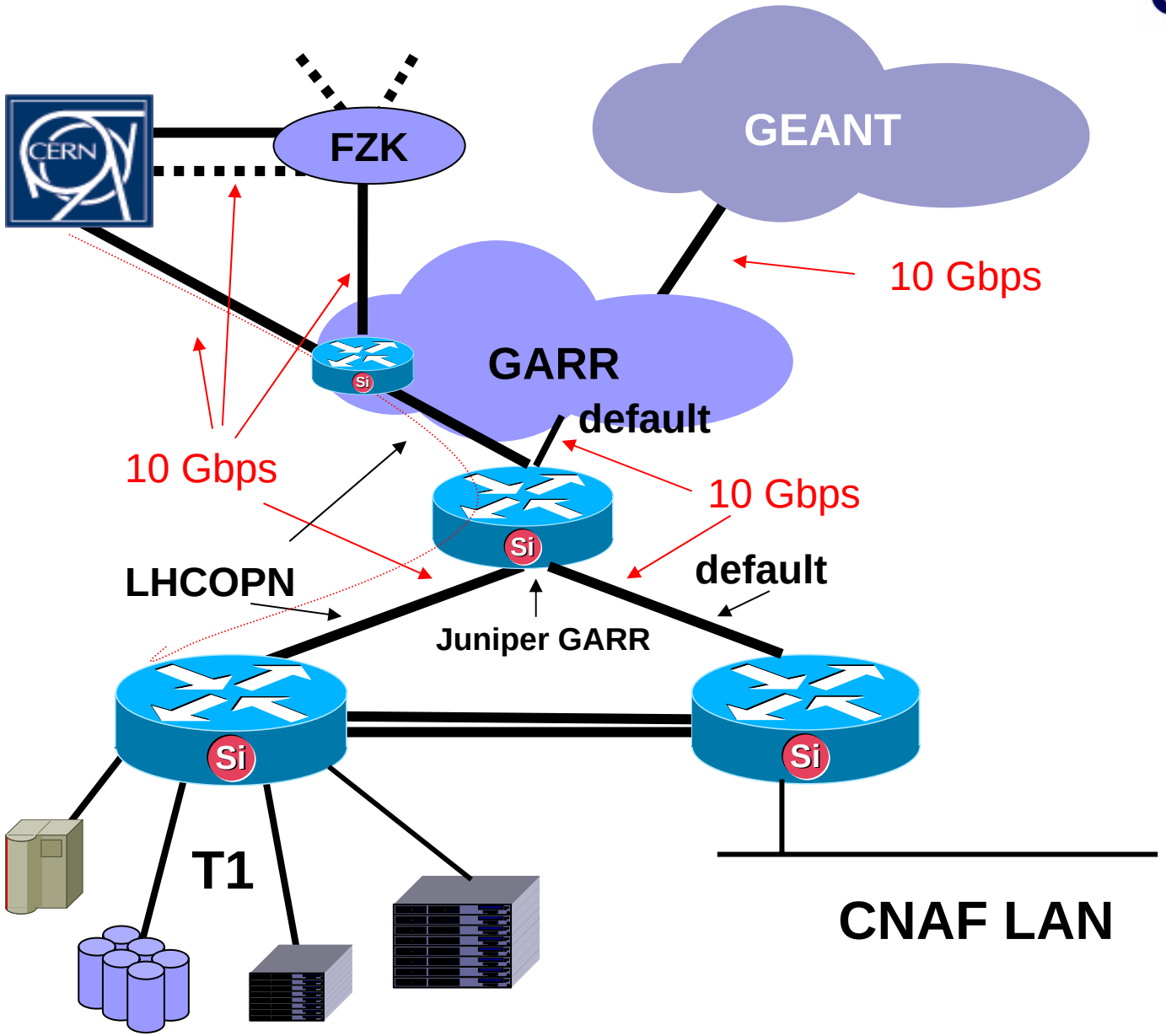
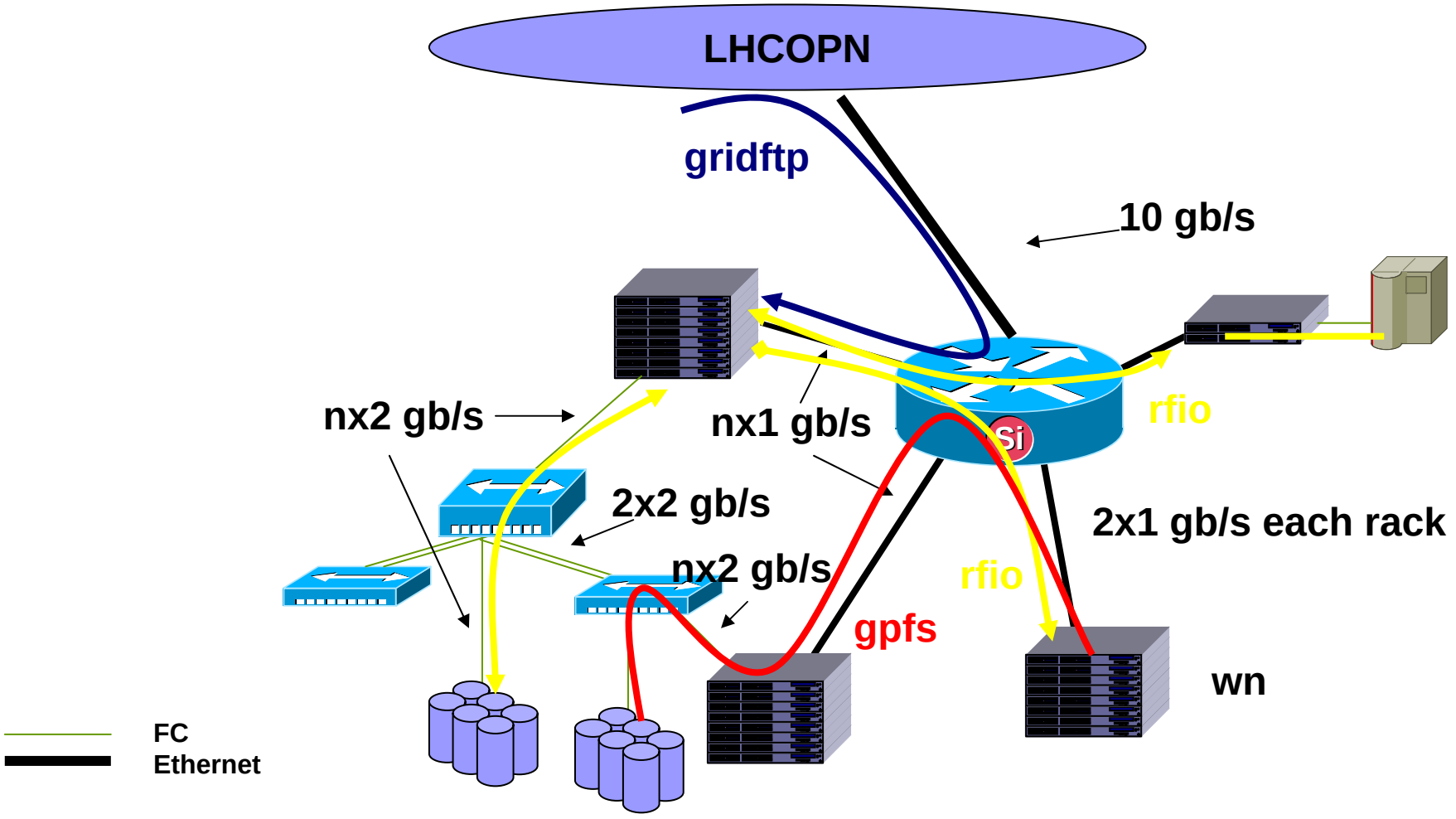
The INFN Tier-1 Storage Implementation

Luca dell'Agnello
INFN-CNAF
February, 6 2007

WAN connectivity

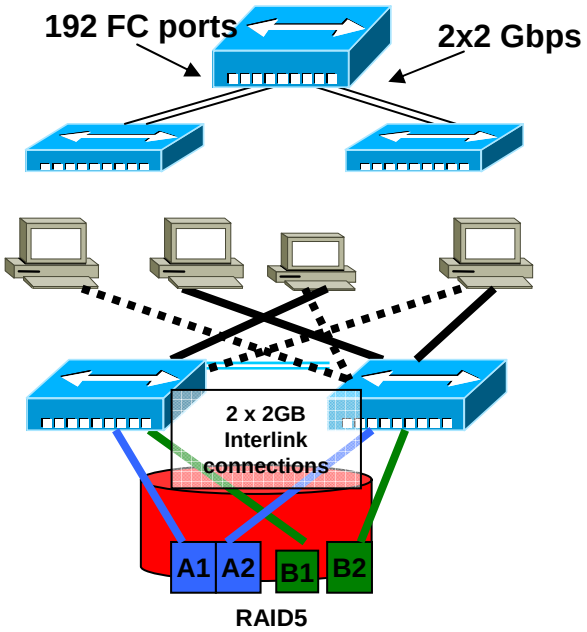


Storage layout



Storage & Mass storage

- Total disk space at Tier1: ~ 600 TB raw (~ 500 TB net disk space)
  - Completely allocated
    - New storage (400 TB raw) arrived – prod. December
  - Mainly based on SATA-FC technology (for general use)
    - Good performances and scalability
    - Raid 5 + hot spare
    - Possibility to have redundant path to the data
- Access via gridftp, rfio, *native GPFS*
  - Still some NFS storage (to be phased off)
- HMS based on CASTOR (v2)
  - StorageTek L5500 library (up to 20 drives, 1 PB on line)
    - 6 LTO2 drives (20-30 MB/s) with 1300 tapes
    - 7 (+3) 9940B drives (25-30 MB/s) with 1350 (+600) tapes
  - CASTOR file system hides tape level
    - Es. /castor/cnaf.infn.it/lcg/cms/prova.dat
  - Access to HSM
    - Native access protocol: rfio
    - srm interface for grid fabric available (rfio/gridftp)



Hw evolution (2007)

- New storage system (400 TB) ready for production
  - 3 Clarion CX3-80 (EMC)
    - SATA – FC technology
  - 36 new disk servers
- Increase of disk capacity depending of infrastructural upgrade (cooling and power)
  - Probably additional 200 TB of disk in the meantime
- Increase of # of disk servers
- Tender for new library starting now

CASTOR

- CASTOR v. 2.1.1-9
- 1 stager instance for production (supporting other VOs besides LHC)
  - 30 (25 for LHC VOs) disk servers
    - supporting rfio, gridftp protocols and interconnected to storage via FC
  - 12 tape servers
  - 2 srm 1.1 end-points (1 instance with no tape back-end, no gc)
  - At present 1 pool per storage class for each LHC experiment
    - Only 3 disk servers for each pool (need probably to increase number)
- 1 stager instance for tests
  - 1srm 2.2 end-point for basic tests
  - 3 disk servers (access to internal disks only at the moment)

GPFS

- GPFS cluster including all WNs and GPFS disk servers
  - 100 TB of disk served
  - V 3.1release installed (stable since a few months)
  - 12 disk servers interconnected via FC to SAN
- No SRM 1.1 interface provided
  - Only possible to use as classic SE
- A small cluster for test purpose available

StoRM

- SRM 2.2 interface to POXIS file systems with ACL support (GPFS, XFS)
- Provides support for disk-only storage systems (D1T0)
- It is possible to have different Storage Areas in the same SC instance

e.g.srm://storm02.cr.cnaf.infn.it:8444//srm/managerv2?SFN=/lhcb/SA1/dir/test.txt

  - where /lhcb/SA1 identifies the SA
  - In next version a "targetSpaceToken" will identify the SA

- A test SRM 2.2 instance (StoRM) installed
  - First tests started (using ad hoc clients)

Planned (basic) tests

- Verification (certification) of storage infrastructure (SAN, LAN, servers)
  - Throughput tests for both CASTOR and GPFS
  - Verification of scalability of GPFS
- Comparison tests between xrootd, GPFS, rfio

Storage classes implementation

- Disk0Tape1 is and will be CASTOR
  - Space managed by system
  - Data migrated to tapes and deleted from when staging area full
- Disk1tape1 will be (probably!) CASTOR
  - Space managed by VO (i.e. if disk is full, copy fails)
  - Large buffer of disk with tape back end (and gc with an high threshold?)
- Disk1tape0 also investigating GPFS/StoRM
  - Space managed by VO
  - Open issue: efficiency of data moving to and from CASTOR (some VOs asked for this and also for some sort of backup ☺ at least in the first phase)
- Still need to investigate clearly experiments needs
  - E.g. LAN, WAN differences

StoRM

- **StoRM is a storage resource manager for disk based storage systems.**
  - It implements the SRM interface version 2.x.
  - StoRM is designed to support guaranteed space reservation and direct access (native POSIX I/O call), as well as other standard libraries (like RFIO).
  - StoRM take advantage from high performance parallel file systems. Also standard POSIX file systems are supported.
  - A modular architecture decouples StoRM logic from the supported file system.
  - Strong security framework with VOMS support.

by L. Magnoni

StoRM General Considerations 1/2

- **File system currently supported by StoRM**
  - GPFS from IBM.
  - XFS from SGI.
  - Any other File System with POSIX interface and ACLs support.
- **Light and flexible namespace structure**
  - The namespace of the files managed by StoRM relies upon the underlying file systems.
  - StoRM does not need to query any DB to know the physical location of a requested SURL.

by L. Magnoni

StoRM General Considerations 2/2

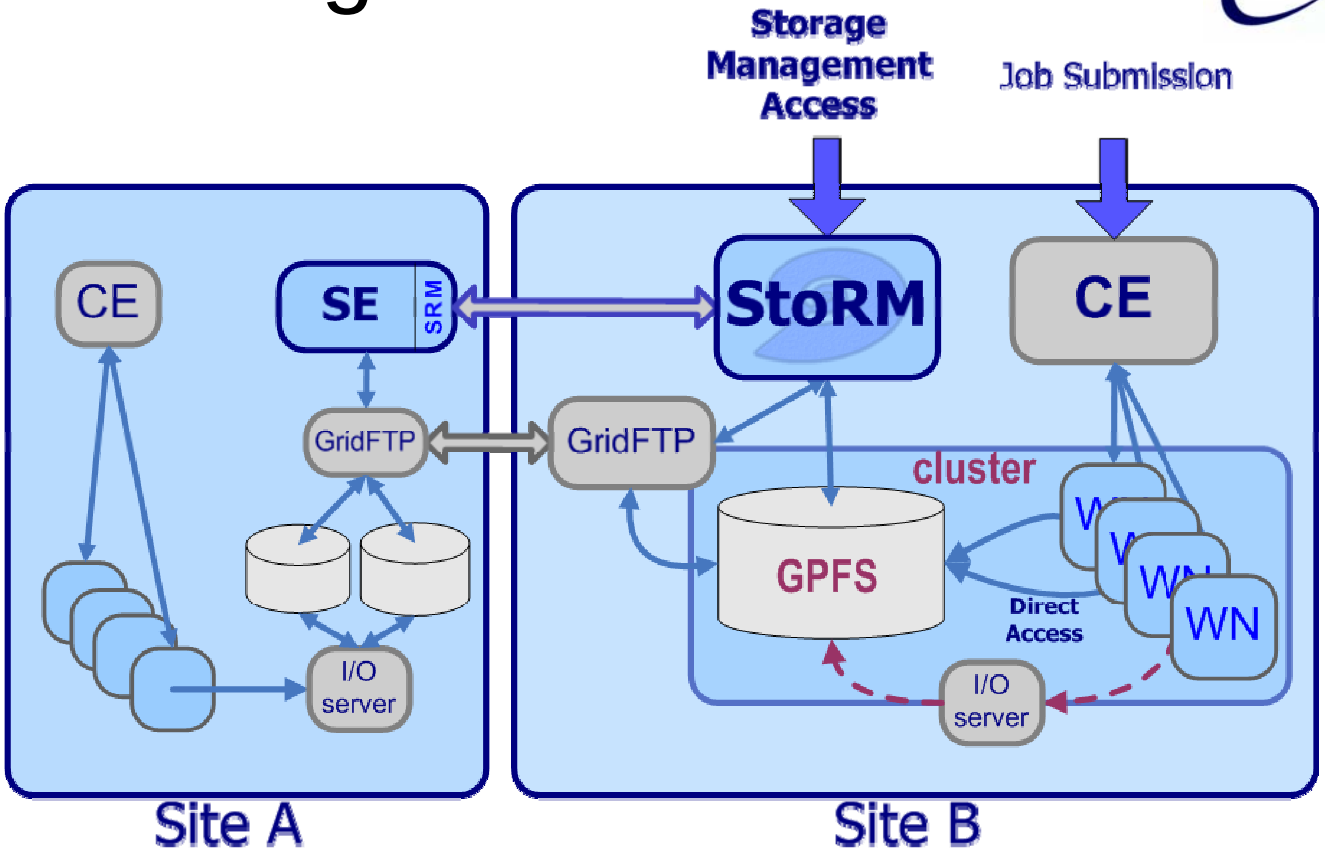
- **ACLs Usage**
  - StoRM enforce ACL entries on physical files for the local user corresponding to the grid-credential.
  - Standard grid applications (such as GridFTP, RFIO, etc.) can access the storage on behalf of the user.
- **Scalability and high availability.**
  - FE, DB, and BE can be deployed in 3 different machines.
  - StoRM is designed to be configured with *n* FE and *m* BE, with a common DB. But more tests are needed to validate this scenario.

by L. Magnoni

StoRM Grid usage scenario

- **StoRM dynamically manages files and space in the storage system.**

- **Applications can directly access the Storage Element (SE) during the computational process.**



File metadata are managed (and stored) by underlying file system. No replica of metadata at application level.. That is a file system job! In this way StoRM gain in performance.

Data access is performed without interacting with an external service, with great performance improvement (POSIX calls). Otherwise, standard data access using I/O Server (such as RFIO) is also fully supported.

by L. Magnoni

StoRM status and SRM issues

- **Status**
  - Migration to SRM v2.2 completed.
  - All functions requested by the SRM WLCG usage agreement are implemented.
  - New version of StoRM available.
- **StoRM SRM tests**
  - StoRM is involved in interoperability tests made by SRM-WG, the results are available here: http://sdm.lbl.gov/srm-tester/v22-progress.html
  - StoRM is involved also in other SRM tests made with S2 test suite: http://gdrb02.cern.ch:25000/srms2test/scripts/protos/srm/2.2/basic/s2_logs/

by L. Magnoni