Vue d'ensemble des services de calcul et de stockage au CCIN2P3

Fabio Hernandez
fabio@in2p3.fr

Workshop l'analyse au CCIN2P3

Lyon, 17 avril 2008

dapnia
cea
saclay

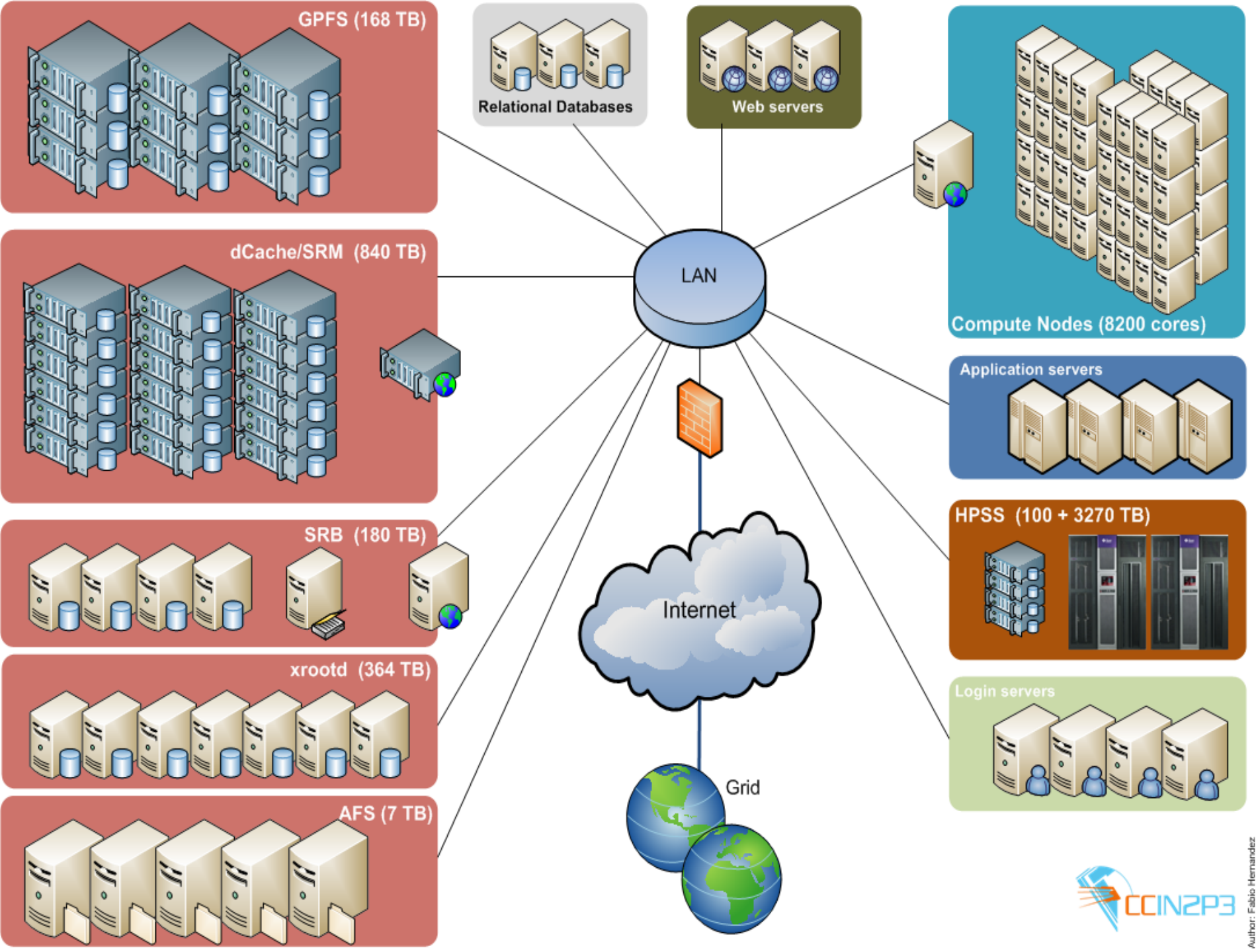
CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

Préambule

- Objectifs
  - Présenter une vue d'ensemble des services proposés par le CCIN2P3 aux expériences
  - Affiner notre compréhension des besoins en infrastructure de calcul et de stockage pour l'analyse des données

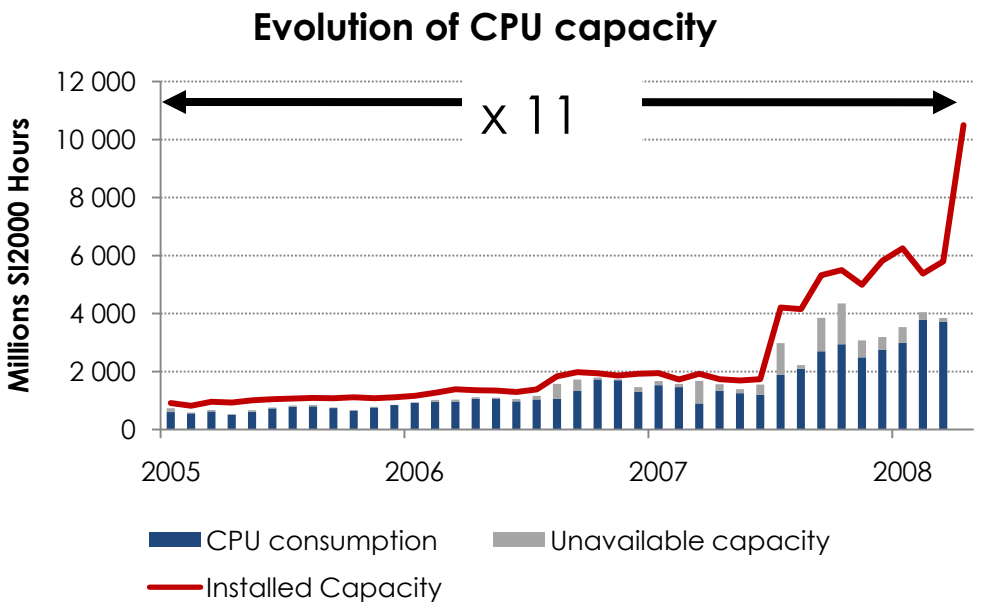
Table des matières

- Vue d'ensemble des services
- Services de stockage de données
- Accès aux données pour l'analyse
- Conclusions
- Questions & commentaires


GPFS (168 TB)
Relational Databases
Web servers
Compute Nodes (8200 cores)
dCache/SRM (840 TB)
LAN
Application servers
SRB (180 TB)
HPSS (100 + 3270 TB)
Internet
xrootd (364 TB)
Login servers
Grid
AFS (7 TB)
CCIN2P3
Author: Fabio Hernandez
Last Updated: 2008-04-14

Service de calcul batch

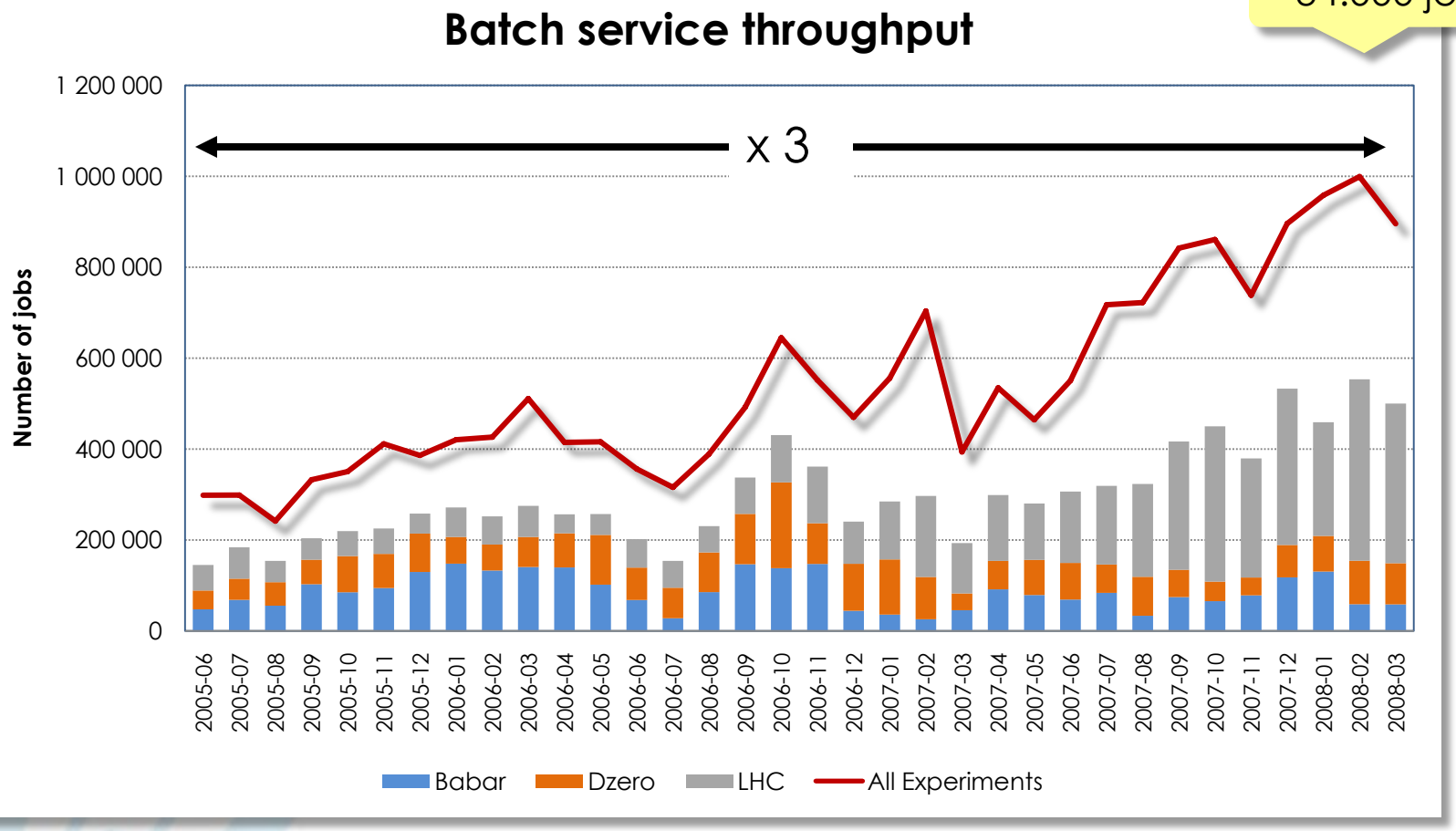
- 2 fermes utilisables par toutes les expériences supportées
  - Anastasie
    - *Généraliste*
    - *1192 machines, 8216 cœurs*
    - *Puissance: 14 M SI2000*
  - Pistoo
    - *Jobs parallèles*
    - *50 machines, 304 cœurs*
    - *Puissance: 0,4 M SI2000*



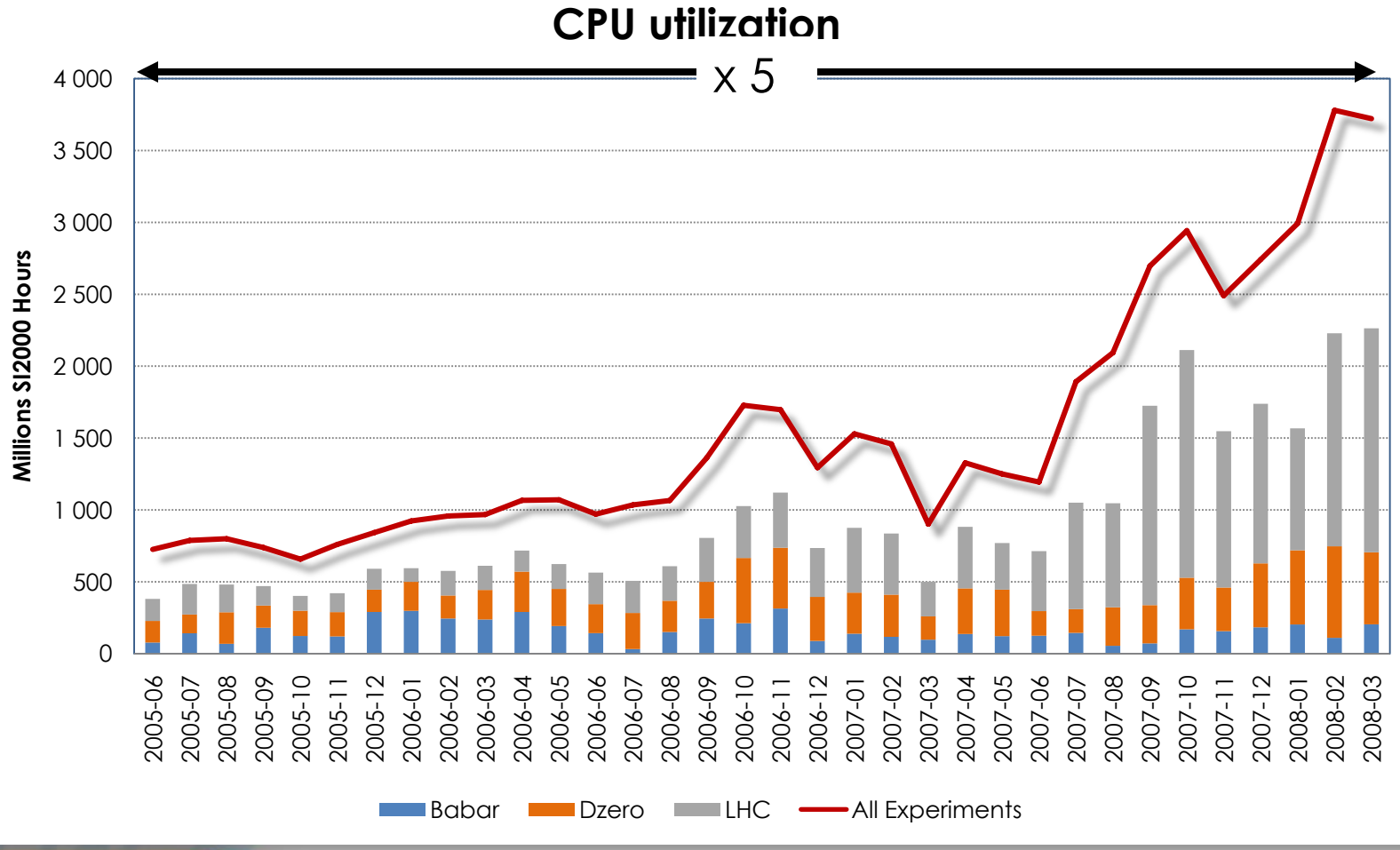
Tient compte des 400 machines en cours de mise en exploitation

Service de calcul batch (suite)

34.000 jobs/jour



Service de calcul batch (suite)



Data storage services

- Personal and group files
  - AFS (Andrew File System)
    - *Networked file system for individual and group files, experiment software, software areas of (grid) virtual organisations and system software (compilers, public domain and commercial software, …)*
    - *POSIX interface*
    - *Origin: Transarc, then IBM, then community (OpenAFS)*

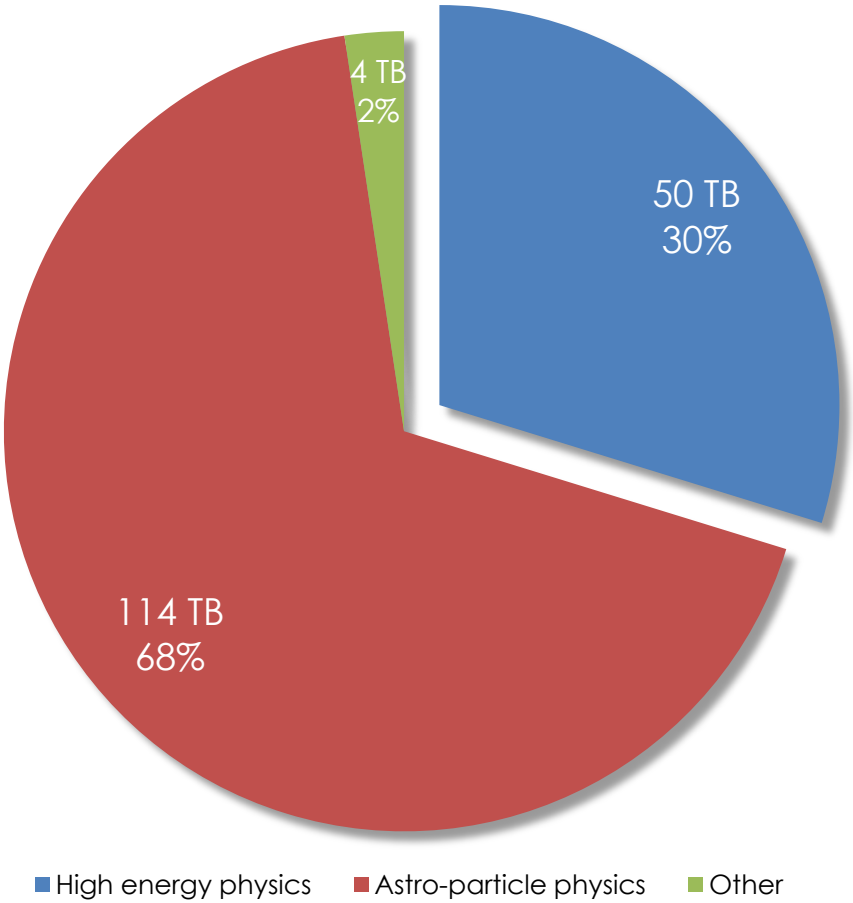
Data storage services (cont.)

- Data files
  - SRB (Storage Resource Broker)
    - *Middleware to aggregate into a hierarchical common name space files physically stored on heterogeneous multi-organizational storage systems (file, tape, etc.)*
    - *SRB presents the user with a single file hierarchy for data distributed across multiple storage systems (local or remote)*
    - *Access protocol: proprietary client commands and API*
    - *Origin: San Diego Supercomputing Center*

Data storage services (cont.)

- Data files (cont.)
  - GPFS
    - *Networked file system, designed for high-performance intra-site highly concurrent access*
    - *Used for medium-term storage of data which primary copy is or should be elsewhere (for instance, tape)*
      - Distributed (over several servers) and shared (by several users and by all compute nodes) disk working space
      - Files under /sps namespace are managed by GPFS
    - *POSIX interface*
    - *Origin: IBM*

/sps allocation



Total allocation:
168 TB

Used space:
138 TB (69 M files)

Source: http://ccspsmon.in2p3.fr/fi

Data storage services (cont.)

- Data files (cont.)
  - GPFS (cont.)
    - *Chunk size (unit of exchange between compute node and file server): 1MB*
      - As a consequence, acessing small blocks of data (e.g. 16K) is not efficient
    - *The (high) number of files makes some meta-data operations expensive*
    - *Mechanisms for space management*
      - Quotas: per user (individual) or groups (experiments)
      - Access control and delegation: ACLs not as flexible as the AFS' ones – extended POSIX interface
    - *Home-grown mechanisms of disk space housekeeping available on experiment's demand*

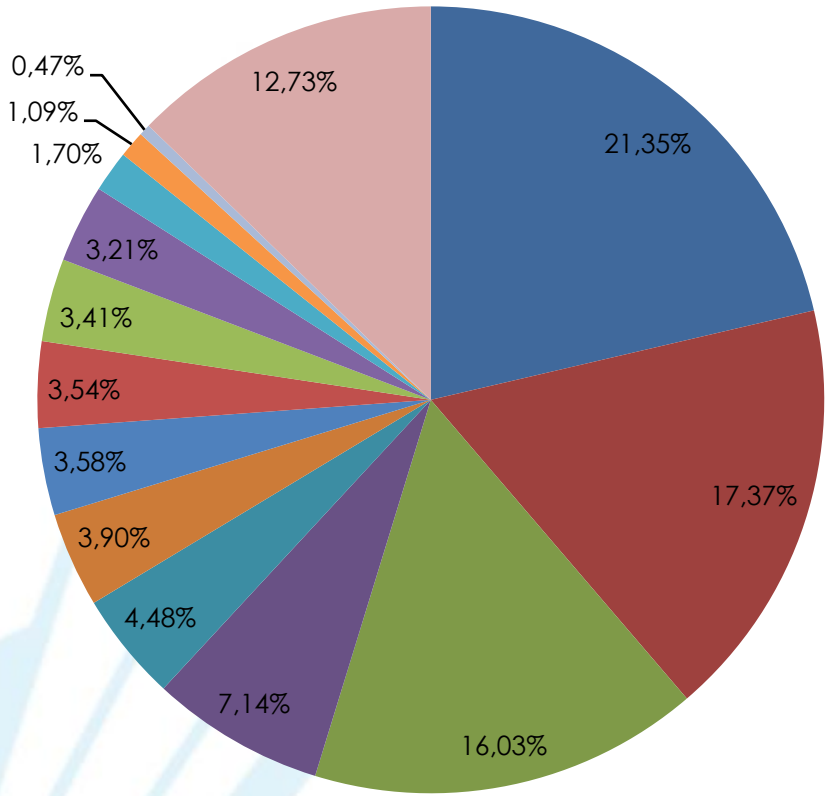
Data storage services (cont.)

- Data files (cont.)
  - GPFS (cont.)
    - *Planned capacity increase in 2008: additional 640 TB*
      - Some additional restrictions: unit of space allocation becomes 4TB for an efficient usage of the selected hardware
    - *Expected availability in production: mid-May*
    - *You can check your experiment usage of GPFS at*
      - http://ccspsmon.in2p3.fr/

Usage of /sps by HEP experiments

Type of files in /sps - HEP experiments

(% of number of files)

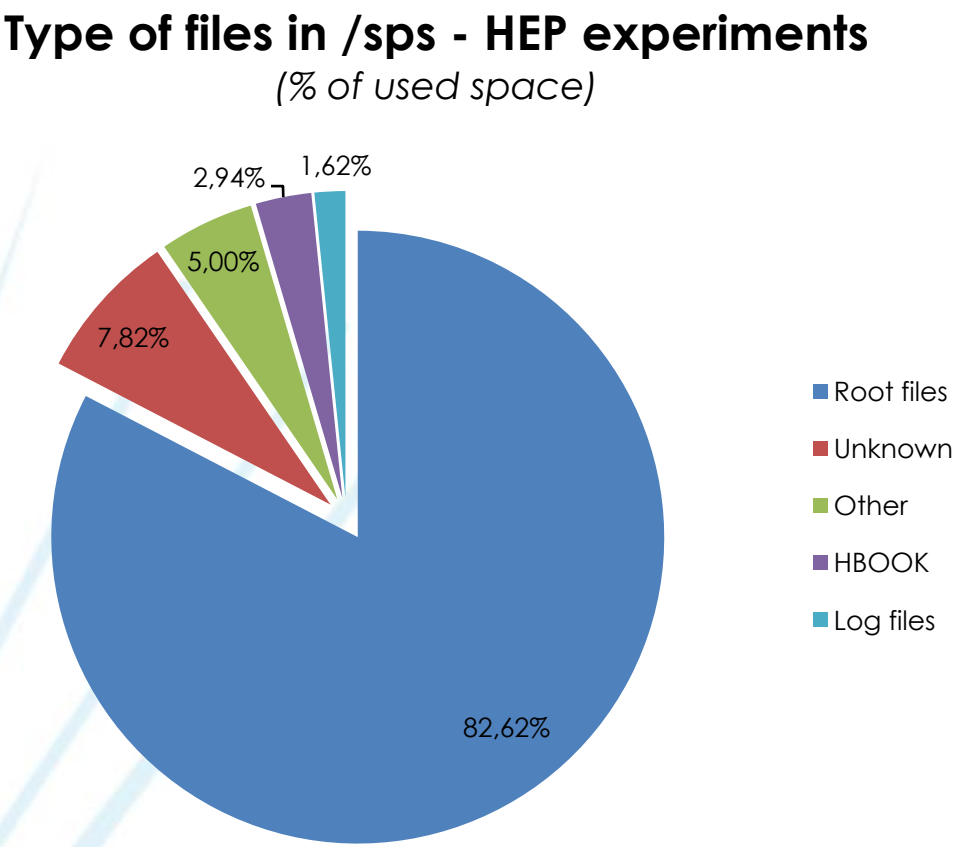


12M files stored in /sps by HEP experiments

/sps is used to store a large variety of file types

Source: http://ccspsmon.in2p3.fr/fi

Usage of /sps by HEP experiments (cont.)



82% of the space used by HEP experiments is taken by ROOT files

Source: http://ccspsmon.in2p3.fr/fi

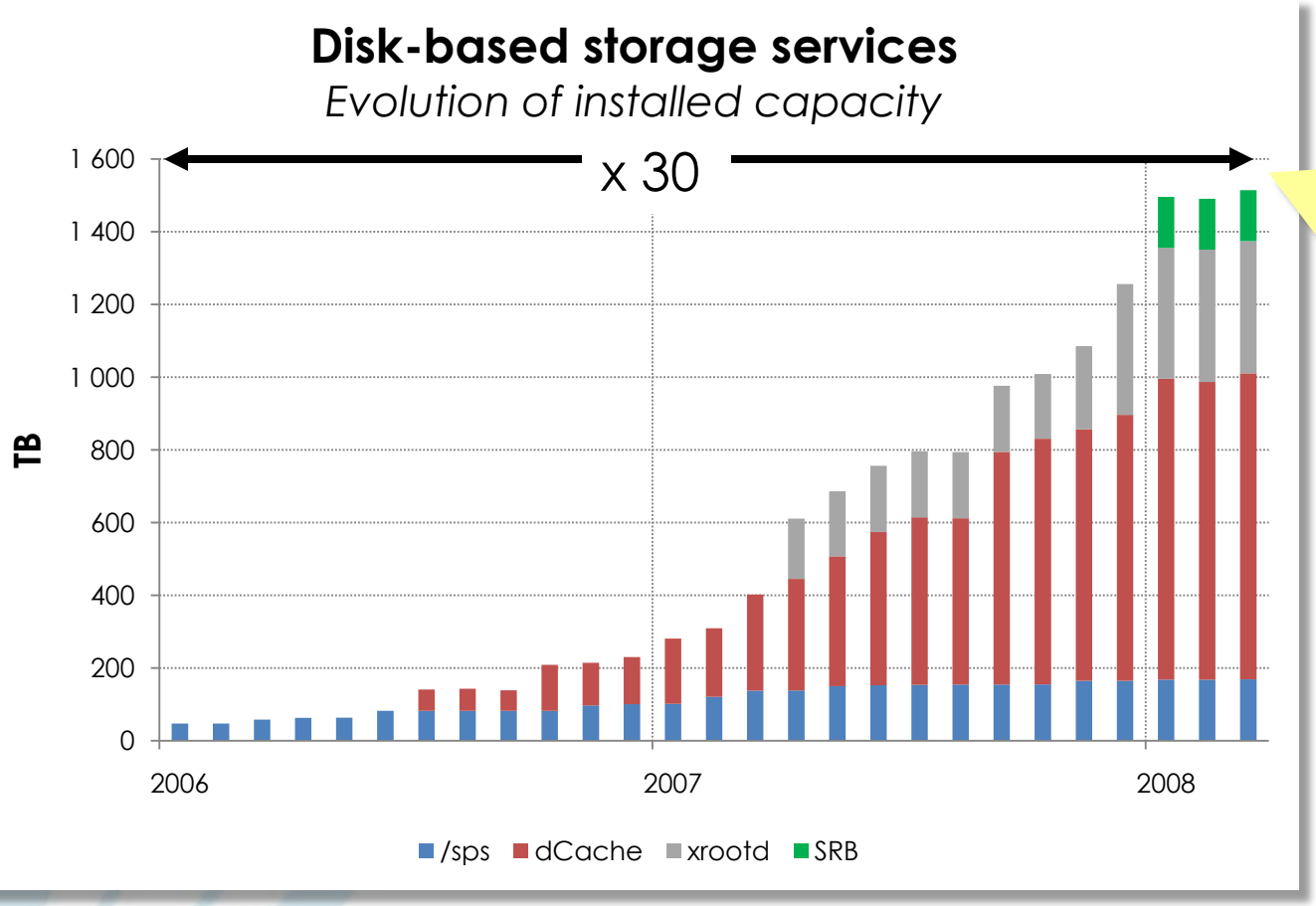
Data storage services (cont.)

- Data files (cont.)
  - Xrootd (a.k.a. Scalla)
    - *Designed for low latency high-bandwidth (byte-level) data access*
    - *Federation of several file servers into a common namespace*
    - *Designed to be efficient for random access of (relatively) small chunks of data in a fault-tolerant way*
    - *Proprietary POSIX-like API (ROOT-enabled)*
    - *Origin: SLAC*

Data storage services (cont.)

- Data files (cont.)
  - dCache/SRM
    - *System for storage and retrieval of data, distributed among a large number of heterogeneous server nodes, under a single virtual filesystem tree with a variety of standard access methods*
    - *Features: space management, pool attraction, dataset replication, hot spot determination and recovery from disk or node failures*
    - *Implements the SRM interface*
      - Required for every storage element of the EGEE/LCG grid
    - *Interfaced to the site's mass storage system (HPSS)*
    - *Origin: DESY and Fermilab*

Stockage sur disque



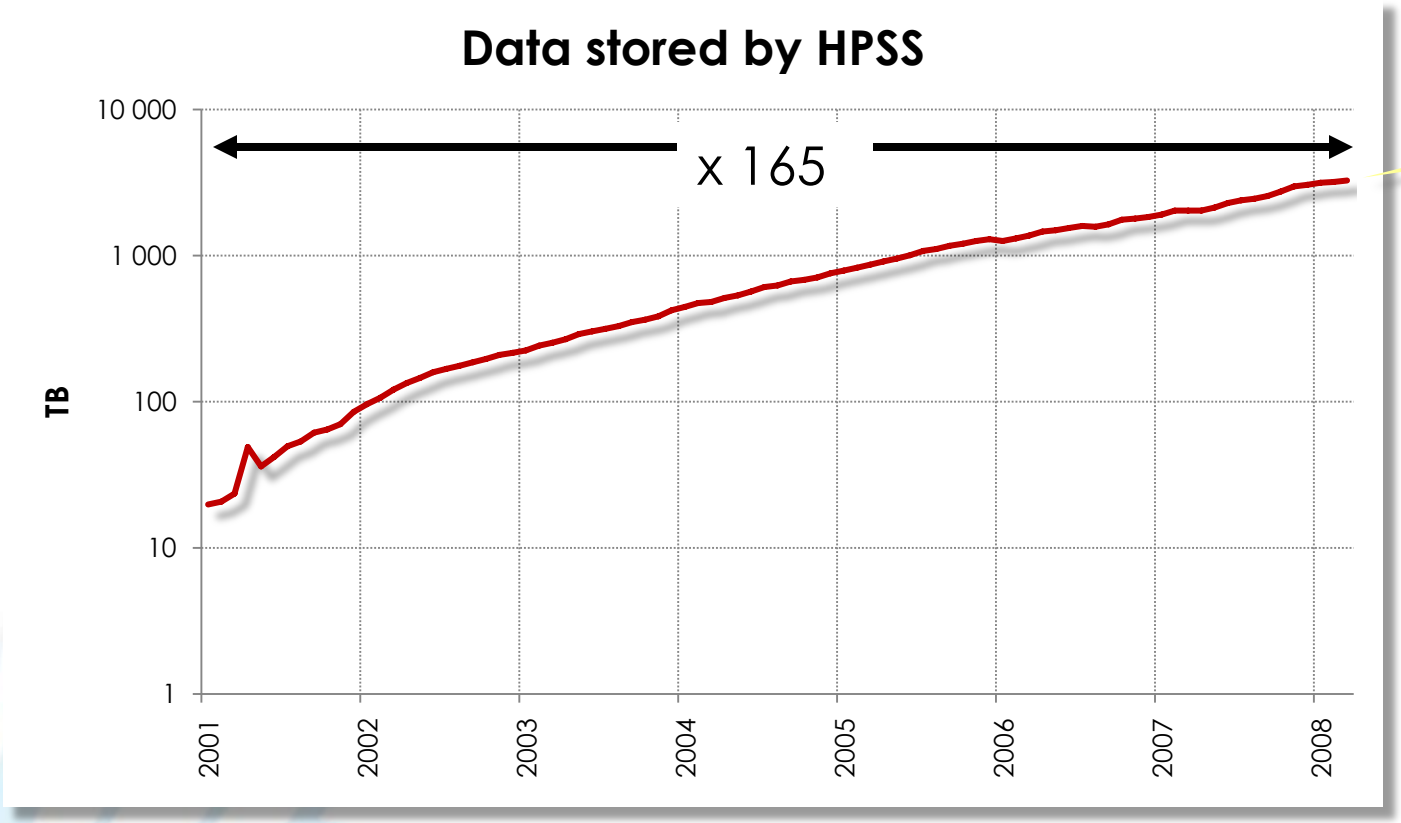
Notes:

- AFS, disque cache HPSS et disque local aux worker nodes non compris
- Données xrootd et SRB non disponibles avant les périodes affichés

Data storage services (cont.)

- Data files (cont.)
  - HPSS (High Performance Storage System)
    - *Hierarchical storage system*
      - software that manages petabytes of data on disk and automated tape libraries
    - *Features: scalable capacity, scalable I/O performance, incremental growth*
    - *Main repository of data files for the site*
      - Grid-enabled through dCache/SRM and SRB
    - *Well suited for sequential access of files in the gigabyte zone*
    - *Origin: US DoE laboratories and IBM*

Stockage de masse

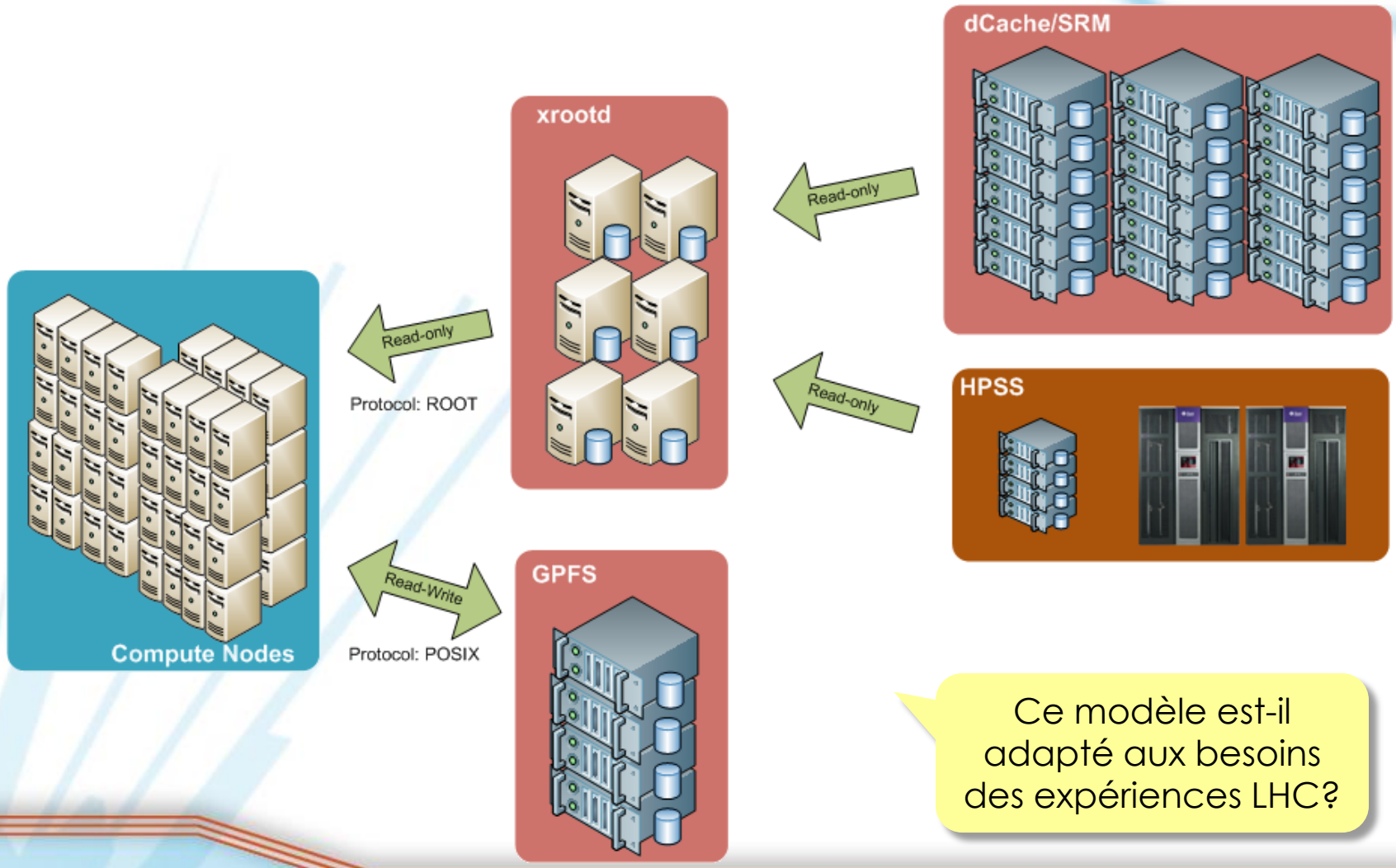


3,2 PB
20 M fichiers

Evolutions planifiées en 2008

- Stockage disque
  - 500 TB supplémentaires pour le cache disque HPSS
  - 640 TB supplémentaires pour GPFS
  - 3 PB supplémentaires pour dCache/xrootd (principalement LHC)
- Stockage sur bande
  - 2 robots et 30 dérouleurs supplémentaires
- Réseau local
  - Vers une généralisation des liens à 1 Gbps pour les nœuds de calcul, n x 1 Gbps pour les serveurs de stockage et 10 Gbps entre routeurs

Analysis – data flow



Ce modèle est-il adapté aux besoins des expériences LHC?

Autres services

- Gestionnaires de bases de données relationnelles
  - Oracle principalement
    - *Forte croissance ces dernières années*
  - MySQL et PostgreSQL également supportés à des niveaux de service différents
- Hébergement de sites web
- Services institutionnels
  - EDMS, Indico, visio-conférence, webcast,…

Conclusions

- Croissance soutenue des différents services pendant ces dernières années
- Diversification de l'offre de services de stockage, en particulier sur disque
  - Augmentation considérable de la capacité
- Intégration des différents services de stockage pour répondre aux besoins des activités des expériences
  - Simulation, re-processing, analyse, …
- Votre contribution est essentielle pour nous aider à comprendre vos besoins, en particulier pour l'analyse

Questions & Comments

Remerciements

- Merci à Loïc Tortay, Ghita Rahal et Benoit Delaunay pour leurs commentaires