/cern.ch/compass/data/2016/raw/W14/cdr12046-275518.raw -S compasscdr
[2017-11-09 18:40:06,884] INFO [prepare_on_castor:73] /castor/cern.ch/compass/data/2016/raw/W14/cdr12046-275518.raw 1568539855@castorns STAGED
[2017-11-09 18:40:06,884] INFO [prepare_on_castor:86] File /castor/cern.ch/compass/data/2016/raw/W14/cdr12046-275518.raw staged, going to update status
[2017-11-09 18:40:06,887] INFO [prepare_on_castor:93] Job 600024 updated at 2017-11-09 18:40:01.427251
```

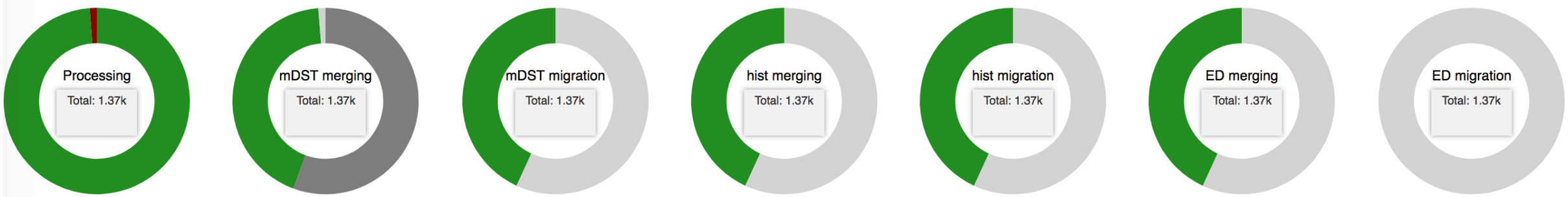
```
[2017-11-16 00:23:02,177] INFO [<module>:28] Starting /srv/compass/prodsys/periodic_tasks/check_castor_dump_status.pyc
[2017-11-16 00:23:02,199] INFO [<module>:31] pid: 7166
[2017-11-16 00:23:02,206] INFO [check_files_on_castor:46] Getting evntdmp chunks with castor evntdmp status sent
[2017-11-16 00:23:02,561] INFO [check_files_on_castor:48] Got list of 1 chunks
[2017-11-16 00:23:02,561] INFO [check_files_on_castor:49] Getting productions with castor evntdmp status sent
[2017-11-16 00:23:03,234] INFO [check_files_on_castor:51] Got list of 1 prods
[2017-11-16 00:23:03,234] INFO [check_files_on_castor:54] Going to request list of files on castor for task 107
[2017-11-16 00:23:03,234] INFO [check_files_on_castor:56] nsls -l /castor/cern.ch/compass/generalprod/testcoral/dvcs2016P09t2r13/mergedDump/slot0/
[2017-11-16 00:23:06,691] INFO [check_files_on_castor:60] Successfully read files on castor for task 107
[2017-11-16 00:23:06,691] INFO [check_files_on_castor:68] mrwxr--r--   1 na58dst1 vy               2037741028 Nov 15 21:27 evtdump0-275678.raw
[2017-11-16 00:23:06,691] INFO [check_files_on_castor:69] Found "m" file for task id 107 run number 275678 chunk number 0, evtdump0-275678.raw
[2017-11-16 00:23:06,691] INFO [check_files_on_castor:71] Going to update jobs of the chunk as migrated
[2017-11-16 00:23:06,698] INFO [check_files_on_castor:74] Job status_castor_evntdmp changed to finished for task 107 run number 275678 chunk number 0
[2017-11-16 00:23:06,698] INFO [check_files_on_castor:85] done
```

6.1: PanDA monitoring

Covers all activity during production/task/job lifecycle

The summary for the **dvcs2016P09t2r13MB** production started on 23 Nov 2017. The total number of chunks is 1368. The average walltime of a finished job is 146 minutes. Built 12:34 Actual version

Tasks for dvcs2016P09t2r13MB | Jobs for dvcs2016P09t2r13MB



Show 50 entries Search:

Run	Number of chunks	Defined	Sent	Running	Failed	Finished	Status of mDST merging	X-checked	mDST migration	Status of histogram merging	Histogram migration	Status of event dump merging	Event dump migration
275518	405	-	-	-	10	395	-	no	-	-	-	-	-
275603	337	-	-	-	-	337	finished	yes	finished	finished	finished	finished	-
275678	373	-	-	-	7	366	-	no	-	-	-	-	-
275744	253	-	-	-	-	253	finished	yes	finished	finished	finished	finished	-

Showing 1 to 4 of 4 entries Previous 1 Next

6.2: PanDA monitoring

Job attribute summary	Sort by count, alpha
attemptnr (9)	1 (4) 2 (8006) 3 (3468) 4 (1521) 5 (919) 6 (278) 7 (76) 8 (13) 11 (8)
computingsite (1)	CERN_COMPASS_PROD (14293)
destinationse (1)	local (14293)
jobstatus (8)	activated (243) defined (1) failed (2176) finished (7824) holding (164) running (99) sent (3770) starting (16)
minramcount (1)	0-1GB (14293)
priorityrange (2)	1000:1099 (13) 3000:3099 (14280)
prodsourcelabel (1)	prod_test (14293)
produsername (1)	Artem Petrosyan (14293)
taskid (6)	108 (1969) 109 (1606) 110 (1965) 111 (2834) 112 (2226) 113 (3693)
transformation (2)	DDD filtering (14280) merging dump (13)

Overall error summary

Category:code	Attempt list	Nerrors	% of job selection	Sample error description
jobdispatcher:102	jobs	2175	15.22	Sent job didn't receive reply from pilot within 30 min
transformation:1	jobs	1	0.01	Unspecified error, consult log file

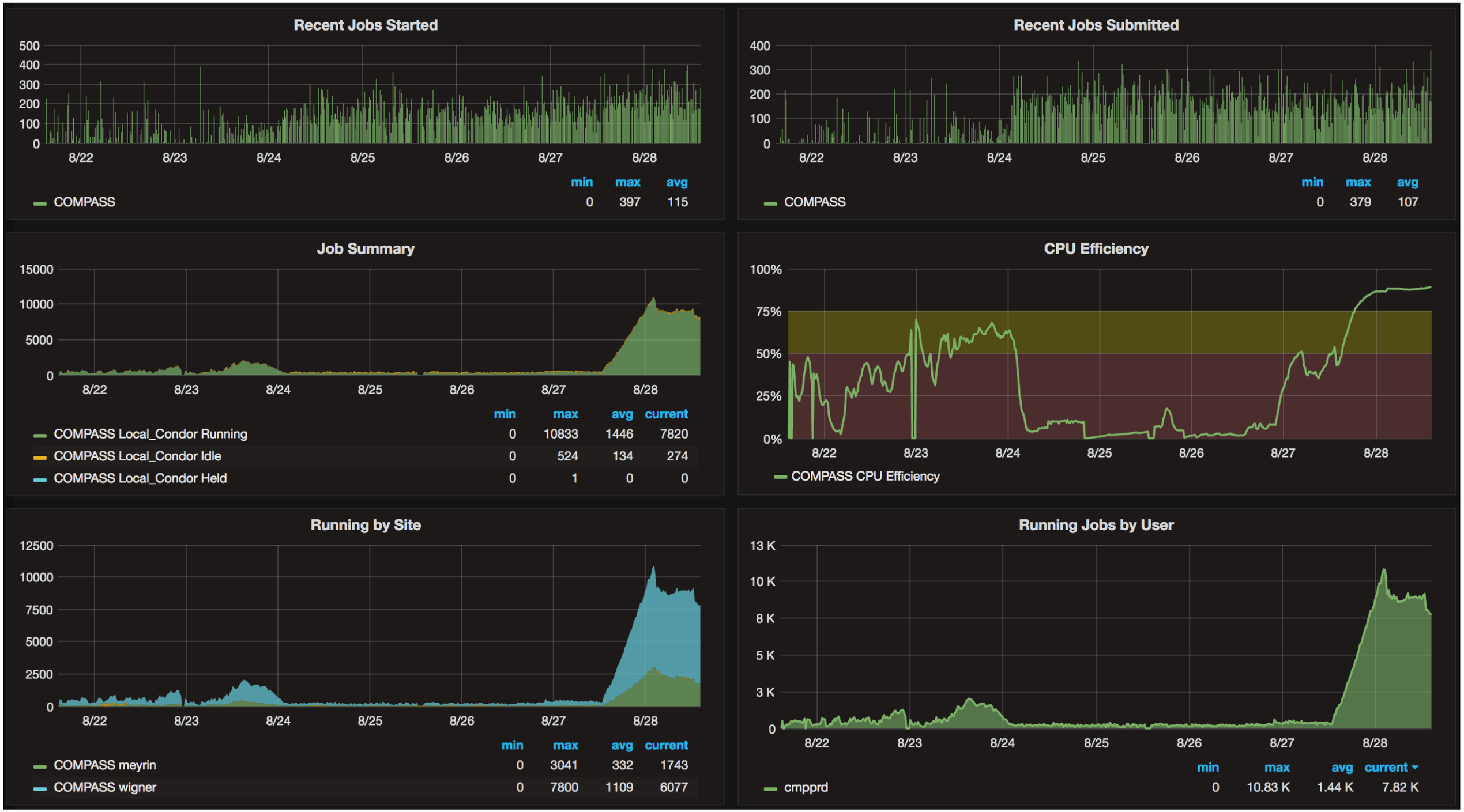
6.3: AutoPyFactory monitoring

Factory view

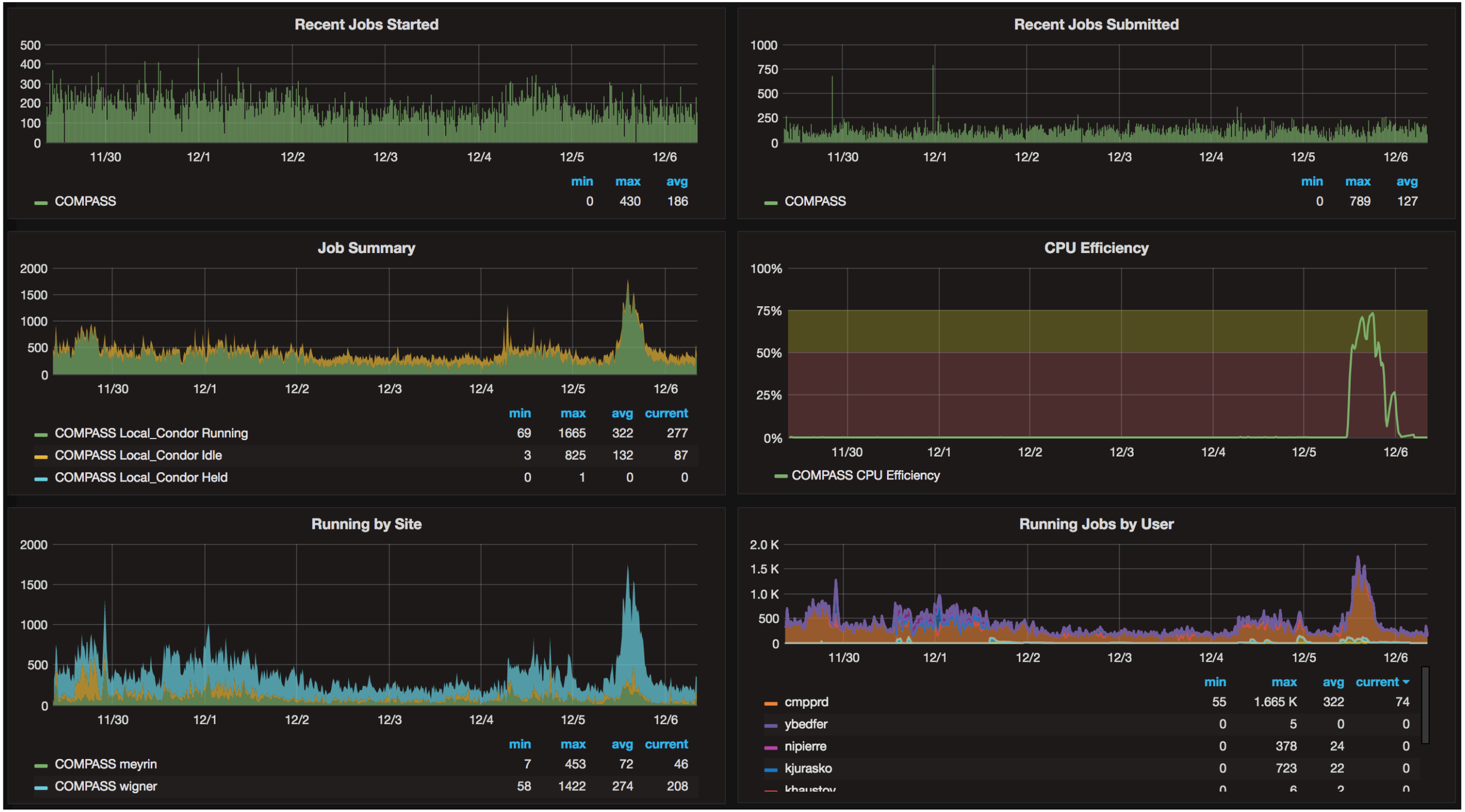
Factory	JINR-pandawms
Version	2.4.9
Last startup	2 days ago
Email	artem.petrosyan@jinr.ru
Activity	86
Links	logs queues.conf

Factory label	last msg
CERN_COMPASS_PROD-ce301-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce302-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce401-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce402-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce403-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce404-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce405-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce406-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce407-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce408-cern-ch	6 mins ago
CERN_COMPASS_PROD-ce503-cern-ch	4 mins ago
CERN_COMPASS_PROD-ce504-cern-ch	seconds ago
CERN_COMPASS_PROD-ce505-cern-ch	1 min ago
CERN_COMPASS_PROD-ce506-cern-ch	3 mins ago
CERN_COMPASS_PROD-ce507-cern-ch	4 mins ago
CERN_COMPASS_PROD-ce508-cern-ch	3 mins ago
CERN_COMPASS_PROD-condorce01-cern-ch	4 mins ago
CERN_COMPASS_PROD-condorce02-cern-ch	5 mins ago
CNAF_COMPASS_PROD-ce04-lcg-cr-cnaf-infn-it	6 mins ago
JINR_COMPASS_PROD-lcgce12-jinr-ru	6 mins ago
JINR_COMPASS_PROD-lcgce21-jinr-ru	6 mins ago
TRIESTE_COMPASS_PROD-ce1-ts-infn-it	6 mins ago

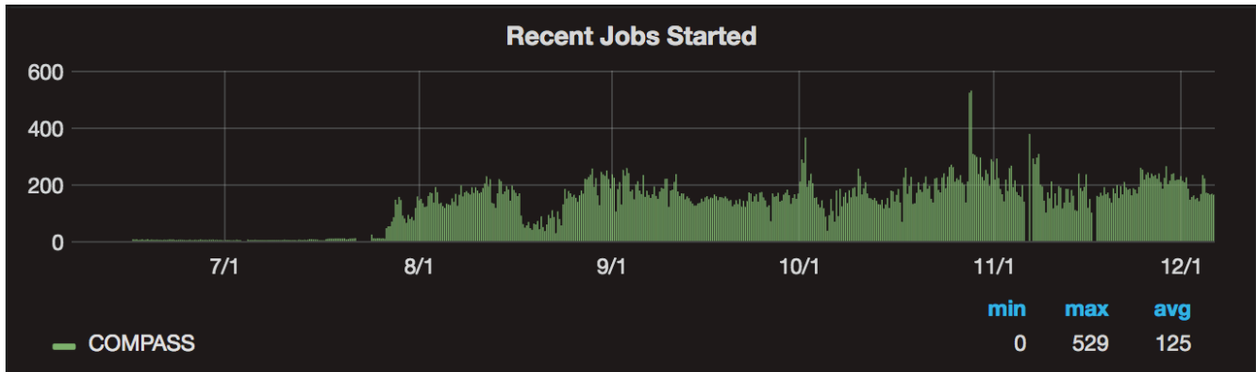
6.4: CERN Condor monitoring

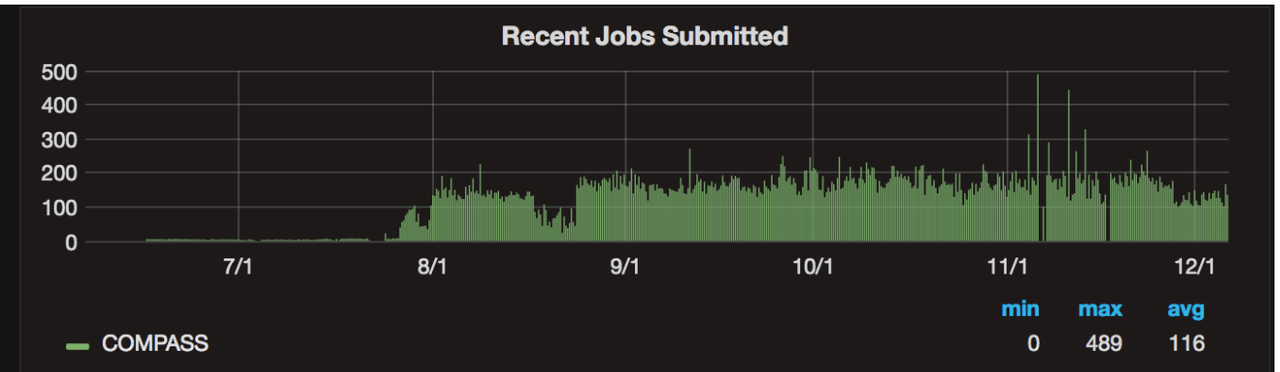


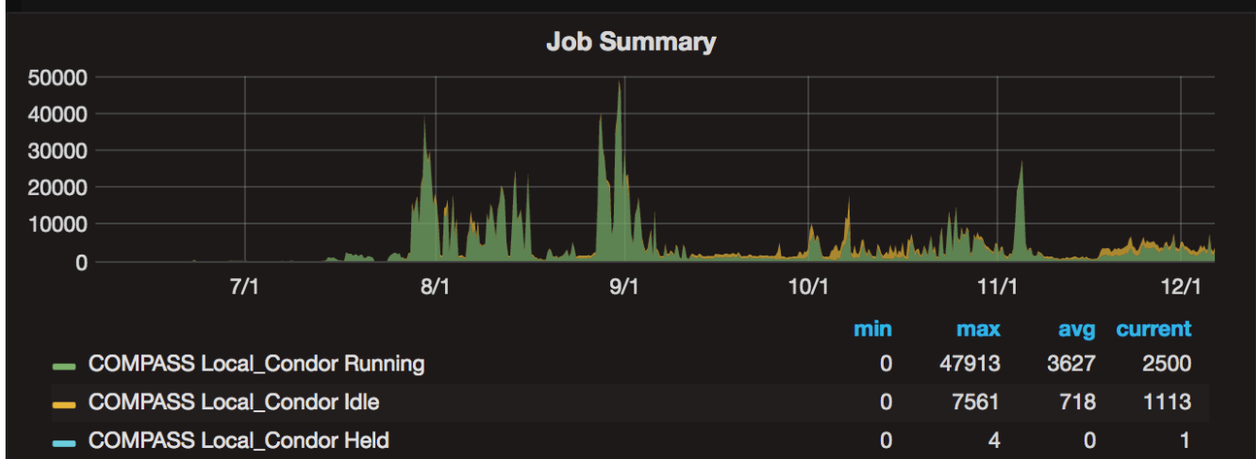
6.4: CERN Condor monitoring

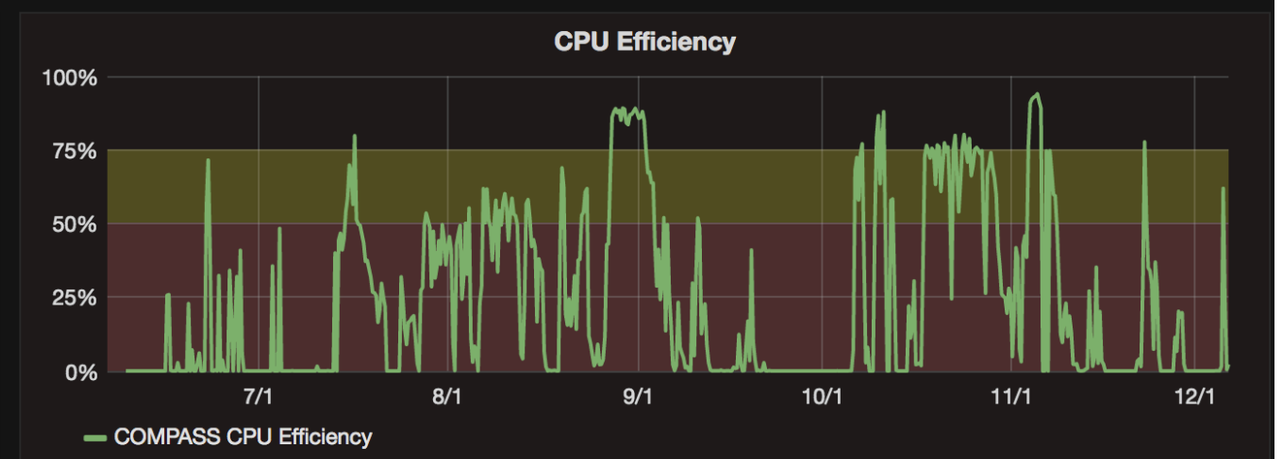


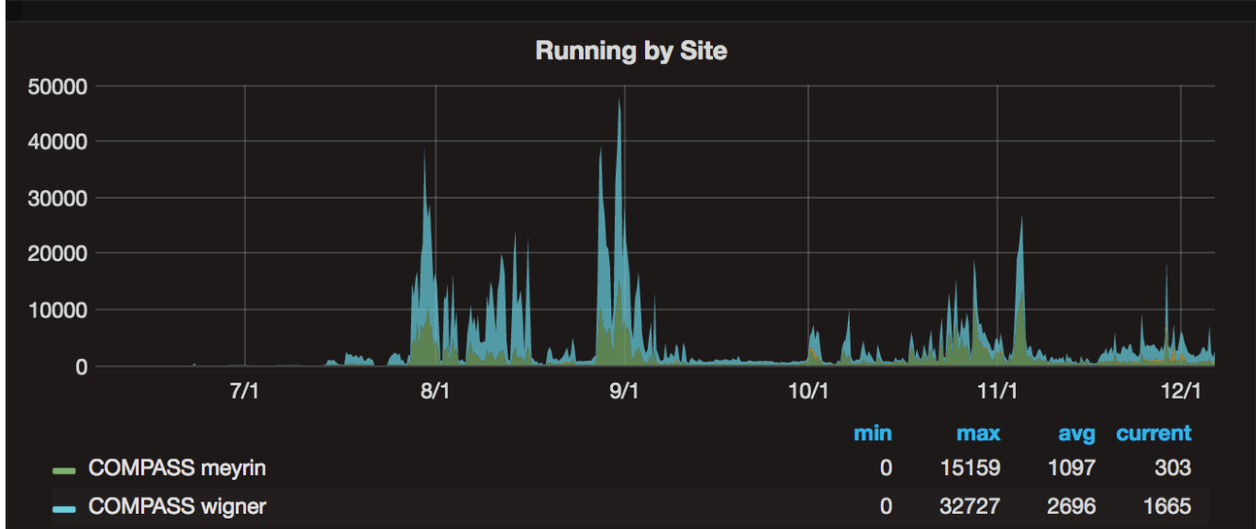
6.4: CERN Condor monitoring

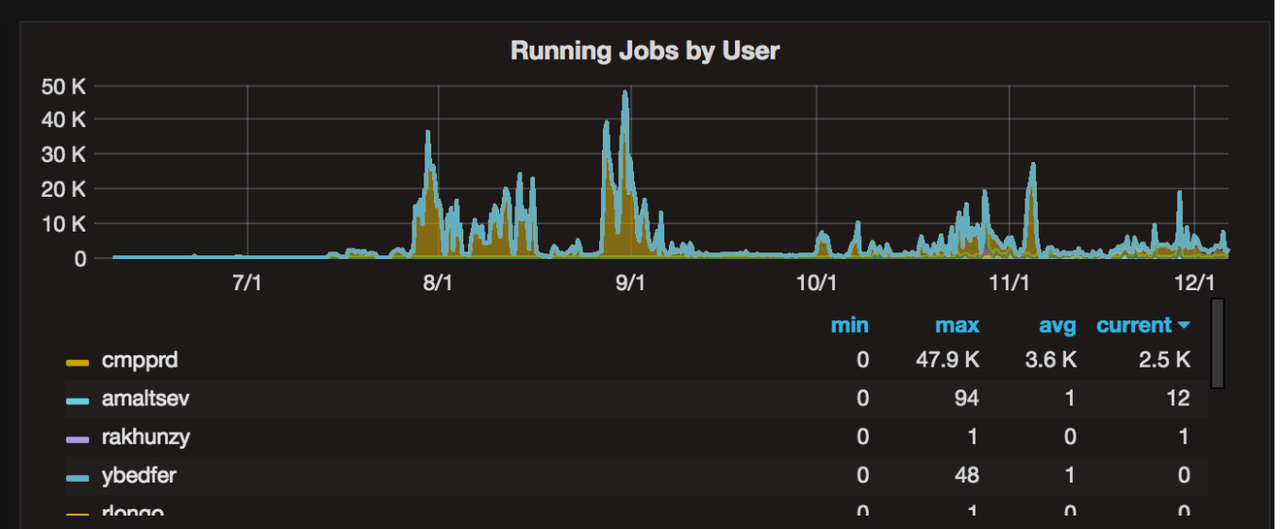












6.5: JINR T2 jobs per VO stats

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

Stats since August

- ~600 000 chunks processed
- ~100TB of merged data produced and migrated to Castor
- ~2 500 000 jobs processed since August: reco, ddd filtering, merging of mDST, hist and event dumps
- ~3 000 000 processing hours
- ~ 3 500 000 wall-time, time spent on computing nodes, including stage-in and stage-out

Summary

- COMPASS Grid Production System provides automated processing from definition till archiving
- Key features:
  - Production management is done via Web UI, which allows to define a task, send, follow and manage it at any step
  - Via PanDA layer jobs may be sent to any type of computing resource: Condor, LSF, PBS, etc.
  - Rich monitoring
- Positive side effects:
  - COMPASS software moves to CVMFS
  - We almost got rid of AFS

Next steps

- Production on BlueWaters HPC
  - SW on BW is ready
  - Data is there already
  - There are setups on facilities like BW in ATLAS
- Extend data management component
  - We may achieve storage and transfer protocol independence
  - Data transfers to and from any endpoint
- Add more (and existing) computing sites to the system, may be dedicated to particular type of processing
- MC workflow
- Users analysis