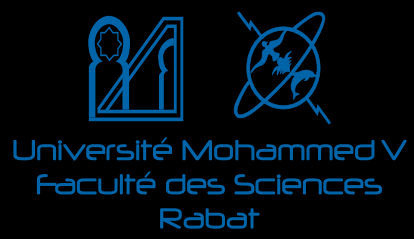


ATLAS workload management system: PanDA

Farida Fassi
Faculty of Sciences Mohammed V
University in Rabat, Morocco

Computing training course
I-COOP+ 2016 project: COOPB20247
3-14, July 2017. IFIC-Valencia, Spain

Distributed Analysis workflow

ATLAS specific
Front-end interface
GANGA, PATHENA

Grid middleware
EGEE/EGI, NorduGrid/ARC,
OSG/VDT

1.

2.

3.

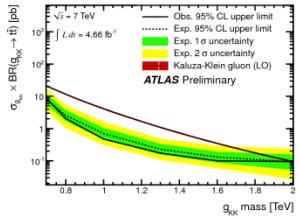
ATLAS specific software
Workload Management
System (PanDa);
Data Management (DDM)

Computing & Storage
resources

4.

User outputs

5.



6.

Publication

Production and Distributed Analysis system

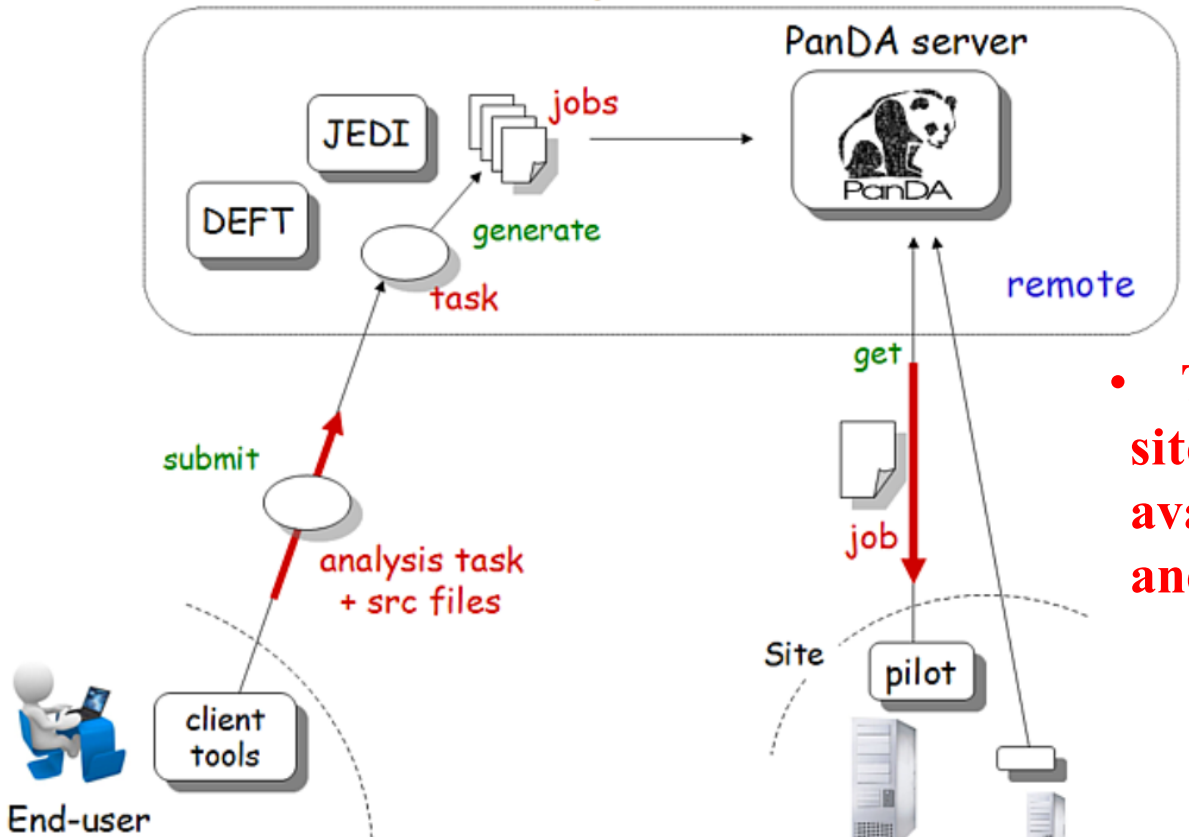
- Production and Distributed Analysis system (PanDA) is the Workload Management System (WMS) to run jobs on Grid.
  - PanDA is an unified system for Production and User Analysis capable of operating at LHC data processing scale.
  - PanDA makes distributed resources optimally accessible by all users.

PanDA Brief Story

2005: Initiated for US ATLAS (BNL and UTA)
2006: Support for analysis
2008: Adopted ATLAS-wide
2009: First use beyond ATLAS
2011: Dynamic data caching based on usage and demand
2012: BigPanDA
2014: Network-aware brokerage
2014 : Job Execution and Definition (JEDI) -dynamic job management
2014: JEDI- based Event Service
2015: New ATLAS Production System, based on PanDA/JEDI
2015 :Manage Heterogeneous Computing Resources: Cloud computing, HPC, etc
2016: PanDA beyond HEP : LSST, BlueBrain

How PanDA works?

- How PanDA works?
- Uses a Pilot model to pull jobs from central queue once a suitable resource found.
- Pilot factories
  - continually submit jobs to available computing resources.



- **The tasks are routed to sites based on the availability of relevant data and processing resources.**

DEFT component

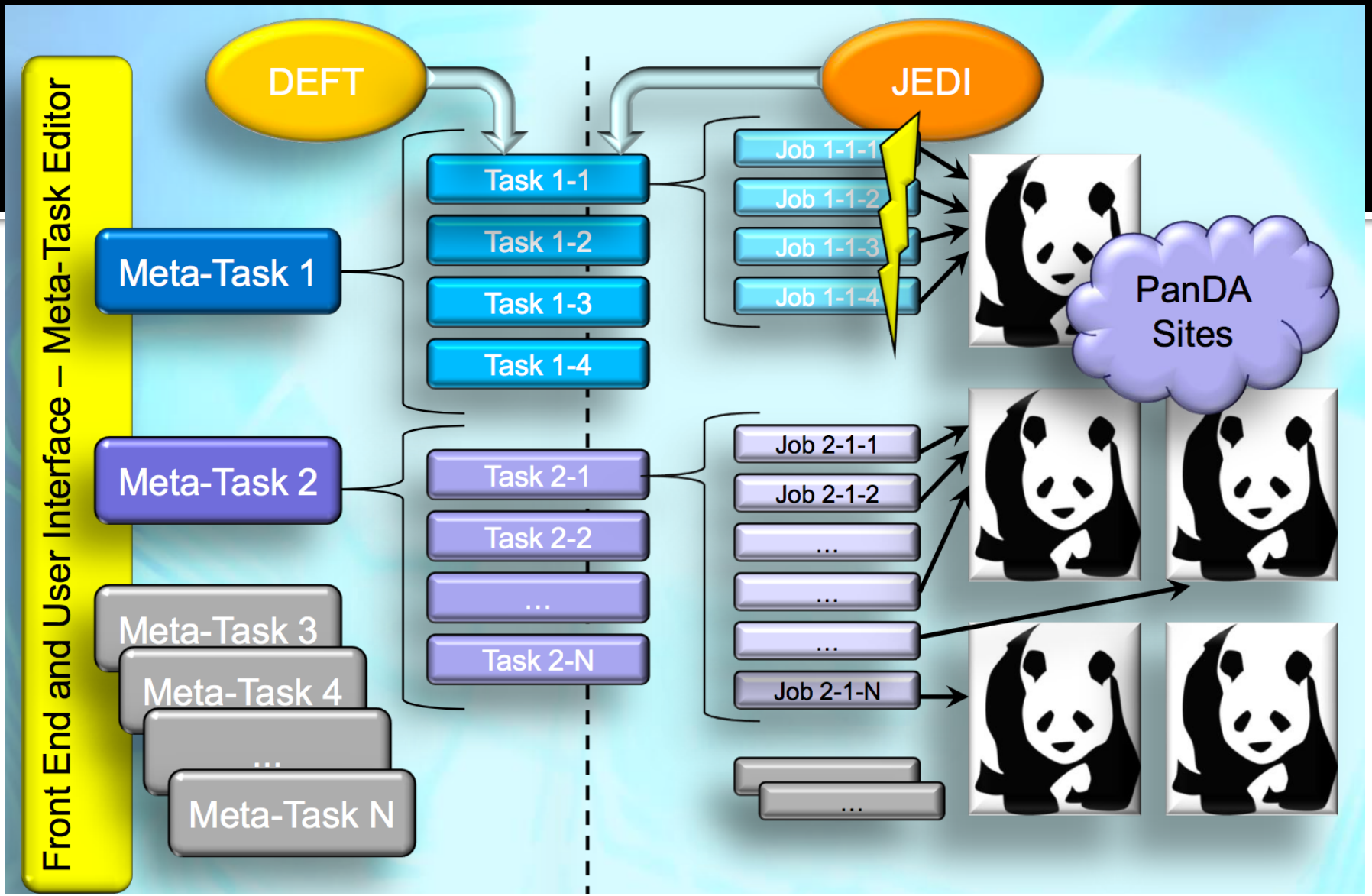
- **DEFT (Database Engine for Task)**
  - **Handles production requests and tasks**
    - For both production and analysis

- **DEFT is responsible for formulating the Meta-Tasks**
- Meta-Tasks can include chains of tasks and task groupings
  - **Completing with all necessary parameters.**

- **DEFT provides the interface** for each Meta-Task definition, management and monitoring throughout its lifecycle.

JEDI component

- **JEDI (Job Execution and Definition Interface)**
  - **Dynamically splits workload for optimal usage of resources**
- **JEDI is using the task definitions formulated in DEFT**
  - **To define and submit individual jobs to PanDA**
  - **To keep track of their progress and handle re-tries of failed jobs,**
  - **To perform job redirection**
- **JEDI interfaces data management services in order to properly aggregate and account for data generated by individual jobs (i.e. general dataset management)**



Tasks are submitted to the system and jobs are dynamically generated on behalf of users

Task, Job, Event

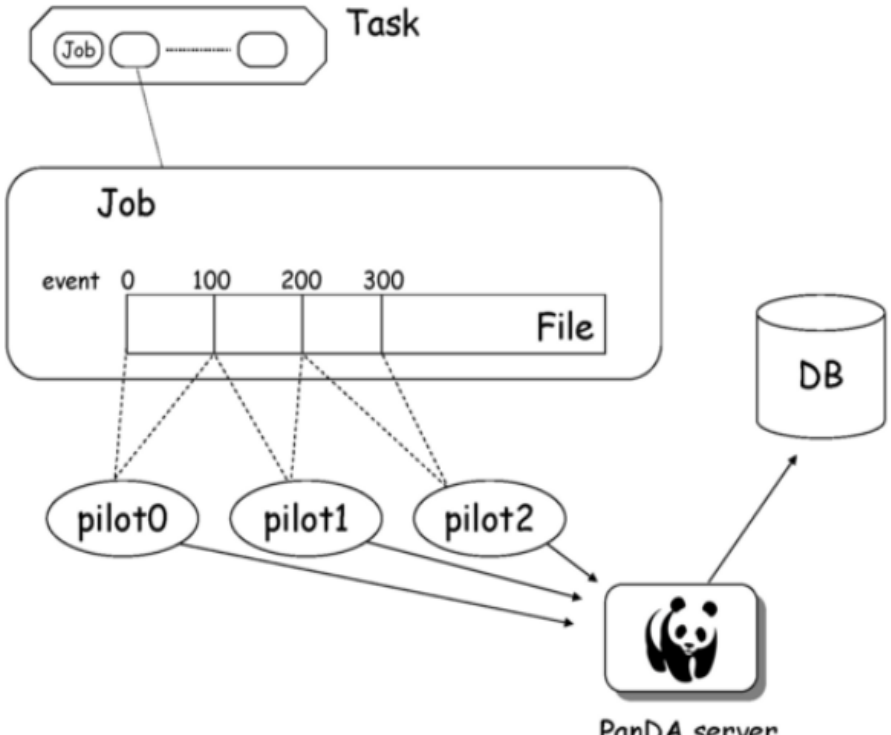
- **Meta-Task: task composed by multiple tasks**
- **Task: a collection of jobs.**
  - **Is used as a main unit of computation**
- **Job: small data processing tasks.**
- Splitting of task into jobs is similar to splitting in packets for networks
- **Event: smallest unit in ATLAS data and processing**
- **Users do not care about transient job failures**
  - **Completion time and low overall loss are the metrics**

Scout jobs functionality

- **Job metrics are unknown when tasks are defined:**
  - **total size of output files, local disk usage, total execution time, Input/Output intensity, etc**
- **Real values are collected using scout jobs**
  - A small number (~10) of jobs are generated for each task with minimum input chunks
- **Job parameters are optimized using job metrics for the rest of input**

Event Level Processing

- Job is split in many events chunks that can run separately
  - Different resources have different abilities
- **Pilot runs get Job to request work from Panda.**
- **A payload is returned from Panda which can be normal work**



- **Pilot parses the payload.**
- **Pilot can run one or more jobs**
- **Pilot automatically selects different processes for different jobs.**

PanDA Client Tools

- ***PanDA Client*** consists of several tools for job execution on the grid and bookkeeping
  - ***pathena*** – for submitting athena jobs to PanDA
  - ***prun*** – for general jobs (e.g. ROOT and Python scripts) to PanDA
  - ***psequencer*** allows for a sequence of different tasks to be submitted (e.g. an analysis job followed by the transfer of the output back to the local machine)
  - ***pbook*** is a bookkeeping tool for all PanDA analysis jobs (used e.g. for job retries and kills)

What is pathena?

- Client tool for PanDA used to submit user-defined jobs from the command line
- Works on the athena runtime environment
- A consistent user-interface to athena

When you run athena with

```
$ athena jobOptions.py
```

all you need to do to submit a job to the grid is

```
$ pathena jobOptions.py [--inDS inputDataset] --outDS outputDataset
```

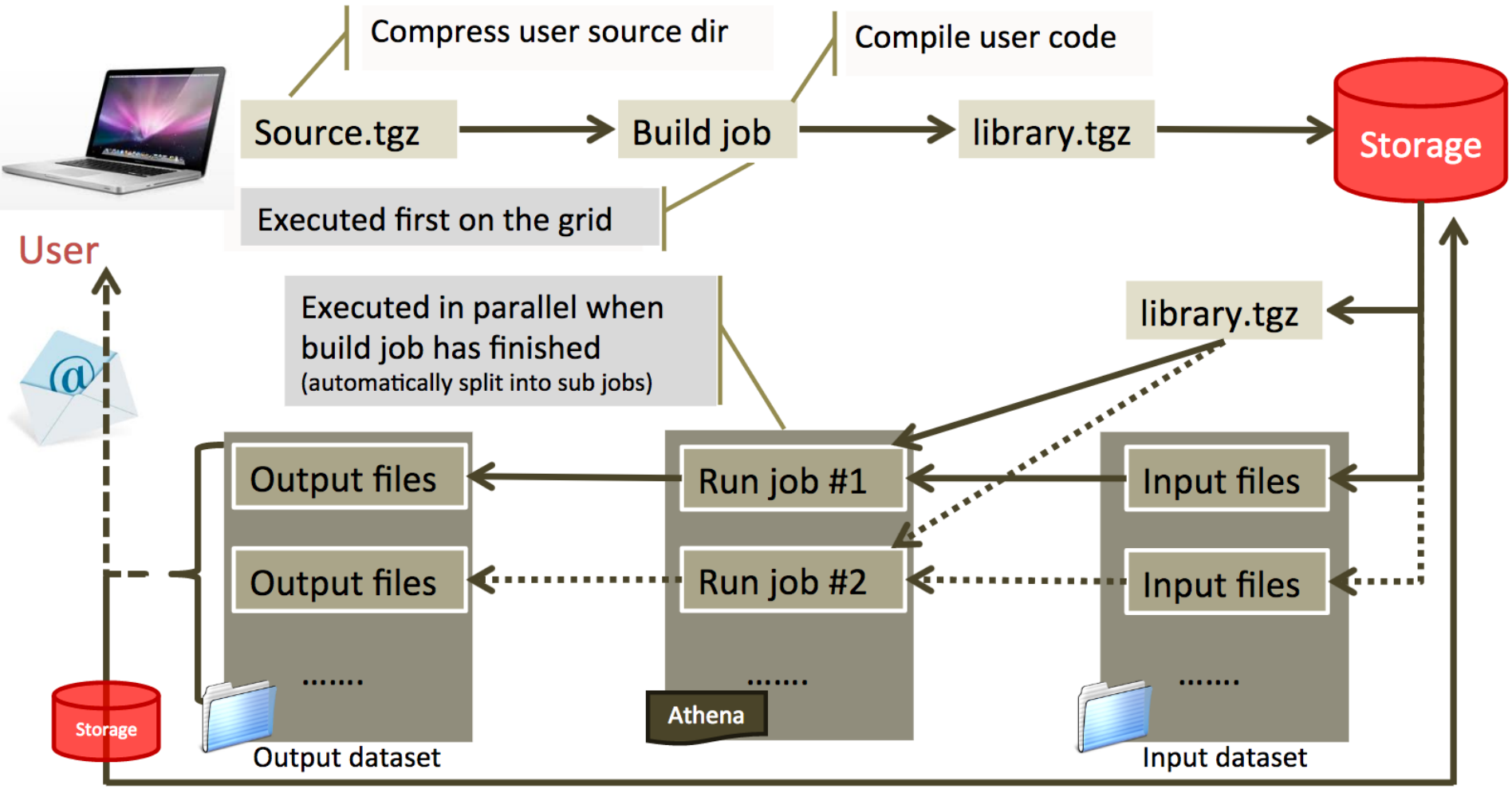
where **inputDataset** is a dataset which contains input files (optional), and **outputDataset** is a dataset which will contain output files (required)

- Simple to use, has advanced capabilities for complex needs (see `pathena -h`)

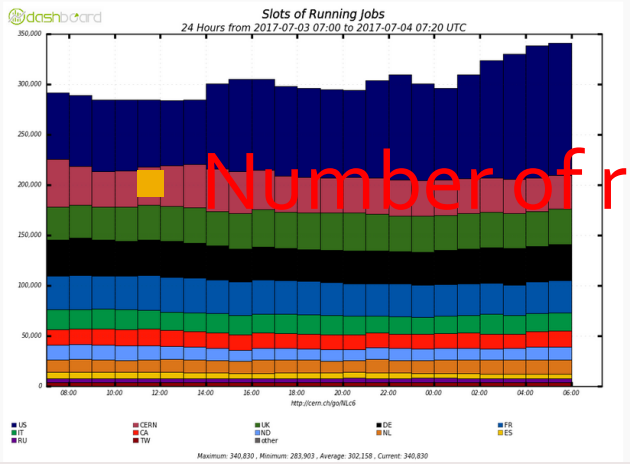
What is prun?

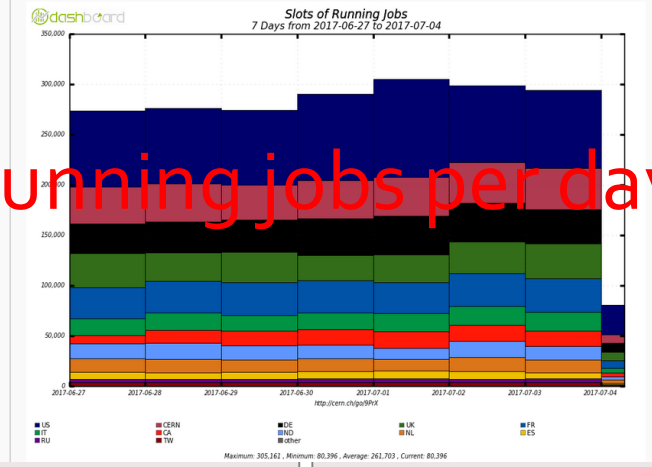
- **`prun`** (PanDA run) is a tool for submitting general jobs to PanDA, e.g. ROOT and Python scripts
- ATLAS analysis has two stages
  - Run athena on input files to produce some output (pathena)
  - Run ROOT, Python, or shell scripts to produce final plots (prun)
- PanDA part of the hands-on tutorial will start with sending a simple “Hello world” job to the grid and learning how to monitor the job and find the output!

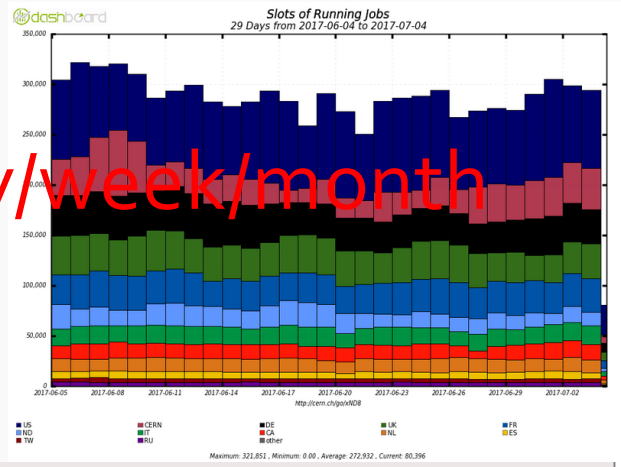
What happens when a job is submitted to the grid?



Monitoring PanDA jobs







Number of running jobs per day/week/month

Search

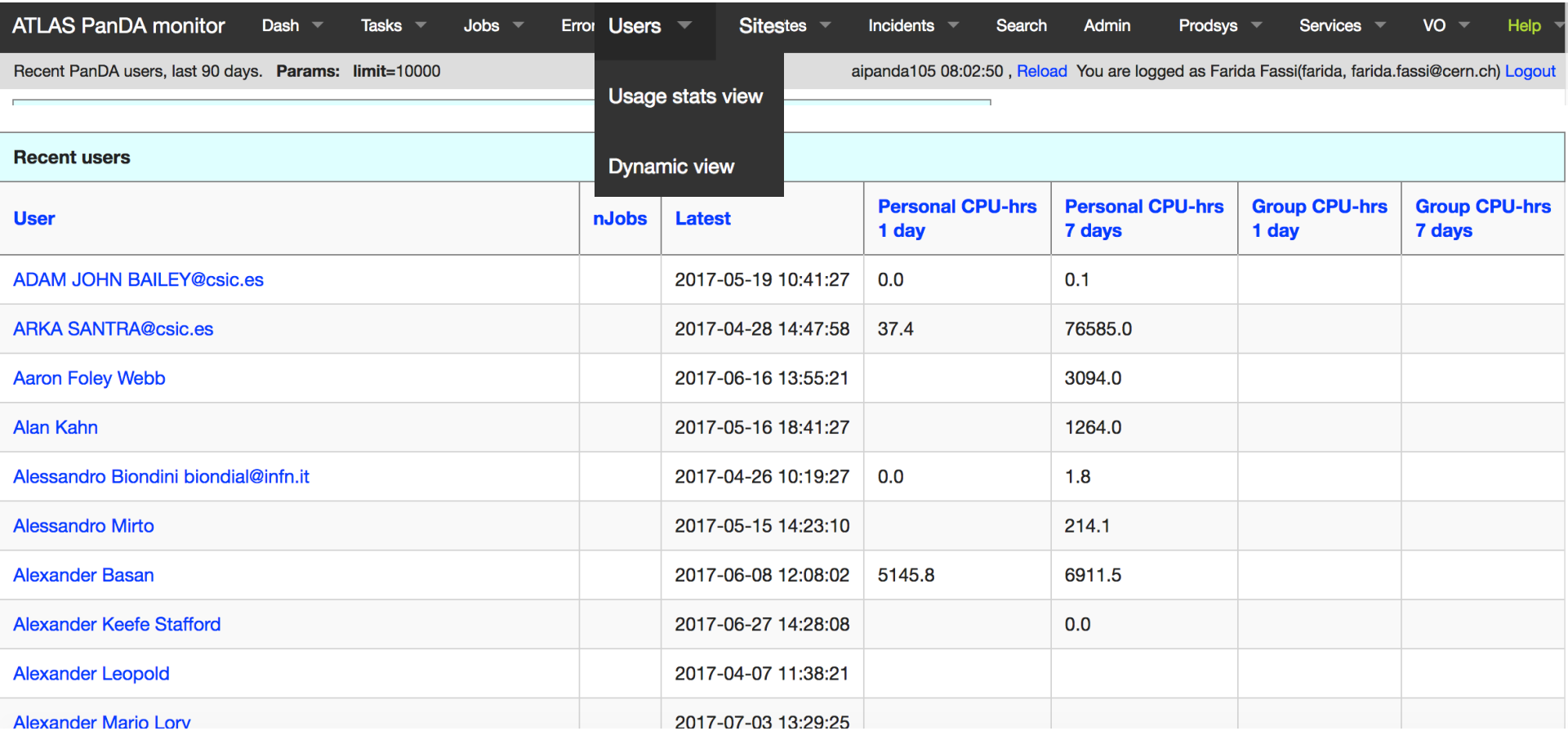
PanDA job ID or name		Submit
Batch ID		Submit
Task ID		Submit
Task name		Submit
Request ID		Submit

Search

News

- 20150318: Memory infomation added to jobs and tasks pages
- 20150316: RW metric added to dashboard
- 20150205: Responce time of tasks display improved
- 20150205: Wildcard search of jobs on job parameters added (ATLASPANDA-133)
- 20150205: Dataset information added to JSON responce (ATLASPANDA-109)
- 20150205: Wildcard search on jobs added (ATLASPANDA-40)
- 20141229: Main page plots show all jobs by cloud and activity
- 20141219: curl dumps of job params. See job list page help.
- 20141216: Request ID shown for jobs, range search added
- 20141215: Sort by duration option for job lists
- 20141211: Error code links in error summary changed to job lists
- 20141211: RequestID, requestID range added to search
- 20141210: Use transformation json metadata for error info

Tasks



User	nJobs	Latest	Personal CPU-hrs 1 day	Personal CPU-hrs 7 days	Group CPU-hrs 1 day	Group CPU-hrs 7 days
ADAM JOHN BAILEY@csic.es		2017-05-19 10:41:27	0.0	0.1		
ARKA SANTRA@csic.es		2017-04-28 14:47:58	37.4	76585.0		
Aaron Foley Webb		2017-06-16 13:55:21		3094.0		
Alan Kahn		2017-05-16 18:41:27		1264.0		
Alessandro Biondini biondial@infn.it		2017-04-26 10:19:27	0.0	1.8		
Alessandro Mirto		2017-05-15 14:23:10		214.1		
Alexander Basan		2017-06-08 12:08:02	5145.8	6911.5		
Alexander Keefe Stafford		2017-06-27 14:28:08		0.0		
Alexander Leopold		2017-04-07 11:38:21				
Alexander Mario Lory		2017-07-03 13:29:25				

task attribute summary

Task attribute summary, 16 tasks	
corecount (1)	1 (16)
eventservice (1)	ordinary (16)
gshare (1)	Analysis (16)
processingtype (1)	panda-client-0.5.84-jedi-athena (16)
ramcount (1)	2-3GB (16)
reqid (16)	3283 (1) 3286 (1) 3324 (1) 3342 (1) 3359 (1) 3388 (1) 3416 (1) 3433 (1) 3609 (1) 3610 (1) 3611 (1) 3612 (1) 3617 (1) 3618 (1) 3619 (1) 3621 (1)
status (3)	broken (3) done (7) finished (6)
status (3)	broken (3) done (7) finished (6)

tasks details

16 tasks, sorted by jeditaskid Show in task list page

ID Parent	Jobset	Task name TaskType/ProcessingType Campaign Group User Logged status	Task status Nfiles	Input files finish% fail% Nfinish Nfail	Modified	State changed	Priority
11551906	3621	user.mgrandi.data17_13TeV.00326468.physics_Main.merge.AOD.f829_m1812-20170627-135614/ anal/panda-client-0.5.84-jedi-athena mario grandi no scout jobs succeeded	broken 200	25% 50	2017-06-27 14:27:55	2017-06-27 14:27:55	1000
11551737	3619	user.mgrandi.data17_13TeV.00325789.physics_Main.merge.AOD.f827_m1807-20170627-135503/ anal/panda-client-0.5.84-jedi-athena mario grandi	finished 11	9% 90% 1 10	2017-06-28 09:37:38	2017-06-28 09:37:38	1000
11551571	3618	user.mgrandi.data17_13TeV.00326439.physics_Main.merge.AOD.f828_m1812-20170627-135349/ anal/panda-client-0.5.84-jedi-athena mario grandi	done 61	100% 61	2017-06-28 11:09:06	2017-06-28 11:09:06	1000
11551415	3617	user.mgrandi.data17_13TeV.00326446.physics_Main.merge.AOD.f828_m1812-20170627-135236/ anal/panda-client-0.5.84-jedi-athena mario grandi	finished 200	97% 2% 195 5	2017-06-28 13:03:06	2017-06-28 13:03:06	1000
11551158	3612	user.mgrandi.data17_13TeV.00325790.physics_Main.merge.AOD.f827_m1807-20170627-134842/ anal/panda-client-0.5.84-jedi-athena mario grandi no scout jobs succeeded	broken 200	25% 50	2017-06-27 13:52:19	2017-06-27 13:52:19	1000
11551057	3611	user.mgrandi.data17_13TeV.00326695.physics_Main.merge.AOD.f832_m1812-20170627-134658/ anal/panda-client-0.5.84-jedi-athena mario grandi	done 200	100% 200	2017-06-28 12:11:52	2017-06-28 12:11:52	1000

11590533 task: user.cwiel.tauIDSig_MC15C_Z_NOMINAL.301057.PowhegPythia8EvtGen_AZNLOCTEQ6L1_DYtautau_4500M5000_v0/

Task ID	Jobset	Type	WorkingGroup	User	Destination	Task status	Nevents | used	HS06*sec Expected Total done failed	Ninputfiles | finished | failed	Average maxpss	Created	Modified	Cores	Priority
11590533	20	analy		Christian Wiel	TOMERGE	failed	50000 | 0 (0%)	None 45894 45894 0	5 | 0 (0%) | 5 (100%)		2017-07-03 12:24:41	2017-07-03 14:51:24	1	1000

clickable

!! All links are here → Go to | Show jobs | Jump to | Open plot

States of jobs in this task (merge jobs excluded) Show all jobs Switch to nodrop mode

defined	waiting	pending	assigned	throttled	activated	sent	starting	running	holding	transferring	finished	failed	cancelled	merging	closed
											1	1			

link to jobs and their log files

Job list Sort by PandaID, time since last state change, ascending mod time, descending mod time, priority, attemptnr, ascending duration, descending duration

PanDA ID Attempt#	Owner Group	Request Task ID	Transformation	Status	Created	Time to start d:h:m:s	Duration d:h:m:s	Mod	Cloud Site	Priority	Maximum PSS	Job info
click 3483196851 Attempt 3	Christian Wiel	20 11590533	runGen-00-00-02	failed	2017-07-03 14:42:50	0:0:00:08	0:0:02:49	2017-07-03 14:50:58	IT ANALY_INFN-T1 online no active blacklisting rules defined	983	6.00	**trans, 220:** Proot: An exception occurred in the user analysis code

Job name: user.cwiel.tauIDSig_MC15C_Z_NOMINAL.301057.PowhegPythia8EvtGen_AZNLOCTEQ6L1_DYtautau_4500M5000_v0/.3482976847 #3

Datasets: **In:** mc15_13TeV:mc15_13TeV.301057.PowhegPythia8EvtGen_AZNLOCTEQ6L1_DYtautau_4500M5000.merge.AOD.e3649_s2576_s2132_r7772_r7676_tid08036356_00
Out: panda.um.user.cwiel.tauIDSig_MC15C_Z_NOMINAL.301057.PowhegPythia8EvtGen_AZNLOCTEQ6L1_DYtautau_4500M5000_v0_hist.144050801

Job information | logfiles | pilot job stdout, stderr, batch log | pilot records | Action logger | Action logger (es-atlas) | child jobs | Memory and IO plots

check athena_stdout.txt (for pathena)/prun_stdout.txt (for prun) and pilotlog.txt (all commands issued by the pilot) from this link

ATLAS PanDA monitor | Dash | Tasks | Jobs | Errors | Users | Sites | Incidents | Search | Admin | Prodsys | Services | VO | Help

PanDA sites

5 08:36:43 , Reload You are logged as Farida Fassi(farida, farida.fassi@cern.ch) Logout

Clouds, sites
CA
CERN
DE
ES
FR
IT
ND
NL
RU
TW
UK

cloud info

Jump to site attribute summary, site list

Clouds TLo, THi are transfer timeouts for low and high priority jobs

Cloud	Tier 1	Status	Comment	TL da	oints	Free space TB	Space updated
CA	TRIUMF	online	LFC.migration 11-29 10:56 Cedric	4/2	_DATADISK,TRIUMF-LCG2_DATATAPE,TRIUMF-E,TRIUMF-LCG2_HOTDISK,TRIUMF-,TRIUMF-LCG2_PERF-*	347	01-23 04:26
	MCP sites (home cloud sites in bold): **ANALY_AUSTRALIA ANALY_MCGILL ANALY_SCINET ANALY_SFU T2_MCORE CA-SCINET-T2 CA-SCINET-T2_MCORE CA-VICTORIA-WESTGRID-T2 CA-VICTORIA-WEST** LRZ-LMU_MUC1_MCORE RRC-KI_MCORE RRC-KI_PROD **SFU-LCG2 SFU-LCG2_ES SFU-LCG2_MCOR** UKI-LT2-RHUL_MCORE UKI-LT2-RHUL_SL6 UKI-SOUTHGRID-SUSX_MCORE UKI-SOUTHGRID-SUSX_SL6				**CTORIA Australia-ATLAS Australia-ATLAS_MCORE CA-MCGILL-CLUMEQ-T2 CA-MCGILL-CLUMEQ-** 2-HPC CSCS-LCG2-HPC_MCORE **IAAS IAAS_MCORE** IN2P3-CC_VVL LRZ-LMU_C2PAP_MCORE **M TRIUMF TRIUMF_HIMEM TRIUMF_MCORE TRIUMF_MCORE_LOMEM** Taiwan-LCG2-HPC		
CERN	CERN-PROD	online	ADCR.11g.upgrade.finished.elog.33058 11-29 10:58	2/2	DATADISK,CERN-PROD_MCTAPE,CERN-APE,CERN-PROD_HOTDISK,CERN-REP,CERN-PROD_DET-*,CERN-PROD_PERF-*,CERN-*,CERN-PROD_TRIG-*	418	01-22 10:00
	MCP sites (home cloud sites in bold): **ANALY_CERN_CLOUD ANALY_CERN_SLC6 ANALY_CERN_T0_SHO CERN-P1_DYNAMIC_SCORE CERN-PROD CERN-PROD-preprod CERN-PROD-preprod_MCORE CERN CERN-PROD_T0_4MCORE CERN-PROD_T0_SCORE_SHORT** CSCS-LCG2-HPC CSCS-LCG2-HPC_MCO RRC-KI_PROD SFU-LCG2 SFU-LCG2_MCORE Taiwan-LCG2-HPC UKI-LT2-RHUL_MCORE UKI-LT2-RHU				**NSION_HARVESTER CERN-P1 CERN-P1_DYNAMIC_MCORE CERN-P1_DYNAMIC_MCORE_LOWMEM D_CLOUD_MCORE CERN-PROD_HI CERN-PROD_PRESERVATION CERN-PROD_PUB_SCORE_SHORT** _C2PAP_MCORE LRZ-LMU_MUC1_MCORE ROMANIA14_MCORE ROMANIA16_MCORE RRC-KI_MCORE X_MCORE UKI-SOUTHGRID-SUSX_SL6 UNI-DORTMUND		

Sites (PanDA resources) (MCP sites are below)

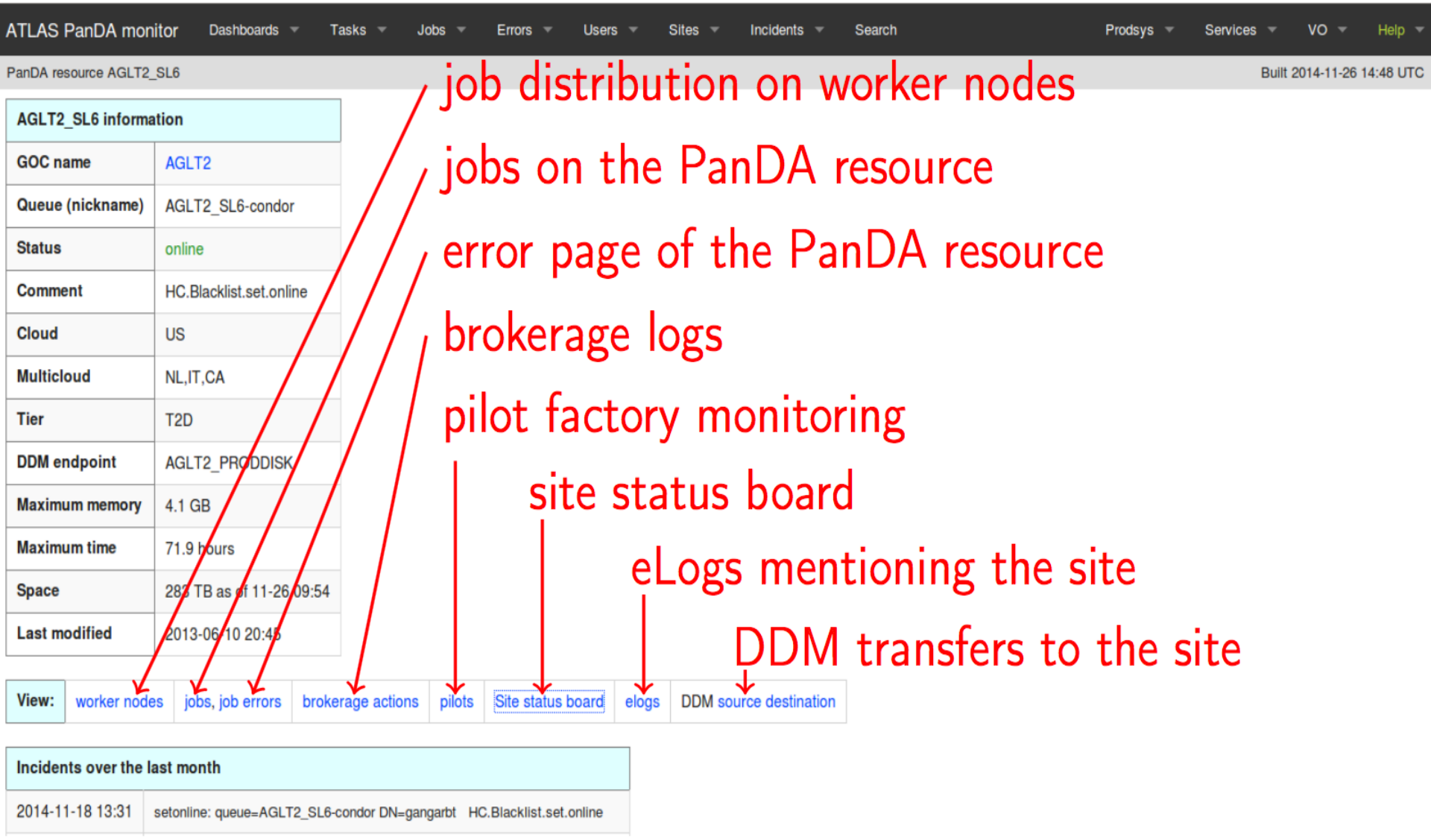
PanDA resource details

PanDA resource Queue name (where different)	GOC site name	Cloud	Status	Tier	Max mem (MB)	Max time (hr)	Comment
ANALY_CERN_CLOUD	CERN-PROD	CERN	online	T0	6000	48.0	no active blacklisting rules defined
ANALY_CERN_SLC6	CERN-PROD	CERN	online	T0	6000	72.0	no active blacklisting rules defined
ANALY_CERN_T0_SHORT	CERN-PROD	CERN	online	T0	6000	72.0	no active blacklisting rules defined
ANALY_CERN_TEST	CERN-PROD	CERN	brokeroff	T0	6000	72.0	Test API 2
BOINC-ES	BOINC	CERN	online	T3	48000	3.0	set online
BOINC-TEST	BOINC	CERN	brokeroff	T3	48000	0.5	Test queue
BOINC_CHECKPOINT	BOINC	CERN	brokeroff	T3	7500	1111.1	in test
BOINC_MCORE	BOINC	CERN	brokeroff	T3	7500	1111.1	Schedconfig initialization
CERN-EXTENSION_HARVESTER	CERN-PROD	CERN	online	T3D	20000	96.0	harvester at increased scale
CERN-EXTENSION_MCORE	CERN-PROD	CERN	test	T3D	20000	96.0	test changes
CERN-EXTENSION_TEST	CERN-PROD	CERN	brokeroff	T3D	48000	96.0	allow tobias to run tests
CERN-P1 CERN-P1-OpenStack	CERN-PROD	CERN	online	T3	6000	50.0	no active blacklisting rules defined
CERN-P1_DYNAMIC_MCORE	CERN-PROD	CERN	online	T3	36000	96.0	no active blacklisting rules defined

Site attribute summary	
allowdirectaccess (2)	False (23) True (4)
allowfax (2)	False (24) True (3)
category (4)	analysis (4) multicloud (13) production (21) test (3)
cloud (1)	CERN (27)
comment_field (10)	Real.Prod.start.only.preassigned.tasks (1) Schedconfig initialization (2) Test API 2 (1) Test queue (1) allow tobias to run tests (1) harvester at increased scale (1) in test (1) no active blacklisting rules defined (17) set online (1) test changes (1)
copytool (3)	lcgcp2 (2) mv (4) xrdcp (21)
faxredirector (2)	atlas-xrd-eu.cern.ch:1094 (3)
gocname (3)	BOINC (4) CERN-PROD (21) HELIX_NEBULA (2)
nickname (27)	ANALY_CERN_CLOUD (1) ANALY_CERN_SLC6 (1) ANALY_CERN_T0_SHORT (1) ANALY_CERN_TEST (1) BOINC-ES (1) BOINC-TEST (1) BOINC_CHECKPOINT (1) BOINC_MCORE (1) CERN-EXTENSION_HARVESTER (1) CERN-EXTENSION_MCORE (1) CERN-EXTENSION_TEST (1) CERN-P1-OpenStack (1) CERN-P1_DYNAMIC_MCORE (1) CERN-P1_DYNAMIC_MCORE_LOWMEM (1) CERN-P1_DYNAMIC_SCORE (1) CERN-PROD-all-prod-CEs (1) CERN-PROD-preprod (1) CERN-PROD-preprod_MCORE (1) CERN-PROD_CLOUD (1) CERN-PROD_CLOUD_MCORE (1) CERN-PROD_HI (1) CERN-PROD_PRESERVATION (1) CERN-PROD_PUB_SCORE_SHORT (1) CERN-PROD_T0_4MCORE (1) CERN-PROD_T0_SCORE_SHORT (1) HELIX_NEBULA_ATOS (1) HELIX_NEBULA_EGI (1)
region (2)	CERN (23) Nordugrid (4)
retry (1)	False (27)
status (3)	brokeroff (7) online (19) test (1)
tier (3)	T0 (14) T3 (10) T3D (3)
timefloor (3)	None (22) 0 (4) 60 (1)

site attribute summary

Site check


ATLAS PanDA monitor
Dashboards
Tasks
Jobs
Errors
Users
Sites
Incidents
Search
Prodsys
Services
VO
Help
PanDA resource AGLT2_SL6
Built 2014-11-26 14:48 UTC
AGLT2_SL6 information
GOC name
AGLT2
Queue (nickname)
AGLT2_SL6-condor
Status
online
Comment
HC.Blacklist.set.online
Cloud
US
Multicloud
NL,IT,CA
Tier
T2D
DDM endpoint
AGLT2_PRODDISK
Maximum memory
4.1 GB
Maximum time
71.9 hours
Space
283 TB as of 11-26 09:54
Last modified
2013-06-10 20:45
View:
worker nodes
jobs, job errors
brokerage actions
pilots
Site status board
elogs
DDM source destination
Incidents over the last month
2014-11-18 13:31
setonline: queue=AGLT2_SL6-condor DN=gangarbt HC.Blacklist.set.online
job distribution on worker nodes
jobs on the PanDA resource
error page of the PanDA resource
brokerage logs
pilot factory monitoring
site status board
eLogs mentioning the site
DDM transfers to the site

Distributed Analysis Support Team (DAST)

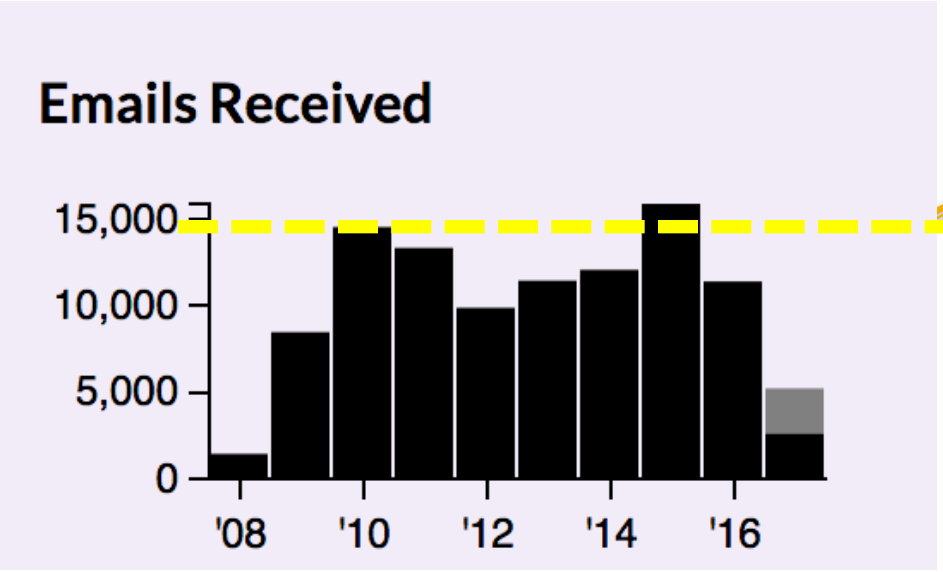
- DAST provides the first contact point to help thousant of Grid users.
  - DAST deals with all kind of the distributed analysis-related-issues.
  - an efficient user support is crucial to get physics results fast.

  - DAST plays a key role to solve these users-related-issues:
    - Panda-clients
    - ATLAS software, Physics Analysis Tools
    - Site service problems
    - DDM-clients, data access at sites and data replication
    - Monitoring system

- Two expert shifters on duty during working hours; one in the North American time zone and one in the European time zone, covering 16 hours/day.

DAST traffic

- **From 1,314 users, we have exchanged 123,813 emails.**
  - Since Oct. 2008 until 2017 more than 10,000 a year.
- DAST continues to be a very successful first-level contact for ATLAS users with Grid analysis issues.



Peaking at more than 15K received e-mails

More information

Where to get help on the Distributed Analysis grid tasks

hn-atlas-dist-analysis-help

- Tutorial
  - https://twiki.cern.ch/twiki/bin/view/AtlasComputing/SoftwareTutorialUsingTheGrid
- Distributed Analysis with Panda (with FAQ)
  - https://twiki.cern.ch/twiki/bin/view/PanDA/PandaTools
  - ***pathena examples***: how to run production transformations, TAG selection, on good run lists, event picking, etc
  - ***prun examples***: how to run CINT macro, C++ ROOT, python job, pyROOT script, skim RAW/AOD/ESD data, merge ROOT files, etc
- Find your jobs in the PanDA monitor
  - http://bigpanda.cern.ch
- PanDA JEDI analysis twiki page
  - https://twiki.cern.ch/twiki/bin/view/PanDA/PandaJediAnalysis