The LHC challenge

(the most demanding project ever of HEP)

UNIVERSITÀ
DEGLI STUDI
DI ROMA
LA SAPIENZA

Fernando Ferroni
Università di Roma "La Sapienza"
INFN Sezione di Roma

INFN
Istituto Nazionale
di Fisica Nucleare

Pisa GARR_05 "la rete daPERtutto" 11-05-2005

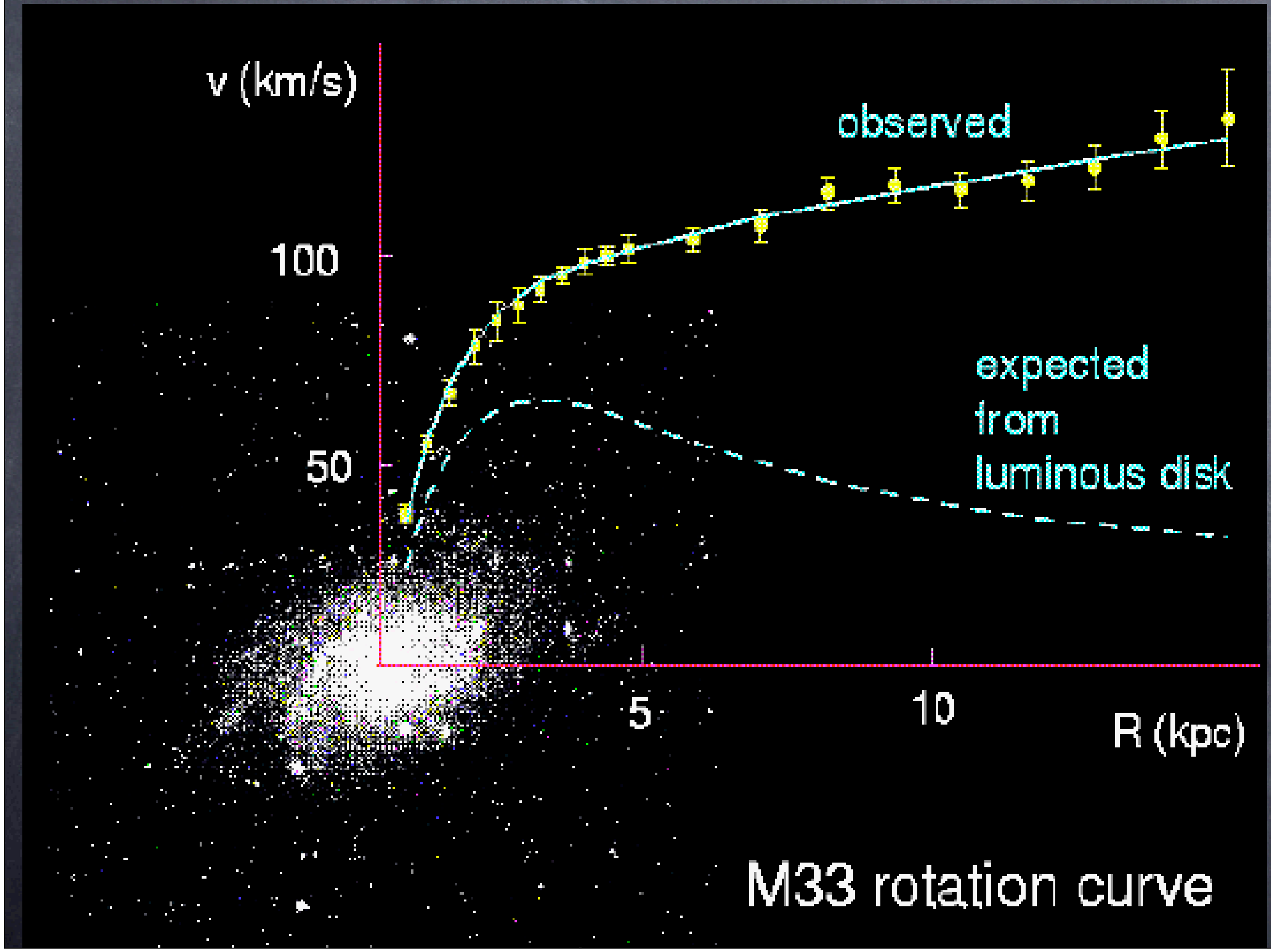
Outline

- The LHC project
- The LCG model
- The Experiment Data Flow
- Network Implication
- Conclusions

The 3 challenges for LHC

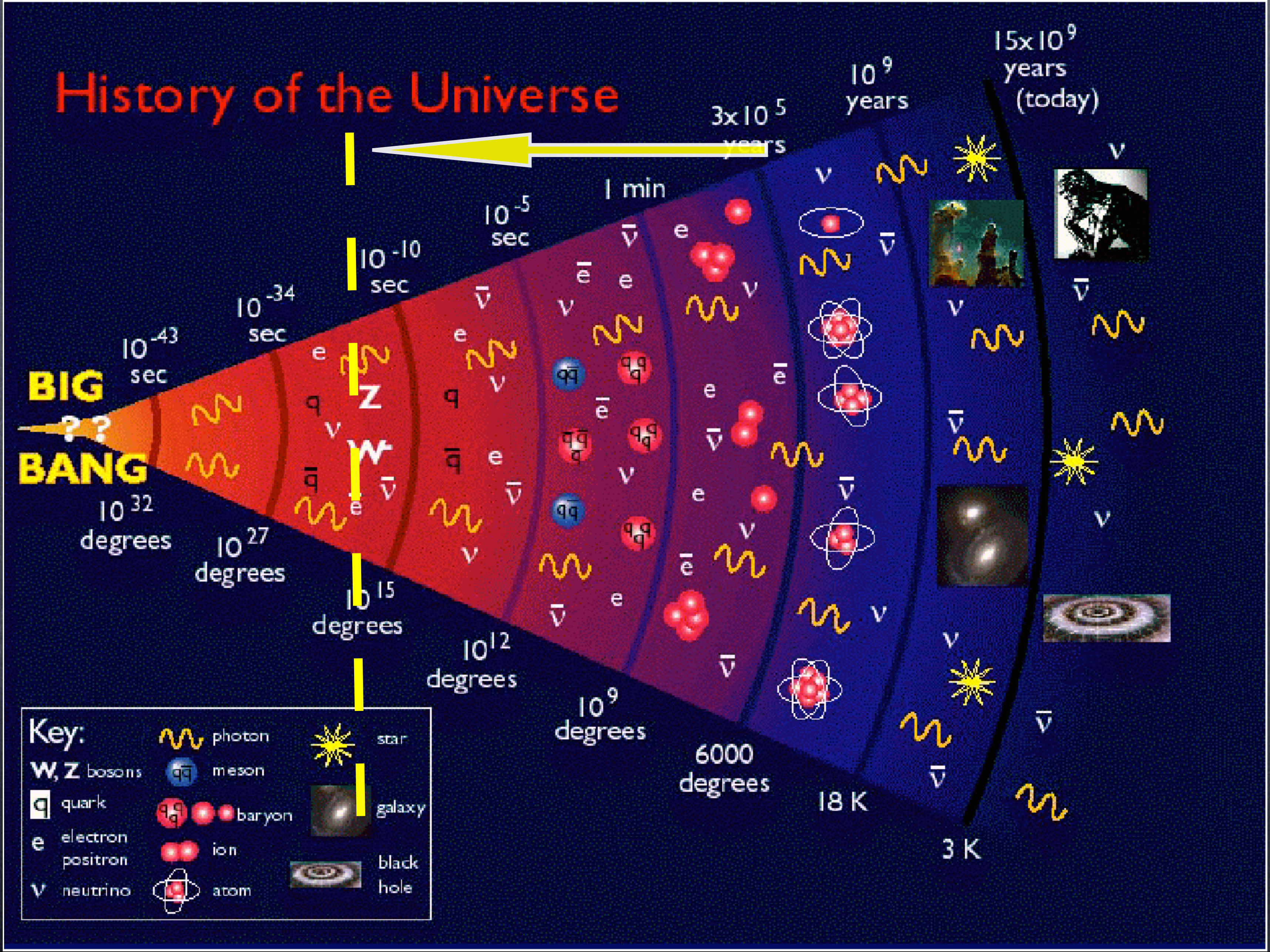
- Complete the Standard Model finding the Higgs boson
- Goes beyond the Standard Model and give an answer to the problem of Dark Matter
- Sail in the unknown land of the energy frontier

There is matter in the universe but it is not the same that makes our body



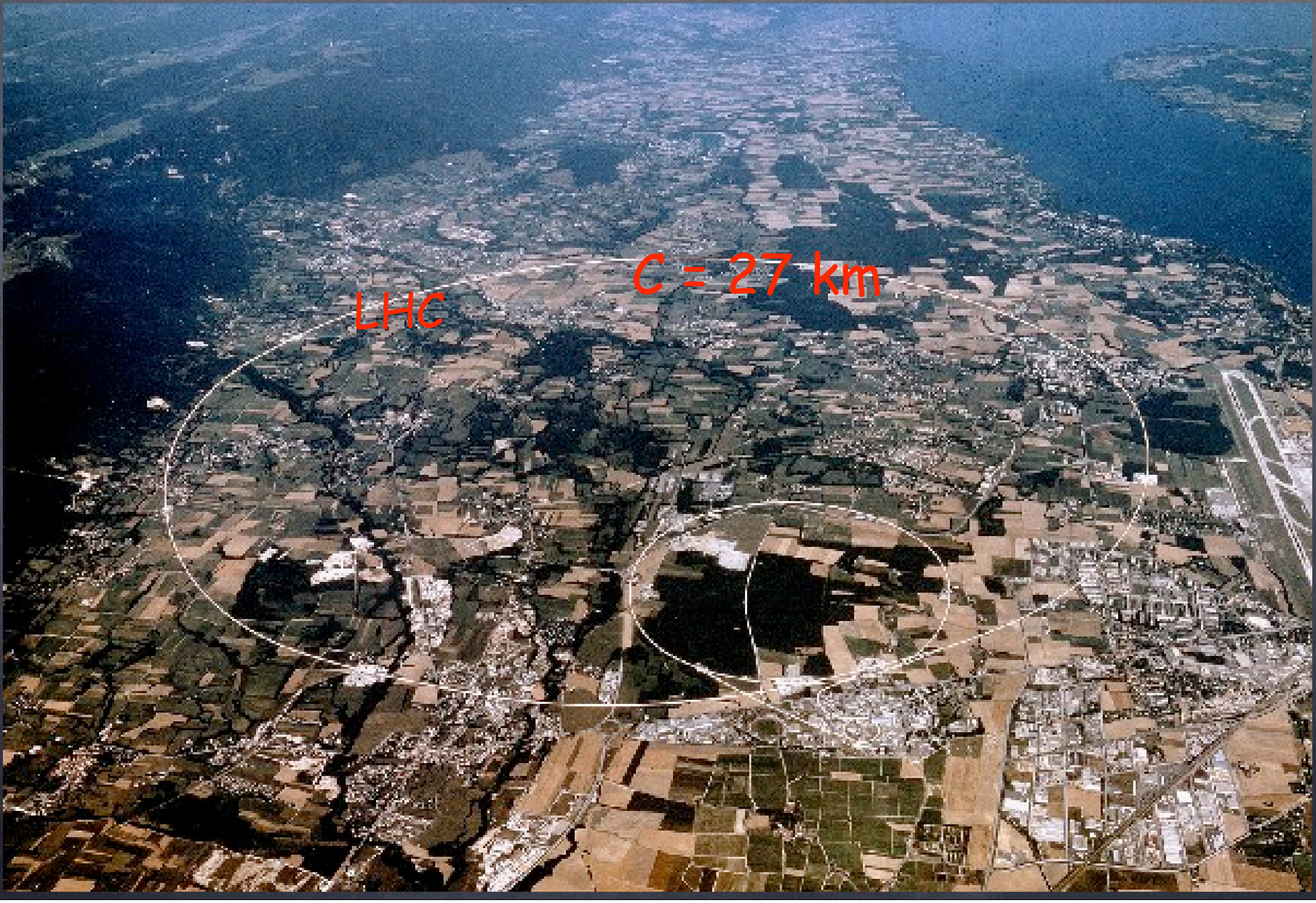
As you can see it is dark

The long trip back

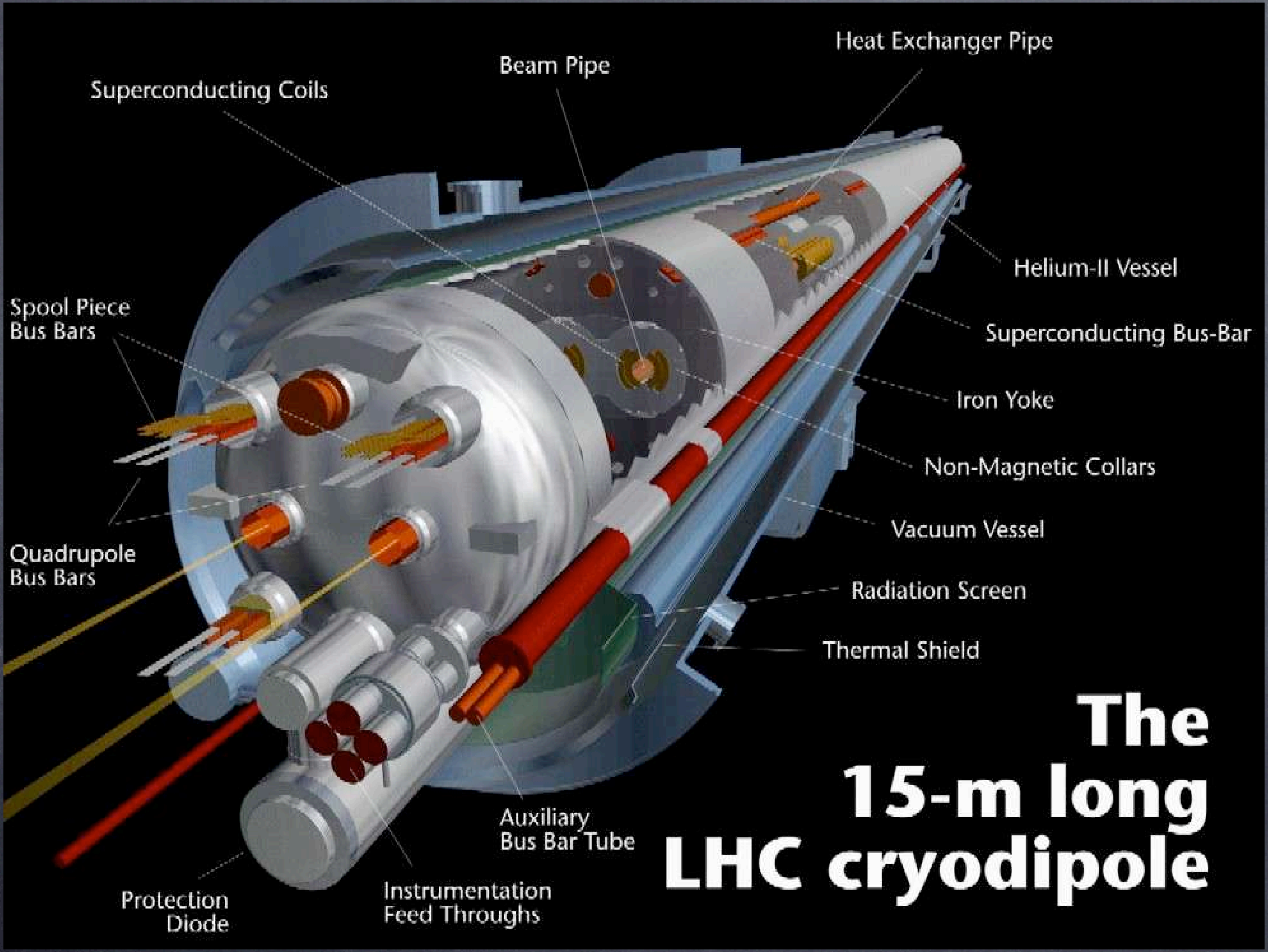


Large Hadron Collider

CERN

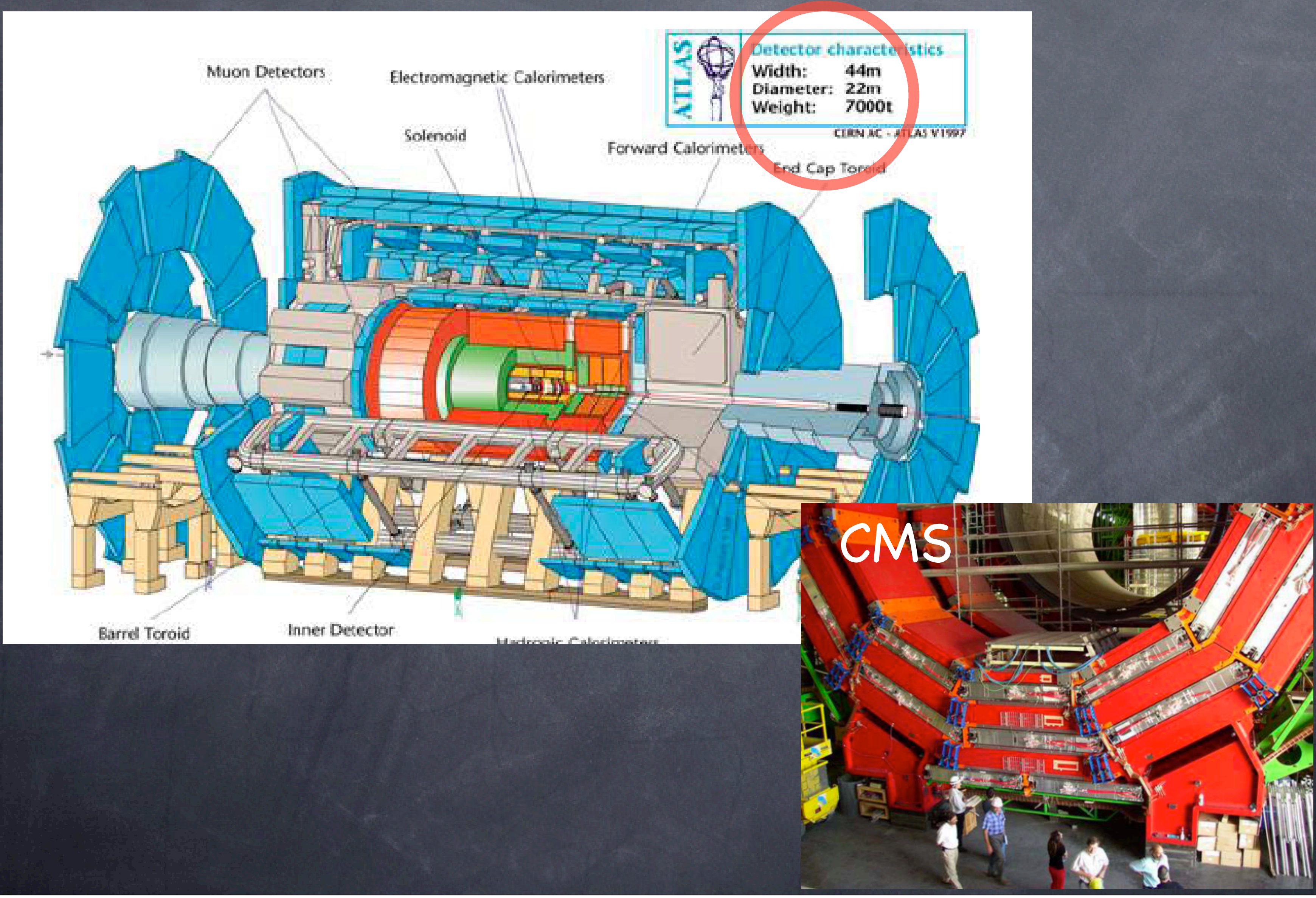


'Il grande freddo'



1232 dipoles
M= 24 ton
L=14 m
B=8.3 Tesla
T= 1.9 Kelvin

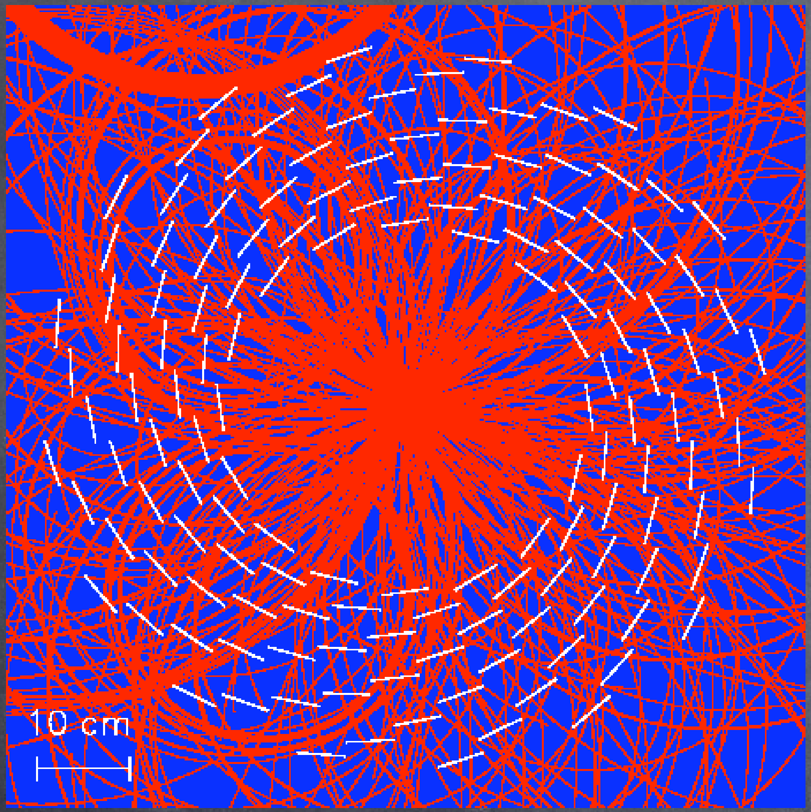
Detectors as big as cathedrals

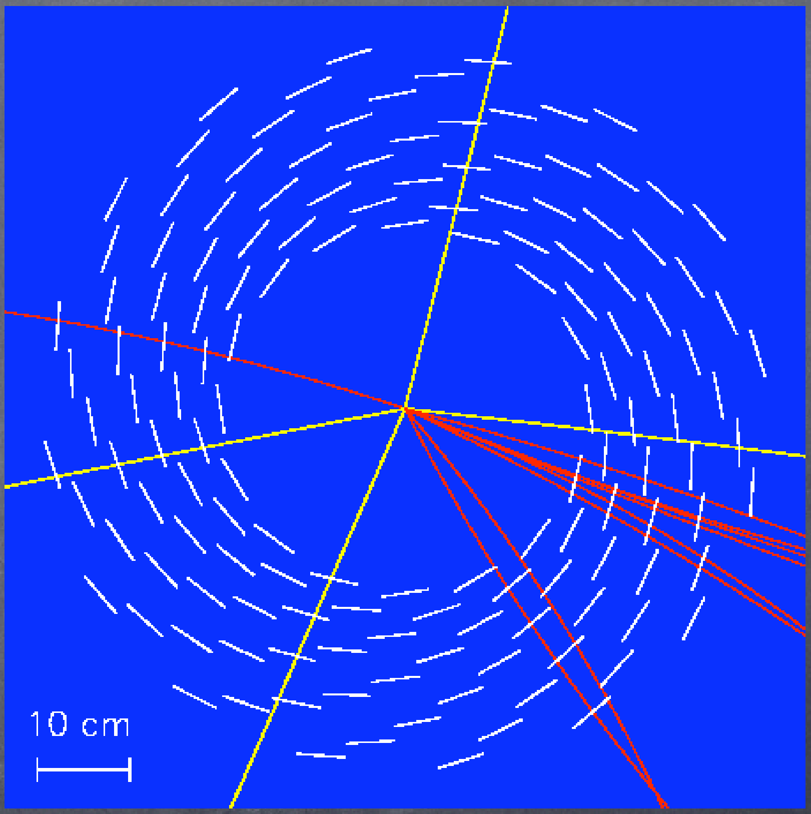


Impressive, isn't it ?

ATLAS Status: 10 December 2004

The (A)Intelligence task



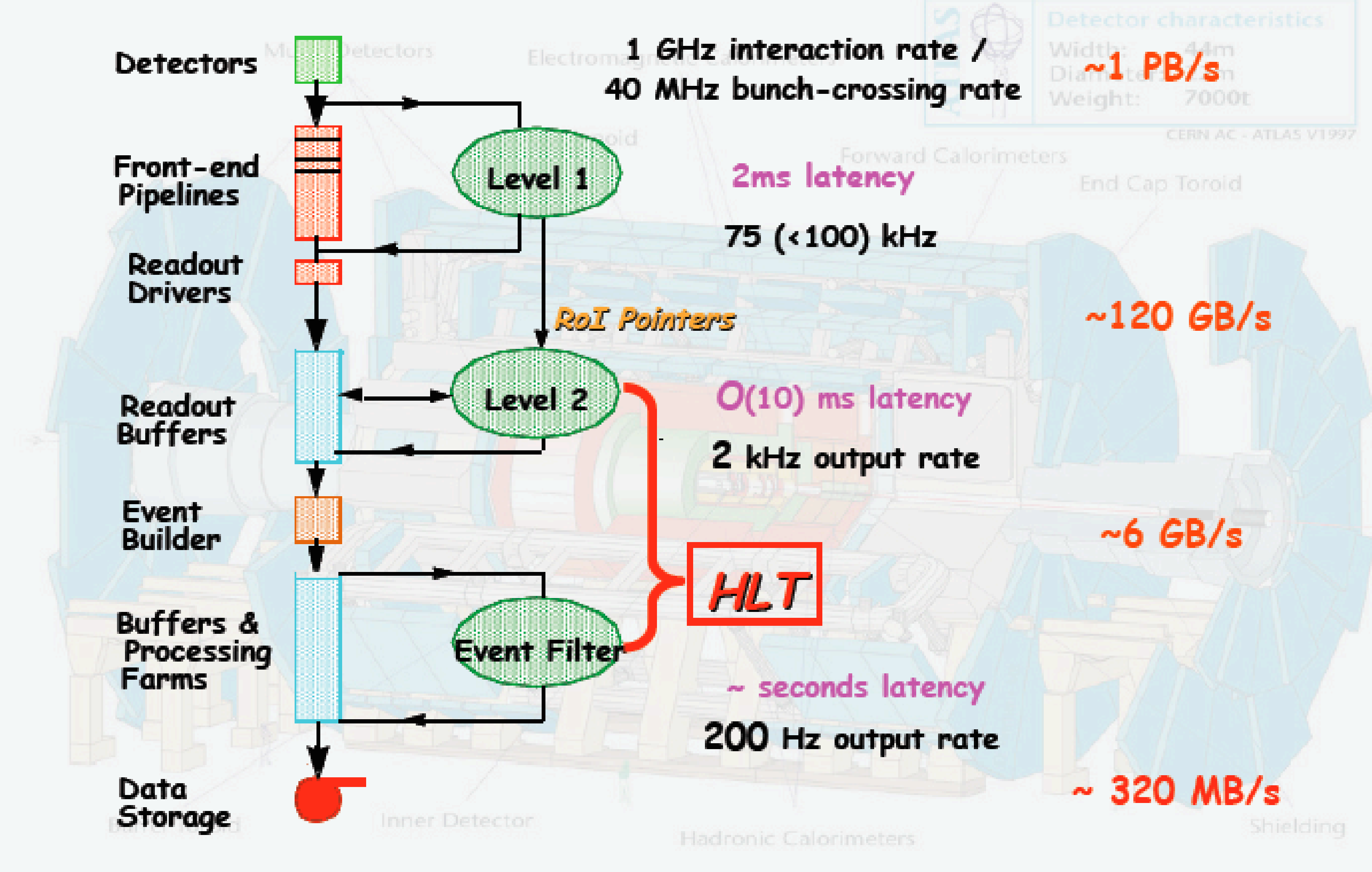


Find that bunch of interesting tracks amongst a couple of hundreds with a machine that fires every 25ns and perhaps produces what you are looking for once an hour

The computing challenge

- Record the interesting events after a dramatic choice (multi-layered trigger)
- Calibrate the detector (on/off line)
- Reconstruct the event (from Rare to Well Done)
- Create the streams for different physics channels and make them available for analysts
- Keep learning and reprocess for better quality
- MonteCarlo as much as you can

The trigger challenge



Detectors

Front-end Pipelines

Readout Drivers

Readout Buffers

Event Builder

Buffers & Processing Farms

Data Storage

Level 1

Level 2

Event Filter

RoI Pointers

HLT

1 GHz interaction rate /
40 MHz bunch-crossing rate

~1 PB/s

2ms latency

75 (<100) kHz

~120 GB/s

O(10) ms latency

2 kHz output rate

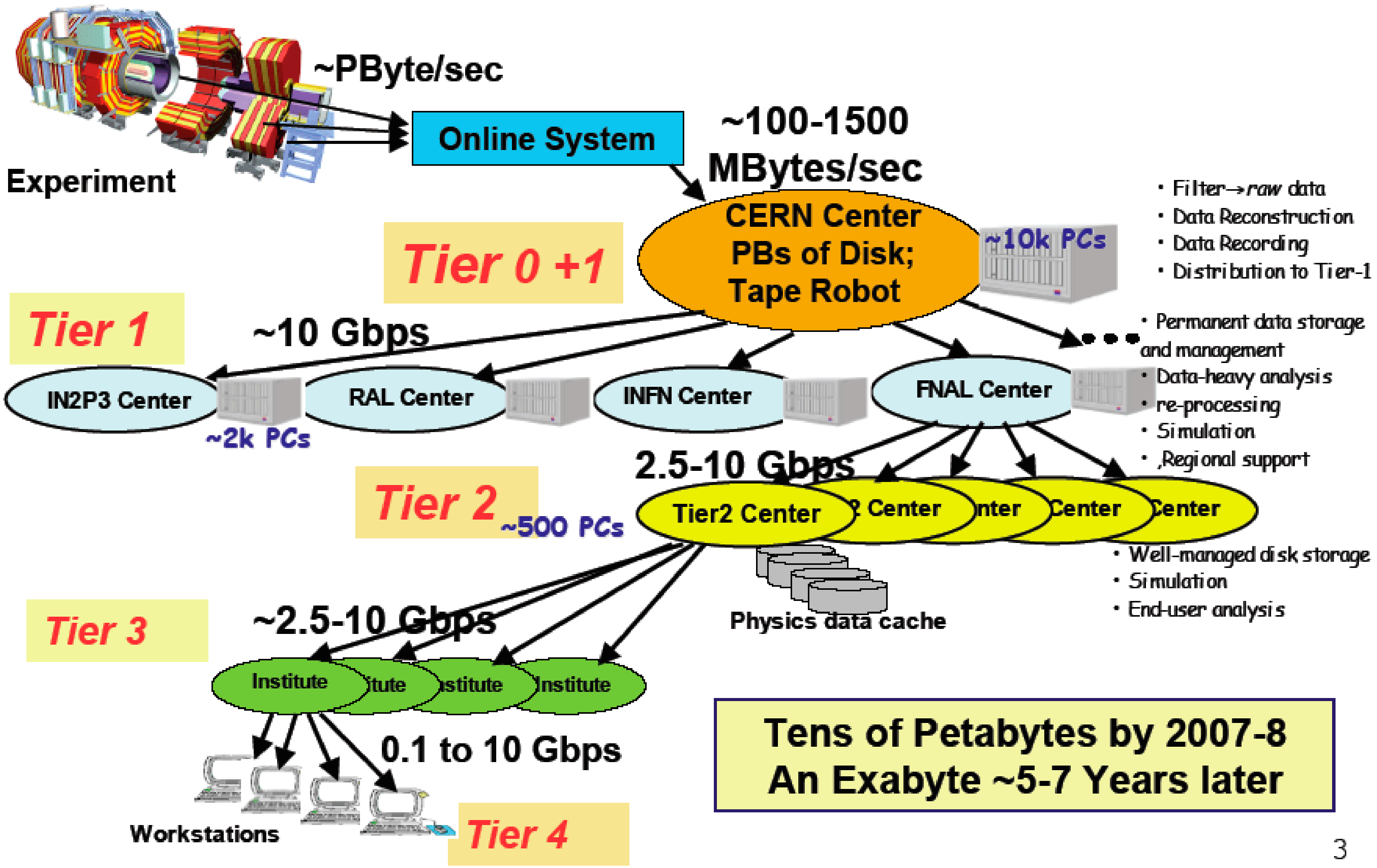
~6 GB/s

~ seconds latency

200 Hz output rate

~ 320 MB/s

The genesis of the Tiers



~PByte/sec

Experiment

Online System

~100-1500 MBytes/sec

Tier 0 +1

CERN Center
PBs of Disk;
Tape Robot

~10k PCs

- Filter→*raw* data
- Data Reconstruction
- Data Recording
- Distribution to Tier-1

Tier 1

~10 Gbps

IN2P3 Center

~2k PCs

RAL Center

INFN Center

FNAL Center

- Permanent data storage and management
- Data-heavy analysis
- re-processing
- Simulation
- ,Regional support

2.5-10 Gbps

Tier 2

~500 PCs

Tier2 Center

2 Center

nter

Center

Center

Physics data cache

- Well-managed disk storage
- Simulation
- End-user analysis

Tier 3

~2.5-10 Gbps

Institute

itute

nstitute

Institute

0.1 to 10 Gbps

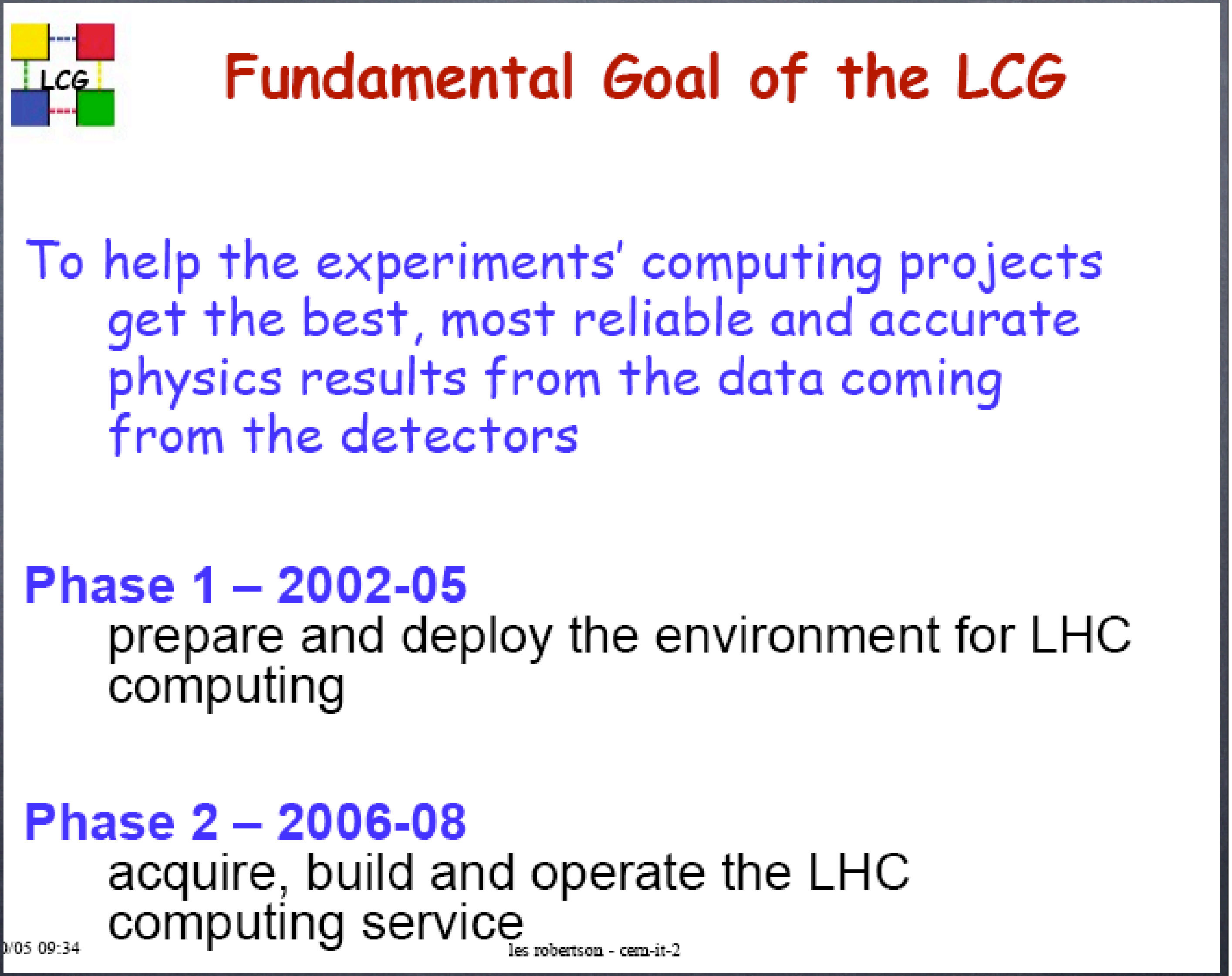
Workstations

Tier 4

Tens of Petabytes by 2007-8
An Exabyte ~5-7 Years later

CERN: where LCG was born

The driving force behind the establishment of LCG is the need for most of the funding agencies:
a) to profit of UE funding
b) to keep most the expenses at home
c) to form computing engineers at home



Fundamental Goal of the LCG

To help the experiments' computing projects get the best, most reliable and accurate physics results from the data coming from the detectors

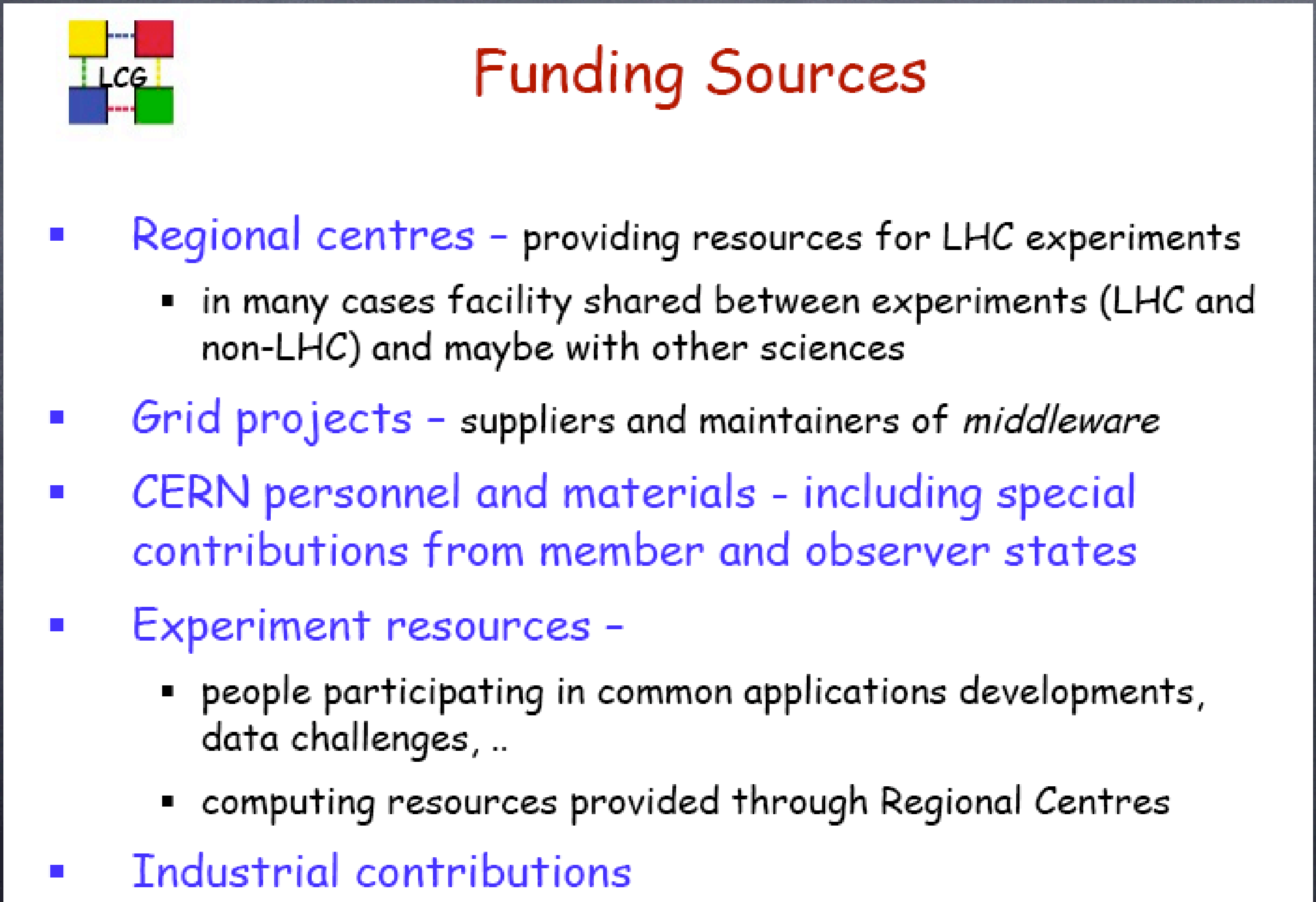
Phase 1 – 2002-05
prepare and deploy the environment for LHC computing

Phase 2 – 2006-08
acquire, build and operate the LHC computing service

/05 09:34

les robertson - cern-it-2

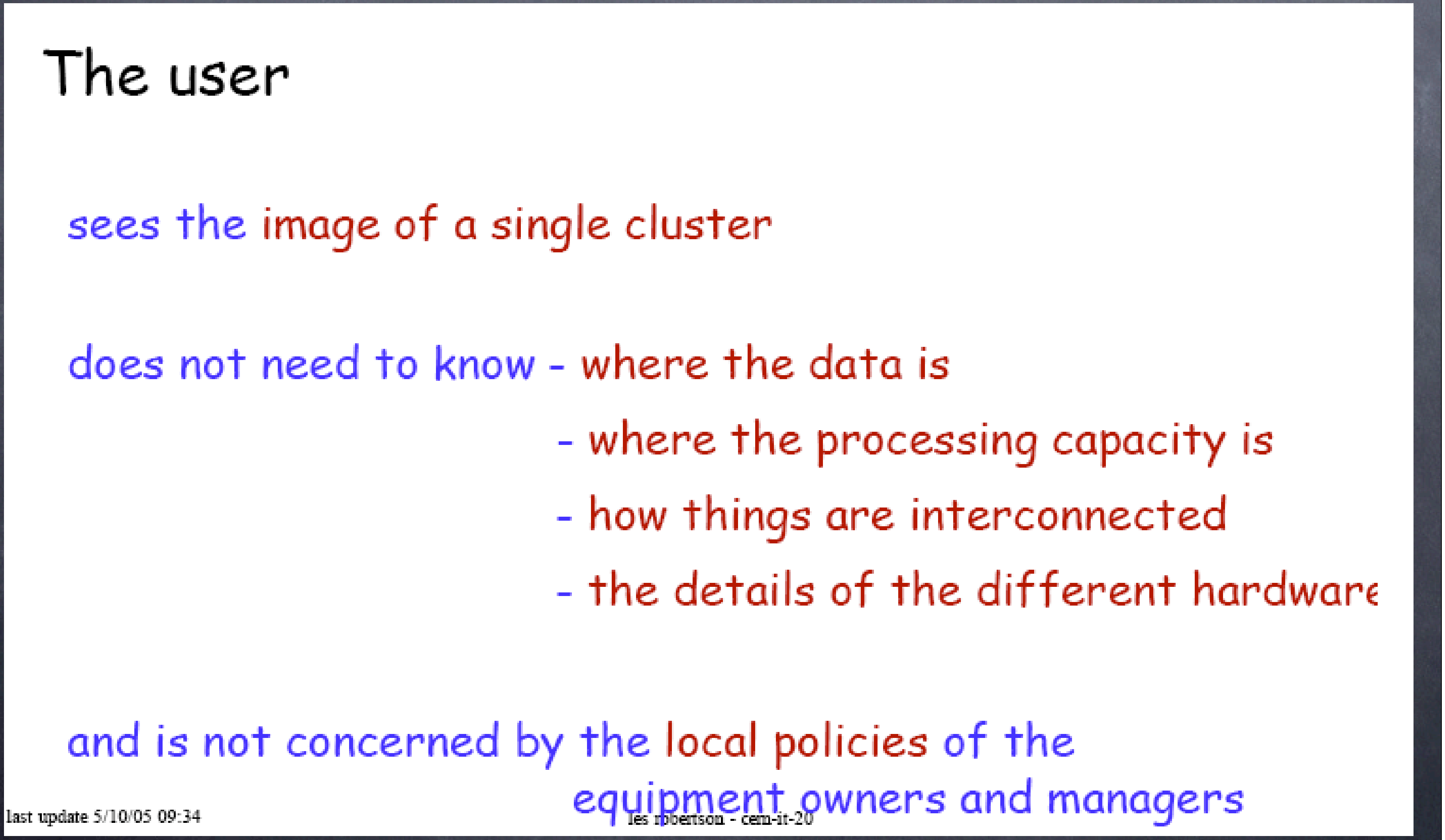
indeed a complex project (management-wise)



Funding Sources

- Regional centres – providing resources for LHC experiments
  - in many cases facility shared between experiments (LHC and non-LHC) and maybe with other sciences
- Grid projects – suppliers and maintainers of *middleware*
- CERN personnel and materials - including special contributions from member and observer states
- Experiment resources –
  - people participating in common applications developments, data challenges, ..
  - computing resources provided through Regional Centres
- Industrial contributions

The GRID in a nutshell



The user

sees the image of a single cluster

does not need to know - where the data is

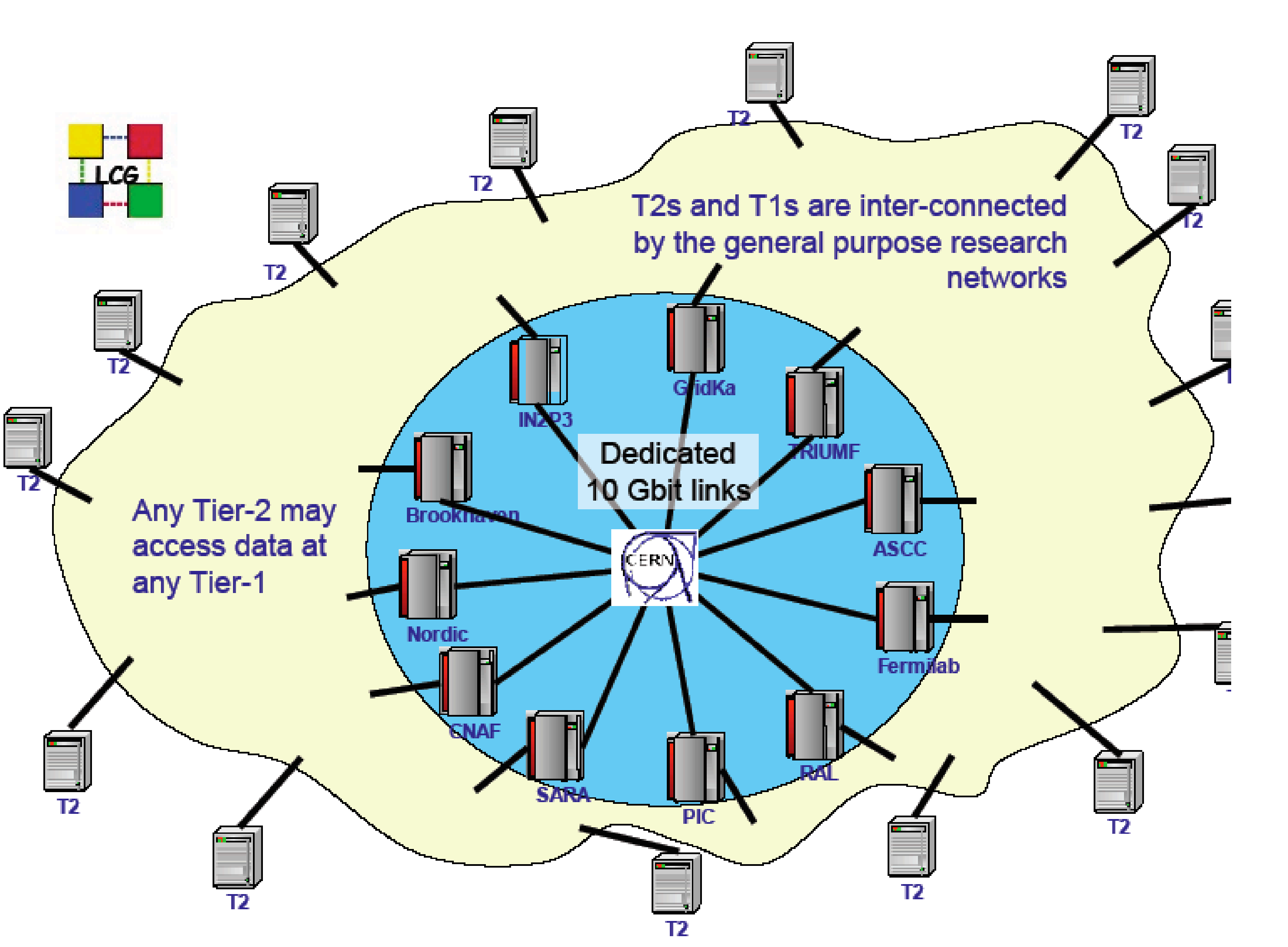
- where the processing capacity is
- how things are interconnected
- the details of the different hardware

and is not concerned by the local policies of the equipment owners and managers

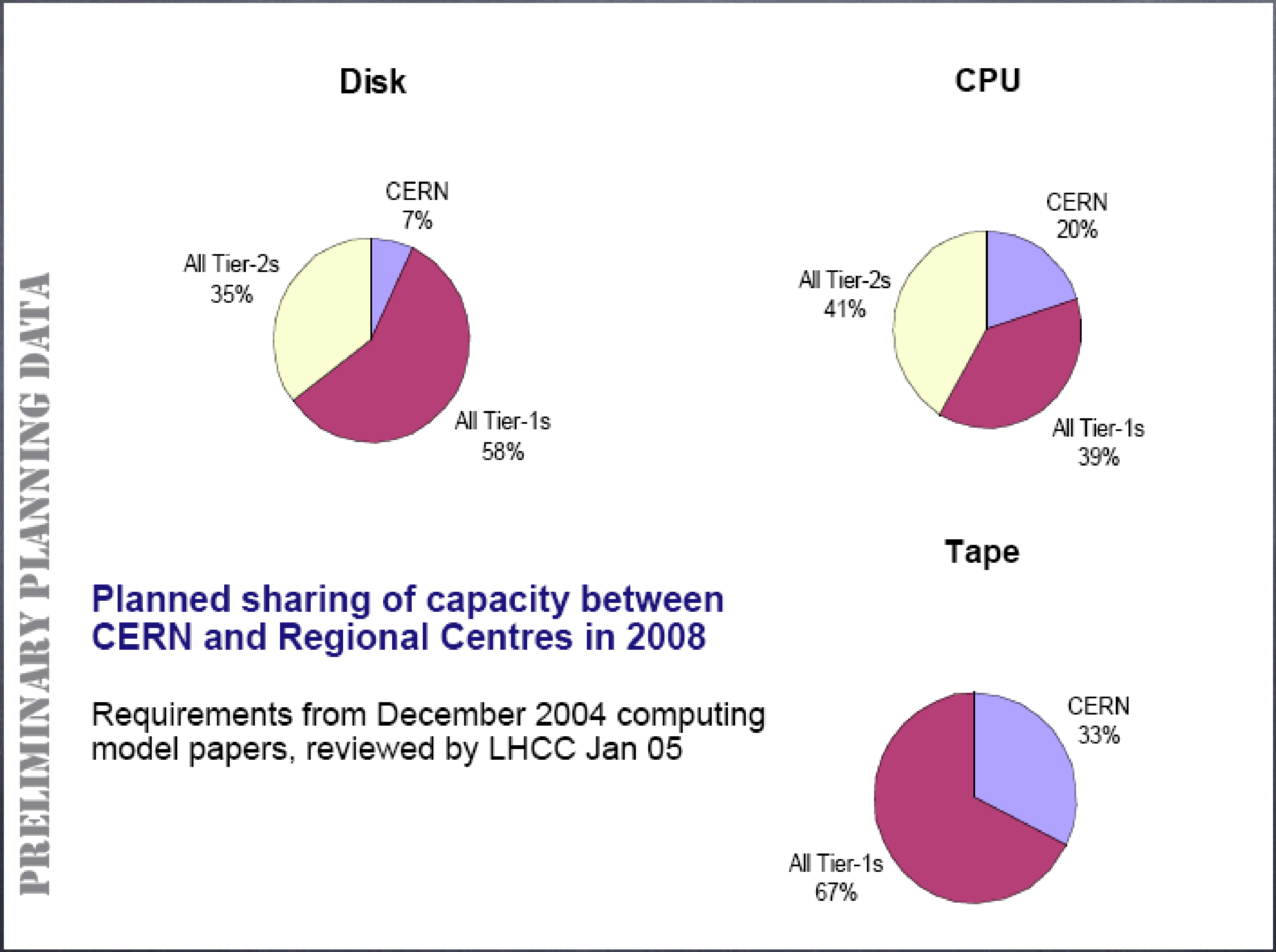
last update 5/10/05 09:34

les robertson - cern-it-20

The architecture of the LHC Grid Computing



The system is Copernican



PRELIMINARY PLANNING DATA

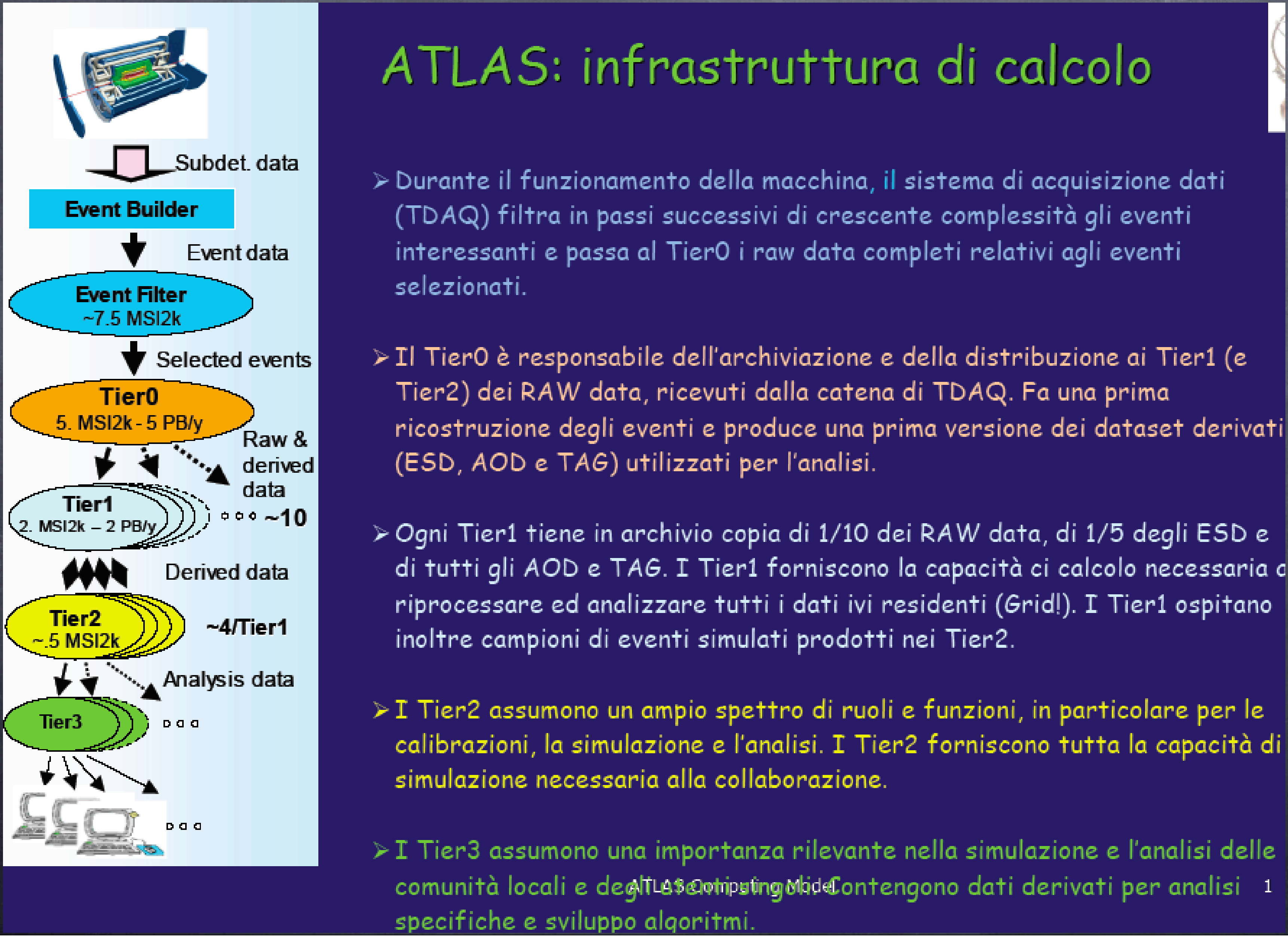
Planned sharing of capacity between CERN and Regional Centres in 2008

Requirements from December 2004 computing model papers, reviewed by LHCC Jan 05

Computing model

- Data are pre-allocated to a (some) given site (Tier)
  - And therefore moved there as soon as possible via Network
- Processing is mostly done where the data are
  - Choice of site is made via Grid-tools (Information System, Catalogs, Resource Broker, etc.)
- Data custodial, serving and processing is assigned to the different Tiers
  - Examples:
    - Re-processing is done at the Tier1s
    - Analysis is mostly done at the Tier2s (and Tier3s)
    - RECO (Reconstructed) and RAW data are distributed among the Tier1s
    - AOD (Analysis Object Data) and Skims are at all Tier1s and sub-samples at the Tier2s and Tier3s
    - Etc.
- User jobs are submitted via a LCG-UI (User Interface)
  - UIs are at all the Tier2s and Tier3s (maybe also at the Tier1s)

Le divisioni funzionali



ATLAS: infrastruttura di calcolo

- Durante il funzionamento della macchina, il sistema di acquisizione dati (TDAQ) filtra in passi successivi di crescente complessità gli eventi interessanti e passa al Tier0 i raw data completi relativi agli eventi selezionati.

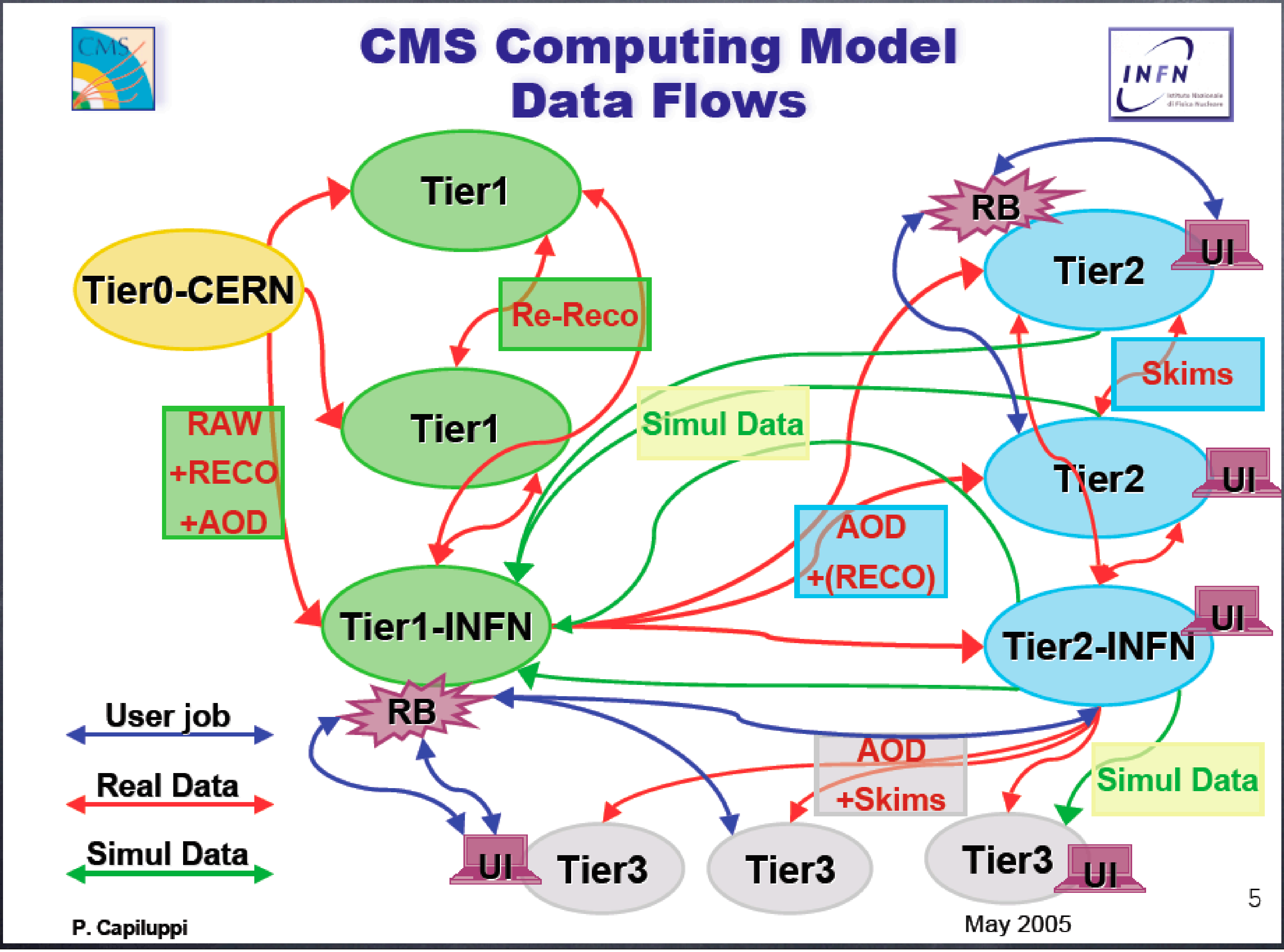
- Il Tier0 è responsabile dell'archiviazione e della distribuzione ai Tier1 (e Tier2) dei RAW data, ricevuti dalla catena di TDAQ. Fa una prima ricostruzione degli eventi e produce una prima versione dei dataset derivati (ESD, AOD e TAG) utilizzati per l'analisi.

- Ogni Tier1 tiene in archivio copia di 1/10 dei RAW data, di 1/5 degli ESD e di tutti gli AOD e TAG. I Tier1 forniscono la capacità ci calcolo necessaria a riprocessare ed analizzare tutti i dati ivi residenti (Grid!). I Tier1 ospitano inoltre campioni di eventi simulati prodotti nei Tier2.

- I Tier2 assumono un ampio spettro di ruoli e funzioni, in particolare per le calibrazioni, la simulazione e l'analisi. I Tier2 forniscono tutta la capacità di simulazione necessaria alla collaborazione.

- I Tier3 assumono una importanza rilevante nella simulazione e l'analisi delle comunità locali e degli utenti singoli. Contengono dati derivati per analisi specifiche e sviluppo algoritmi.

Data flow



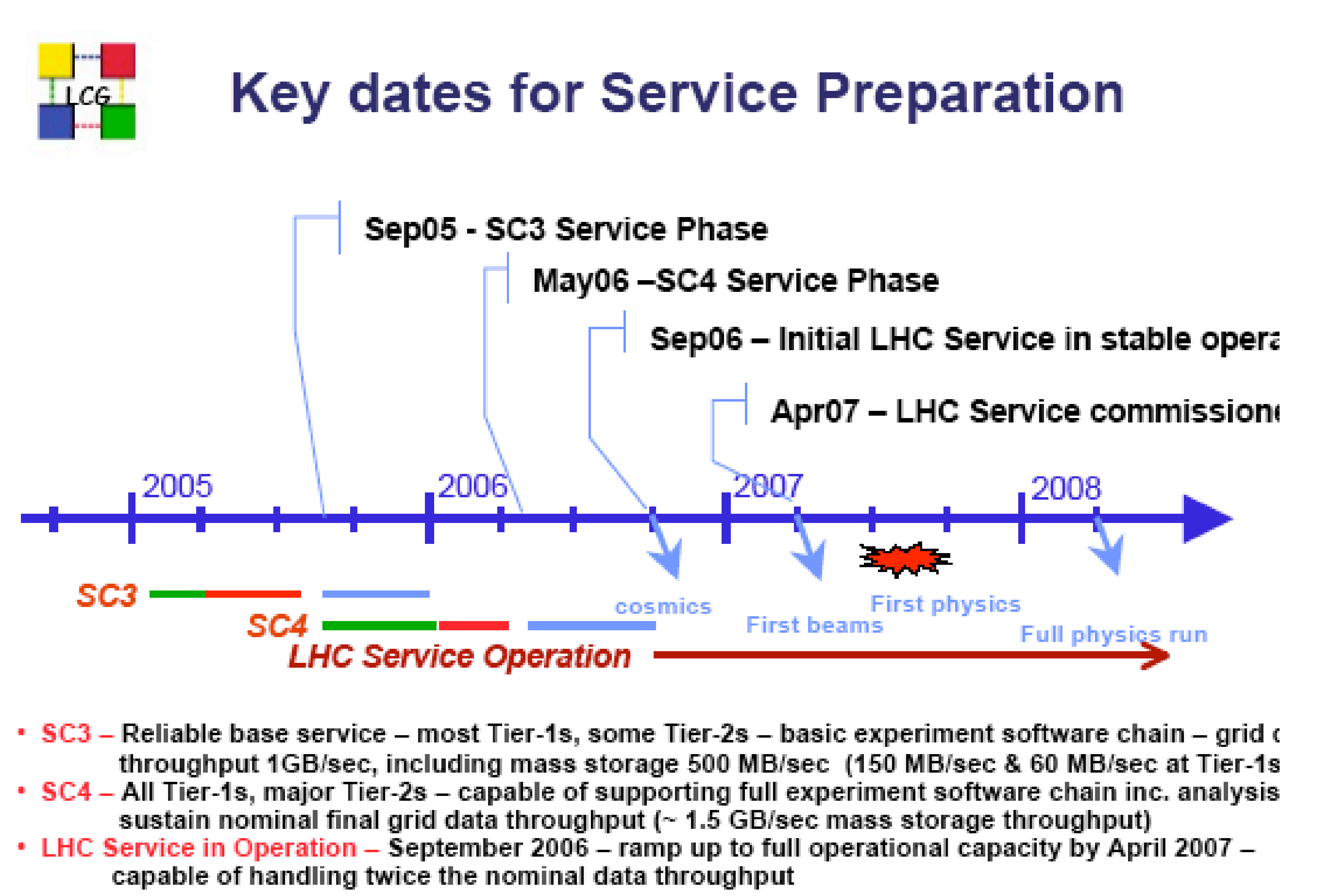
CMS Computing Model
Data Flows

P. Capiluppi

May 2005

5

The time schedule



Key dates for Service Preparation

Sep05 - SC3 Service Phase

May06 –SC4 Service Phase

Sep06 – Initial LHC Service in stable opera

Apr07 – LHC Service commission

2005 2006 2007 2008

SC3

SC4

LHC Service Operation

cosmics

First beams

First physics

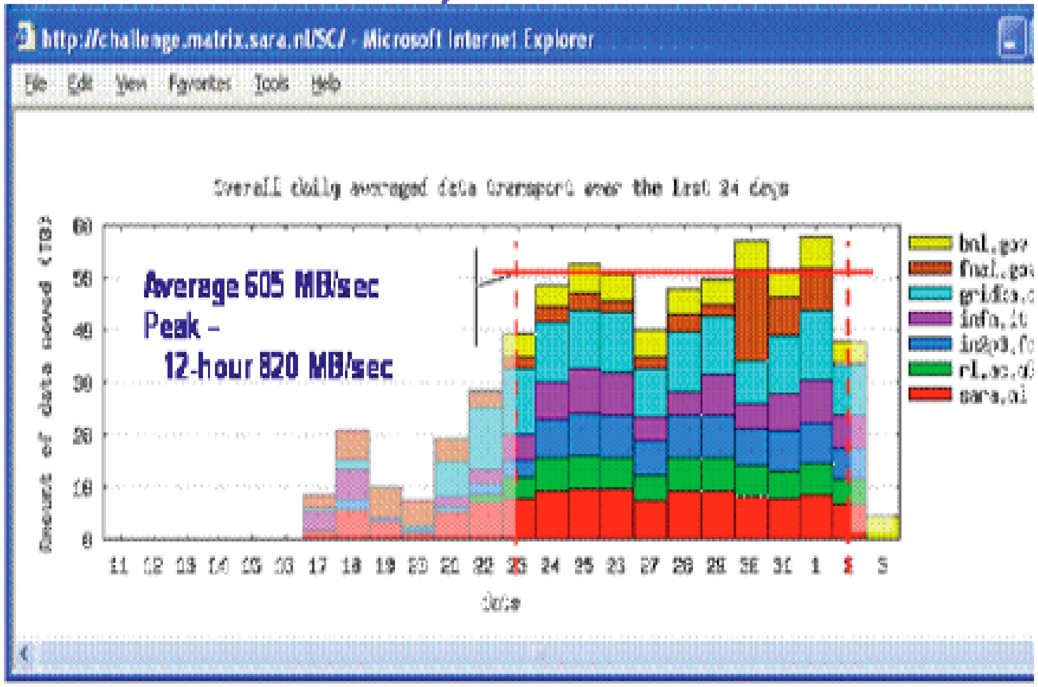
Full physics run

- SC3 – Reliable base service – most Tier-1s, some Tier-2s – basic experiment software chain – grid throughput 1GB/sec, including mass storage 500 MB/sec (150 MB/sec & 60 MB/sec at Tier-1s
- SC4 – All Tier-1s, major Tier-2s – capable of supporting full experiment software chain inc. analysis sustain nominal final grid data throughput (~ 1.5 GB/sec mass storage throughput)
- LHC Service in Operation – September 2006 – ramp up to full operational capacity by April 2007 – capable of handling twice the nominal data throughput

Testing the network ahead

Service Challenge 2

- Data distribution from CERN to Tier-1 sites
- Original target - sustain daily average of 500 MByte/sec from CERN to at least 5 Tier-1 sites for one week by the end of April
- Target raised to include 7 sites and run for 10 days

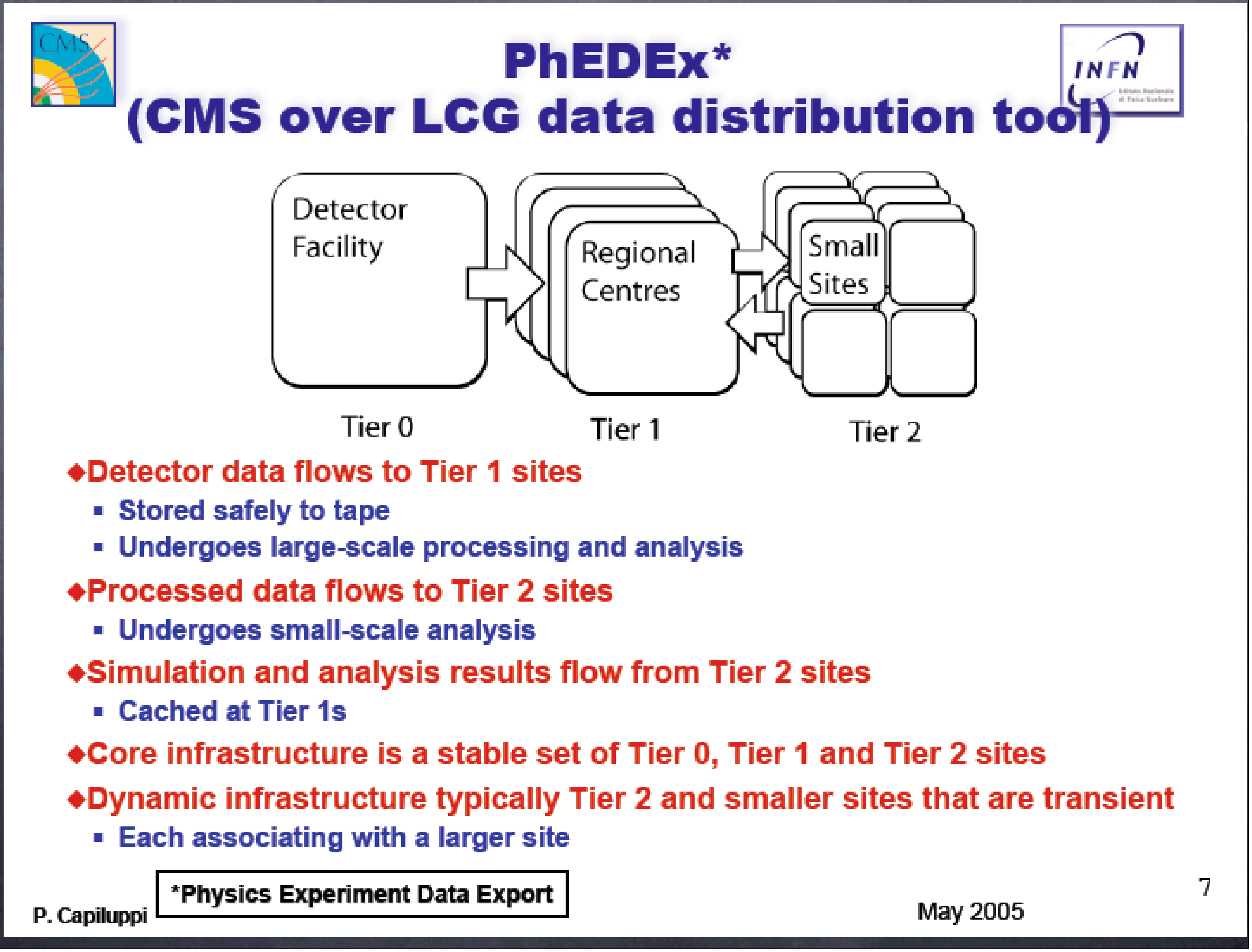

http://challenge.matrix.sara.nl/SC/ - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Overall daily averaged data transport over the last 24 days
Amount of data moved (TB)
0
10
20
30
40
50
60
11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 1 2 3
date
bnl.gov
fnal.go
gridka.
infn.it
in2p3.f
rl.ac.u
sara.nl

Average 605 MB/sec
Peak -
12-hour 820 MB/sec

- BNL, CCIN2P3, CNAF, FNAL, GridKa, RAL, NIKHEF/SARA
- Achieved on 2 April -
  -- average 600 MB/sec
  -- peak 820 MB/sec
- 500 MB/sec is 30% of the data distribution throughput required for LHC

tools for moving data around

is this a parameter of your game ?



PhEDEx*
(CMS over LCG data distribution tool)

- ◆Detector data flows to Tier 1 sites
  - Stored safely to tape
  - Undergoes large-scale processing and analysis
- ◆Processed data flows to Tier 2 sites
  - Undergoes small-scale analysis
- ◆Simulation and analysis results flow from Tier 2 sites
  - Cached at Tier 1s
- ◆Core infrastructure is a stable set of Tier 0, Tier 1 and Tier 2 sites
- ◆Dynamic infrastructure typically Tier 2 and smaller sites that are transient
  - Each associating with a larger site

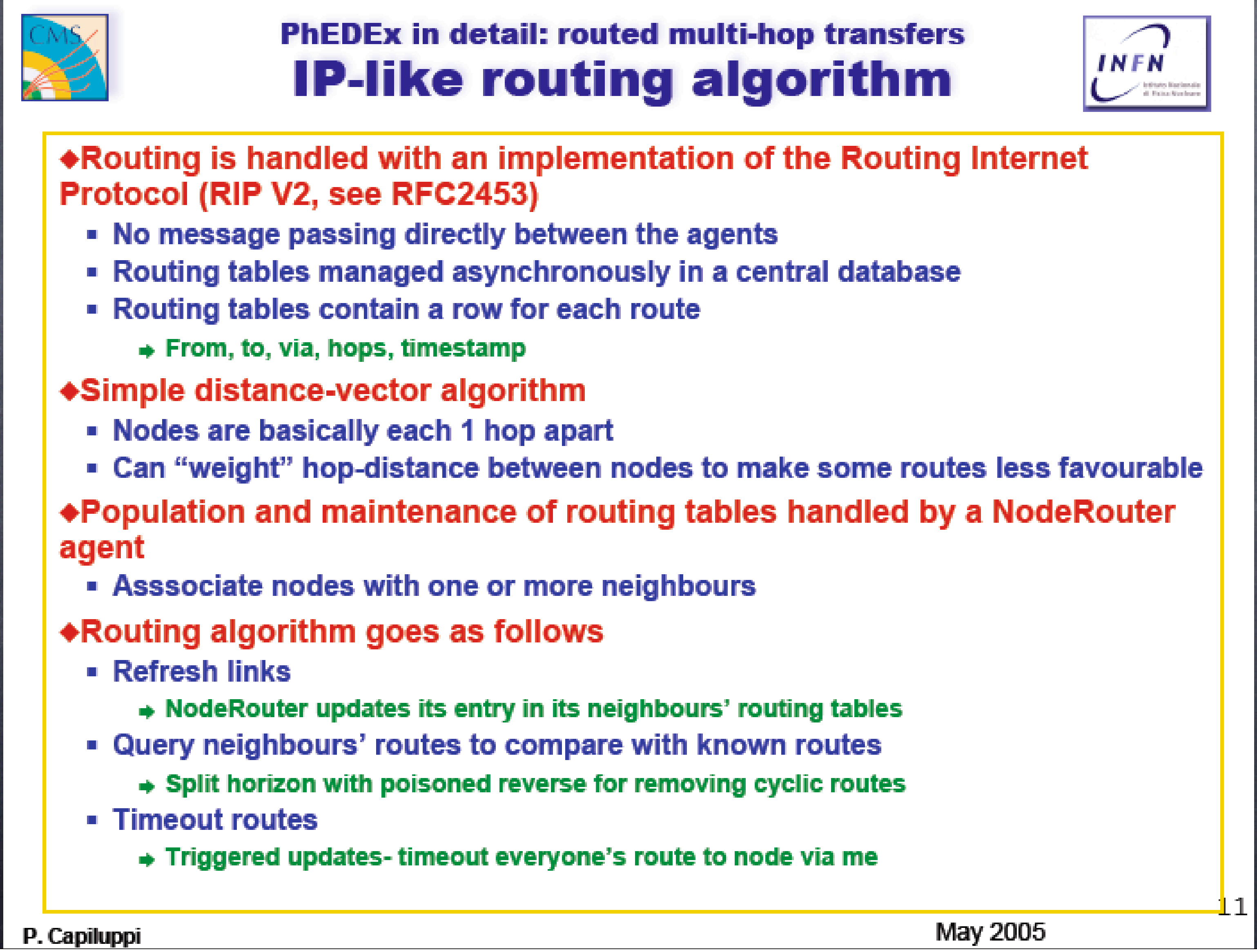
*Physics Experiment Data Export

P. Capiluppi

May 2005

7

for those really interested



PhEDEx in detail: routed multi-hop transfers

IP-like routing algorithm

- Routing is handled with an implementation of the Routing Internet Protocol (RIP V2, see RFC2453)
  - No message passing directly between the agents
  - Routing tables managed asynchronously in a central database
  - Routing tables contain a row for each route
    - From, to, via, hops, timestamp
- Simple distance-vector algorithm
  - Nodes are basically each 1 hop apart
  - Can “weight” hop-distance between nodes to make some routes less favourable
- Population and maintenance of routing tables handled by a NodeRouter agent
  - Asssociate nodes with one or more neighbours
- Routing algorithm goes as follows
  - Refresh links
    - NodeRouter updates its entry in its neighbours’ routing tables
  - Query neighbours’ routes to compare with known routes
    - Split horizon with poisoned reverse for removing cyclic routes
  - Timeout routes
    - Triggered updates- timeout everyone’s route to node via me

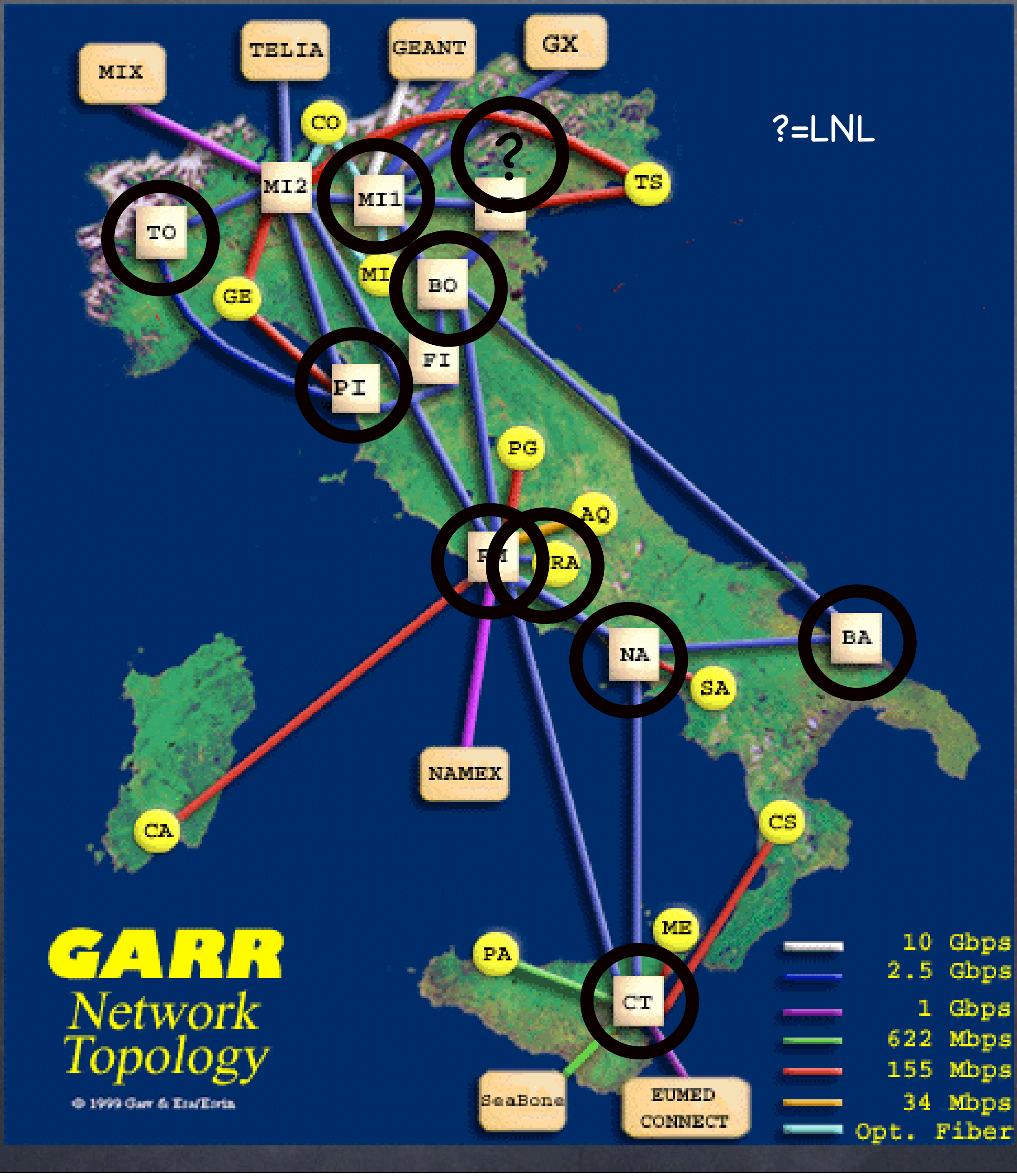
P. Capiluppi

May 2005

Uno sguardo in Italia

Indovina, indovinello dove e' il Tier2 piu' bello ?

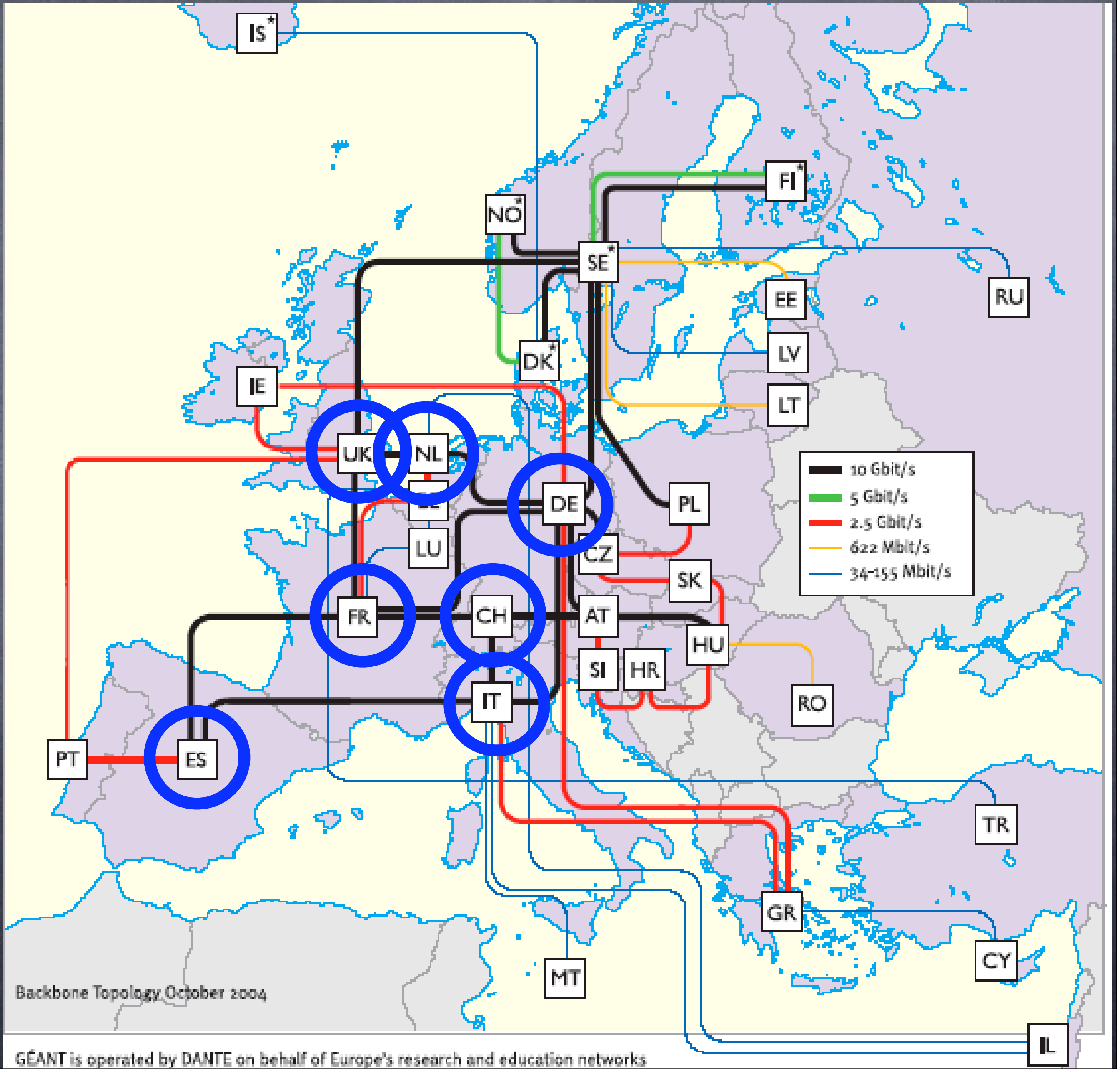
Posso immaginare link preferenziali CNAF-TO-CT per Alice, CNAF-LNL-PI-RM-BA per CMS, CNAF-RM-MI-NA-LNF per ATLAS ma non leggo il futuro e ci saranno evoluzioni. Se la rete e' flessibile non vedo problemi



?=LNL

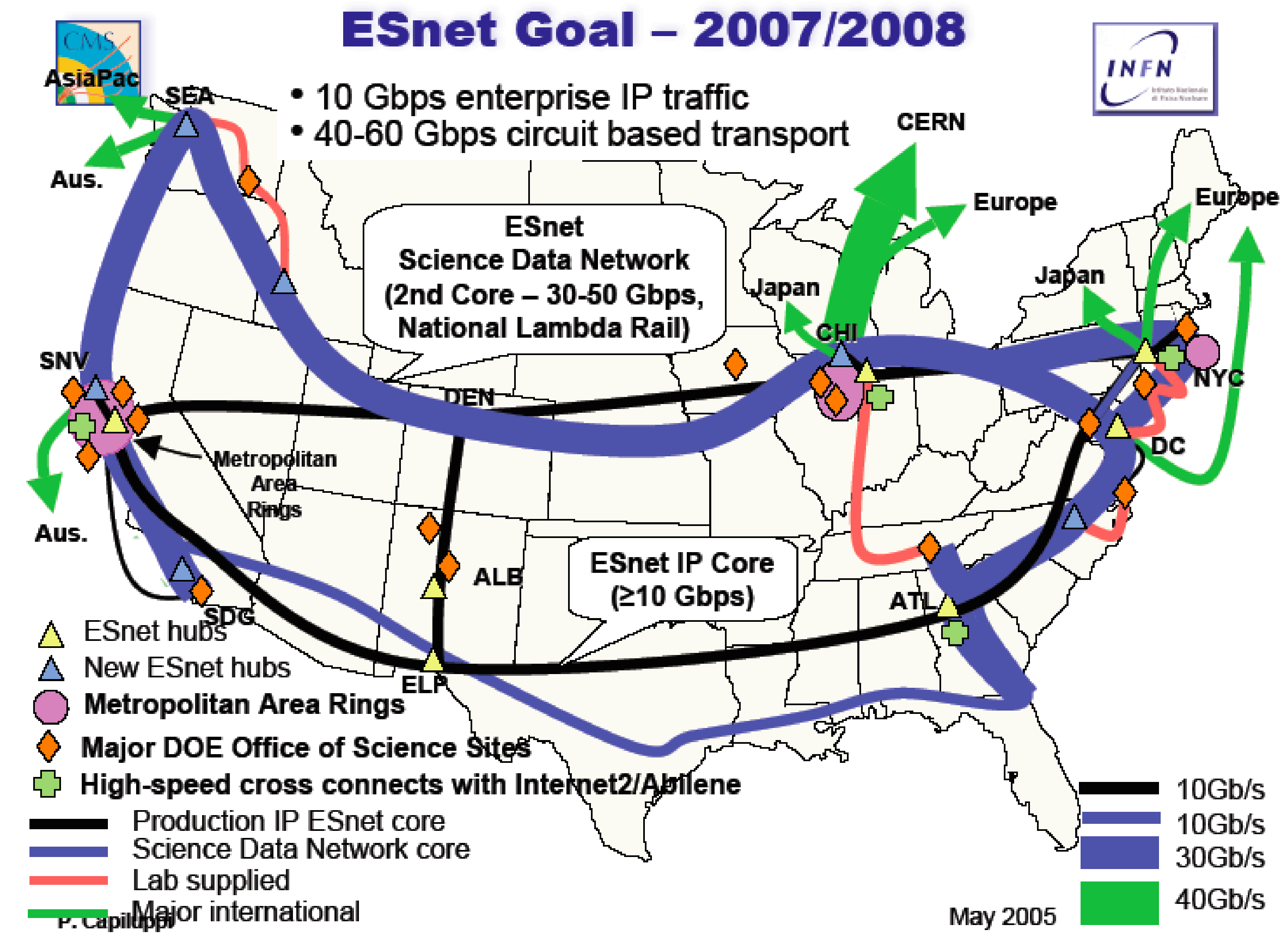
and another in Europe

My understanding is that those nodes shall never fail and most of the combination should be allowed



and further away

most of you
know this
much better
than me



ESnet Goal – 2007/2008

- 10 Gbps enterprise IP traffic
- 40-60 Gbps circuit based transport

ESnet
Science Data Network
(2nd Core – 30-50 Gbps,
National Lambda Rail)

ESnet IP Core
(≥10 Gbps)

Metropolitan
Area
Rings

△ ESnet hubs

▲ New ESnet hubs

● Metropolitan Area Rings

◆ Major DOE Office of Science Sites

✚ High-speed cross connects with Internet2/Abilene

━ Production IP ESnet core

━ Science Data Network core

━ Lab supplied

━ Major international

━ 10Gb/s

━ 10Gb/s

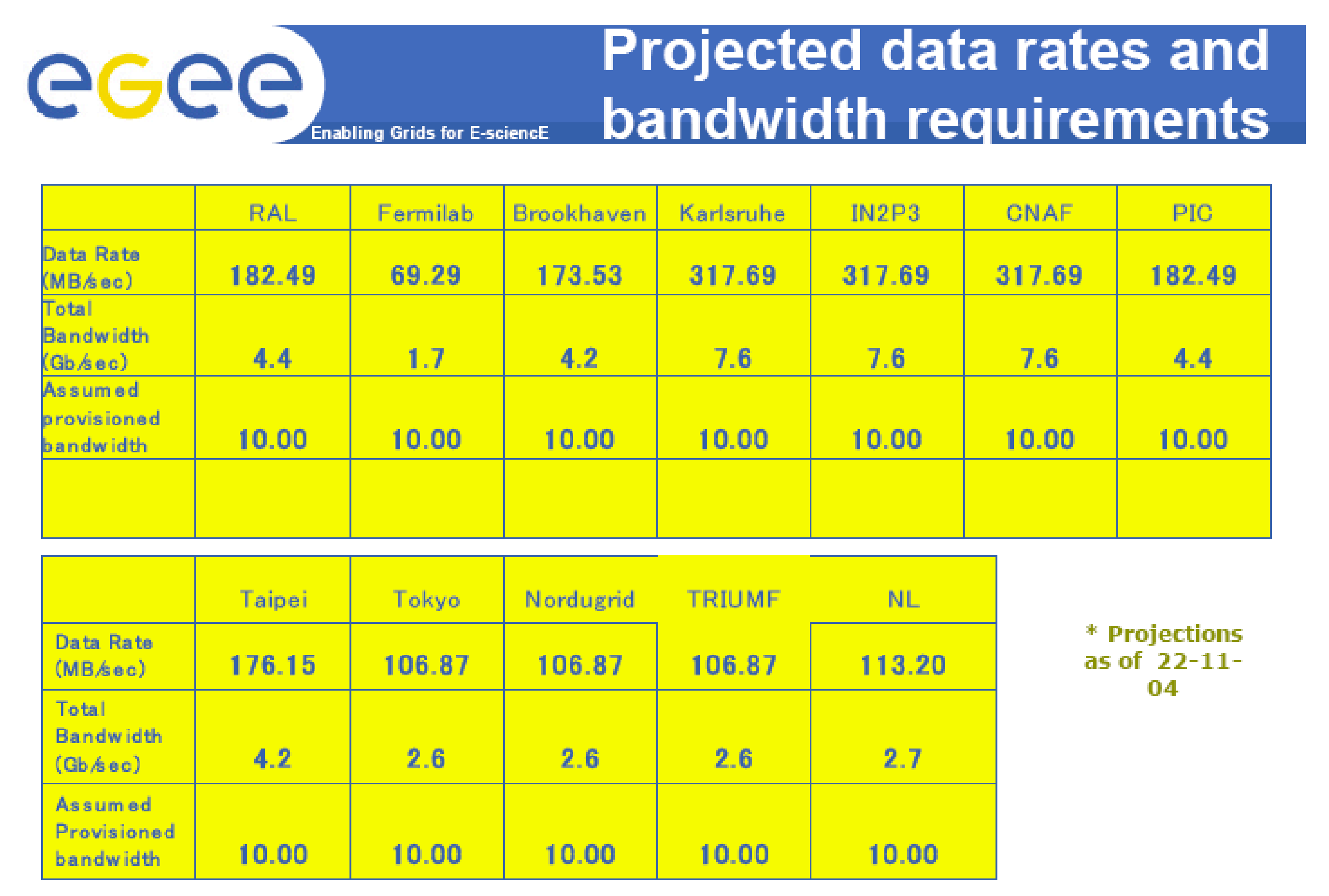
━ 30Gb/s

━ 40Gb/s

May 2005

P. Capiluppi

one point of view to summarize



Enabling Grids for E-sciencE

Projected data rates and bandwidth requirements

	RAL	Fermilab	Brookhaven	Karlsruhe	IN2P3	CNAF	PIC
Data Rate (MB/sec)	182.49	69.29	173.53	317.69	317.69	317.69	182.49
Total Bandwidth (Gb/sec)	4.4	1.7	4.2	7.6	7.6	7.6	4.4
Assumed provisioned bandwidth	10.00	10.00	10.00	10.00	10.00	10.00	10.00

	Taipei	Tokyo	Nordugrid	TRIUMF	NL
Data Rate (MB/sec)	176.15	106.87	106.87	106.87	113.20
Total Bandwidth (Gb/sec)	4.2	2.6	2.6	2.6	2.7
Assumed Provisioned bandwidth	10.00	10.00	10.00	10.00	10.00

* Projections as of 22-11-04

Conclusions

- Computing model of LHC experiments are getting close to reality
- Impact on network reasonably quantified
- Important details still missing
- Flexibility and evolution far to be accounted for (this is the worst news for network providers)
- Past experience imprinted on physicists the concept that network is transparent