LCG Workshop

Computing Fabric

Summary

Bernd Panzer-Steindel
LCG Fabric Area Manager

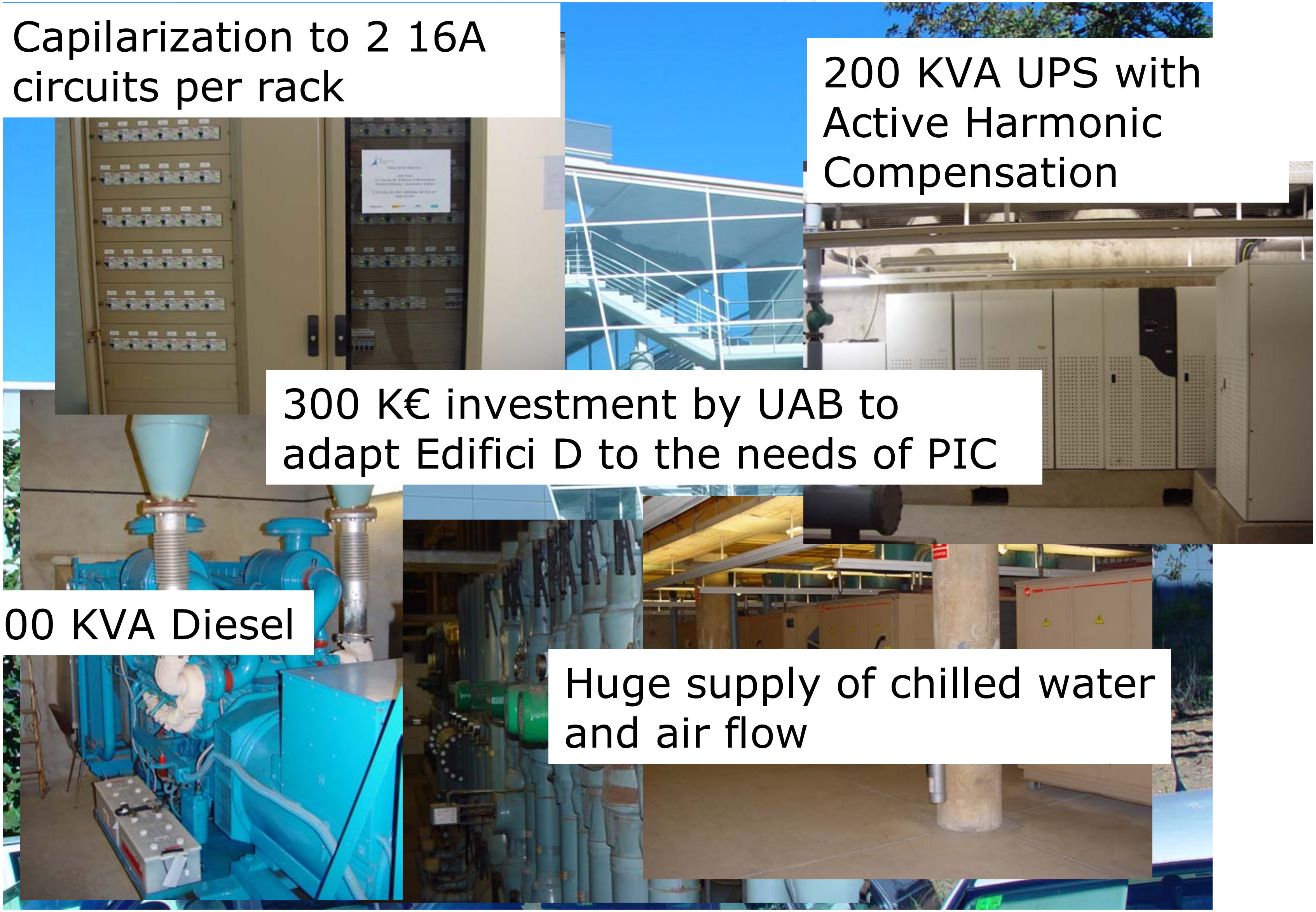
10 reports from different Tier1 center about their current status and their expansion plans

IN2P3 (France)
Nikhef (Netherland)
RAL   (UK)
Fermilab (US)
Brookhaven (US)
Tokio (Japan)
PIC (Spain)
Karlsruhe (Germany)
Tokio (Japan)
CERN

covering topics like : infrastructure (electricity, space, cooling), purchasing cpu, disk, tape and network resources, developments, problems

Cooling and Power

Good physical infrastructure



Capilarization to 2 16A circuits per rack

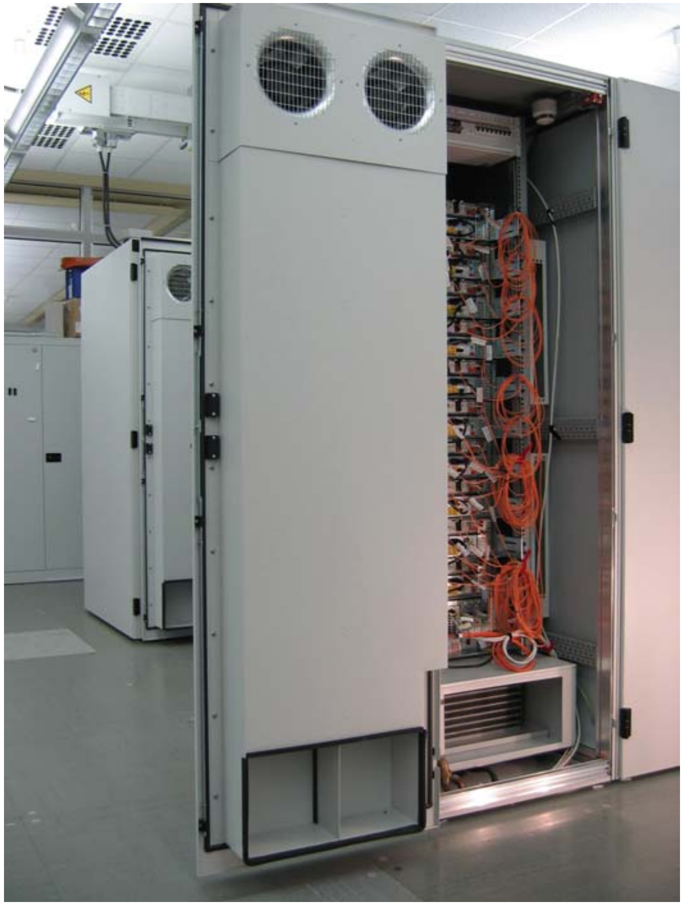
200 KVA UPS with Active Harmonic Compensation

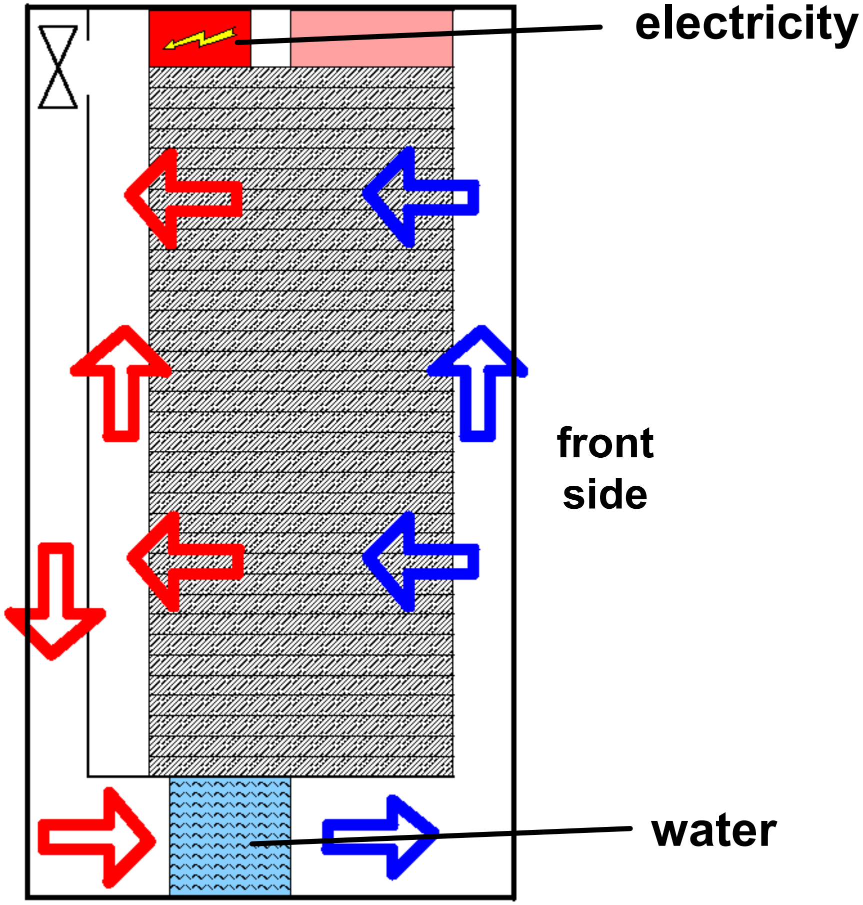
300 K€ investment by UAB to adapt Edifici D to the needs of PIC

500 KVA Diesel

Huge supply of chilled water and air flow

Equipment cabinet with water cooling





Power

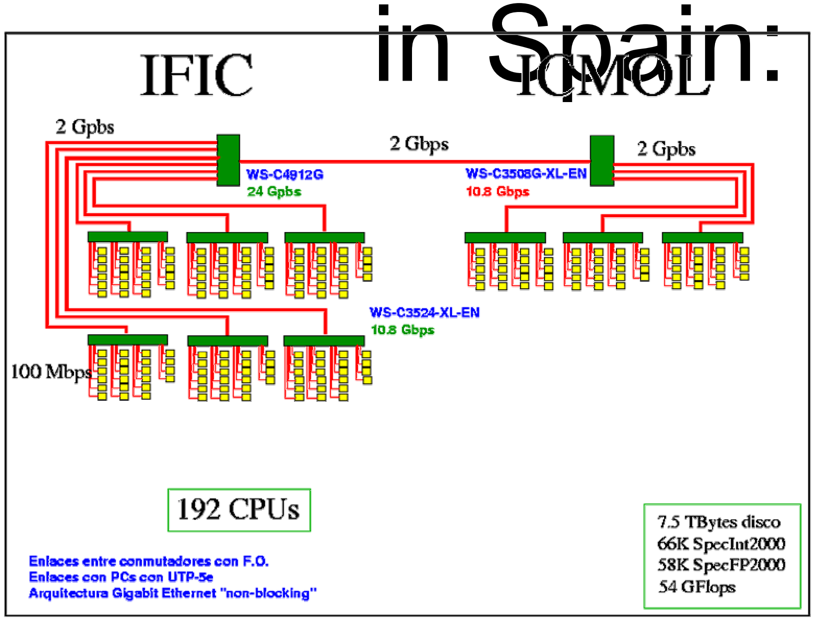
- What is the future for processor power? Inspite of reported "power budgets" per processor class, consumption seems to rise with each generation.
- CERN plans for 2.5MW active load; building consumption more like 5.5-6MW.
  - But with a 50% overcapacity in the low voltage distribution for flexibility.
  - Machine room & UPS consumption monitored by us (data stored in Lemon repository).
- Power factor as important as power.
  - Increased harmonics lead to unbalanced 3-phase system.
  - Fortunately EU directives seem to have led to an improvement from ~0.7 to ~0.9, even 0.95.
    - We now reserve space for filters but don't include these in the baseline solution.

CPU Systems

Cluster

- +1000 processors (90% Linux Redhat 7.2)
- Job submissions : BQS
- Parallel computation

An example of another large facility in Spain: IFIC-Valencia



IFIC

ICMOL

2 Gpbs

2 Gbps

2 Gpbs

WS-C4912G
24 Gpbs

WS-C3508G-XL-EN
10.8 Gbps

WS-C3524-XL-EN
10.8 Gbps

100 Mbps

192 CPUs

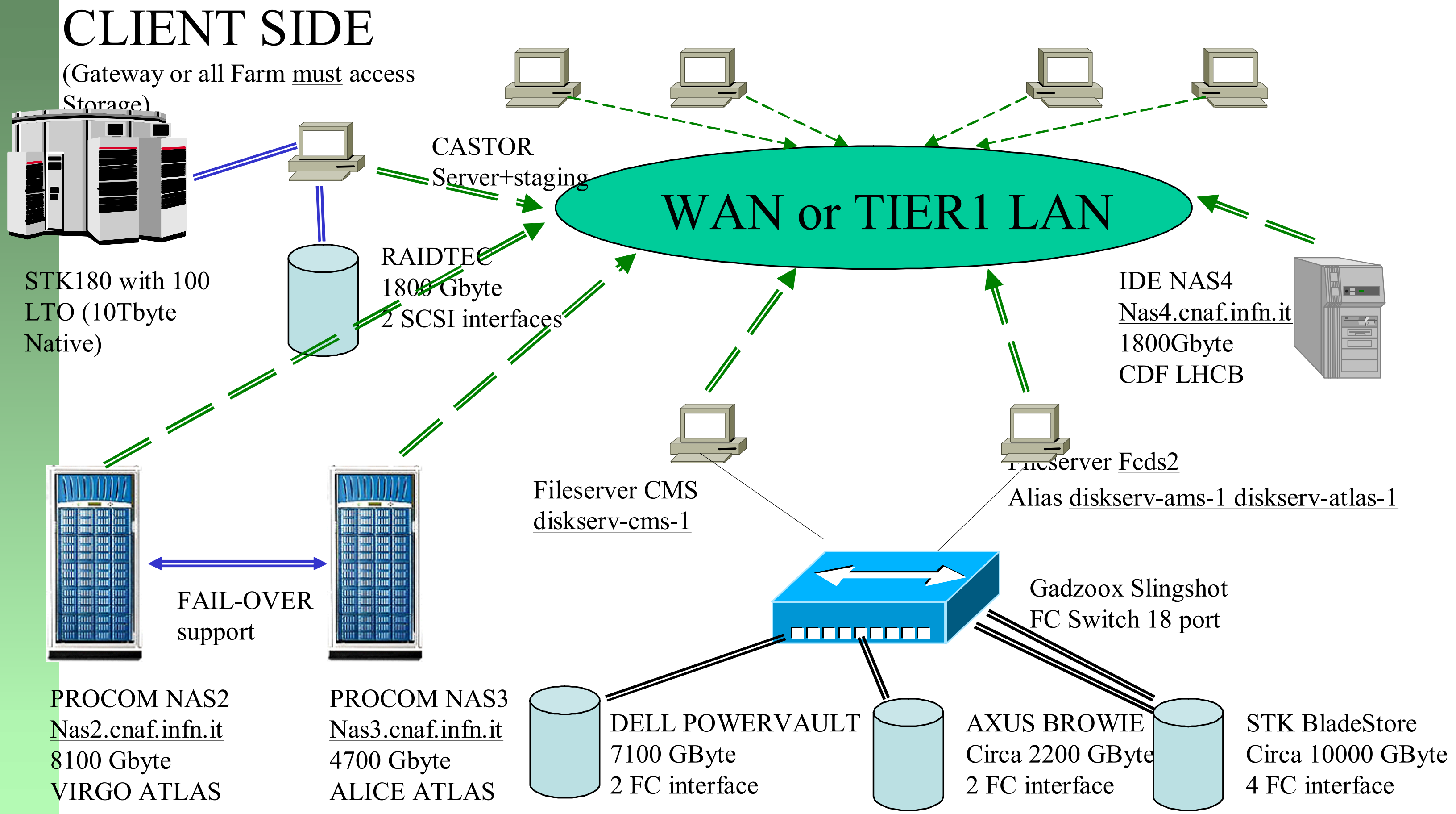
Enlaces entre conmutadores con F.O.
Enlaces con PCs con UTP-5e
Arquitectura Gigabit Ethernet "non-blocking"

7.5 TBytes disco
66K SpecInt2000
58K SpecFP2000
54 GFlops

HP Blade Server 2

Storage

STORAGE resource



CLIENT SIDE

(Gateway or all Farm must access Storage)

CASTOR Server+staging

STK180 with 100 LTO (10Tbyte Native)

RAIDTEC
1800 Gbyte
2 SCSI interfaces

WAN or TIER1 LAN

IDE NAS4
Nas4.cnaf.infn.it
1800Gbyte
CDF LHCB

Fileserver CMS
diskserv-cms-1

Fileserver Fcds2
Alias diskserv-ams-1 diskserv-atlas-1

FAIL-OVER support

Gadzoox Slingshot
FC Switch 18 port

PROCOM NAS2
Nas2.cnaf.infn.it
8100 Gbyte
VIRGO ATLAS

PROCOM NAS3
Nas3.cnaf.infn.it
4700 Gbyte
ALICE ATLAS

DELL POWERVAULT
7100 GByte
2 FC interface

AXUS BROWIE
Circa 2200 GByte
2 FC interface

STK BladeStore
Circa 10000 GByte
4 FC interface

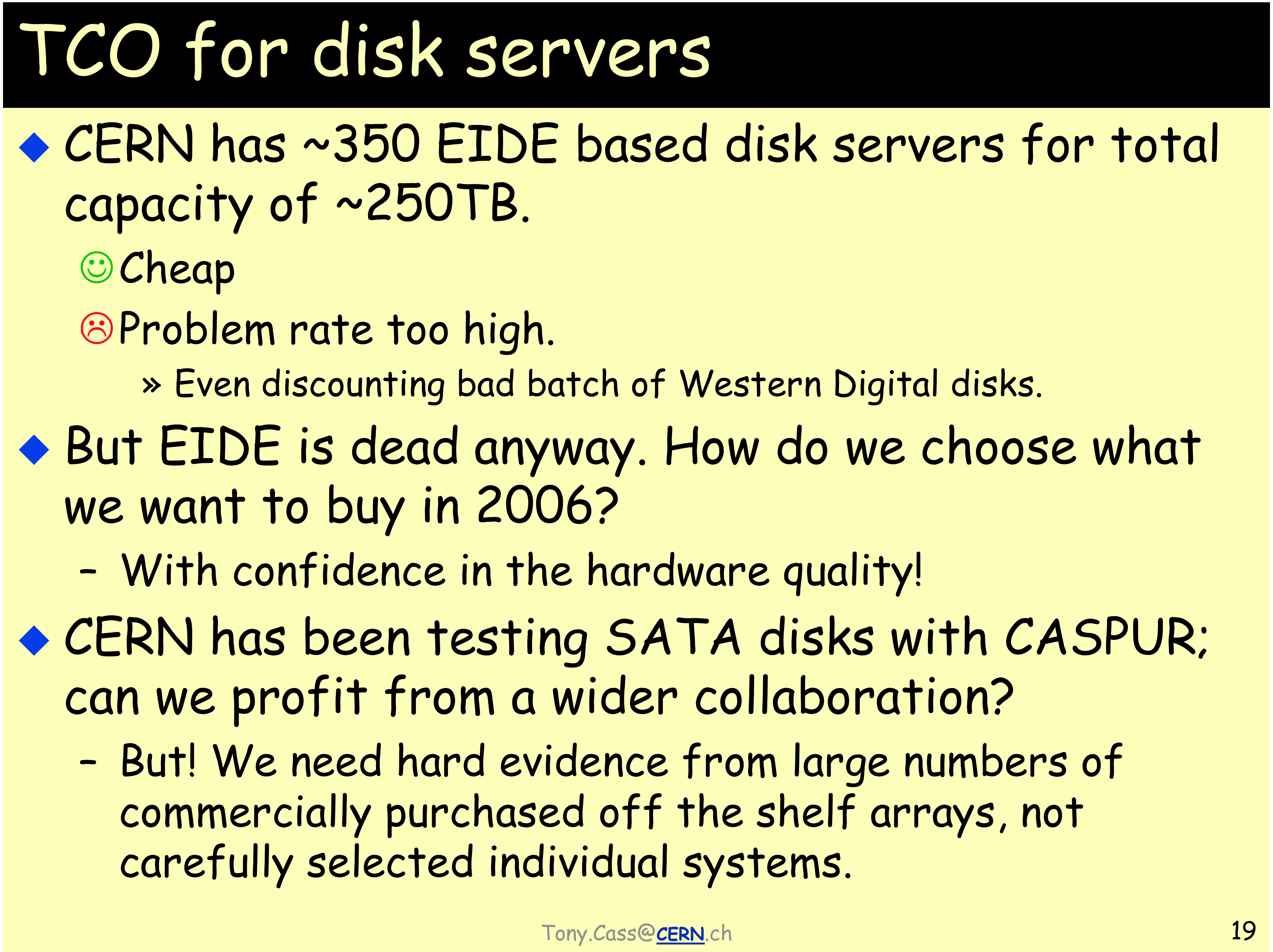
Present Hardware - Disk

- 11 Linux rack mount servers providing ~40TB IDE disk
  - 11 dual 2.4GHz P4 HT Xeon servers with PCIx (1GB RAM), each with:
  - 2 Infortrend IFT-6300 arrays, each with:
  - 12 Maxtor 200GB Diamondmax Plus 9 drives per array, most configured as 11+1 spare in RAID 5 => ~2TB/array.
- 26 Linux rack mount servers providing ~44TB IDE disk
  - 26 dual 1.266GHz P3 servers (1GB RAM), each with:
  - 2 Accusys arrays, each with:
  - 12 Maxtor 80GB drives -1.7TB disk per server.
- 3 Linux tower servers providing ~4.8TB IDE disk
  - 3 Athlon MP 2000+ single processor tower servers, each with:
  - 1 x 3ware 7500-8 with 8 Maxtor DiamondMax Plus 9 as RAID5
- 2 Linux servers providing 300Gb SCSI RAID 5 (to be deployed).
- Solaris server with 4.5TB
- 3 x Ultra10 Solaris servers (being phased out)
- AFS Cell – 1.3TB, AIX +Transarc – migrate to Linux + OpenAFS server during 2004.

Storage

GridKa

- online data stored in NAS (40 TB) and SAN (130 TB)
- NAS boxes have 16 EIDE disks and 3Ware controllers
  - problems with 3ware controllers
- SAN cluster file system (GPFS) exported via NFS to the WNs
  - high availability through multiple redundant servers
  - load balancing via automounter program map
  - since introduction of above: CPU/Wall clock time nears 1
- planned offering of (x)rootd on file servers



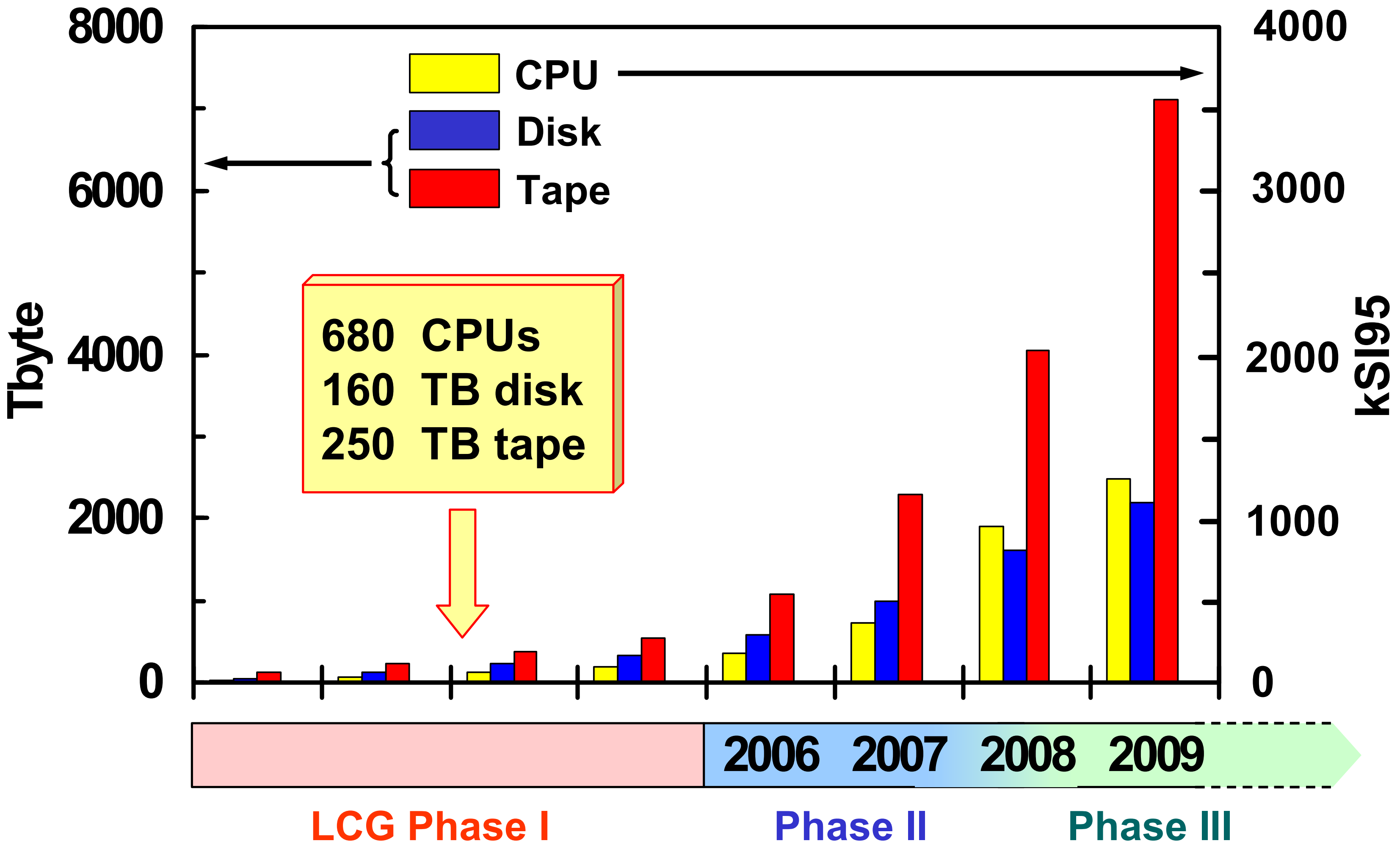
TCO for disk servers

- CERN has ~350 EIDE based disk servers for total capacity of ~250TB.
  - ☺ Cheap
  - ☹ Problem rate too high.
    - » Even discounting bad batch of Western Digital disks.
- But EIDE is dead anyway. How do we choose what we want to buy in 2006?
  - – With confidence in the hardware quality!
- CERN has been testing SATA disks with CASPUR; can we profit from a wider collaboration?
  - – But! We need hard evidence from large numbers of commercially purchased off the shelf arrays, not carefully selected individual systems.

Tony.Cass@CERN.ch

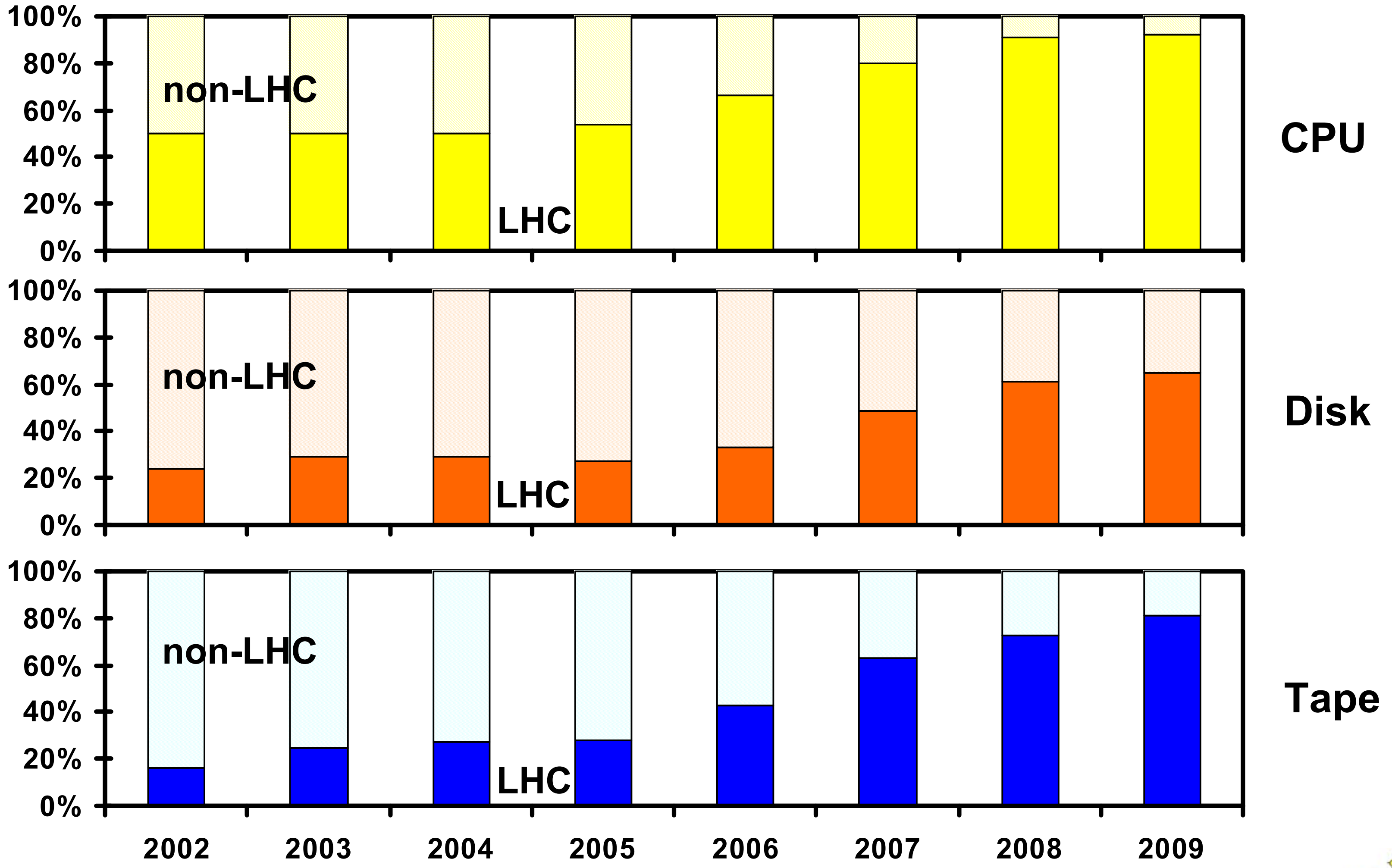
Upgrade Plans

GridKa planned resources



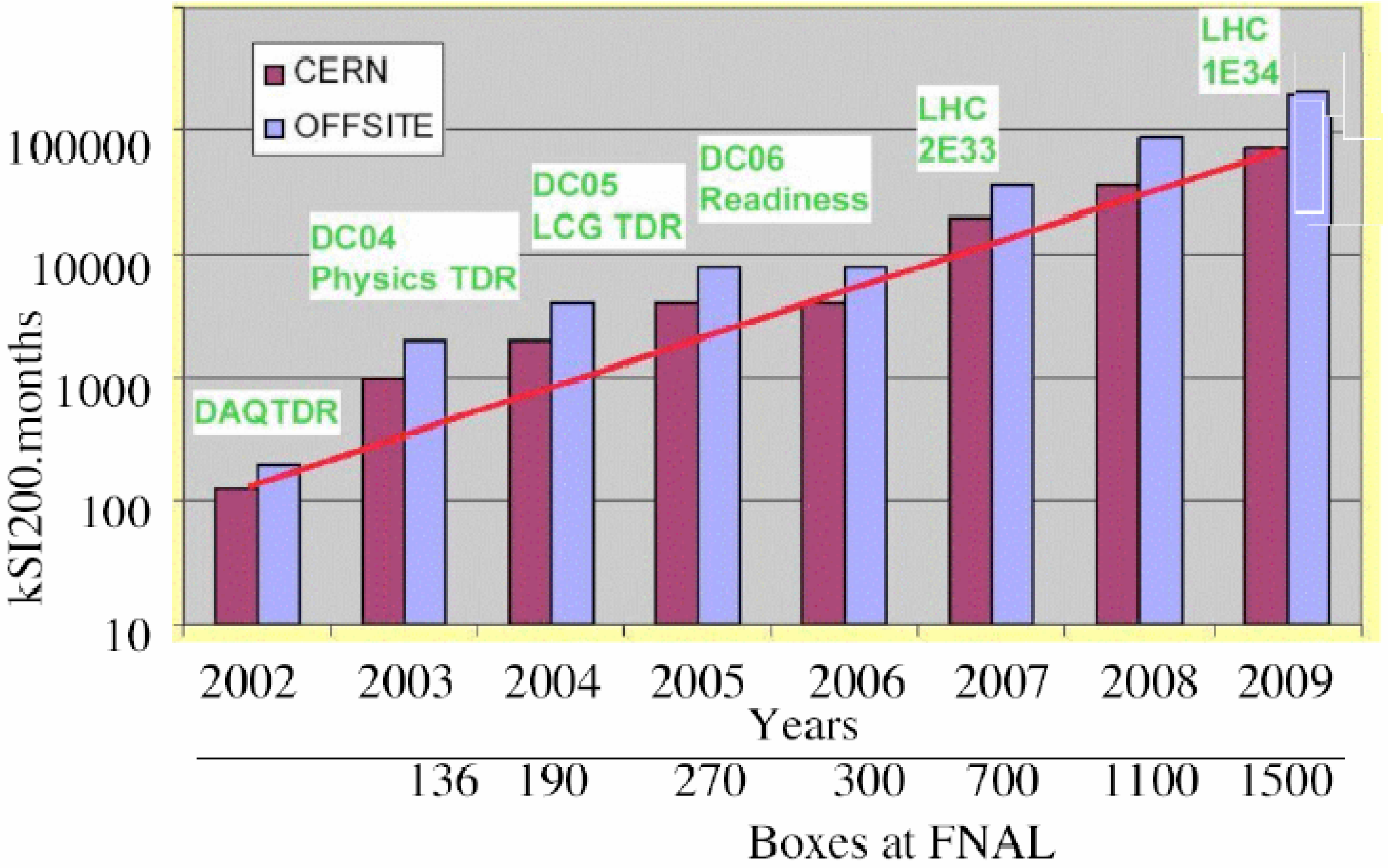
680 CPUs
160 TB disk
250 TB tape

Distribution of planned resources at GridKa



Expected Ramp of the Facility

We are using the following CMS estimates for required computing at a function of time



FNAL Tier1 Represents about 10% of the total

- Roughly on schedule for 2003 and 2004

Future Fermilab Facility Upgrades

Networking

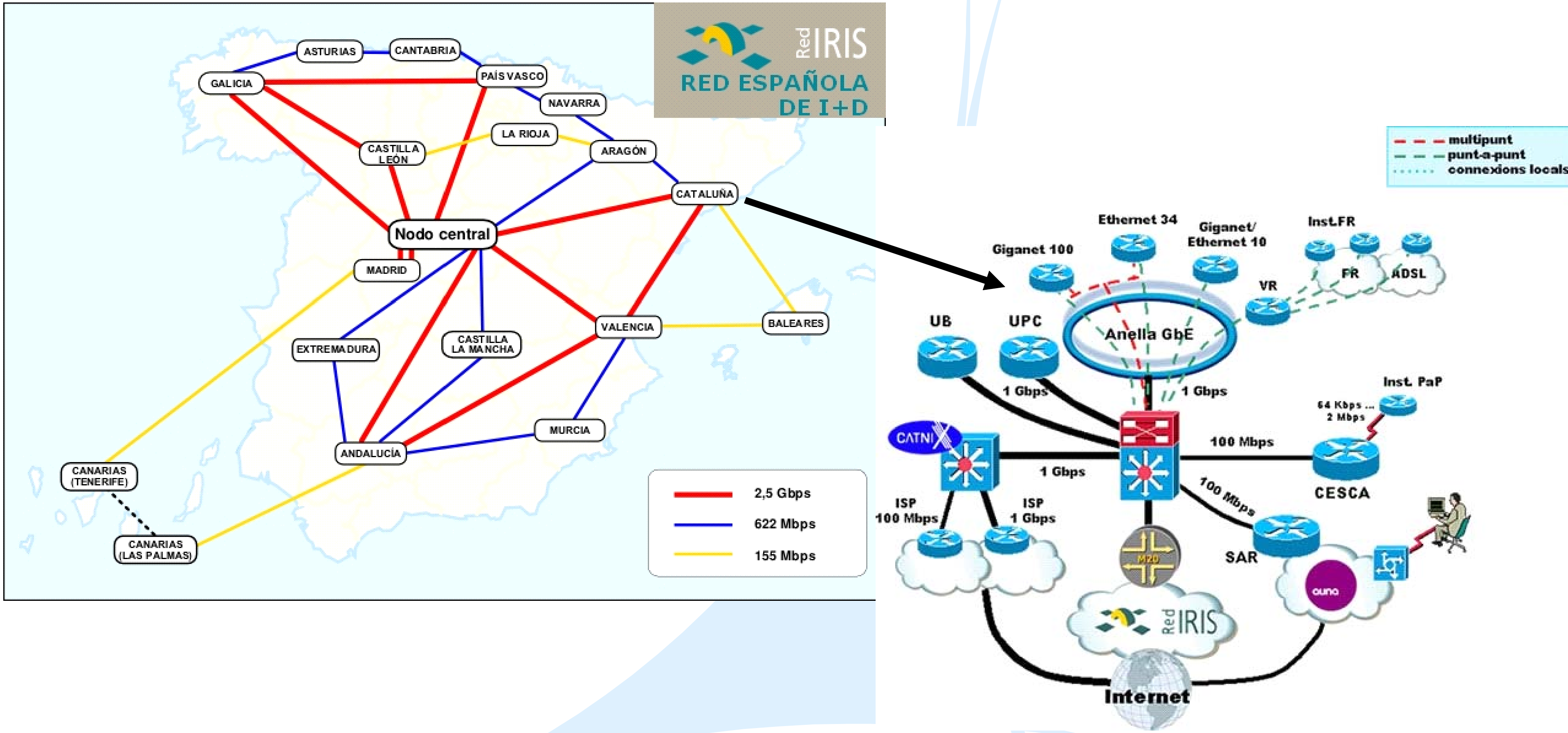
- Fermilab is physically close to Starlight in Chicago (60km)
  - From there the DOE supported link from Starlight to CERN provides ~10GB.
- The current Fermilab link is 622Mbit/s
  - Primarily network traffic is from Tevatron detectors off-site
- DOE has a long term strategy for a Metro ring with high performance and availability
- For a research network and improve access, Fermilab is arranging a fiber connection to StarLight
  - Contracts are in place
  - We hope to see light in the fiber before then end of the year.
  - It should provide a good short term and long term network solution for US-CMS

Short term plan

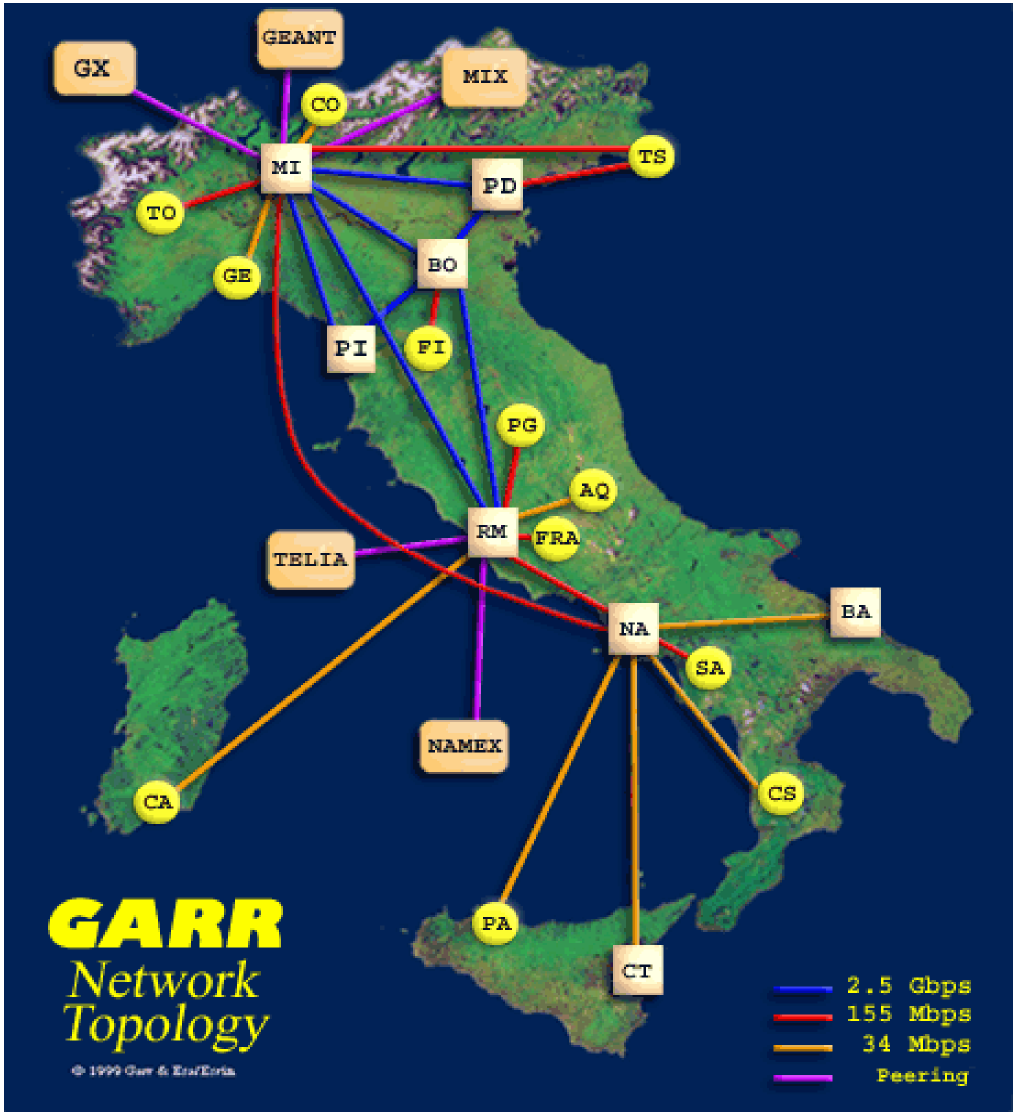
- Network Connectivity
  - CERN-Tokyo to 10Gbps now
  - Tokyo-Taipei connectivity study soon
- PC Farm / Mass Storage
  - ~100TB fiber-channel disks installed
  - Hierarchical storage study soon (IBM LTO2)
  - PC farm upgrade
- LCG-2
  - 60 Nodes (Dual 2.8GHz Xeon)
  - ~30TB Disk Space

Wide Area Network

WAN:
Anella Científica + RedIRIS + GEANT

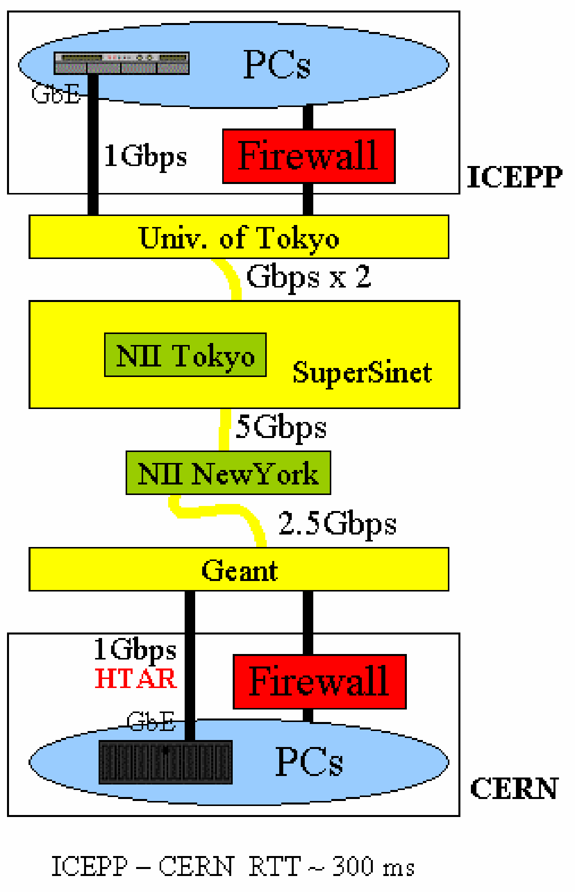


GARR Topology



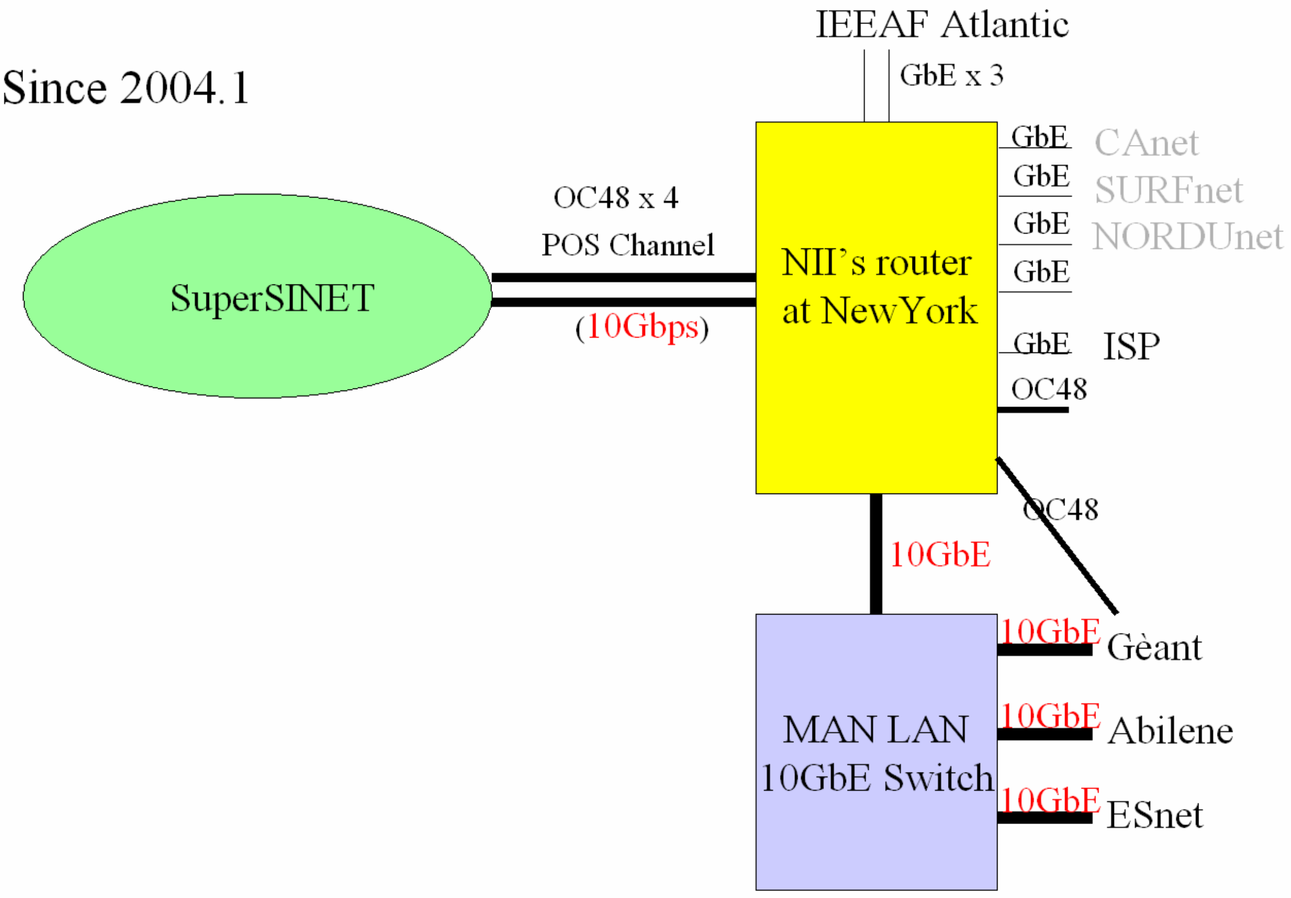
Connection to CERN

- 2.5Gbps x 2 (=5Gbps) to NY (2003)
- 2.5Gbps x 4 Now



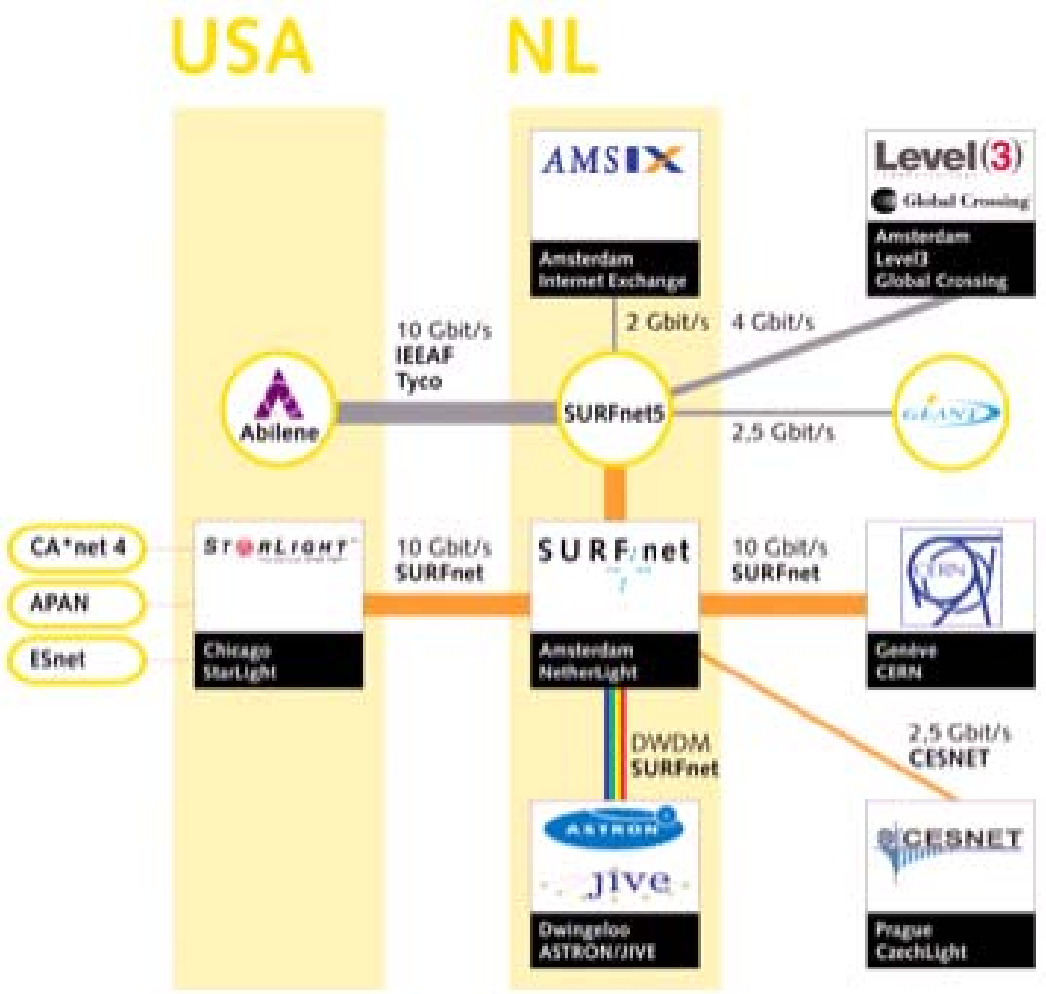
ICEPP – CERN RTT ~ 300 ms

Since 2004.1

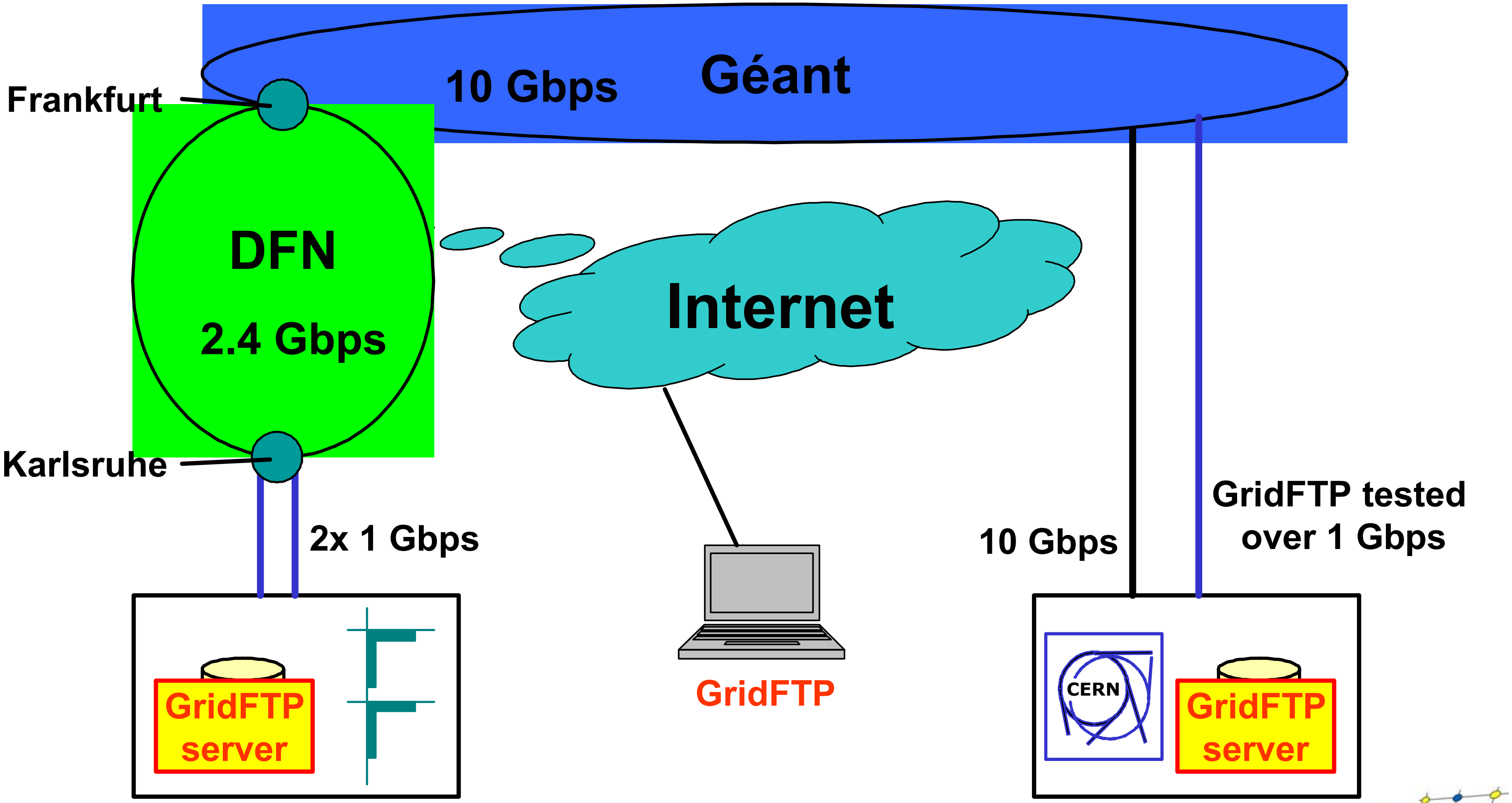


Infrastructure

- Network connectivity via SURFnet5
  - 10 Gbit/s to:
    - Abilene
    - STARLIGHT
    - CERN
    - GEANT
  - 4 Gbit/s to:
    - Level3 (GBX)
  - 2.5 Gbit/s to:
    - CESNET
  - 2 Gbit/s to:
    - AMSIX

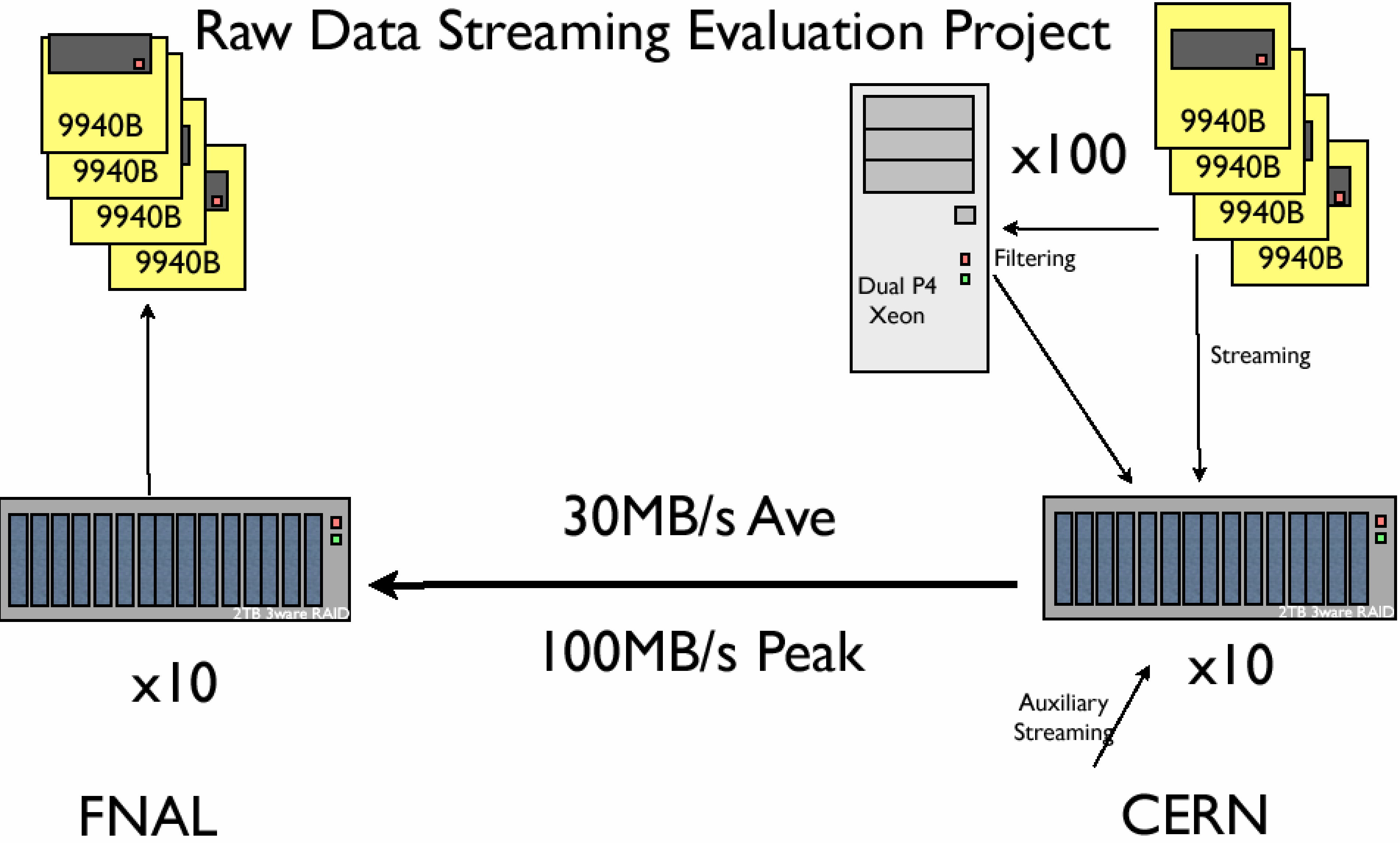


WAN connectivity and Gigabit test with CERN



Data Streaming Project

Raw Data Streaming Evaluation Project



Summary

‘Technical’ coupling of the Tier 0/Tier 1 centers

independent developments

Basic infrastructure (box size, electricity, cooling)

Cluster management   Batch systems

sharing experience

Filesystems, repositories (software,calibration,metadata,etc.)

Mass storage   Equipment quality, stability   Large disk pools

common activities

Operating system (Linux version x)   Local security

Grid middleware

Mass storage interfaces

synchronization

Online raw data and ESD copy, WAN

dependency level

Common developments (a few examples) :

- disk storage evaluations (SATA disks with fibre channel attachments)
- benefits of Hyperthreating
- reliability/stability of components (disks, memory, controller)

Hepix was and is still a major place to exchange information and experience but the 6 month time-frame seems to be too long, thus we agreed to use a dedicated mailing list (to be started asap) to foster more peer-to-peer communications between the Tier 1 centers on selected and focused topics, this could also lead to a more concentrated ‘voice’ of the Tier1 centers about policies

Common issues (a few examples) :

- Scheduling policies of batch systems → middleware
  (there will be different systems : PBS, BQS, LSF, TORQUE, etc)
  the word optimal means different things to different communities

- Security : opening the firewalls, outbound connectivity

- Software installation procedures are site dependent, software needs to be packaged correspondingly

Centers are independent units with their own individual
‘boundary’ conditions : funding sources, history, user community, etc.
which effect their way of selecting hardware and software
They have to provide a reliable and efficient service to a mostly mixed user community

→ requires more flexibility and adaptability from the middleware and experiment software
(this is of course also a matter of reasonable compromises..)

In general there is the feeling that there is a lack of
understanding/communication between software ‘developers’
and service implementers