

Chapter 6

Case Study: Big Science

1 Particle Physics

GridPP organisation was set in the UK up to handle the challenge of providing computing resources for the **L**arge **H**adron **C**ollider at its detectors.

The major detectors are **ATLAS**, **CMS** and **ALICE**.

The characteristics are enormous data rates sustained over many years.

A world-wide decentralised structure for funding and personnel.



First meeting May 2001

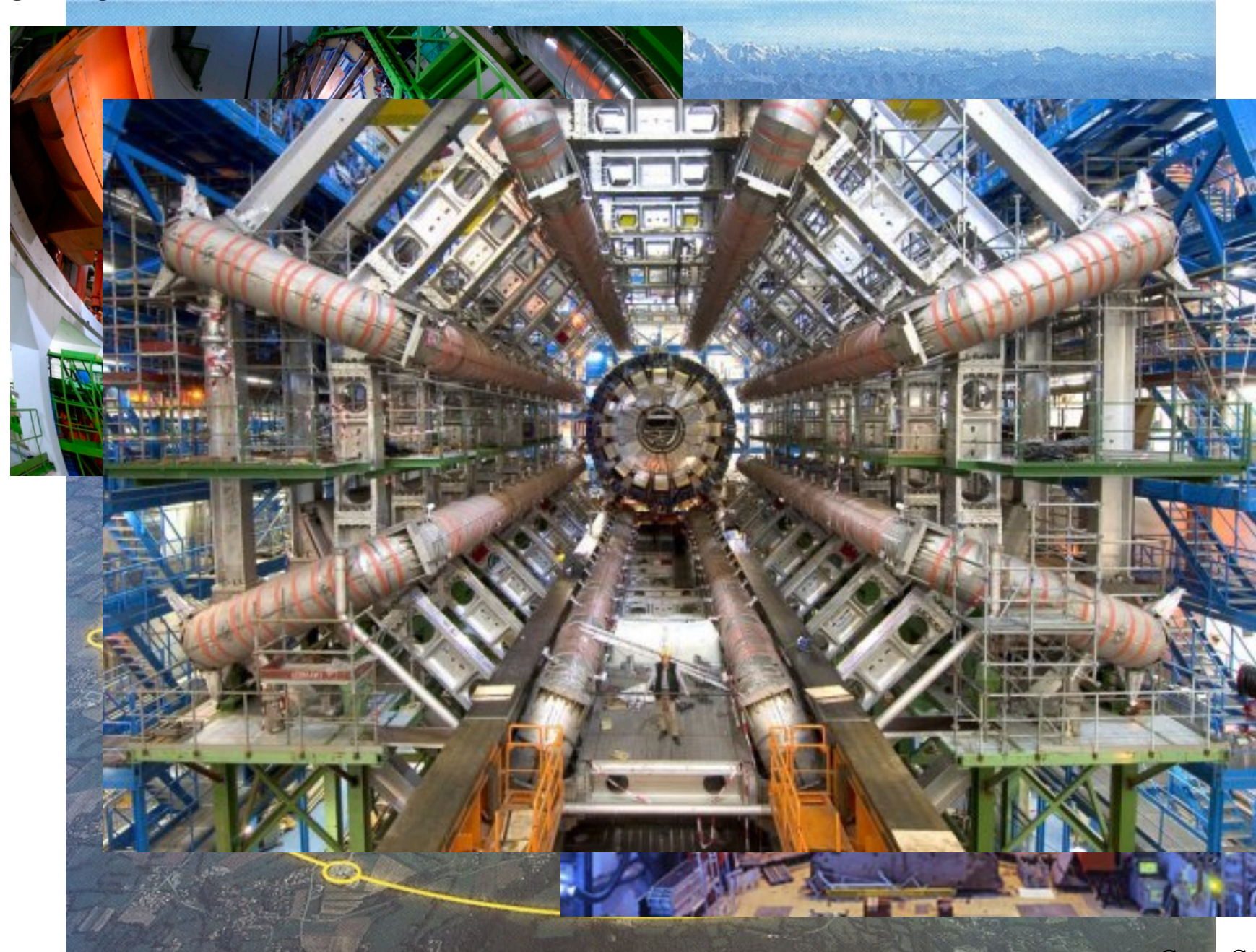
2 GridPP institutes

GridPP included every university with a particle physics group in the UK.

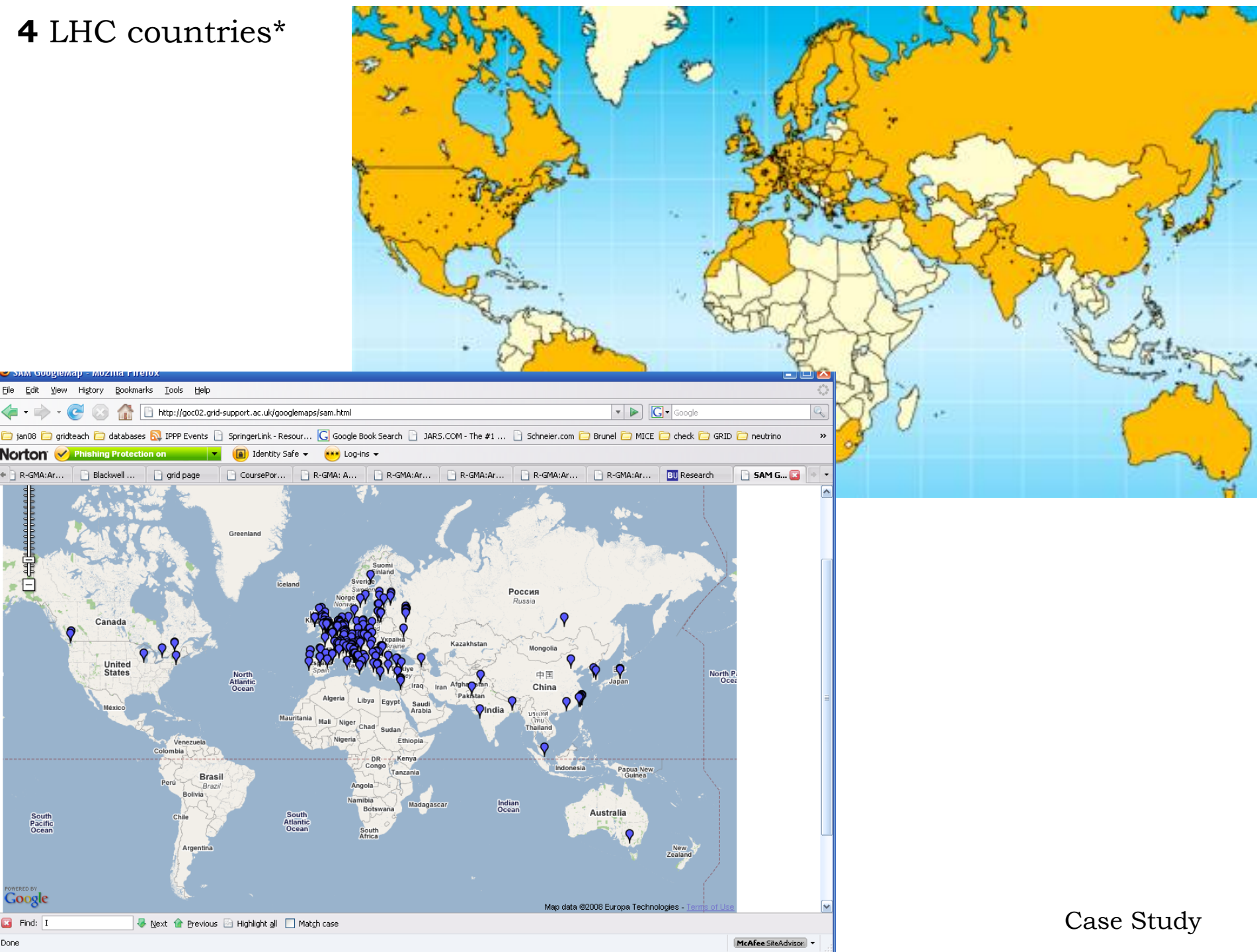


To users all over the world this looks very like a single resource.

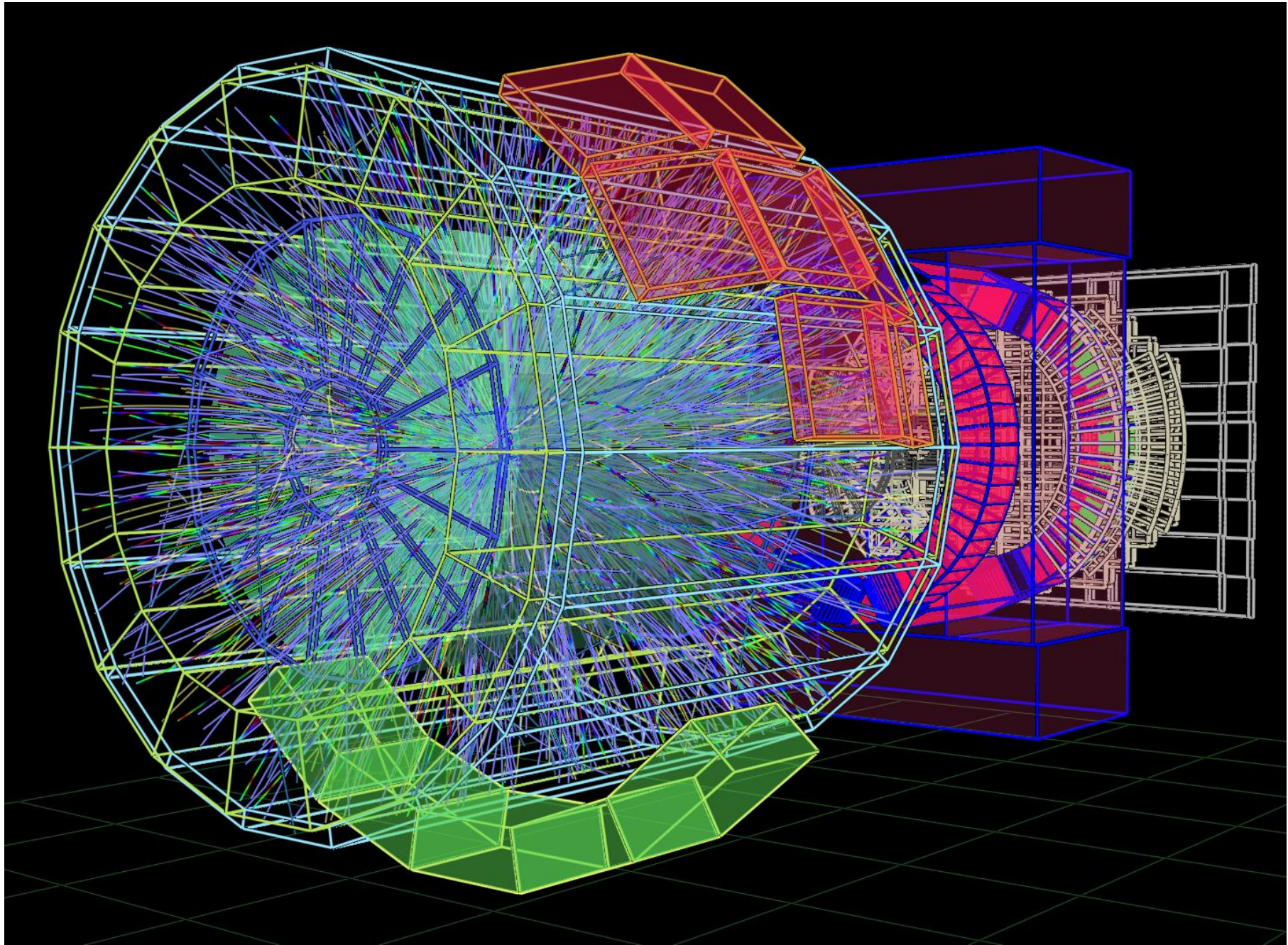
3 LHC*



4 LHC countries*



5 LHC event



6 LHC data rate

2,835 bunches of 10^{11} protons per bunch spaced by 25ns

Each experiment at LHC will see
800 million events per second.

We will run for 10 million seconds per year (4 months
24 hours a day)

$8,000,000,000,000,000 = 8 \cdot 10^{15}$ events/detector/year
The protons travel at 0.999997828 c

A total of $> 10^8$ electronic channels

LHC will produce $2 \cdot 10^{-4}$ $H \rightarrow \gamma\gamma$ per second

A $2 \cdot 10^{-4}$ needle in a $8 \cdot 10^8$ Haystack

Data is 40 TBytes per second from the detector
Special hardware reduces this to 10-100 GB/s
Embedded processors 1-10 GBytes/s
Processor Farm reduces to 100-200Mbytes/s

Total data to tape/disk is $(1-8) \cdot 10^{15}$ bytes/year.

How to analyse all this information?

Even if you use the Amazon cloud ... problems remain.
How to follow the progress of 100,000 jobs per day

A needle in a haystack is
easy

eine Nadel in einem
Heuhaufen finden

Case Study

Higgs rate

7 Data transfer rates

Sustained programme by CERN to increase the speed of data transfer.

Raw rate GBits/second and period of time.

80 Gbits/s had been sustained for “many hours” by a CERN team. Over international distances.

With 200 Mbytes/second/experiment.

8 GBits/second from CERN during data taking.
(2 CDs/second).

Data export problem solved.

Needs engineering.

Using a sever client programme written in JAVA
Java is not slow.

“FDT is capable of reading and writing at disk speed over wide area networks (with standard TCP).

It is written in Java.

FDT is based on an asynchronous, flexible multithreaded system.

Uses independent threads to read and write on each physical device.

Transfers data in parallel on multiple TCP streams.

Restores the files from buffers asynchronously.

Resumes a file transfer session without loss, when needed.”

FDT logo

8 CMS data challenge

Since 2003, CMS, Atlas, ALICE and LHCb have run annual data challenges.

Create monte carlo data streams which are structurally identical to the real data streams.

The stream them off site to institutes all over the world.

Run 24 hour/day for many weeks processing and increasing fraction of the “real data rate”

CMS already has largest database in the world (*already*) 2 Petabytes, all accessible within a few minutes.

Using 2000 CPUs world wide

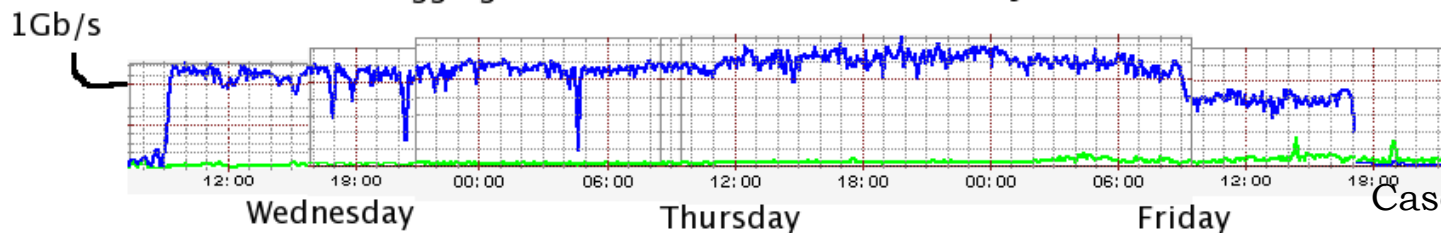
Running 10,000-100,000 jobs per day.

Rutherford Appleton Lab has been moving 200 Mbits per second.

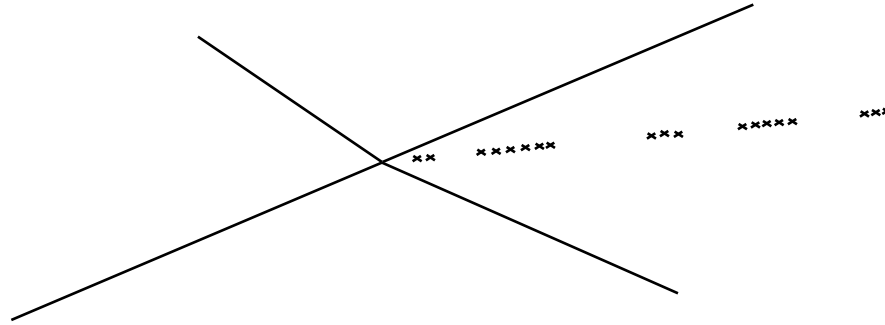
World wide 12 Gbits per second. 24 hours per day for several months.

Managed and aggregate 900 Mbits for 48 hours

48 Hour Aggregate Test from RAL to Tier2s over SJ4 and UKLIGHT



9 Data



At each collision a large number of elementary particles are created – 100's

They move rapidly from their point of creation out in all directions.

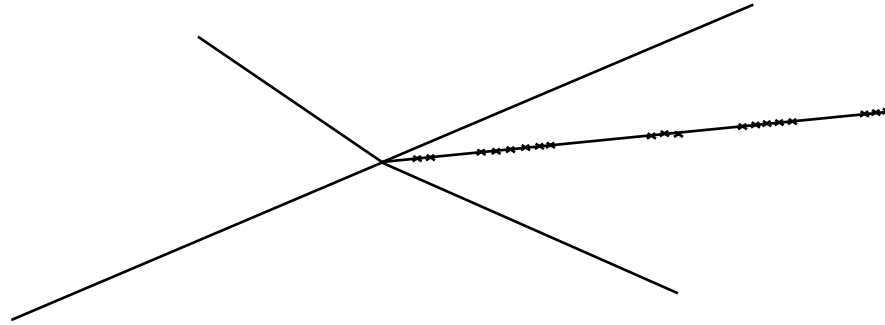
They pass through a number of detectors depositing energy as they go.

The detectors measure the energy deposits.

From this data we can determine how much energy the particles deposit along their paths and their positions in space at a large number of points with high precision *~10 microns*

10

Reconstruction*



From the energy deposited along the track calculate the total energy at the point of creation.

From the energy deposits along the track calculate the space positions. This already involves significant computing power in taking the signal from one of the 100 billion channels. Using “calibration” measurements to translate this into an energy, and survey information to identify the position this channel corresponds to.

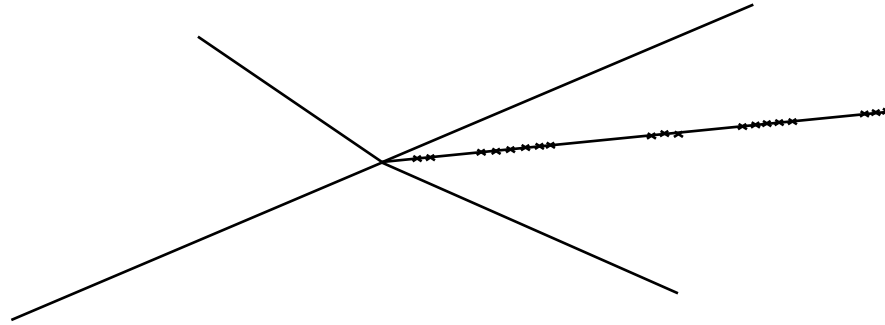
The “points” associated with a single track must be identified. Sophisticated pattern recognition.

Best track calculated from these points by fitting algorithm.

Truly expensive in time.

Kalman filter

11 Monte Carlo



Test Beams

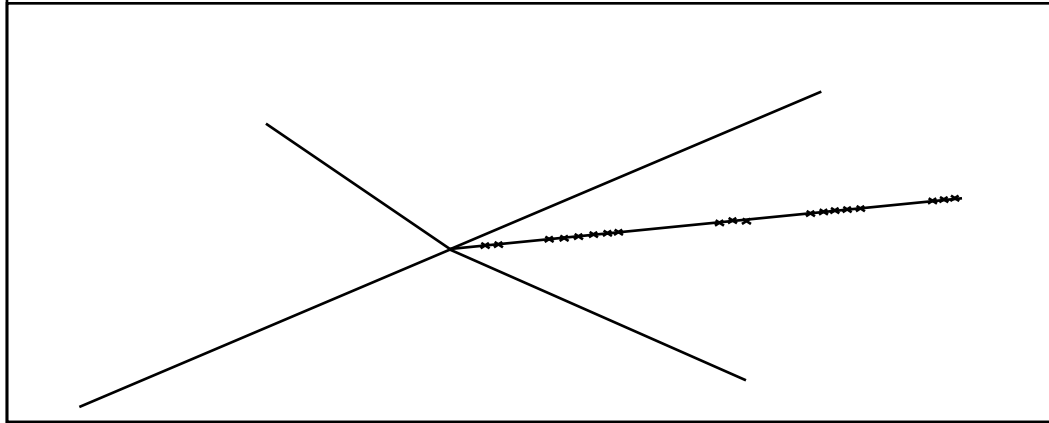
The response of the individual detectors to single particles is measured as part of the detector construction process.

In order to interpret the collisions we need to model what happens when one of these complex “events” occurs inside the detector.

So we create a simulated event (Monte Carlo) with a suitable set of energy and momenta and track it through the detectors, calculating how much energy is deposited at each step and hence what will be measured by the detectors.

The time to model one event is much greater than the time to analyse it.

12 Monte Carlo(i)



Test Beams

These simulated events are then “analysed” as if there were real events and the energy and path of all the particles is deduced from this simulated data.

The “measurements” are then compared with the input to confirm that we are able to accurately “reconstruct” such complex events.

Each event can be seen as a different experiment and so each distinct configuration needs to be simulated and measured.

At least as many simulated as real events.

Different models are compared with data by simulating the “events” with different values of the parameters.

At least as many monte carlo events as real events. As much space to store and as much CPU time to analyse.

Monte carlo data stream identical with real data stream.

The accuracy of a measurement may be limited by the amount of MC which can be generate

Case Study

13 Summary

Real events occur at multi-hertz.
They take some fraction of a minute to analyse.
How long depends on event complexity.

Monte carlo events takes more than 3 minutes to create! They take the same length of time to analyse as the real events.

So say 4 minutes per event on a single PC. Need to produce tens per second.

We need to distribute the data around the world to analyse.

Then distribute the reconstructed events to the physicists laptops for them to analyse.

We need to keep track of all these many events.
what stage have the reached?
where are they stored?

As the data is analysed, the *calibration* of the detectors can be improved, leading to repeated reconstruction of events, each time with improved accuracy!
Which version of the calibration/alignment did a particular dataset use?

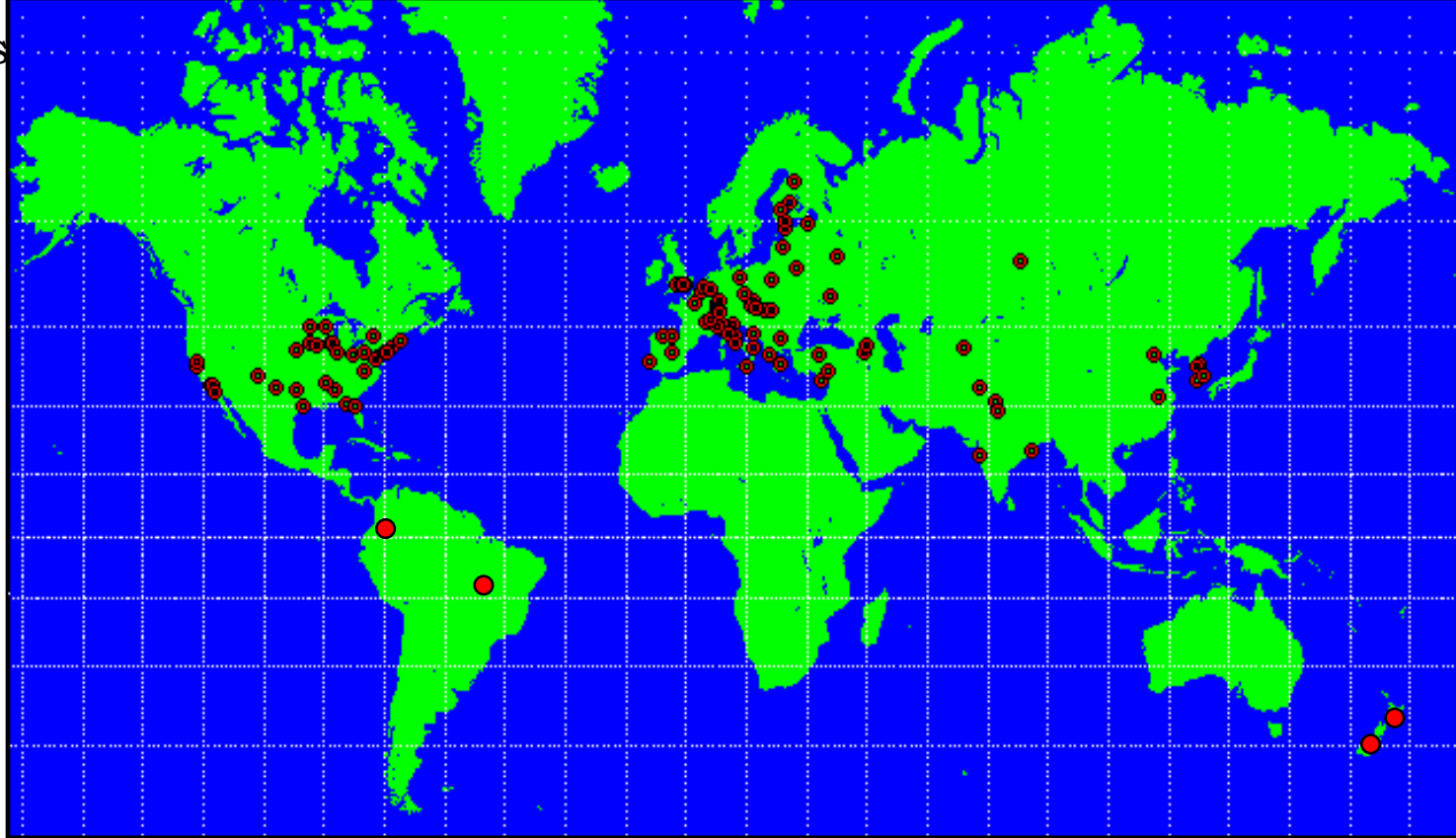
All data created in a volume around 20m on a side.
Distributed world wide

All data created in as volume
not much bigger than a
normal house.

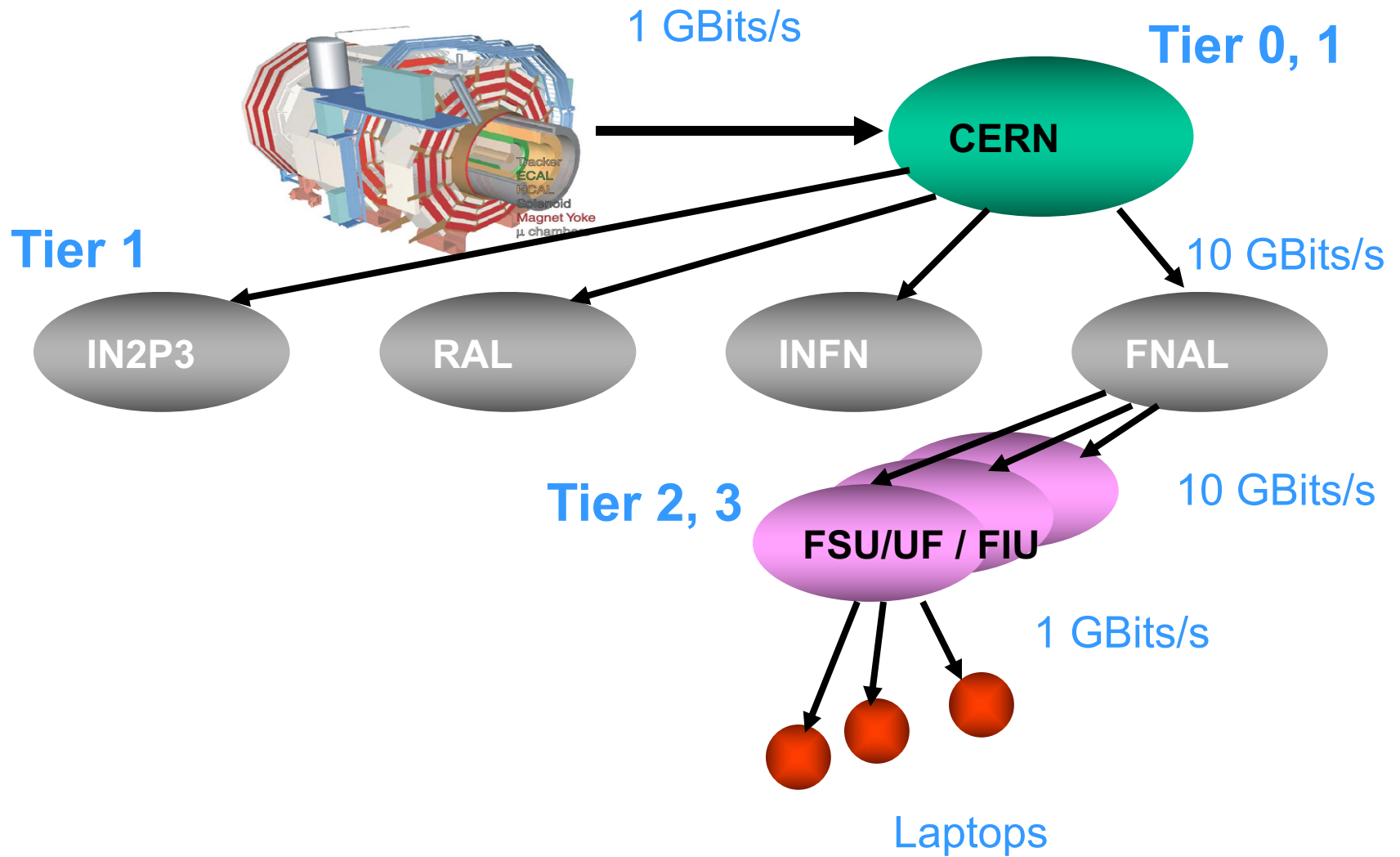
Case Study

Databases

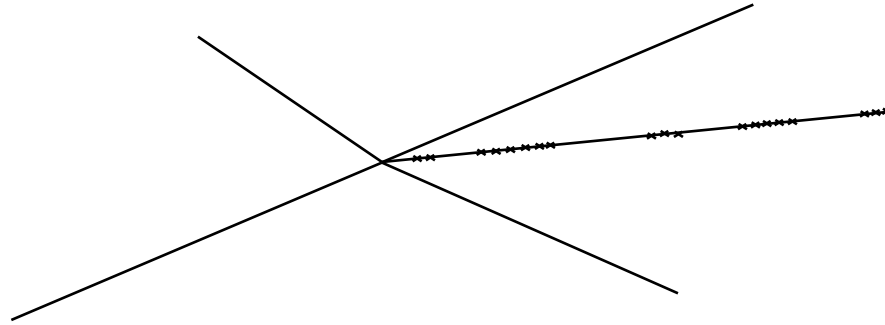
14 CMS institutes



15 CMS Data flow



16 Monte Carlo



Monte Carlo simulation is in general easily parallelisable.

Particle physics code is embarrassingly parallel.

Events generated independently – and reconstructed independently.

Send the data for one event to a processor and run the code and return the processed event.

No communication is required between the processes running the different events.

Pseudo random number generator is providing different sequences on different machines.

Management software for distribution and merge on return.

All Monte Carlo

Discuss this problem in detail, common to parallel monte carlo

Case Study

17 Computing organisation

CERN decided that the data would be available at four levels

Tier 0

The CERN site, all data will be stored at the CERN site, although not all Monte Carlo data needs to be shipped there.

Tier 1

Each country will have a tier 1 site, large subsets of the data will be copied to every country's tier 1. Physicists in a country will access the data from their local tier 1 site. The sites will provide some user support.

Tier 2

Universities and research labs will provide tier 2 resources. Only data of interest to that institute's scientists will be copied to the tier 2 sites. User support is expected to be much more limited.

Tier 3

Departmental machines, the machines on the scientist's desk.

Design early 1990's. Late 1990's grid invented and identified as delivery mechanism

Now laptops

Digital sky

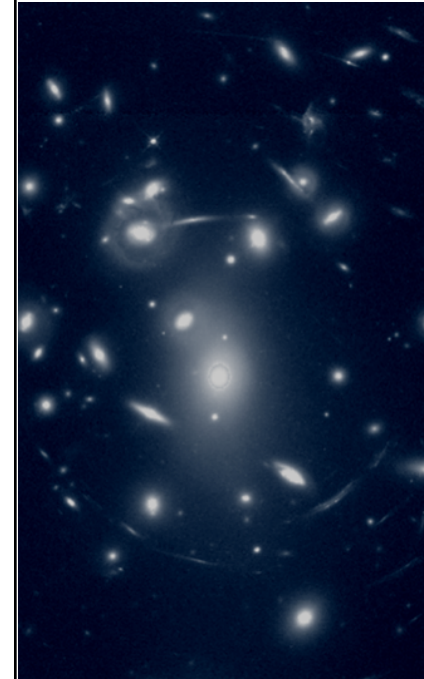
Many astronomical surveys – whole sky
40,000 square degrees
1/2 trillion pixels
1 TB x multi-wavelengths
1 billion sources

Want to be able to access all surveys for a source.
Multi-databases, different local implementation and
management domains.

No central management.
No common data organisation.
Widely separated databases.

The solution is to create a middleware stack which
allows access to systems in a unified manner,
including access to a system which looks like a single
domain.

The answer is to use X.500 certificates.



Large scale computing



Cloud
Economies of scale



WLCG
Worldwide LHC Computing Grid

Grid

Many management domains