CERN

openlab for DataGrid applications

Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

The LHC Computing Challenge

Preparing the computing solutions for the Large Hadron Collider at CERN

Sverre Jarp, IT Division, CERN

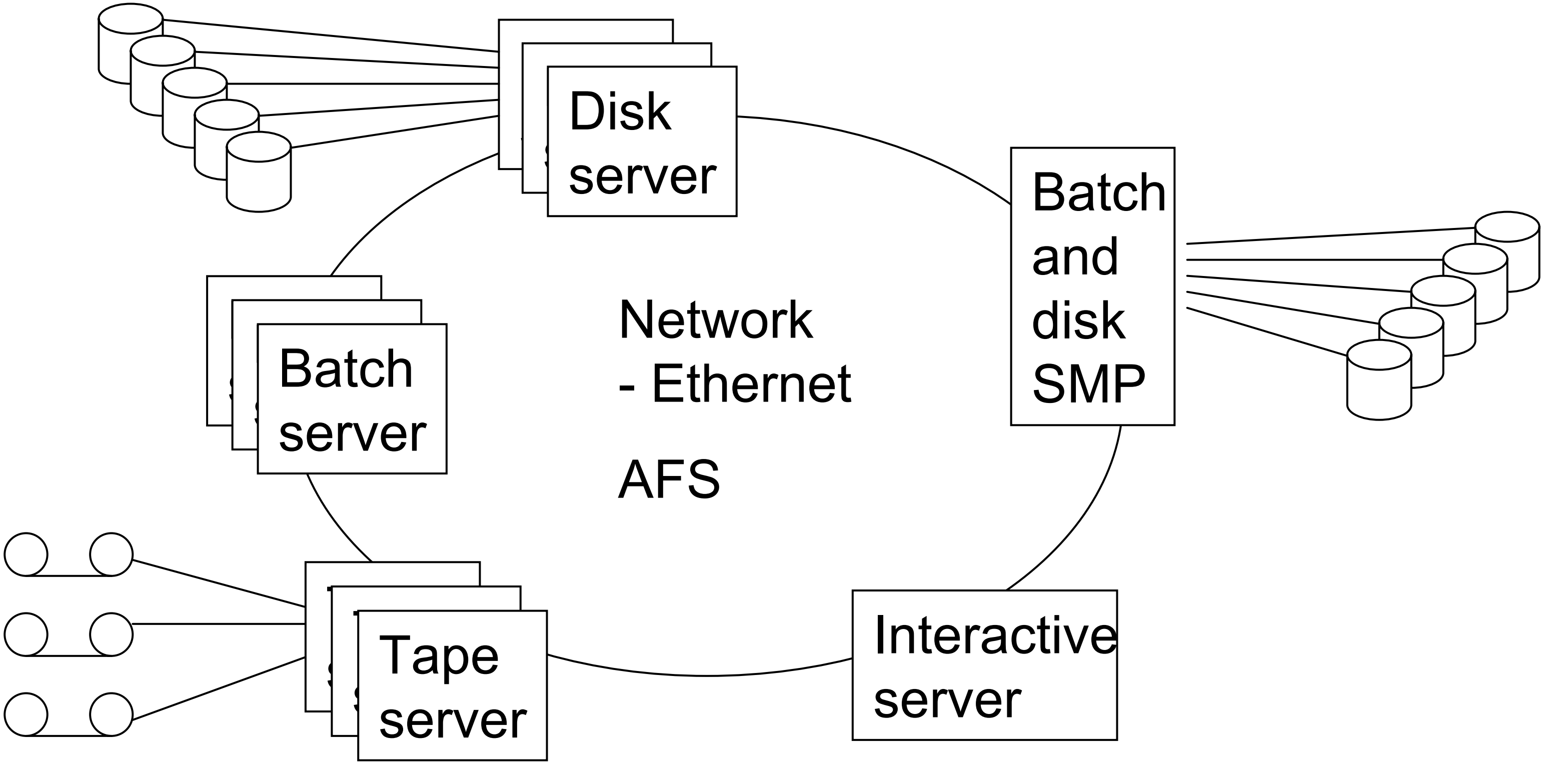
High Energy Physics Computing Characteristics

- Independent events (collisions of particles)
  - trivial (read: pleasant) parallel processing
- Bulk of the data is read-only
  - versions rather than updates
- Meta-data in databases linking to "flat" files
- Compute power measured in SPECint (not SPECfp)
  - But good floating-point is important
- Very large aggregate requirements:
  - computation, data, input/output
- Chaotic workload –
  - research environment - physics extracted by iterative analysis, collaborating groups of physicists

  → Unpredictable → unlimited demand

SHIFT architecture

(Scalable Heterogeneous Integrated Facility)



Disk server

Batch and disk SMP

Batch server

Network - Ethernet

AFS

Tape server

Interactive server

In 2001 SHIFT won the 21st Century Achievement Award issued by Computerworld

CERN's Computing Environment (today)

- High-throughput computing (based on reliable "commodity" technology)
  - More than 1500 (dual processor) PCs with Red Hat Linux
  - About 3 Petabytes of data (on disk and tapes)

IDE Disk servers

- Cost-effective disk storage: ~10 CHF/GB

The LHC Challenge

The Large Hadron Collider (LHC) has 4 Detectors:

Requirements:

Storage –

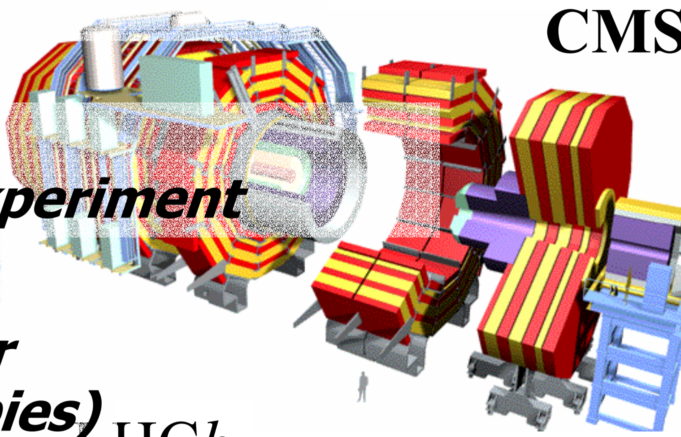
Raw recording rate: up to 1 GB/s per experiment

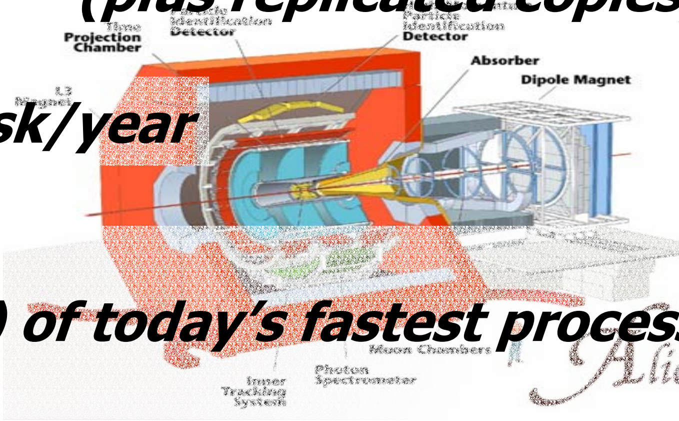
Accumulating data at 10 Petabytes/year (plus replicated copies)

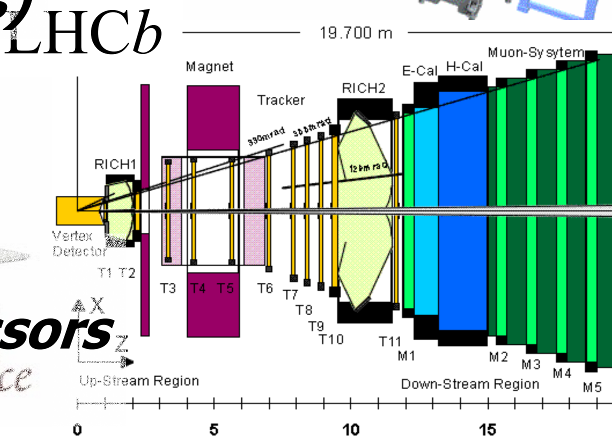
2 Petabytes of disk/year

Processing –

50,000 (100,000) of today's fastest processors





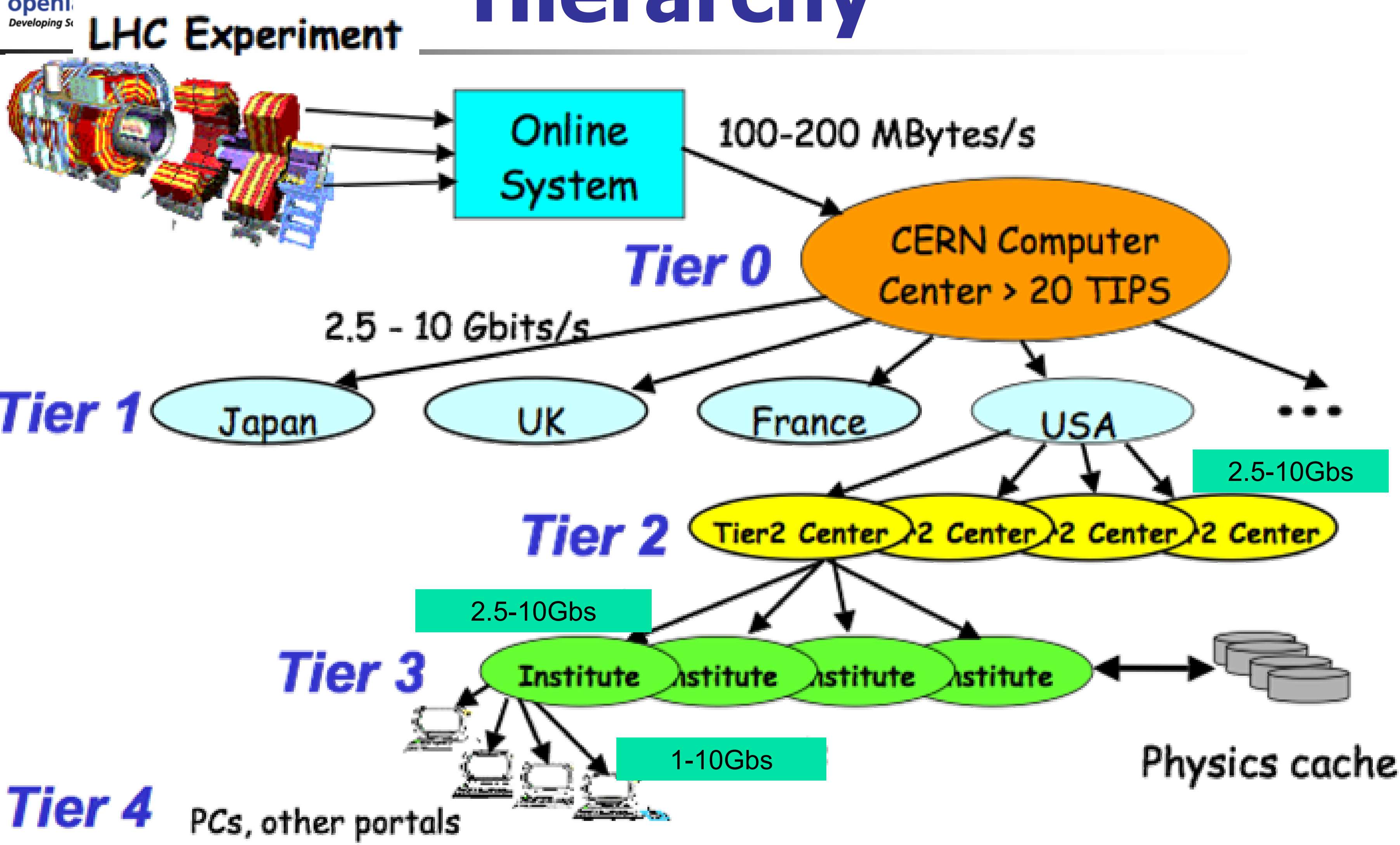


LHC Computing Plan

1 – Build the “fabric”

2 – Interconnect sites in a Grid

The LHC Data Grid Hierarchy



LHC Experiment

Online System

100-200 MBytes/s

Tier 0

CERN Computer Center > 20 TIPS

2.5 - 10 Gbits/s

Tier 1

Japan

UK

France

USA

• • •

2.5-10Gbs

Tier 2

Tier2 Center

Tier2 Center

Tier2 Center

Tier2 Center

2.5-10Gbs

Tier 3

Institute

Institute

Institute

Institute

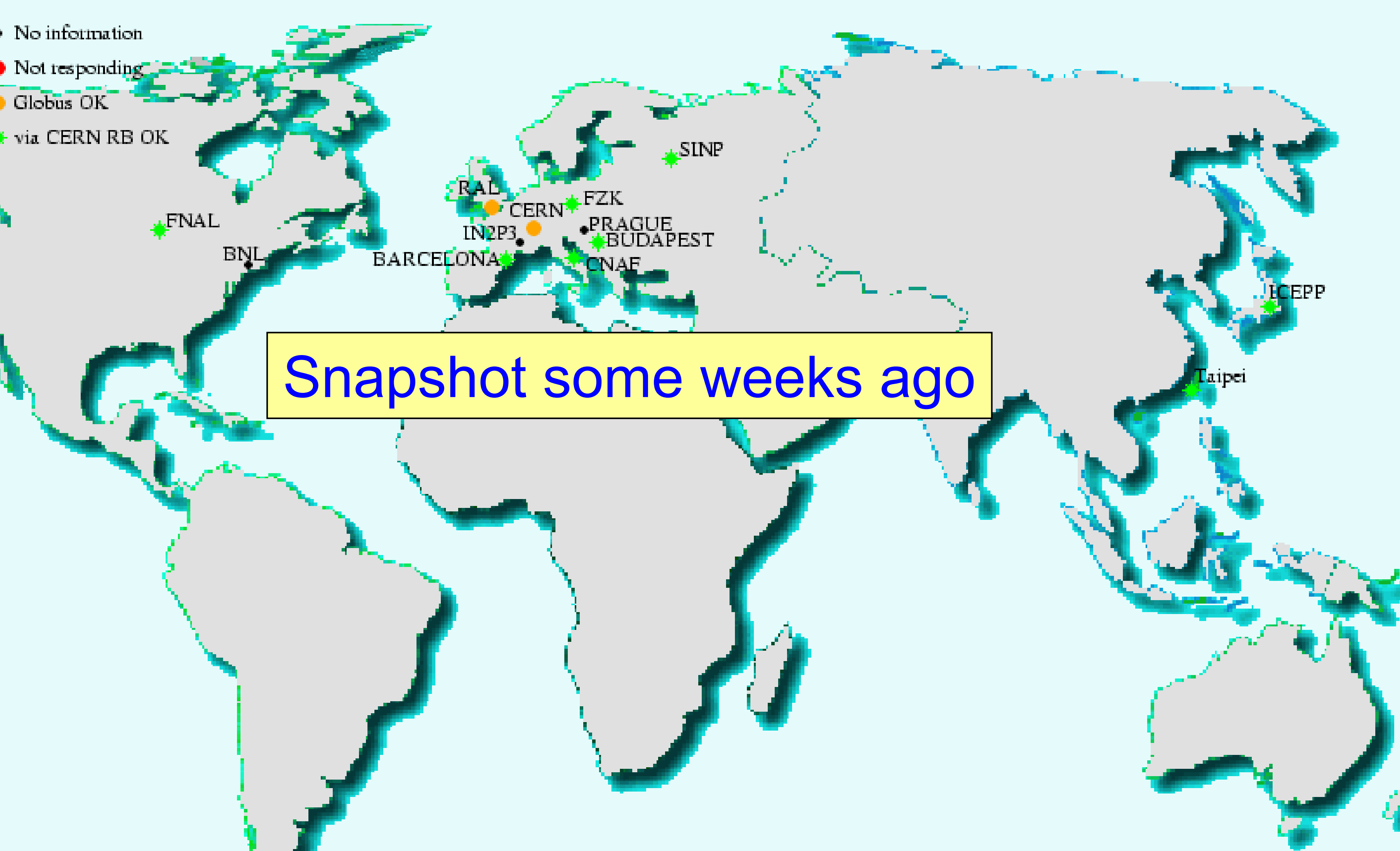
1-10Gbs

Physics cache

Tier 4

PCs, other portals

Current sites in LCG-1



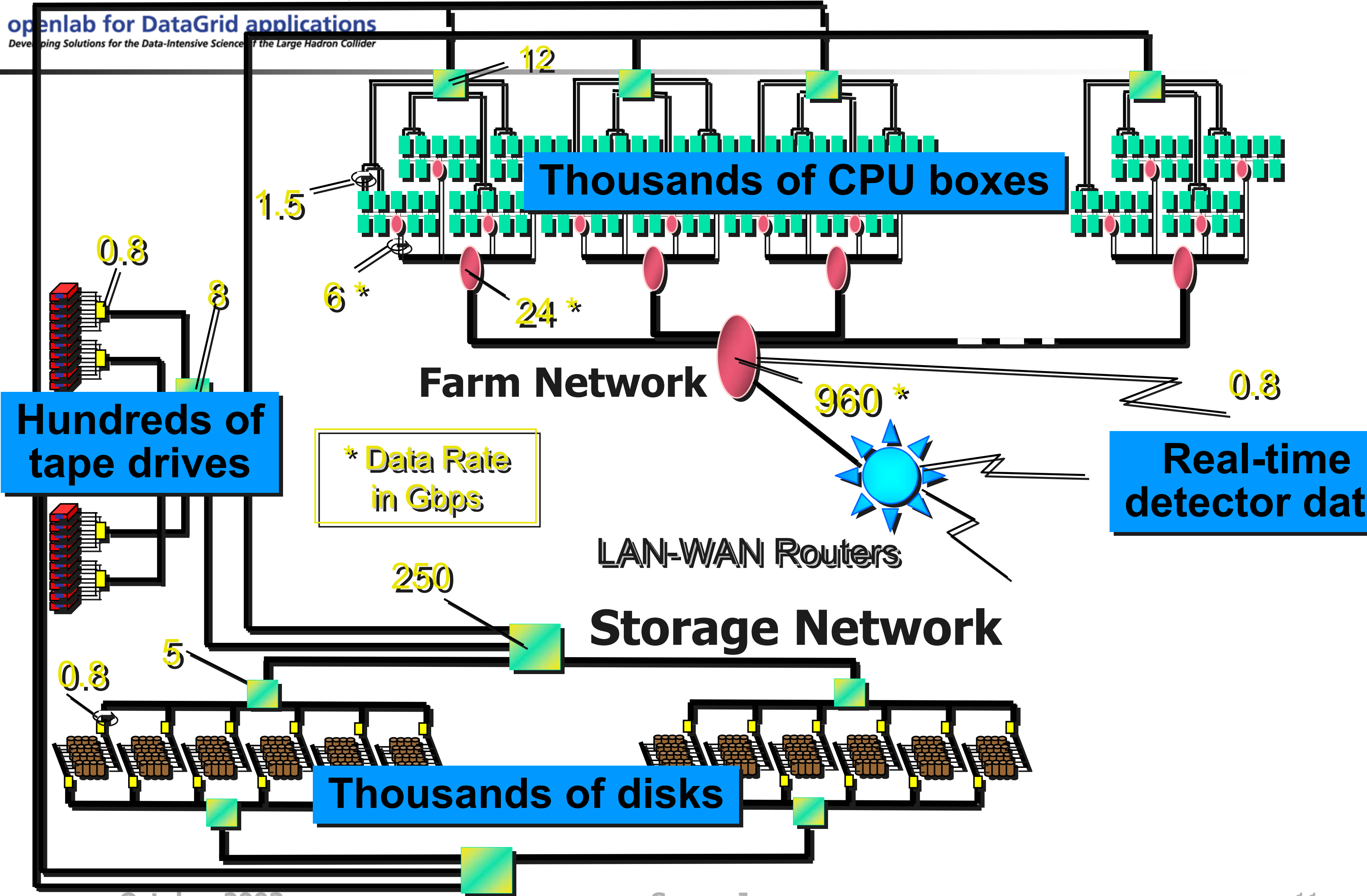
Snapshot some weeks ago

CERN

openlab for DataGrid applications

Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

Proposed Computer Fabric Architecture



Thousands of CPU boxes

Farm Network

Hundreds of tape drives

* Data Rate in Gbps

Real-time detector dat

LAN-WAN Routers

Storage Network

Thousands of disks

CERN openlab

openlab: The technology focus of CERN/IT

- **Industrial Collaboration:**
  - **Enterasys, HP, IBM, and Intel are our partners**
  - **Stop Press: ORACLE just joined**
  - **Technology aimed at the LHC era:**
    - **Network switches at 10 Gigabits**
    - **41 rack-mounted HP servers**
    - **82 Itanium-2 processors**
    - **StorageTank storage system**

Itanium cluster in detail

- **Software integration:**
  - 32 nodes + development nodes
  - Fully automated kickstart installation
  - Red Hat Advanced Workstation 2.1
  - OpenAFS 1.2.7, LSF 5.1
  - GNU, Intel, ORC Compilers
      - ORC (Open Research Compiler, used to belong to SGI)
  - CERN middleware: Castor data mgmt
  - CERN Applications
    - Porting, Benchmarking, Performance improvements
  - Database software
    - Oracle 10g

Program porting status

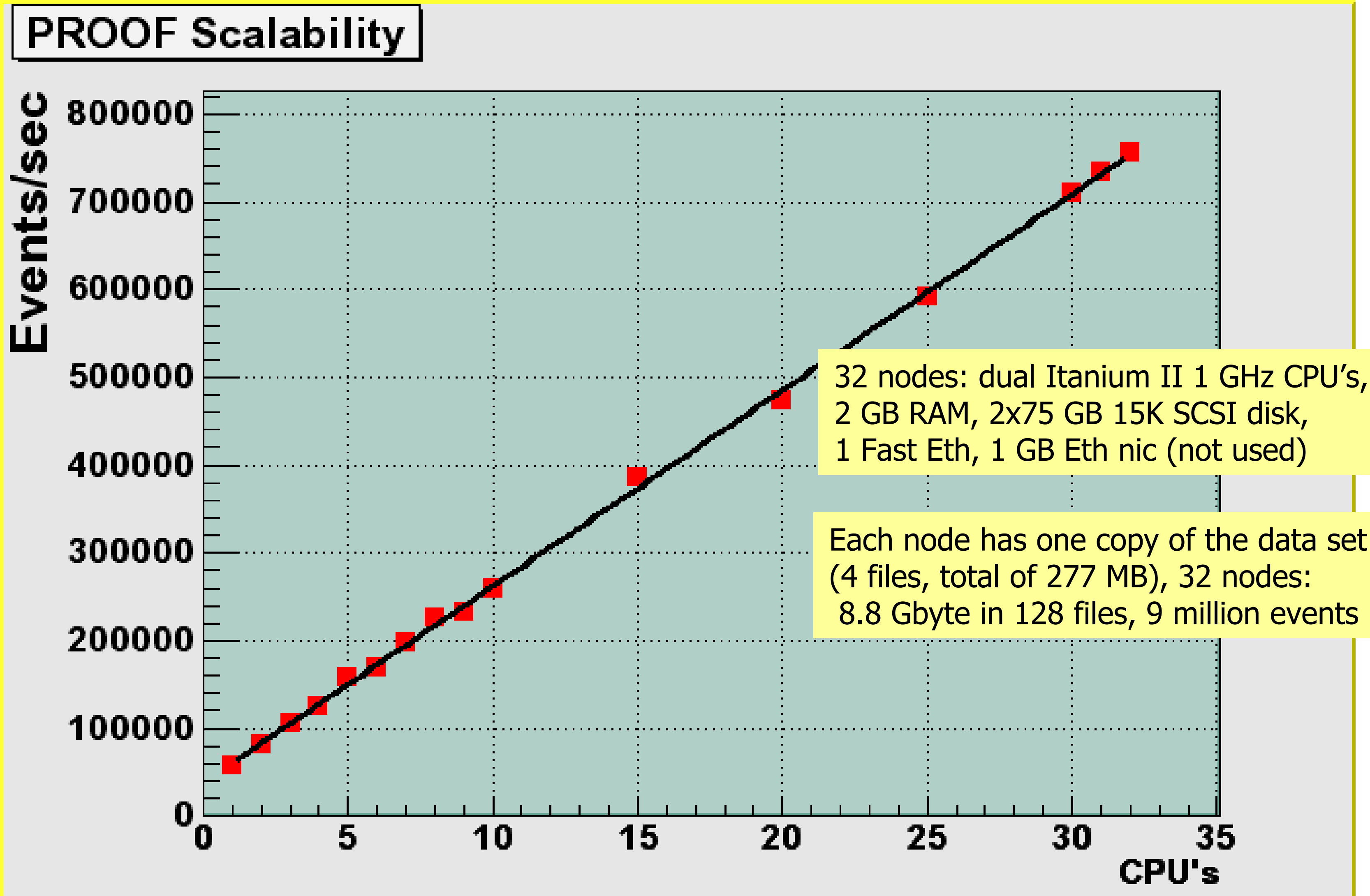
- Ported to 64-bits:
  - Castor (data management subsystem)
    - GPL. Certified by authors.
  - ROOT (C++ data analysis framework)
    - Own license. Binaries both via gcc and ecc. Certified by authors.
  - CLHEP (class library for HEP)
    - GPL. Certified by maintainers.
  - GEANT4 (C++ Detector simulation toolkit)
    - Own license. Certified by authors.
  - CERNLIB (all of CERN's FORTRAN software)
    - GPL. In test.
      - Zebra memory banks are I*4
  - ALIROOT (entire ALICE framework)
- Being ported:
  - LCG software from VDT/EDG
    - GPL-like license.

opencluster

- Current planning:
  - Cluster evolution:
    - Late 2003: Move to 64 nodes (with "Madison" @ 1.5 GHz)
      - Two more racks
    - 2004: Possibly 128 nodes, next generation processors
  - Redo all relevant tests
    - Application benchmarks
      - Also: New compiler versions
    - Network challenges
    - Scalability tests
  - Other items
    - Infiniband tests
    - Serial-ATA disks w/RAID

Make the cluster available to all relevant LHC Data Challenges:
Alice "online" currently using 24 nodes

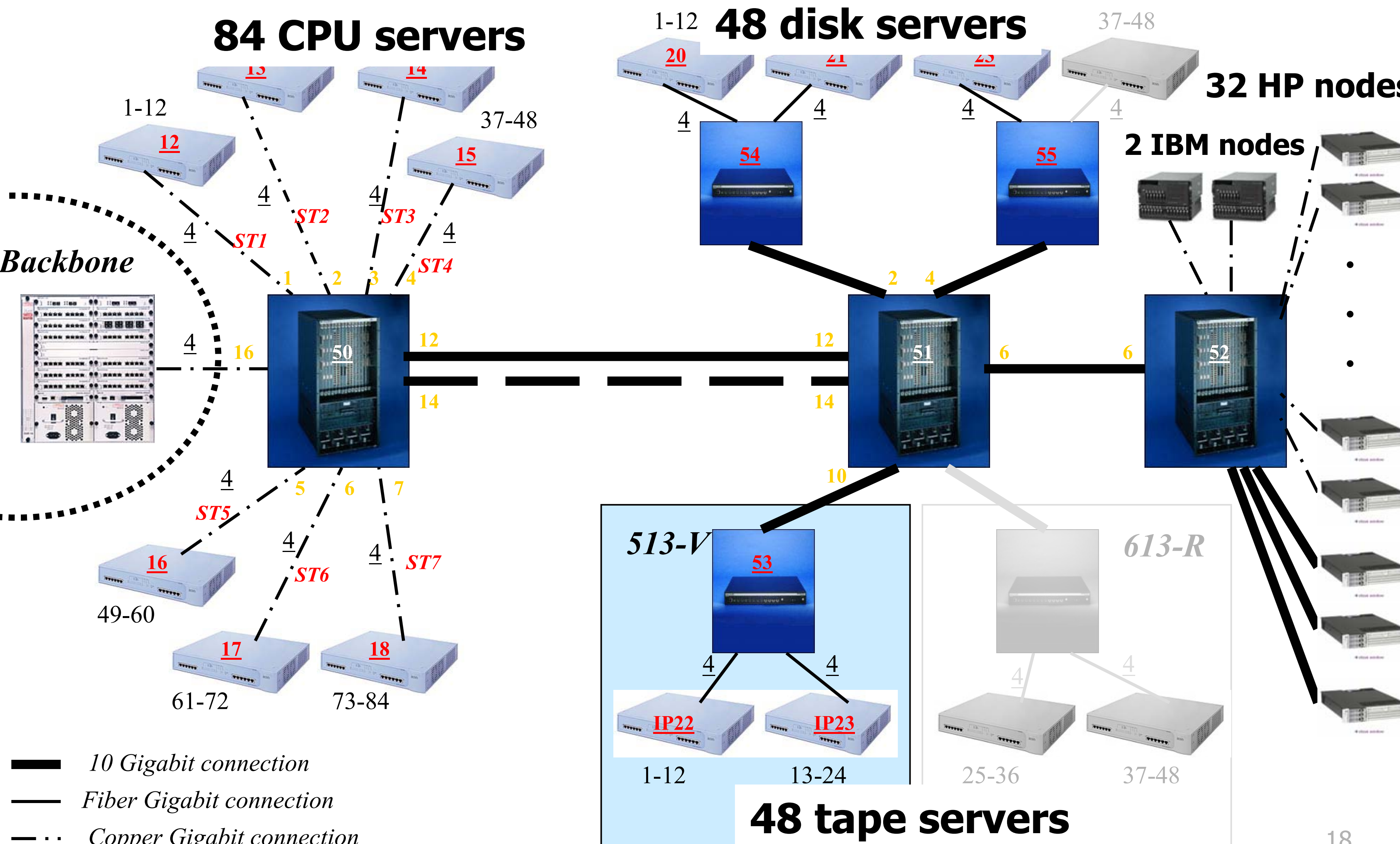
PROOF Scalability
(Presented at CHEP2003)



32 nodes: dual Itanium II 1 GHz CPU's,
2 GB RAM, 2x75 GB 15K SCSI disk,
1 Fast Eth, 1 GB Eth nic (not used)

Each node has one copy of the data set
(4 files, total of 277 MB), 32 nodes:
8.8 Gbyte in 128 files, 9 million events

Enterasys 2Q 2003


84 CPU servers

48 disk servers

32 HP nodes

2 IBM nodes

Backbone

513-V

613-R

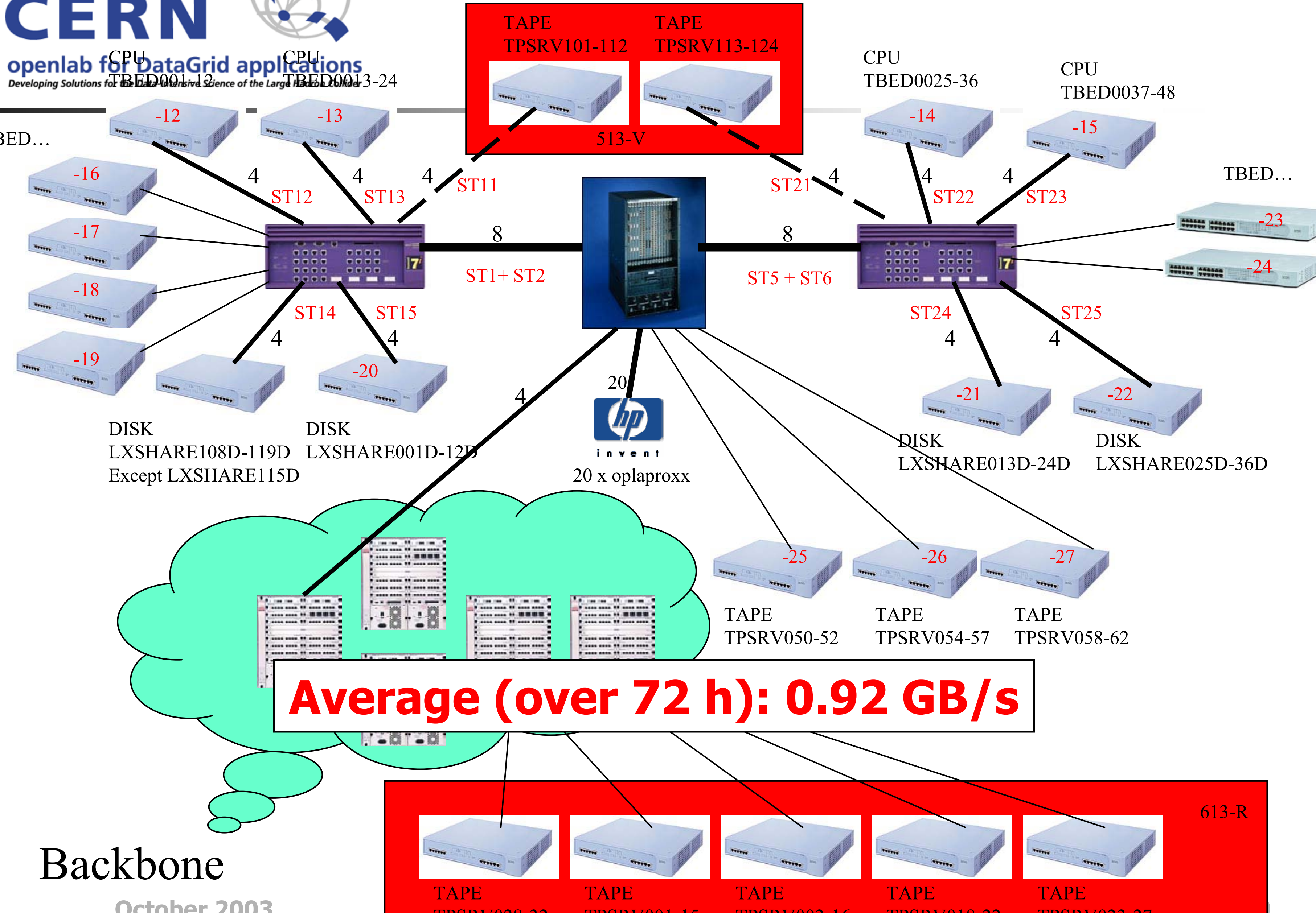
48 tape servers

▬ 10 Gigabit connection

— Fiber Gigabit connection

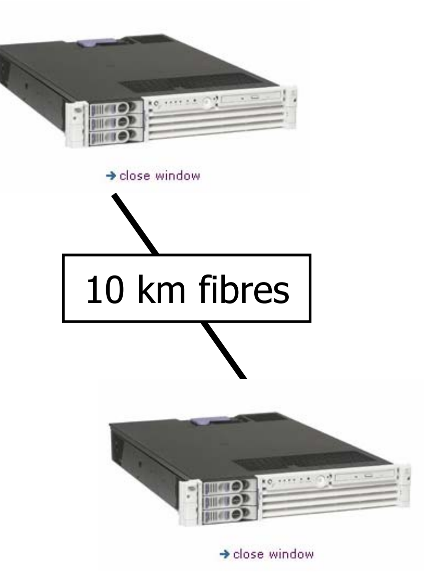
–·· Copper Gigabit connection

“1 GB/s to tape” challenge



Average (over 72 h): 0.92 GB/s

Status of 10 GbE challenge



- **Successful back-to-back tests:**
  - **Peak of 523 MB/s with 12 streams**
    - **Without ANY tuning**
  - **Peak of 755 MB/s single stream**
    - **With intensive tuning**
  - **10 km fibers used**
  - **Current limitation is the PCI-X bus**
    - **Absolute maximum thought to be 800 MB/s**
  - **Good validation of Intel NICs, HP chipset (zx1) and PCI-X bus**

Also: testing with IA-32 Xeon 2.4 GHz

IPv4 record setup: 5.44 Gbps

Internet2 Landspeed Record

(category TCP/IPv4 single stream)
Established on October 1 2003 by Caltech and CERN within the DataTAG project framework, using iperf
7’067 Km of network: Geneva-Chicago
5.44 Gbits/sec (1.1 Terabyte of data transferred in 26 minutes)

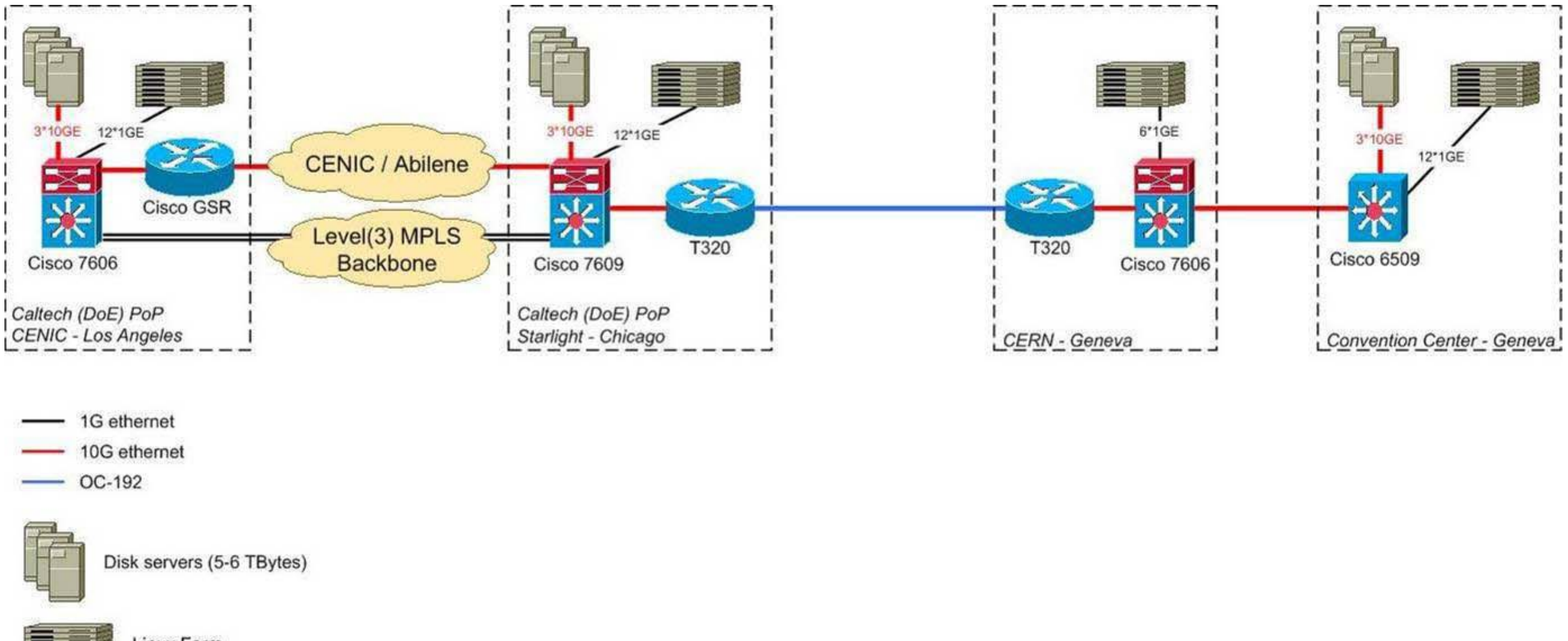
Results: 38’420.54 Terabit-meters/second

Hardware

Chicago: Dual Intel® Xeon™ processors, 3.06GHz, 2 GB RAM
SuperMicro X5DPE Motherboard (Intel E7501 chipset)
Geneva: HP RX2600, Dual Itanium2 1.5GHz, 4GB RAM
10 GbE interfaces: Intel Pro/10 GbE-LR

Software & Setup

Standard Linux (Kernel 2.6.0-test5)
MTU set to ~9000 bytes



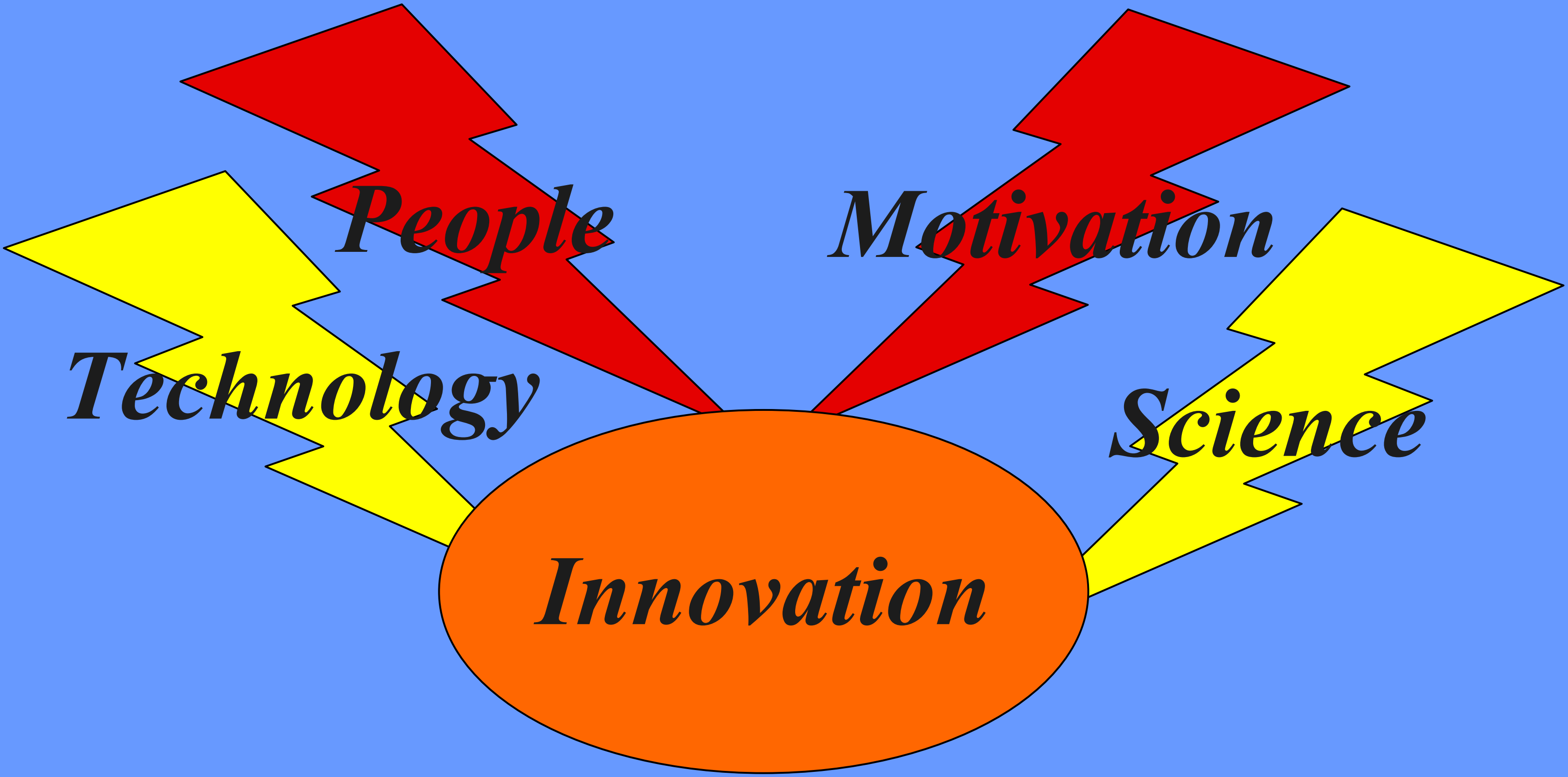
IBM StorageTank plans

- Storage Tank file system initial usage tests
- Establish a set of standard performance marks
  - raw disk speed
  - disk speed through iSCSI
  - file transfer speed through iSCSI & Storage Tank
- Storage Tank replacing Castor disk servers ?
  - Tape servers reading/writing directly from/to Storage Tank file system
- "CMS" challenge:
  - random access @ 400 MB/s on a 100 GB data set, from some 200 servers.

Opencluster and the Grid

- VDT 1.8 installed (contains Globus 2.2.4)
  - Native 64 bit version
  - First tests with Globus + LSF have begun
- Active porting of EDG 2.0 software started
- Joint project with CMS
  - Integrate opencluster alongside EDG testbed
  - Porting, Verification
    - Relevant software packages (hundreds of RPMs)
    - Understand chain of prerequisites
    - Exploit possibility to leave control node as IA-32
- Interoperability with LCG-1 testbeds
- Integration into existing authentication and virtual organization schemes
- GRID benchmarks
    - To be defined
    - Certain scalability tests already in existence

CERN
“Where the Web was born…”®


People
Motivation
Technology
Science
Innovation