

# CMS

## **What have we learned from building the LHC (CMS) DAQ systems.**

**S. Cittolin PH-CMD. CERN Openlab meeting. 3-4 March 2009**

**DAQ at LHC overview**

**CMS systems**

**Project timeline**

**CMS experience**



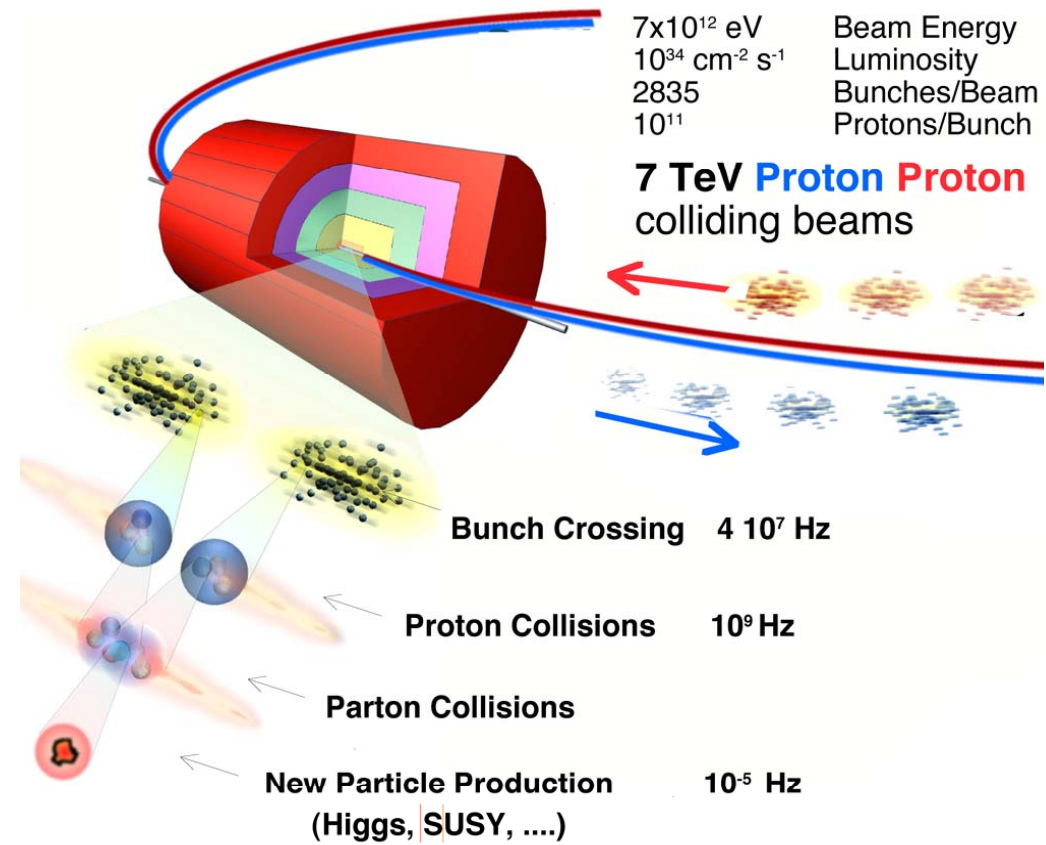
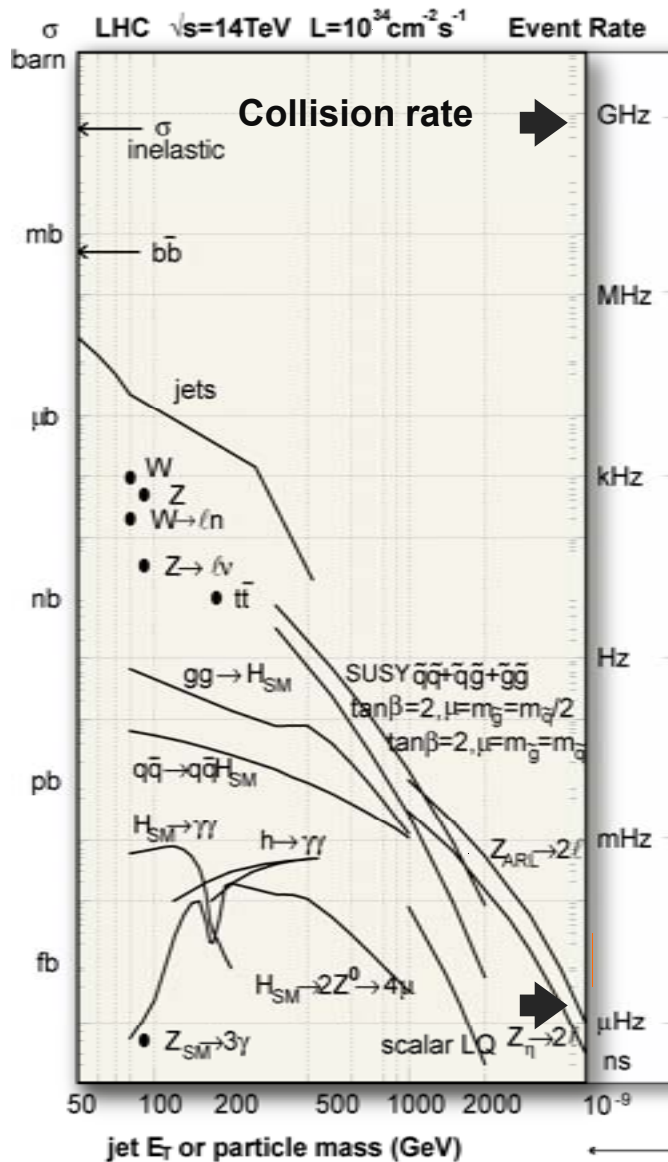
# Trigger and DAQ overview



Collisions at LHC  
The four experiments  
Readout and event selection  
Trigger levels architecture



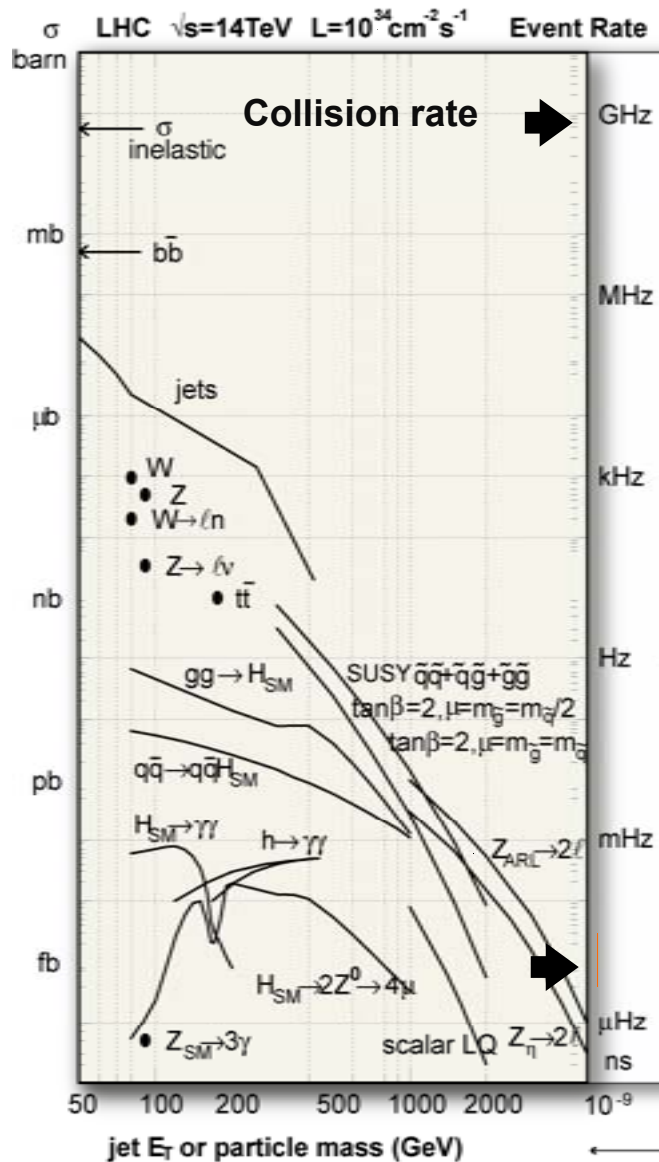
# Proton-proton collisions at LHC



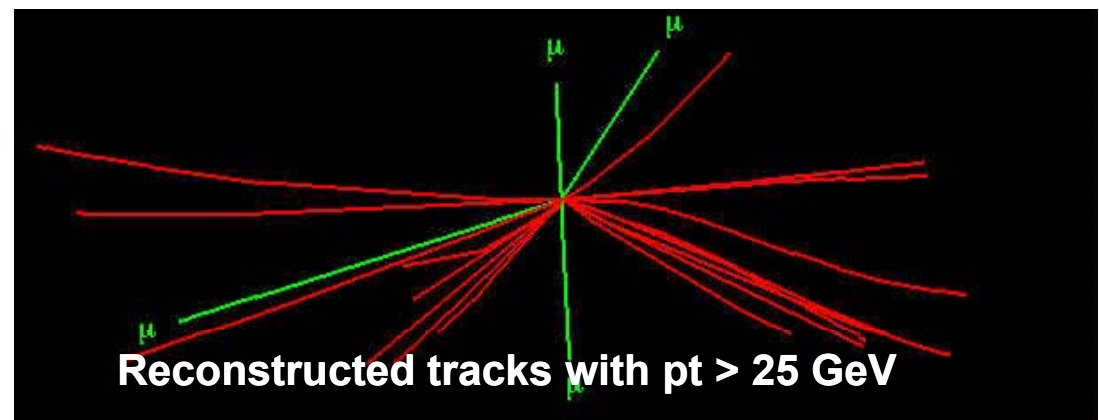
**Collision Rates:**  $\sim 10^9$  Hz  
**Event Selection:**  $\sim 1/10^{13}$

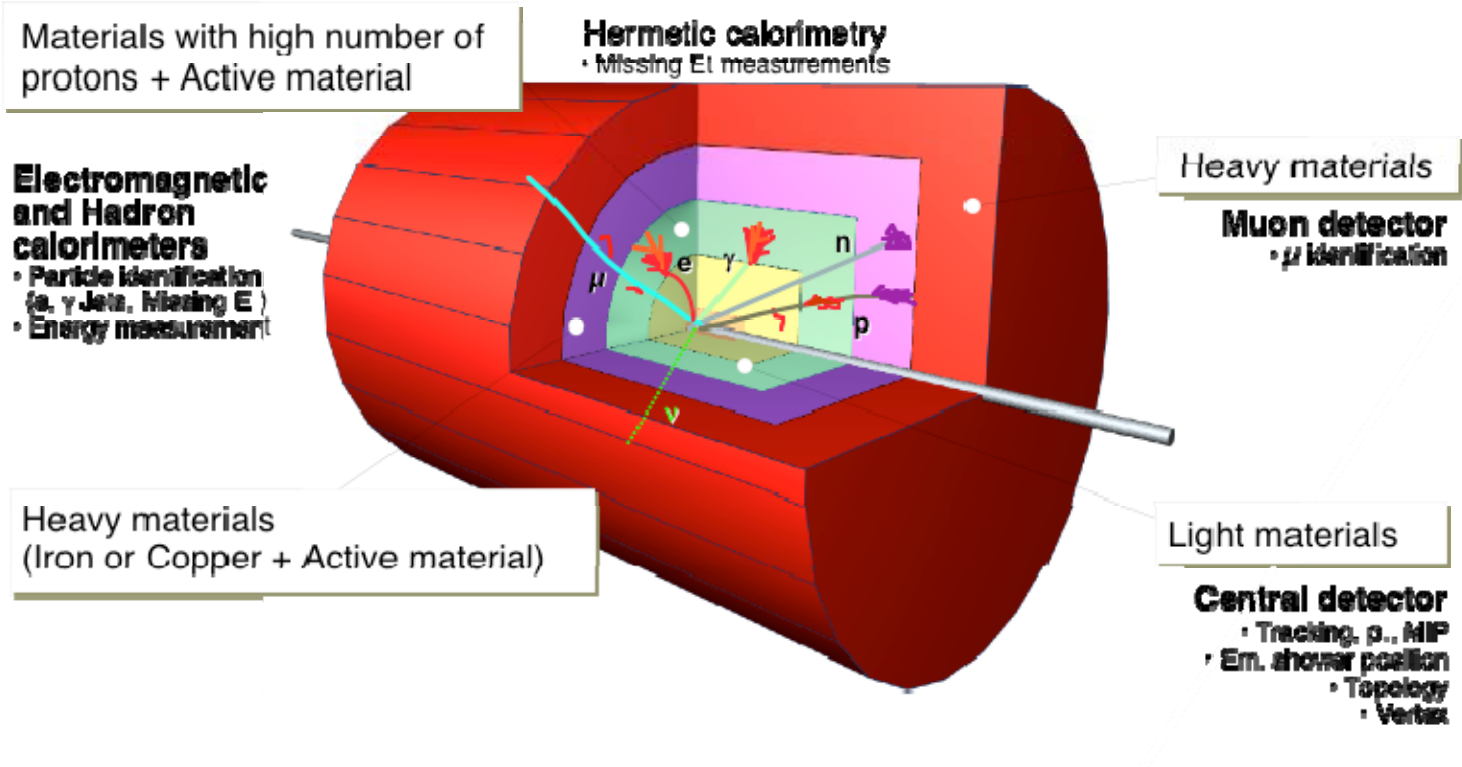


# Data detection and event selection



<b>Detector granularity</b>	$\sim 10^8$ cells
<b>Event size:</b>	$\sim 1$ Mbyte
<b>Processing Power:</b>	$\sim$ Multi-TFlop





**Each layer identifies and enables the measurement of the momentum or energy of the particles produced in a collision**





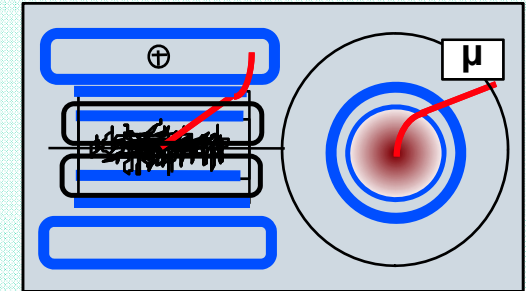
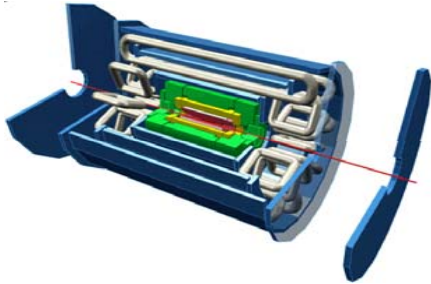
# The Experiments



## ATLAS

### Study of pp collisions

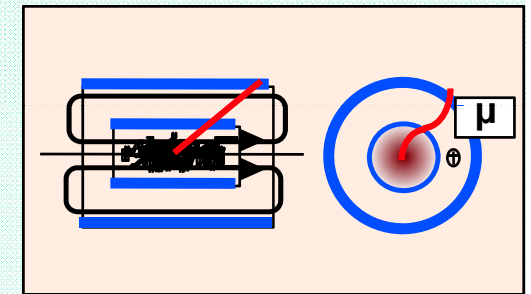
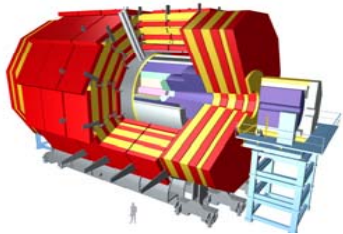
Tracker: Si ( Pixel and SCT), TRT  
 Calorimeters: LAr, Scintillating Tiles  
 Muon System: MDT, RPC, TGC, CSC,  
 Magnets: Solenoid and Toroid



## CMS

### Study of pp & heavy ion collisions

Tracker: Si ( Pixel, Strips, Discs)  
 Calorimeters: BGO, Brass Scintillators, Preshower  
 Muon System: RPC, MDT, CSC,  
 Supraconducting solenoid

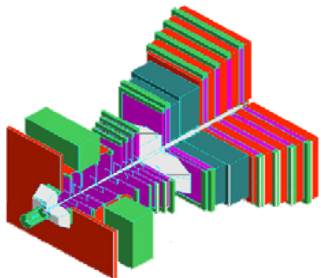


Magnetic character

## ALICE

### Study of heavy ion collisions

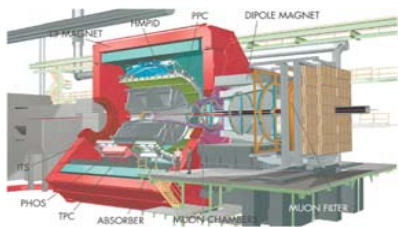
Tracker: Si (ITS), TPC, Chambers, TRD, TOF  
 Particle Id: RICH, PHOS (scintillating crystals)  
 RPC, FMD (forward mult.; Si) ZDC (0 degree cal)  
 Magnets: Solenoid, Dipol



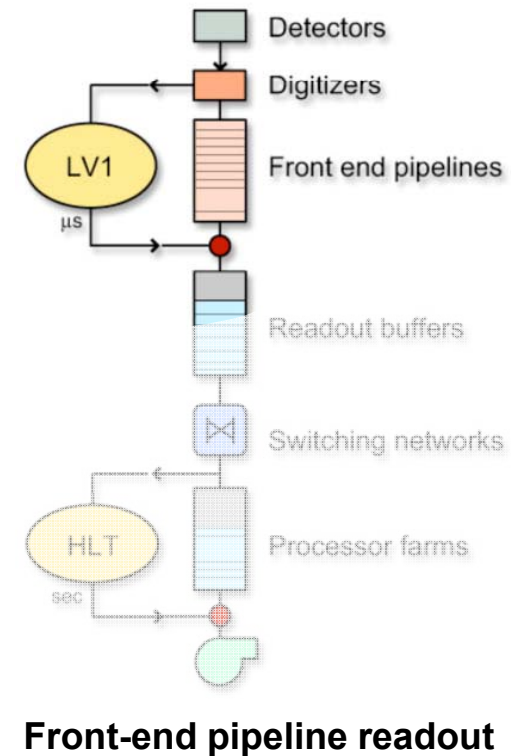
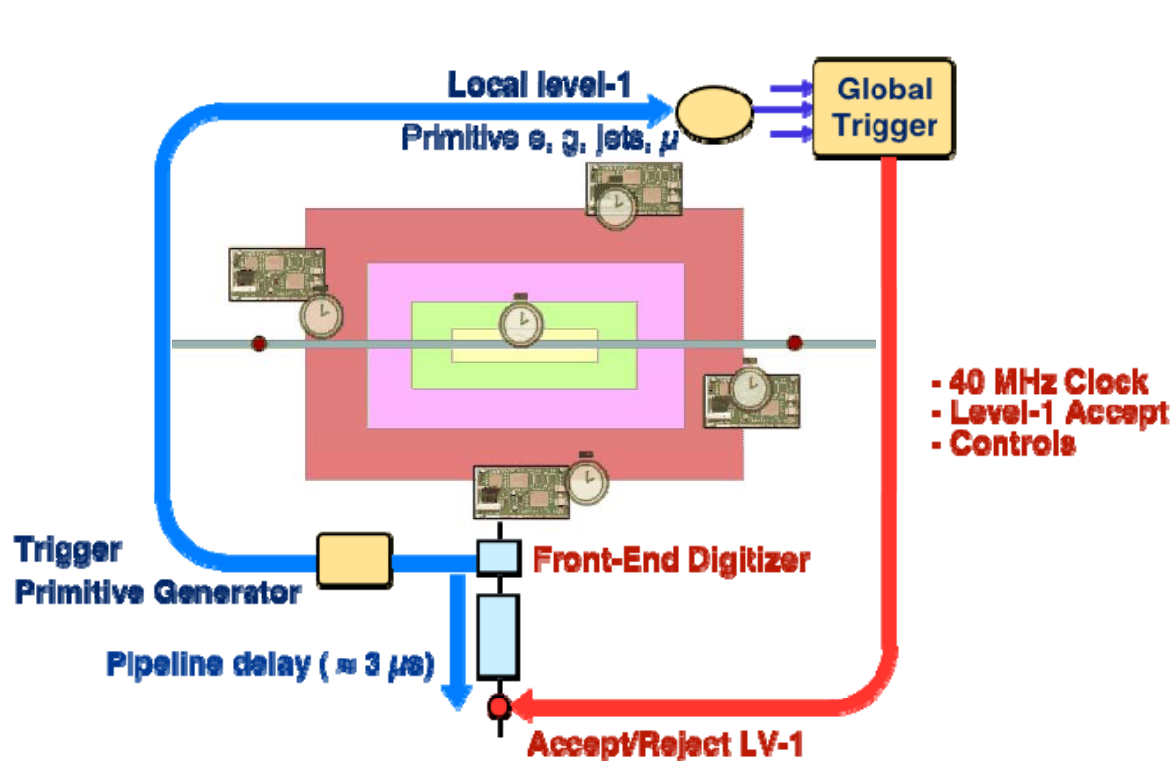
## LHCb

### Study of CP violation in B decays (pp)

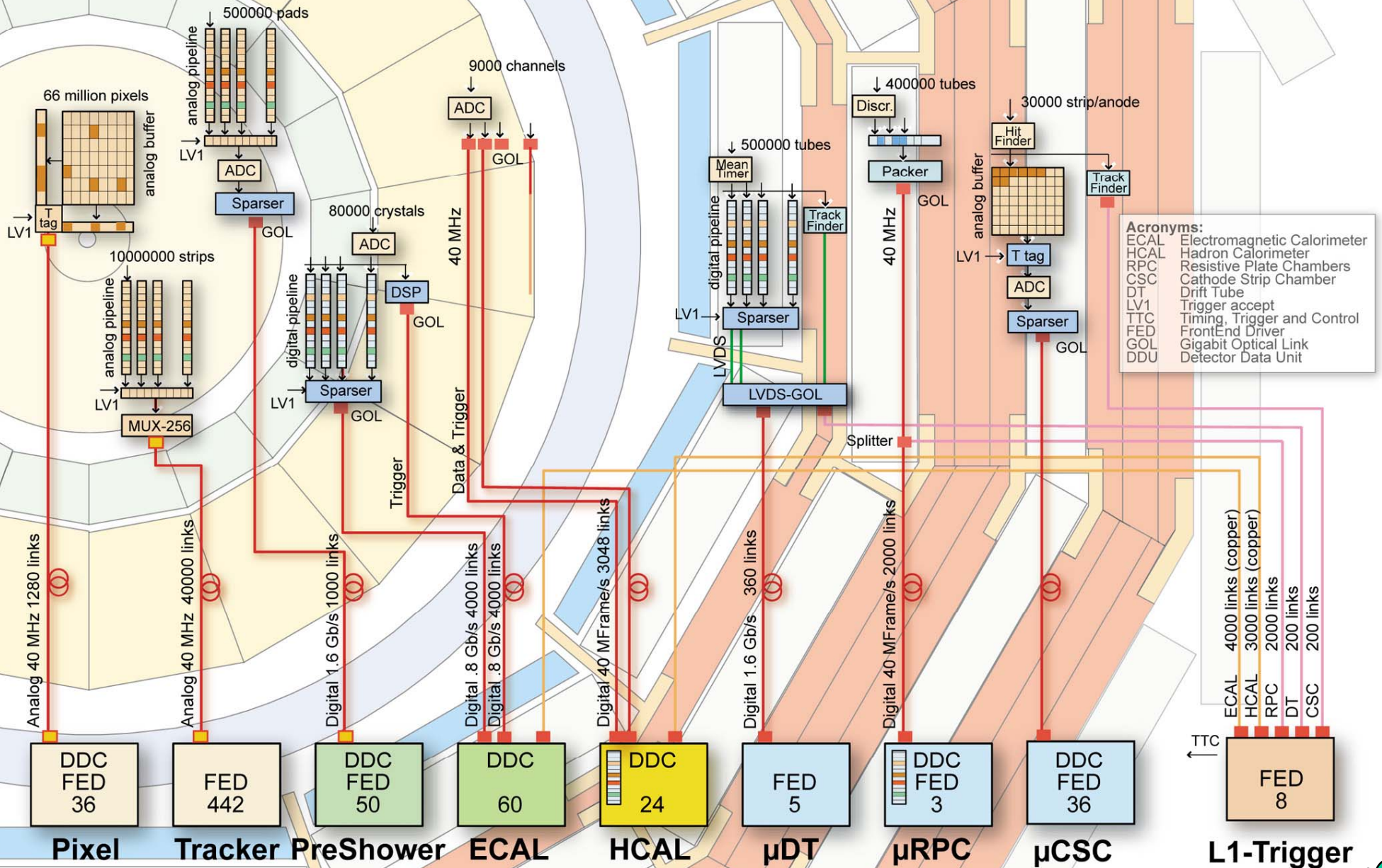
Tracker (Si, Velo), 2 RICH, 4 Tracking stations (Straw-Tubes, Si), SPD (scintill. Pads), Preshower, ECAL (lead scintillator) HCAL (steel scintillator), Muon stations (MWPCs)



High precision ( $\sim 100\text{ps}$ ) timing, trigger and control distribution  
 40 MHz digitizers and 25ns pipeline readout buffers  
 40 MHz Level-1 trigger (massive parallel pipelined processors)  
**Multi-level event selection architecture**



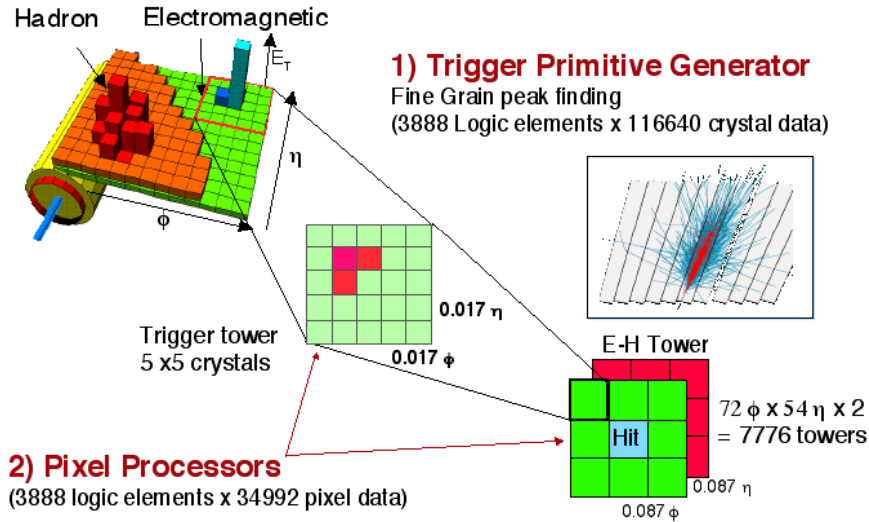
# CMS front-end readout systems





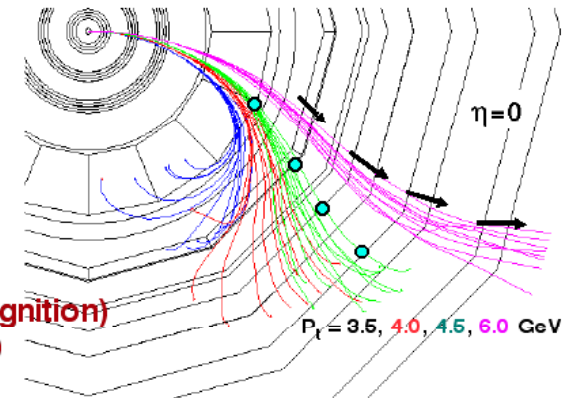


# Level-1 trigger systems. Pipelines massive parallel



Trigger based on tracks in external muon detectors that point to interaction region

- Low- $p_t$  muon tracks don't point to vertex
- Multiple scattering
- Magnetic deflection
- Two detector layers
- Coincidence in "road"



Detectors:

- RPC (pattern recognition)
- DT(track segment)

## Trigger Primitive Generator

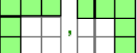



Fine grain Flag Max of (  )

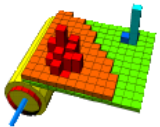
## Pixel Processor

$E_t$  cut  + Max (  ) > Threshold

Longitudinal cut (H/E)  /  < 0.05

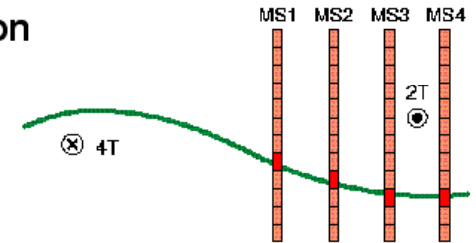
Neighbors longitudinal cut  /  < 2 GeV

One of (  ,  ,  ,  )  
↓  
**ISOLATED ELECTRON**



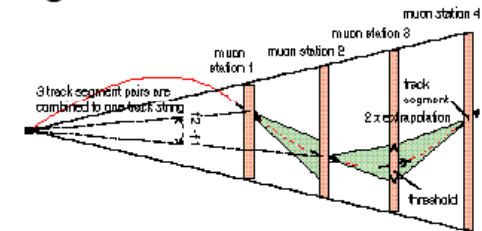
## RPC pattern recognition

- Pattern catalog
- Fast logic



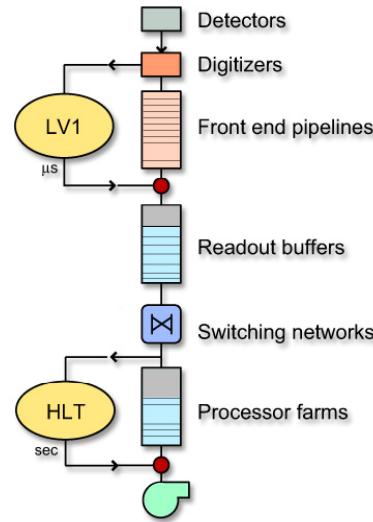
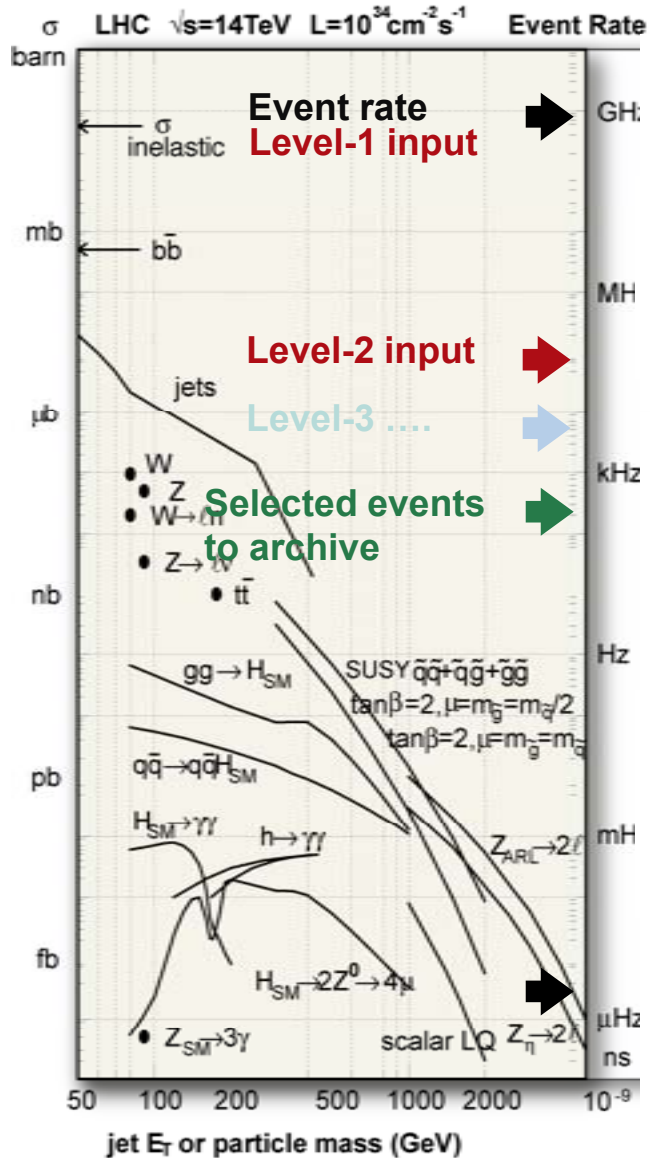
## DT and CSC track finding:

- Finds hit/segments
- Combines vectors
- Formats a track
- Assigns  $p_t$  value





# Multi-level trigger DAQ architecture



## On-line requirements

<b>Event rate</b>	<b>1 GHz</b>
<b>Event size</b>	<b>1 Mbyte</b>
<b>Level-1 Trigger input</b>	<b>40 MHz</b>
<b>Level-2 Trigger input</b>	<b>100 kHz</b>
<b>Mass storage rate</b>	<b>~100 Hz</b>
<b>Online rejection</b>	<b>99.999%</b>
<b>System dead time</b>	<b>~ %</b>

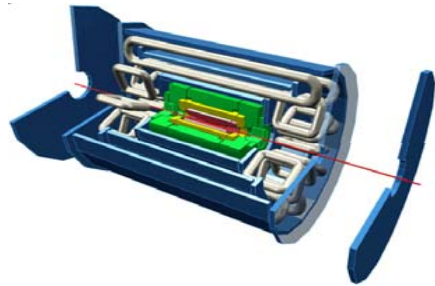
## DAQ design issues

<b>Data network bandwidth (EVB)</b>	<b>~ Tb/s</b>
<b>Computing power (HLT)</b>	<b>~ 10 Tflop</b>
<b>Computing cores</b>	<b>~ 10000</b>
<b>Local storage</b>	<b>~ 300 TB</b>

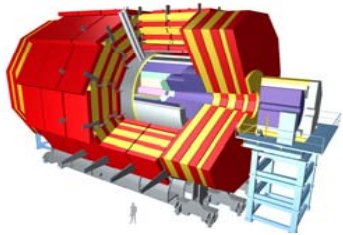
- Minimize custom design
- Exploit data communication and computing technologies
- DAQ staging by modular design (scaling)



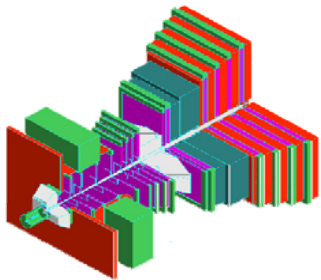
# LHC trigger and DAQ summary



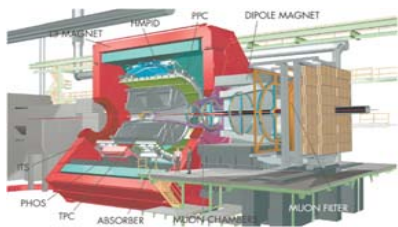
No.Levels Trigger	Level-0,1,2 Rate (Hz)	Event Size (Byte)	Readout Bandw.(GB/s)	HLT Out MB/s (Event/s)
<b>3</b>	LV-1 <b><math>10^5</math></b> LV-2 <b><math>3 \times 10^3</math></b>	<b><math>1.5 \times 10^6</math></b>	<b>4.5</b>	<b>300</b> ( $2 \times 10^2$ )



<b>2</b>	LV-1 <b><math>10^5</math></b>	<b><math>10^6</math></b>	<b>100</b>	<b>O(1000)</b> ( $10^2$ )
----------	-------------------------------	--------------------------	------------	---------------------------



<b>2</b>	LV-0 <b><math>10^6</math></b>	<b><math>3 \times 10^4</math></b>	<b>30</b>	<b>40</b> ( $2 \times 10^2$ )
----------	-------------------------------	-----------------------------------	-----------	-------------------------------



<b>4</b>	Pb-Pb <b>500</b> p-p <b><math>10^3</math></b>	<b><math>5 \times 10^7</math></b> <b><math>2 \times 10^6</math></b>	<b>25</b>	<b>1250</b> ( $10^2$ ) <b>200</b> ( $10^2$ )
----------	--	--	-----------	---

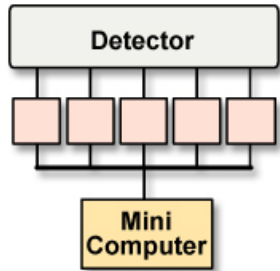


# LHC DAQ architecture



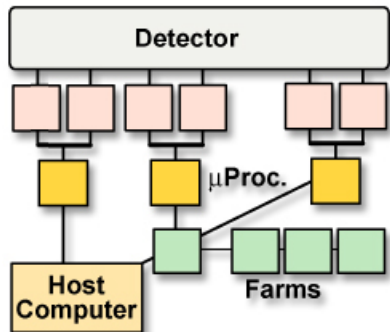
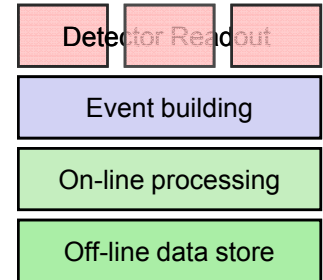
DAQ technologies  
DAQ systems at LHC





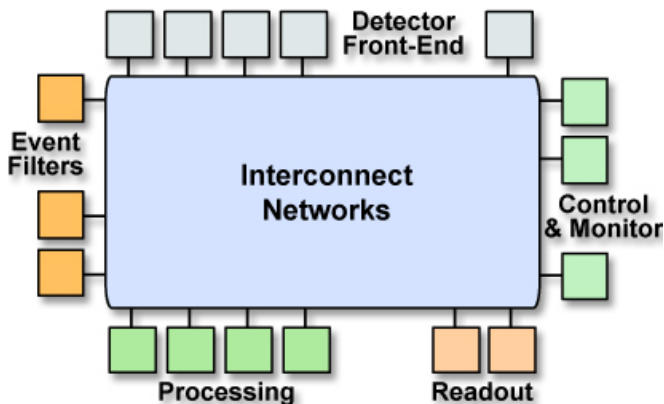
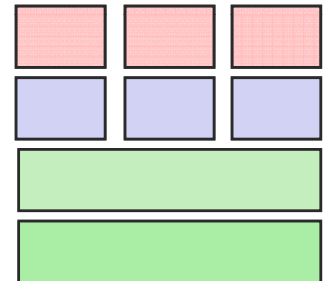
## PS:1970-80: Minicomputers

Readout custom design  
 First standard: CAMAC  
 Software: noOS, Assembler  
 • **kByte/s**



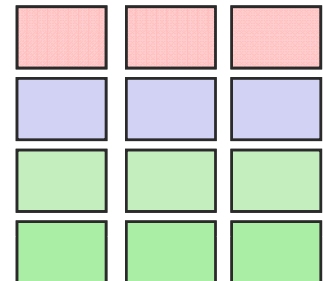
## p-p/LEP:1980-90: Microprocessors

HEP proprietary (Fastbus), Industry standards (VME)  
 Embedded CPU, servers  
 Software: RTOS, Assembler, Fortran  
 • **MByte/s**



## LHC: 200X: Networks/Clusters/Grids

PC, PCI, Clusters, point to point switches  
 Software: Linux, C,C++,Java,Web services  
 Protocols: TCP/IP, I2O, SOAP,  
 • **TByte/s**

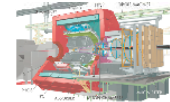
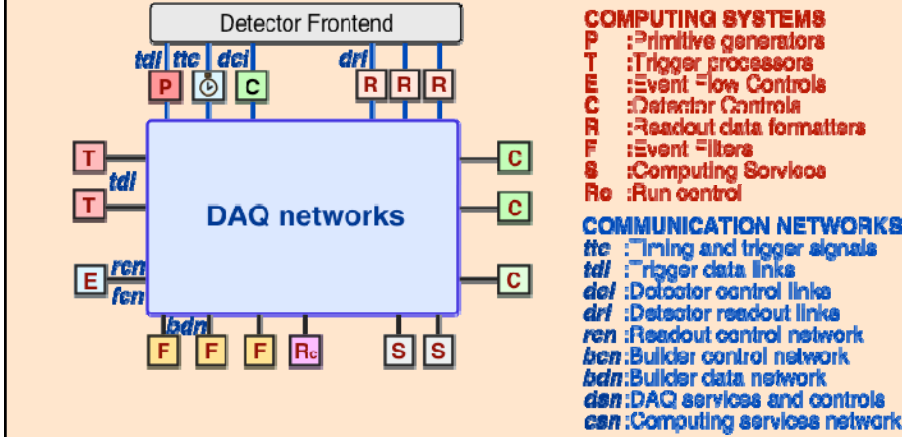




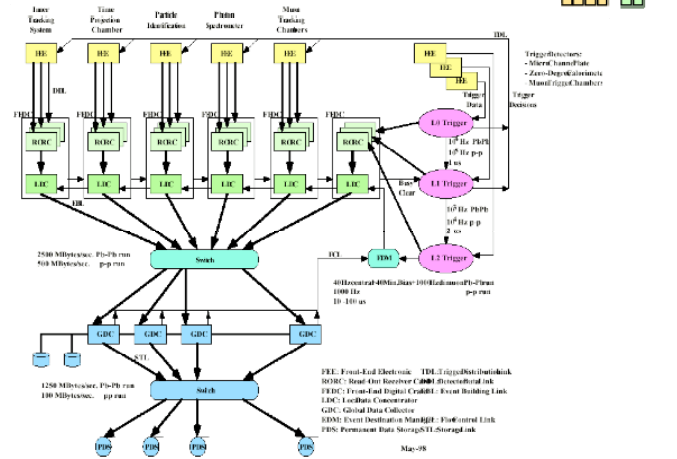
# LHC trigger and data acquisition systems



## LHC DAQ : A computing & communication network

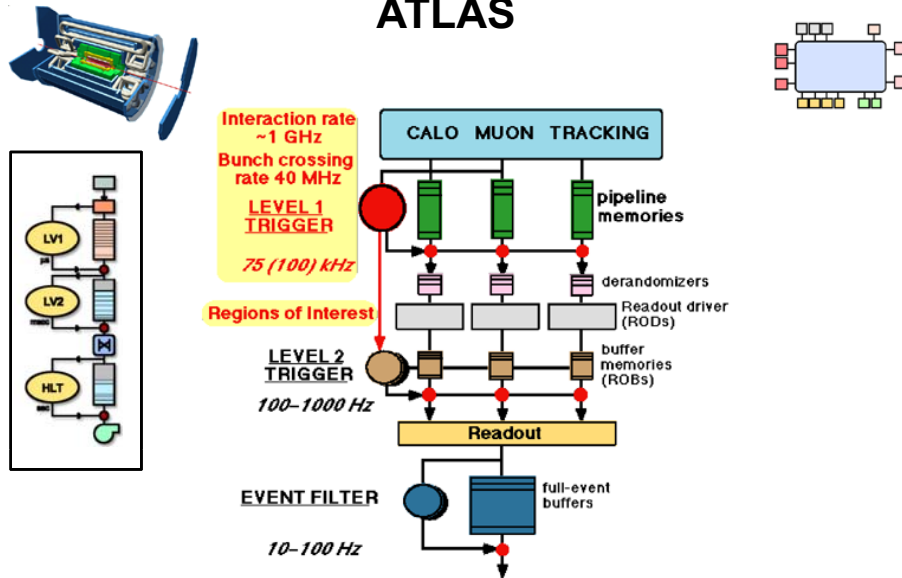


## Alice

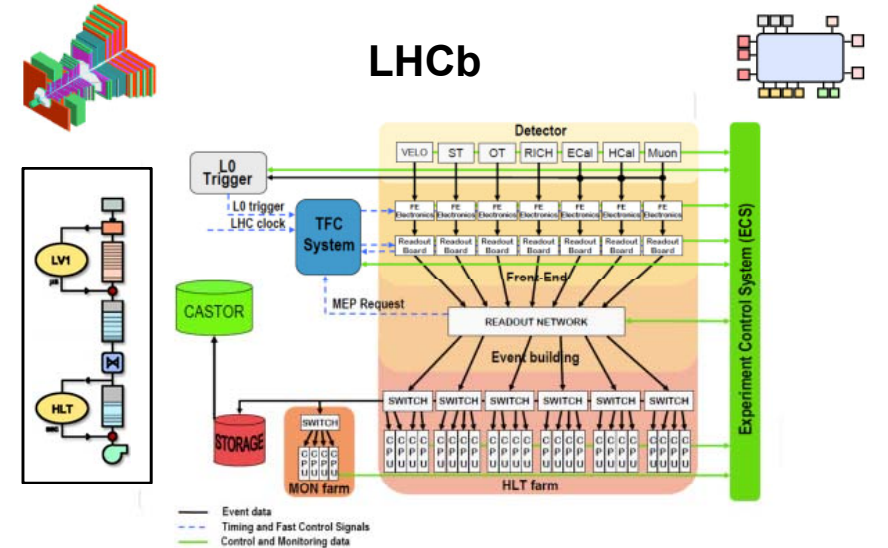


A single network cannot satisfy at once all the LHC requirements, therefore present LHC DAQ designs are implemented as multiple (specialized) networks

## ATLAS

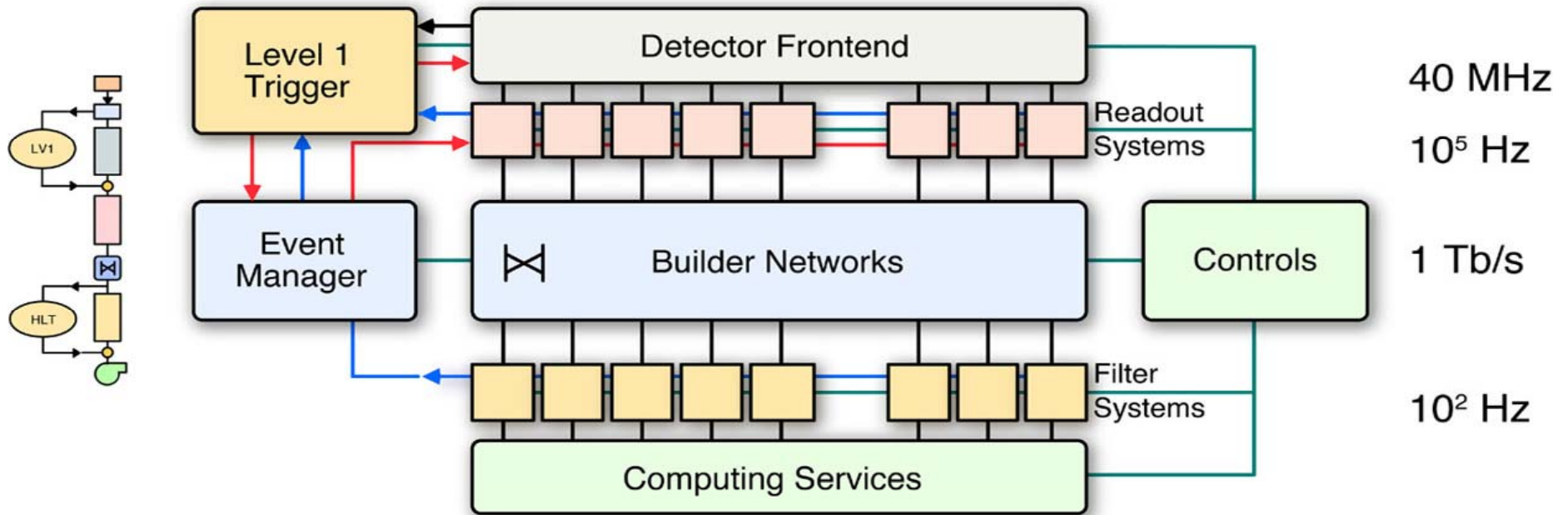


## LHCb





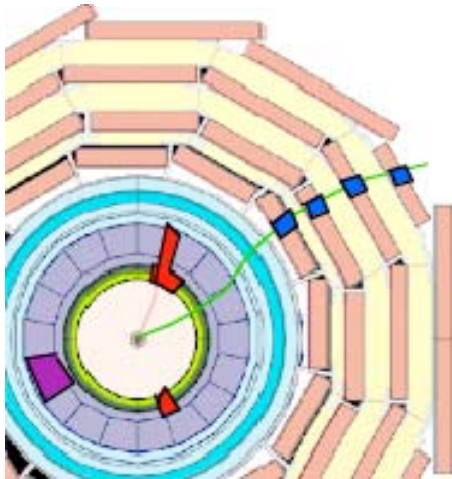
# CMS DAQ baseline structure



Collision rate	40 MHz	Readout concentrators/links	512 x 4 Gb/s
<b>Level-1 Maximum trigger rate</b>	<b>100 kHz</b>	<b>Event Builder bandwidth max.</b>	<b>2 Tb/s</b>
<b>Average event size</b>	<b>≈ 1 Mbyte</b>	<b>Event filter computing power</b>	<b>≈ 10 TeraFlop</b>
Flow control&monitor	≈ 10 <sup>6</sup> Mssg/s	Data production	≈ Tbyte/day
		Processing nodes	≈ Thousands



# Two trigger levels

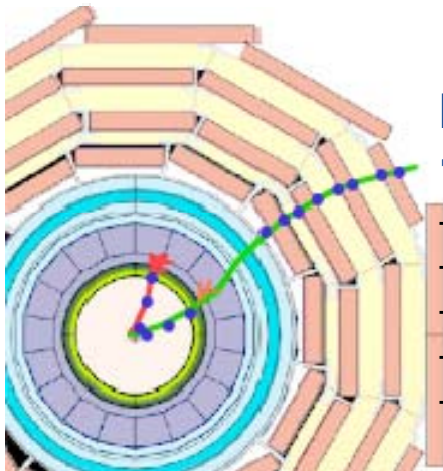
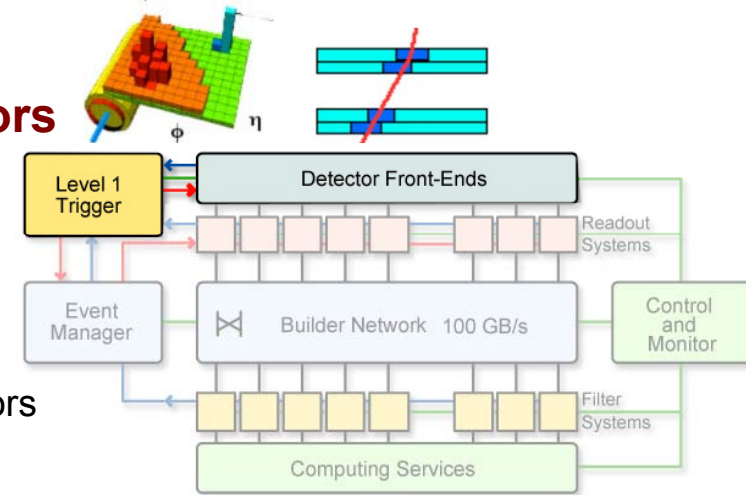


## Level-1: Massive parallel processors 40 MHz synchronous

- Particle identification:
- high pT electron, muon, jets, missing ET
- Local pattern recognition and energy
- evaluation on prompt macro-granular information from calorimeter and muon detectors

99.99 % rejected:

0.01 Accepted

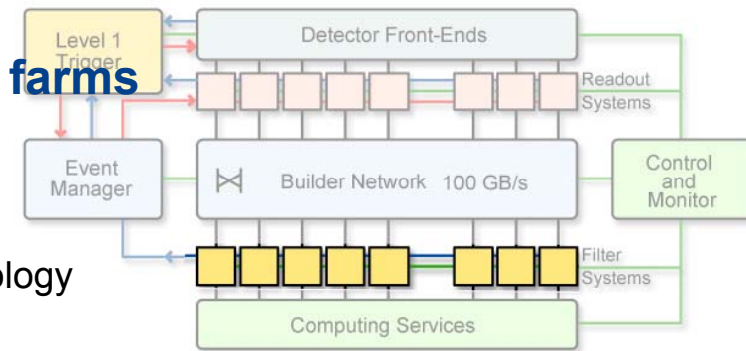


## Level-2: Full event readout into PC farms 100 kHz asynchronous farms

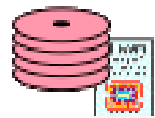
- Clean particle signature
- Finer granularity precise measurement
- Kinematics. effective mass cuts and event topology
- Track reconstruction and detector matching
- Event reconstruction and analysis

99.9 % rejected:

0.1 Accepted



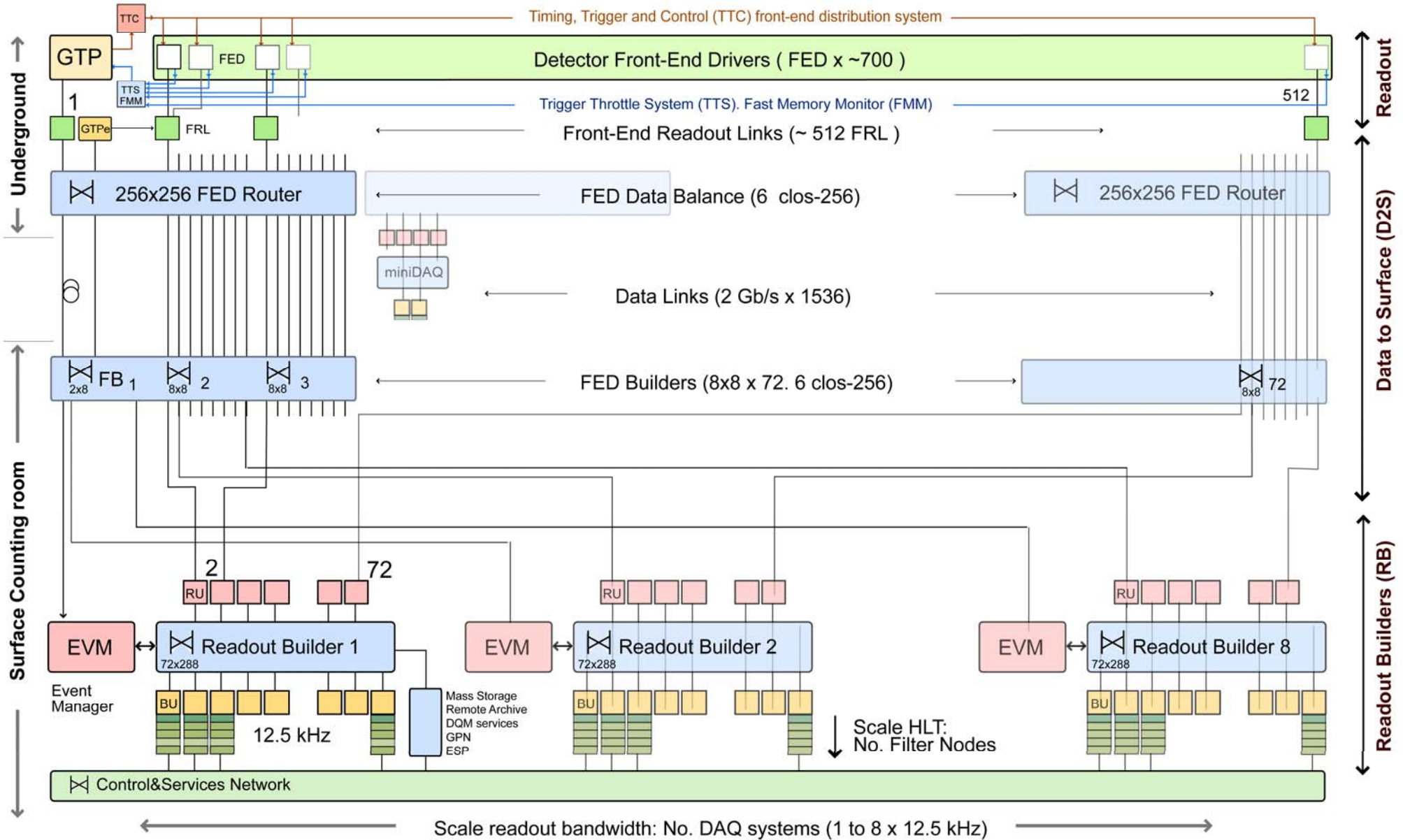
100-1000 Hz. Mass storage  
Reconstruction and analysis.



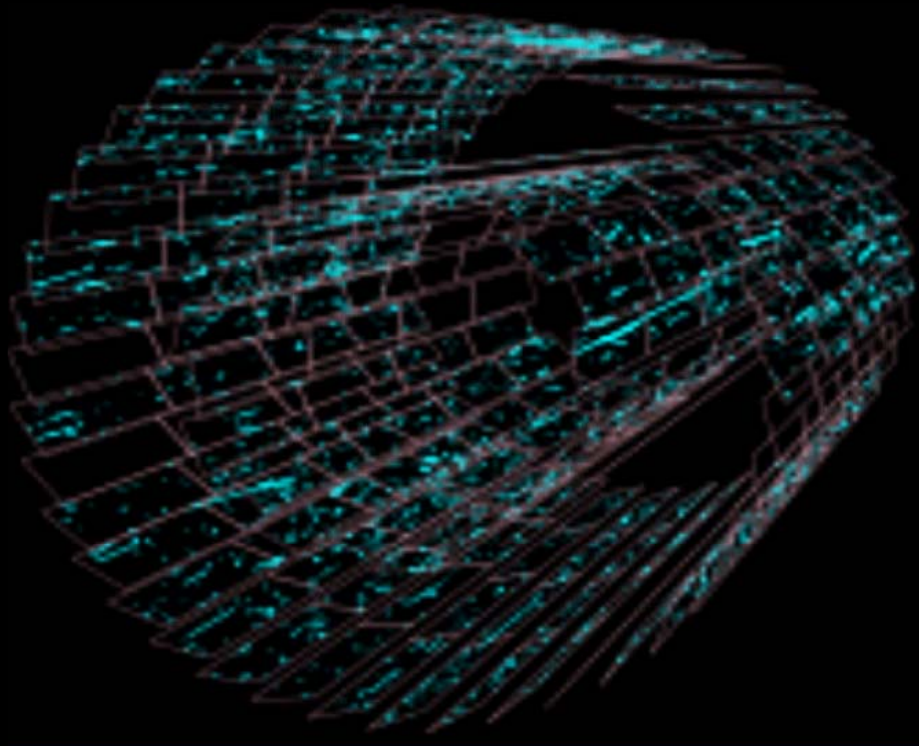
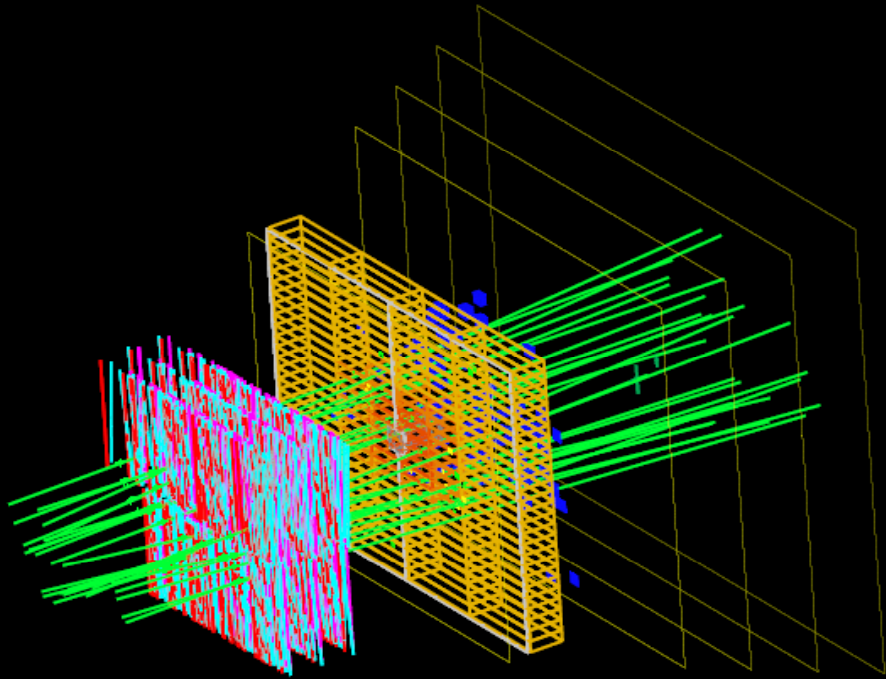
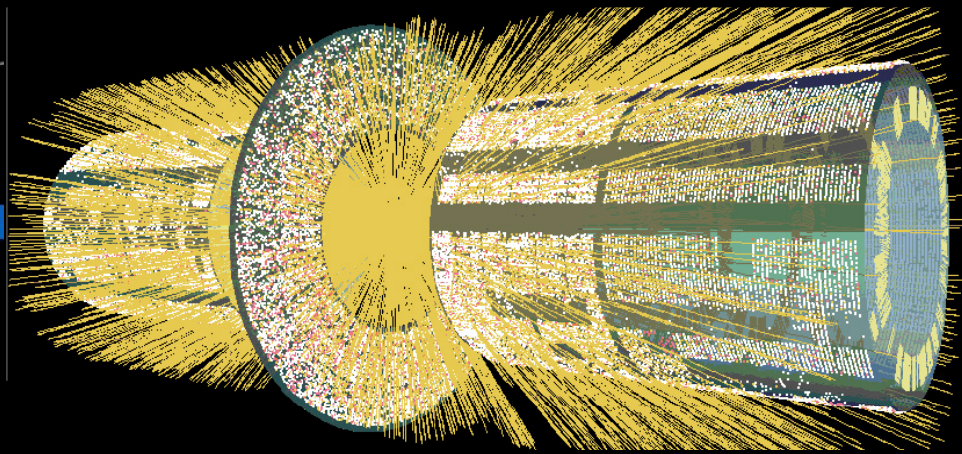
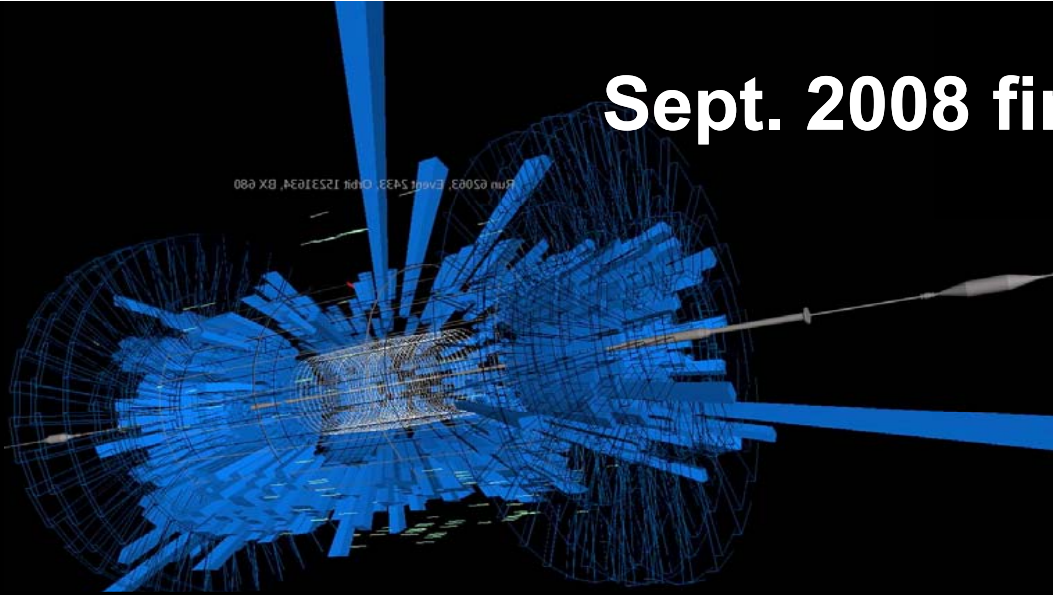




# 8-fold DAQ structure



# Sept. 2008 first events





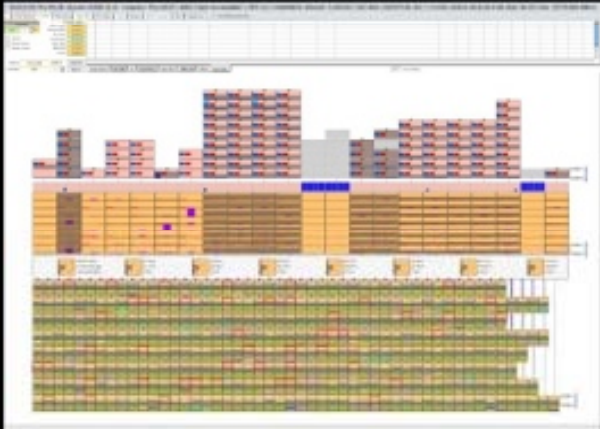


# March 09 Technical Global Run



26/03/09 Thu 09:29. Session 42682 [1:2] <toppro> Thu 09:27 | DAQ "state not available"

## FU states



## Data to Surface

Sub-System	State	FRL	FED	IN
ECAL	X	54	54	44
HCAL	X	33	33	32
TRG	X	5	5	5
CSC	X	8	8	8
RPC	X	3	3	3
DT	X	5	5	5
	X	0	0	0
	X	0	0	0
	X	0	0	0
	X	0	0	0
	X	0	0	0
	X	0	0	0
	X	0	0	0
	X	0	0	0
	X	0	0	0
	X	0	0	0

## DAQ items

	FED	FRL	EVM	RU	BU	FU	SM
#Tot.	108	108	8	168	379	758	8
# InFla.	97	100	8	168	379	758	8
# Enabl.	97	97	8	168	379	758	8
# Dead	0	0	0	0	51	0	0
dt(s)	1	1	0	1	3028	2	1
Late.(s)	34	34	35	34	35	35	35
Slice 1	1	21	45	90	1		
Slice 2	1	21	49	98	1		
Slice 3	1	21	45	90	1		
Slice 4	1	21	49	98	1		
Slice 5	1	21	47	94	1		
Slice 6	1	21	46	92	1		
Slice 7	1	21	48	96	1		
Slice 8	1	21	50	100	1		

## Data Flow

Missing flashes

#Lum.Seg.

Random ON  
Physics ON  
CalibCyc ON

#Lv1(GT)  1 FEDCRC

Rate (kHz)

Pending Lv1  95 BX-Lv1 #

#Frag. in RU Max  FBI occ. % Max  Min

Bnw (MB/s)  Ev. size (kB)  FBO occ. % Max  Min

Events in BU  Pending Req.  Rec.-Disc.

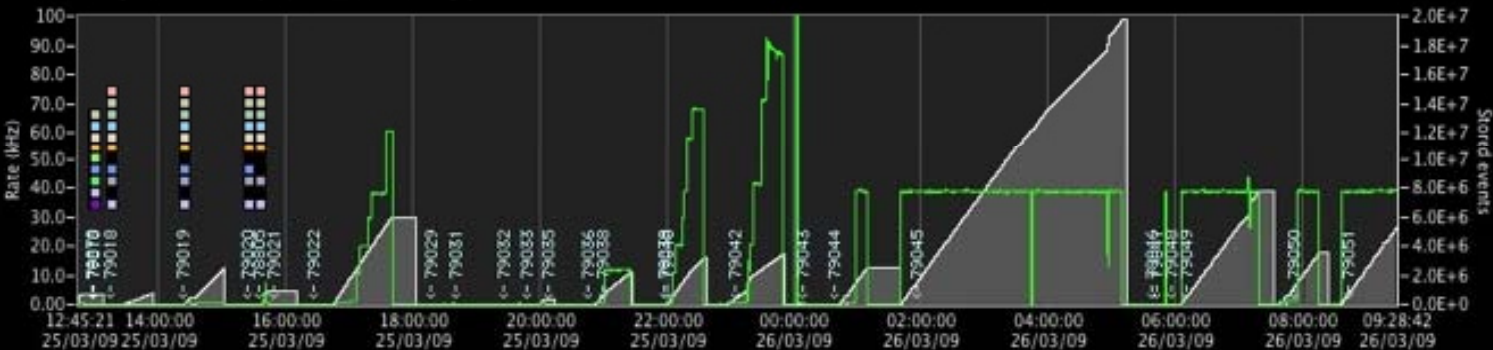
#Active FUs  #P> 21.8 18691 P.M.-m 797 A.M.-m

Accepted

Stored events (all streams)

Fill 41.7 40.7 39.2 8.5

[Rate(kHz) | Stored events] / Time



FMURL <http://cmsrc-top:10000/urn:rcms-fm:fullpath=/toppro/PublicGlobal/levelZeroFM,group=levelZeroFM,owner=toppro>



# CMS experience



DAQ project timeline  
Industry trends/DAQ  
Hardware/Software components  
DAQ at Super LHC





# LHC/CMS-DAQ project timeline



1990 Design of experiment

**1992 CMS Letter of Intent**

**1994 Technical Design Report**

1996

◆ 1998

2000 Trigger Technical Design Report

◆ 2002 DAQ Technical Design Report

2004

◆ 2006 Magnet test Global Run

◆ 2008 Circulating beam Global Run

◆ 2009 Colliding beams

## Research and Development (DRDC)

Trigger, Timing and Control distribution (TTC)  
Readout prototypes (FPGA, PC, IOP-200 MB/s )  
Networks (ATM, Fiber Channel, xyz..)

CMS 2-level triggers design

## Event Builder Demonstrators

FPGA/PC data concentrators  
8x8 Fiber channel EVB  
32x32 Myrinet EVB  
64x64 Ethernet EVB, PC driven

## Final Design Pre-series

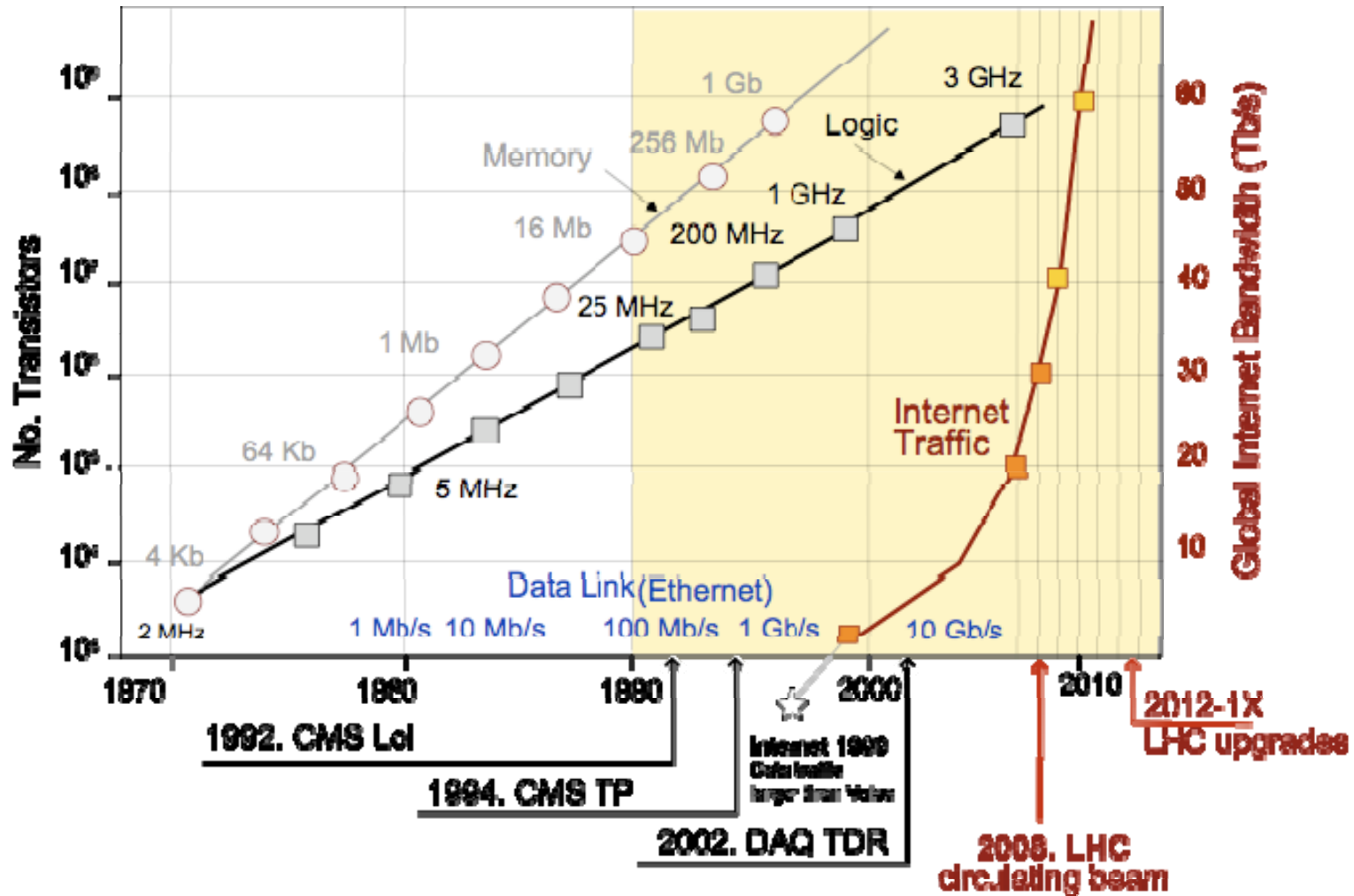
64x64 Myrinet/Ethernet

## Construction and commissioning

1024 2 Gb/s D2S Myrinet links and routers  
8x80x(80x7) GbEthernet EVB/HLT  
10000 on-line cores



# Computing and communication trends



Lesson 2. Moore law confirmed

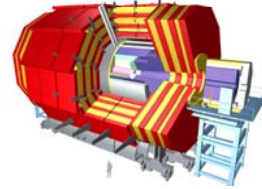
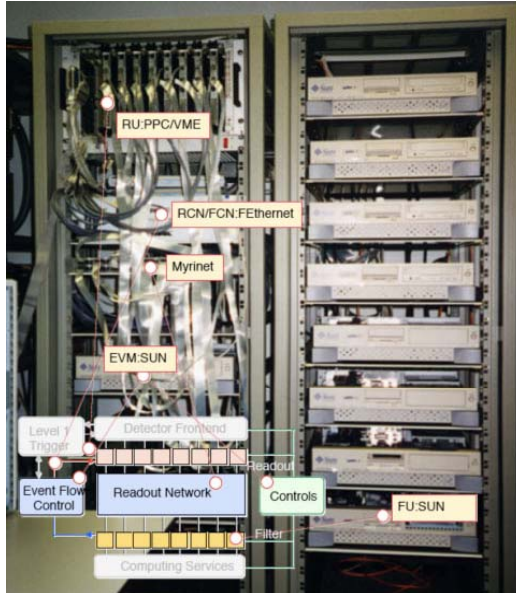




# Two trajectories



### 1997 CMS 4x4 FC-EVB



### 1997 GOOGLE first cluster



### 2008 Cessy CMS HLT center 10<sup>4</sup> cores, 2 Tb/s maximum bandwidth

### 2008 One of Google data centers 10<sup>6</sup> cores





## Global Internet traffic (Cisco forecasts)



<b>US Consumer (PB per month)</b>	<b>2007</b>	<b>2008</b>	<b>2009</b>	<b>2010</b>	<b>2014</b>
Web, email, transfer	710	999	1336	1785	
P2P	1747	2361	3075	3981	
Gaming	131	187	252	324	
Video Communications	25	37	49	70	
VoIP	39	56	72	87	
Internet Video to PC	647	1346	2196	3215	
Internet Video to TV	99	330	756	1422	
Business	1469	2031	2811	3818	
Mobile	26	65	153	345	
<b>Total global traffic (Pb/M)</b>	<b>4884</b>	<b>7394</b>	<b>10666</b>	<b>14984</b>	
<b>Global Internet traffic (Tb/s)</b>	<b>20</b>	<b>30</b>	<b>40</b>	<b>60</b>	
<b>Total US traffic (Tb/s)</b>	<b>3</b>	<b>4</b>	<b>6</b>	<b>8</b>	
<b>Google US traffic (Tb/s)</b>	<b>0.3</b>	<b>0.7</b>	<b>1.5</b>	<b>3</b>	
<b>CMS Maximum bandwidth (Tb/s)</b>	<b>1</b>	<b>2</b>	<b>2</b>	<b>2</b>	<b>&gt;10</b>





## Data communication

### Custom

- 6000 1 to 1 Optical trigger primitive readout 1 Gb/s (Rad hard)
- 60000 1 to 1 Optical analog front-end readout 40 Mb/s (Rad hard)
- 1000 1 to N Optical fast signal distribution tree 40 MHz
- 1000 N to 1 Copper Leaves tree signals collection system
- 800 1 to 1 Copper detector readout LVDS 4 Gb/s links

### Proprietary

- 1024 1 to 1 Optical full duplex data links (Myrinet 2.5 Gb/s)
- 2056 N to N Optical routers. FED builders (Myrinet)
- 1024 PCI dual 2.5 Gb/s optical link (Myrinet 2000)

### Commercial standard

- 4120 N to N Copper Ethernet switches (Force10)
- 800 PCI card quad GbE copper link (Silicom)

## Data processing

### Custom

- All sub-detector digitizers, data concentrator, on detector controls
- Trigger processors logic cards

### Proprietary

- 100 Water cooled racks HLT computing rooms (CIAT)

### Commercial standard

- 300 PC Intel Dual-CPU. Front-end VME/PCI controllers (Dell)
- 700 PC Intel Dual-CPU Dual-Core. DAQ nodes RU-BU (Dell 2950)
- 900 PC Intel Dual-CPU Quad-Core. High Level Trigger (Dell 1950)
- 100 PC servers (Dell). 300 Tbyte mass storage
- VME and PCI crates, PCI express, Field busses



# DAQ costs



## Project construction costs

### R&D, Prototypes and pre-series 2 8%

- 120 SuperMicro PCs
- 256 port Myrinet switch and interfaces
- 512 ports Ethernet switch and interfaces

### Detector readout links 1 4%

- 800 Front-end-PCI interfaces
- 100 Fast Monitor Modules (FMM)

### D2S 2 Tb/s 5 20%

- 300 VME controller PCs and PCI crates
- 2048 port Myrinet routers
- 1024 dual 2.5 Gb/s Myrinet interfaces
- USC-SCX optical cables

### EVBU 100 kHz 4 16%

- 640 RUBU Dell 2950 Dual CPU Dual core
- 4120 port GbEthernet switches and interfaces

### HLT 50 kHz 2.5 10%

- 740 Dell 1950 Dual CPU Quad core

### Infrastructures 2 8%

- 120 Dell 1950 servers
- 300 TB local Mass storage. Remote archive link
- Racks W.cooled, Service networks, Controls

### HLT 100kHz 2.5 10%

### Total 19 76% (of the requested budget in 2003)

## Maintenance and Operation costs

### Custom and proprietary M&O

**25% spares** are acquired for long term maintenance of custom designed boards and non standard equipment (e.g. Myrinet)

### Commercial standards M&O

HTL PCs are replaced every **3 years**  
 Data flow PCs, network and storage disks are replaced every **4 years**. All other servers are replaced every **5 years**

### System administration

Dedicated manpower to administrate and maintain the PC farms and networks

Lesson 5 Thanks Moore. Should we buy systems and services?



## Operating Systems

Linux SLC4/SLC5, Window

## Languages

C++, Java, Perl, Unix Shells, XML, HTML, Java Script

## Databases

Oracle, MySQL, File System

## GUI

Web Browsers, HTML, DHTML, LabView, Qt, Applets, JFree Chart (Java), ROOT

## Protocols

TCP, HTTP, CGI, I2O (binary for data flow), XDR (binary for monitoring),  
SOAP(XML + binary attachments), SMI, DMI, PVSSII, log4j

## Software Maintenance and Documentation

Quattor, elog, Media wiki, Twiki, CVS, Source Forge, Savannah

## DAQ Core framework and components

System and communication services, Hardware access facilities and device drivers  
Interface to external systems (e.g. DCS, computing services), DAQ monitoring

## DAQ applications

FED builder, Event Builder, HLT framework support, Storage manager, DB support,

## Run Control and Monitor System

Configuration, control and monitoring (> 10000 processes). Interface to operators (GUI, script) and to DCS  
Remote access, security

## Detector Controls

Detector DCS coordination. Common tools development&support,Framework and central DCS system,  
DAQ infrastr

Lesson 6. Configuration, control and operation of complexity is an issue



# Control room



**Cessy: Master&Command control room**



**Fermilab: Remote Operations Center**



**Meyrin: CMS DQM Center**



**CR: Any Internet access.....**



**Security is an issue**





## Luminosity increase (2012-16) will require

- New front-end electronics and readout links
- Higher **level-1 selection power** (to maintain 100 kHz max. output)
- **Event builder (>10 Tb/s)** with an order of magnitude higher

## The upgrade programme will include:

- **New Front-End digitizers**, new rad hard data links and a new timing and trigger distribution system (distribute event type, HLT destination etc.).
- All very front-end systems and selection logic will still be based on custom design. However **new telecommunication technologies** (e.g. TCA etc.) can be employed to interconnect data concentrators, level-1 logic modules and to interface the detector readout with commercial standards.
- **Data to Surface links (10 Tb/s)** has to be replaced (2005 proprietary technology life time and 10 time the speed). Likely with standards e.g. 1000x10Gb/s data links (not yet a Moore law for data links)
- Event data fragment will be tagged with trigger type and HLT destination. Event builder and High Level Trigger will be **embedded in an single data network** (real-time internet clusters/grid like?) which includes local/central data archives and off-line

Lesson 7. the best DAQ R&D is the completion and operation of the current system



## 1. 12 yeas of R&D (too much?)

the project has lasted more or less a man generation from design to implementation...

## 2. Moore law confirmed

## 3. Will we buy computing power and network bandwidth?

New kind of commodities. CPU power, memory, mass storage and bandwidth are becoming commercial products..

## 4. Custom/Standards attention

Pay attention to maintenance and replacement issues. Survey new standards in the field of telecommunication, server packages, data centers, cooling etc.

## 5. Less cost, thanks Moore. Buy more from services in the future?

The process of procurement, installation and commissioning of the last HLT farm took about 10 months (because administrative rules, tender, reliability of components etc.). System management and maintenance for Cluster, Network and DataBase can be centralized?

## 6. Configuration and control of complexity is an issue

Data taking efficiency depends on the real-time system performances but also on the prompt handling of on-line resources. E.g. all experiments need long time (minutes) to cold-start and configure their DAQ system (> 10000 processes), Fault tolerant systems, fast recovery etc....

Distributed control rooms. Master, command, monitor and security

## 7. DAQ best R&D is the completion and operation of the current system

The upgrade will be mainly upgrade of network and servers following the M&O expenditure profile. Real new improvement will come from Point 5 issues and the operation experience of the current system

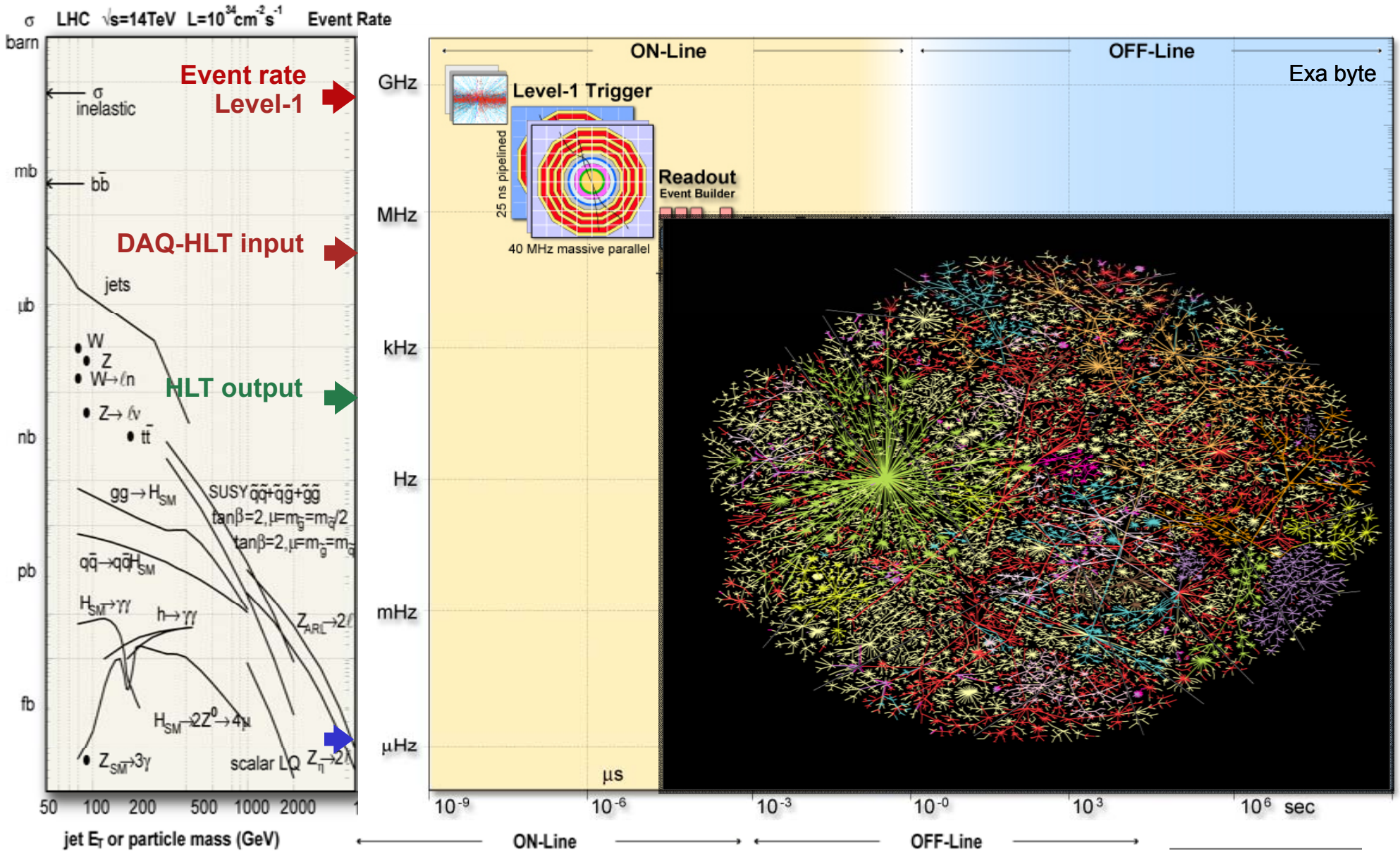


**extra**





# DAQ data flow and computing model







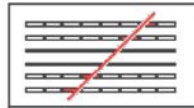
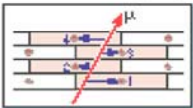
# CMS design parameters and DAQ requirements



## Detectors

### MUON BARREL

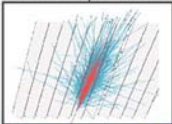
Drift Tube Chambers (DT)



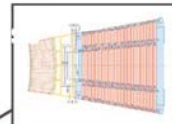
Resistive Plate Chambers (RPC)

### CALORIMETERS

ECAL Scintillating PbWO<sub>4</sub> Crystals



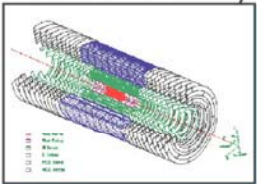
HCAL Scintillator brass sandwich



IRON YOKE

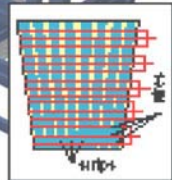
SUPERCONDUCTING COIL

### TRACKERS



Pixels  
Silicon Microstrips

MUON ENDCAPS



Cathode Strip Chambers (CSC)  
Resistive Plate Chambers (RPC)

Total weight : 12,500 t  
Overall diameter : 15 m

Overall length : 21.6 m  
Magnetic field : 4 Tesla

Detector	Channels	Control	Ev. Data
Pixel	60000000	1 GB	50 (kB)
Tracker	10000000	1 GB	650
Preshower	145000	10 MB	50
ECAL	85000	10 MB	100
HCAL	14000	100 kB	50
Muon DT	200000	10 MB	10
Muon RPC	200000	10 MB	5
Muon CSC	400000	10 MB	90
Trigger		1 GB	16

**Event size**  
**Max LV1 Trigger**  
**Online rejection**  
**System dead time**

**1 Mbyte**  
**100 kHz**  
**99.999%**  
**~ %**