Disk and File Systems

12 Mars 2002

Philippe GAILLARDON
IN2P3 Data Center

Introduction

- This talk is based on our actual experience of Mass Storage.
- Its aim is not to provide a definitive solution for the LHC but to outline the key points we are facing.

Disk and File Systems

- Disk
  - Disk Only?
  - Hierarchical (Disk/Tapes)?
- File Systems
  - One UNIX File System?
  - One Name Space?

Disk versus Hierarchical

- Choice of hierarchical system at IN2P3
  - Cost
  - Volume availability
  - May lead to disastrous performances
- Common points
  - Sharing with many user hosts (hundreds to thousands)
  - Sharing with many servers
    - Static or dynamic access to several servers
    - Disk (and tape) drives must cooperate: Fiber Channel solutions look promising
    - The user network shall have increased performances

File System versus Name Space System

- File System
  - Solutions based on a unique File System can't be imagined for Pbytes volumes (recovery, performances….)
  - File System can be only a simulated file system
- Name space system
  - The adressability is at file level
  - The access must be as transparent as possible for the user applications
- Many solutions exist in the HEP community

IN2P3 today's solution

- HSM solution based on HPSS
  - Disk and Tape Hierarchy
    - BABAR Objectivity: 65 TB / 20 TB disk out of HPSS
    - Others: 45 TB / 1.7 TB disk
    - Total: 110 TB
  - One File-Name Space
- with RFIO Access
  - Developed RFIO 64 bits with CERN
- Function very close to CASTOR

Which experiments?

- BABAR
  - Objectivity: 65 TB / 20 TB on disk
  - Other (of which analysis and user space): 1.5 TB
- Astrophysics
  - EROS: 9 TB
  - AUGER: 16.5 TB
- LHC
  - CMS (2TB), ATLAS (1.8TB), LHCB (1.6TB), ALICE (1.3TB)
- Other
  - D0 (5.7 TB), VIRGO (0.9 TB)…

Performances

- Dynamic data path between user host and the right data server
  - Best achieved by RFIO/HPSS readlist/writelist (10 MB/s)
  - RFIO streaming mode has good results (5 to 8 MB/s)
  - The basic read/write performances are poor (1 to 2 MB/s)
- Several RFIO servers bound to an HPSS disk server
  - Static server name resolution at the moment

Miscellaneous topics

- Access rights
  - The UNIX-style permissions are inadequate
- Identity
  - All is based on uid/gid. This seems difficult to change.
- Quality of Service
  - It’s achieved thru the COS (Class of Service) for HPSS. It differs from other implementations based on directory tree.
- Statistics
  - The statistics provided by HPSS and RFIO are insufficient

User view of the Mass Storage

- Have a user data base (or book keeping)
  - Cost of search operations
    - High for searching in large tree directory
    - Inadequate/prohibitive for seeking in files
  - Associate file names with specific physics-significant fields and management fields.
  - The Mass Storage is used only as a data store
- Transparency for applications
  - Source is not always available, RFIO API is not so simple
  - We are developing a transparent access thru BYPASS (WYSCONSIN University)

Conclusion

- Announced volumes require
  - Cooperation of data servers or Fiber Channel drives
  - Name Server support
- Use of several Mass Storage in HEP
  - Don’t provide too imbedded solution (physics/mass storage)
  - Promote user usage of data book keeping
- Documentation
  - http://doc.in2p3.fr/hpss/
  - http://doc.in2p3.fr/doc/public/products/rfio/rfio.html