The DESY GridLab

detailed evaluation of local protocols

D.Ozerov, DESY
for GridLab-team

04*05=20!!, HEPIX, Darmstadt

GridLab People

Y.Kemp, D.O. (DESY) setup, (help to) run tests

P.Fuhrmann(DESY/dCache.org) ideas, coordination, motivation, presentations

T.Mkrtchyan(DESY/dCache.org) pnfs client, (time)pressure

J. Elmsheuser (LMU, ATLAS) HammerCloud

H.Stadie (DESY, CMS) CMS test jobs

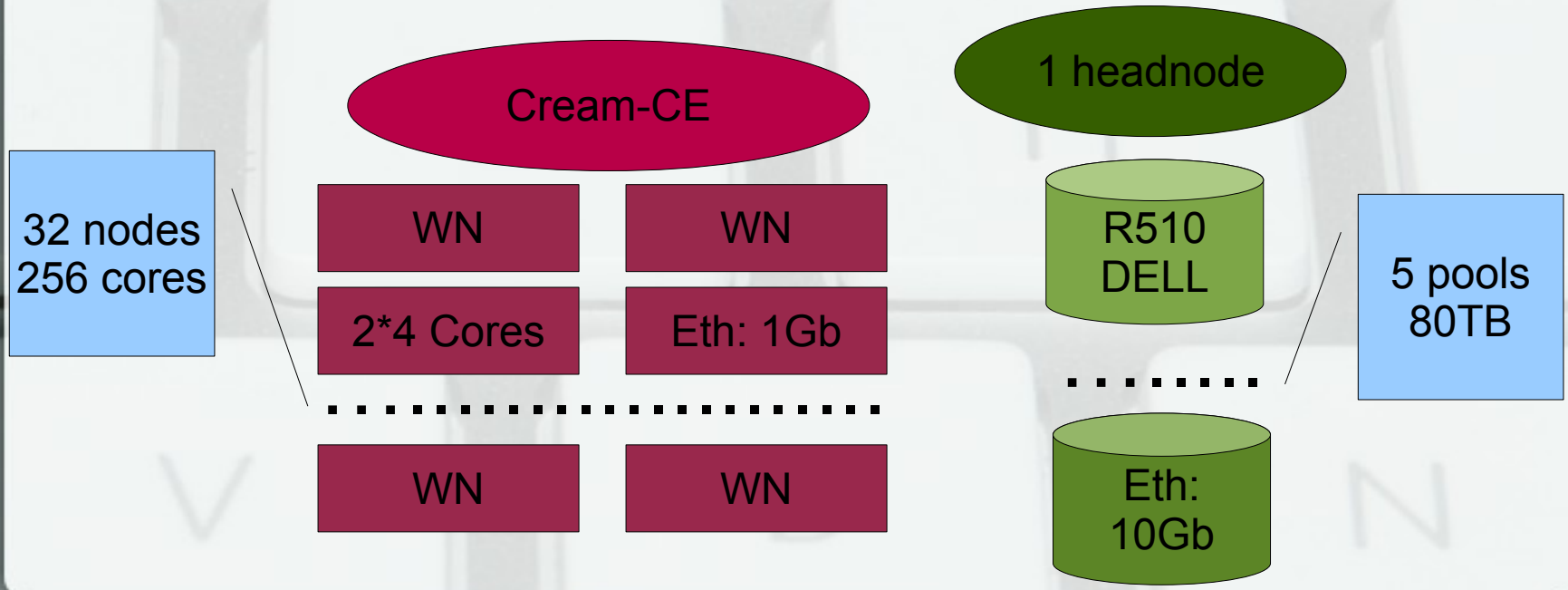
F.Legger (LMU, ATLAS) Real Atlas Analysis Jobs in HC

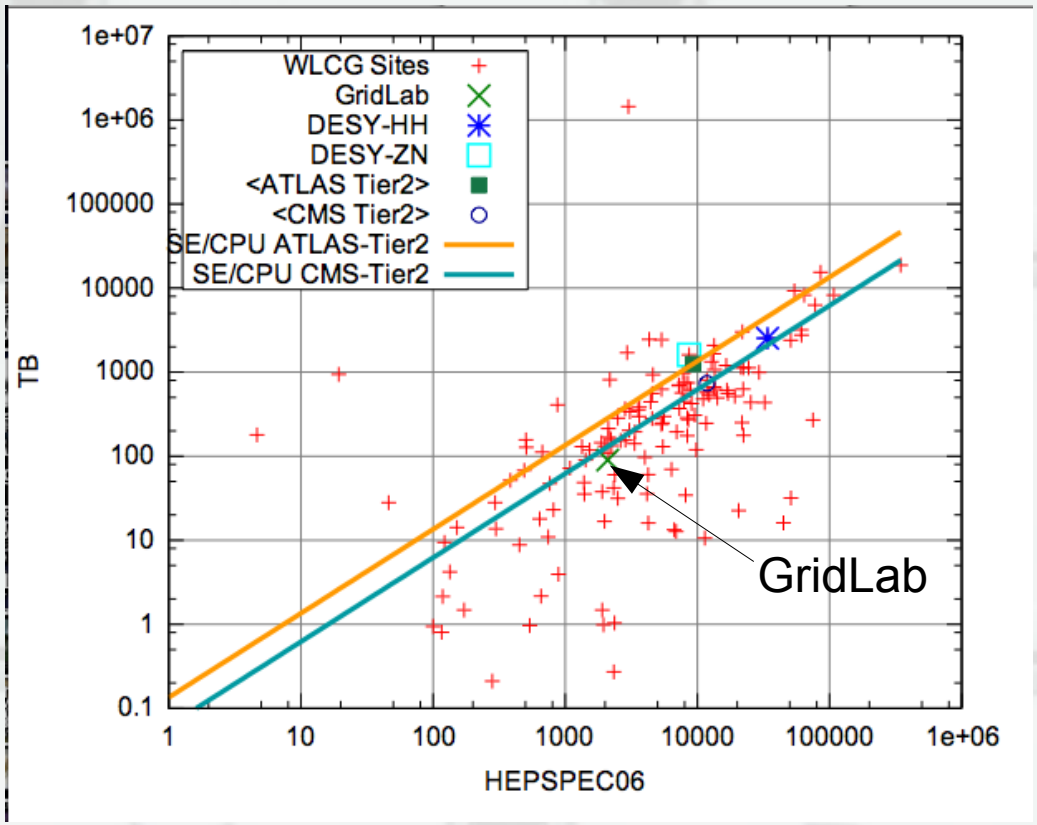
G.Behrmann,T.Zangerl (dCache.org) Enthusiasts for root:// in dcache

S.Kalinin (Wuppertal) First external tester with full access

GridLab hardware

- CPU: re-use old wn's, part of production grid
- Storage : current building blocks of DESY-HH grid





Status: April 2011, Gstat2.0
>160 WLCG Sites
GridLab in the middle of the list in CPU and Storage

Disclamer: obvious errors (>Exabyte Storage, too little CPU) are responsibilities of SiteAdmins to publish correct information

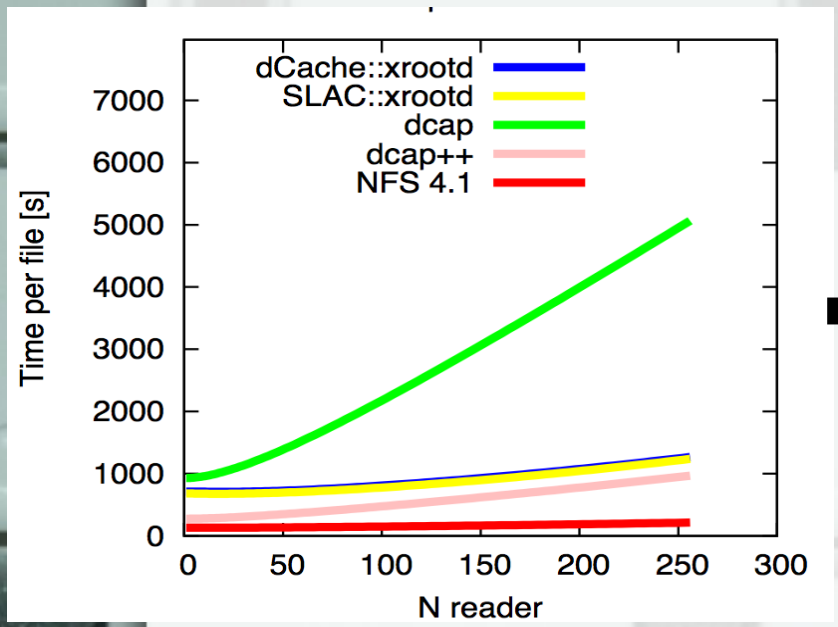
Tested Protocols

Protocol	Application Cache		Client cache
	no	TTreeCache	
Nfs4.1	V	V	OS
dcap	V	V	Vector read
dcap++	V	V	Smart block caching
Root:// (dcache,xrootd)	V	V	Vector read

Reading all info from file

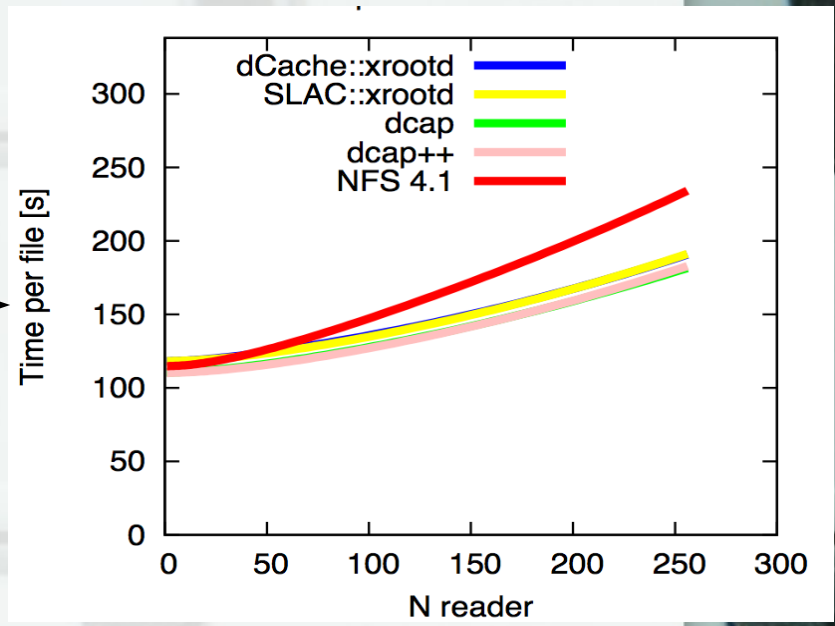
Ideal storage

Files read from memory on pool machines



Switch ON TTreeCache

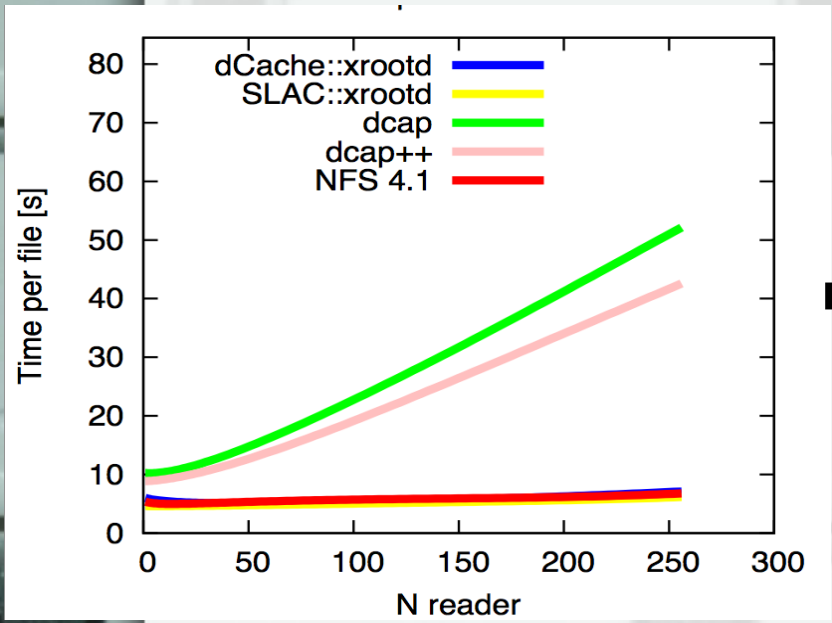
(event caching)



>10 times improvement with TTreeCache for slow protocol

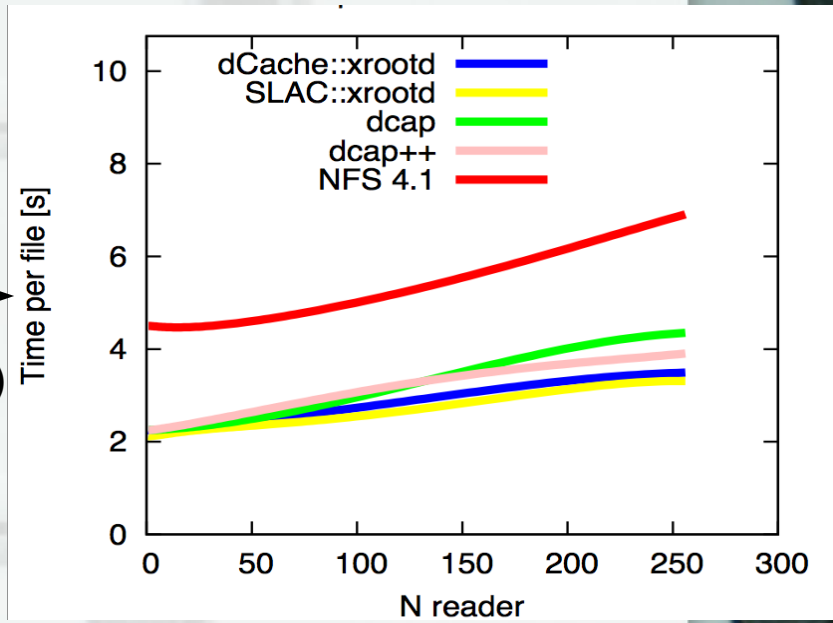
Reading only 2 branches

Ideal storage



Switch ON
TTreeCache

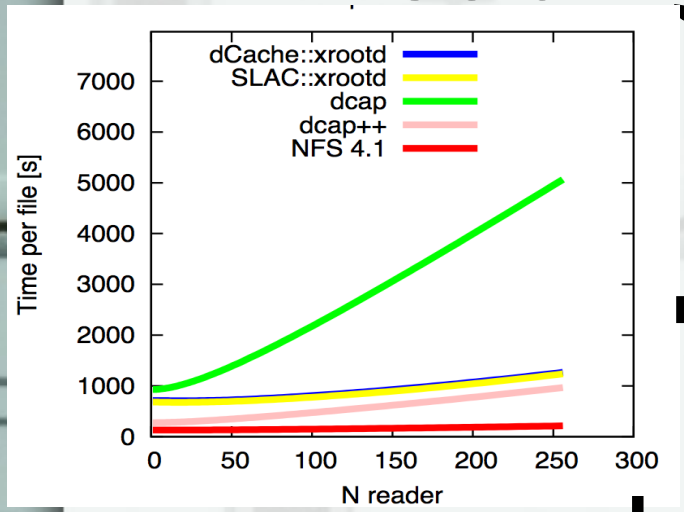
(event caching)



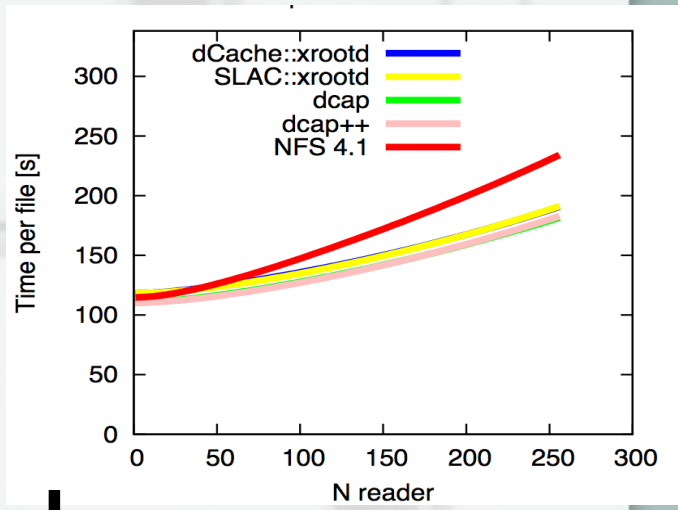
Similar improvement when reading only partial info from the file

Reading all info from file

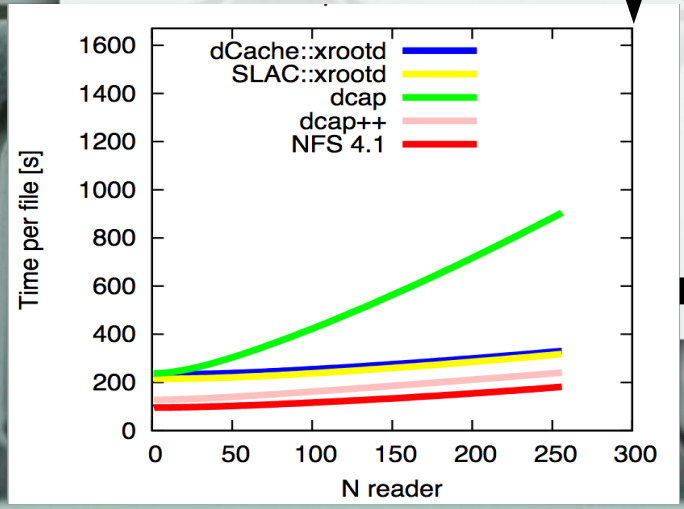
Ideal storage



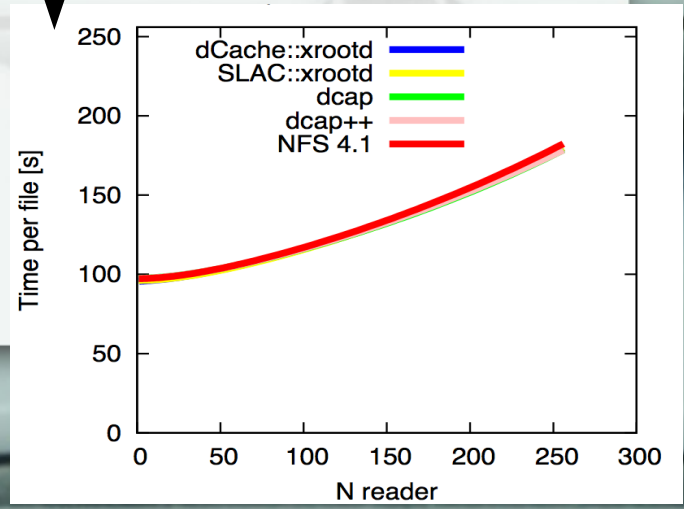
Switch ON TTreeCache



Re-write root files using optimised basket sizes

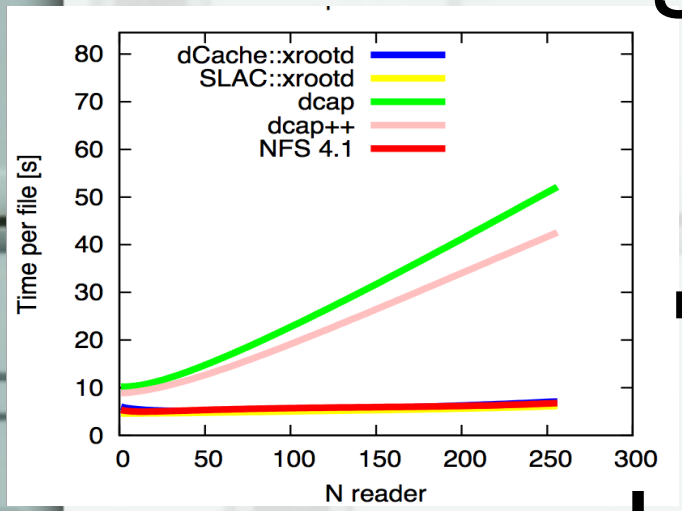


Switch ON TTreeCache

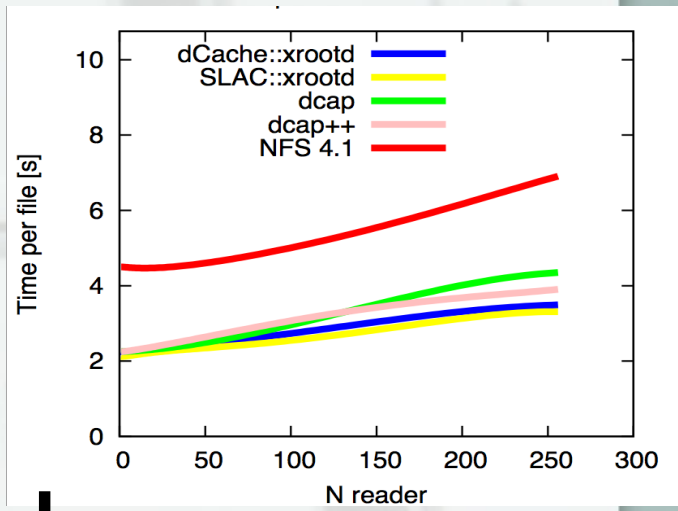


Reading only 2 branches

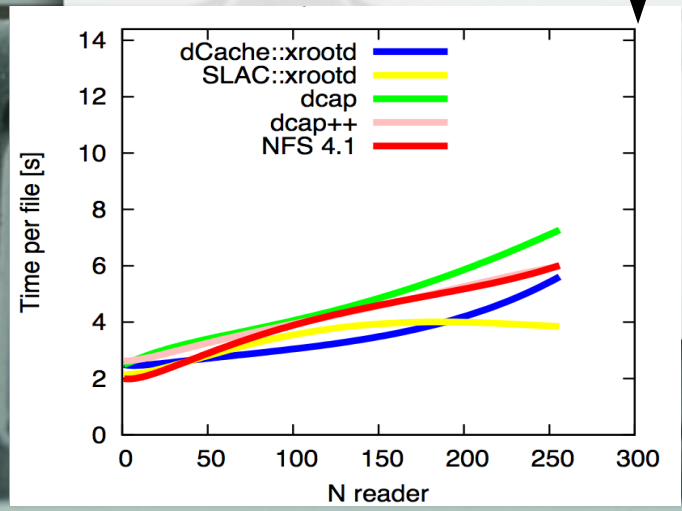
Ideal storage



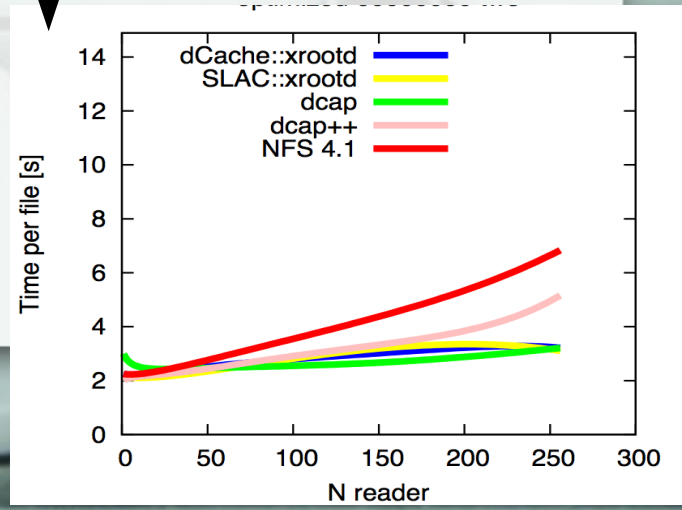
Switch ON TTreeCache



Re-write root files using optimised basket sizes



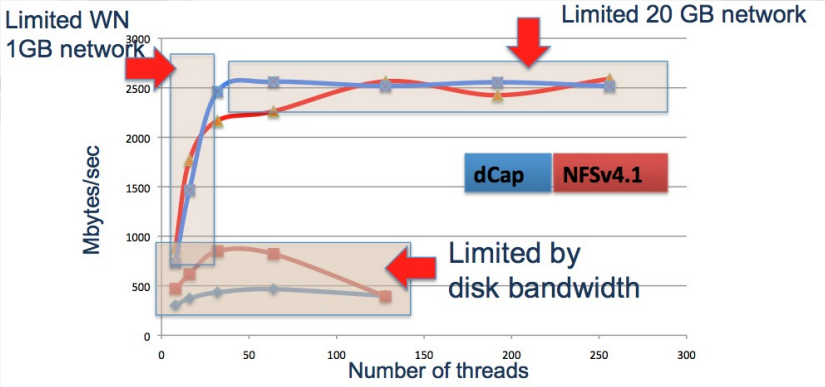
Switch ON TTreeCache



Performance depends on:

- File organisation
- Protocols used

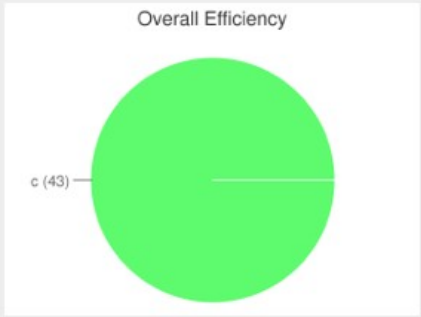
- ✔ root://dcache = root://xrootd
- ✔ Standard protocol (nfs4.1) is doing good job

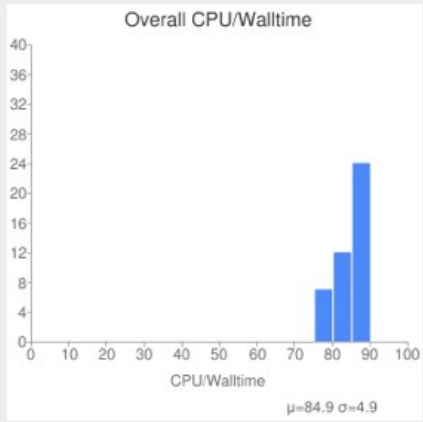


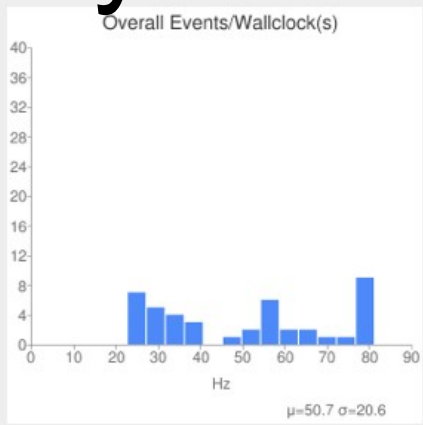
But difference between ideal and real storage is huge.
Look at reality

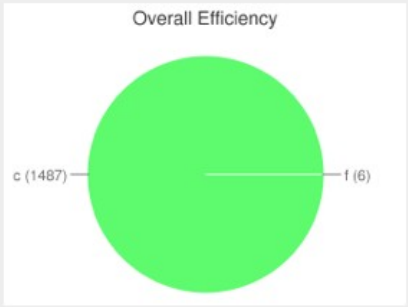
HC Standard AOD Analysis

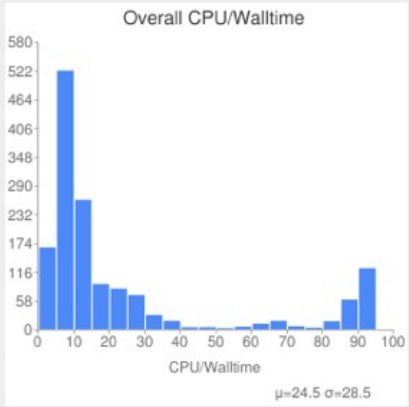
dcap
1 job
at any time

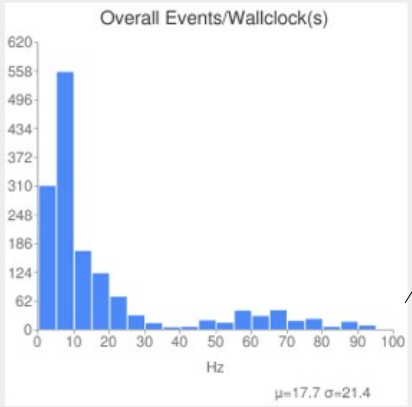






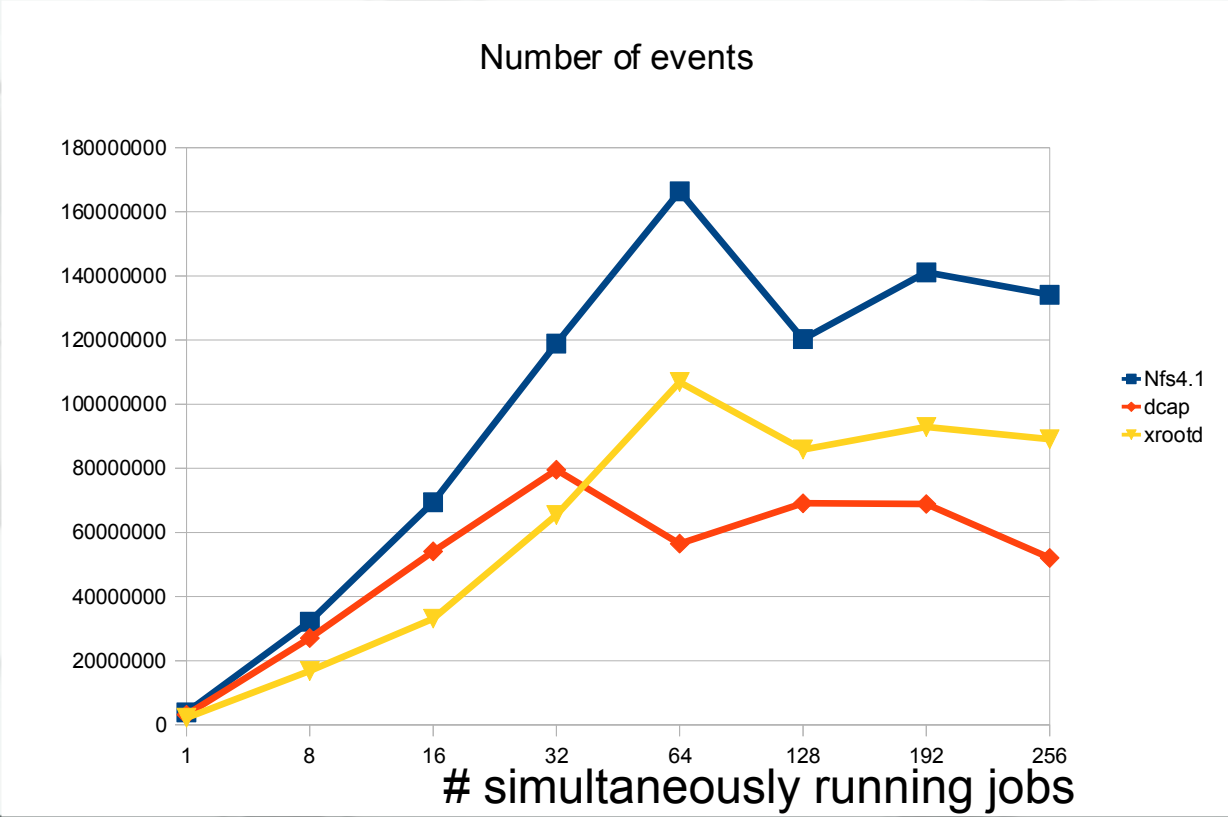




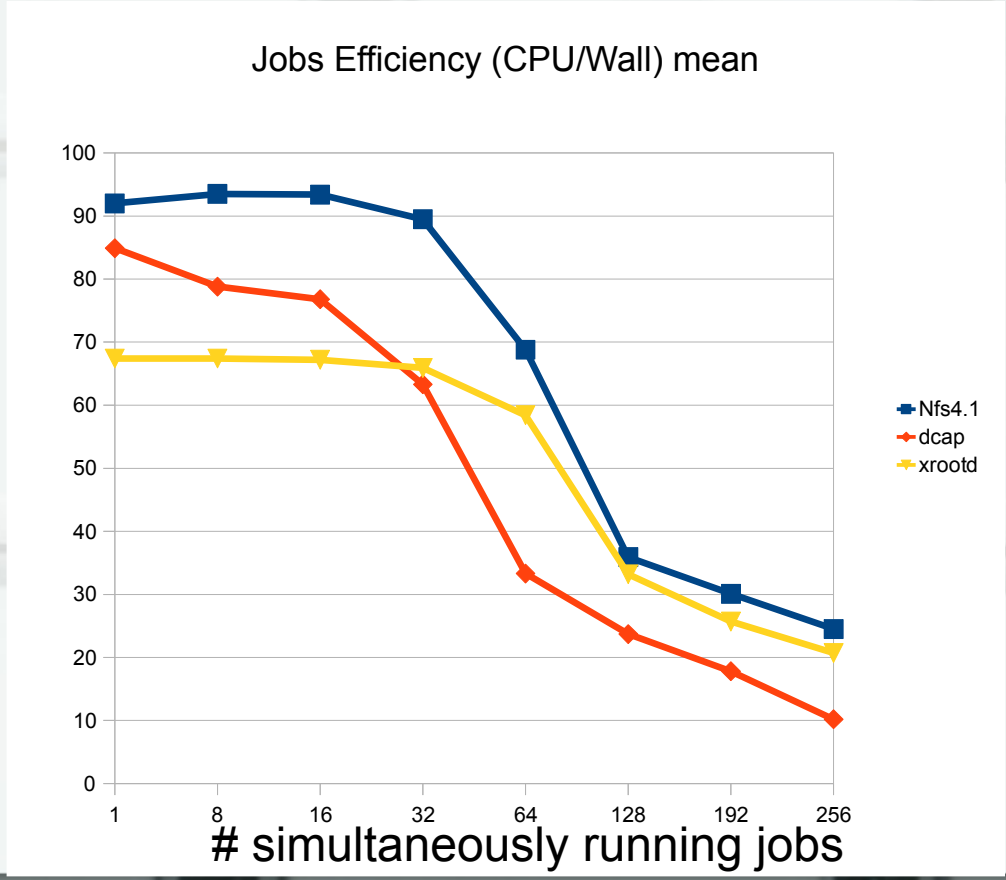


Nfs4.1
256 jobs
at any time

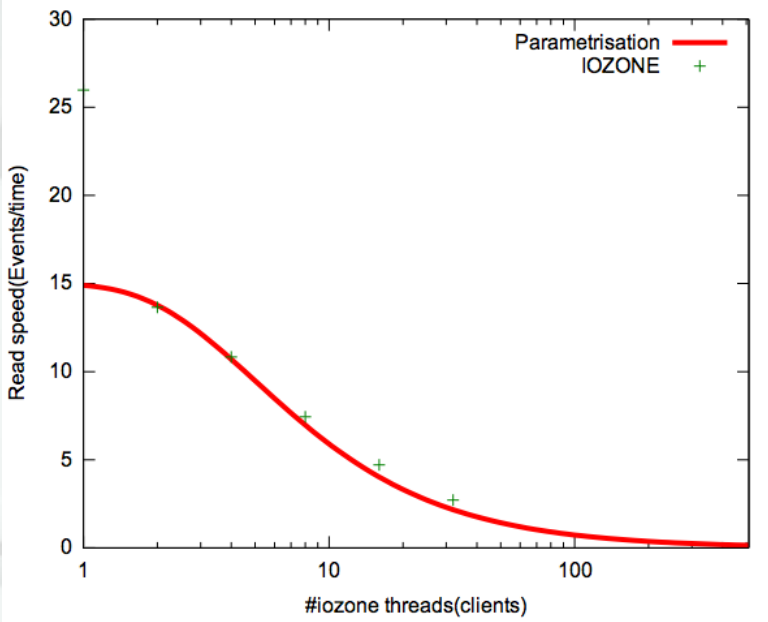
HC: Processed events in 24 hours (only finished jobs)

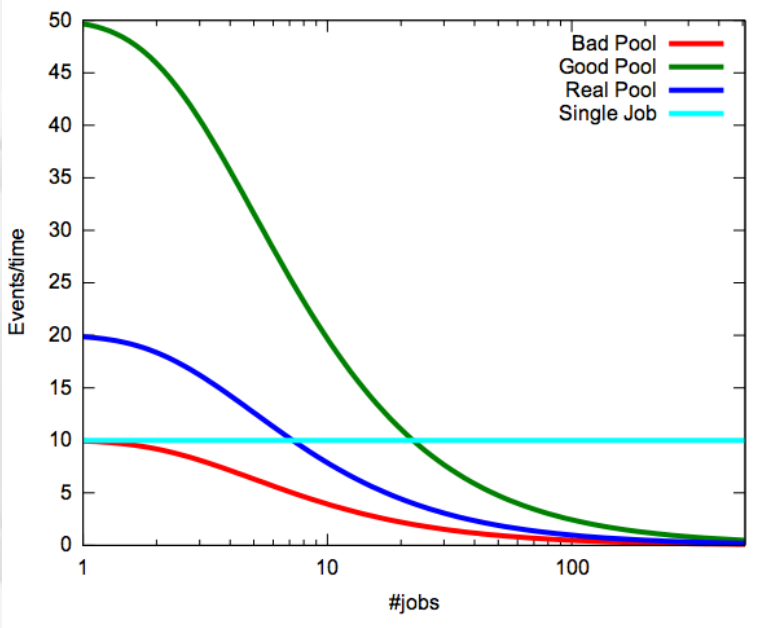


HC: jobs efficiency



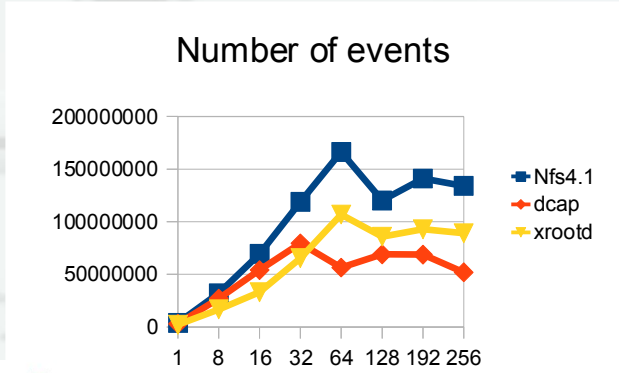
Model



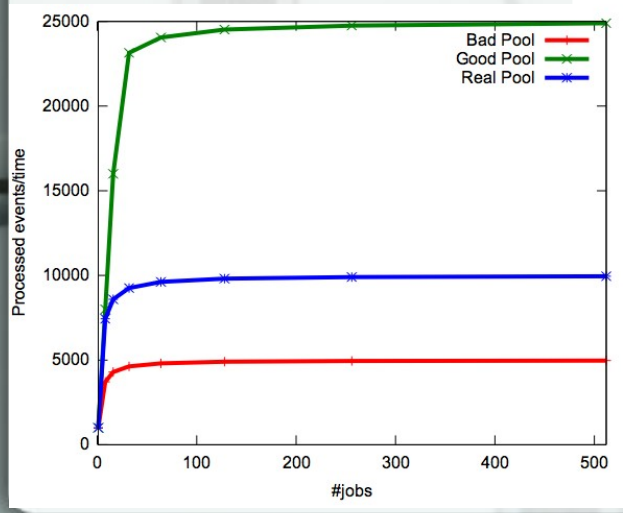


Input to model: job can read events with the limited speed (not more than)
Pool performance decreasing with number of clients
(current: is the iozone results for random reads)

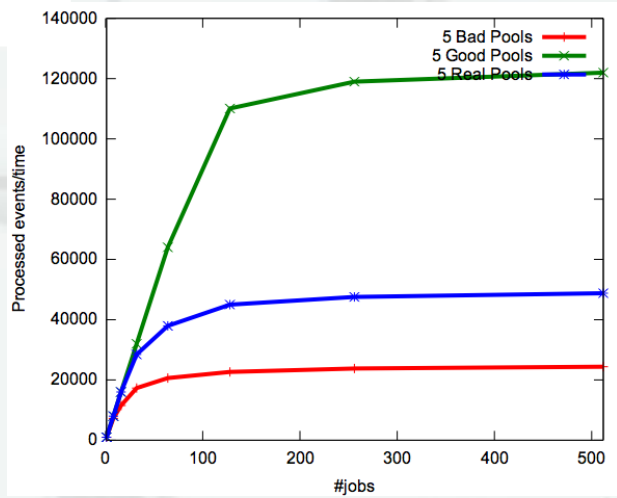
Processed Events



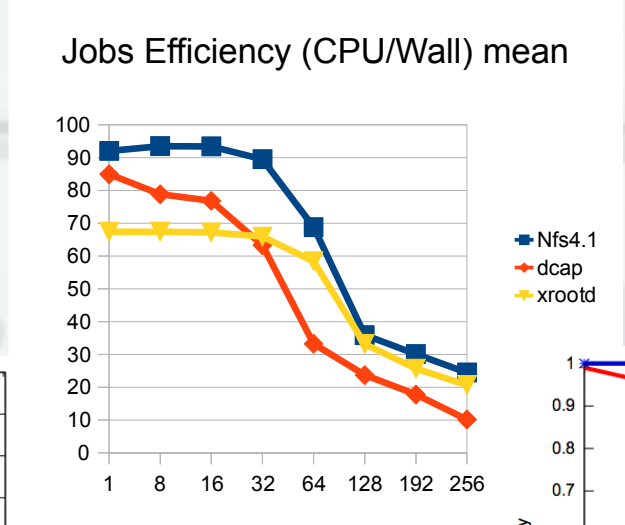
Once bottleneck is reached – it's contraproductive to send more jobs of the same type to the same storage

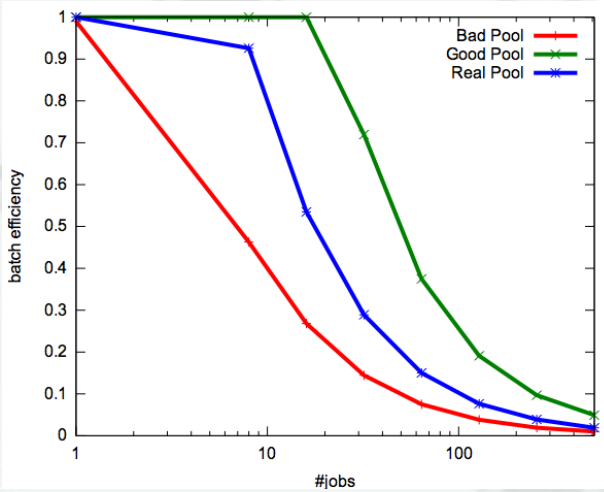


Shape reproduced



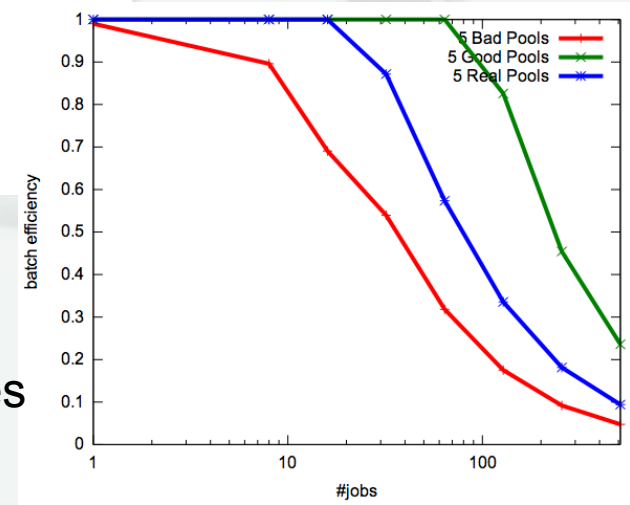
Jobs efficiency





Shape reproduced

Need realistic and free from atlas-services overhead example



Understanding Storage

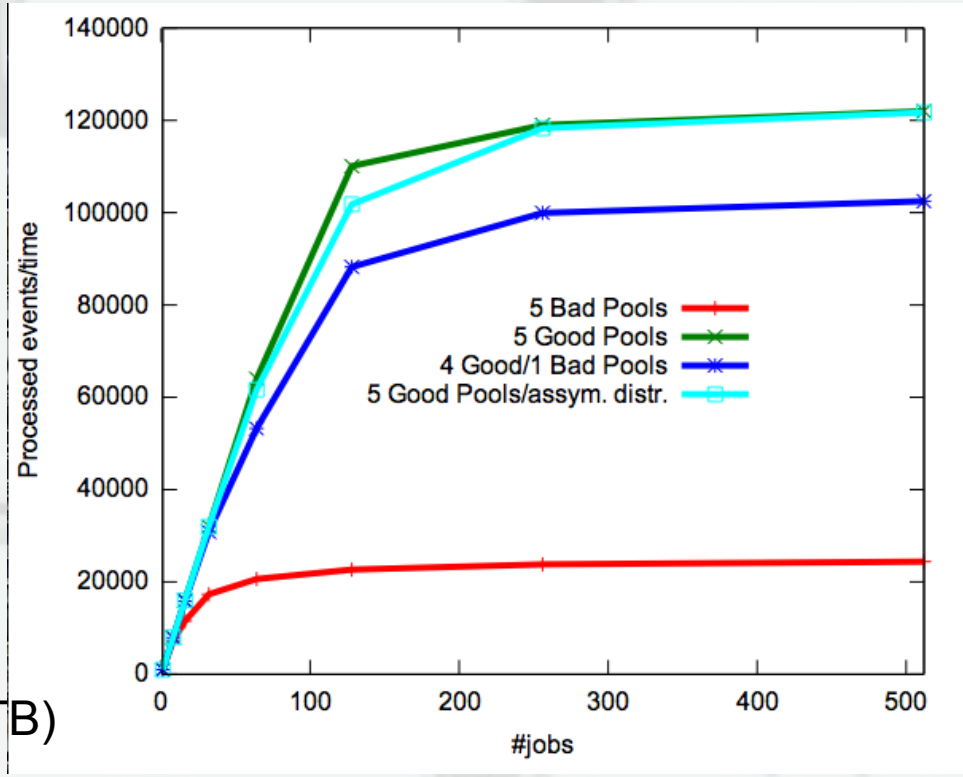
Config – N pools. What is worse to have:

1(few) bad(old) pools?

or

pools with non-uniform files distrbution?

Assymetric performance pool is bad. Compensate by the pool size (performance/TB)



GridLab activity

- NFS4.1 demonstrator (finished – NFS4.1 is the reality)
- Test of different hardware boxes
- CERNVM-fs test
- Testing dCache ideas (large testbed)
- Developing new HC tests (F.Legger)
- Open for your ideas – access via cream-ce
- …..

GridLab publications/presentations

- P.Fuhrmann, “NFS4.1 Initiative”, HEPIX'10 Cornell
- Y.Kemp et al. “NFS4.1 evaluation” CHEP2010 (presen.; publ.)
- G.Behrmann et al. “xrootd in dCache” CHEP2010 (poster; publ.)
- P.Fuhrmann, “EMI,dCache and standards”, LBNL seminar
- P.Fuhrmann, “Report on NFS4.1” GDB 01/2011
- P.Fuhrmann, EGI User Forum; dCache Workshop
- S.Kalinin, “xRootd”, dCache Workshop, Gottingen

Conclusion

- GridLab is a realistic size testbed
- NFS4.1 demonstrator activity finished (WAN test)
- Understanding the protocols need effort from both sides – storage provider and application developers
- NFS4.1 protocol is not worse than HEP made specific
- root://dcache = root://xrootd (confirming HEPiX StGr)
- More results with realistic analysis jobs