LHC experimental data: From today's Data Challenges to the promise of tomorrow

B. Panzer – CERN/IT,
F. Rademakers – CERN/EP,
P. Vande Vyvre - CERN/EP

Academic Training CERN

Today

- **Day 1 (Pierre VANDE VYVRE)**
  - Outline, main concepts
  - Requirements of LHC experiments
  - Data Challenges
- **Day 2 (Bernd PANZER)**
  - Computing infrastructure
  - Technology trends
- **Day 3 (Pierre VANDE VYVRE)**
  - Data acquisition
- **Day 4 (Fons RADEMAKERS)**
  - Simulation, Reconstruction and analysis
- **Day 5 (Bernd PANZER)**
  - Computing Data challenges
  - Physics Data Challenges
  - Evolution

Data Challenges

Day 5

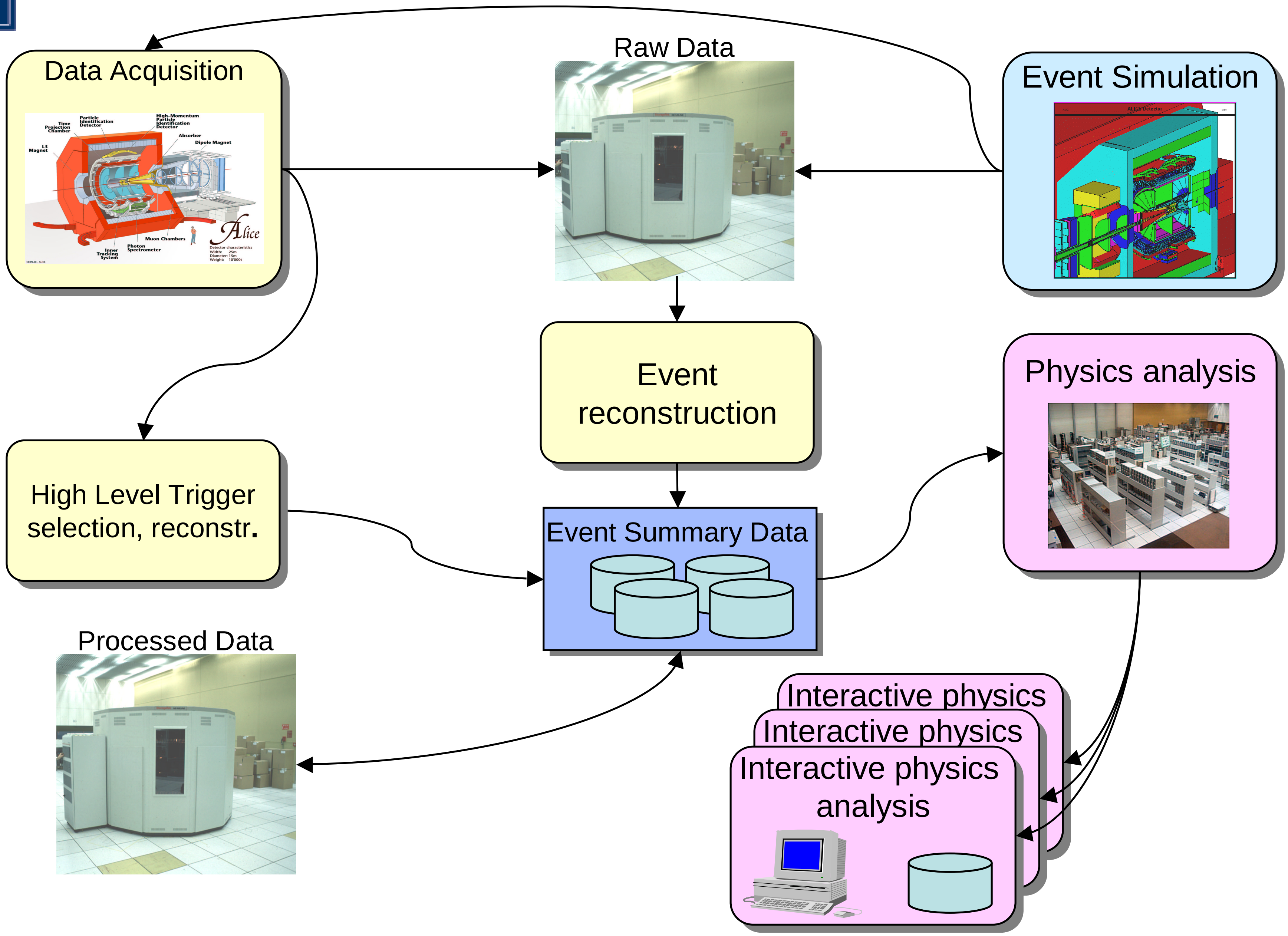
Academic Training CERN 12-16 May 2003

Bernd Panzer-Steindel CERN-IT

Outline

- Motivation
- Physics Data Challenges
- Computing Data Challenges
- Summary

Experiment dataflow



Data Acquisition

Raw Data

Event Simulation

Event reconstruction

Physics analysis

High Level Trigger selection, reconstr.

Event Summary Data

Processed Data

Interactive physics

Interactive physics

Interactive physics analysis

Considerations

- **current state of performance, functionality and reliability is good and**
  technology developments look still promising

  → **more of the same for the future !?!?**

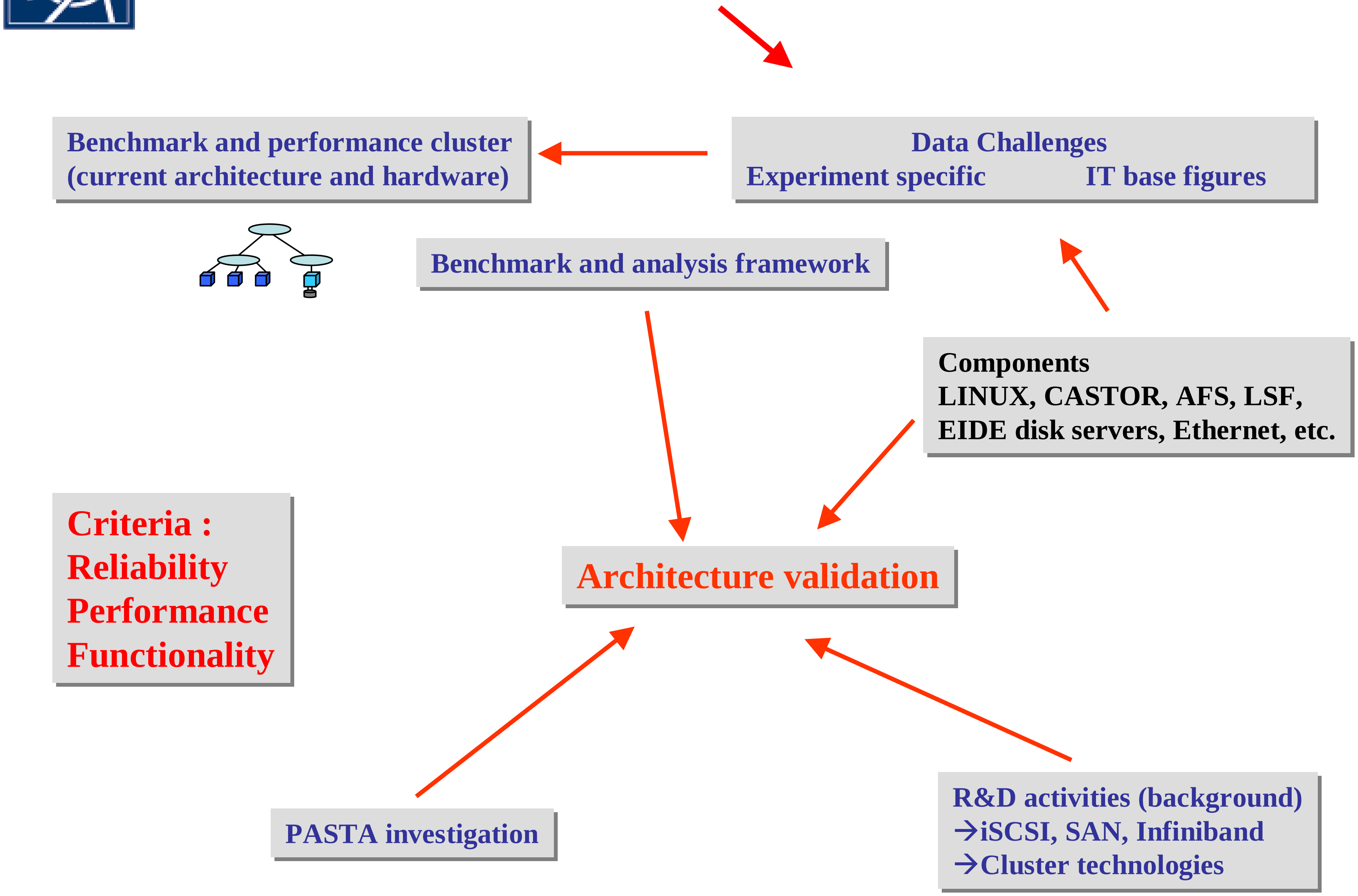
How can we be sure that we are following the right path ?

How to adapt to changes ?

Timescales

2003	major verification of the architecture
2004	further verification or verification of a DIFFERENT architecture
mid 2005	IT Computing Technical Design Report, architecture decided
end 2005	purchasing procedure starts, 10-15 million SFr value
Q3 2006	Installation of disk, cpu and tape resources
Q2 2007	first data taking

Computing model of the Experiments



Benchmark and performance cluster
(current architecture and hardware)

Data Challenges
Experiment specific IT base figures

Benchmark and analysis framework

Components
LINUX, CASTOR, AFS, LSF,
EIDE disk servers, Ethernet, etc.

Criteria :
Reliability
Performance
Functionality

Architecture validation

PASTA investigation

R&D activities (background)
→iSCSI, SAN, Infiniband
→Cluster technologies

Strategy

- continue and expand the current system

BUT do in parallel :

- R&D activities
  SAN versus NAS, iSCSI, IA64 processors, ….
- technology evaluations
  infiniband clusters, new filesystem technologies,…..
- Data Challenges to test scalabilities on larger scales
  “bring the system to it’s limit and beyond “
  we are very successful already with this approach, especially with the “beyond” part
- watch carefully the market trends

Challenges

1. **Status of the current system**

   Is the stability of the equipment acceptable ?
   stress test the equipment ?
   where and what are the weak points / bottlenecks ?

2. **Physics Data Challenges**

   test the bookkeeping, organization and management of data processing

3. **Computing Data Challenge**

   scalability of software and hardware in the fabric

   try to verify whether the current architecture would survive the anticipated load in the LHC area.

Data Challenge Areas

Physics Data Challenges

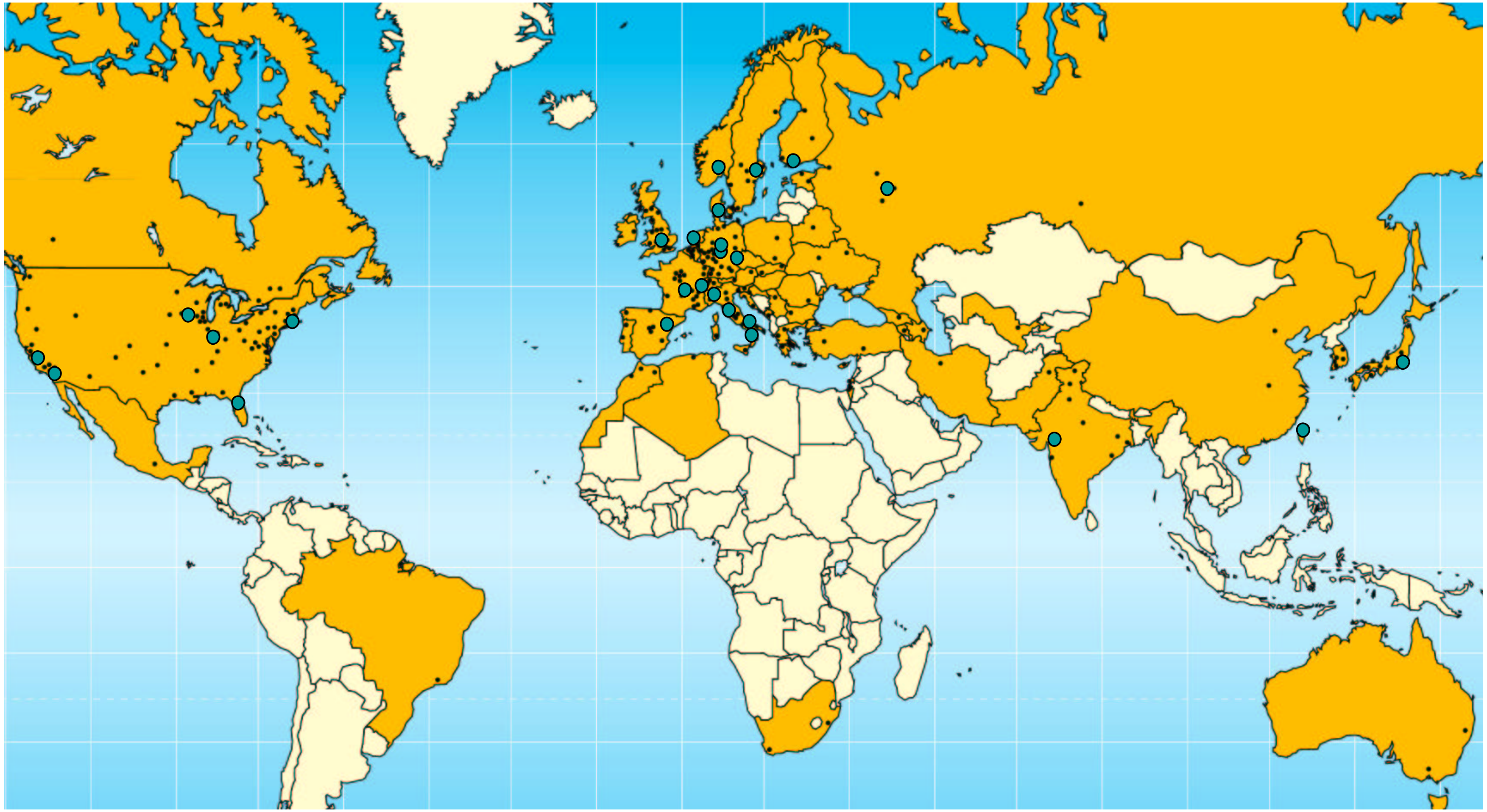
- user applications from the 4 experiments
- experiment infrastructure for processing (bookkeeping, verifications, ….)
- Wide Area Network connections and network protocol software
- GRID middleware to ‘connect’ different sites for task distribution and data exchange

Computing Data Challenges

- system application software to organize site specific distribution of tasks and load balancing
  site hierarchical storage management (HSM) system
- network to couple the hardware components
- operating system + system software
- hardware : processors, nodes, storage

Physics Data Challenges

Centres taking part in the LCG prototype service (2003-05)



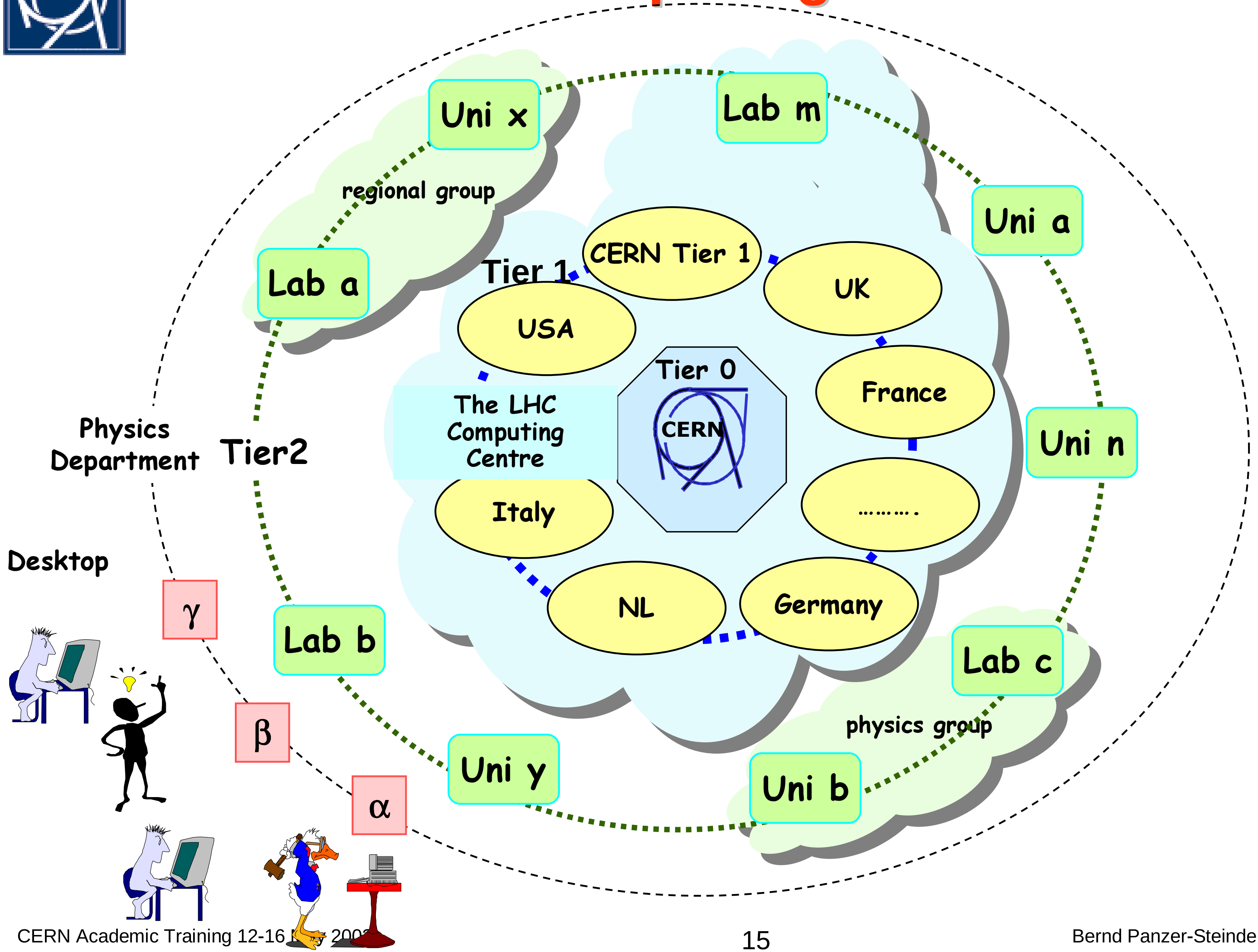
around the world → around the clock

The LHC Computing Environment

Multi-Tier Model

- Tier 0: CERN
  - data recording and reconstruction, repository for all data.
- Tier 1: CERN and a small number of centres
  - full range of services including managed mass storage, user support and high bandwidth networking
  - full copy of the event summary data (ESD), sample of the raw data
  - full range of analysis activities, with emphasis on data-intensive batch processing
- Tier 2:
  - reliable batch and interactive services, supported by good networking to Tier 1 centres
  - substantial data storage - for analysis and simulation
- Tier 3+:
  - Local facilities, with the emphasis on interactive analysis and simulation.

LHC Computing Model



Middleware

- each experiment has already it’s own pre-GRID version of a distributed production environment
  → bookkeeping
  → distribution of jobs
  → tracking of jobs and problems
  → collecting the output
  → distribution of the software environment
  → transfer of input and output data

- continue to use and improve these programs

- adapt and integrate the new GRID middleware produced by the different GRID projects

DataGrid in Numbers

People

>350 registered users

12 Virtual Organisations

16 Certificate Authorities

>200 people trained

278 man-years of effort

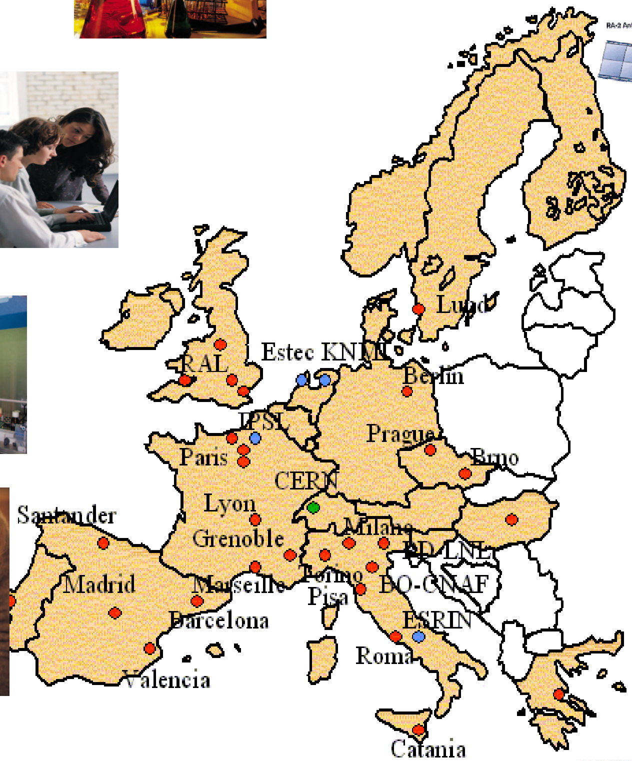
100 years funded

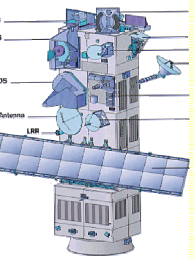
Software

50 use cases

18 software releases

>300K lines of code





Testbeds

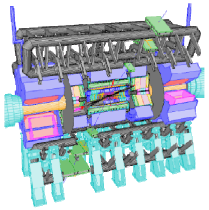
>15 regular sites

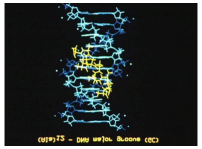
>10'000s jobs submitted

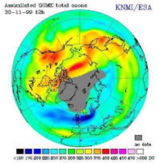
>1000 CPUs

>5 TeraBytes disk

3 Mass Storage Systems







GEANT

Scientific applications

5 Earth Obs institutes

9 bio-informatics apps

6 HEP experiments

Deploying the LHC Grid

Three Stages

2003 –

- Establish the LHC grid as a reliable, manageable, permanently available service including the Tier 1 and many Tier 2 centres
- Serve as one of the computing facilities used for simulation campaigns during 2H03

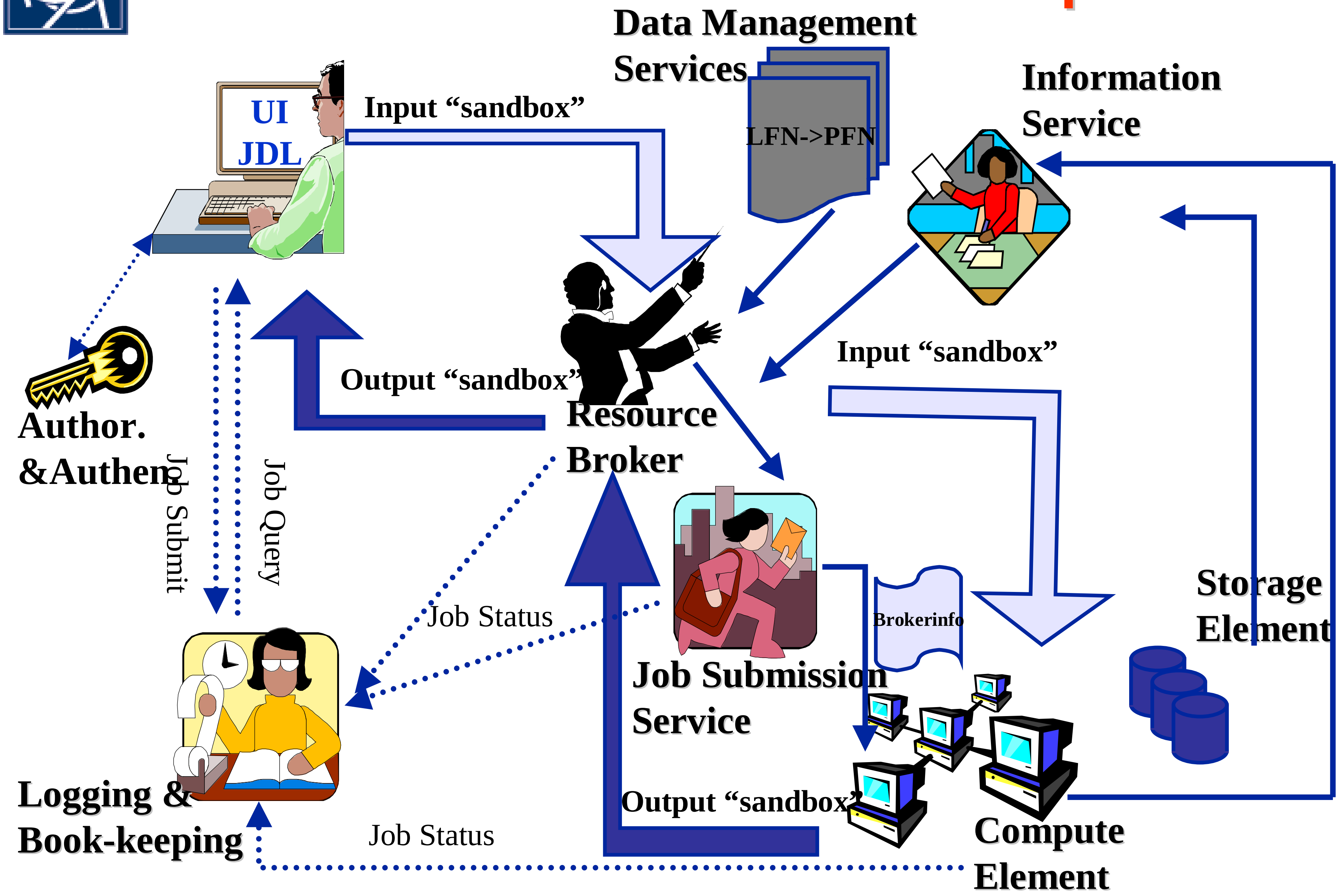
2004 –

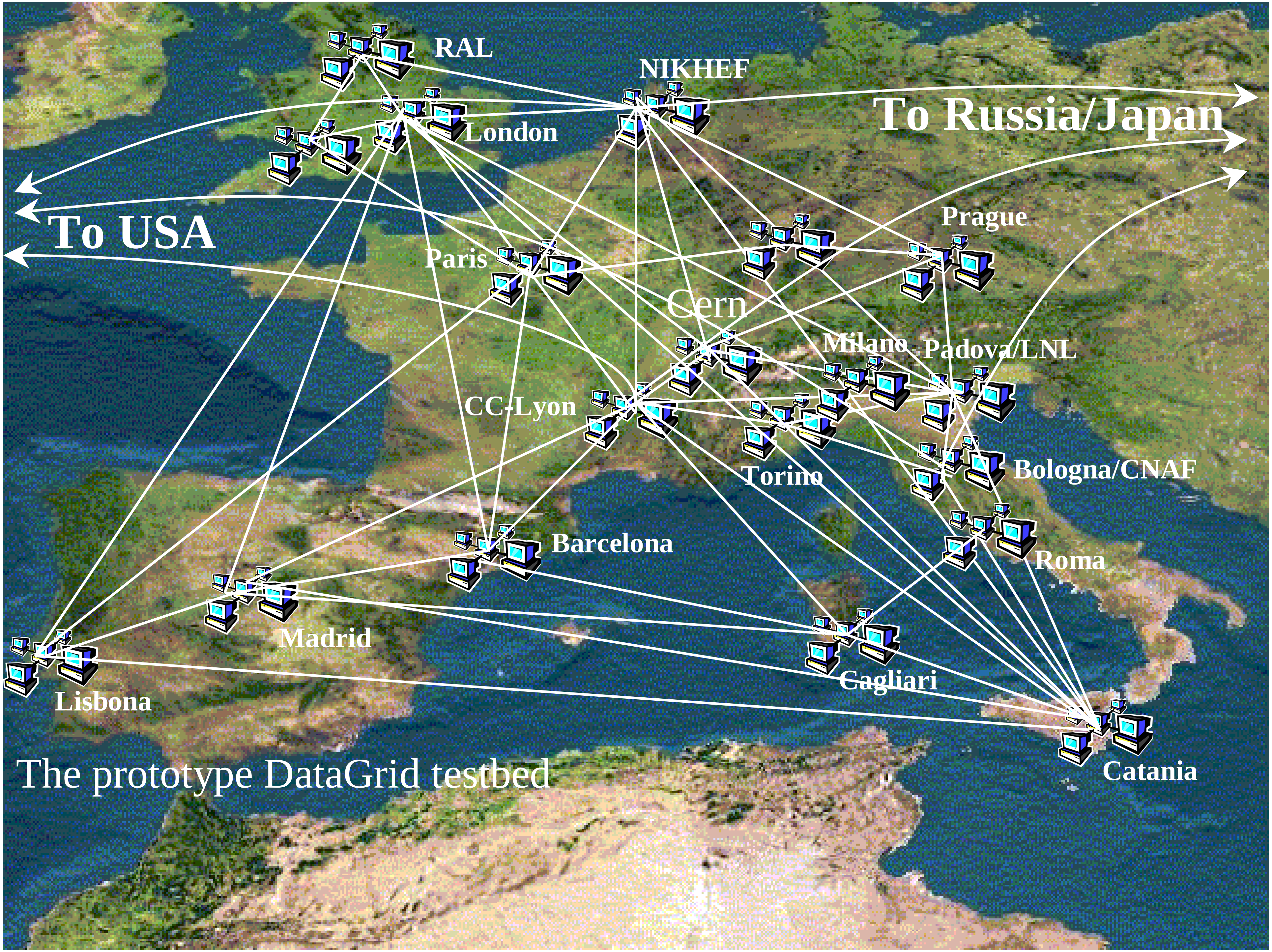
- Stable service for batch analysis
- Scaling and performance tests, commissioning of operations infrastructure
- Computing model tests – 4 collaborations
  Tier 0 – Tier 1 – Tier 2 – Tier 3 → Computing TDRs at end 2004

2005 –

- Full prototype of initial LHC service – second generation middleware
  - validation of computing models (4 collaborations)
  - validation of physical implementation – technology, performance, scaling
- LCG TDR – sizing/cost/schedule for the initial LHC service – July 2005

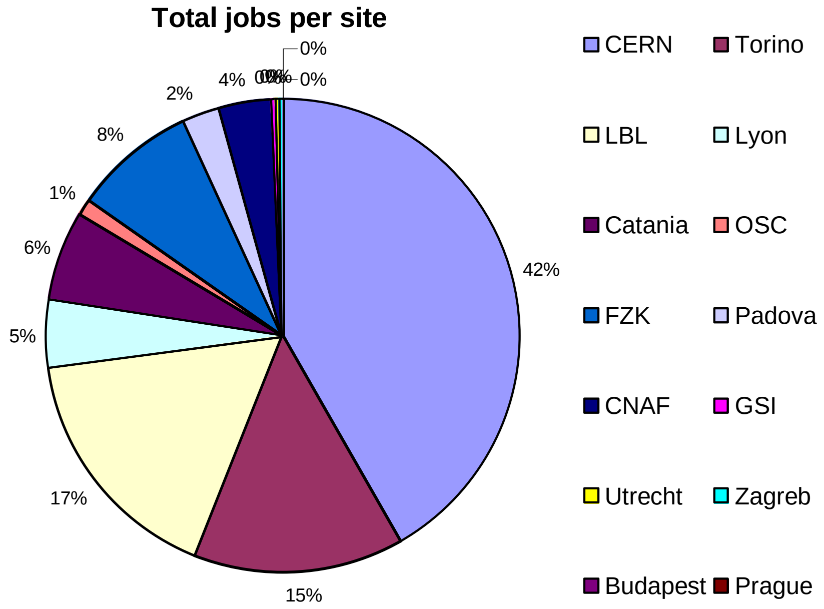
A Job Submission Example

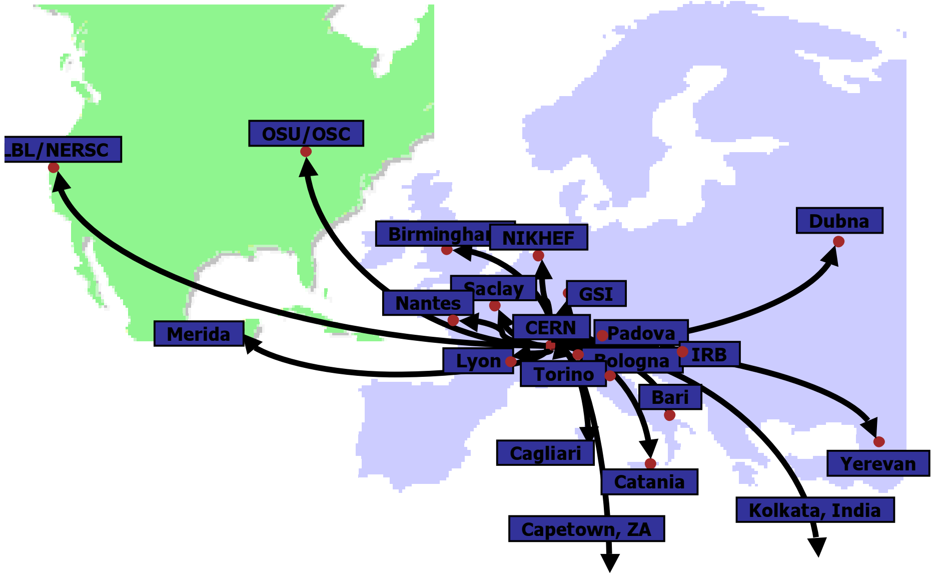




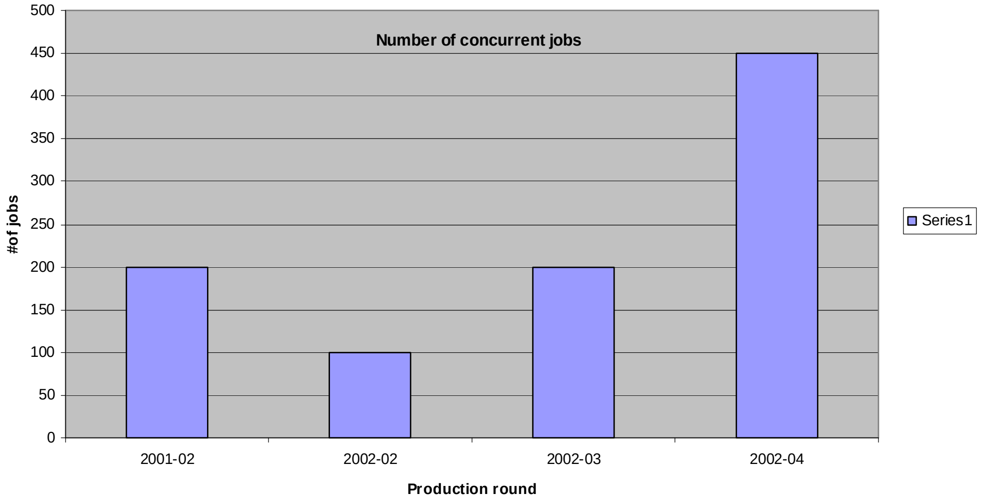
The prototype DataGrid testbed

AliEn progress

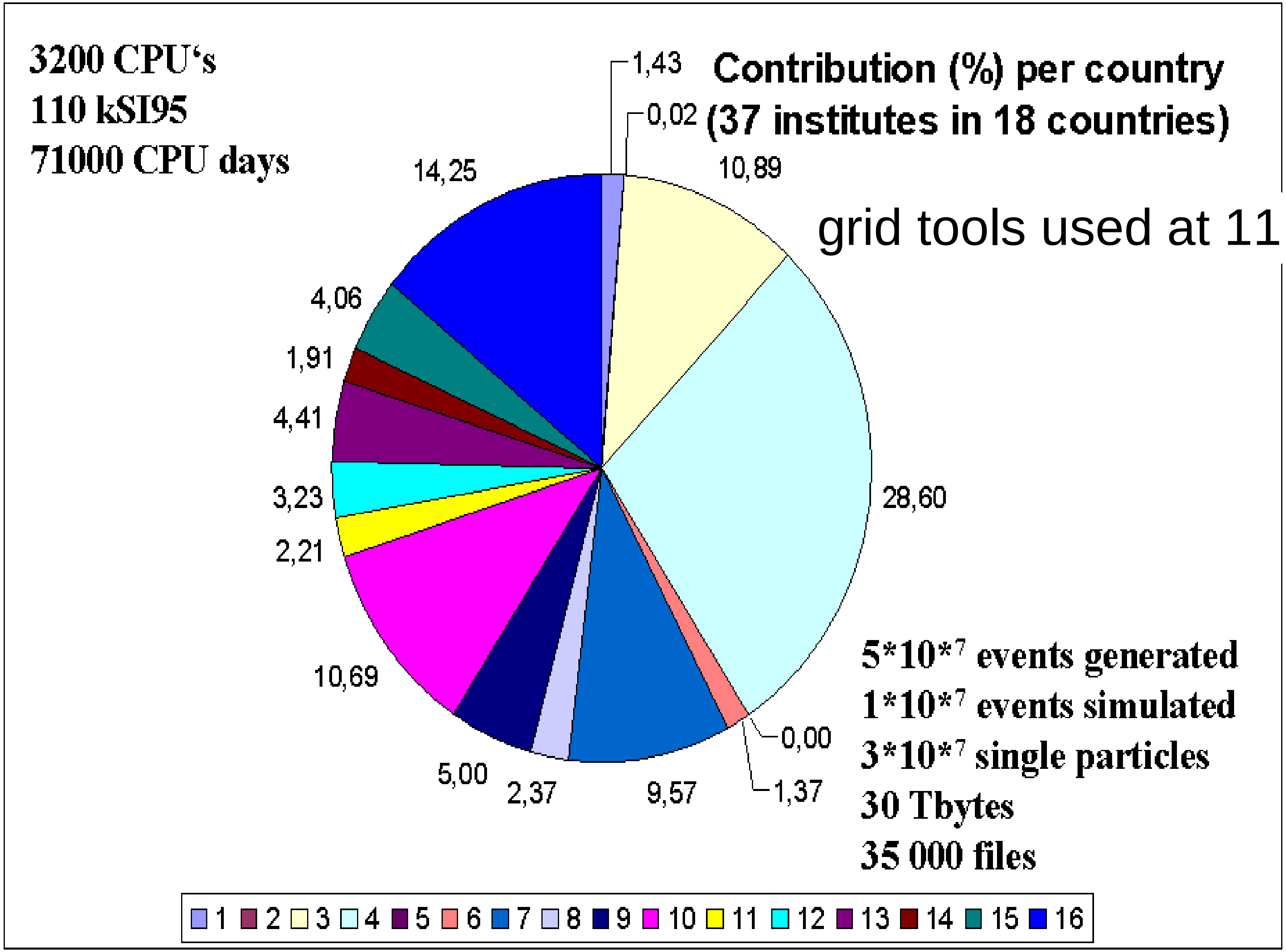




- 32 (was 28) sites configured
- 5 (was 4) sites providing mass storage
- 12 production rounds
- 22773 jobs validated, 2428 failed (10%) (PPR production)
- Up to 450 concurrent jobs
- 0.5 operators



ATLAS DC1 Phase 1 : July-August 2002



3200 CPU's
110 kSI95
71000 CPU days

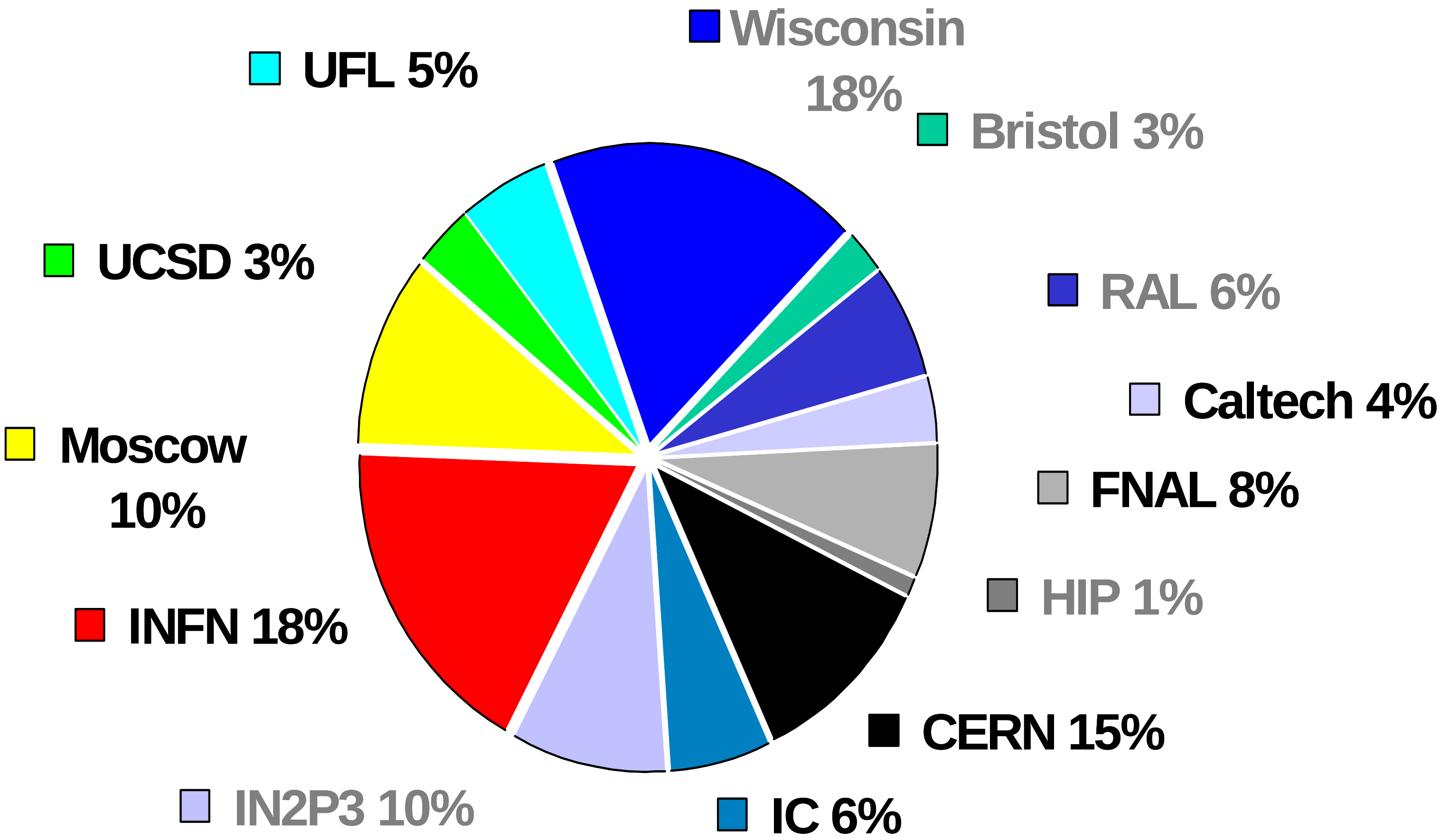
Contribution (%) per country
(37 institutes in 18 countries)

grid tools used at 11 sites

5*10*^{7} events generated
1*10*^{7} events simulated
3*10*^{7} single particles
30 Tbytes
35 000 files

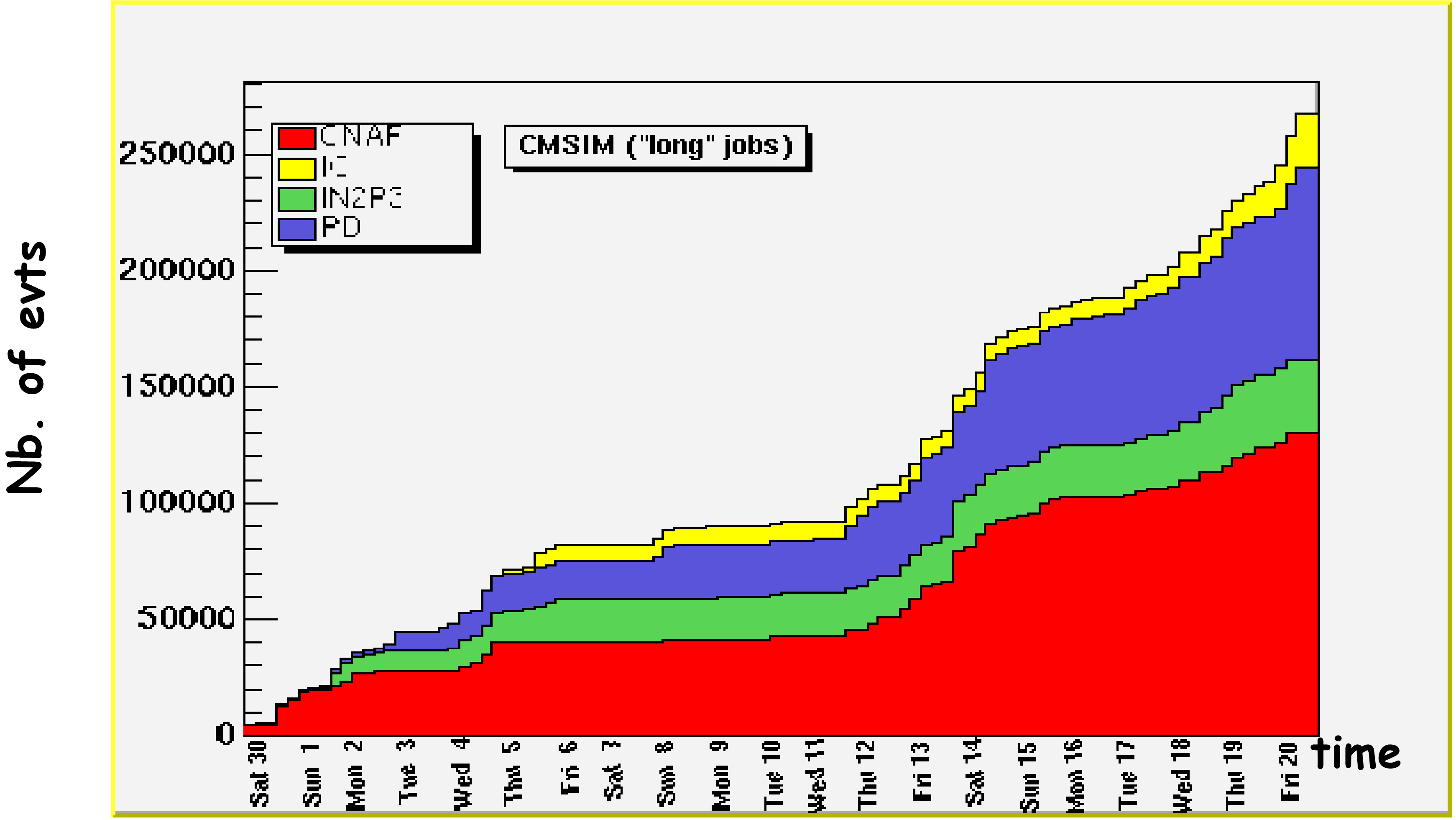
Spring02: CPU Resources

Most Resources not at CERN
(CERN not even biggest Single Resource)



6 million events
~20 sites

CMS event production in December 2002 using EDG software and applications



CMS/EDG Summary of Stress Test
Preliminary Analysis

Short jobs

After Stress Test – Jan 03

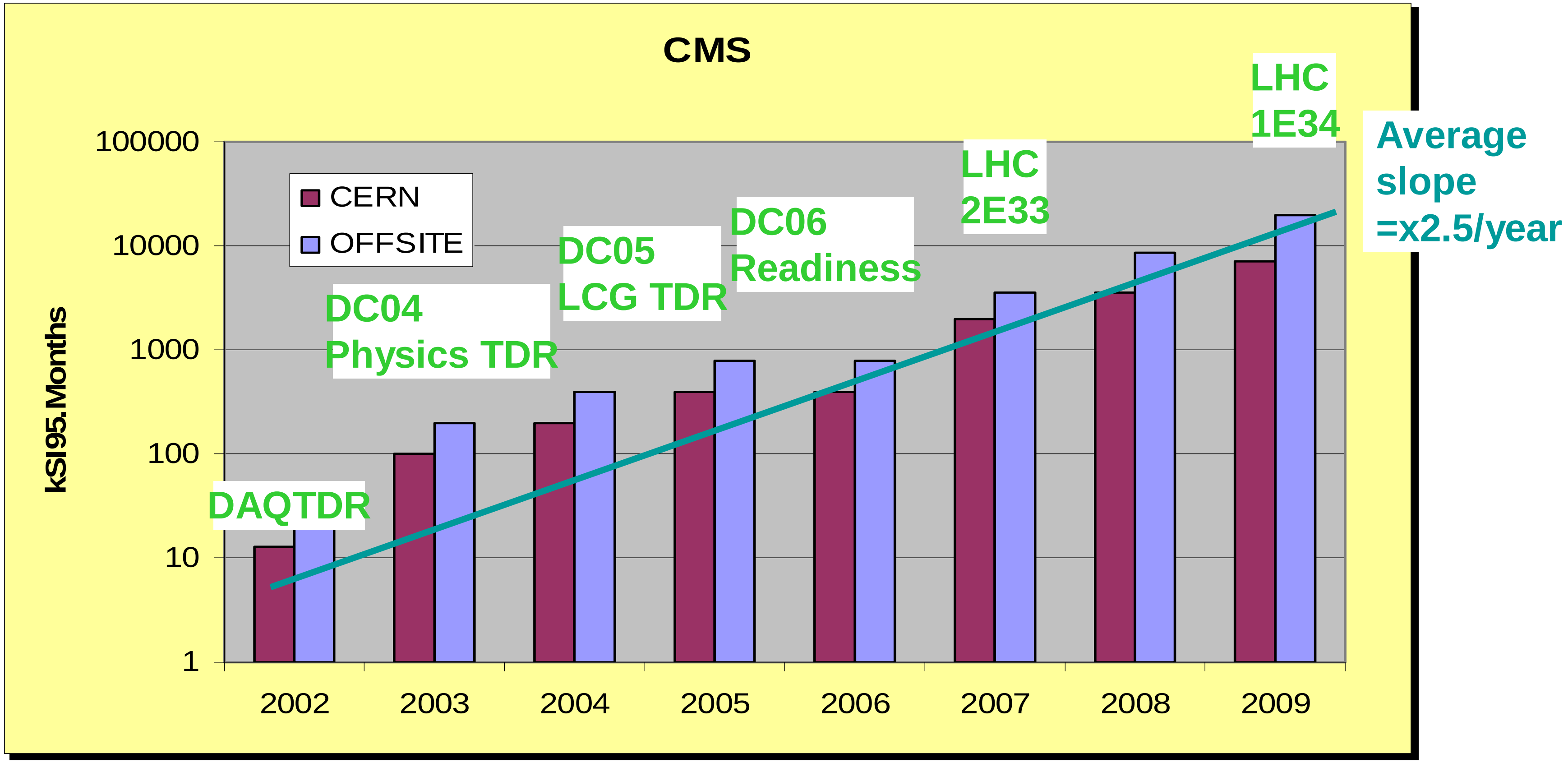
CMKIN jobs			
Status	EDG evaluation	CMS evaluation	EDG ver 1.4.3
Finished Correctly	5518	4601	604
Crashed or bad status	818	1099	65
Total number of jobs	6336	5700	669
Efficiency	0.87	0.81	0.90

Long jobs

After Stress Test – Jan 03

CMSIM jobs			
Status	EDG evaluation	CMS evaluation	EDG ver 1.4.3
Finished Correctly	1678	2147	394
Crashed or bad status	2662	934	104
Total number of jobs	4340	3081	498
Efficiency	0.39	0.70	0.79

CPU Power Ramp Up



Time shared Resources

Dedicated CMS Resources

CMS Data Challenges

- **2000/2001**

  Verify code, bring up production worldwide, prepare for DAQ TDR
- **2002**

  DAQ TDR massive production and analysis
- **2003/4 (DC04)**

  First Year of Physics TDR, GEANT4 in Production

  New Persistency, First truly GRID dependant challenge

  Verify model and components for CMS Computing TDR
- **2004/5 (DC05)**

  Verify LCG2 Prototype in time for LCG TDR
- **2005/6 (DC06)**

  Final Readiness Check, all Software and Computing systems
- **2007**

  First Data. Ready for new Physics in first few fb^{-1}

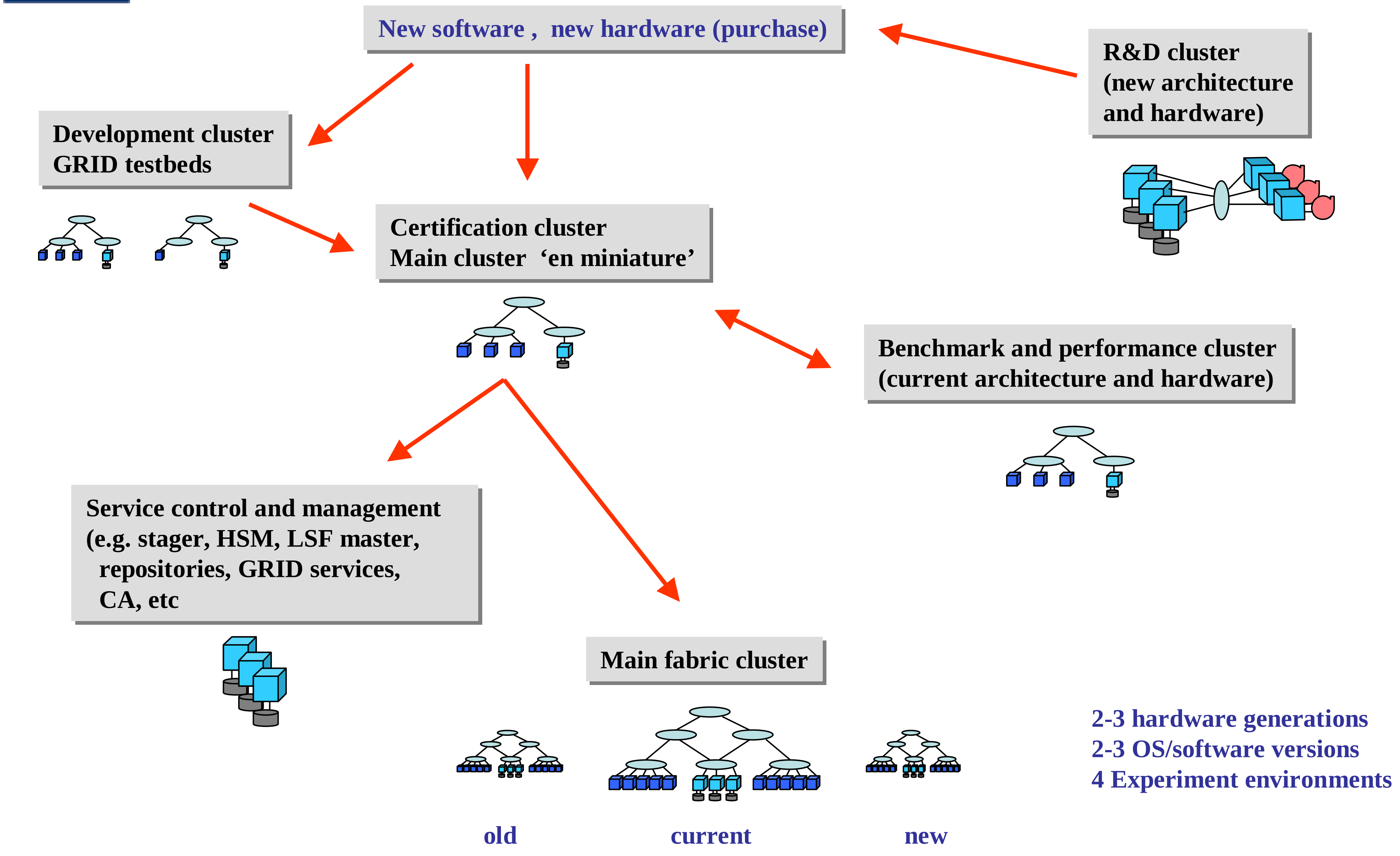
Computing Data Challenges

Current CERN Fabrics architecture

based on :

- In general on commodity components
- Dual Intel processor PC hardware for CPU, disk and tape Server
- Hierarchical Ethernet (100, 1000, 10000) network topology
- NAS disk server with EIDE disk arrays
- RedHat Linux Operating system
- Medium end tape drive (linear) technology
- OpenSource software for storage (CASTOR, OpenAFS)

General Fabric Layout



New software , new hardware (purchase)

R&D cluster
(new architecture
and hardware)

Development cluster
GRID testbeds

Certification cluster
Main cluster ‘en miniature’

Benchmark and performance cluster
(current architecture and hardware)

Service control and management
(e.g. stager, HSM, LSF master,
repositories, GRID services,
CA, etc

Main fabric cluster

2-3 hardware generations
2-3 OS/software versions
4 Experiment environments

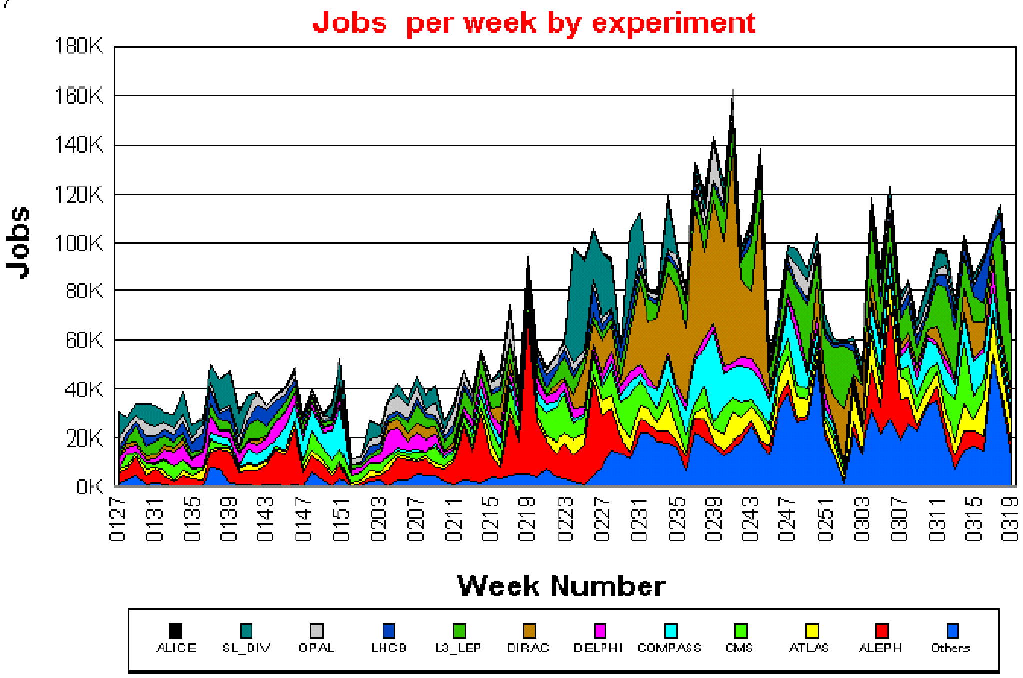
Farm Status

~1000 nodes running batch jobs at ~ 65% cpu utilization

Stability :

~10 reboots per day == 1%
0.7 Hardware interventions per day
(mostly disk problems)

→Average job length ~ 2.3 h,
→ 3 jobs per nodes
== Loss rate is 0.3 %



Jobs per week by experiment

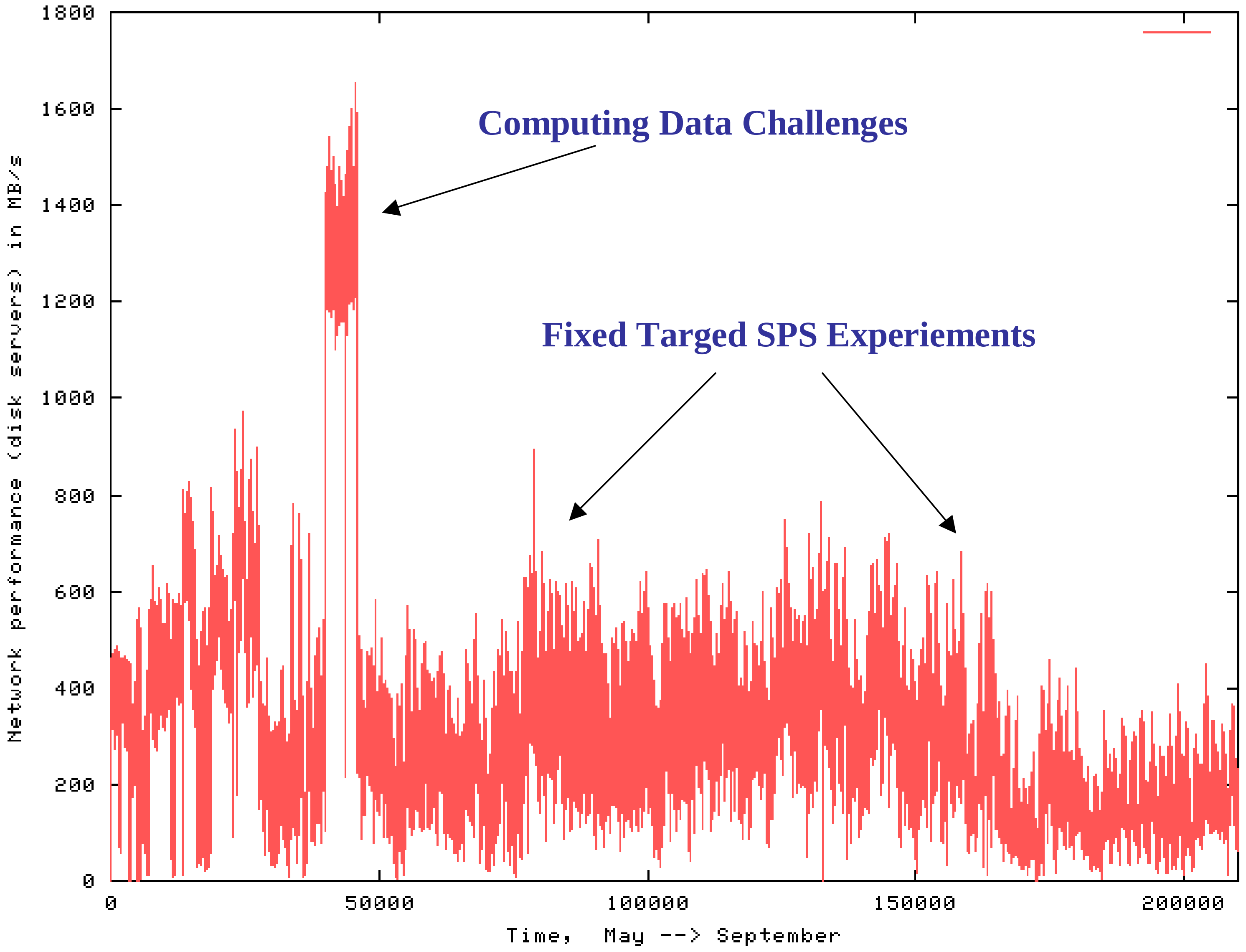
Storage Status

- **Disk stress tests :**
  30 servers with 232 disks running for 30 days I/O tests (multiple streams per disk, random+sequential) ~ 3 PB → 4 disk server crashes and one disk problem (~> 160000 disk MTBF)

- **Stability :**
  About 1 reboot per week (out of ~200 disk servers in production)and ~one disk error per week (out of ~3000 disks in production)

- **Tape system Stability :**
  About one intervention per week on one drive
  bout 1 tape with recoverable problems per 2 weeks(to be send to STK HQ)

Aggregate disk server Network traffic

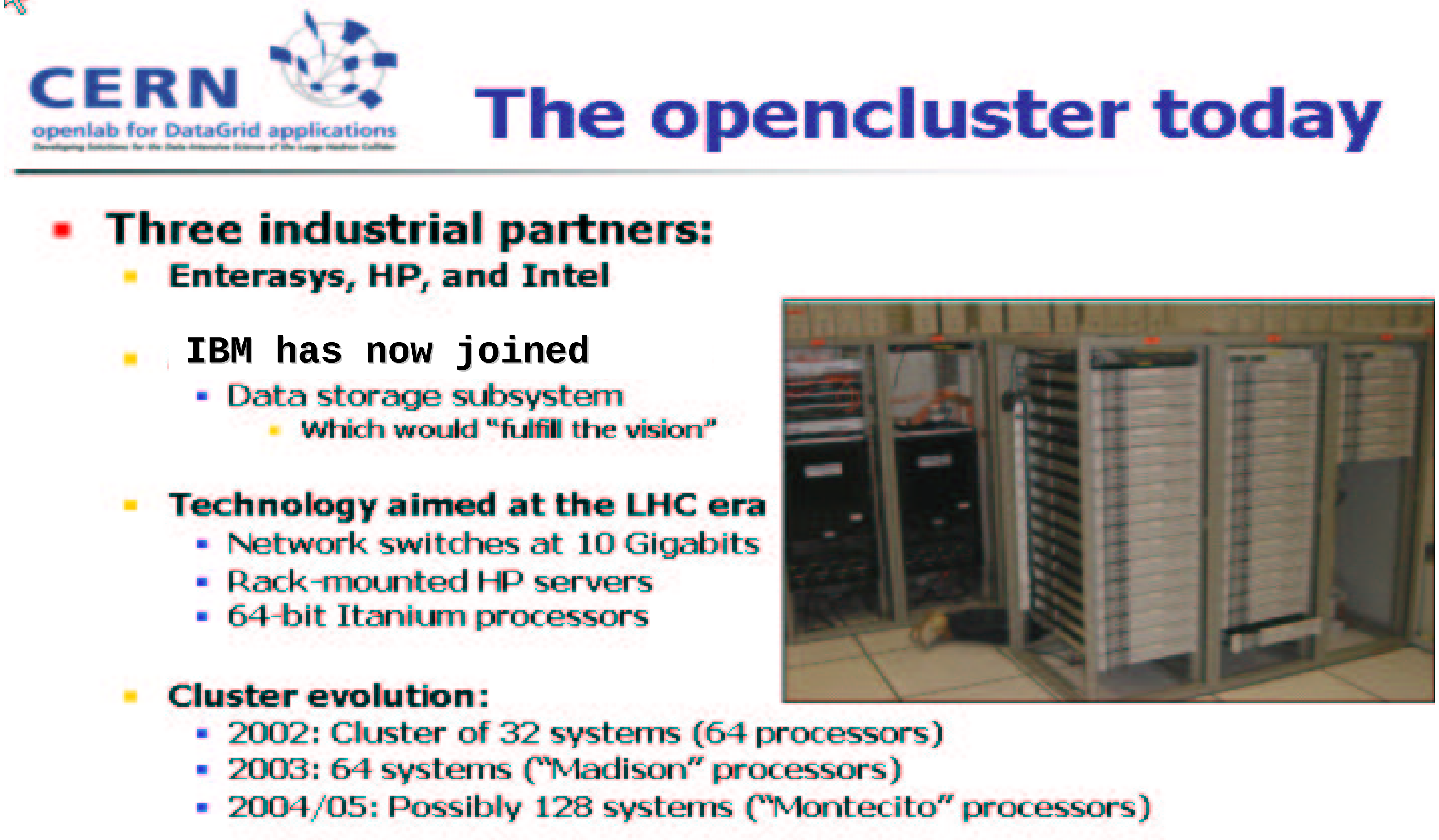


Computing Data Challenges

Fixed Targed SPS Experiements

read+write traffic in 2002

CERN openlab

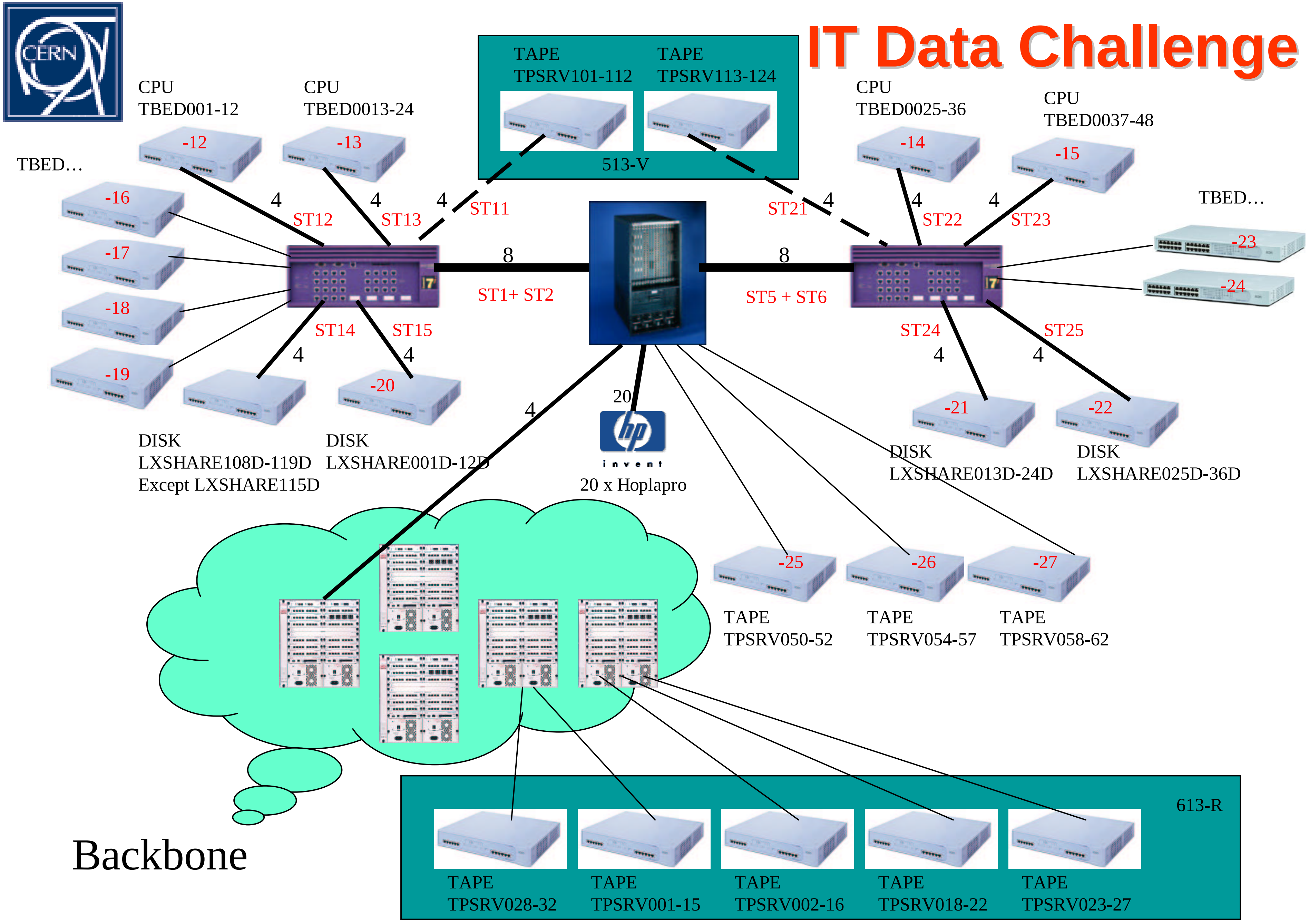


CERN openlab for DataGrid applications

The opencluster today

- Three industrial partners:
  - Enterasys, HP, and Intel
  - IBM has now joined
    - Data storage subsystem
      - Which would "fulfill the vision"
  - Technology aimed at the LHC era
    - Network switches at 10 Gigabits
    - Rack-mounted HP servers
    - 64-bit Itanium processors
  - Cluster evolution:
    - 2002: Cluster of 32 systems (64 processors)
    - 2003: 64 systems ("Madison" processors)
    - 2004/05: Possibly 128 systems ("Montecito" processors)

IT Data Challenge



TAPE
TPSRV101-112

TAPE
TPSRV113-124

513-V

CPU
TBED001-12

CPU
TBED0013-24

CPU
TBED0025-36

CPU
TBED0037-48

TBED…

TBED…

-12 -13 -14 -15 -16 -17 -18 -19 -20 -21 -22 -23 -24 -25 -26 -27

ST11 ST12 ST13 ST14 ST15 ST21 ST22 ST23 ST24 ST25

ST1+ ST2

ST5 + ST6

4 4 4 4 4 4 4 4 4 4 4

8 8

20

20 x Hoplapro

DISK
LXSHARE108D-119D
Except LXSHARE115D

DISK
LXSHARE001D-12D

DISK
LXSHARE013D-24D

DISK
LXSHARE025D-36D

TAPE
TPSRV050-52

TAPE
TPSRV054-57

TAPE
TPSRV058-62

613-R

TAPE
TPSRV028-32

TAPE
TPSRV001-15

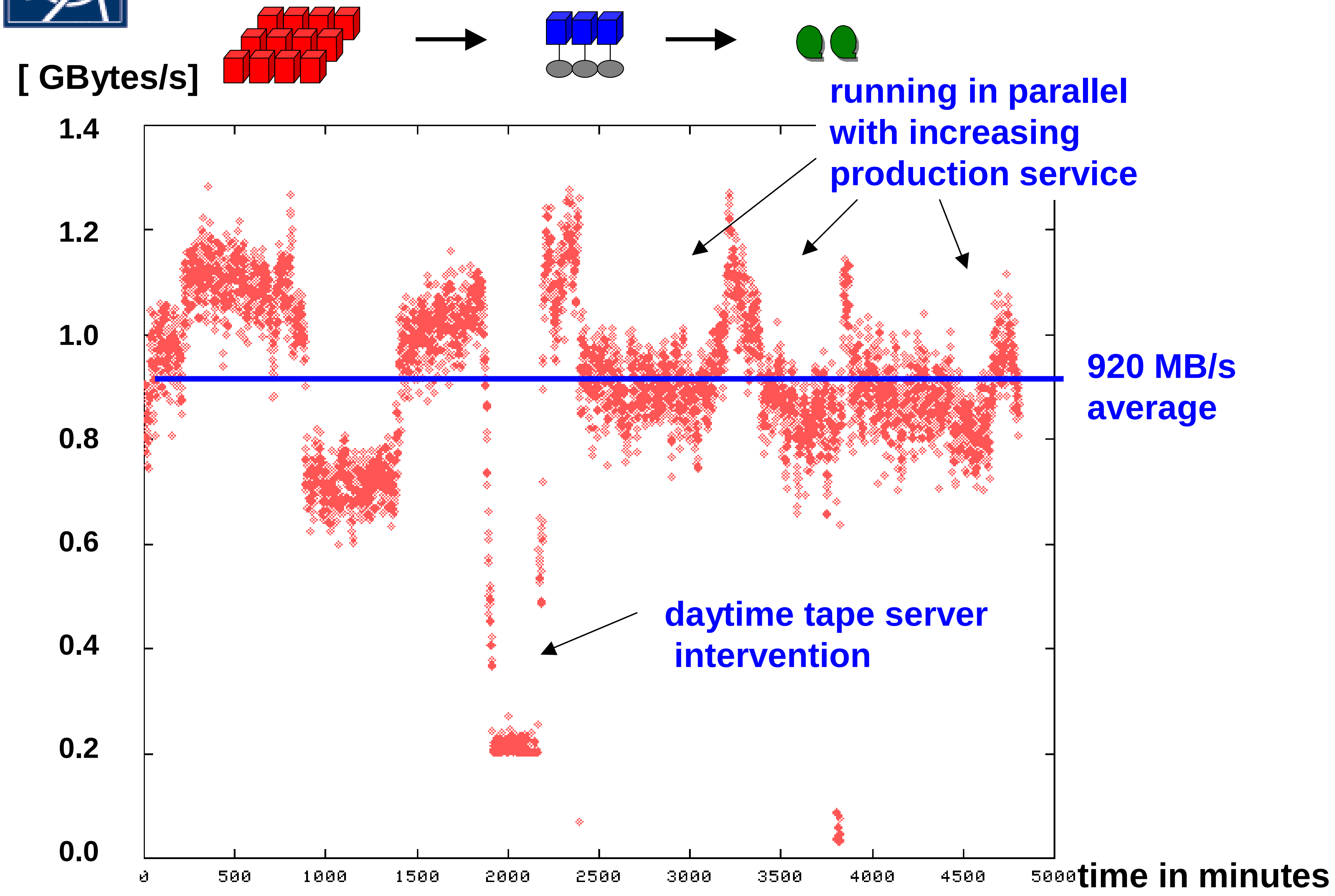
TAPE
TPSRV002-16

TAPE
TPSRV018-22

TAPE
TPSRV023-27

Backbone

IT Data Challenge performance



running in parallel with increasing production service

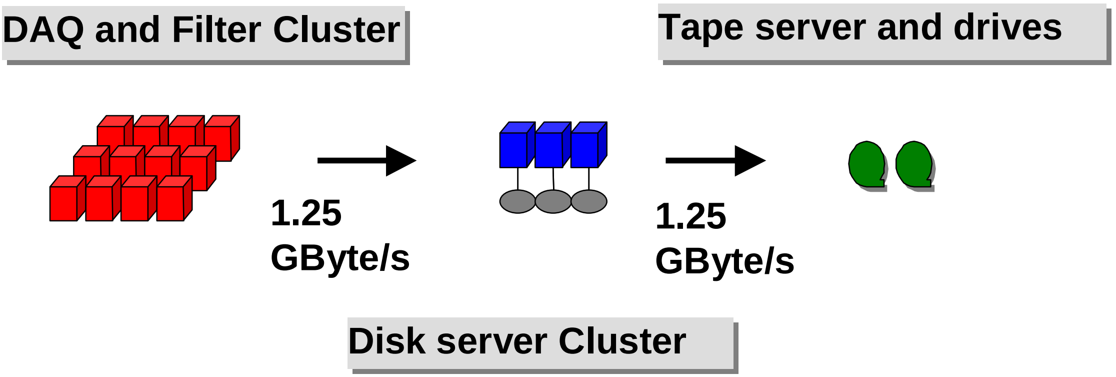
920 MB/s average

daytime tape server intervention

ALICE-IT Computing Data Challenges

- focused on CDR (Central Data Recording)
- challenges 1-2 times a year since more than 3 years
- increasing performance goals
- the first 2 DC's were extremely challenging with many problems
  continuously improved the performance, but also the methods and the ways of working together

ALICE, Central Data Recording, CERN Tier 0



DAQ and Filter Cluster

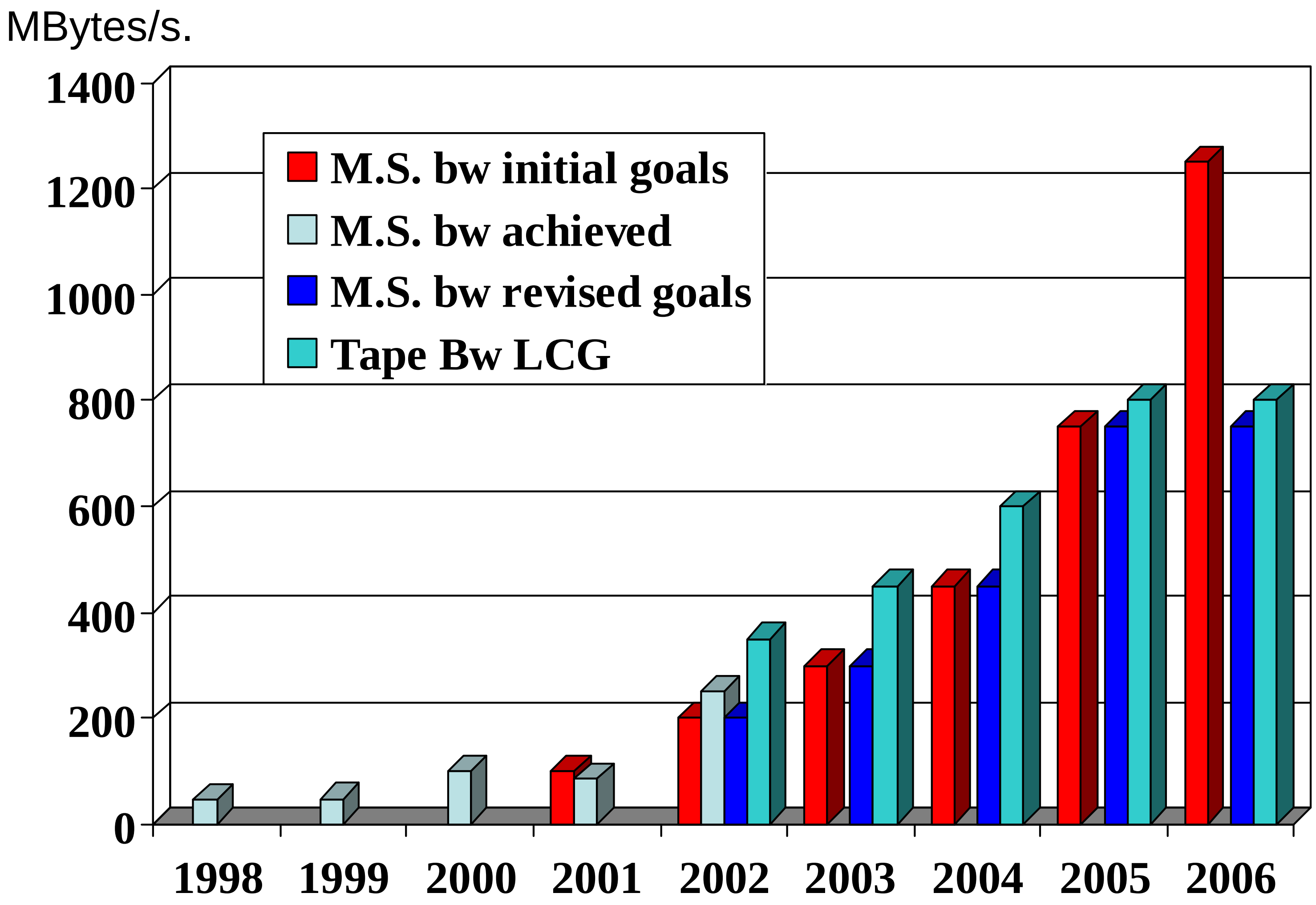
Tape server and drives

1.25 GByte/s

1.25 GByte/s

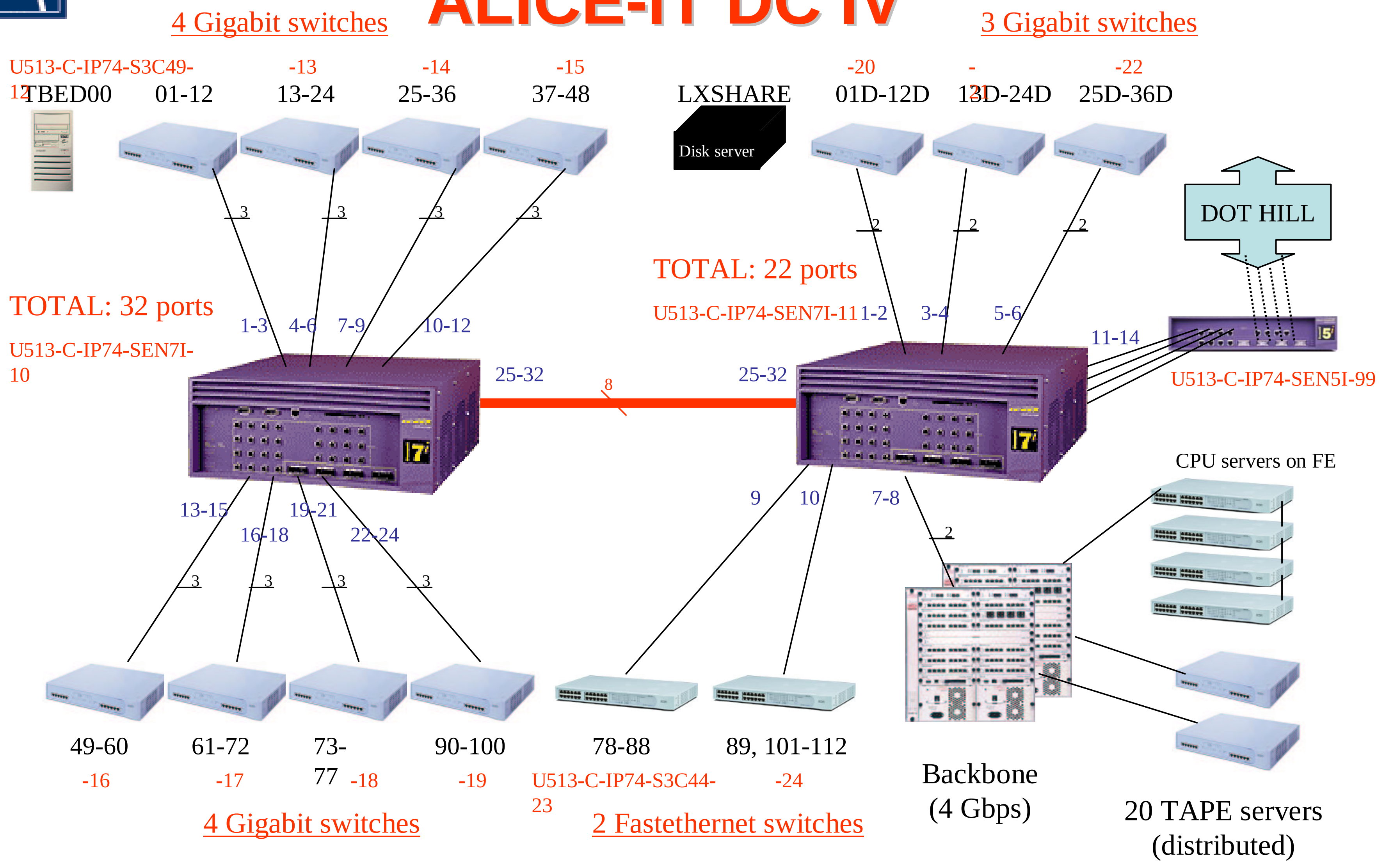
Disk server Cluster

ALICE DC – MSS Bw



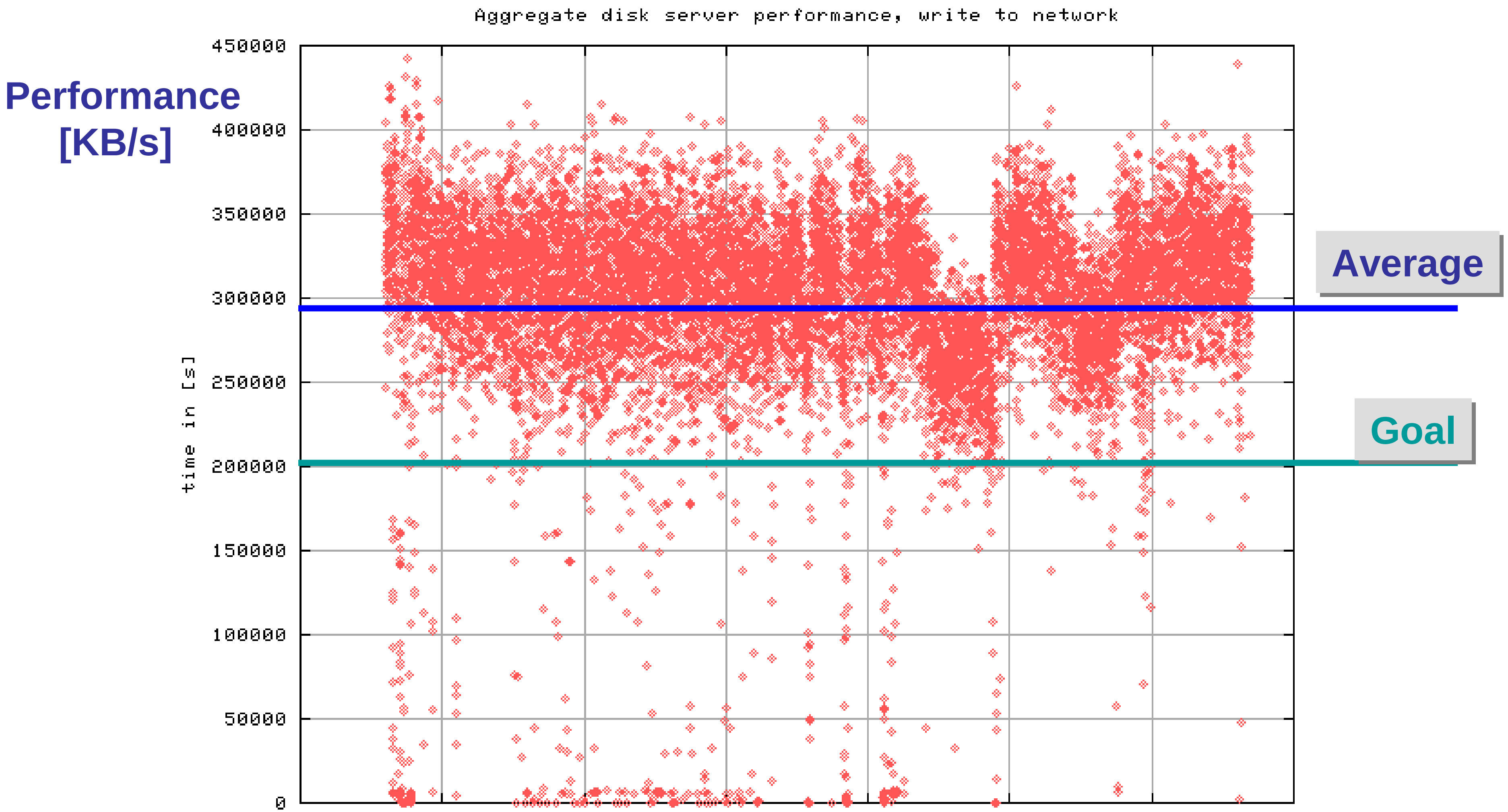
Hardware and Network Topology

ALICE-IT DC IV



Total: 192 CPU servers (96 on Gbe, 96 on Fe), 36 DISK servers, 20 TAPE servers

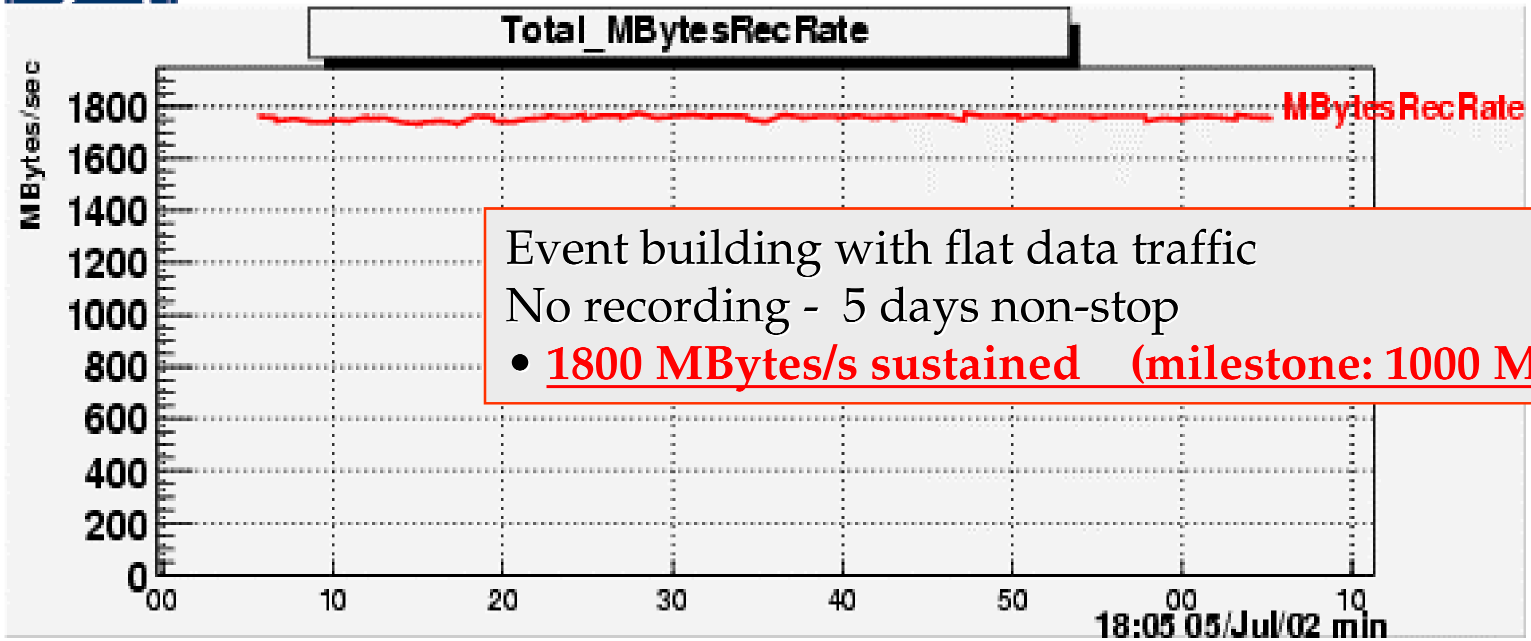
ALICE-IT DC IV



Time, from Friday 6th to Friday 13th in December 2002

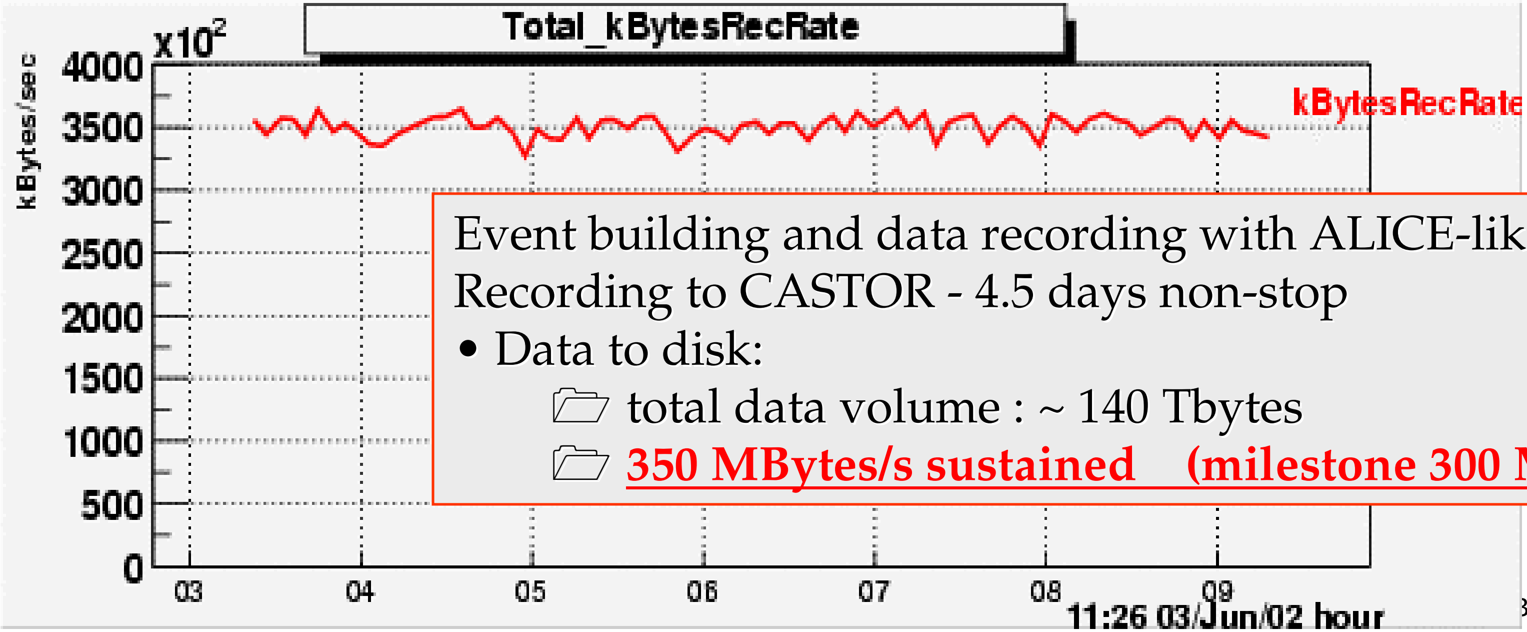
Aggregate disk server performance in 40s time intervals → writing to tape

ADC IV performances – Period 1



Event building with flat data traffic
No recording - 5 days non-stop

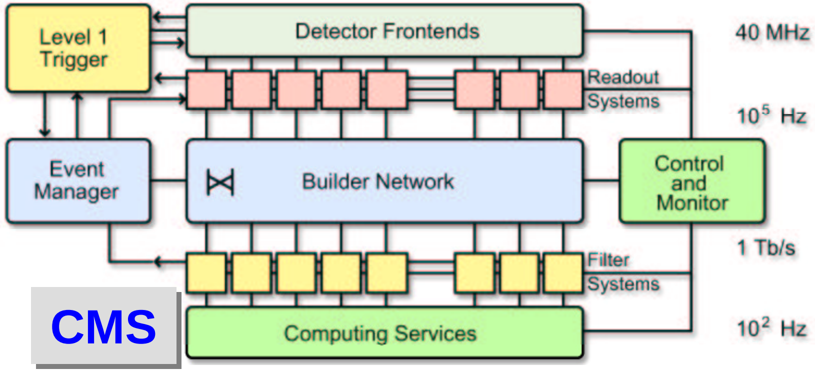
- 1800 MBytes/s sustained (milestone: 1000 Mbytes/s)

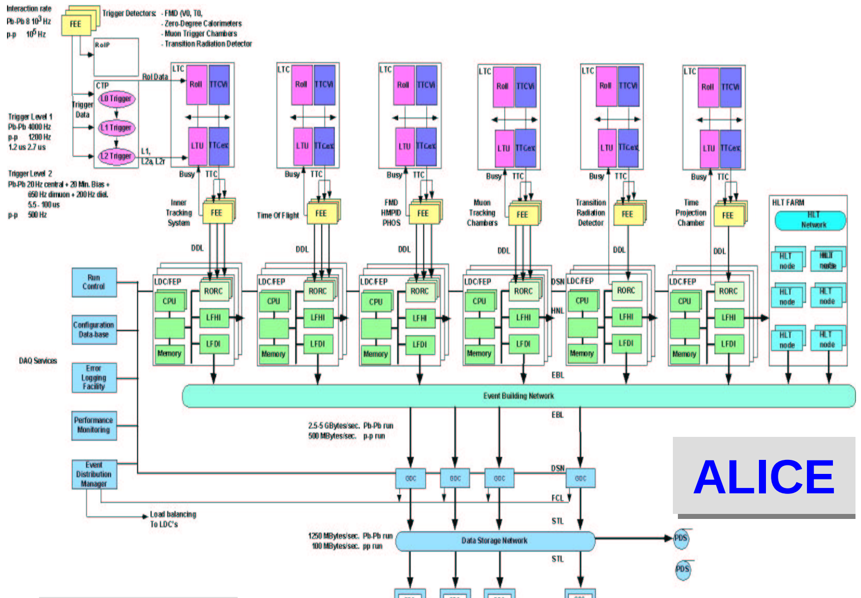


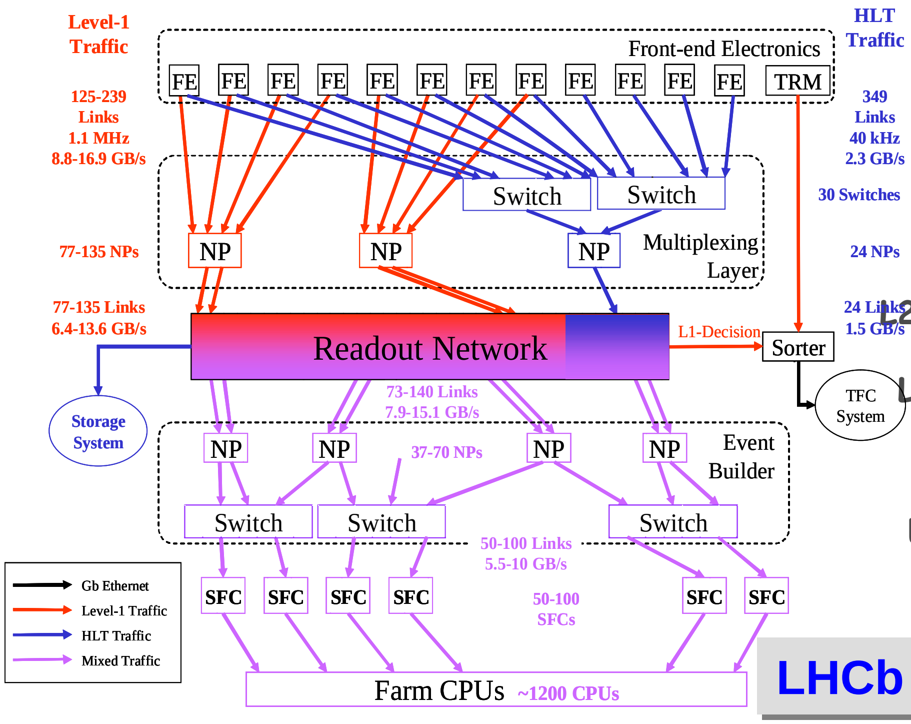
Event building and data recording with ALICE-like data traffic
Recording to CASTOR - 4.5 days non-stop

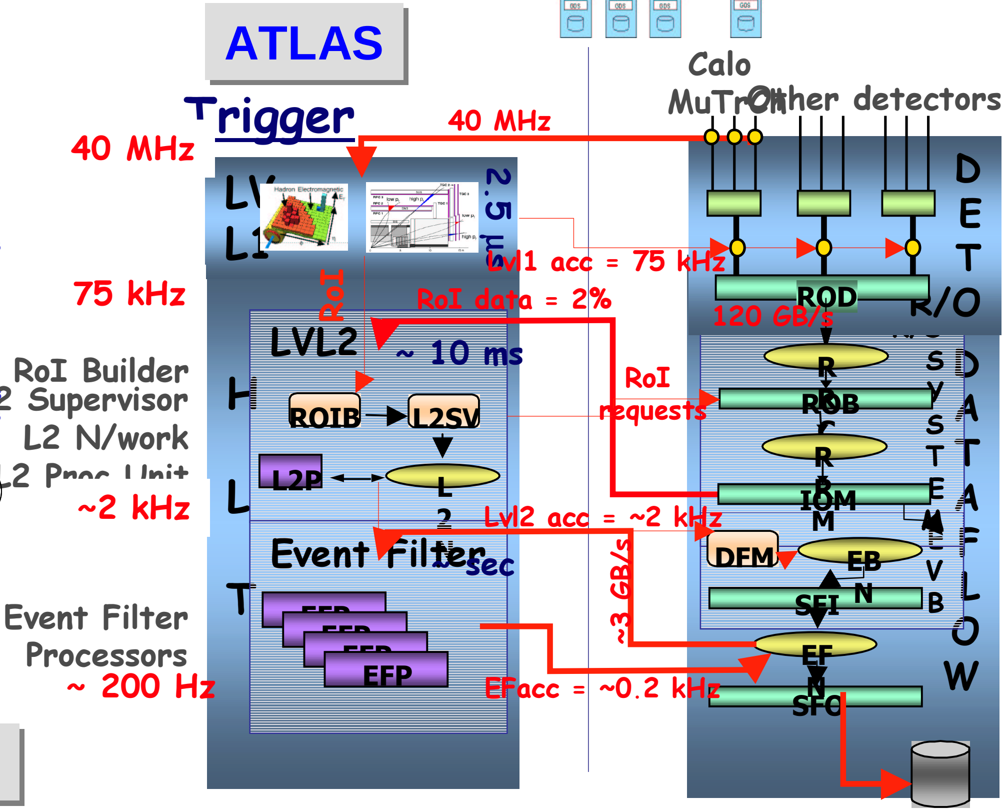
- Data to disk:
  - total data volume : ~ 140 Tbytes
  - 350 MBytes/s sustained (milestone 300 MBytes/s)

Event building and filter farms

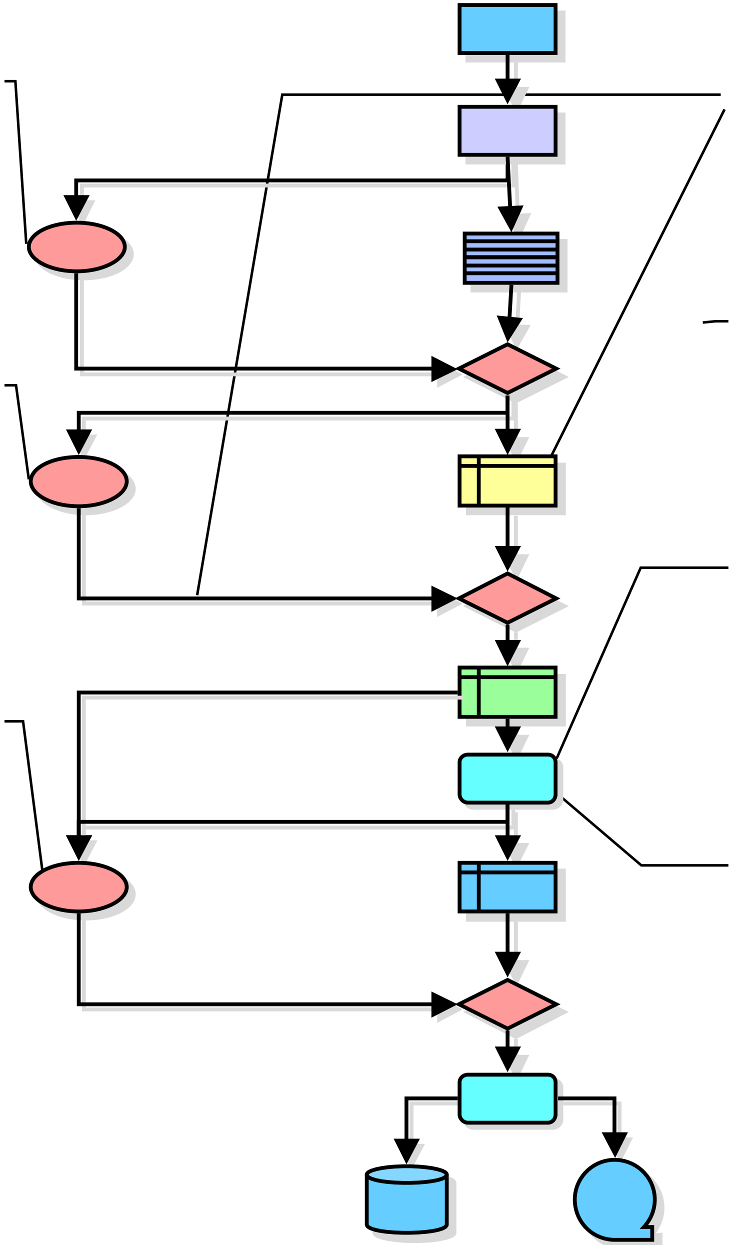








“R&D humanum est” (1)



- RD-27
  First-level trigger systems for LHC experiments.

- RD-11
  EAST Embedded architectures for second-level triggering in LHC experiments

- LCB_005
  Event Filter Farm

- RD-12
  Readout system test benches.

- RD-13
  A scalable data taking system at a test beam for LHC.

- RD-24
  Applications of the scalable coherent interface to data acquisition at LHC (SCI).

- RD-31
  NEBULAS: An asynchronous self-routing packet-switching network architecture for event building in high rate experiments (ATM).

Summary

- a large amount of ongoing activities in the area of offline and online computing to prepare for the LHC start
- quite confident in the current models and installations
- the next 2 years will require to try and test alternatives and to be confident for the purchasing exercise in 2006
- we have made very good progress in all areas, but we can by no means ‘relax’ !!

  expect still lots of surprises during the Data Challenges
- very tight and constructive collaboration between the Experiments and IT

Data Challenge Motto

“ seeing is believing “

“ you have shown scalability only when you have done the installation, not when you have predicted it ! ”

“ you will encounter any imaginable problem + the ones you have not even dreamed of ”

“ your invisible assistant is called Murphy ”