In collaboration with

NFS 4.1 Demonstrator Milestone 2 report

Presented by : Patrick Fuhrman

DPM information provided by Ricardo Rocha (CERN)

All the work has been done by (in alphabetic order)
Tanja Baranova (dCache.org)
Jean-Philippe Baud (CERN)
Johannes Elmsheuser (LMU Munich)
Yves Kemp (DESY)
Maarten Litmaath (CERN)
Tigran Mkrtchyan (dCache.org)
Dmitri Oserov (DESY)
Ricardo Rocha (CERN)
Andrea Sciaba (CERN)
Hartmut Stadie (DESY, CMS)

Content

- Check against my promises from London
- High level goals of milestone II (due next week).
- Status
  - Community
  - Development
  - Testing/Evaluation
  - Kernel
- Next steps
- Conclusion

Reminder : Why NFS 4.1

Why NFS 4.1 ?

- See "*11 reasons you should care*" by Gerd Behrmann @

  http://www.dcache.org/manuals/

- Properly defined standard
- Security part of the protocol
- Adopted by industry heavyweights : IBM, EMC , NetApp …
- We don't have to care about the client(s).
- Don't force sites to run WLCG specific storage software,

  which no other community can use at that site.



Milestone Review

- Setting up production like hardware
  - See later

OK

- Collecting a realistic 'test suite'
  - Basic file system (cat , cp …)
  - Atlas Hammercloud
  - CMS Analysis
  - ROOT tests (involving ROOT people)

OK

Milestones Review

- Running on regular NFS 4.1 implementation
  - DESY :
    - trying NetApp (January)
    - Pillar Systems (after visit in Nov)
  - CERN : after CHEP'10

IN PREPARATION

- NFS 4.1 dCache server
  - Build production size system
  - Stability tests (servers and client)
  - Performance testing

OK



Milestones Review

- NFS 4.1 DPM server
  - Build prototype system
  - Functionality tests

OK

- COMMON efforts
  - Wide area dCache/DPM — PARTIALLY
  - Security (Kerberos) — OK
  - Share test suites and test installations — NO
  - Presentations at CHEP (dCache/DPM) — NEXT WEEK

Goals of milestone II

- Finish development.
- Provide a NFS 4.1 enabled kernel for SL5.
- Provide production size test infrastructure.
- Setup test infrastructure for ‘realistic analysis’ use cases
  - ✓Regular file system I/O
  - ✓Experiment use cases
  - ✓ROOT expert use cases
- Prove stability on production level size and use-cases.
- First performance measurements.
- Attract more volunteers

Growing community

- StoRM : CNAF/INFN (Mirco, Riccardo) committed to join the NFS 4.1 group.
- dCache : PIC (Gerard) is successfully testing the dCache NFS 4.1 implementation and is providing valuable feedback.
- DPM : planning to cooperate with UK sites for larger scale testing.

Growing community

- CMS
  - Hartmut Stadie (DESY) : for CHEP'10 already
  - Official test-suite by Leonardo Sala will become available soon
- ATLAS
  - Hammer-cloud by Johannes Elmsheuser, and Daniel van der Ster.
- ROOT
  - Rene provided first version of test-suite

Status : development

PEOPLE

Task	People
Server	Tigran, Ricardo, Tanja, (Jean-Philippe)
Linux Kernel	Tigran, Ricardo
Testbeds	DESY : Yves Kemp, Dmitri Ozerov; CERN : Andrea Sciaba, Maarten Litmaath : PIC : Gerard Bernabeu (PIC)
Hammercloud	Johannes Elmsheuser, Daniel van der Ster
CMS Analysis	Hartmut Stadie (CMS, DESY)
ROOT support	Rene
Volunteer	Gerard (PIC), Andrea S., Maarten L.

Status : development

SERVER

Functionality	dCache (Tigran & Tanja)	DPM (Ricardo R.)
NFS 4.1 (sessions, gssapi)	DONE	DONE
pNFS	DONE	Finalizing
Kerberos	DONE	DONE
X509	Missing	Missing

Client (Linux Kernel)

Tigran cooked a 2.6.36_rc3 kernel with a NFS4.1 and pNFS driver for SL5 plus an RPM with modified ‘mount’ tools.

http://www.dcache.org/chimera/x86_64/

Status : evaluation (dCache only)

Hardware setup @ DESY

Storage and CPU Power

Amount	Type	CPU Xenon®	RAM GB	Cores	Network GBit	Disk TBytes
1	Headnode	5160@3.00GHz	8	4	1	
5	Pool	5520@2.27GHz	12	16	10	12 * 2
16 or 32	Workernode	5150@2.66GHz	4	8	1	

Network

Software : CREAM CE; dCache 1.9.10

Setup :

- About 50 % CPU Tier II
- About 20 % Storage Tier II

Status : evaluation (dCache only)

Stability

Main message of this milestone.

- CFEL Production Transfers from SLAC to DESY
  - 13 TBytes over 10 days
  - 100 GBytes average file size
  - No crash
- Un-taring Linux Kernel into NFS 4.1
  - No crash
- High-latency test
  - Recursive 'ls –l' over 60.000 files via DSL from home.
  - Finished w/o problem.
- 128 Processes writing into the same file
  - Client nodes get stuck
  - Server was still ok

Status : evaluation (dCache only)

Please kindly notice :

The NFS 4.1 evaluation for dCache, described subsequently, has been done by Yves Kemp and Dmitri Ozerov from DESY with great support from Johannes Elmsheuser, Daniel van der Star and Hartmut Stadie, as a preparation for their presentation at CHEP'10.

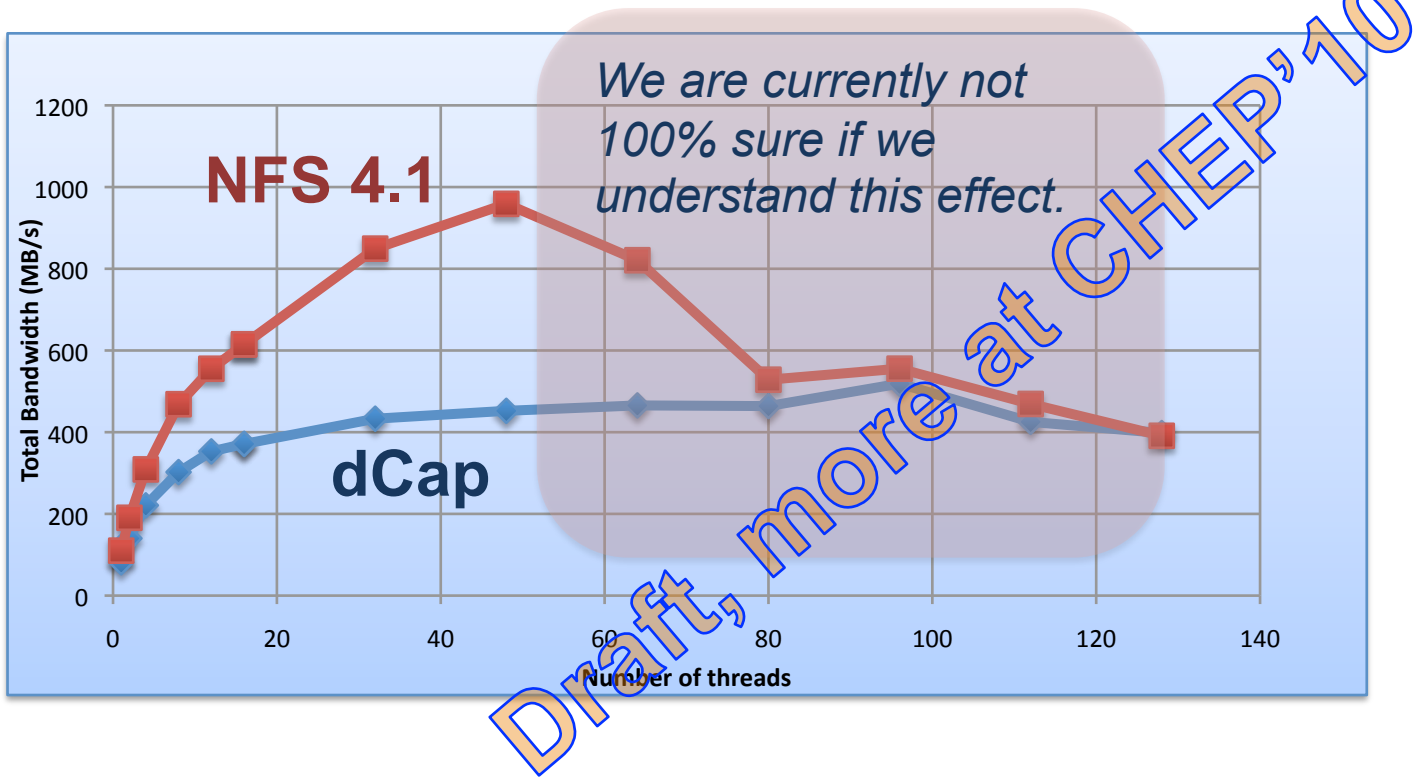
With the following slides, I'm presenting what has been done, NOT the results themselves.

Please make yourself familiar with the original work, either by visiting their talk at CHEP or by getting the publication afterwards.

Status : evaluation (dCache only)

Simple I/O

- Simple ‘cat .. >/dev/null’ or dccp to /dev/null.
- Files with ‘random’ numbers. So, no compression effects.
- No caching (read once)



Status : evaluation (dCache only)

Hammercloud (just one typical example)

Problems, which had to be solved first :

- CE is 'hidden'
- Files are not in catalogue

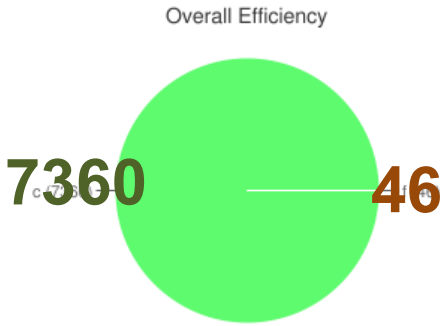
- 8248 jobs
- Cancelled after 4 days.

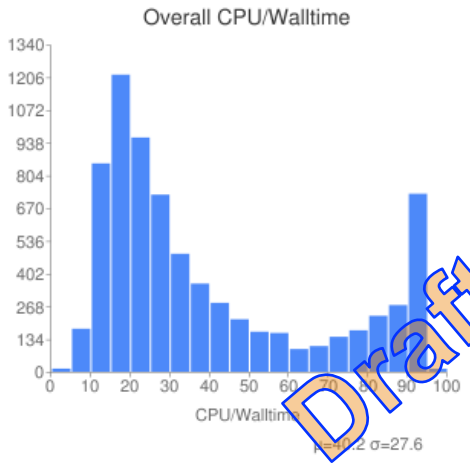
Input type: DQ2_LOCAL
Output DS: user.elmsheus.hc.10001206.*
Input DS Patterns: /data/hammercloud/atlas/inputfiles/tier3pfns/r12/
Ganga Job Template: /data/hammercloud/atlas/inputfiles/muon1566/muon1566_cream_tier3_special1.tpl
Athena User Area: /data/hammercloud/atlas/inputfiles/muon1566/MuonTriggerAnalysis_1566.tar.gz
Athena Option file: /data/hammercloud/atlas/inputfiles/muon1566/MuonTriggerAnalysis_1566.py
Test Template: 45 (stress) - Muon 15.6.6 CREAM Tier3 DESY-HH dCache Test

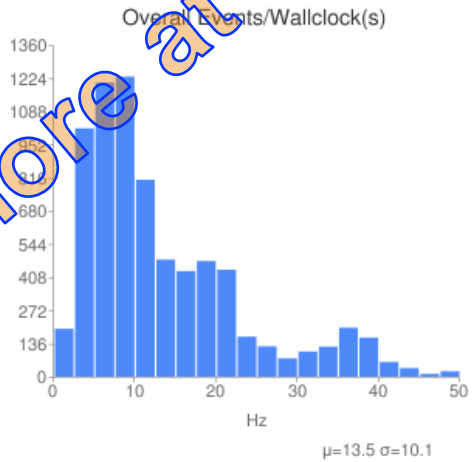
Setup

CPU = 50% Tier 2
Storage = 20% Tier 2

We expect results to be even better if
Storage ≈ CPU





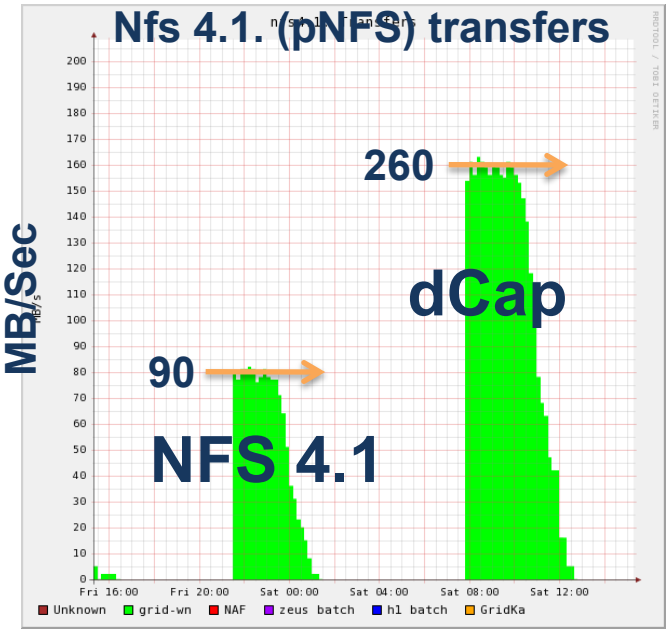


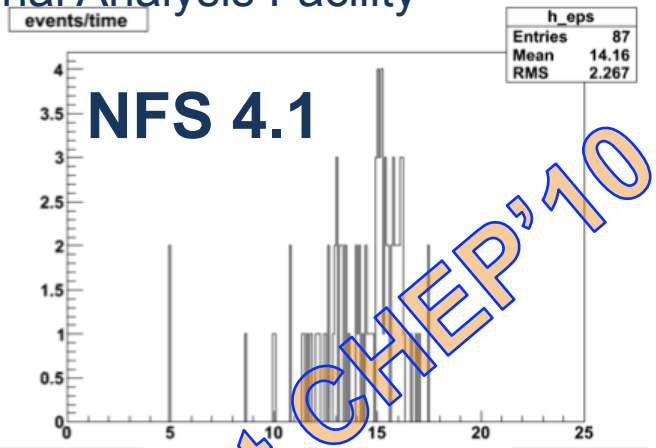
Draft, more at CHEP'10

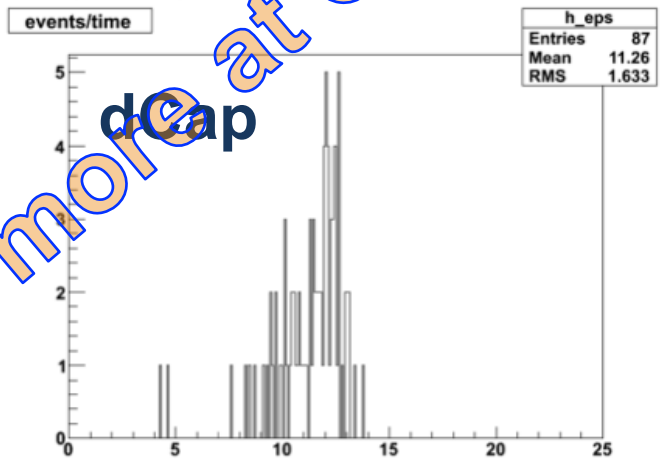
Status : evaluation (dCache only)

CMS analysis One job per core / 32 cores

- Muon dataset : 1.7 TB in 308 Files.
- RECO files, test is stripping a PAT Ntuple out of the CMSSW framework.
- Most prominent use-case on the DESY National Analysis Facility
- Not much CPU, nearly only I/O







Draft, more at CHEP'10

Status : evaluation (dCache only)

Artificial ROOT tests

Results not yet available (hopefully at CHEP'10)

Kernel availability

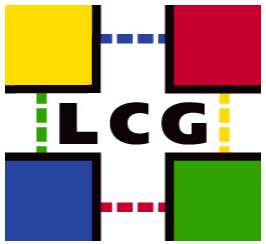
- Kernel used for evaluation : 2.6.36_rc3
- NFS 4.1 (pNFS) kernels expected in SL6.(>2)
- 2.6.36 back-port to SL5 available from DESY
  - Plus ‘mount tools’ RPM.
  - Kernel will very likely not cover all hardware setups.
- With a Joined Effort (e.g. CERN, FNAL, DESY), we would be able to provide an SL5 with NFS 4.1 (pNFS) kernel within months. (If we really want)


Next Steps

- More details at CHEP'10 by Yves and Dmitri.
- More investigation with various different ROOT setups.
- Working with the CMS official test-case.
- Investigating X509 Certificate/Proxy security.
- Wide area transfer evaluation.
- Setting up a regular NFS 4.1 (pNFS) system e.g. : NetApp and Pillar.
- Evaluation by the HEPIX working group.
- Trying to find groups as guinea-pigs for NFS4.1 production.

Conclusions

- Well in time with the “Demonstrator process”
- Stability is much better than expected : Production ready.
- Kernel situation : short term solution for SL5 would be available, if we want.
- Performance already comparable with existing solutions.
- Nevertheless : more evaluation on ROOT framework interaction needed.
- Efforts will continue within the EMI/dCache.org framework.
- You want to volunteer ?
  - Get dCache 1.9.10 from dCache.org
  - Get nfs enabled kernel : http://www.dcache.org/chimera/x86_64/



To stay in touch, please visit :

https://twiki.cern.ch/twiki/bin/view/EMI/EmiJra1DataDetailsNFS41

Thank you

EMI is partially funded by the European Commission under Grant Agreement INFSO-RI-261611