Overview of the history of LHC DAQ systems

DAQ@LHC workshop, 12 March 2013
S. Cittolin (University of California San Diego)

- **Design issues: Architectures**
  - Physics, rates and requirements
  - Front-end, event selection levels
  - Readout networks
- **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

1990-1995
Research and
Development

1996 LHC project approved
1998
2000 Trigger Technical Design Report
2001 LHC cost review
2002 DAQ Technical Design Report

1996-2002
Prototypes and Demonstrators

2003-2005
Final Design. Choice of technologies

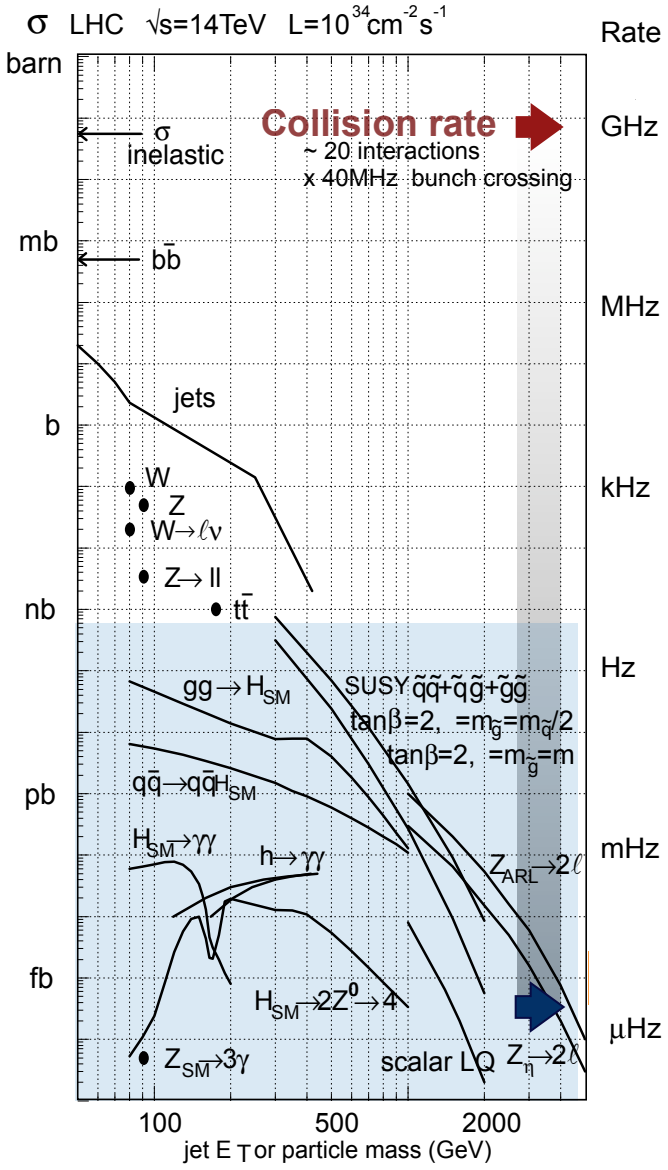
2006 Magnet test Global Run
2005
2008 Circulating beam Global Run
2009 Colliding beams
2010 Start physics runs

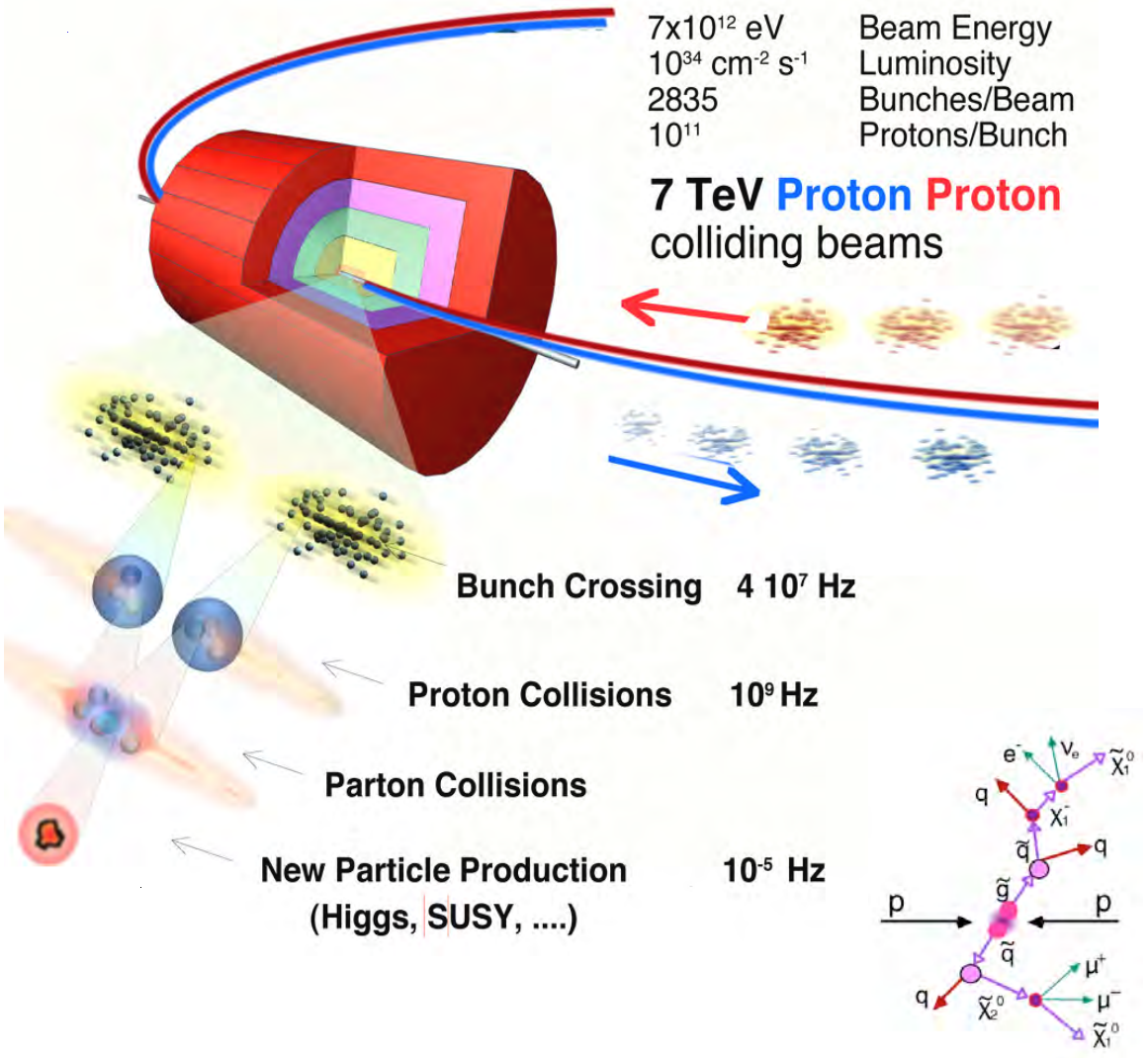
2006-2008
Construction
and commissioning

DAQ design issues at LHC (1990-2010)

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

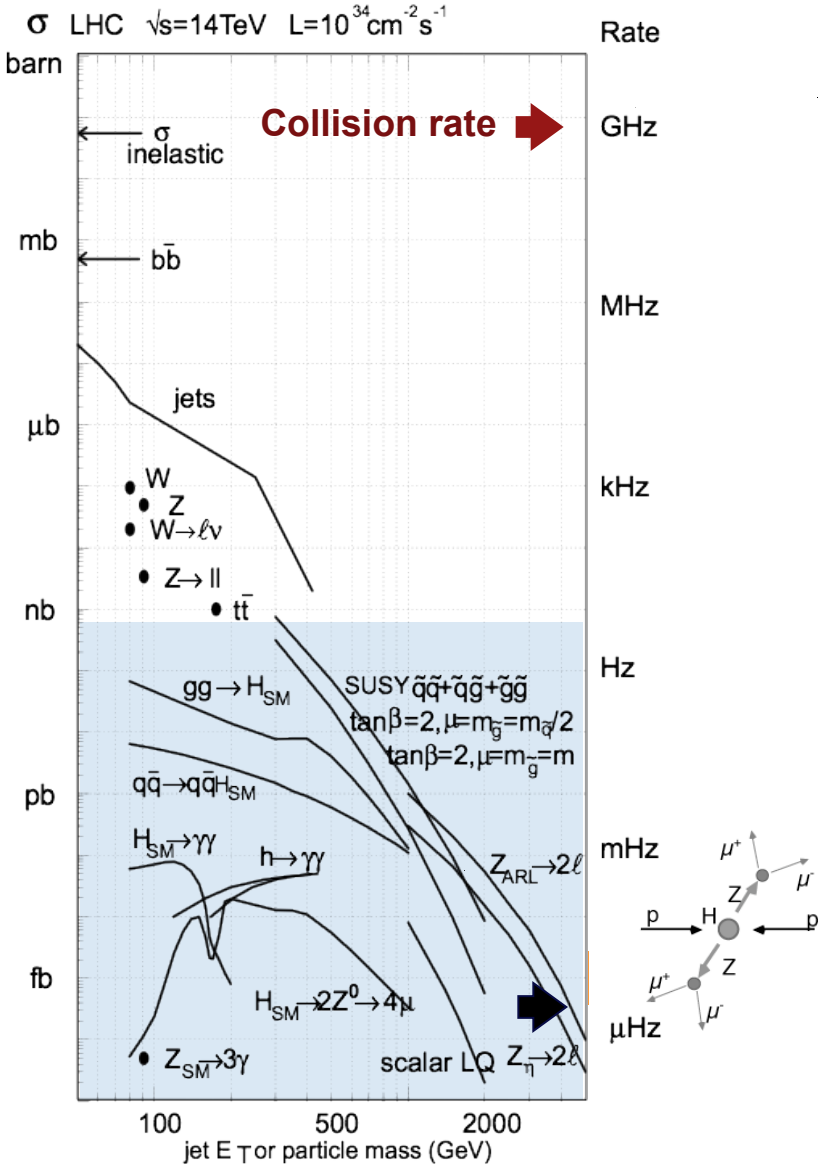
Proton-proton collisions at LHC. Searching issue



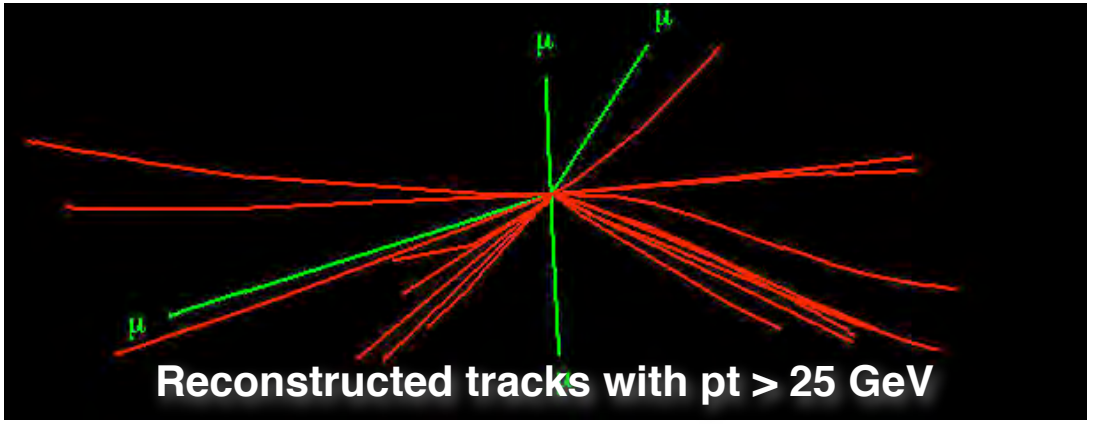


Collision Rate: ~10^9 Hz. Event Selection: ~1/10^{13}

Data detection, event identification and event analysis

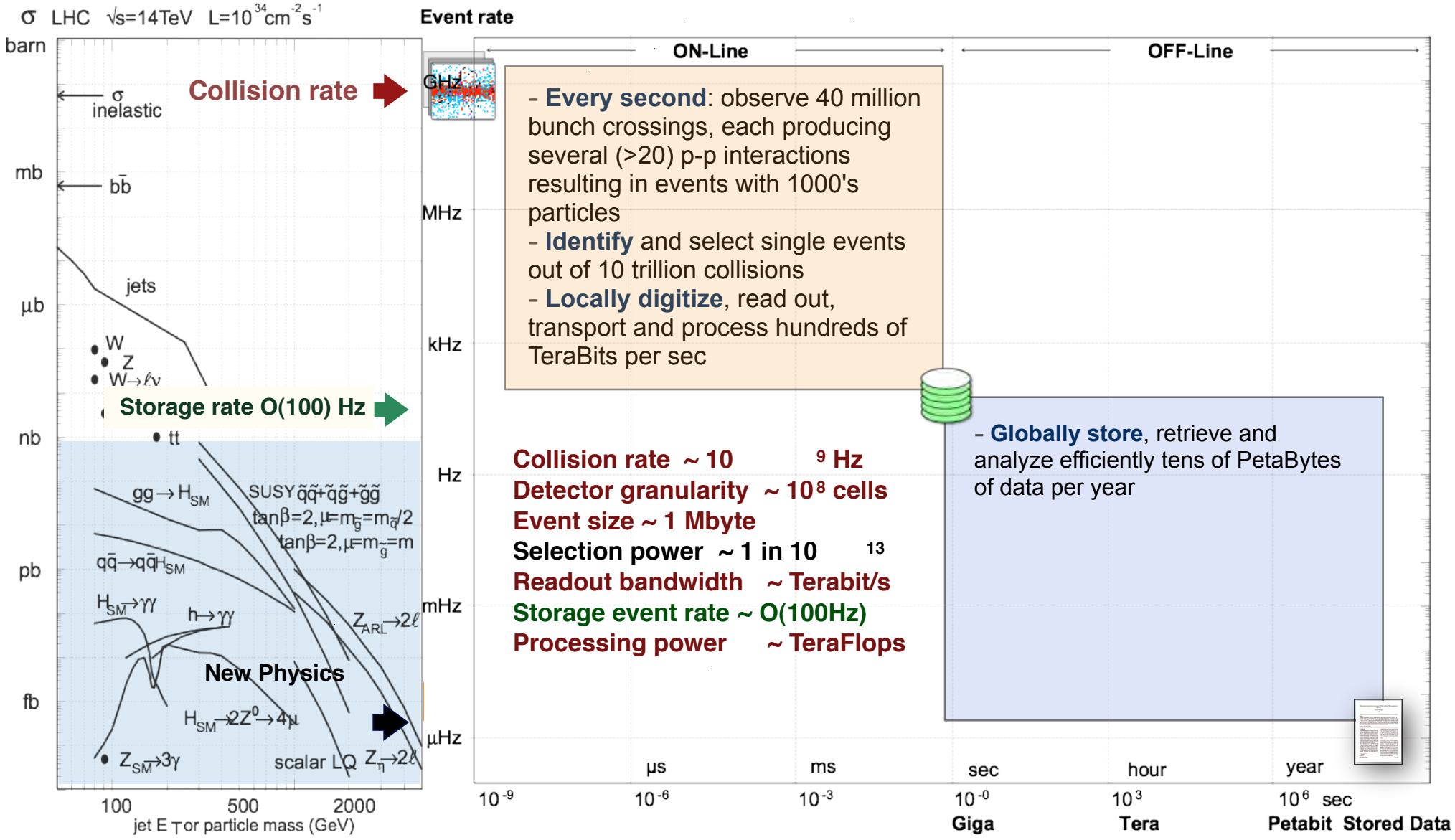


Detector granularity ~ 10^8 cells
Event size: ~ 1 Mbyte
Store and analyse data 10's of PetaBytes/year



Data processing power: tens of TFlops

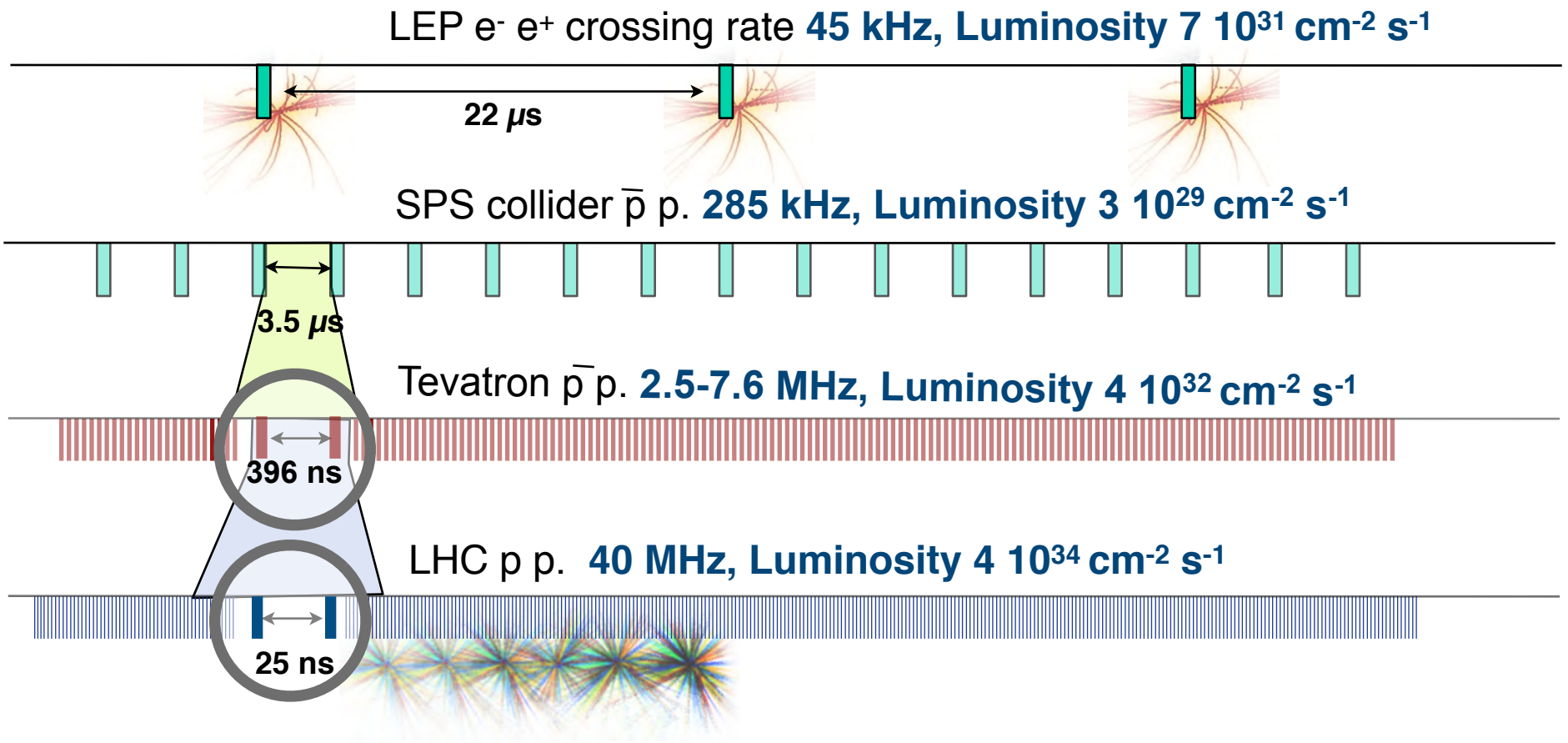
Physics at LHC: overall data handling requirements



DAQ design issues at LHC (1990-2010)

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

Colliders bunch crossing frequencies

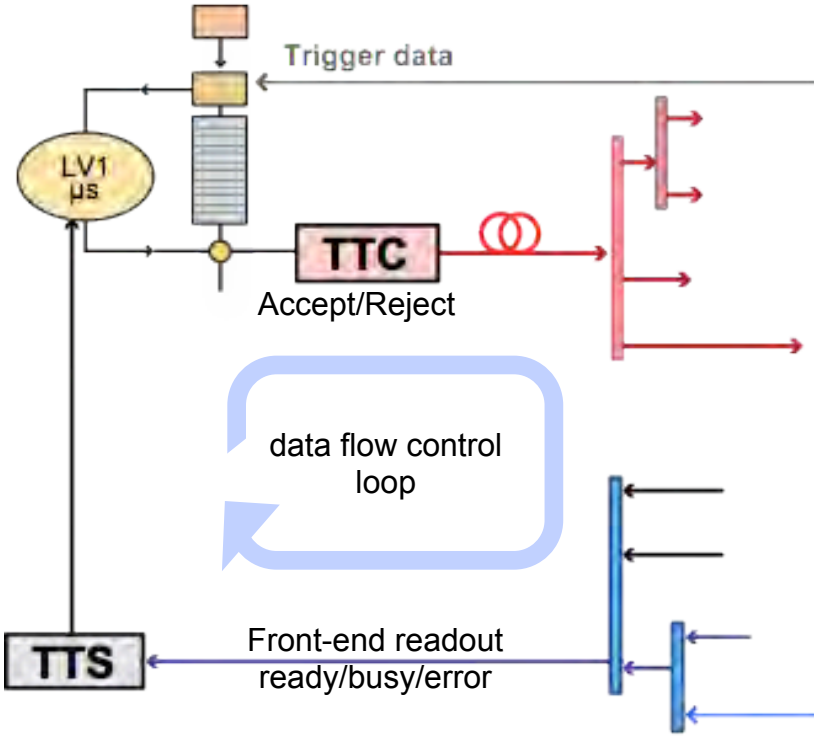


- **25 ns** defines an overall time constant for signal integration, DAQ and trigger.
- The rate of the collisions (**40 MHz**) is (was) not affordable by any data taking system.
- The off-line computing budget and storage capacity limit the **output rate** (~**100 Hz**)

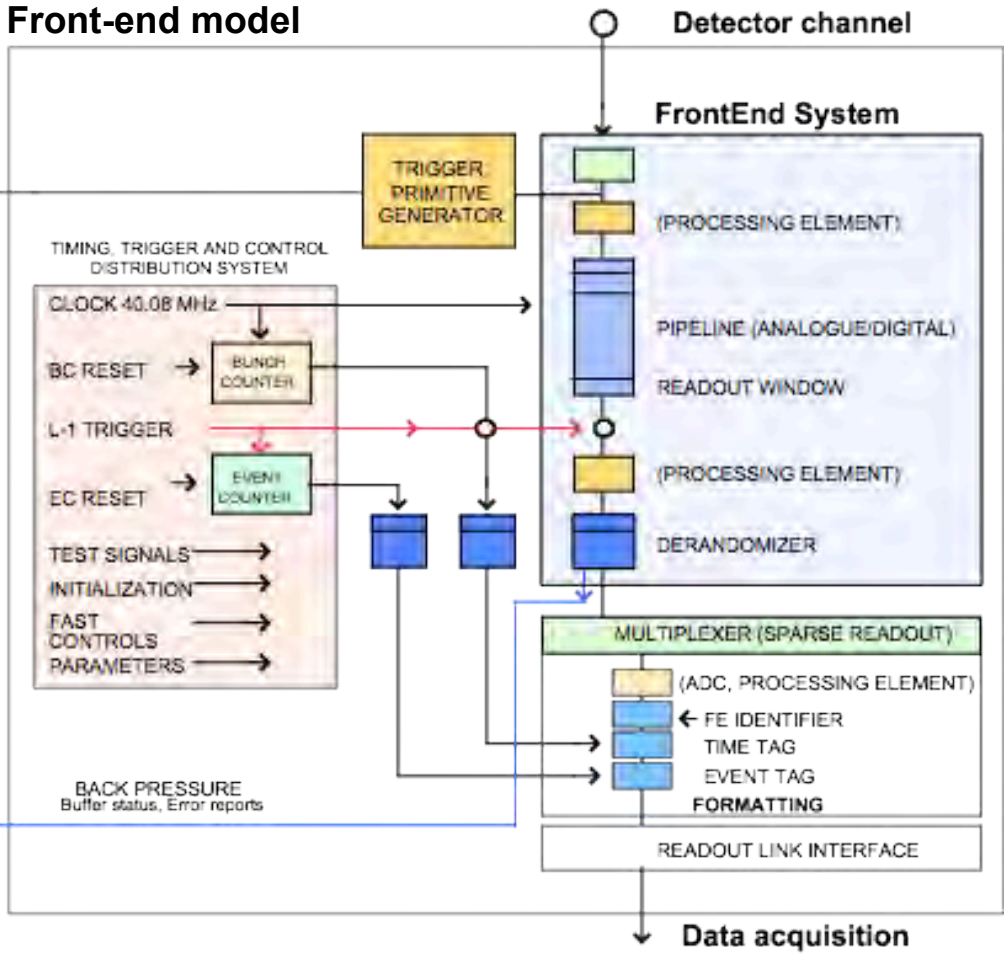
Front-end generic model and timing system

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

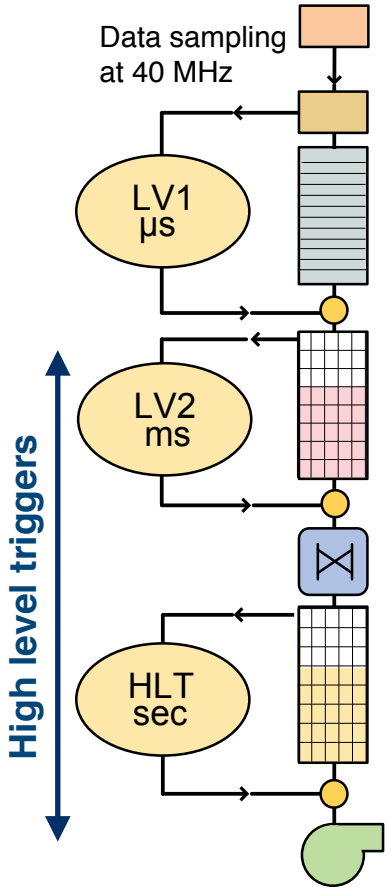


TTS. Trigger, Throttle System



On-line multi-levels trigger generic architecture

Successively more complex decisions are made on successively lower data rates

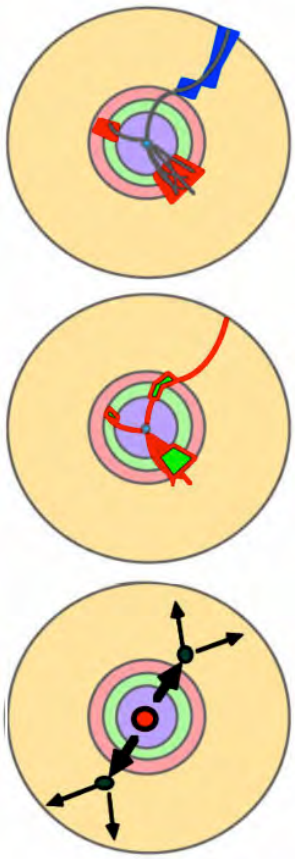


10^{-7} s

10^{-6}

10^{-3}

10^{0}



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

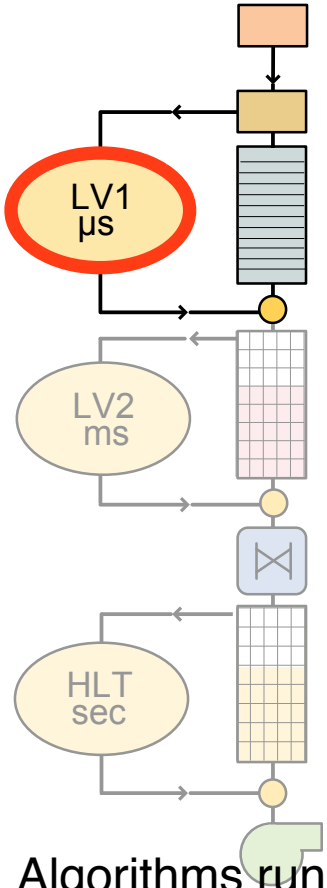
On-line global selection:

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**

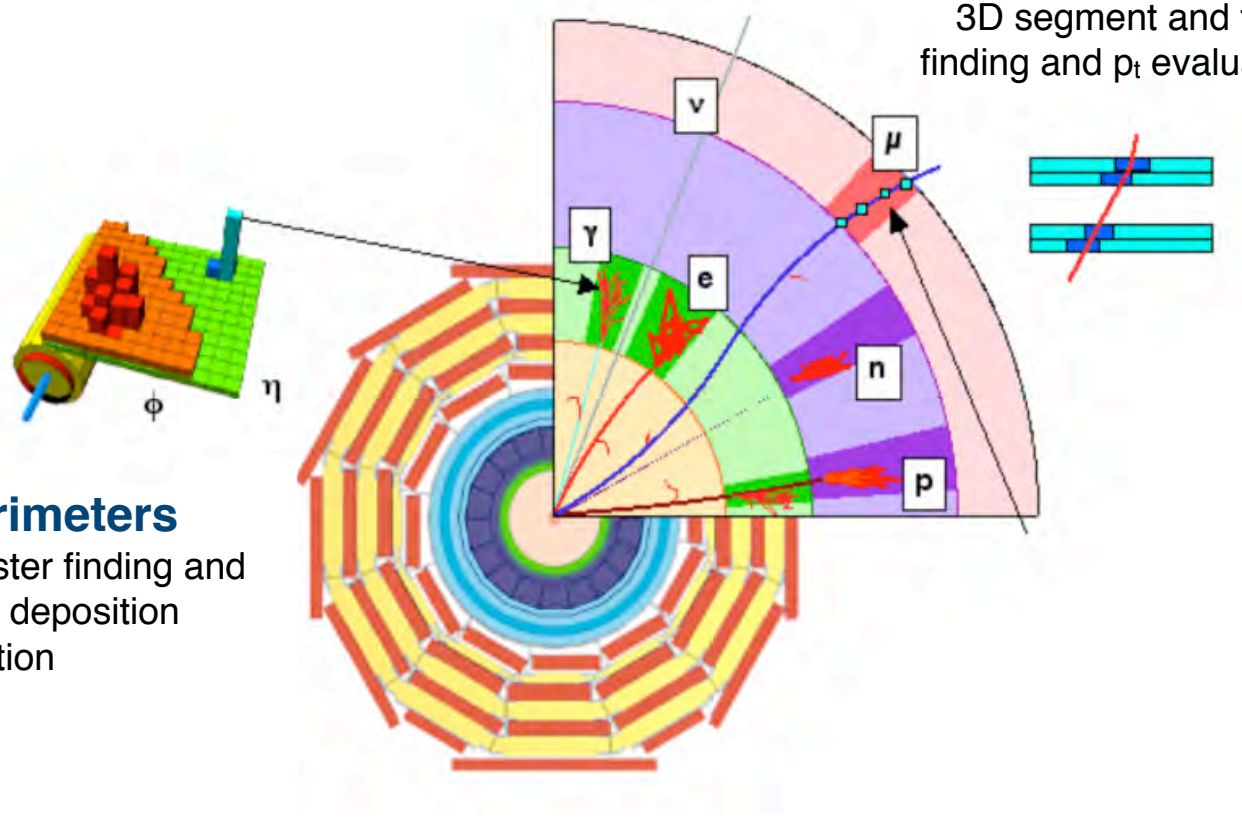
Level-1 trigger. Particle identification

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



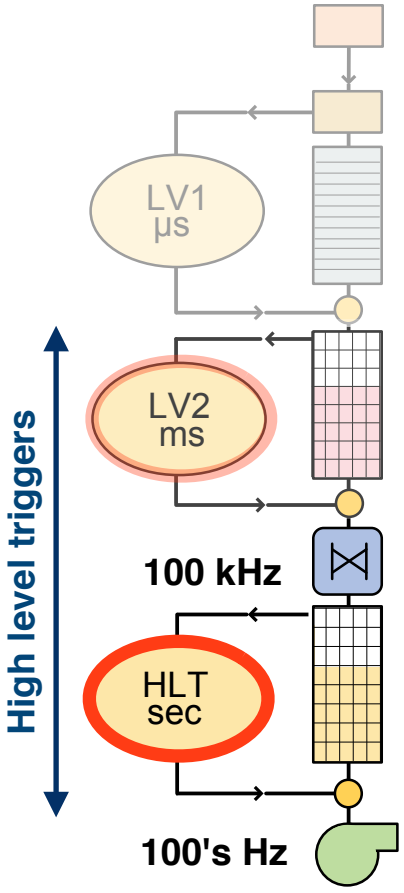
Calorimeters

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.**

High Level Triggers (HLT) requirements and operation

HLT algorithms have **the full event data** available and **no limitation on complexity**.
(In CMS a single physical step (**HLT**) after L1 is used to achieve a rejection factor of ~1000)



Main requirements:

- Input after level-1 at **maximum event rate of 100 kHz**
- Selection must be inclusive based on the presence on one or more objects above p_T/E_T thresholds
- All algorithms/processors run off-line code
  L2: muon+ calorimeter only.
  L3: use full information including tracking
- Run on **farm of commercial CPUs**
- Code **runs in a single processor,** which analyzes one event at a time
- HLT has access to **full event data** (full granularity and resolution)

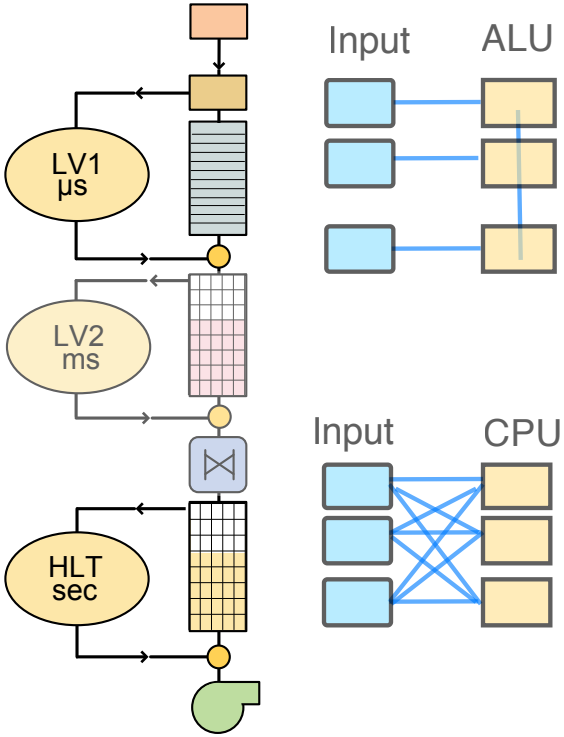
Only limitations:

- **CPU time (TeraFlops needed)**
- **Output selection rate (~10^2 .. 10^3 Hz)**
- Precision of calibration constants

Triggering: architectures and technologies

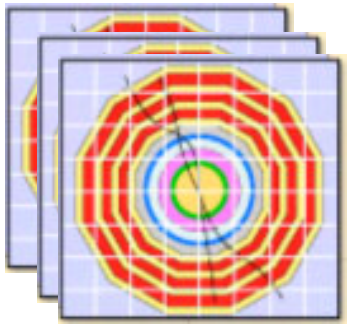
Level-1 trigger architecture. **Massively parallel: One event -> Multi-processors**

- High (fixed interconnections), Short (fixed) latency. Pipelined simple ALUs

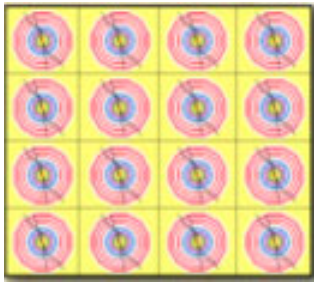


DATA PROCESSING

Synchronous pipelined

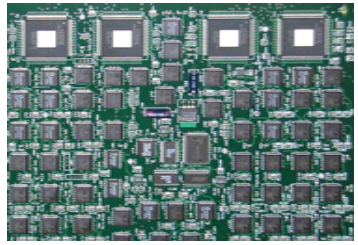


Asynchronous clusters



IMPLEMENTATION

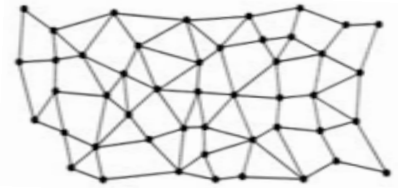
Custom design
(ASIC, FPGA, LVDS..)



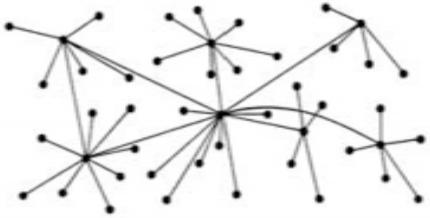
Commodities
PCs and networks

INTERCONNECTION

Distributed



Decentralised

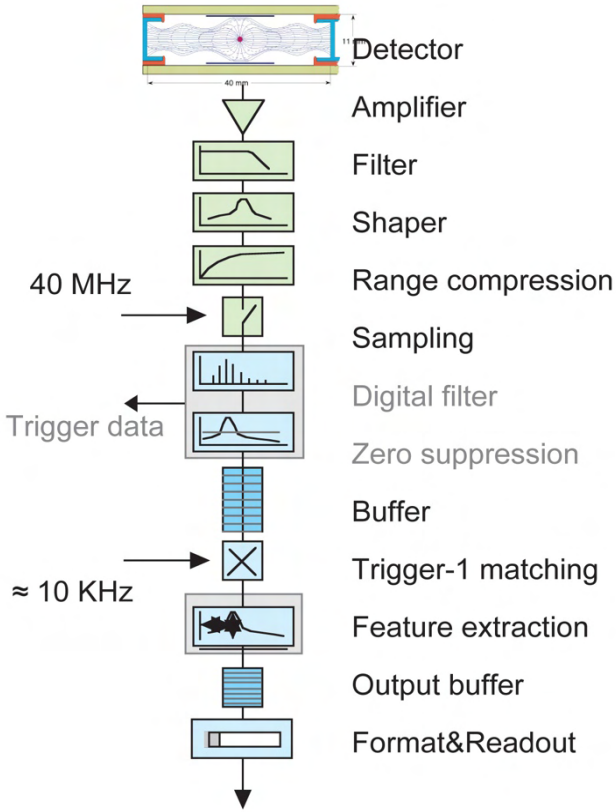


High Level Triggers architecture. **Cluster structure: One event -> One processor**

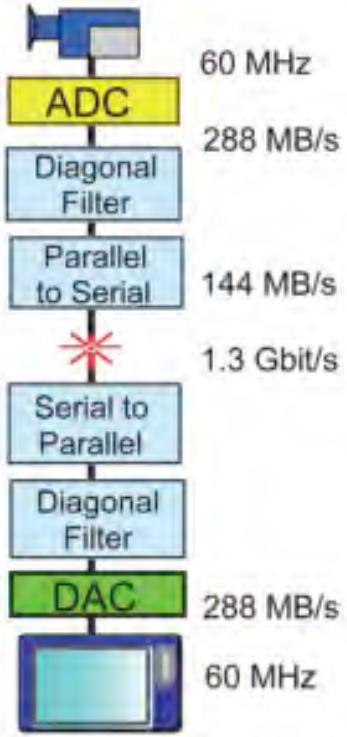
- Loose coupling, large latency. Node high power

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.

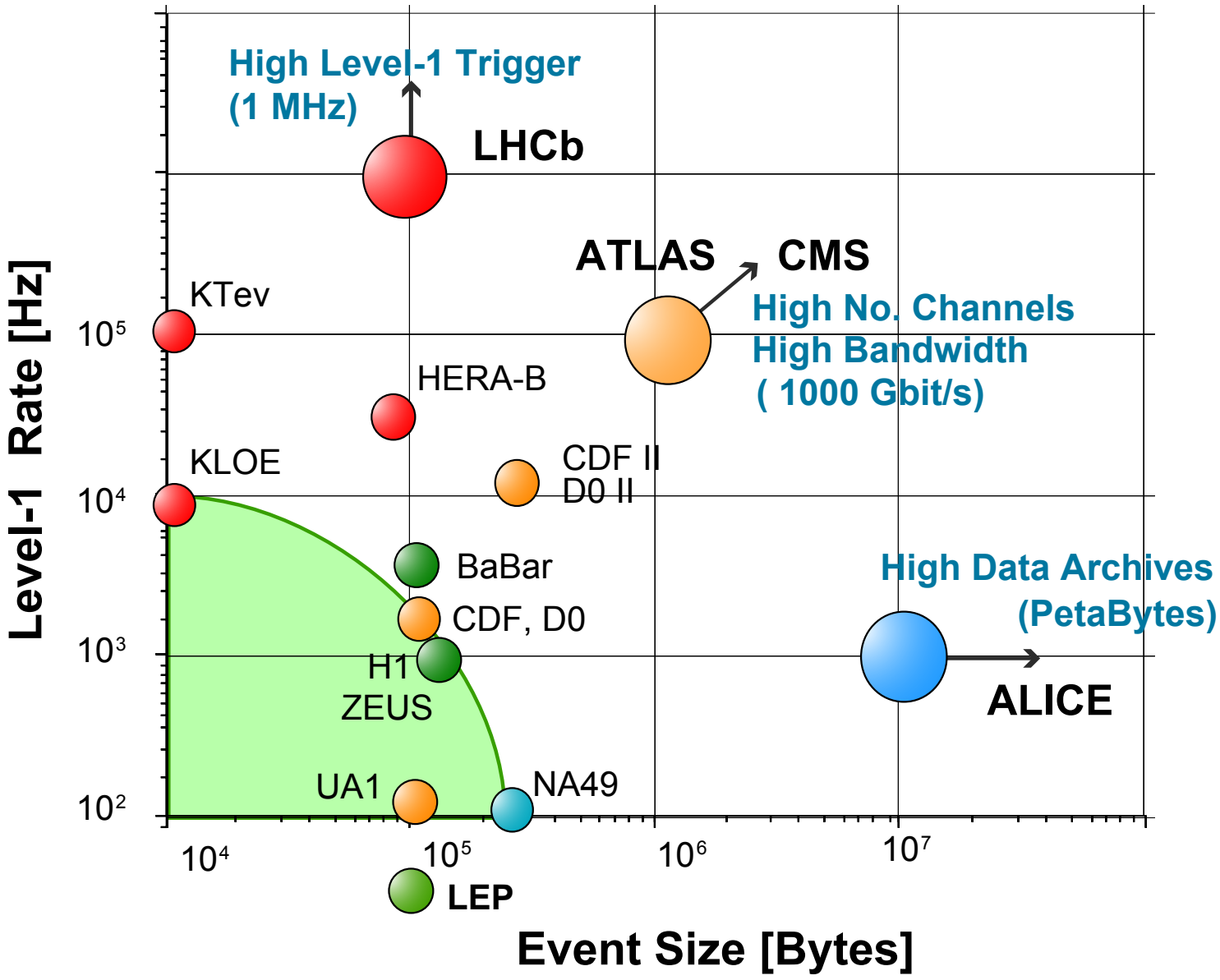
- The telecommunication industry has been one of the major customers for the development of this technology.

- 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.

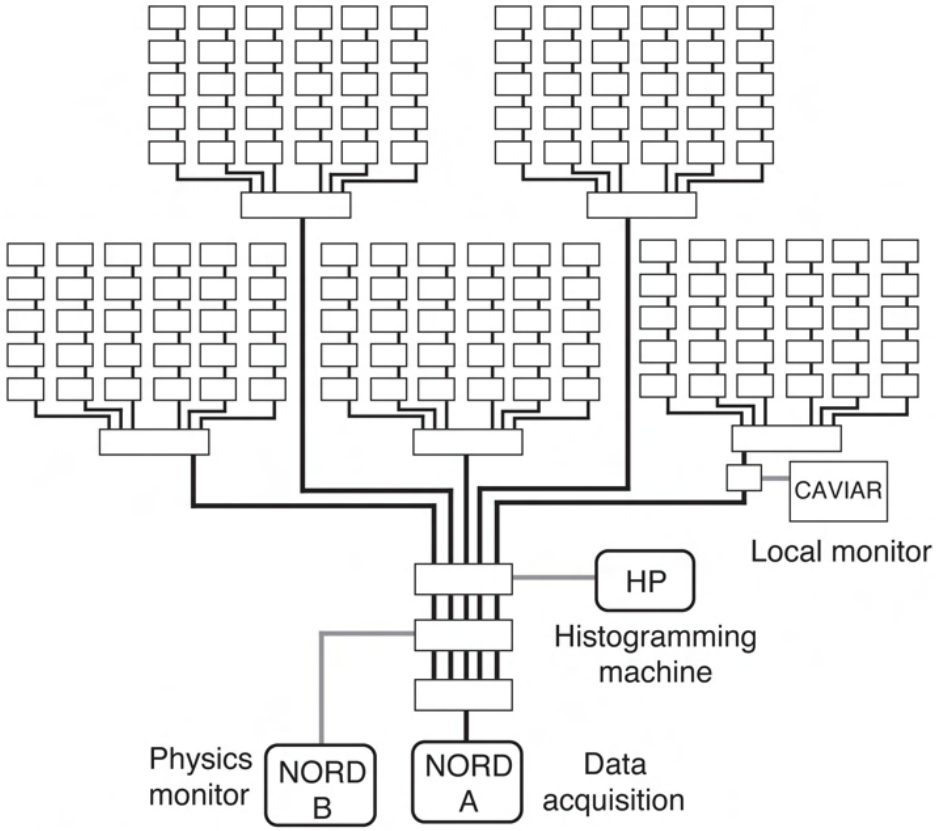
DAQ design issues at LHC (1990-2010)

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

HEP experiments Level-1 rate / data volume trends

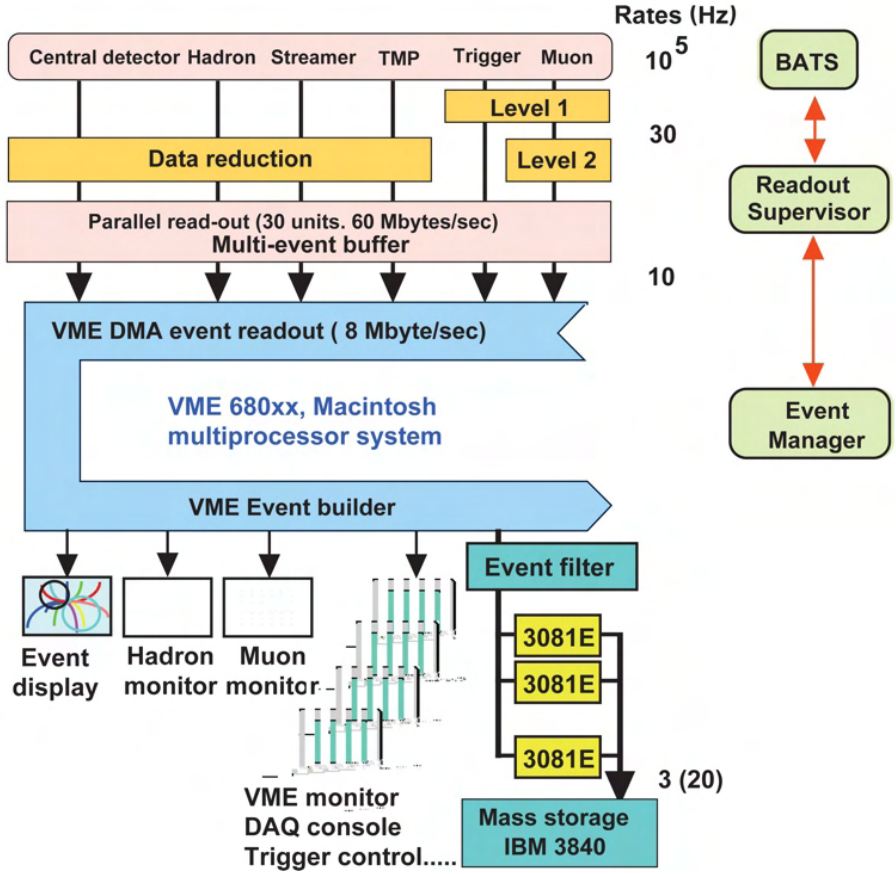


1978-1989. UA1 DAQ system



1981-84

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

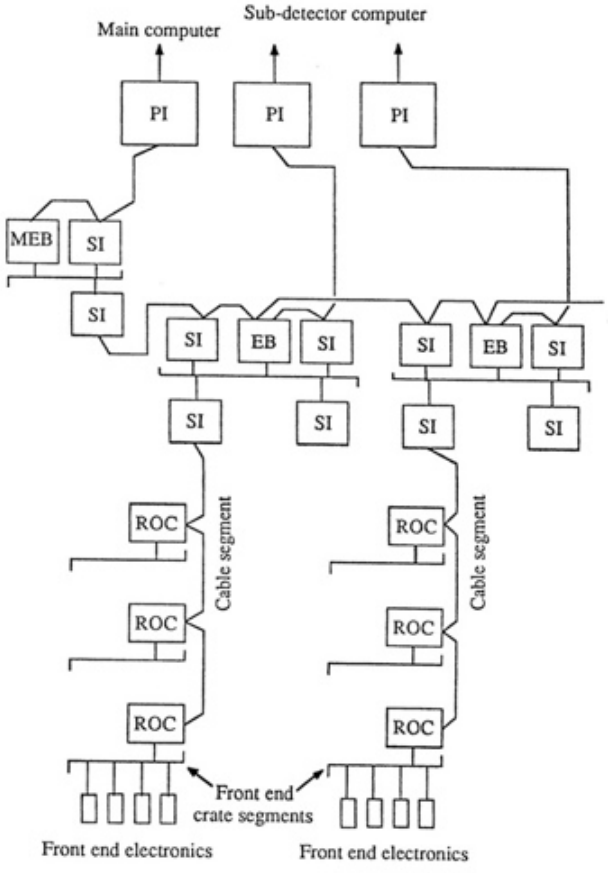


1985-1989
VME, IBM-emulators, Desktops

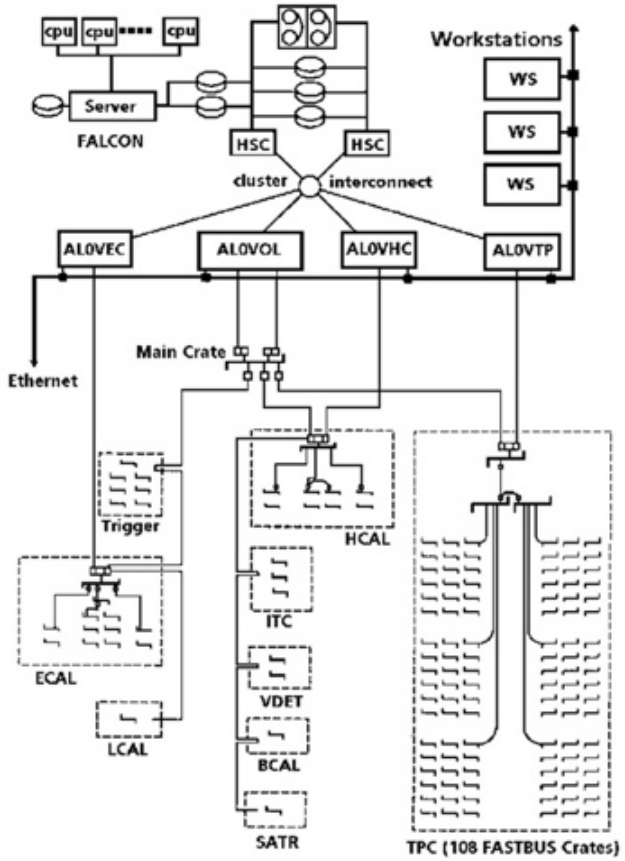
Proprietary/Standards: CAMAC, embedded µP, custom-CPU, VME

1989-2001 LEP DAQ systems

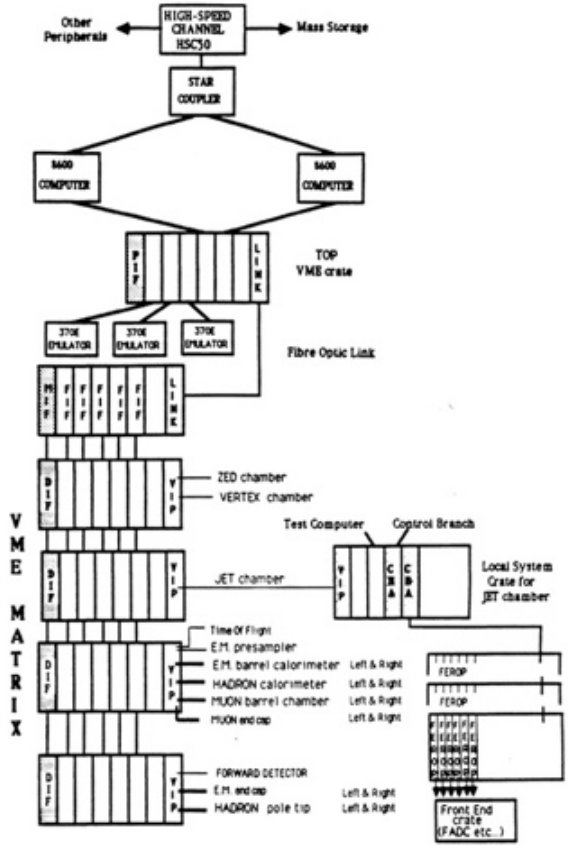
Aleph



Delphi

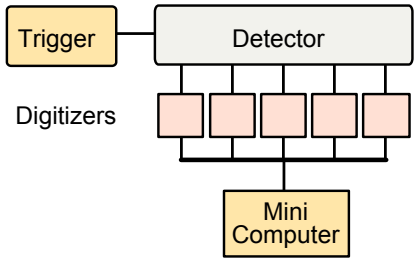


Opal



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

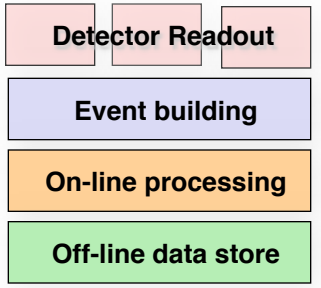
TDAQ architecture evolution

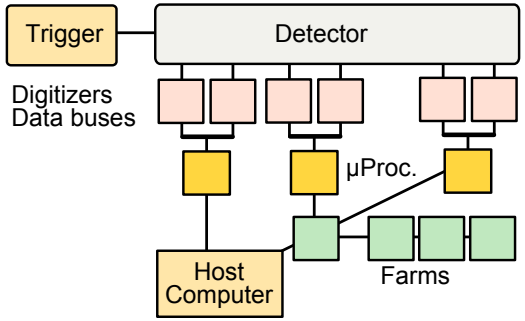


1970-80. PS/ISR/SPS: Minicomputers

Readout custom design
First standard: CAMAC
Software: no-OS, Assembler

- **kByte/s, kFlop**

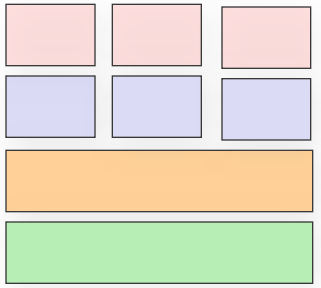


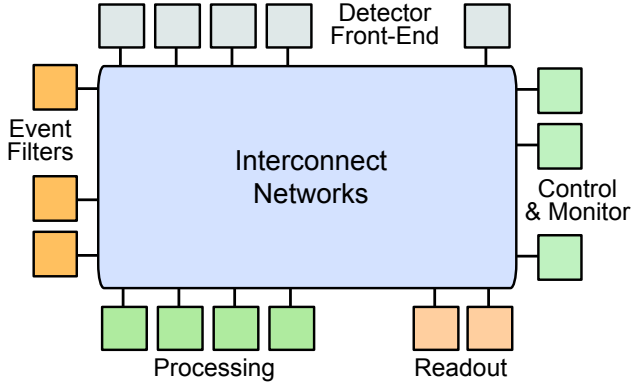


1980-90. p-p/LEP: Microprocessors

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

- **MByte/s, MFlop**

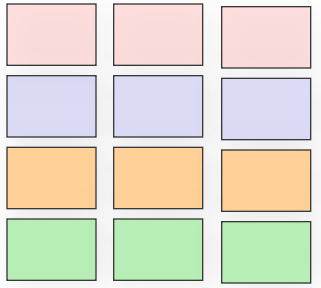




2000. LHC: Networks/Clusters/Grids

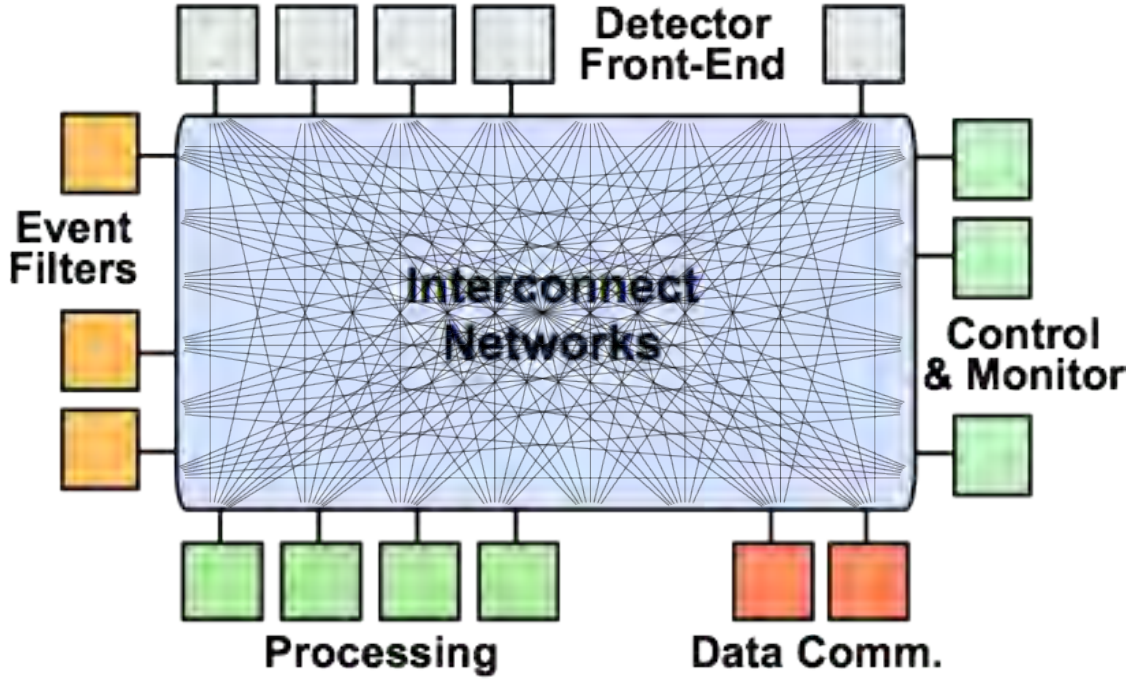
PC, PCI, Clusters, point to point switches
Software: Linux, C,C++,Java,Web services
Protocols: TCP/IP, I2O, SOAP,

- **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

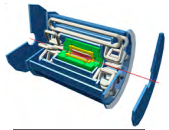


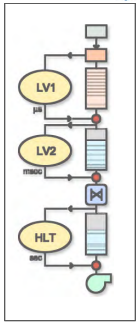
At the time of the finalization of the system design (2002-03), **a single network technology could not satisfy at once all the LHC requirements. The LHC DAQ designs had to adopt multiple specialized networks instead.**

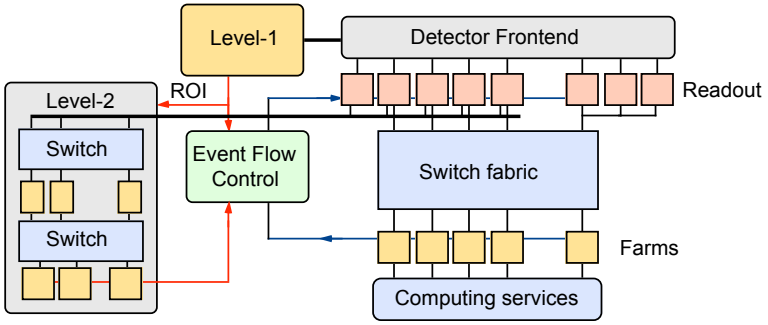
TDAQ data network designs

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**

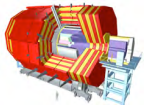
ATLAS

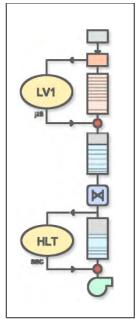


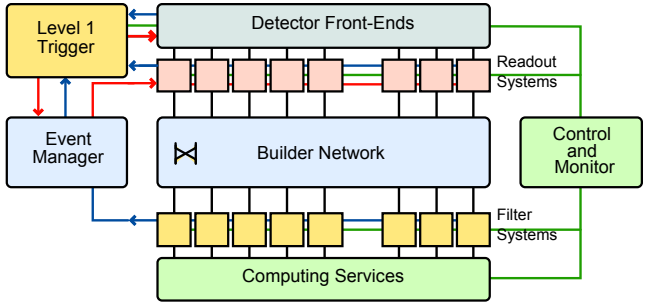




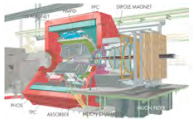
CMS

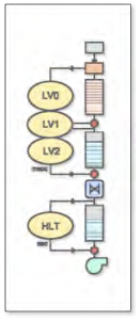


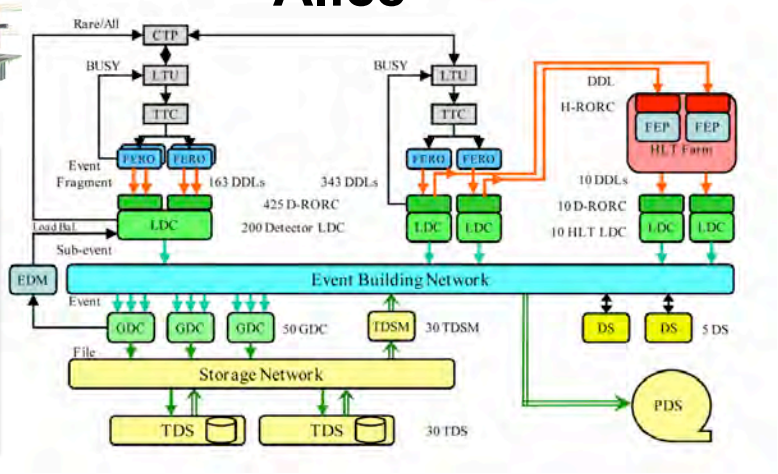




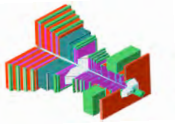
Alice

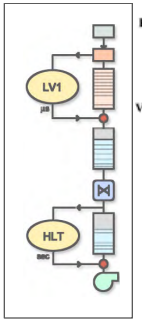


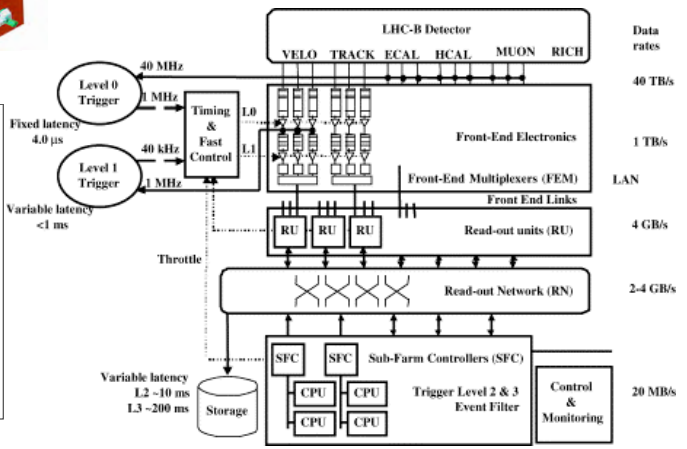




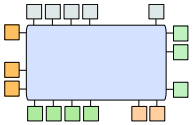
LHCb

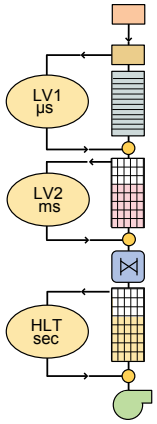


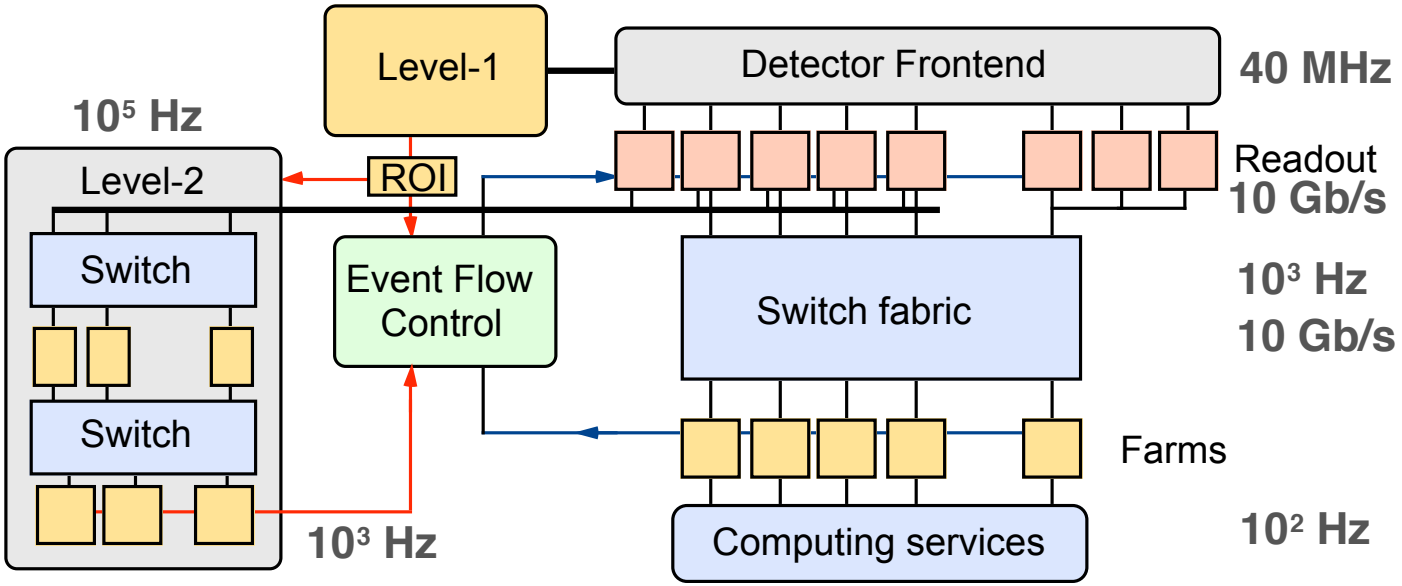




Three levels ATLAS TDAQ







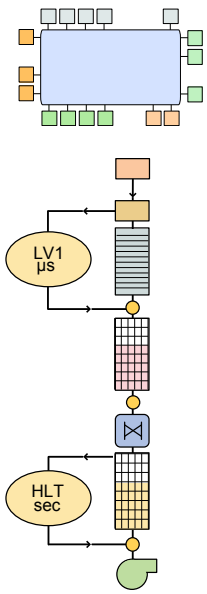
ATLAS LVL2 trigger refines the selection of candidate objects compared to LVL1, using full-granularity information from all detectors, including the inner tracker which is not used at LVL1. In this way, the rate can be reduced to ~1kHz. The data can be accessed selectively by the **LVL2 trigger which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

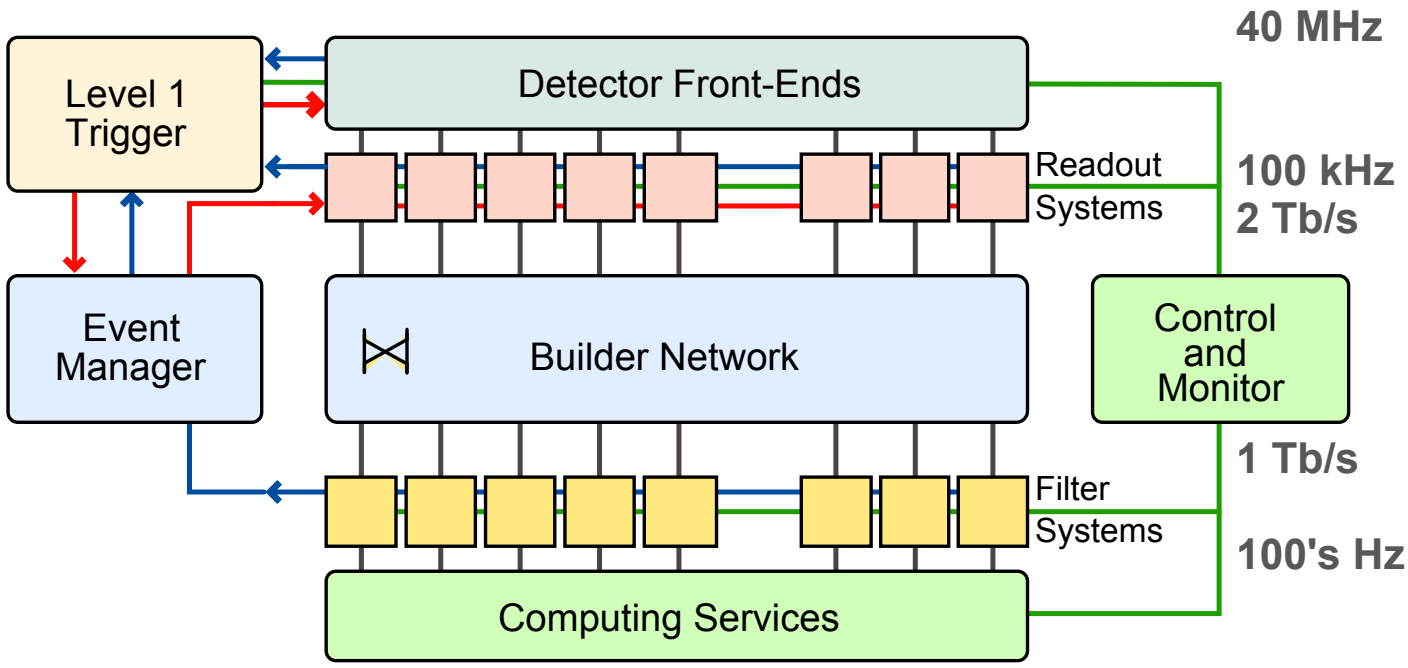
Collision rate	40 MHz
Level-1 Maximum trigger rate	**100 kHz**
Average event size	**≈ 1.5 Mbyte**
Flow control&monitor	≈ 10^6 Msg/s

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



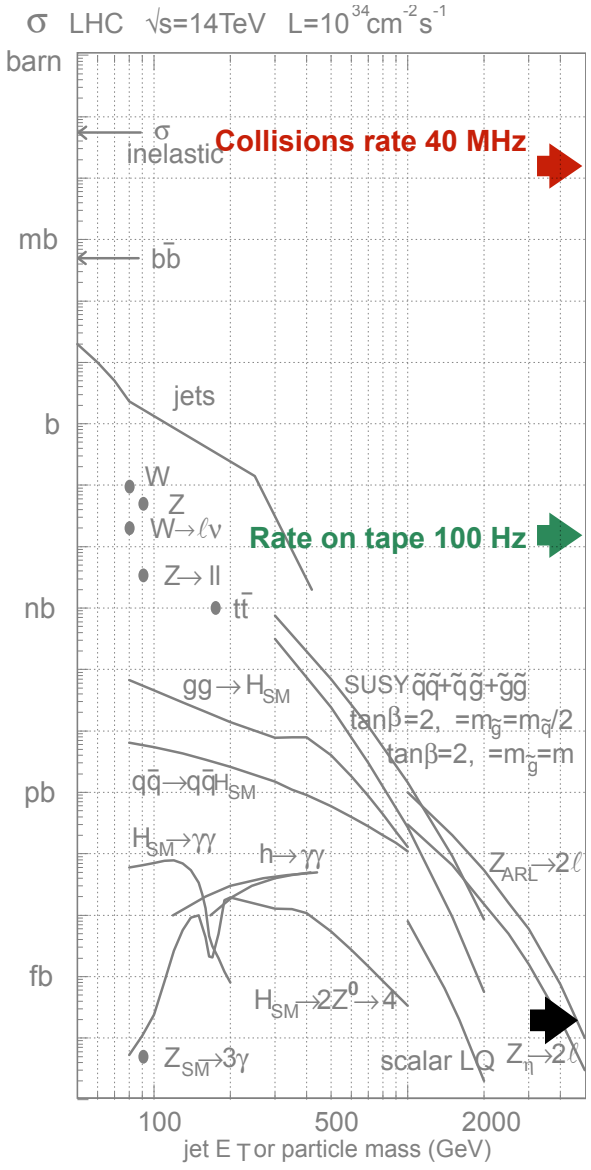


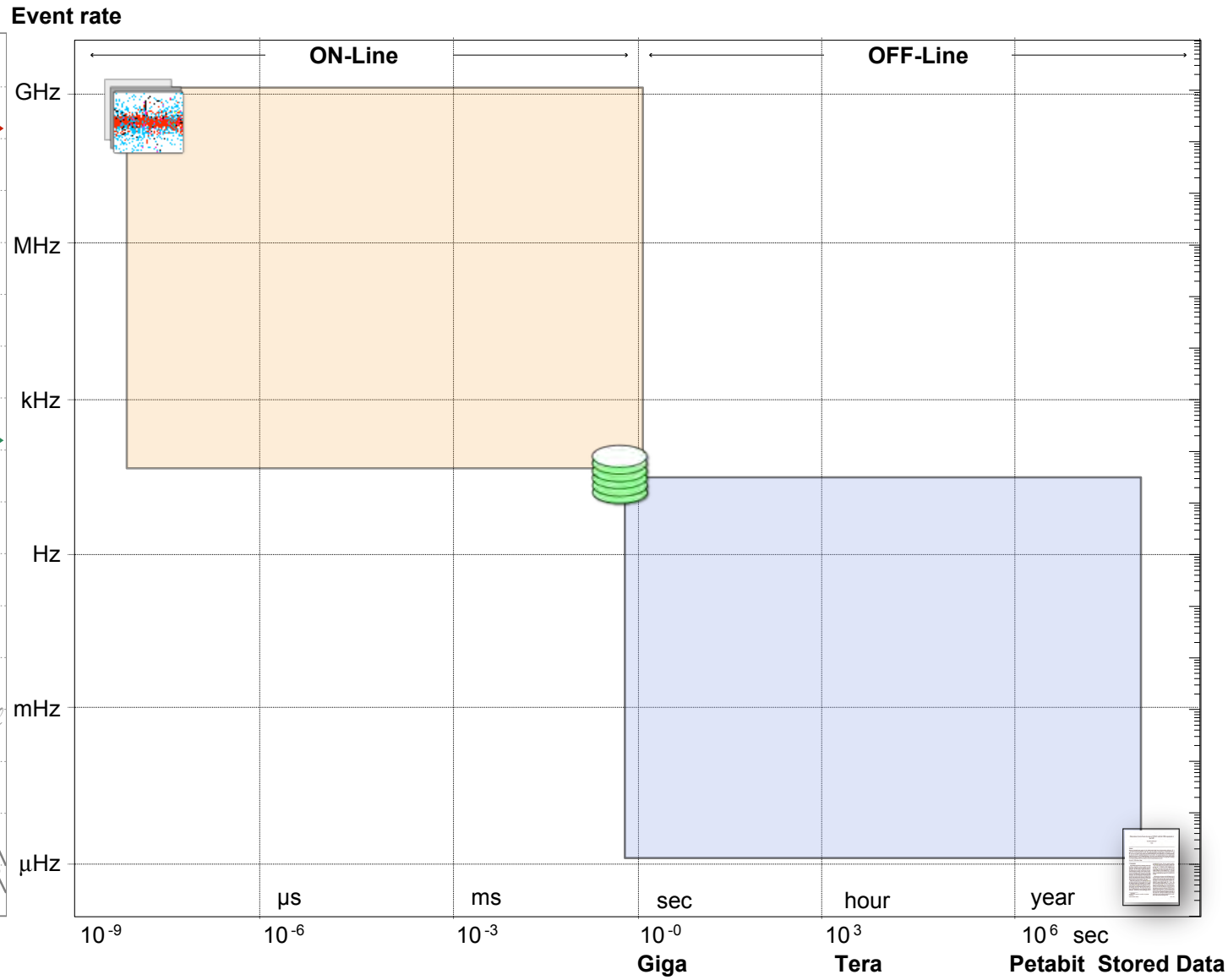
Collision rate	40 MHz
Level-1 Maximum trigger rate	**100 kHz**
Average event size	**≈ 1 Mbyte**
Flow control&monitor	≈ 10^6 Msg/s

Readout concentrators/links	**512 x 4 Gb/s**
Event Builder bandwidth max.	**2 Tb/s**
Event filter computing power	**≈ 10-20 TeraFlop**
Event Builder GBE ports	**> 4000**
Data production	≈ Tbyte/day
Processing nodes	≈ x Thousands

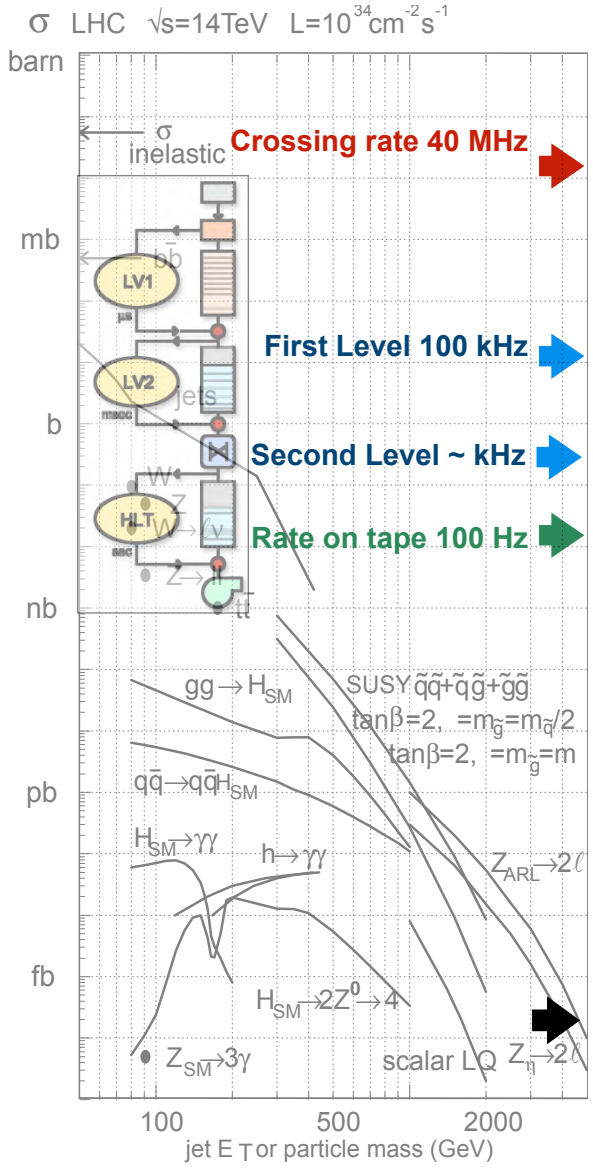
Proprietary/Standards: Front-end, VME, PC servers, Networks, Protocols, OS

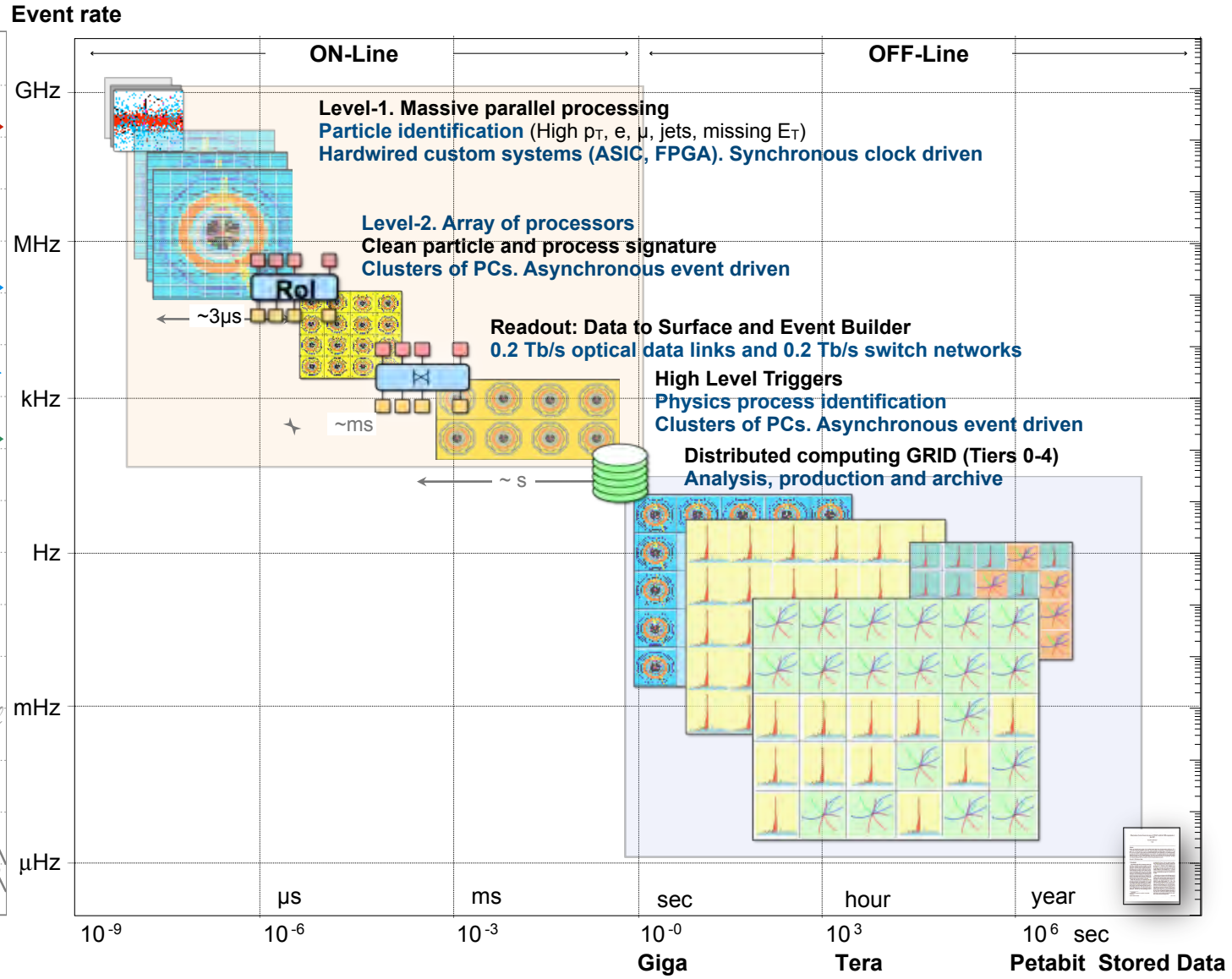
On-line rate decimation and data flow



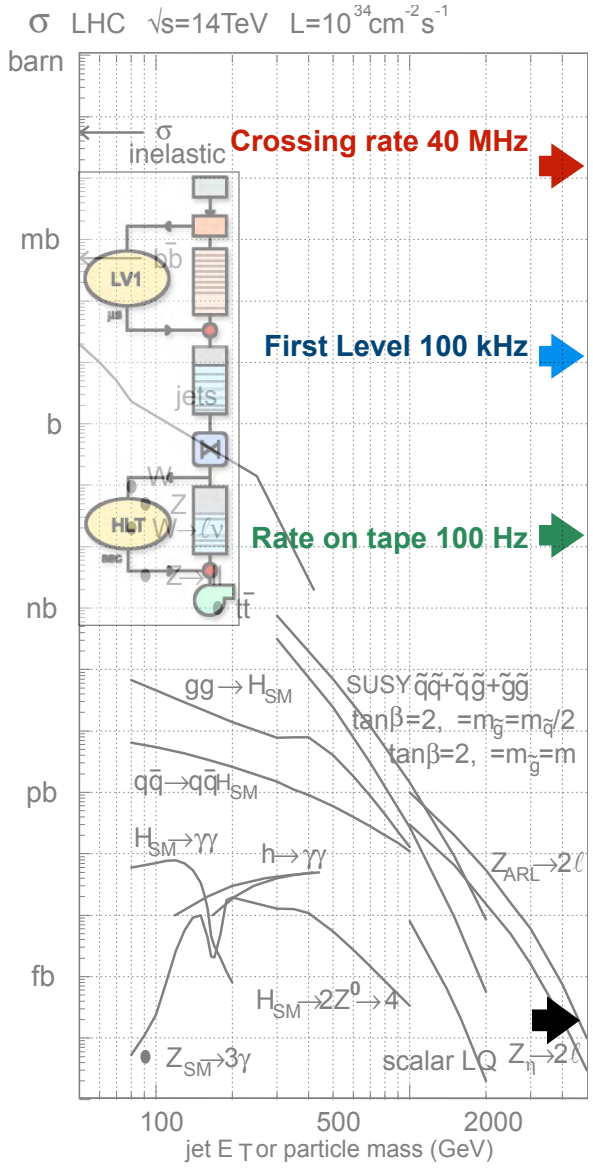


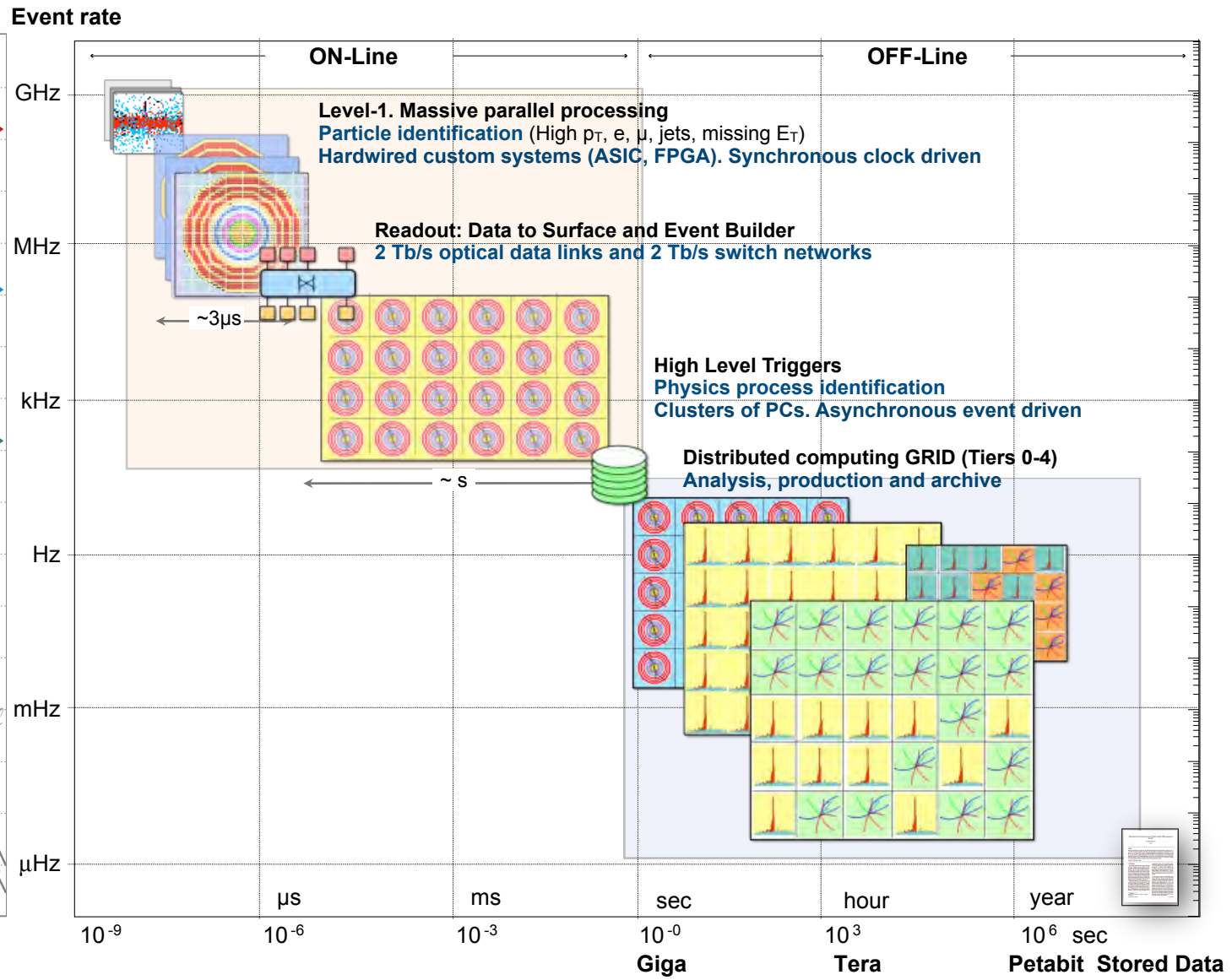
ATLAS: On-line trigger levels and event building



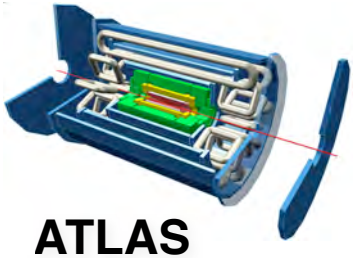
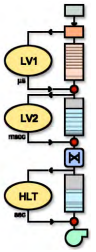
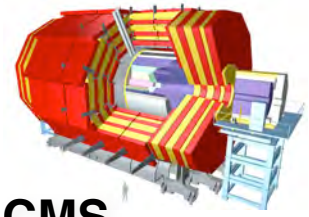
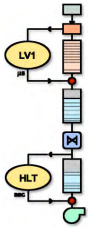
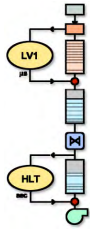
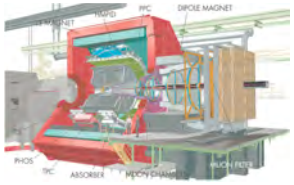
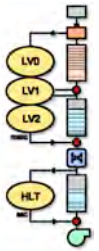


CMS: On-line trigger levels and event building





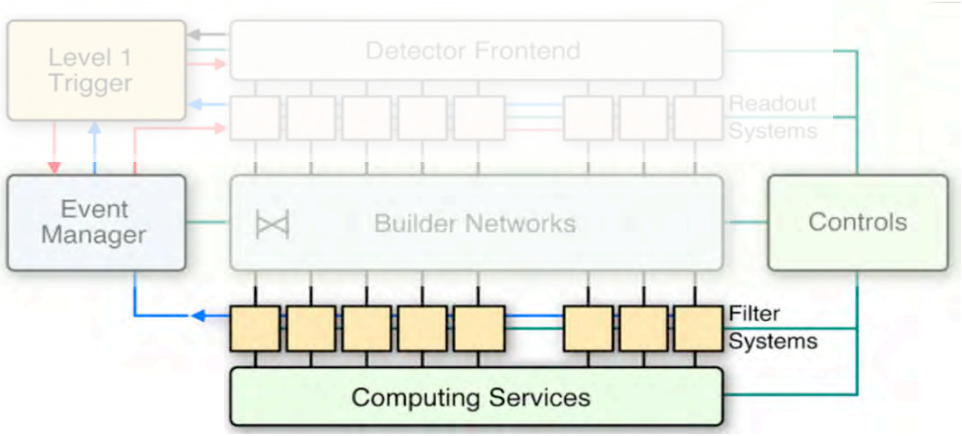
LHC experiments TDAQ summary

	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 1.5x10 LV-2 $3x10^3$	6	4.5	300 ($2x10^2$)
 CMS	2 	LV-1 10^5 10	6	100	O(1000) (10^2)
 LHCb	2 	LV-0 10^6 3x10	4	30	40 ($2x10^2$)
 ALICE	4 	Pb-Pb 500	$5x10^7$	25	1250 (10^2)
		p-p 10^3 2x10	6	200	(10^2)

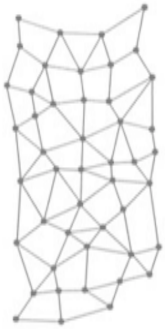
Computing and networking

- Scale free systems

Architecture issue I: Scale-free HLT parallel processing

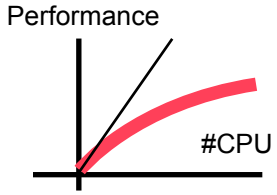


Massive Parallel Processing (MPP)

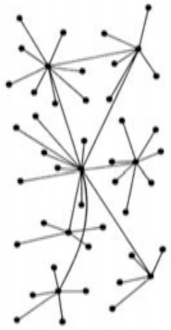


ONE Event, ALL processors

- **Distributed network**
- Programming complexity
- Single points of failure
- Low latency
- **Exponential scaling**

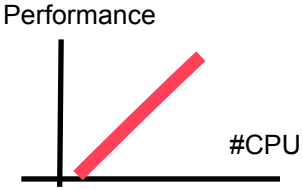


Cluster of processors (CPU farm)



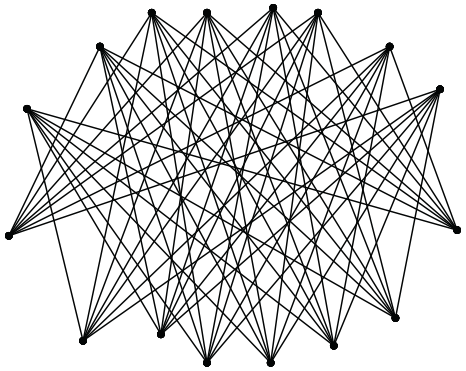
ONE Event, ONE processor

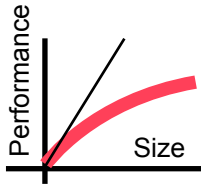
- **Decentralised network**
- Sequential programming
- 100 kHz, 10000 cores
- High latency (large memory)
- **Scale free**

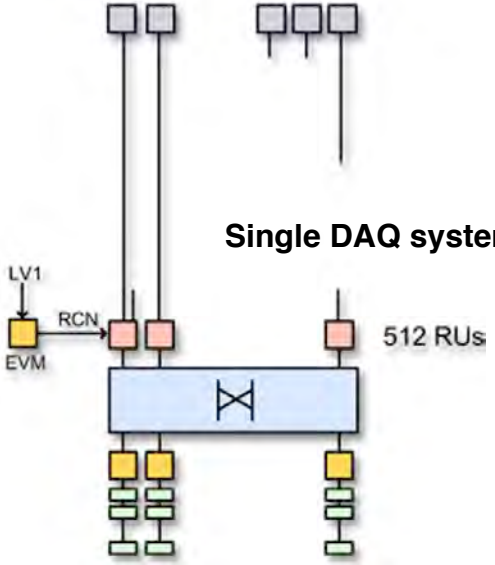


Architecture issue II. Scale-free event builder

Exponential network expansion

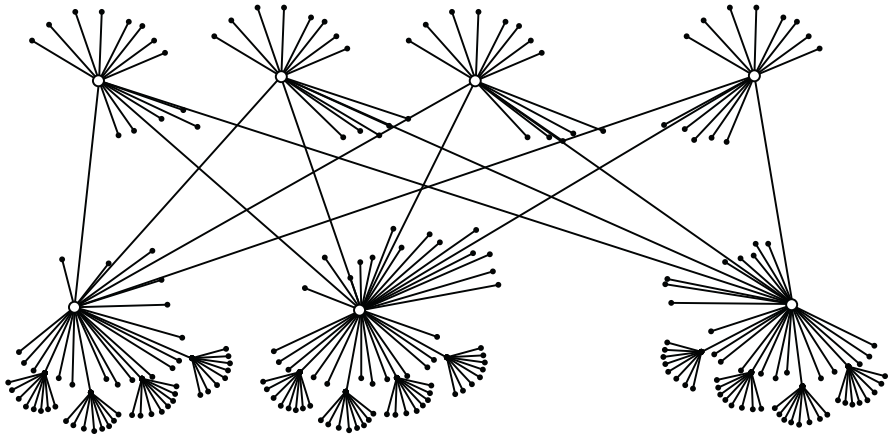


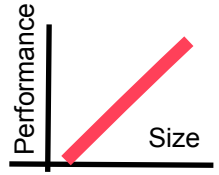


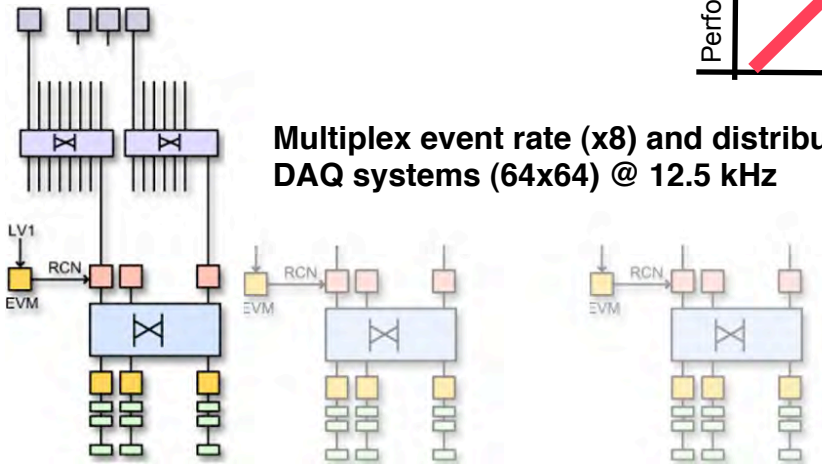


Single DAQ system (512x512) @ 100 kHz

Scale-free network expansion

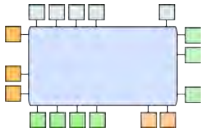






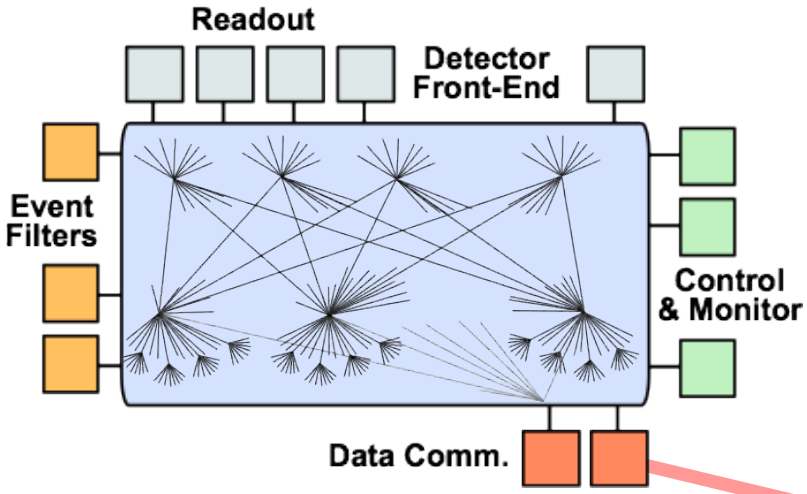
Multiplex event rate (x8) and distribute to 8 DAQ systems (64x64) @ 12.5 kHz

LHC On/Off-line data handling summary



No technology today provides the functionality and performance required by the overall throughput and Of/Off-line computing.

Factorize the problem: splitting **On-line (TDAQ) and Off-line (Analysis)**

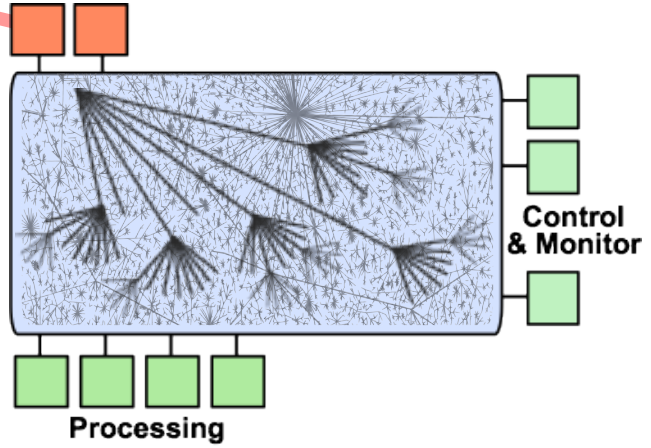


On-line (TDAQ)

-Custom logic
-Front-end readout by custom link
-Data readout by dedicated networks
-Optimized network interconnections
-Local processing units (HLT)
-Local data storage

Off-line (Storage & Analysis)

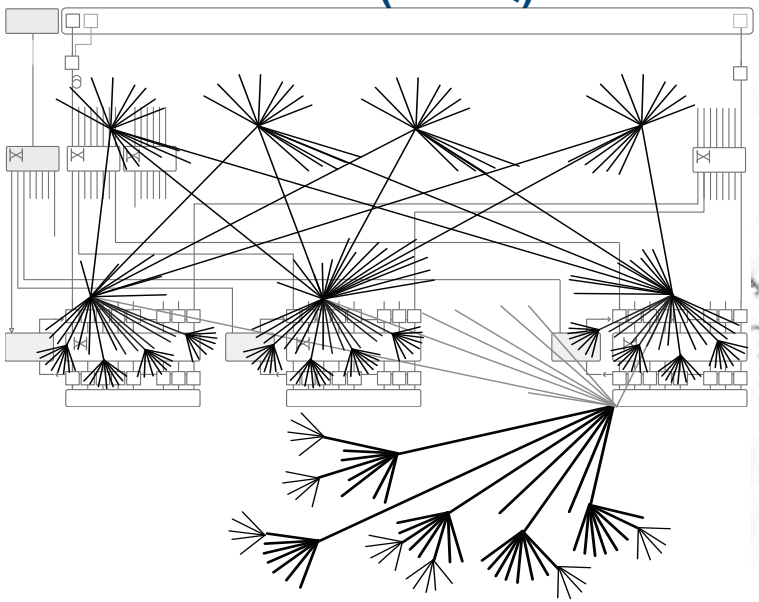
-Centralized permanent data archive
-Decentralized processing and storage
-Distributed physics analysis
-Interconnection by public networks



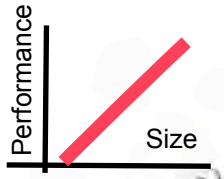
Scale-free networks

On/Off-line **TDAQ (and GRID) systems are, by construction, scale-free systems;** they are capable of operating efficiently, taking advantage of any additional resources that become available or as they change in size or volume of data handled.
Other complex systems. e.g. **the Word Wide Web, show the same behavior.** This is the result of the simple mechanism that allows networks to expand by the addition of new vertices which are attached to existing well-connect vertices.

On-line (TDAQ)

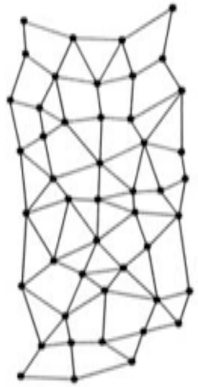


Off-line (GRID)

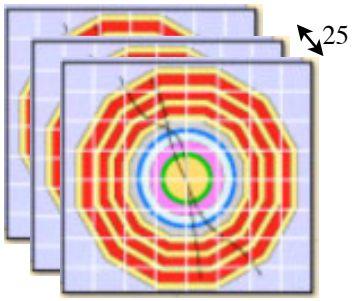


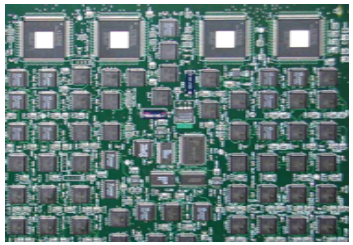
Scale-free internet (2002 snapshot)

Processing architecture and network topologies

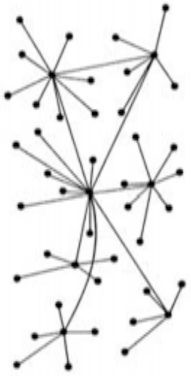


Level 1: Massive parallel pipelined processors. CLOCK DRIVEN
Implementation: Custom design (ASIC, FPGA)

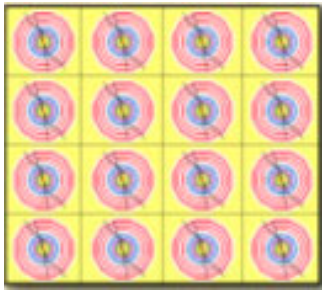




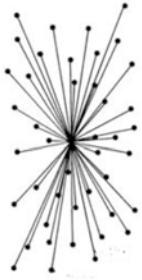
Lv-1 processor, low Latency (μs)
synchronous 40 MHz, 128 cells depth
ONE event ALL processors
Pipeline memory for each channel
Radiation & power issues



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

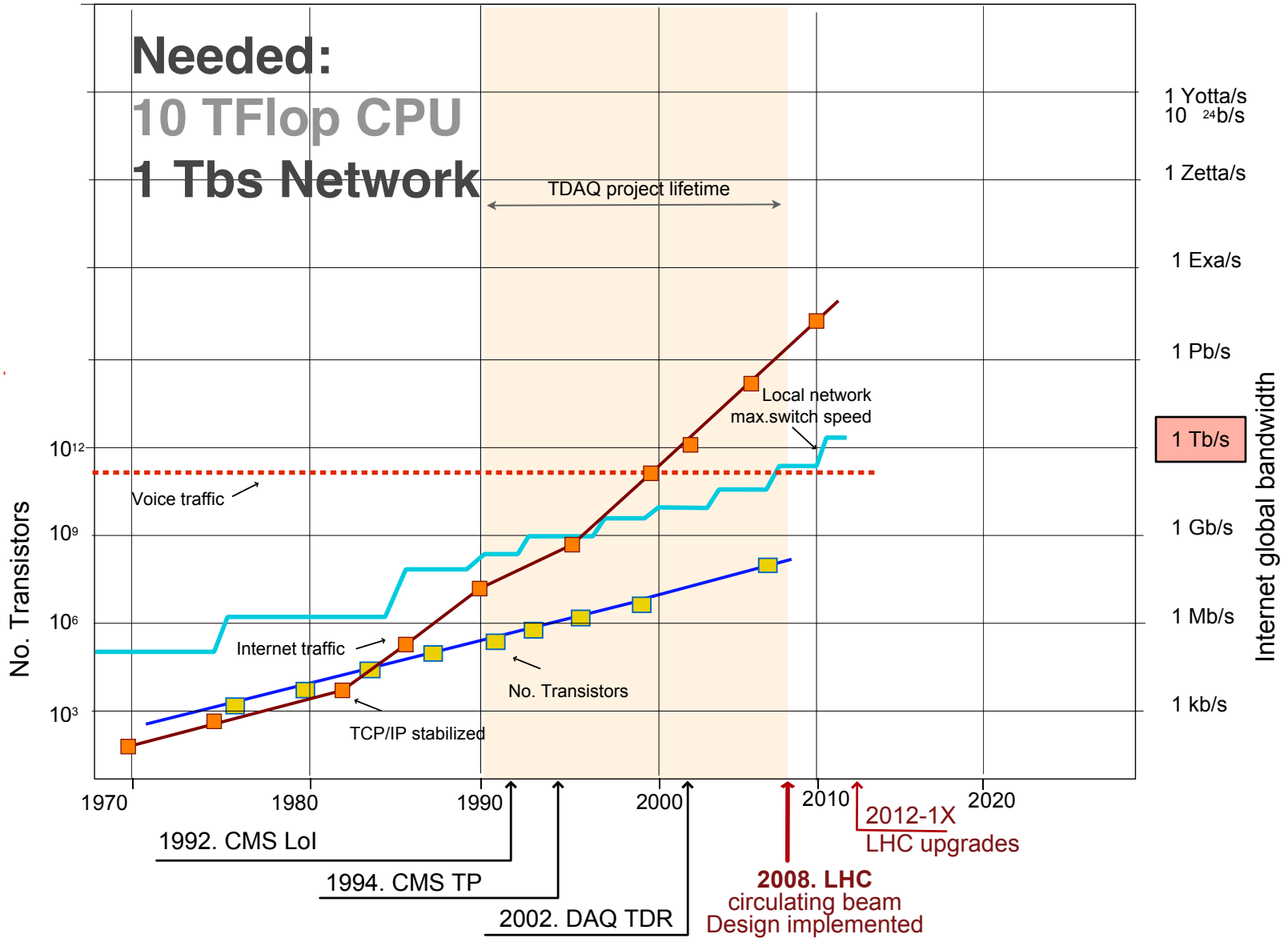


OFF-line. Data source: Centralised
Data analysis and storage distributed GRID. CLOUD...
Implementation: Commodities (Servers, links, networks)

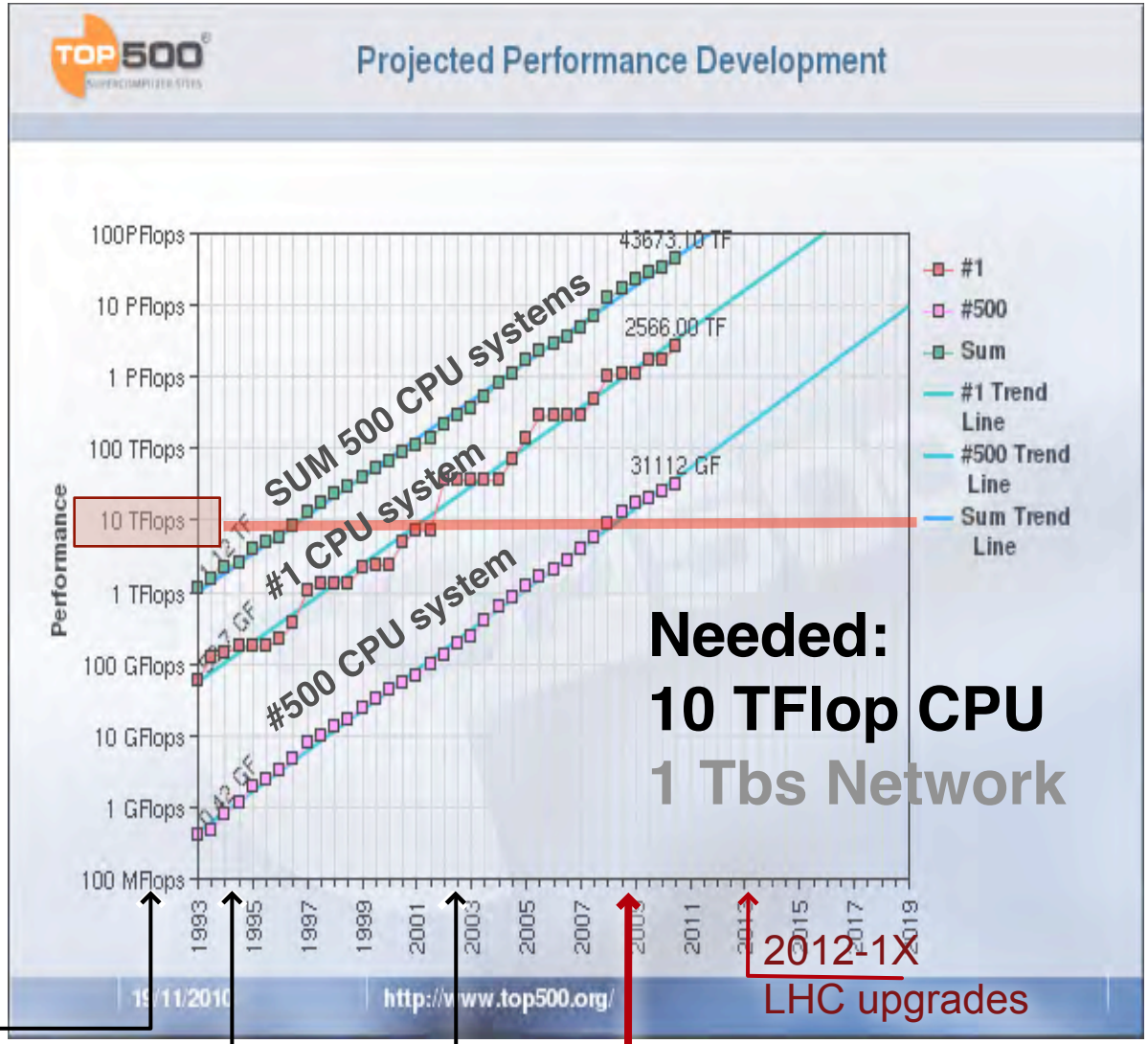
Design issues: Technologies

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

Data communication. Network and Internet traffic trends



Computing power trends



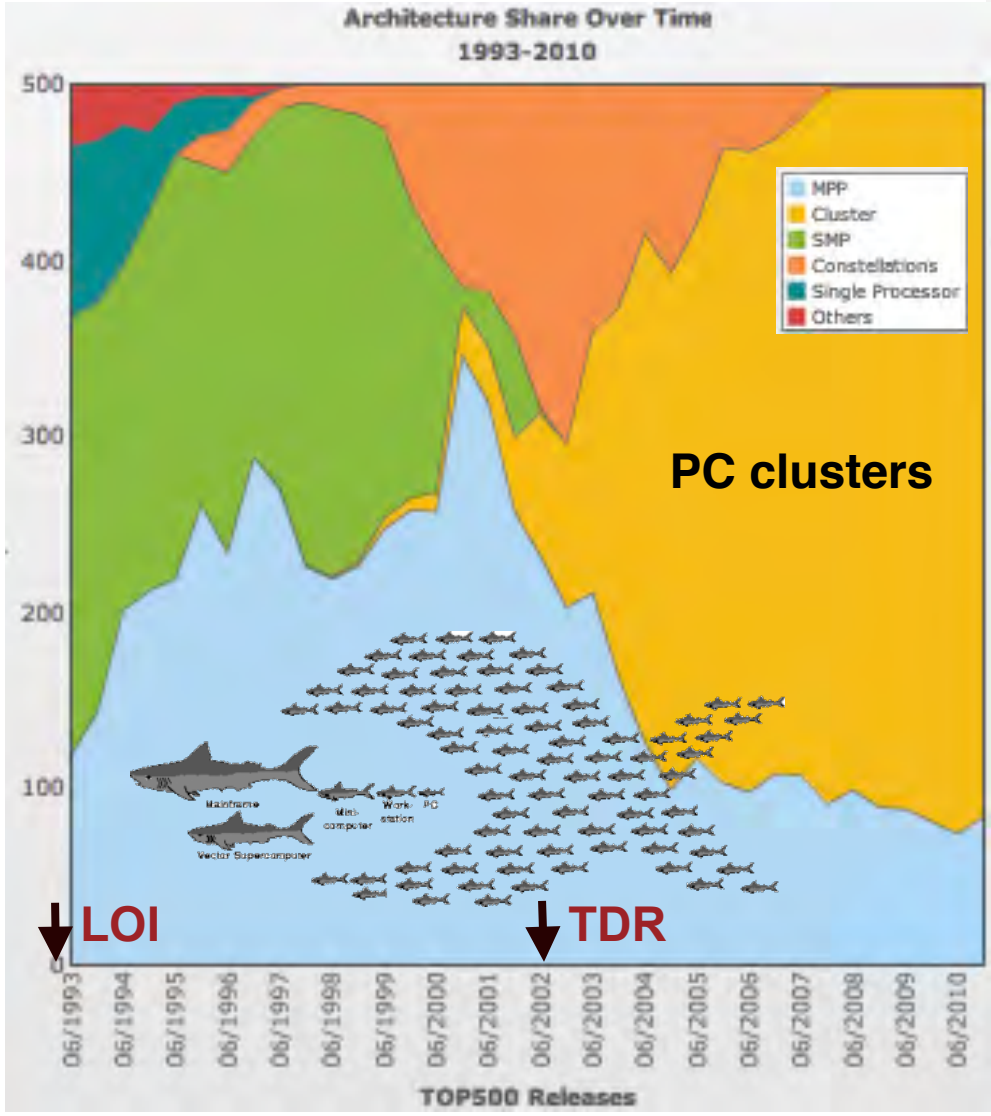
1992. CMS LoI

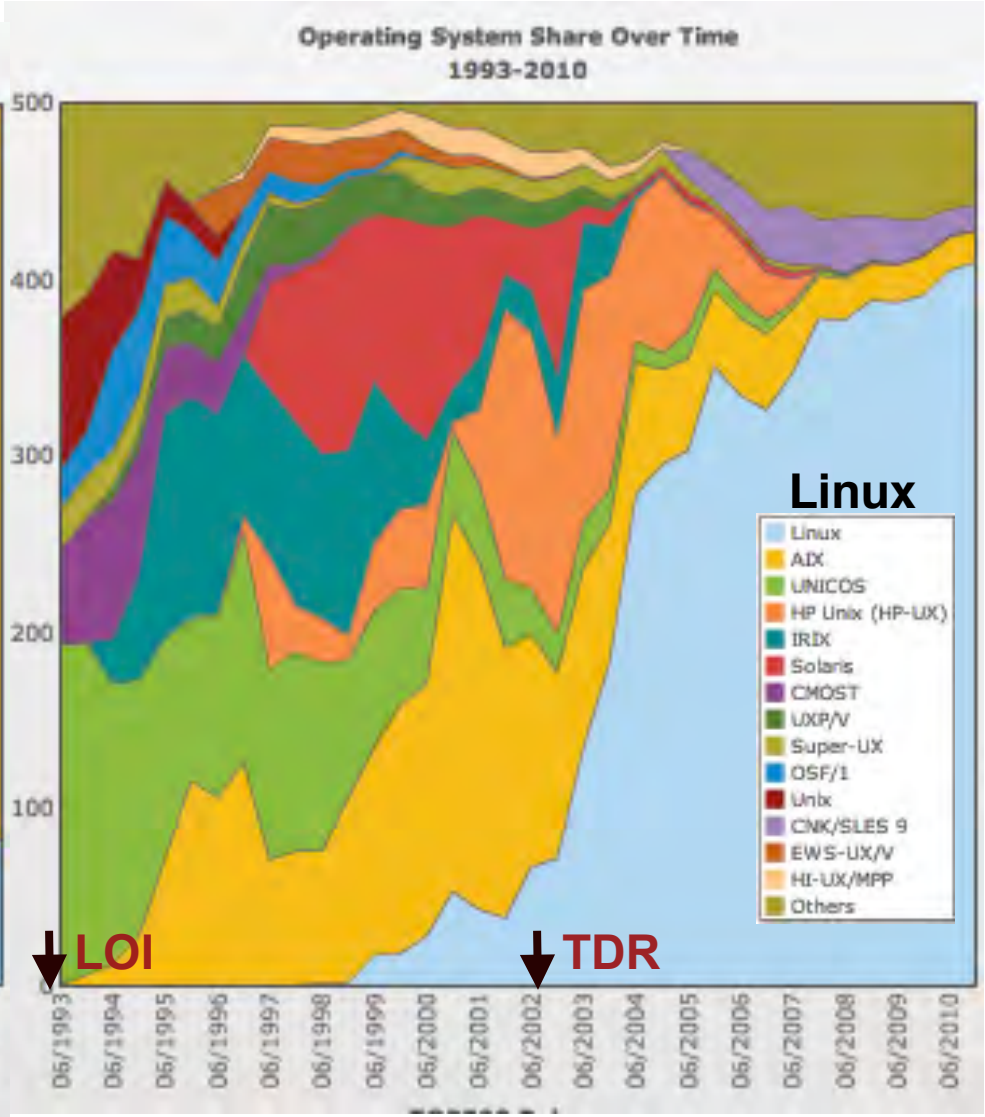
1994. CMS TP

2002. DAQ TDR

2008. LHC
circulating beam

Top500 supercomputer architectures and operating systems history





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.

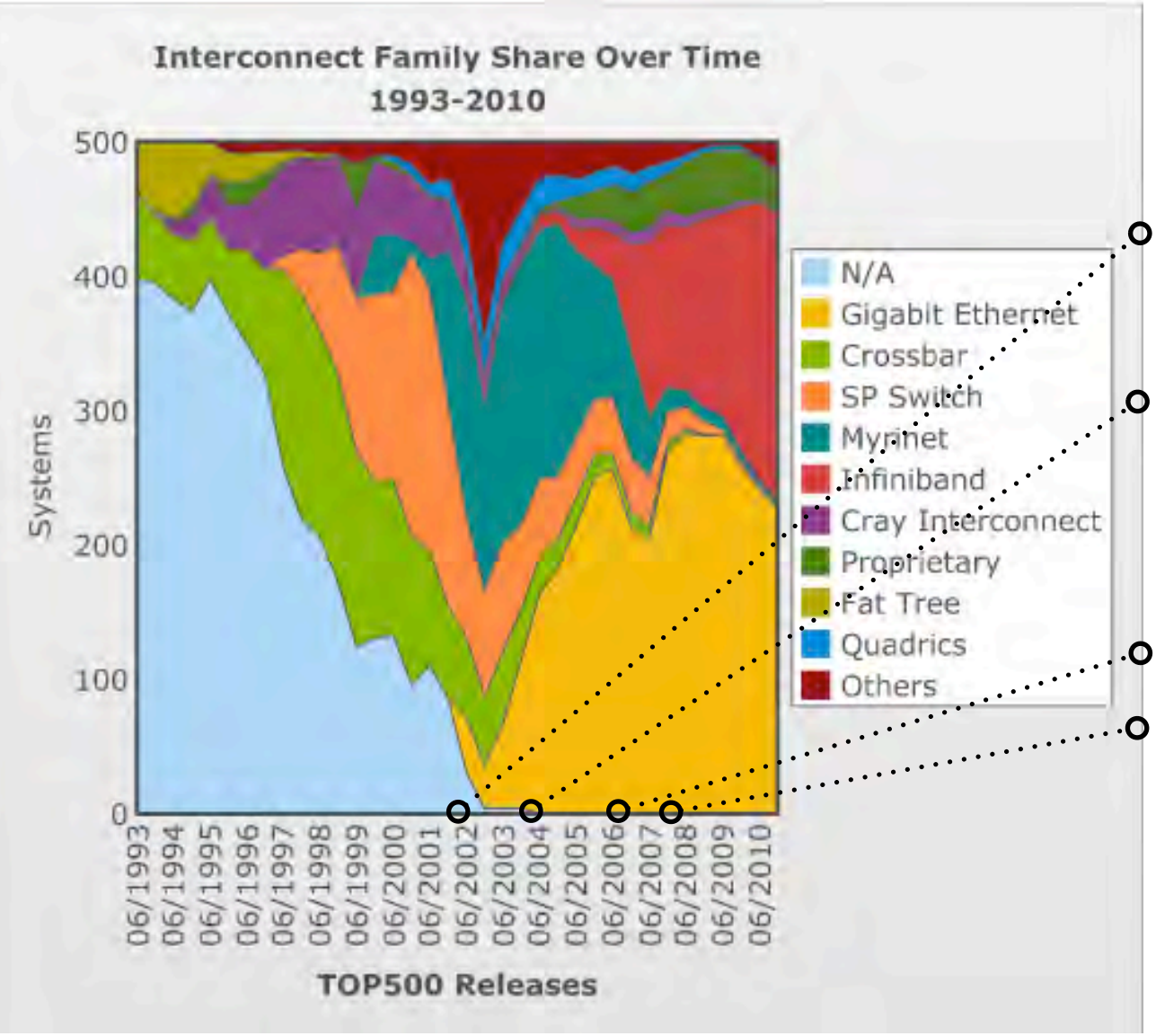
Since 90's:

- 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 interconnection technologies history 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

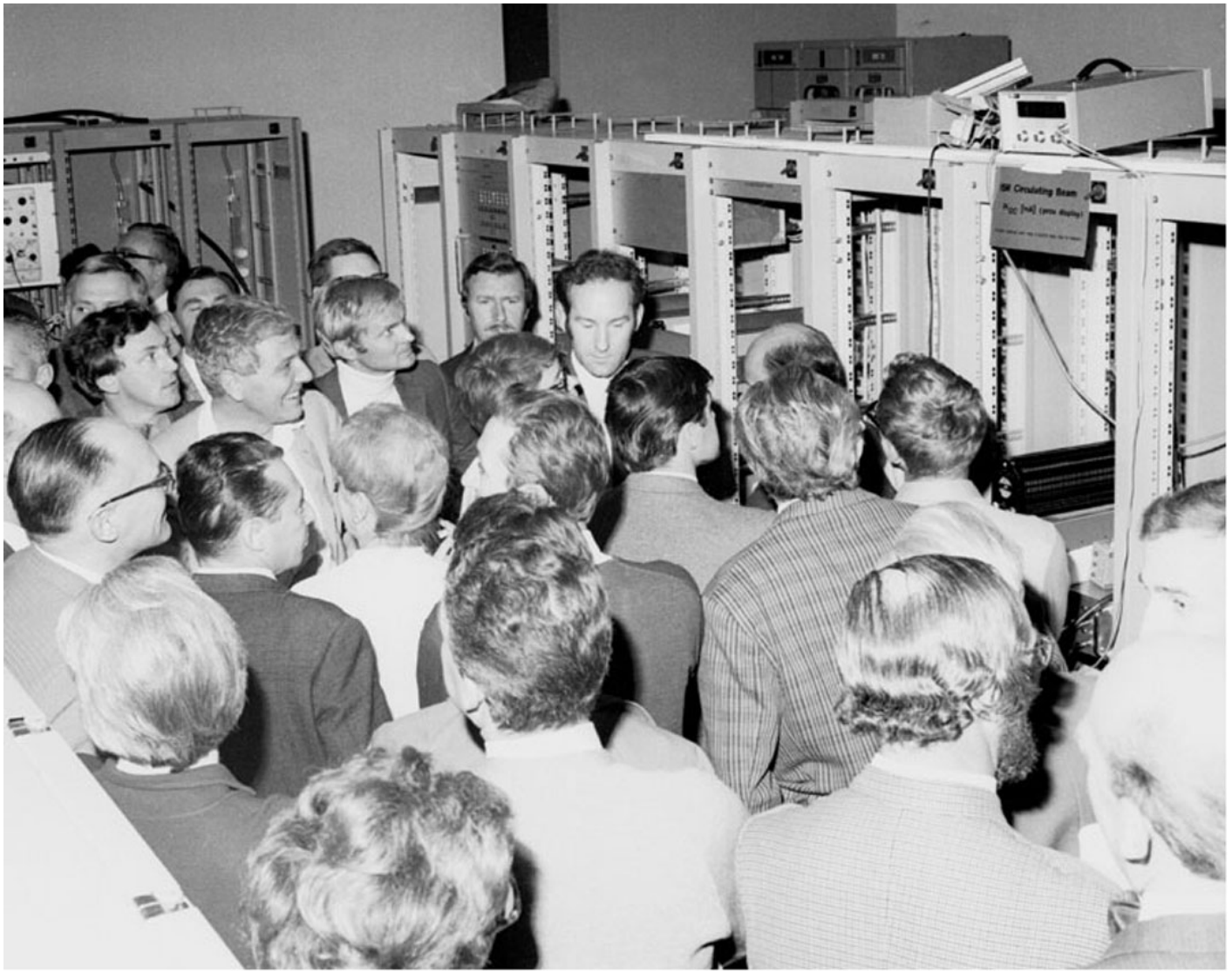
2008 Commissioning

Unpredicted

- Collaborative work
- Network&Computing fusion

Experiment control rooms

ISR. 1970
CR info tools:
Coaxial Cables
Teletype
Telephone

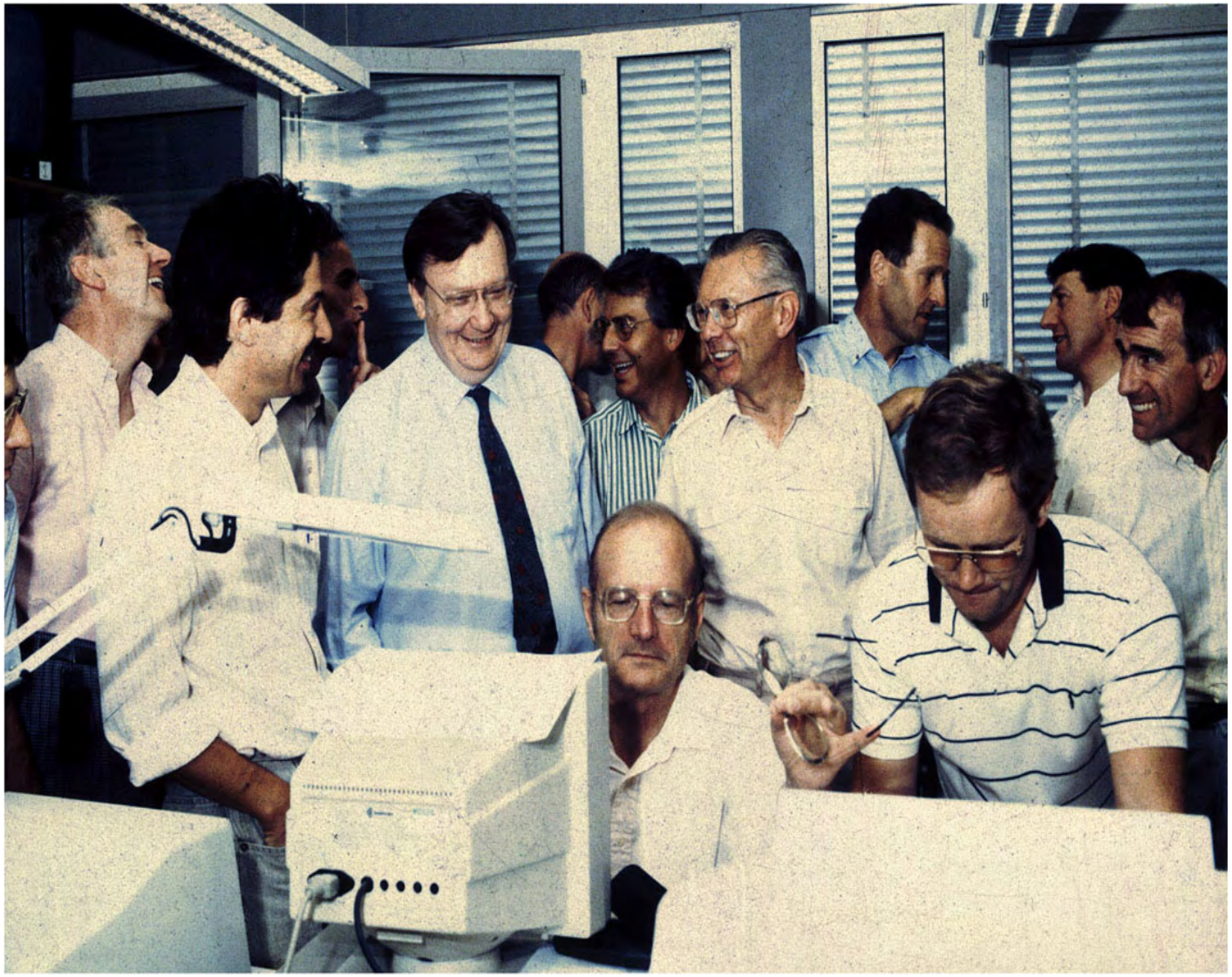


ISR 1970. Voltmeter display, no terminal

Experiment control rooms

ISR. 1970
CR info tools:
Coaxial Cables
Teletype
Telephone

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



1980 P-Pbar. A lot of persons in front of one screen

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

1990 LEP. A lot of screens in front of one person

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

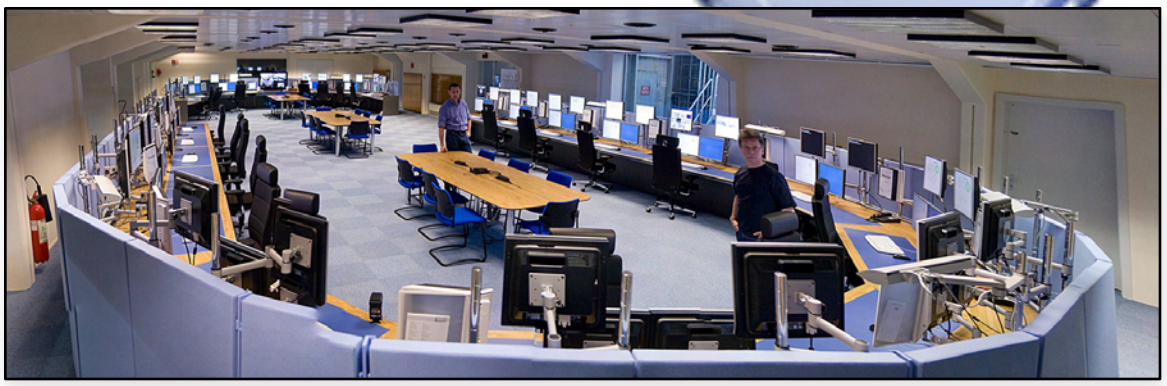
2010 LHC. The person is on the screen

Experiment control and monitor system and WWW services

Cessy: Master&Command control room

Fermilab: Remote Operations Centre

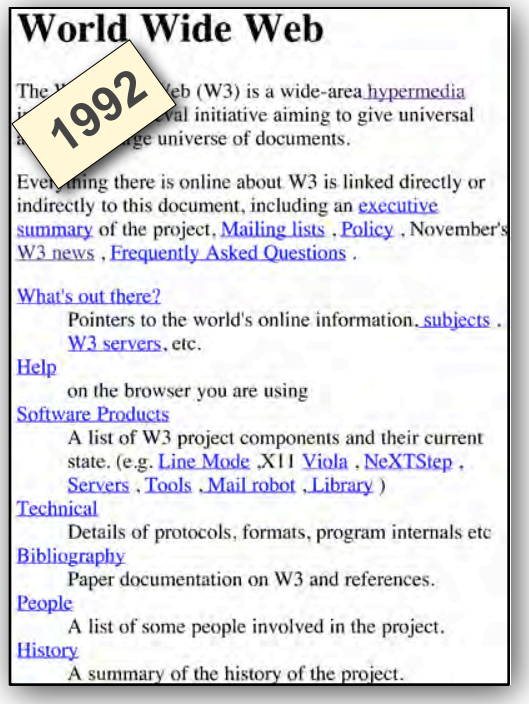
Meyrin: CMS DQM Centre



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**

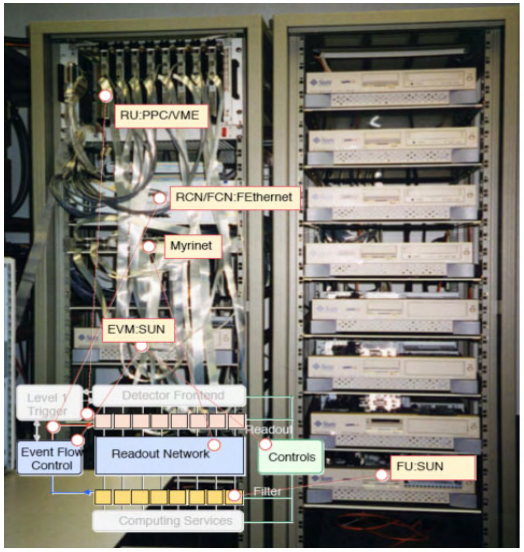
Hard-to-predict in the 90's: The World Wide Web



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,...)

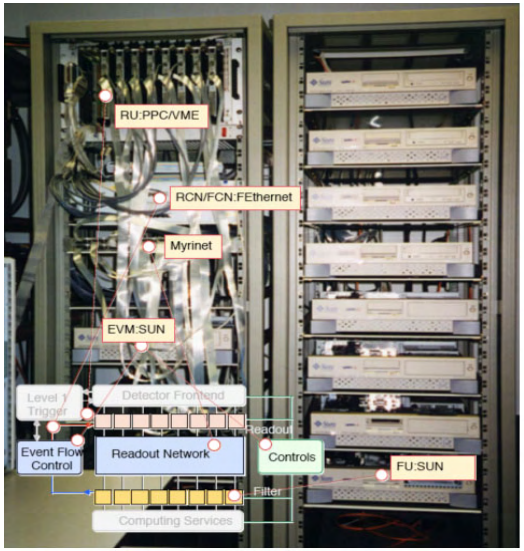
Hard-to-predict in the 90's (II): the same model elsewhere



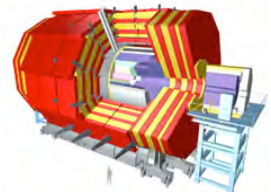
1997 CERN.
A LHC event builder prototype

1997 Stanford.
A Web search engine prototype

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 centers 10^5 cores

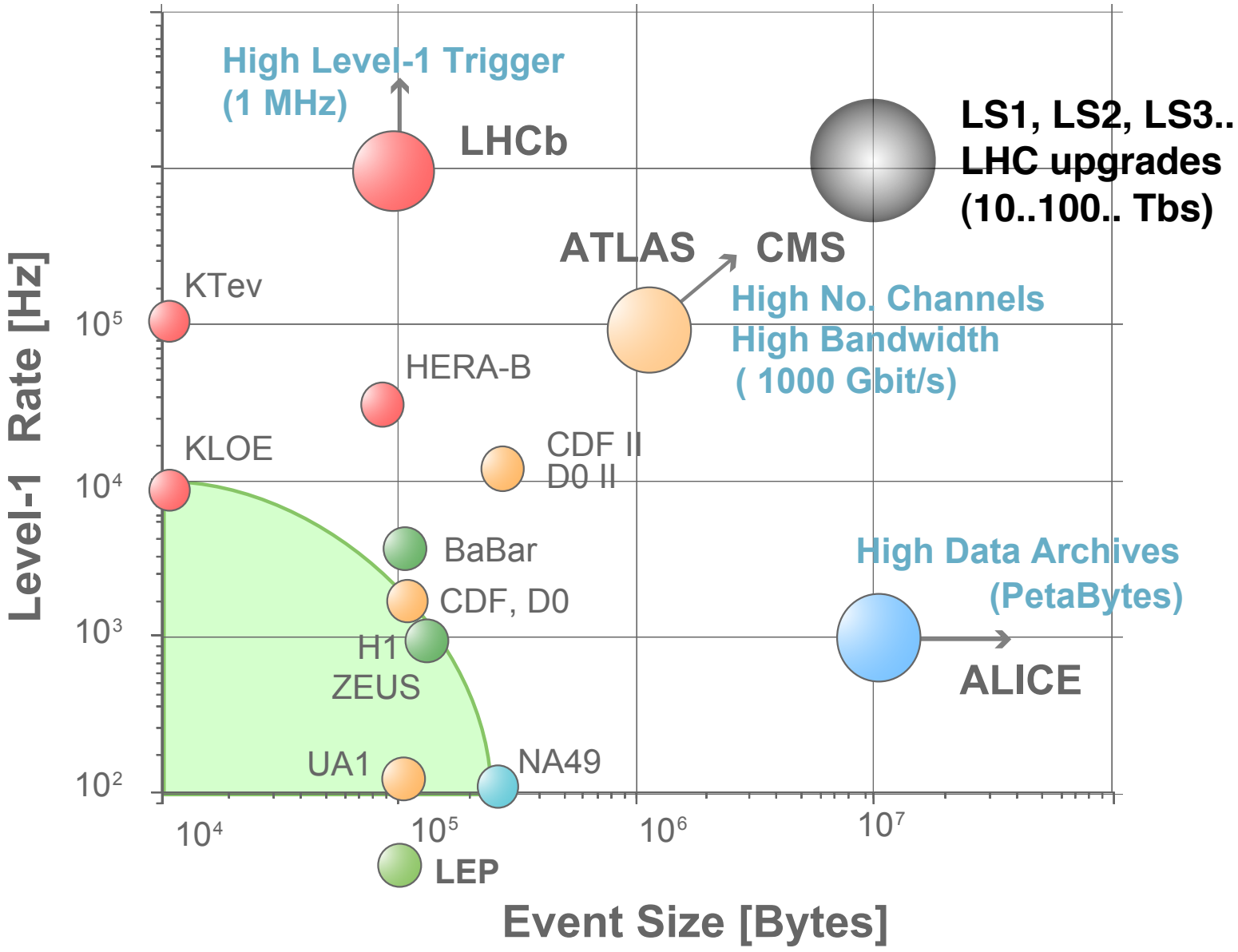


2008 One of Google data center 10^6 cores

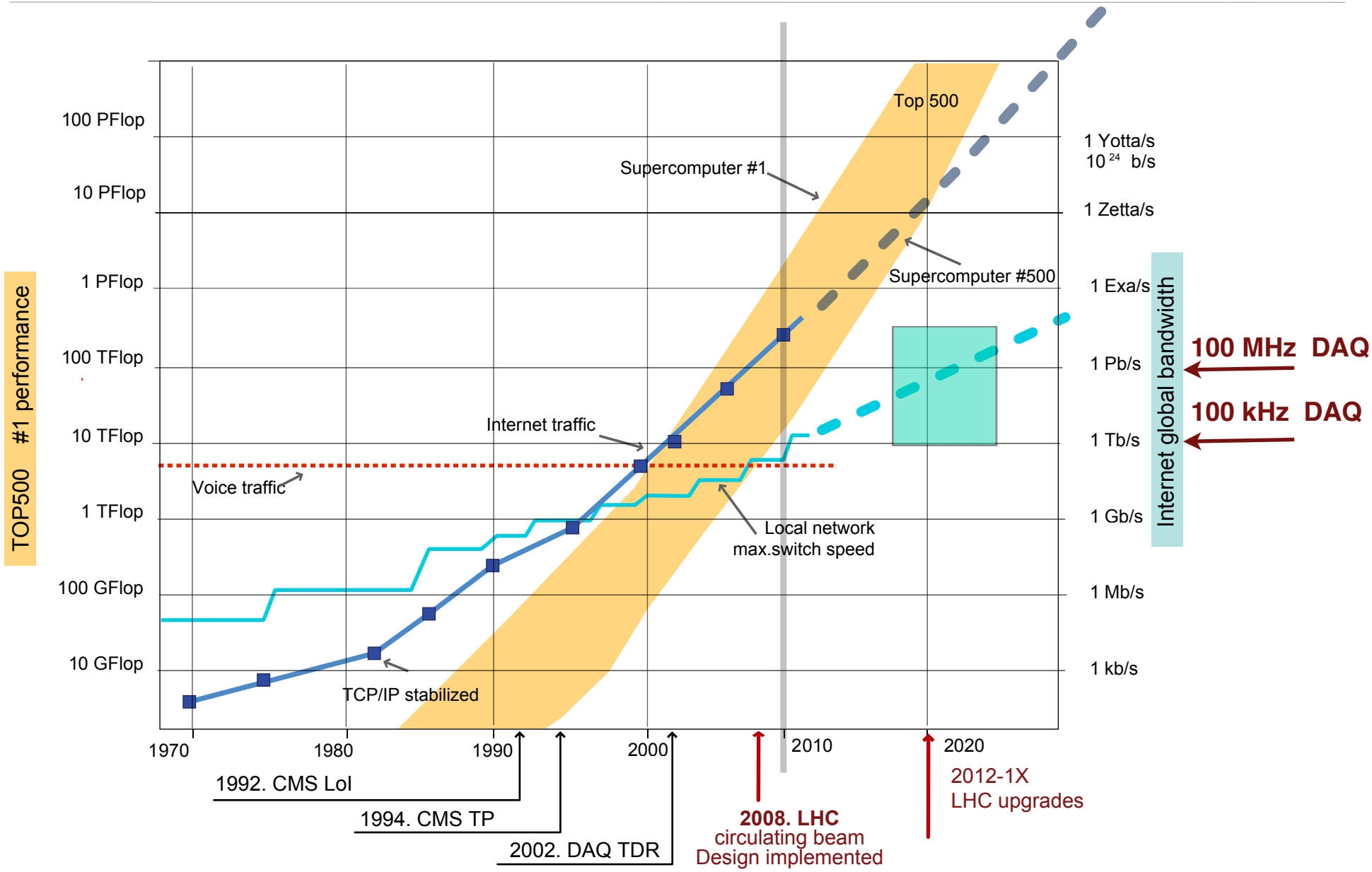
What next, higher:

- Rates, bandwidth, selection power

HEP experiments Level-1 rate / data volume trends



Data communication. Network and Internet traffic trends



TDAQ design guidelines

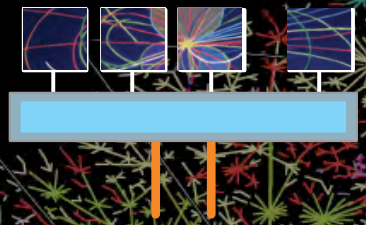
- **Confine custom design to specialized front-end interfaces**
  - Front-end critical electronics, space, features,radiation hard, power consumption
  - Level-1 processors and specialised data links (TTC, analog readout etc.)
  - Readout interface to detector electronics

- **Rely on hardware and software industry standards**
  - Custom/standards (PC clusters, VME, PCI, GBE networks, Web, C++, TCP/IP, SOAP, I2O, slow control industrial infrastructure)

- **Maximally scaling architecture**
  - Exploit **technology evolution**
  - **Scale-free modular system** (simpler controls, error handling, smaller basic units)
  - Cost optimization via **staged installation**

- **Invest in the advance of communication and processing technologies**
  - Computing (**100 kHz Readout, HLT by PC farms**)
  - Communication (**Terabit/s networks, GB/s memories**)

DAQ@WEB services



CLOCK-DRIVEN systems are local (front-end sensors, timing control and first level trigger)

High-level triggers

Analysis and production

Data archive

All EVENT-DRIVEN tasks might be based on Internet hardware and software services.
The required performances are anticipated by the data processing and data communication trends

Controls and monitoring