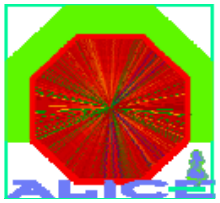


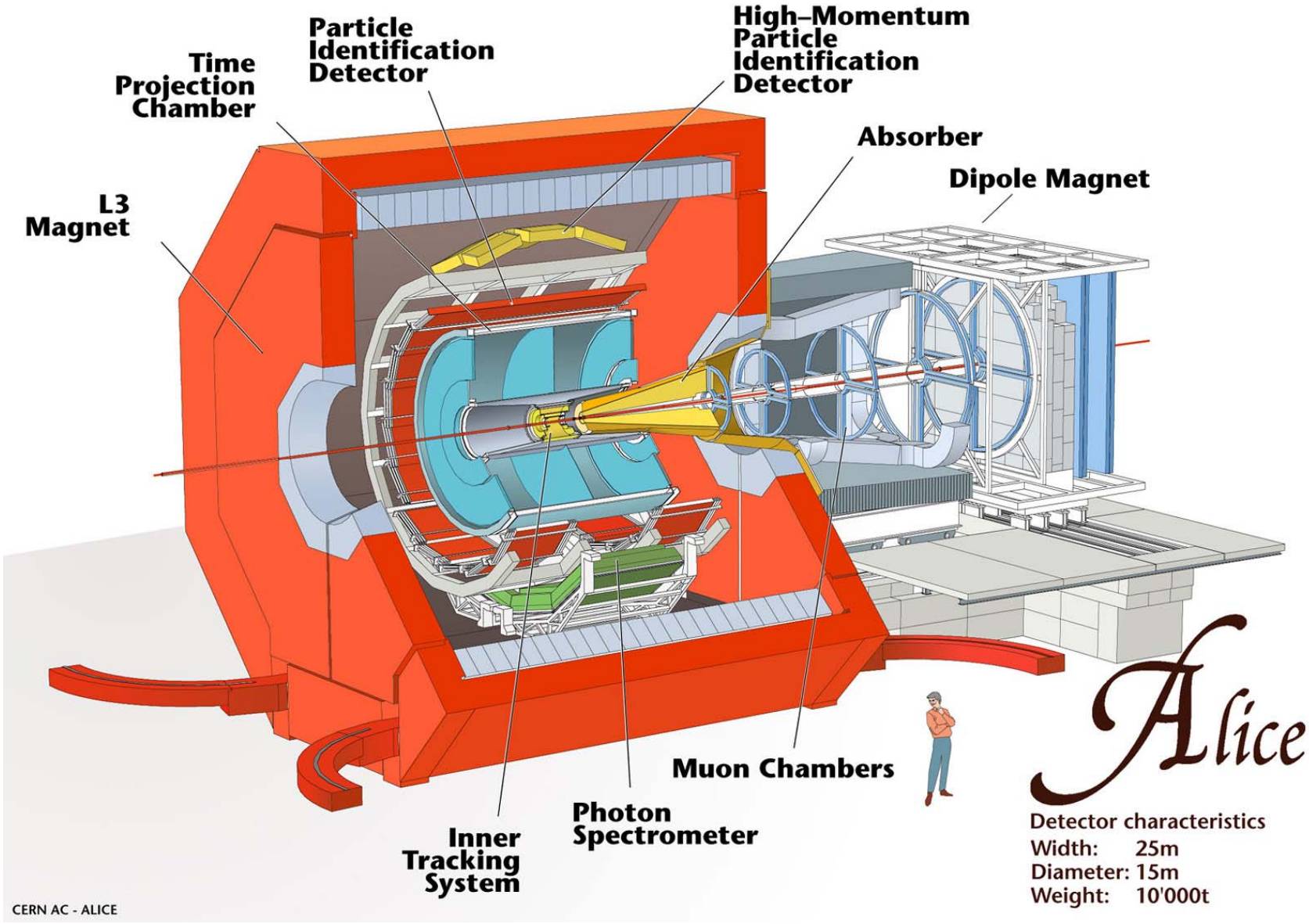
Challenges of Data Acquisition at the LHC

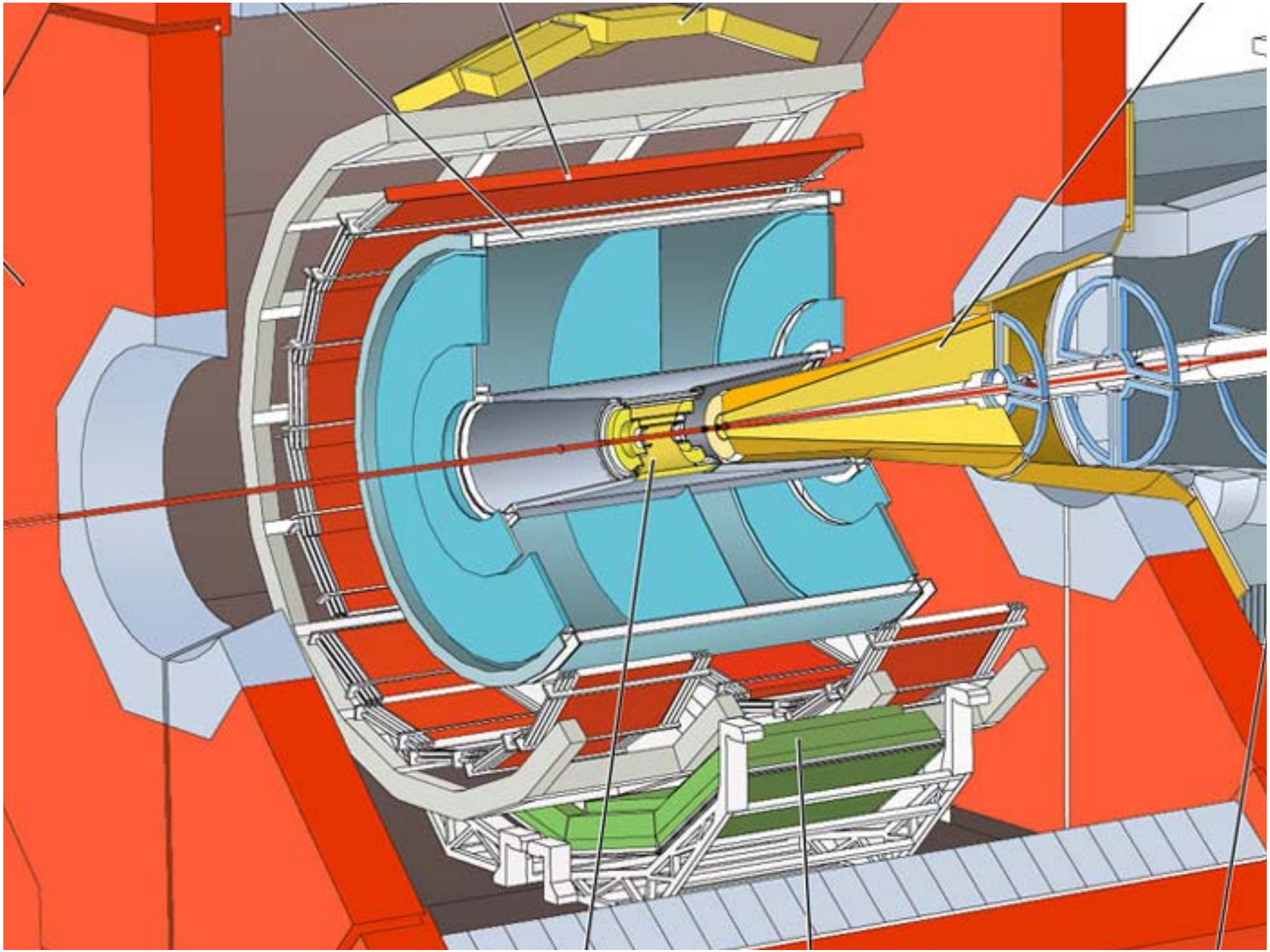
UK DTI Delegation
30-June-2004

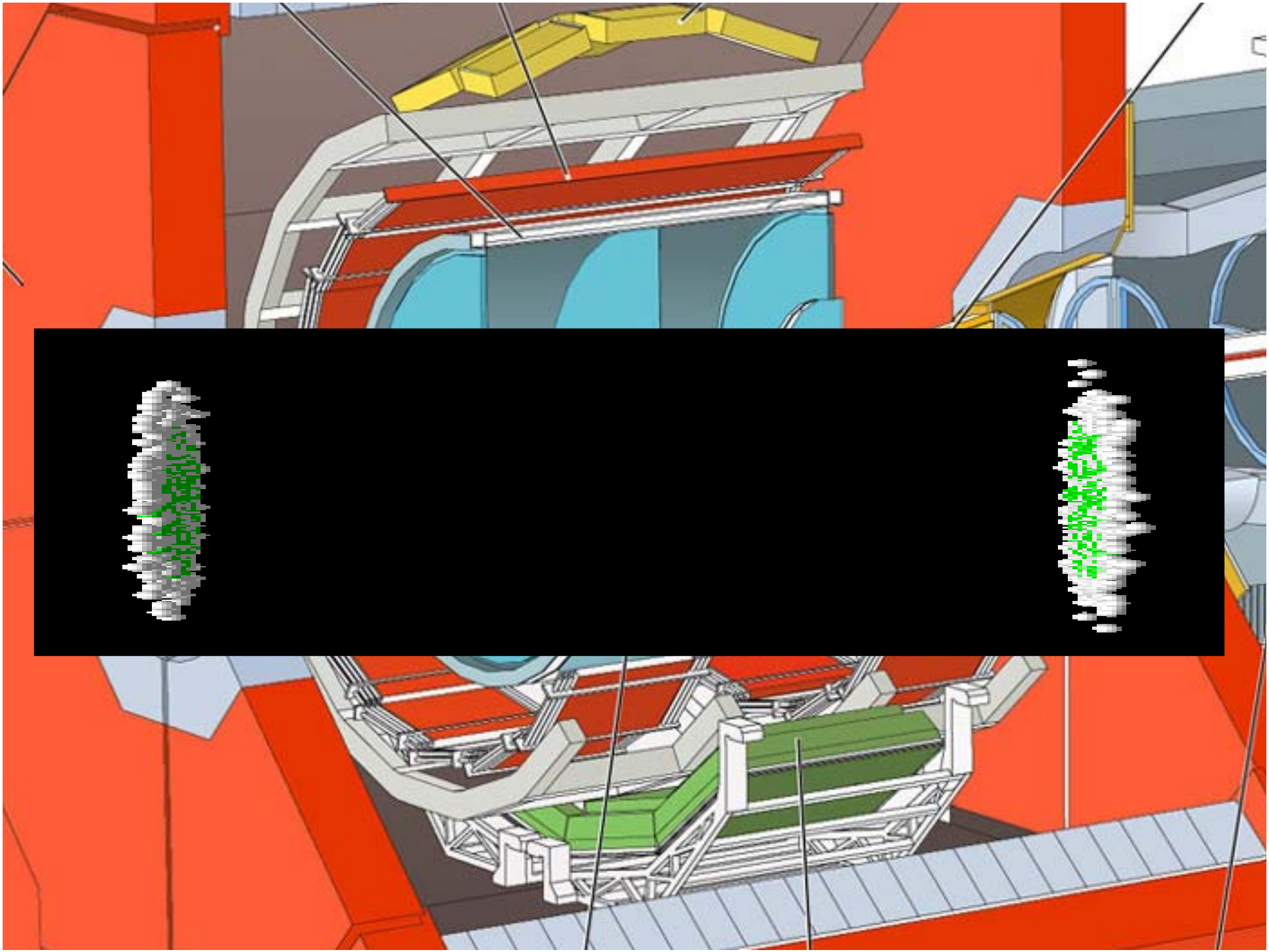
P.Vande Vyvre - CERN/PH

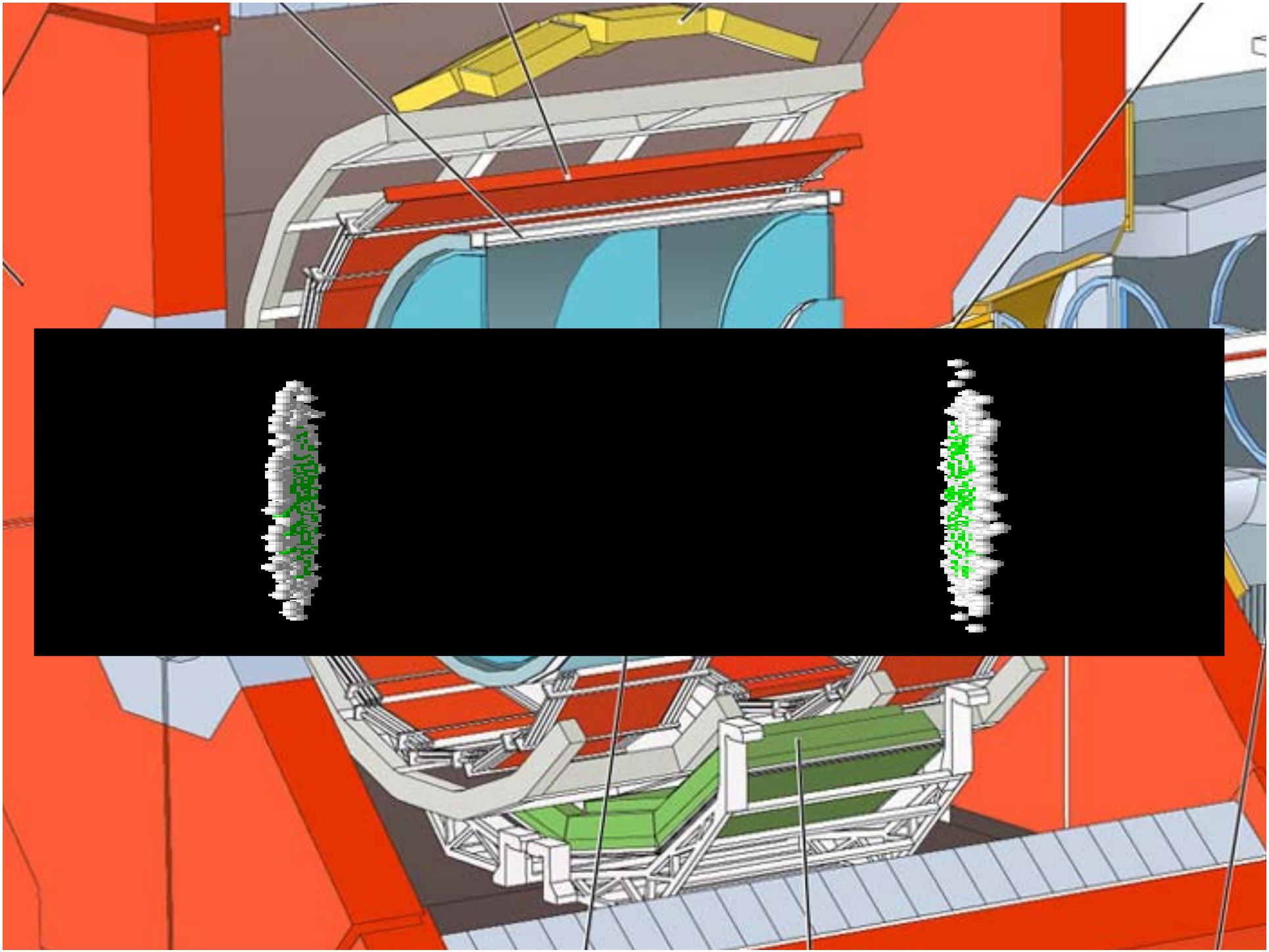
Outline

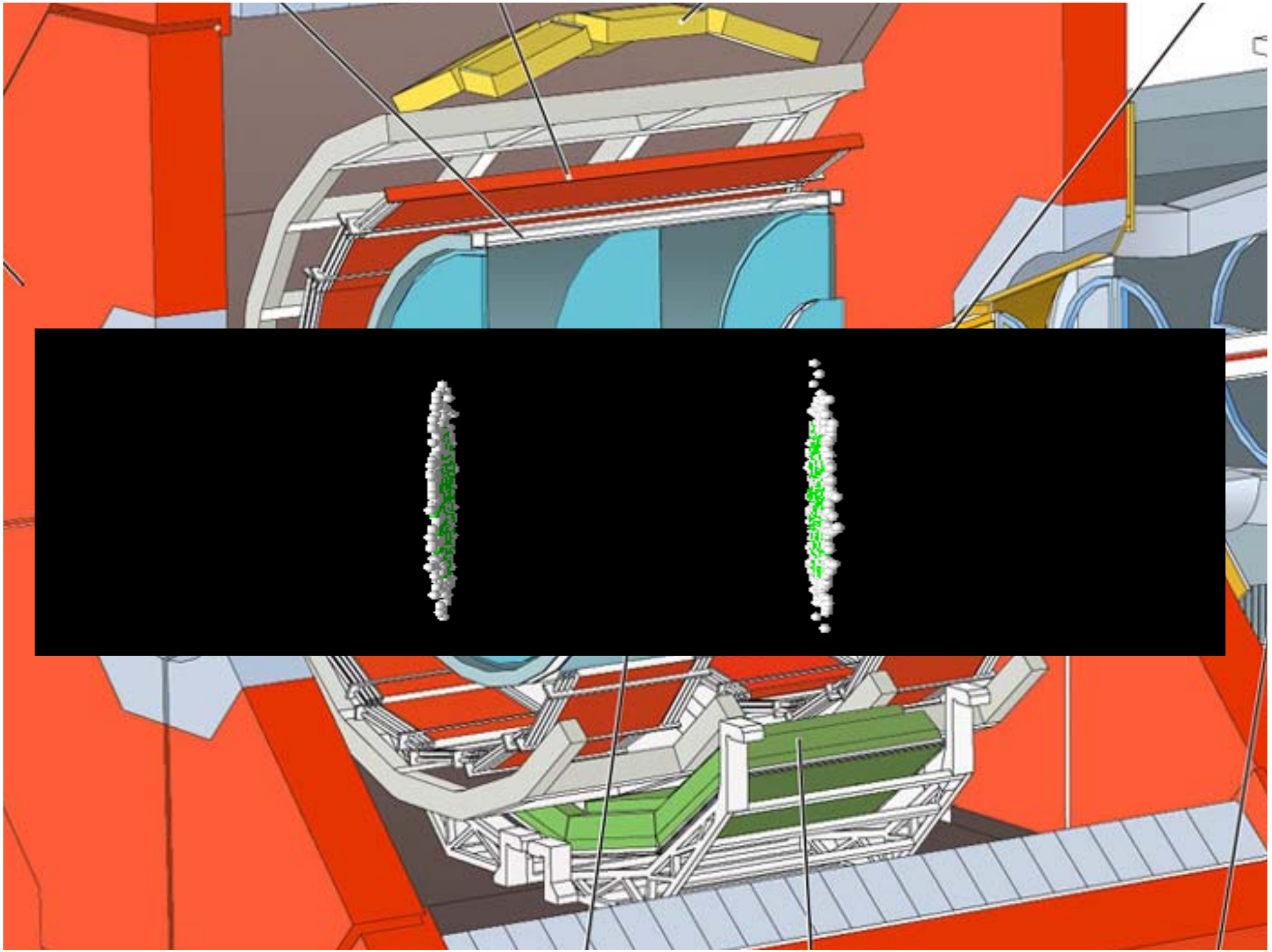
- Trigger and Data Acquisition
  - Main principles
  - Challenge at LHC
- Trigger and DAQ
  - Logical model
  - Main elements:
    - Baseline
    - Selection criteria
    - Industry involvement
- Conclusions

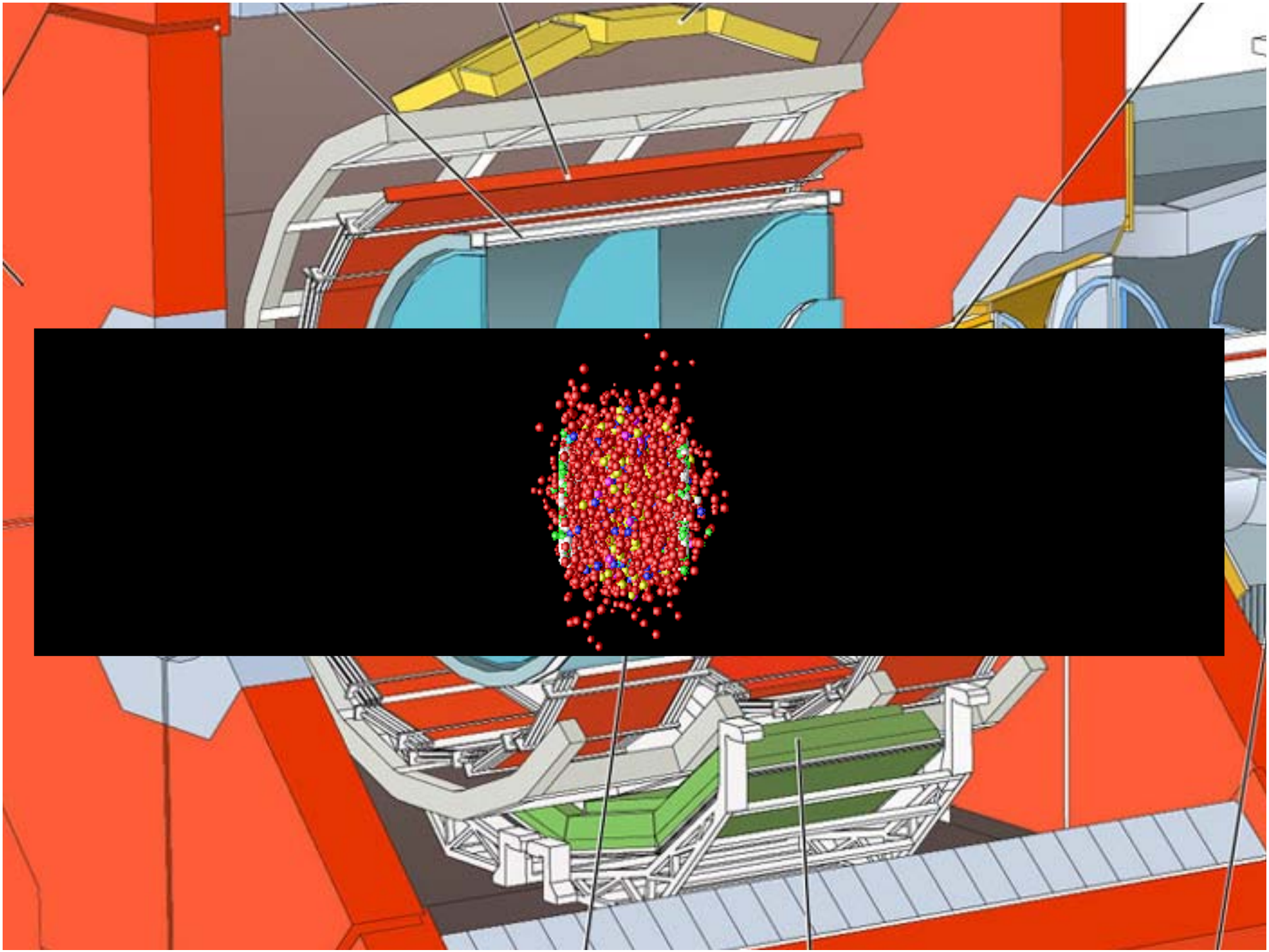

Time Projection Chamber
Particle Identification Detector
High–Momentum Particle Identification Detector
Absorber
Dipole Magnet
L3 Magnet
Muon Chambers
Photon Spectrometer
Inner Tracking System
CERN AC - ALICE
Alice
Detector characteristics
Width: 25m
Diameter: 15m
Weight: 10'000t

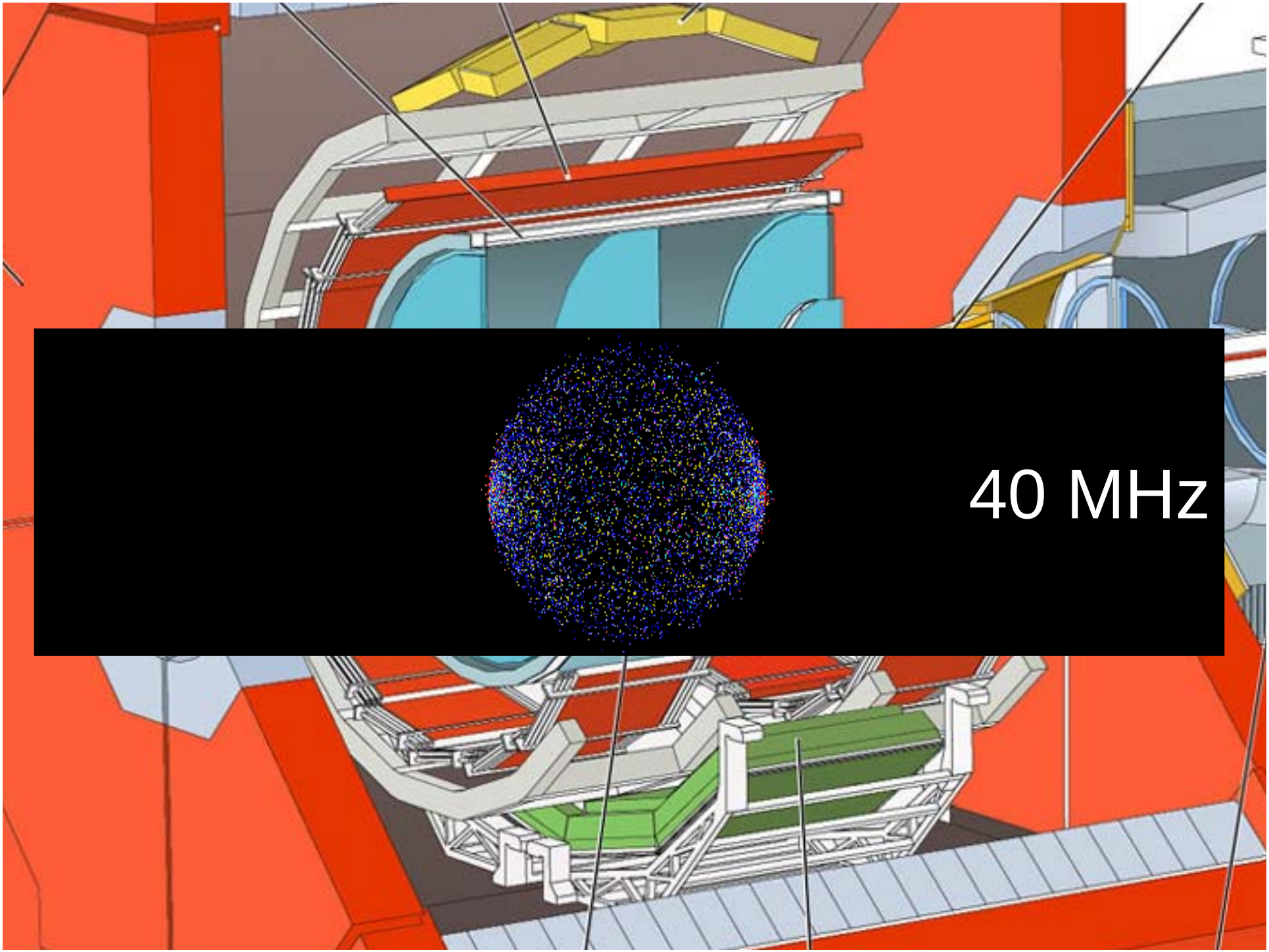




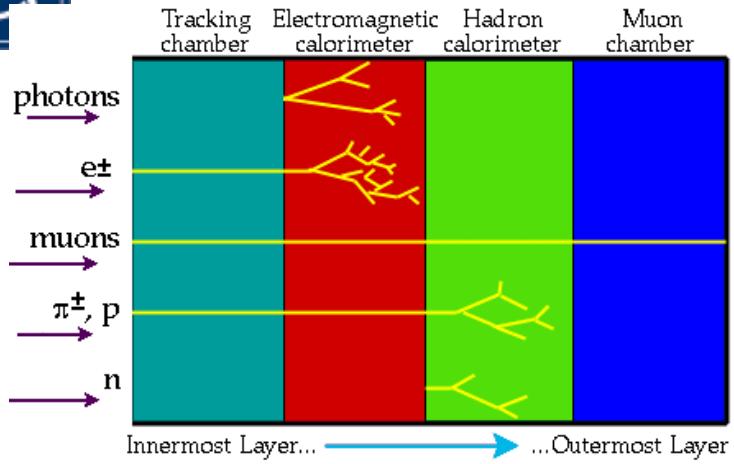







40 MHz

Trigger

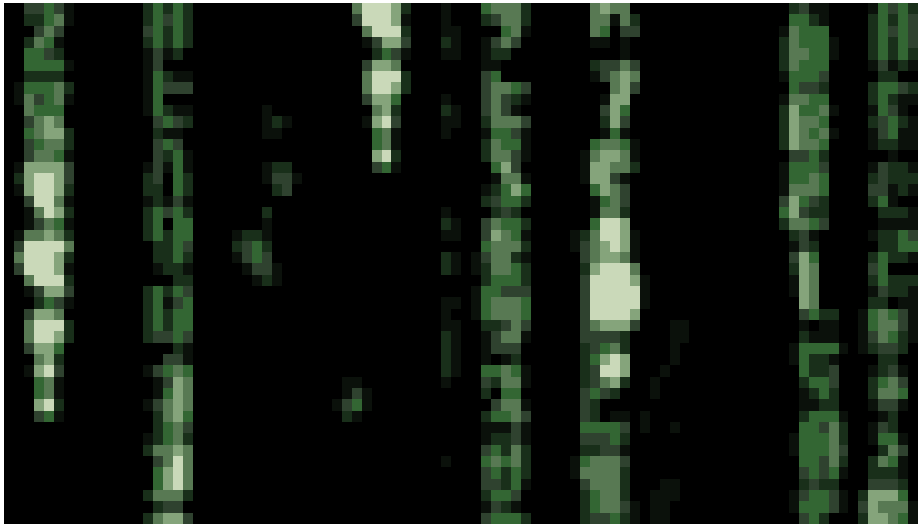


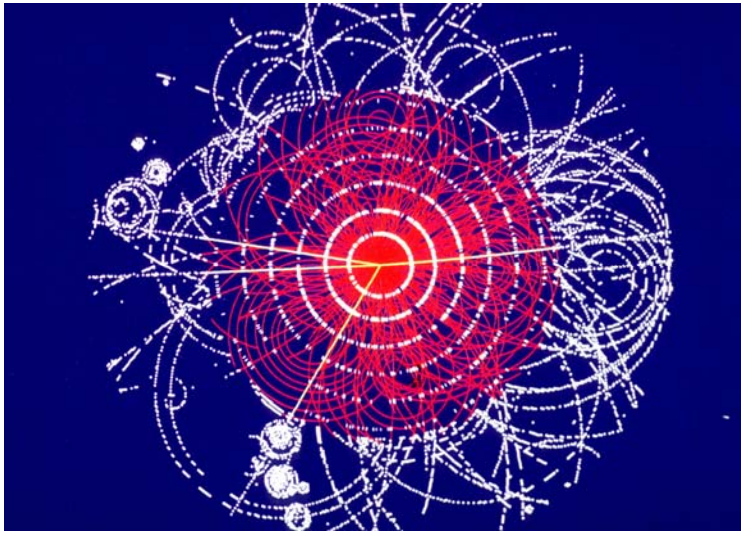
Multi-level trigger system

Reject background

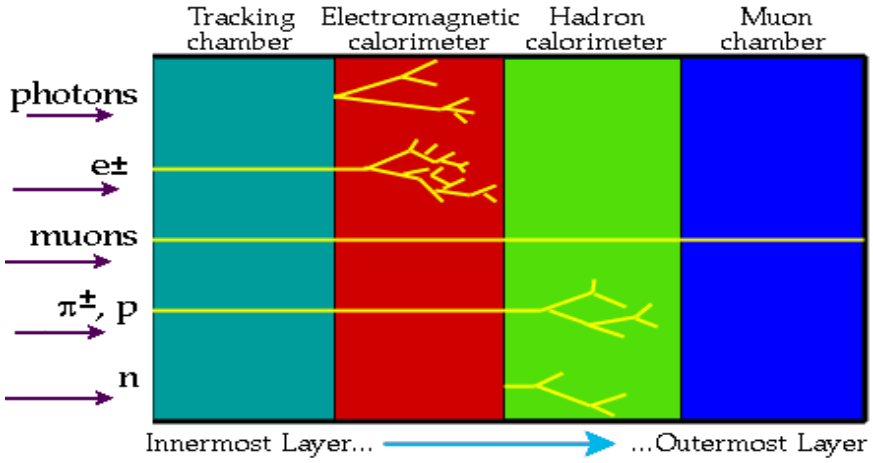
Select most interesting collisions

Reduce total data volume





Data Acquisition (DAQ)

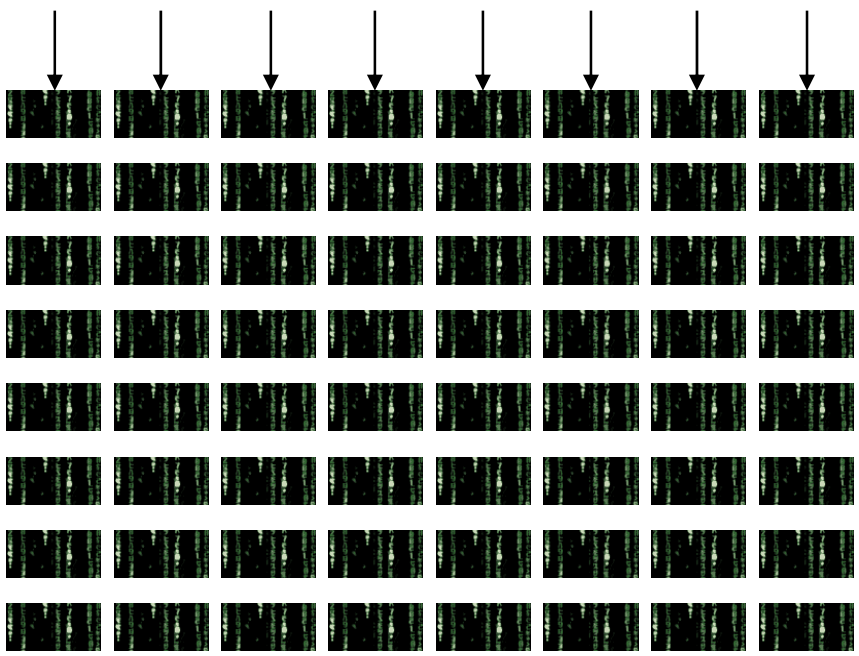


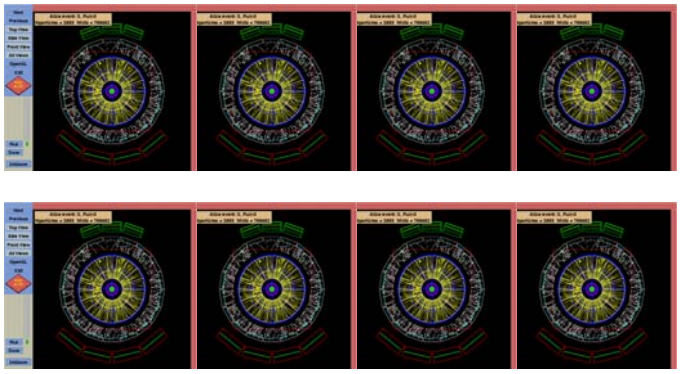
Acquire data from 1000's of sources

Reassemble all the data pertaining to the same physics event

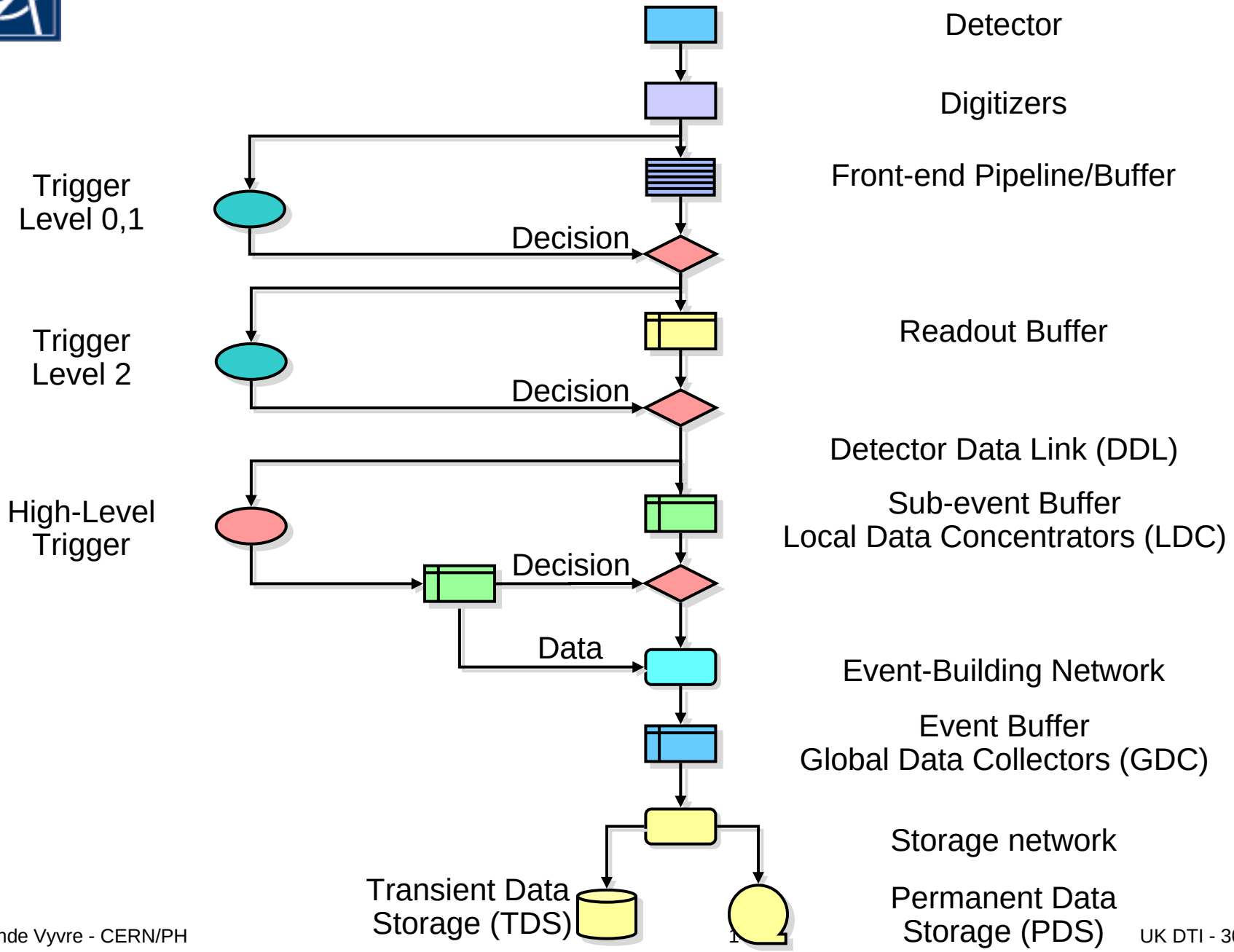
Control the data taking

Monitor the data quality

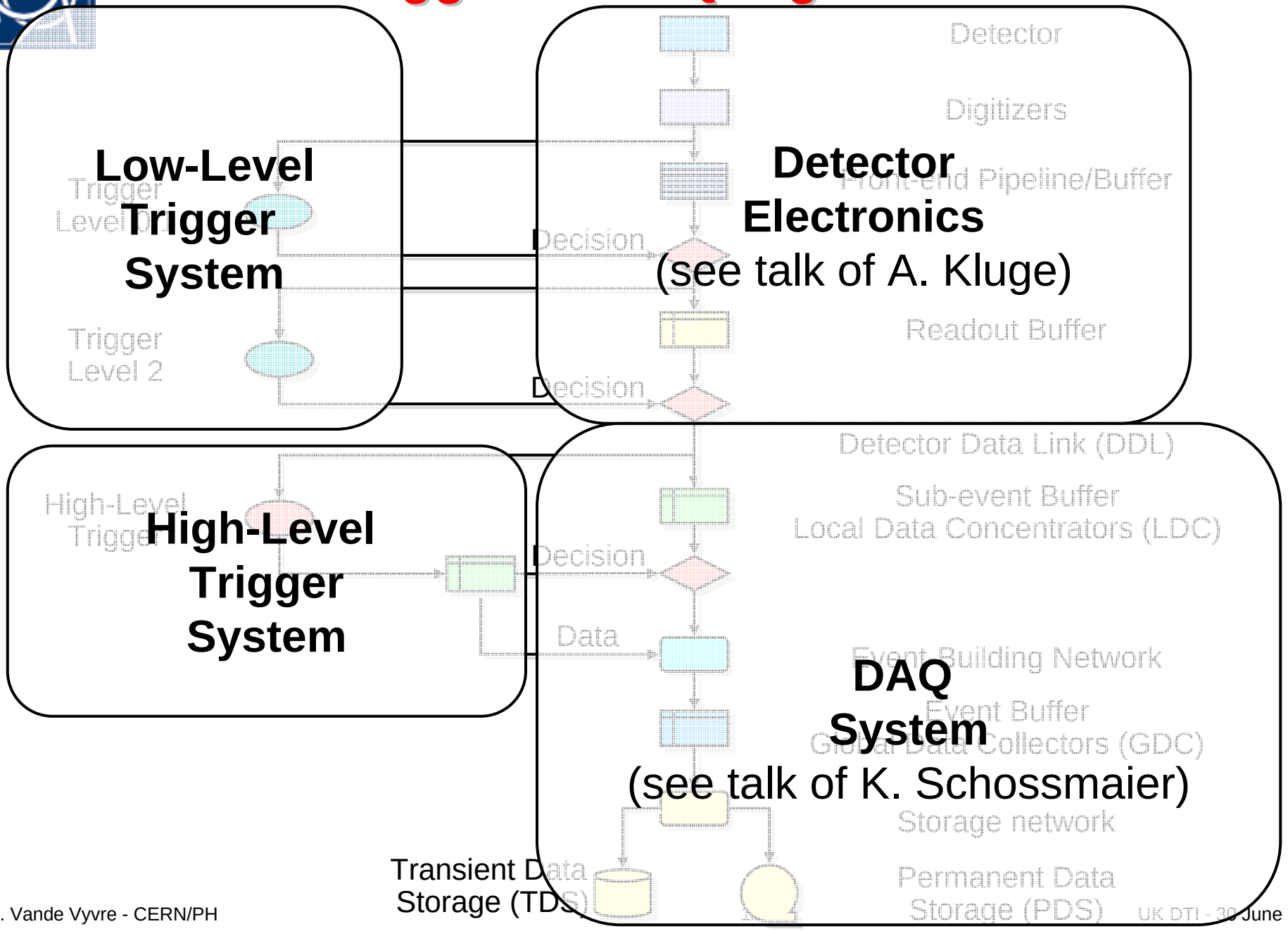




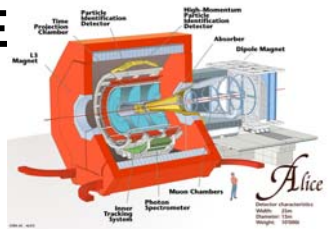
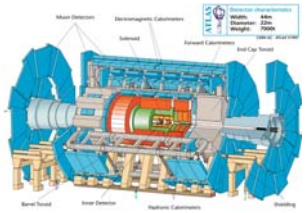
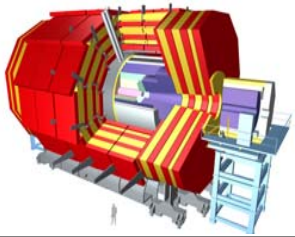
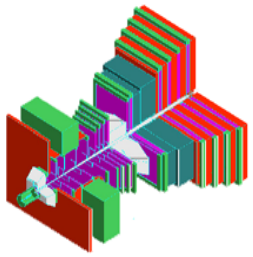
Trigger & DAQ logical model



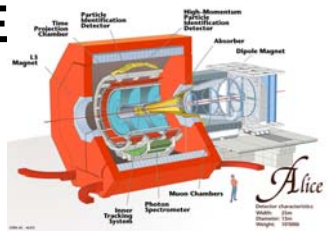
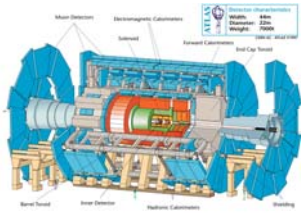
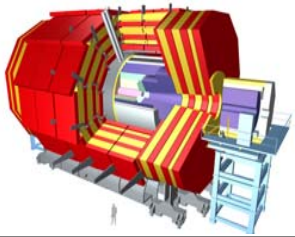
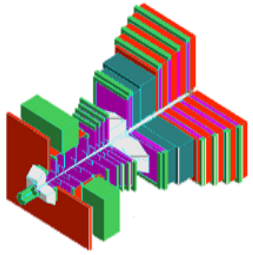
Trigger & DAQ logical model



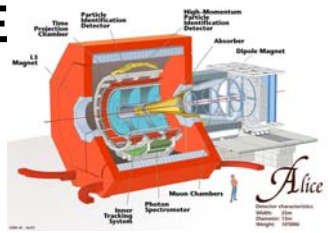
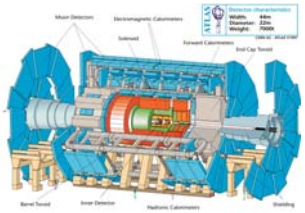
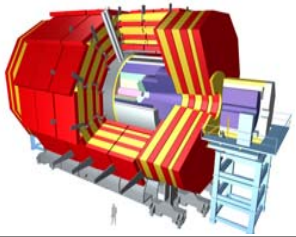
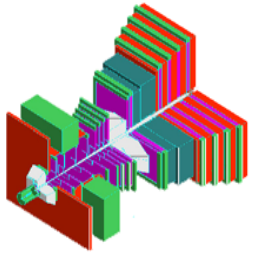
Trigger @ LHC

		# Trigger Levels	Rate First Level Trigger (Hz)	Rate To MSS (Hz)
ALICE		4	$6x10^3$	200
ATLAS		3	10^5	100
CMS		2	10^5	100
LHCb		3	10^6	200

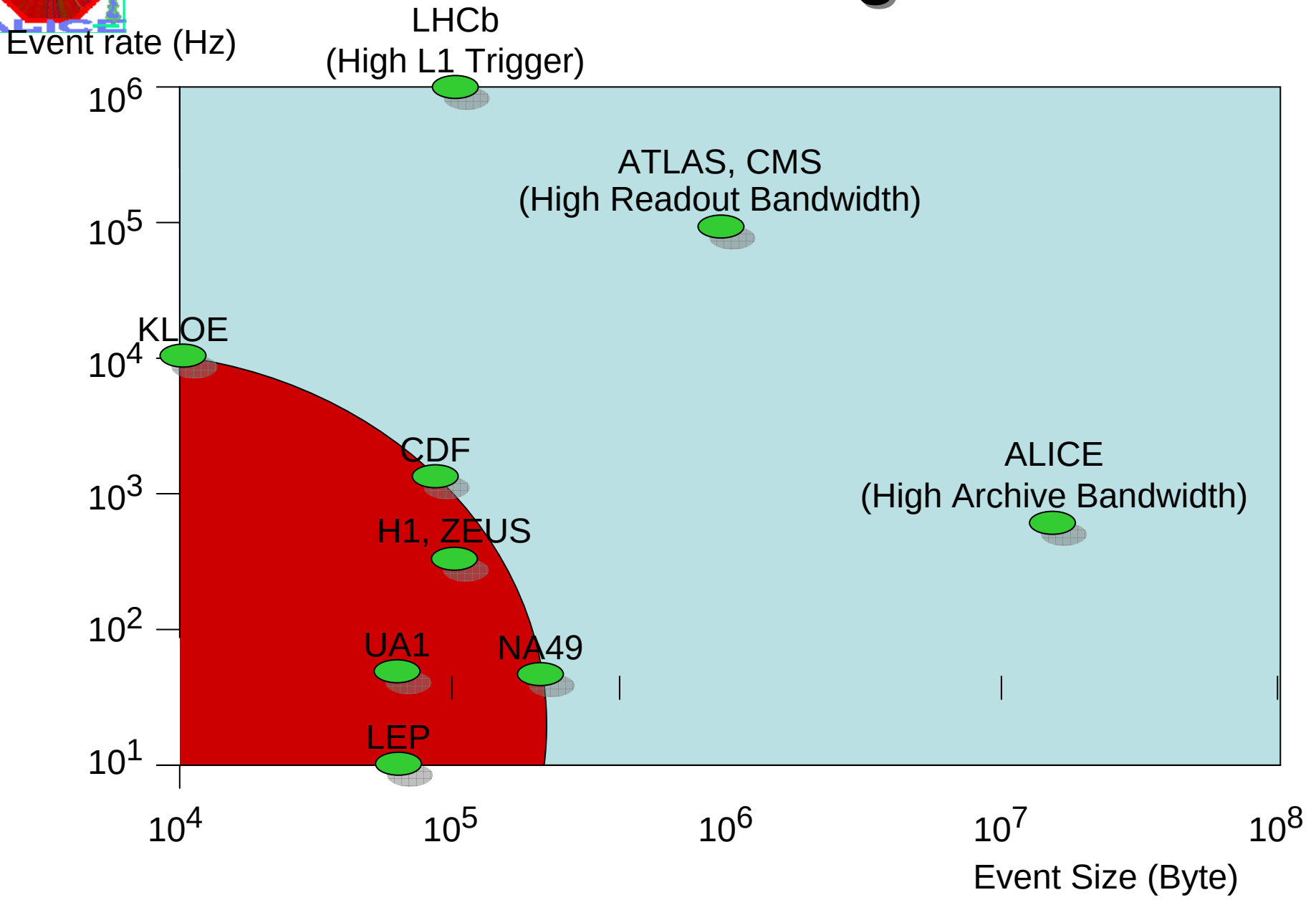
DAQ @ LHC

		Event Size (Byte)	Readout (HLT input) (Events/s.)	(GB/s)
ALICE 	**Pb-Pb**	**5×10^{7}**	**2×10^{3}**	**25**
ATLAS 	**pp**	**10^{6}**	**2×10^{3}**	**10**
CMS 	**pp**	**10^{6}**	**10^{5}**	**100**
LHCb 	**pp**	**2×10^{5}**	**40×10^{4}**	**4**

Mass Storage @ LHC

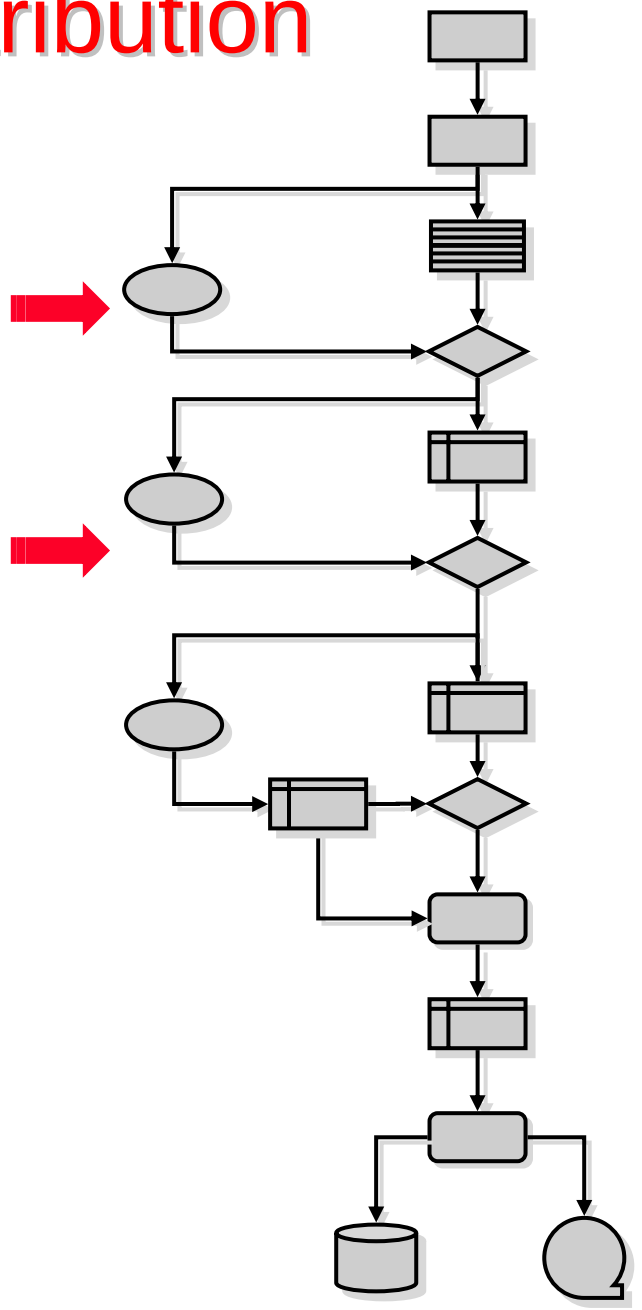
			Readout (HLT output) (Events/s.)	(MB/s)	Data archived Total/year (PBytes)
ALICE		Pb-Pb	200	1250	2.3
ATLAS		pp	100	100	6.0
CMS		pp	100	100	3.0
LHCb		pp	200	40	1.0

LHC Challenge

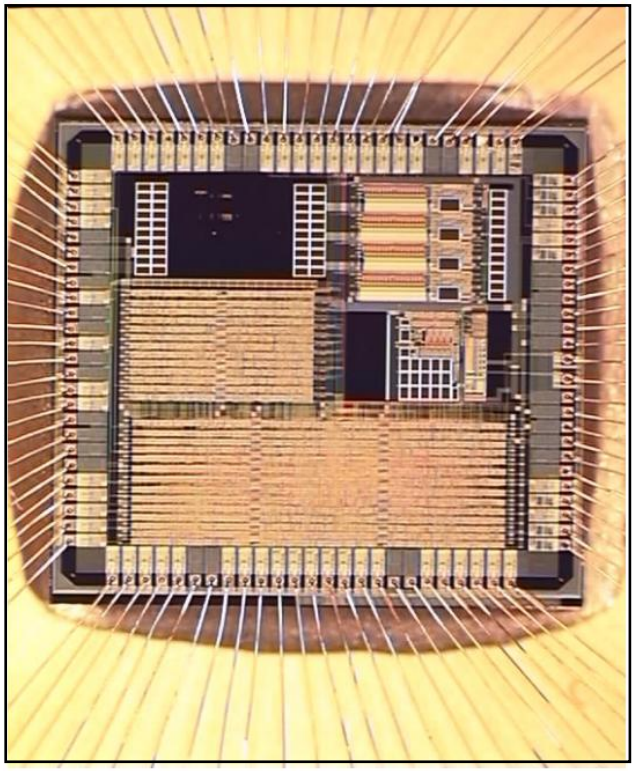


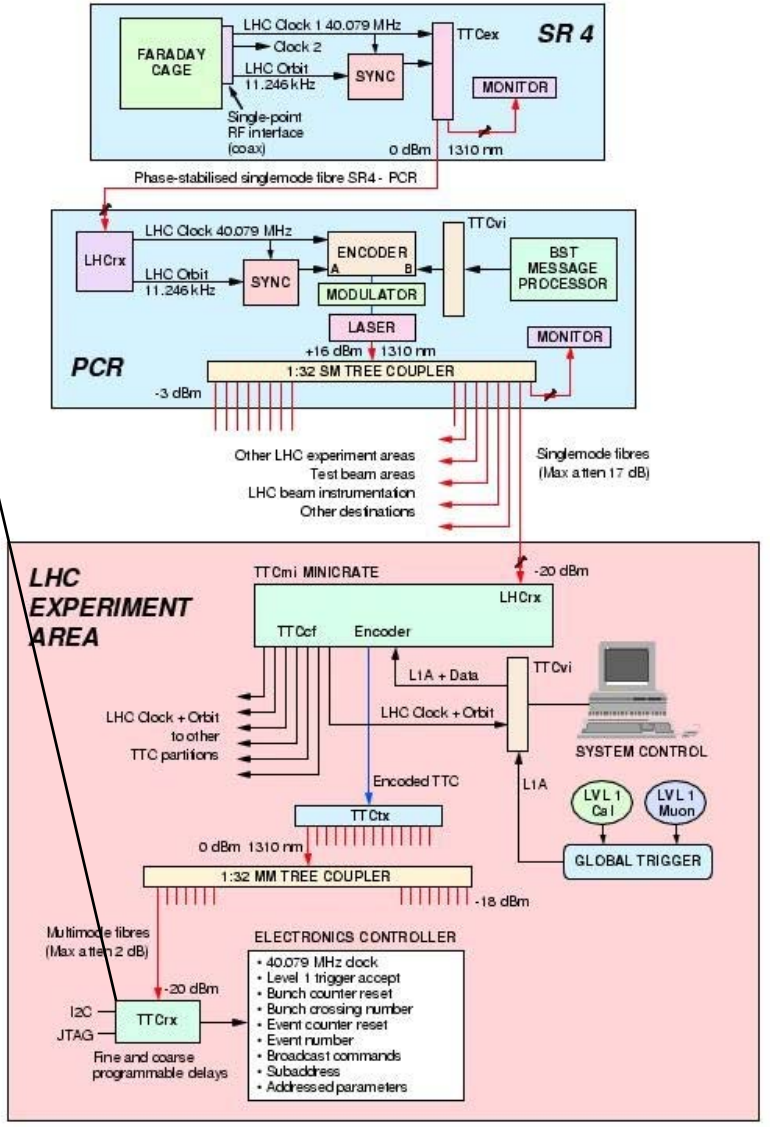
Trigger & Timing distribution

- Transfer from TRG to electronics
  - LHC clock 40 MHz
  - TRG decisions
- One to many
- Massive broadcast (100's to 1000's)
- Optical, Digital
  - HEP development
  - HEP-specific microelectronic components
  - Industrial production
- Future
  - Higher clock
  - Larger number of destinations



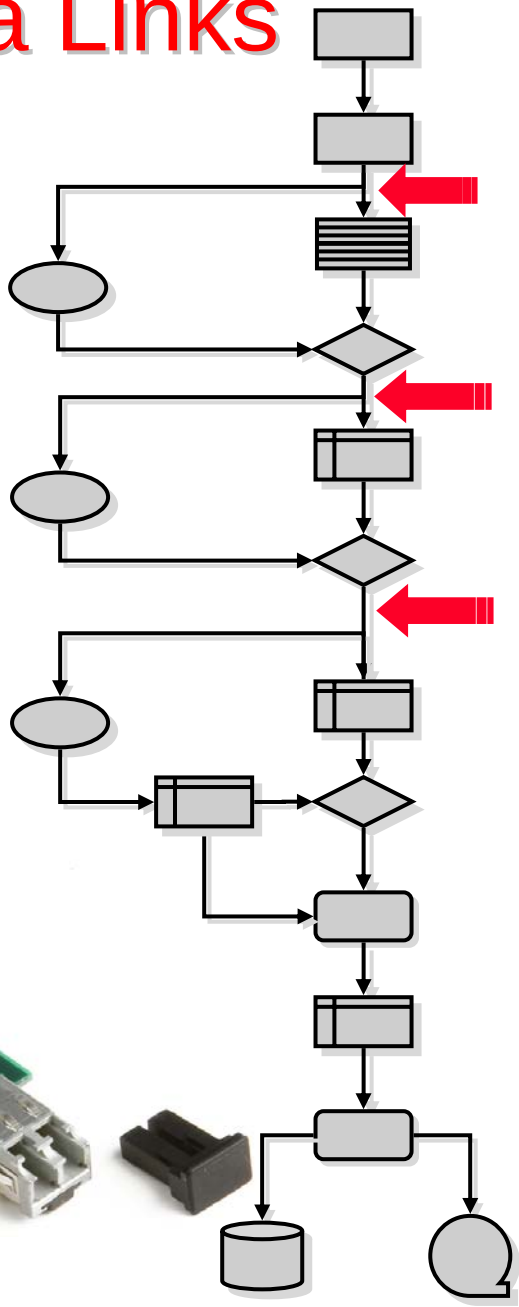
LHC Trigger & Timing distribution





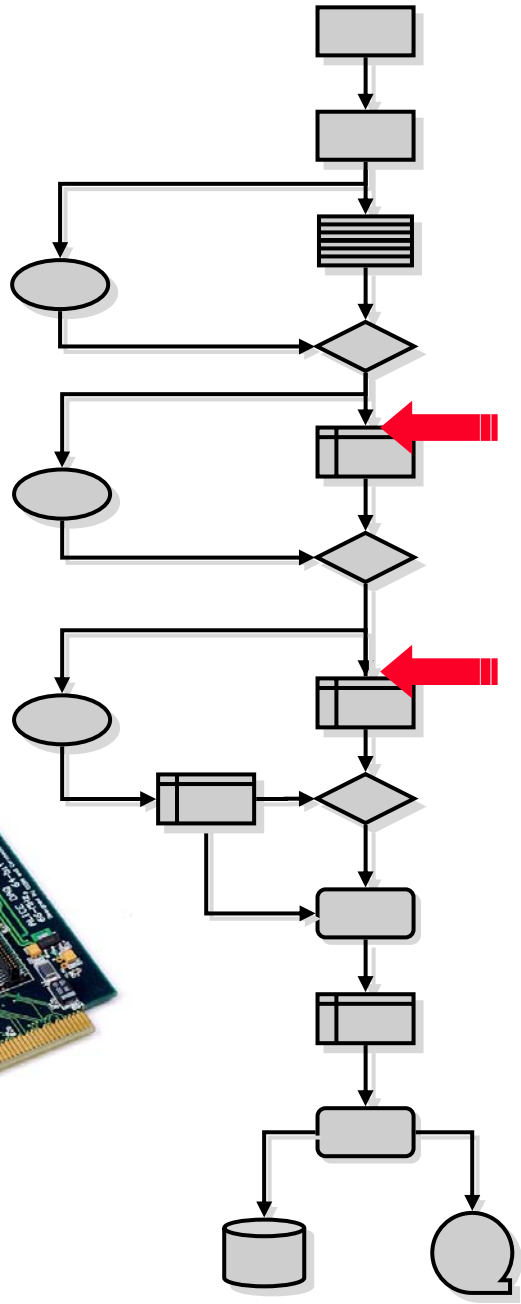
Detector & Readout Data Links

- Interface and data-transfer detector/DAQ
- Point-to-point
- Massive parallelism (100's to 1000's)
- Analog: HEP-specific components
- Digital
  - HEP developments based on commodity components
  - ASIC (Radiation hard) or FPGA (Radiation tolerant)
  - Fiber Channel or Gig. Ethernet: 1, 2.1 or 2.5 Gb/s
  - Industrial production
- Future
  - Optical component and FPGA for 10 and 40 Gb/s
  - DWDM up to 1 Tb/s



Links Adapters

- Adapter for 1 or a few links to PC I/O bus
- A few-to-one
- Massive parallelism (100's to 1000's)
- Physical interface realized by
  - Commodity Custom chip
  - IP core (VHDL code synthesized in FPGA)
  - HEP development
  - Industrial production
- Future: I/O bus evolution



Link and adapter performance

- PCI and PCI-X busses
- No large local memory. Fast transfer to PC memory

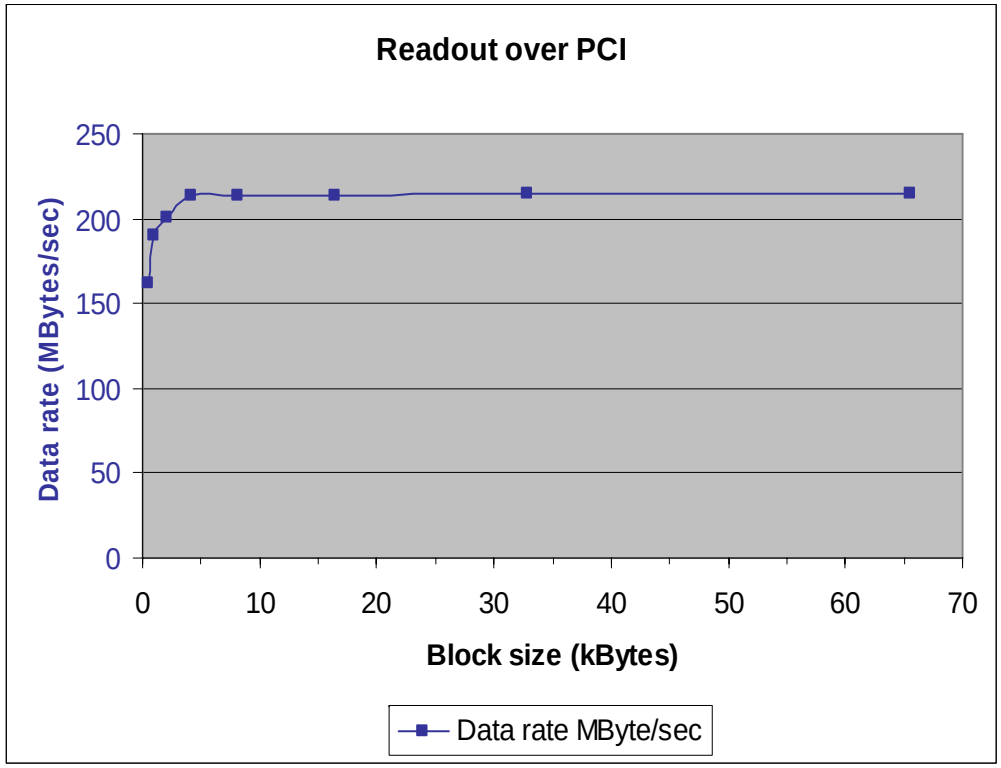
200 MB/s sustained

Total PCI load: 92 %

Data transfer PCI load: 83 %

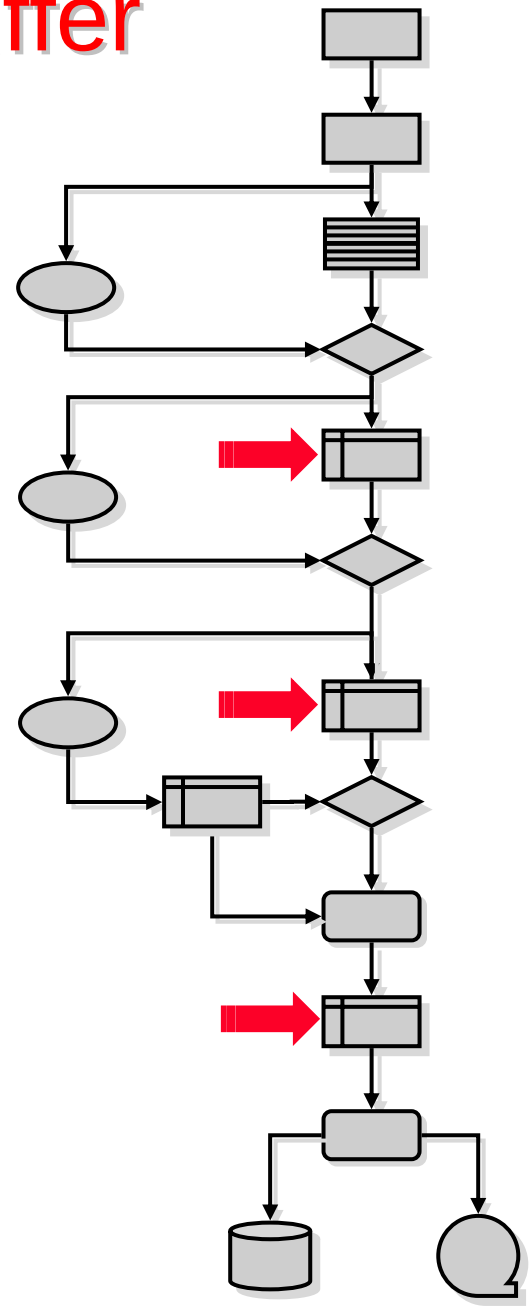
Lots of bw available.

Major fraction available to end application.



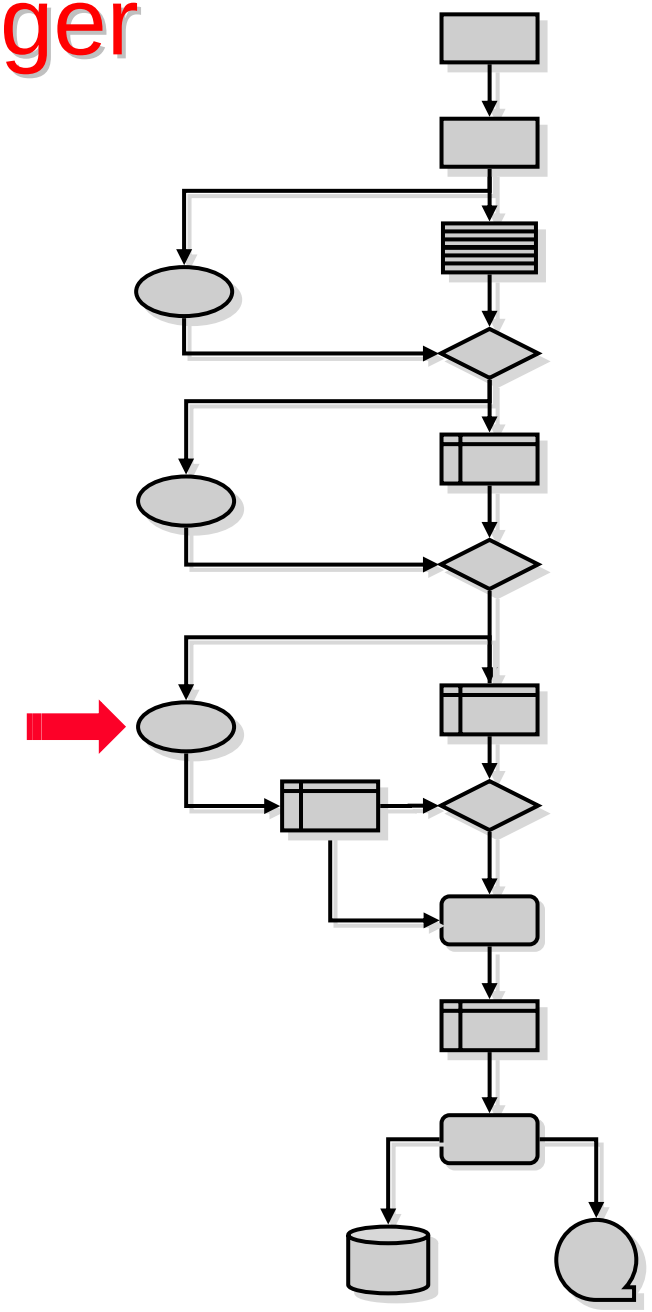
Subevent & event buffer

- Baseline:
  - Fast dual-port memories
  - Electronics racks are over
    Extensive use of dual-CPUs PCs
- Key parameters:
  - Cost/performance
  - Performance: I/O and memory bandwidth
- Partnering for test of motherboards
  (See talk of K. Schossmaier)
- Future
  - Faster memory clock
  - Wider data bus



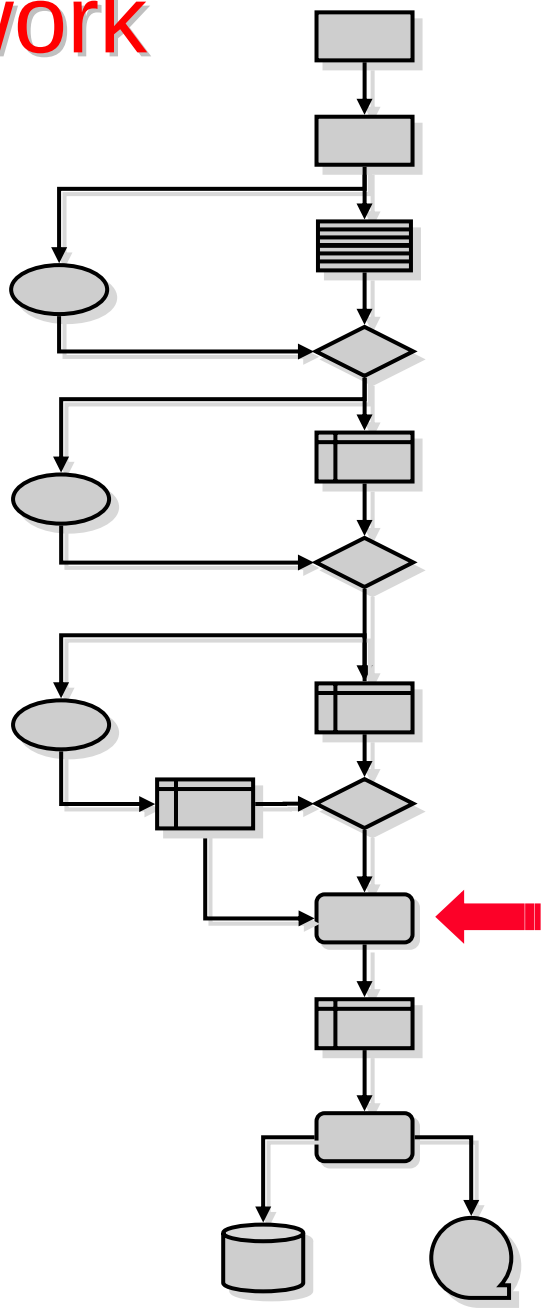
High-Level Trigger

- Baseline:
  - Function: fast dual-port memories and data processing
  - Dual-CPUs PCs
- Key parameters:
  - Cost/performance
  - Performance: memory bandwidth & CPU performance
- Future
  - Faster CPU clock
  - Multi CPUs chips (3G, human I/O)
  - Wider data bus



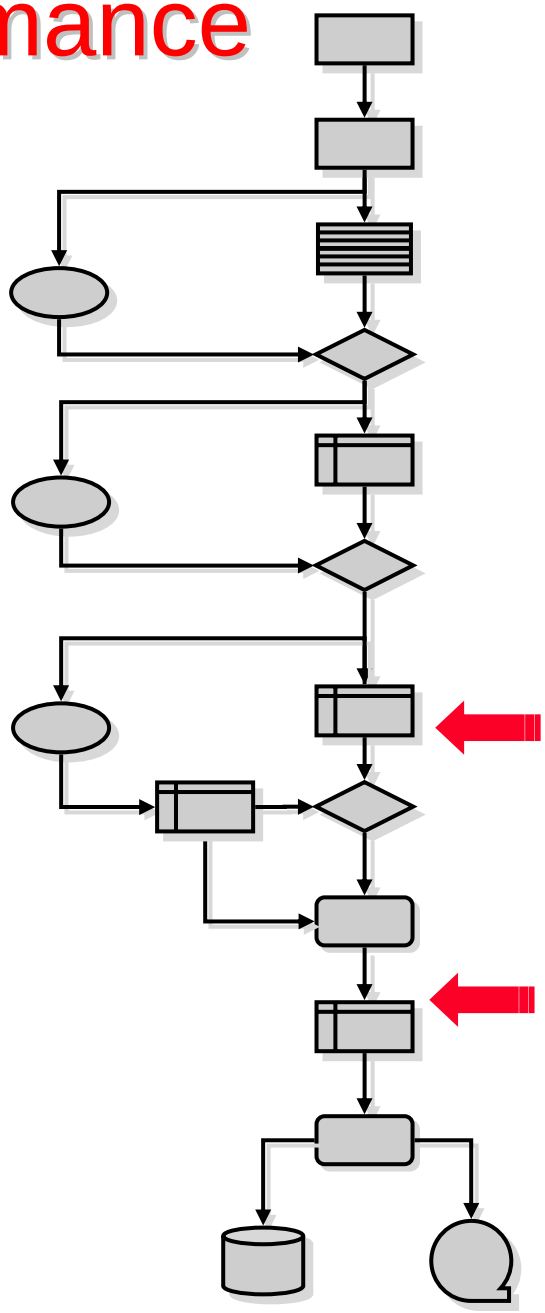
Event Building Network

- Baseline:
  - Adopt broadly exploited standards (Switched Ethernet)
  - Adopt a performing commercial product (Myrinet)
- Motivations for switched Ethernet:
  - Performance of Gigabit Ethernet switches already adequate
    2 Tbit/s of aggregate bandwidth
  - Use of commodity items: network switches and interfaces
  - Easy (re)configuration and reallocation of resources
- Partnering for test of network equipment
- Key parameters:
  - Cost/performance
  - Scalability
- Future: 40 or 100 Gbit/s network

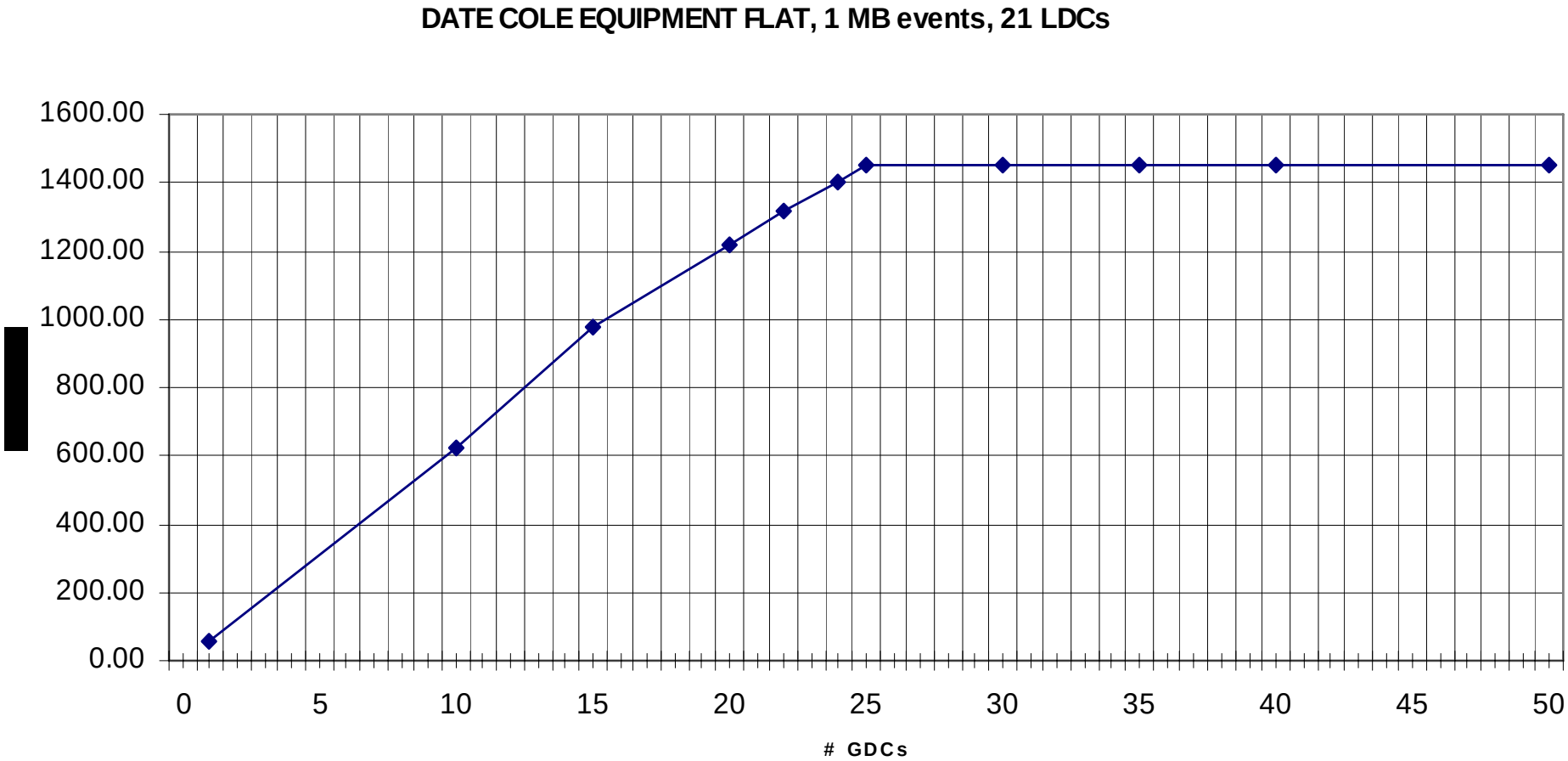


Ethernet NIC's Performance

- Gigabit Ethernet
  - New generation of PC motherboard includes 2 Gbit Eth ports
  - Active market with several players
    - **3Com, Broadcom, Intel, NetGear**
    - **Fast evolution since 3 years**
    - **BW: from 50 to 110 MB/s**
    - **CPU usage: 150 to 60 %**
- TCP/IP Offload Engine (TOE)
  - Dedicated processor to execute IP stack
- 10 Gigabit Ethernet
  - Up to 700 MB/s

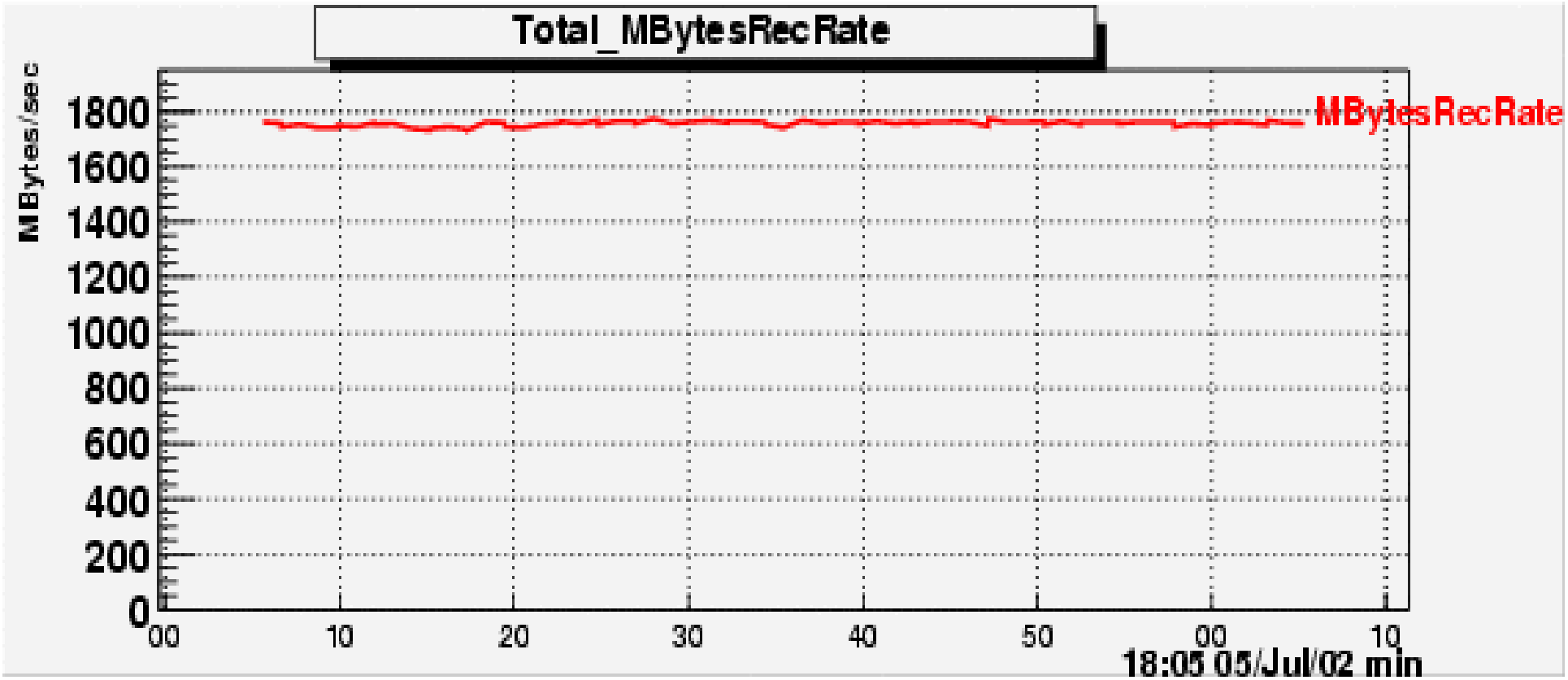


Scalability of network-based event building



- Large integrated tests (Data Challenges)
- Reliability and scalability of the whole system

Performance of network-based event building

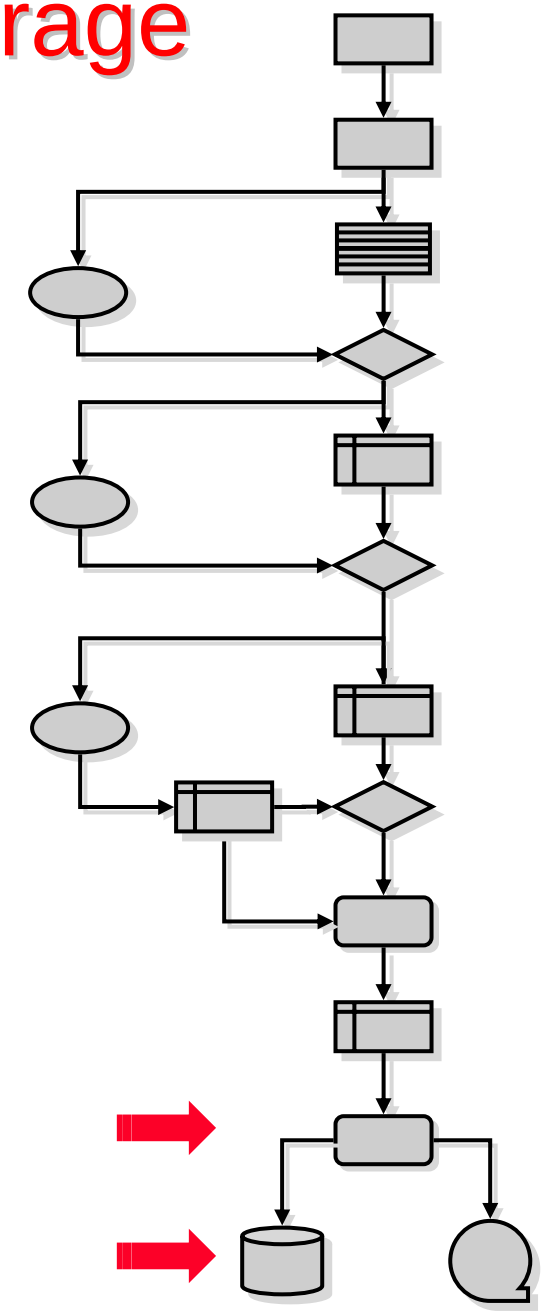


5 days non-stop

- 1750 MBytes/s sustained (goal was 1000)

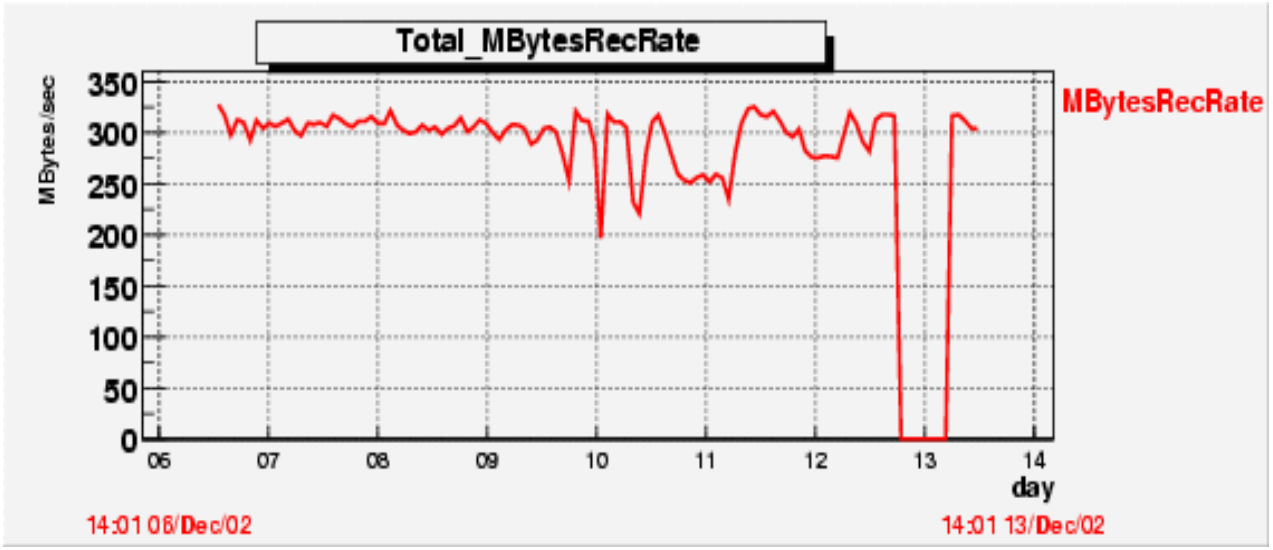
Transient Data Storage

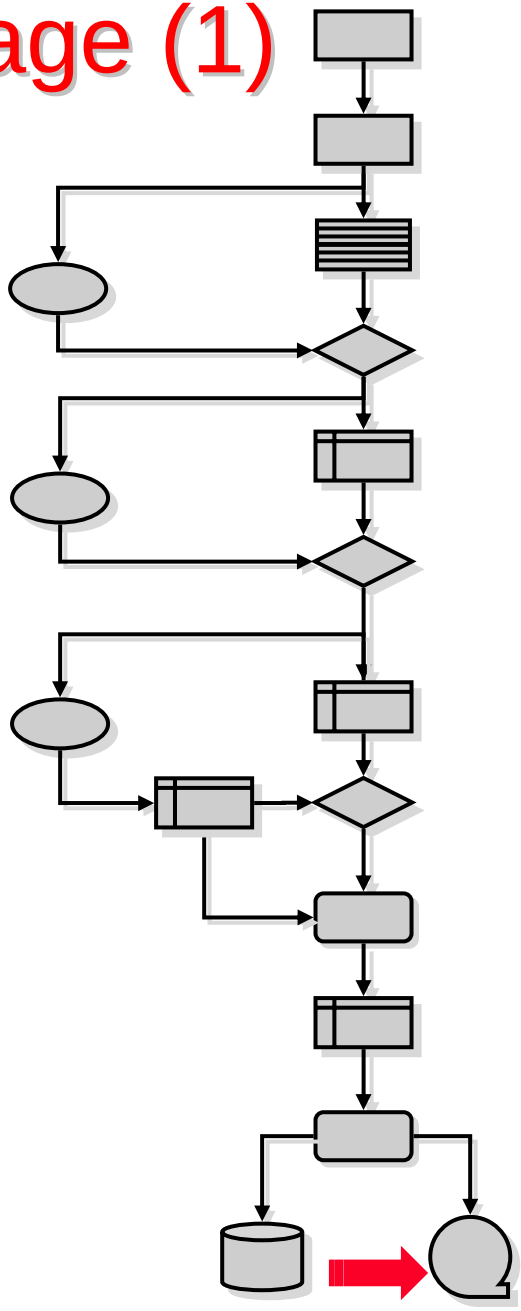
- Transient Data Storage
- Before archiving to tape, if any
- Baseline
  - Storage arrays of commodity disks
  - Box attachment: Fibre Channel
  - Disk attachment: IDE or serial-ATA
  - RAID-level
- Partnering for test of equipment (See talk of K. Schossmaier)
- Key selection criteria:
  - Cost/performance
  - Bandwidth/box
  - Robustness



Permanent Data Storage (1)

- Infinite storage at very low cost
- 1 realistic solution: magnetic tape
- Critical areas
  - High Energy Physics peculiar use of tapes
  - Infrastructure hidden by a hierarchical storage management sw
  - Limited market, different application
  - Limited competition, no real alternative
- Demonstrated solution for LHC
  - 15 parallel streams





Permanent Data Storage (2)

Tape Drive

STK 9940A	10 MB/s 60 GB/Volume SCSI
STK 9940B	30 MB/s 200 GB/Volume Fibre Channel

Tape Library

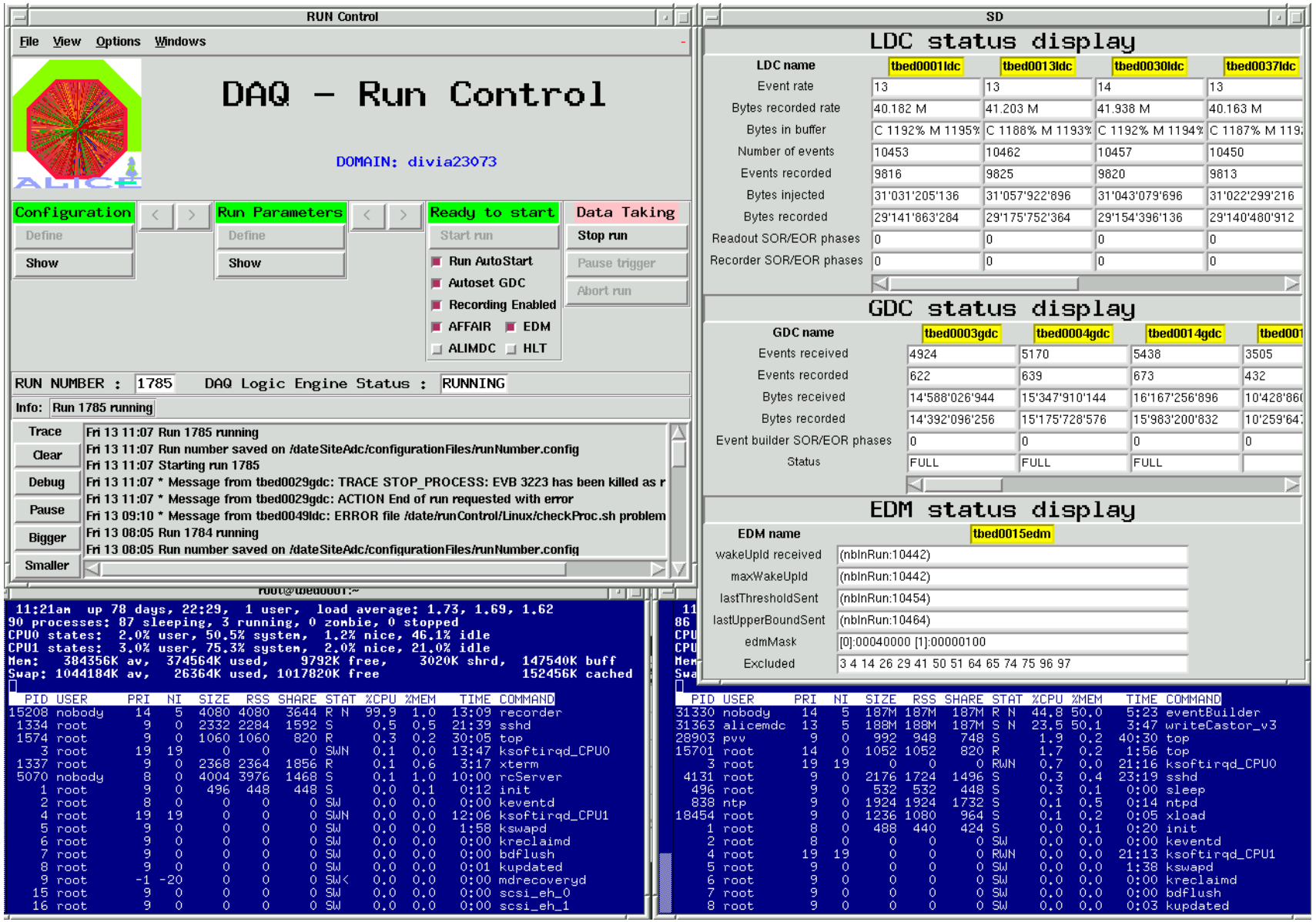
Several tape drives of both generations

Permanent Data Storage (3)

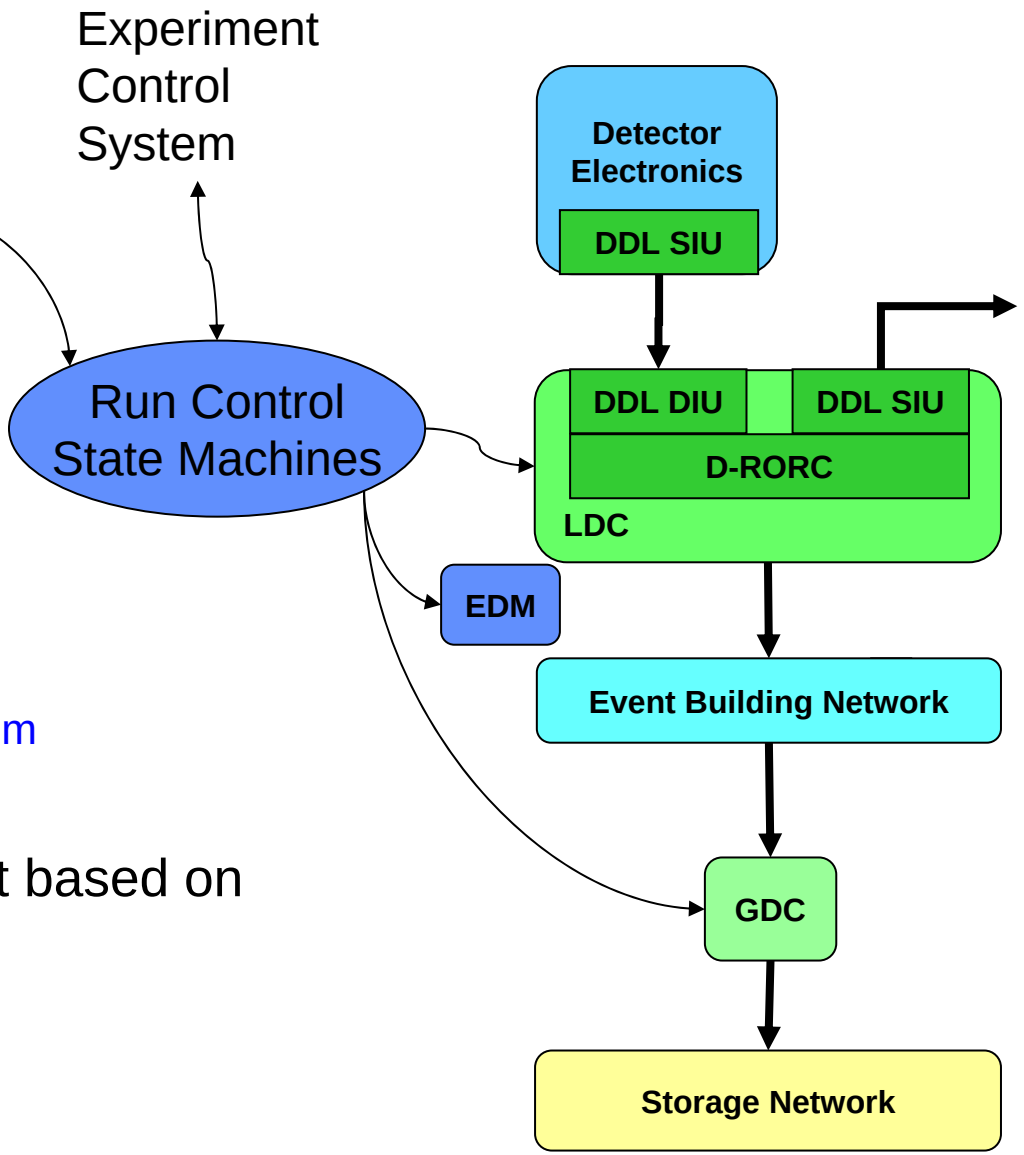
DAQ Software Framework

- DAQ Software Framework
  - Common interfaces for detector-dependant applications
  - Address all configurations and all phases from the start
- DAQ Software
  - Complete ALICE DAQ software framework in 3 packages:
    - **DATE:**
      - **Data-flow: detector readout, event building**
      - **System configuration, control (1000's of programs to start, stop, synchronize)**
    - **AFFAIR: Performance monitoring**
    - **MOOD: Data quality monitoring**
  - Production-quality releases
  - Evolving with requirements and technology ⇒ home-development
- Key issues
  - Scalability (1 to 1000, demonstrate it)
  - Support and documentation

Data Flow - DATE



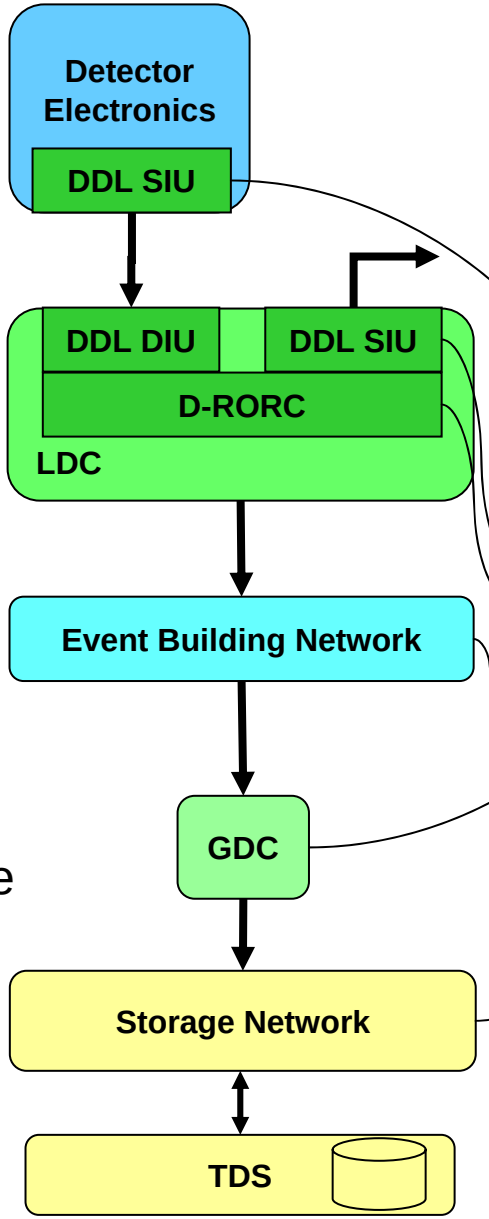
Control - DATE

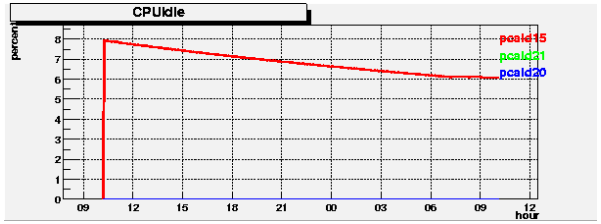


- DATE software
  - Operator console
  - State machines
  - Control of distributed system
- Home-made development based on free software

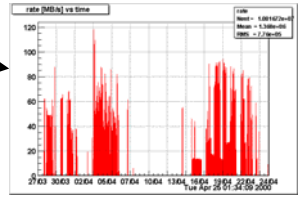
Performance Monitoring - AFFAIR

- Home-made development
- Based on free software

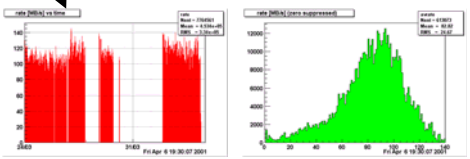




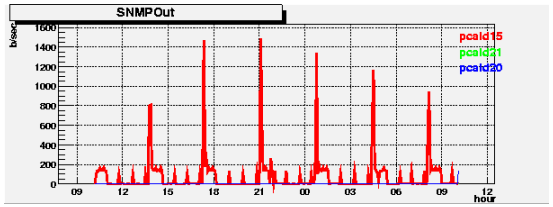
Fabric monitoring



DATE performances



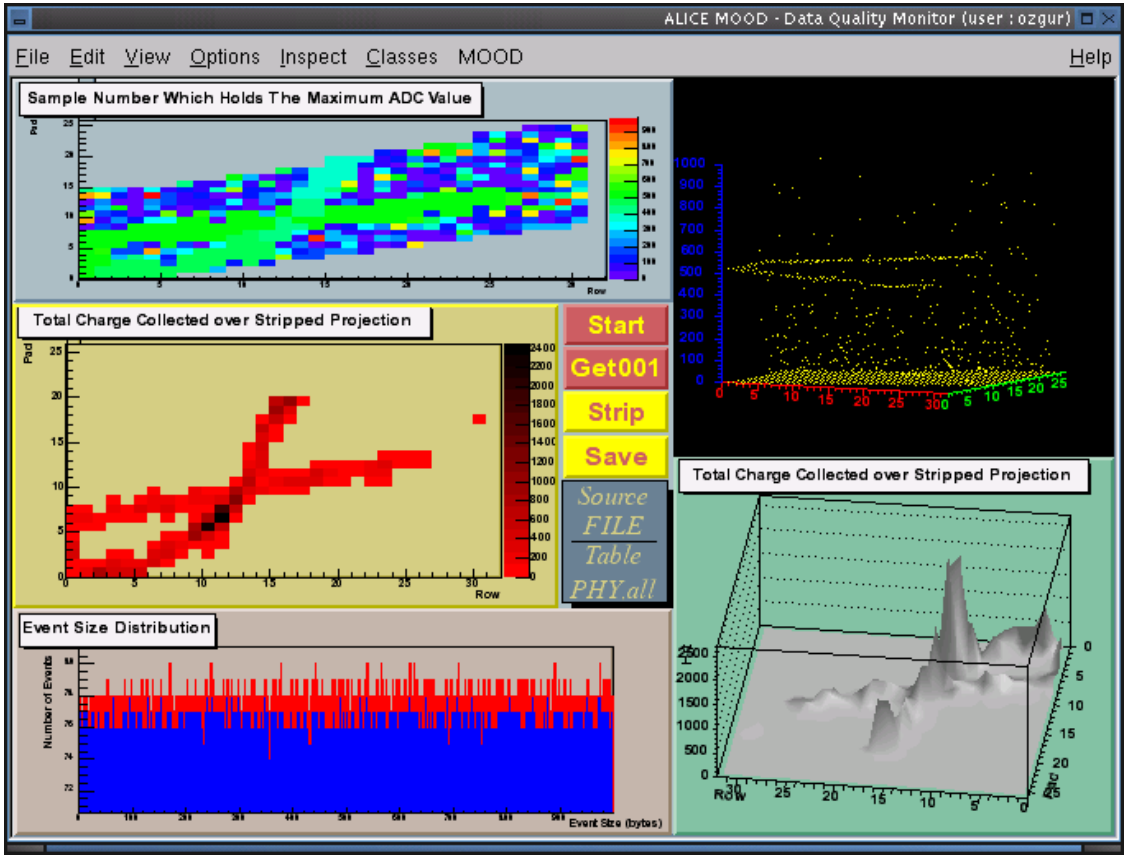
ROOT I/O performances



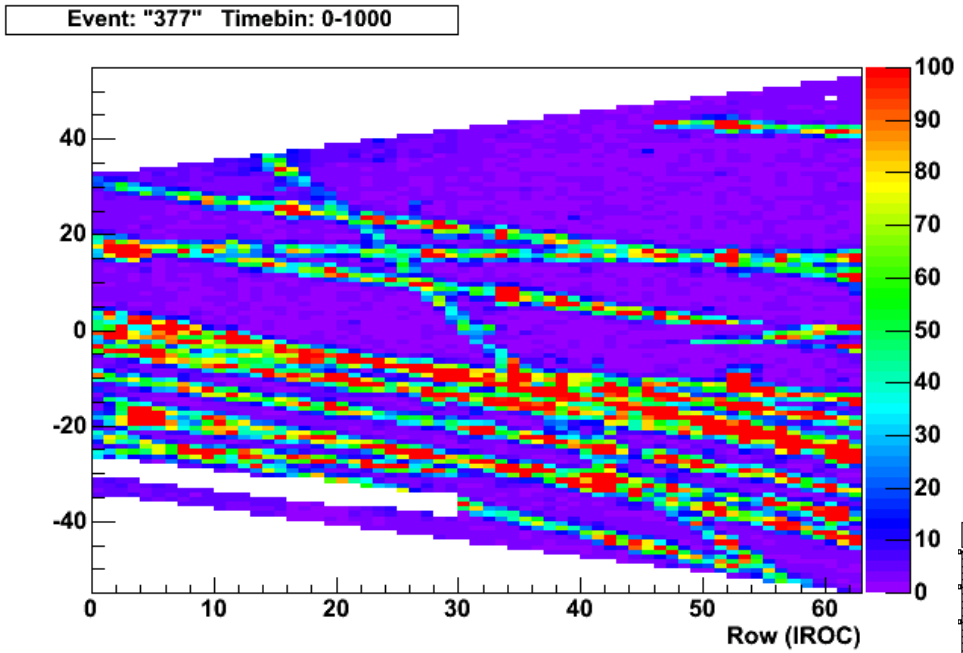
CASTOR performances

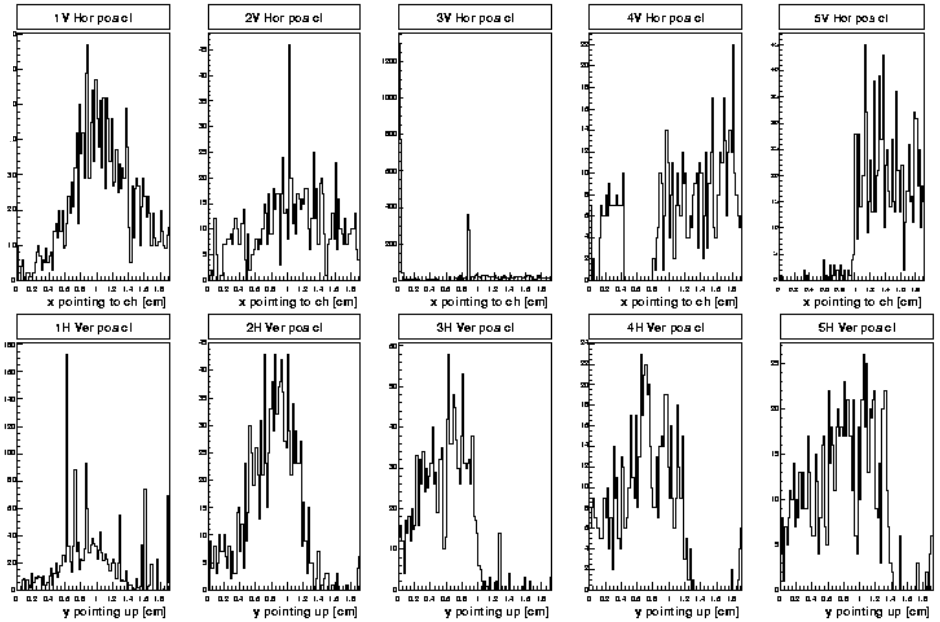
Data quality monitoring - MOOD

- MOOD framework
  - Interfaces to detector code
  - Software development in all institutes
- Applications:
  - Raw data integrity
  - Detector performance



Data quality monitoring - MOOD





Conclusions

- TRG and DAQ for HEP: huge performance needs
- Hardware
  - Needs achieved with moderate budget by intensive usage of commodity equipment
  - Special needs (higher performance or special environment): home-built development
- Large software development
  - Home-development:
    - Special applications
    - Flexibility
    - Involvement of institutes