



LANCASTER
UNIVERSITY



ATLAS - STEP09 and beyond

Peter Love
Department of Physics
Lancaster University

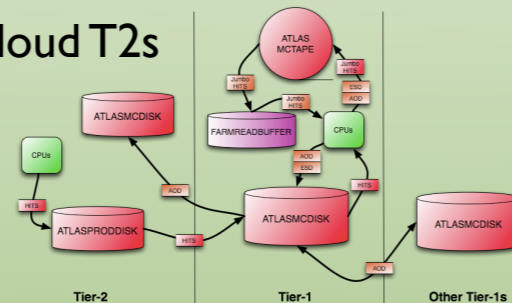
Content

- What occurred during STEP09?
- What lessons were learnt?
- Beyond STEP09
- Summary

STEP09 metrics

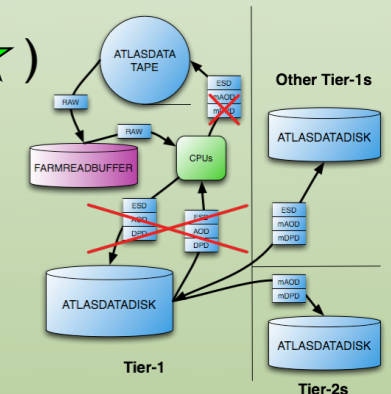
Simulation production

- T2 HITS uploaded to T1
- merge HITS at T1 and archive to tape
- reconstruct at T1 and produce merged AOD
- deploy to sister T1 and cloud T2s
- reco 50% of generated



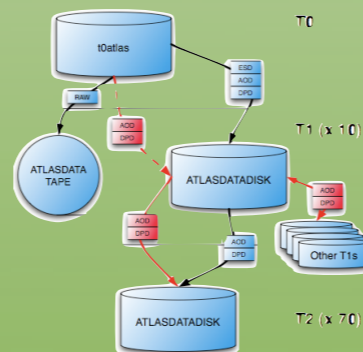
Reprocessing

- reprocessing at Tier-1s RAW->ESD
- 400Hz and 1000Hz metric (★)
- >99.9% of files reprocessed
- merged and distributed



Data distribution

- 14h, 200Hz run cycle
- all data at Tier-1 within allocated period
- participating T2 no outage >24h



Analysis

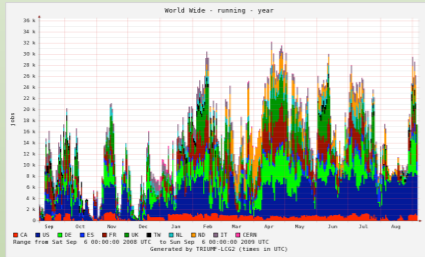
- HammerCloud tool to load CPU slots including usual user analysis
- occupy 50% of ATLAS capacity
- 85% efficiency, event gen 15Hz
- stressed WN access to local storage

- participation from all Tier-1, no outage >12h
- participating T2 no outage >24h

STEP09 - condensed summary

Simulation production

- 100k jobs/day, 12M events during STEP09
- various simulation types - works well



tasktype	defined	assigned	waiting	activated	running	holding	transferring	success	failure	efficiency
simul	528	3859	0	28846	25805	4547	24436	1229723	204439	85.7%
pile	18	2279	154	6846	4349	760	1109	615194	65704	90.4%
reco	1	1664	320	761	2398	335	353	481085	96498	83.3%
merge	35	9	137	551	82	91	407	78489	45852	63.1%
evgen	0	0	5	31	10	16	16	64462	17453	78.7%
digit	0	0	0	262	0	0	0	17872	488	97.3%
reprocessing	0	0	0	8	0	0	0	0	0	-
total	582	7811	611	37279	32665	5743	26321	2486825	430434	85.2%

Reprocessing

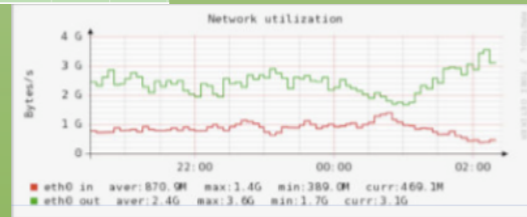
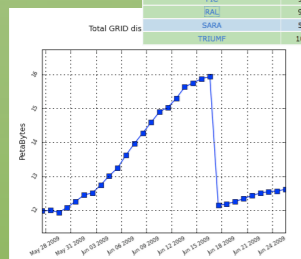
- Psuedo reprocessing at Tier-1s, writing small files
- tested number of files/hr written to tape
- RAL golden green star performance

TI	Base Target	Result	Comment
ASGC	10 000	4 782	Many batch system and basic setup problems
BNL + SLAC	50 000	99 276	Also ran high priority validation and other tasks
CNAF	10 000	29 997 ★	
FZK	20 000	17 954	Big tape system problems pre-STEP; no CMS
LYON	30 000	29 187	Very late start due to tape system upgrade, then good
NDGF	10 000	28 571 ★	
PIC	10 000	47 262 ★	
RAL	20 000	77 017 ★	
SARA	30 000	28 729	Tape system performance very patchy
TRIUMF	10 000	32 481 ★	Also ran high priority validation and other tasks

Data distribution

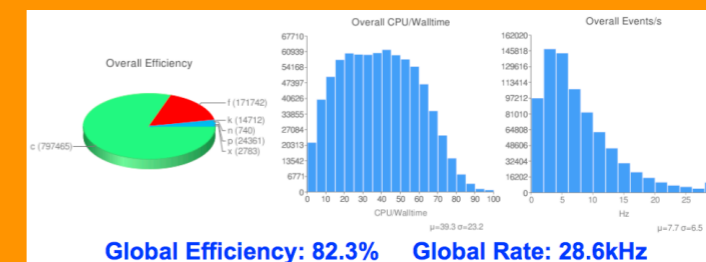
- 4PB transferred, 5.5GB/s peak rates

Cloud	Efficiency	Throughput	Successes	Datasets	Files	Transfer
ASGC	99%	397 MB/s	6286	833	6299	59
BNL	84%	1128 MB/s	23903	855	24000	4698
CERN	100%	334 MB/s	5712	133	5692	10
CNAF	98%	561 MB/s	7638	382	7652	150
FZK	85%	556 MB/s	6852	381	6858	1200
LYON	96%	620 MB/s	8510	749	8506	325
NDGF	84%	137 MB/s	913	102	907	170
PIC	93%	429 MB/s	3365	908	3361	256
RAL	99%	838 MB/s	13324	1656	13726	125
SARA	53%	262 MB/s	3186	134	3171	2882
TRIUMF	100%	297 MB/s	4704	169	4706	13



Analysis

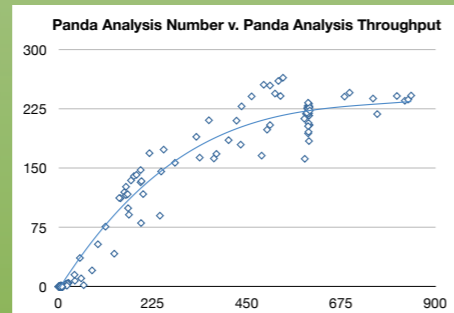
- 4 analysis types, 2 backends, 3 file access types
- 1 million jobs, 82% success, 26.3 billion events
- mean event rate 7.7Hz, cpu/wall 0.39
- PanDA 85% eff. with event rate > metric
- WMS 75% eff. with event rate < metric
- Many issues, many tool improvements



Detailed WLCG post-mortem here: <http://indico.cern.ch/conferenceTimeTable.py?confId=56580> (These pics from Graeme and Dan's ATLAS talk)

STEP09 - Tier2 lessons (i)

- STEP09 analysis summary
<http://gangarobot.cern.ch/st/step09summary.html>
- STEP09 feedback from sites (UK reports from Glasgow, Lancaster, Sheffield)
<https://twiki.cern.ch/twiki/bin/view/Atlas/Step09Feedback>
- Glasgow - data import from RAL was affected by loaded disk servers due to analysis access
- Glasgow - received large .stream dataset, causing a backlog on 1Gb/s JANET link, managed to complete transfer during 48h grace period



- Lancaster - many large .stream datasets, expected 30TB, actual 45TB. Required major intervention to meet the data transfer metric
- Lancs/Liverpool hit by software install problems, trashed all jobs from 3/4 analysis types - fixed late in 2nd week

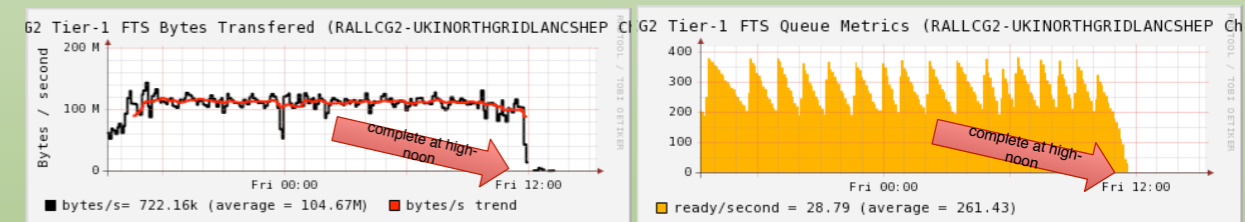
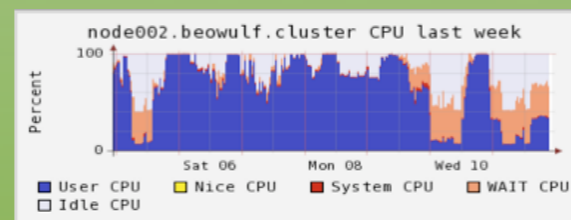


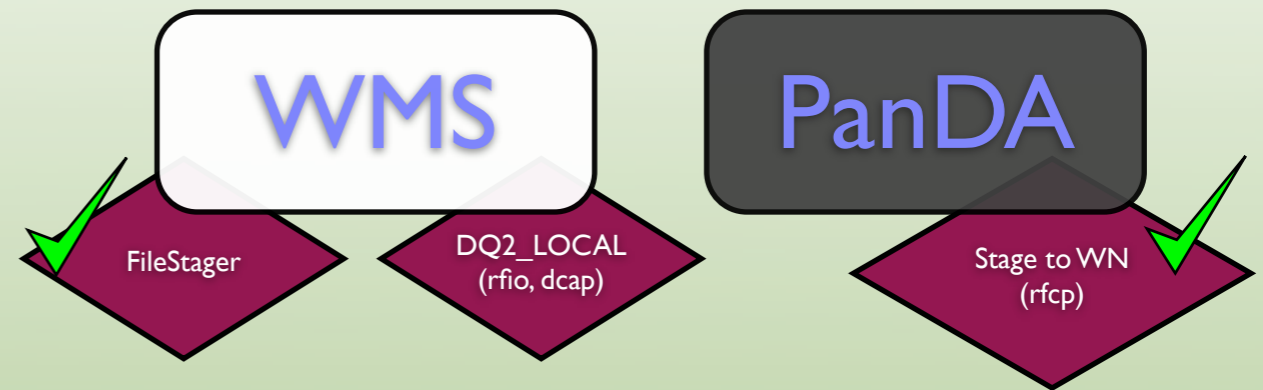
Figure 4. RAL traffic over 100MB/s, completing transfers on last day of STEP09

- JANET links may need upgrading
- Flexible FTS config required, responsive team at RAL



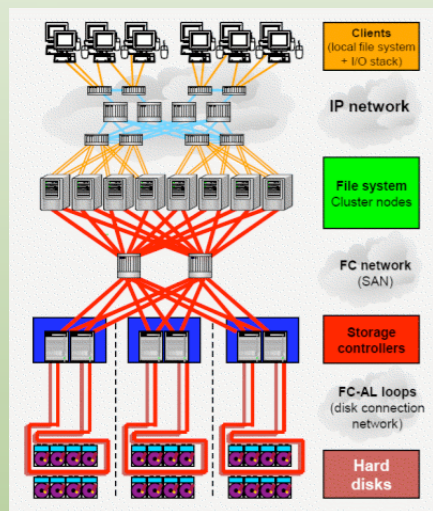
STEP09 - Tier2 lessons (ii)

- Submission via WMS or panda pilots
- Twelve T2 sites in UK: 1/3 good, 1/3 mediocre, 1/3 could do better
- random reads clobber disk server, cause LAN congestion
- STEP illustrated difficulty in isolating behaviour of different job types, valuable experience for sys-admins prior to data taking
- reassess sites before data taking in order to optimize data placement shares
- sites: know your limits and throttle



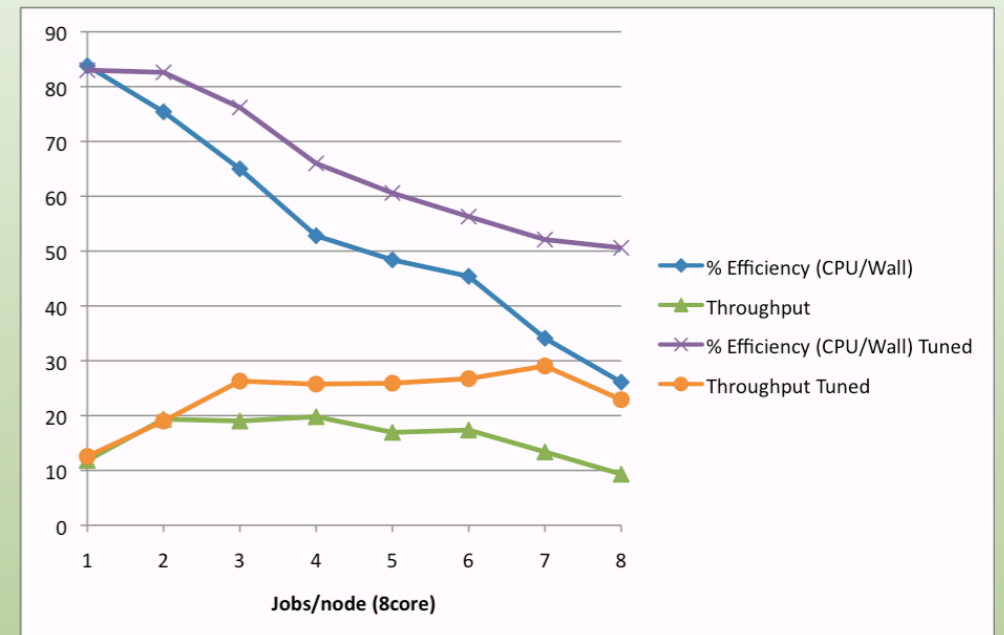
Token	Share (%)	Expected (TB)	Reserved (TB)	Used(TB)	Free(TB)	Usage	Updated
UKI-LT2-QMUL_DATADISK	16	17.9	75.2	0.2	75.0		2 hours, 19 minutes ago
UKI-LT2-RHUL_DATADISK	24	26.9	94.6	17.8	76.8		2 hours, 18 minutes ago
UKI-LT2-UCL-CENTRAL_DATADISK	4	4.5	17.6	2.1	15.5		2 hours, 17 minutes ago
UKI-NORTHGRID-LANCS-HEP_DATADISK	28	31.4	59.4	18.8	40.6		2 hours, 13 minutes ago
UKI-NORTHGRID-LIV-HEP_DATADISK	12	13.4	33.0	14.9	18.1		2 hours, 12 minutes ago
UKI-NORTHGRID-MAN-HEP1_DATADISK	16	17.9	18.7	1.9	16.8		2 hours, 22 minutes ago

Tuning @ Liverpool



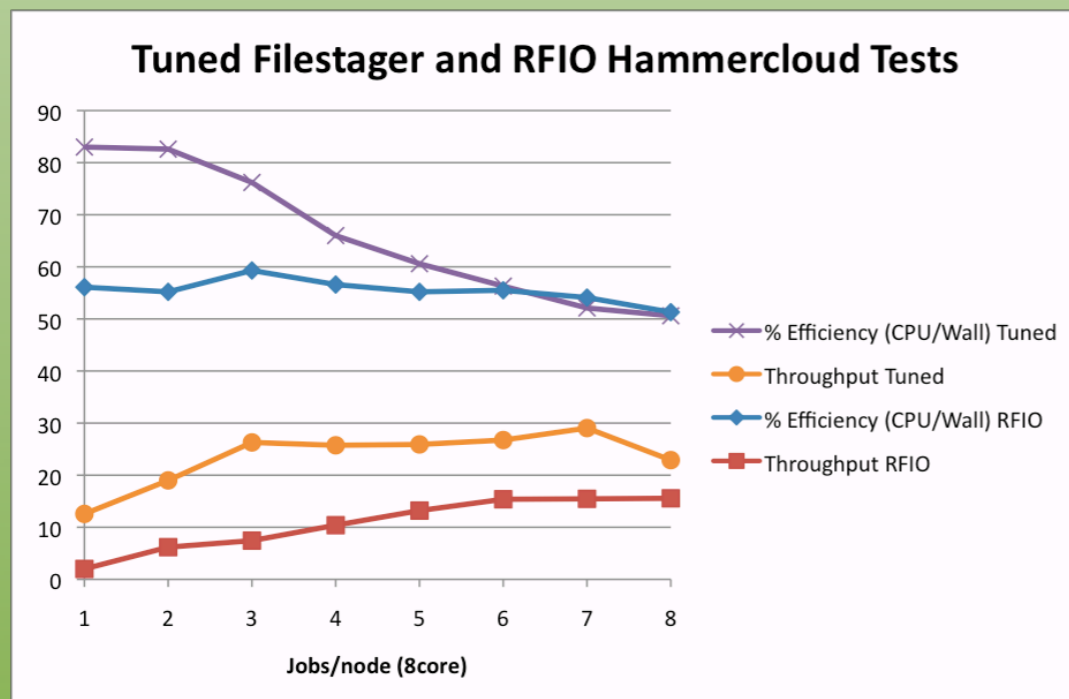
Liverpool has extensive tuning, very informative for all sites

- FileStager
- large merged AOD cannot be cached
- writes clobber WN disk, blocking reads
- RAID0, separate sys disk, repartition, page cache flushing timeout



TERENA storage benchmarking HOWTO

https://wiki.terena.org/index.php/TF_Storage_Benchmarking_Howto



- HammerCloud test #582
- comparing tuned filestager (above) to RFIO with 4kB read-ahead buffer
- FileStager gives better throughput than RFIO (in LIV case)
- many jobs leads to high IOWAIT on disk servers, lower efficiency on WN
- Concluded FileStager giving better throughput
- Discussion tomorrow...

Beyond STEP

- SL5
- Pilot jobs glexec



SEPT09 and tuning of sites

- New LFC instance at RAL for UK cloud
- Ongoing developments with production system
- Review token spaces, sites will be asked to make changes over next months
- Keep on tuning!

Spacetoken	% (CM)	%	Spacetoken	% (CM)	%
DATADISK	50	25	DATADISK	25	
DATATAPE	5	5	MCDISK	50	
MCDISK	25	25	GROUPDISK	10	
MCTAPE	5	5	SCRATCHDISK	10	
GROUPDISK	15	10	PRODDISK	5	
SCRATCHDISK	0	10			
(Stage)	0	20			

Beyond STEP09 - SL5

- Scientific Linux version 5
 - sites have been asked to install SL5 WNs 'as soon as possible', separate queues please
 - ATHENA SL4 kits run on SL5 machines, Metapackage rpm available to sort compatibilities
 - disable SELinux 'execcheap' protection (releases <15.4.0)
 - full SELinux compatibility being worked on, difficult to backport solution
 - Outputfile differences traced to run-time library
SLC4: libm.so.2.3.4 SLC5: libm.so.2.5
 - SIT group plan to complete large-scale validation of SL5/gcc43 before data taking
 - Migration guide here: <https://twiki.cern.ch/twiki/bin/view/Atlas/SL5Migration>
- RAL upgrading this month

Pilots and glexec

- development of glexec, LCMAPs SCAS plugin
- testing at various sites, Lancaster has glexec/SCAS
- ATLAS Pilot ongoing development and testing
- to be deployed?
- continue with Pilot role at selected sites

Beyond STEP09 - SEPT09



- Planned September 2009 re-run of STEP exercise
 - many different real users running own jobs on large datasets
 - sites should consider consequences to batch scheduler and LAN load
 - need to nail down sites' throughput, as this determines the fraction of data deployed to each site
- SEPT09 to determine global analysis throughput (improve on 30kHz global STEP09)

ADC developments

- ADC continuously being improved - some news from Software Week
Copious amounts of information here:
<http://indico.cern.ch/conferenceDisplay.py?confId=50976>
- ganga robot site testing
- HammerCloud improvements, admin interface
- ATLAS Grid information system (AGIS)
- ProdSys procedure, file loss cleanup actions to be improved
- <http://dashb-atlas-prodsys.cern.ch/dashboard/request.py/overview>
- Tier-3 proposals
- Pilot refactoring, pilot factory consolidation

Summary

- STEP09 huge effort, fortunately long past...
- ...SEPT09 desired
- Main ATLAS activity T0->T1->T2 robust
- Understanding of T2 performance greatly improved
- Continue with performance studies
- Normal analysis usage still unclear, expect surprises

On target.



Reference

- ATLAS STEP09 washup meeting
<http://indico.cern.ch/conferenceDisplay.py?confId=62853>
- Graeme's GDB talk on plan for STEP'09
<http://indico.cern.ch/conferenceDisplay.py?confId=45475>
- STEP09 analysis summary
<http://gangarobot.cern.ch/st/step09summary.html>
- STEP09 feedback from sites (UK reports from Glasgow, Lancaster, Sheffield)
<https://twiki.cern.ch/twiki/bin/view/Atlas/Step09Feedback>
- WLCG STEP09 post-mortum
<http://indico.cern.ch/conferenceTimeTable.py?confId=56580>
- TERENA storage benchmarking HOWTO
https://wiki.terena.org/index.php/TF_Storage_Benchmarking_Howto
- ATLAS software and computing week
<http://indico.cern.ch/conferenceDisplay.py?confId=50976>