ALICE O^2 Project

Pierre VANDE VYVRE for the O^2 project

07-Oct-2013 – Bangkok, Thailand

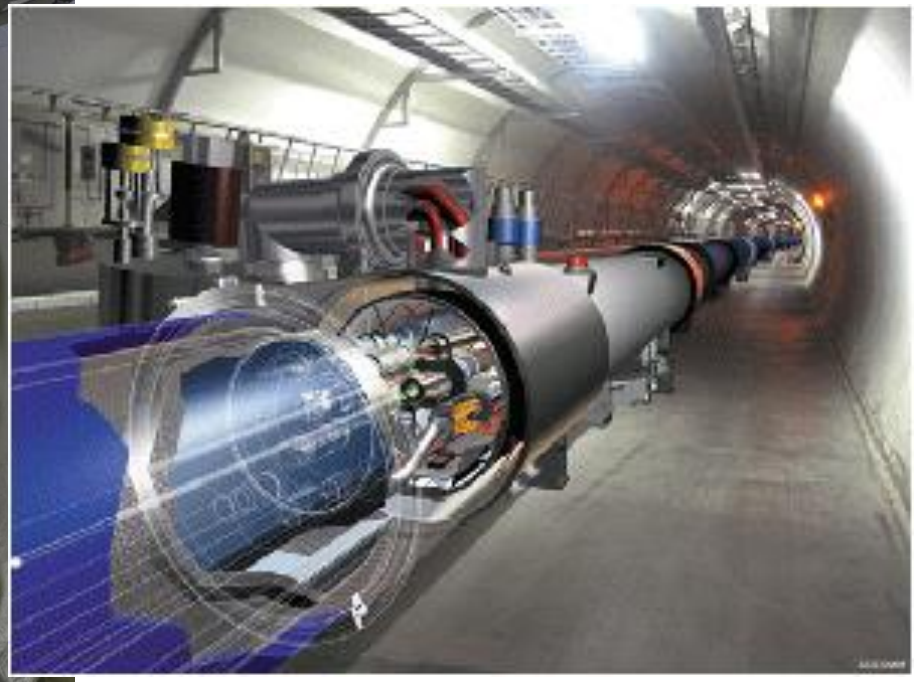
Outline

- High Energy Physics (HEP)
  - Large Hadron Collider (LHC)
  - ALICE experiment
  - Data selection and acquisition
- ALICE experiment upgrade

- O^2 project
  - Requirements
  - Big data
  - Computing Working Groups
  - Next steps

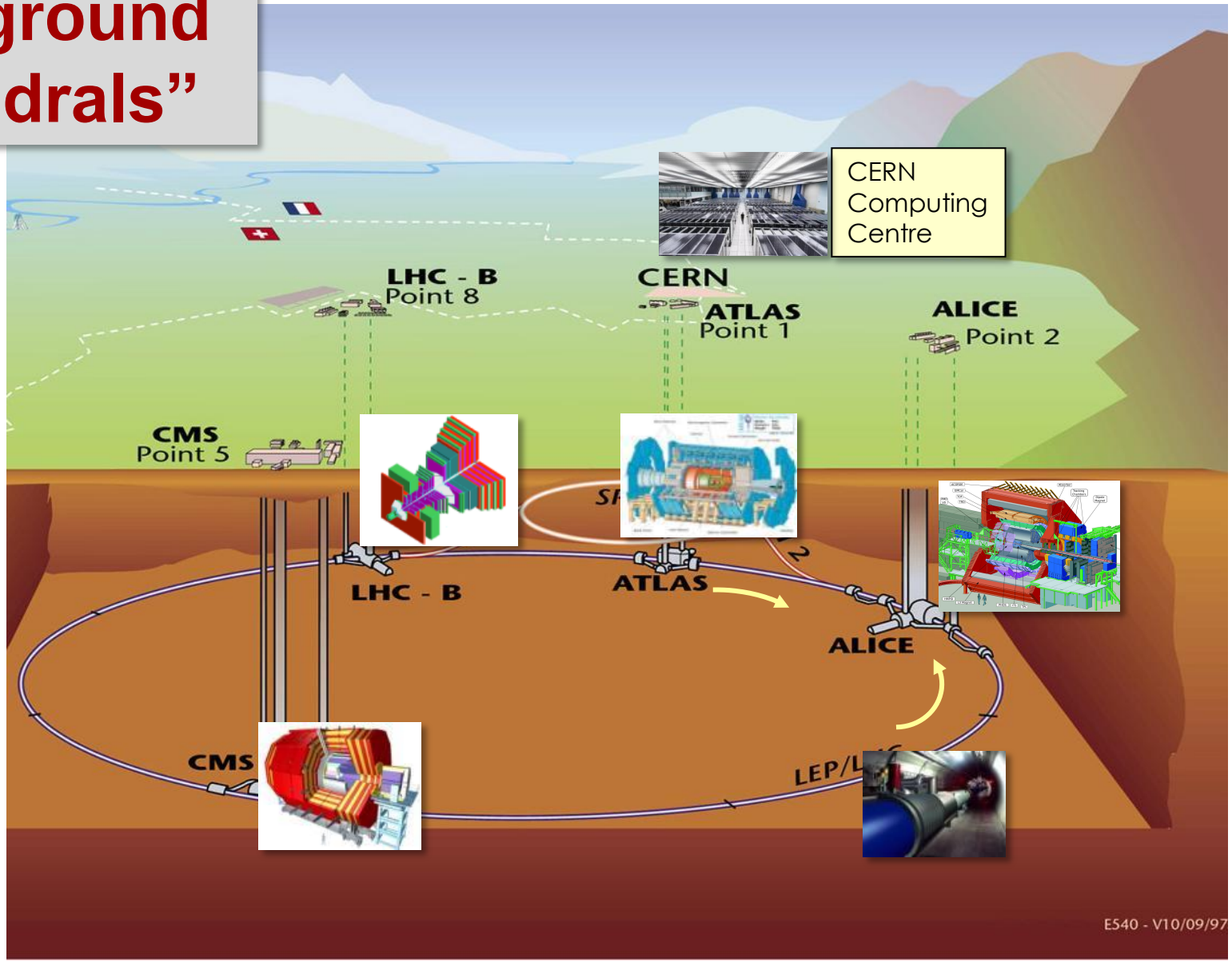
Large Hadron Collider (LHC) at CERN

Inside the LHC machine

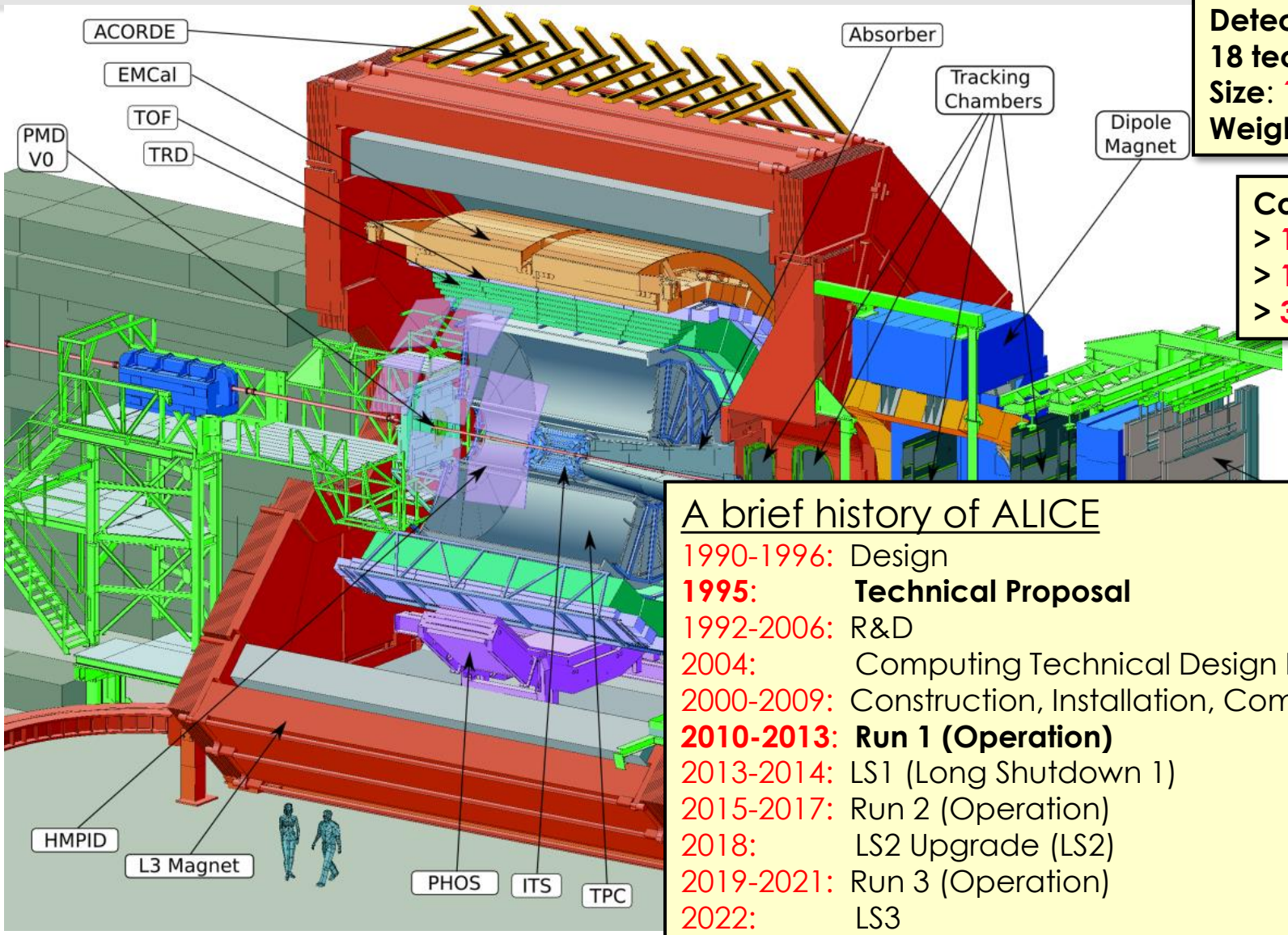


Experiments

Underground "cathedrals"



ALICE Experiment

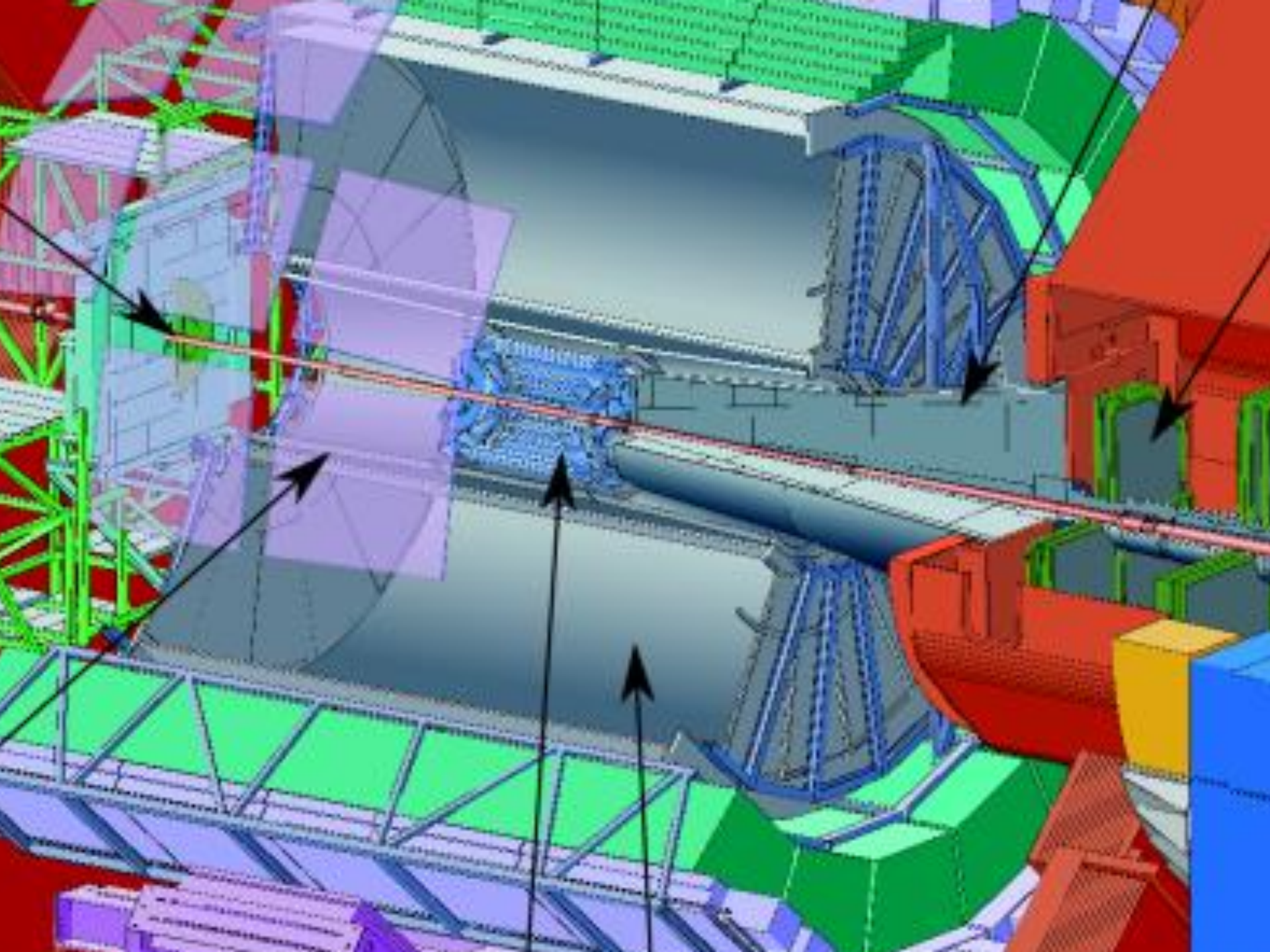


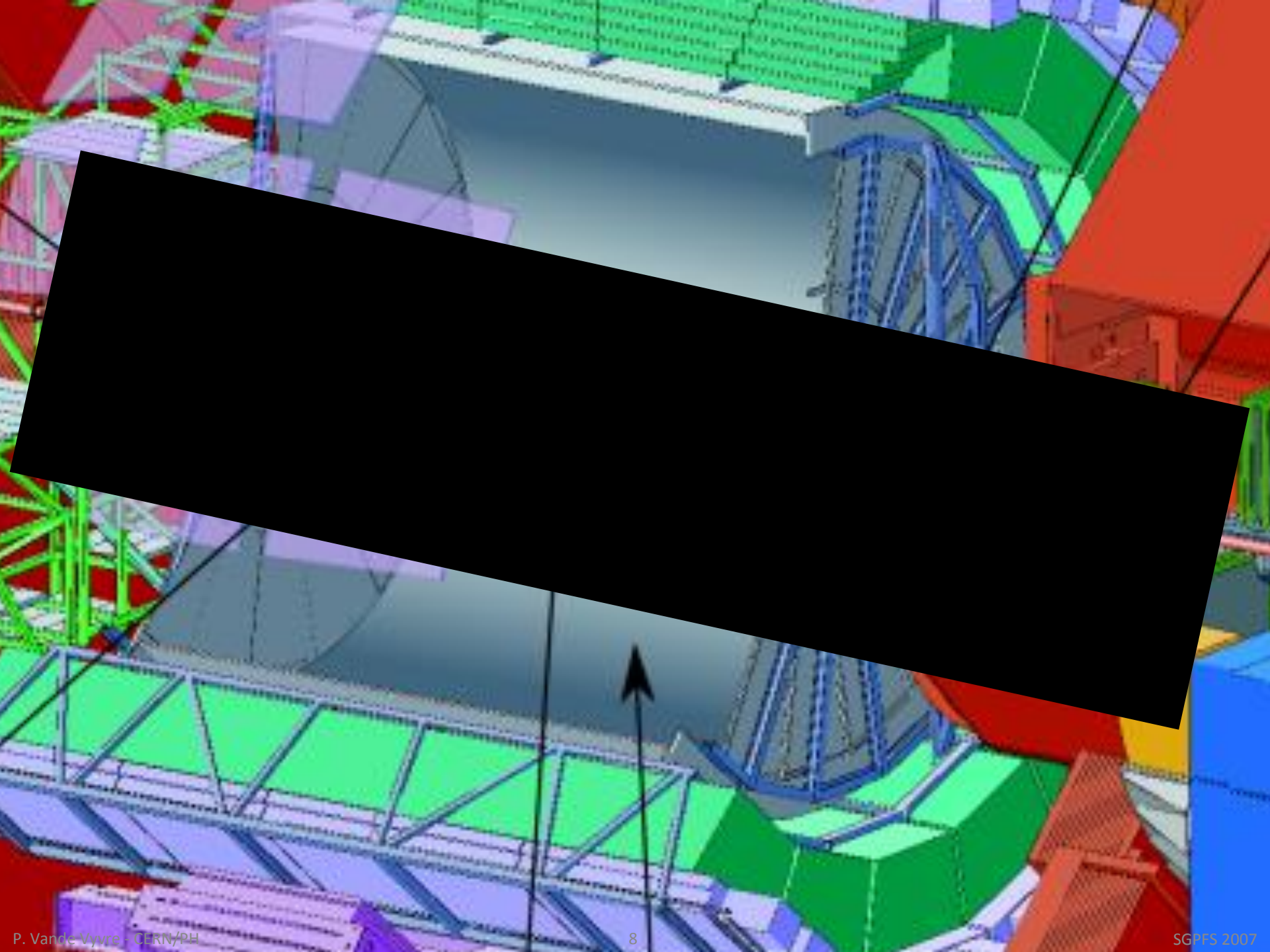
Detector:
18 technologies
Size: **16 x 26** meters
Weight: **10,000** tons

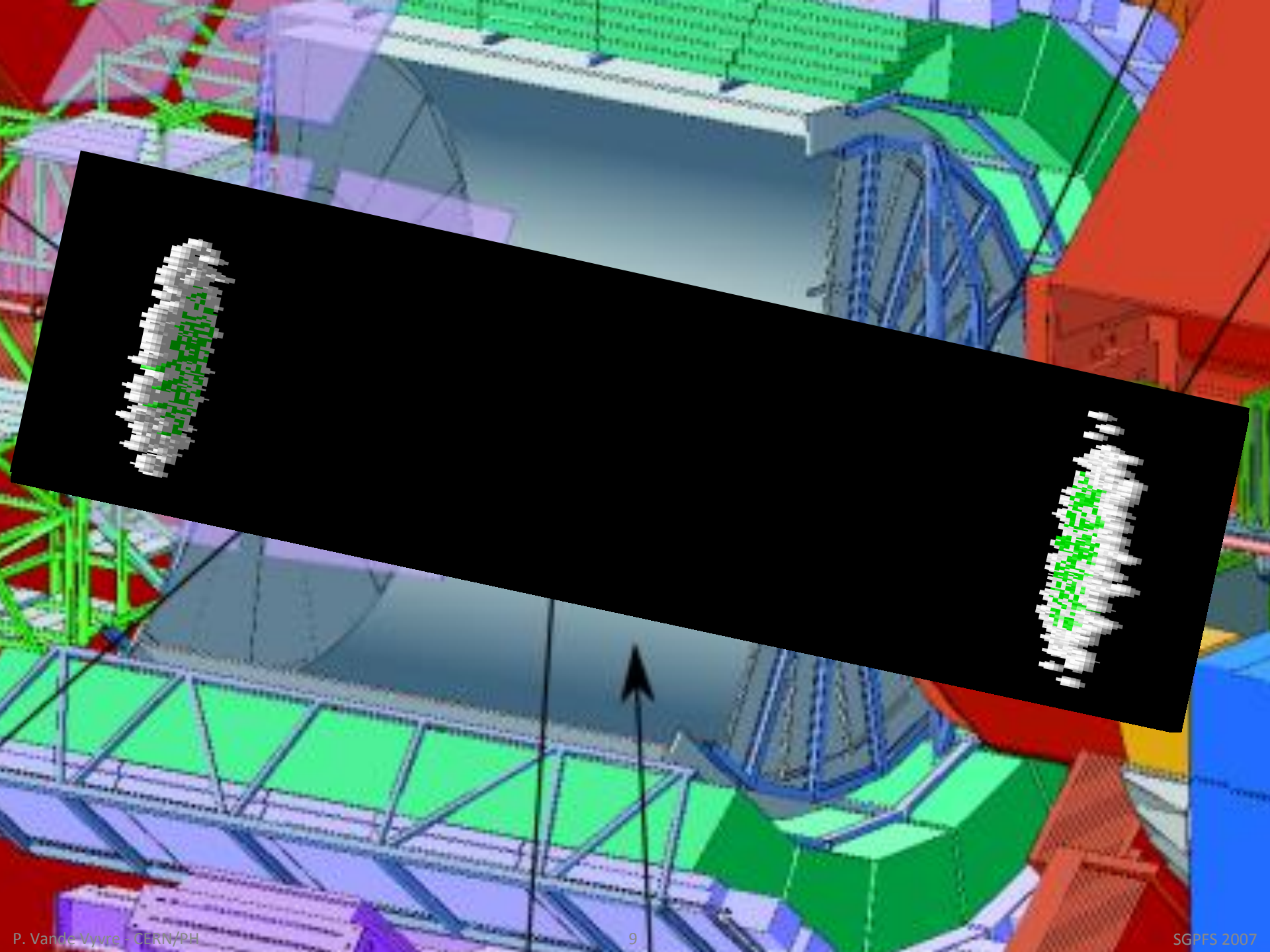
Collaboration:
> 1000 Members
> 100 Institutes
> 30 countries

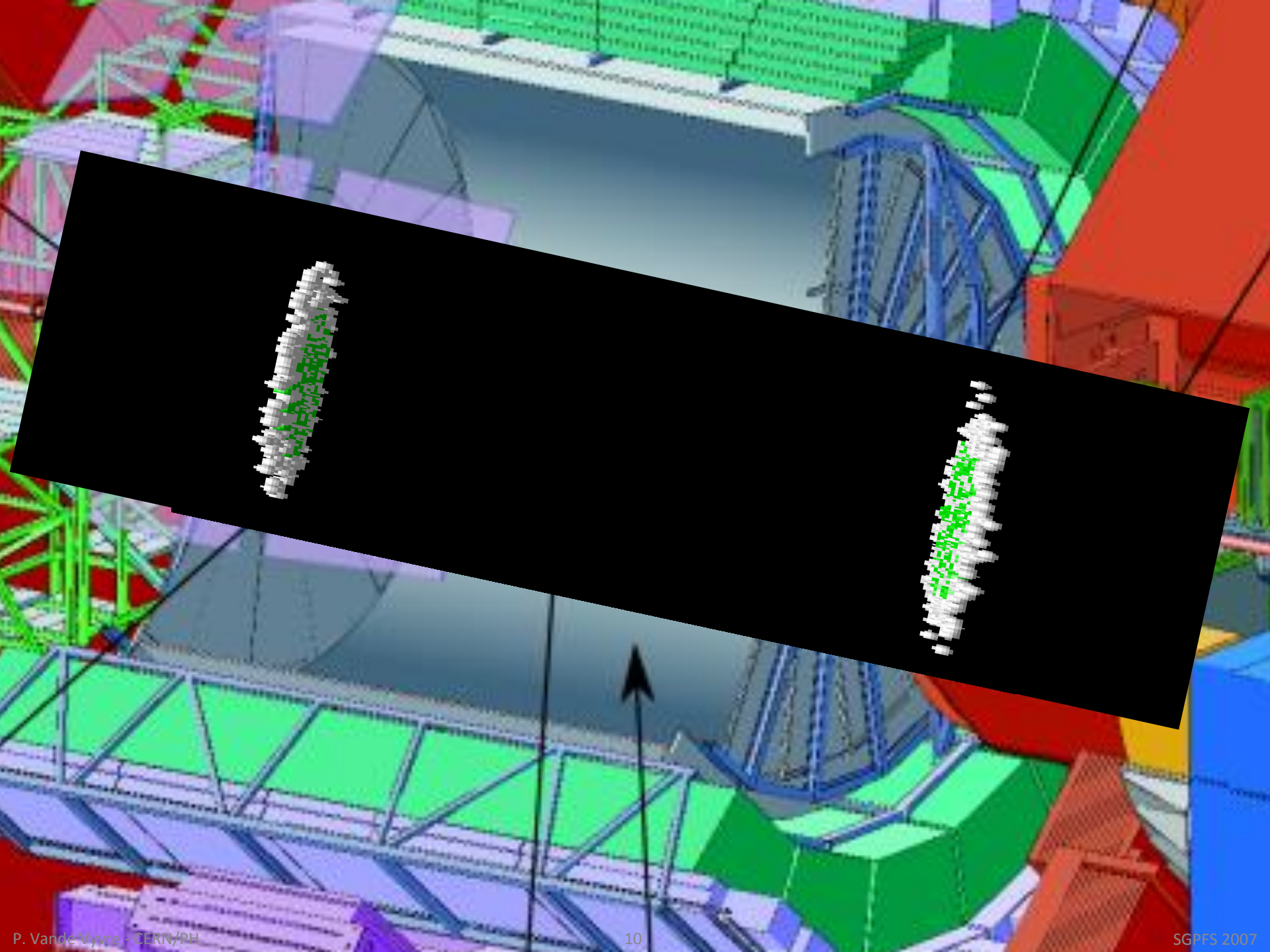
A brief history of ALICE

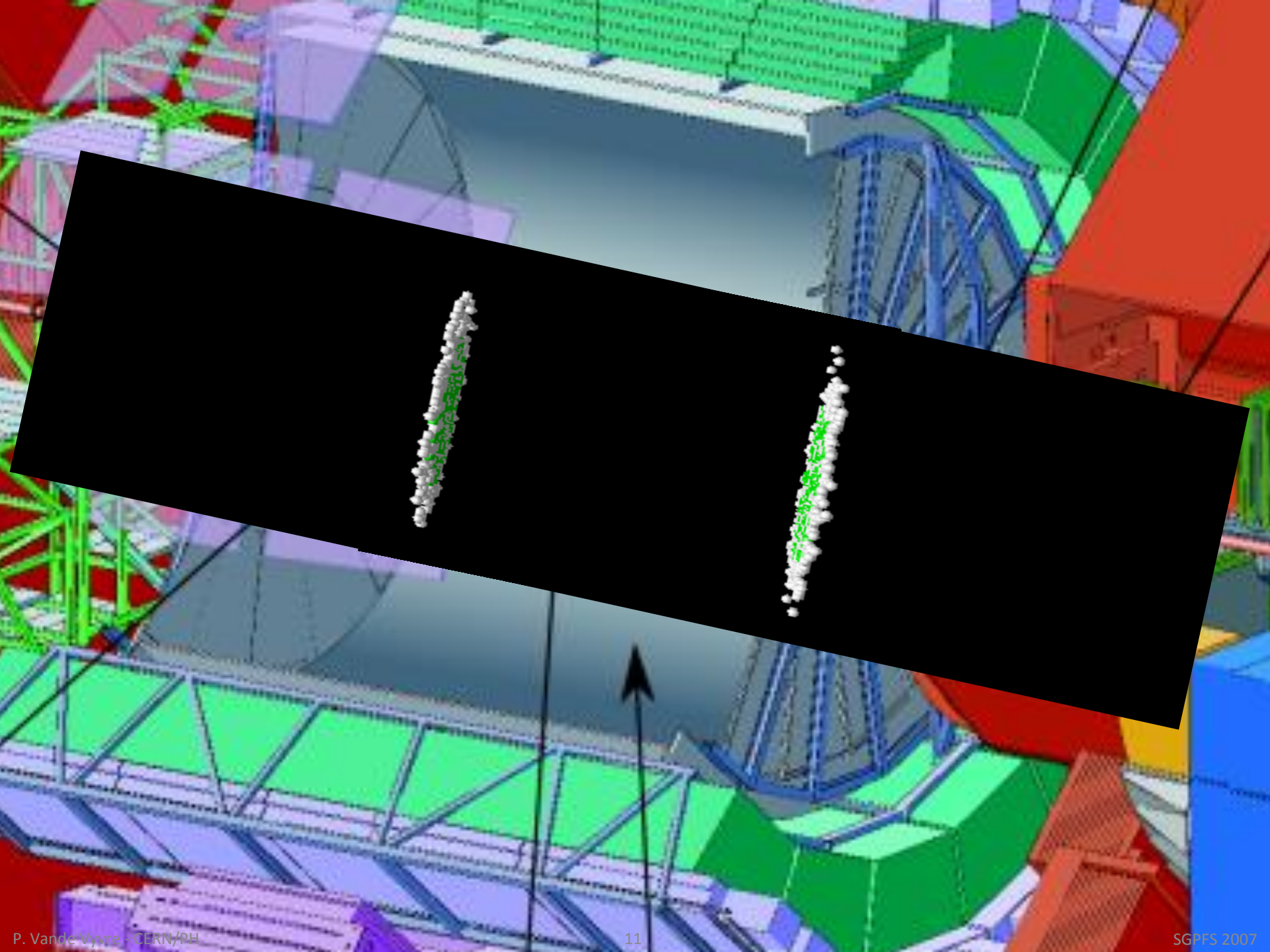
1990-1996: Design
1995: **Technical Proposal**
1992-2006: R&D
2004: Computing Technical Design Report
2000-2009: Construction, Installation, Commissioning
2010-2013: **Run 1 (Operation)**
2013-2014: LS1 (Long Shutdown 1)
2015-2017: Run 2 (Operation)
2018: LS2 Upgrade (LS2)
2019-2021: Run 3 (Operation)
2022: LS3
2023-**2025**: Run 4 (Operation)
Project lifetime of 35 years

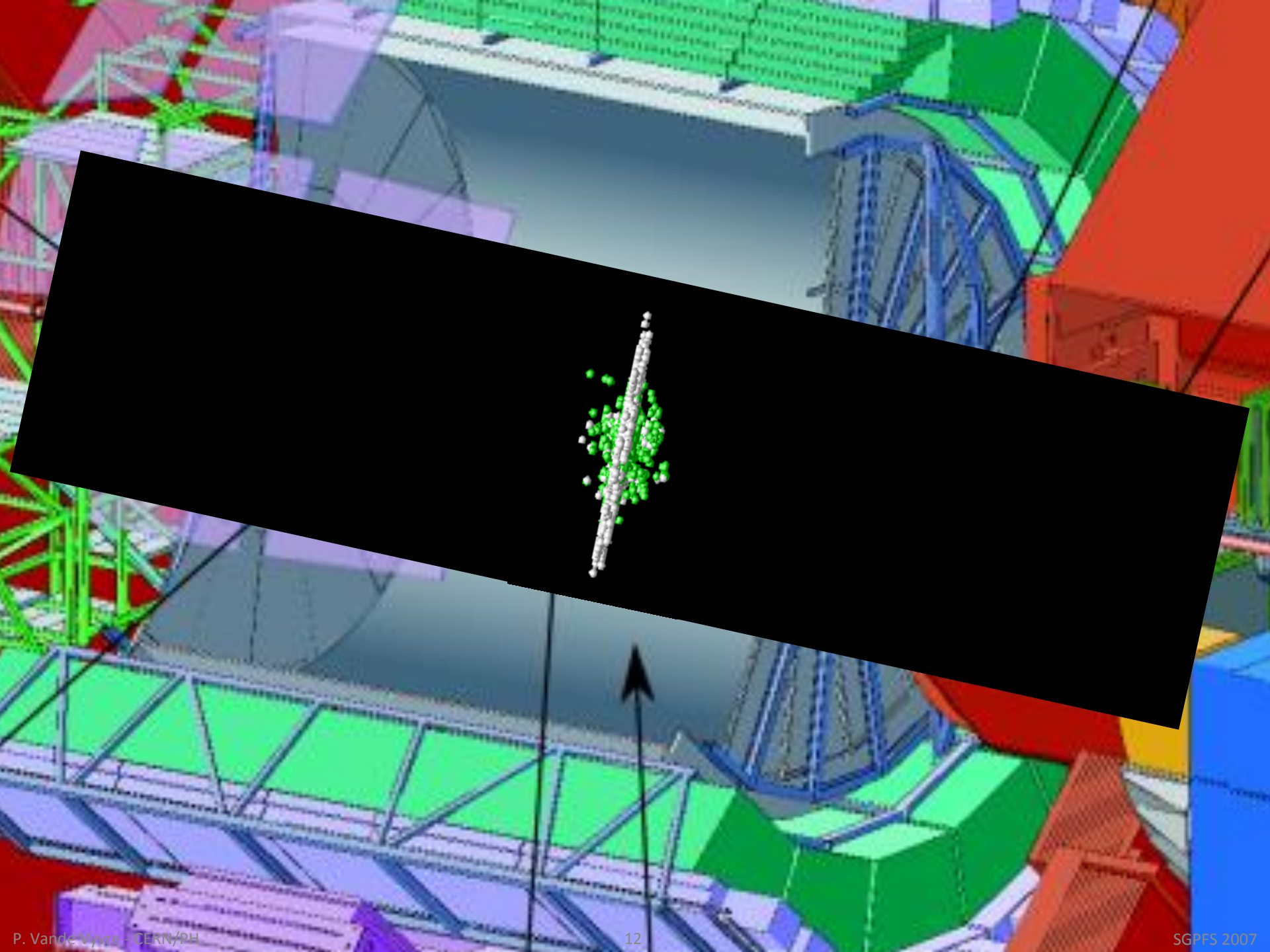


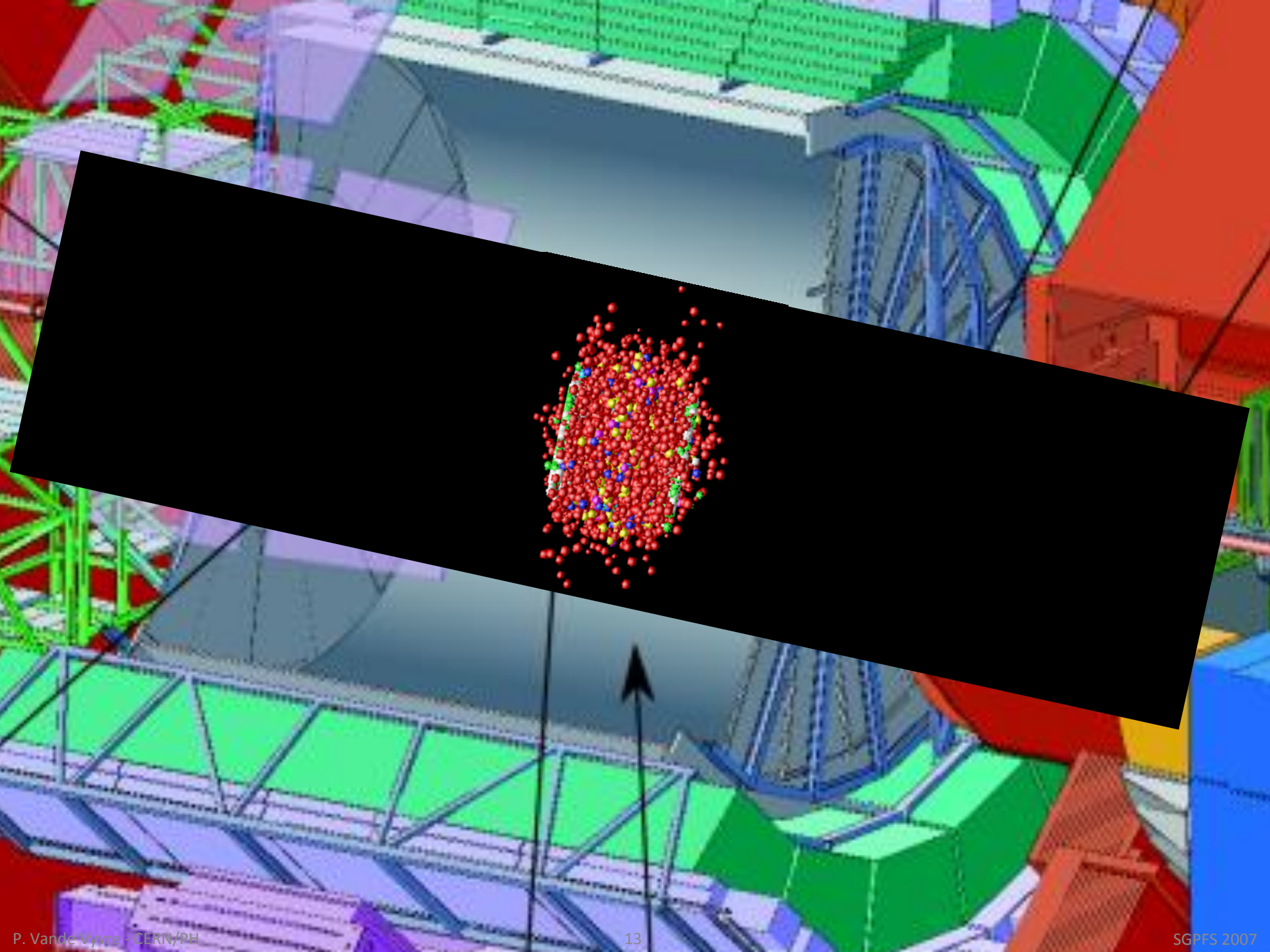


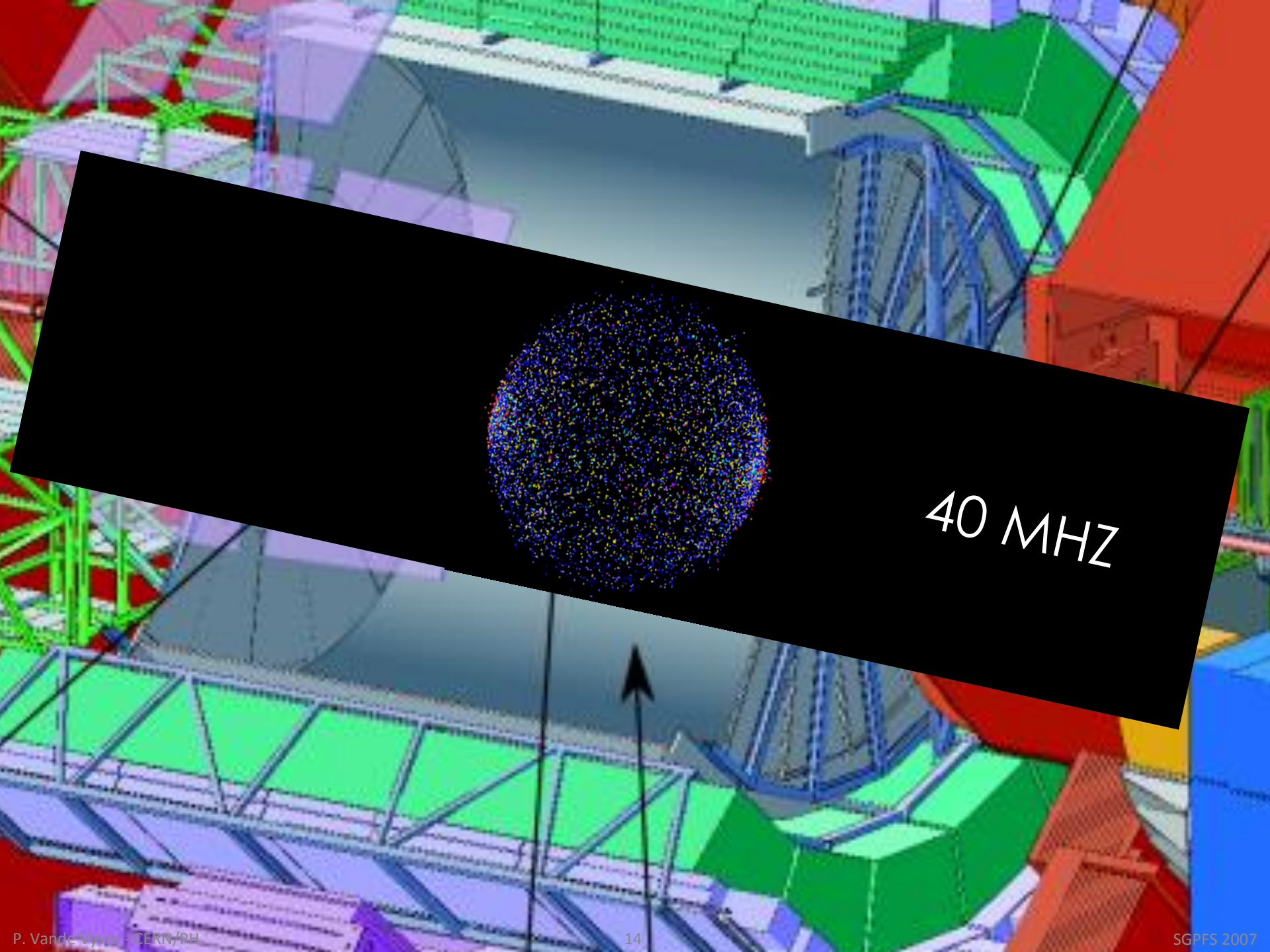




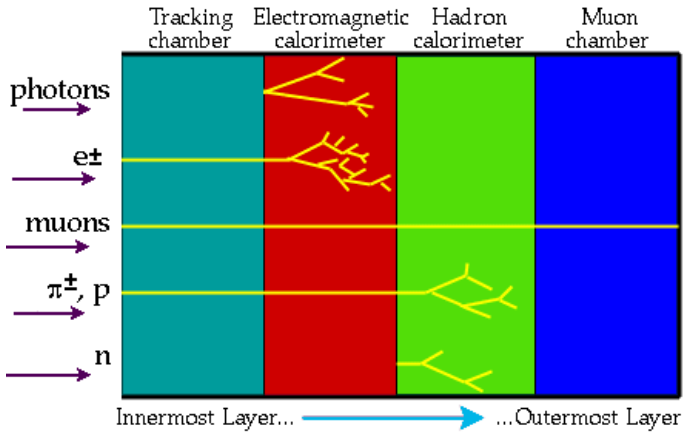




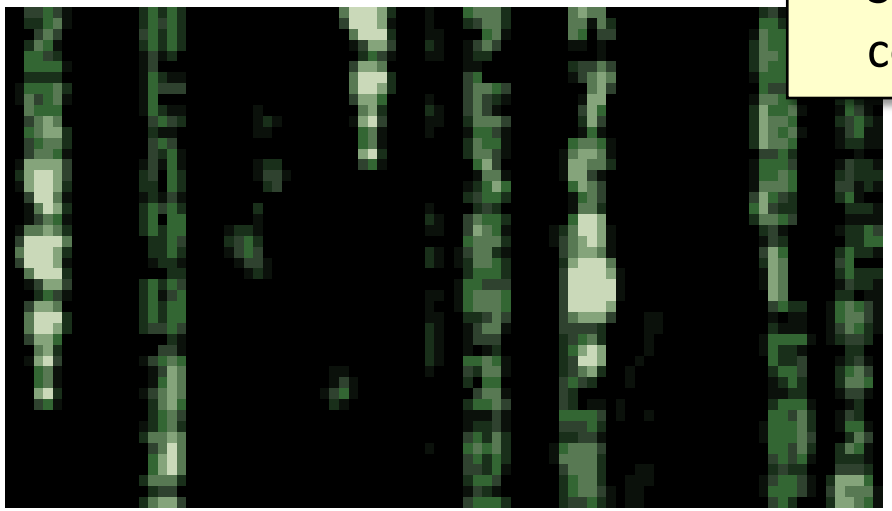


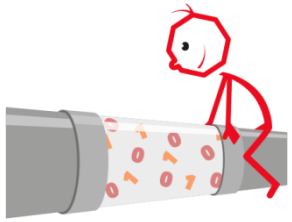

40 MHZ

Data Deluge

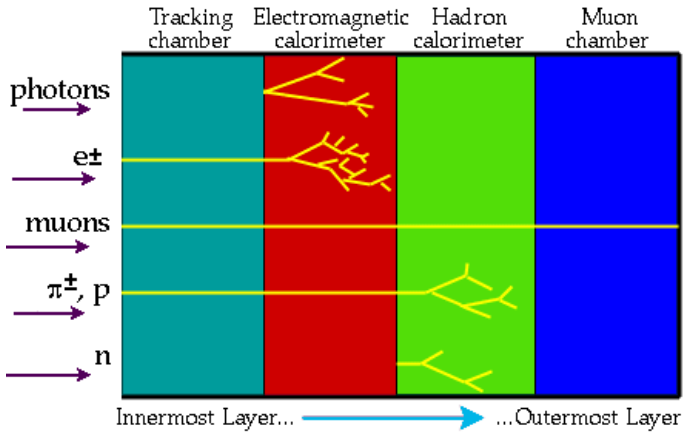


- Beam crossing at 40 MHz
  600 million potential collisions/second
- 100-10'000 "interesting" events/second
- 10-500 million measurement channels
- 1-50 Mbytes of data
  after first level of data compression
- Severe selection is mandatory to keep the computing costs to a reasonable level



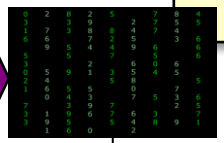


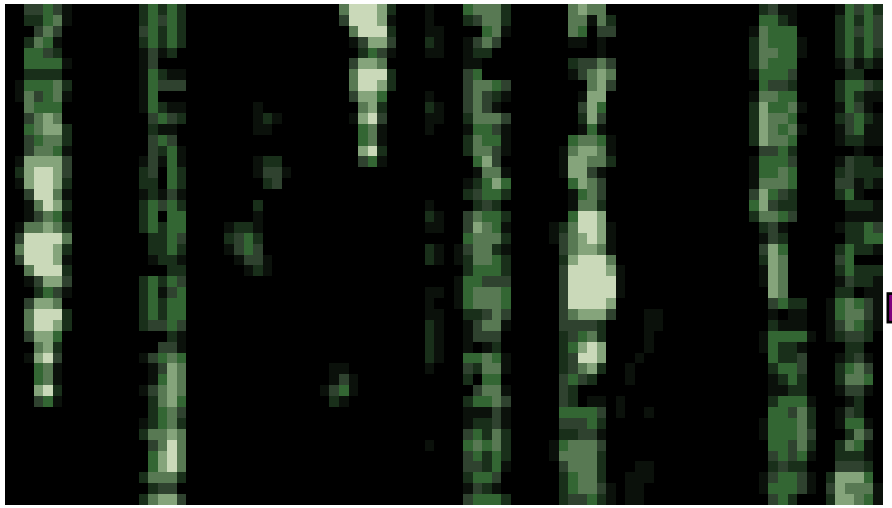
Data Selection

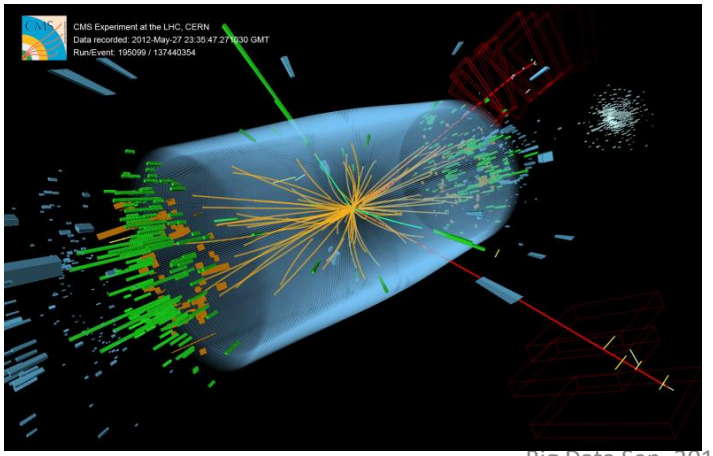


Multi-level trigger system (40 MHz → a few kHz)

- Reject background
- Select most interesting interactions
- Custom computer to reduce the total data volume

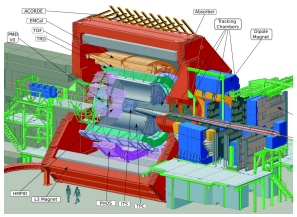


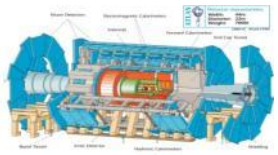


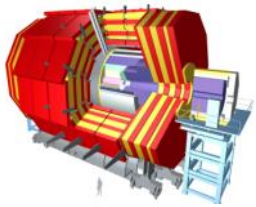


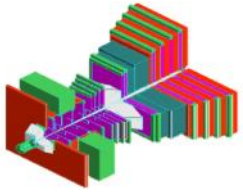
Lots of data anyway !

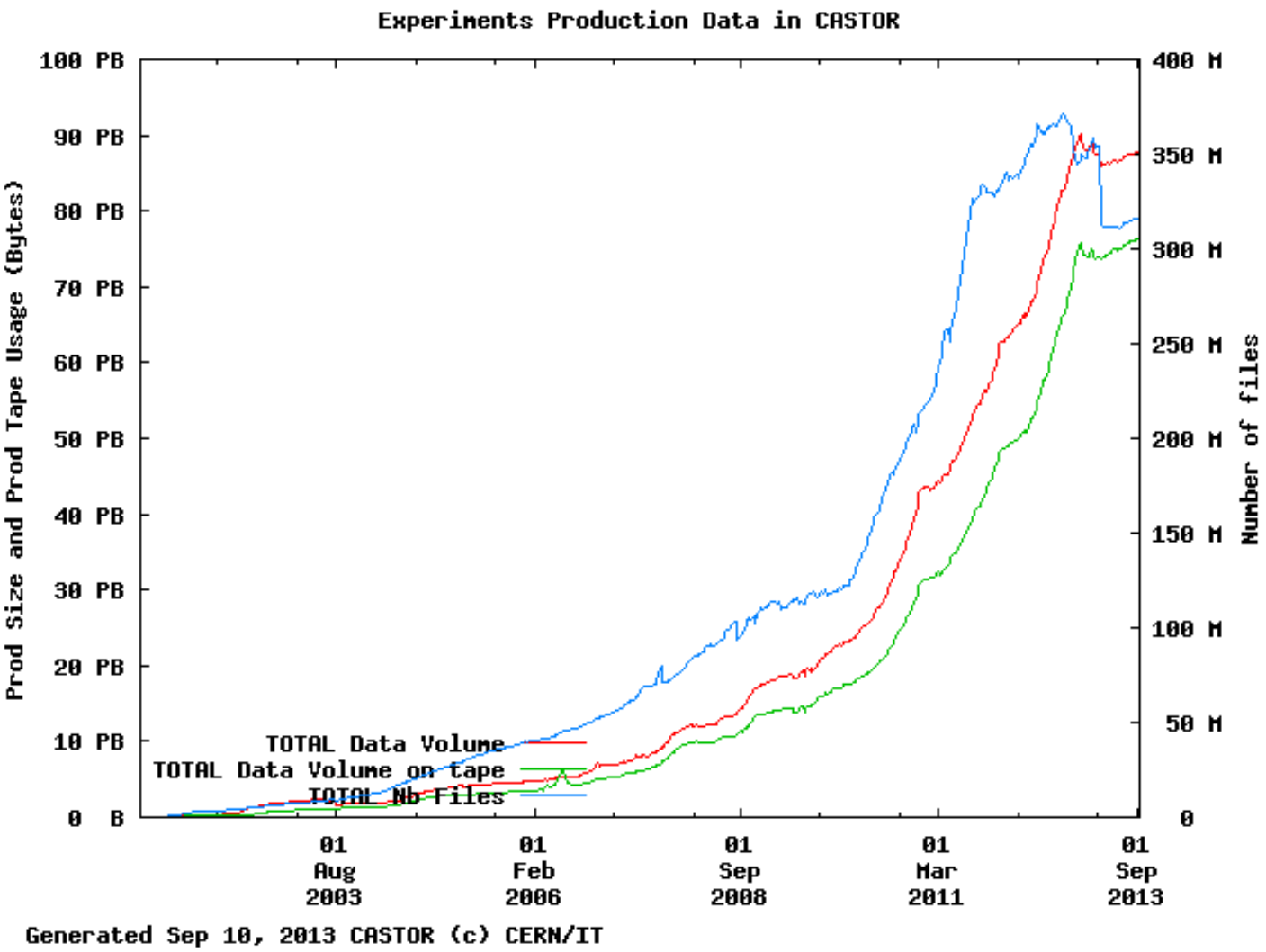
	Beam Type	Recording (Mass Storage)	Data Archived
ALICE			
ATLAS			
CMS			
LHCb			



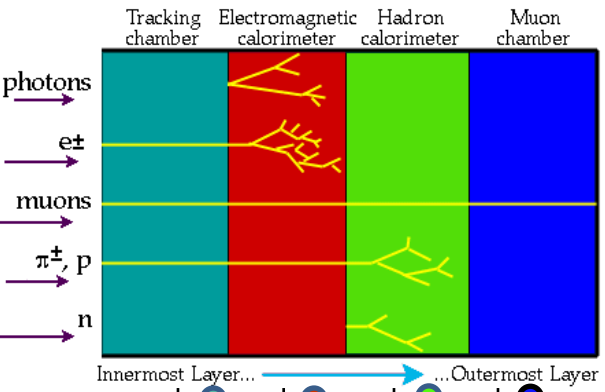




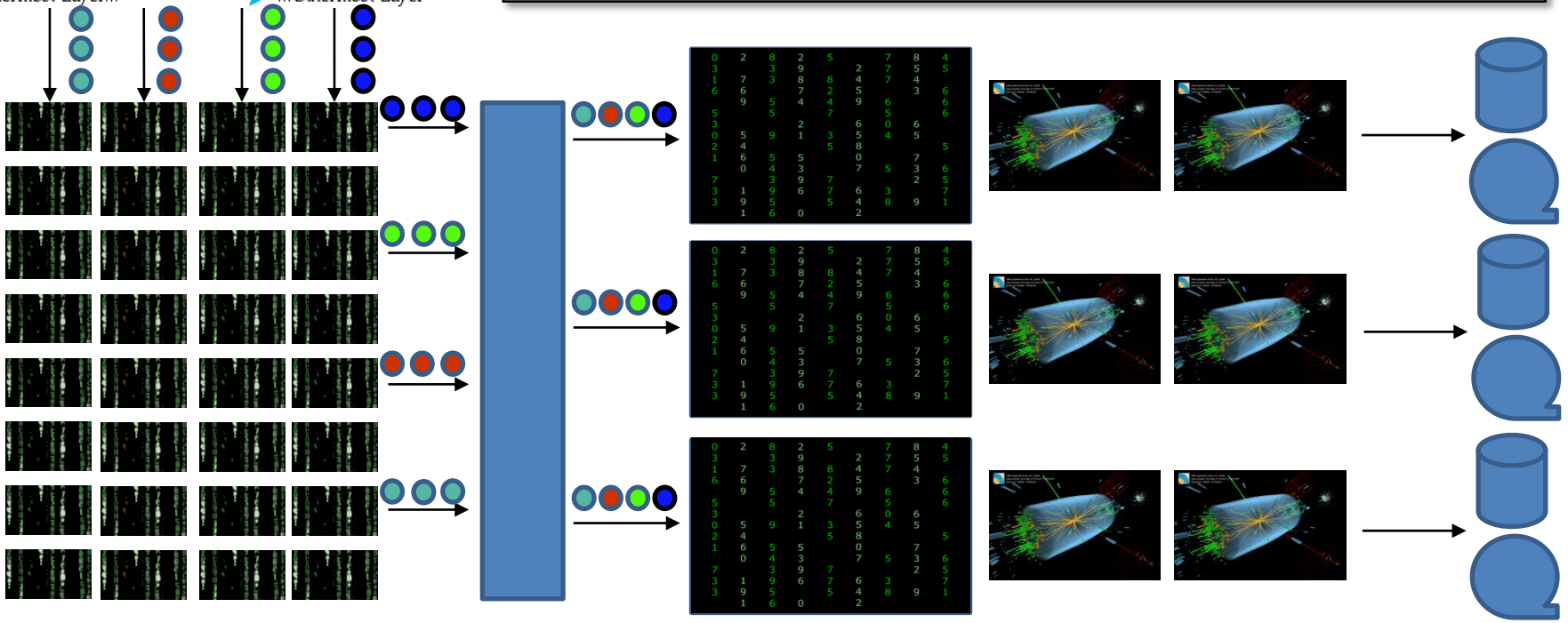




Data Acquisition Design Concepts

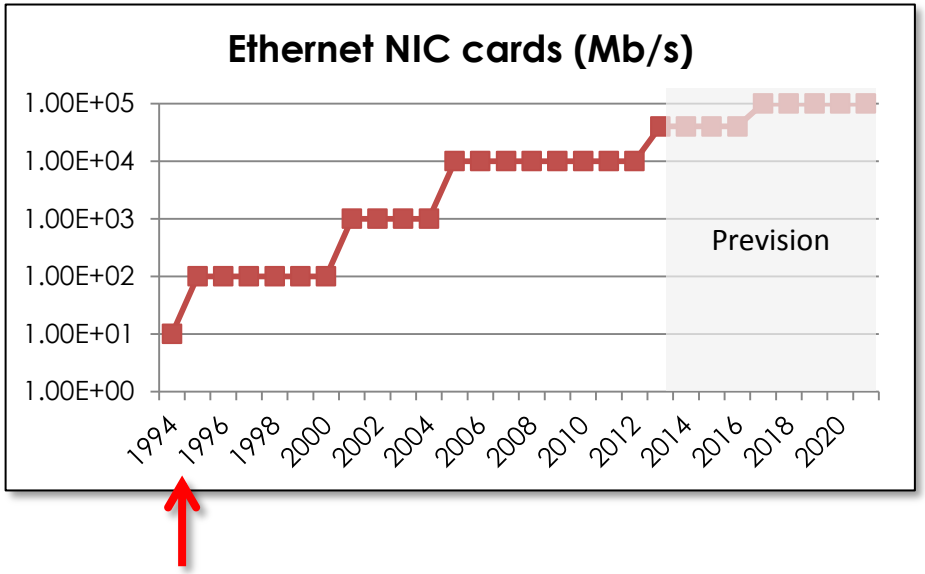


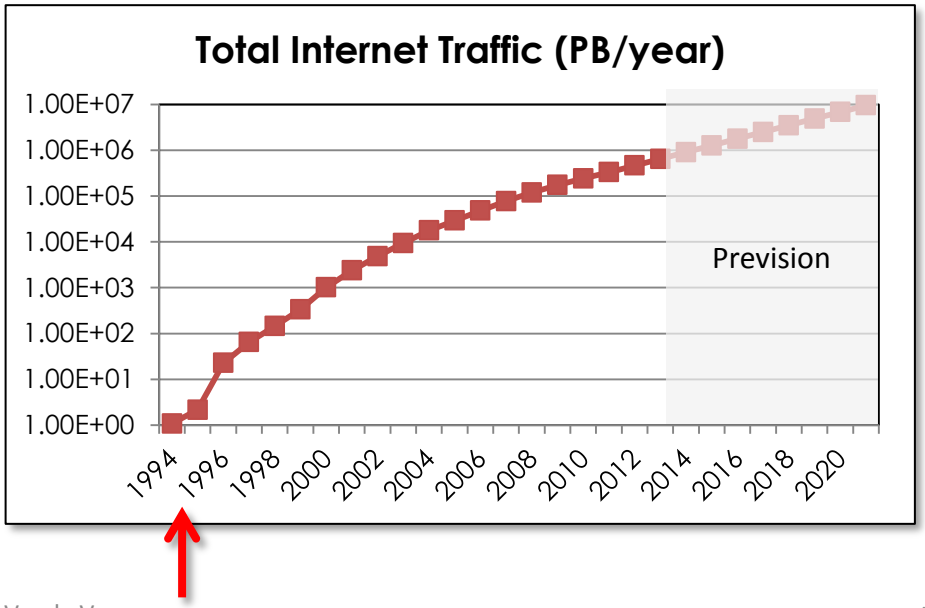
- Acquire data of tens of millions of channels
- Store them in a matrix of hundreds of memories
- Multiplex to a computer farm
- Assemble and store the data pertaining to the same particle collision



Memory matrix | Multiplexer | Computer Farm | Complete events

Challenge #1: Data Collection

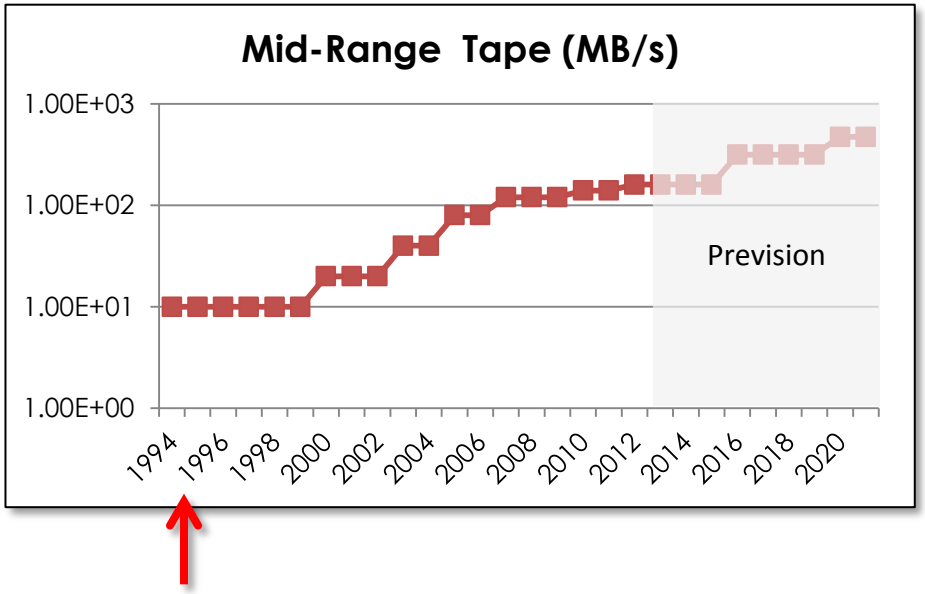


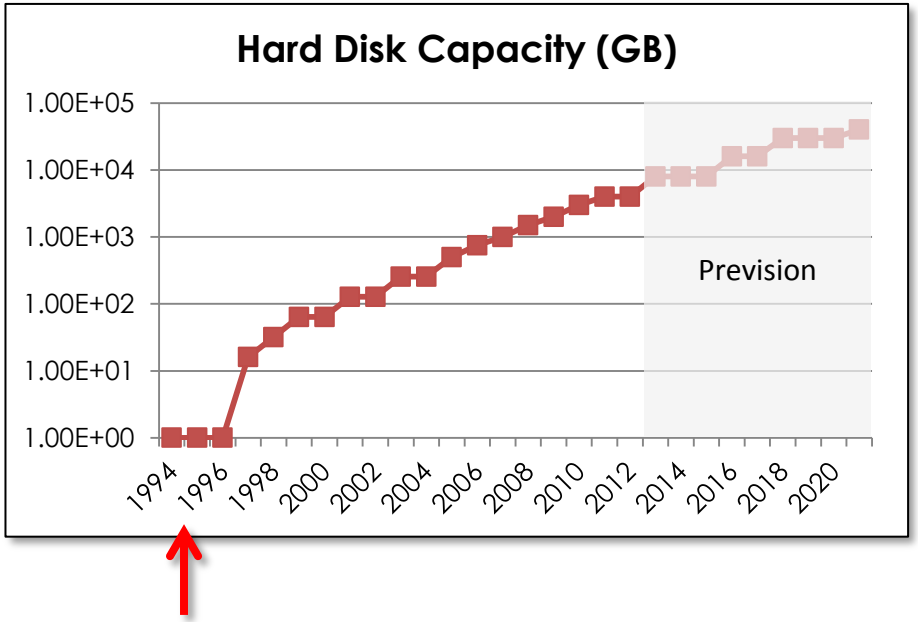


- Memories: initially custom and now using PC's central memory
- In 1995 (ALICE 1 GB/s, 1PB/yr)
- Multiplexing: many data sources and data destinations
- Big issue with ad-hoc projects during the R&D phase (1990-96)
- Fast switched Ethernet products delivered in 1995
- Bandwidth increased and prices dropped thanks to the huge market triggered by the explosion of Internet

- Data memories, transport and multiplexing based on commodity products

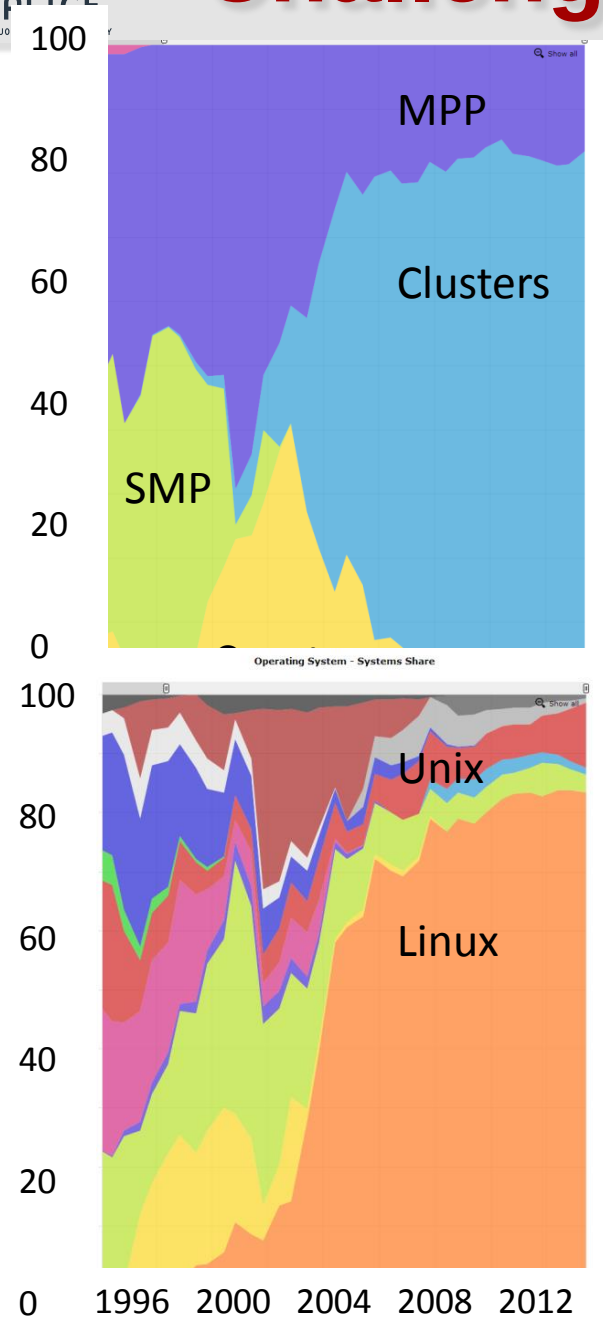
Challenge #2: Data Storage





- HEP has traditionally relied on tape for data storage
- In 1995 (ALICE 1 GB/s)
  - 1 GB/s → 100 tapes drives
  - 1 GB/ set as a reasonable limit
- Since then
  - Tape devices have slowly improved
  - Cheap disk capacity has exploded sustained by the PC market
  - Large disk storage bandwidth obtained by parallelism (RAID)
- Today
  - The yearly dataset on disk
  - Data storage using commodity products

Challenge #3: Data Processing



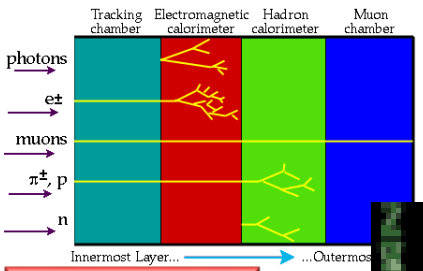
- Massive use of the highest performance computing facilities available to extract the physics from the data.
- 60-80's : **supercomputers** (CDC, Cray and IBM).
- End of the 80's : use of the inexpensive **micro-processors and PC farms** improving price/performance by an order of magnitude compared with the supercomputers.
- De-facto standard : clusters of Linux PC's

Top 500

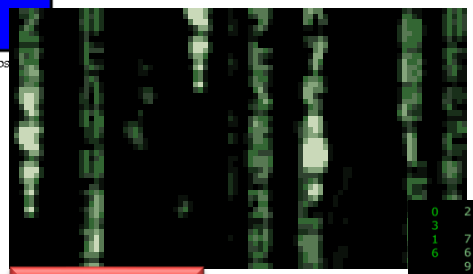
- Top: fraction of systems per architecture
- Bottom: fraction of systems per OS

- In the 2000's: insufficient funding for the computing at CERN required for the LHC experiments.
- No hardware breakthrough on the horizon.
- Experiments: scientists from very many institutes spread around the world, many with their own computing capabilities.
- Challenge: integrate of these diverse facilities to provide a coherent computing service available through the **Data Grid**.

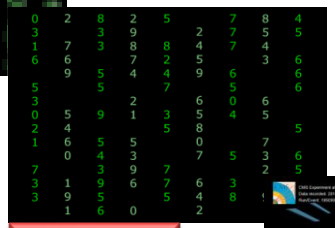
From Particles to Articles



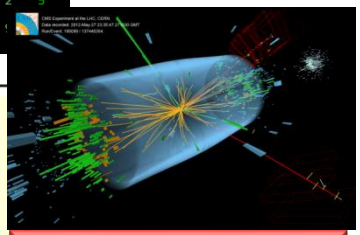
Particles



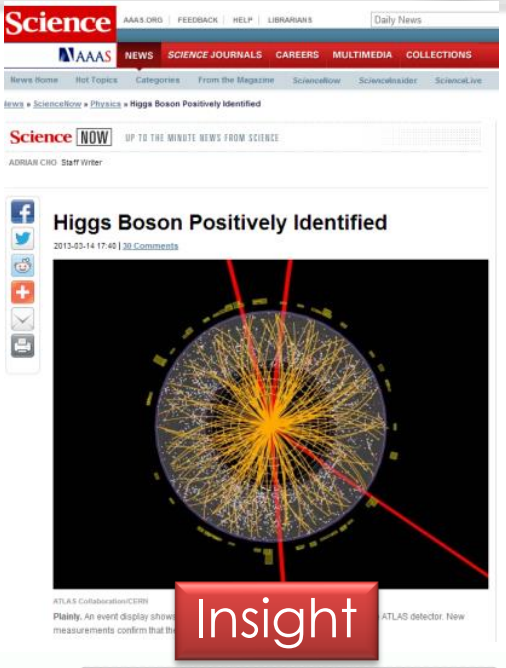
Signal



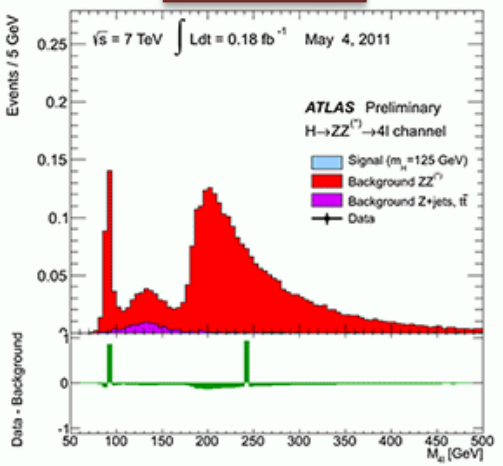
Data



Information



Insight

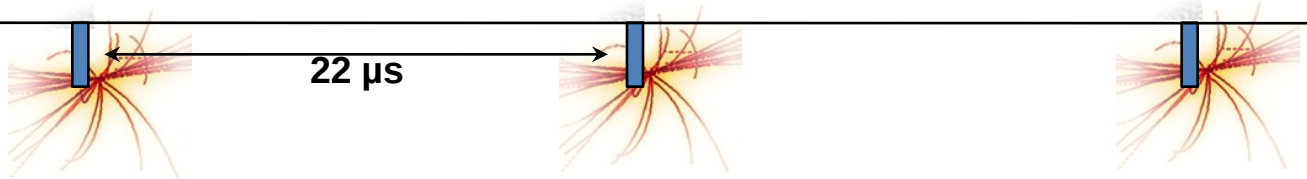


Knowledge

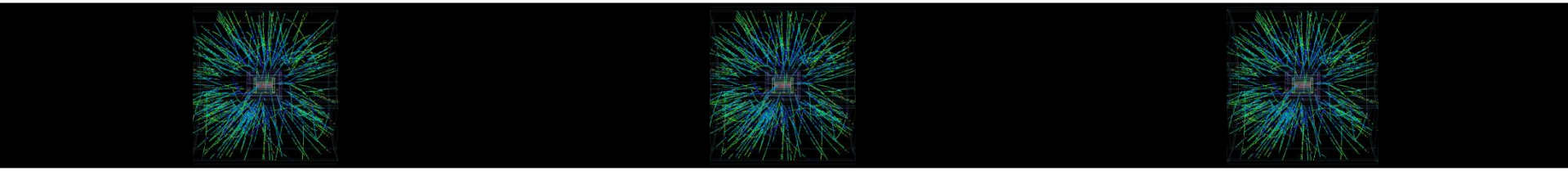
- Inspect all collisions.
- Select the most interesting ones.
- Store the related data and make them available for analysis on the Data Grid.
- Challenges solved thanks to commodity products from the computing industry and to a distributed approach for data processing.
- HEP extracting Insight from Big Data

Increase of LHC luminosity

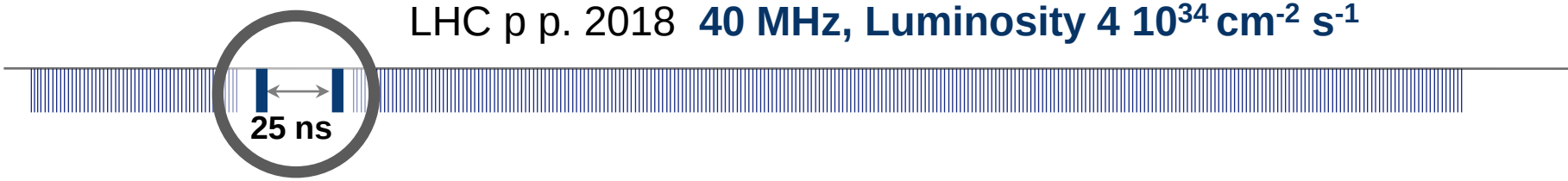
LEP e^- e^+ crossing rate **45 kHz, Luminosity 7 10^{31} cm^{-2} s^{-1}**

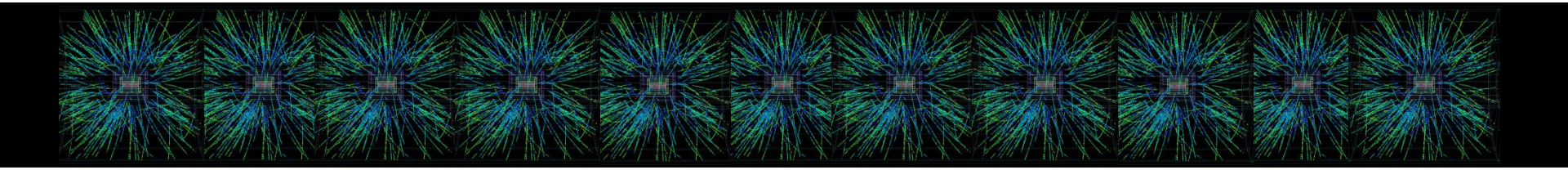


LHC p p. now **40 MHz, Luminosity 7 10^{33} cm^{-2} s^{-1}**



LHC p p. 2018 **40 MHz, Luminosity 4 10^{34} cm^{-2} s^{-1}**

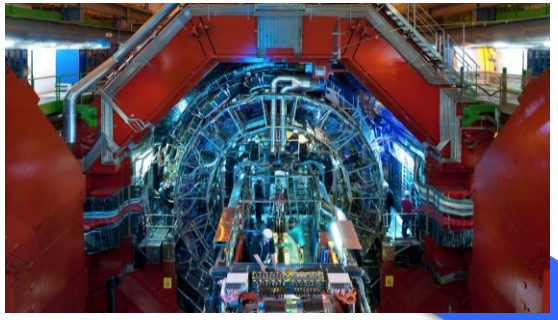




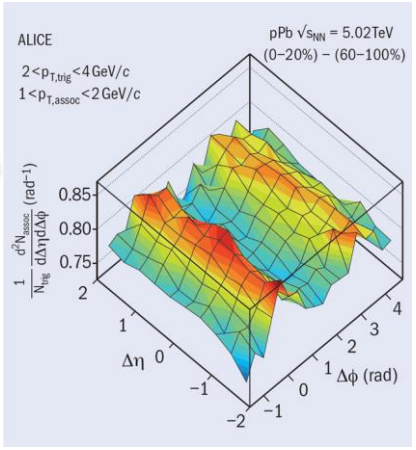
ALICE Upgrade

- Now: reducing the event rate from 40 MHz to ~1 kHz
  - Select the most interesting particle interactions
  - Reduce the data volume to a manageable size
- After 2018:
  - Much more data (x100) because
    - Higher interaction rate
    - More violent collisions → More particles → More data (1 TB/s)
    - Physics topics require measurements characterized by very small signal-over-background ratio → large statistics
    - Large background → traditional triggering or filtering techniques very inefficient for most physics channels.
    - Read out all particle interactions (PbPb) at the anticipated interaction rate of 50~kHz.
  - No more data selection
    - Continuous detector read-out → Data less structured than in the past
    - Read-out and process all interactions with a standard computer farm ~1'500 nodes with the computing power expected by then
- Total data throughput out of the detectors: 1 TB/s

O² Project

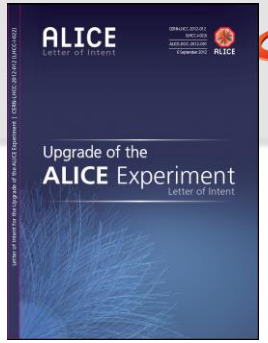


From Detector Readout to Analysis:
What is the "optimal" computing architecture?



Overall Schedule

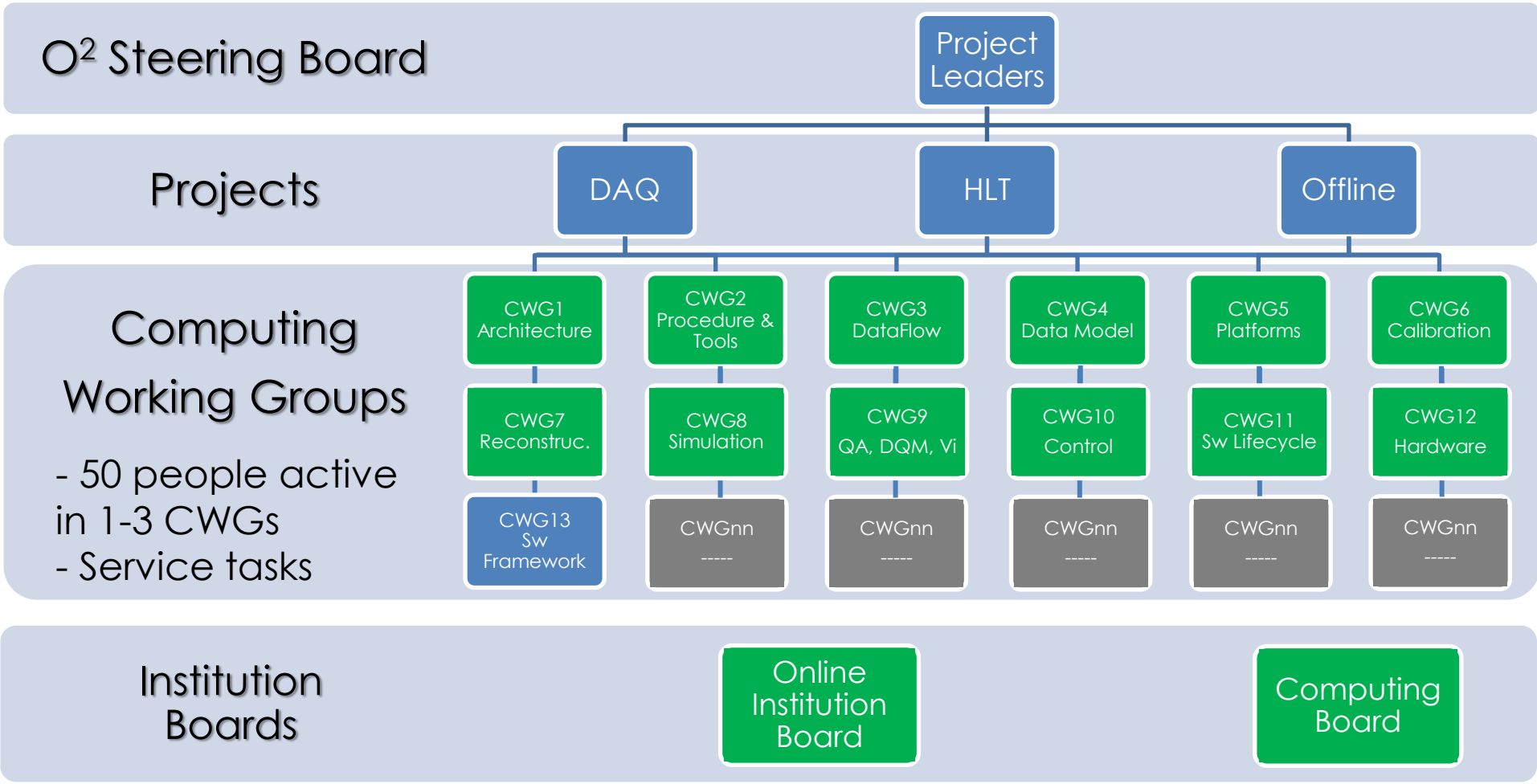
- Sep 2012 ALICE Upgrade LoI
- Jan 2013 Report of the DAQ-HLT-Offline software panel on “ALICE Computer software framework for LS2 upgrade”
- Mar 2013 O^2 Computing Working Groups
- Sep 2014 O^2 Technical Design Report



O2 Computing Working Groups

O^2 Technical Design Report

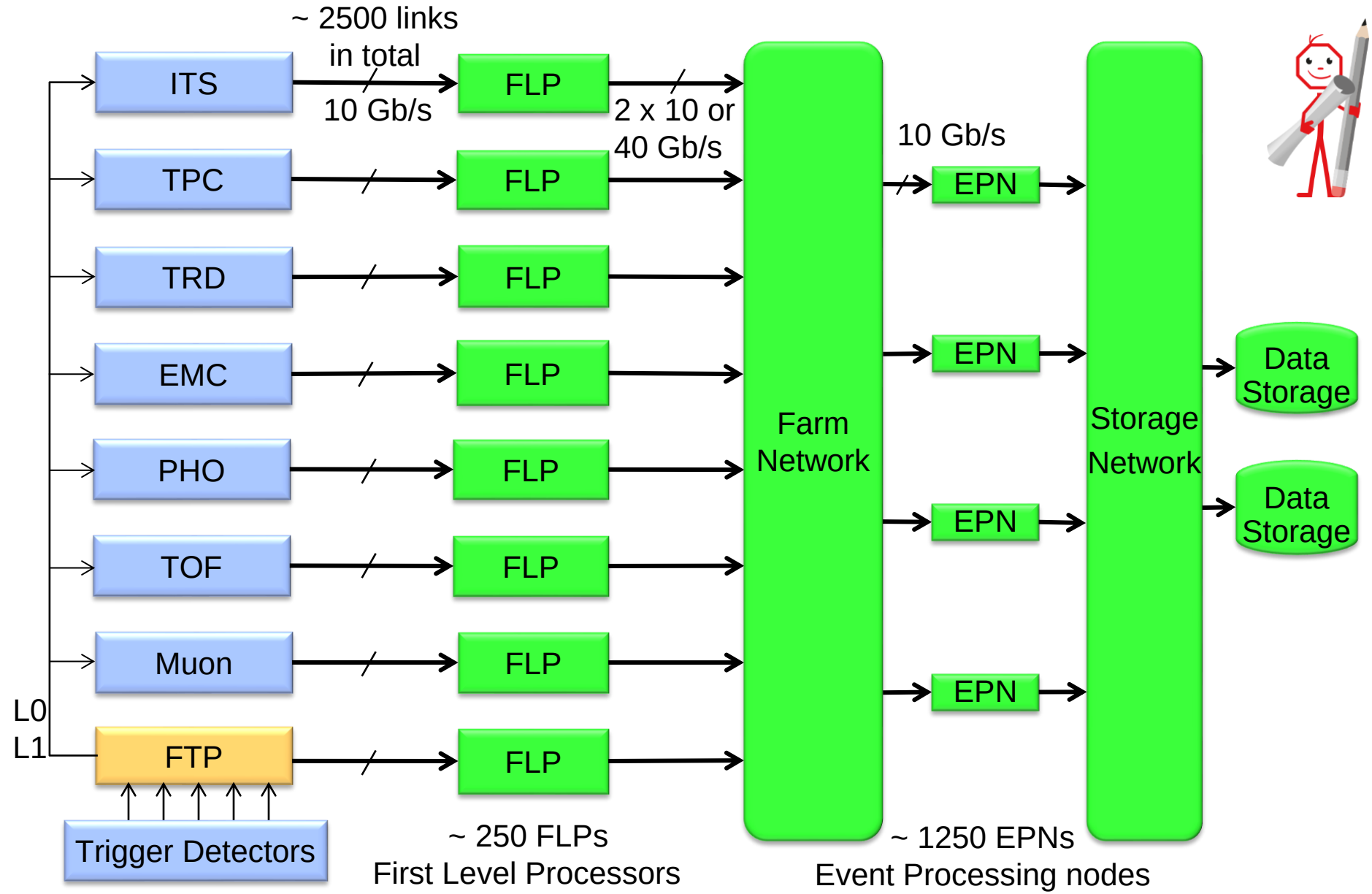
O² Project


O² Steering Board
Project Leaders
Projects
DAQ
HLT
Offline
Computing Working Groups
- 50 people active in 1-3 CWGs
- Service tasks
CWG1 Architecture
CWG2 Procedure & Tools
CWG3 DataFlow
CWG4 Data Model
CWG5 Platforms
CWG6 Calibration
CWG7 Reconstruc.
CWG8 Simulation
CWG9 QA, DQM, Vi
CWG10 Control
CWG11 Sw Lifecycle
CWG12 Hardware
CWG13 Sw Framework
CWGnn
CWGnn
CWGnn
CWGnn
CWGnn
Institution Boards
Online Institution Board
Computing Board

O² Institutes

- Institutes
  - FIAS, Frankfurt, Germany
  - IIT, Mumbay, India
  - Jammu University, Jammu, India
  - IPNO, Orsay, France
  - IRI, Frankfurt, Germany
  - Rudjer Bošković Institute, Zagreb, Croatia
  - SUP, Sao Paulo, Brasil
  - University Of Technology, Warsaw, Poland
  - Wiegner Institute, Budapest, Hungary
  - CERN, Geneva, Switzerland
- Looking for more people
  - Need people with computing skills and from detector groups
- CWG's membership is neither closed nor rigid:
  - New members more than welcome to join

O² Hardware System from LoI

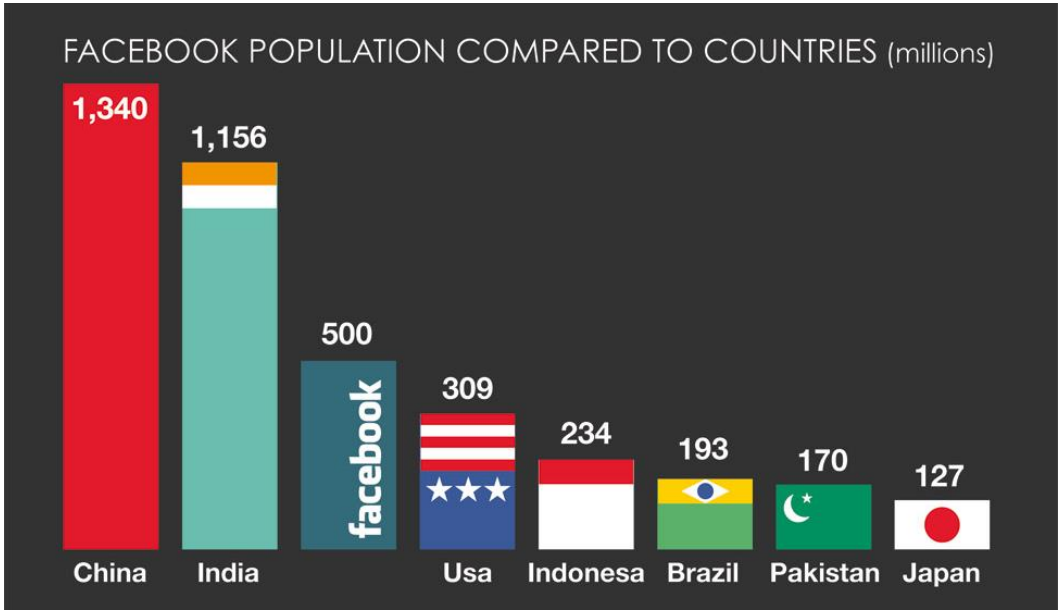


Internet : an inflating universe…

- HEP is not alone
  in the computing universe !

- 1 ZB/year in 2017 (Cisco)
- 35 ZB in 2020 (IBM)
- 1 ZB = 1'000 EB = 1'000'000 PB

- Number of users
  (Kissmetrics)



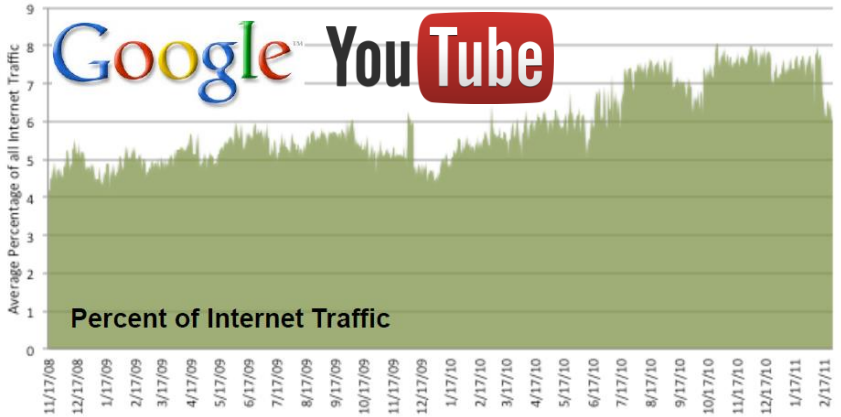
…with a few very large galaxies !

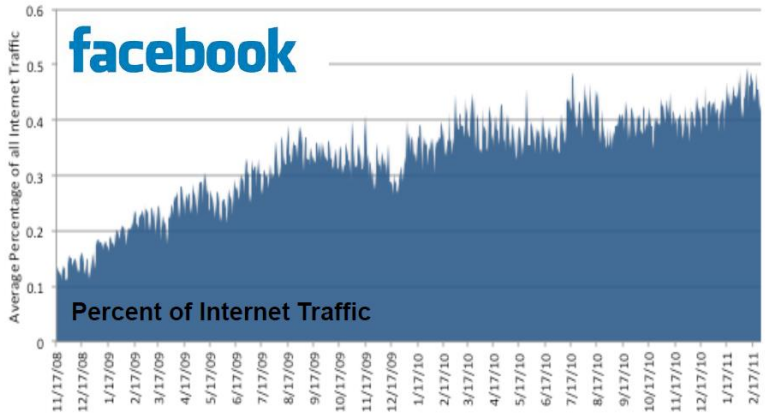
- "Hyper giants": the 150 companies that control 50% of all traffic on the web (Arbor Networks)

- Google :
  100 billion searches/month,
  38'500 searches/second
- YouTube:
  6 billion hours of video are watched each month

- Facebook
  350 millions photos uploaded/day

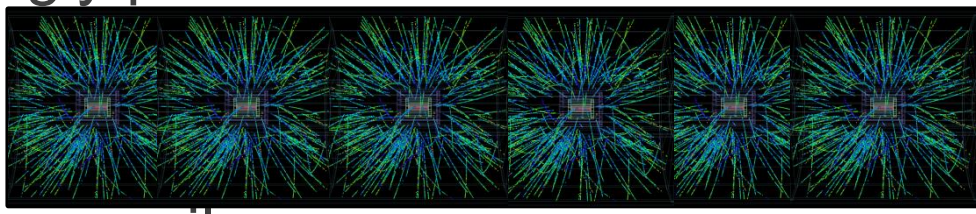
- HEP should definitely try to navigate in the wake of the Big Data hyper giants





Big Data approach

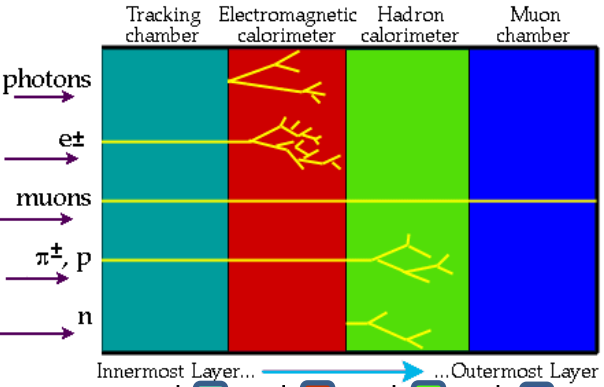
- Very large data sets
  - High Energy Physics data are inherently and embarrassingly parallel… but

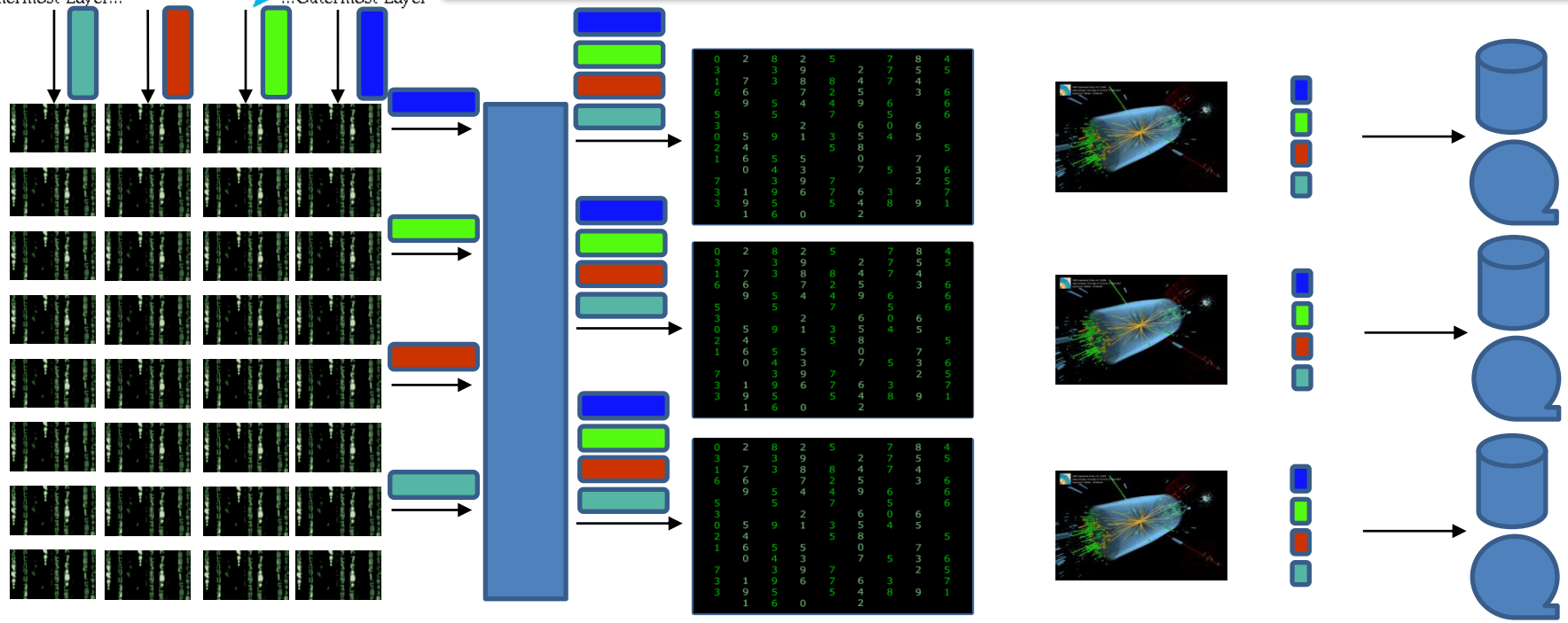


  - At the luminosity targeted for the upgrade there will some pile-up
    → Continuous dataflow
  - Good calibration often requires high statistics and therefore some central database.
- Issues to become a Big Data shop
  - Lots of legacy software not designed for this paradigm
  - Fraction the work into small independent manageable tasks
  - Merge results

Big Data approach

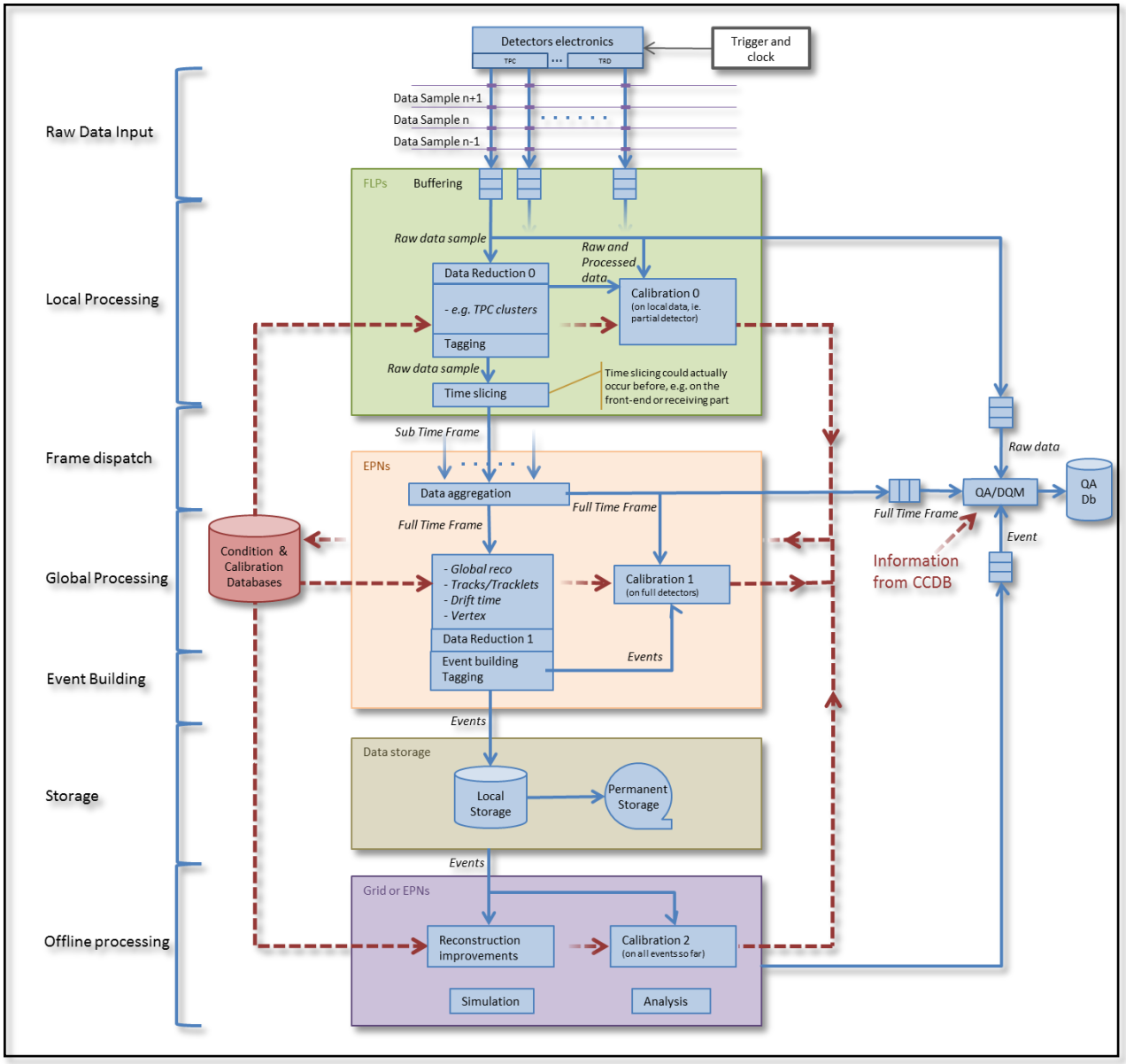
- Continuous detector reading
- Replace events with time windows (100 ms ~5'000 events)
- Self sufficient small dataset ?
- Calibrate and reconstruction online →
  Reduce data volume, Structure the data, Faster results
- Prototyping with ZeroMQ and Zookeeper





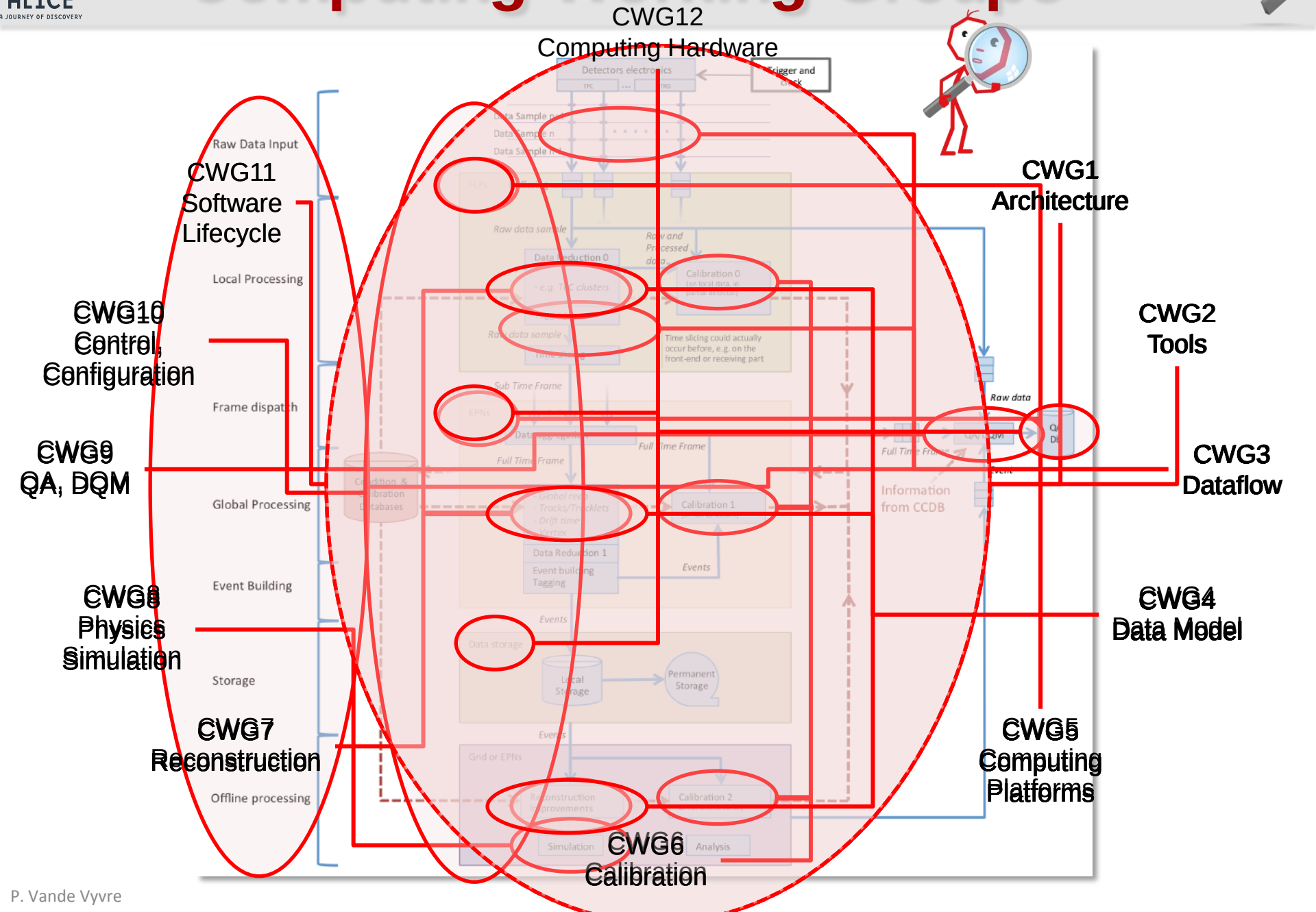
Memory matrix | Multiplexer | Computer Farm | Complete events

Dataflow Model



Model from the Software Framework Panel

Computing Working Groups



Next steps

- Intensive period of R&D :
  - Collect the requirements: ITS and TPC TDRs
  - System modeling
  - Prototyping and benchmarking
- Technology and time are working with us
  - New options
  - Massive usage of commercial equipment very appealing
- TDR
  - October ‘13:
    - Define table of content
    - Establish editorial board
  - December ‘13:
    - System Requirement Document
    - High-level dataflow model
    - Computing platforms benchmarks
    - Networking benchmark
  - June ‘14
    - Software framework architecture
  - Sep ‘14
    - TDR

Thanks !