“A prototype for INFN TIER-1 Regional Centre”

Luca dell’Agnello
INFN – CNAF, Bologna
Hepix Meeting
Catania, April 19 2002

INFN – TIER1 Project

- Computing facility for INFN HNEP community
  - Usage by other countries will be regulated by a Mutual Agreements
- Multi-Experiment TIER1
  - LHC experiments (ALICE, ATLAS, CMS, LHCb)
  - VIRGO
  - CDF (in a near future)
- Resources assigned to Experiments on a Yearly Plan.
- Location: INFN-CNAF, Bologna (Italy)
  - one of the main nodes of GARR
- TIER2, TIER3 under development at other places
- INFN-TIER1 is a prototype!
  - 4th quarter 2003: End of project
  - Winter 2004: experimental phase revision and new master plan
  - 2004: TIER1 becomes fully operational

Experiments needs

- LHC experiments
  - CERN (TIER 0)
    - Mass Storage: 10 Peta Bytes (10^{15} B)/yr
    - disk: 2 PB
    - CPU: 2 MSI95 (PC today ~ 30SI95)
  - Multi-Experiment TIER1
    - Mass Storage: 3 PB/yr
    - disk: 1.5 PB
    - CPU: 1 M SI95
  - Networking Tier 0 --> Tier 1: 2 Gbps
- Other experiments (VIRGO, CDF)
  - To be defined

Services

- Computing servers (CPU FARMS)
- Access to on-line data (Disks)
- Mass Storage/Tapes
- Broad-band network access and QoS
- System administration
- Database administration
- Experiment specific library software
- Helpdesk
- Coordination with TIER0, other TIER1s and TIER2s

Issues

- Technical staff
  - Recruiting & Training
- Resource management
  - Minimization of manual operations
- Sharing of resources (network, CPU, storage, HR) among experiments
  - Resource use optimization
- Compatibility between tests and production activity
  - Technological tests for Tier-1
  - Prototype phase (LHC experiments)
  - Production phase (VIRGO)
- Integration with (Data)grid framework
  - interoperation
  - Common tool development and test

HR Resources

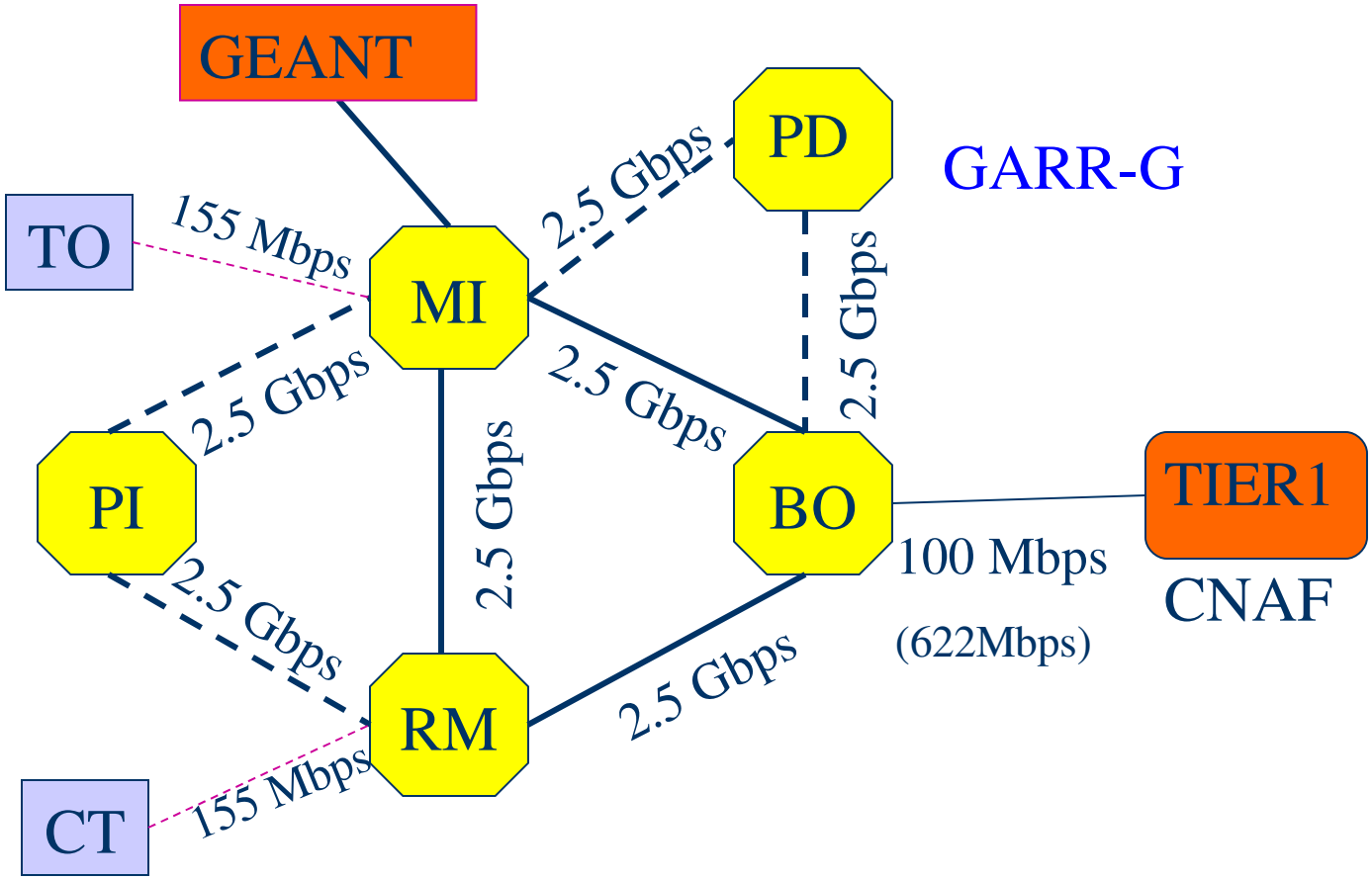
PERSONNEL

Type	N.	New	Outsource
Manager	1		
Deputy	1		
LHC Experiments Software	2		
Programs, Tools, Procedures	2	2	
FARM Management & Planning	2	2	
ODB & Data Management	2	1	
Network (LAN+WAN)	2	2	
Other Services (Web, Security, etc.)	2	1	
Administration	2	1	
System Managers & Operators	6		6
Total	22	9	6

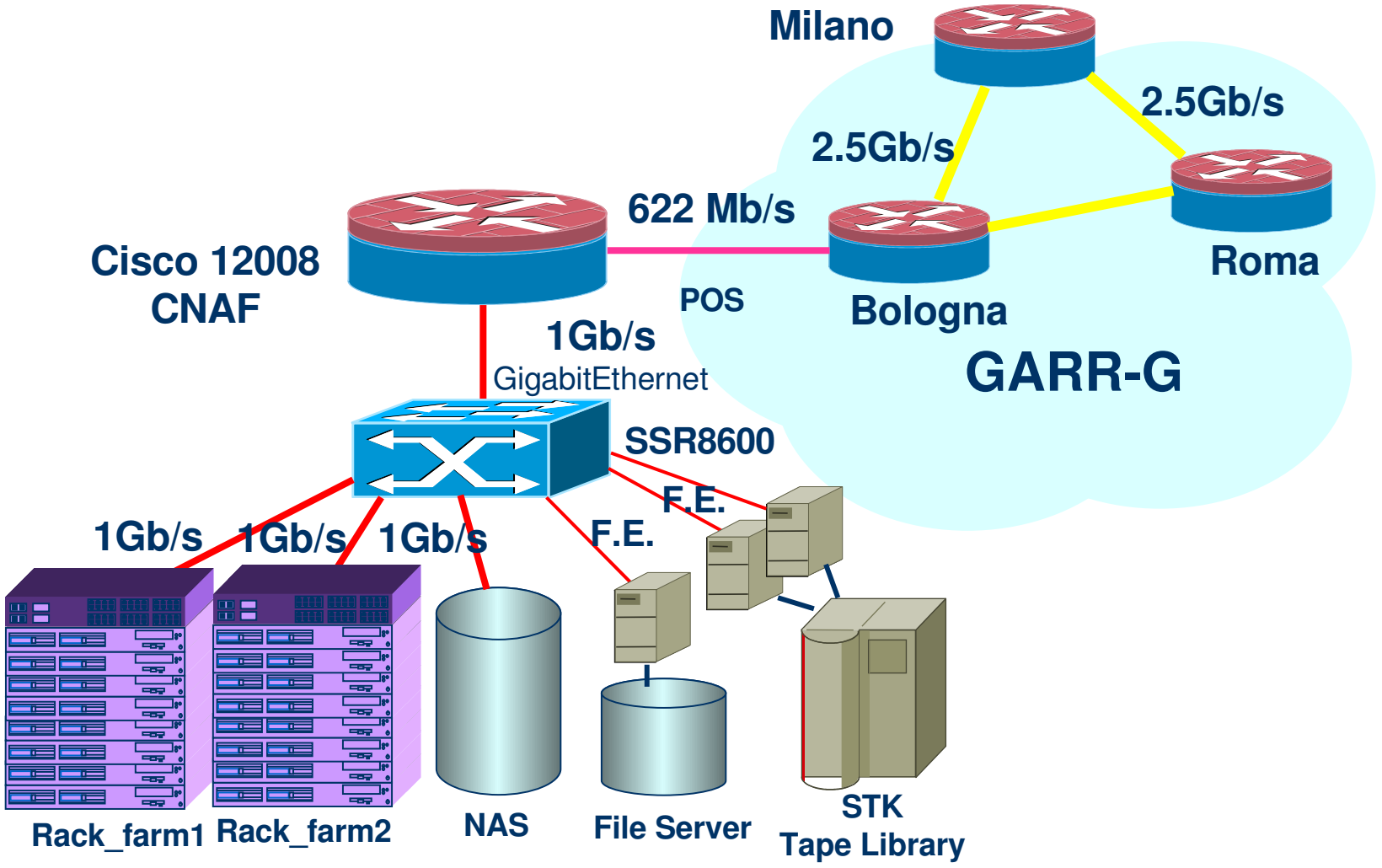
Networking (1)

- New GARR-G Backbone with 2.5 Gbps F/O lines already in place.
- CNAF-TIER1 access is now 100 Mbps and will be 622 Mbps in a few weeks
  - Gigapop is colocated with INFN-TIER1
- Many TIER2 are now 34 Mbps and will migrate soon to 155 Mbps.
- International Connectivity via Geant: 2.5 Gbps access in Milano and 2x2.5 Gbps links of Geant with US (Abilene+commodity) already in place.

Networking (2)



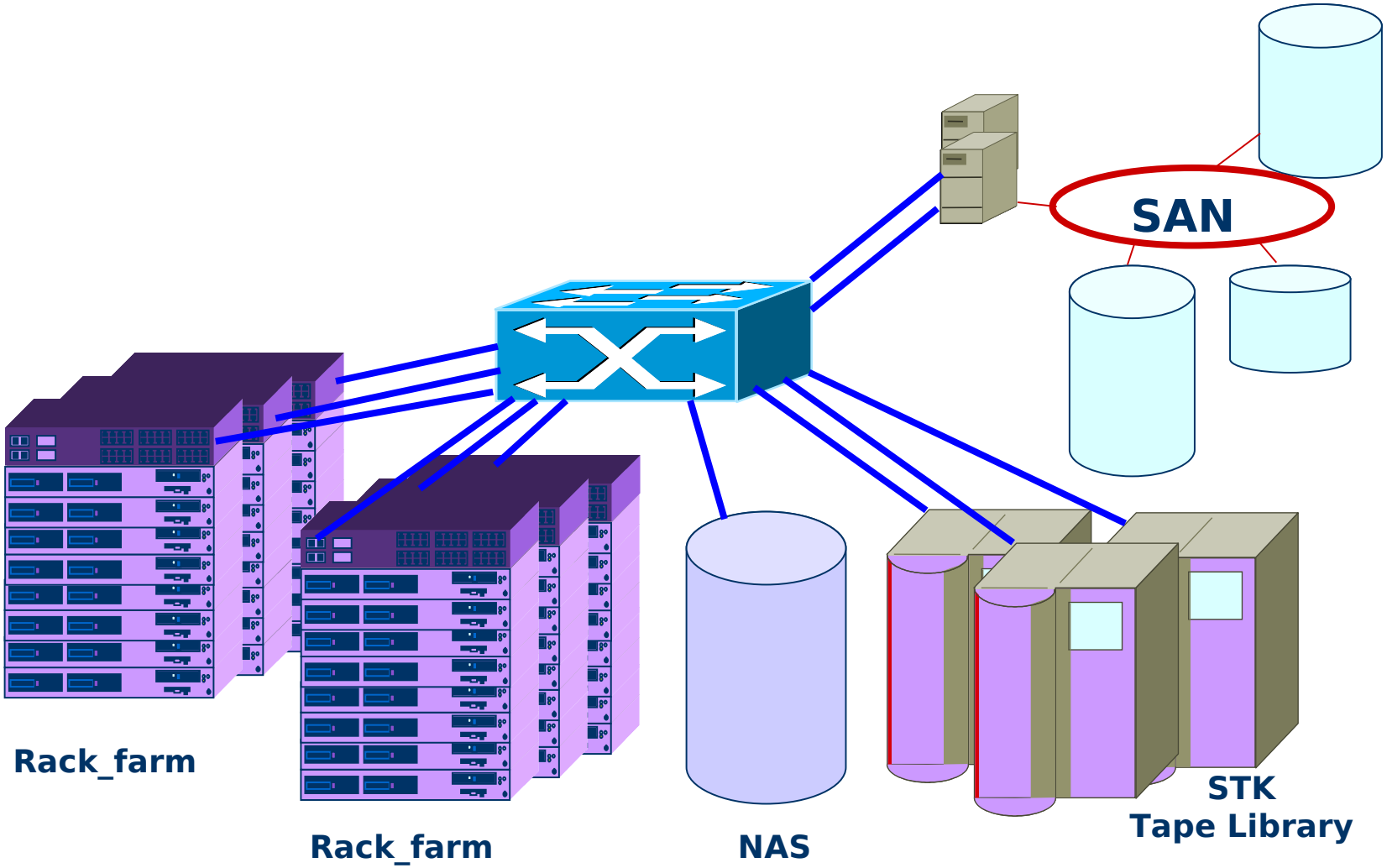
Interconnection to Internet (near future)



LAN

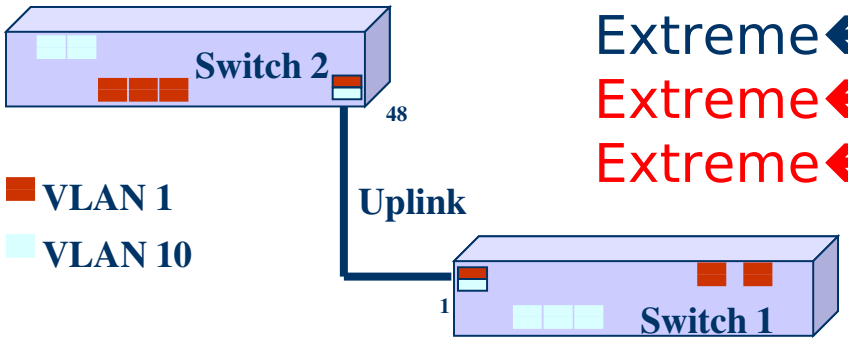
- Large amount of resources to be interconnected at high speed...
  - CPU's connected to rack FE switch with Gb uplink to core switch
  - Disk servers connected via GE to core switch
  - Foreseen upgrade to rack Gb switch/10G core switch (2003 ?)
- .... and shared among experiments
  - Possibility to reallocate each resource at every moment
  - Avoid recabling (or physical moving) of hw to change the topology
- Level 2 isolation of farms
  - Aid for enforcement of security measures

Tier1 LAN model layout



Vlan Tagging (1)

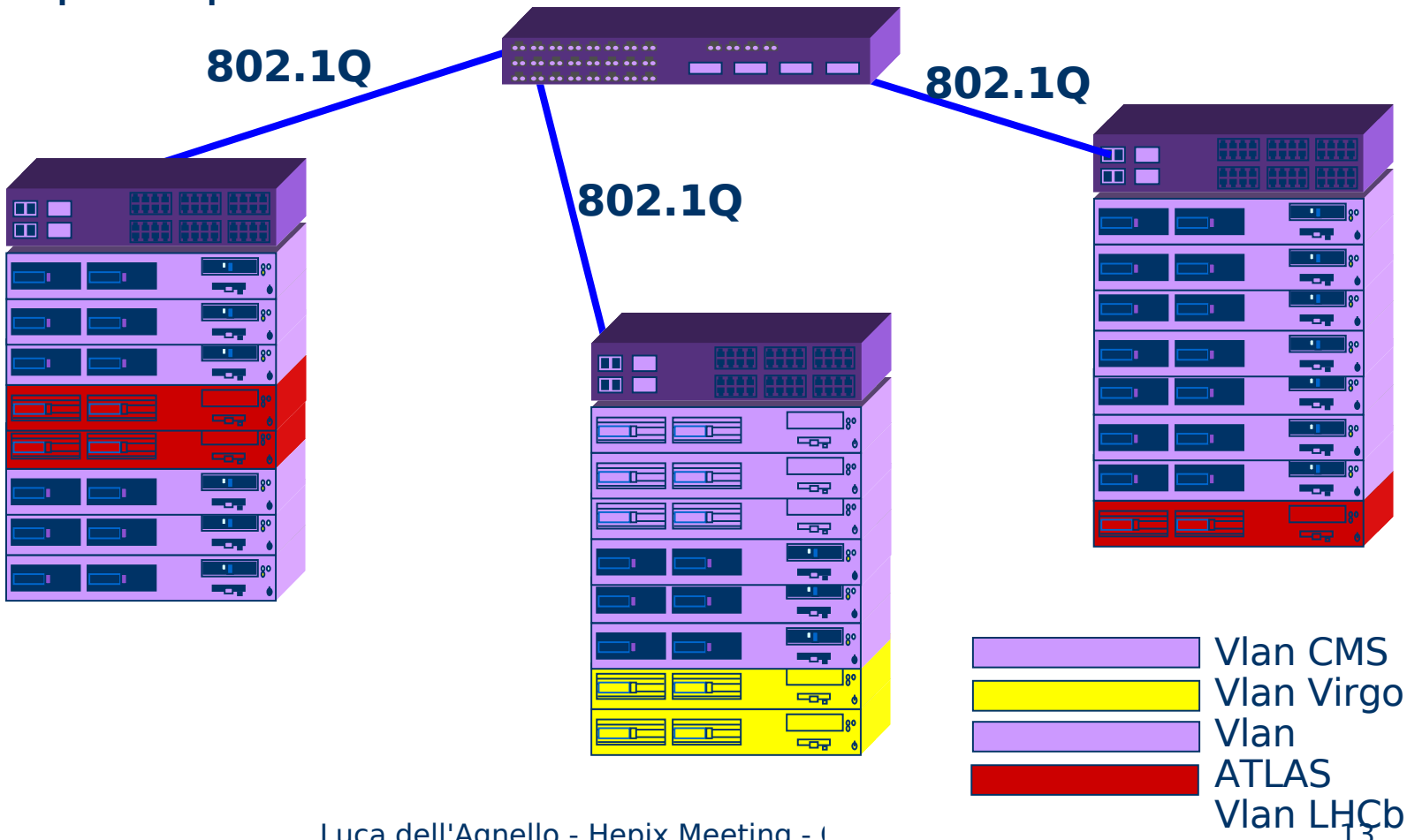
- Possible solution for complete granularity
  - To each switch port is associated one VLAN identifier
  - Each rack switch uplink propagates VLAN informations
  - VLAN identifiers are propagated across switches
  - Each farm has its own VLAN
- Independent from switch brand (Standard 802.1q)
- First interoperability tests show viability of solution



Estreme←→Extreme OK!
Extreme←→Enterasys OK!
Extreme←→Cisco tests ongoing
Extreme←→HP tests ongoi

Vlan Tagging (2)

- 1 Vlan per experiment
- 1 Uplink per rack



Computing units

- Basic unit
  - Intel CPU with Redhat Linux (Datagrid framework)
    - Different requirements from various experiments
      - RedHat 6.2 (moving to RedHat 7.2) + experiment specific libraries
  - 1U rack-mountable dual processor servers
    - 800 MHz - 1.4 GHz
    - 2 FE interfaces
    - 512 MB – 2 GB RAM
- Rack unit (what we buy)
  - 40 1U dual processor servers
  - 1 Fast Ethernet switch with Gigabit uplink to main switch (to be upgraded in a next future)
  - Remote control via KVM switch (tests with Raritan ongoing)
- A new bid (1 rack) is in progress

Brand choice

- “Certification” of platforms
  - Mechanical and assembling aspects
  - Hardware characteristics (e.g. cooling system, PCI slots number, max RAM etc..)
  - Hardware benchmarks (I/O etc..)
  - PXE protocol support
    - Installation tests
  - Compatibility with RedHat Linux
- Tested platforms:
  - Proliant 360 (COMPAQ) 5
  - Netfinity X330 (IBM) 12
  - PowerEdge 1550 (DELL) 48
  - INTEL (various OEM’s)
  - SIEMENS

Computing units issues (1)

- Coexistence Datagrid/Datatag test-beds – “traditional” installations
  - Need to develop tools to manage non-grid servers
- Dynamic (re)allocation of server pools as experiments farms
  - Automatic procedure for installation & upgrade
    - LCFG (developed by Datagrid WP4)
      - Central server for configuration profiles
      - Use of standard services (NFS, HTTP)
      - Only RedHat 6.2 currently supported
      - First boot from floppy
    - LCFG+PXE protocol (only a quick patch!)
      - No floppy needed

Computing units issues (2)

- Resource database and management interface
  - Under development (php + mysql)
  - Hw servers characteristics
  - Sw servers configuration
  - Servers allocation
  - Possibly interface to configure switches and prepare LCFG profiles
- Monitoring system for central control
  - Ganglia (+ lmsensors)
  - Proprietary system (e.g. DELL) under consideration
  - Generic tool using SNMP under development
- Batch system (non Datagrid servers)
  - PBS
  - Condor

Storage (1)

- Comparison between NAS and SAS architectures
  - SAS has Linux limits (e.g. problem with large volumes) but can be customized
  - NAS with proprietary OS can be optimized
- Choice of Fiber Channel
  - Flexibility
- Access to disk servers via Gigabit Ethernet
  - Access via NFS v.3 (AFS considered, NFS v.4 in a near future)
  - Tests for HA (fail-over) ongoing
- Legato Networker for user Backup (on L180)

Storage (2)

- Use of staging sw system for archive
  - Installed and tested CASTOR under Linux 7.2 configuration CLIENT/SERVER with a a single tape connected
    - Waiting for the Storagetek CSC Toolkit for starting test with the ACSLS software and the STK180 Library
- Near future tests:
  - Study of volume managers tools for better space organization and server fail-over (Veritas, GFS,GPFS…)
  - Study of a SAN solution (F.C. switches)
  - Integration of the NAS SAS solutions
  - Test and comparison between disk solutions (IDE Raid array, SCSI and F.C. Raid array)
  - Tests with TAPE DISPATCHER staging software (developed by A. Maslennikov and R. Gurin)

Storage resources (1)

- NAS Procom
  - 2 TB *raw*
  - NFS v.3
  - 24 FC disks (72 GB)
    - Upgrade to 16 TB in a 2-3 months
  - Network interface FE, GE
  - One RAID 5 volume
- SAS Raidtec
  - 2 TB *raw*
  - 12 SCSI disks (180 GB)
  - Raid controller

Storage resources (2)

- SAN Dell Power Wault
  - 8 TB *raw*
  - Fiber channel
  - To be installed
- Library STK L180
  - 180 slots, 4 drives LTO, 2 drives 9840
    - 100 tapes LTO (100 GB)
    - 180 tapes 9840 (20 GB)

Security issues (1)

- Need a centralized control for resource access
  - 1 FTE required
- DataGrid AAA infrastructure based on PKI
  - Authentication via certificates (well established procedure)
  - Authorization currently in evolution
    - presently via “gridmap-file”
    - CAS (tests ongoing) does not seem to solve all issues
    - LCAS (tests ongoing) to map authorization at local level
    - Some “glue” needed
- Interim solutions
  - Ssh, sftp, bbftp
  - Bastion host at present
  - Kerberos (v. 5) in a few weeks if suitable (presently in test)

Security issues (2)

- LAN access control
  - Packet filters on border router
    - Use of more sophisticated firewalls to be considered
  - Limit traffic to known active services
  - Centralized log for automatic filtering
  - NIDS under consideration
    - Requires manpower!
- Servers configuration
  - Completely under our control
  - Use on-board firewall
    - Filter all unnecessary ports
  - Upgrade of vulnerable packages
    - RedHat Network Alerts, CERT alerts etc..

Conclusions

- INFN-TIER1 has began an experimental service.....
  - VIRGO, CMS, ATLAS, LHCb
  - Datagrid test-bed
  - Datatag test-bed
- .... but we are still in a test phase
  - Study and tests technological solutions
- Main goals of the prototype are:
  - Train people
  - Adopt standard solutions
  - Optimize resource use
  - Integration with the GRID