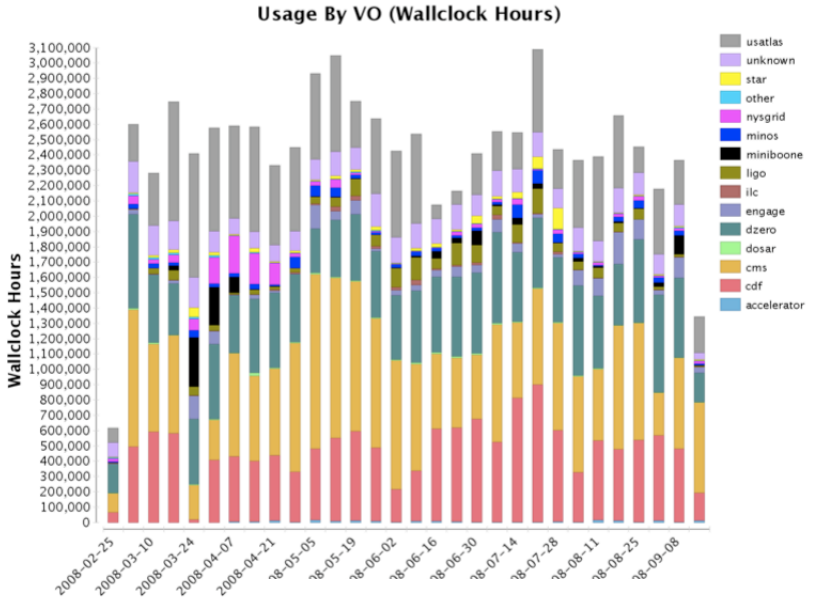
Panda, a Pilot-based workflow manager

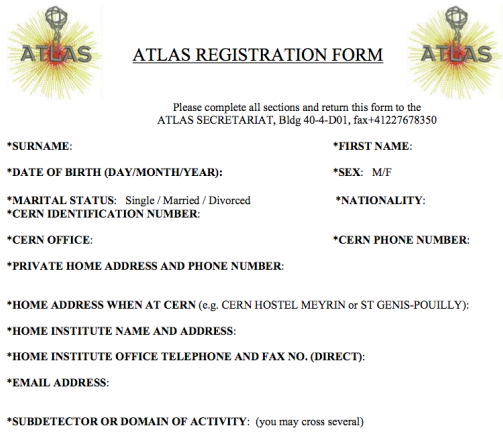
New Mexico Grid School – April 8, 2009

Marco Mambelli – University of Chicago
marco@hep.uchicago.edu

The ATLAS VO

- Virtual Organization in OSG (and other Grids)
  - In OSG since the beginning
  - https://twiki.grid.iu.edu/bin/view/VO/ATLAS
  - https://lcg-voms.cern.ch:8443/vo/atlas/vomrs
- Collaboration for the ATLAS experiment in the LHC at CERN
  - http://atlas.ch/
  - http://atlas.web.cern.ch/Atlas/ATLASreg_form.pdf





ATLAS REGISTRATION FORM

Please complete all sections and return this form to the ATLAS SECRETARIAT, Bldg 40-4-D01, fax+41227678350

***SURNAME**:

***FIRST NAME**:

***DATE OF BIRTH (DAY/MONTH/YEAR):**

***SEX**: M/F

***MARITAL STATUS**: Single / Married / Divorced

***NATIONALITY**:

***CERN IDENTIFICATION NUMBER**:

***CERN OFFICE**:

***CERN PHONE NUMBER**:

***PRIVATE HOME ADDRESS AND PHONE NUMBER**:

***HOME ADDRESS WHEN AT CERN** (e.g. CERN HOSTEL MEYRIN or ST GENIS-POUILLY):

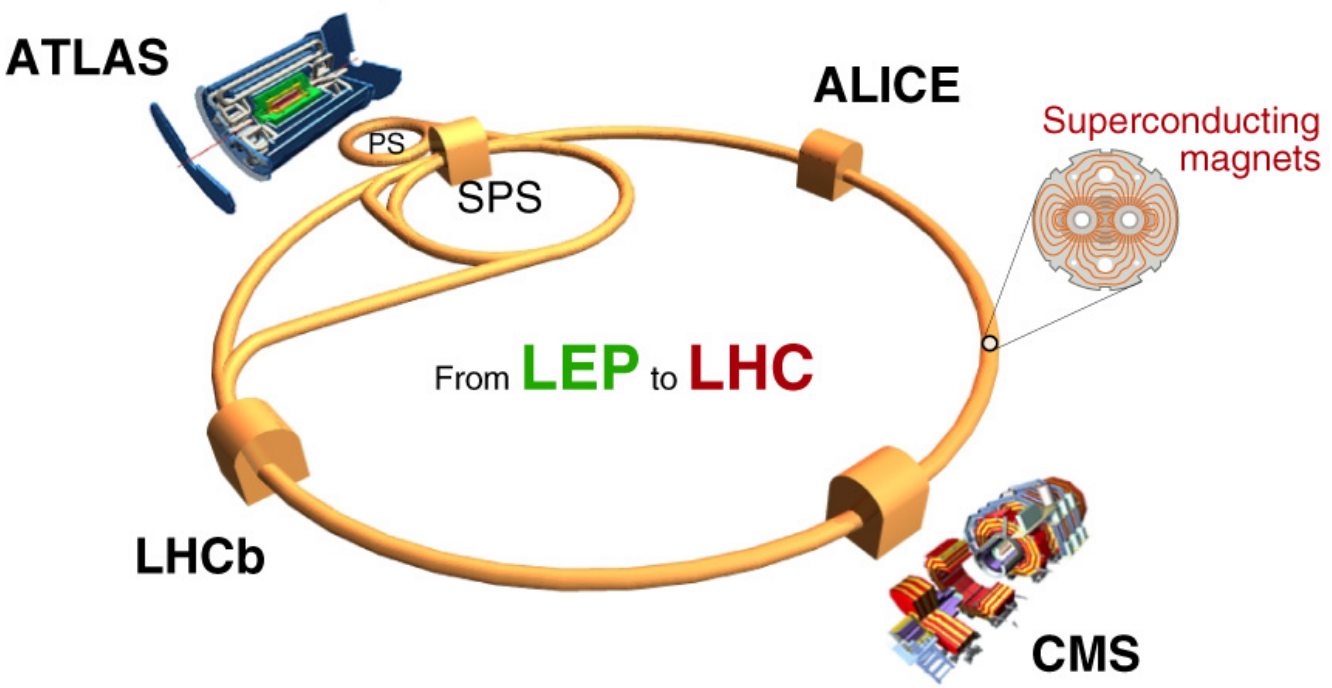
***HOME INSTITUTE NAME AND ADDRESS**:

***HOME INSTITUTE OFFICE TELEPHONE AND FAX NO. (DIRECT)**:

***EMAIL ADDRESS**:

***SUBDETECTOR OR DOMAIN OF ACTIVITY**: (you may cross several)

LHC experiment at CERN

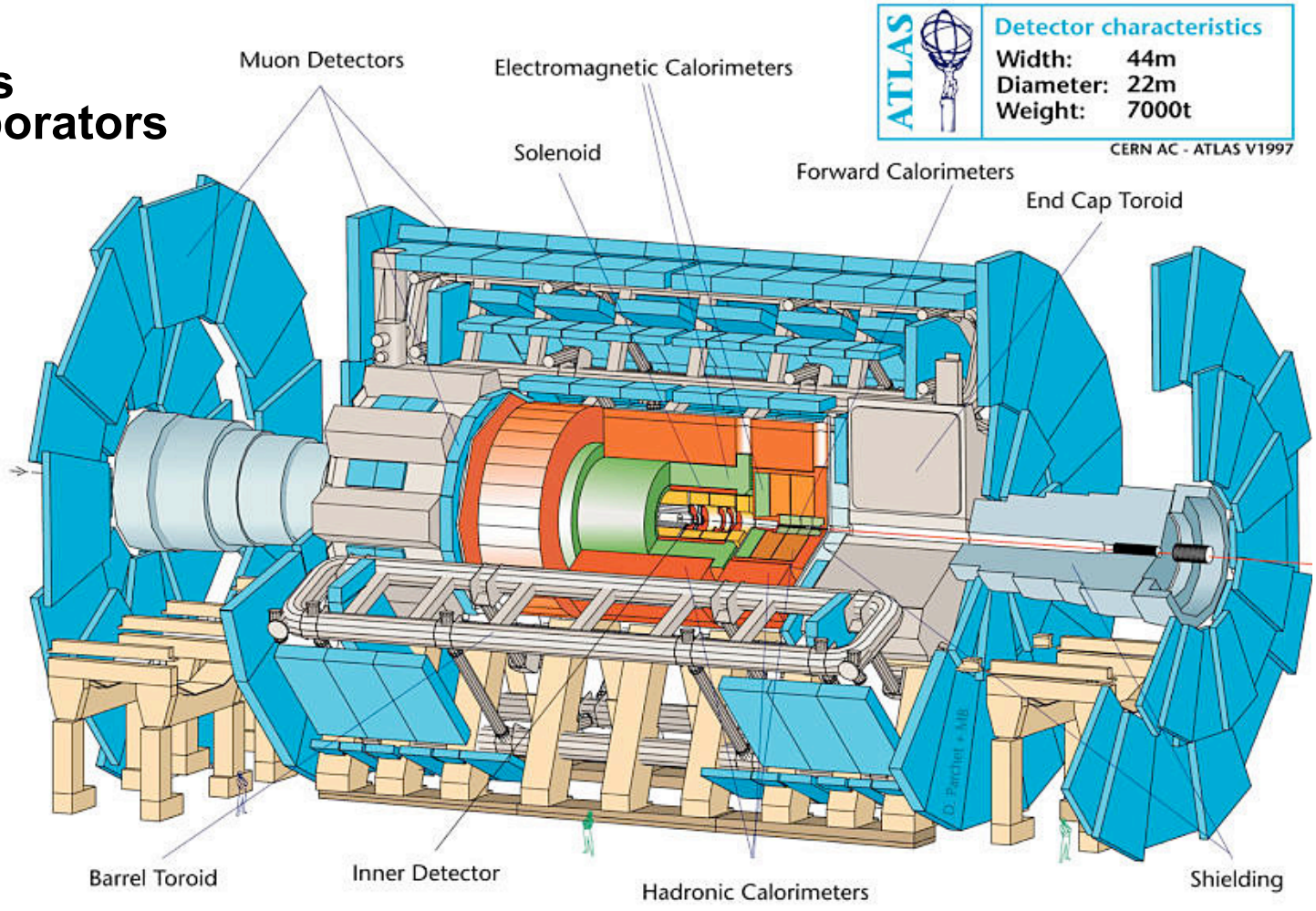


	Beams	Energy	Luminosity
LEP	e^+ e^-	200 GeV	10^{32} $cm^{-2}s^{-1}$
LHC	p p	14 TeV	10^{34}

http://www.youtube.com/watch?v=j50ZssEojtM

The ATLAS experiment

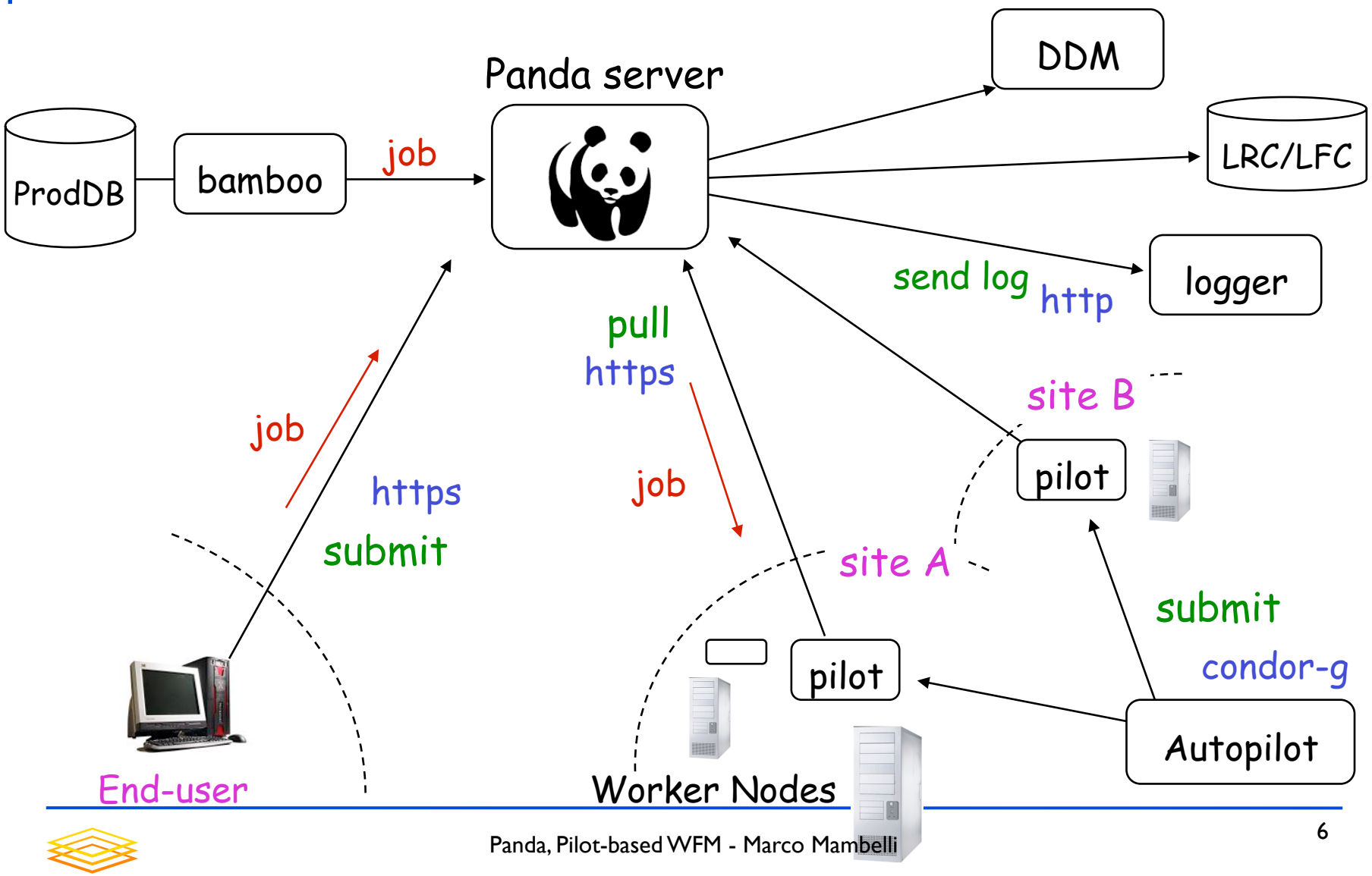
37 Countries
167 Institutes
~2000 Collaborators



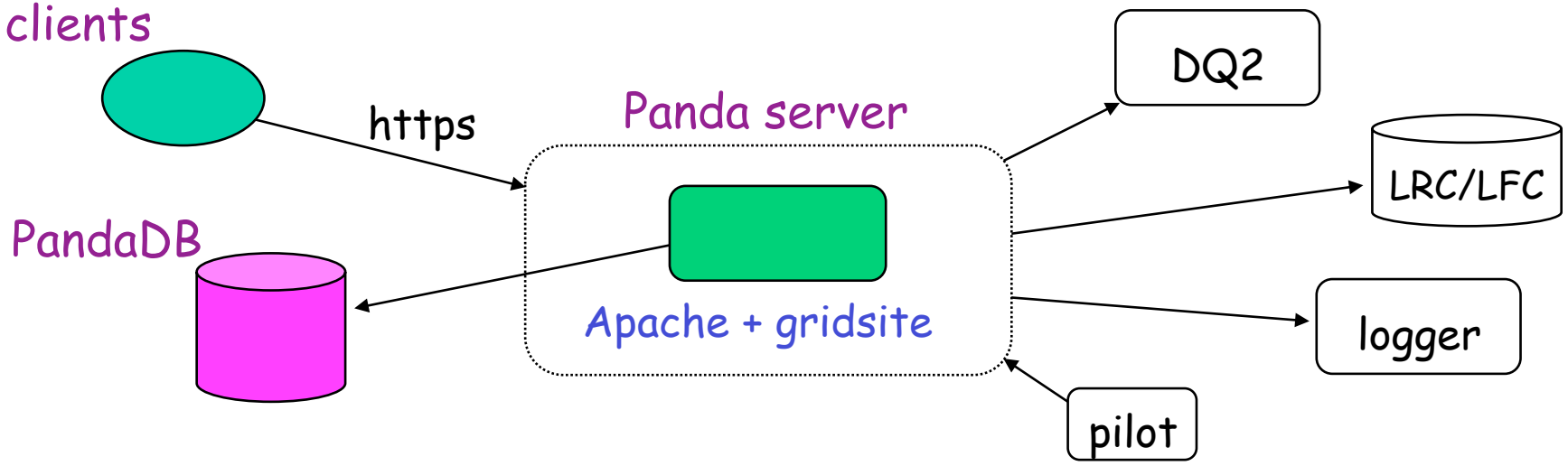
PANDA

- PANDA = Production ANd Distributed Analysis system
  - Designed for analysis as well as production for High Energy Physics
  - Works both with OSG and EGEE middleware
- A single task queue and pilots
  - Apache-based Central Server
  - Pilots retrieve jobs from the server as soon as CPU is available → late scheduling
- Highly automated, has an integrated monitoring system
- Integrated with ATLAS Distributed Data Management (DDM) system
- Not exclusively ATLAS: has its first OSG user in CHARMM (Chemistry at HARvard Molecular Mechanics)

Panda System

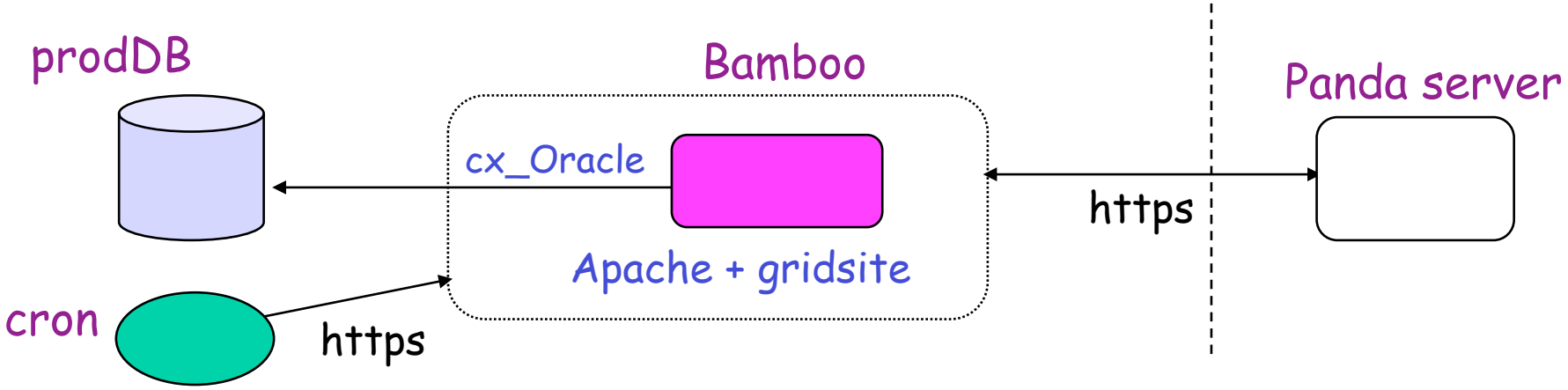


Panda Server



- Central queue for all kinds of jobs
- Assign jobs to sites (brokerage)
- Setup input/output datasets
  - Create them when jobs are submitted
  - Add files to output datasets when jobs are finished
- Dispatch jobs

Bamboo

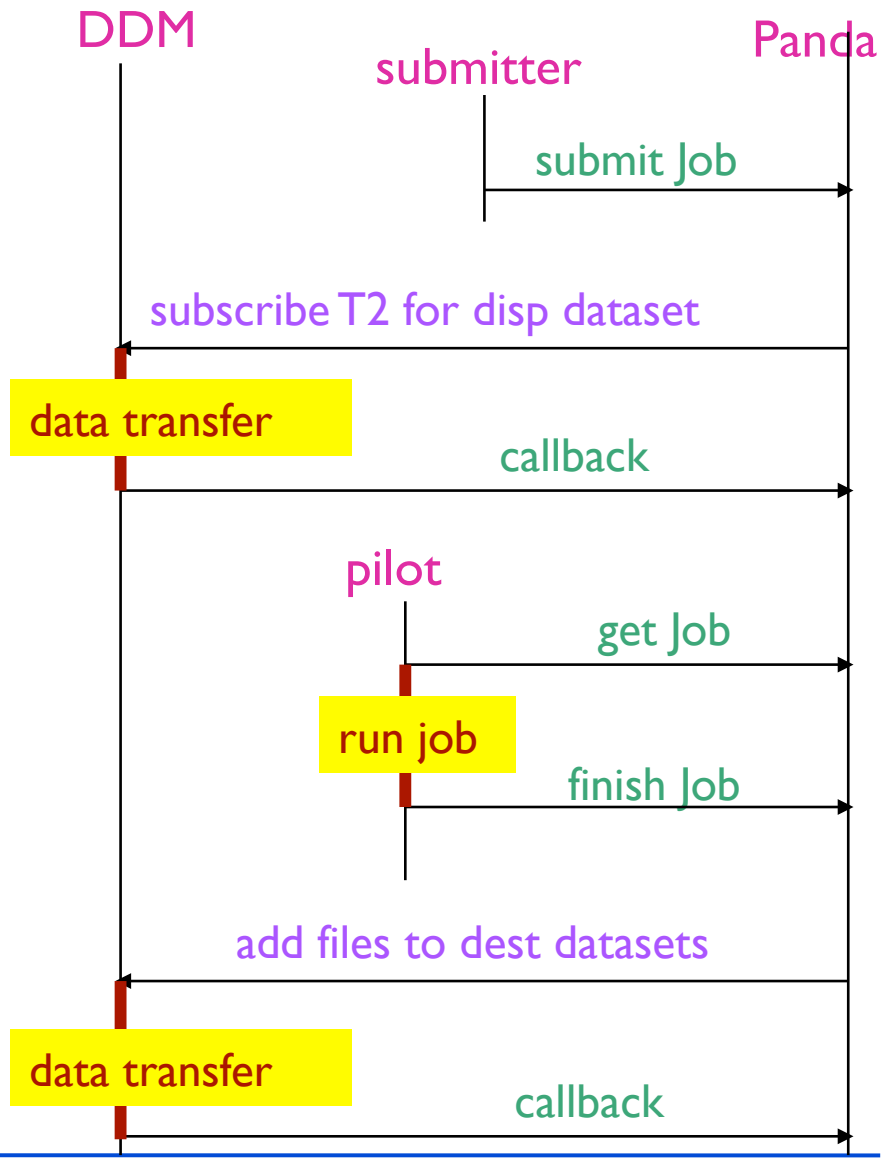


- Get jobs from prodDB to submit them to Panda
- Update job status in prodDB
- Assign tasks to clouds dynamically
- Kill TOBEABORTED jobs

- A cron triggers the above procedures every 10 min

Panda Job Timeline

- Rely on ATLAS DDM
  - Panda sends requests to DDM
  - DDM moves files and sends notifications back to Panda
  - Panda and DDM work asynchronously
- Dispatch input files to execution sites and aggregate output files to destination
- Jobs get 'activated' when all input files are copied, and pilots pick them up
  - Pilots don't have to transfer data (asynchronous)
  - Data-transfers and Job-executions can run in parallel



How the pilot works

- Sends the several parameters to Panda server for job matching (HTTP request)
  - CPU speed
  - Available memory size on the WN
  - List of available ATLAS releases at the site
- Retrieves an `activated' job (HTTP response of the above request)
  - activated → running
- Runs the job immediately because all input files should be already available at the site
- Sends heartbeat every 30min
- Copy output files to local Storage Element and register them to Local Replica Catalog

Pilot vs ATLAS Job

Pilot

- Submitted by factories
  - remote submit hosts
  - local cluster factories
- Managed by factories
- Python code to support ATLAS Job execution
- Submitted continuously
- Partially accounted
  - no big deal if some fail

ATLAS Job

- Submitted by users or production managers (Bamboo)
- Managed by Panda Server
- Runs Athena software (ATLAS libraries)
- Submitted when needed
- Fully accounted
  - error statistics are important

Some monitoring resources

- The following pages present some monitoring example
- Screenshots are just example pages, actual content varies
- URLs are one of the possible URLs providing a similar page
  - e.g. queries may vary the actual Site or Time interval
- Main URLs:
  - DDM Dashboard: http://dashb-atlas-data-test.cern.ch/dashboard/request.py/site
  - Panda Monitor: http://panda.cern.ch:25880/ or http://panda.atlascomp.org/?redirect=pandamon
    (hostname may change since there are multiple servers)
- Take time to navigate Panda Monitor and the Dashboard

Panda Monitor: production dashboard

http://panda.cern.ch:25880/server/pandamon/query?dash=prod

Panda monitor
Times are in UTC

Wiki Bugs
Shift elog CERN elog

Panda Production Operations Dashboard

Panda shift guide calendar mailing list
ADCoS twiki calendar mailing list
Production task support mailing list

Click for help

Jobs - search
Recent running, activated, waiting, assigned, defined, finished, failed jobs
Select analysis, prod, install, test jobs

Quick search
Job
Dataset
Task request
Task status
File

Summaries
Blocks: days
Errors: days
Nodes: days
Daily usage

Tasks - search
Generic Task Req
EvGen Task Req
CTBsim Task Req
Task list
New Tag
Bug Report

Datasets - search
Dataset browser
Aborted MC datasets
Panda subscriptions

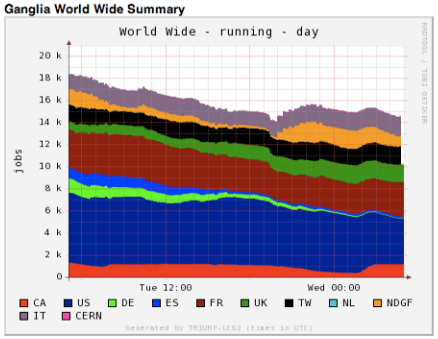
Datasets Distribution
DDM Req
Req list
AODs

Servers: BNL:OK BNLdev:OK CERN:OK Logger:OK Bamboo:OK

Active tasks: CA:14 CERN:1 DE:15 ES:4 FR:22 IT:12 NL:1 TW:15 UK:6 US:18
Bamboo task brokerage, job submissions, status over last 12 hours

Jobs updated >12 hrs ago: activated:5950 running:none
Jobs updated >36 hrs ago: transferring:12

Ganglia World Wide Summary



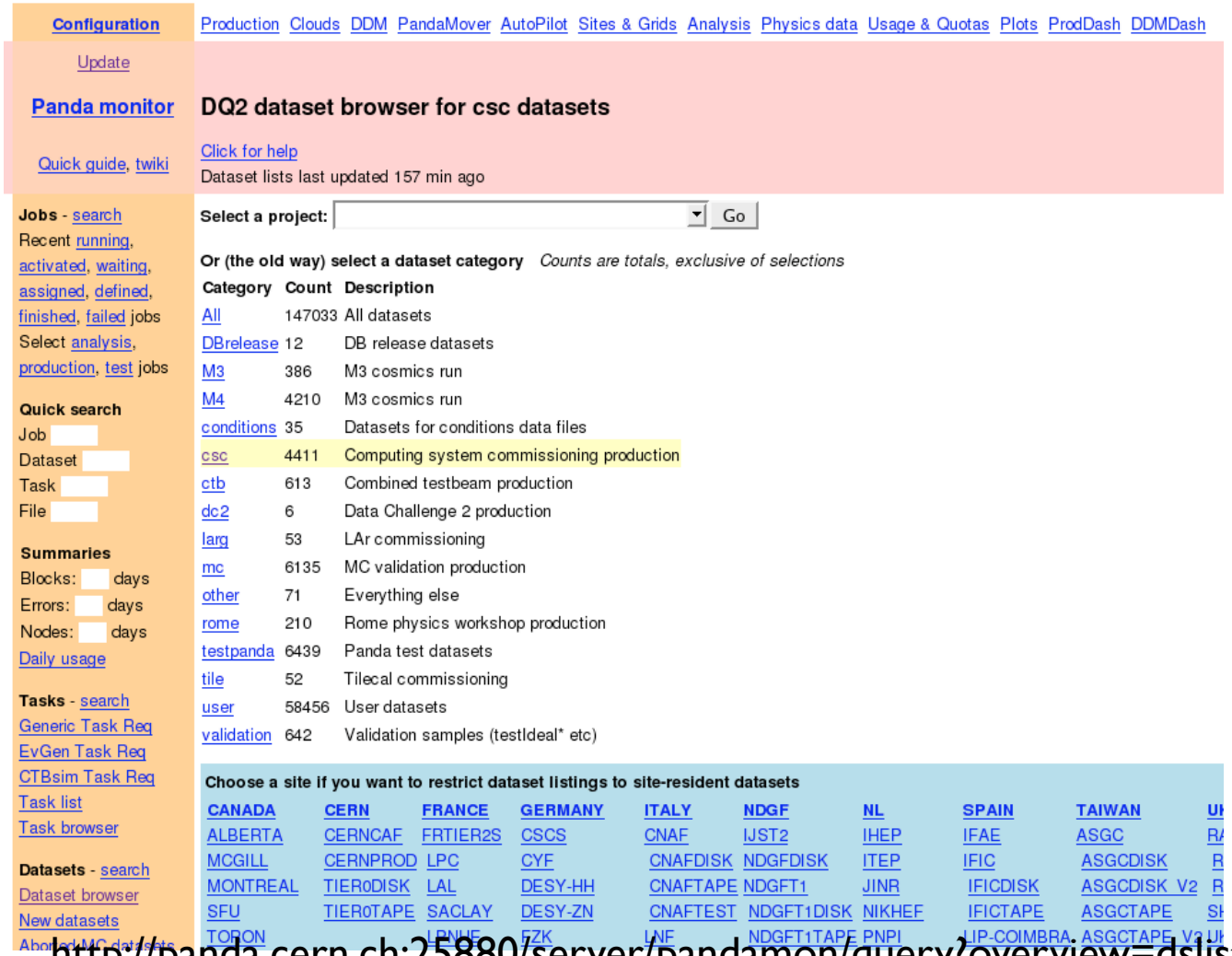
Space available at sites:

Site	GB	As of (UTC)
US		
AGLT2-condor	6921	09-17 02:19
ANALY_AGLT2-condor	6933	09-17 01:10
ANALY_MWT2-pbs	66815	09-17 04:27
ANALY_NET2-pbs	41342	09-17 04:21
ANALY_SLAC-lsf	16401	09-17 04:50
ANALY_SWT2_CPB-pbs	144839	09-17 01:08
BU_ATLAS_Tier2	41342	09-17 03:39
BU_ATLAS_Tier2o	41342	09-17 04:49
HU_ATLAS_Tier2	41342	09-17 02:40
IU_OSG-pbs	76176	09-17 04:49
MWT2_IU-pbs	76176	09-17 04:41
MWT2_UC-pbs		
OU_OCHEP_SWT2-		
SLACXRD-lsf		
SWT2_CPB-pbs		
UC_ATLAS_MWT2-		
Other SEs reporting		
ALBERTA-LCG2_PF		
FZK-LCG2_MCDISK		
IN2P3-CC_MCDISK		
INFN-T1_MCDISK		
NDGF-T1_MCDISK		
PIC_MCDISK		
RAL-LCG2_MCDISK		
SARA-MATRIX_MC		

IT	566	8372	09-17 04:50	255	0	15	0	3417	0	1776	50	443 / 0	2003	668	25% 0% 25%
NL	2	228	09-16 23:24	688	0	227	0	0	0	0	0	0 / 0	0	1	100% 0% 100%
UK	870	8439	09-17 04:50	804	0	0	0	2140	0	1599	46	1236 / 0	2905	513	15% 0% 15%
US	1526	13514	09-17 04:50	1275	0	5	0	152	0	4113	230	2526 / 12	6263	225	3% 0% 3%
TW	468	7643	09-17 04:50	241	0	10	0	3241	1	1624	29	267 / 0	2444	27	1% 0% 1%

US Sites	Nodes	Jobs	Latest	Pilots (3hrs)	defined	assigned	waiting	activated	sent	running	holding	transferring	finished	failed	tot	trf	other
Site Name	1526	13514	09-17 04:50	1275	0	5	0	152	0	4113	230	2526 / 12	6263	225	3%	0%	3%
AGLT2	114	2009	09-17 04:50	197	0	0	0	0	0	529	48	935 / 0	397	100	20%	0%	20%
BNL_ATLAS_1	318	2580	09-17 04:50	184	0	0	0	1	0	914	46	0 / 0	1605	14	1%	0%	1%
BNL_ATLAS_2	0	0		0	0	0	0	0	0	0	0	0 / 0	0	0			
BU_ATLAS_Tier2	12	44	09-17 04:46	0	0	0	0	13	0	22	6	2 / 0	0	1	100%	0%	100%
BU_ATLAS_Tier2o	85	845	09-17 04:50	85	0	0	0	0	0	299	5	88 / 0	429	24	5%	5%	0%
GLOW-ATLAS	47	160	09-17 04:40	0	0	0	0	73	0	1	7	76 / 0	2	1	33%	0%	33%
HU_ATLAS_Tier2	24	74	09-17 04:44	83	0	5	0	0	0	0	0	0 / 0	0	69	100%	0%	100%
IU_OSG	17	232	09-17 04:50	19	0	0	0	6	0	124	0	41 / 0	60	1	2%	0%	2%
LTU_CCT	0	0		0	0	0	0	0	0	0	0	0 / 0	0	0			
MWT2_IU	118	1569	09-17 04:50	283	0	0	0	0	0	608	32	166 / 0	762	1	0%	0%	0%
MWT2_UC	184	2916	09-17 04:50	14	0	0	0	21	0	761	20	567 / 0	1545	2	0%	0%	0%
OU_OCHEP_SWT2	60	455	09-17 04:50	102	0	0	0	0	0	161	6	69 / 0	219	0	0%	0%	0%

Panda Monitor: Dataset browser



Configuration | Production | Clouds | DDM | PandaMover | AutoPilot | Sites & Grids | Analysis | Physics data | Usage & Quotas | Plots | ProdDash | DDMDash

Update

Panda monitor

Quick guide, twiki

DQ2 dataset browser for csc datasets

Click for help

Dataset lists last updated 157 min ago

Jobs - search
Recent running, activated, waiting, assigned, defined, finished, failed jobs
Select analysis, production, test jobs

Quick search
Job
Dataset
Task
File

Summaries
Blocks: days
Errors: days
Nodes: days
Daily usage

Tasks - search
Generic Task Req
EvGen Task Req
CTBsim Task Req
Task list
Task browser

Datasets - search
Dataset browser
New datasets

Select a project: Go

Or (the old way) select a dataset category *Counts are totals, exclusive of selections*

Category	Count	Description
All	147033	All datasets
DBrelease	12	DB release datasets
M3	386	M3 cosmics run
M4	4210	M3 cosmics run
conditions	35	Datasets for conditions data files
csc	4411	Computing system commissioning production
ctb	613	Combined testbeam production
dc2	6	Data Challenge 2 production
larg	53	LAr commissioning
mc	6135	MC validation production
other	71	Everything else
rome	210	Rome physics workshop production
testpanda	6439	Panda test datasets
tile	52	Tilecal commissioning
user	58456	User datasets
validation	642	Validation samples (testIdeal* etc)

Choose a site if you want to restrict dataset listings to site-resident datasets

CANADA	CERN	FRANCE	GERMANY	ITALY	NDGF	NL	SPAIN	TAIWAN
ALBERTA	CERNCAF	FRTIER2S	CSCS	CNAF	IJST2	IHEP	IFAE	ASGC
MCGILL	CERNPROD	LPC	CYF	CNAFDISK	NDGFDISK	ITEP	IFIC	ASGCDISK
MONTREAL	TIER0DISK	LAL	DESY-HH	CNAFTAPE	NDGFT1	JINR	IFICDISK	ASGCDISK_V2
SFU	TIER0TAPE	SACLAY	DESY-ZN	CNAFTEST	NDGFT1DISK	NIKHEF	IFICTAPE	ASGCTAPE
TORON		LPNHE	FZK	LNF	NDGFT1TAPE	PNPI	LIP-COIMBRA	ASGCTAPE_V2

http://panda.cern.ch:25880/server/pandamon/query?overview=dslist

Panda Monitor: error reporting

Panda monitor
Times are in UTC

Wiki Bugs
Shift elog CERN elog

Jobs - search
Recent running, activated, waiting, assigned, defined, finished, failed jobs
Select analysis, prod, install, test jobs

Quick search
Job
Dataset
Task request
Task status
File

Summaries
Blocks: days
Errors: days
Nodes: days
Daily usage

Tasks - search
Generic Task Req
EvGen Task Req
CTBsim Task Req
Task list
New Tag
Bug Report

Datasets - search
Dataset browser
Aborted MC datasets
Panda subscriptions

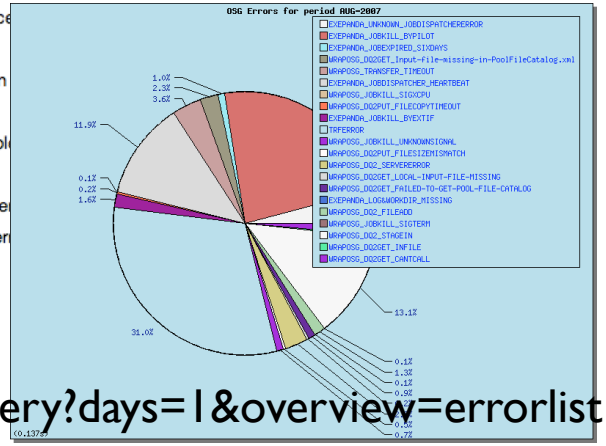
Datasets Distribution
DDM Req
Req list

Panda job error summary for last 24 hours (1.0 days)

All CEs and jobs. Show production, analysis, test, **all** jobs/CEs

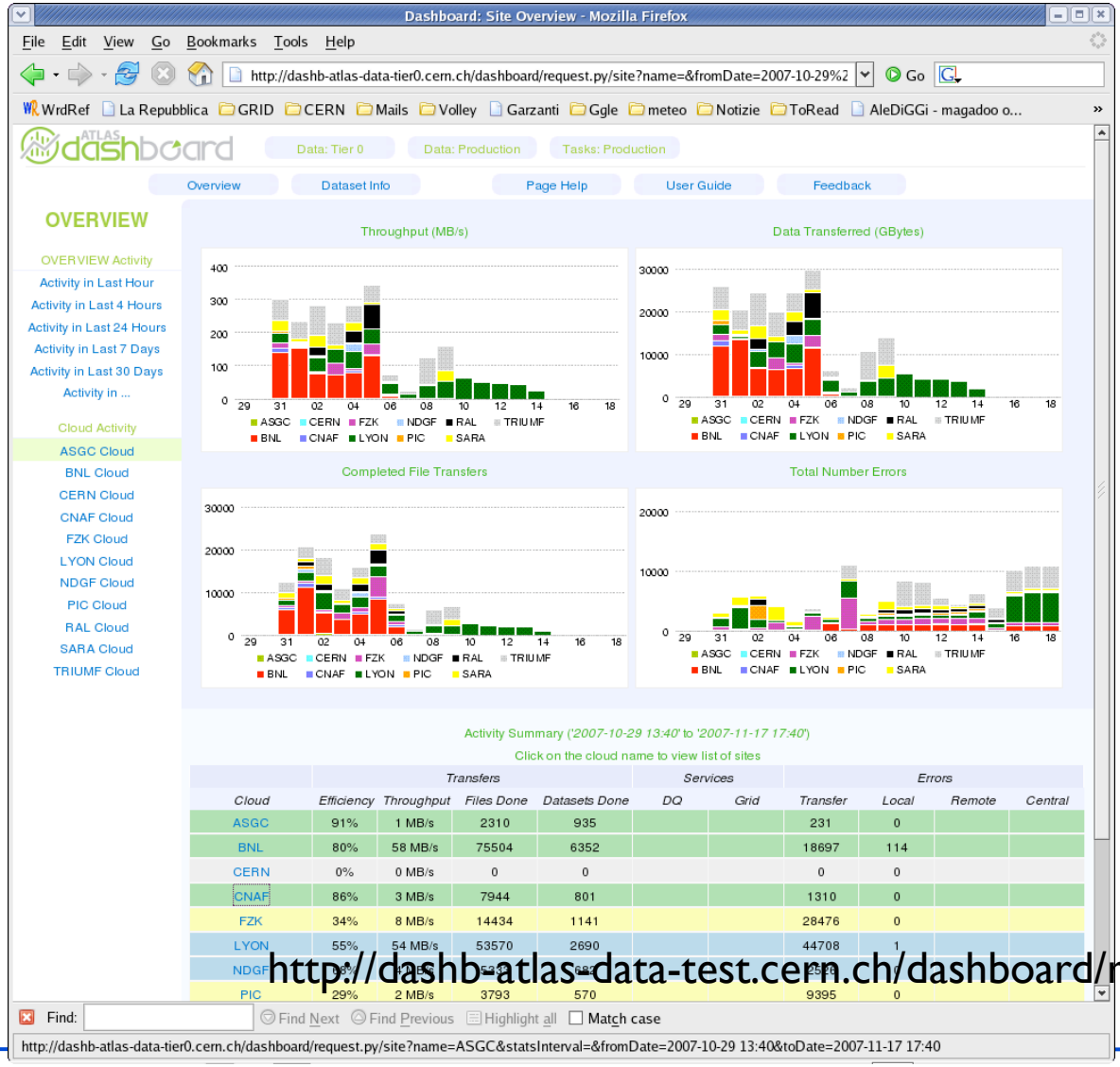
Job wall time: 317553 hrs Error losses: trans: 9971 (3.1%) panda: 8458 (2.7%) ddm: 3329 (1.0%) other: 1317 (0.4%)

Error type (type count)	Count	CPU-hrs	Latest	Code:	Description
All	defined:708 failed:8358 (17.1%)	assigned:264	waiting:0	activated:19020	sent:0 running:10202 holding:1599 transferring:5359 finished:40421
brokerageErrorCode (120)	120	0.0	09-17 13:11	100:	*Unknown error code*
ddmErrorCode (6)	1	0.0	09-16 18:14	100:	DQ2 server error
ddmErrorCode (6)	5	14.0	09-17 13:54	200:	Could not add output files to dataset
exeErrorCode (1114)	2	2.6	09-16 13:45	1101:	LRC registration error: Connection refused
exeErrorCode (1114)	1	0.9	09-16 20:13	1114:	Put error: Failed to import LFC python module
exeErrorCode (1114)	4	30.3	09-16 18:57	1131:	Put function can not be called for staging out
exeErrorCode (1114)	31	13.6	09-17 04:50	1132:	LRC registration error (consult log file)
exeErrorCode (1114)	7	14.8	09-17 10:25	1133:	Put error: Fetching default storage URL failed
exeErrorCode (1114)	1	26.2	09-15 10:22	1135:	Could not get file size in job workdir
exeErrorCode (1114)	875	7494.9	09-16 22:32	1137:	Put error: Error in copying the file from job workdir to localSE
exeErrorCode (1114)	13	159.2	09-16 15:34	1154:	Failed to register log file
exeErrorCode (1114)	6	58.6	09-16 15:20	1155:	Failed to move output files for lost job
exeErrorCode (1114)	1	11.8	09-14 15:01	1176:	Pilot has no child proce
exeErrorCode (1114)	1	22.1	09-15 07:10	1211:	Missing installation
exeErrorCode (1114)	3	51.7	09-17 13:52	60000:	segmentation violation
exeErrorCode (1114)	117	399.1	09-17 13:44	60010:	segmentation fault
exeErrorCode (1114)	5	92.2	09-17 10:47	61200:	ServiceManager Unabl
exeErrorCode (1114)	6	107.3	09-17 13:36	62600:	AthenaCrash
exeErrorCode (1114)	30	94.8	09-17 10:27	64100:	Transform output file e
exeErrorCode (1114)	11	52.2	09-17 12:15	69999:	Unknown Transform er



http://panda.cern.ch:25880/server/pandamon/query?days=1&overview=errorlist

DDM Dashboard: overview



http://dashb-atlas-data-test.cern.ch/dashboard/request.py/site

?

!

Client-Server Communication

- HTTP/S-based communication (curl+grid proxy+python)
- GSI authentication via mod_gridsite
- Most of communications are asynchronous
  - Panda server runs python threads as soon as it receives HTTP requests, and then sends responses back immediately. Threads do heavy procedures (e.g., DB access) in background → better throughput
  - Some are synchronous