Analysis at T2 activity in STEP09

interim assessment as much understanding is in progress, Friday picture could be a little different

Sanjay Padhi, James Letts, Thomas Kress, Dave Evans (non-FNAL), Dashboard and CRAB developers, Stefano Belforte, Frank Wuerthwein

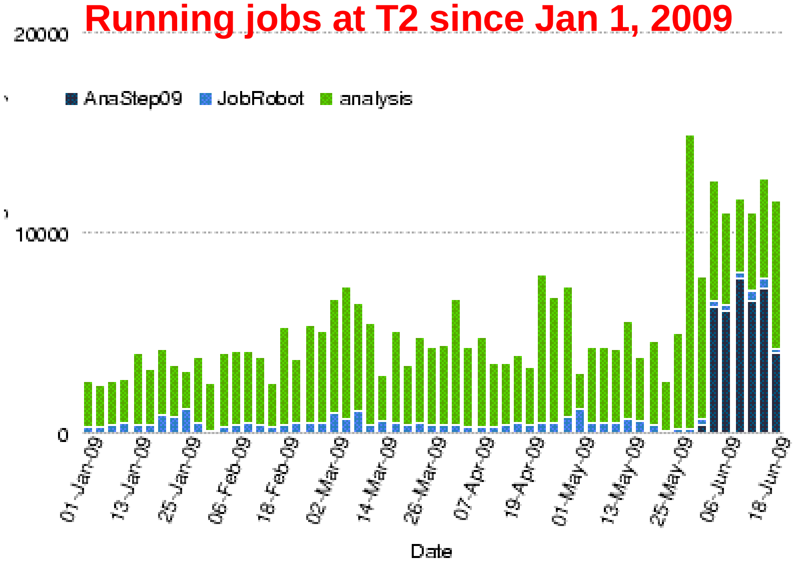
Goals

- From twiki:
  - ☞ https://twiki.cern.ch/twiki/bin/view/CMS/Step09
  - ☞ https://twiki.cern.ch/twiki/bin/view/CMS/Step09T2

1. demonstrate analysis at a scale using all pledged resources at T2
   - Close to 16,000 pledged slots, about 50% for analysis
2. Explore, validate, extend monitoring tools
   - Monitoring the totality of jobs to understand the available monitoring capabilities, and investigate the fairshare situation at T2 sites.
3. explore data placement
   - measure how (much) the space granted to physics groups is used
   - Move dataset aroung as we expect to do in operations
   - Monitor effect on job submission
4. Derive more clues for Analysis Operations

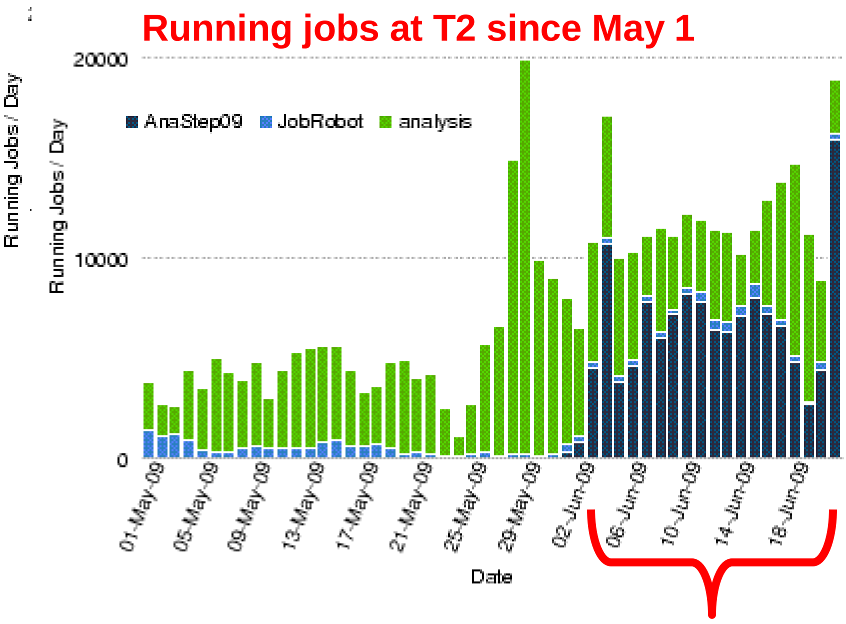
1.The Jobs

- Strong user analysis activity since long



Running jobs at T2 since Jan 1, 2009

- Are we about to hit a wall ?

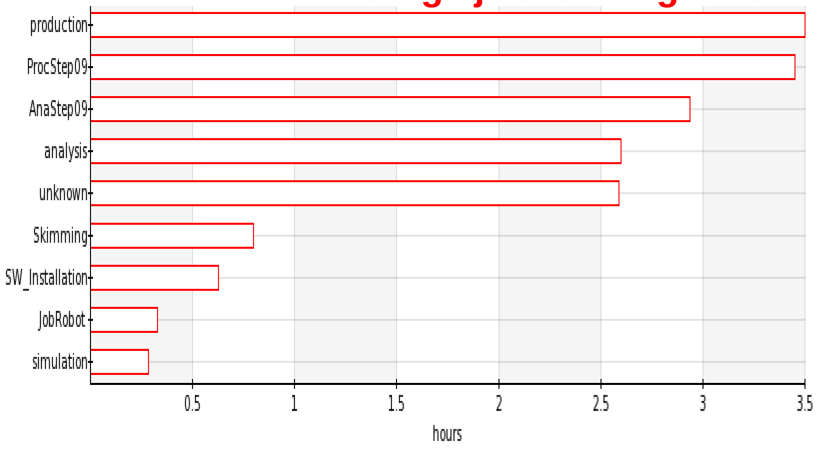


Running jobs at T2 since May 1

June 2 – 22

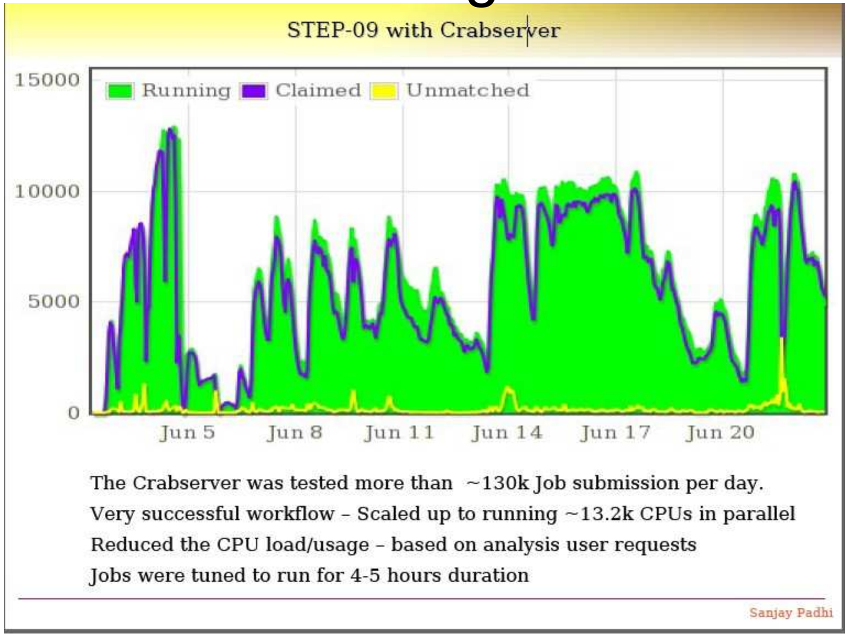
- AnaStep09 jobs are “typical”

Average job running time



AnaStep09 submission

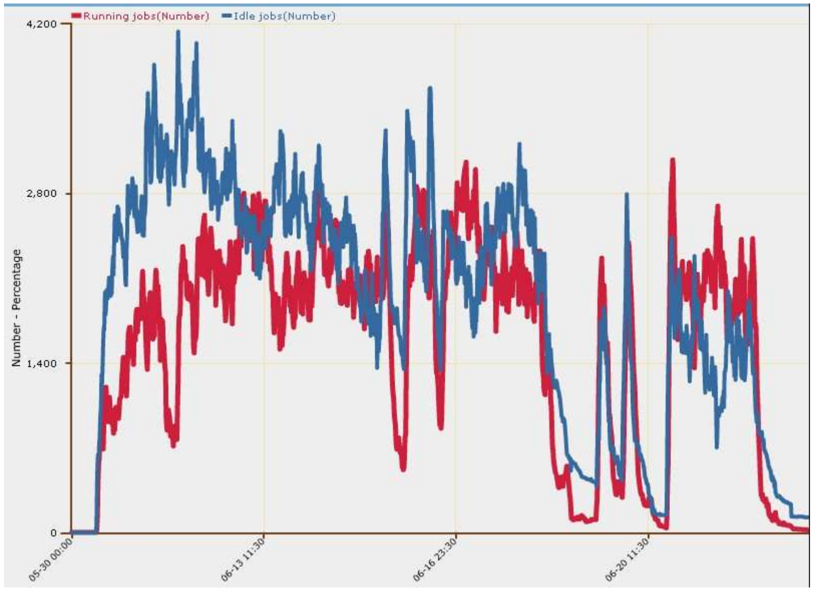
UCSD-CS + glideInWMS



The Crabserver was tested more than ~130k Job submission per day.
Very successful workflow – Scaled up to running ~13.2k CPUs in parallel
Reduced the CPU load/usage – based on analysis user requests
Jobs were tuned to run for 4-5 hours duration

Sanjay Padhi

Bari-CS + gLite WMS

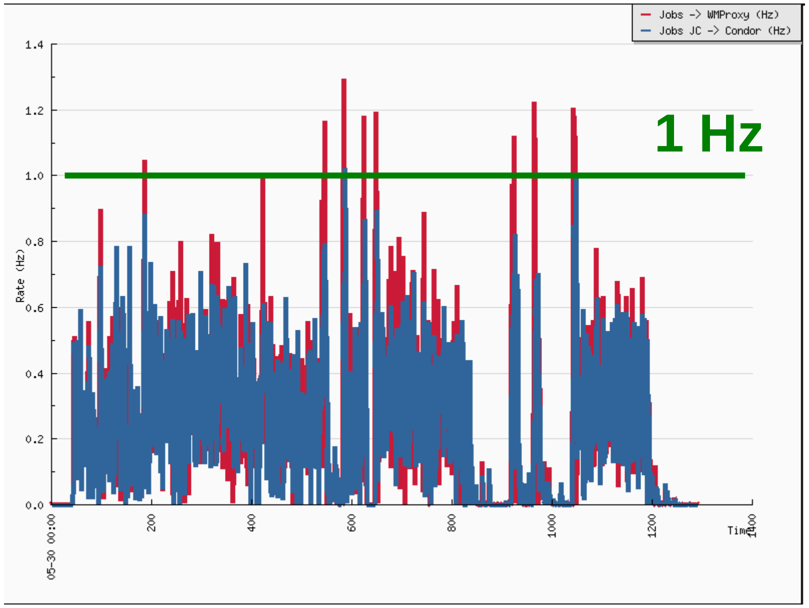


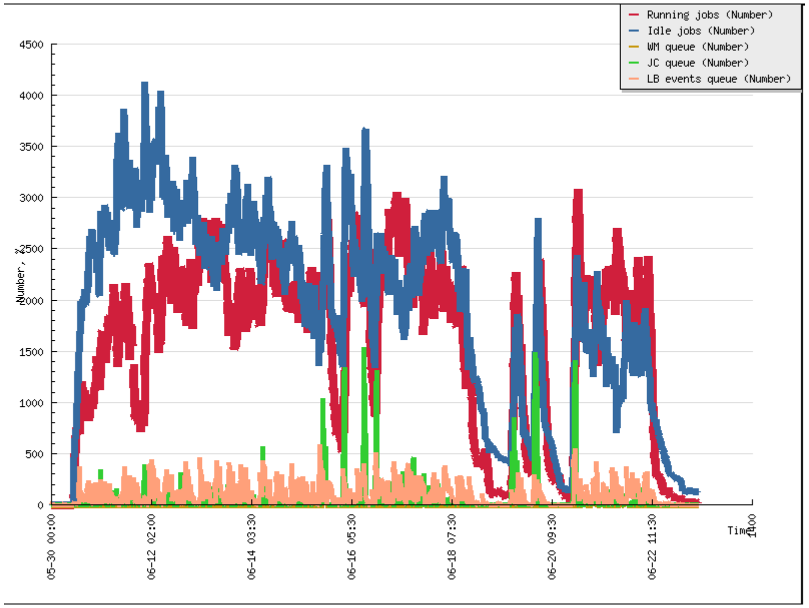
- Structure partly due to automatic submission getting stuck where a user would have Ctl-C-ed sooner
- Partly responding to sites load and need to start/stop to put changes in

- Submitted to about 40 T2's
  - Also used several T3's
    - T3_IT_Padova, T3_IT_Perugia, T3_UK_London_QMUL, T3_UK_London_RHUL, T3_US_Colorado, T3_US_Omaha, T3_US_TTU, T3_US_UMD

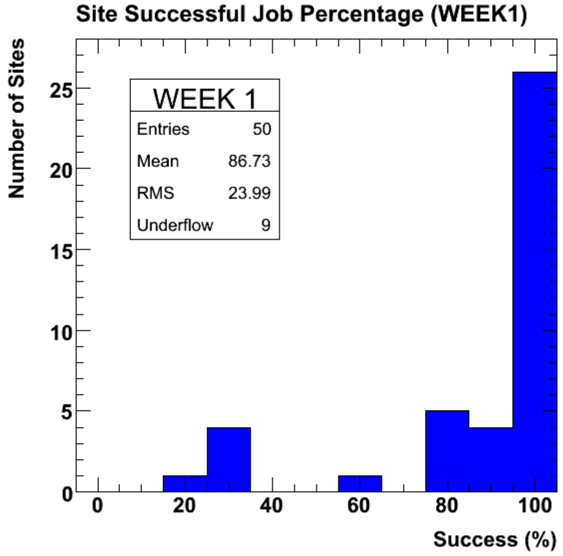
Side note on new gLite 3.2 WMS

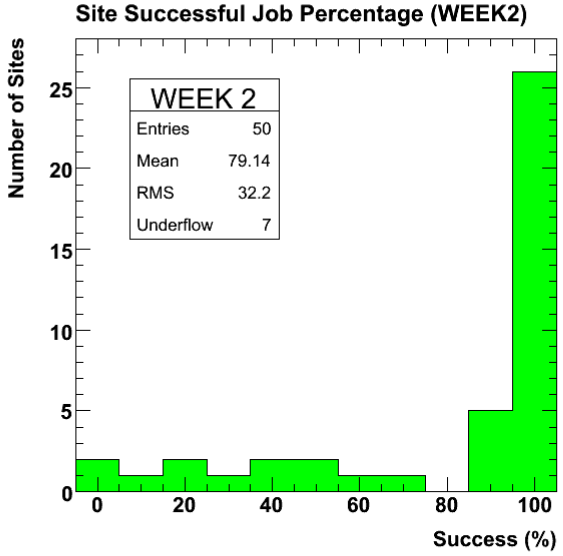
- Only one gLite WMS used by Bari Crab Server (was not limiting)
- Ran stabily at 0.3 Hz (30K jobs/day), peacks at 1Hz
  - No internal queue buildup
  - Submission time to WMS comparable to CrabServer (30sec per task)
- Should not need more WMS'es then we have now
  - Job submission throughput is not an issue anymore
- There are new frontiers

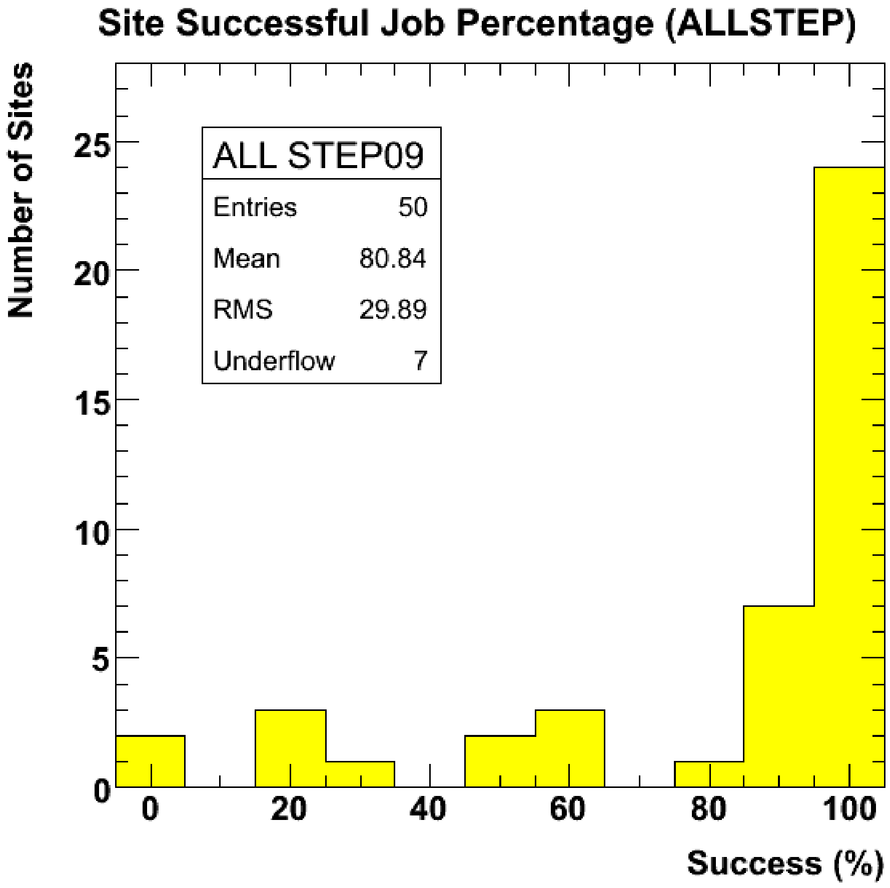




In the end jobs did fairly well


Site Successful Job Percentage (WEEK1)


Site Successful Job Percentage (WEEK2)

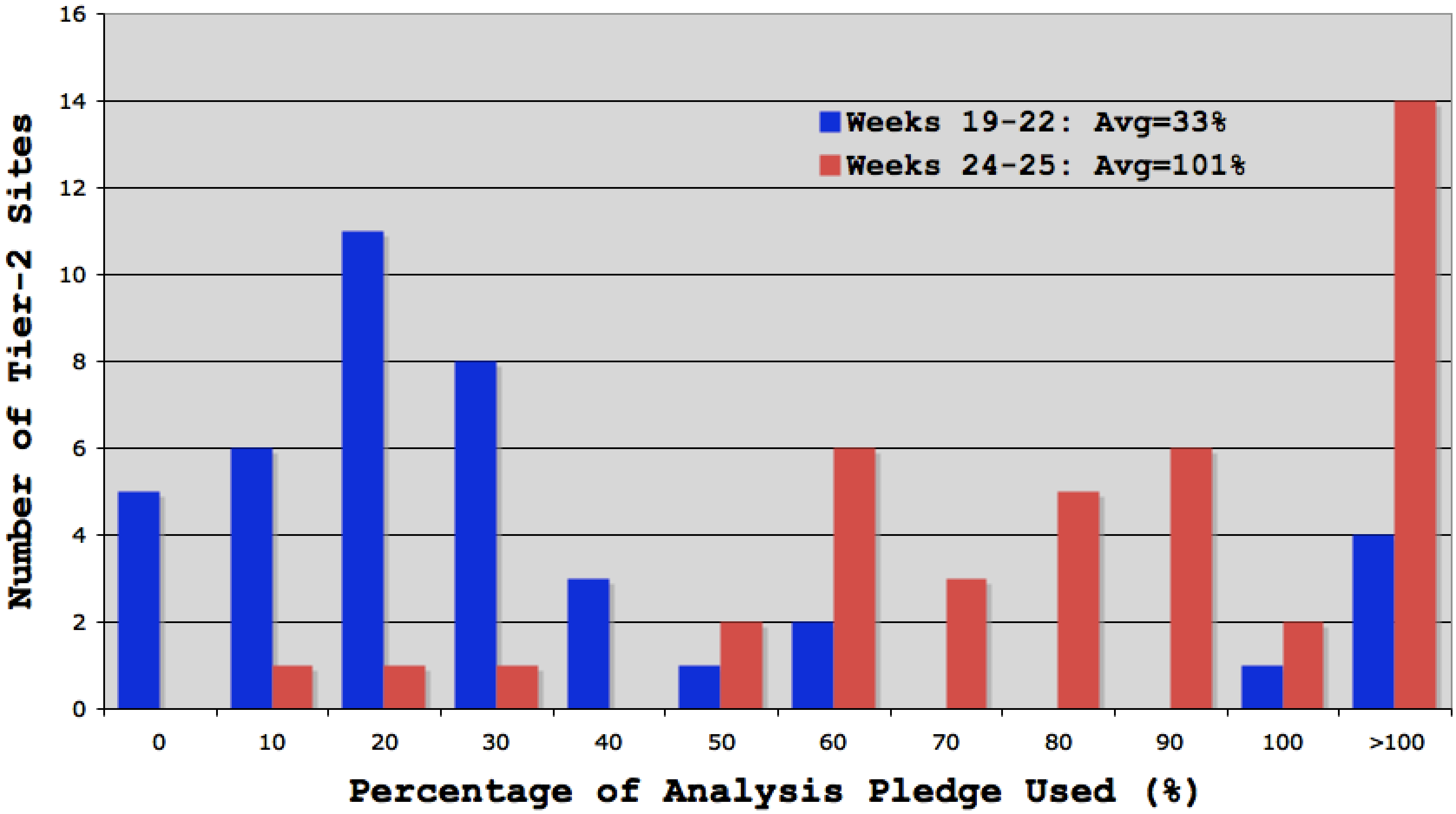

Site Successful Job Percentage (ALLSTEP)

Most errors are file read (exit code 8020)

2. Monitor (the big picture)

- Need to get pictures like this from dashboard UI

Analysis Usage Relative to Pledge
Before and During Step09



More on monitoring

- We need to test every tool/feature at large scale before users do
  - Stating the obvious
- Everytime we do it, we learn diagnostics is not sufficient
  - Why jobs were not submitted ? Resubmitted ?
  - Which file gave problems ?
  - Which WN's gave problems ?
  - What do those CRAB job states exactly mean ?
- Also need to do in "systematic/pedantic" way, not as user who is happy with any Q&D hack that works

- In order to help/support more users (i.e. understand their problems so that they can be fixed) need more operation oriented WM tools
  - For the tool "generation" that users will be using with first data
  - Before more functionalities are brought in
  - List of "desiderata" is under preparation (many things are the well known ones since long time anyhow)

Lessons from running jobs

- We doubled the load and “none” noticed
  - ~2h jobs vs. 15’ JobRobot (keep more files open)
    - ☞ a couple of years ago it would have killed several sites
  - Goal was not to break the system ➔ submission throttled down
  - May explain why did not hit 100% everywhere
  - Surely we could have submitted more jobs

- CRAB Server OK under load, scaling verified up to ~130K j/day

- Evidence of hot-data problems
  - Three users reading same dataset caused troubles
  - Expect more of this, also with stageout

- CRAB Server not so good when something goes bad
  - Poor error reporting (esp. for failed submissions)
  - Internal tracking could be improved
  - Communications can hang

Ongoing work this week

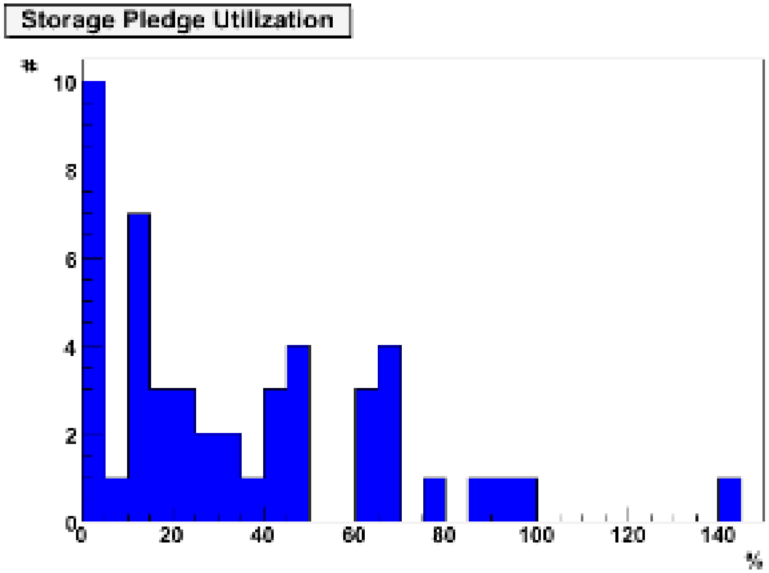
- Verify monitoring from dashboard against glideInWMS own monitoring and CRAB logs
  - Lot's of progress done already on better reporting, esp. for glideIn

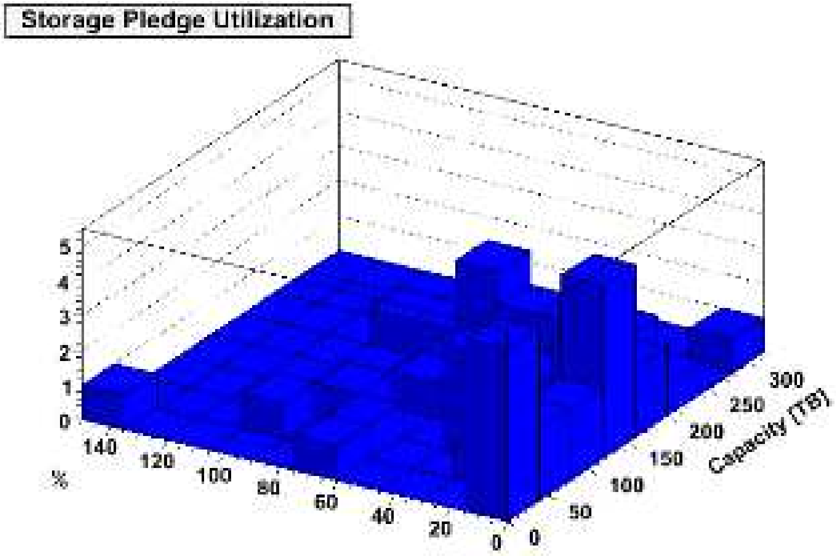
- Verify that fair share was working at sites

- Understand (on selected sites) why we did not fill pledges

3. Storage

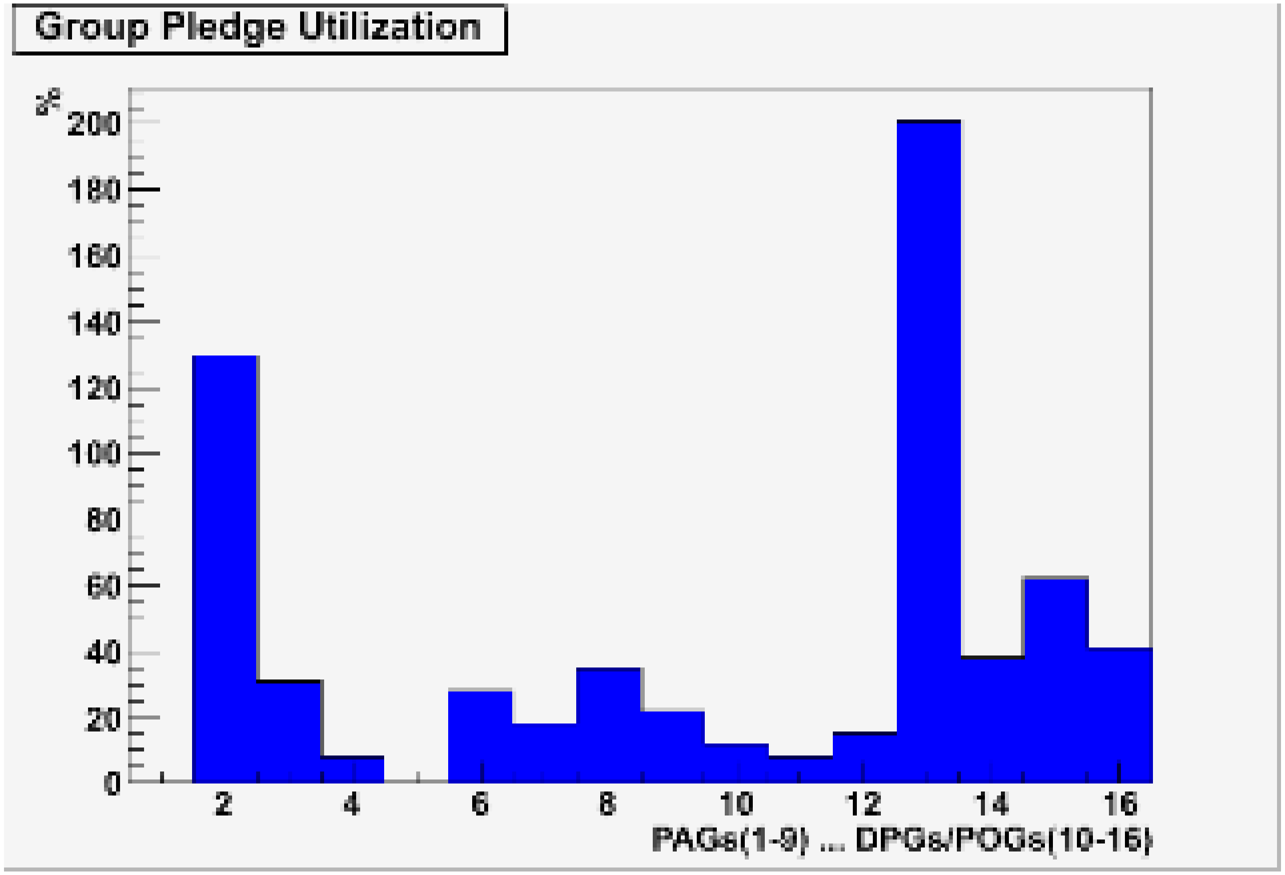
- Using information from PhEDEx/DBS we start to grab control of what is going on
- https://twiki.cern.ch/twiki/bin/view/CMS/Step09StorageMetrics
- basic ingredients are there, presentation/(dis)aggregation tools still need development





Storage usage by groups

- There is space
  - some have been faster in grabbing it
  - Conclusions (if any) need some thought



Moving data around

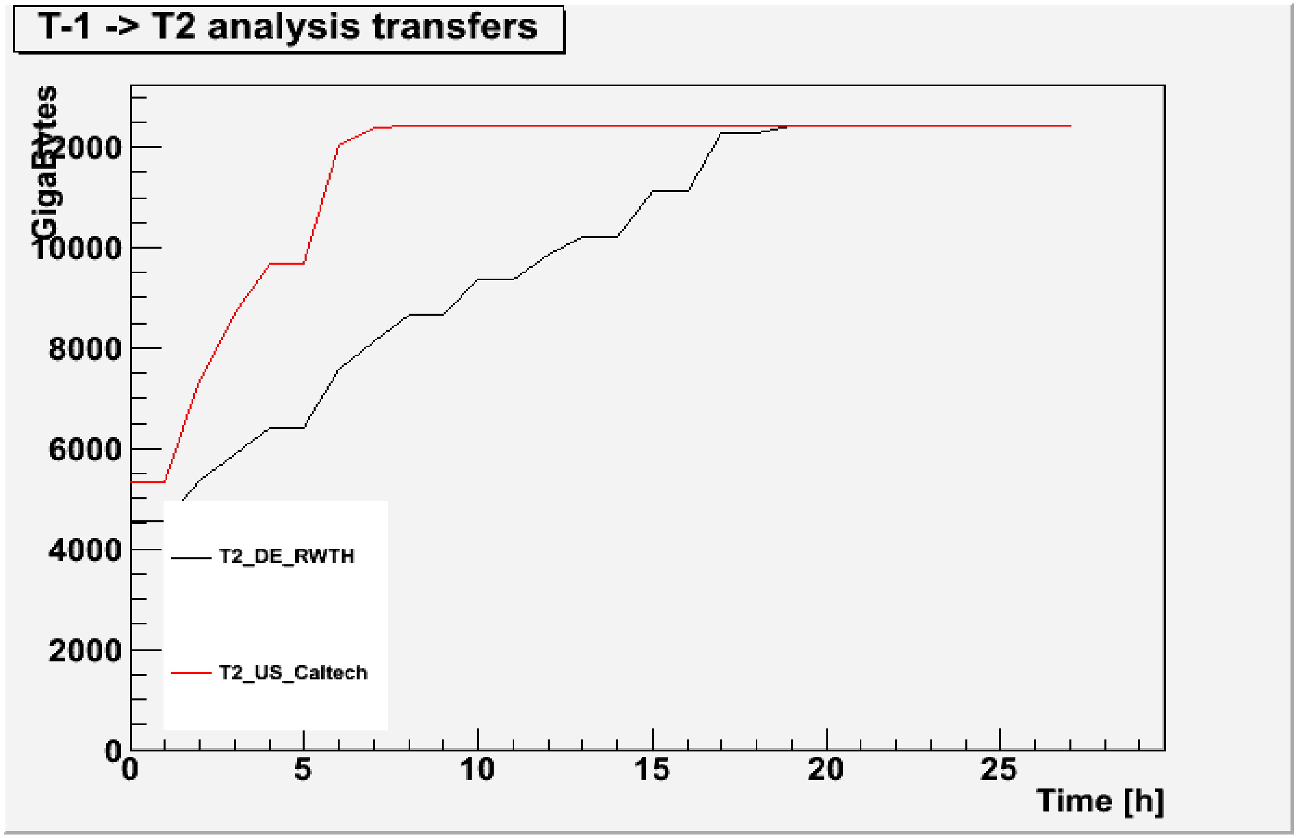
- Only used commissioned links
- Three cathegories of sites emerged

- ~3TB/hour, reliably and solidly
  - Can use as cache for fast action, replicate hot data JIT

- ~300GB/hour
  - Aligned with Computing TDR

- Data move very slowly, intermittently, not at all, not completely
  - Bandwidth limitations ? Malconfigured storage ? Non-attentive admins ?
  - May be difficult to use those T2's in real life
  - More investigation is needed on these cases (effort commensurate to expected gain)

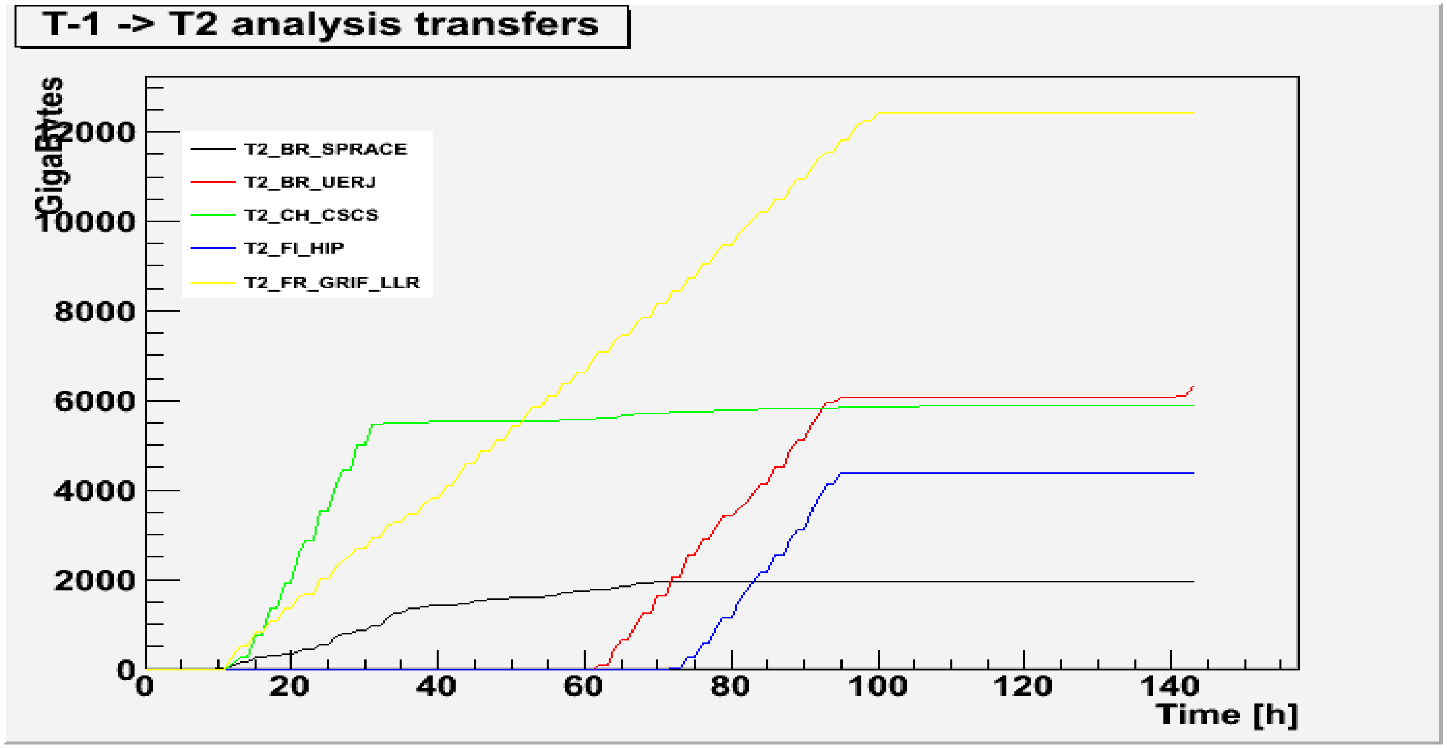
Fast sites



Reference sites with 10 Gbit/s network, very reliable and fast transfers, no tail issues

From FNAL one big 13 TB sample
(First few hours of history missing)

Normal sites



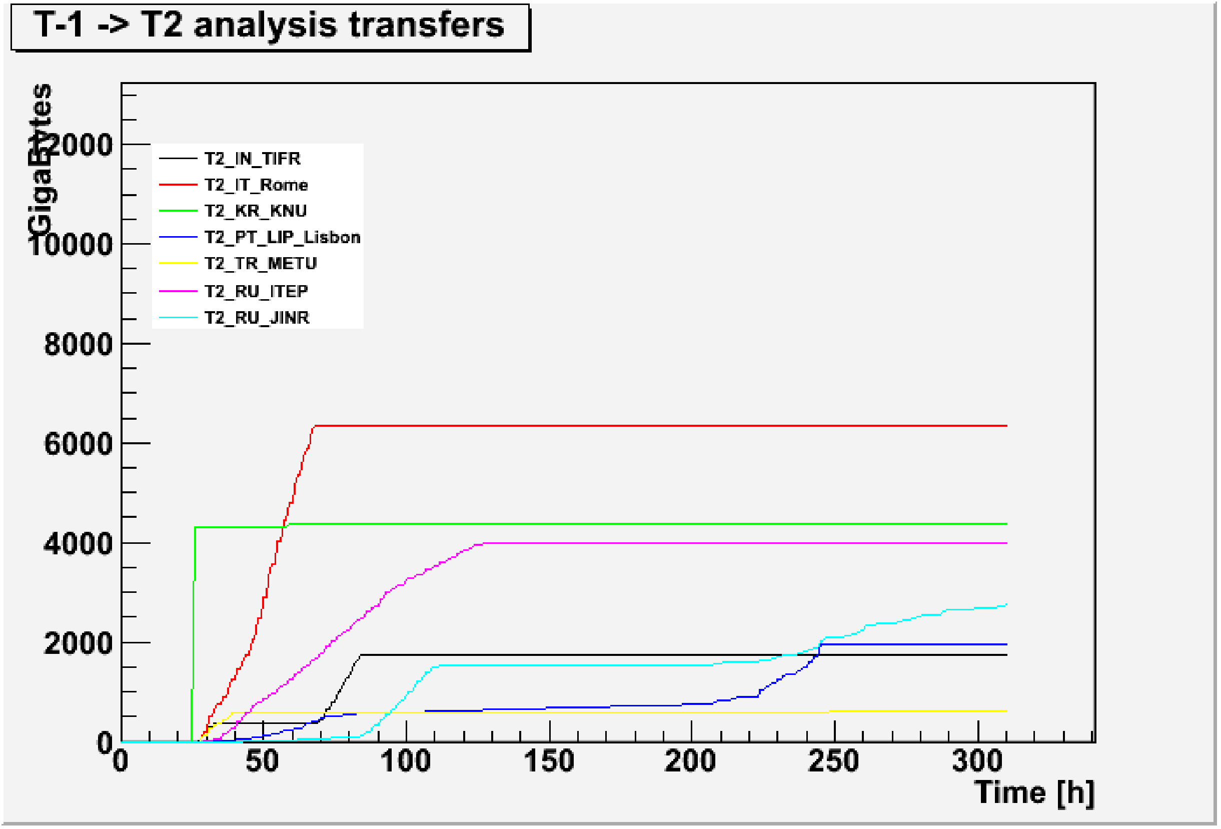
Note that not all sites had to transfer same amount of data!

FI_HIP accepted Phedex request rather late

For BR_UERJ it took long from immediate Phedex approval to transfer start

A few tail issues

Sites needing more follow up



Note that not all sites had to transfer same amount of data!

Impressive transfer speed for KR_KNU (real ?) but it took a few days to be fully registered in DBS

Slow stepwise transfer for PT_LIP_Lisbon

TR_METU, RU_JINR did not yet reach target (4 TB) after 2 weeks

IN_TIFR (target 13 TB) stuck

Lessons from data placement exercise

- Starting to look at new things: disk usage at T2
- Available information seems adequate
  - But too early to claim nothing more will be asked to offline
- Data can be moved at the expected rate
- But not really to every site

- Should we revisit "commissioned link" criteria ?

- Ongoing work this week:
  - Are there tails in dataset transfers ? (when was last file copied)
  - Correlation of dataset/jobs: do sites with more data have more jobs ? How much are datasets used ?
  - Did jobs follow the data (or users use white/black list so much that they do not notice) ?

TEMPORARY CONCLUSION

- Findings
  - Users will be able to run more analysis jobs when LHC starts
  - Confidence on infrastructure (sites, schedulers, submission tools) has increased
  - Monitoring is not up to the task yet
  - CRAB not enough operation oriented yet
  - Storage access will be largest source of pain

- Work in progress
  - Correlate site CPU/Disk usage
  - Do data replica pull jobs with them ?
  - Fair share at sites

- Work ahead
  - Private data publishing
  - Stage out
  - Push job submission to the limit (make users unhappy for a day ?)