EGEE Production – DDM Experience

Rod Walker, TRIUMF

- Definition and Executor types
- DDM interactions
- Errors
- Operations
- Conclusions

Definition and executor types

- EGEE
  - LCG sites
  - WLCG except NG and OSG
- Several ways to submit jobs to EGEE
  - gLite WMS – Lexor executor(not gLite CE's)
  - CondorG WMS – CondorG executor(“)
  - Condor Glide-ins – Cronus executor

Common Executor

- EGEE executors differ only in WMS part
  - i.e. submit,getstatus, getoutput
  - Db access via Eowyn, wrapper on WN and DDM interaction all the same
  - Migration to DDM was easy (but late) and 0.3 should be same.

DDM interactions

- Tasks assignment
  - To a T1 cloud. Input and output data is subscribed there, and jobs heavily weighted to run in cloud
  - Evgen output subscribed to all T1 disk
  - Very occasionally datasets have not been defined.
  - DDM services were no problem.

DDM interaction(cont.)

- Job submission
  - Need to get guid for each input lfn
  - Executor lists dataset lfn:guid and maintains 10k pair cache.
    - popular inputs always in cache, e.g. DbRelease
    - Some cache turnover, but typically 1 lookup per active task per executor instance. Not persistent at restart.
  - DDM service interruption, barely noticeable due to cache

DDM interaction(cont.)

- No DDM from worker node(later)
- Job validation
  - attach log file to dataset for all jobs inc.failed
  - attach output files to datasets for validated jobs
  - Bulk, pseudo-atomic attach
    - Need all output files to either attach or not(i.e. atomic operation)
    - For all finished jobs, gather lfn:guid pairs per dataset. Run 1 attach per dataset
    - Only validate jobs when ALL attaches are successful – retry indefinitely

DDM interaction(cont.)

- Validation is main dependency on DDM
  - attach can take up to a few minutes when in bad shape (before Xmas)
  - Only bulk per dataset, so must run N times
  - Not atomic per job, so 1 failure delays all job validation, e.g. missing dataset
  - So far not a major problem
  - Would like single command to attach files per job and return success/failure per job.

WN data handling

- Stage-in and stage-out are via lcg-utils
  - lcg-cp, lcg-cr and external services LFC, BDII
  - adding ability to get sfn from OSG,NG then lcg-cp
  - Would like DDM stage-in and out function
    - Timeouts, failovers and retries and not our problem

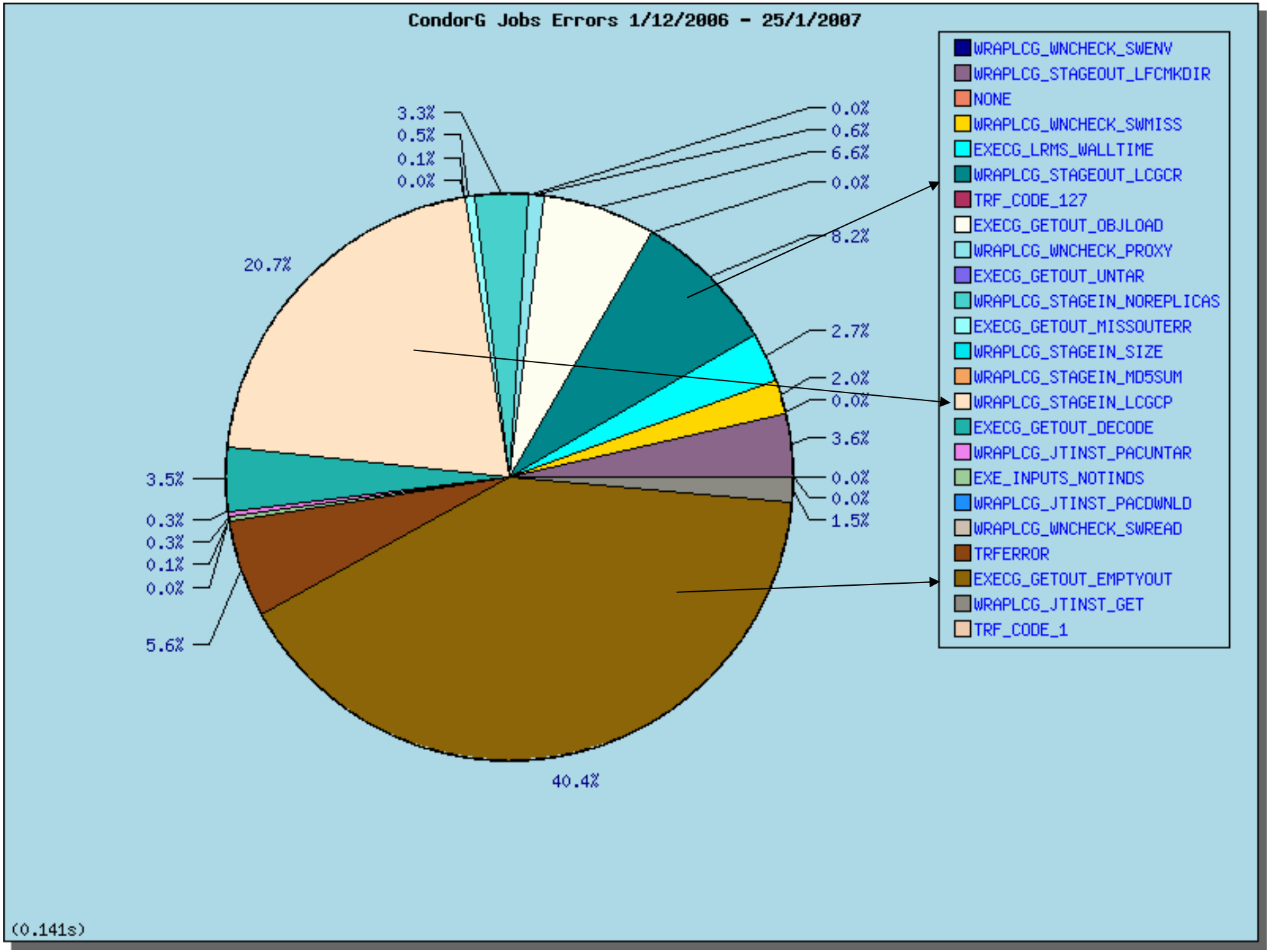
- Stage-in
  - Evgen replicated to all T1's and 2 copies on assigned cloud
    - 9 complete evgen datasets on T1's,T0 – never!
    - If SRM problem then can`t get file no matter how local you are.
    - Now look in all LFC's for all replicas and try them all, starting with the closest (domain/country)
    - Introduce dependency on all LFCs
      - Lfc api hang will hang all jobs
      - Several episodes but under control with timeouts
  - SRM problems dominate

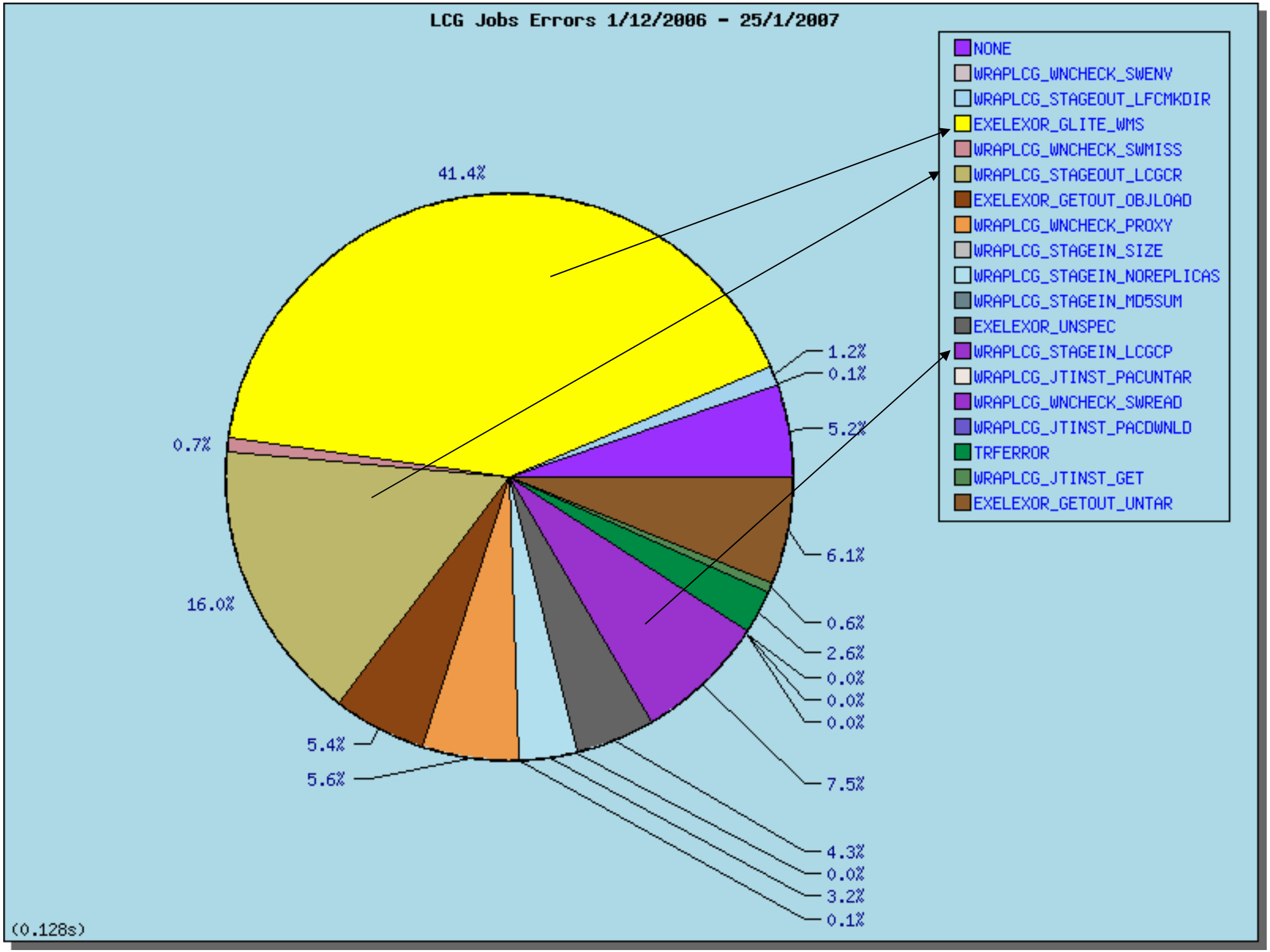
- Stage-in methods
  - Now we only use lcg-cp srm://..
    - srmcp similar
      - No BDII dependency but ok for srm2.2 compatibility?
  - Avoid srm
    - dcap,rfcp – bypasses srm throttling?
    - Or wait for srm2.2 to fix everything?
    - Cache files on NFS area – 1GB evgen used 50 times
    - Use Xrootd? Parrot?
  - Suggest: cache and wait for srm2.2 and in meantime improve site notification
    - Failures logged in prod Db

- Stage-out
  - lcg-cr to any SRM in assigned cloud
    - First to local SE, then T1, then any
  - T1 LFC is single failure point(rare)
  - ‘thin’ clouds with few SRMs have problems
    - T0, NL – DPM acl issue makes some thinner(IT)
  - if necessary add last resort failover to other T1
    - Spoils cloud model – rely on DDM more.

Errors

- Some effort to better categorize errors


CondorG Jobs Errors 1/12/2006 - 25/1/2007
WRAPLCG_WNCHECK_SWENV
WRAPLCG_STAGEOUT_LFCMKDIR
NONE
WRAPLCG_WNCHECK_SWMISS
EXECG_LRMS_WALLTIME
WRAPLCG_STAGEOUT_LCGCR
TRF_CODE_127
EXECG_GETOUT_OBJLOAD
WRAPLCG_WNCHECK_PROXY
EXECG_GETOUT_UNTAR
WRAPLCG_STAGEIN_NOREPLICAS
EXECG_GETOUT_MISSOUTERR
WRAPLCG_STAGEIN_SIZE
WRAPLCG_STAGEIN_MD5SUM
WRAPLCG_STAGEIN_LCGCP
EXECG_GETOUT_DECODE
WRAPLCG_JTINST_PACUNTAR
EXE_INPUTS_NOTINDS
WRAPLCG_JTINST_PACDWNLD
WRAPLCG_WNCHECK_SWREAD
TRFERROR
EXECG_GETOUT_EMPTYOUT
WRAPLCG_JTINST_GET
TRF_CODE_1
3.3%
0.5%
0.1%
0.0%
0.0%
0.6%
6.6%
0.0%
8.2%
20.7%
2.7%
2.0%
0.0%
3.6%
3.5%
0.0%
0.0%
1.5%
0.3%
0.3%
0.1%
0.0%
5.6%
40.4%
(0.141s)


LCG Jobs Errors 1/12/2006 - 25/1/2007
NONE
WRAPLCG_WNCHECK_SWENV
WRAPLCG_STAGEOUT_LFCMKDIR
EXELEXOR_GLITE_WMS
WRAPLCG_WNCHECK_SWMISS
WRAPLCG_STAGEOUT_LCGCR
EXELEXOR_GETOUT_OBJLOAD
WRAPLCG_WNCHECK_PROXY
WRAPLCG_STAGEIN_SIZE
WRAPLCG_STAGEIN_NOREPLICAS
WRAPLCG_STAGEIN_MD5SUM
EXELEXOR_UNSPEC
WRAPLCG_STAGEIN_LCGCP
WRAPLCG_JTINST_PACUNTAR
WRAPLCG_WNCHECK_SWREAD
WRAPLCG_JTINST_PACDWNLD
TRFERROR
WRAPLCG_JTINST_GET
EXELEXOR_GETOUT_UNTAR
41.4%
1.2%
0.1%
5.2%
0.7%
6.1%
16.0%
0.6%
2.6%
0.0%
0.0%
0.0%
5.4%
7.5%
5.6%
4.3%
0.0%
3.2%
0.1%
(0.128s)

- Immediate Grid job failures
  - Huge number of these – 40% both gLite & CG
    - Immediate(no cpu wasted) but annoying and messes up priorities. Stresses WMSs and executor.
  - Symptom is empty stdout and stderr
    - Nothing to work from
    - Seems to affect some sites more than others
      - Suspect CE load – switch “Production” to Busy” when load>N
    - Recent investigations by FR team found 1 reason
      - Stale ssh keys in user area
      - Could be other reasons – blackhole WNs
    - Need to let sites know stats – failure vs time
    - gLite CE will not have this problem(but others!)

Shift Organization

- Share workload and experience
  - run executor, check/report errors
  - check/repair data movement
    - Monitor subscriptions – overlap with DDM ops
    - Not really happening due to lack of instructions
      - lack of tools. Check multiple LFCs, LRCs
    - French model more successful
      - Cloud model for ops too
      - T2 cache clean-up
    - Clean-up is important to close datasets

Panda on EGEE

- Pilot removes WMS issues
- Input data is on local SRM
  - Unlike most US sites, no NFS access
    - Test hypothesis that proximity is secondary factor
  - More DDM operation load
    - Most sites have no ATLAS computing person
    - T2 cache turnover important (but also for non-panda)
  - Likely to cherry pick best sites but maybe no bad thing – result will be interesting whatever.

Conclusions

- DDM problems addressed in 0.3
  - bulk attach, performance
  - T2 cache turnover, catalogue/SRM consistency
- Stagein/out – SRM2.2 or bust?
  - investigate caching on NFS.
  - Xrootd maybe tested for analysis