# INFN Tier1

Andrea Chierici
INFN – CNAF, Italy
LCG Workshop
CERN, March 23-24 2004

# INFN – Tier1

- INFN computing facility for HNEP community
  - Location: INFN-CNAF,  Bologna (Italy)
    - One of the main nodes on GARR network
  - Ending prototype phase last year, now fully operational
  - Personnel: ~ 10 FTE's
    - ~ 3 FTE's dedicated to experiments
- Multi-experiment
  - LHC experiments, Virgo, CDF, BABAR, AMS, MAGIC, ...
  - Resources dynamically assigned to experiments according to their needs
- Main (~50%) Italian resource for LCG
  - Coordination with Tier0 and other Tier1 (management, security, etc...)
  - Coordination with other Italian tiers
  - Participation to grid test-beds (EDG, EDT, GLUE)
  - Participation to experiments data challenge
  - ROC + CIC in EGEE (deployment in progress)

# Logistics

- Recently moved to a new location (last January)
  - Hall in the basement (-2nd floor)
  - ~ 1000 $m^2$ of total space
    - Computers
    - Electric Power System (UPS, MPU)
    - Air conditioning system
    - *Garr GPop*
  - Easily accessible with lorries from the road
  - Not suitable for office use (remote control)

# Electric Power

- 380 V three-phase distributed to all racks
  - rack power controls output 3 independent 220 V lines for computers
    - 16 A or 32 A
  - 3 APC power distribution modules (24 outlets each)
    - Completely programmable (allows gradual servers switching on)
    - Remotely manageable via web
- 380 V three-phase for other devices (tape libraries, air conditioning, etc…)
- Uninterruptible Power Supply (UPS)
  - Located into a separate room (conditioned and ventilated)
  - 800 KVA (~ 640 KW)
- Electric Power Generator
  - 1250 KVA (~ 1000 KW)
    - ➔ up to 80-160 racks

# Cooling & Air Conditioning

- RLS (Airwell) on the roof
  - ~ 700 KW
  - Water cooling
  - Need "booster pump" (20 mts T1 ←→ roof)
  - Noise insulation
- 1 Air Conditioning Unit (uses 20% of RLS refreshing power and controls humidity)
- 9 (→ 12) Local Cooling Systems (Hiross) in the computing room ~ 30 KW each

# Networking (1)

- Network infrastructure using optical fibres (~ 20 Km)
  - To insure a better electrical insulation on long distances
  - To ease adoption of new (High Performances) transmission technologies
  - To decrease the number of cables in and out from each rack
  - Local links with UTP (copper) cables
- LAN has a "classical" star topology
  - GE core switch (Enterasys ER16)
  - Servers directly connected to GE switch (mainly fibre)
    - Disk servers connected via GE to core switch
    - Some servers concentrated with copper cables to a dedicated switch
  - Farms up-link via GE trunk to core switch

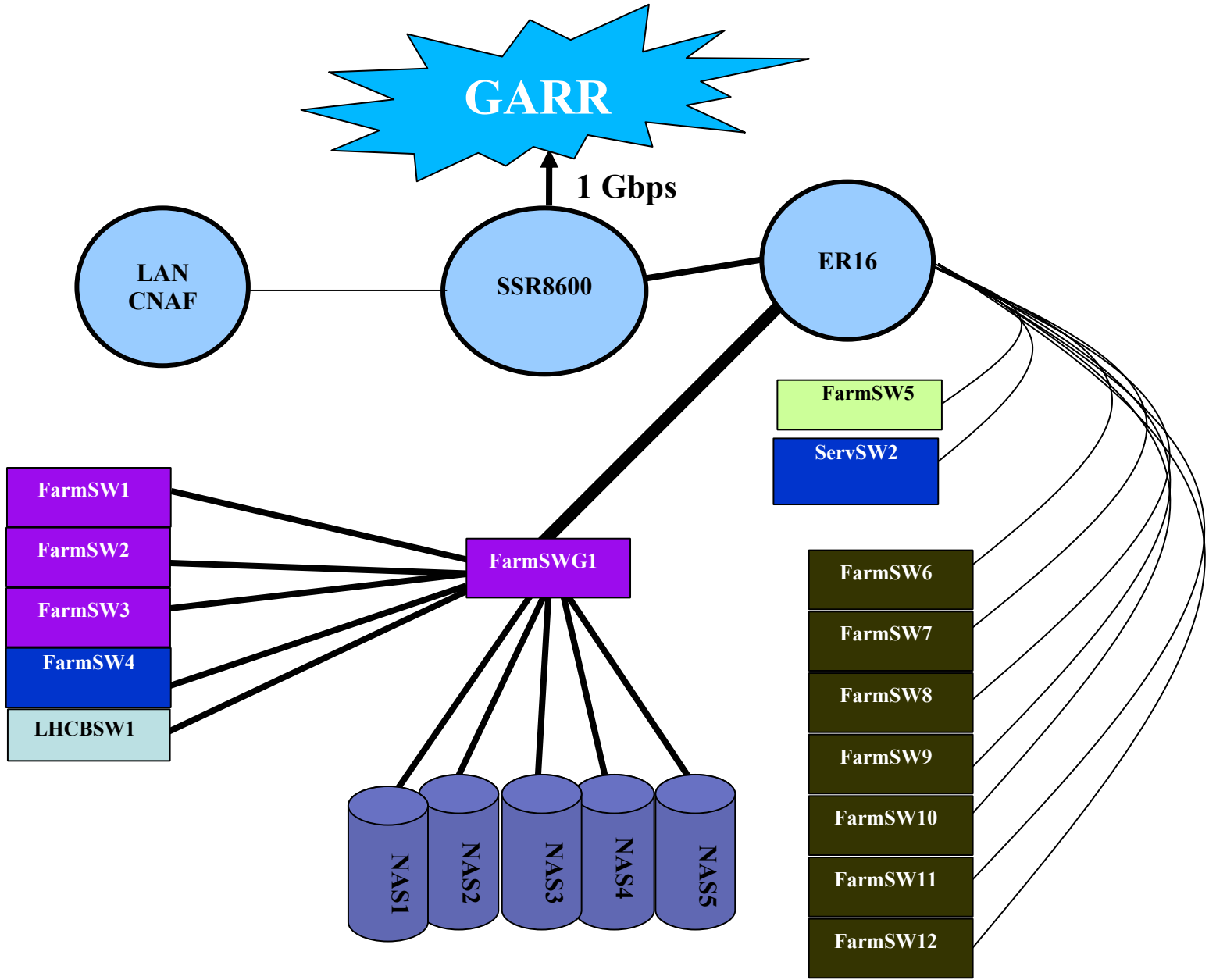
# Networking (2)

- WN's connected via FE to rack switch (1 switch per rack)
  - Not a single brand for switches (as for wn's)
    - 3 Extreme Summit 48 FE + 2 GE ports
    - 3 3550 Cisco 48 FE + 2 GE ports
    - 8 Enterasys 48 FE 2GE ports
  - Homogeneous characteristics
    - 48 FastEthernet ports
    - Support of main standards (e.g. 802.1q)
    - 2 Gigabit up-links (optical fibers) to core switch
  - Moving to 1U 48 GE ports switches (10 Gbps ready)
    - Extreme Summit 400 (to be installed)
- CNAF interconnected to GARR-G backbone at 1 Gbps.
  - Giga-PoP co-located
  - GARR-G backbone at 2.5 Gbps
  - 2 x 1 Gbps test links to CERN, Karlsruhe

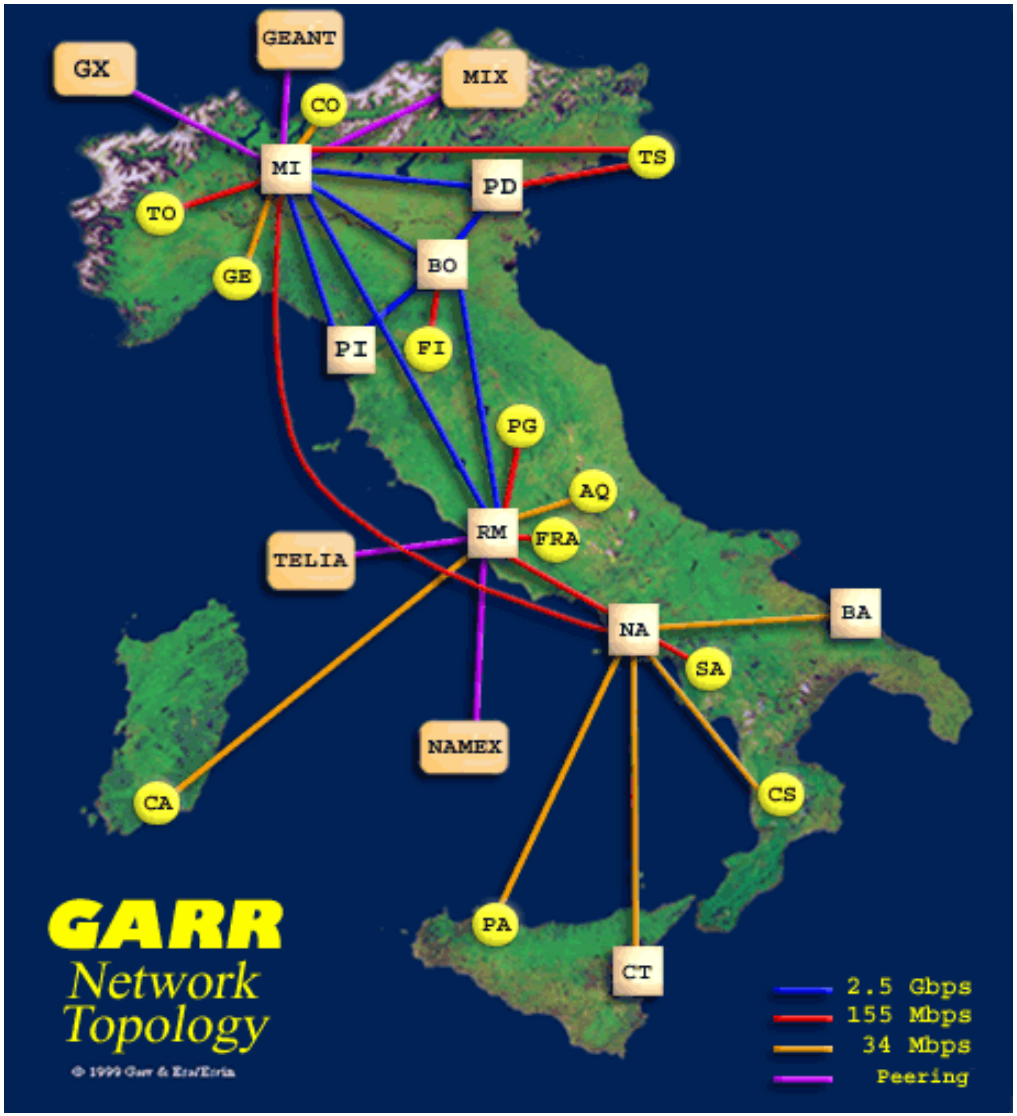
# L2 Configuration

- VLAN's defined across switches
  - Independent from switch brand (Standard 802.1q)
- Solution adopted for complete granularity
  - Each switch port has one VLAN identifier associated
  - Each rack switch uplink propagates VLAN information
  - VLAN identifiers are propagated across switches
  - Each farm has its own VLAN
  - Avoid recabling (or physical moving) of machines to change farm topology
- Level 2 isolation of farms
- Possibility to define multi-tag ports (for servers)

# Network layout



# GARR Topology



# Computing units

- ~ 320 1U rack-mountable Intel dual processor servers
  - 800 MHz – 2.4 GHz
  - ~ 240 wn's (~ 480 CPU's) available for LCG
- 350 1U bi-processors Pentium IV 3.06 GHz to be shipped April 2004
- Testing Opteron farm
- OS: Linux RedHat 7.3
  - Experiment specific library software
- Goal: have generic computing units experiments independent
  - Taken LCG WN as a starting point
  - Experiment specific library software in standard places (e.g. /opt/exp_software/cms)
  - Simple storage mount points (e.g. /LHCb/01)

# Installation issues

- Centralized installation system
  - LCFG (EDG WP4)
  - Integration with a central Tier1 db (see below)
  - Each farm on a distinct VLAN
    - Moving from a farm to another implies just changes in IP address (not name)
  - Unique dhcp server for all VLAN's
  - Support for DDNS (cr.cnaf.infn.it)
- Investigating Quattor for future needs

# Tier1 Database

- Resource database and management interface
  - Hw servers characteristics
  - Sw servers configuration
  - Servers allocation
  - Postgres database as back end
  - Web interface (apache+mod_ssl+php)
- Possible direct access to db for some applications
  - Monitoring system
  - Nagios
- Interface to configure switches and interoperate with installation system
  - Vlan tags
  - dns
  - dhcp

# Batch system

- torque+maui (http://www.supercluster.org)
  - 50% of resources statically assigned to experiments (1 VO = 1 queue)
  - 50% of resources in "overflow" queue shared among experiments
  - Still looking for the "ultimate configuration" to maximize CPU utilization and to simply management
- Problems with single LCG gatekeeper
  - Publishing feed queue (1 CE) fails to report correct information in IS
  - Now several queues announced
    - LCFGng does not support this configuration

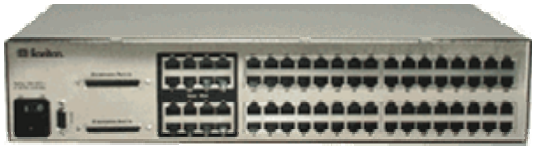
# Monitoring/Alarms

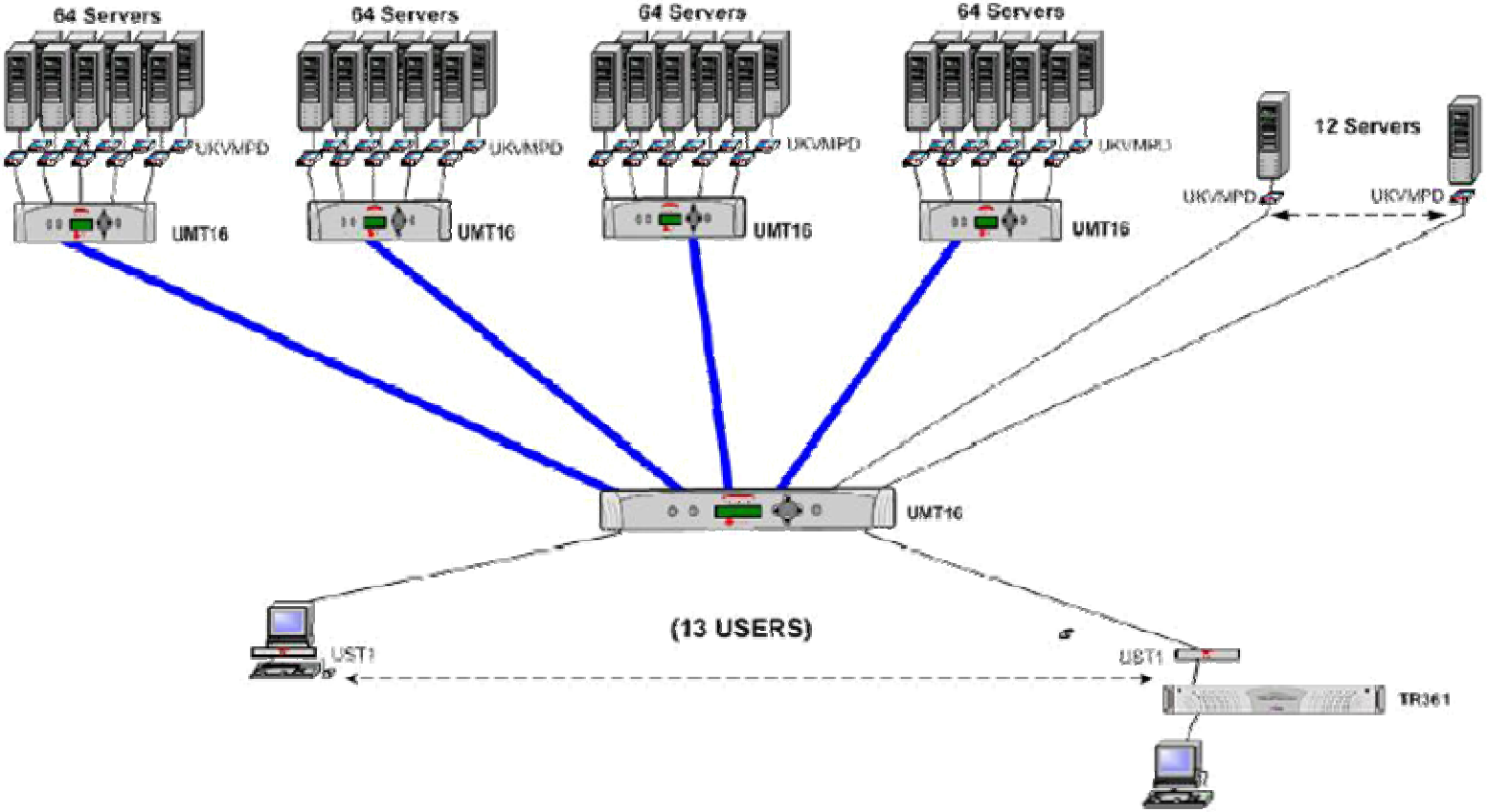
- Monitoring system developed at CNAF
  - Centralized collector
  - ~100 variables collected every 5 minutes
    - Data archived on flat file
      - In progress: XML structure for data archives
  - User interface: http://tier1.cnaf.infn.it/monitor/
    - Next release: JAVA interface (collaboration with D. Galli, LHCb)
- Critical parameters periodically checked by ***nagios***
  - Connectivity (i.e. ping), system load, bandwidth use, ssh daemon, pbs, etc…
  - User interface: http://tier1.cnaf.infn.it/nagios/
  - In progress: configuration interface

# Remote console control

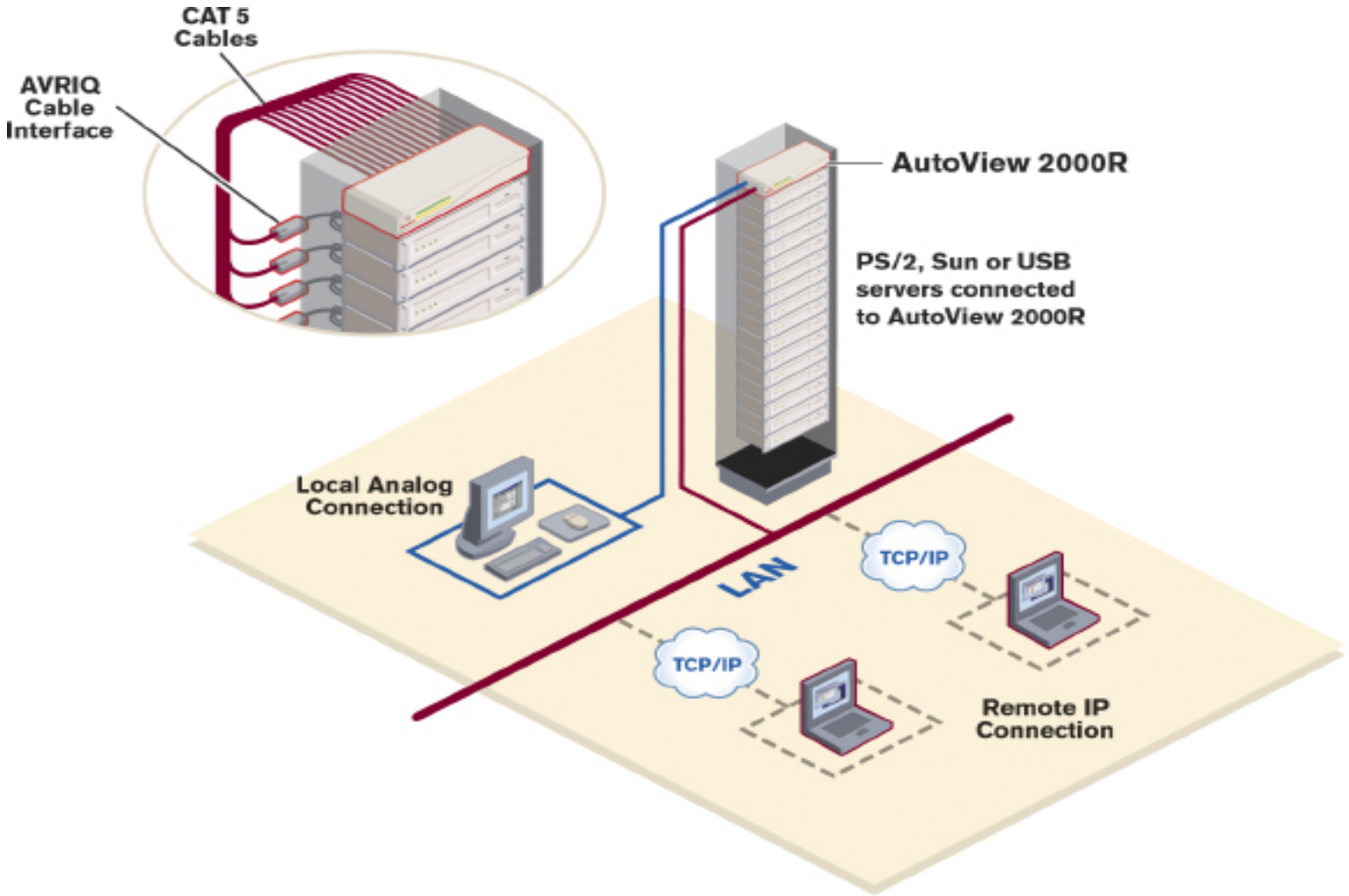
- Paragon UTM8 (Raritan)
  - 8 Analog (UTP/Fiber) output connections
  - Supports up to 32 daisy chains of 40 nodes (UKVMSPD modules needed)
  - Costs: 6 KEuro + 125 Euro/server (UKVMSPD module)
  - IP-reach (expansion to support IP transport) evaluted but not used
- Autoview 2000R (Avocent)
  - 1 Analog + 2 Digital (IP transport) output connections
  - Supports connections up to 16 nodes
    - Optional expansion to 16x8 nodes
  - Compatible with Paragon ("gateway" to IP)
- Evaluating Cyclades Alterpath KVM via serial line: cheaper
- APC Network Power Controls: allows remote and scheduled power cycling via snmp calls or web
  - Remote control via IP up to 24 outlets/module

# Raritan




64 Servers
64 Servers
64 Servers
64 Servers
UKVMPD
UKVMPD
UKVMPD
UKVMPD
UMT16
UMT16
UMT16
UMT16
12 Servers
UKVMPD
UKVMPD
UMT16
(13 USERS)
UST1
UST1
TR361

# Avocent


CAT 5
Cables
AVRIQ
Cable
Interface
AutoView 2000R
PS/2, Sun or USB
servers connected
to AutoView 2000R
Local Analog
Connection
LAN
TCP/IP
TCP/IP
Remote IP
Connection

# Storage (1)

- Access to on-line data: DAS, NAS, SAN
  - 50 TB
  - Data served via NFS v3
- Several hw technologies used (EIDE, SCSI, FC)
- Ongoing study of large file system solutions (>2TB) and load balancing/failover architectures
  - GPFS (load balancing, large file systems)
    - Not that easy to install and configure!
- High Availability for NFS servers
    - RedHat AS 2.1 (Redhat AS 3.0 soon)
    - Tested and ready for production

# Storage (2)

- "SAN on WAN" tests  (collaboration with CASPUR)
  - Iscsi
  - FC over IP on WAN
  - Peak throughput: 100 MB/s (RAID 0)
- Tests with PVFS (LHCb, Alice)
  - Easy to configure and install
  - Not yet production quality software
- SAN as architectural choice
  - Currently ~ 30 TB available
  - Soon ~ 200 TB more

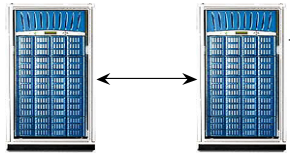
# DISK resources at TIER1

Direct SCSI Data access

RaidTec Flex Array HI 160
Ultra 160 2 LVD SCSI Channels,
**TOT: 2 TB**

-LINUX 7.2 Compaq 360 raid 1 server
-CASTOR
-Adaptec SCSI 3 39160 LVD 64 bit 160 MB/s dual port VHDCI output port.

TCP/IP Fast Ethernet 100 or 1Gb/s optical



Procom NAS 2 heads
4 Gigabit Ethernet Interface
**TOT: 17 TByte (raw)**

TCP/IP Gigabit Ethernet 1Gb/s Optical

NETWORK

DELL PowerVaul 660F dual Fibre Channel interface
+ 5 DELL PowerVault 224F Fibre Channel Disk Unit
**TOT: 8 Tbyte (raw)**

Fibre Channel Copper 1 Gbit/s

SAN

-1 LINUX DELL 1650 server
Qlogic 2300 Fibre Channel Copper Adapter
-Gigabit Ehernet Adapter Optical

TCP/IP Gigabit Ethernet 1Gb/s Copper

# DISK resources at TIER1 (2)

Assembled NAS 4 IDE disk
3ware interface
2 Gigabit Ethernet Interface
**TOT: 2.5TByte (raw)**

2 TCP/IP Gigabit
Ethernet 1Gb/s
Optical

AXUS BROWIE
2 Fibre Channel
interface
**TOT: 2.5 TByte (raw)**

Fibre
Channel
Copper
2 Gbit/s

SAN

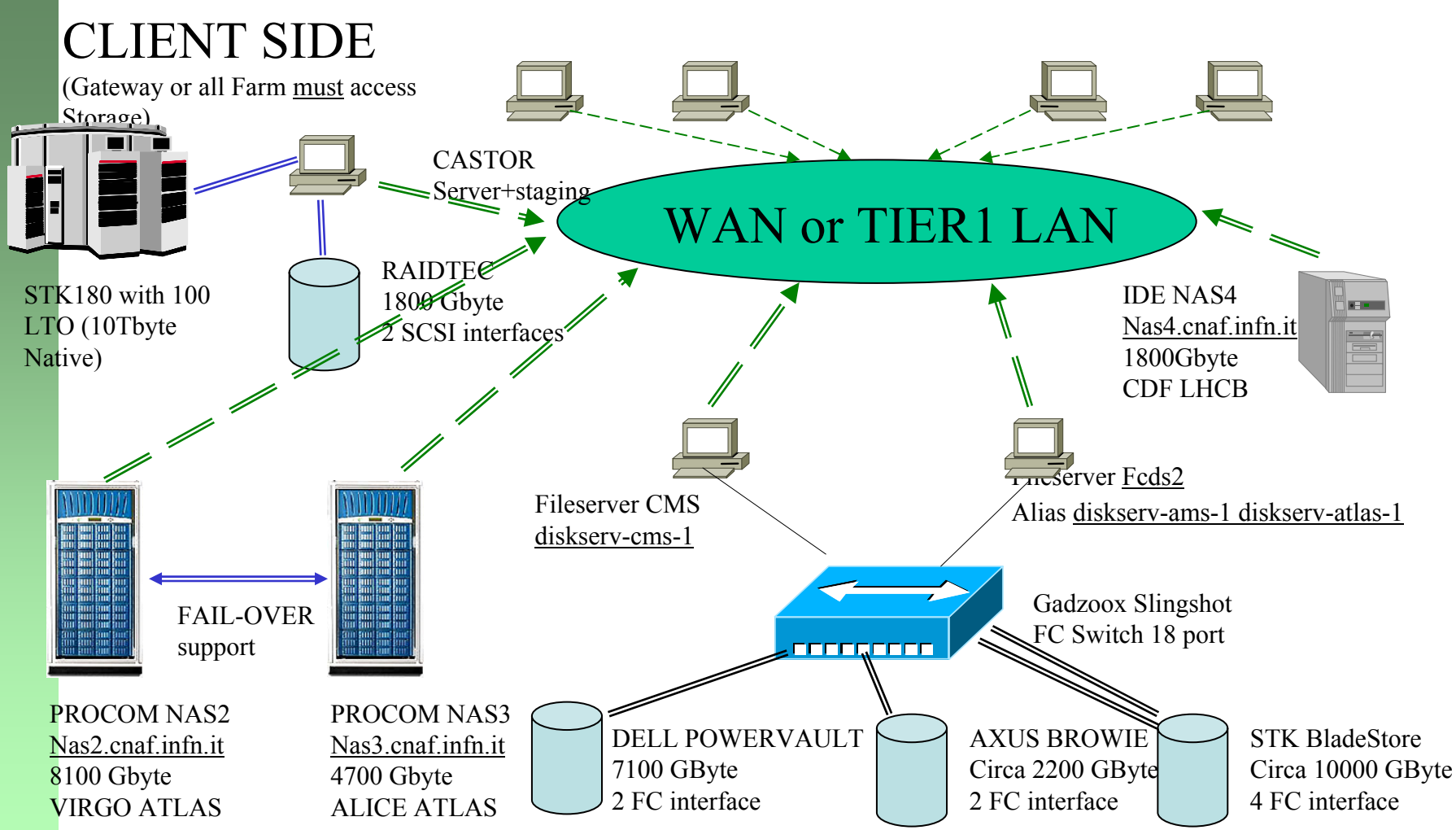
NETWORK

-1 LINUX DELL 1550 server
-LSI LOGIC Fibre Channel
Copper Adapter
-Gigabit Ehernet Adapter
Optical

STK BLADESTORE
4 Fibre Channel
interface
**TOT: 12TByte (raw)**

Fibre
Channel
optical
2 Gbit/s

# STORAGE resource



# CASTOR resource at TIER1 (1)

- 1 COMPAQ DL360 Redhat 7.2 Machine (castor) running CASTOR 1.6.1.3 services + ORACLE database (Nsdaemon, vmgrdaemon, Cupvdaemon, vdqmdaemon, Cdbdaemon, msgdaemon, stgdaemon, tpdaemon, rtcpd, rfiod, ctpdaemon) with SCSI connection with 2 LTO-1 drive on STK L180

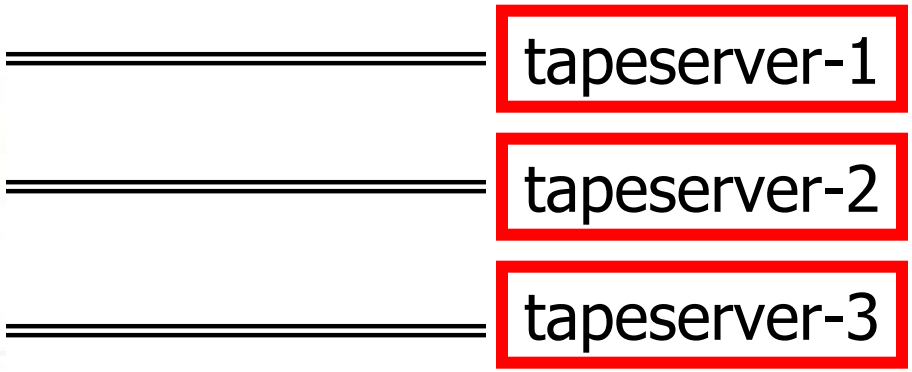
- DELL 1650 RedHat 7.2 Machine (castor-1) running 2 CASTOR services (rtcpd, tpdaemon) + ORACLE Recovery Manager Database (needed for LEGATO backup of the db) with SCSI connection with 2 LTO-1 drive on STK L180

- Both Machines run the STK CDC Dev. Toolkit (ssi)

# CASTOR resource at TIER1 (2)

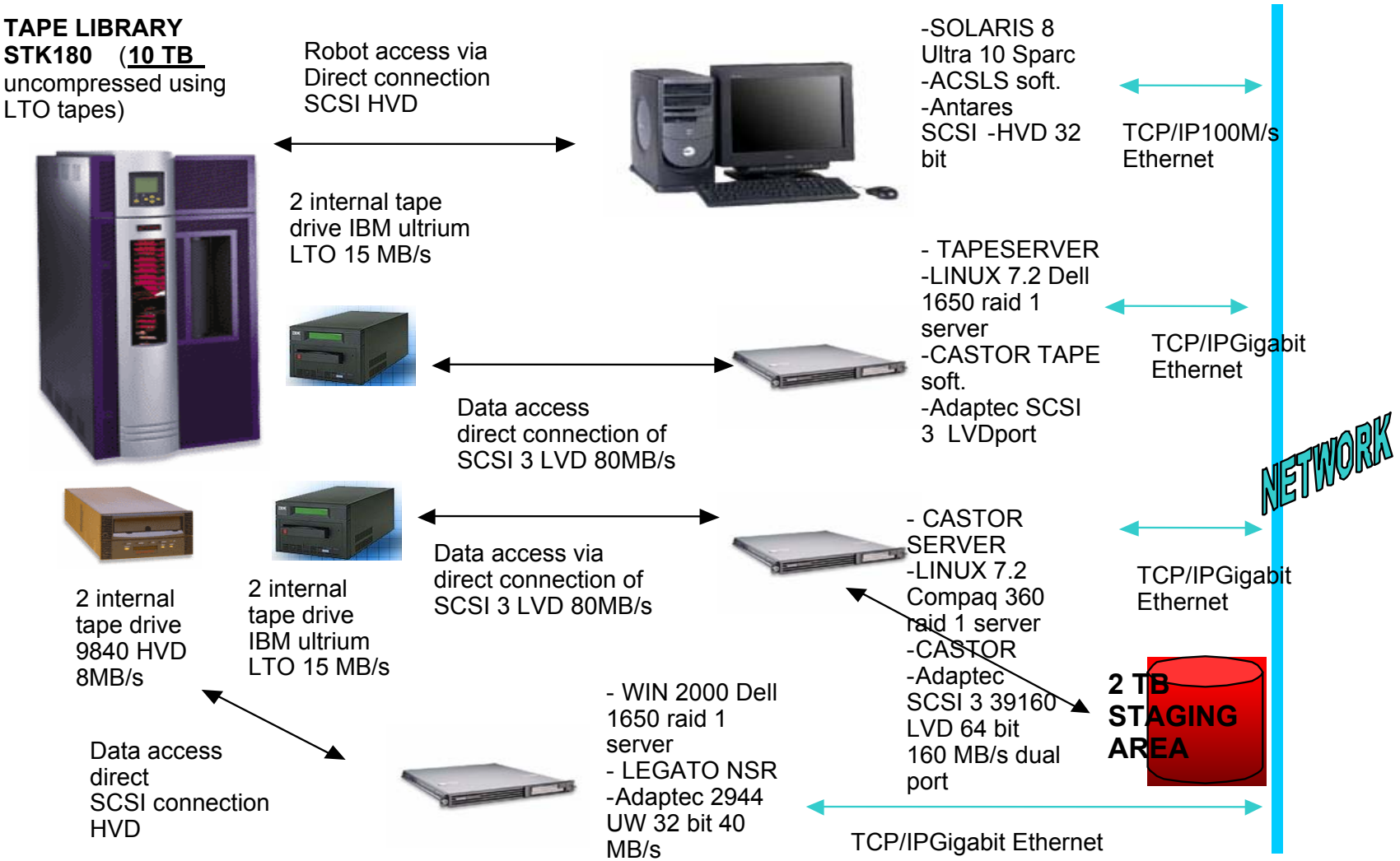
- 1 SUN SPARC ULTRA 10 SOL.v8 running ACSLS 6.0 (with ext. SCSI disk)
- 1 RAIDTEC HI 160 SCSI  2Tbyte RAID5 array disk
- 4 SCSI LTO-1 IBM DRIVE with 100 Maxell LTO tapes (+ 50) on the STK L180
- The "old" Castor installation is built on a central machine used also as a disk server (with stager) and as a tape server
  - A different machine used as another tape server

# CASTOR resource at TIER1 (3)




- STK L5500 robot with 2000 active slots (out of 5000) and 6 LTO-2 FC Drives
- 1300 LTO-2 (Imation) Tapes
- Sun Blade v100 with 2 internal ide disks with software raid-0 running ACSLS 7.0
- Point to point F.C. connection to 6 F.C. HBA on 3 dedicated machines (tape servers)

# CASTOR resources



# Summary & conclusions (1)

- INFN-TIER1 ended prototype phase
  - But still testing new technological solutions
- Moved resources to the final location
- Several things still to fix
- Starting integration with LCG
  - CNAF Tier1 is an LCG2 Core site (Jan 2004)
    - Supporting italian Tier2 (Catania, Legnaro, Roma, Milano)
- Wish to collaborate more with CERN and other Tiers to improve
  - CASTOR
  - RLS
  - Farm utilization

# Summary & conclusions (2)

- Participating to Data Challenges
  - ~ 240 computing servers (~ 480 CPU's)
- CMS DC04
  - ~ 70 computing nodes allocated
  - ~ 4M events (40% of Italian commitment) 15+60 (Tier0) TB of data (July to December 03)
  - Analysis of simulated events (January to February 04)
  - Interoperation with Tier0 (CERN) and Tier2 (LNL)