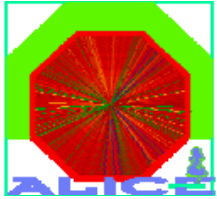


Challenges of Data Acquisition at the LHC

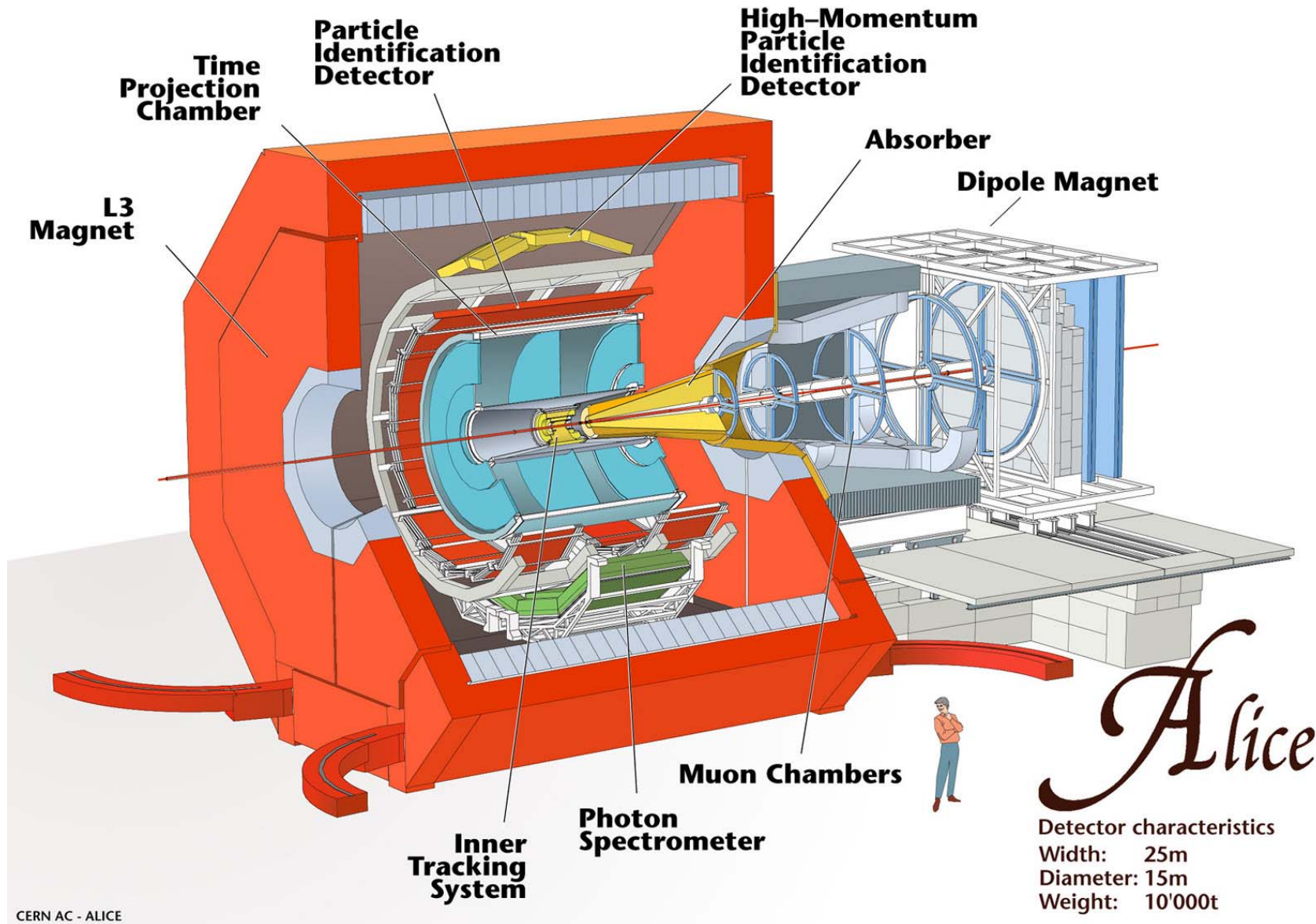
UK DTI Delegation
30-June-2004

P.Vande Vyvre - CERN/PH

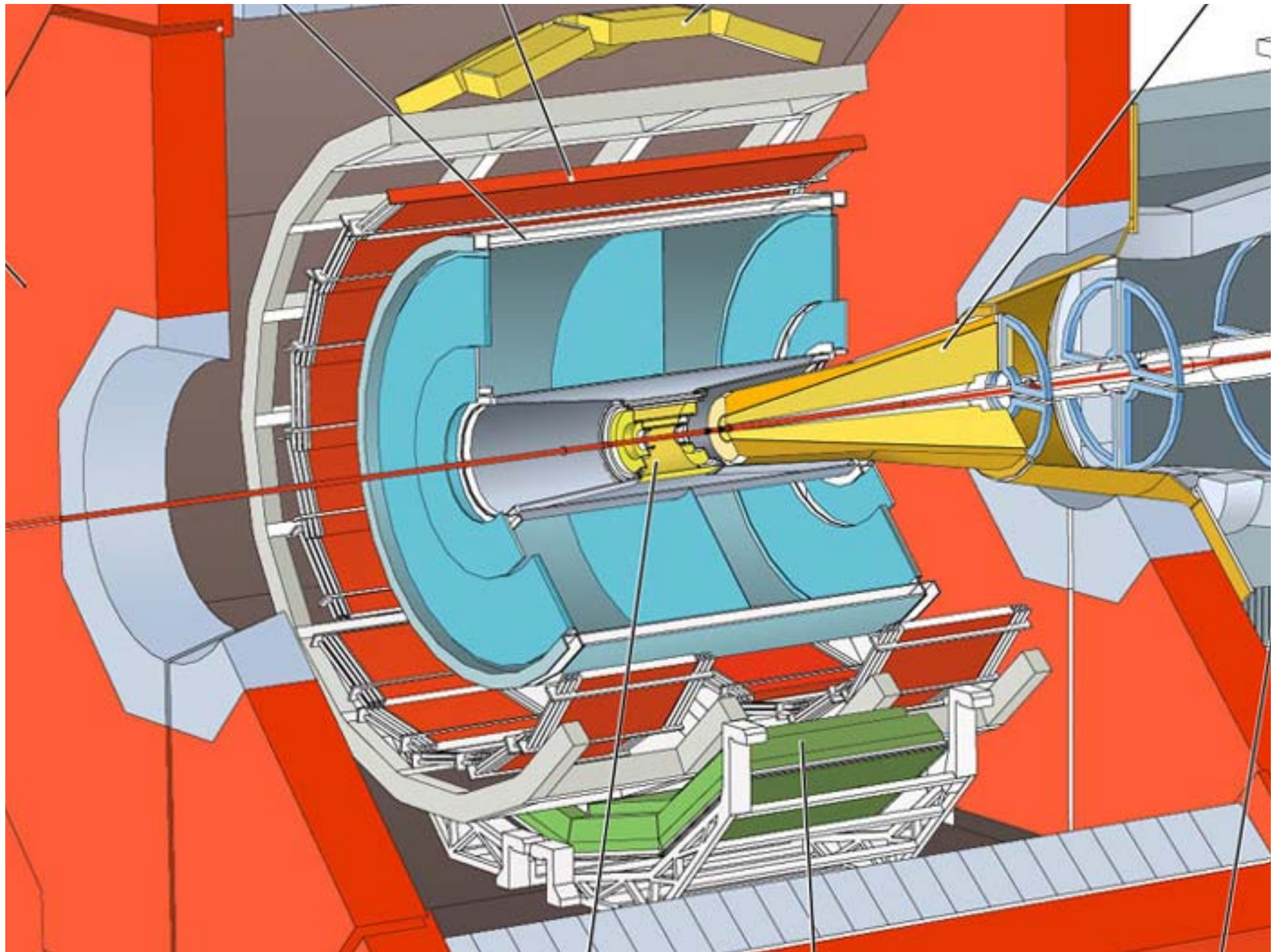


Outline

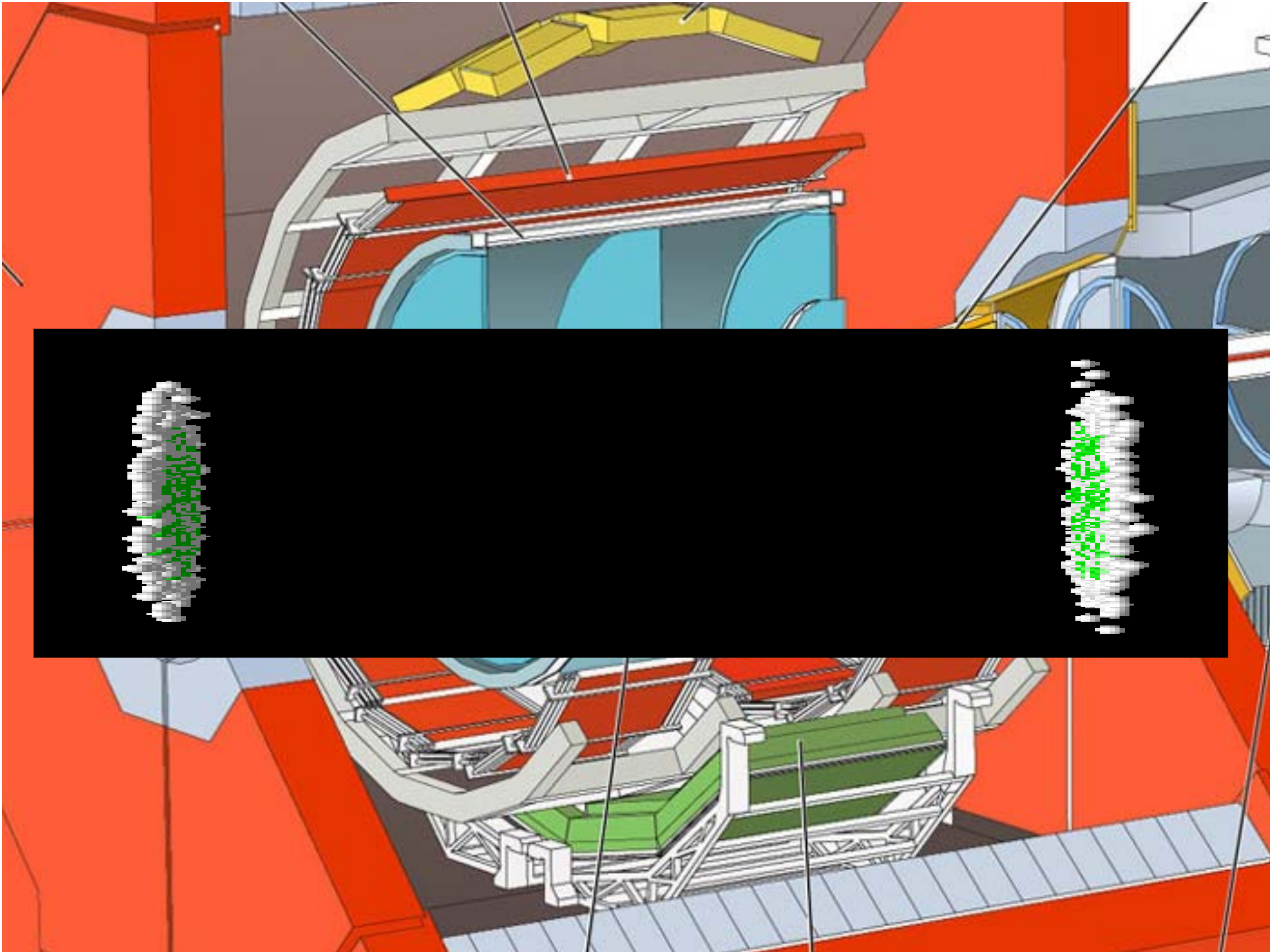
- Trigger and Data Acquisition
 - Main principles
 - Challenge at LHC
- Trigger and DAQ
 - Logical model
 - Main elements:
 - Baseline
 - Selection criteria
 - Industry involvement
- Conclusions

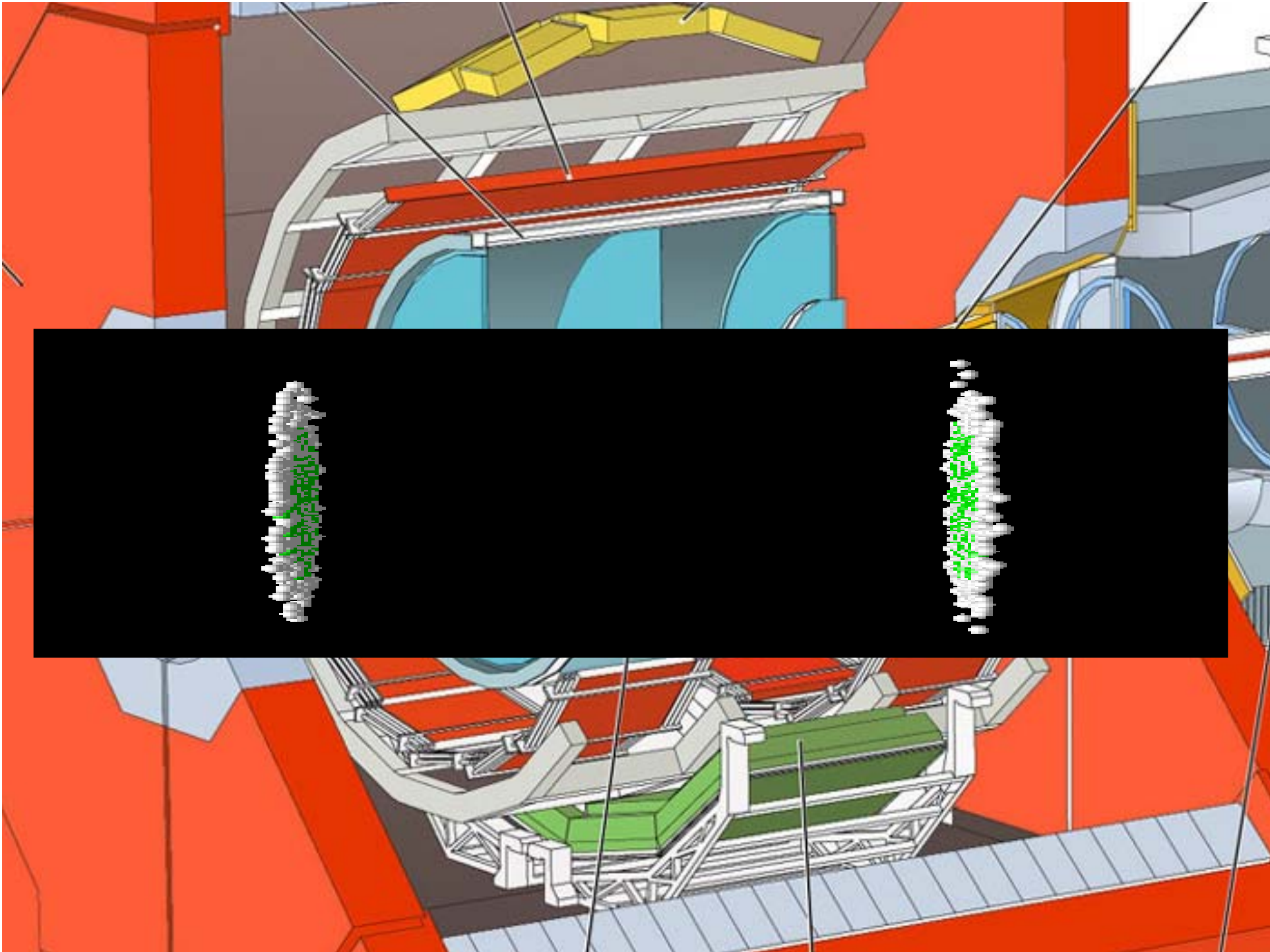


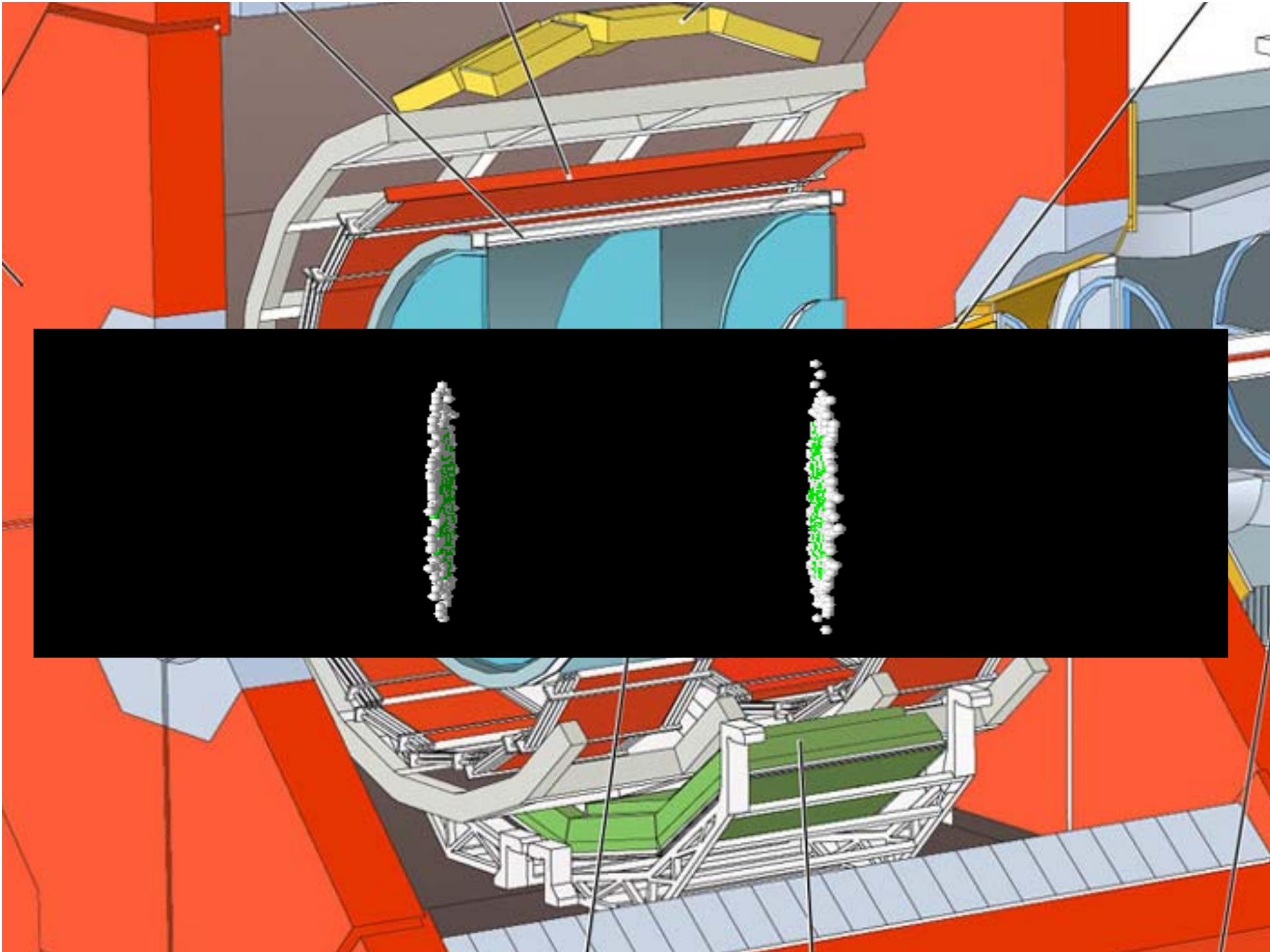
CERN AC - ALICE

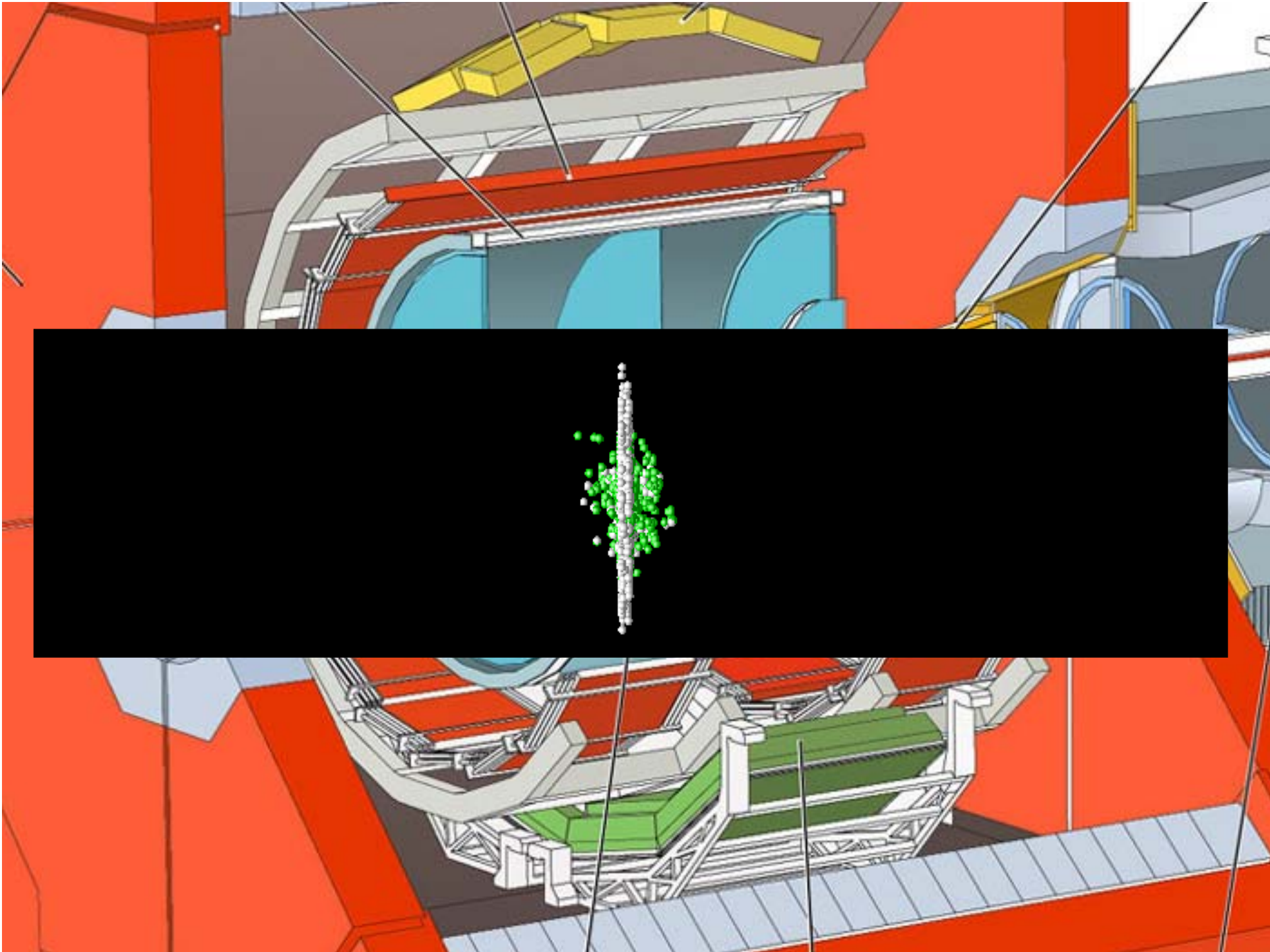


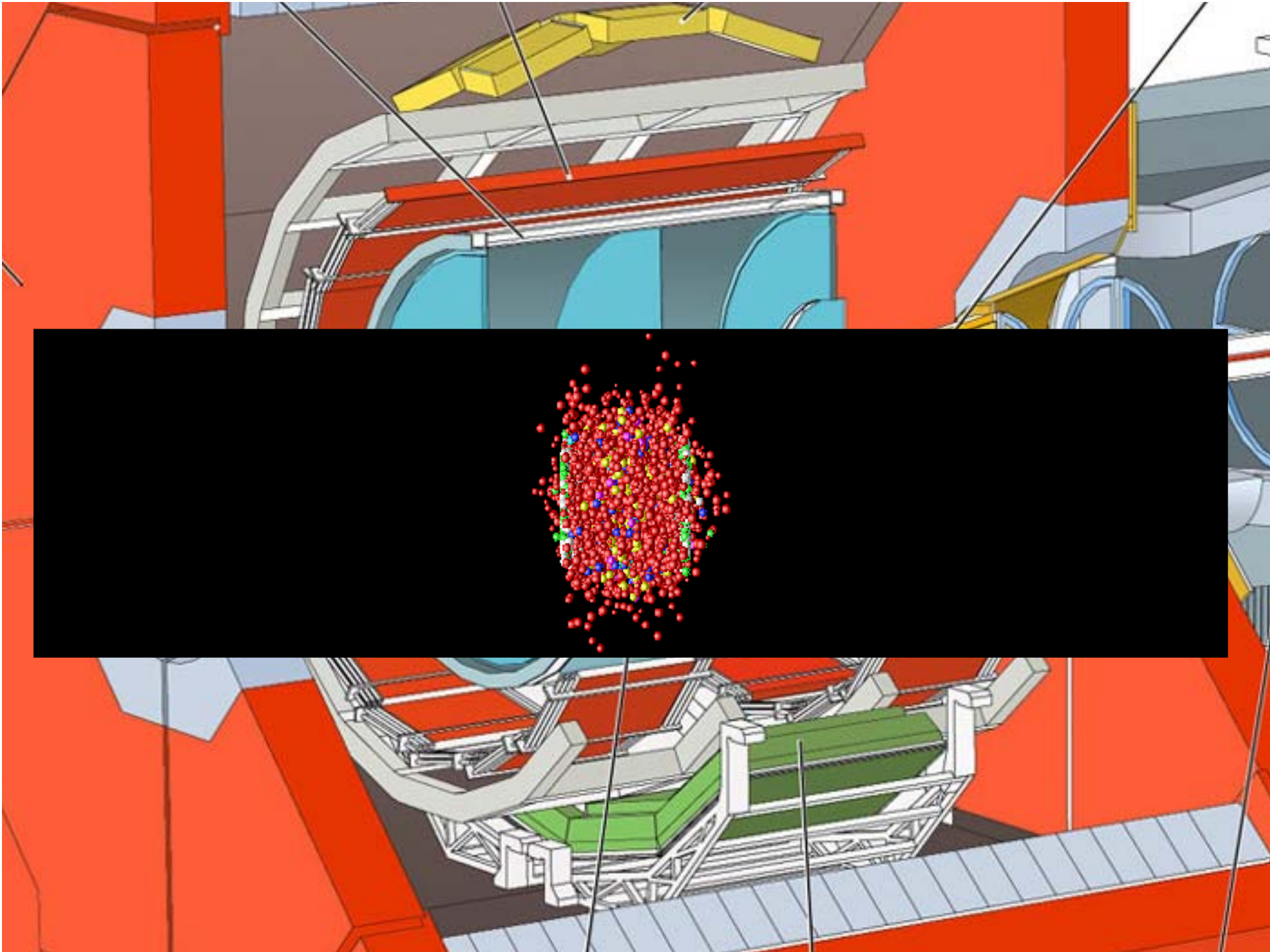


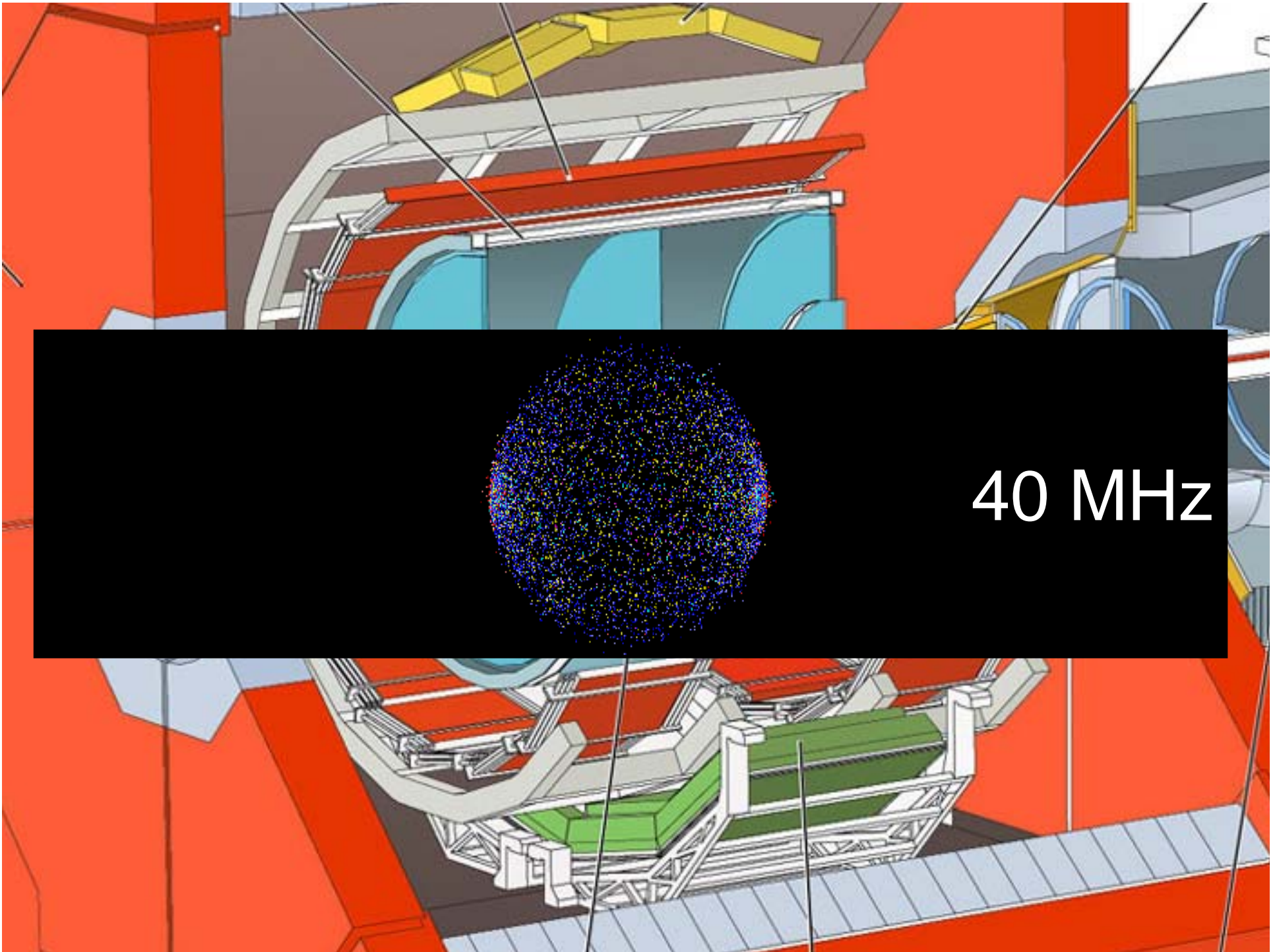








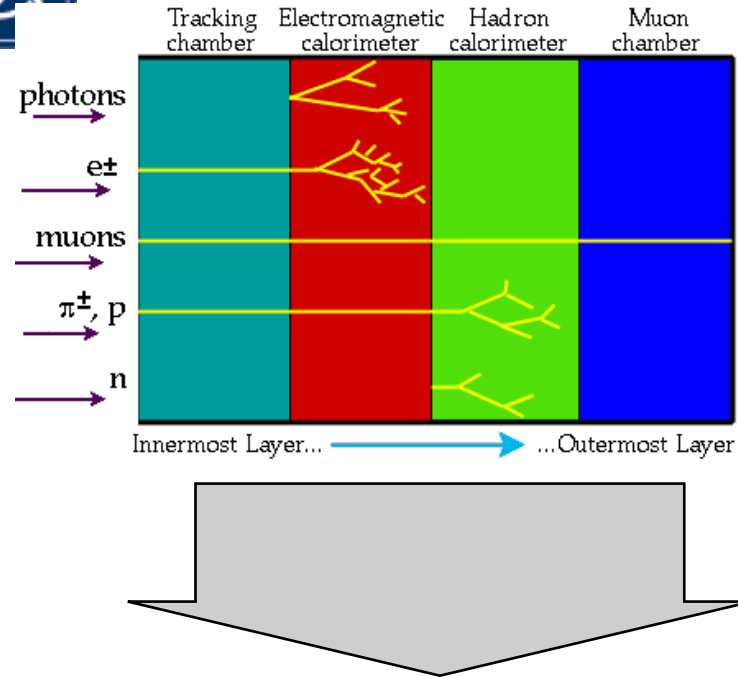




40 MHz



Trigger

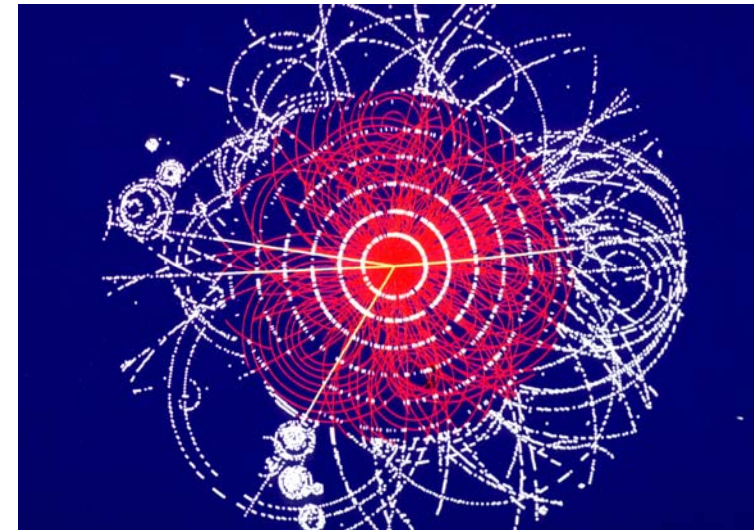
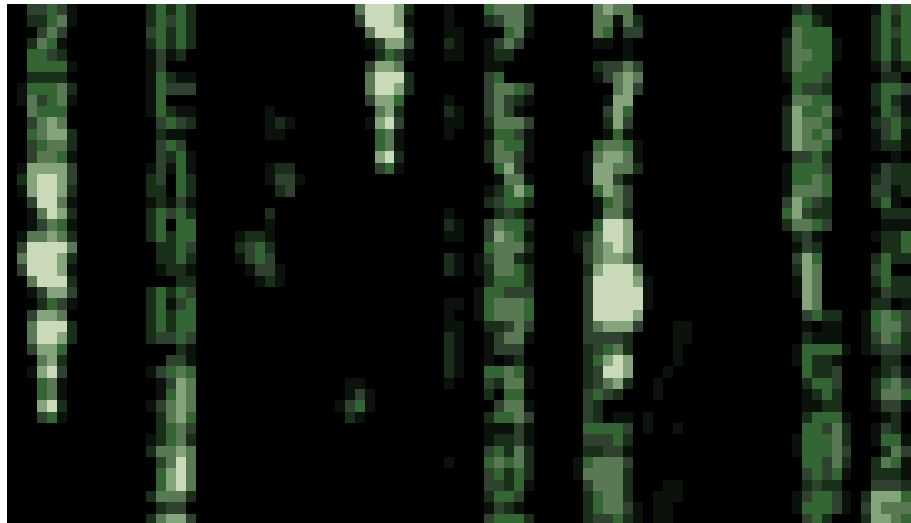


Multi-level trigger system

Reject background

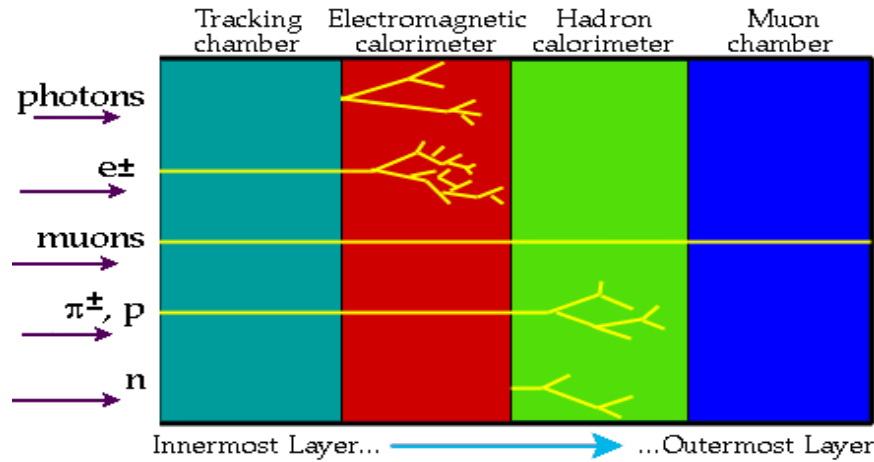
Select most interesting collisions

Reduce total data volume

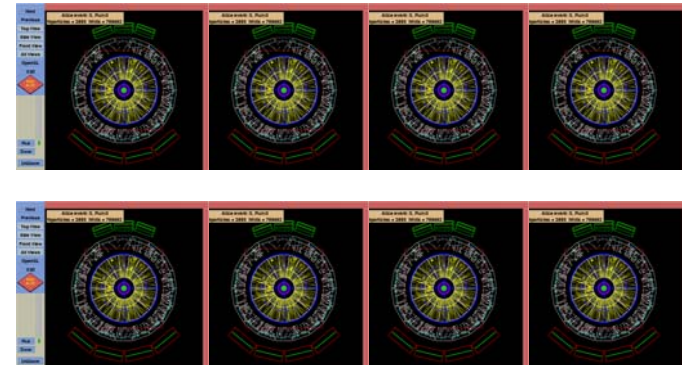
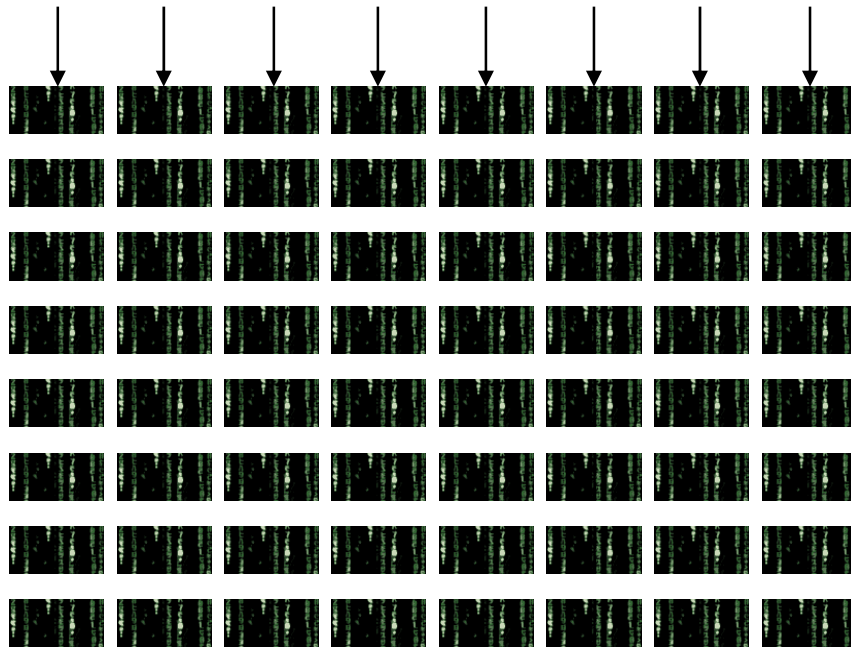




Data Acquisition (DAQ)

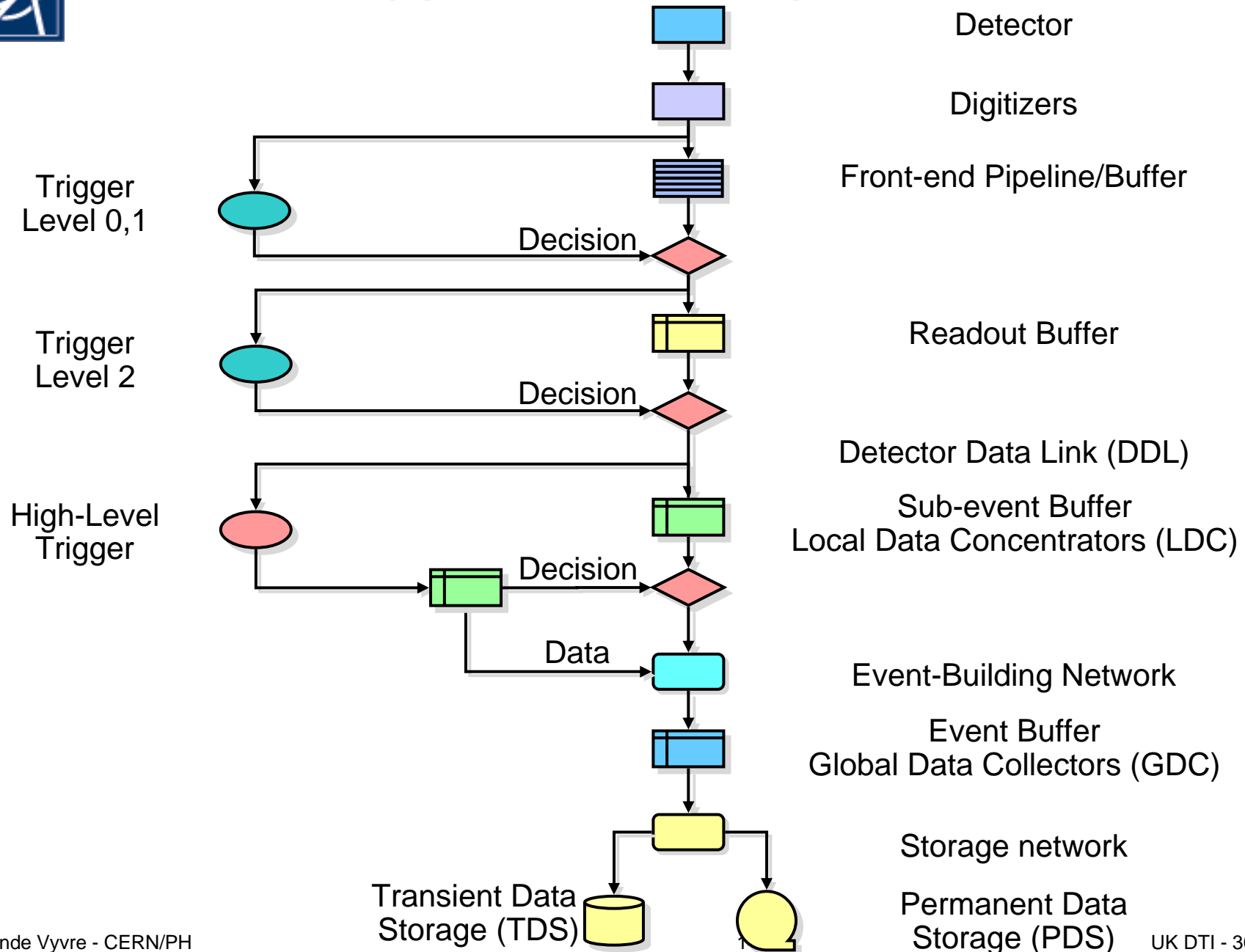


Acquire data from 1000's of sources
Reassemble all the data pertaining to the same physics event
Control the data taking
Monitor the data quality



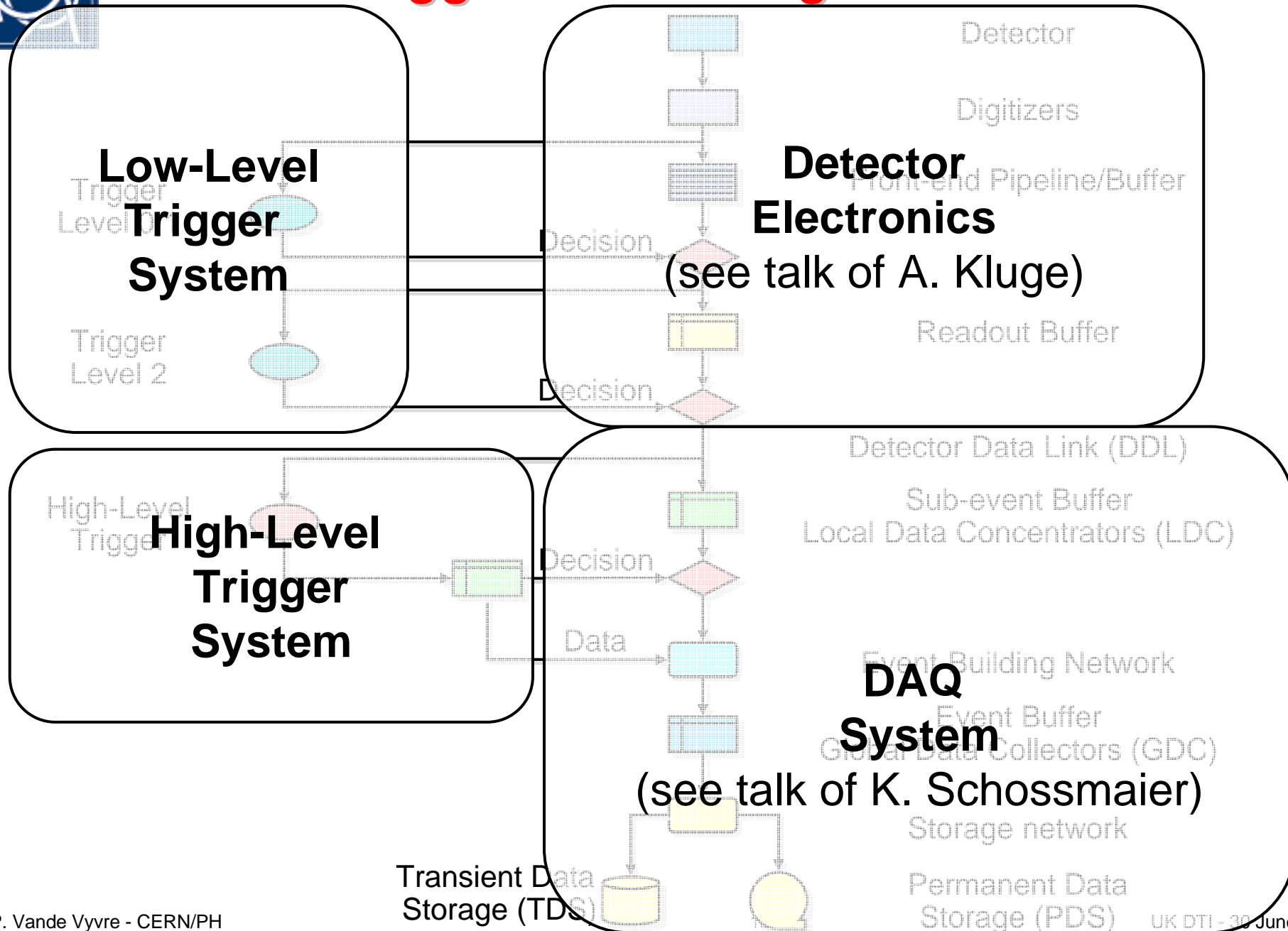


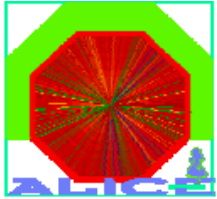
Trigger & DAQ logical model



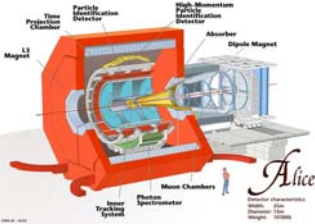
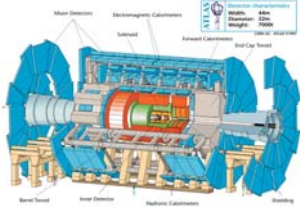
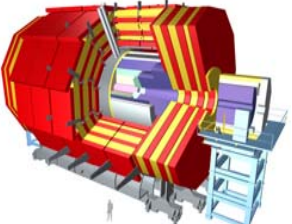
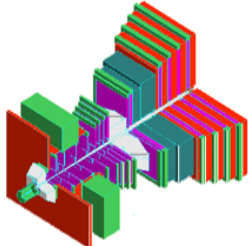


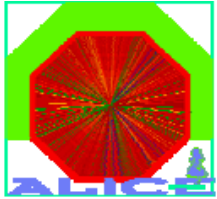
Trigger & DAQ logical model



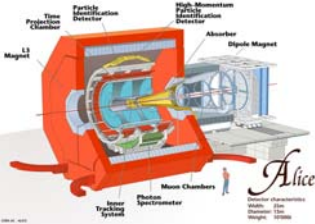
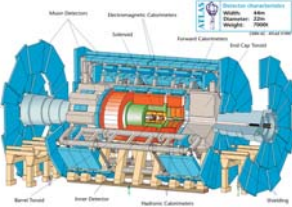
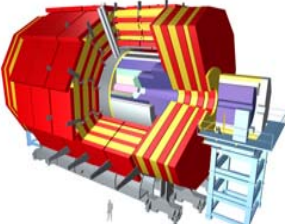
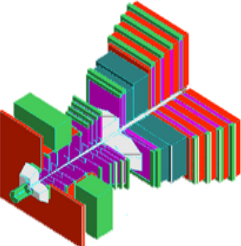


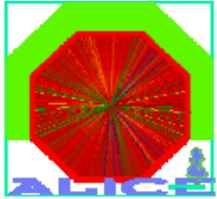
Trigger @ LHC

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

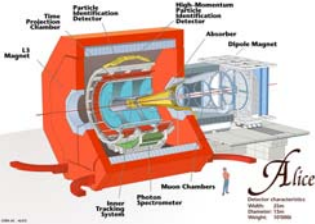
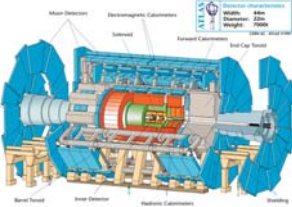
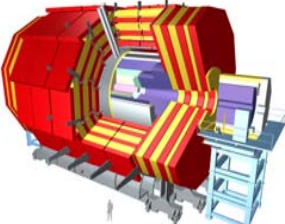
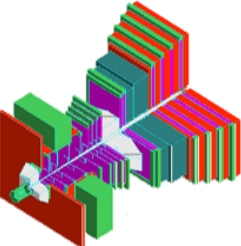


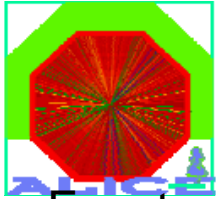
DAQ @ LHC

		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

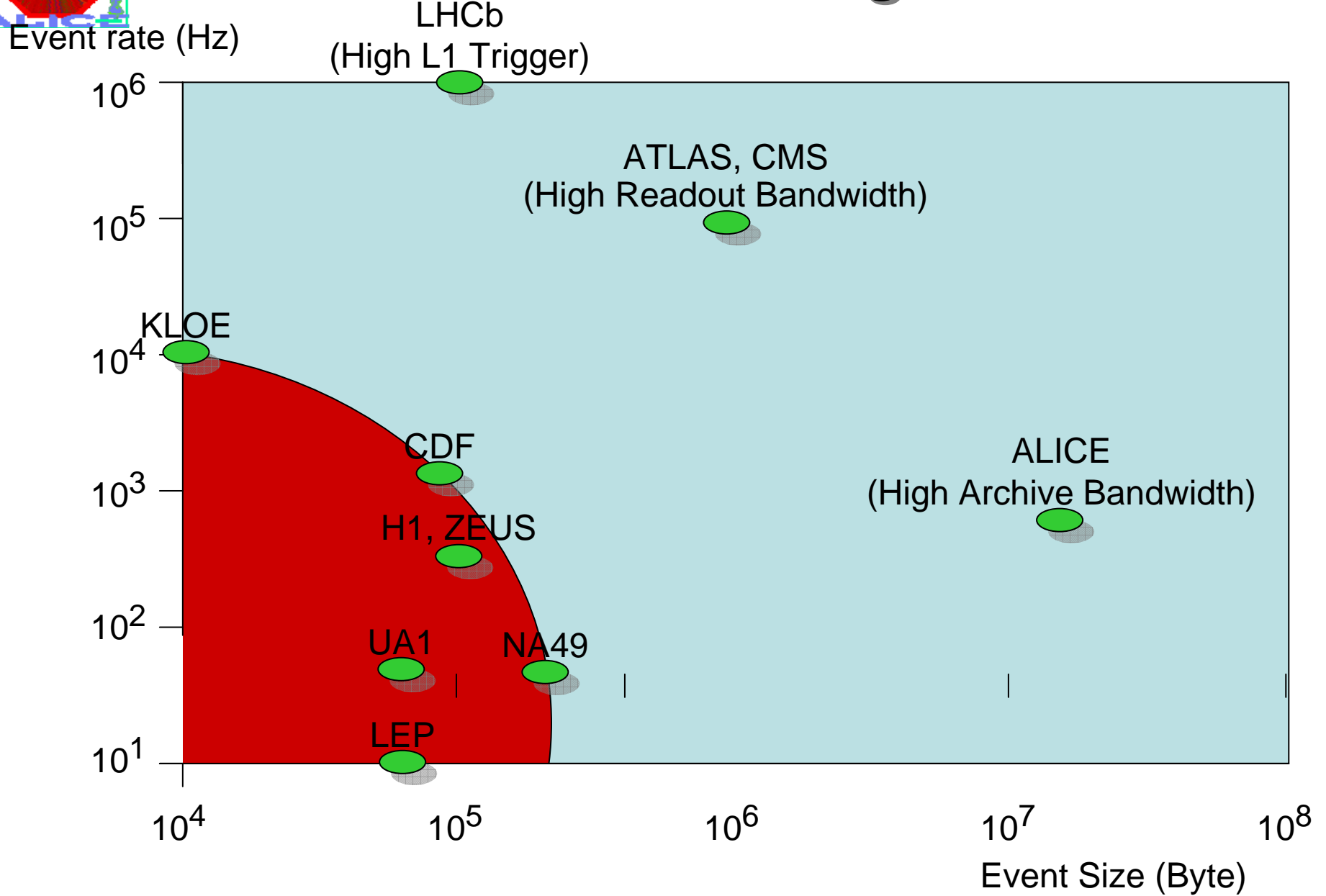


Mass Storage @ LHC

		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



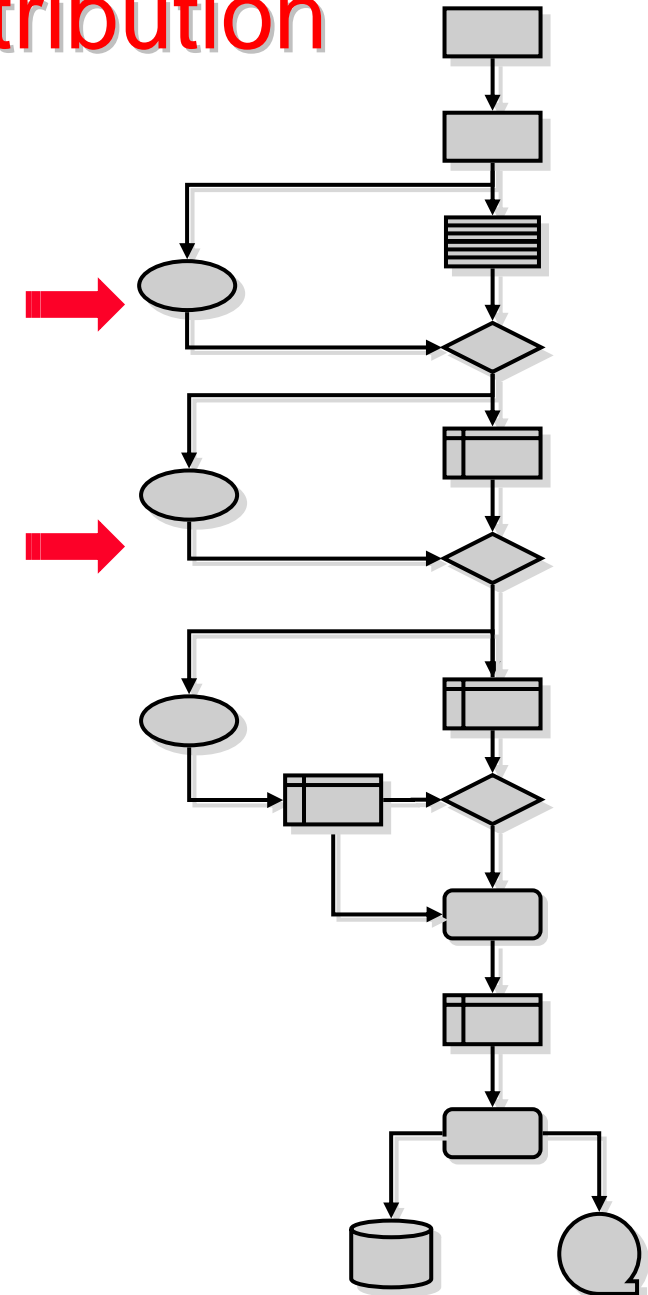
LHC Challenge





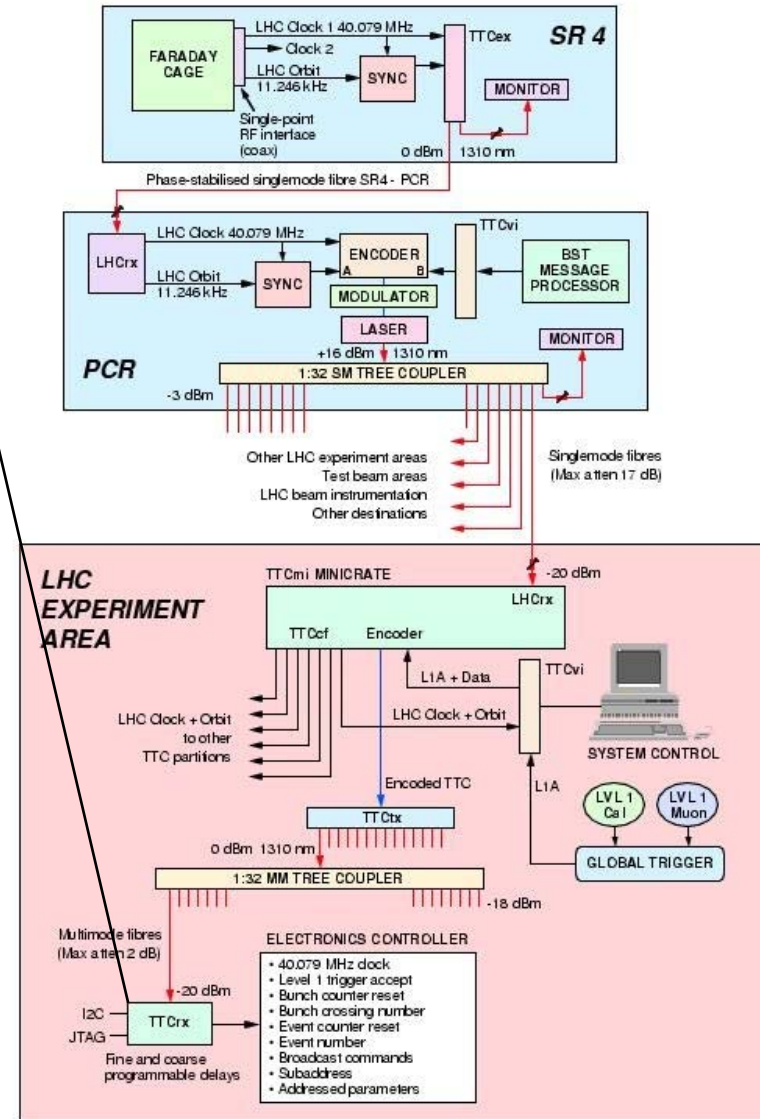
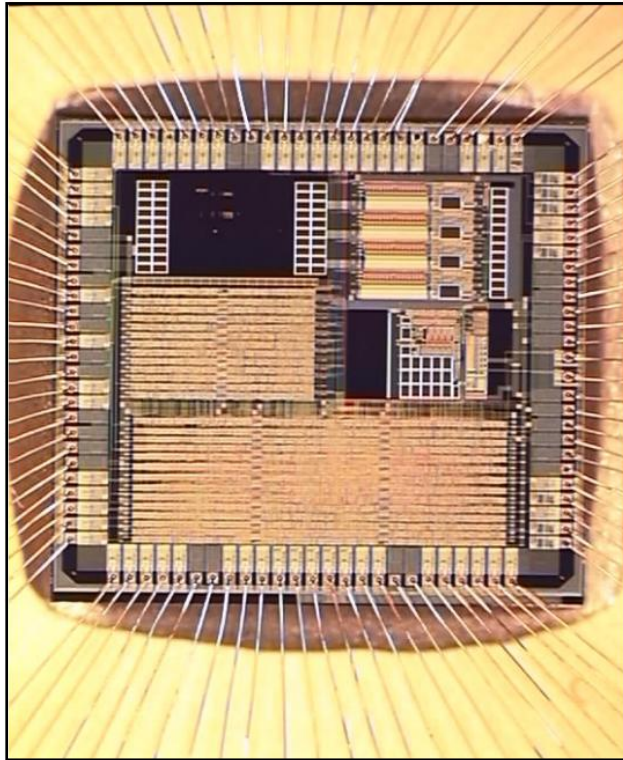
Trigger & Timing distribution

- ◆ 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





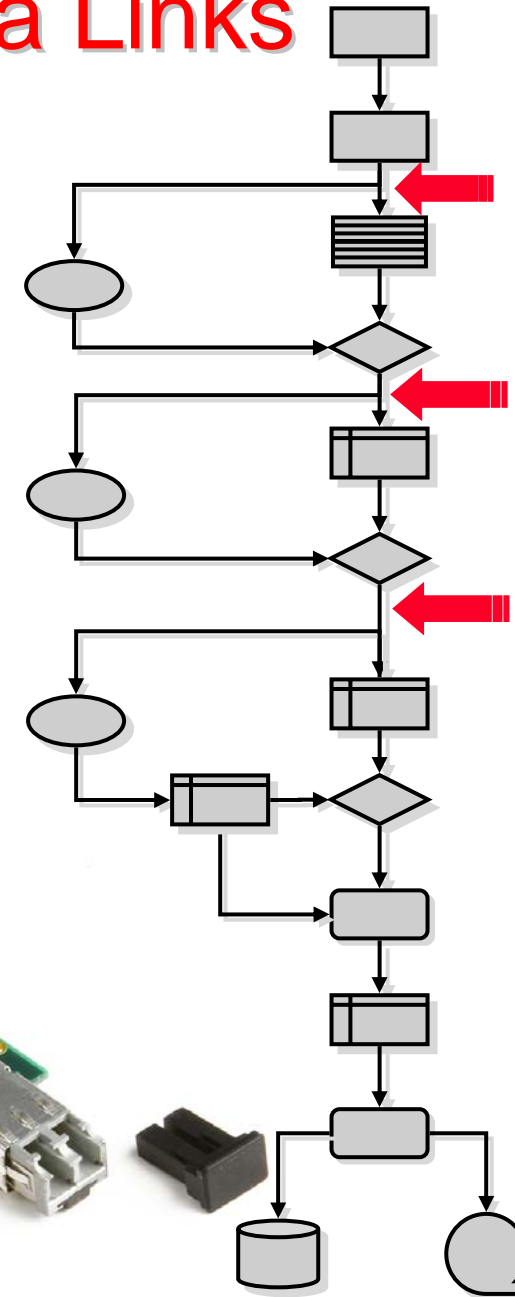
LHC Trigger & Timing distribution





Detector & Readout Data Links

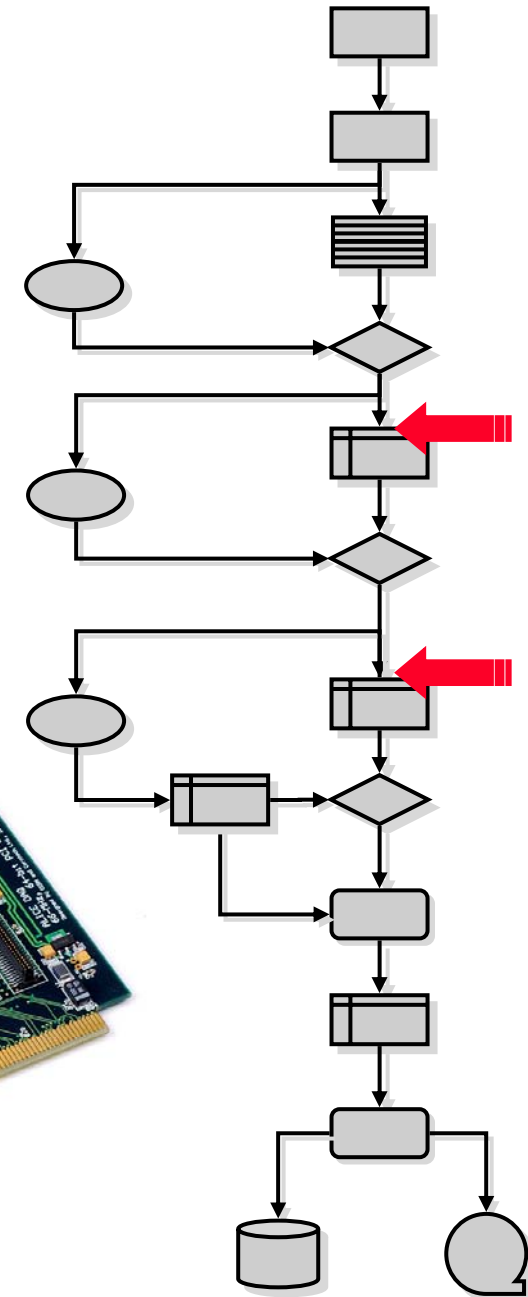
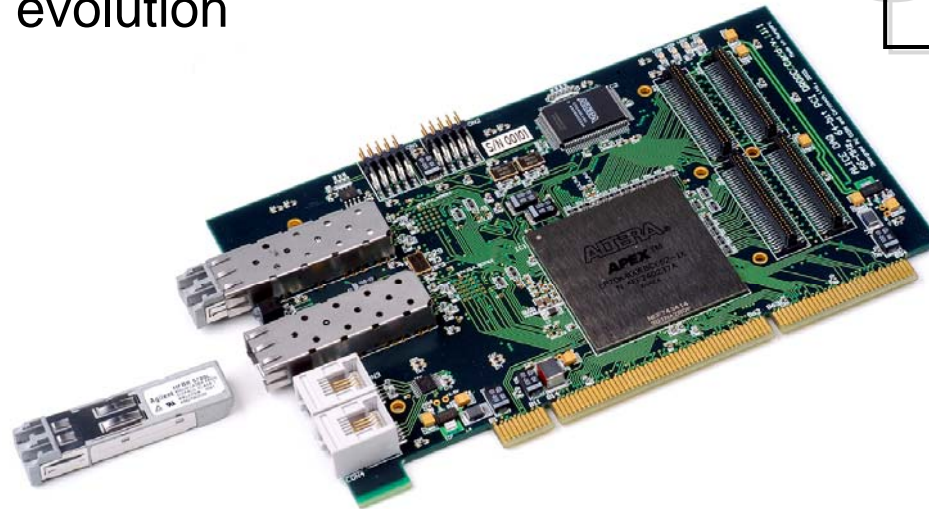
- ◆ Interface and data-transfer detector/DAQ
- ◆ Point-to-point
- ◆ Massive parallelism (100's to 1000's)
- ◆ Analog: HEP-specific components
- ◆ Digital
 - HEP developments based on commodity components
 - ASIC (Radiation hard) or FPGA (Radiation tolerant)
 - Fiber Channel or Gig. Ethernet: 1, 2.1 or 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

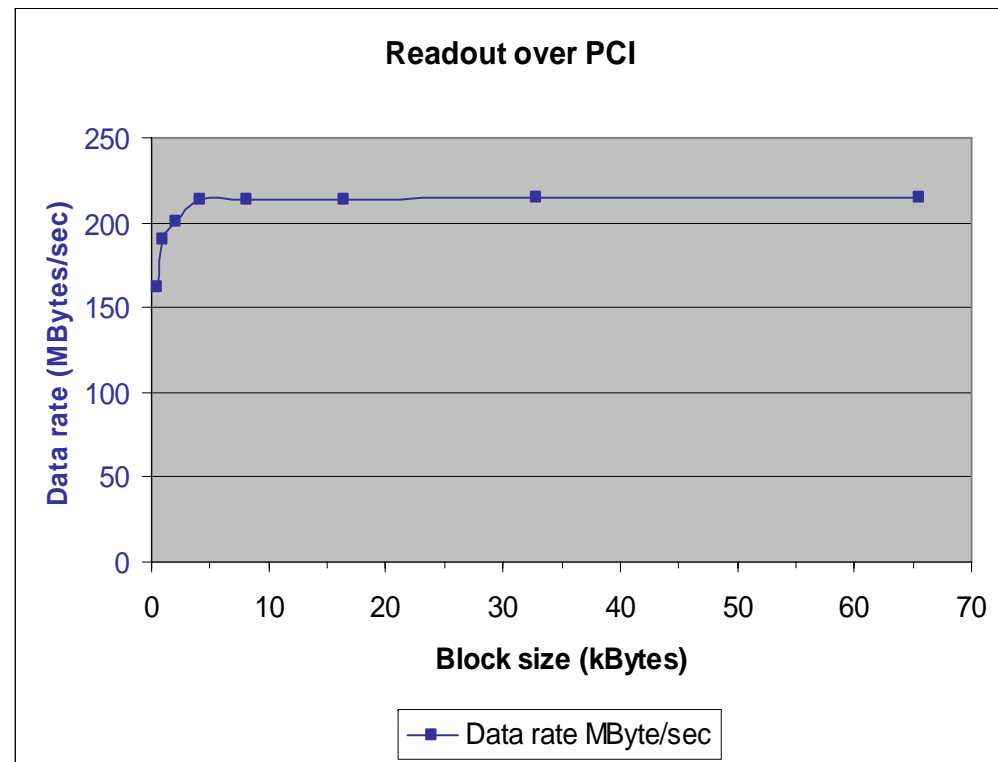
200 MB/s sustained

Total PCI load: 92 %

Data transfer PCI load: 83 %

Lots of bw available.

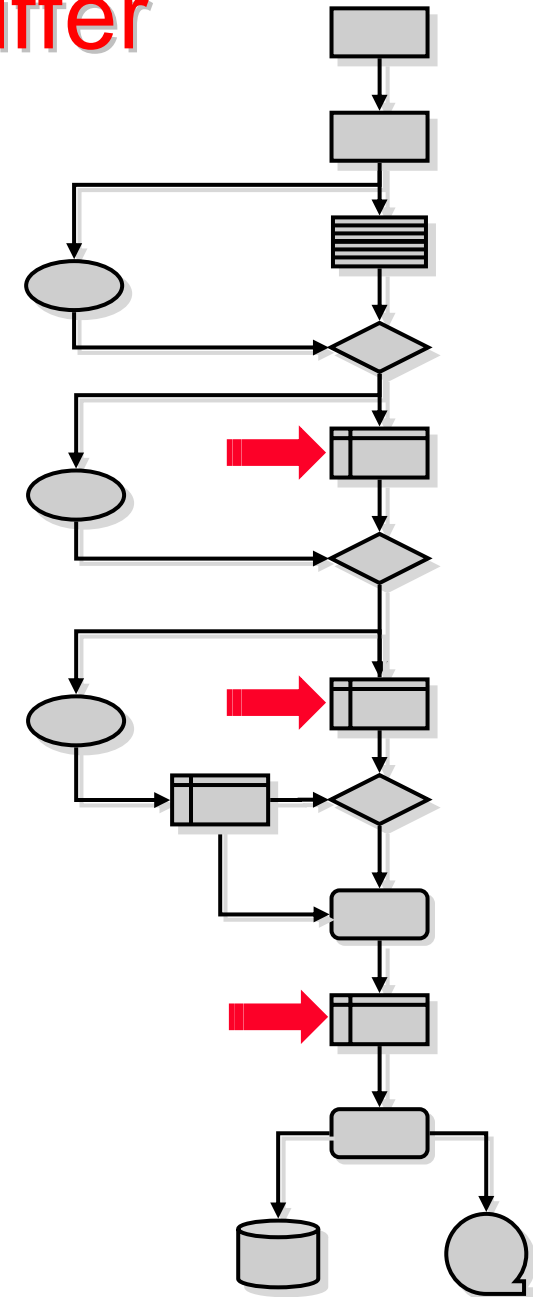
Major fraction available to end application.





Subevent & event buffer

- ◆ **Baseline:**
 - Fast dual-port memories
 - Electronics racks are over
 - Extensive use of dual-CPU 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



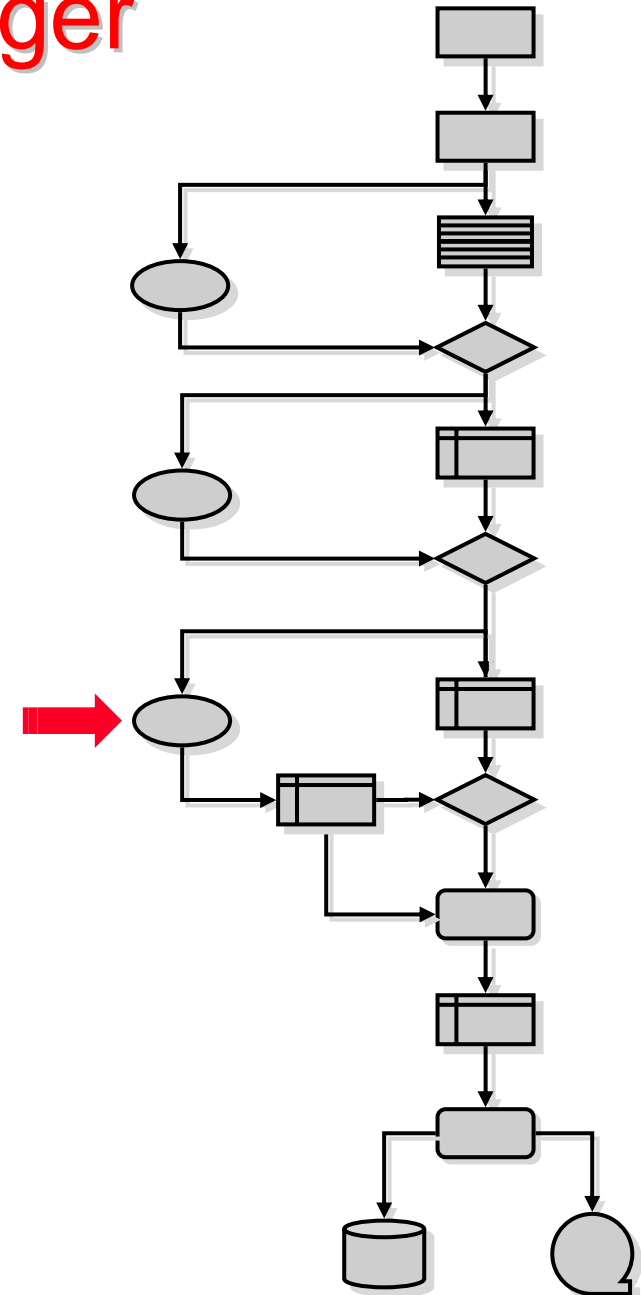


High-Level Trigger

- ◆ **Baseline:**
 - Function: fast dual-port memories and data processing
 - Dual-CPU 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

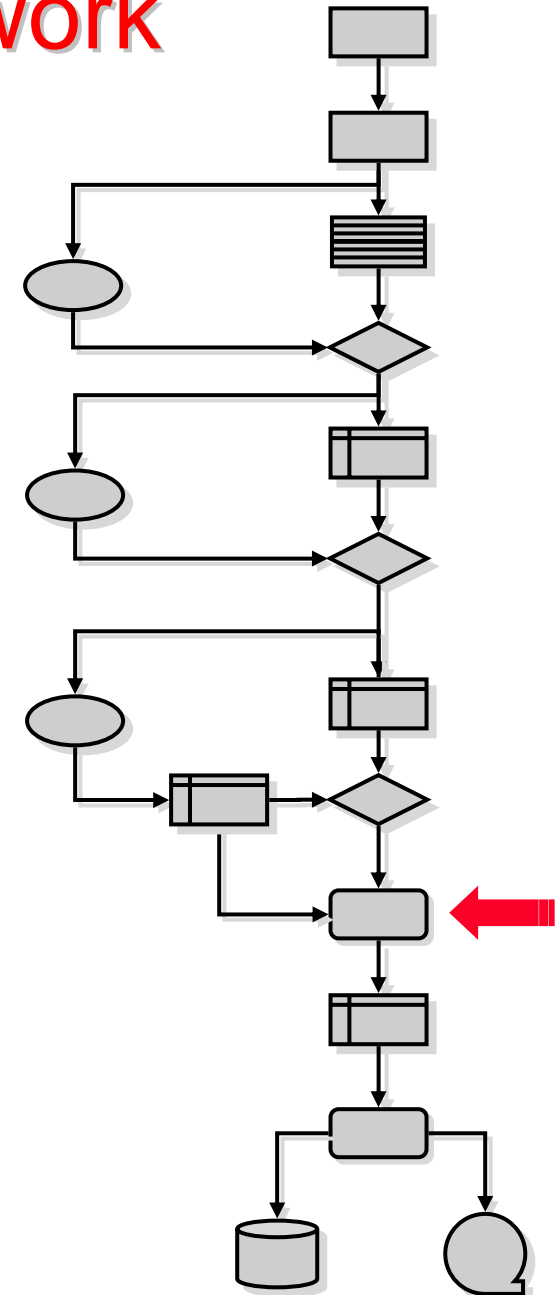
- ◆ **Baseline:**
 - Adopt broadly exploited standards (Switched Ethernet)
 - Adopt a performing commercial product (Myrinet)

- ◆ **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

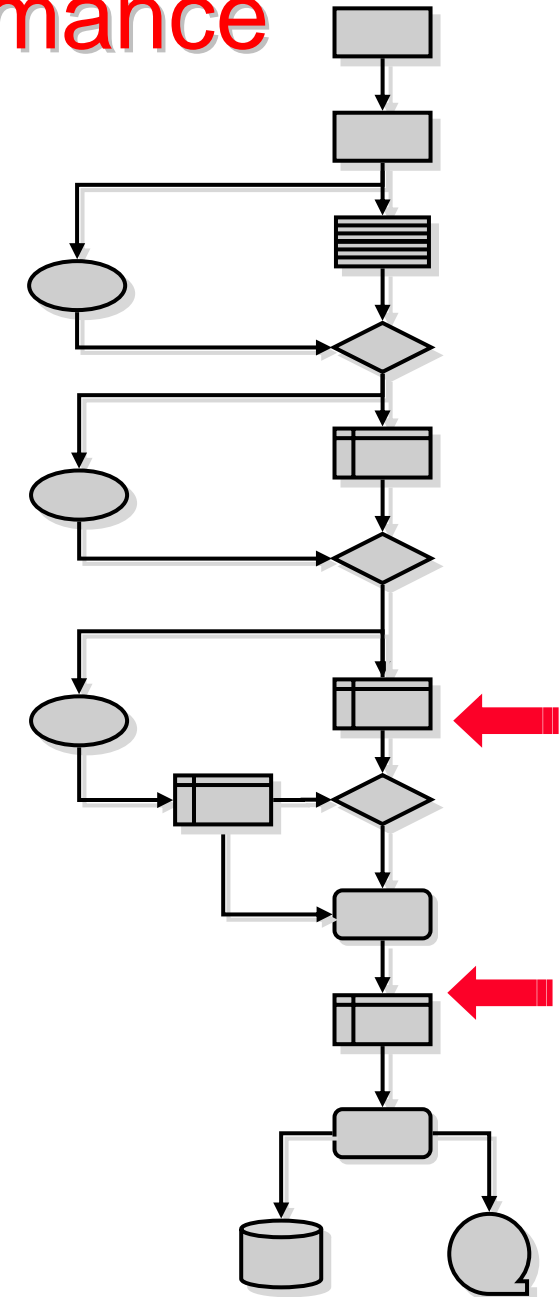
- ◆ **Future: 40 or 100 Gbit/s network**





Ethernet NIC's Performance

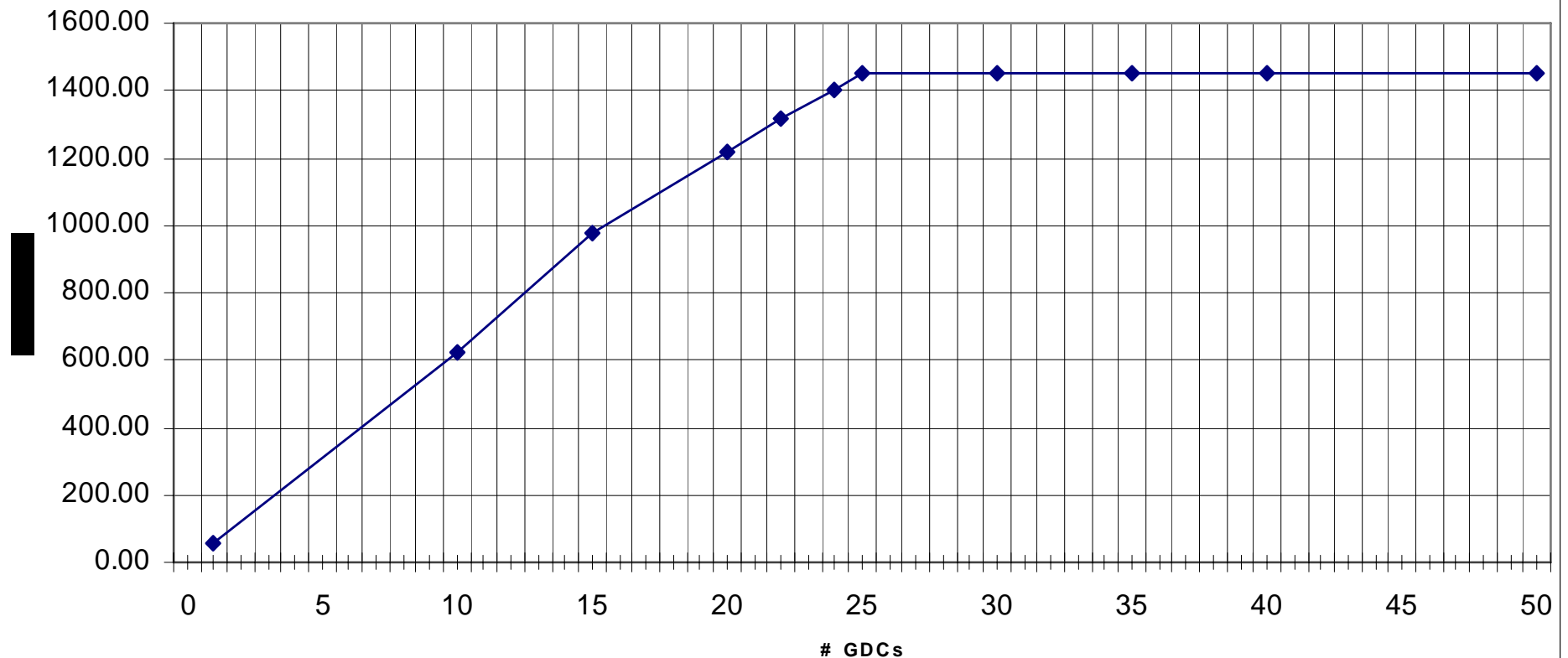
- ◆ Gigabit Ethernet
 - New generation of PC motherboard includes 2 Gbit Eth ports
 - ◆ Active market with several players
 - ◆ 3Com, Broadcom, Intel, NetGear
 - ◆ Fast evolution since 3 years
 - ◆ BW: from 50 to 110 MB/s
 - ◆ CPU usage: 150 to 60 %
- ◆ TCP/IP Offload Engine (TOE)
 - ◆ Dedicated processor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





Scalability of network-based event building

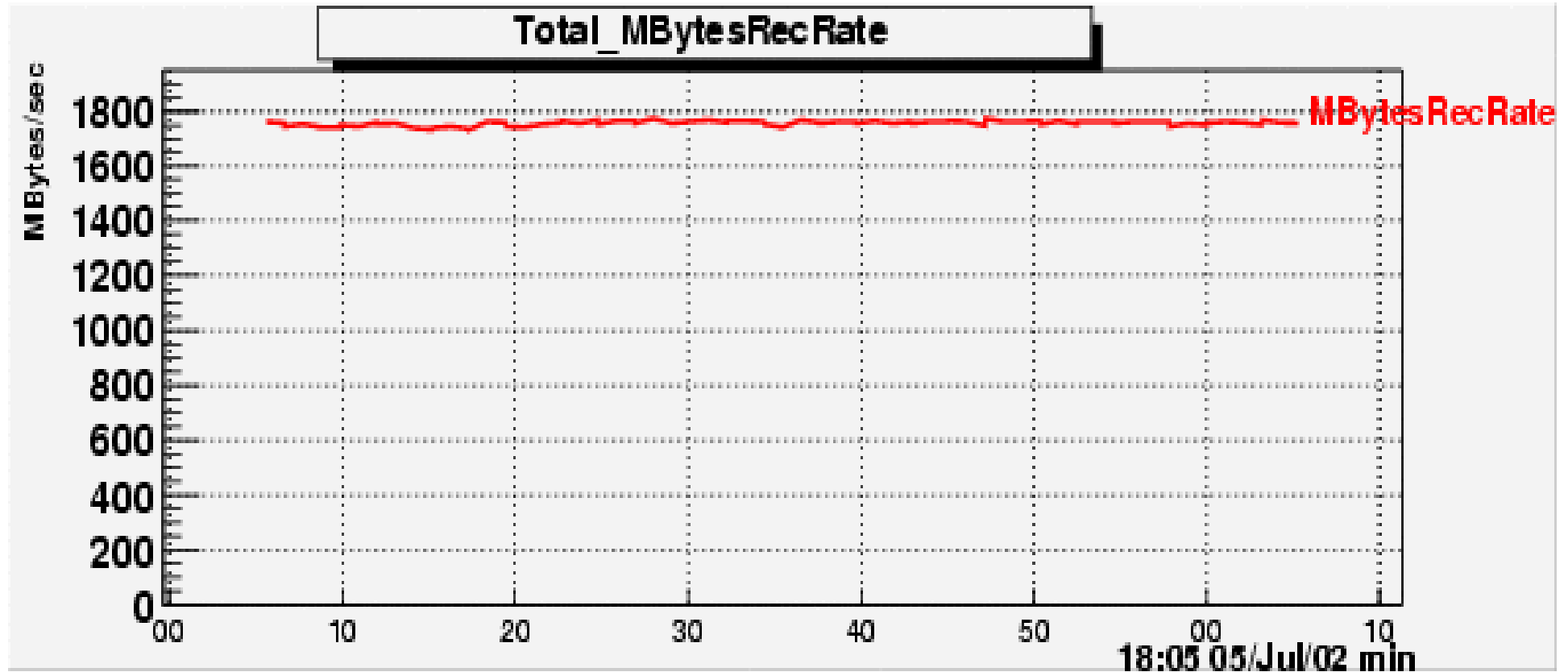
DATE COLE EQUIPMENT FLAT, 1 MB events, 21 LDCs



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



Performance of network-based event building

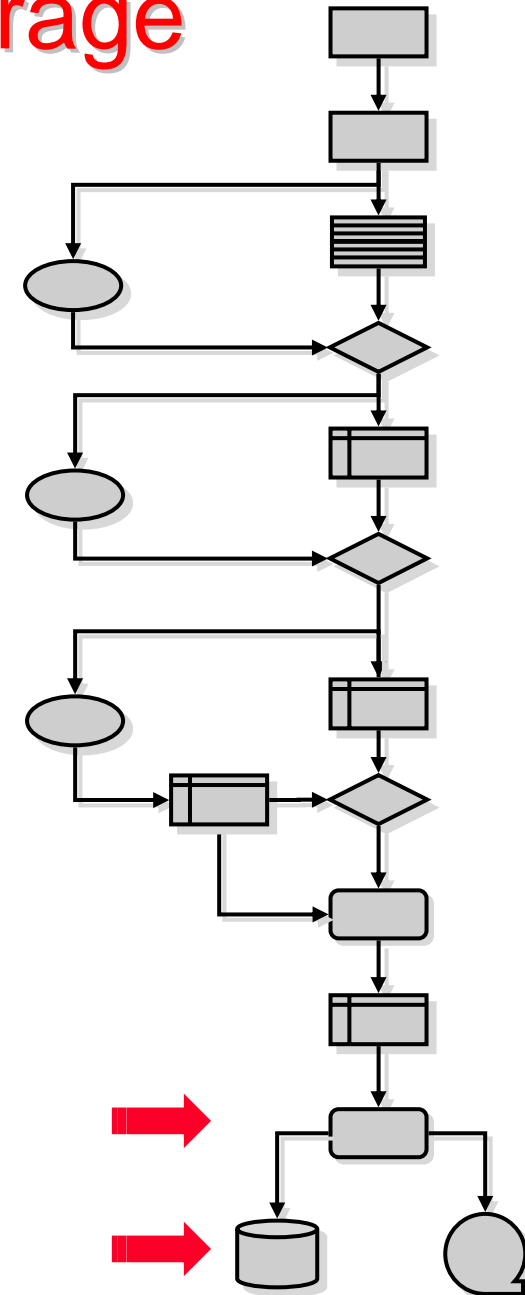


- 5 days non-stop
- 1750 MBytes/s sustained (goal was 1000)



Transient Data Storage

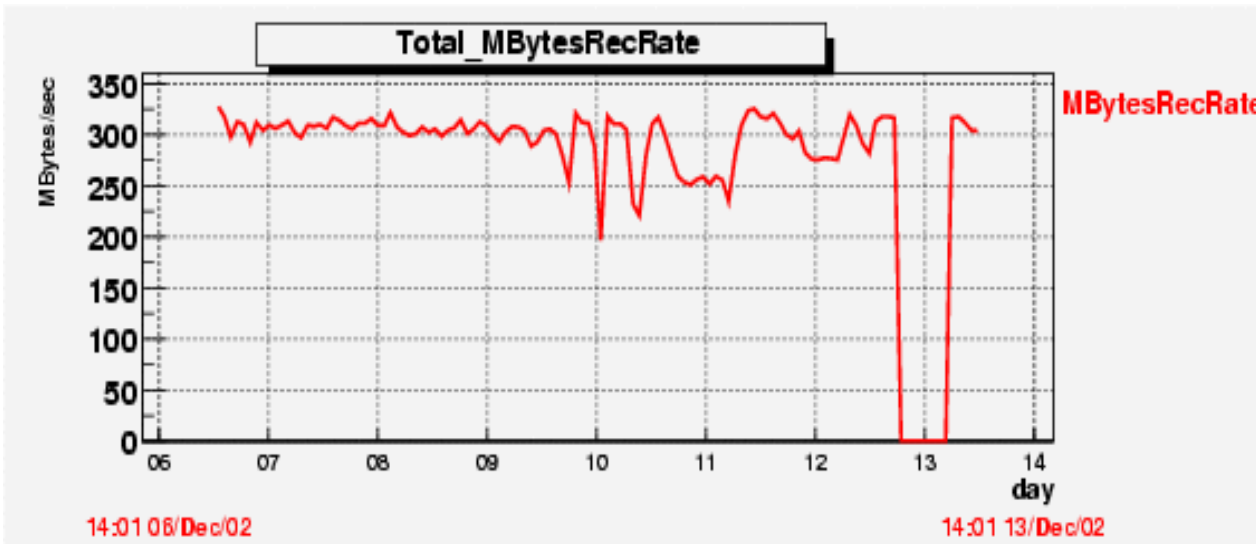
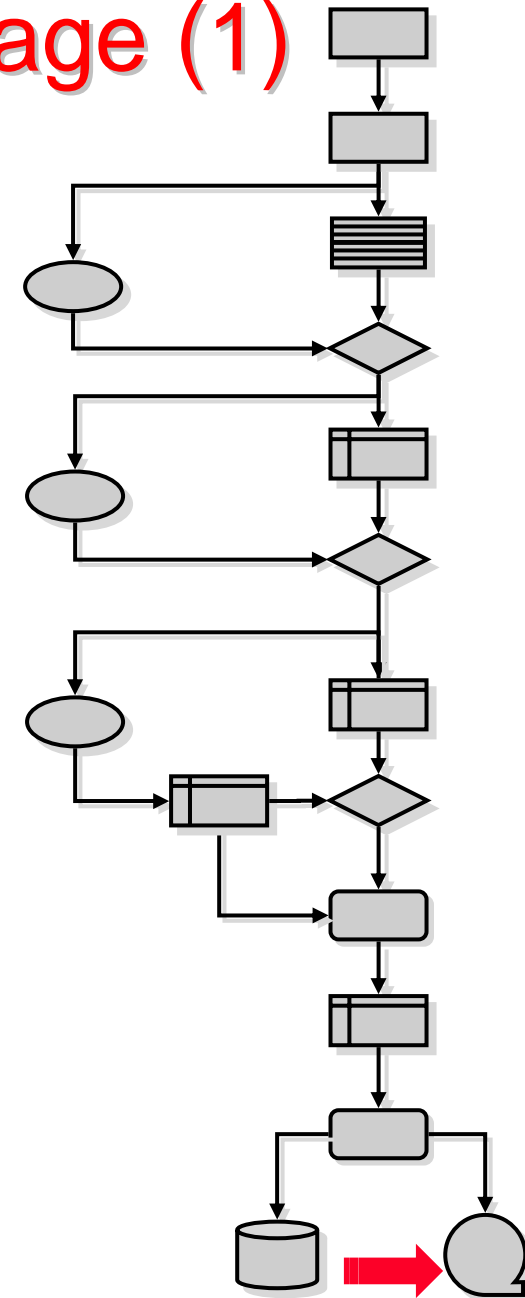
- ◆ Transient Data Storage
- ◆ Before archiving to tape, if any
- ◆ Baseline
 - Storage arrays of commodity disks
 - Box attachment: Fibre Channel
 - Disk attachment: IDE or serial-ATA
 - RAID-level
- ◆ Partnering for test of equipment (See talk of K. Schossmaier)
- ◆ Key selection criteria:
 - Cost/performance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

DAQ - Run Control

DOMAIN: divia23073

Configuration
Run Parameters
Ready to start
Data Taking

Define

Define

Start run

Stop run

Show

Show

Run AutoStart

Autoset GDC

Recording Enabled

AFFAIR

EDM

ALIMDC

HLT

RUN NUMBER : 1785 DAQ Logic Engine Status : RUNNING

Info: Run 1785 running

Trace

Fri 13 11:07 Run 1785 running

Clear

Fri 13 11:07 Run number saved on /dateSiteAdc/configurationFiles/runNumber.config

Fri 13 11:07 Starting run 1785

Debug

Fri 13 11:07 * Message from tbed0029gdc: TRACE STOP_PROCESS: EVB 3223 has been killed as r

Fri 13 11:07 * Message from tbed0029gdc: ACTION End of run requested with error

Pause

Fri 13 09:10 * Message from tbed0049ldc: ERROR file /date/runControl/Linux/checkProc.sh problem

Bigger

Fri 13 08:05 Run 1784 running

Smaller

Fri 13 08:05 Run number saved on /dateSiteAdc/configurationFiles/runNumber.config

```

11:21am up 78 days, 22:29, 1 user, load average: 1.73, 1.69, 1.62
90 processes: 87 sleeping, 3 running, 0 zombie, 0 stopped
CPU0 states: 2.0% user, 50.5% system, 1.2% nice, 46.1% idle
CPU1 states: 3.0% user, 75.3% system, 2.0% nice, 21.0% idle
Mem: 384356K av, 374564K used, 9792K free, 3020K shrd, 147540K buff
Swap: 1044184K av, 26364K used, 1017820K free, 152456K cached
          
```

PID	USER	PRI	NI	SIZE	RSS	SHARE	STAT	%CPU	%MEM	TIME	COMMAND
15208	nobody	14	5	4080	4080	3644	R N	99.9	1.0	13:09	recorder
1334	root	9	0	2332	2284	1592	S	0.5	0.5	21:39	sshd
1574	root	9	0	1060	1060	820	R	0.3	0.2	30:05	top
3	root	19	19	0	0	0	SWN	0.1	0.0	13:47	ksoftirqd_CPU0
1337	root	9	0	2368	2364	1856	R	0.1	0.6	3:17	xterm
5070	nobody	8	0	4004	3976	1468	S	0.1	1.0	10:00	rcServer
1	root	9	0	496	448	448	S	0.0	0.1	0:12	init
2	root	8	0	0	0	0	SW	0.0	0.0	0:00	keventd
4	root	19	19	0	0	0	SWN	0.0	0.0	12:06	ksoftirqd_CPU1
5	root	9	0	0	0	0	SW	0.0	0.0	1:58	kswapd
6	root	9	0	0	0	0	SW	0.0	0.0	0:00	kreclaimd
7	root	9	0	0	0	0	SW	0.0	0.0	0:00	bdflush
8	root	9	0	0	0	0	SW	0.0	0.0	0:01	kupdated
9	root	-1	-20	0	0	0	SW<	0.0	0.0	0:00	mdrecoveryd
15	root	9	0	0	0	0	SW	0.0	0.0	0:00	scsi_ah_0
16	root	9	0	0	0	0	SW	0.0	0.0	0:00	scsi_ah_1

SD

LDC status display

LDC name	tbed0001ldc	tbed0013ldc	tbed0030ldc	tbed0037ldc
Event rate	13	13	14	13
Bytes recorded rate	40.182 M	41.203 M	41.938 M	40.163 M
Bytes in buffer	C 1192% M 1195%	C 1188% M 1193%	C 1192% M 1194%	C 1187% M 1191%
Number of events	10453	10462	10457	10450
Events recorded	9816	9825	9820	9813
Bytes injected	31'031'205'136	31'057'922'896	31'043'079'696	31'022'299'216
Bytes recorded	29'141'863'284	29'175'752'364	29'154'396'136	29'140'480'912
Readout SOR/EOR phases	0	0	0	0
Recorder SOR/EOR phases	0	0	0	0

GDC status display

GDC name	tbed0003gdc	tbed0004gdc	tbed0014gdc	tbed0015gdc
Events received	4924	5170	5438	3505
Events recorded	622	639	673	432
Bytes received	14'588'026'944	15'347'910'144	16'167'256'896	10'428'860'800
Bytes recorded	14'392'096'256	15'175'728'576	15'983'200'832	10'259'840'000
Event builder SOR/EOR phases	0	0	0	0
Status	FULL	FULL	FULL	

EDM status display

EDM name	tbed0015edm
wakeUpId received	(nblnRun:10442)
maxWakeUpId	(nblnRun:10442)
lastThresholdSent	(nblnRun:10454)
lastUpperBoundSent	(nblnRun:10464)
edmMask	[0]:00040000 [1]:00000100
Excluded	3 4 14 26 29 41 50 51 64 65 74 75 96 97

```

11:21am up 78 days, 22:29, 1 user, load average: 1.73, 1.69, 1.62
90 processes: 87 sleeping, 3 running, 0 zombie, 0 stopped
CPU0 states: 2.0% user, 50.5% system, 1.2% nice, 46.1% idle
CPU1 states: 3.0% user, 75.3% system, 2.0% nice, 21.0% idle
Mem: 384356K av, 374564K used, 9792K free, 3020K shrd, 147540K buff
Swap: 1044184K av, 26364K used, 1017820K free, 152456K cached
          
```

PID	USER	PRI	NI	SIZE	RSS	SHARE	STAT	%CPU	%MEM	TIME	COMMAND
31330	nobody	14	5	187M	187M	187M	R N	44.8	50.0	5:23	eventBuilder
31363	alicemdc	13	5	188M	188M	187M	S N	23.5	50.1	3:47	writeCastor_v3
28903	pvv	9	0	992	948	748	S	1.9	0.2	40:30	top
15701	root	14	0	1052	1052	820	R	1.7	0.2	1:56	top
3	root	19	19	0	0	0	RWN	0.7	0.0	21:16	ksoftirqd_CPU0
4131	root	9	0	2176	1724	1496	S	0.3	0.4	23:19	sshd
496	root	9	0	532	532	448	S	0.3	0.1	0:00	sleep
838	ntp	9	0	1924	1924	1732	S	0.1	0.5	0:14	ntpd
18454	root	9	0	1236	1080	964	S	0.1	0.2	0:05	xload
1	root	8	0	488	440	424	S	0.0	0.1	0:20	init
2	root	8	0	0	0	0	SW	0.0	0.0	0:00	keventd
4	root	19	19	0	0	0	RWN	0.0	0.0	21:13	ksoftirqd_CPU1
5	root	9	0	0	0	0	SW	0.0	0.0	1:38	kswapd
6	root	9	0	0	0	0	SW	0.0	0.0	0:00	kreclaimd
7	root	9	0	0	0	0	SW	0.0	0.0	0:00	bdflush
8	root	9	0	0	0	0	SW	0.0	0.0	0:03	kupdated

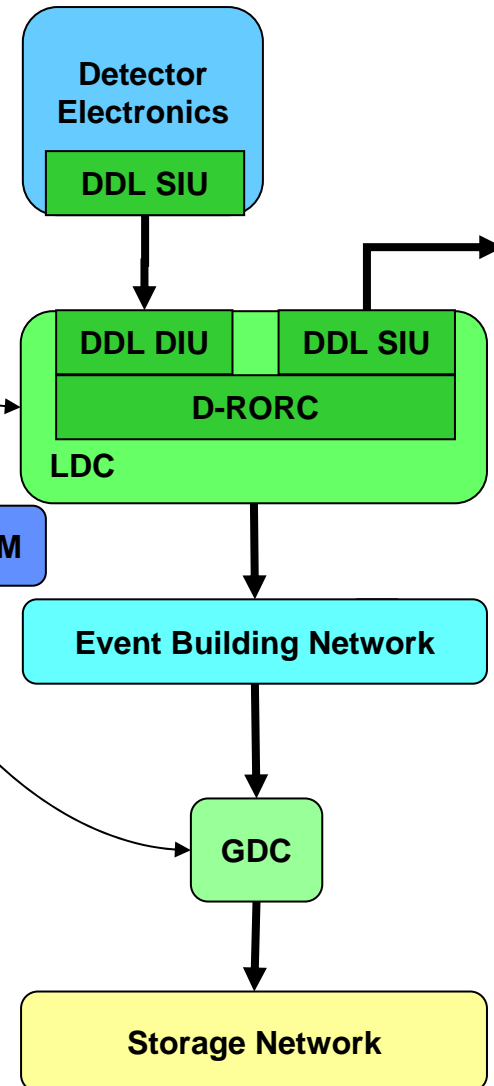


Control - DATE



Experiment
Control
System

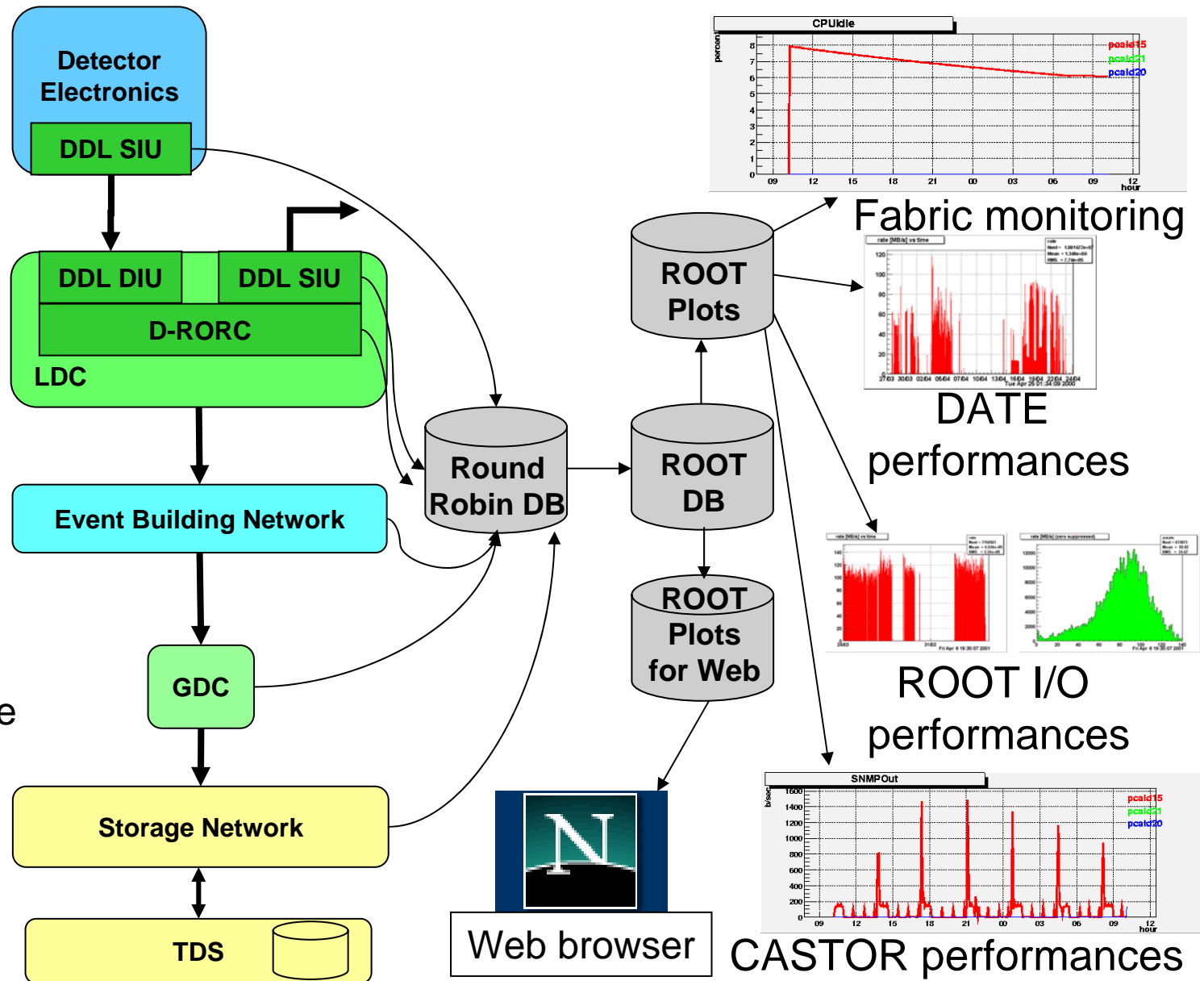
Run Control
State Machines



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



Performance Monitoring - AFFAIR

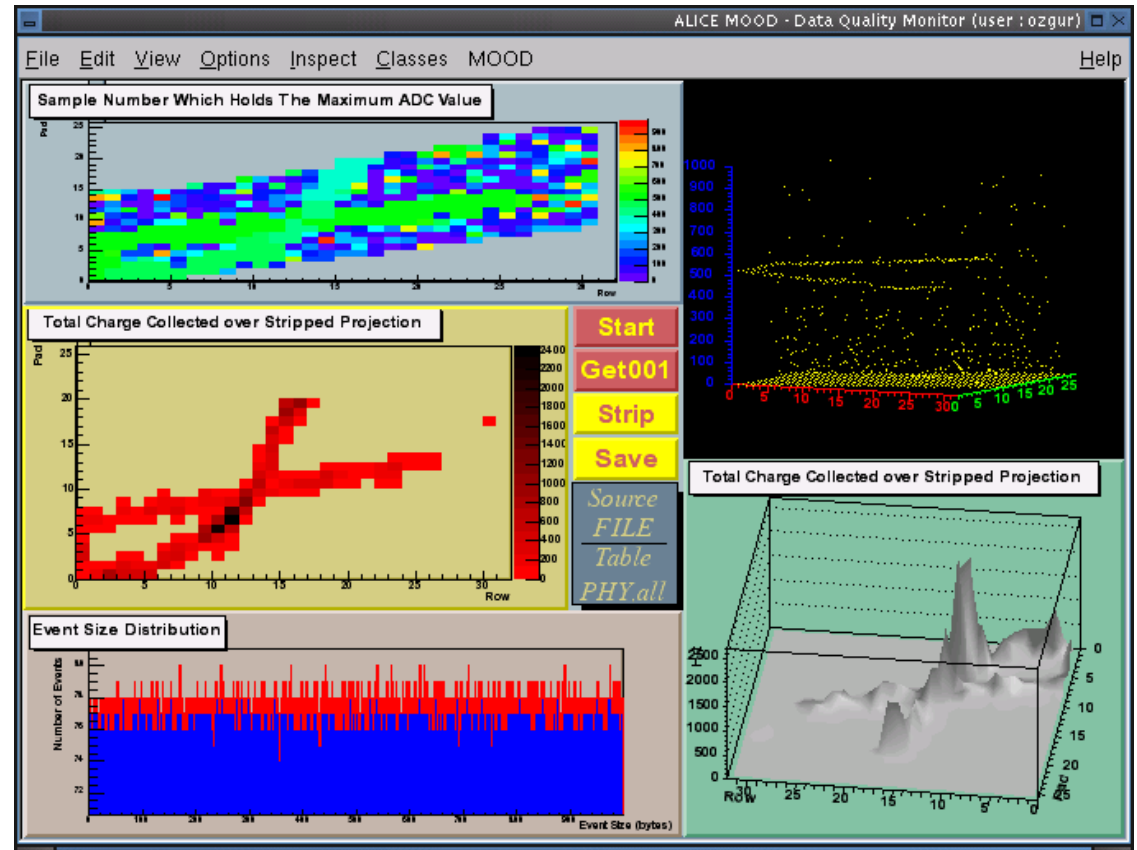


- ◆ Home-made development
- ◆ Based on free software



Data quality monitoring - MOOD

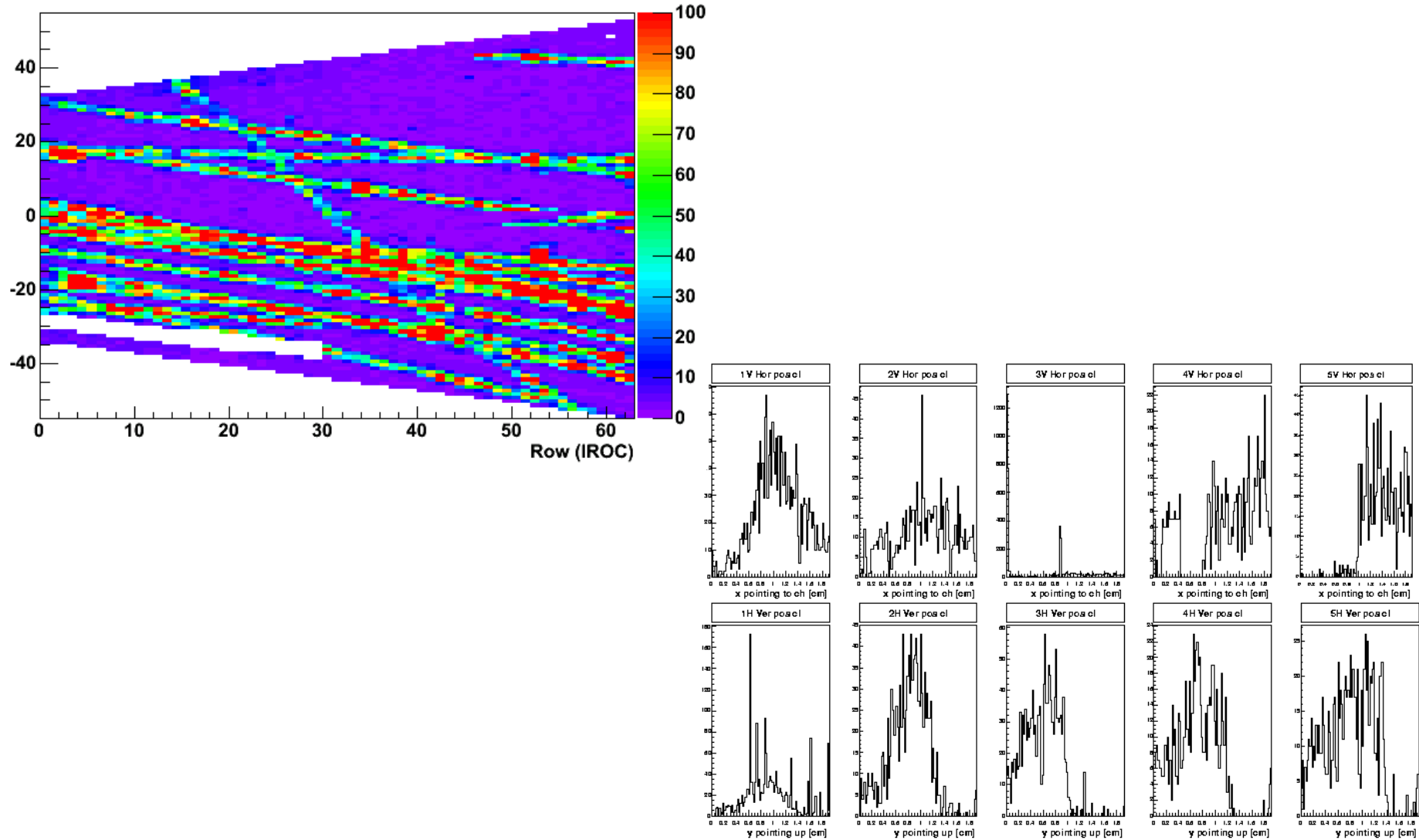
- ◆ MOOD framework
 - Interfaces to detector code
 - Software development in all institutes
- ◆ Applications:
 - Raw data integrity
 - Detector performance

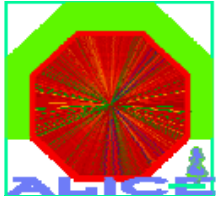




Data quality monitoring - MOOD

Event: "377" Timebin: 0-1000





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