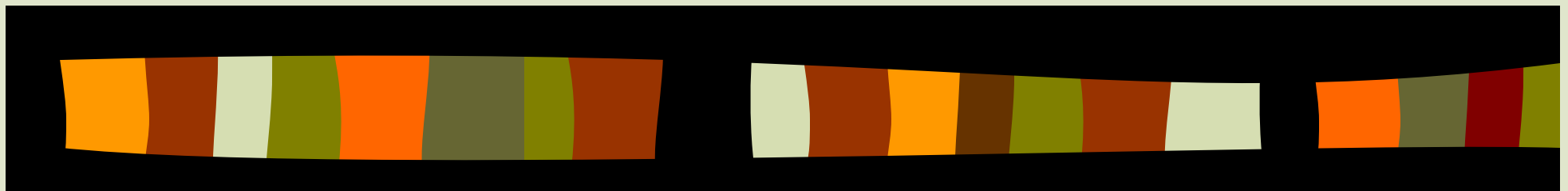


IN2P3 Site Report

John O'Neill, Philippe Gaillardon, Rolf Rumler

Centre de Calcul de l'IN2P3

<http://doc.in2p3.fr/hpss/>

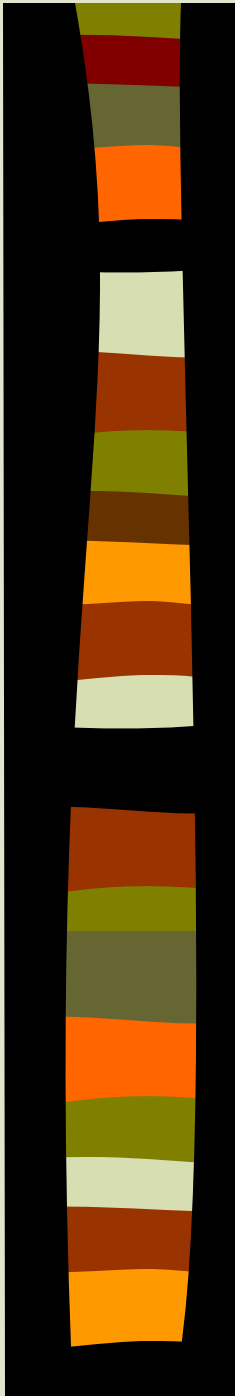


HPSS User Forum

18-20 June 2002, Indianapolis

Site report overview

- Introduction
- Our unique HPSS interface
- Access profile and ensuing concerns
 - Exception handling
 - Performance
 - Control
- Some specific points
 - File families
 - Future
- Summary



Who are we?

- IN2P3 = IN²P³ = National Institute of Nuclear Physics and Particle Physics: ~3300 people, ~1700 permanent scientists
- Institute of the CNRS (National Center for Scientific Research: ~26000 people, 12000 scientists), similar to American NSF
- 18 laboratories + 1 Computing Center (~45 people)
- Computing for *all* French particle physics, including those of CEA



History

- In production since October 1999
- Initially installed for Babar experiment (SLAC) with unique tape-only COS (disks under Objectivity)
- Now have ~150 TB stored for ~ 20 different experiments in particle and astro-particle physics
- Current use statistics at doc.in2p3.fr/hpss/HPSSexps_loop.php



Configuration

- 1 core server
- 7 tape movers, 21 Storagetek 9840s (SCSI, FC)
- 3 disk movers, 2.5 TB
- 90 TB tape only, 60 Tb disk + tape, 2.5×10^6 files
- Ethernet10 control network, GigE data network
- All AIX
- Foresee +15 drives and +2 TB disk by end 2002



Unique interface: RFIO

- Developed at CERN in early '90s
- Many versions since then
- Currently joint development by IN2P3 and CERN
- 64-bit, HPSS-knowledgeable (setcos, readlist/writelist)



RFIO (2)

- Available commands: rfcop, rfdir, rfstat, rfmkdir, rfrename, rfrm
- Standard C API + readlist/writelists, setcos
- C++ streaming interface
- Studying transparent i/f for closed-source programs
- Daemon runs on each disk mover and on 1 tape mover (tape-only COS)
- Permissions: correspondence between Unix perm bits and ACLs

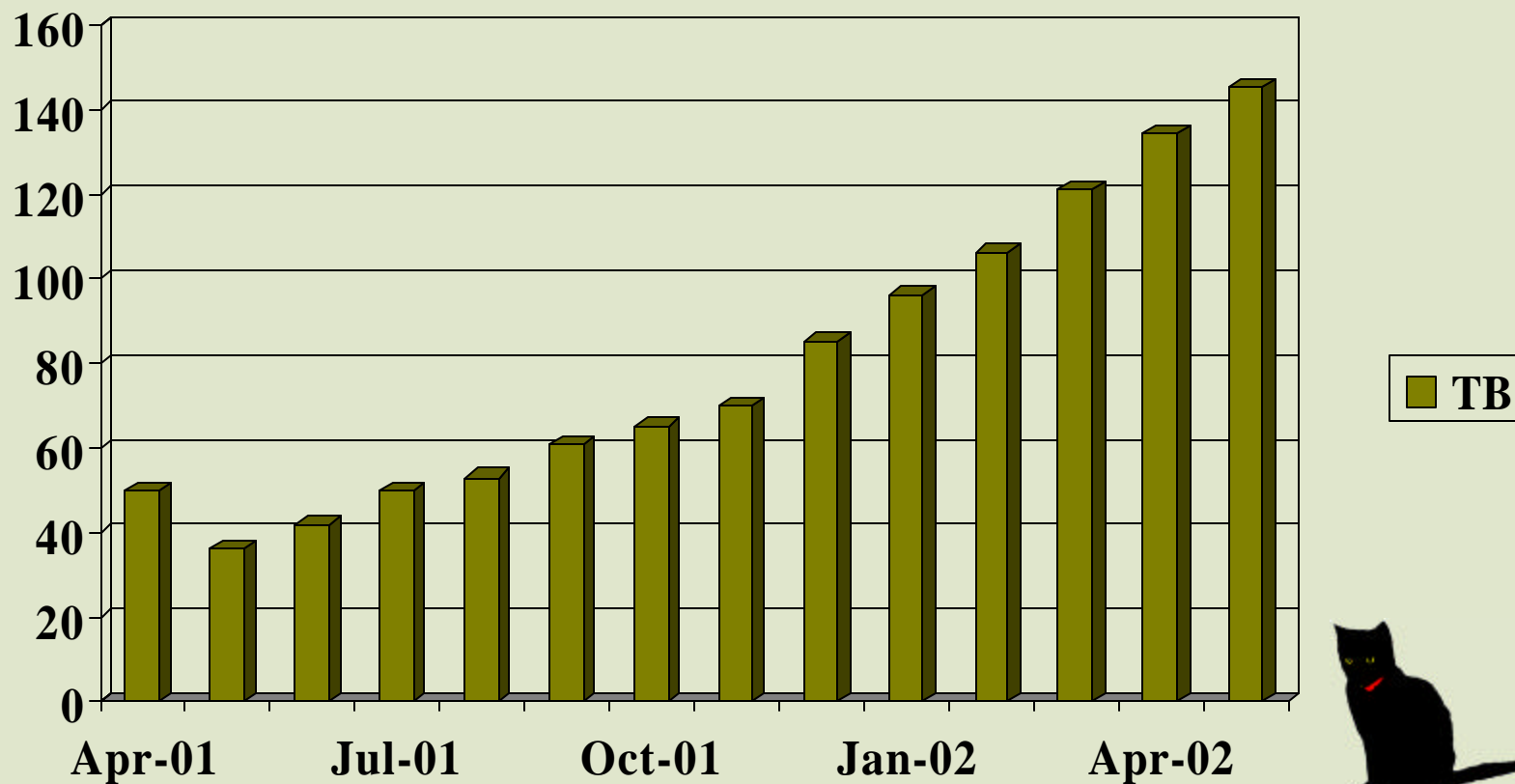
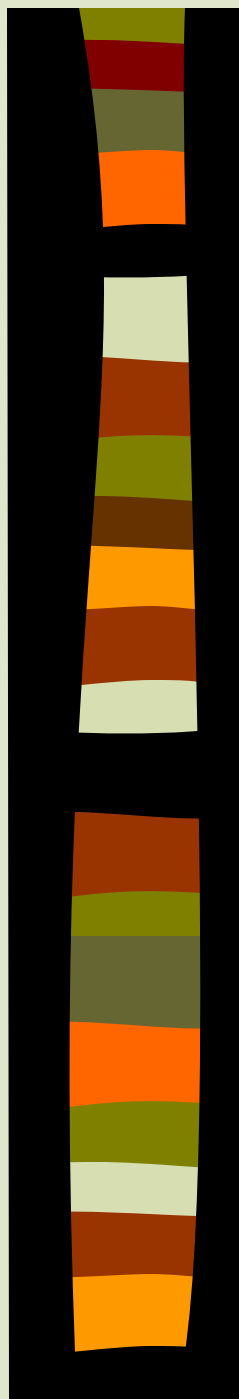


bbftp

- “Babar” ftp (developed at IN2P3)
- Uses readlist/writelist interface to HPSS via RFIO
- Authentication by ssh, Grid certificate, other
- Parallel streams over line
- Can saturate any line

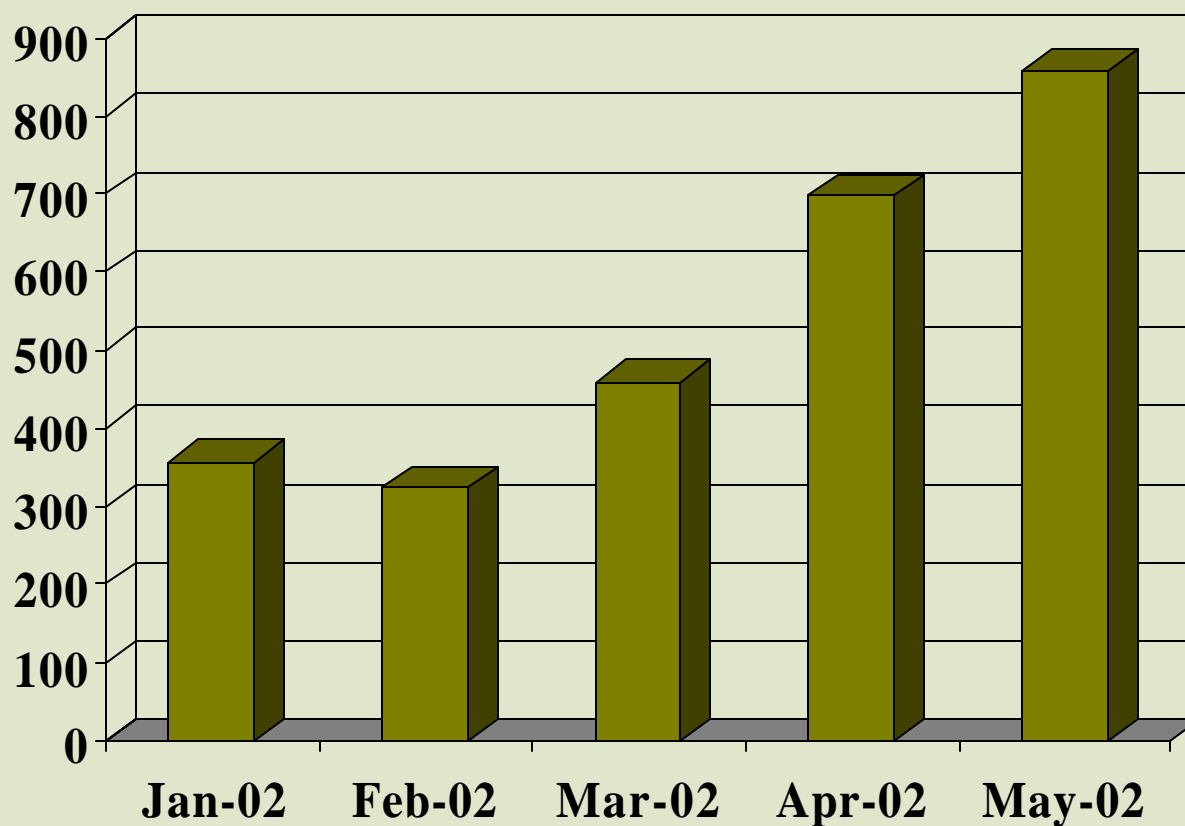


Data storage rate



Files accessed per month

(currently $\sim 2.8 \times 10^6$ files in all)



■ Accesses (K)



Access profile

- Access by hundreds of machines (all by rfiio):
 - local batch and interactive
 - remote transfer from around the world
 - access different for each experiment
 - no parallelism in access methods
- 7x24 operation
- Are we the only site with such a usage profile?



Access consequences

- Many HPSS optimization techniques not useable (striped tapes, e.g.)
- Leads to various concerns
 - hardware exception handling (especially for tapes)
 - performance
 - control



Tapes

- HPSS doesn't handle tapes that well: e.g., error during write → EOM, no log sense, *etc.*
- Doesn't handle drive problems well, either
- No useful drive status window



Performance concerns

- SFS (CPU, many MRA files)
- Utilities dumpppv_pvl, scrub (ls -R) extremely slow
- Storagetek PVR (+PVL+SSI) can lead to long mount waits (>4 minutes) due to query_mount in critical section; now down to 2 minutes
- Just upgraded robot to increase mount rate
- Seeing 250 mounts/hour with 2 silos (therefore. passthru)



Control

- hpssadm
 - Provides longed-for bulk configuration
 - Initialization verrrrrry slow (Security necessary? Option?) → limits usefulness in scripts
 - Various needs: authorization domains, macro facility, more options, etc.
- Need dynamic configuration (disks, drives, COS, SCs, etc.) without stopping production



EROS experiment - file families

- Capture sky segments on given date; study one segment across different dates (write order != read order)
- File families should be useful, with one family per segment
- But, migration may mount >1 tape per FF, which spread out data over several tapes
- Change request opened for mount limit per family



On the horizon

- Scratch tape pool (common to all SCs), à la TSM, etc.
- Anxiously waiting to give the *coup de grace* to Sammi
- Concerns about future DB (access SQL, conversion, disk space, DB2 only, ...)
- SAN + Fibre-channel integration: We need automatic recovery and would like load balancing (MDM approach)



Summary: wish list

- Better DB performance for small, frequently-accessed files
- Improved handling of tape and drive exceptions
- Improved (STK) PVR performance
- Dynamic configuration
- Open SQL DB access
- Less latency at startup of hpssadm
- Per-file-family mount limit
- Global scratch-tape pool
- SAN and Fibre-channel for recovery



The future

- More drives, tapes and disks
- More users, too → more accesses
- Pressure to install CERN CASTOR system
- Current license to beginning 2005
- Expecting good things from HPSS 5.1 next year
- Need to handle numerous, small files
- “The future is yet to come.” (Ionesco)



More wishes

- Commands: suspend, quiesce/drain
- Simpler installation/upgrade
- Messages: more precise, less repetition, tracking problems
- Cartridge-drive scheduling
- Disk management: finding and deleting “suspicious” files



Some links

- HPSS at IN2P3
 - <http://doc.in2p3.fr/hpss/>
- Storage use
 - http://doc.in2p3.fr/hpss/HPSSexps_loop.php
- RFIO documentation (in French)
 - <http://doc.in2p3.fr/doc/public/products/rfio/rfio.html>
- Bbftp documentation
 - <http://doc.in2p3.fr/bbftp/>



Au revoir



<http://doc.in2p3.fr/hpss/>