Assessment of Analysis Failures, DAST Perspectives

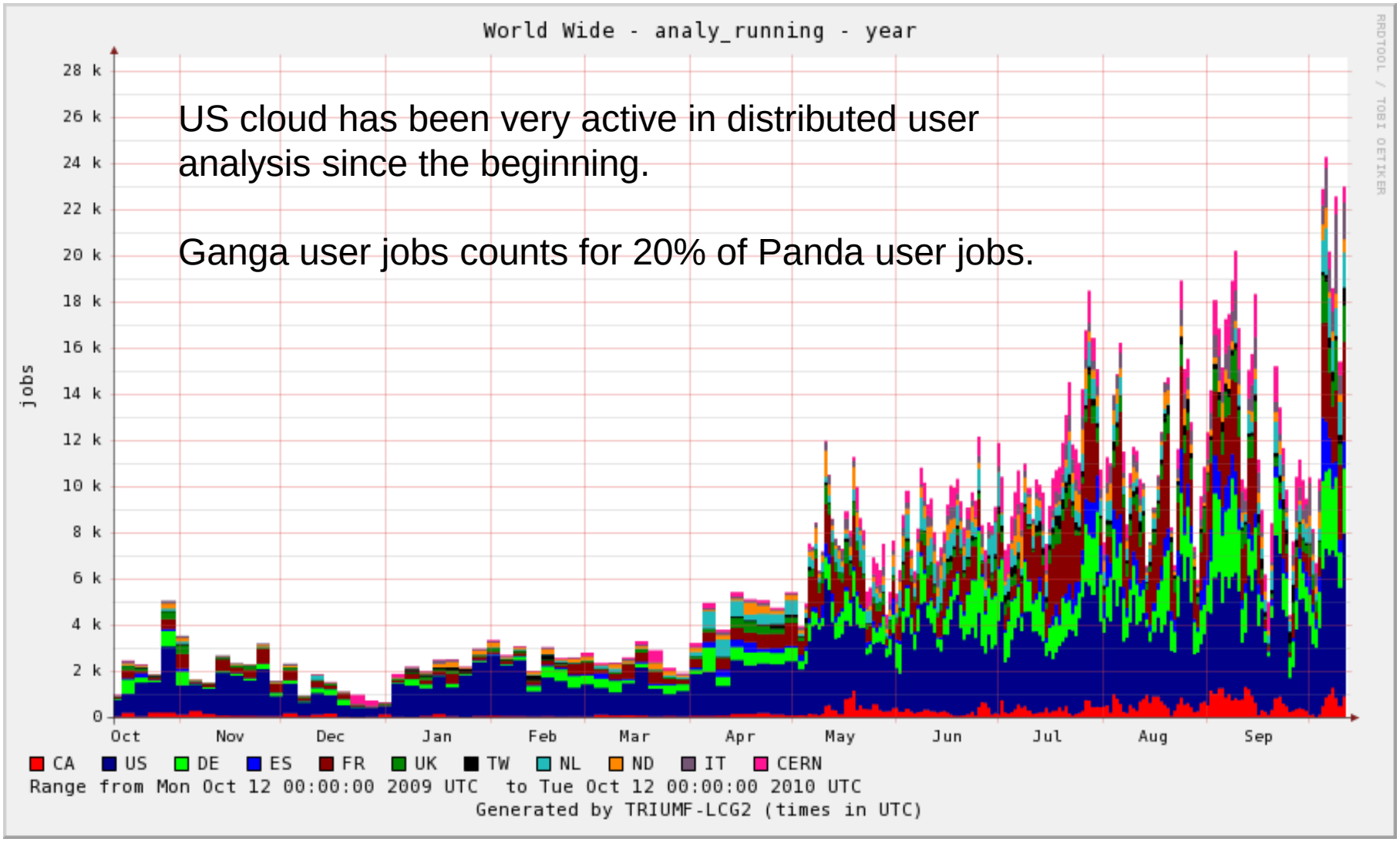
Nurcan Ozturk

University of Texas at Arlington

US ATLAS Distributed Facility Workshop at SLAC

12-13 October 2010

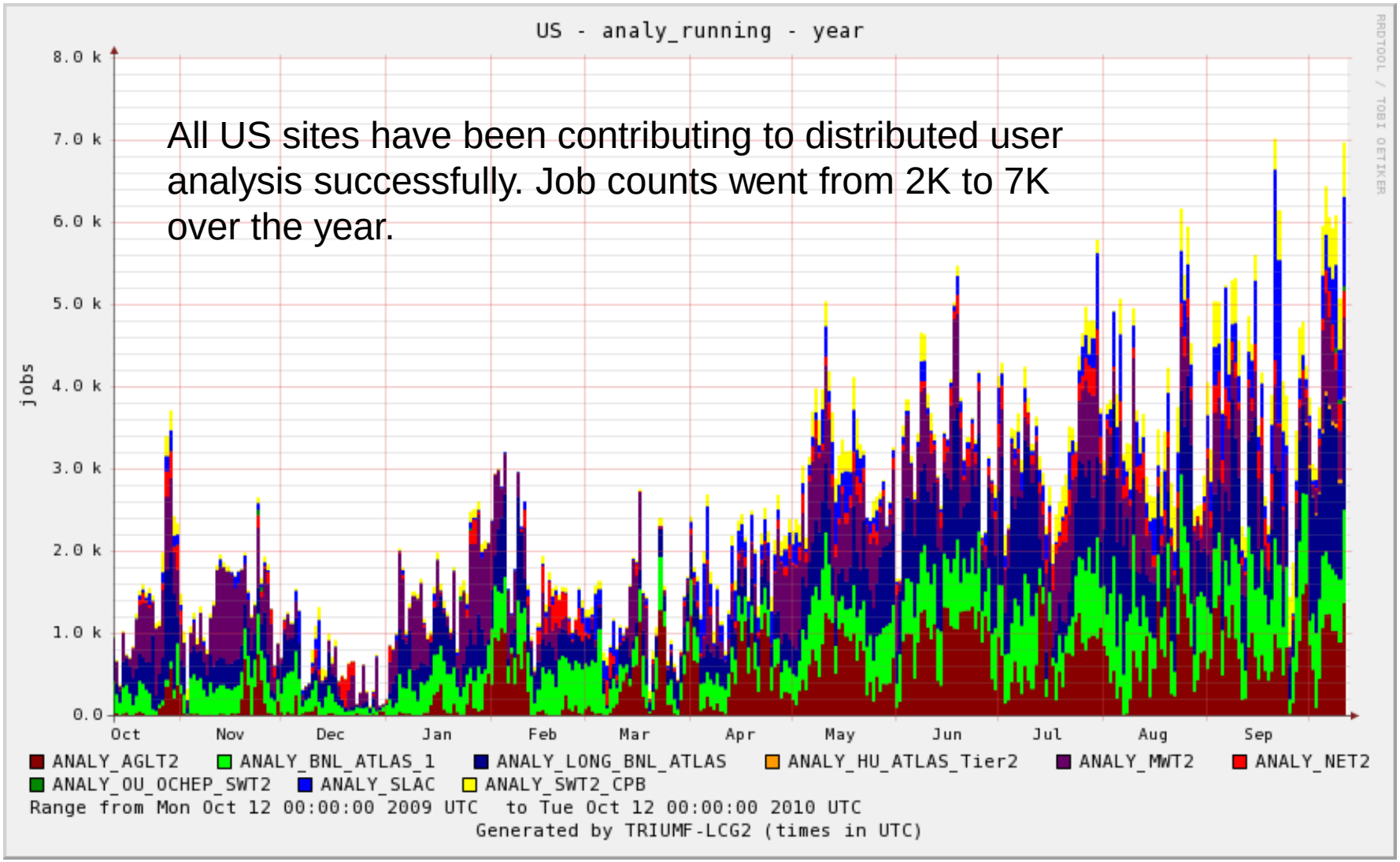
Distributed Analysis Panda User Jobs in All Clouds



US cloud has been very active in distributed user analysis since the beginning.

Ganga user jobs counts for 20% of Panda user jobs.

Distributed Analysis Panda User Jobs in US Cloud



Analysis job summary in all clouds, last 12 hours (@12.50pm CERN time)

Cloud	Pilots	Latest	defined	assigned	waiting	activated	sent	running	holding	transferring	finished	failed	cancelled	%fail
ALL			30989	0	0	63402	4	19407	2107	13	157435	26867	28939	15%
CA	532	10-12 10:20	23	0	0	6645	1	793	61	0	2685	402	11	13%
CERN (brokeroff)	1198	10-12 10:20	68	0	0	1367	0	998	143	0	4027	288	9	7%
DE	3547	10-12 10:20	7400	0	0	9446	0	2976	543	0	28783	2924	2545	9%
ES	831	10-12 10:20	1186	0	0	3905	0	1192	273	0	9625	1975	73	17%
FR	4883	10-12 10:20	4306	0	0	9805	1	5415	184	0	16496	3453	2990	17%
IT	1123	10-12 10:20	3643	0	0	985	0	1235	47	0	6593	1689	5703	20%
ND	194	10-12 10:20	6452	0	0	0	1	151	144	13	3946	1633	49	29%
NL	1304	10-12 10:20	623	0	0	8067	0	1064	83	0	6525	5345	1945	45%
TW	344	10-12 10:20	3	0	0	5849	0	99	3	0	8599	489	2486	5%
UK	1556	10-12 10:20	1740	0	0	2930	0	1370	231	0	8177	1116	1536	12%
US	2602	10-12 10:20	5545	0	0	14403	1	4114	395	0	61979	7553	11592	11%

Analysis job error report, last 12 hours
Job wall time: 286395 hrs Error losses: trans: 12282 (4.3%) panda: 24156 (8.4%) ddm: 8623 (3.0%) athena: 24464 (8.5%) user: 202 (0.1%) other: 8338 (2.9%)

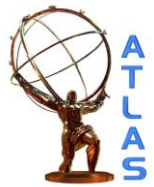
US cloud has been running the most jobs among all clouds. A stable cloud with low job failure rate. Note: BNL queues are being drained for the dCache intervention scheduled for today.

Analysis job summary in US cloud, last 12 hours (@12.50pm CERN time)

US sites	Pilots	Latest	defined	assigned	waiting	activated	sent	running	holding	transferring	finished	failed	cancelled	%fail
ANALY_AGLT2	520	10-12 10:20	867	0	0	2995	0	1749	94	0	7917	11	0	0%
ANALY_ANLASC (offline)			0	0	0	0	0	0	0	0	0	0	0	
ANALY_BNL_ATLAS_1	2	10-12 10:20	1459	0	0	30	0	0	1	0	7152	4950	2151	41%
ANALY_BNL_LOCAL	2	10-12 09:00	0	0	0	0	0	0	0	0	0	0	0	
ANALY_DUKE (brokeroff)	3	10-12 10:20	0	0	0	0	0	0	0	0	0	0	0	
ANALY_GLOW-ATLAS (offline)			0	0	0	0	0	0	0	0	0	0	0	
ANALY_HU_ATLAS_Tier2	40	10-12 10:20	0	0	0	851	0	50	0	0	52	203	0	80%
ANALY_IllinoisHEP	26	10-12 10:20	0	0	0	117	0	43	4	0	279	0	500	0%
ANALY_LONG_BNL_ATLAS	638	10-12 10:20	1326	0	0	1530	0	752	9	0	9781	450	316	4%
ANALY_LONG_BNL_LOCAL	2	10-12 09:00	0	0	0	0	0	0	0	0	0	0	0	
ANALY_MWT2	430	10-12 10:20	506	0	0	3637	0	622	80	0	19782	607	367	3%
ANALY_MWT2_X (brokeroff)	17	10-12 10:20	0	0	0	0	0	0	0	0	0	0	0	
ANALY_NET2	190	10-12 10:20	0	0	0	2035	0	219	43	0	2733	113	2	4%
ANALY_OU_OCHEP_SWT2	47	10-12 10:20	0	0	0	1055	0	44	5	0	274	0	0	0%
ANALY_SLAC (offline)	289	10-12 10:20	1086	0	0	1428	0	1	0	0	7159	463	4	6%
ANALY_SLAC_LMEM (offline)	4	10-12 08:30	0	0	0	0	0	0	0	0	0	0	0	
ANALY_SWT2_CPB	363	10-12 10:20	300	0	0	725	1	634	159	0	6850	756	8252	10%
ANALY_Tufts_ATLAS_Tier3 (brokeroff)	29	10-12 10:20	0	0	0	0	0	0	0	0	0	0	0	
ANALY_UTA_T3 (brokeroff)			0	0	0	0	0	0	0	0	0	0	0	
ANALY_WISC-ATLAS (offline)			0	0	0	0	0	0	0	0	0	0	0	

BNL queues are being drained for the dCache intervention scheduled for today.

Quick Look at BNL and HU Site Failures



ANALY_BNL_ATLAS_1	defined:1459	assigned:0	waiting:0	activated:30	sent:0	running:0	holding:1	transferring:0	finished:7152	failed:4950	(40.9%)
pilotErrorCode (4049)	2	1.9	10-12 02:02	1099:	Get error: Staging input file failed						
pilotErrorCode (4049)	1	1.7	10-12 00:05	1104:	User work directory too large						
pilotErrorCode (4049)	3969	1642.0	10-12 06:00	1165:	Put error: Local output file missing	<-- User code problem					
pilotErrorCode (4049)	77	115.8	10-12 00:56	1169:	Put error: LFC registration failed						
transExitCode (901)	1	0.2	10-12 02:06	20:	*Unknown error code*						
transExitCode (901)	900	73.3	10-12 05:50	40:	Athena crash - consult log file	<-- User code problem					

ANALY_HU_ATLAS_Tier2	defined:0	assigned:0	waiting:0	activated:851	sent:0	running:50	holding:0	transferring:0	finished:52	failed:198	(79.2%)
pilotErrorCode (31)	27	33.8	10-11 23:32	1099:	Get error: Staging input file failed						
pilotErrorCode (31)	4	9.2	10-11 23:30	1151:	Get error: Input file staging timed out						
transExitCode (167)	167	100.9	10-12 01:04	40:	Athena crash - consult log file	<-- User code problem					

Closer Look to Analysis Job Error Report (1)

Analysis job error report, last 12 hours
Job wall time: 279524 hrs Error losses: trans: 11597 (4.1%) panda: 24176 (8.6%) ddm: 9039 (3.2%) athena: 23437 (8.4%) user: 239 (0.1%) other: 8188 (2.9%)

- trans:
  - Unspecified error, consult log file
  - TRF_SEGVIO - Segmentation violation
- panda:
  - Put error: Error in copying the file from job workdir to localSE
  - Lost heartbeat
  - No space left on local disk
  - Get error: Staging input file failed
  - Put error: Failed to add file size and checksum to LFC
  - Payload stdout file too big
  - Exception caught by pilot
  - Put error: Failed to import LFC python module
  - Bad replica entry returned by lfc_getreplicas(): SFN not set in LFC for this guid
  - wget command failed to download trf
  - Missing installation
  - etc

"panda' errors (pilotErrorCode) are relevant for the attention of the US Facility, however difficult to observe a pattern.

Closer Look to Analysis Job Error Report (2)

- ddm:
  - Could not add output files to dataset
- athena:
  - Athena ran out of memory, Athena core dump
  - ATH_FAILURE - Athena non-zero exit
  - Athena core dump or timeout, or conddb DB connect exception
  - Athena crash - consult log file
- user:
  - User work directory too large
- other:
  - Unknown error code

Note: above is my classification of errors, have not double checked with Torre as to how they are classified in the analysis job error report on Panda monitor.

User problems reported to DAST for US sites

DAST – Distributed Analysis Support Team

- DAST started in September 2008 for a combined support of pathena and Ganga users.
- First point of contact for distributed analysis questions.
- All kinds of problems are discussed in the DA help forum (hn-atlas-dist-analysis-help@cern.ch) not just pathena and Ganga related ones:
  - athena
  - physics analysis tools
  - conditions database access
  - site/service problems
  - dq2-* tools
  - data access at sites
  - data replication
  - etc
- DAST helps directly by solving the problem or escalating to relevant experts

Team Members

EU time zone	NA time zone
Daniel van der Ster	Nurcan Ozturk (now in EU time zone)
Mark Slater	Alden Stradling
Hurng-Chun Lee	Sergey Panitkin
Bjorn Samset	Bill Edson
Christian Kummer	Wensheng Deng
Maria Shiyakova	Shuwei Ye
Jaroslava Schovancova	Nils Krumnack
Manoj Jha	Woo Chun Park
Elena Oliver Garcia	
Karl Harrison	
Frederic Brochu	
Daniel Geerts	
Carl Gwilliam	
Mohamed Gouighri	
Borge Gjelsten	
Katarina Pajchel	
Eric Lancon	

red: trainee

A small team in NA time zone, most people are the ones who are fully/partly supported by the USATLAS Physics Support and Computing Program, thus close ties to USATLAS Facilities and Support.

User problems reported to DAST for US sites (1)

By looking at the recent DAST shift reports, here are some of the issues reported by users on US sites:

- A wrong update concerning the analysis cache 15.6.13.1.1. Affected several sites including US ones.
- Error at BNL**: pilot: Get error: lsm-get failed.** The pool hosts the input files was not available due to machine reboot. The pool is back shortly after.
- MWT2 site admin reporting user analysis failures (HammerCloud and user code problems). Very useful for heads-up and understanding the site performances better.
- Production cache 16.0.1.1 missing at MWT2. Installed by Xin, available at several analysis sites but not all, no indication of usage in physics analyses at: http://atlas-computing.web.cern.ch/atlas-computing/projects/releases/status/

User problems reported to DAST for US sites (2)

- dq2-get error, "lcg_cp : Communication error on send’ at MWT2 – Transient error.

- Why do Australian grid certs not work on BNL? dq2-get difficulties by some Australian users. Shifter has not escalated this issue to BNL yet

- Tier3 issue – User can't use WISC_LOCALGROUP disk as a DaTRI source. DDM people reported that it is in the works.

User File Access Pattern (1)

- Pedro examined the user access pattern at BNL. His observations as follows:
- Took a 1.5h sample, 29 files per second were accessed just for reading. 56% of the files are user library files. This suggested that users are splitting jobs into many subjobs (each subjob needs to access the library file).
- Outcome: Lots of reads and writes to the userdisk namespace database, the writes clog up and timeouts on the whole pnfs namespace. Heavy usage of the same file, possibility to break the storage server, the files will not be accessible since there is no redundancy (this is not a hotdisk area). In this case he estimated a failure of more than 2/sec and depending on the problem, the recovery can take 2min-15min (240-1800 job failures).
- Solution: If this continues to be the practice among users, he needs to add more changes on the local site mover and add some extra 'spices' to be able to scale the whole system.(except the data availability since for that lib files would need to go to HOTDISK by default or some new HOTUSERDISK token).

User File Access Pattern (2)

- I looked at the user jobs Pedro gave as examples.

user	gregor	mdavie	csandova	zmeng
jobsetID	2354	3909	639	3879
# subjobs	4999	2295	3884	1390
#input file/job	4 ESD	1 D3PD	1 ESD	1 ESD
average run time/job	3h	3' to 37'	13’ to 3h	14’ to 50’
average input file size	800 MB	200 MB (some 6 MB)	3 GB	3 GB
average output file size	60 KB	300-900 KB	200 MB	130 MB

User File Access Pattern (3)

- Conclusion: Users are splitting into many subjobs and their output file size is small, then they have difficulty to download them by dq2-get. As discussed in DAST help list last week, the file-look up can be as long as 30s in DDM system, so users have been advised to merge their output files at the source and then download.
- The default limit on the input file size is 14GB in the pilot. So these users could use to run on more input files in one job.They are probably splitting more to be on the safe side.
- How could we advise users? Can pathena print a message when the jobis splitted into too many subjobs?
- This needs to be further discussed.

Storage space alerts for SCRATCHDISKs

DAST receives notifications from DQ2 system. An example follows from today. NET2 is often in the list. Is there a problem with cleaning the SCRATCHDISK?

Date: Tue, 12 Oct 2010 10:06:25 +0200

From: ddmusr01@cern.ch

To: atlas-project-adc-operations-analysis-shifts@cern.ch,
fernando.harald.barreiro.megino@cern.ch,
alessandro.di.girolamo@cern.ch, dan@vanderster.com

Subject: [DQ2 notification] Storage space alerts for SCRATCHDISKs

Site	Free(TB)	Total(TB)
IN2P3-LPSC_SCRATCHDISK	0.359 (16%)	2.199
NET2_SCRATCHDISK	0.047 (8%)	0.537

For questions related to this alarm service: atlas-dq2-dev@cern.ch

How US Cloud Provide Support to DAST

- Site issues are followed by GGUS tickets:
  - Good response from site admins
  - Some site admins also watch for issues reported in the DA help forum for their sites, thus prompt support.
- General questions:
  - atlas-support-cloud-us@cern.ch
  - Good response, Yuri often keeps an eye on this list.
- Release installation issues at US sites:
  - Xin Xhao
- Also private communications with Paul Nilsson, Torre Weanus on the pilot and Panda monitor issues (not only for US sites of course).

Summary

- US cloud continues to be a successful cloud in ATLAS distributed user analysis.
- Good US site performances, one of the least problematic analysis sites in ATLAS.
- User analysis jobs fail with diffferent errors. The dominant ones are classifed as “panda” and “athena” on Panda monitor. “athena’ errors often refer to user code problems. “panda” errors vary, difficult to observe a pattern or classify as common problem.
- Good US cloud support to DAST. No operational issues to report in the US Facility, in particular from the DAST point of view.
- User access pattern needs to be further discussed for better site/storage performances.
- I have not mentioned here the new Panda Dynamic Data Placement (PD2P) mechanism which helped greatly users (long waiting period in the queue) and sites (storage). I expect Kaushik will cover this topic.