The Italian Tier-1: INFN-CNAF

Andrea Chierici,
on behalf of the INFN Tier1

3° April 2006 – Spring HEPIX

Introduction

- Location: INFN-CNAF, Bologna (Italy)
  - one of the main nodes of the GARR network
  - Hall in the basement (floor -2): ~ 1000 m^2 of total space
  - Easily accessible with lorries from the road
  - Not suitable for office use (remote control mandatory)
- Computing facility for the INFN HENP community
  - Partecipating to LCG, EGEE, INFNGRID projects
- **Multi-Experiment TIER1 (22 VOs, including LHC experiments, CDF, BABAR, and** others)
  - Resources are assigned to experiments on a yearly basis

Infrastructure (1)

- Electric power system (1250 KVA)
  - UPS: 800 KVA (~ 640 KW)
    - needs a separate room
    - Not used for the air conditioning system
  - Electric Generator: 1250 KVA (~ 1000 KW)
    - **Theoretically suitable for up to 160 racks (~100 with 3.0 GHz Xeon)**
    - 220 V mono-phase (computers)
      - 4 x 16A PDU needed for 3.0 GHz Xeon racks
    - 380 V three-phase for other devices (tape libraries, air conditioning, etc…)
  - Expansion under evaluation
- **The main challenge is the electrical/cooling power needed in 2010**
  - Currently, we have mostly Intel Xeon @ 110 Watt/KspecInt, with quasi-linear increase in Watt/SpecInt
  - Next generation chip consumption is 10% less
    - E.g. Opteron Dual Core ~factor -1.5-2 less ?

Infrastructure (2)

- Cooling
  - RLS (Airwell) on the roof
    - ~530 KW cooling power
    - Water cooling
    - Need “booster pump” (20 mts T1 ←→ roof)
    - Noise insulation needed on the roof
  - 1 UTA (air conditioning unit)
    - 20% of RLS refreshing power and controls humidity
  - 14 UTL (local cooling systems) in the computing room (~30 KW each)
- New control and alarm systems (including cameras to monitor the hall)
  - Circuit cold water temperature
  - Hall temperature
  - Fire
  - Electric power transformer temperature
  - UPS, UTL, UTA

WN typical Rack Composition

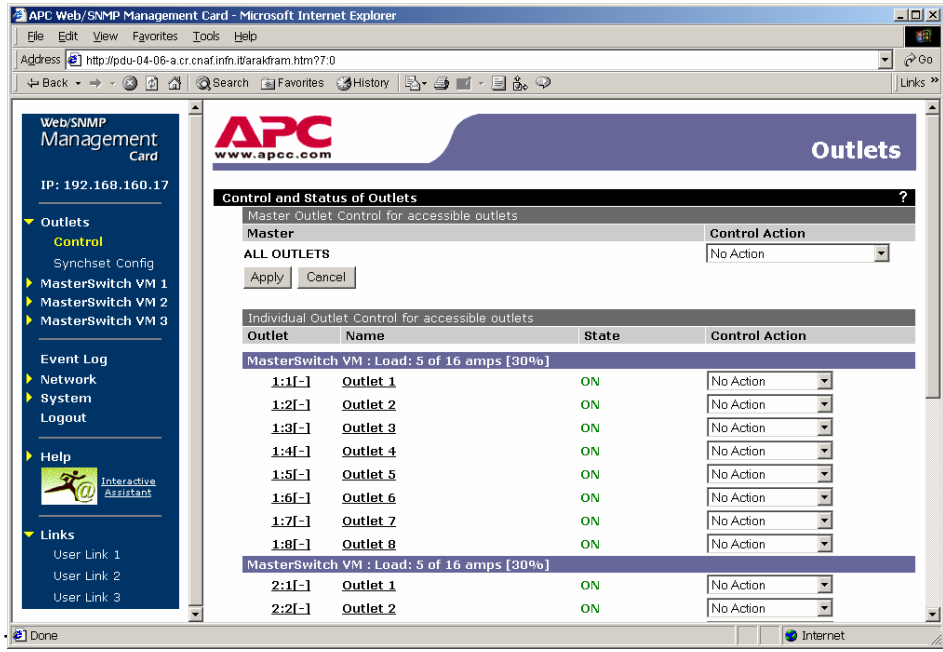
- Power Controls (3U)
  - Power switches
- 1 network switch (1-2U)
  - 48 FE copper interfaces
  - 2 GE fiber uplinks
- ~36 1U WNs
  - Connected to network switch via FE
  - Connected to KVM system

Remote console control

- Paragon UTM8 (Raritan)
  - 8 Analog (UTP/Fiber) output connections
  - Supports up to 32 daisy chains of 40 nodes (UKVMSPD modules needed)
  - IP-reach (expansion to support IP transport) evaluated but not used
  - Used to control WNs
- Autoview 2000R (Avocent)
  - 1 Analog + 2 Digital (IP transport) output connections
  - Supports connections up to 16 nodes
    - Optional expansion to 16x8 nodes
  - Compatible with Paragon (“gateway” to IP)
  - Used to control servers
- IPMI
  - New acquisitions (Sunfire V20z) have IPMI v2.0 built-in. IPMI is expected to take over other remote console methods in the middle term

Power Switches

- 2 models used:
  - “Old” APC MasterSwitch Control Unit AP9224 controlling 3 x 8 outlets 9222 PDU from 1 Ethernet
  - “New” APC PDU Control Unit AP7951 controlling 24 outlets from 1 Ethernet
- “zero” Rack Unit (vertical mount)
- Access to the configuration/control
- menu via serial/telnet/web/snmp
- Dedicated machine using APC Infrastructure Manager Software
- Permits remote switching-off of resources in case of serious problems



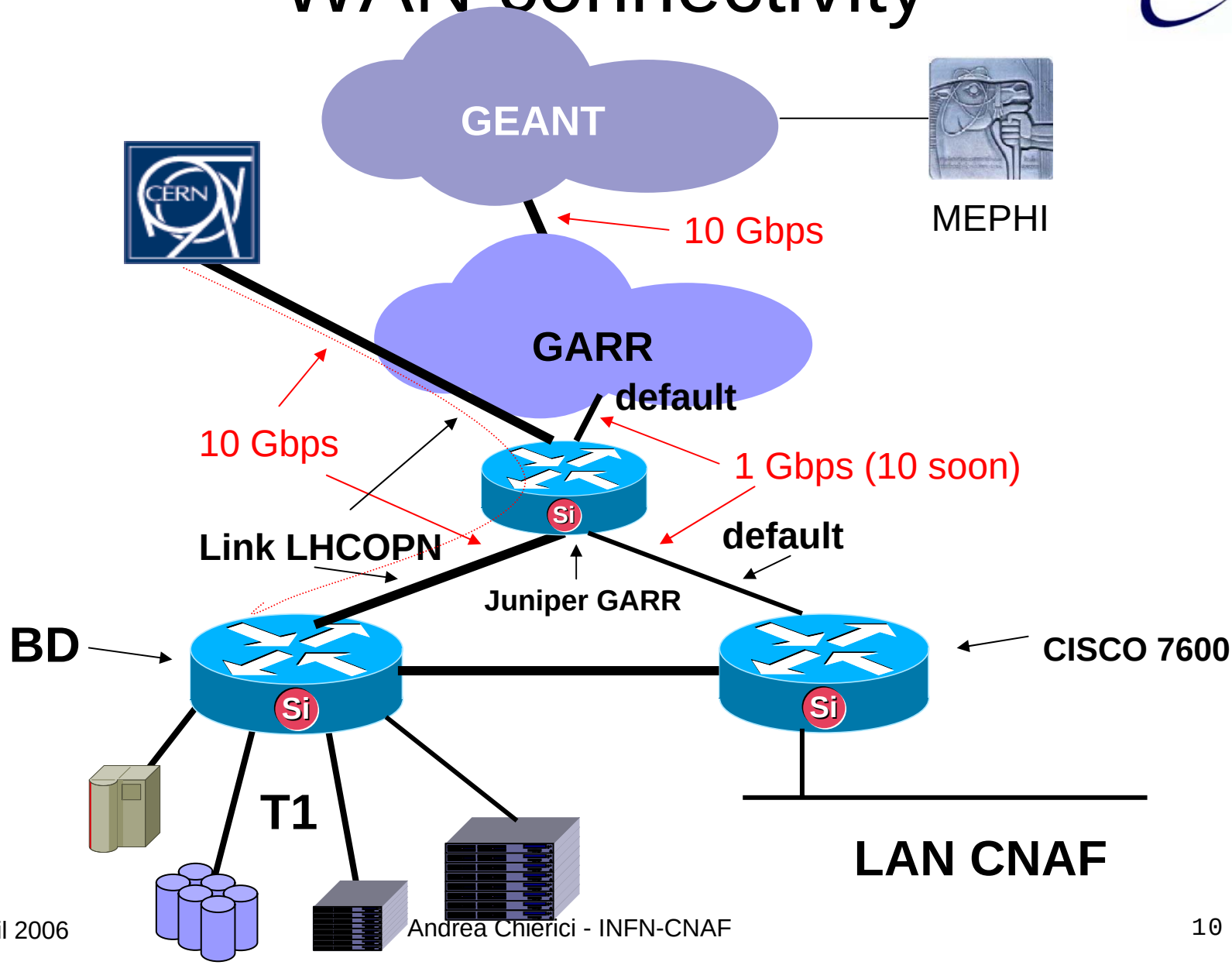
Networking (1)

- Main network infrastructure based on optical fibres (~20 Km)
- LAN has a “classical” star topology with 2 Core Switch/Router (ER16, BD)
  - Migration to Black Diamond 10808 with 120 GE and 12x10GE ports (it can scale up to 480 GE or 48x10GE) soon
  - Each CPU rack equipped with FE switch with 2xGb uplinks to core switch
  - Disk servers connected via GE to core switch (mainly fibre)
    - Some servers connected with copper cables to a dedicated switch
  - VLAN’s defined across switches (802.1q)

Networking (2)

- 30 rack switches *(14 switches 10Gb Ready):* several brands, homogeneous characteristics
  - 48 Copper Ethernet ports
  - Support of main standards (e.g. 802.1q)
  - 2 Gigabit up-links (optical fibres) to core switch
- CNAF interconnected to GARR-G backbone at 1 Gbps + 10 Gbps for SC4
  - GARR Giga-PoP co-located
  - SC link to CERN @ 10 Gbps
  - New access router (Cisco 7600 with 4x10GE and 4xGE interfaces) just installed

WAN connectivity



Hardware Resources

- CPU:
  - ~600 XEON bi-processor boxes 2.4 – 3 GHz
  - 150 Opteron biprocessor boxes 2.6 GHz
    - ~1600 KSi2k Total
    - Decommissioned ~100 WNs (~150 KSi2K) moved to test farm
  - New tender ongoing (800 KSI2k) – exp. delivery Fall 2006
- Disk:
  - FC, IDE, SCSI, NAS technologies
  - 470 TB raw (~430 FC-SATA)
    - 2005 tender: 200 TB raw
  - Requested approval for new tender (400 TB) – exp. Delivery Fall 2006
- Tapes:
  - Stk L180 18 TB
  - Stk 5500
    - 6 LTO-2 with 2000 tapes → 400 TB
    - 4 9940B with 800 tapes → 160 TB

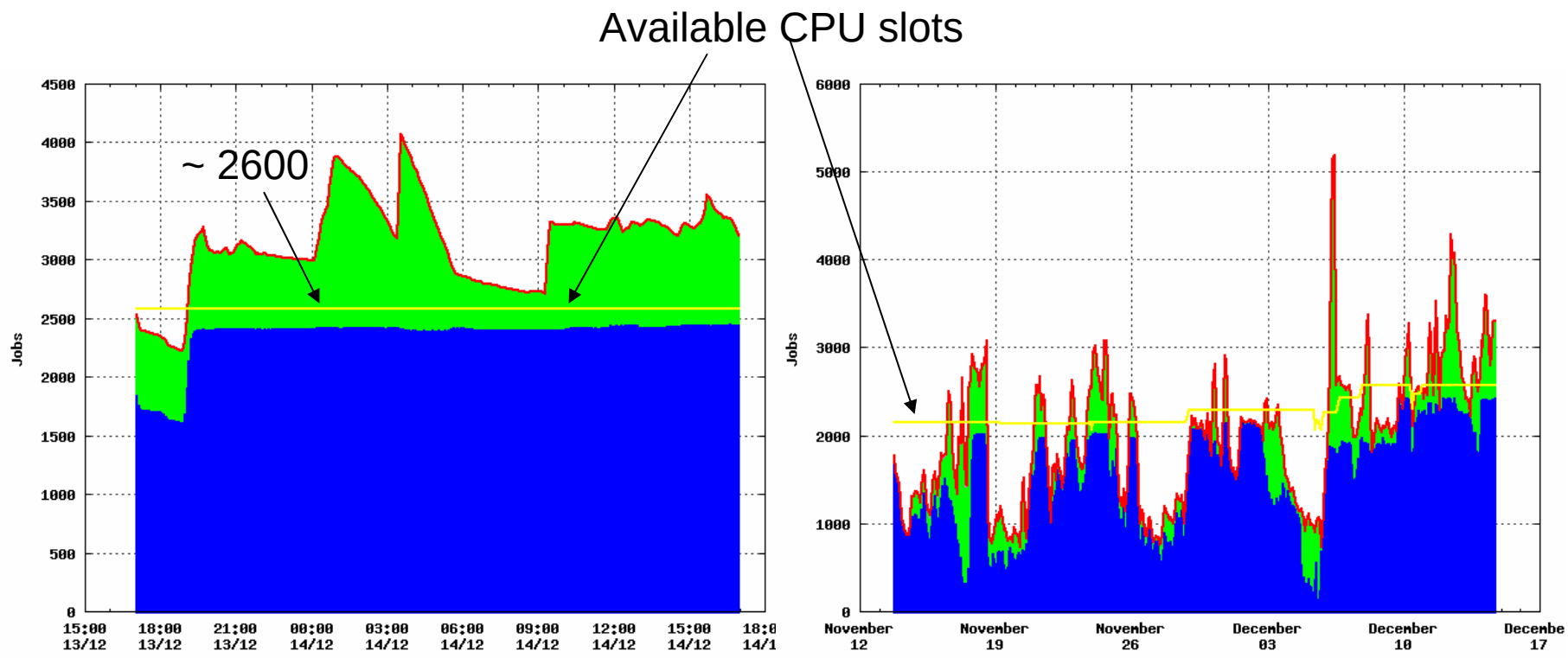
CPU Farm

- Farm installation and upgrades centrally managed by Quattor
- 1 general purpose farm (~750 WNs, 1600 KSI2k)
  - SLC 3.0.x, LCG 2.7
  - Batch system: LSF 6.1
    - Accessible both from Grid and locally
  - ~2600 CPU slots available
    - 4 CPU slots/Xeon biprocessor (HT)
    - 3 CPU slots/Opteron biprocessor
  - **22** experiments currently supported
    - Including special queues like infngrid, dteam, test, guest
  - 24 InfiniBand-based WNs for MPI on a special queue
- Test farm on phased-out hardware (~100 WNs, 150 KSI2k)

LSF

- At least one queue per experiment
  - Run and Cpu limits configured for each queue
- Pre-exec script with e-mail report
  - Verify software availability and disk space on execution host on demand
- Scheduling based on fairshare
  - Cumulative CPU time history (30 days)
- No resources granted
  - *Inclusion of legacy farms completed*
  - Maximization of CPU slots usage

Farm usage



Last day

Last month

See presentation on monitoring and accounting on Wednesday for more details

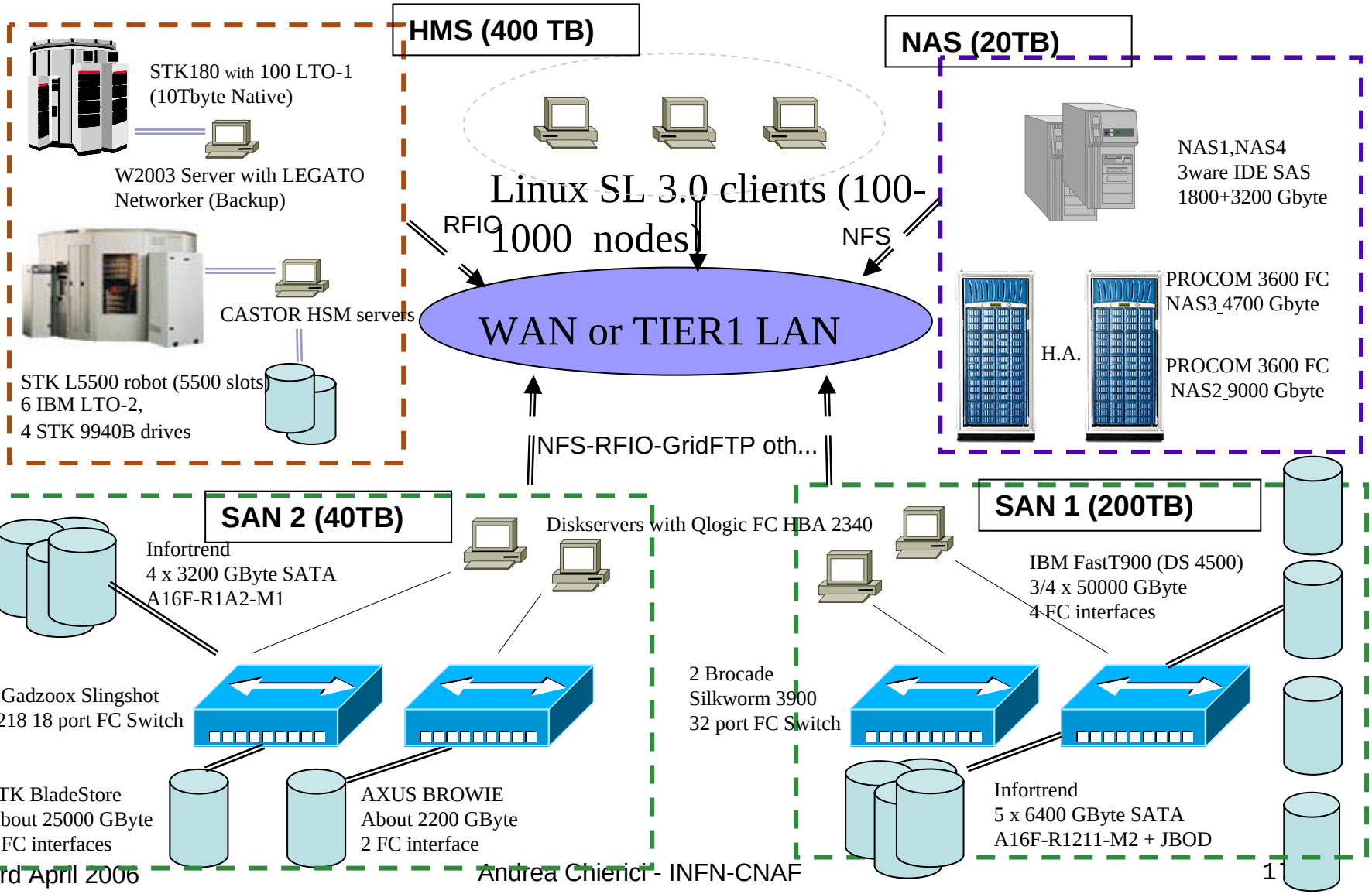
User Access

- T1 users are managed by a centralized system based on kerberos (authc) & LDAP (authz)
- Users are granted access to the batch system if they belong to an authorized Unix group (i.e. experiment/VO)
  - Groups centrally managed with LDAP
  - One group for each experiment
- Direct user logins not permitted on the farm
- Access from the outside world via dedicated hosts
  - *New anti-terrorism law making access to resources more complicated to manage*

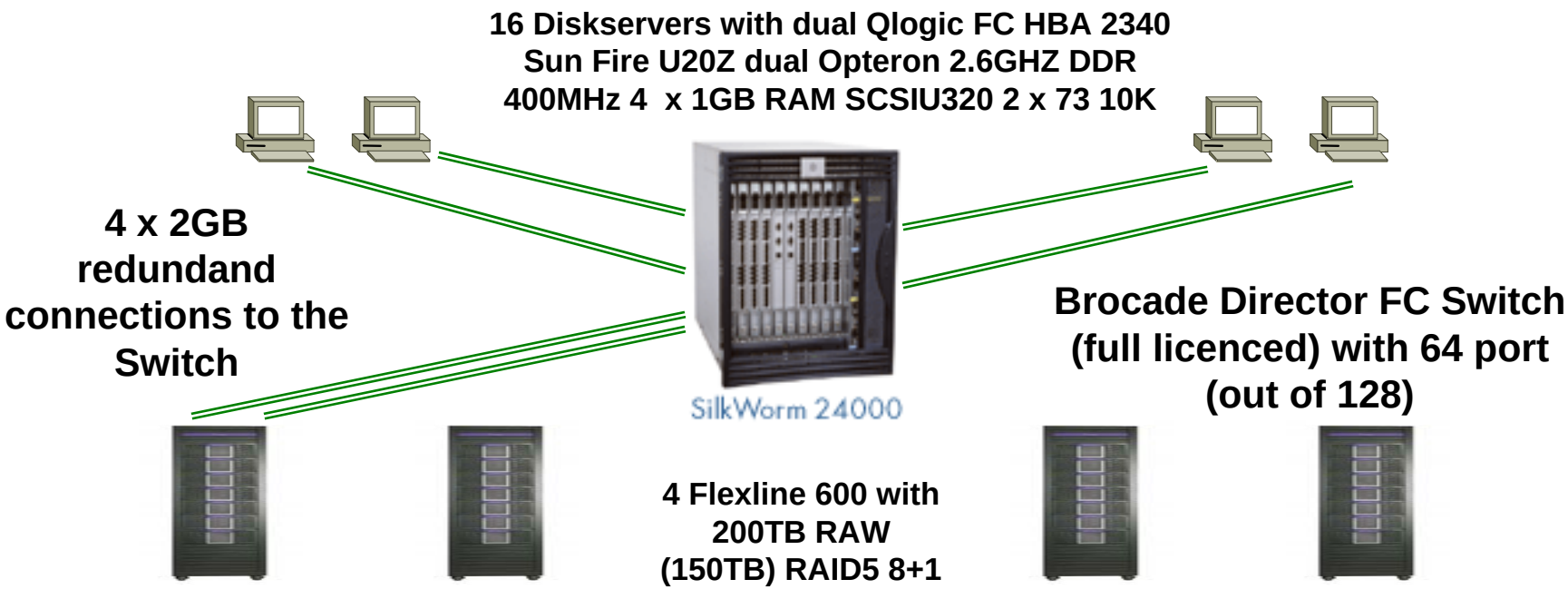
Grid access to INFN-Tier1 farm

- Tier1 resources can still be accessed both locally and via grid
  - Actively discouraging local access
- Grid gives opportunity to access transparently not only Tier1 but also other INFN resources
  - You only need a valid X.509 certificate
    - INFN-CA (http://security.fi.infn.it/CA/) for INFN people
  - Request access on a Tier1 UI
  - More details on http://grid-it.cnaf.infn.it/index.php?jobsubmit&type=1

Storage: hardware (1)

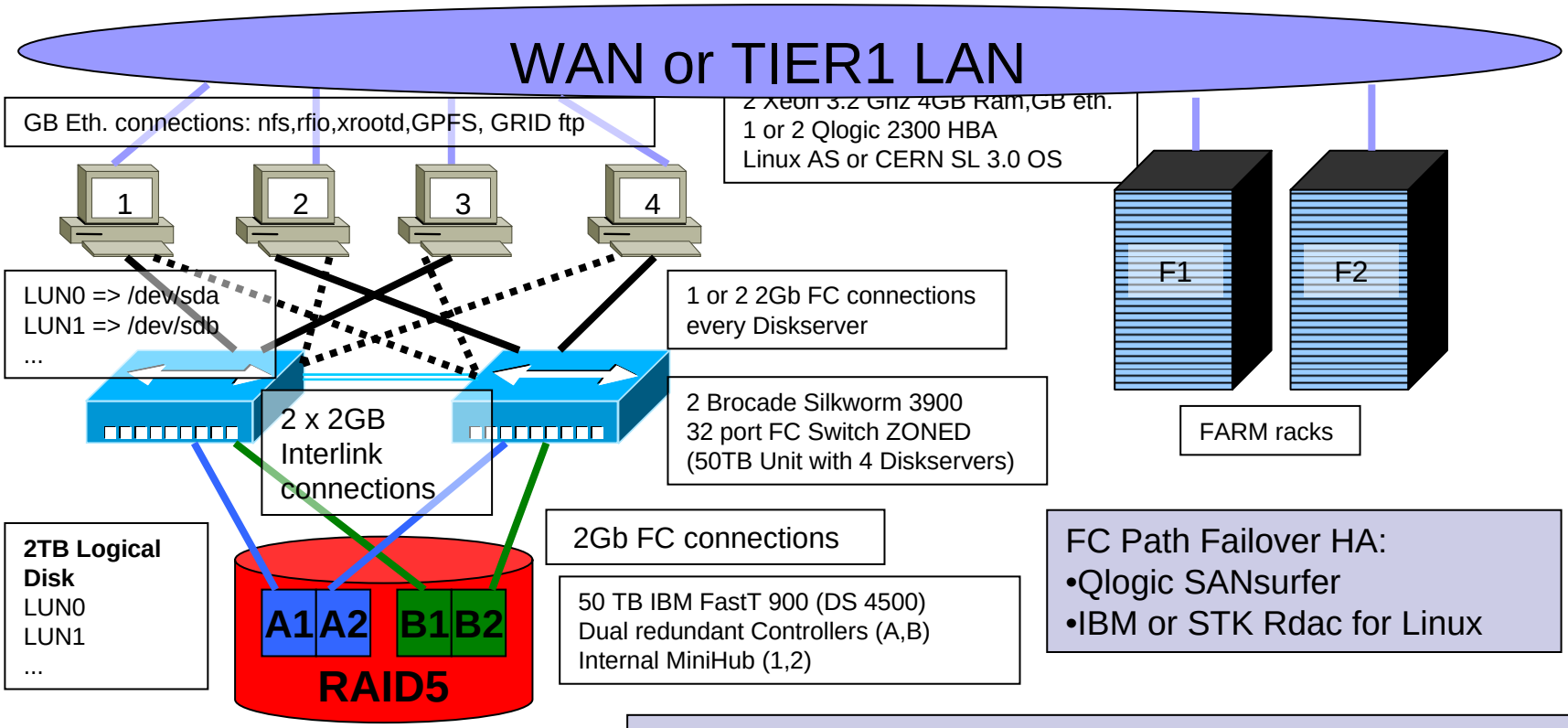


Storage: hardware (2)



- All problems now solved (after many attempts!)
  - Firmware upgrade
- Aggregate throughput 300 MB/s for each Flexline

DISK access



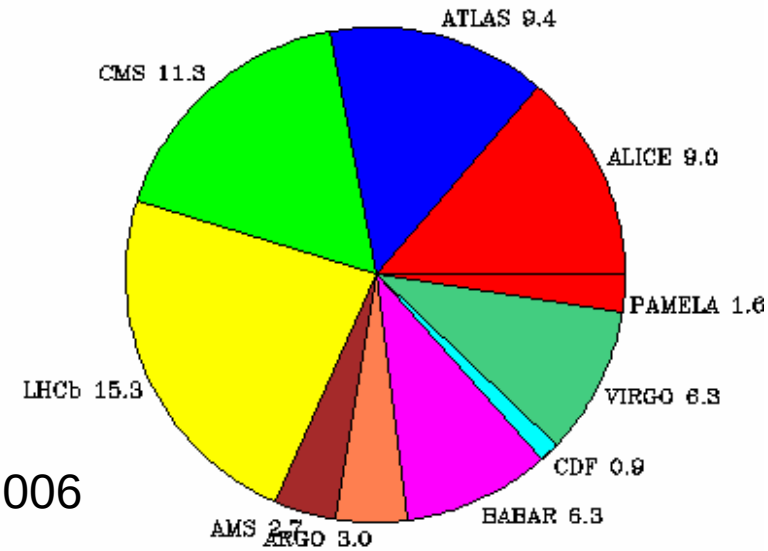
4 Diskservers every 50TB Unit: every controller can perform a maximum of 120MByte/s R-W

Application HA:
•NFS server, rfio server with Red Hat Cluster AS 3.0(*)
•GPFS with configuration NSD Primary Secondary
/dev/sda Primary Diskserver 1; Secondary Diskserver2
/dev/sdb Primary Diskserver 2; Secondary Diskserver3
(*) tested but not used in production yet

CASTOR HMS system (1)

- STK 5500 library
  - 6 x LTO2 drives
  - 4 x 9940B drives
  - 1300 LTO2 (200 GB) tapes
  - 650 9940B (200 GB) tapes
- Access
  - CASTOR file system hides tape level
  - Native access protocol: rfio
  - srm interface for grid fabric available (rfio/gridftp)
- Disk staging area
  - Data migrated to tapes and deleted from staging area when full
- Migration to CASTOR-2 ongoing
  - CASTOR-1 support ending around Sep 2006

Capacity



CASTOR disk space

CASTOR HMS system (2)

STK L5500 2000+3500 mixed slots

6 drives LTO2 (20-30 MB/s)

4 drives 9940B (25-30 MB/s)

1300 LTO2 (200 GB native)

650 9940B (200 GB native)

Point to Point FC 2Gb/s connections

8 tapeserver

Linux RH AS3.0

HBA Qlogic 2300

1 CASTOR (CERN)Central Services server RH AS3.0

1 ORACLE 9i rel 2 DB server RH AS 3.0

Sun Blade v100 with 2 internal ide disks with software raid-0 running ACSLS 7.0 OS Solaris 9.0

WAN or TIER1 LAN

6 stager with diskserver RH AS3.0

15 TB Local staging area

8 or more rfio diskservers

RH AS 3.0 min 20TB staging area

SAN 2

SAN 1

Indicates Full redundancy FC 2Gb/s connections (dual controller HW and Qlogic SANsurfer Path Failover SW)

Other Storage Activities

- dCache testbed currently deployed
  - 4 pool servers w/ about 50 TB
  - 1 admin node
  - 34 clients
  - 4 Gbit/sec uplink
- GPFS currently under stress test
  - Focusing on [LHCb] analysis jobs, submitted to the production batch system
    - 14000 jobs submitted, ca. 500 in simultaneous run state, all jobs completed successfully. 320 MByte/sec effective I/O throughput.
  - IBM support options still unclear
  - See presentation on GPFS and StoRM in the file system session.

DB Service

- Active collaboration with 3D project
- One 4-nodes Oracle RAC (test environment)
  - OCFS2 functional tests
  - Benchmark tests with Orion, HammerOra
- Two 2-nodes Production RACs (LHCb and ATLAS)
  - Shared storage accessed via ASM, 2 Dell PowerVault 224F, 2TB raw
- Castor2: 2 single instance DBs (DLF and CastorStager)
- One Xeon 2,4 with a single instance database for Stream replication tests on 3D testbed
- Starting deployment of LFC, FTS, VOMS readonly replica