The data acquisition and reduction challenge at LHC

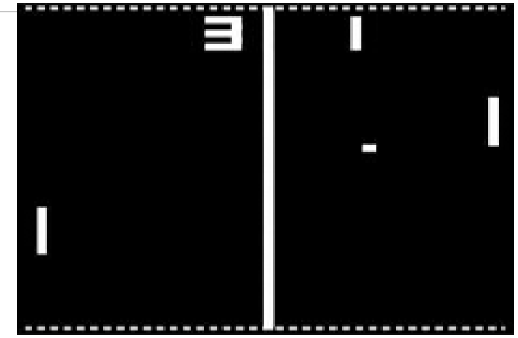
Thessaloniki, ISOTDAQ 2013
S. Cittolin (University California San Diego)

- **Introduction**
  - Data handling requirements at LHC
- **Design issues: Architectures**
  - Front-end, event selection levels
- **Design issues: Technologies**
  - Project history and technologies trends
  - Predicted and unpredicted evolutions
- **Conclusion**

LHC&TDAQ project timeline (the time of a generation)

1984 Lausanne workshop (LEP/LHC)
1990 Design of experiments
1992 CMS Letter of Intent
1994 Technical Design Report
1996 LHC project approved
1998
2000 Trigger Technical Design Report
2001 LHC cost review
2002 DAQ Technical Design Report
2006 Magnet test Global Run
2005
2008 Circulating beam Global Run
2009 Colliding beams
2010 Start physics runs

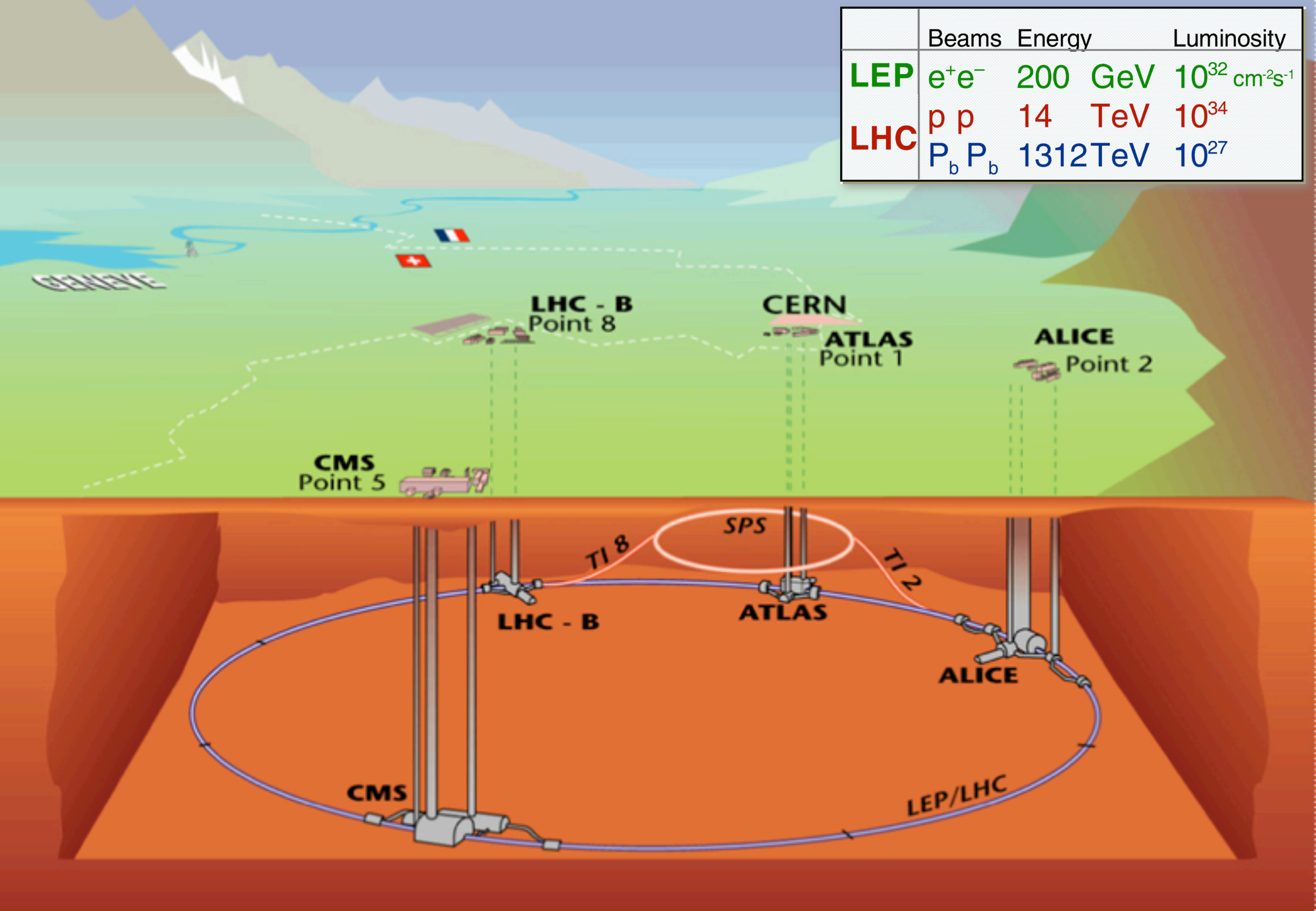
1990-1995
Research and Development



1996-2002
Prototypes and Demonstrators

2003-2005
Final Design. Choice of technologies

2006-2008
Construction
and commissionin

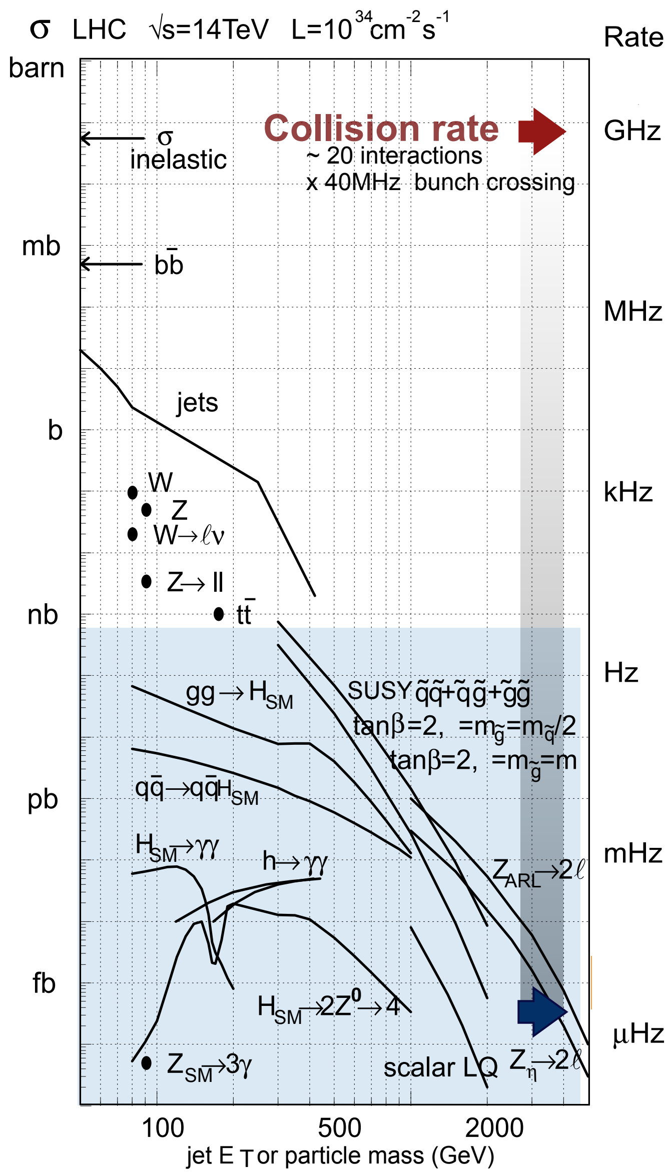


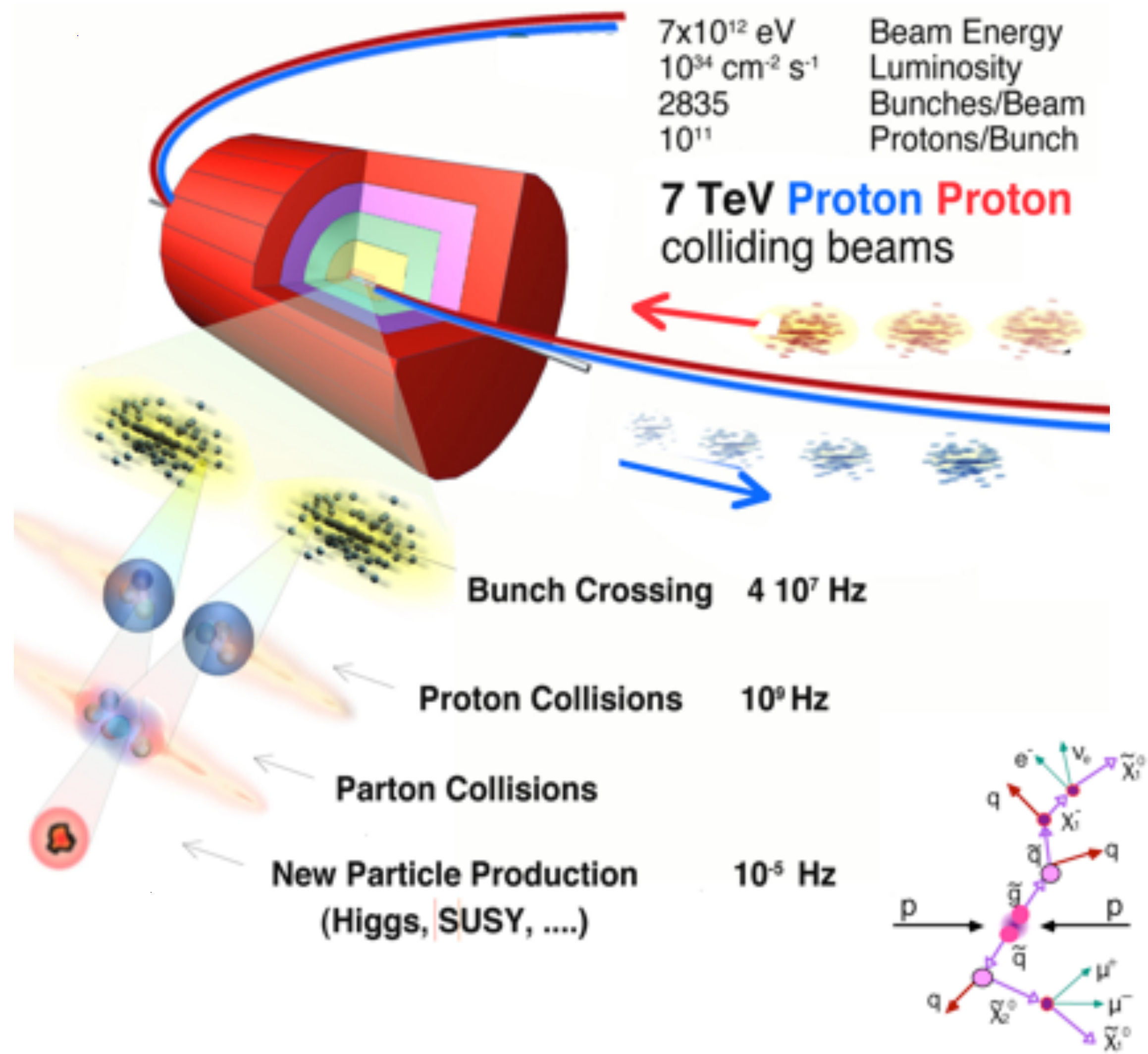
	Beams	Energy		Luminosity
LEP	e^+e^-	200	GeV	10^{32} $cm^{-2}s^{-1}$
LHC	p p	14	TeV	10^{34}
	$P_b P_b$	1312	TeV	10^{27}

DAQ design issues at LHC (1990-2010)

- **Physics and rates**
- Experiments
- Collisions and detector front-end
- Event selection levels
- DAQ readout network

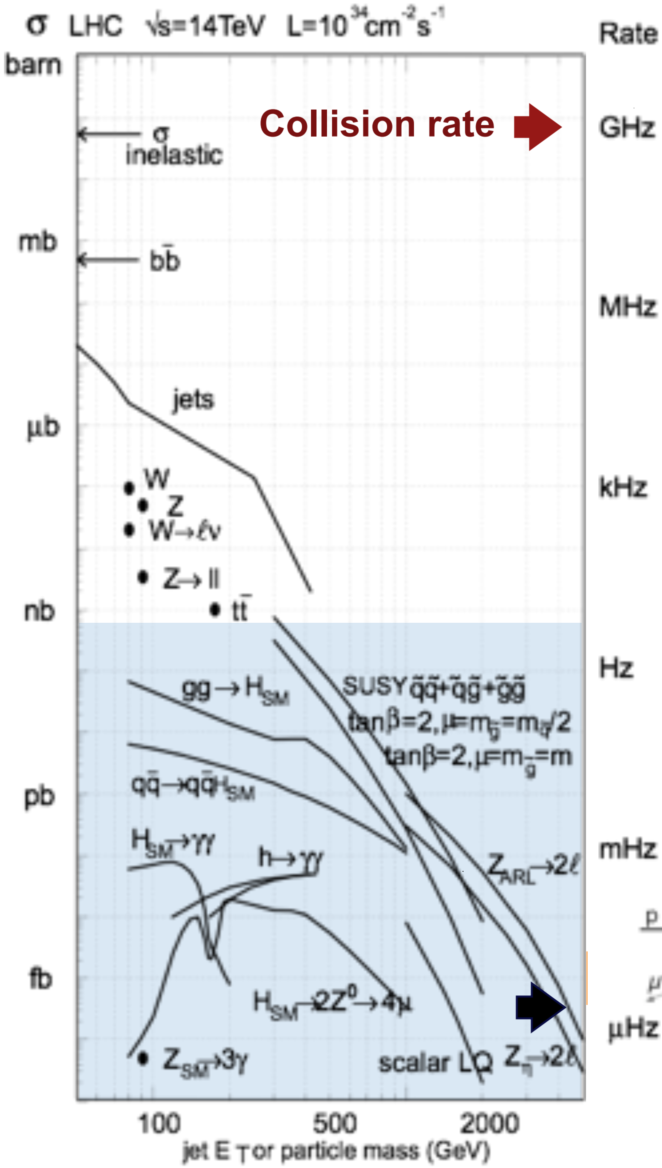
Proton-proton collisions at LHC. Searching issue

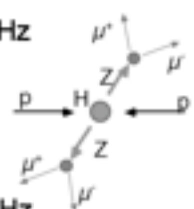


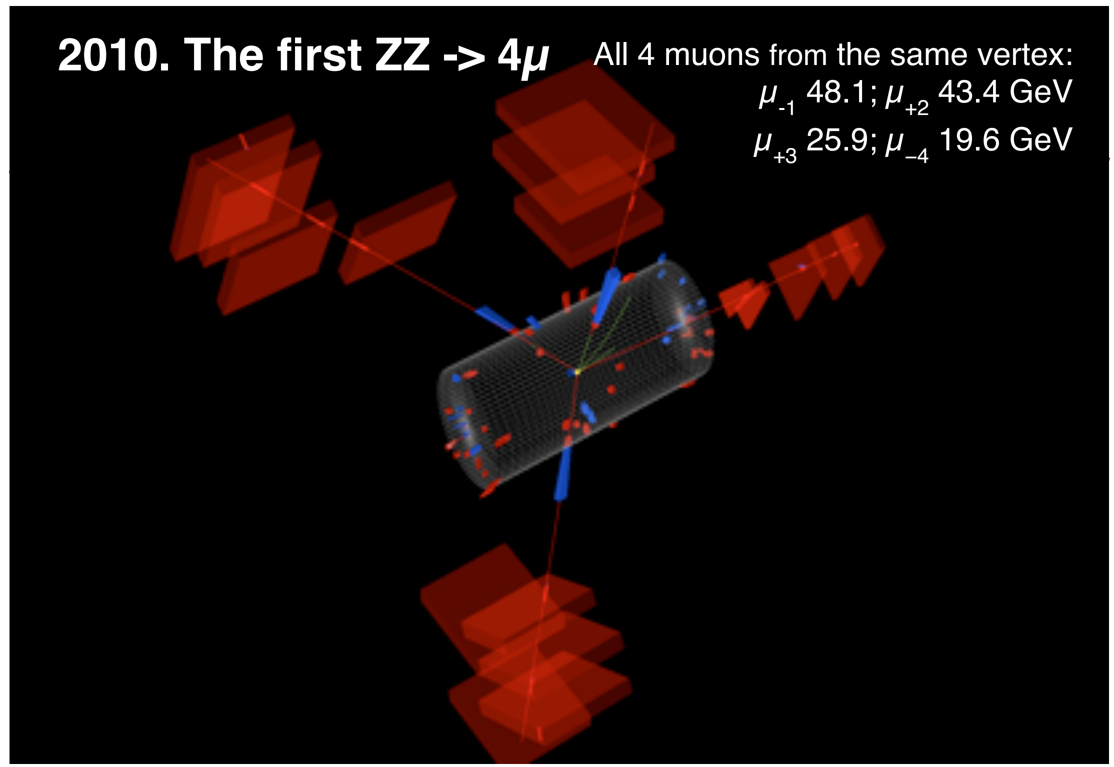


Collision Rate: ~10^9 Hz. Event Selection: ~$1/10^{13}$

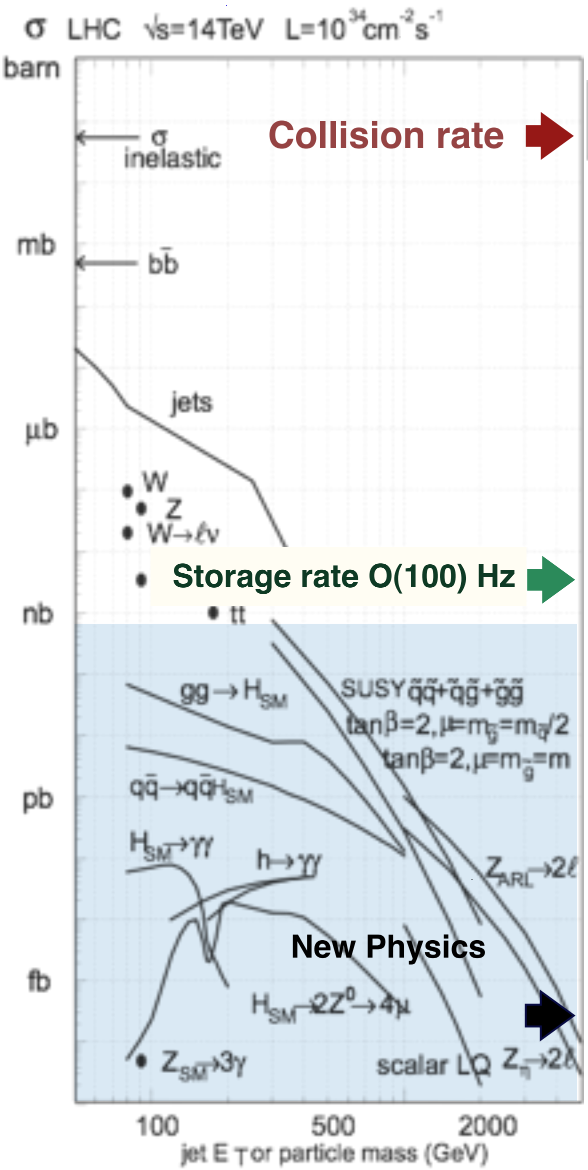
Data detection, event identification and event analysis

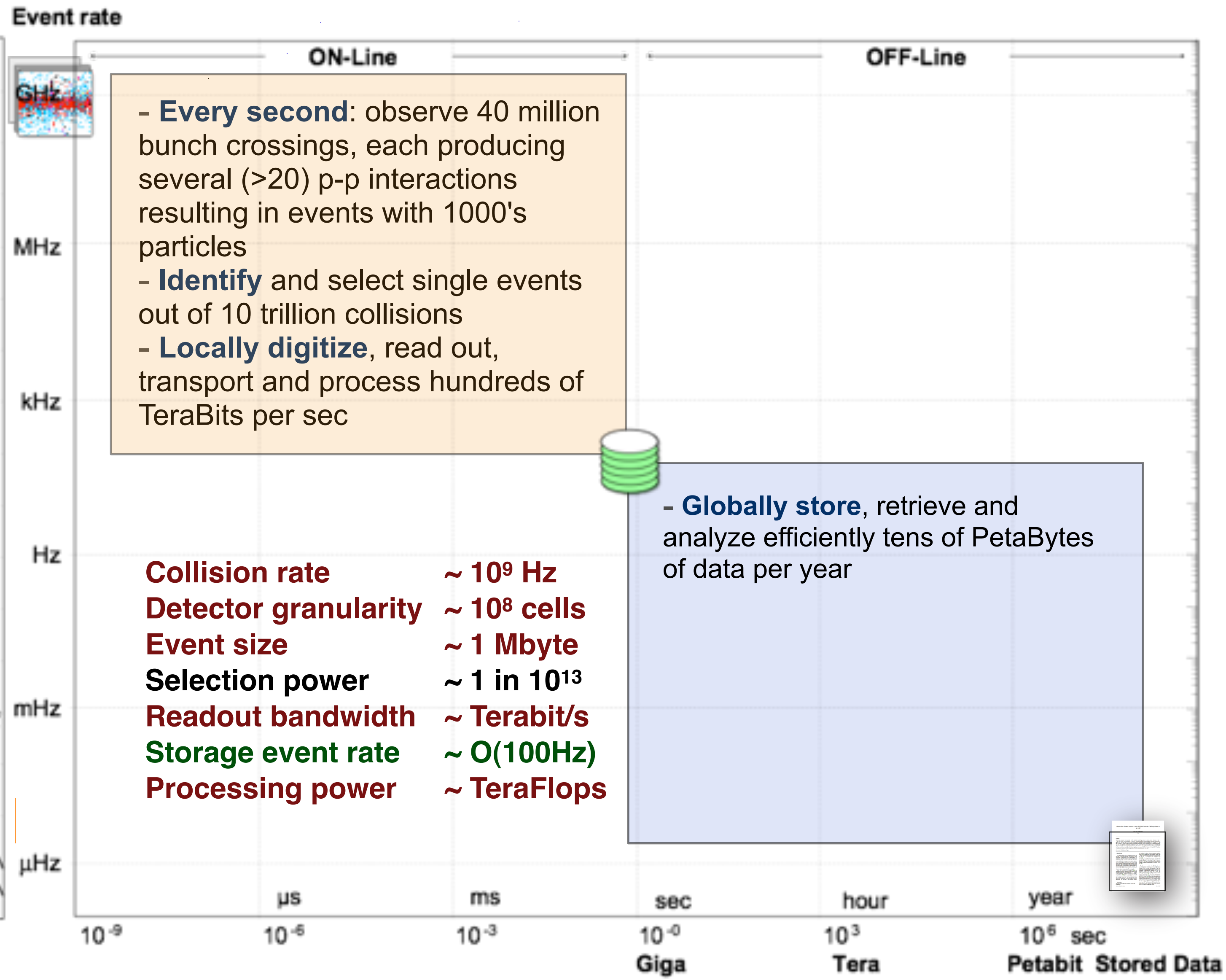






Physics at LHC: overall data handling requirements





- **Every second**: observe 40 million bunch crossings, each producing several (>20) p-p interactions resulting in events with 1000's particles
- **Identify** and select single events out of 10 trillion collisions
- **Locally digitize**, read out, transport and process hundreds of TeraBits per sec

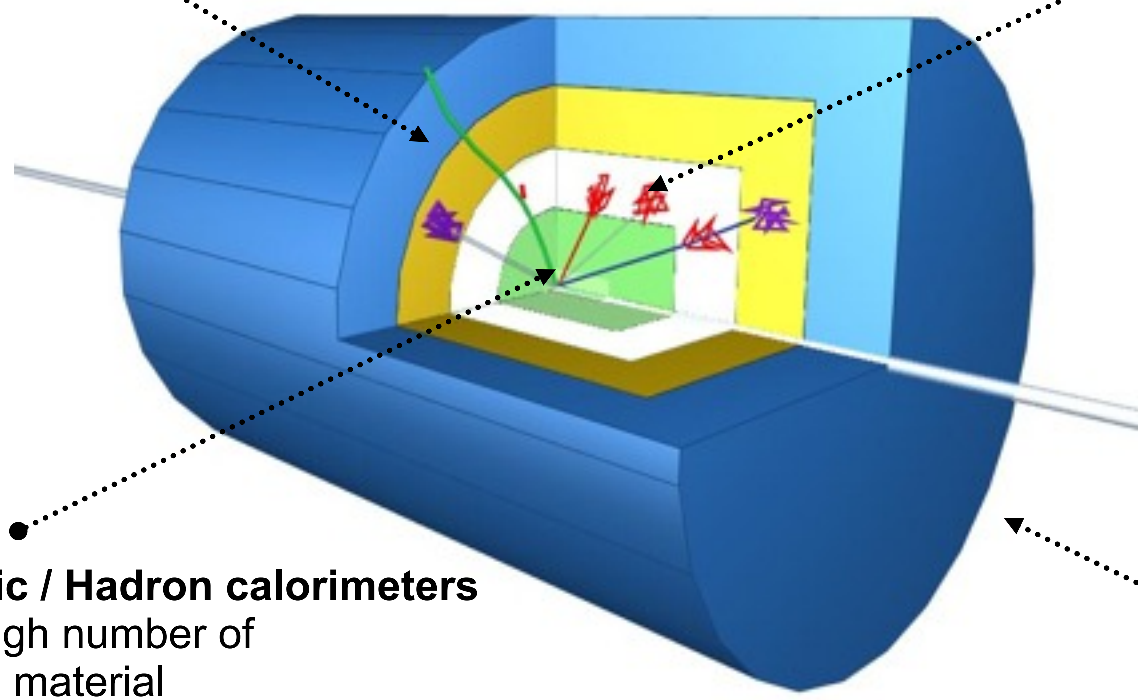
- **Globally store**, retrieve and analyze efficiently tens of PetaBytes of data per year

Collision rate	**~ 10^9 Hz**
Detector granularity	**~ 10^8 cells**
Event size	**~ 1 Mbyte**
Selection power	**~ 1 in 10^{13}**
Readout bandwidth	**~ Terabit/s**
Storage event rate	**~ O(100Hz)**
Processing power	**~ TeraFlops**

DAQ design issues at LHC (1990-2010)

- Physics and rates
- **Experiments**
- Collisions and detector front-end
- Event selection levels
- DAQ readout network

General purpose detector at p-p colliders



Muon detectors
Heavy materials
μ identification

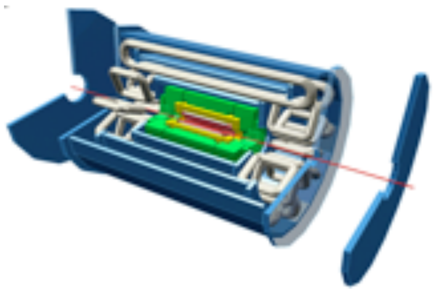
Central detectors
Light materials
Tracking, pt, MIP
Em. shower position
Topology
Interaction Vertex

Electromagnetic / Hadron calorimeters
Materials with high number of protons + Active material
Particle identification (e, gamma, Jets, Missing Et)
Energy measurements

4π calorimetry
Hermetic calorimetry
Missing Et measurements

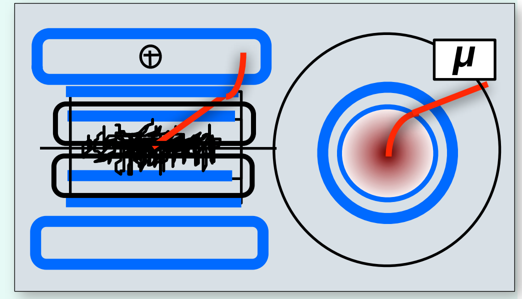
Each layer identifies and enables the measurement of the momentum or energy of the particle produced in a collision

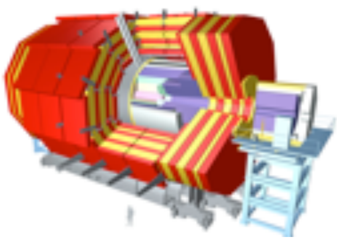
The Experiments



ATLAS

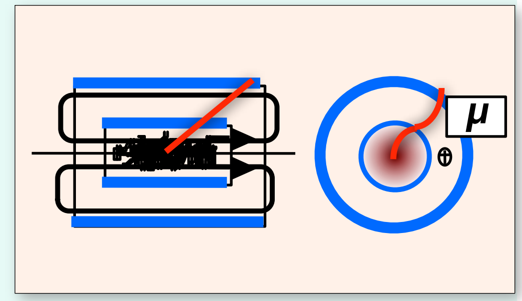
Study of **pp** collisions
Tracker: Si(Pixel and SCT), TRT
Calorimeters:LAr, Scintillating Tiles
Muon System: MDT, RPC, TGC, CSC,
Magnets: Solenoid and Toroid





CMS

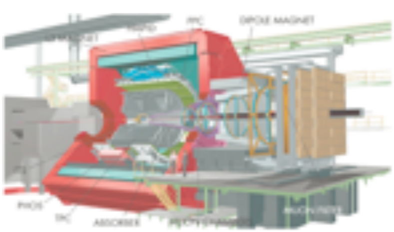
Study of **pp & heavy ion collisions**
Tracker: Si (Pixel, Strips, Discs)
Calorimeters: BGO, Brass Scintillators, Preshower
Muon System: RPC, MDT, CSC,
Supraconducting solenoid



Magnetic character

LHCb

Study of **CP violation in B decays** (pp)
Tracker (Si, Velo), 2 RICH, 4 Tracking stations (Straw-Tubes, Si), SPD (scinitll. Pads), Preshower, ECAL (lead scintillator) HCAL(steel scintillator), Muon stations (MWPCs)



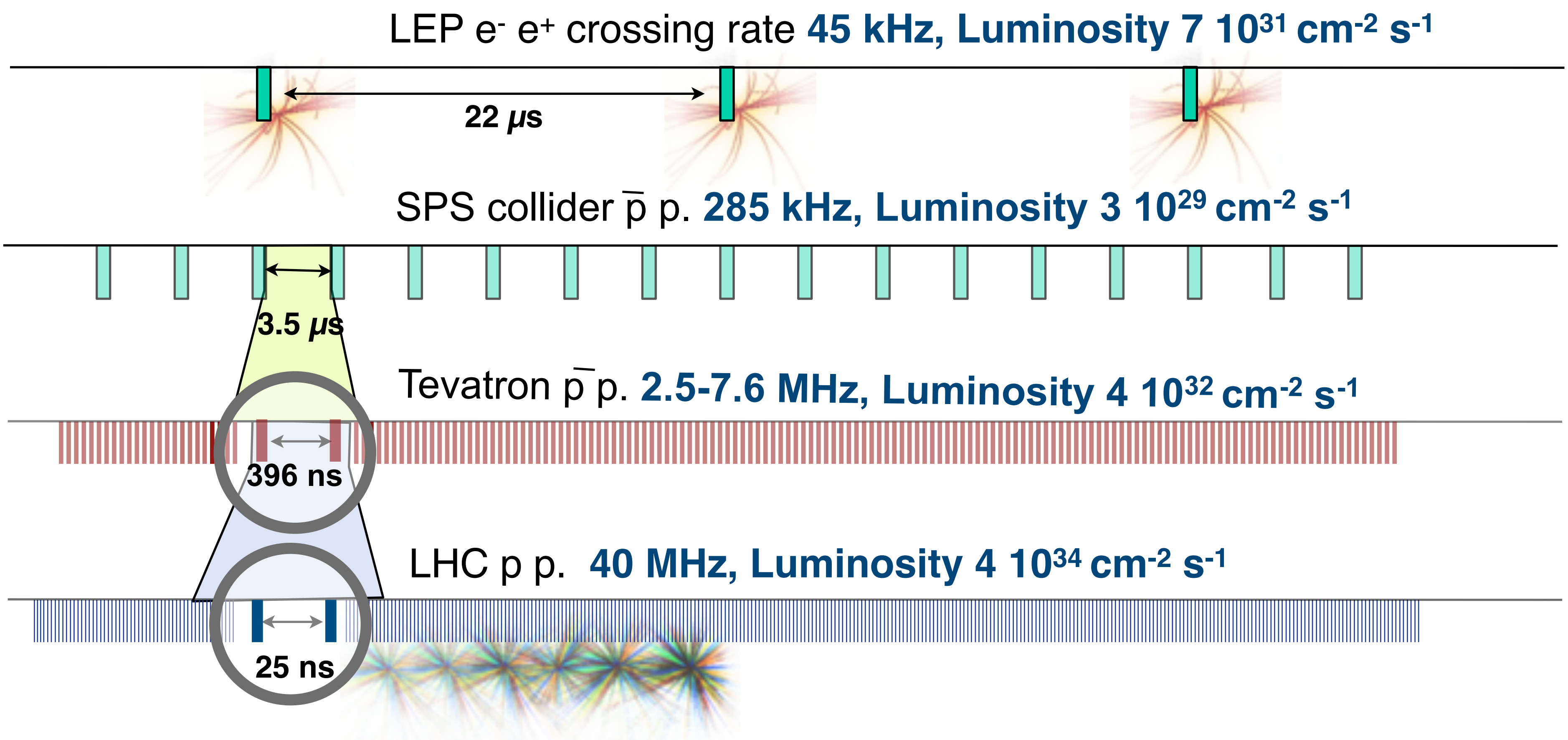
ALICE

Study of **heavy ion collisions**
Tracker: Si (ITS), TPC, Chambers, TRD, TOF
Particle Id: RICH, PHOS (scintillating crystals)
RPC, FMD(froward mult.; Si) ZDC (0 degree cal)
Magnets: Solenoid, Dipol

DAQ design issues at LHC (1990-2010)

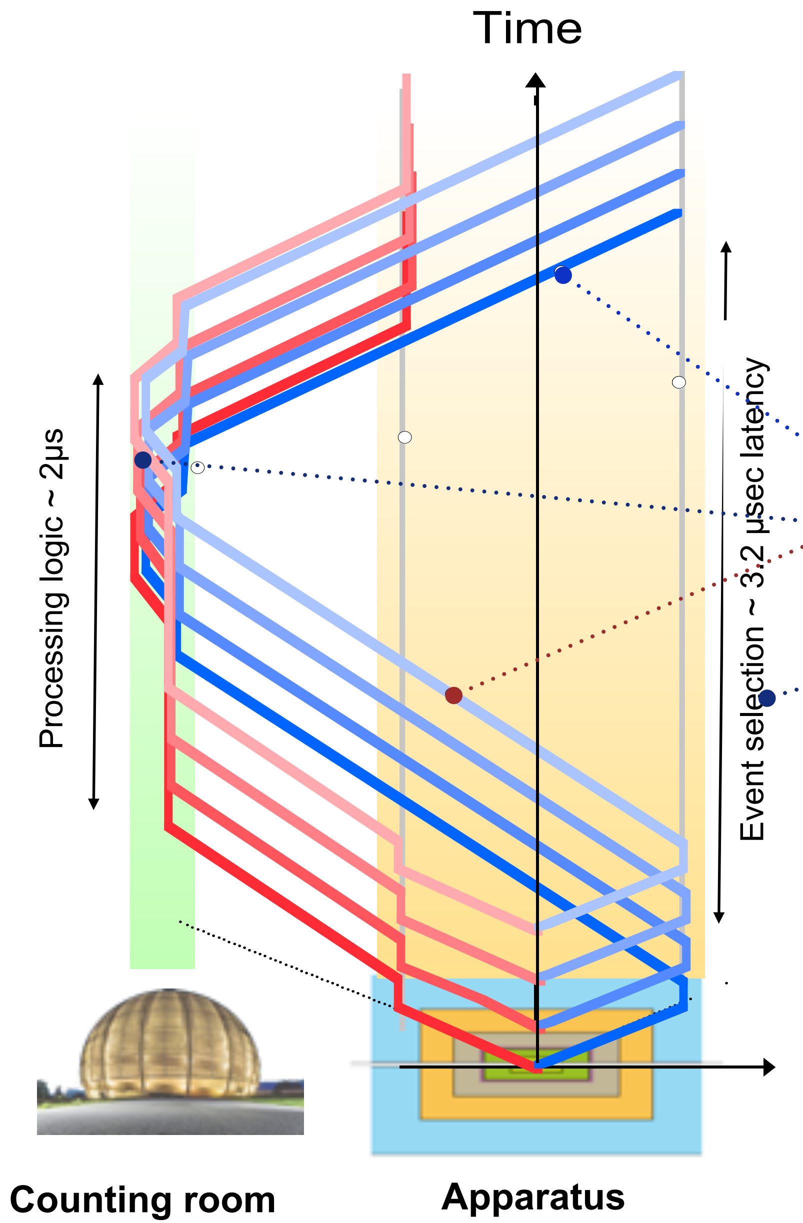
- Physics and rates
- Experiments
- **Collisions and detector front-end**
- Event selection levels
- DAQ readout network

Colliders bunch crossing frequencies

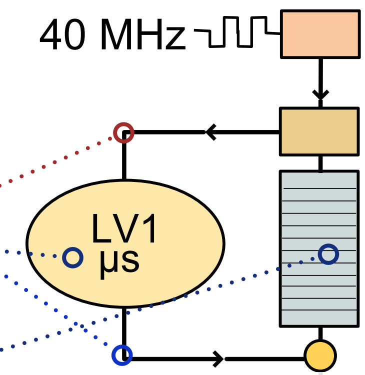


- **25 ns** defines an overall time constant for signal integration, DAQ and trigger.
- The rate of the collisions (**40 MHz**) is (was) not affordable by any data taking system.
- The off-line computing budget and storage capacity limit the **output rate** (~**100 Hz**)

Data signal propagation and front-end structure



The **front-end** system includes preamplifier, shaper, digitizer and the buffers to hold the data for the duration of the level-1 trigger decision and the signal propagation delay (~ 3 μs)

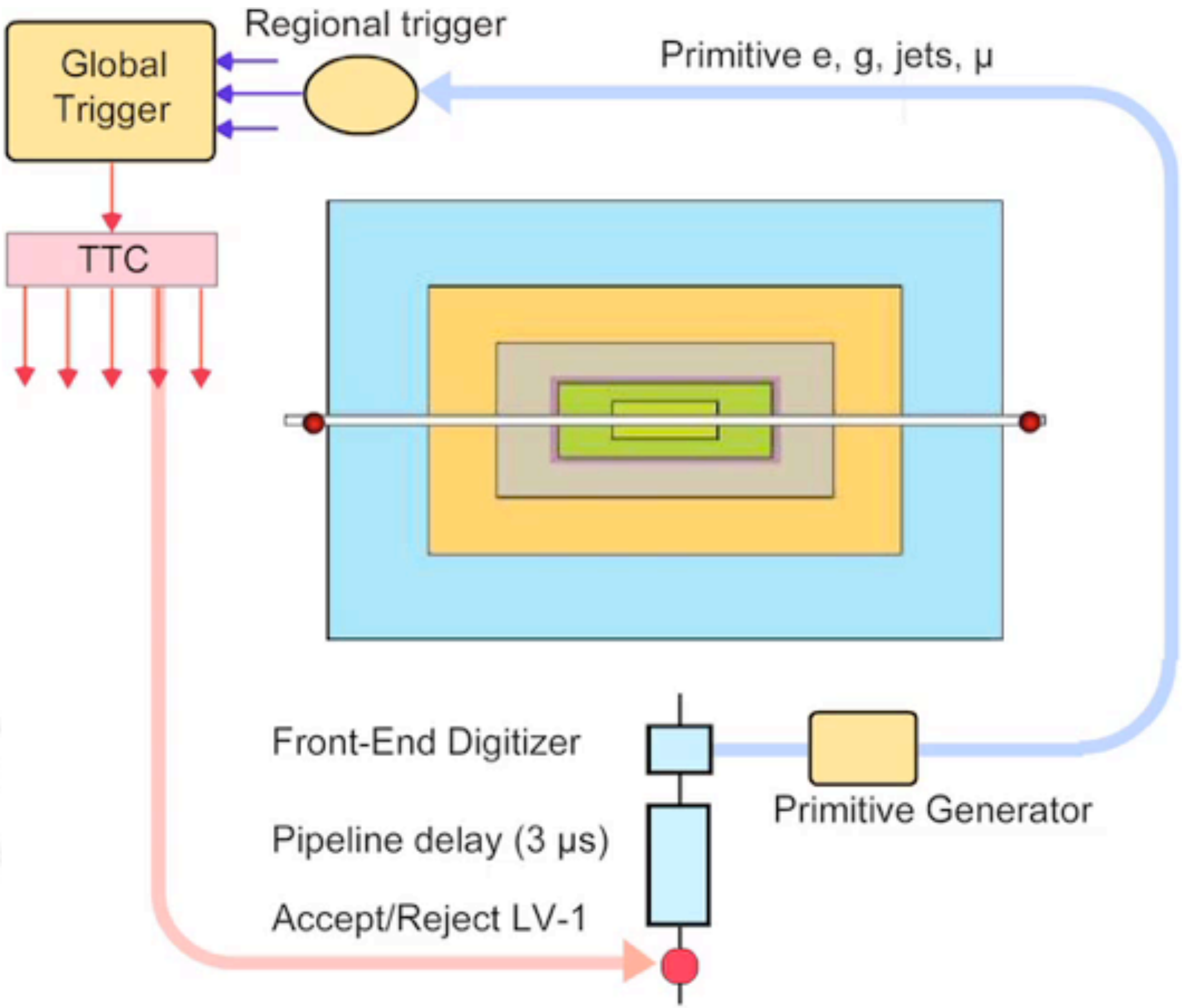


The **Level-1** trigger is a centralised multi-step logic system collecting, buffering and processing sub-set of detector data every 25 ns.

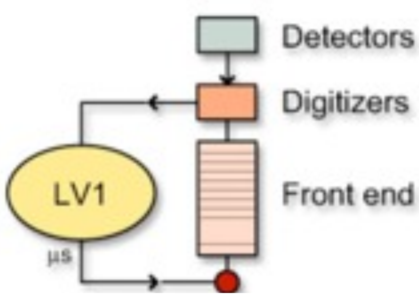
Front-end time budget (3.2μs = 128 Bunch crossings)

Signals to central logic	18bx	450 ns
Muon/Calorimeter logic	90bx	2250 ns
Back to detectors	18bx	450 ns
total latency	**≤128bx**	**3.2 μs**

Front-end timing and readout structure



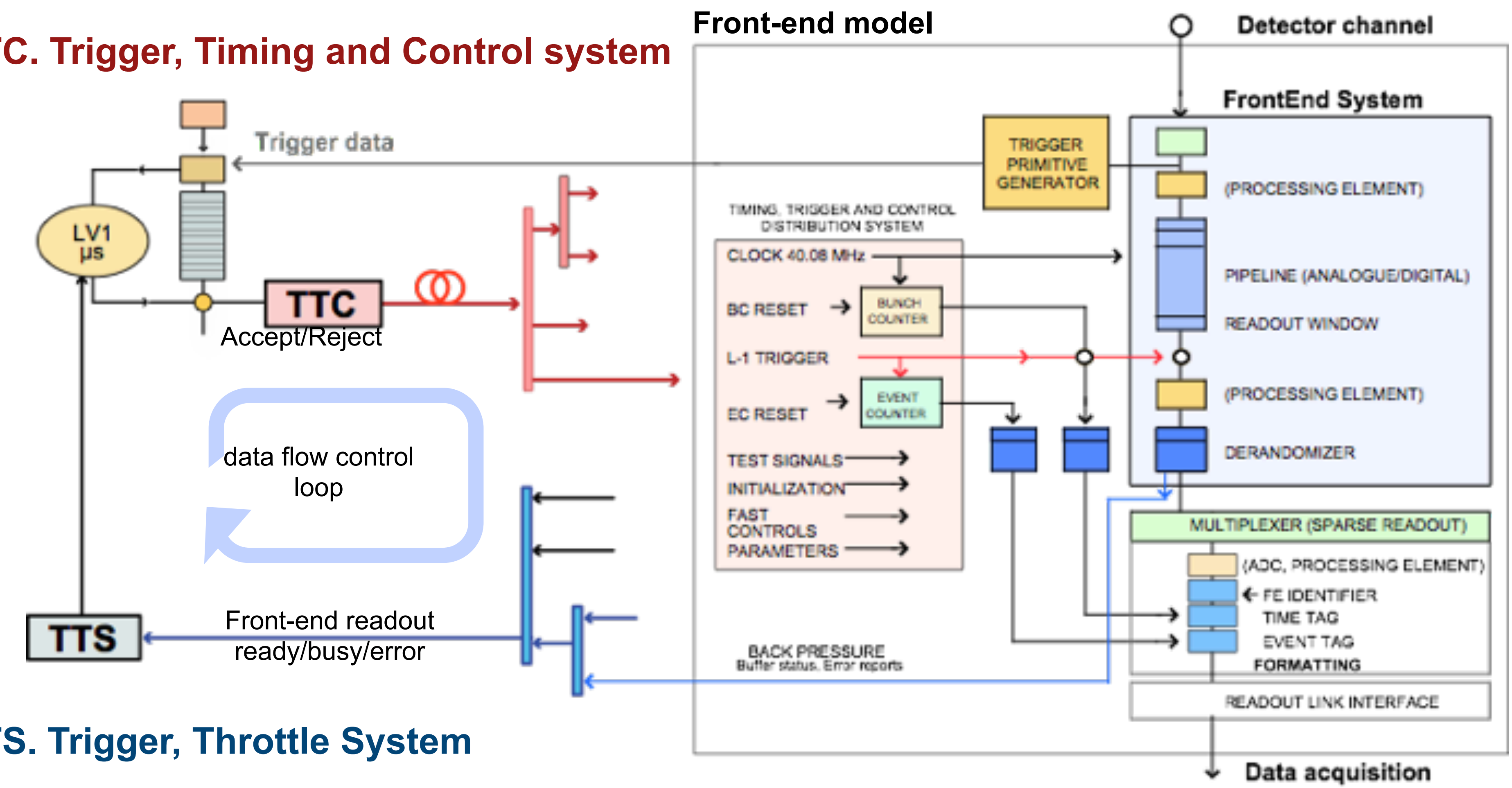
TTC. A multichannel **optical distribution system broadcasts the LHC 40 MHz clock** and the Global Trigger signals to several thousand destinations



Front-end generic model and timing system

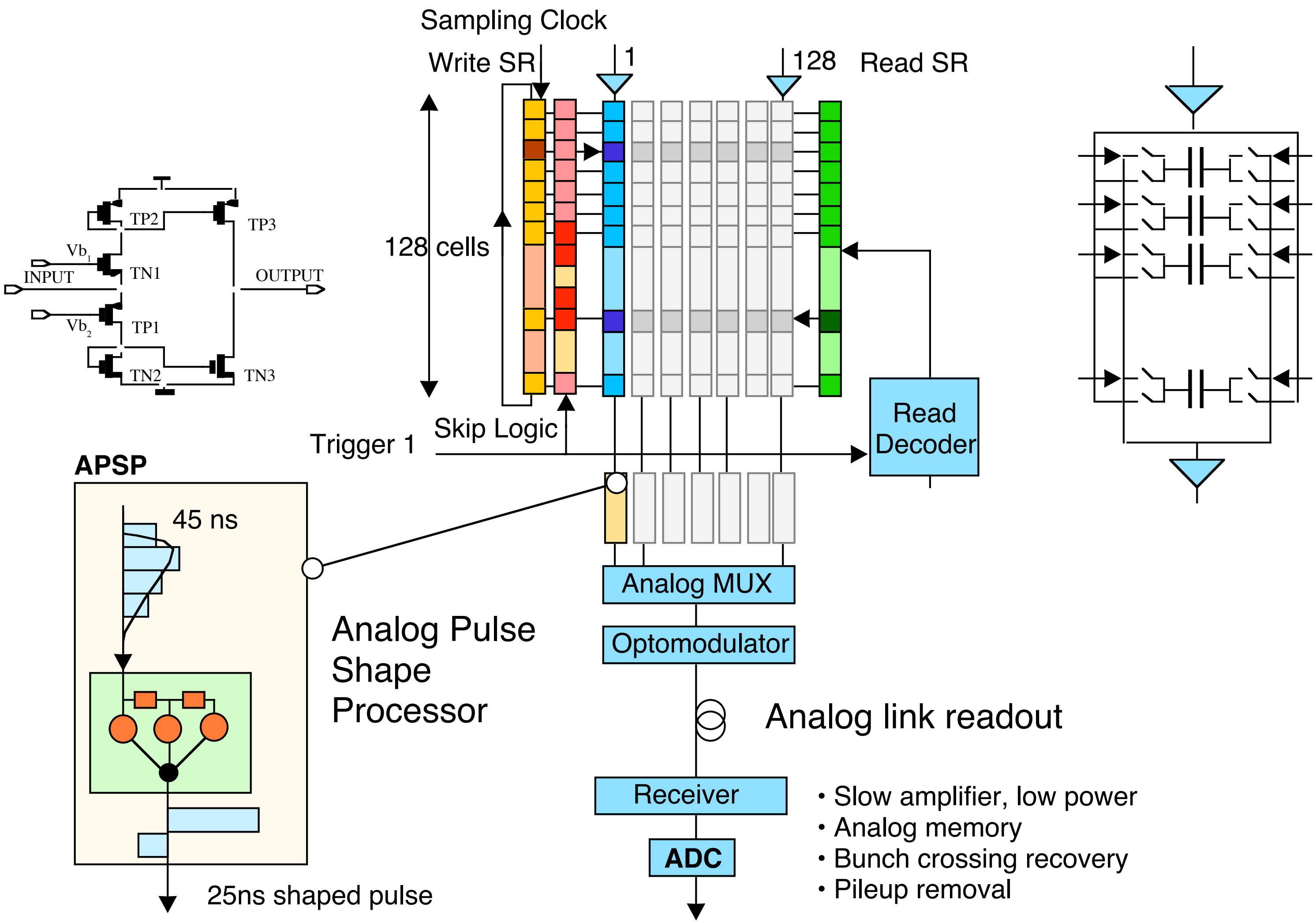
TTC. A multichannel **optical distribution system broadcasts the LHC 40 MHz clock** and the Global Trigger signals to several thousand destinations

TTC. Trigger, Timing and Control system



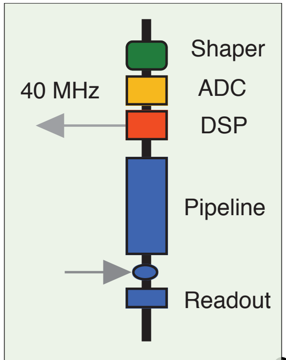
TTS. Trigger, Throttle System

Analog synchronous front-end: Silicon strip



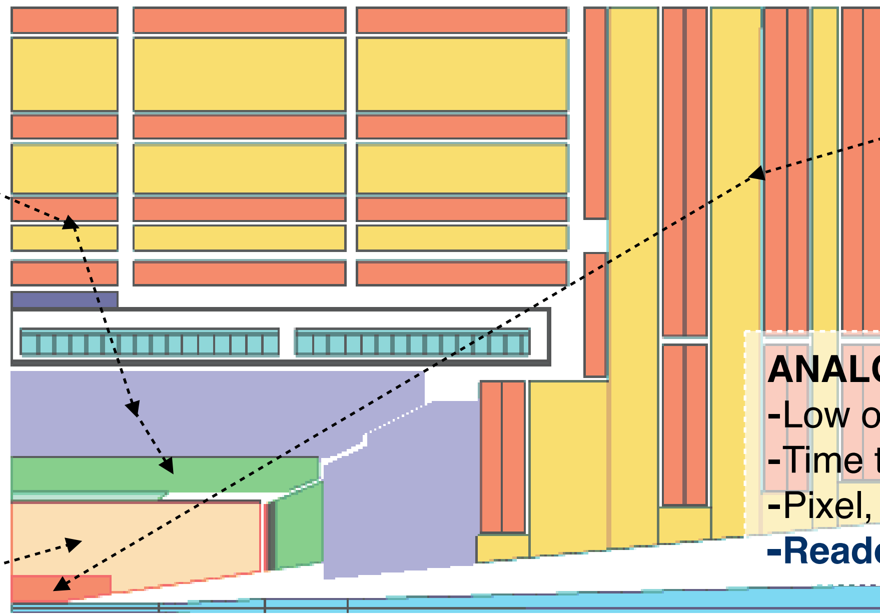
- Slow amplifier, low power
- Analog memory
- Bunch crossing recovery
- Pileup removal

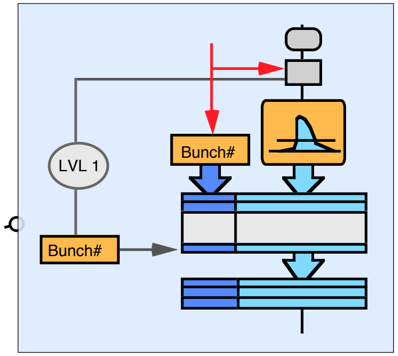
Analog/Digital front-end systems (CMS)



500.000 ch.

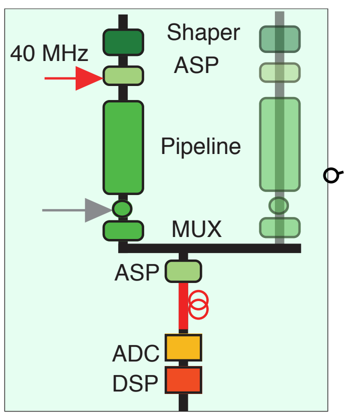
DIGITAL (25 ns) synchronous pipeline
- Large dynamic range (15bits)
- Digital filtering (high power consumption)
- ECAL, HCAL Calorimeters, DT, RPC Barrel Muons
- **Readout: OPTICAL digital** links





60.000.000 ch.

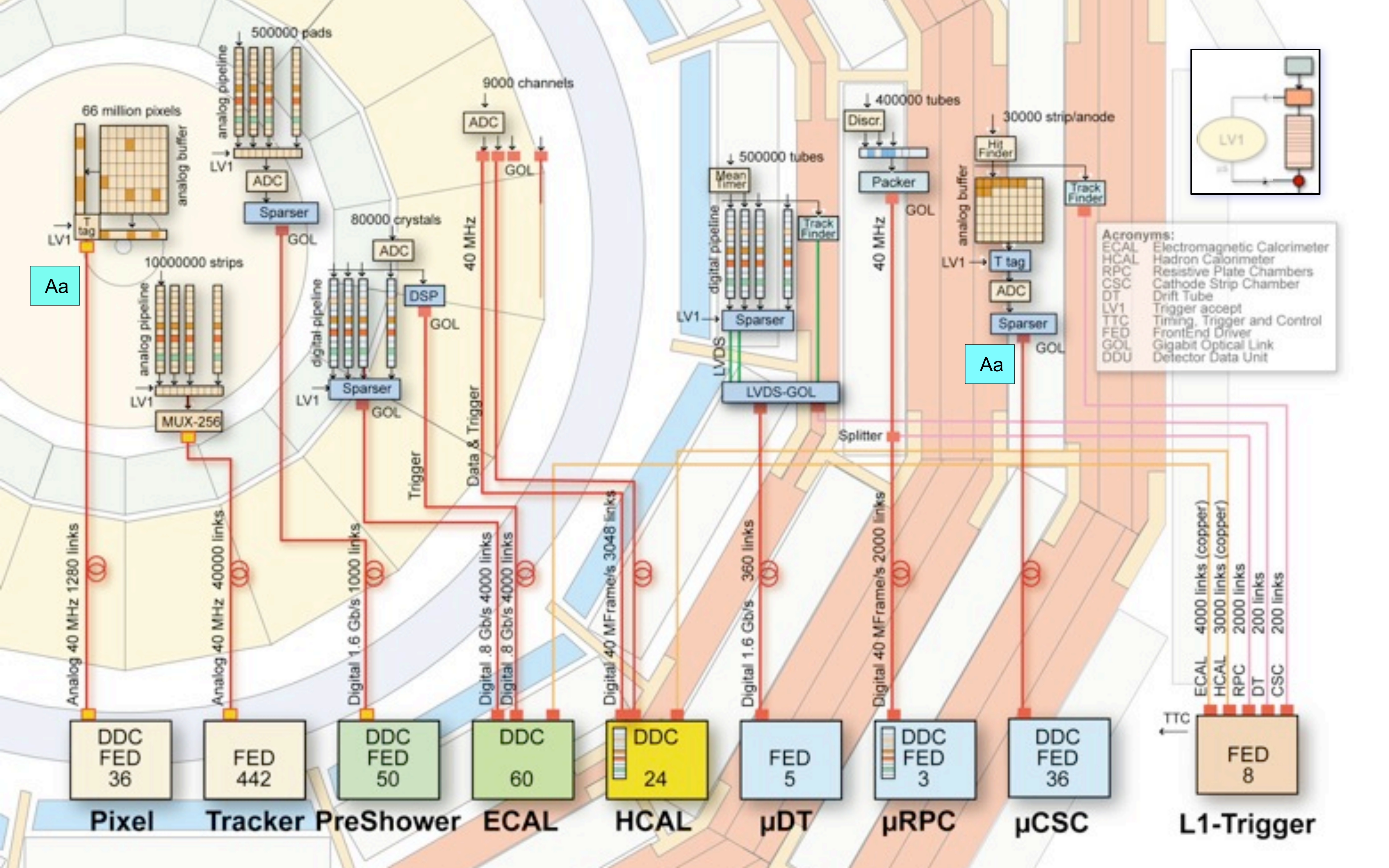
ANALOG/DIGITAL asynchronous buffer
- Low occupancy. High No.channels
- Time tag identification system
- Pixel, CSC forward muons
- **Readout: OPTICAL&COPPER**



10.000.000 ch.

ANALOG (25 ns) synchronous pipeline
- Low dynamic range and resolution(≤ 10-bits)
- Low power consumption, radiation tolerant,
- High number of channels.Tracker, Pre-shower
- **Readout: OPTICAL analog** links

CMS front-end systems

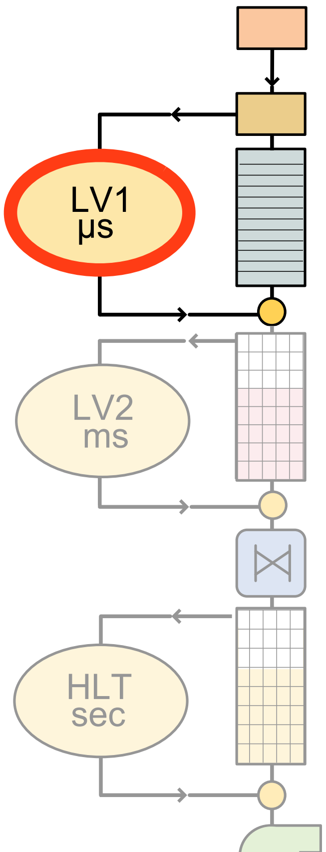


DAQ design issues at LHC (1990-2010)

- Physics and rates
- Experiments
- Collisions and detector front-end
- **Event selection levels**
- DAQ readout network

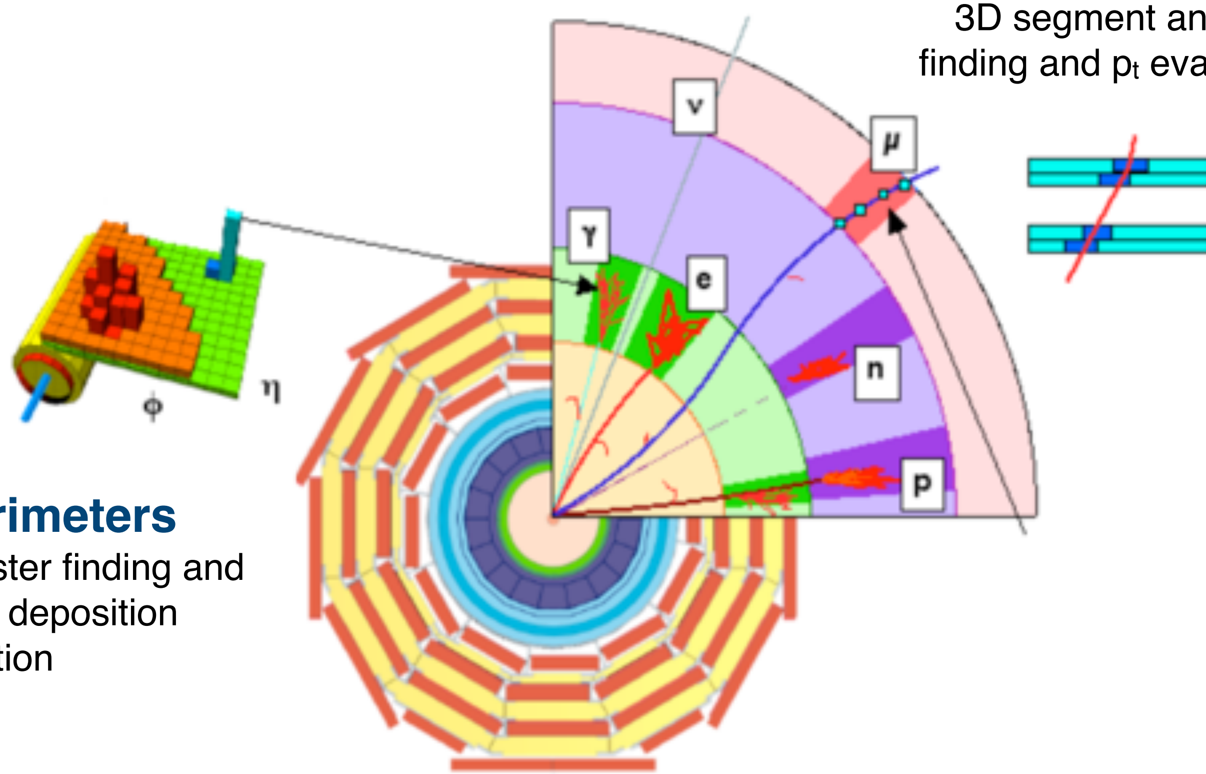
Level-1 trigger. Particle identification

Use signals from fast detectors (calorimetry and muon systems) to identify: **high p_t electron, muon, jets, missing E_T**



Muon systems

3D segment and track finding and p_t evaluation



Calorimeters

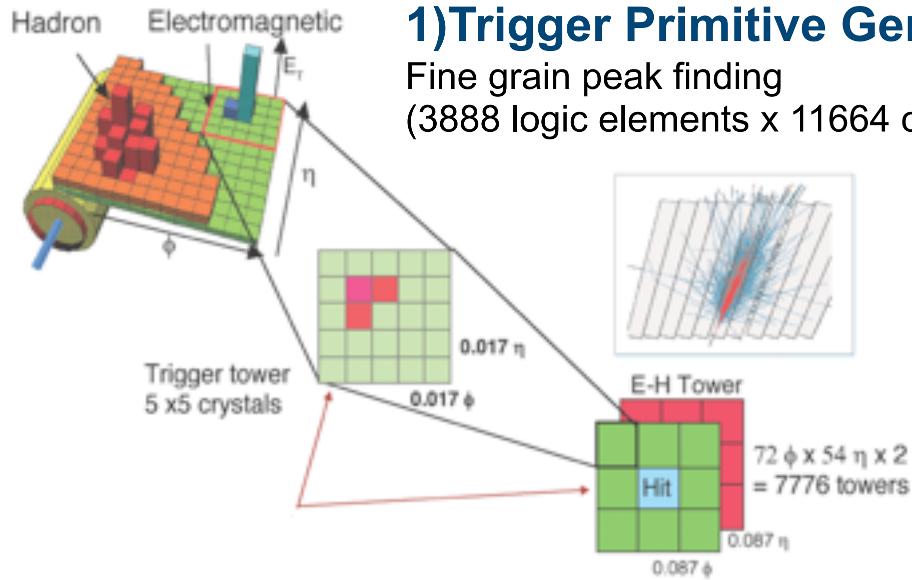
2D cluster finding and energy deposition evaluation

Algorithms run on local **calorimeter and muon coarse data.**

With **new data every 25 ns and decision latency ~µs**

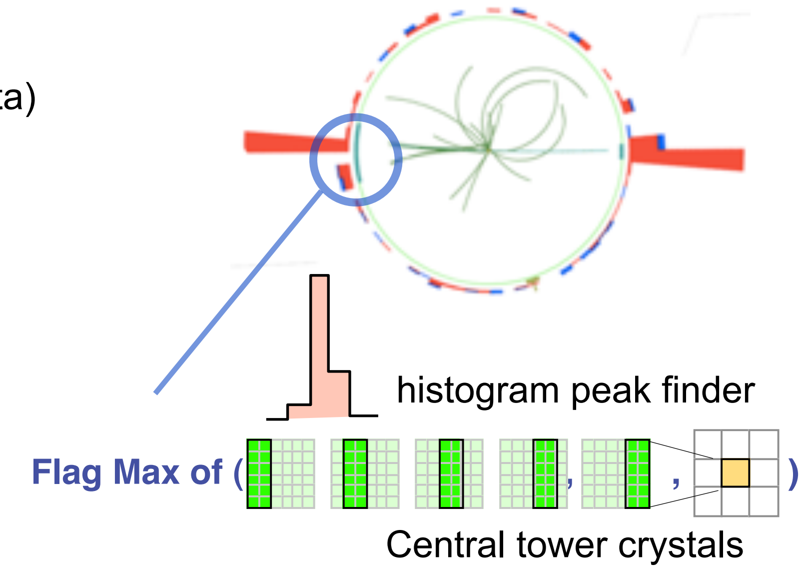
Special-purpose hardware reduces event rate (to be read out) **from 40 MHz to 100 kHz.**

Level-1: Z -> e^+e^-



1)Trigger Primitive Generator

Fine grain peak finding
(3888 logic elements x 11664 crystal data)



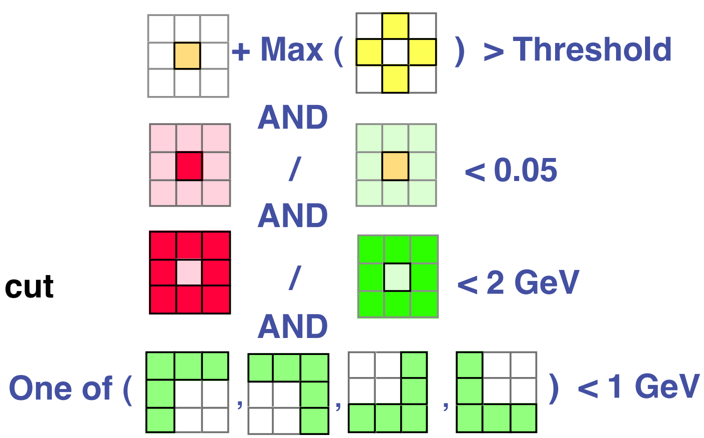
2) Pixel Processor

3888 logic elements x 34992 pixel data)

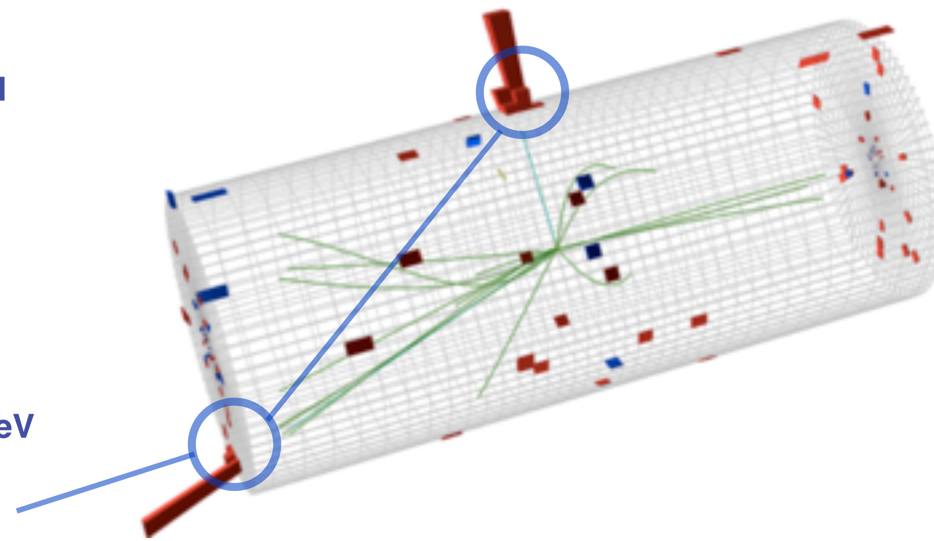
E_t cut

Longitudinal cut (H/E)

Neighbors longitudinal cut

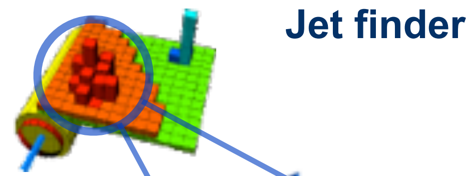


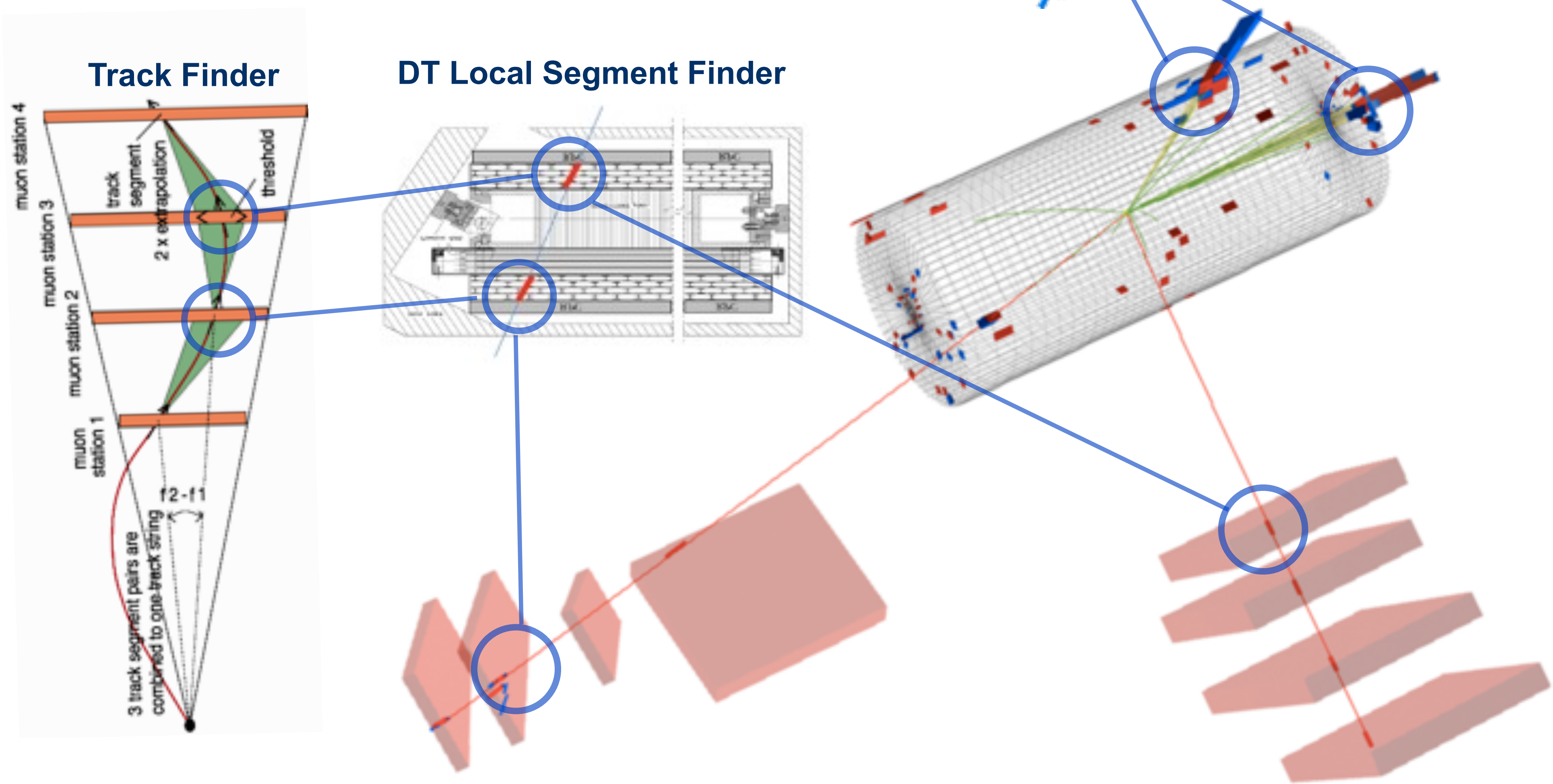
ISOLATED ELECTRON



DT and CSC track finding:

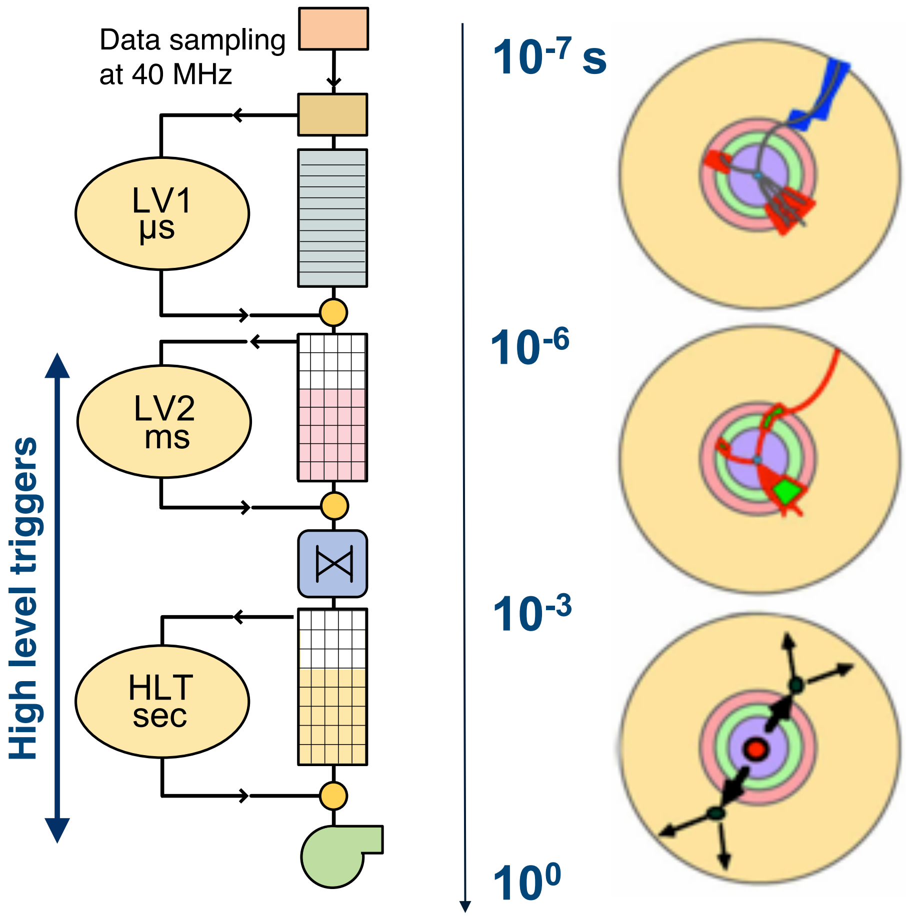
Based on track segments identification by of mean timer logic followed by full Track Finder reconstruction by associating multiple plane vectors to allowed particle trajectories





On-line multi-levels trigger generic architecture

Successively more complex decisions are made on successively lower data rates



Level-1 input: 40 MHz output: 100 kHz

Particle identification (High p_T, e, μ, jets, missing E_T)

- Local pattern recognition
- Energy evaluation on prompt macro-granular information

Level-2 input: 100 kHz. Output 1 KHz

Clean particle signature (Z, W, ...)

- Finer granularity, precise measurement
- Kinematics, effective mass cuts and event topology
- Track reconstruction and detector matching

Level-3 input 1 kHz. Output: O(100Hz)

Physics process identification

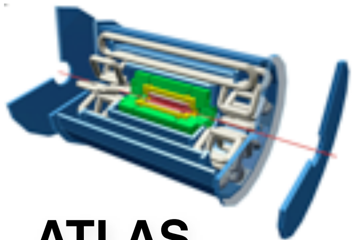
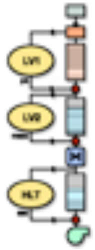
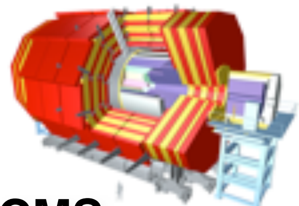
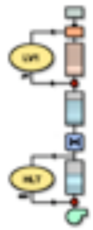
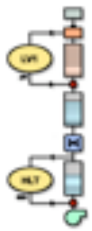
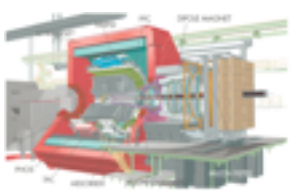
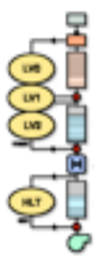
- Event reconstruction and analysis

On-line global selection:

99.999 % rejected, 0.001 % selected

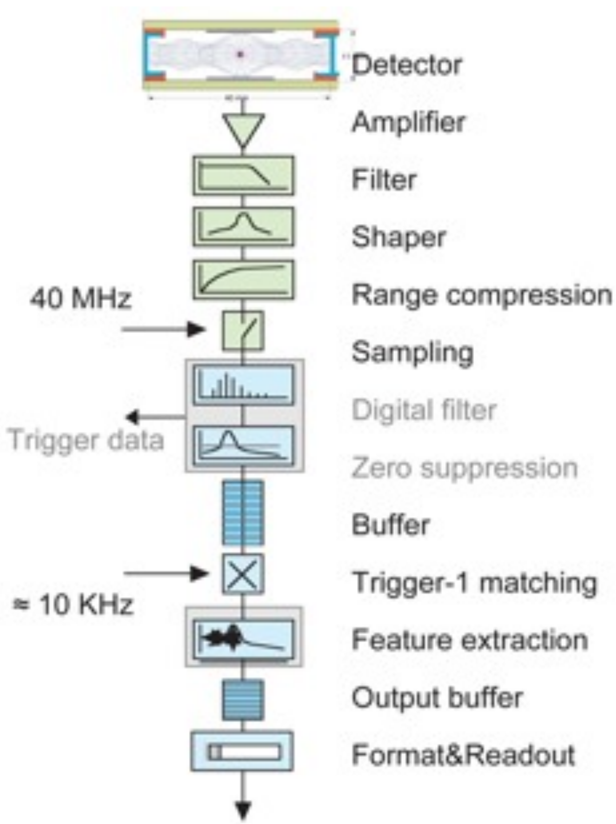
Readout and trigger dead-time must be kept at minimum (typically of the order of few %)
The trigger system has to maximise the collection of data for physics process of interest at all levels, since **rejected events are lost for ever**

LHC experiments TDAQ summary

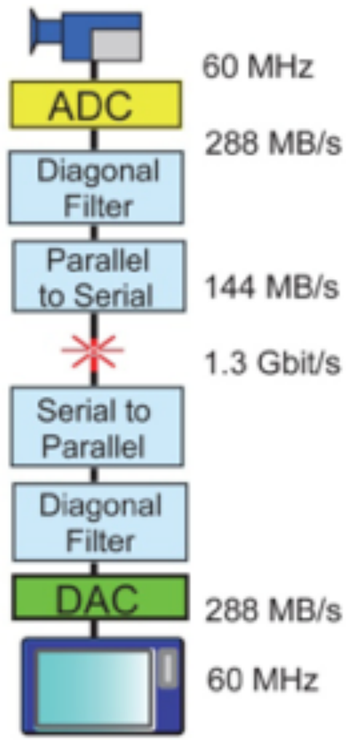
	Trigger No. Levels	Level-0,1,2 Rate (Hz)	Event Size (Byte)	Readout Bandw.(GB/s)	HLT Out MB/s (Event/s)
 **ATLAS**	**3** 	LV-1 10^5 LV-2 3×10^3	1.5×10^6	4.5	300 (2×10^2)
 **CMS**	**2** 	LV-1 10^5	10^6	100	O(1000) (10^2)
 **LHCb**	**2** 	LV-0 10^6	3×10^4	30	40 (2×10^2)
 **ALICE**	**4** 	Pb-Pb 500 p-p 10^3	5×10^7 2×10^6	25	1250 (10^2) 200 (10^2)

Technical challenges I: Digital signal processing

1990. LHC detector channel



1990. HDTV chain



One HDTV = One LHC channel

Analog bandwidth	~ 100 MHz
Digital resolution	12_14 bits
Digital bandwidth	~ 1 Gb/s

Since early 80's:

- Digital Signal Processing (**DSP**) has become pervasive at all levels in our society.

- **DSP** as a technology has become the primary growth driver for the entire semiconductor market.

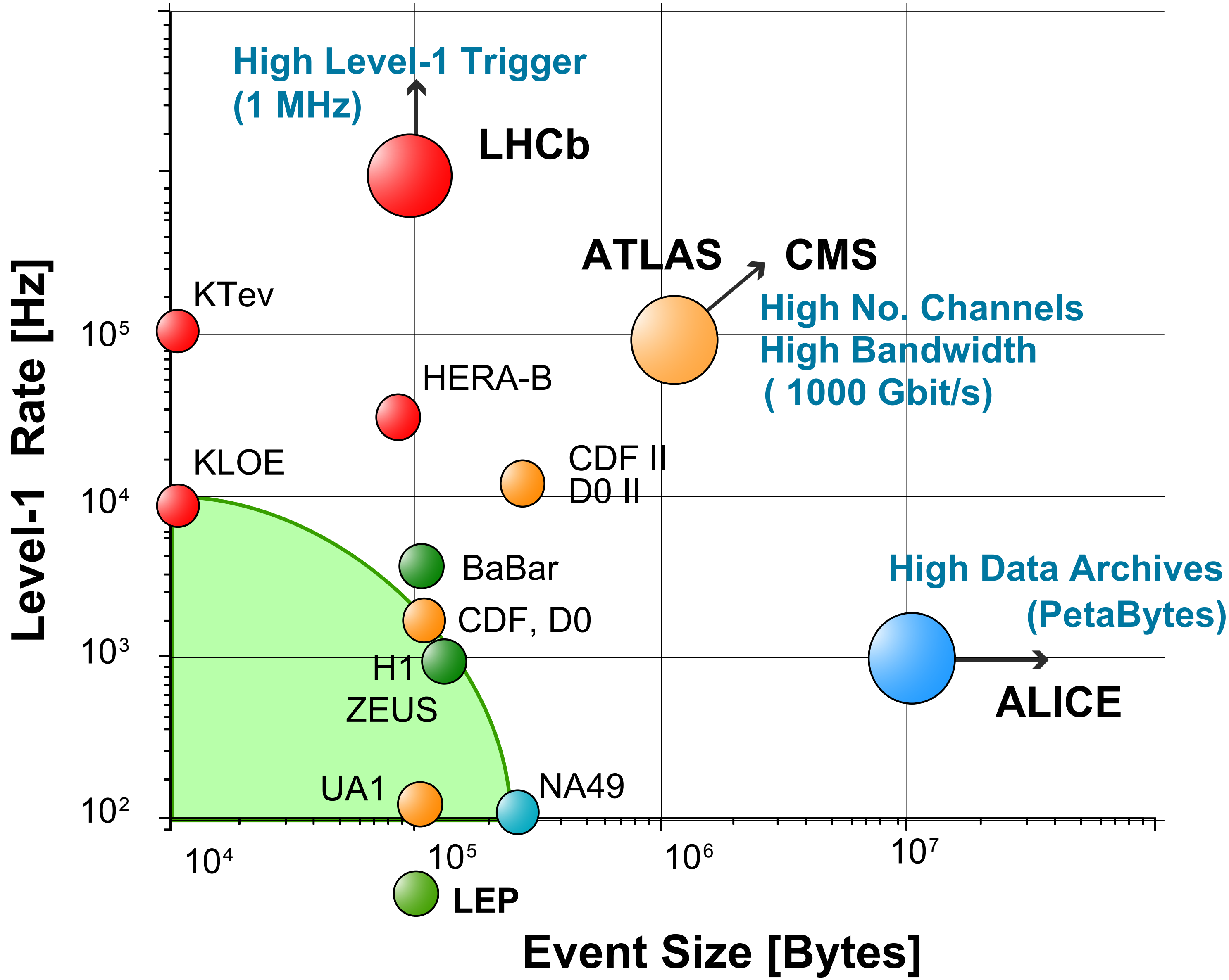
- The telecommunication industry has been one of the major customers for the development of this technology.

- Analog to digital converters (**ADC**)
- Multiply accumulator (**MAC**)
- GHz **optical links** and Laser LED
- Finite Impulse Response (**FIR**) digital filters and vector processing are today the **building blocks of any LHC detector readout chain** as well.

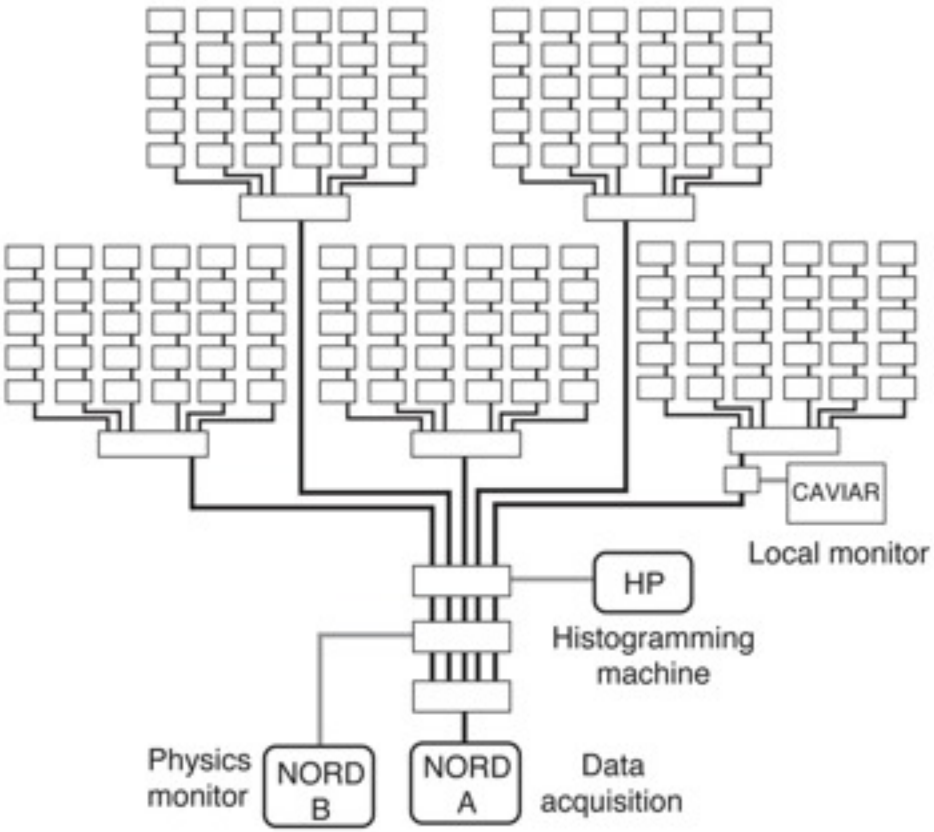
DAQ design issues at LHC (1990-2010)

- Physics and rates
- Experiments
- Collisions and detector front-end
- Event selection levels
- **DAQ readout network**

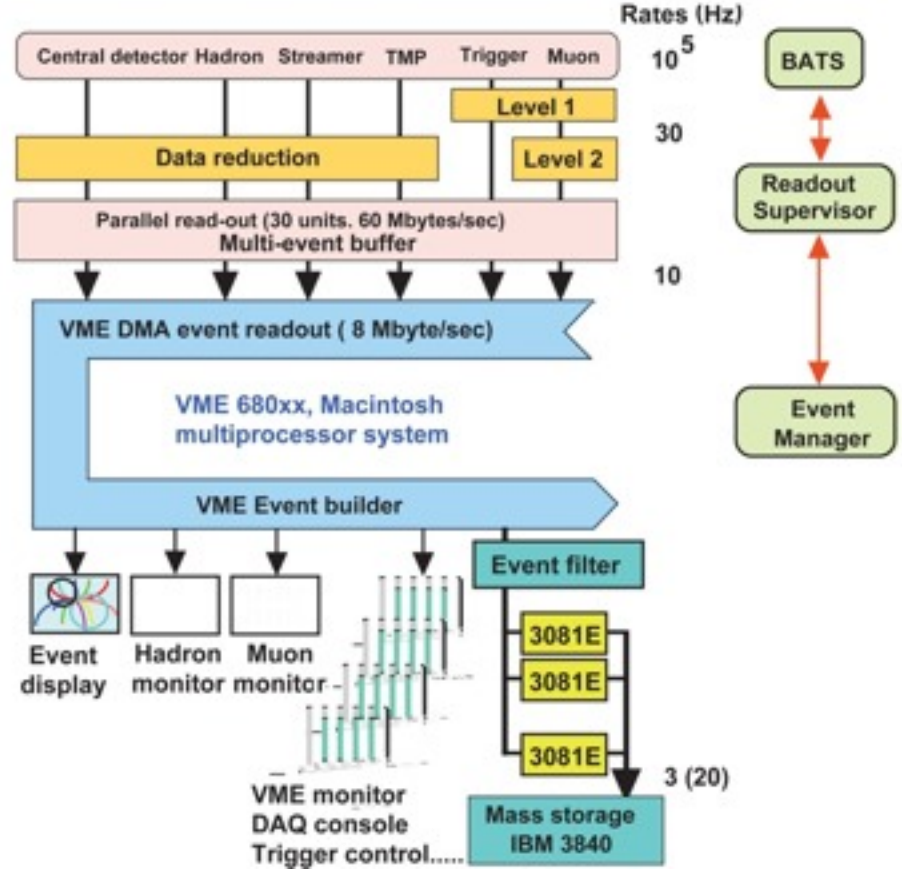
HEP experiments Level-1 rate / data volume trends



1978-1989. UA1 DAQ system



1981-84
- Remus data acquisition (≈200 CAMAC crates)
- rate on tape ≈ 1Hz (event size ≈100 Kbyte)

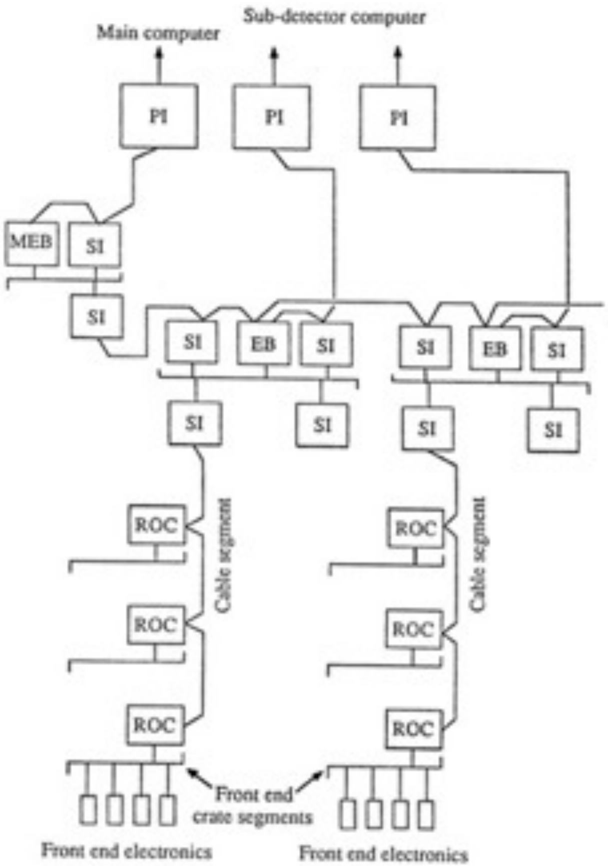


1985-1989
VME, IBM-emulators, Desktops

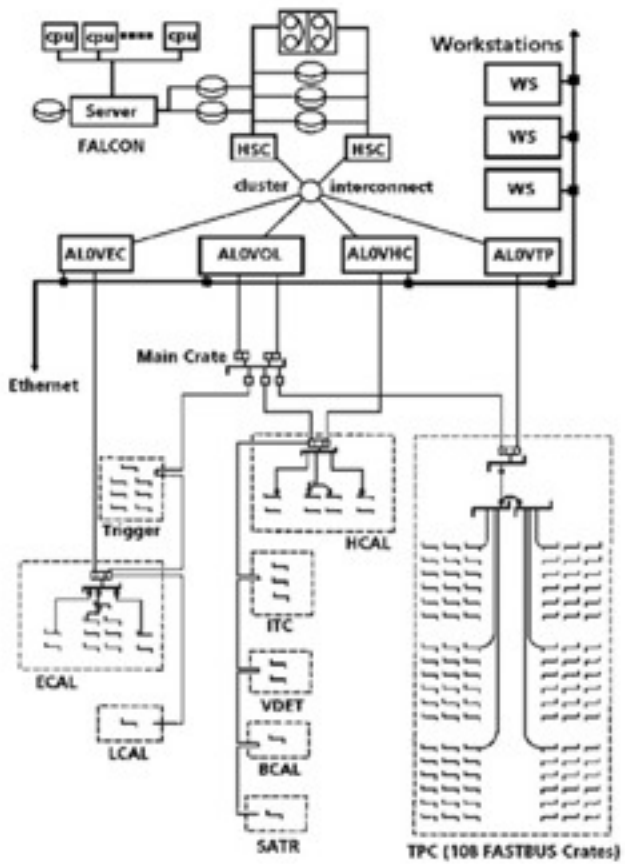
Proprietary/Standards: CAMAC, embedded μP, customCPU, VME

1989-2001 LEP DAQ systems

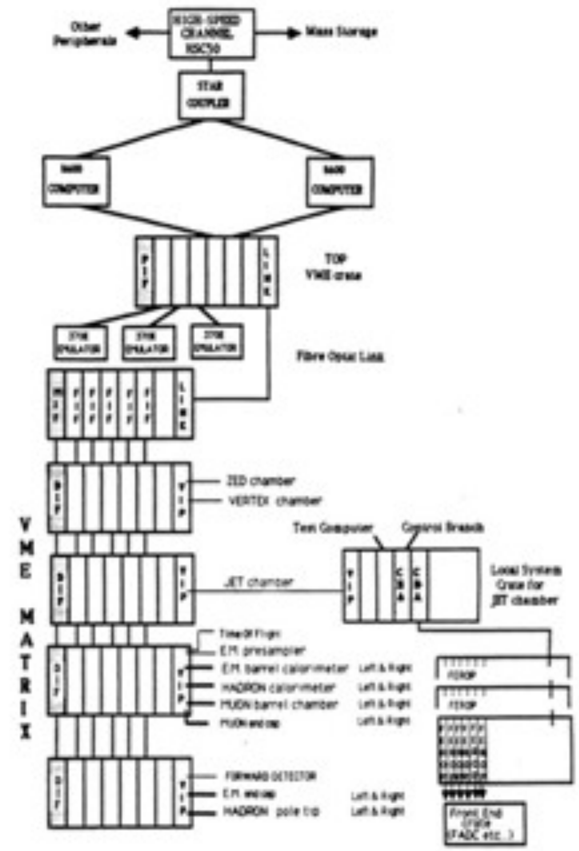
Aleph



Delphi

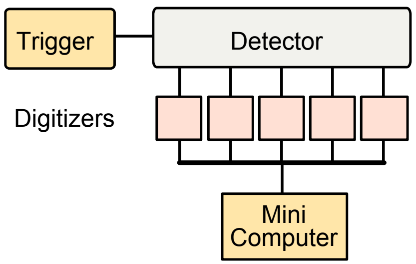


Opal



Proprietary/Standards: CAMAC, FASTbus, µp, VME, servers

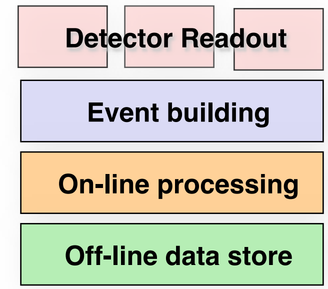
TDAQ architecture evolution

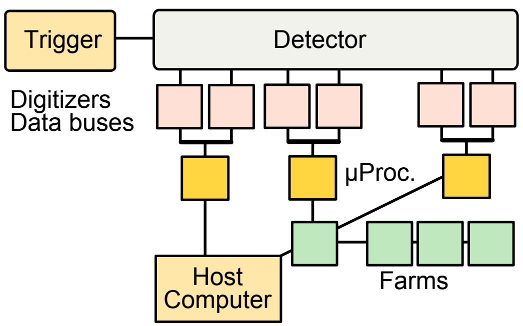


1970-80. PS/ISR/SPS: Minicomputers

Readout custom design
First standard: CAMAC
Software: noOS, Assembler

- **kByte/s, kFlop**

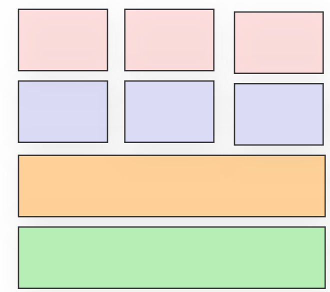


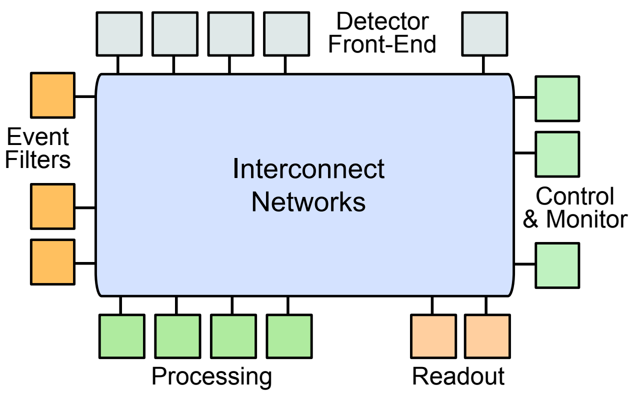


1980-90. p-p/LEP: Microprocessors

HEP proprietary (Fastbus), Industry standards (VME)
Embedded CPU, servers
Software: RTOS, Assembler, C, Fortran

- **MByte/s, MFlop**

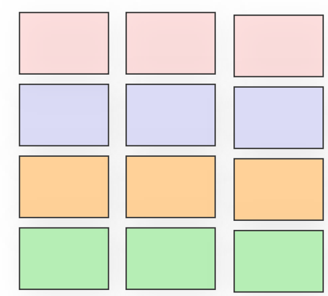




2000. LHC: Networks/Clusters/Grids

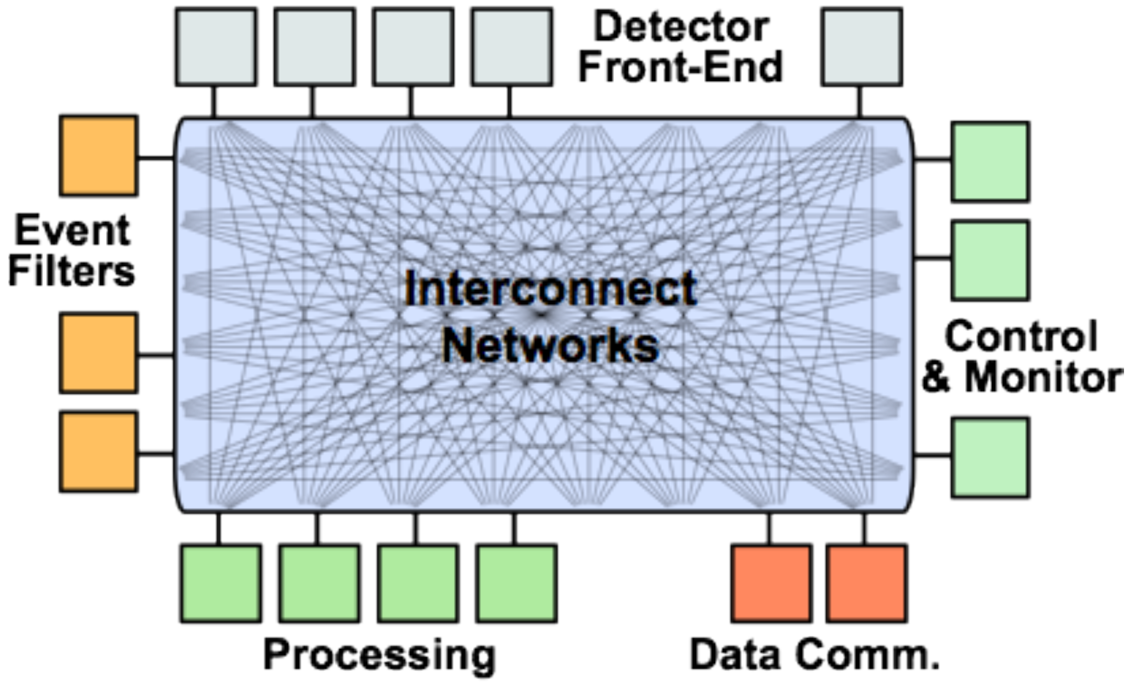
PC, PCI, Clusters, point to point switches
Software: Linux, C,C++,Java,Web services
Protocols: TCP/IP, I2O, SOAP,

- **TByte/s, TFlop**



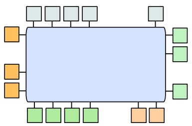
2000's On&Off-line processing and communication model

Consists of buffer memories, processors, communication links, data-flow supervisors, storage and data analysis units. Conceptually, the On/Off-line systems can be seen as a global **network interconnecting all** the data-flow, control and processing units



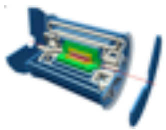
At the time of the finalization of the system design (2002-03), **a single network technology could not satisfy at once all the LHC requirements. The LHC DAQ designs had to adopt multiple specialized networks instead.**

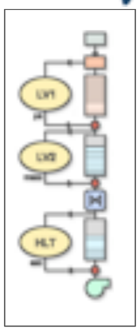
TDAQ data network designs

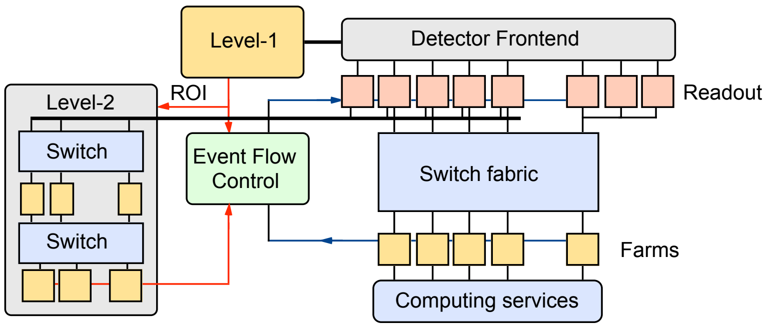


Each LHC experiment developed its own scheme to cut the rate, to process events online and/or optimize the throughput. In a sense, the **systems designed and built are "approximations" of the basic architecture/conceptual design**

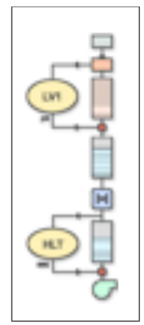
ATLAS

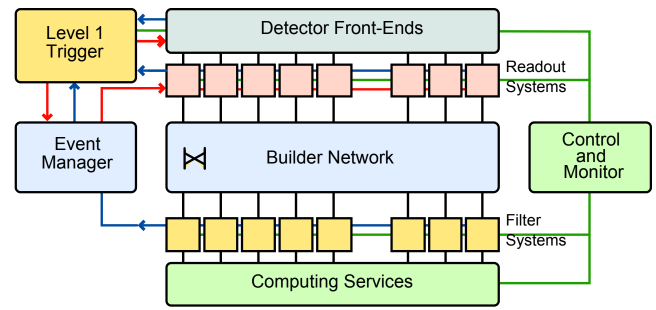




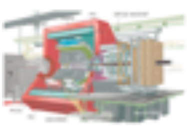


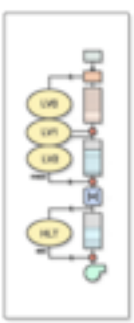
CMS

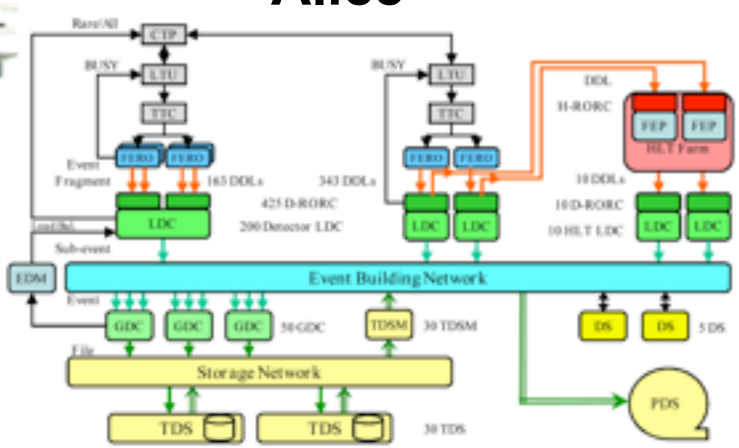




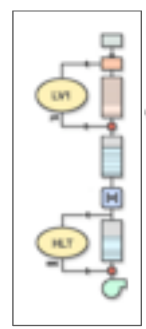
Alice

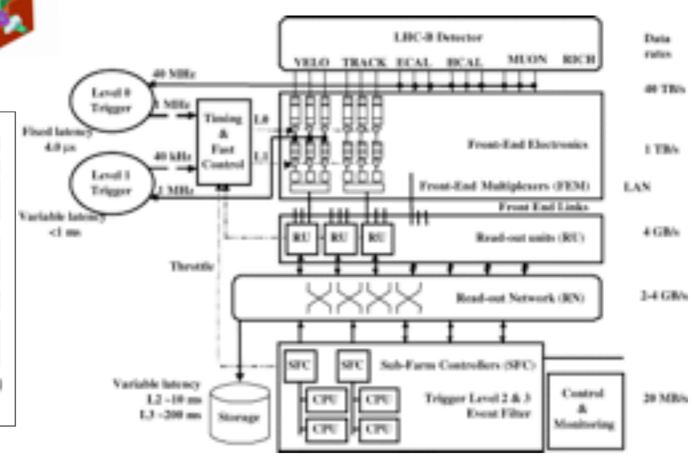




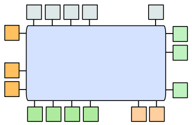


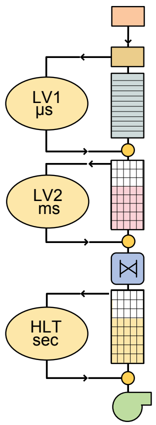
LHCb

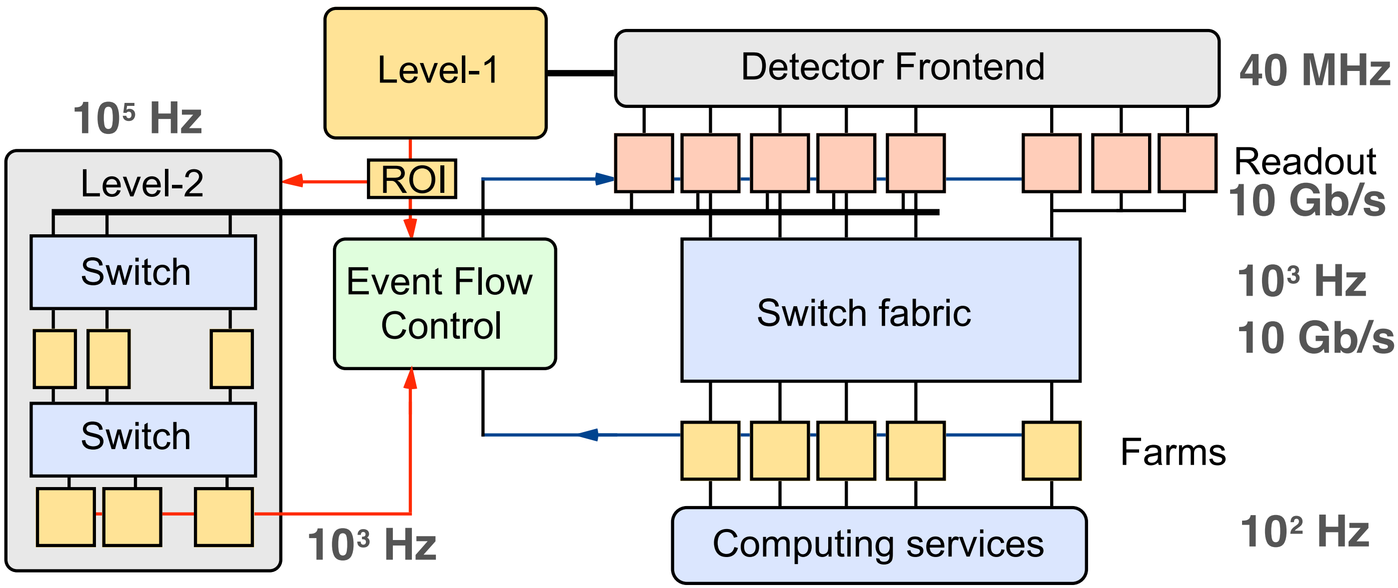




Three levels ATLAS TDAQ







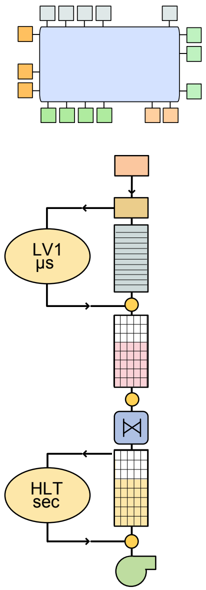
ATLAS LVL2 trigger refines the selection of candidate objects compared to LVL1, using full-granularity information from all detectors, including the inner tracker which is not used at LVL1. In this way, the rate can be reduced to ~1kHz. The data can be accessed selectively by the **LVL2 trigger which uses regions of interest (ROI) defined by the LVL1** trigger

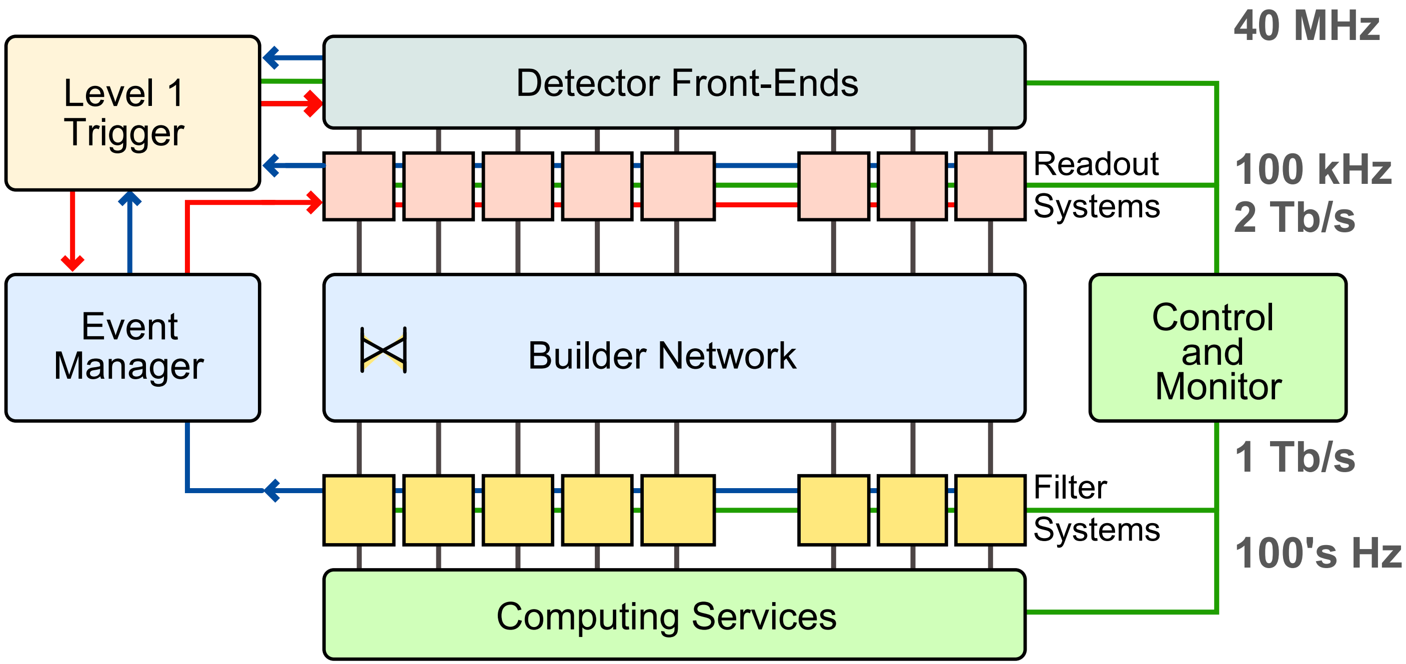
Collision rate	40 MHz
Level-1 Maximum trigger rate	**100 kHz**
Average event size	**≈ 1.5 Mbyte**
Flow control&monitor	≈ 10^6 Msg/s

Readout concentrators/links	**1500 x 1 Gb/s**
Event Builder bandwidth max.	**0.2 Tb/s**
Event filter computing power	**≈ 10-20 TeraFlop**
Event Builder GBE ports	**> 4000**
Data production	≈ Tbyte/day
Processing nodes	≈ x Thousands

Proprietary/Standards: Front-end, VME, PC servers, Networks, Protocols, OS

Two levels CMS TDAQ system

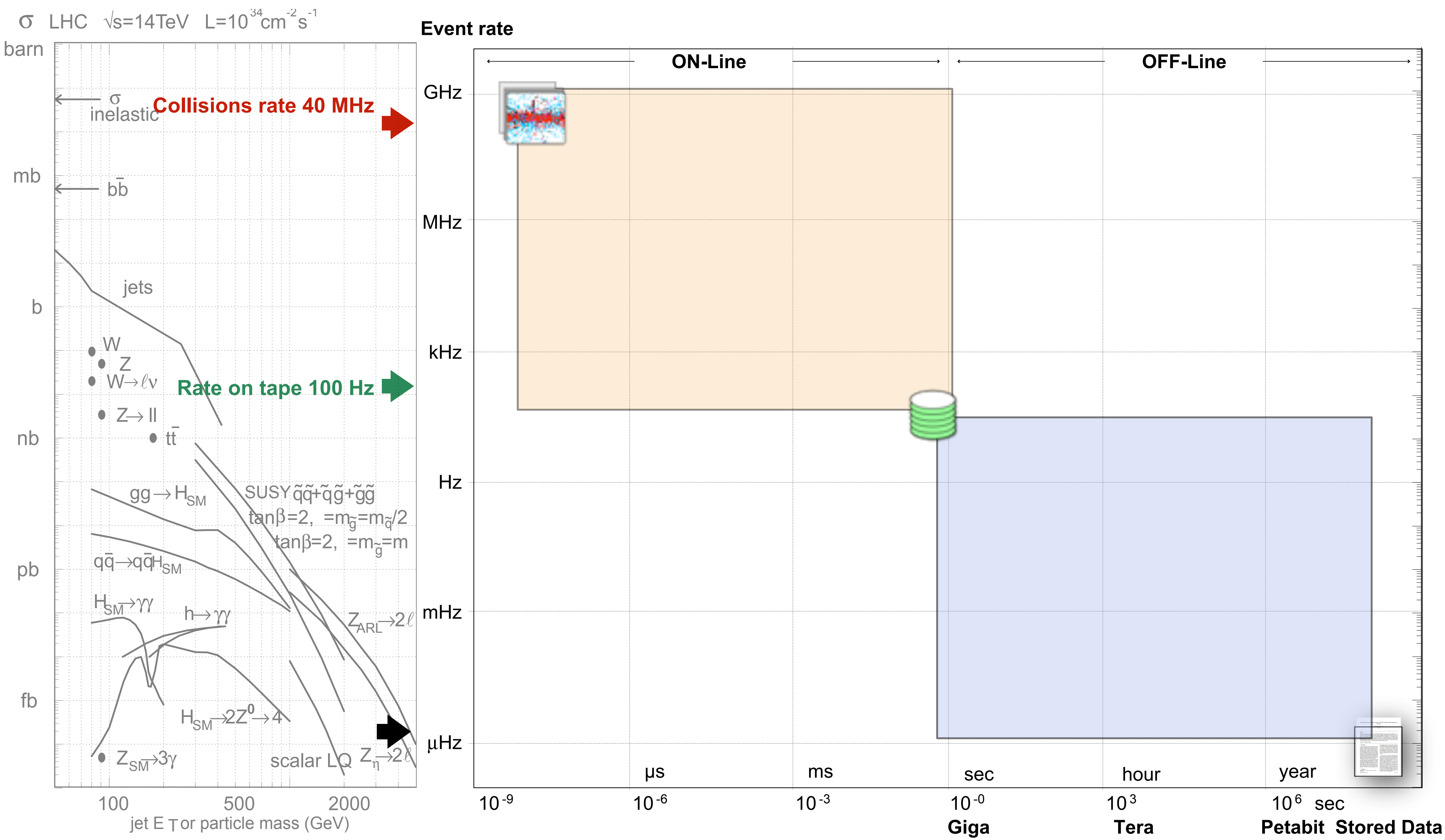




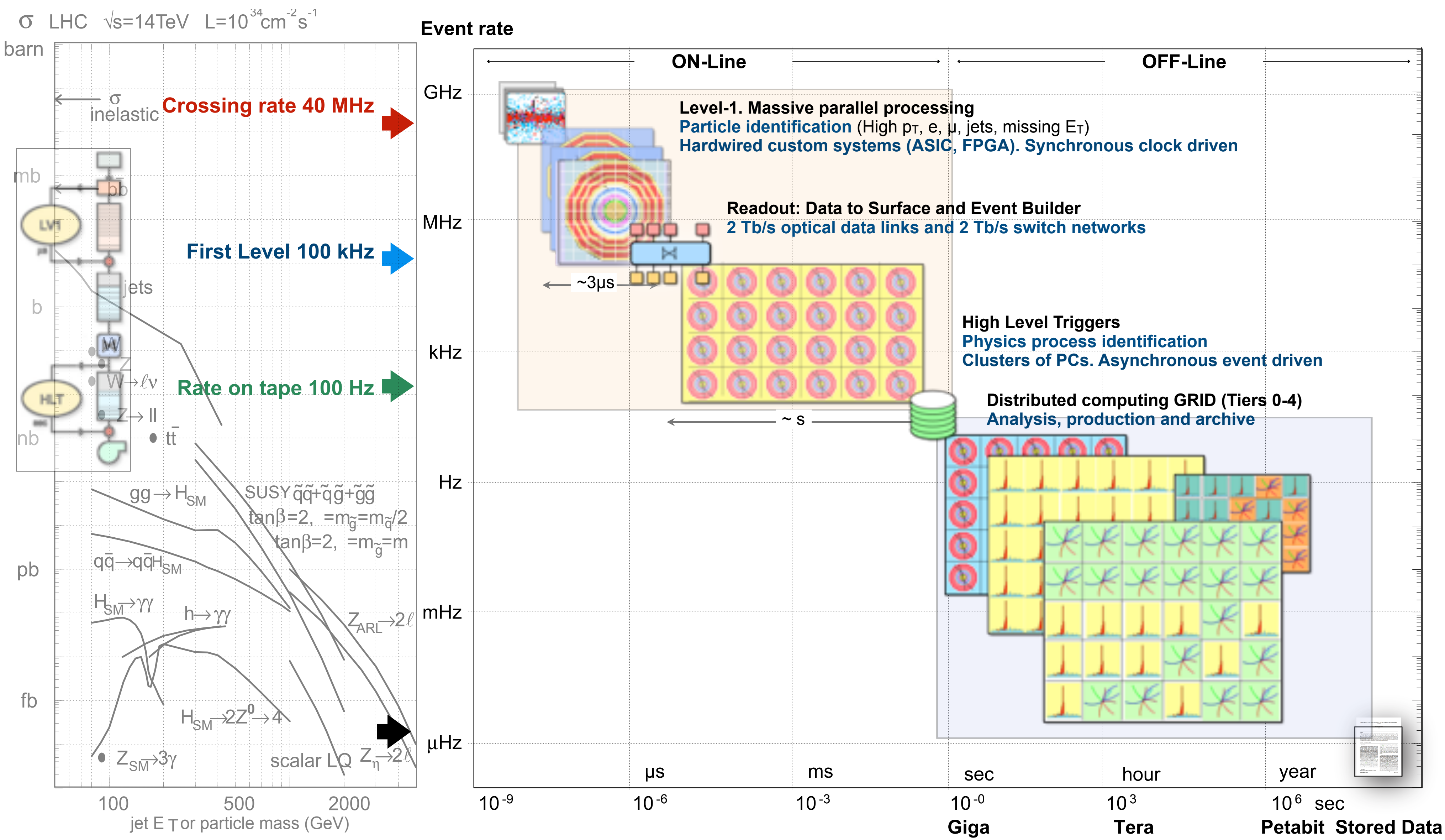
Collision rate	40 MHz	Readout concentrators/links	**512 x 4 Gb/s**
Level-1 Maximum trigger rate	**100 kHz**	**Event Builder bandwidth max.**	**2 Tb/s**
Average event size	**≈ 1 Mbyte**	**Event filter computing power**	**≈ 10-20 TeraFlop**
Flow control&monitor	≈ 10^6 Msg/s	**Event Builder GBE ports**	**> 4000**
		Data production	≈ Tbyte/day
		Processing nodes	≈ x Thousands

Proprietary/Standards: Front-end, VME, PC servers, Networks, Protocols, OS

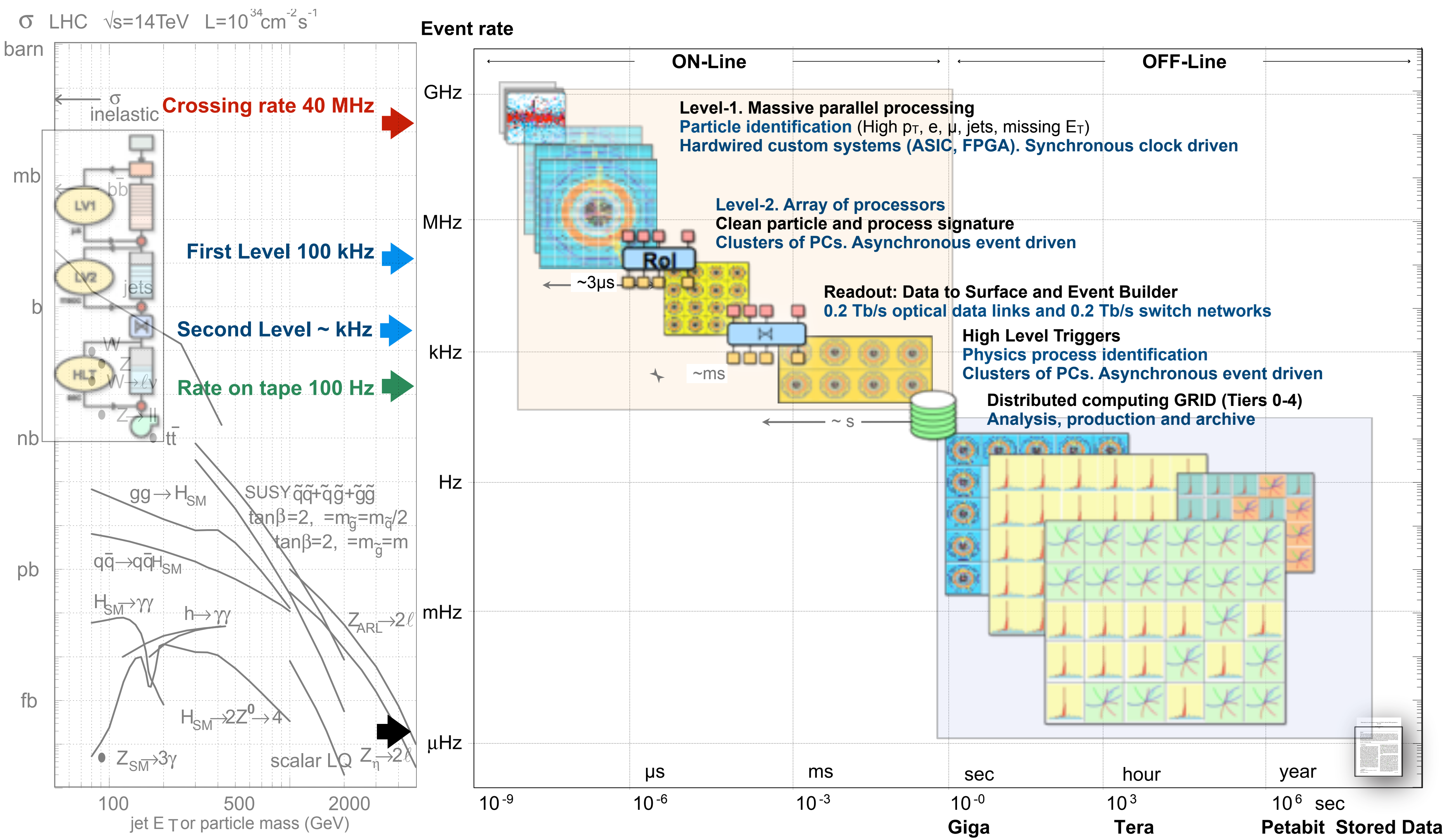
On-line rate decimation and data flow



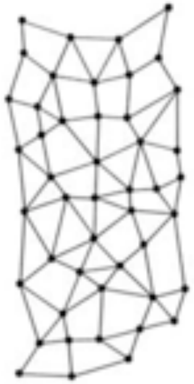
CMS: On-line trigger levels and event building



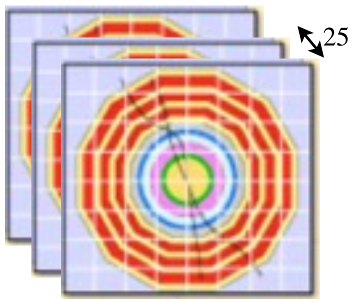
ATLAS: On-line trigger levels and event building

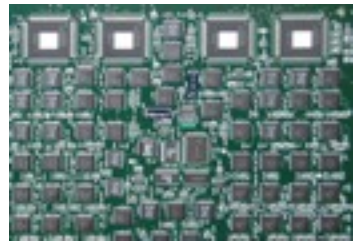


Processing architecture and network topologies

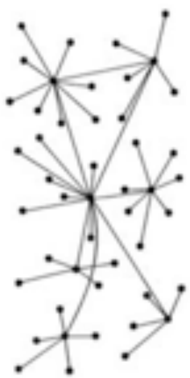


Level 1: Massive parallel pipelined processors. CLOCK DRIVEN
Implementation: Custom design (ASIC, FPGA)





Lv-1 processor, low Latency (μs)
synchronous 40 MHz, 128 cells depth
ONE event ALL processors
Pipeline memory for each channel
Radiation & power issues



Higher levels: Parallel processor clusters. EVENT DRIVEN
Implementation: Commodities (Servers, links, networks)

HLT PC farms, high latency (sec)
asynchronous scale-free expandable
ONE event, ONE processor
Data memory (PC) for each event

OFF-line. Data source: Centralised
Data analysis and storage distributed GRID
Implementation: Commodities (Servers, links, networks)

Design issues: Technologies

- Project history and information technologies trends
- Predicted and unpredicted evolutions

LHC&TDAQ project timeline (the time of a generation)

1984	Lausanne workshop (LEP/LHC)
1990	Design of experiments
1992	**CMS Letter of Intent**
1994	**Technical Design Report**
1996	LHC project approved
1998	
2000	**Trigger Technical Design Report**
2001	LHC cost review
2002	**DAQ Technical Design Report**
2006	Magnet test Global Run
2005	
2008	**Circulating beam Global Run**
2009	**Colliding beams**
2010	**Start physics runs**

1990-1995
Research and Development

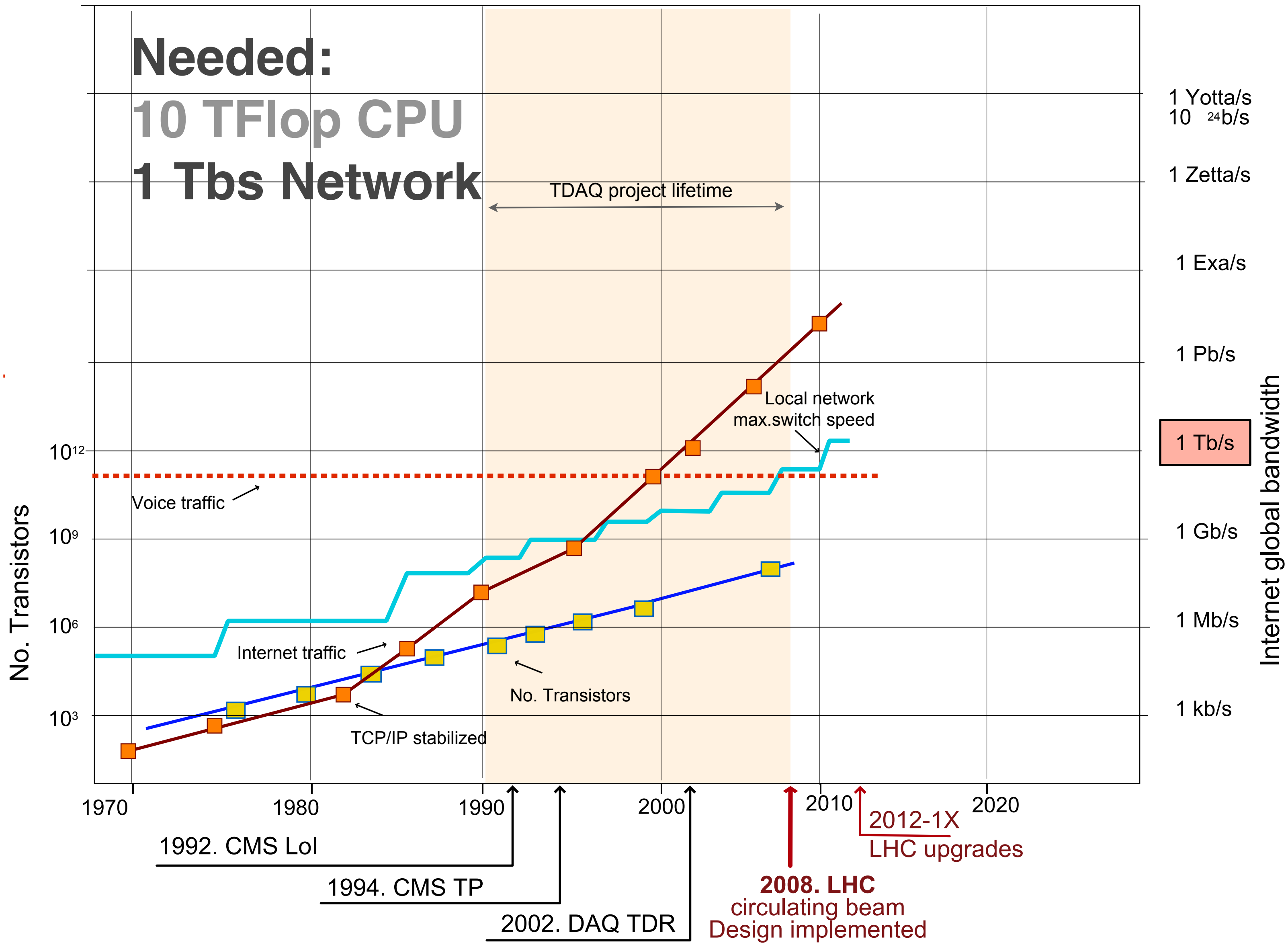
1996-2002
Prototypes and Demonstrators

2003-2005
Final Design. Choice of technologies

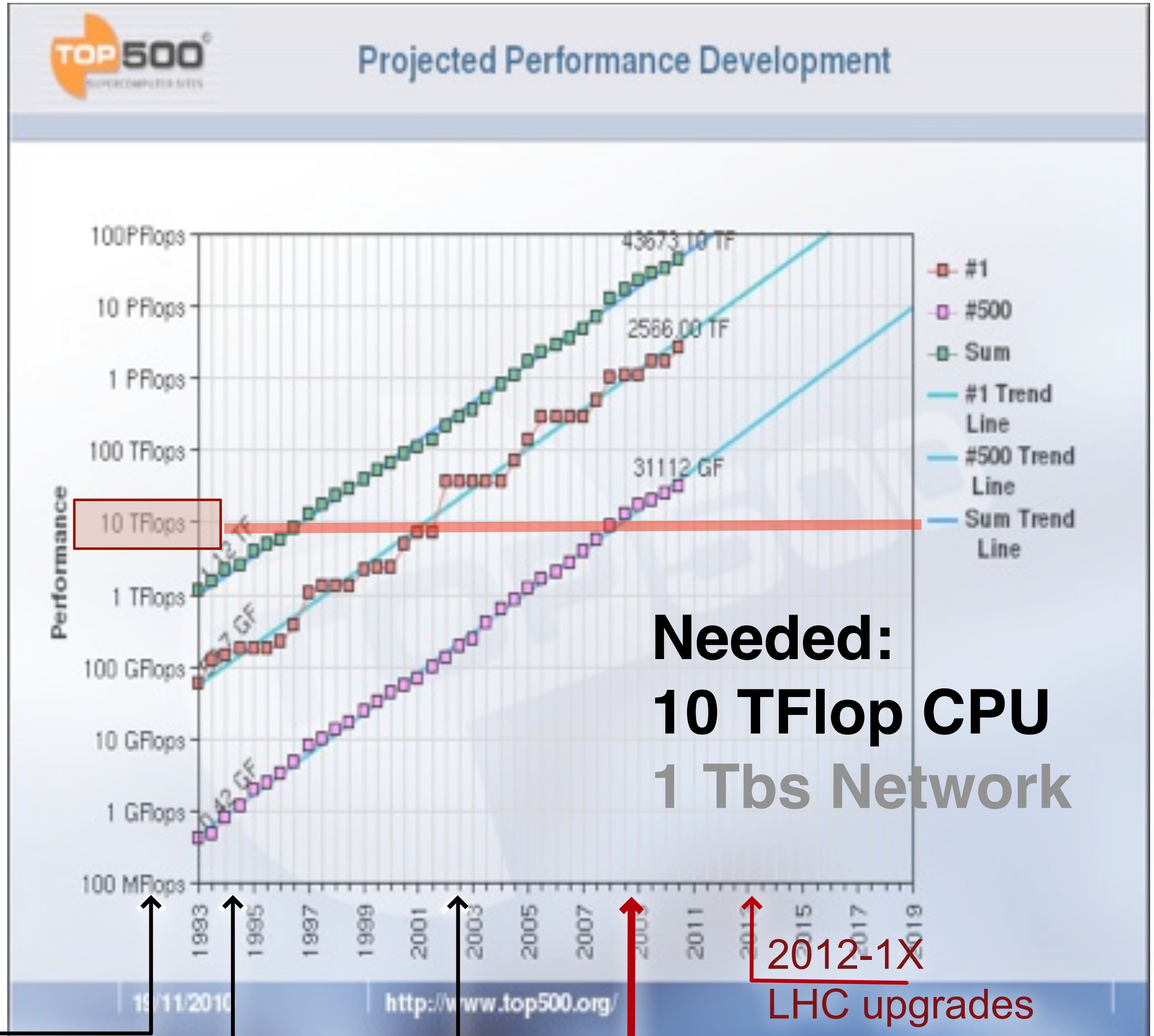
2006-2008
Construction and commissioning

Predicted
- Computing power
- Network bandwidth

Data communication. Network and Internet traffic trends



Computing power trends



1992. CMS LoI

1994. CMS TP

2002. DAQ TDR

2008. LHC
circulating beam

Technical challenges II. Affordable CPU power

1996. According to Linux Magazine, Digital Domain, a production studio located in Venice, California. produced a large number of visual effects for the film Titanic. During the work on Titanic the facility had approximately **350 SGI CPUs, 200 DEC Alpha CPUs and 5 Tbytes of disk** all connected by a 100 Mbit/s network.

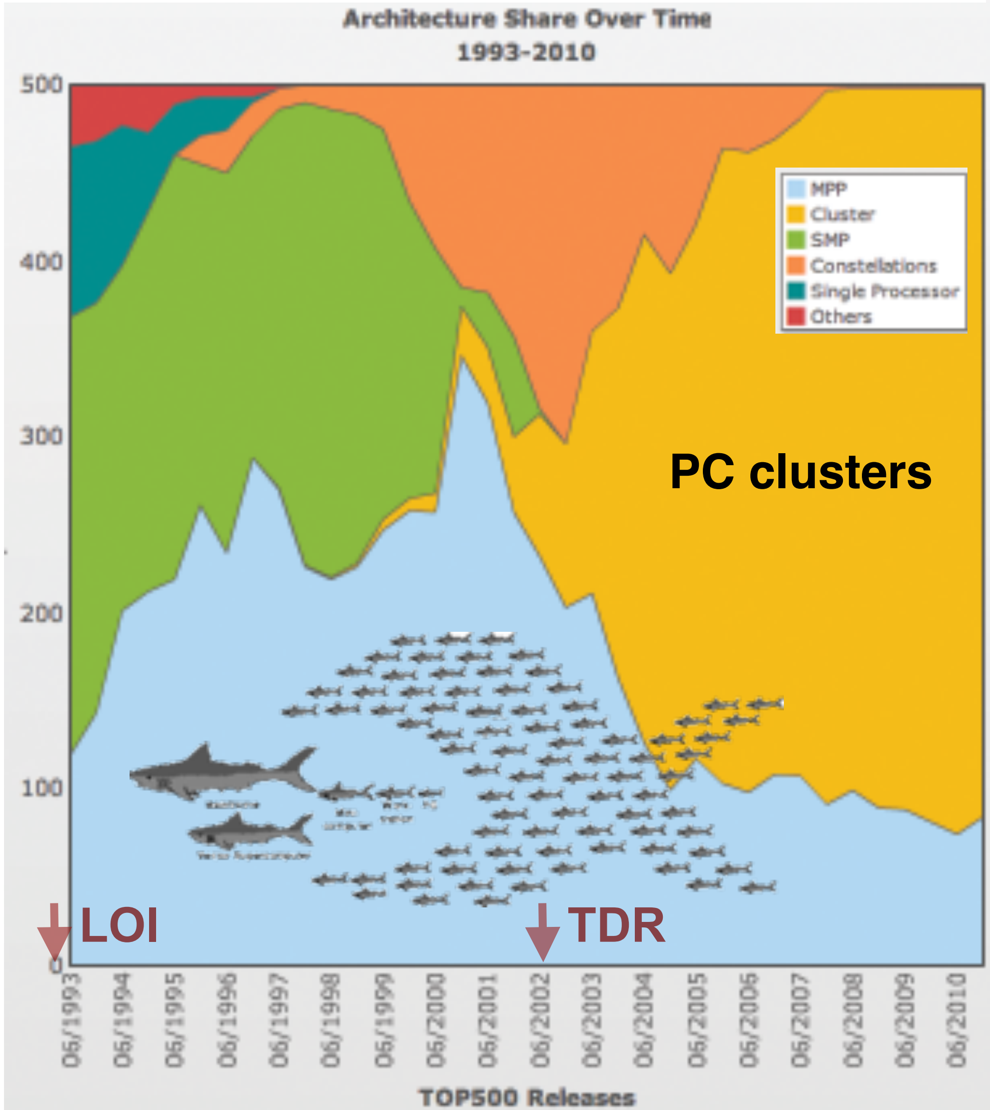
Since 90's:

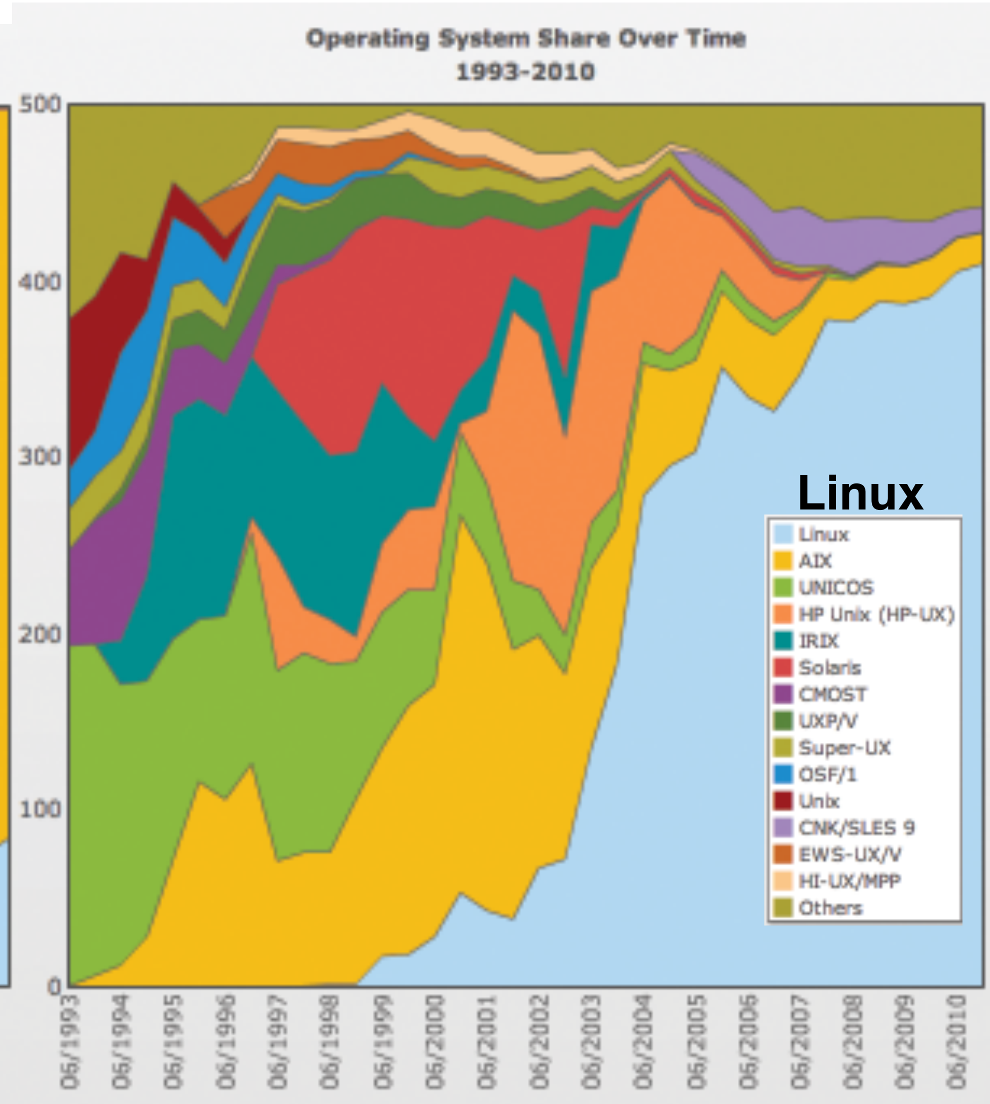
- Large computing power at low cost is made available as **clusters of commodities** (PCs and networks)
- **LINUX** has become the most popular Operating System

CPU estimated in 2002. Total: 4092 s for 15.1 kHz → 271 ms/event. Therefore, a 100 kHz system required about 13 TFLOPs (corresponding to ~30000 CPUs of 2002)

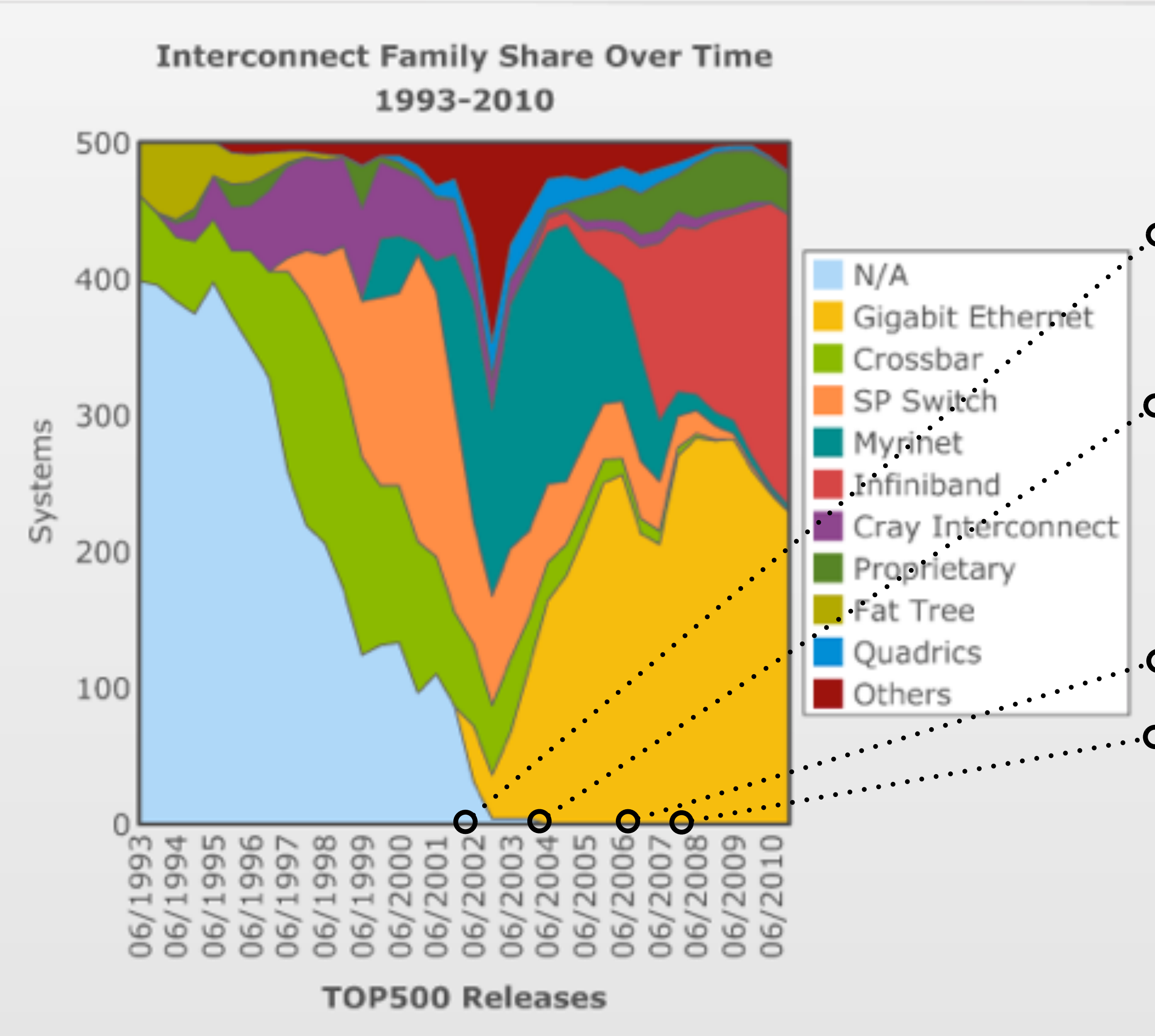
CPU implemented in 2008. The 50% of the HLT system integrated in 2008 consisting of 5000 2.6 GHz CPUs (720 PCs of two quad-core) corresponds to about **10 TFLOPs** in line with the foreseen requirements and in agreement with the Moore law of integrated logic systems (corresponding to a factor 10 in speed every 6 years)

Top500 supercomputer architectures and operating systems





Top500 interconnection technologies and TDAQ decisions



Decision schedule

2002 Data to surface:

- Myrinet used as first layer of readout (FED builder and Data link to surface)

2004 Event builder:

- Gigabit Ethernet routers used for Event builders and DAQ services (controls, mass storage, data link to central Tier0)

2006 Procurements

2007 Construction

2008 Commissioning

Unpredicted

- Collaborative work
- Network&Computing fusion

Experiment control rooms

ISR. 1970
CR info tools:
Coaxial Cables
Teletype
Telephone

P-aP. 1980
CR info tools:
RS 232
Alpha terminal
Video&Telephone

LEP. 1990
CR info tools:
RS 232, Ethernet
Graphics terminals
Video&Telephone

LHC 2010
CR info tools:
Wireless
LAN, WAN
Internet, WWW

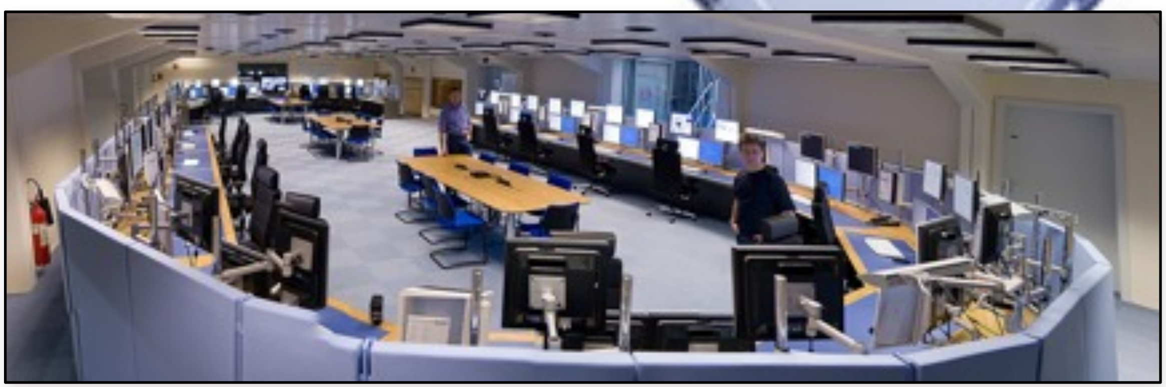
2010 LHC. The person is on the screen

Experiment control and monitor system and WWW services

Cessy: Master&Command control room

Fermilab: Remote Operations Centre

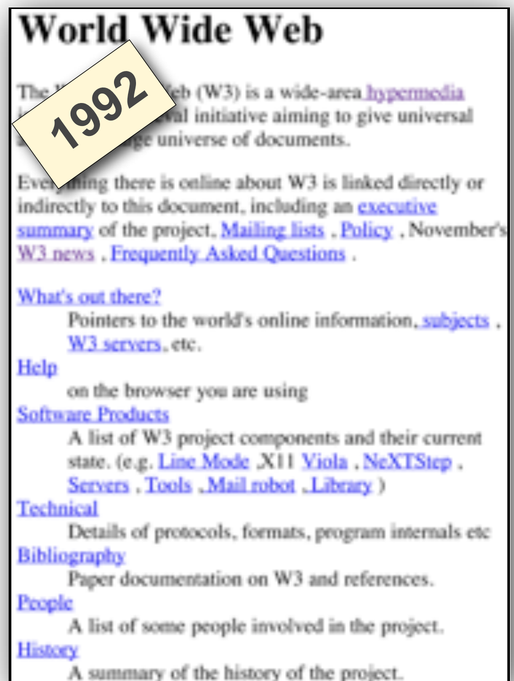
Meyrin: CMS DQM Centre



CR: Any Internet access.....

A general and expandable architecture has been deployed for the **experiments' Run control and monitoring** largely based on the emerging Internet technology developed in the field of **WWW services**

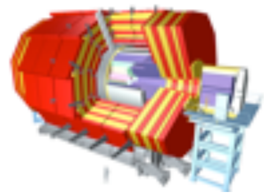
Hard-to-predict in the 90's: The World Wide Web



Since the start of the exploitation of large accelerator laboratories around the world, the design and operation of High Energy Physics experiments have required an ever increasing number of participating institutions and collaborators. From tens of institutions and hundreds participants during the Collider and LEP period up to **hundreds of institutions and thousands scientists** in today LHC experiments.
At the end of 80's with the digitalization of information and the growing support of information infrastructures (computer centers and Internet), a tool was needed to improve the collaboration between physicists and other researchers in the high energy physics community.

The **World Wide Web** originally was intended for this purpose, however fusing together networking, document/information management and interface design it has become in few years the most popular instrument to provide seamless access to any kind of information that is stored in many millions of different geographical locations. In addition, it stimulated the expansion of network infrastructures and the development of new software and hardware services based on common standards (TCP/IP, HTML, , SOAP, XML,.... GRID, CLOUD,...)

Hard-to-predict in the 90's (II): the same model elsewhere



2008 The CMS HLT center on CESSY
and hundreds Off-line GRID computing centres 10^5 cores

2008 One of Google data center 10^6 cores

Conclusion

Computing power evolved as expected -- if not faster -- as needed by experiments at the LHC.

Digital information technology as well as the Internet have generated the drive for the development of higher bandwidth networks, along with the expansion of world-wide infrastructures to interconnect computing and data routing centers.

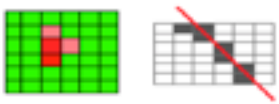
The computing and communications challenges that were posed by experiments in high energy physics (HEP) have not only presented themselves as high-end applications of the most advanced technologies. They have also been a source of inspiration for the development of new ones.

Even more importantly, HEP promises to maintain its dual role of client/motivator.

On-line data communication and processing at LHC

Crossing rate: 40 MHz
Collision rate: 1 GHz

Raw data production
100's of Petabits/s



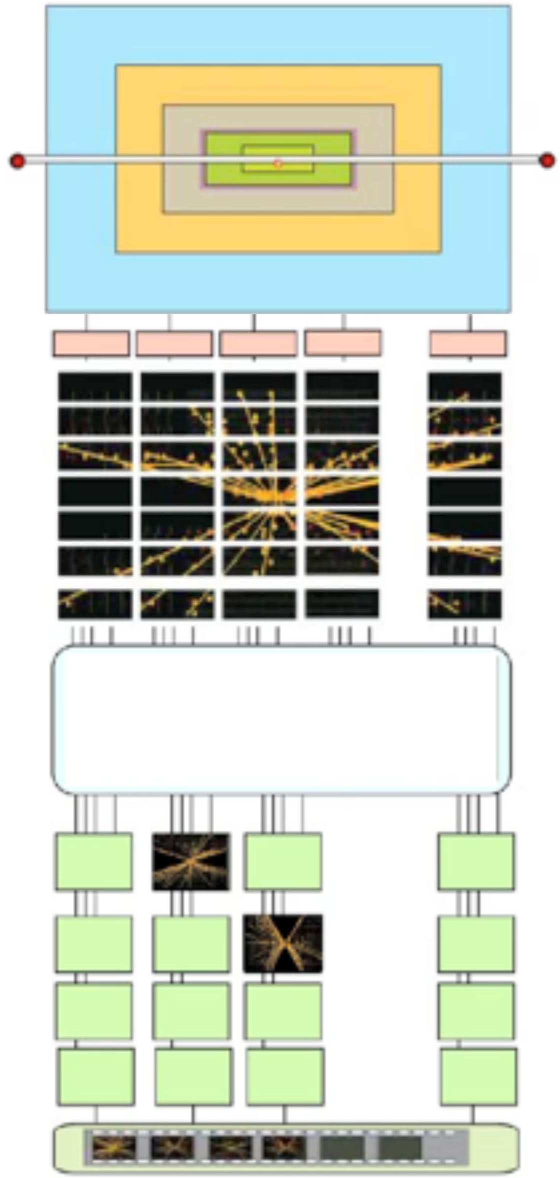
Level-1 Trigger

100 kHz output rate
50 Million fragments/s

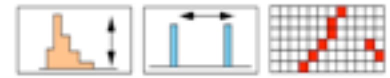
Readout network
2 Terabit/s

Build, process and select
10 TeraFlops
100000 event/s

Store O(100 Hz)
Tens of Gigabits/s



100 Millions instrumented sensors



Billions of (A/D) memory cells

Parallel readout

Hundreds of event fragment readers

Data to Surface
Thousands of Optical links

Event Builder
Switching system with thousands of ports
~1 Megabyte data per event

Event Filter
Thousands of CPU cores

Local mass storage
Hundreds of Terabits