HEPiX FSWG Progress Report

Andrei Maslennikov

Michel Jouvin

GDB, May 2nd, 2007

Outline

- Mandate and work plan
- First achievements
- Next milestones

Credits

- This presentation is based on Andrei Maslenikov's presentation during last HEPiX meeting (DESY, April 25th)
- Full (detailed) presentation available at https://indico.desy.de/materialDisplay.py?contribId=53&sessionId=39&materialId=slides&confId=257

WG Mandate

- The group was commissioned by IHEPCCC in the end of 2006 under umbrella of HEPiX
  - Chairman : Andrei Maslenikov
- Officially supported by the HEP IT managers
- The goal is to review the available file system solutions and storage access methods, and to divulge the know-how among HEP organizations and beyond
  - Not focused on grid
- Timescale : Feb 2007 – April 2008
- Milestones: 2 progress reports (HEPiX Spring 2007, Fall 2007), 1 final report (HEPiX Spring 2008)

Members

- Currently we have 20 people on the list, but only these 15 appeared in the meetings/conf. calls and did something since the group had started:

BNL	R.Petkus
CASPUR	A.Maslennikov (Chair), M.Calori (Web Master)
CEA	J-C.Lafoucriere
CERN	B.Panzer-Steindel
DESY	M.Gasthuber, P.van der Reest
FZK	J.van Wezel, S.Meier
IN2P3	L.Tortay
INFN	V. Sapunenko
LAL	M.Jouvin
NERSC/LBL	C.Whitney
RZG	H.Reuter
U.Edinburgh	G.A.Cowan

- These very people maintain contacts with several other important labs like LLNL, SLAC, JLAB, DKRZ, PNL and others.

Work Plan

- The work plan for the group was discussed and agreed upon during the first two meetings. Accent will be made on shared / distributed file systems.
- We start with an Assessment of the existing file system / data access solutions; at this stage we will be trying to classify the storage use cases
- Next, in the course of the Analysis stage we will try get a better idea of the requirements for each of the classes defined during the previous stage
- This will be followed by the selection of the viable Candidate Solutions for each of the storage classes, followed by a possible Evaluation of some of them on the common hardware
- Then the Final Report with conclusions and practical recommendations will be due, by the Spring 2008 HEPiX meeting

Assessment progress

- Prepared an online questionnaire on deployed file stores
  - http://hepix.caspur.it/storage/questionnaire1.php (hepix/hepix)
- Selected 21 important sites to be covered: Tier-0, all Tier-1 plus several large labs/orgs like CEA, LLNL, DKRZ
- All selected sites were invited to fill the questionnaire for their most relevant file store solutions; at least two areas had to be covered: home directories and the largest available shared filestore

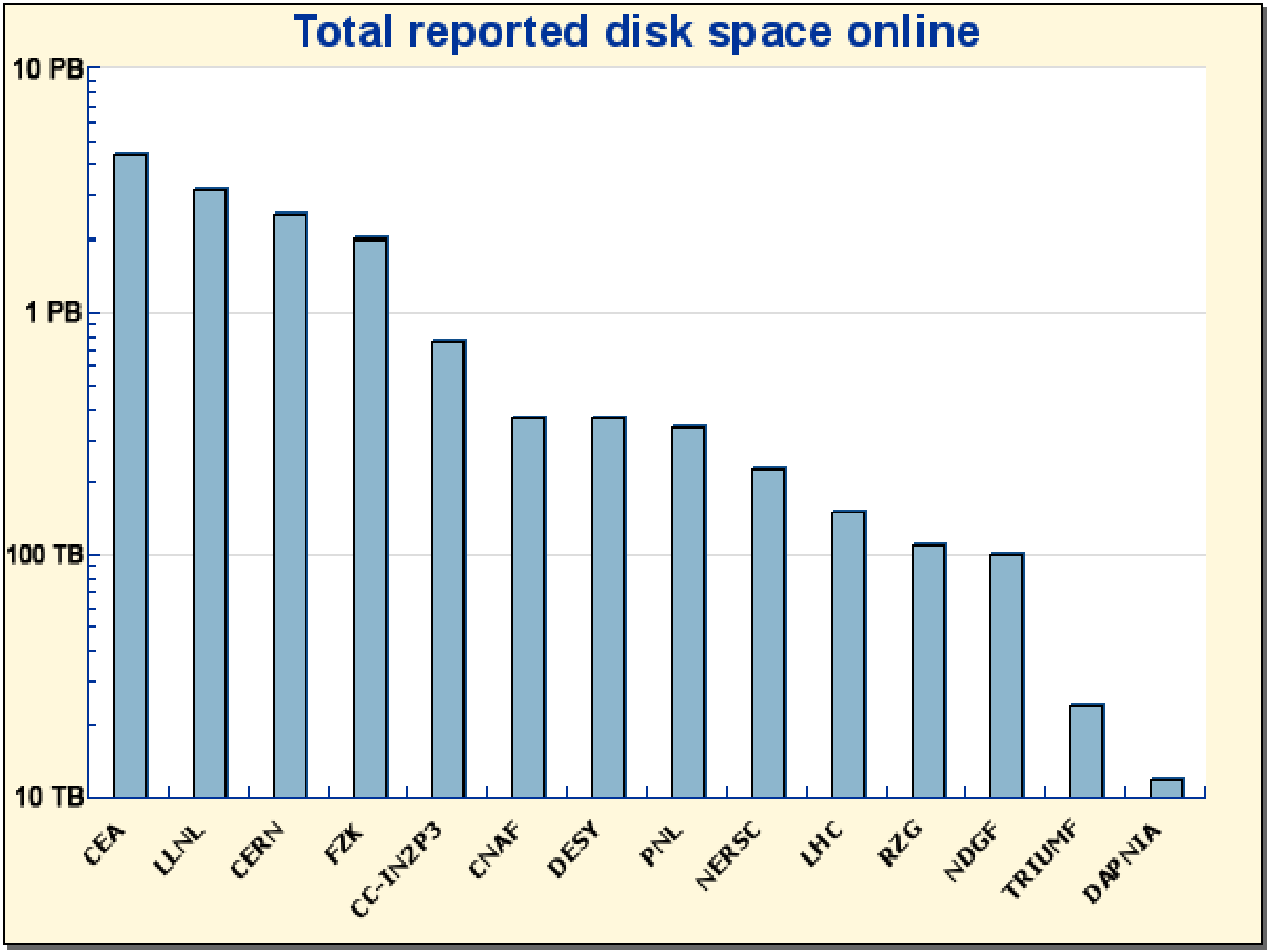
Sites under assessment

- ASGC
- BNL
- CC-IN2P3
- CEA
- CERN
- CNAF
- DAPNIA
- DESY
- DKRZ
- FNAL
- FZK
- JLAB
- LLNL
- NERSC
- Netherlands LHC
- NDGF
- PIC
- PNL
- RAL
- RZG
- SLAC
- TRIUMF

- - Collected / being verified
- - Being collected
- - NO INFO / NO CONTACT

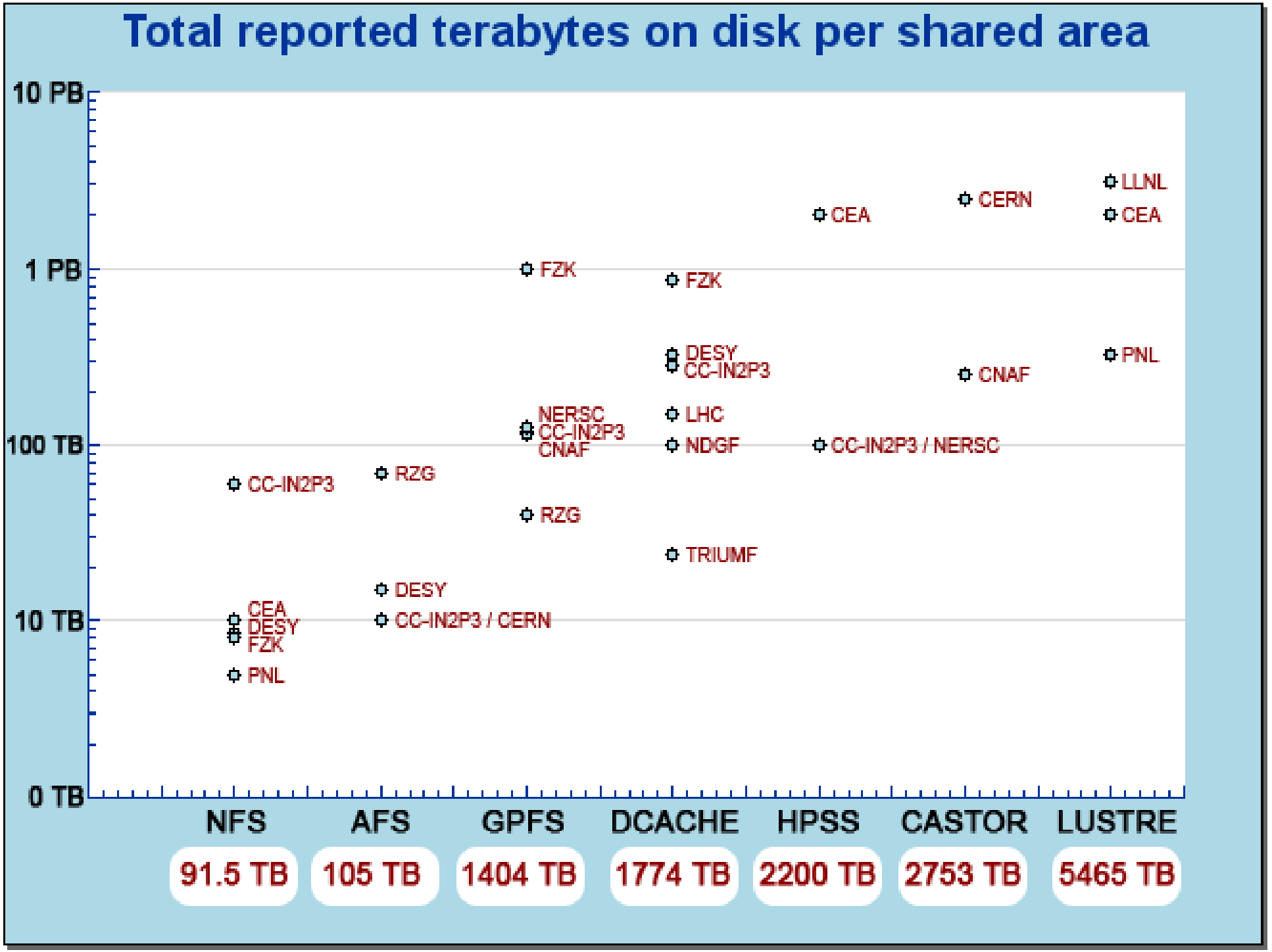
Initial Observations

- The big picture looks a bit chaotic, the reasons to choose this or that storage access platform are often not clear
- The data collected are yet to be verified! Some of the numbers provided by the local “info collectors” appear to be unprecise.
- Moreover, in several cases some fields of the questionnaire were interpreted in different ways by different info collectors. We hence scheduled an effort to clean this up (“normalize”), and to see if the questions should be made in a better form.
- So far we were only able to make a pair of intermediate plots on the basis of data collected over 14 sites out of planned 21, but already these partial infos could tell us something. We only looked at the online disk areas. The slow tier (tape backend) has to be studied seprately, and we still miss plenty of data.
- The total area size online reported is large but may not be called very impressive: 13.7 PB over all sites including the non-HEP organizations (compare with the planned 12-14 PB/year for LHC production).


Total reported disk space online
10 PB
1 PB
100 TB
10 TB
CEA
LLNL
CERN
FZK
CC-IN2P3
CNAF
DESY
PNL
NERSC
LHC
RZG
NDGF
TRIUMF
DAPNIA

File systems / data access solutions in use

- Please note that we are still *very* far away from any conclusions!
- However, here are some facts and thoughts:
  - The large initial list of candidate solutions may probably be reduced to just 7 names: Lustre, GPFS, HPSS, CASTOR, dCache, AFS and NFS
  - AFS and NFS are mostly used for home directories and software repositories and remain very popular
  - Solutions with the HSM function (HPSS, CASTOR and dCache) have similar deployed base in petabytes, and probably have to be compared
  - GPFS and Lustre dominate in the field of distributed file systems and deserve to be compared
  - Lustre has the largest reported installed base (5.5 PB), but not a single HEP organization had ever deployed it !
  - dCache is present in many HEP sites, however CASTOR alone stores more data than all reported dCache areas (NB: we miss data from FNAL)


Total reported terabytes on disk per shared area
10 PB
1 PB
100 TB
10 TB
0 TB
CC-IN2P3
CEA
DESY
FZK
PNL
RZG
DESY
CC-IN2P3 / CERN
FZK
NERSC
CC-IN2P3
CNAF
RZG
DESY
CC-IN2P3
FZK
LHC
NDGF
TRIUMF
CEA
CC-IN2P3 / NERSC
CERN
CNAF
LLNL
CEA
PNL
NFS
AFS
GPFS
DCACHE
HPSS
CASTOR
LUSTRE
91.5 TB
105 TB
1404 TB
1774 TB
2200 TB
2753 TB
5465 TB

Tentative plan until September 2007

- Continue with the data collection and analysis
  - Complete the questionnaire by the Fall 2007
  - Report during the meeting at St Louis
- Reduce the list of solutions to 7 names and create three mini task forces:
  - On home directories / software repositories: AFS, NFS, GPFS(?)
  - On data access solutions with the tape backend: CASTOR, dCache, HPSS
  - On scalable high performance distributed file systems: Lustre, GPFS
- Each of the task forces will have a goal to prepare an exhaustive collection of documentation on the corresponding solutions, describe best practices, provide deployment advice, cost estimates and performance benchmarks
  - All task forces will have to present an interim progress report during the St Louis meeting

Some input for discussion

- This workgroup is open to all sites (HEP- and non-) interested in the storage issues. We appreciate any feedback and would welcome any new active members
- We appeal to FNAL to join us actively (or at least to provide their data on storage, otherwise our report will not be complete)
- Our web site (http://hepix.caspur.it/storage) is open to all universities, research labs and organizations. The access to it is protected by a symbolic password which was widely circulated among HEPiX members and may at any time be obtained via mail. Just send your request to *monica.calori at caspur.it*