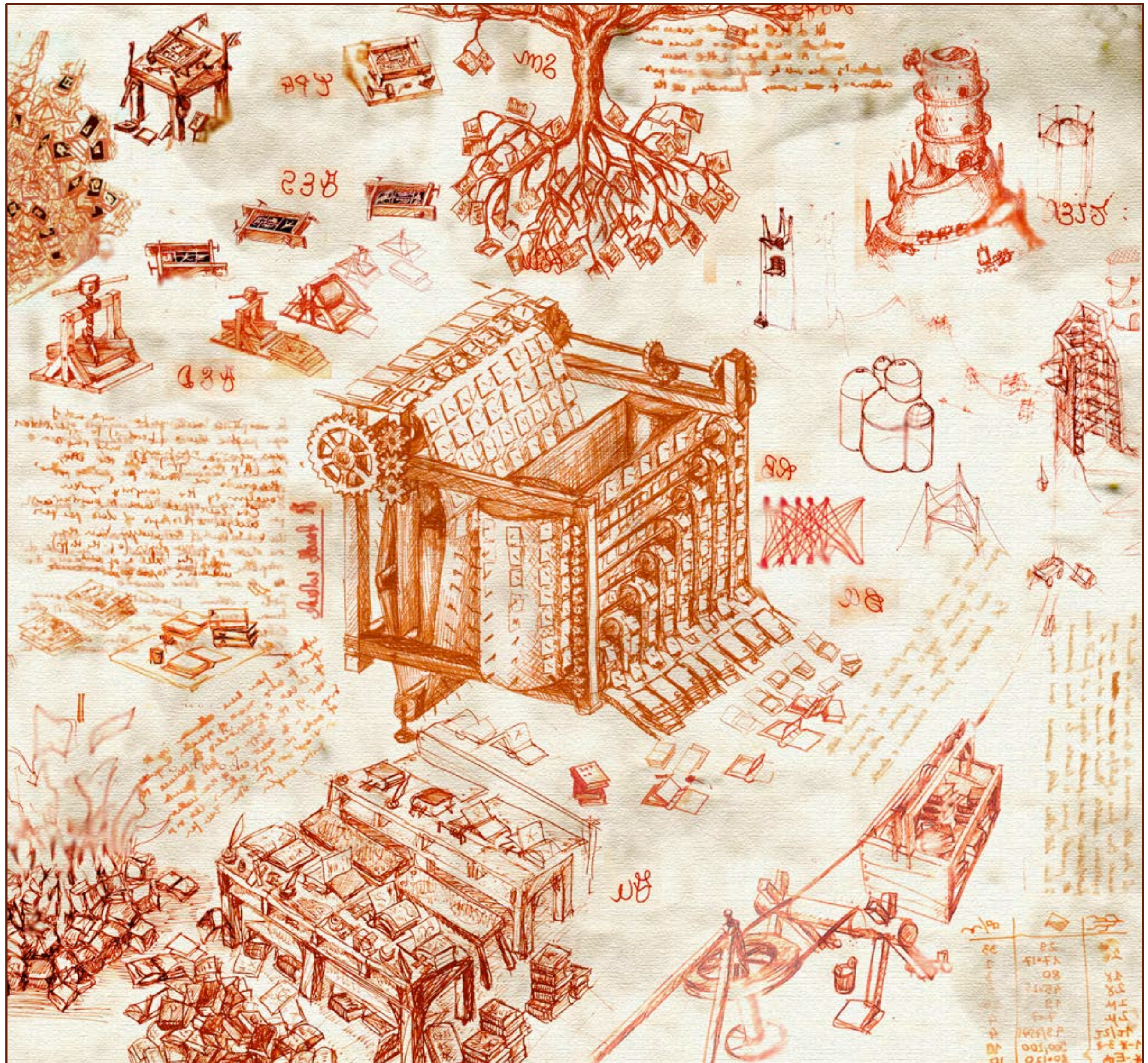
The data acquisition and reduction challenge at LHC

EXCELLENCE IN SCIENCE

Physics at the high energy frontier - the Large Hadron Collider project

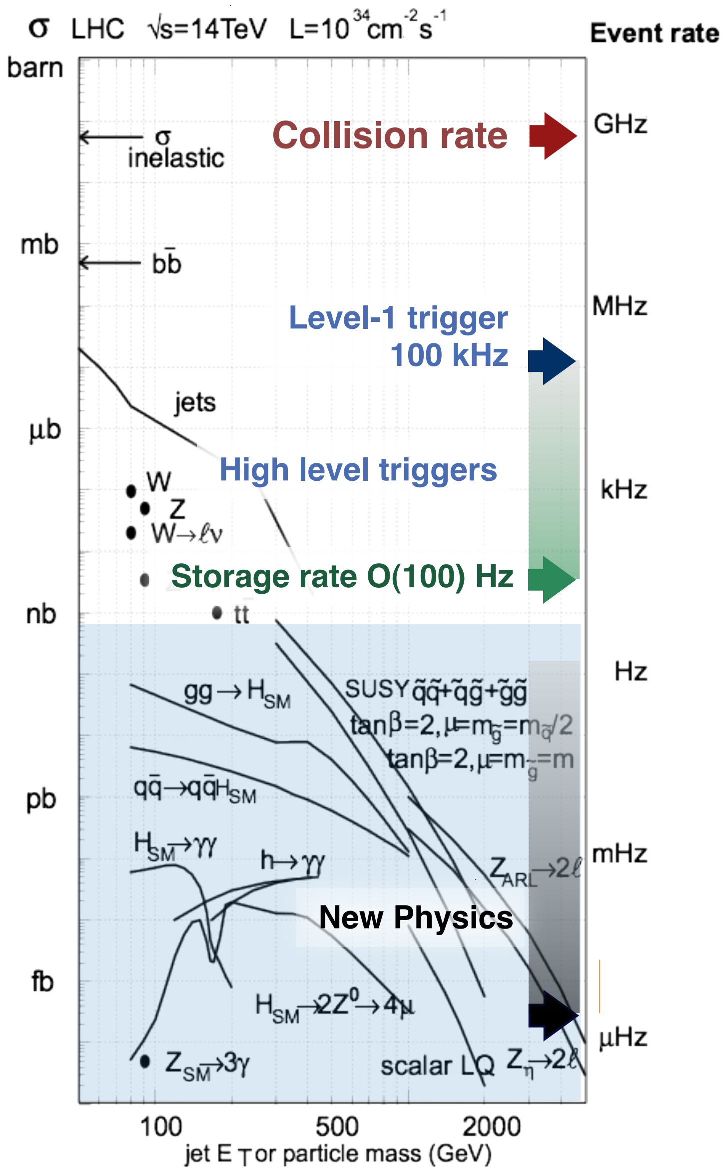
London 17 May 2011

Sergio Cittolin
ETHZ-UCSD-CERN



- **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**

Physics at LHC: overall data handling requirements



At the highest LHC beam intensities:

-**Observe every second** 40 million bunch crossings, each producing several (>20) p-p interactions resulting in events with 1000's particles every bunch crossing

- **Identify** and select single events out of 10 Trillion collisions

- **Locally digitize**, readout, transport and process 100's of TeraBits per second

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

Collision rate	**~ 10^9 Hz**
Detector granularity	**~ 10^8 sensors**
Event size	**~ 1 Mbyte**
Event selection power	**~ 1 in 10^{13}**
Data readout bandwidth	**~ Terabit/s**
Storage event rate	**~ O(100Hz)**
On/Off-line Processing power	**~ TeraFlops**

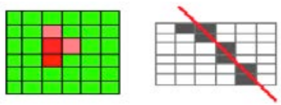
Design issues: Architectures

- Data flow, Front-end, Event selection levels

On-line data communication and processing at LHC

Crossing rate: 40 MHz
Collision rate: 1 GHz

Raw data production
100's of Petabits/s



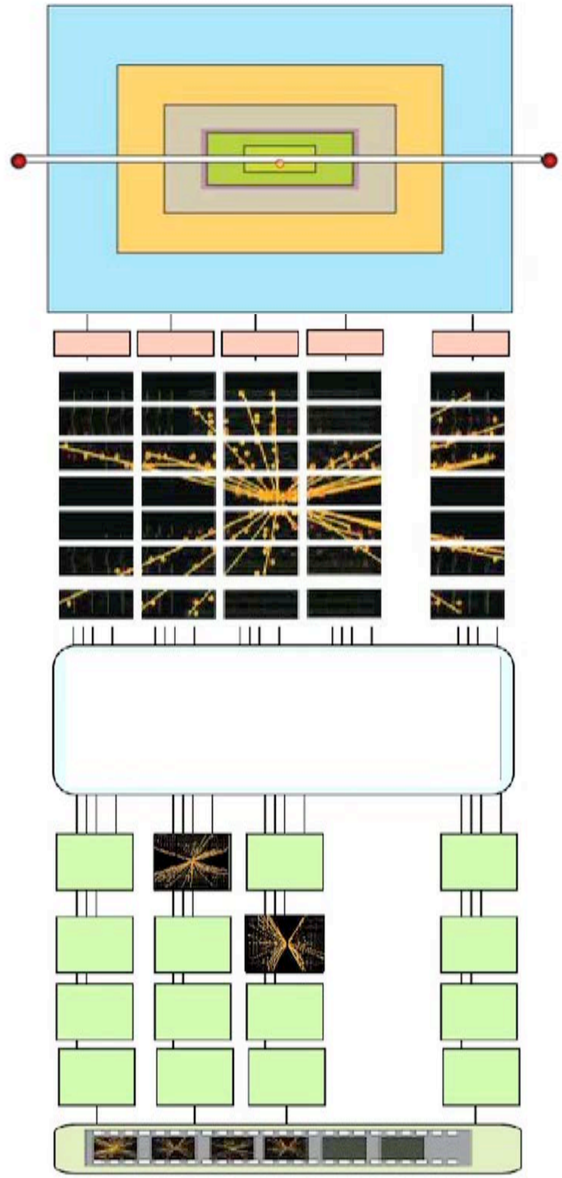
Level-1 Trigger

100 kHz output rate
50 Million fragments/s

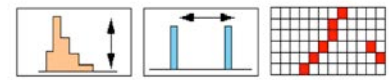
Readout network
2 Terabit/s

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

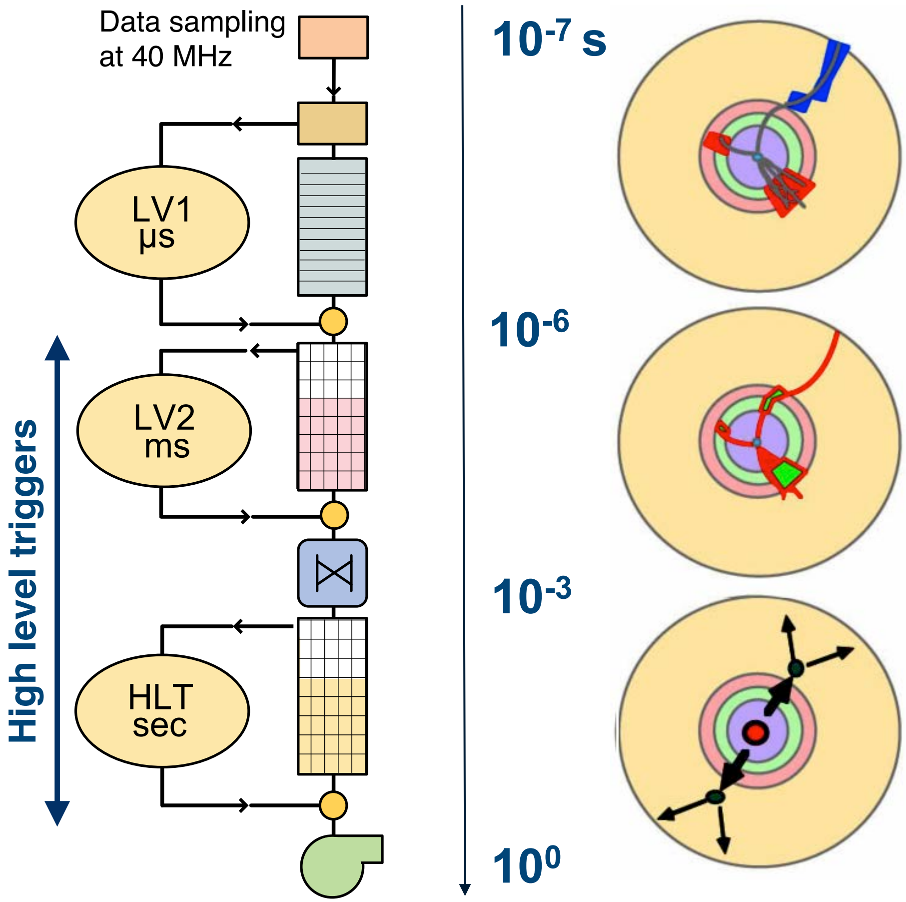
Event Builder
Switching system with thousands of ports
~1 Megabyte data per event

Event Filter
Thousands of CPU cores

Local mass storage
Hundreds of Terabits

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

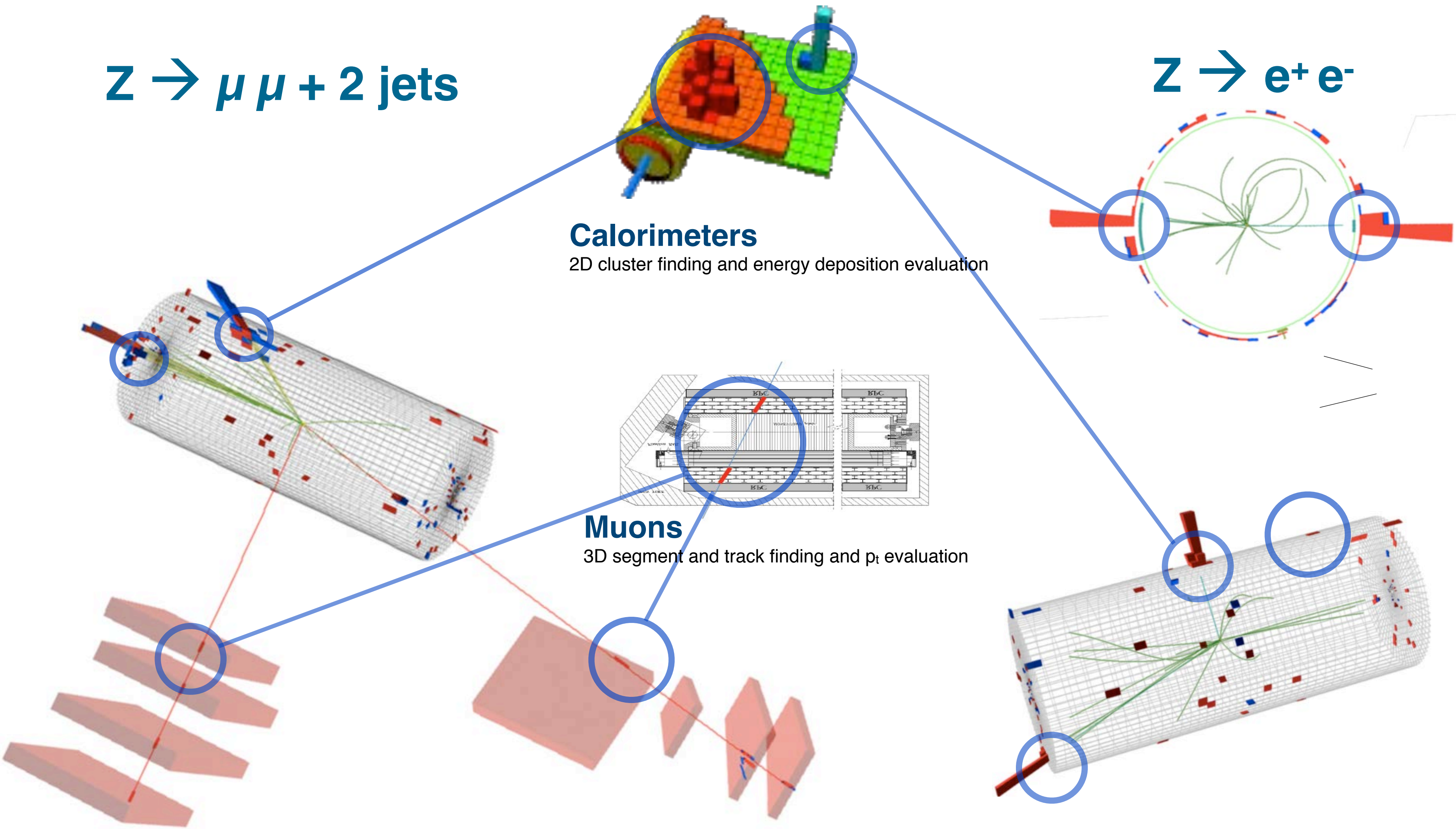
- Event reconstruction and analysis

On-line global selection:

99.999 % rejected, 0.001 % selected

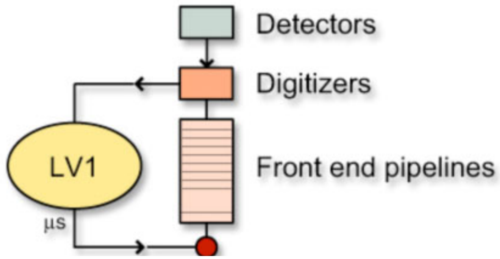
Readout and trigger dead-time must be kept at minimum (typically of the order of few %)
The trigger system has to maximise the collection of data for physics process of interest at all levels, since **rejected events are lost for ever**

Level-1: 2D-clusters and 3D-segments finder logic



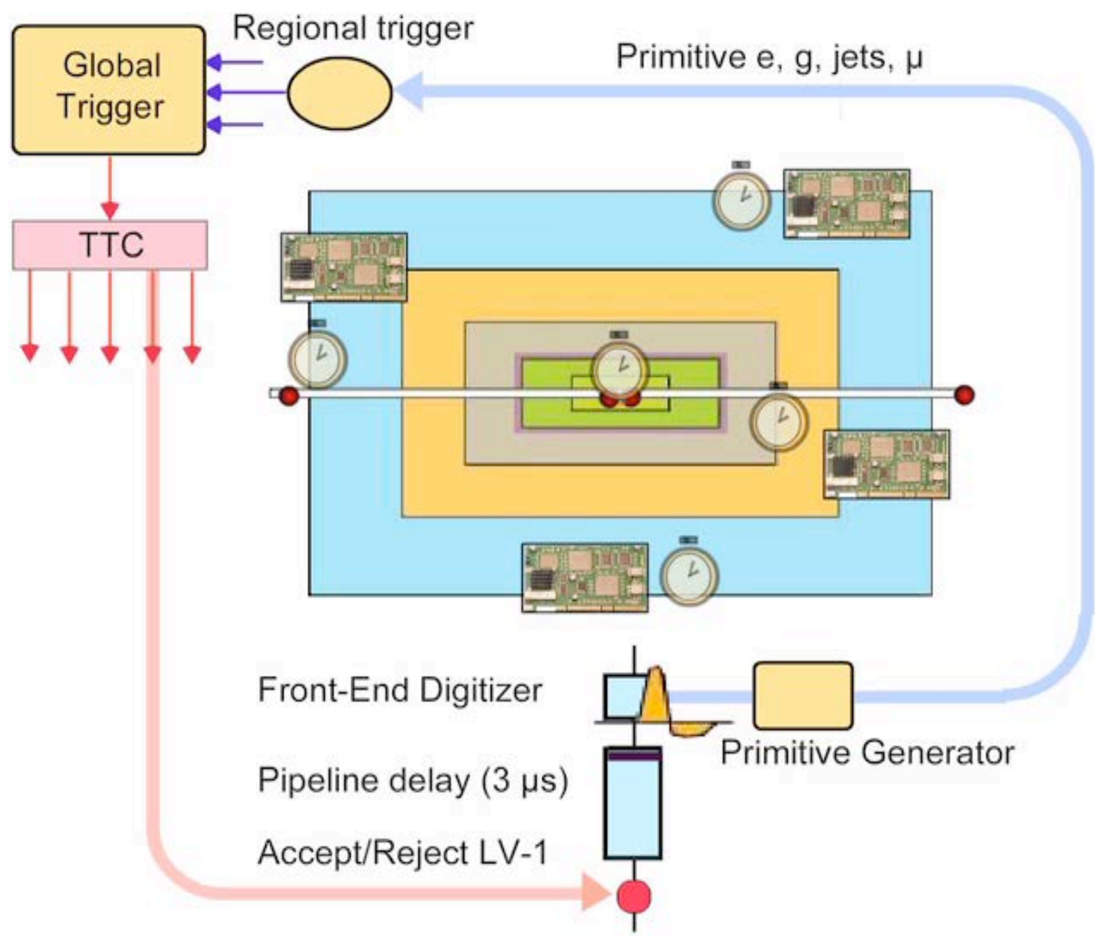
Algorithms run on local **calorimeter and muon coarse data**. With **new data every 25 ns and decision latency ~μs.** Special-purpose hardware reduces event rate (to be read out) **from 40 MHz to 100 kHz.**

Front-end timing and readout 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)

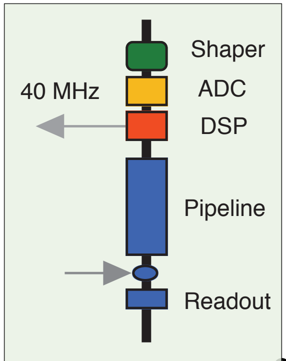
The **Level-1** trigger is a centralized multi-step logic system collecting, buffering and processing sub-set of detector data every 25 ns.



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

Front-end time budget (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**

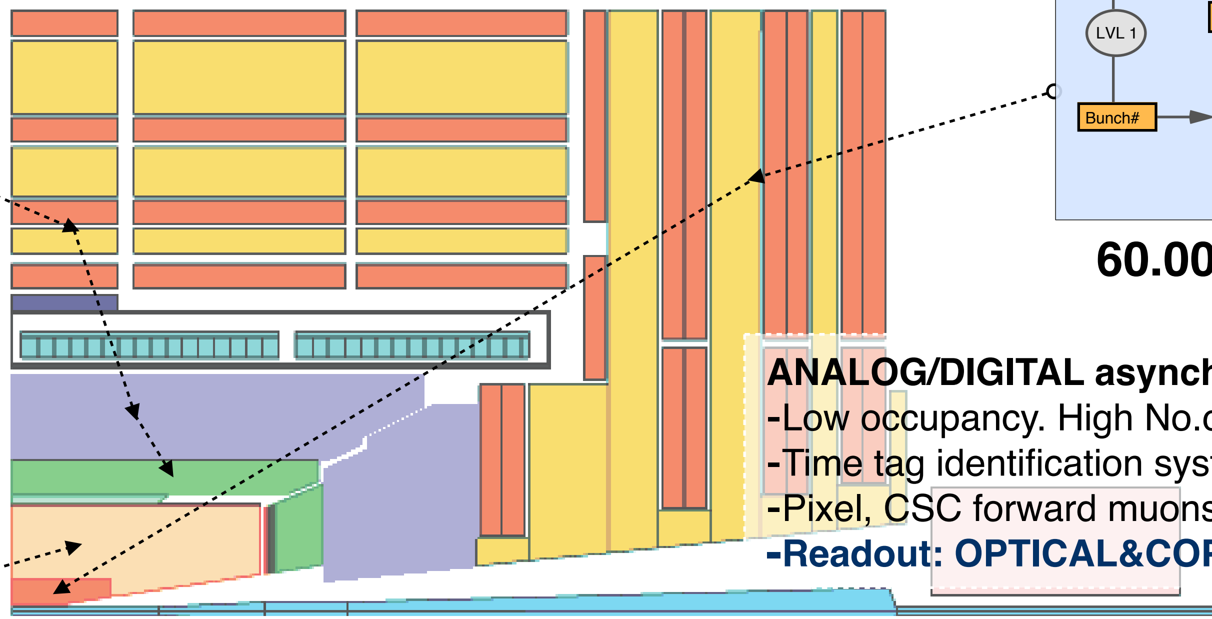
Analog/Digital front-end systems (CMS)

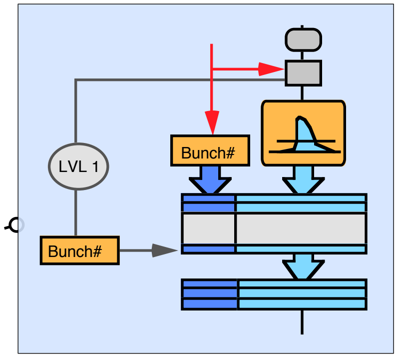


500.000 ch.

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

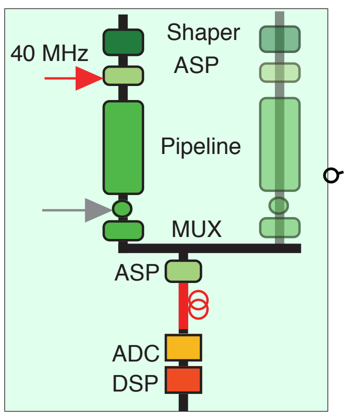




60.000.000 ch.

ANALOG/DIGITAL asynchronous buffer

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



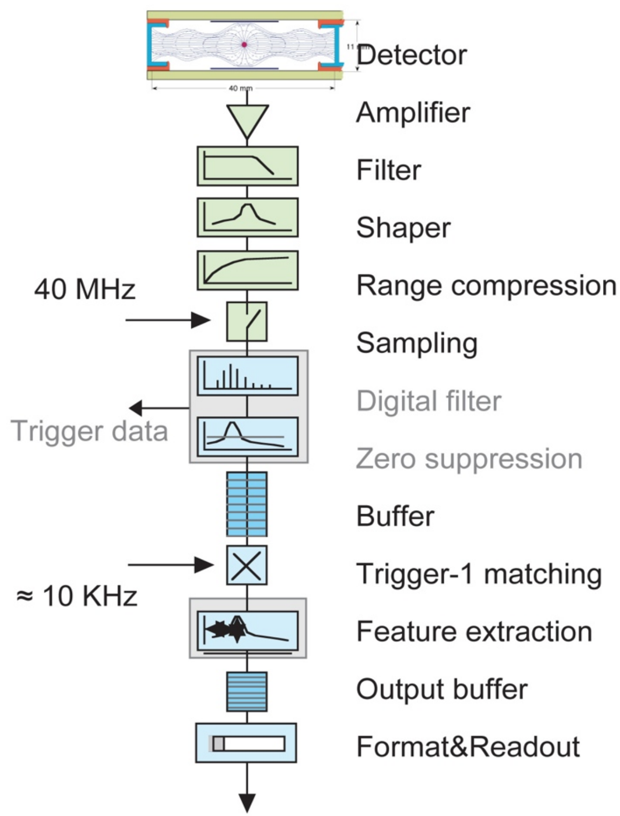
10.000.000 ch.

ANALOG (25 ns) synchronous pipeline

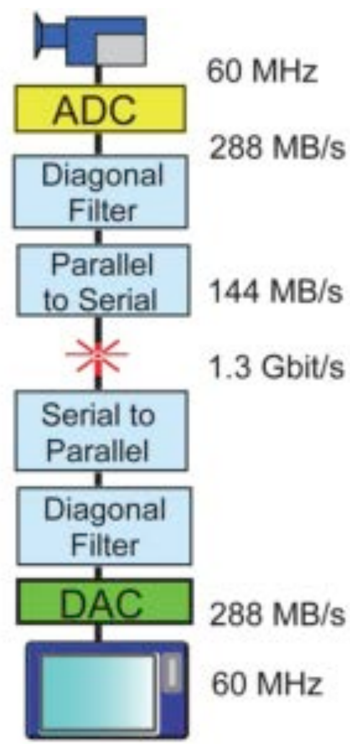
- Low dynamic range and resolution($\leq$ 10-bits)
- Low power consumption, radiation tolerant,
- High number of channels.Tracker, Pre-shower
- **Readout: OPTICAL analog** links

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.

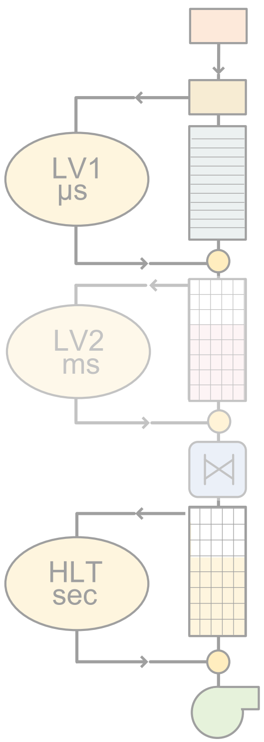
- The telecommunication industry has been one of the major customers for the development of this technology.

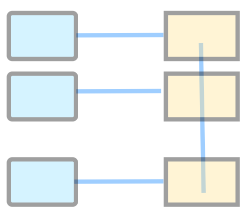
- Analog to digital converters (**ADC**)
- Multiply accumulator (**MAC**)
- GHz **optical links** and Laser LED
- Finite Impulse Response (**FIR**) digital filters and vector processing are today the **building blocks of any LHC detector readout chain** as well.

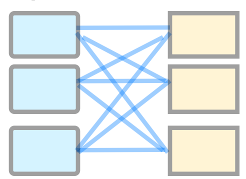
Triggering: architectures and technologies

Level-1 trigger architecture. Massively parallel: One event -> Multi-processors
- High (fixed interconnections), Short (fixed) latency. Pipelined simple ALUs, **Clock driven**

MEMORY (I/O) access

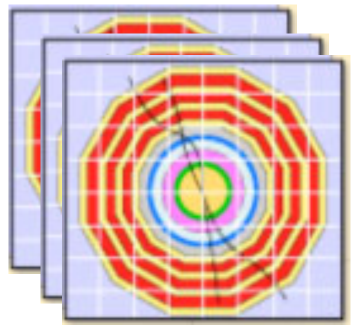




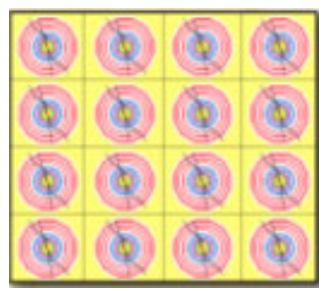


DATA PROCESSING

Synchronous pipelined

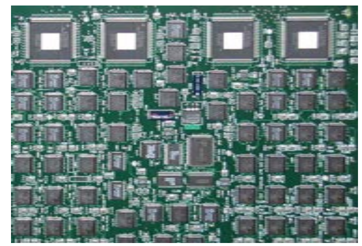


Asynchronous clusters



IMPLEMENTATION

Custom design
(ASIC, FPGA, LVDS..)



Commodities
PCs and networks

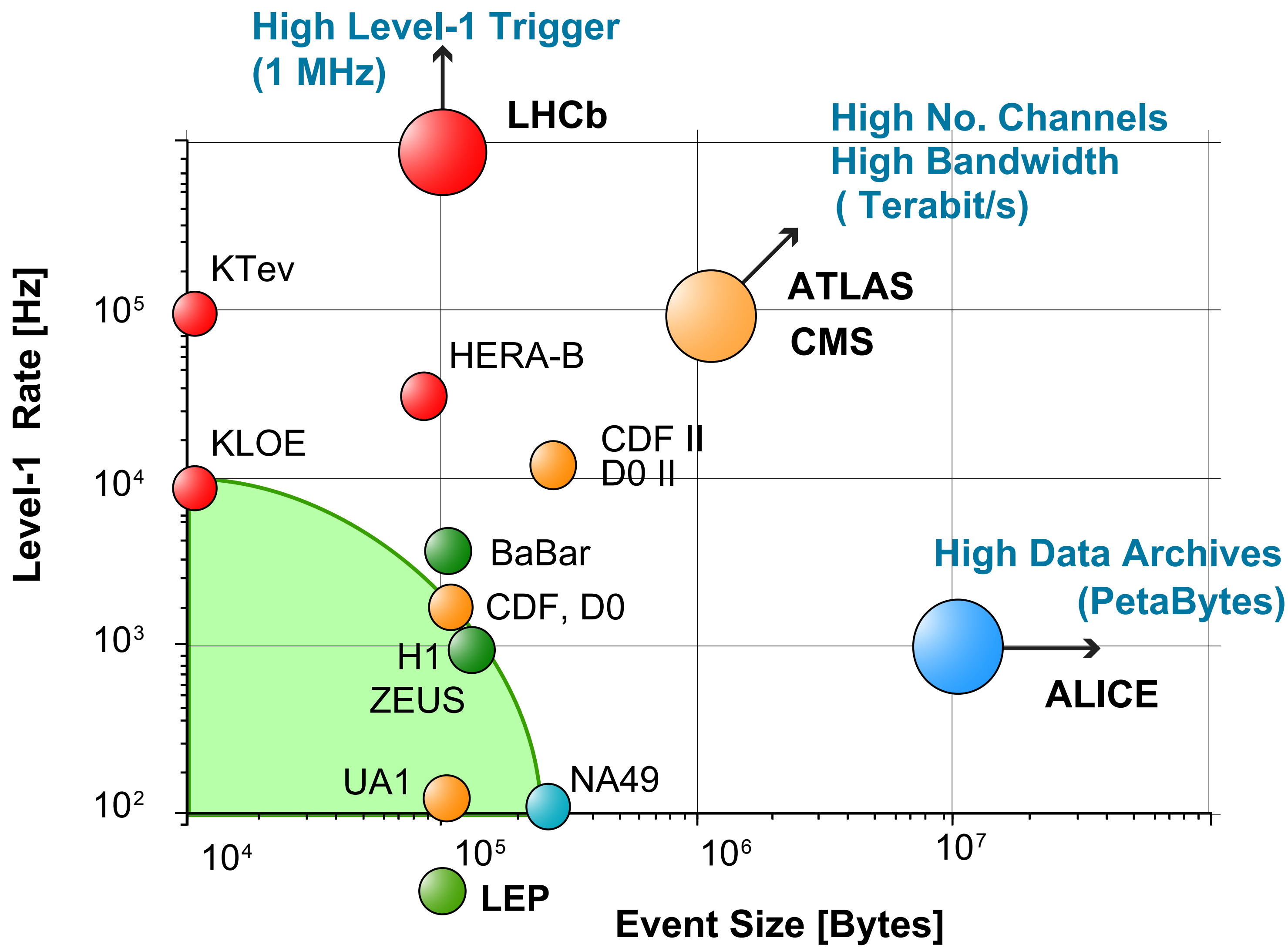
INTERCONNECTION

Distributed

Decentralised

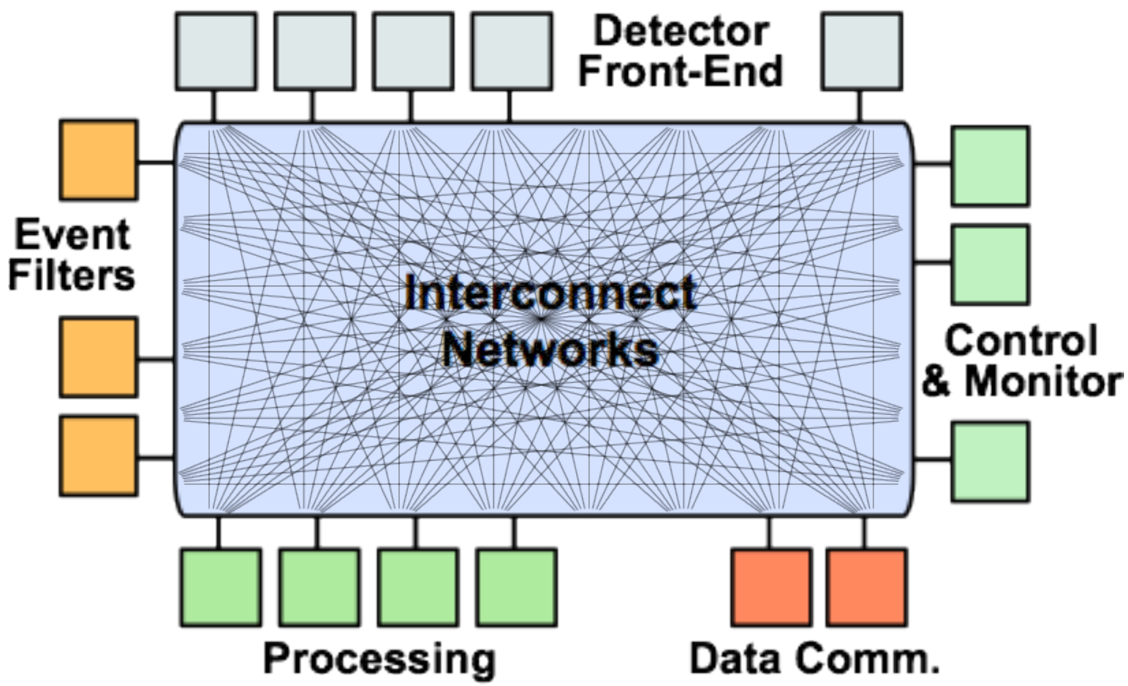
High Level Triggers architecture. Cluster structure: One event -> One processor
- Loose coupling, large latency. Node high power, **Event driven**

HEP experiments Level-1 rate / data volume trends



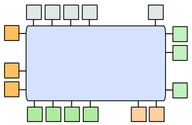
2000's On&Off-line processing and communication model

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



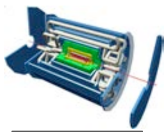
At the time of the finalization of the system design (2002-03), **a single network technology could not satisfy at once all the LHC requirements. The LHC DAQ designs had to adopt multiple specialized networks instead.**

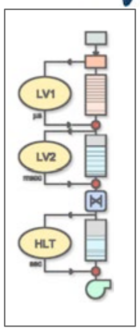
TDAQ data network designs

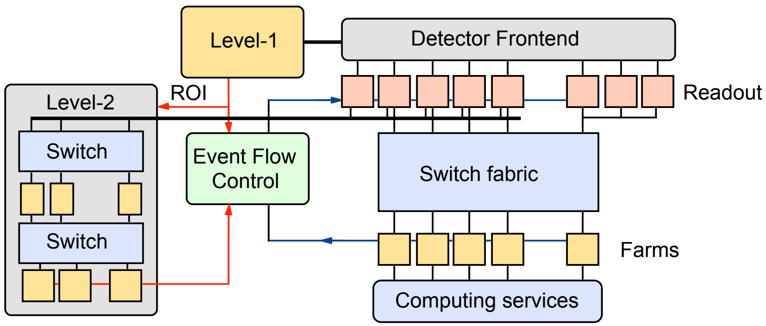


Each LHC experiment developed its own scheme to cut the rate, to process events online and/or optimize the throughput. In a sense, the **systems designed and built are "approximations" of the basic architecture/conceptual design**

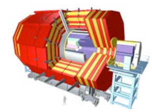
ATLAS

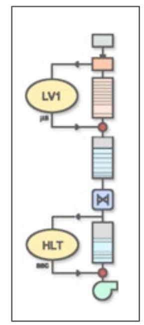


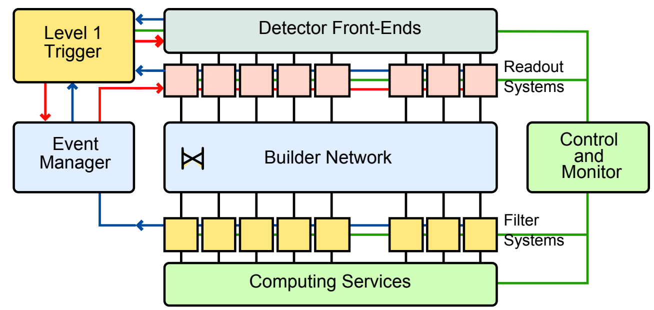




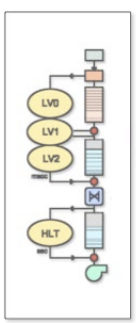
CMS

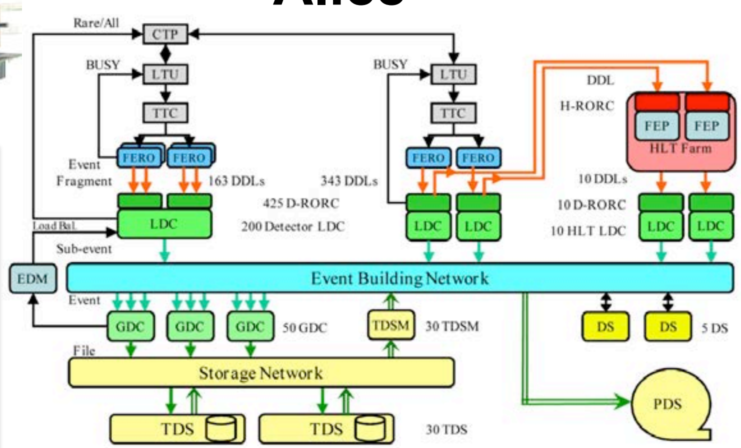




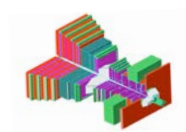


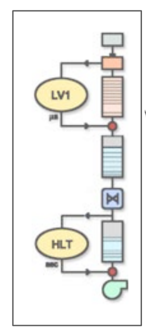
Alice

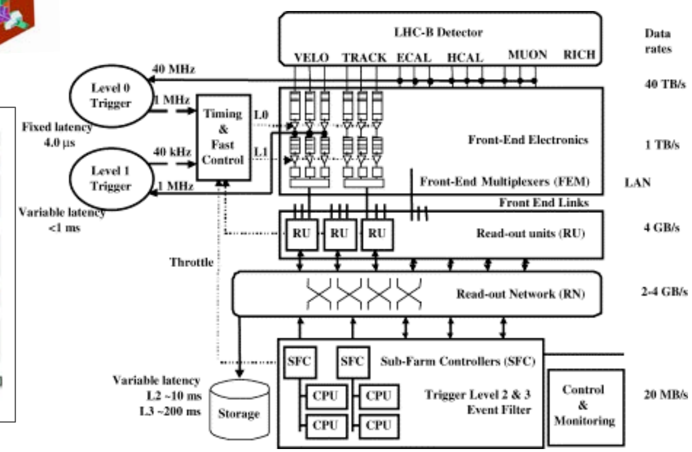




LHCb





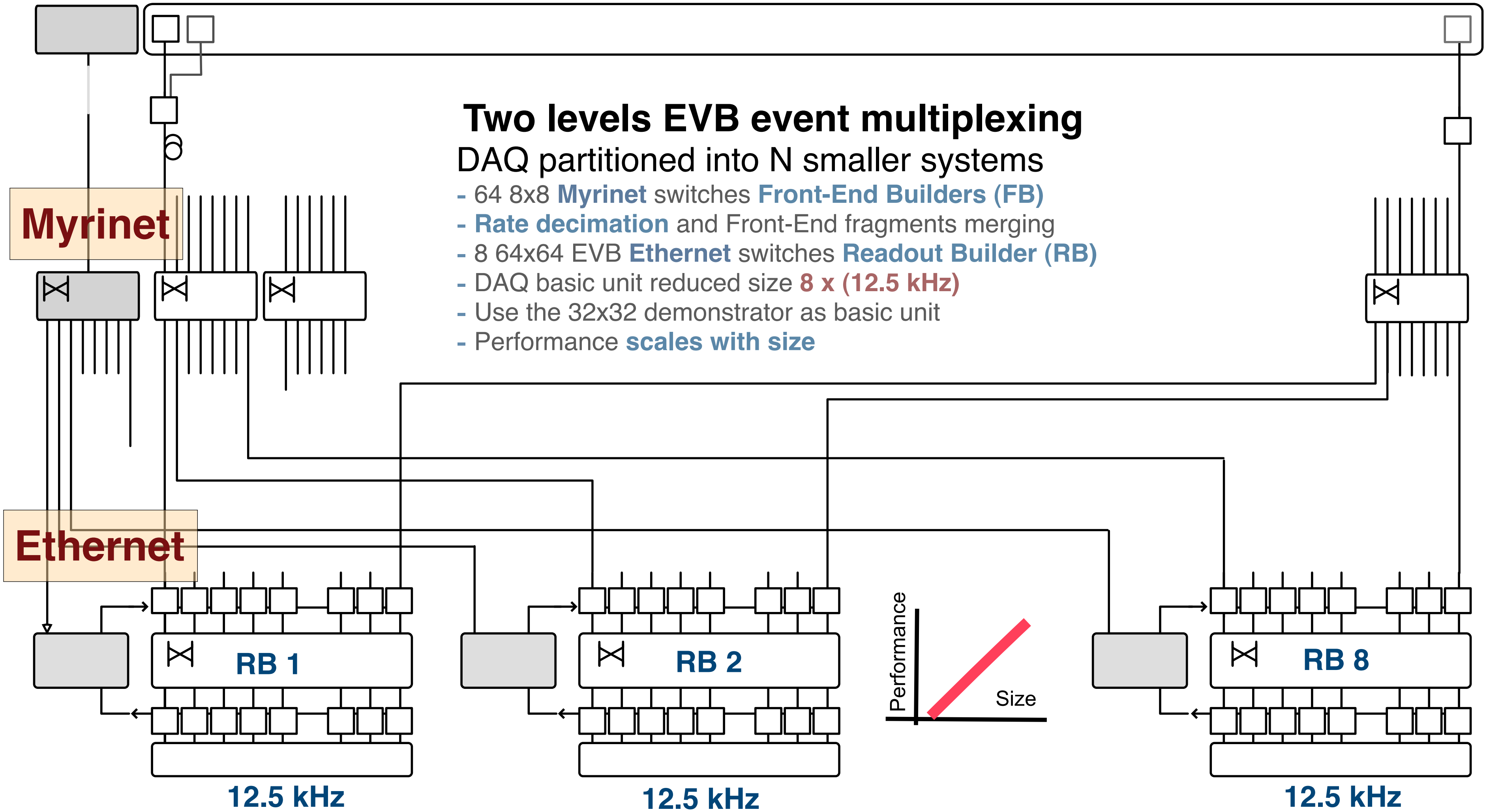


Scale-free 8-fold DAQ system

Two levels EVB event multiplexing

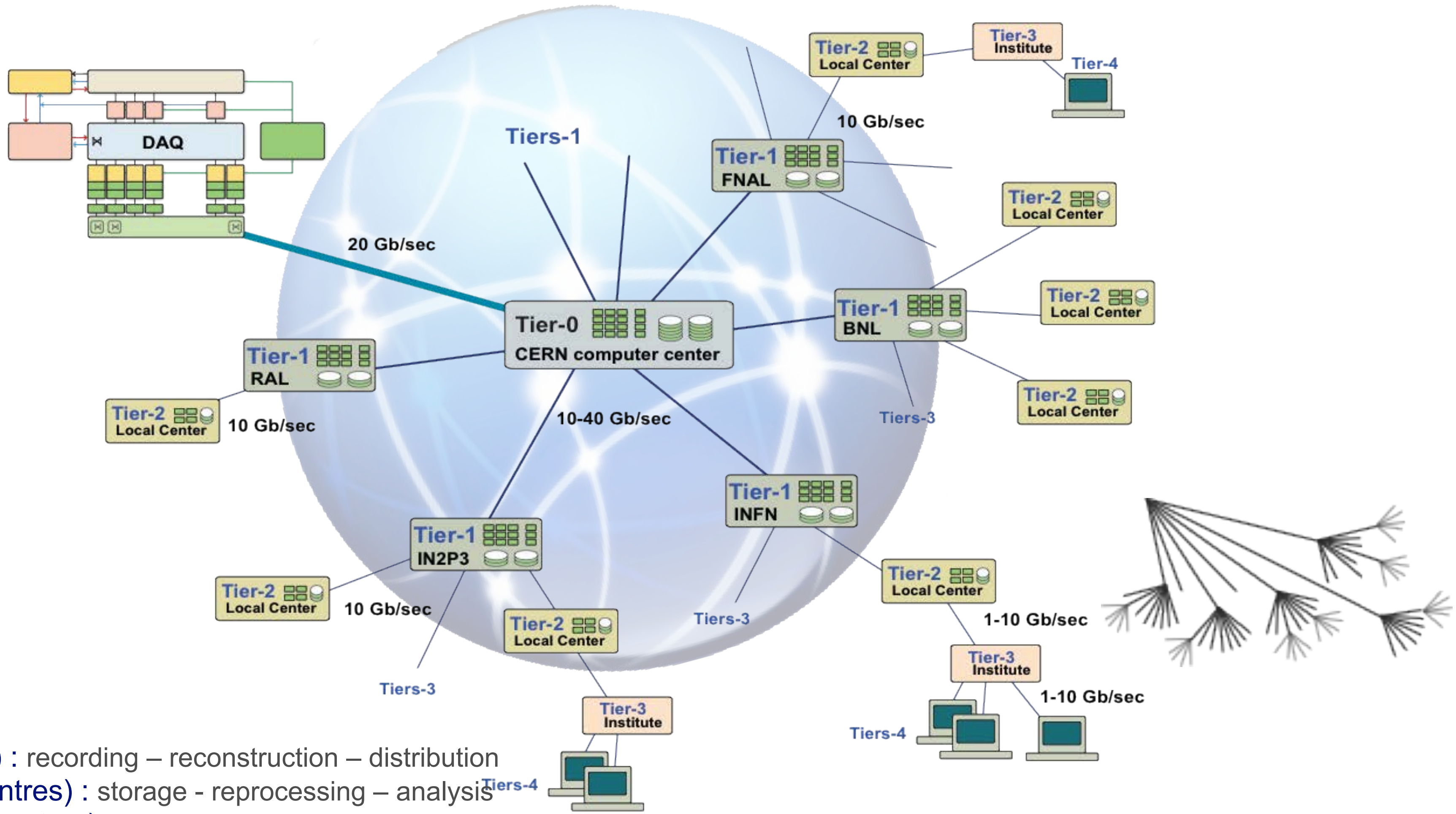
DAQ partitioned into N smaller systems

- 64 8x8 **Myrinet** switches **Front-End Builders (FB)**
- **Rate decimation** and Front-End fragments merging
- 8 64x64 EVB **Ethernet** switches **Readout Builder (RB)**
- DAQ basic unit reduced size **8 x (12.5 kHz)**
- Use the 32x32 demonstrator as basic unit
- Performance **scales with size**



Off-line computing by HEP centers and Internet

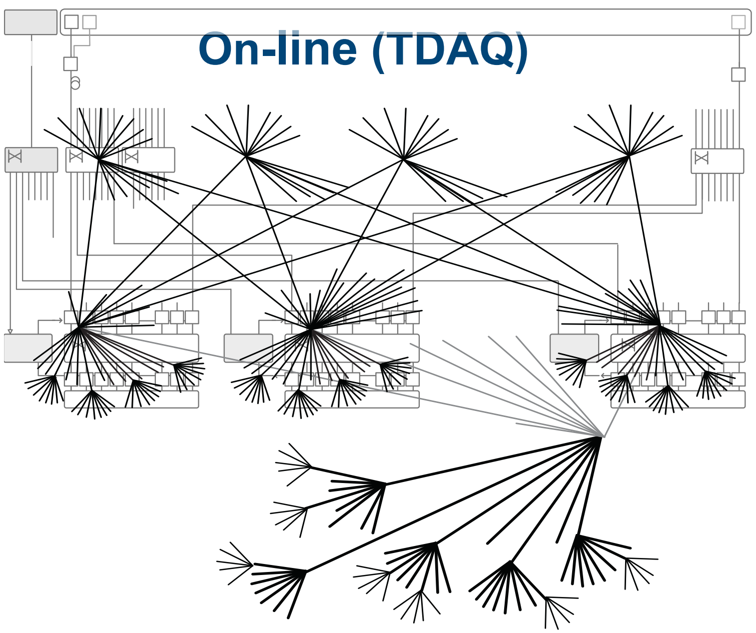
The GRID. A distributed computing infrastructure (~150 kCores, 50 PB disk), uniting resources of HEP institutes around the world to provide seamless access to CPU and storage for the LHC experiments. A common solution for an unprecedented demand (in HEP) of computing power for physics analysis.

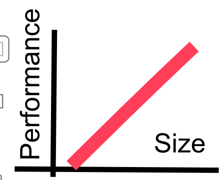


Tier-0 (CERN) : recording – reconstruction – distribution
Tier-1 (~10 centres) : storage - reprocessing – analysis
Tier-2 (~140 centres) : simulation – end-user analysis

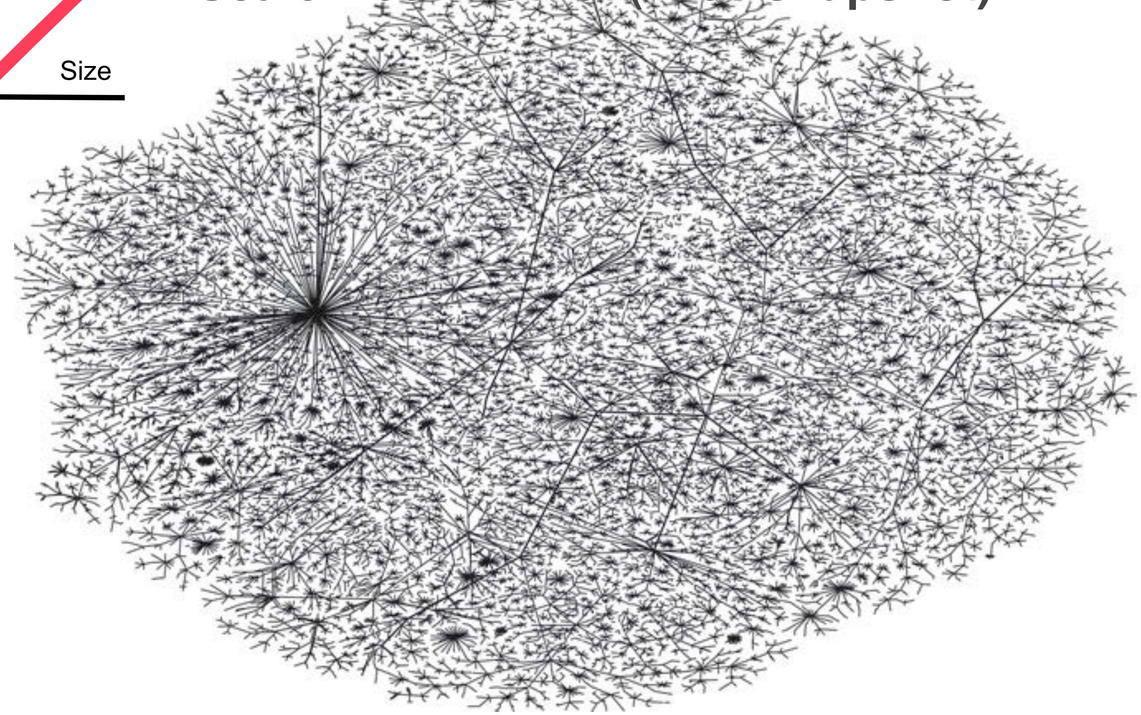
Scale-free networks

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.





Scale-free internet (2002 snapshot)



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

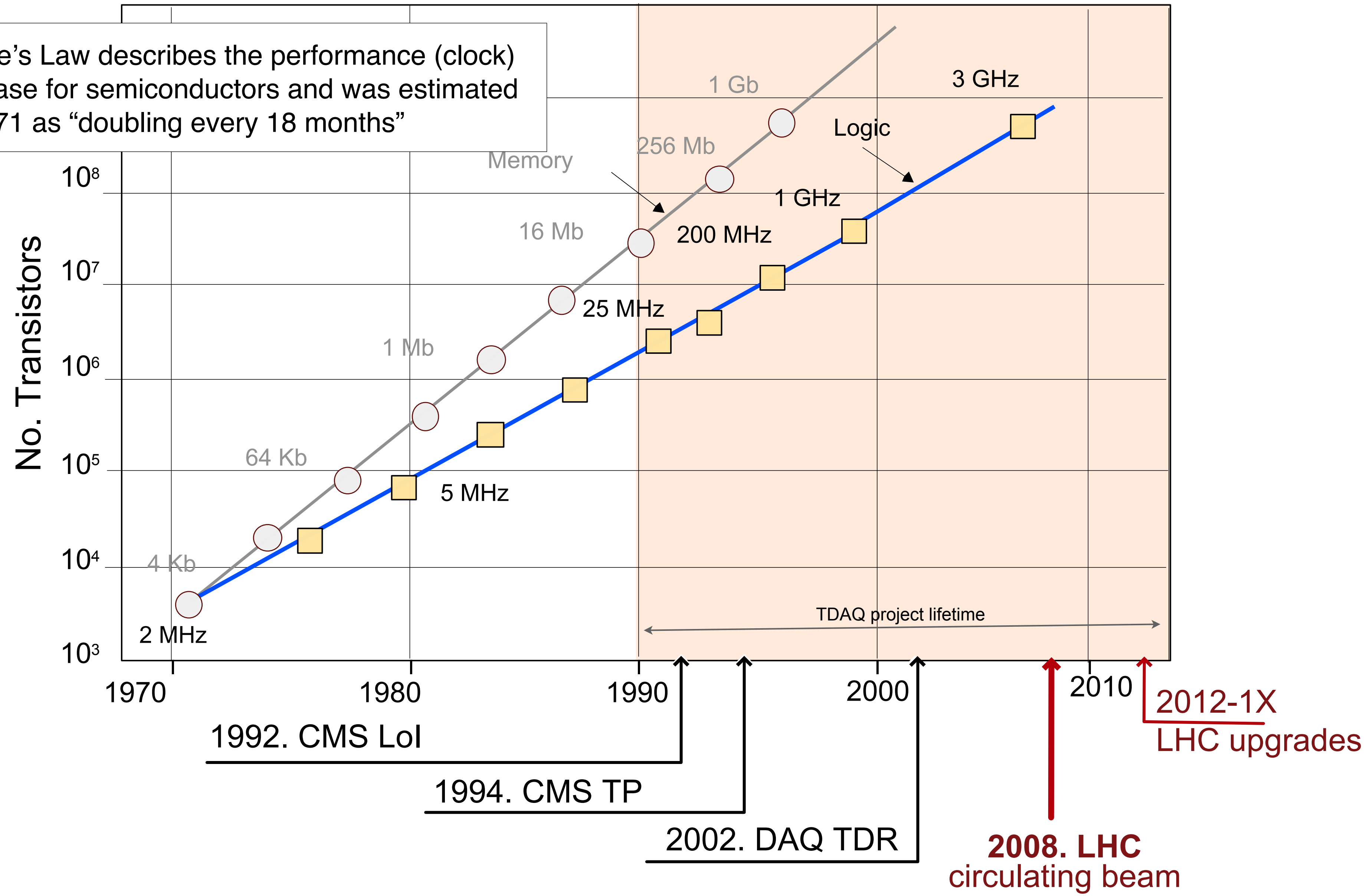
1996-2002
Prototypes and Demonstrators

2003-2005
Final Design. Choice of technologies

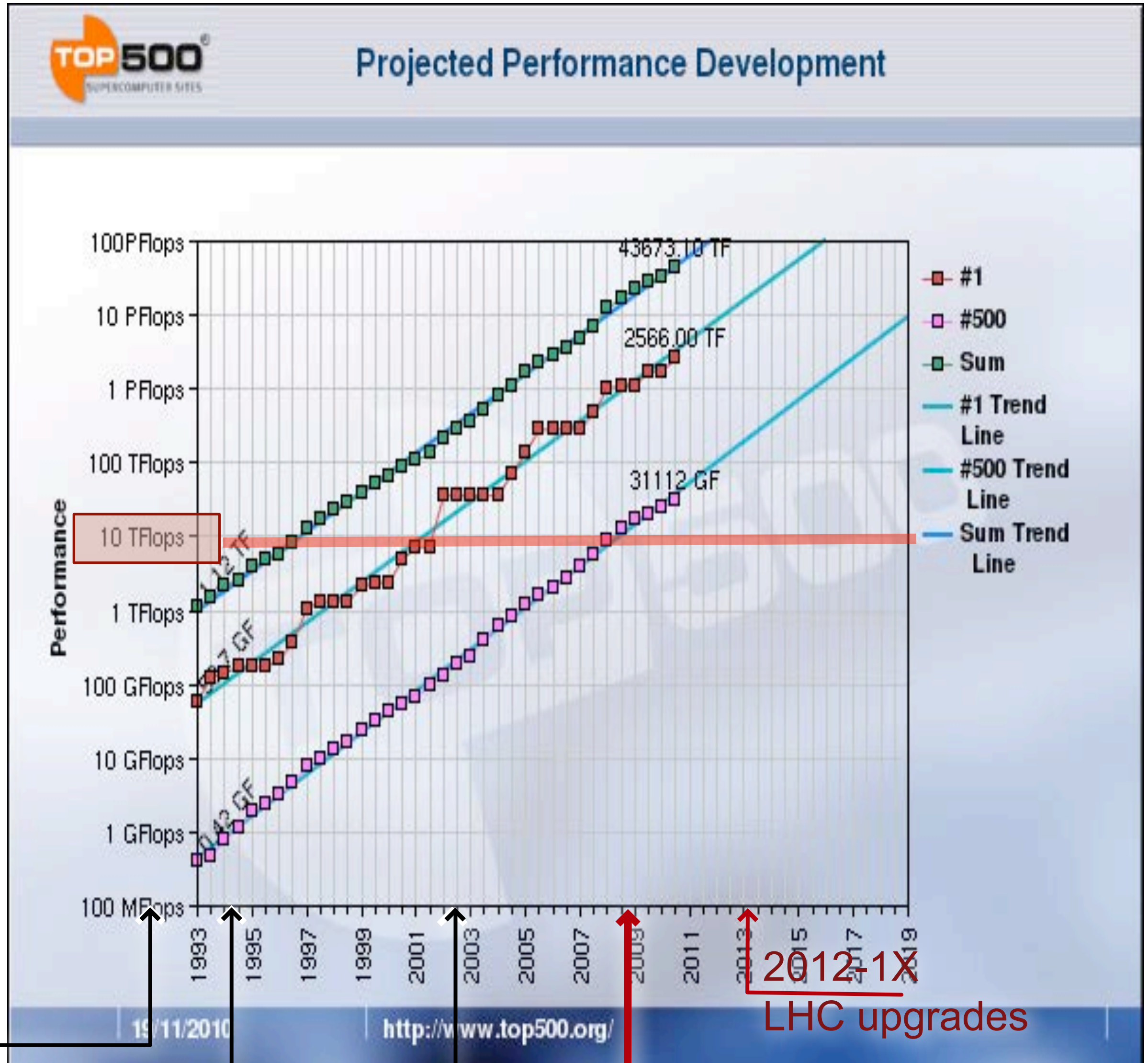
2006-2008
Construction and commissioning

Semiconductor industry trends

Moore's Law describes the performance (clock) increase for semiconductors and was estimated in 1971 as "doubling every 18 months"



Computing power



1992. CMS LoI

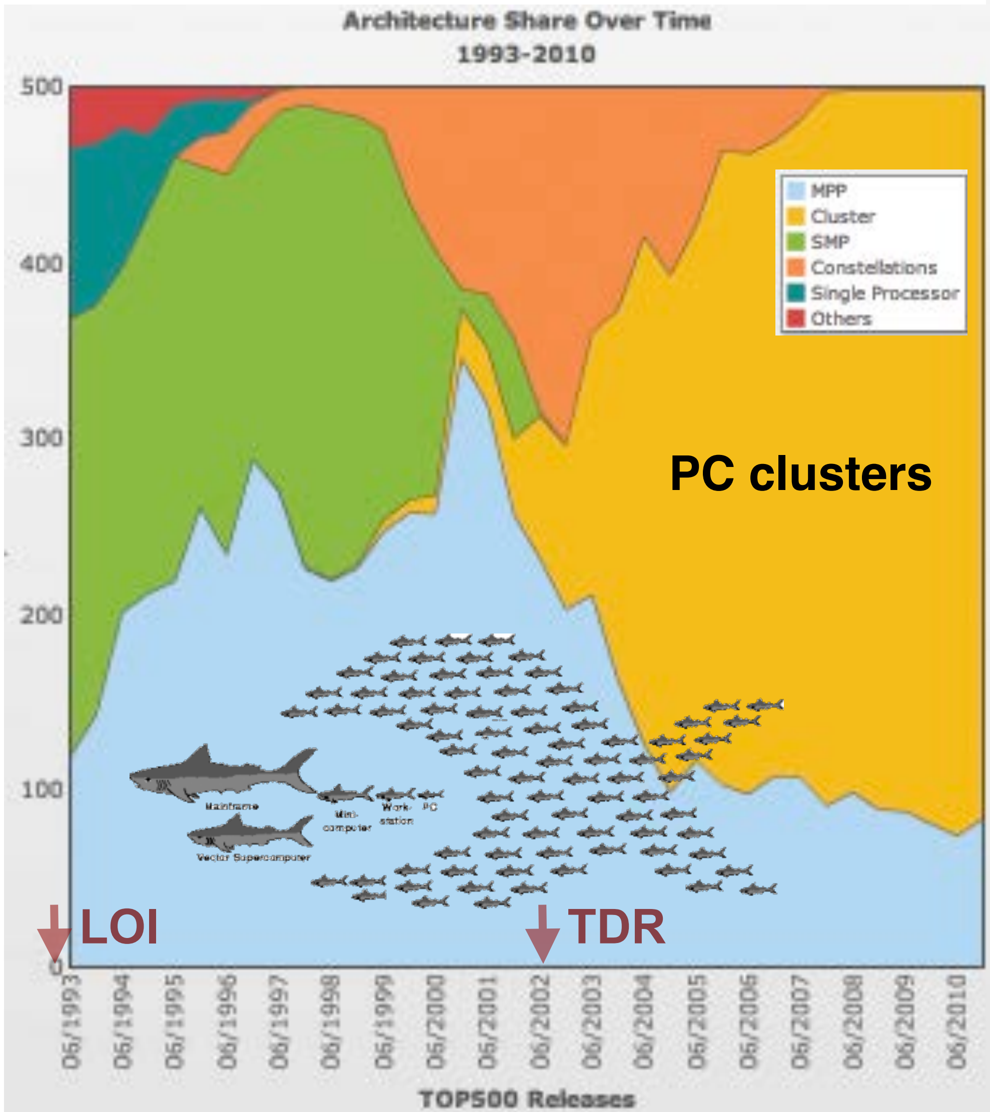
1994. CMS TP

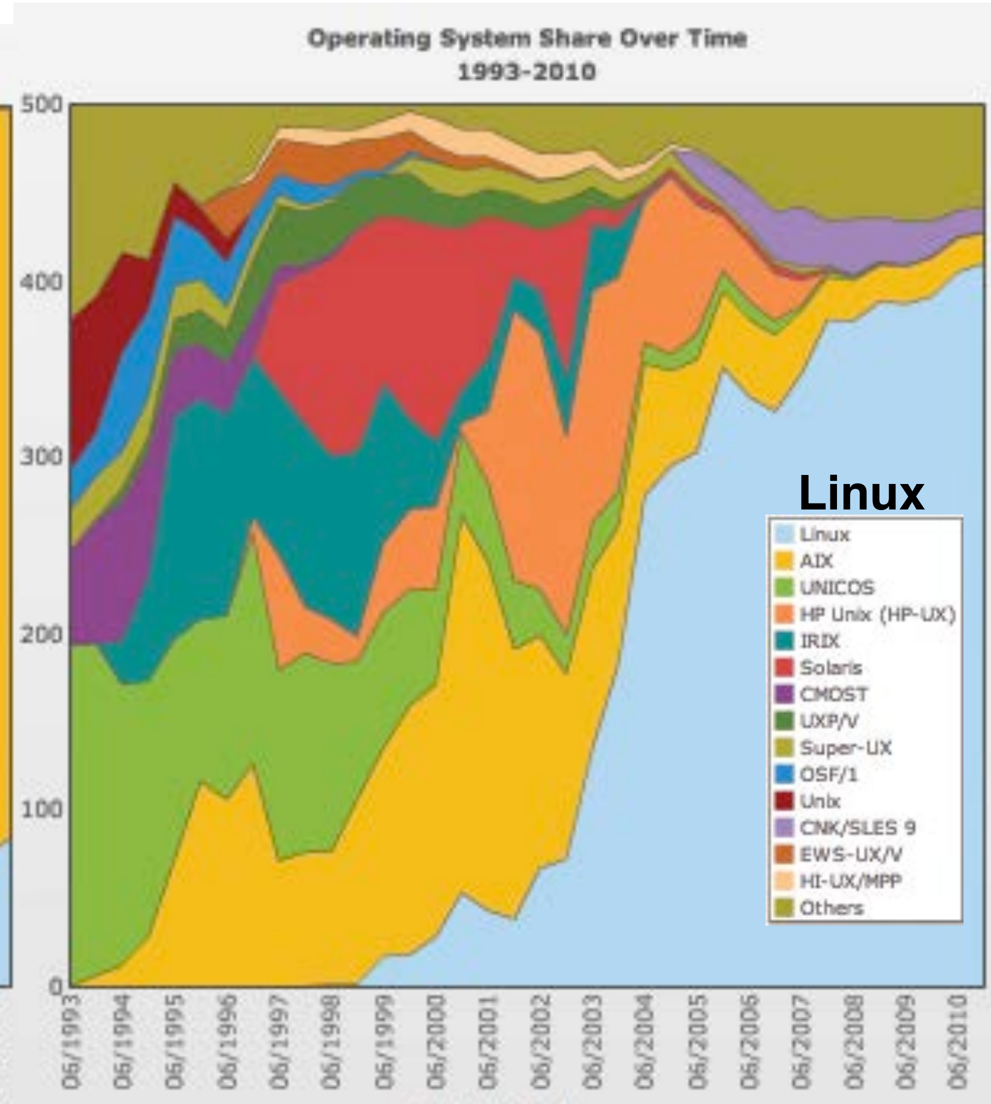
2002. DAQ TDR

2008. LHC circulating beam

2012-1X LHC upgrades

Top500 supercomputer architectures and operating systems





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.

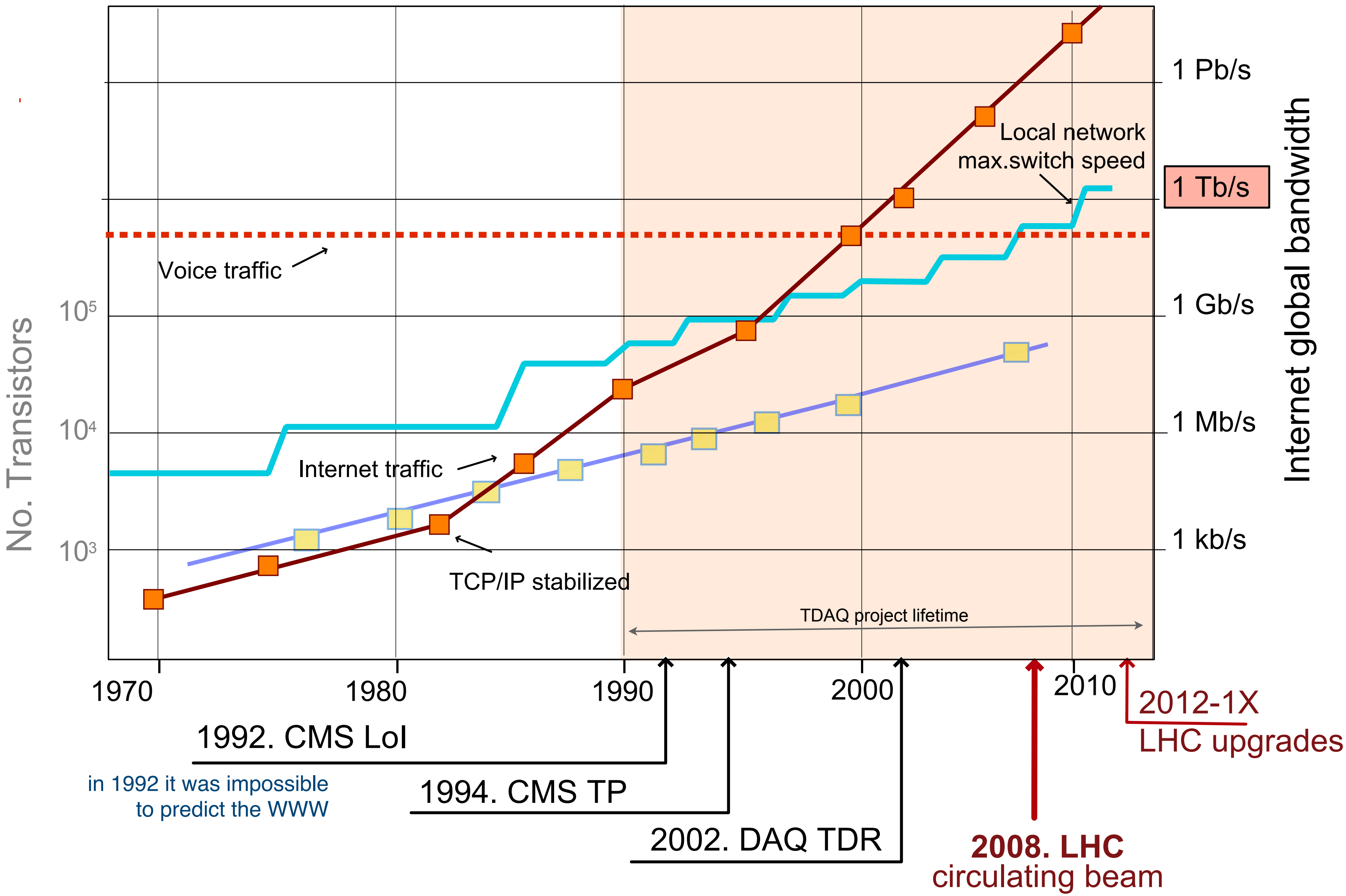
Since 90's:

- 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)

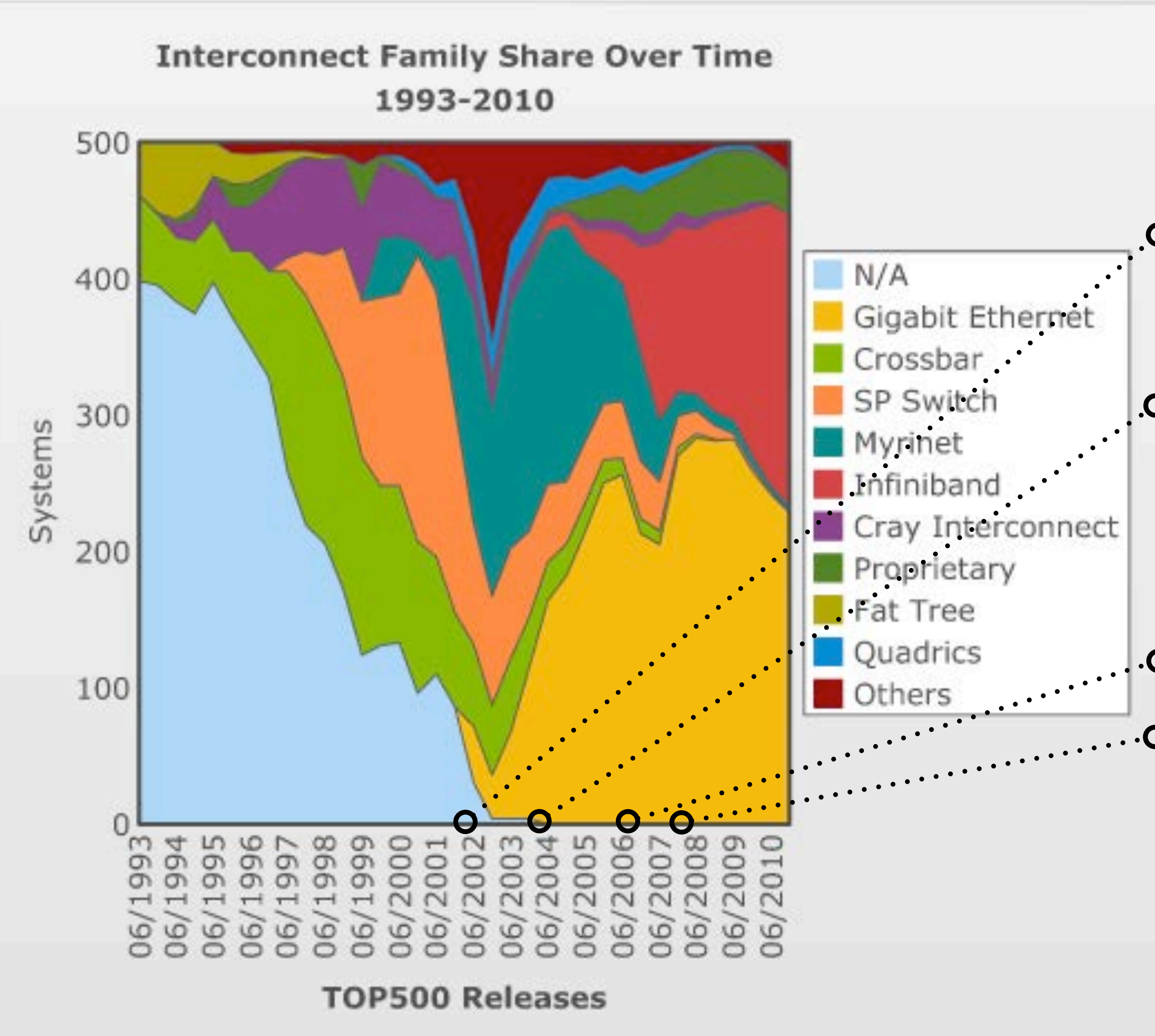
CPU implemented in 2008. The 50% of the HLT system integrated in 2008 consisting of 5000 2.6 GHz CPUs (720 PCs of two quad-core) corresponds to about **10 TFLOPs** in line with the foreseen requirements and in agreement with the Moore law of integrated logic systems (corresponding to a factor 10 in speed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

Data communication. Network and Internet traffic trends



in 1992 it was impossible
to predict the WWW

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

Unexpected

- Collaborative work
- Network&Computing fusion

Experiment control rooms

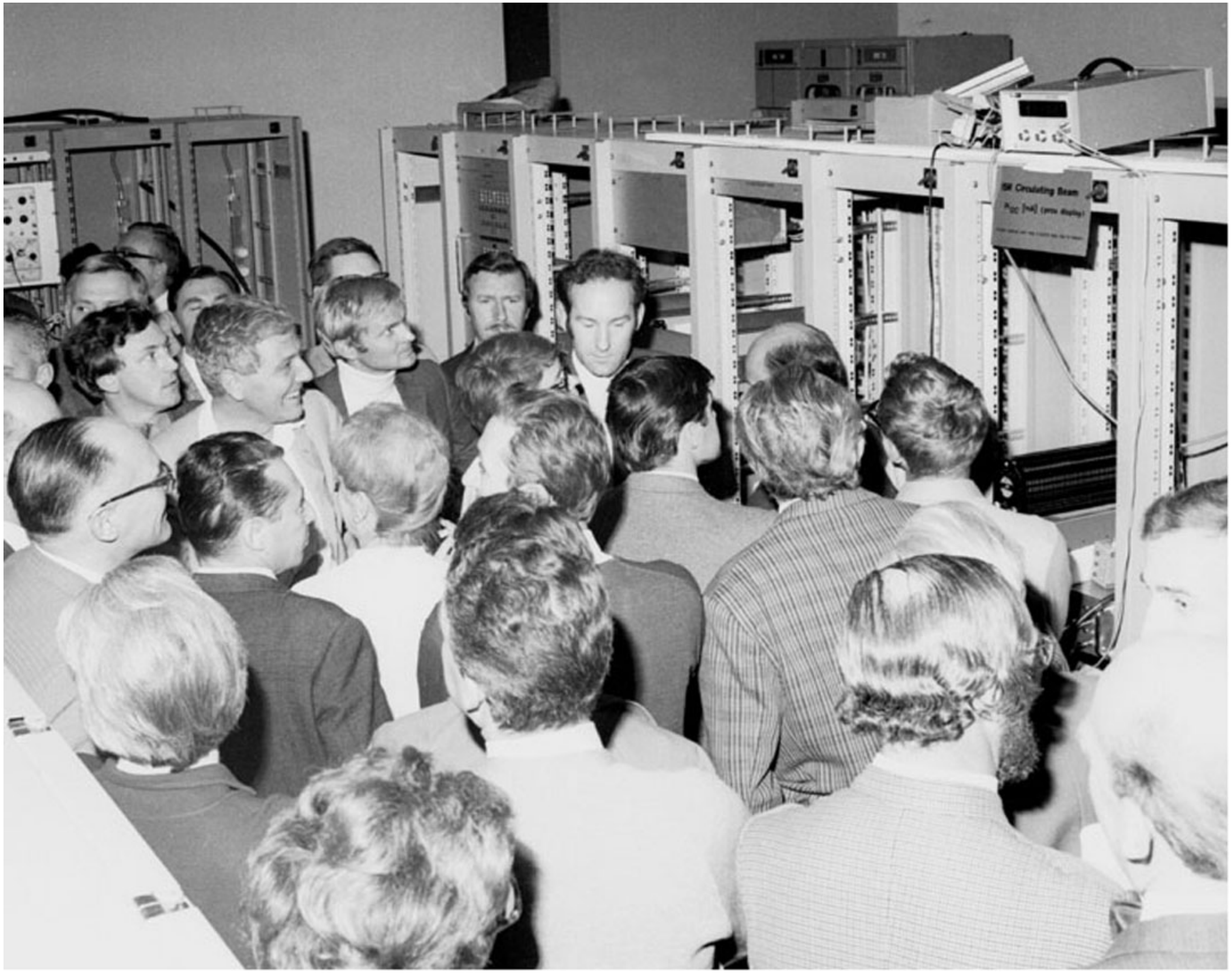
ISR. 1970

CR info tools:

Coaxial Cables

Teletype

Telephone



ISR 1970. Voltmeter display, no terminal

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

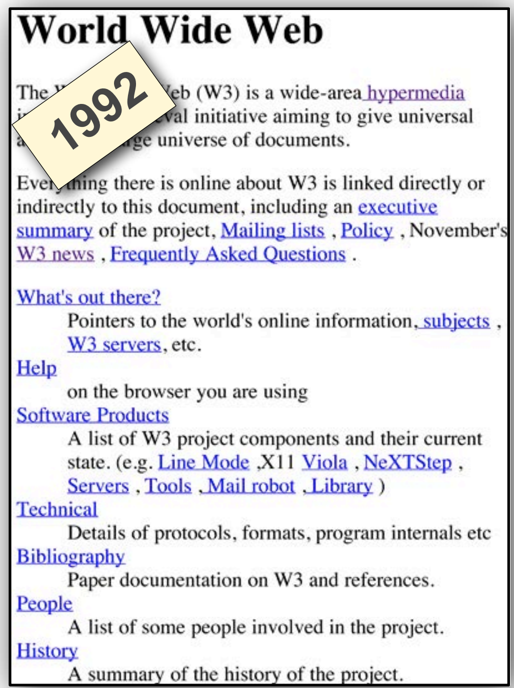
Fermilab: Remote Operations Centre

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



World Wide Web

1992

The W... eb (W3) is a wide-area hypermedia i... al initiative aiming to give universal a... ge universe of documents.

Eve... hing there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
: Pointers to the world's online information, subjects , W3 servers, etc.

Help
: on the browser you are using

Software Products
: A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)

Technical
: Details of protocols, formats, program internals etc

Bibliography
: Paper documentation on W3 and references.

People
: A list of some people involved in the project.

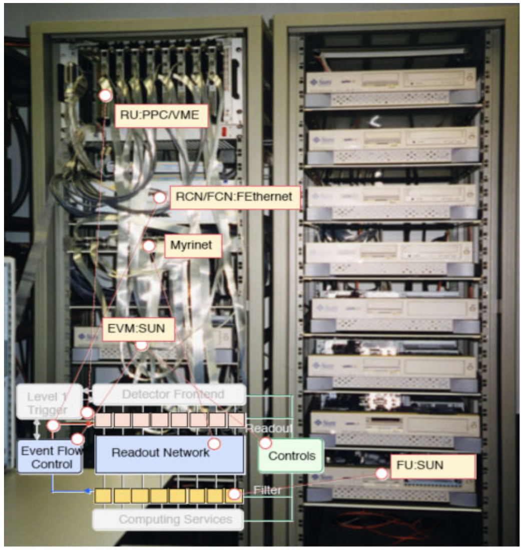
History
: A summary of the history of the project.

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.

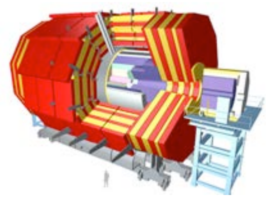
At the end of 80's with the digitalization of information and the growing support of information infrastructures (computer centers and Internet), a tool was needed to improve the collaboration between physicists and other researchers in the high energy physics community.

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

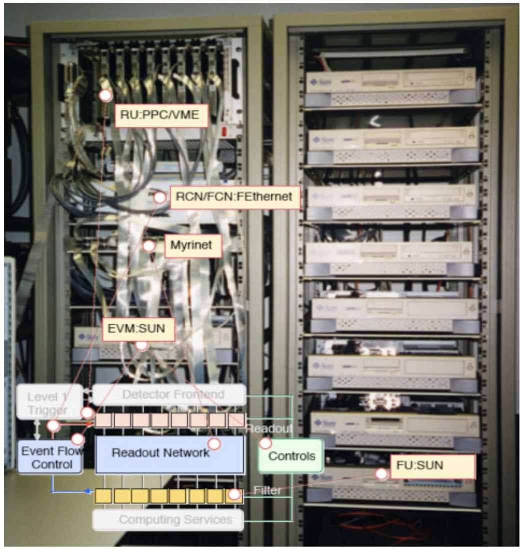


2008 The CMS HLT center on CESSY
and hundreds Off-line GRID computing centres 10^5 cores

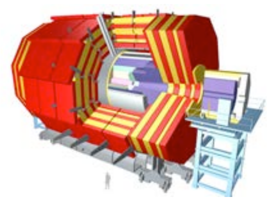


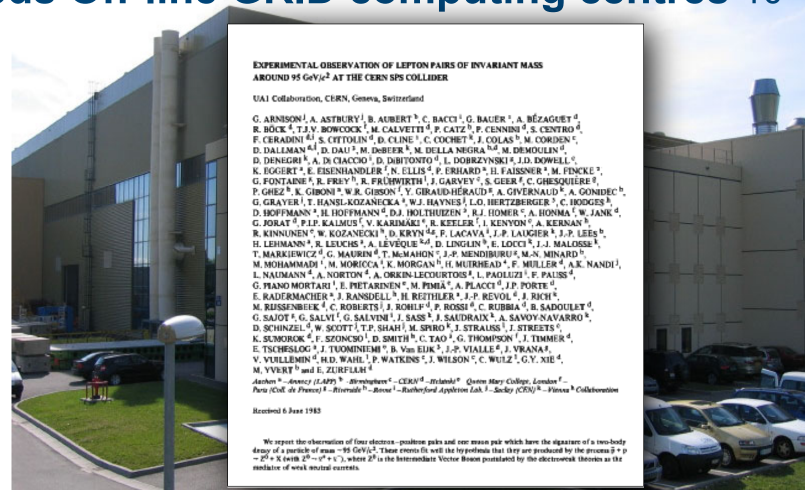
2008 One of Google data center 10^6 cores

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

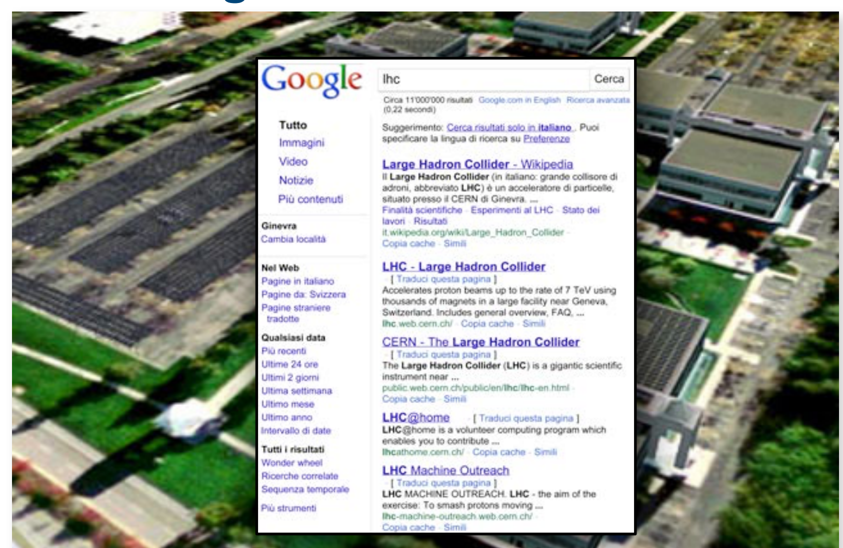


2008 The CMS HLT center on CESSY and hundreds Off-line GRID computing centres 10^5 cores





2008 One of Google data center 10^6 cores



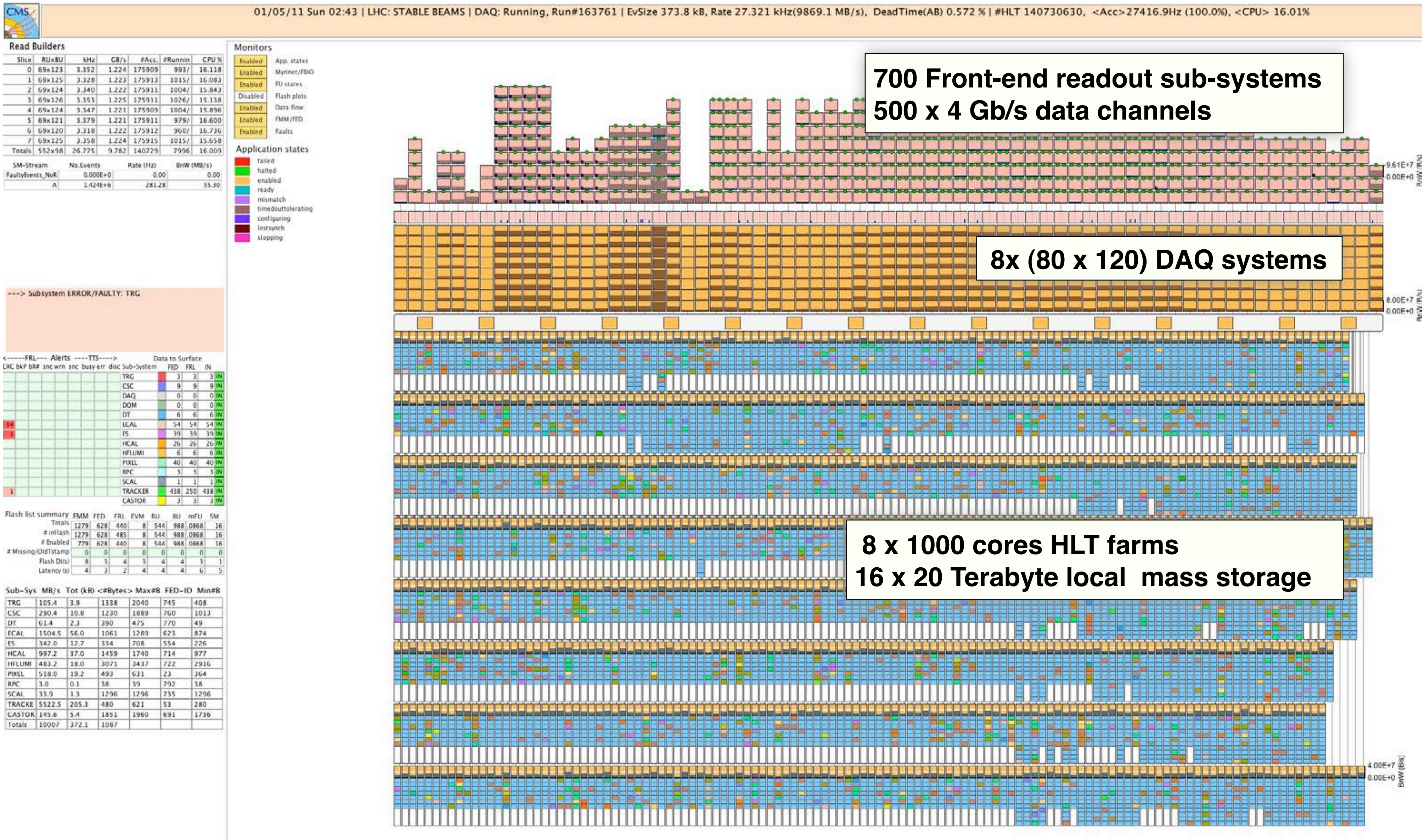
In operation ...

http://cmsonline.cern.ch/daqStatusSCX/aDAQmon/DAQstatusGre.jpg

http://cmsonline.cern.ch/daqStatusSCX/aDAQmon/LayoutB.jpg

http://cmsonline.cern.ch/daqStatusSCX/aDAQmon/wDAQ/aDAQmonScreenDumps1.html

System status display



Data acquisition in operation

Technologies and TDAQ: summary

- Continued rapid processor performance growth following Moore's law
- Network bandwidth grew at rate even higher than that of Moore's Law
- Commodity products everywhere
- Open software model (Linux) became a standard
- Internet explosion included software and hardware services
- Global network services allow scale-free decentralisation of data analysis and storage

However

High technology products limited life time (3-5 years)
Long-term maintenance issues and equipment replacement plans

The LHC program – machine, experiments, computing – has taken 25 years to complete, from conception to operation. It is expected to operate for as long with some major upgrades in the performance of the machine.

25 years is several generations of technology changes which the TDAQ design has to take into account for its maintenance plans and for the next machine upgrade program. Even more use of emerging computing and communication Internet services will (have to) be made

SLHC upgrade

The current TDAQ design should be considered complete, but its implementation is not final. It is expected to change with time, accelerator and experiment conditions to provide the maximum possible flexibility to execute a physics selection on-line.

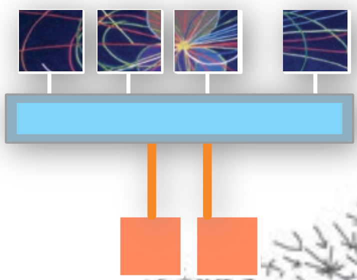
Luminosity increase (2018...) will require

- New front-end electronics and readout links
- Higher **level-1 selection power** (to maintain 100 kHz max. output)
- **Readout network (>10 Tb/s)** with an order of magnitude higher
- **More on-line computing power/mass-storage ....**

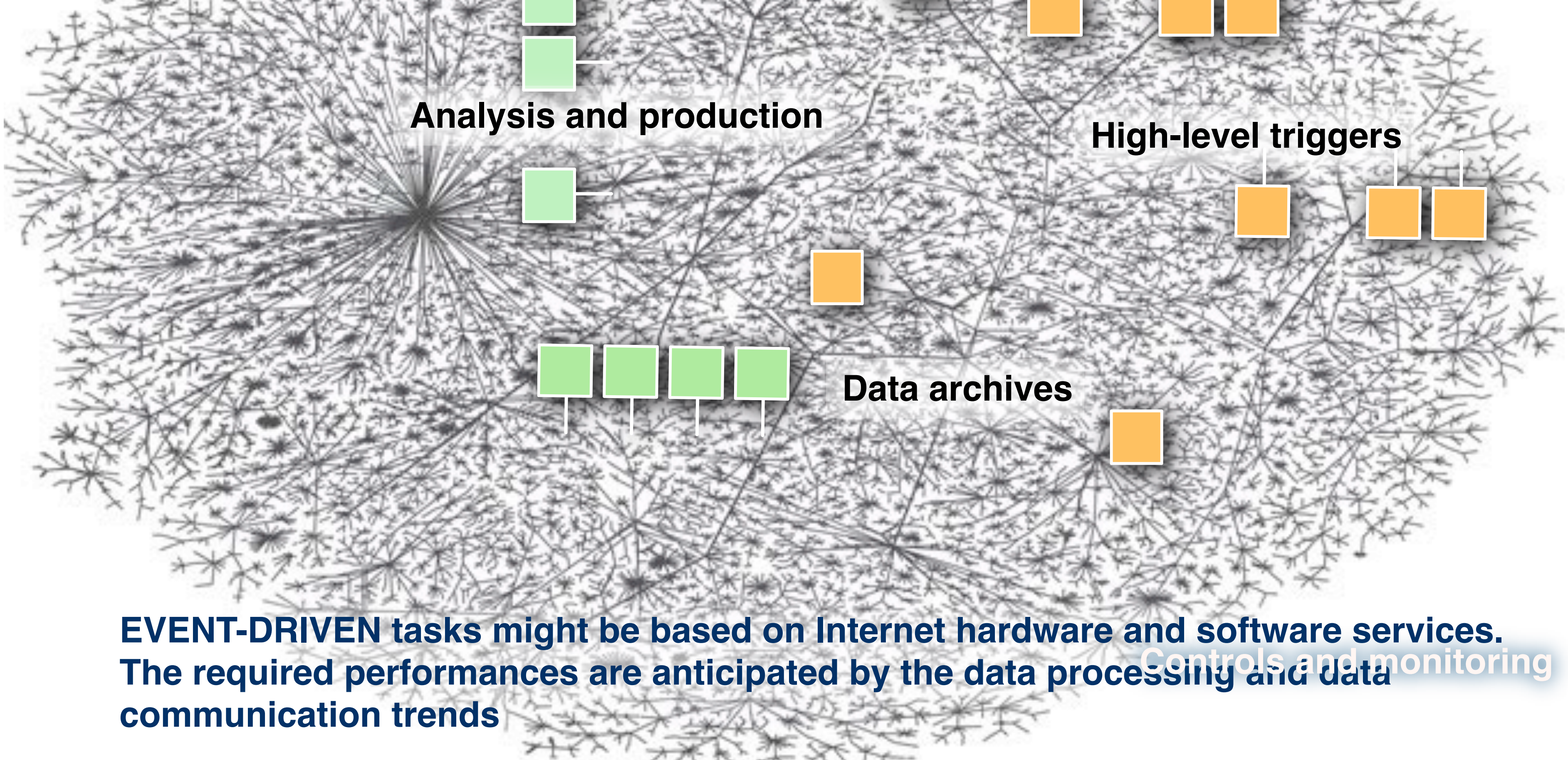
The upgrade program will include:

- 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.
- **Level1 trigger.** Wider scope, new architectures including arrays of logic nodes. Use standards for data communication and central logic
- **Front-End digitizers**, new rad hard data links and a new timing and trigger distribution system (distribute event type, HLT destination etc.).
- **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 fragments 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

Scale-free internet services



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



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

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.