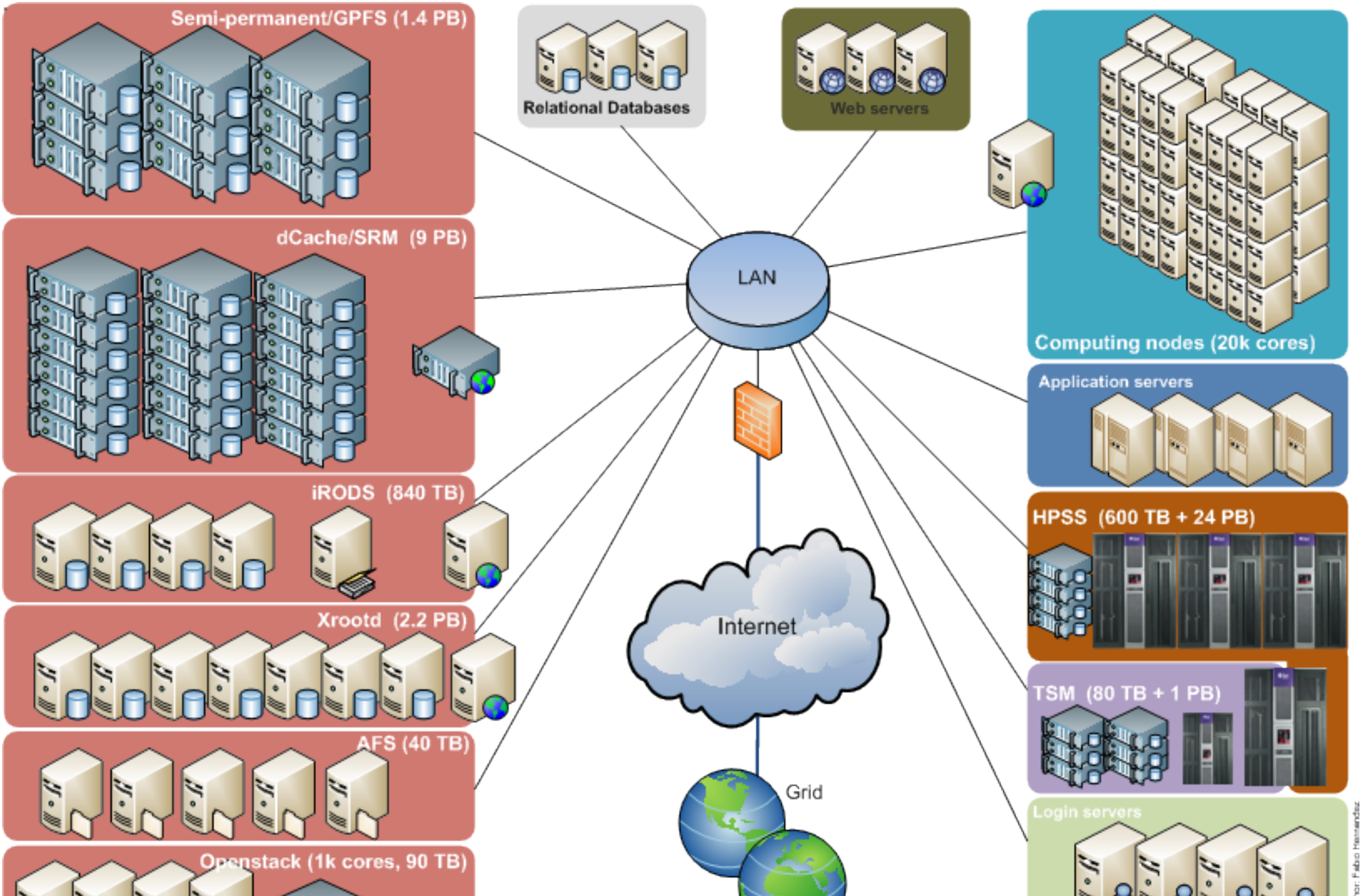
Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

Data Storage at CC-IN2P3

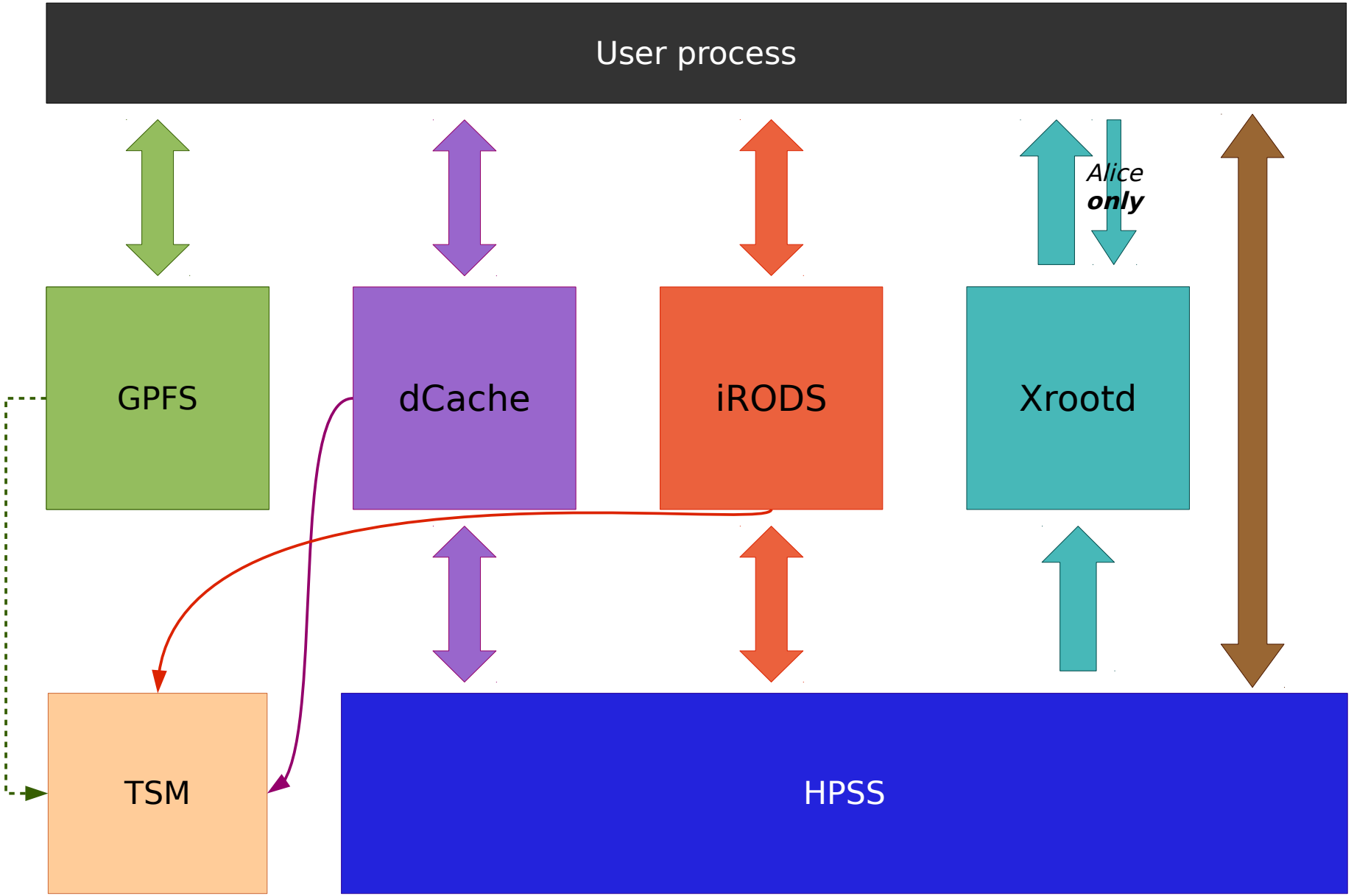
Loïc Tortay

Storage team

- 7 staff members
- In charge of:
  - Storage infrastructure architecture/design (w/ system administration team)
  - Storage applications architecture/design, installation & operation
  - Storage related developments
  - Storage hardware & software testing (w/ sysadmins)
  - Level 2/3 support
  - Backup & Tape infrastructure
  - Other non storage related tasks (Kerberos, Openstack operations, version control systems, ...)


Semi-permanent/GPFS (1.4 PB)
dCache/SRM (9 PB)
iRODS (840 TB)
Xrootd (2.2 PB)
AFS (40 TB)
Openstack (1k cores, 90 TB)
Relational Databases
Web servers
LAN
Internet
Grid
Computing nodes (20k cores)
Application servers
HPSS (600 TB + 24 PB)
TSM (80 TB + 1 PB)
Login servers

CC-IN2P3 Storage stack



HPSS: Mass storage system (HSM)

- Developed by a consortium of IBM & US DoE labs
- Introduced here in 1999 for BaBar Tier-A (SLAC)
- Currently running HPSS 7.3:
  - 24 PiB in 46 millions files
  - 14 disk servers (600 TB)
  - 12 tape servers, 94 (Oracle T10K) tape drives
  - About to start using T10K-D: 8.5 TB (native) per tape
- Local developments:
  - TReqS: Tape mounts scheduling, BBFTP, monitoring tools
- About 1.2 FTE

dCache: EGEE/LCG Storage

- Developed by DESY (Germany) & FNAL (US)
- Large scale read/write disk buffer to an HSM:
  - Running dCache 2.6
  - 170 disk servers, 9 PiB
  - 12 PiB on tape (in HPSS)
  - 58 million files for 8200 unique users (x509 DN)
  - SRM interface used for (LCG) data imports
  - Xrootd & dCap protocols for data access from computing nodes (Atlas moving from dCap to WebDAV)
- About 1.5 FTE
- Formerly involved in the development, now mostly monitoring/operation tools

Xrootd: HEP Specific large scale disk cache

- Developed by SLAC (US), initially for BaBar needs: Efficient & scalable remote access to ROOT files
- Used here mostly as a read-only disk cache for the HSM:
  - 1.2 PB of disk for about a dozen groups & collaborations
  - Except for Alice (LHC): Xrootd is their main storage system (about 1 PB dedicated disk, read/write access)
- Running Xrootd v3 on 40 disk servers
- Part of the worldwide Xrootd federations for CMS & Atlas
- Extremely light management: less than 0.1 FTE
- Involved in the early tests and development

iRODS: Data transfer and management middleware

- Successor to SRB (UCSD, SDSC)
- Part of the development collaboration
- Used as a filesystem like storage application (no VFS integration)
- Applies rules to migrate files transparently to or from an HSM
- Running iRODS 3.3:
  - 15 disk servers (840 TB)
  - 600 TB in 46 million files on disk
  - 8.2PB, 8 million files on tape (HPSS)
  - Used by about 20 groups
- Lightweight management: 0.3 FTE

GPFS: Cluster filesystem

- Proprietary (IBM)
- Running GPFS 3.5:
  - 40 disk servers
  - 800 computing (and login/service) nodes
  - 1.4 PB disk infrastructure
  - 450 million files for 1600 users in 75 groups
- About 0.75 FTE
- No backup (except experimental data)
- Local developments:
  - Monitoring and operation tools
  - Space & quota management (& delegation) tool
  - Batch integration (load management)

TSM: Backup system

- Proprietary (IBM)
- Running TSM 6.3 on AIX servers:
  - 4 disk & tape servers
  - 24 tape drives, 2000 tapes
  - About 1 PB in 700 million files
  - Around 5 TB/per day
- Infrastructure service for CC-IN2P3, IN2P3 and campus labs, not end users laptops
- About 1 FTE
- Local developments:
  - Monitoring tools

- GPFS for Openstack infrastructure(s):
  - VM images (Glance)
  - VM instances
  - 2 clusters, 14 servers, around 40 TB
- Cinder (Openstack block device storage):
  - Ephemeral and persistent block devices for VMs
  - Linux disk servers (DAS), about 25 TB
- Swift (Openstack S3-like storage):
  - Largely unused due to low end-user demand
  - Linux disk servers (DAS), about 25 TB
- Ceph:
  - Plans to (re)test, including as an alternative to GPFS

Storage tests/evaluations

- Mostly hardware before/during procurements:
  - Historically used local "model-based" benchmark (Xio)
  - Considering switching to fio for some/most tests
- Recently:
  - Test of Flash-only storage unit for GPFS metadata infrastructure overhaul (more tests to come)
  - EMC Scale-out storage system (NDA)
- Software too, long collaboration with SLAC:
  - Xrootd large scale tests here before SLAC deployment
  - Medium-large Qserv cluster for LSST, up to 310 storage nodes