Data Management at the Large Hadron Collider

e-Science
CIC Center for Library Initiatives conference 2008

Norbert Neumeister
Department of Physics
Purdue University

Outline

Outline

- Introduction
  - What is Particle Physics
  - Why do we go to the energy frontier
  - The Large Hadron Collider
  - The Compact Muon Solenoid detector
- Computing and Data Handling
  - Motivation and Requirements
  - Processing
  - Storage
  - Data Management
- Summary

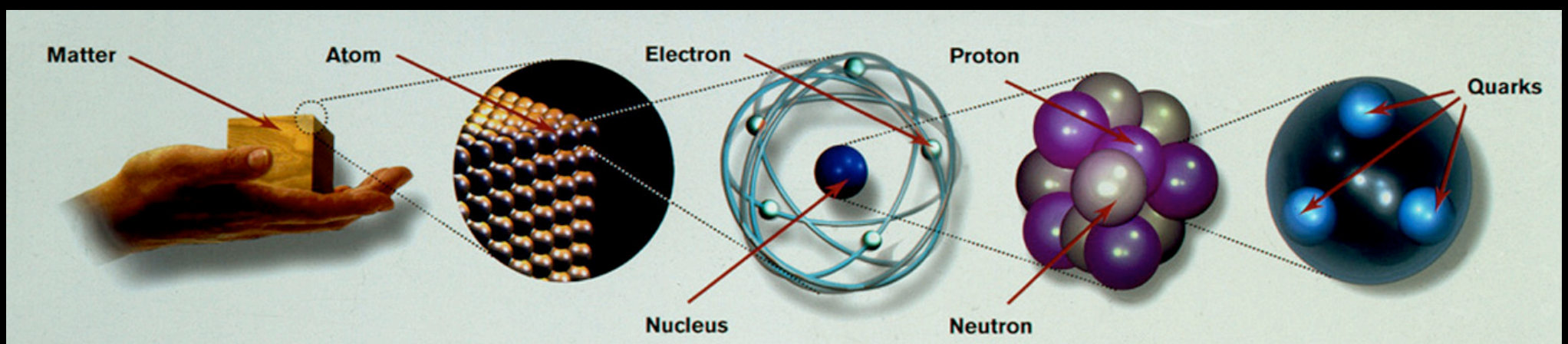
Particle Physics

Aim to answer the two following questions:

- What are the elementary constituents of matter?
- What are the fundamental forces that control their behavior at the most basic level?

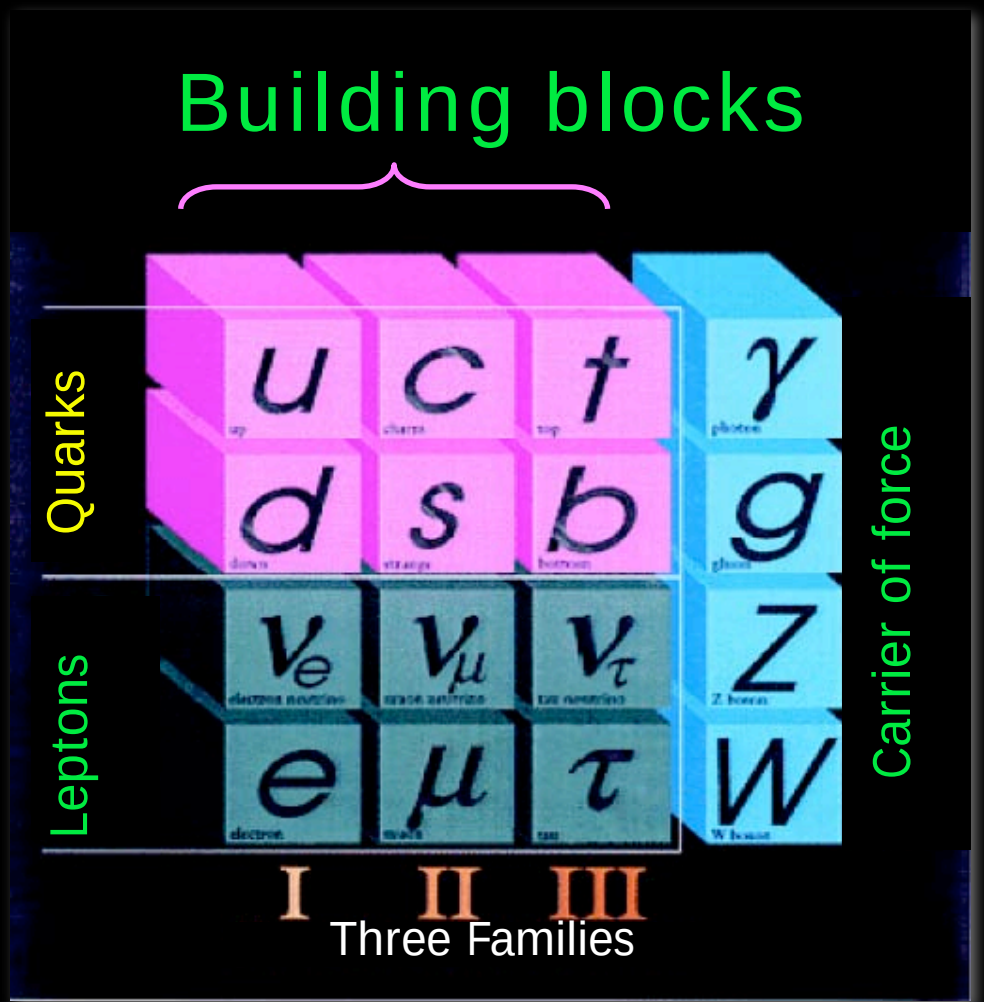
Tools:

- Particle Accelerators
- Particle Detectors
- Computers

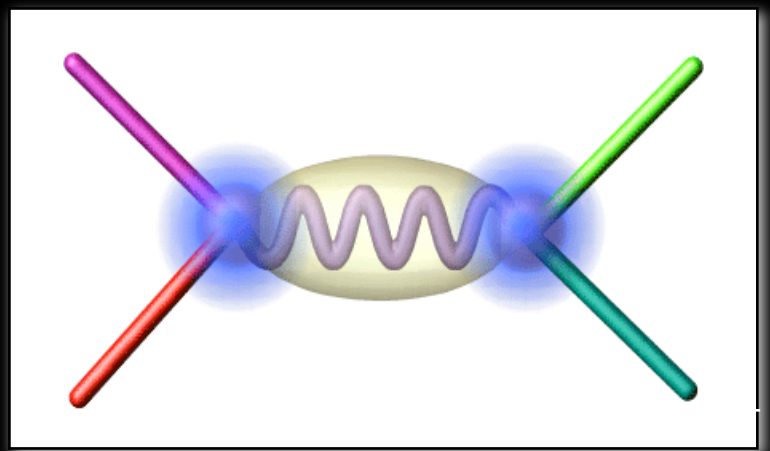


atom	nucleus	nucleon	quark
10^{-10} m	10^{-14} m	10^{-15} m	10^{-18} m

Standard Model of Particle



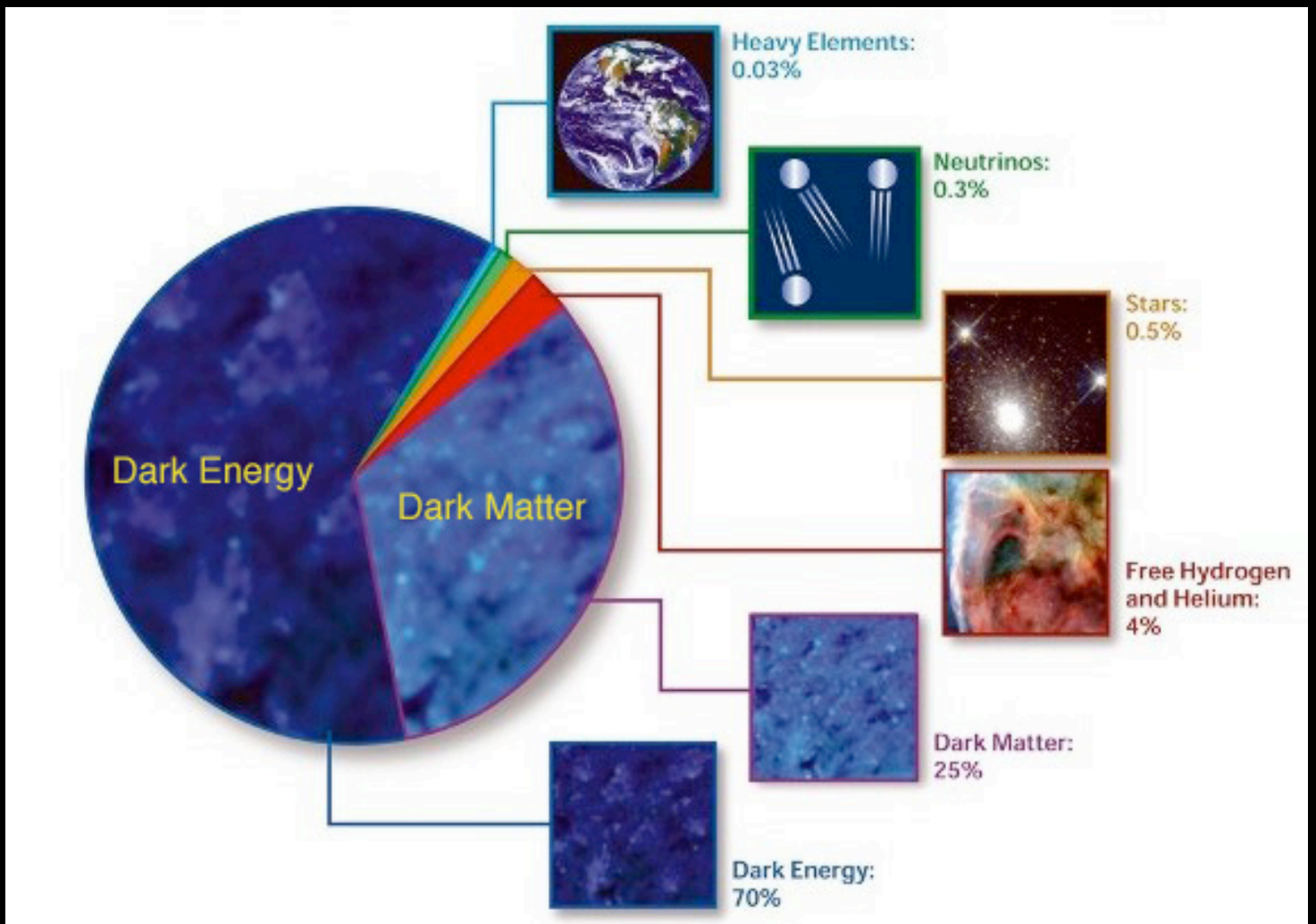
After many years,
no unambiguous evidence of
new physics!



- Very successful model which contains all known particles in particle physics today.
- Describes the interaction between spin 1/2 particles (*quarks* and *leptons*) mediated by spin 1 *gauge bosons* (gauge symmetry).
- The SM has been tested at ‰ level
- All particles discovered, except Higgs Boson

The Universe

Stars and Planets only account for a small percentage of the universe!

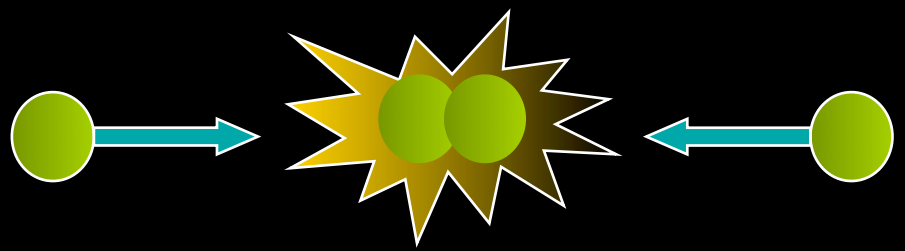


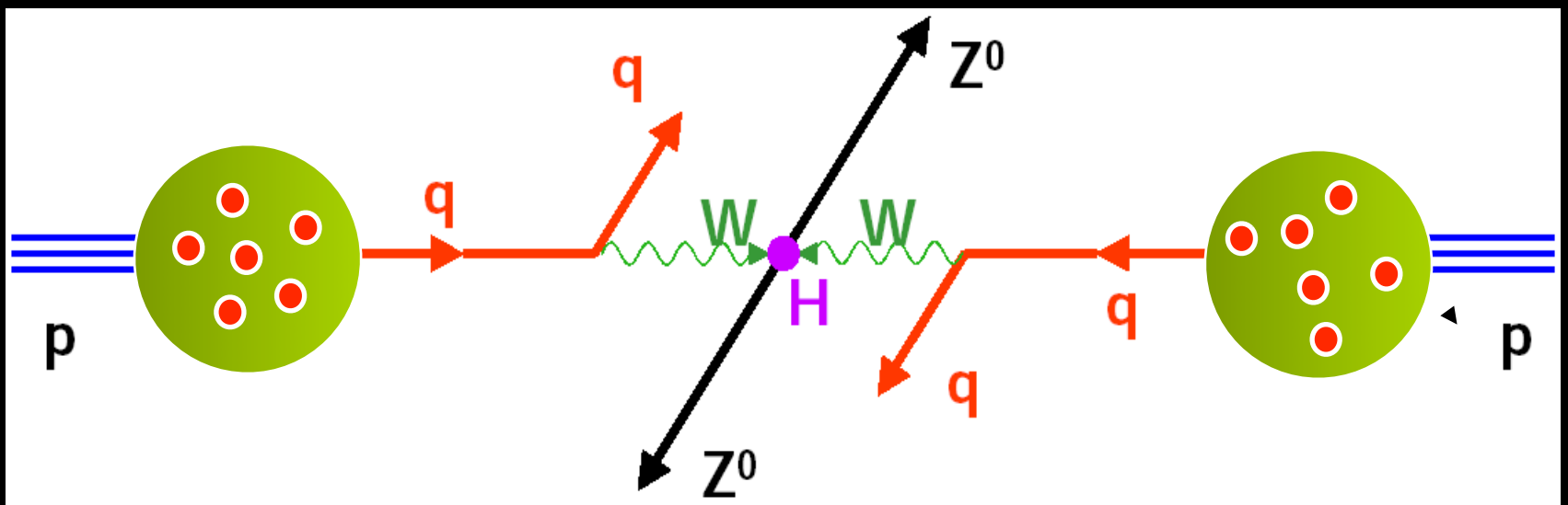
Probing the TeV Energy Scale

- Higher energy: Reproduce conditions of early Universe
- TeV energy scale: Expect breakdown of current calculations unless a new interaction or phenomenon appears
- Many theories, but need data to distinguish between them
- What might we find:
  - The mechanism that generates mass for all elementary particles
    - In Standard Model, masses generated through interaction with a new particle the Higgs
    - Other options possible, but we know that the phenomena occurs somewhere between 100 and 1000 GeV
  - A new Symmetry of Nature
    - Supersymmetry gives each particle a partner
    - Would provide one source of the Dark Matter observed in the Universe
  - Extra Space-Time Dimensions
    - String theory inspired
    - This would revolutionize Physics!

Particle Collisions

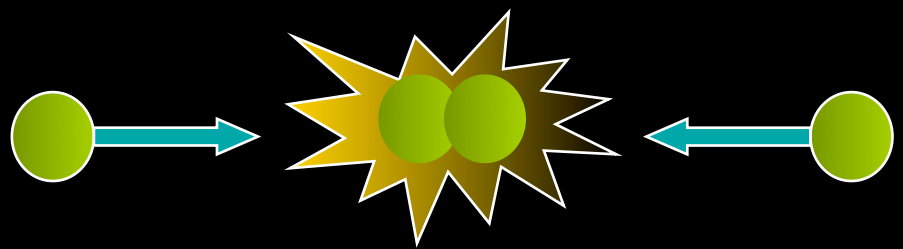
Collider:

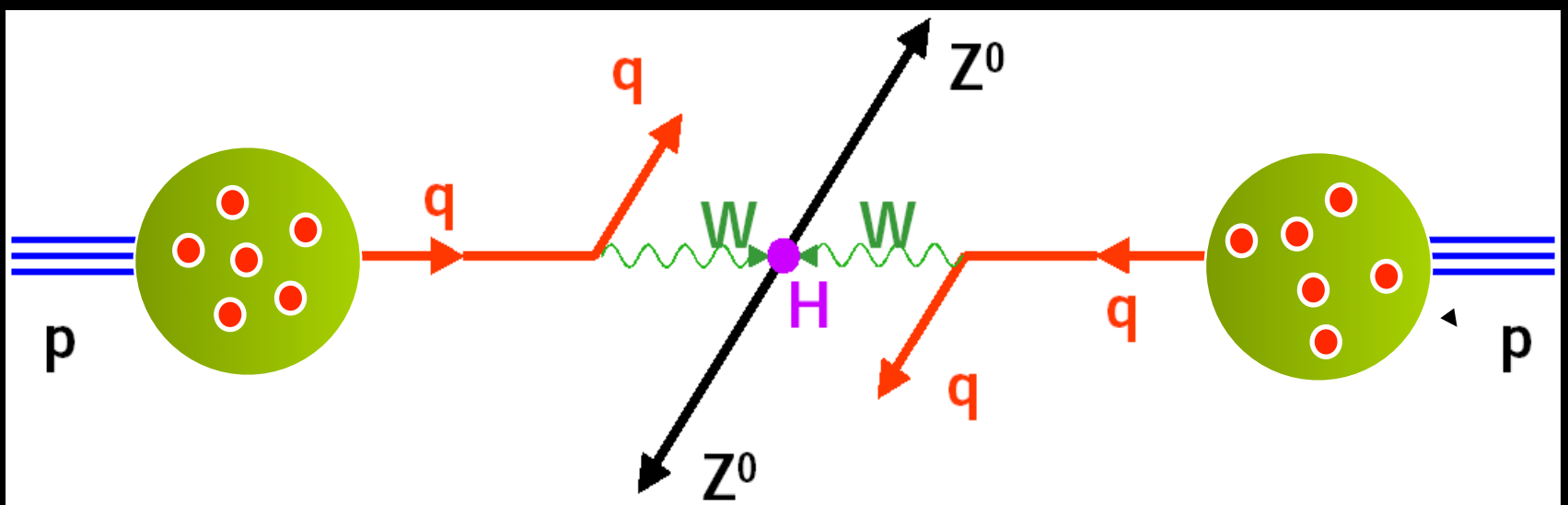




Particle Collisions

Collider:

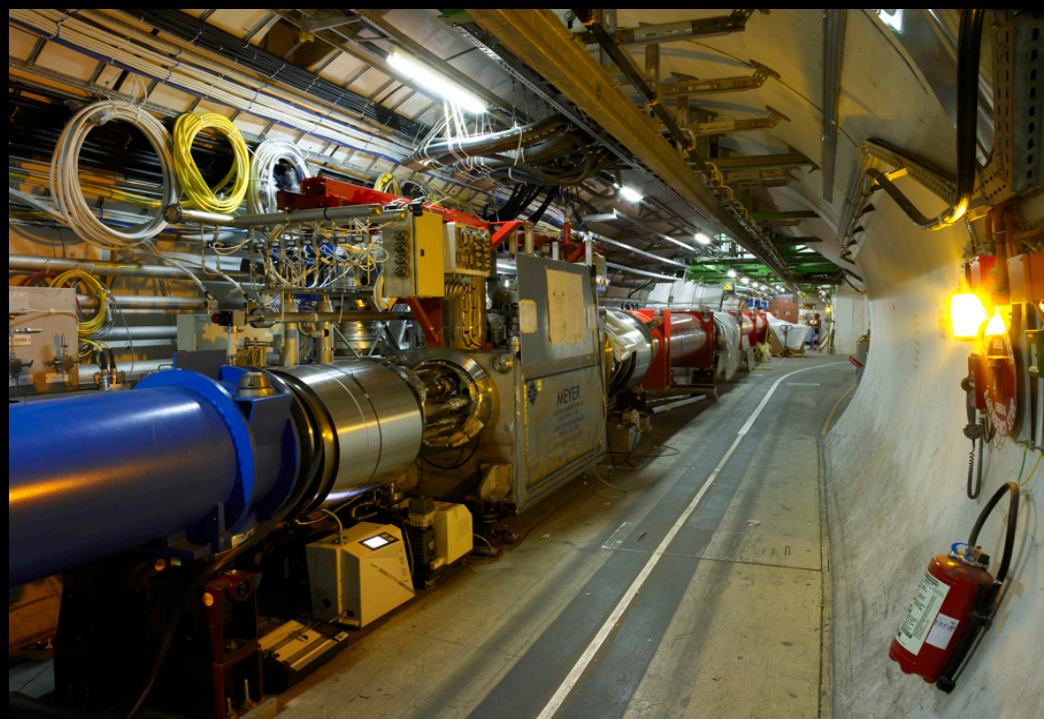




→ **Proton-proton collider with $E_p \geq 7$ TeV**

The Large Hadron Collider

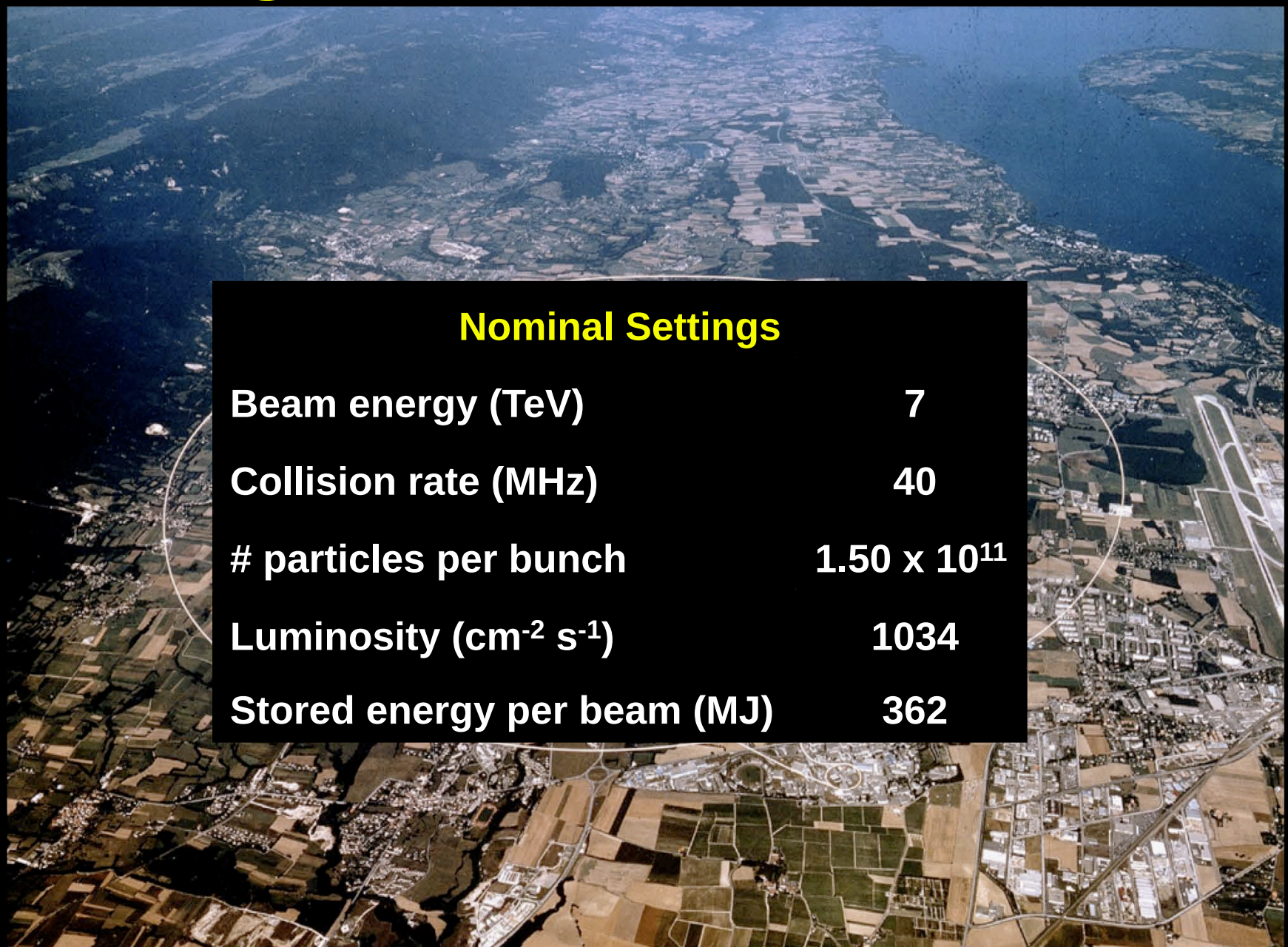
- Energy: 14 TeV
  - (7 x current best)
- Intensity:
  - Initial 10 fb^{-1}/year (5 x current best)
- First Data: Summer 2008
- Official LHC inauguration:
  - 21 Oct. 2008



In general the LHC was designed as a “broadband” discovery machine which aims for the largest possible energy and the largest possible luminosity

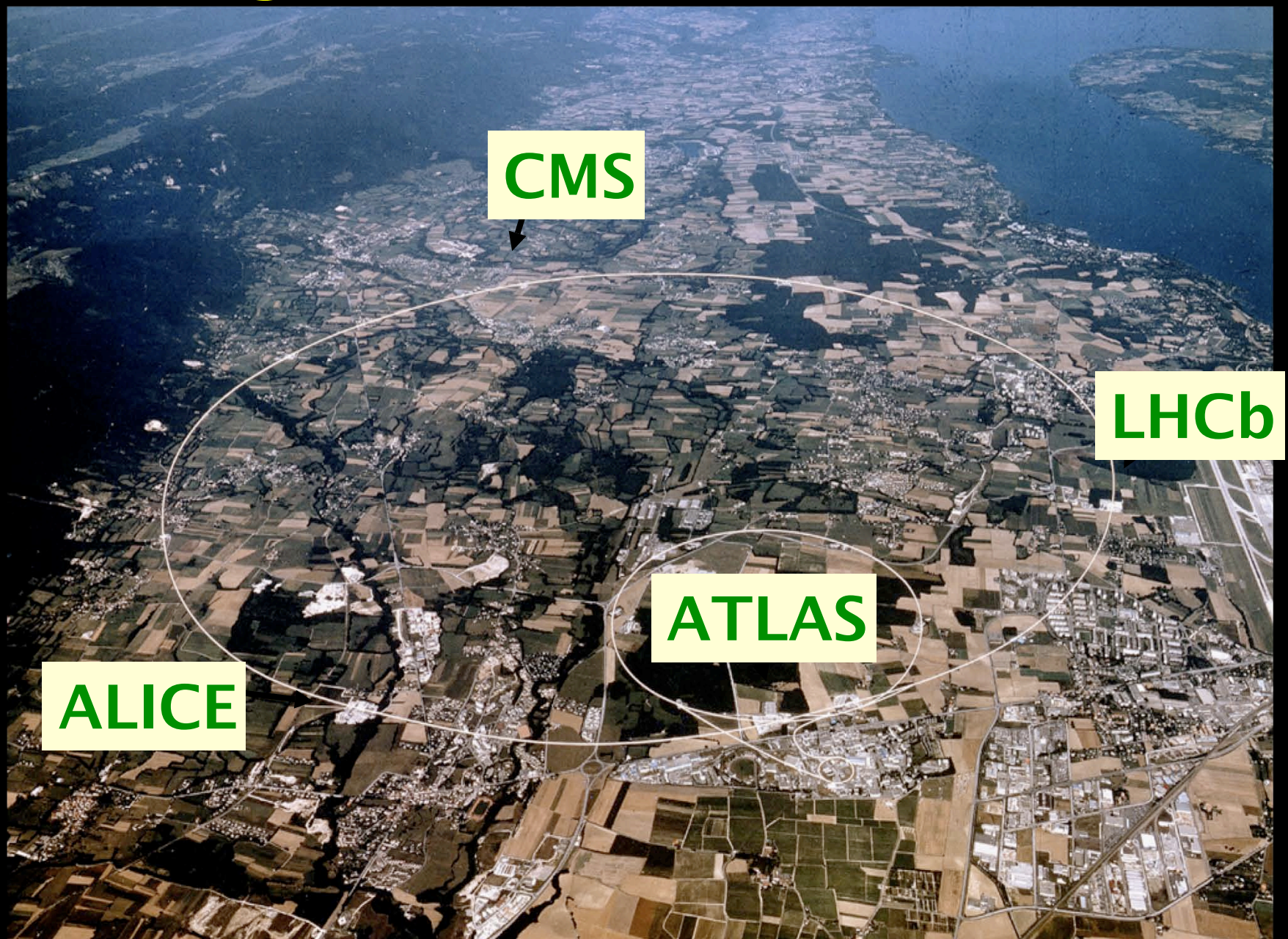
New energy frontier, high luminosity proton proton collider at CERN, Geneva, Switzerland

The Large Hadron Collider

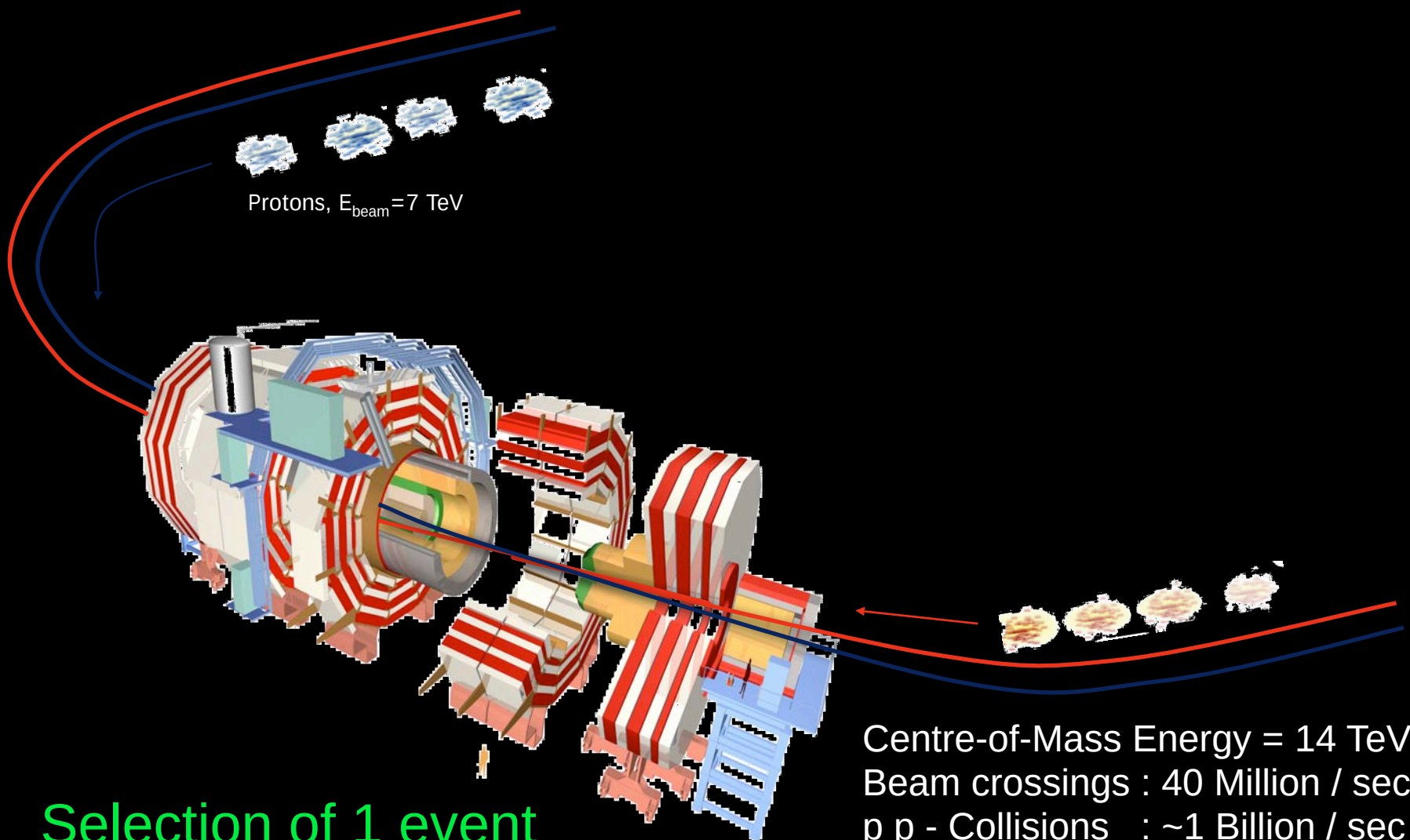


Nominal Settings	
Beam energy (TeV)	7
Collision rate (MHz)	40
# particles per bunch	1.50×10^{11}
Luminosity ($cm^{-2}\ s^{-1}$)	1034
Stored energy per beam (MJ)	362

The Large Hadron Collider



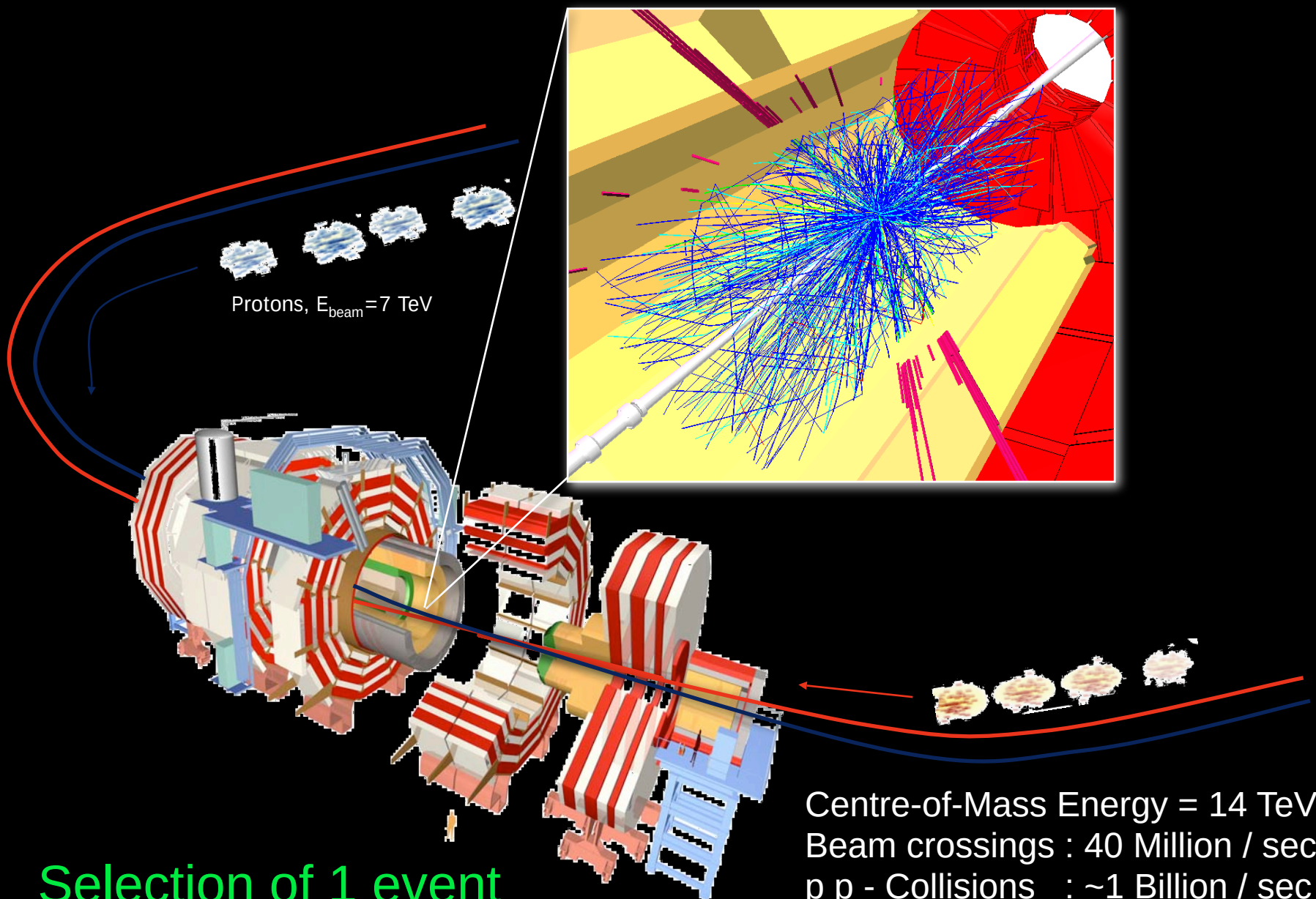
Collisions at the LHC



Selection of 1 event
in 10,000,000,000,000

Centre-of-Mass Energy = 14 TeV
Beam crossings : 40 Million / sec
p p - Collisions : ~1 Billion / sec
Events to tape : ~100 / sec, each 1-2 MByte

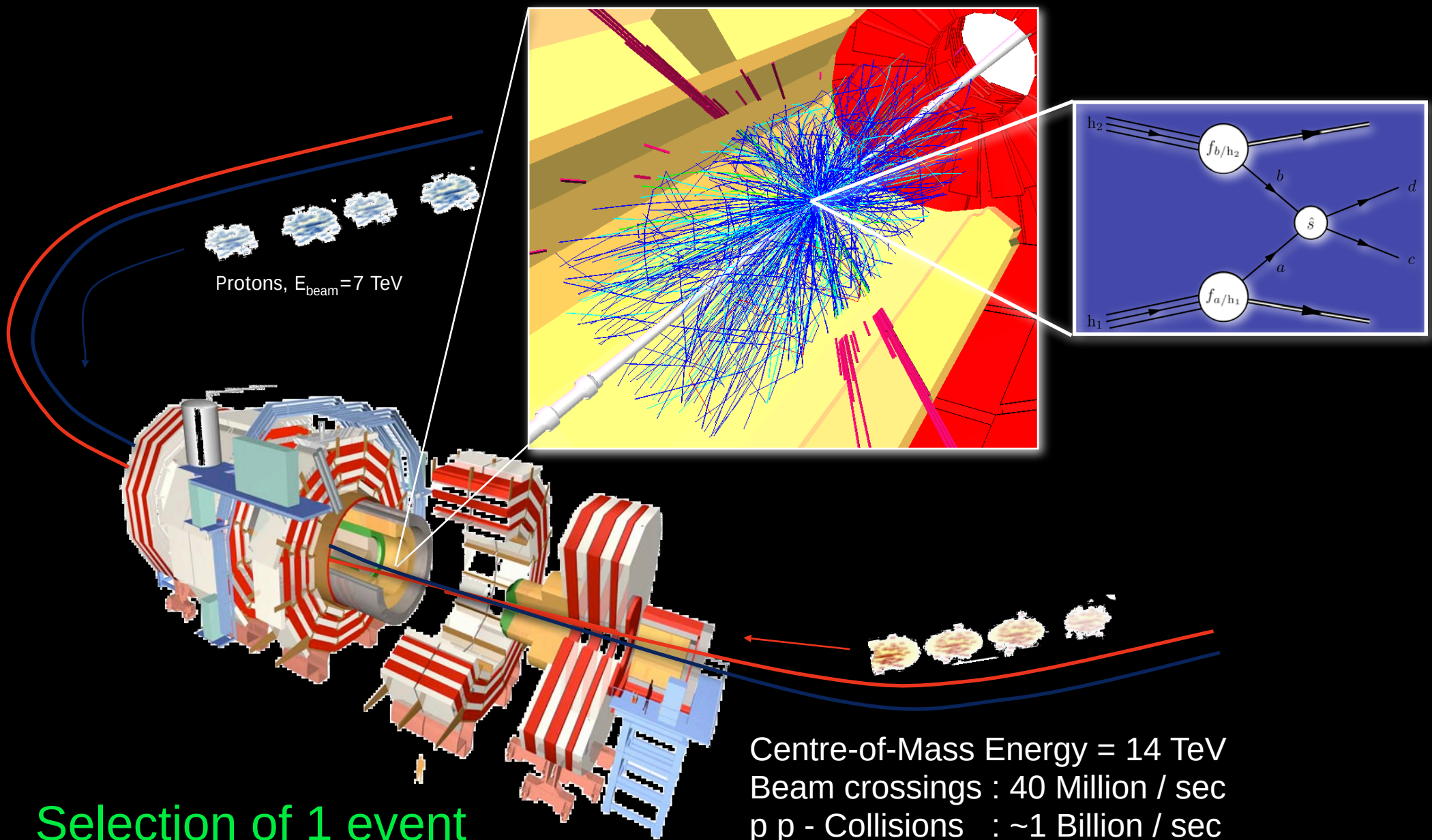
Collisions at the LHC



Selection of 1 event
in 10,000,000,000,000

Centre-of-Mass Energy = 14 TeV
Beam crossings : 40 Million / sec
p p - Collisions : ~1 Billion / sec
Events to tape : ~100 / sec, each 1-2 MByte

Collisions at the LHC

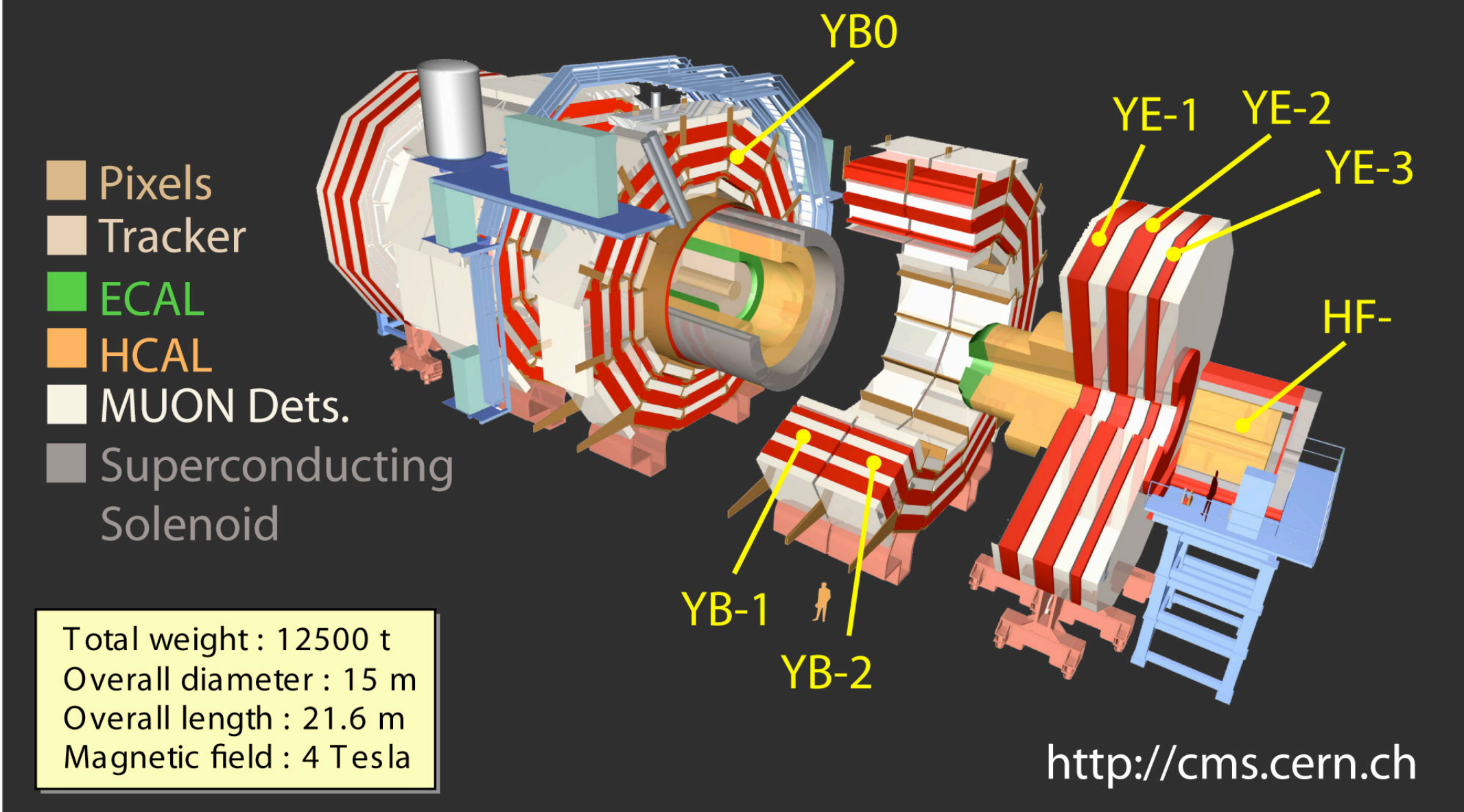


Selection of 1 event
in 10,000,000,000,000

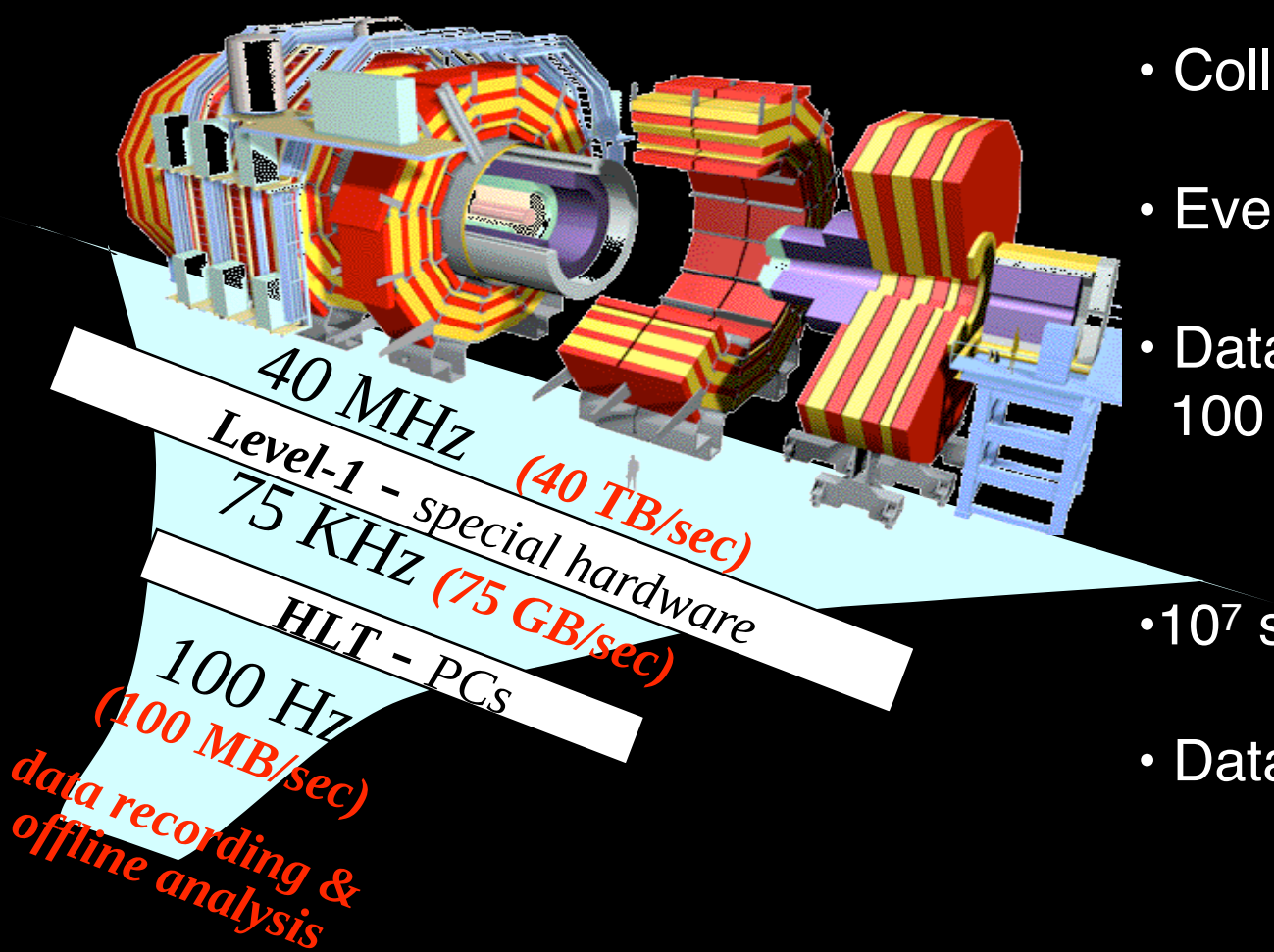
Centre-of-Mass Energy = 14 TeV
Beam crossings : 40 Million / sec
p p - Collisions : ~1 Billion / sec
Events to tape : ~100 / sec, each 1-2 MByte

The CMS Detector

(CMS collaboration: 184 Institutions with about 2880 scientists)

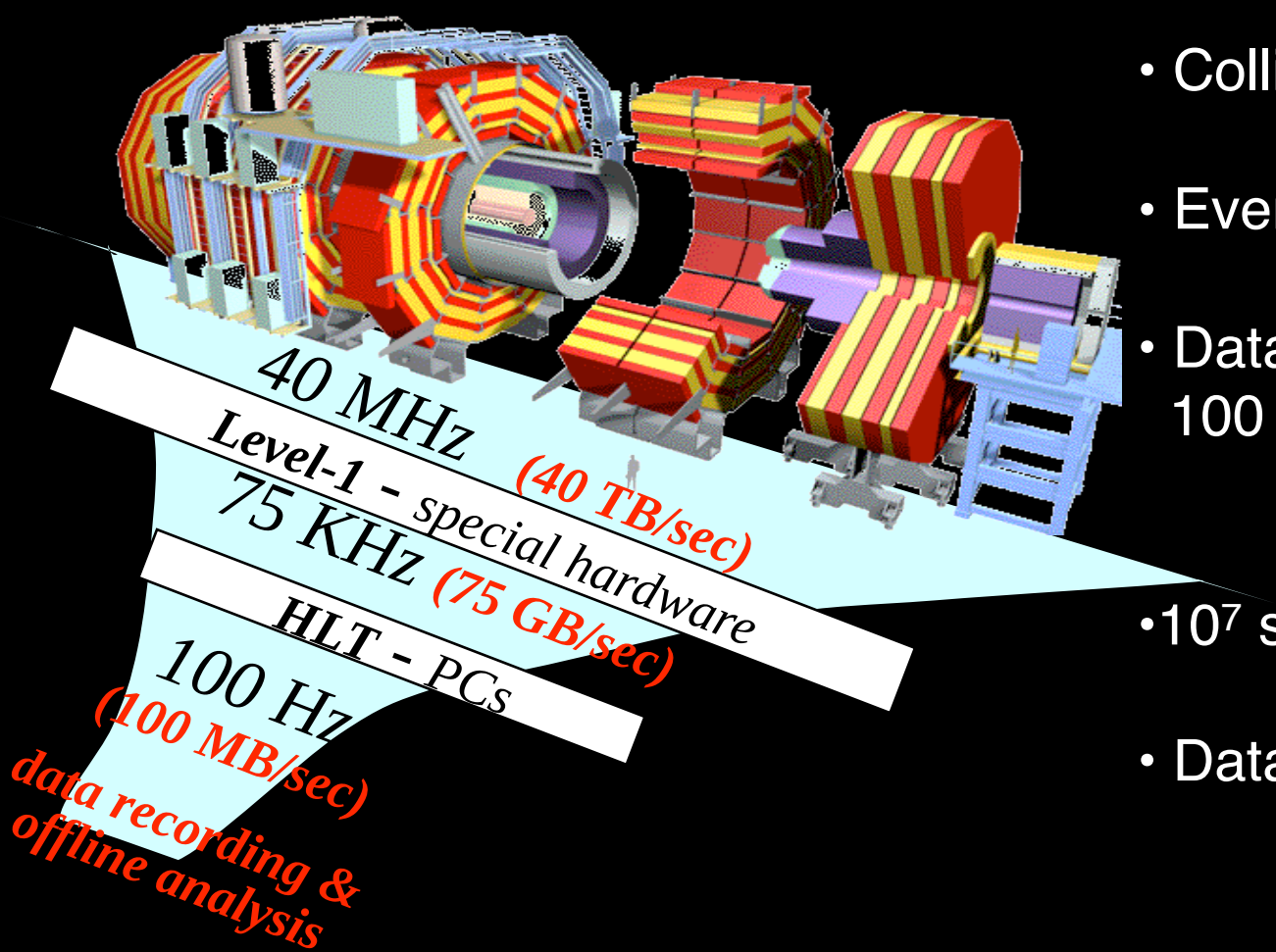


Data Recording



- Collision rate: 40 MHz
- Event size: ≈ 1 MByte
- Data size: 1 MByte/event
  100 events/s
  → 100 MByte/s
- 10^7 s data taking per year
- Data size: 1 PetaByte =
  10^6 GByte per year

Data Recording



- Collision rate: 40 MHz
- Event size: ≈ 1 MByte
- Data size: 1 MByte/event
  100 events/s
  → 100 MByte/s
- 10^7 s data taking per year
- Data size: 1 PetaByte =
  10^6 GByte per year

~ PetaBytes/year
~10^9 events/year
~10^3 batch and interactive users

LHC Data Challenge

- The LHC generates $40 \cdot 10^6$ collisions / s
- Combined the 4 experiments record:
  - 100 interesting collision per second
  - 1 ÷ 12 MB / collision $\Rightarrow$ 0.1 ÷ 1.2 GB / s
  - ~ 10 PB (10^{16} B) per year (10^{10} collisions / y)
  - LHC data correspond to $20 \cdot 10^6$ DVD's / year!
    - Space equivalent to 400.000 large PC disks
  - Computing power ~ 10^5 of today's PC

Balloon (30 Km)

LHC data: DVD stack after 1 year! (~ 20 Km)

Airplane (10 Km)

Mt. Blanc (4.8 Km)

LHC Data Challenge

- The LHC generates $40 \cdot 10^6$ collisions / s
- Combined the 4 experiments record:
  - 100 interesting collision per second
  - 1 ÷ 12 MB / collision $\Rightarrow$ 0.1 ÷ 1.2 GB / s
  - ~ 10 PB (10^{16} B) per year (10^{10} collisions / y)
  - LHC data correspond to $20 \cdot 10^6$ DVD's / year!
    - Space equivalent to 400.000 large PC disks
  - Computing power ~ 10^5 of today's PC

LHC data: DVD stack after 1 year!
(~ 20 Km)

Airplane
(10 Km)

Mt. Blanc
(4.8 Km)

LHC Data Challenge

- The LHC generates $40 \cdot 10^6$ collisions / s
- Combined the 4 experiments record:
  - 100 interesting collision per second
  - 1 ÷ 12 MB / collision $\Rightarrow$ 0.1 ÷ 1.2 GB / s
  - ~ 10 PB (10^{16} B) per year (10^{10} collisions / y)
  - LHC data correspond to $20 \cdot 10^6$ DVD's / year!
    - Space equivalent to 400.000 large PC disks
  - Computing power ~ 10^5 of today's PC

Using parallelism is the only way to analyze this amount of data in a reasonable amount of time

The HEP Environment

- HEP collaborations are quite large
  - Order of 1000 collaborators, globally distributed
  - CMS is one of four Large Hadron Collider (LHC) experiments being built at CERN
- Typically resources are globally distributed
  - Resources organized in tiers of decreasing capacity
  - Raw data partitioned between sites, highly processed ready-for-analysis data available everywhere
- Computing resources in 2008:
  - 34 Million SpecInt2000
  - 11 PetaByte of disk
  - 10 PetaByte of tape
  - Distributed across ~25 countries in ~4 continents

Computing Model

- Tier-0: Host of CMS @ CERN, Switzerland
  - Prompt reconstruction & “back-up” archive
- Tier-1: in 7 countries across 3 continents
  - Distributed “life” archive
  - All (re-)reconstruction & primary filtering for analysis @ Tier-2
- Tier-2: ~50 clusters in ~25 countries
  - All simulation efforts
  - All physics analysis
  - General end-user analysis for local communities or physics groups

Data Organization

- **HEP data are highly structured**
- "event" ~ 1MByte
  - Atomic unit for purpose of science
- File ~ 1Gigabyte
  - Atomic unit for purpose of data catalogue
- Block of files ~ 1Terabyte
  - Atomic unit for purpose of data transfer
- A science dataset generally consists of many blocks with same provenance.
- A science result generally requires analysis of multiple datasets.

Data Management System

- Manage **Data** and **Meta-data** from where it is created
  - from production, online farm, calibration, reconstruction, re-processing
  - to where its being used for physics
- Consider storage/retrieval of run/time dependent non-event data
  - such as calibrations, alignment and configuration
- Consider the system of tapes, disk and networks
  - to support moving, placing, caching, pinning datasets

Data Management System

- Keep it simple!
- Optimize for the common case:
  - Optimize for read access (most data is write-once, read-many)
  - Optimize for organized bulk processing, but without limiting single user
- Decouple parts of the system:
  - Site-local information stays site-local
- Use explicit data placement
  - Data does not move around in response to job submission
  - All data is placed at a site through explicit CMS policy
- Grid interoperability (LCG and OSG)

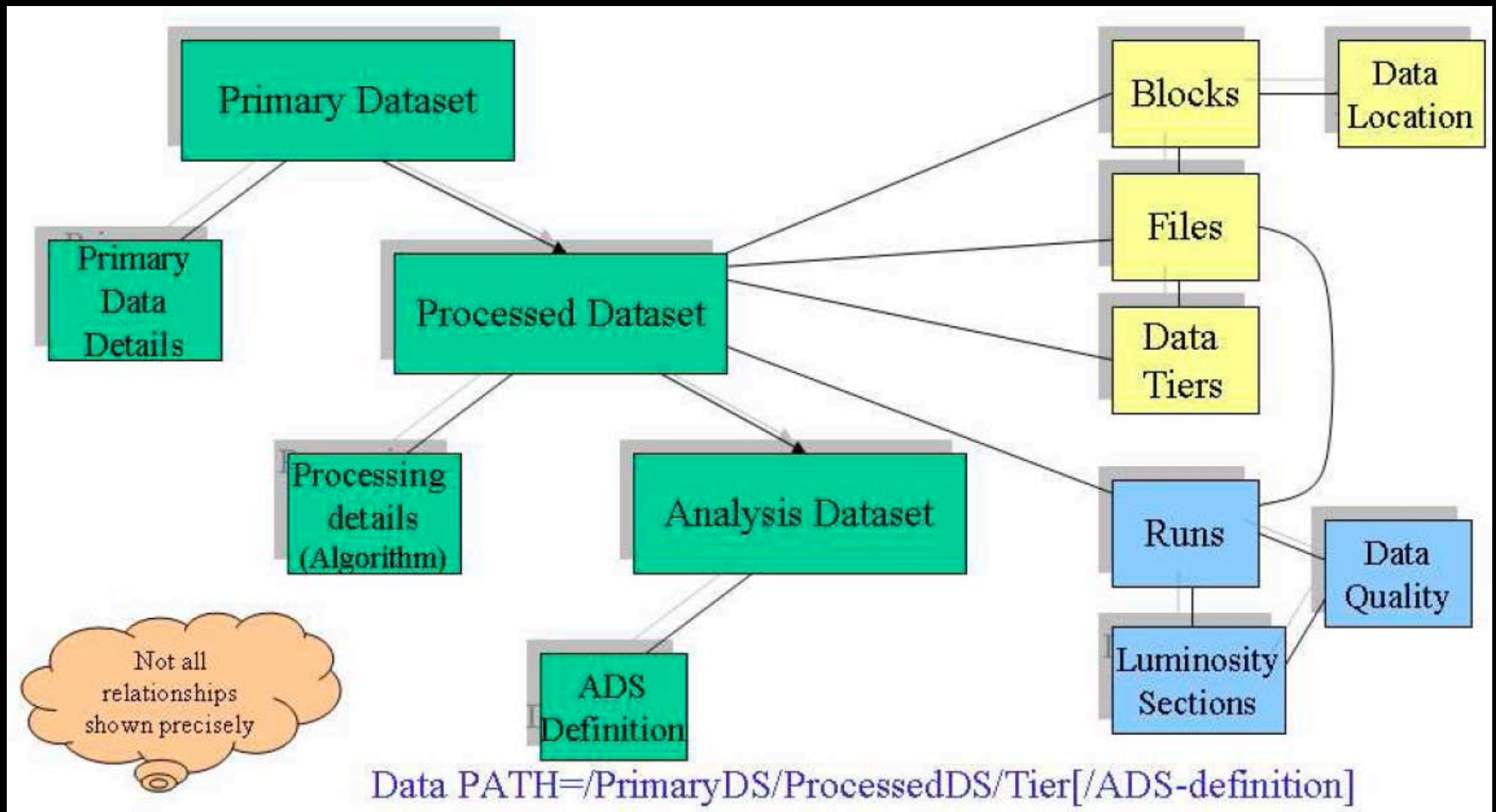
Data Management Components

- Data Bookkeeping Service (DBS)
  - Catalog all event data
  - Provide access to a catalog of all event data
  - Records the files and their processing parentage
  - Track provenance of data
- Data Movement Service (PHEDEX)
  - Move data at defined granularities from multiple sources to multiple destinations
  - Provide scheduling
  - Push vs. Pull mode
  - User Interface

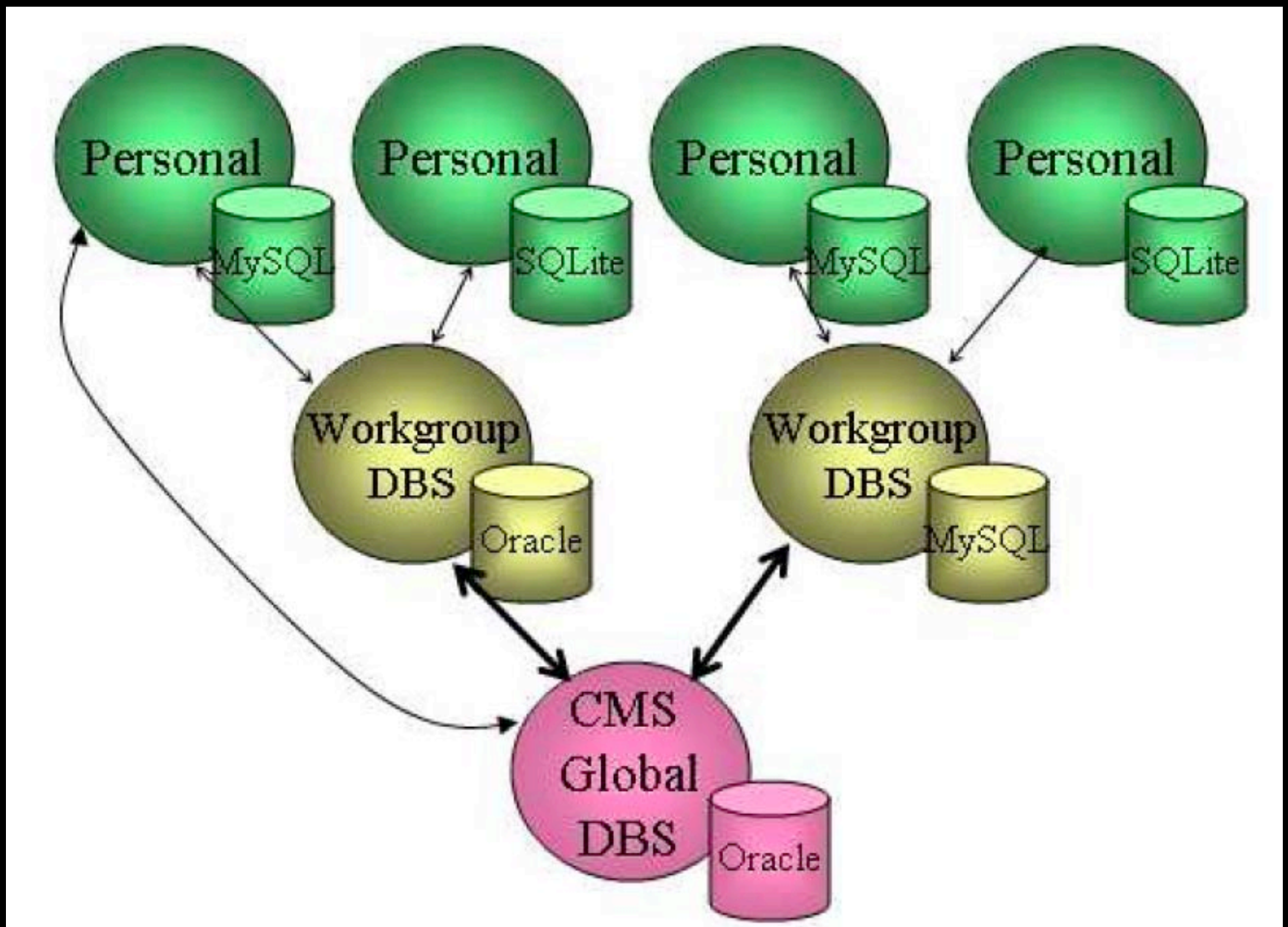
Data Bookkeeping Service

- Keep the dataset catalog for dataset lookups
- Import into and move/track datasets through the system
  - through data movement service
- Implement data placement policies etc.
- Interact with local storage managers
  - as needed
- Support all workflows
- Ensure and verify consistency of distributed data sets
  - local custodianship of Tier-1 centers

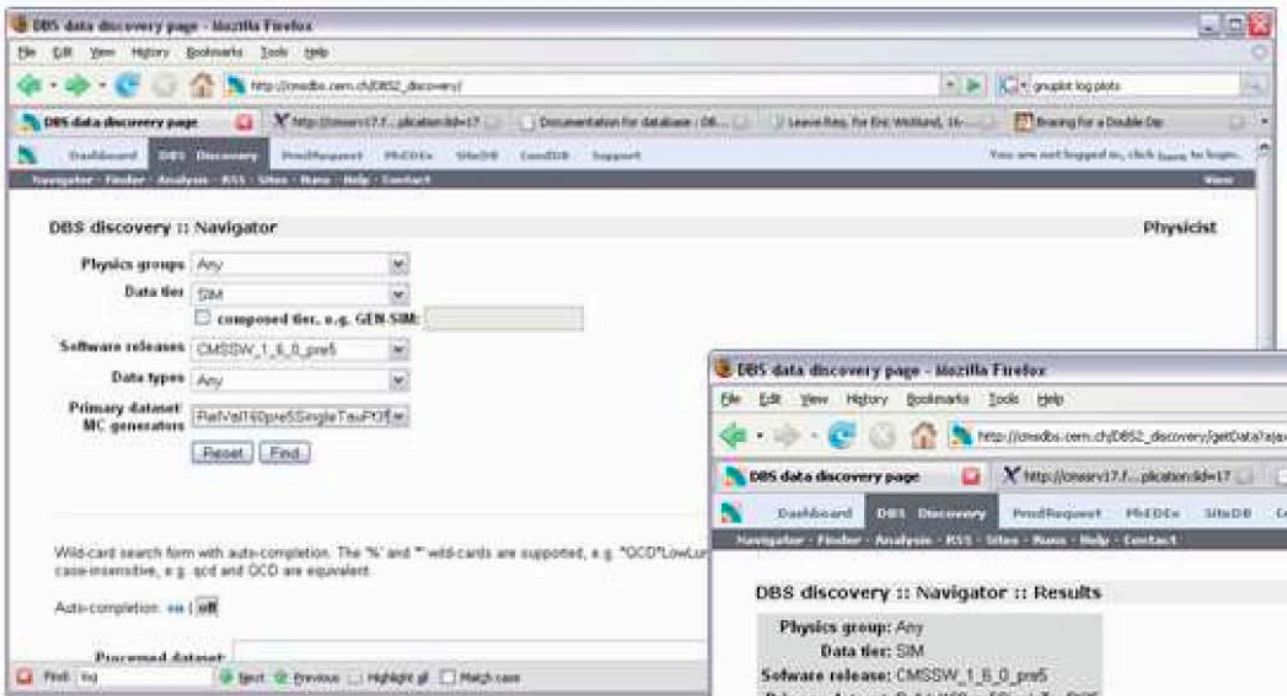
Relationship of the DBS Schema

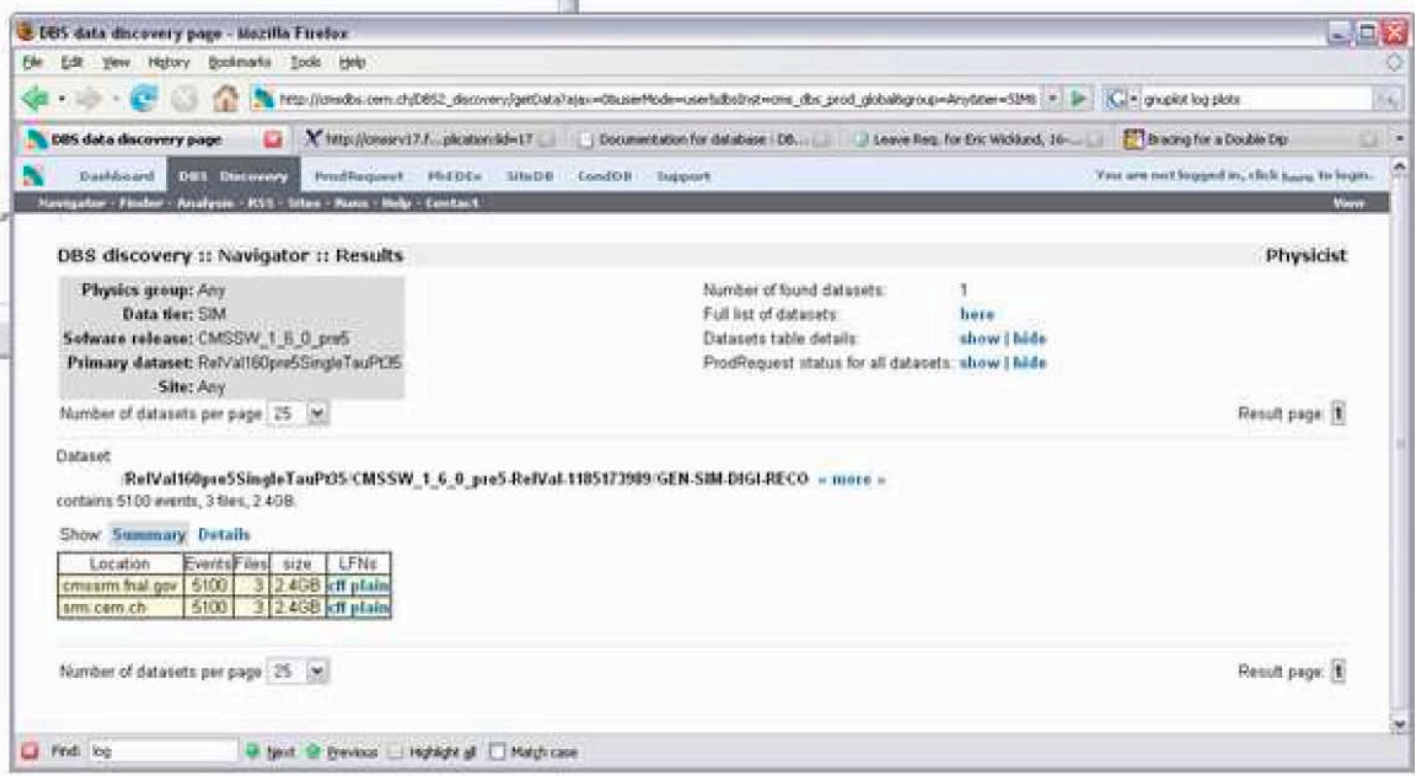


DBS Instance Hierarchy



User Interface





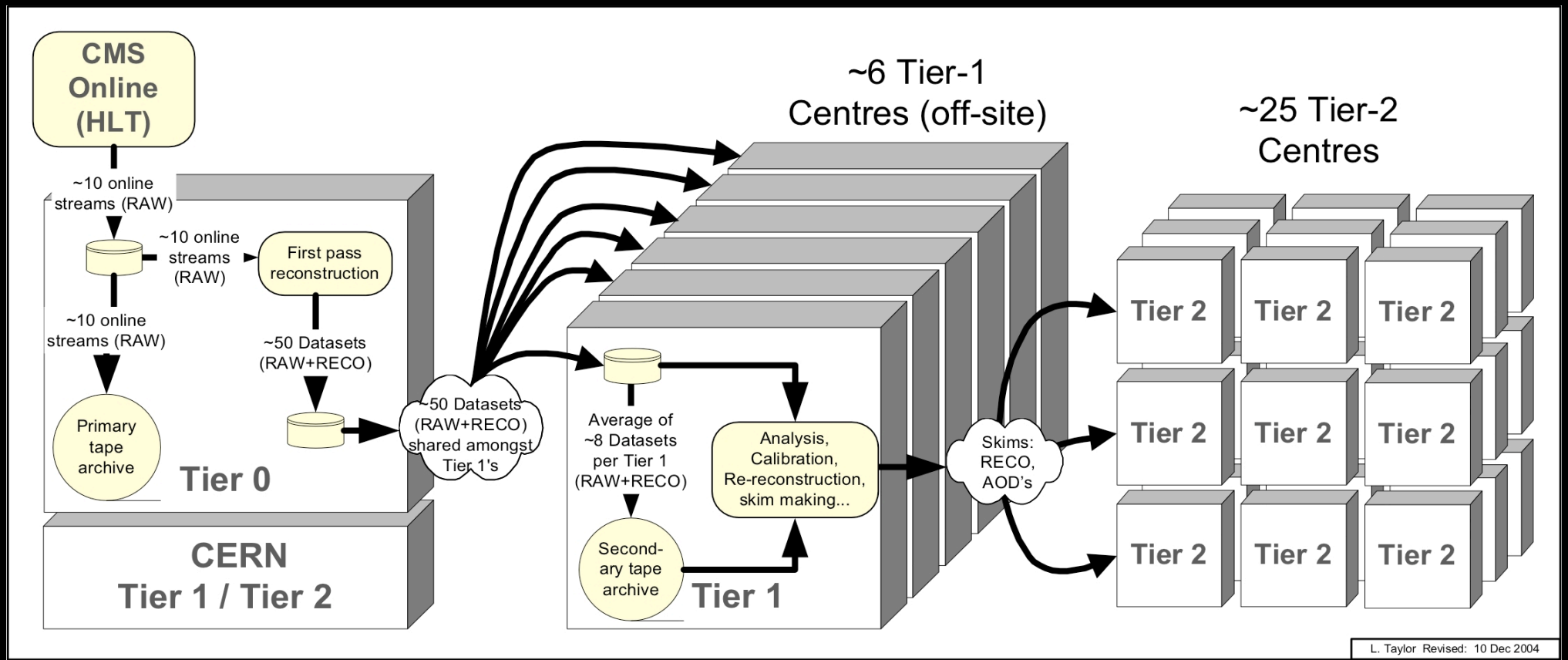
Data Movement Service

- High-Level Functionality
  - Move data at defined granularities from multiple sources to multiple destinations through a defined topology of buffers
  - Provide scheduling information on latency, rate
- Low-Level Functionality
  - Enables buffer management
  - Manages scalable interaction with local resource management
  - Maintains (purely) replica metadata: Filesize, checksums
  - Can manage aggregation of files…
  - Controls introduction of data into the network/Grid
  - Manages node to node transfer
    - Nodes map to location, but also function, providing interface to workflow management
  - Interacts/overlaps with global replica location structure

Data Movement Service Design

- Keep complex functionality in discrete agents
  - Handover between agents minimal
  - Agents are persistent, autonomous, stateless, distributed
  - System state maintained using a modified blackboard architecture
- Layered abstractions make system robust
- Keep local information local where possible
  - Enable site administrators to maintain local infrastructure
  - Robust in face of most local changes
- Draws inspiration from agent systems, “autonomic” and peer-to-peer computing

CMS Data Flow

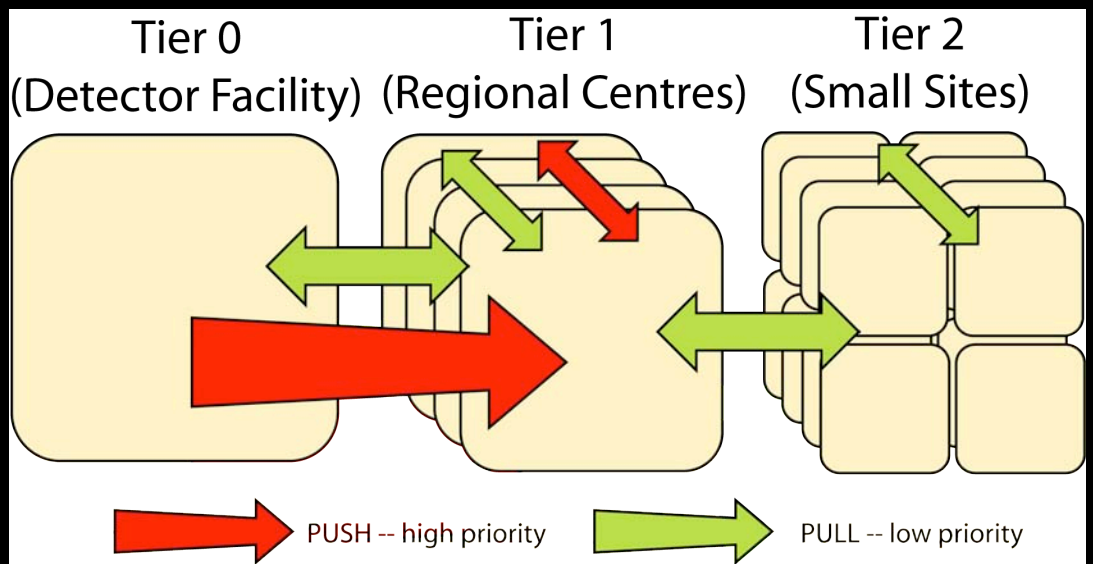


Tier-0:
- Gets data from DAQ
- Prompt reconstruction
- Data archive and distribution to Tier-1's

Tier-1's:
- Making samples accessible for selection and distribution
- Data-intensive analysis
- Re-processing
- Calibration
- FEVT, MC data archiving

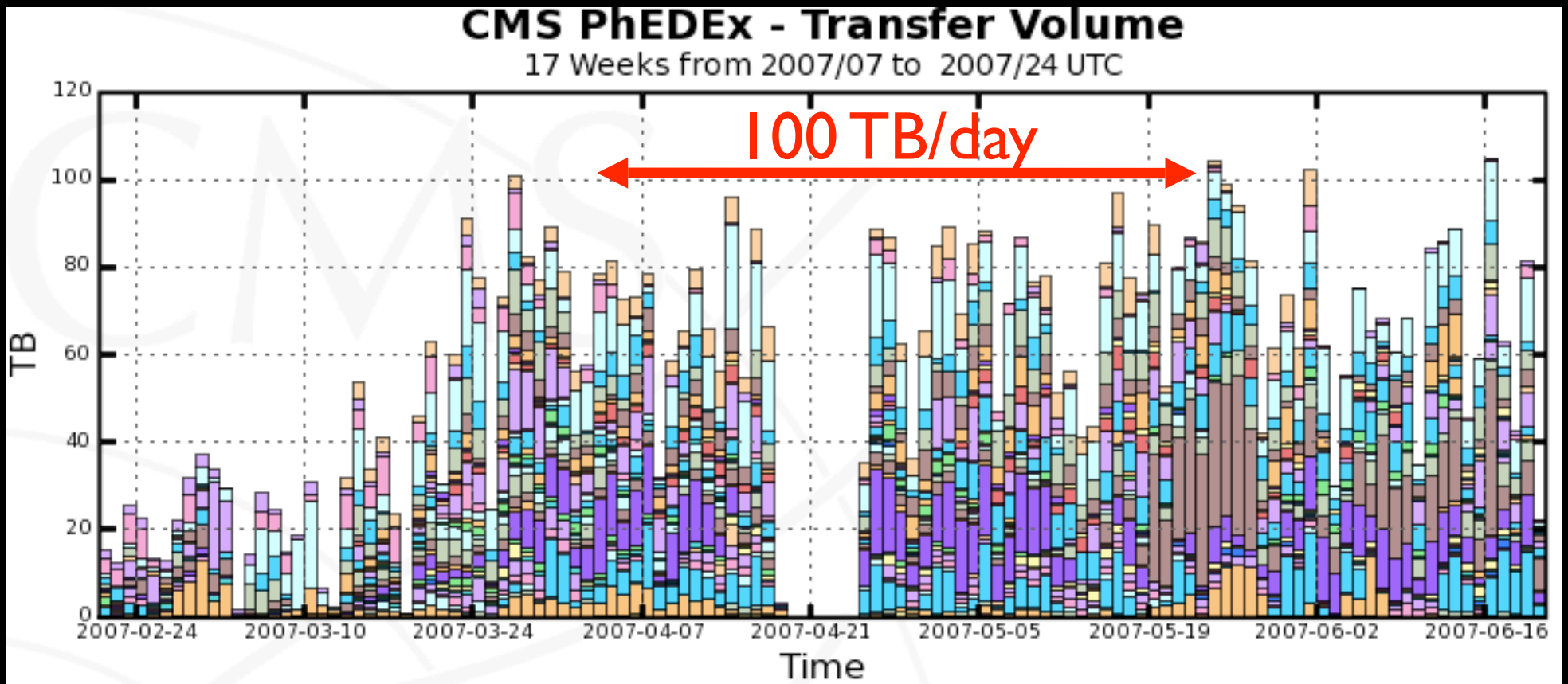
Tier-2's:
- User data analysis
- MC production
- Import skimmed datasets from Tier-1 and export MC data
- Calibration/Alignment

Data Distribution



- Two principle use cases- push and pull of data
  - Raw data is pushed onto the regional centers
  - Simulated and analysis data is pulled to a subscribing site
  - Actual transfers are 3rd party- handshake between active components important, not push or pull
- Maintain end-to-end multi-hop transfer state
  - Can only clean online buffers at detector when data safe at Tier-1
- Policy must be used to resolve these two use cases

Setting the Scale



CMS routinely moves up to 100TB of data a day across its Data Grid of more than 50 sites worldwide.

Summary

- LHC will provide access to conditions not seen since the early Universe
  - Analysis of LHC data has potential to change how we view the world
  - Substantial computing and sociological challenges
- The LHC will generate data on a scale not seen anywhere before
  - Rapid deployment and growth of IT infrastructure across more than 50 institutions in 25 countries
  - LHC experiments will critically depend on parallel solutions to analyze their enormous amounts of data
- A lot of sophisticated data management tools have been developed

Exciting times ahead!