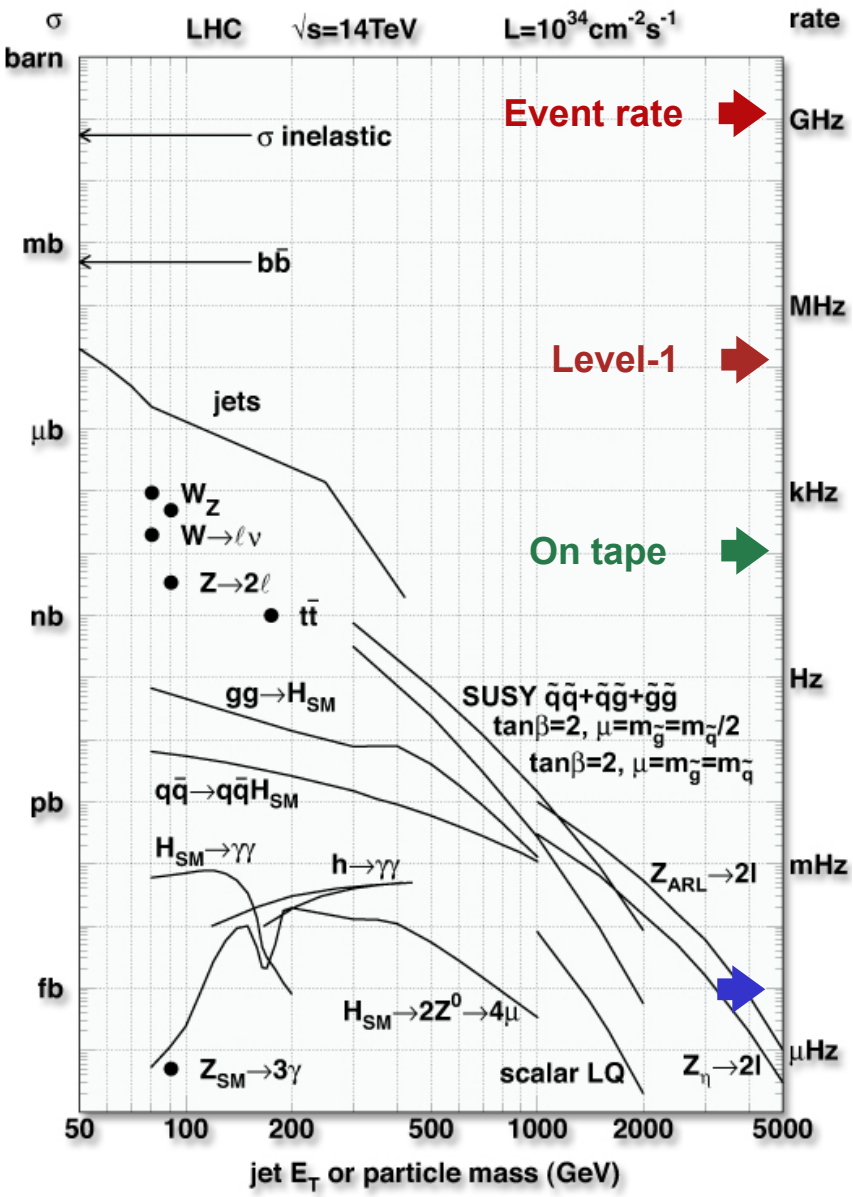
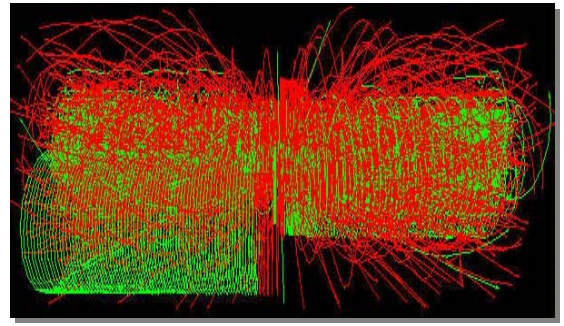
CMS Data Acquisition

S. Cittolin CERN/PH-CMD
Oct. 7 2004 LHC DAYS IN SPLIT

Requirements and Architecture
Trigger and Readout
Event Builder and Technologies
Plans and summary

p-p collisions at LHC



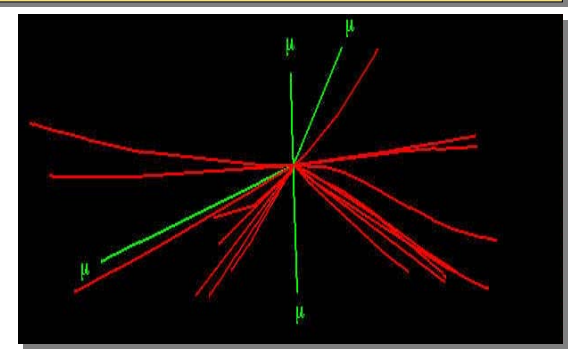


Operating conditions:
Higgs in 4 muons
+ ~20 minimum bias

All charged tracks
with pt > 2 GeV

Event Rates:	**~10^9 Hz**
Event size:	**~1 MByte**

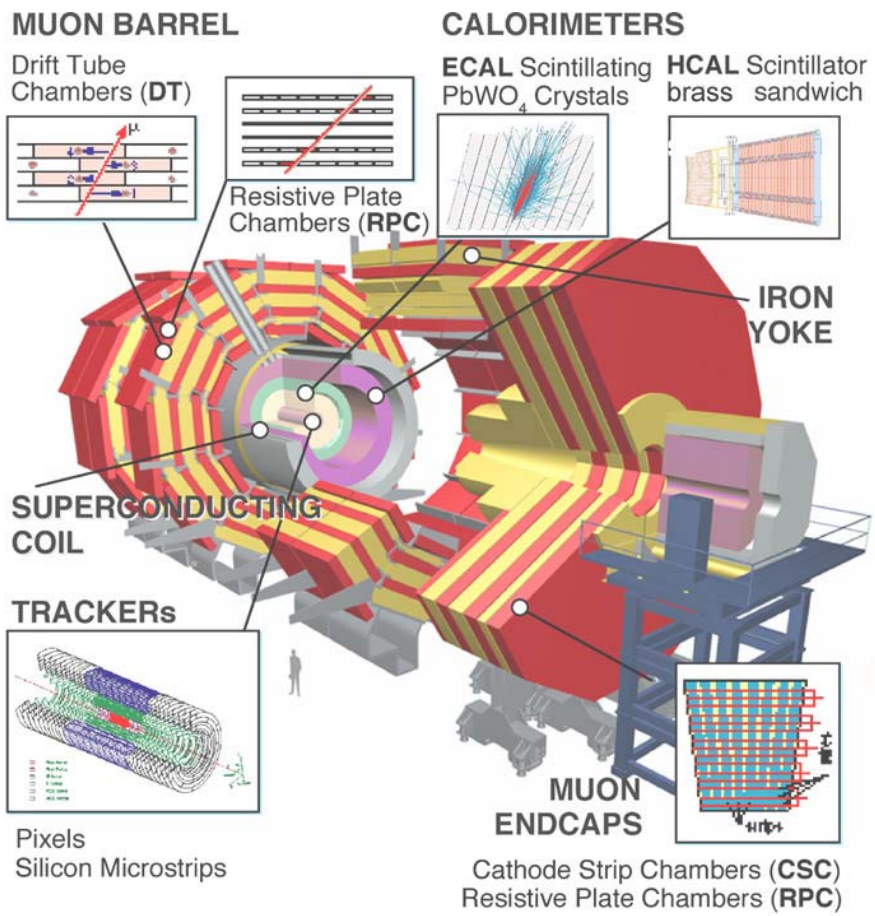
Level-1 Output	**100 kHz**
Mass storage	**10^2 Hz**
Event Selection:	**~$1/10^{13}$**



Reconstructed tracks
with pt > 25 GeV

Requirements and design parameters

Detectors

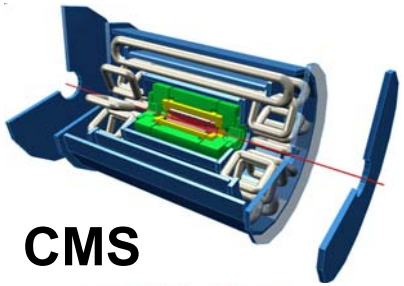
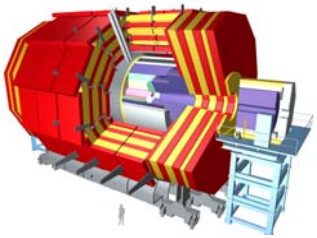
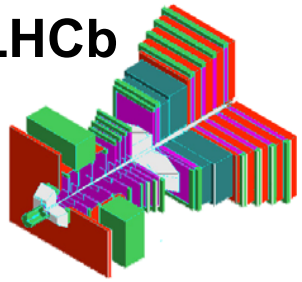
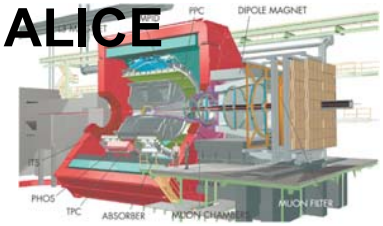


Total weight : 12,500 t
Overall diameter : 15 m
Overall length : 21.6 m
Magnetic field : 4 Tesla

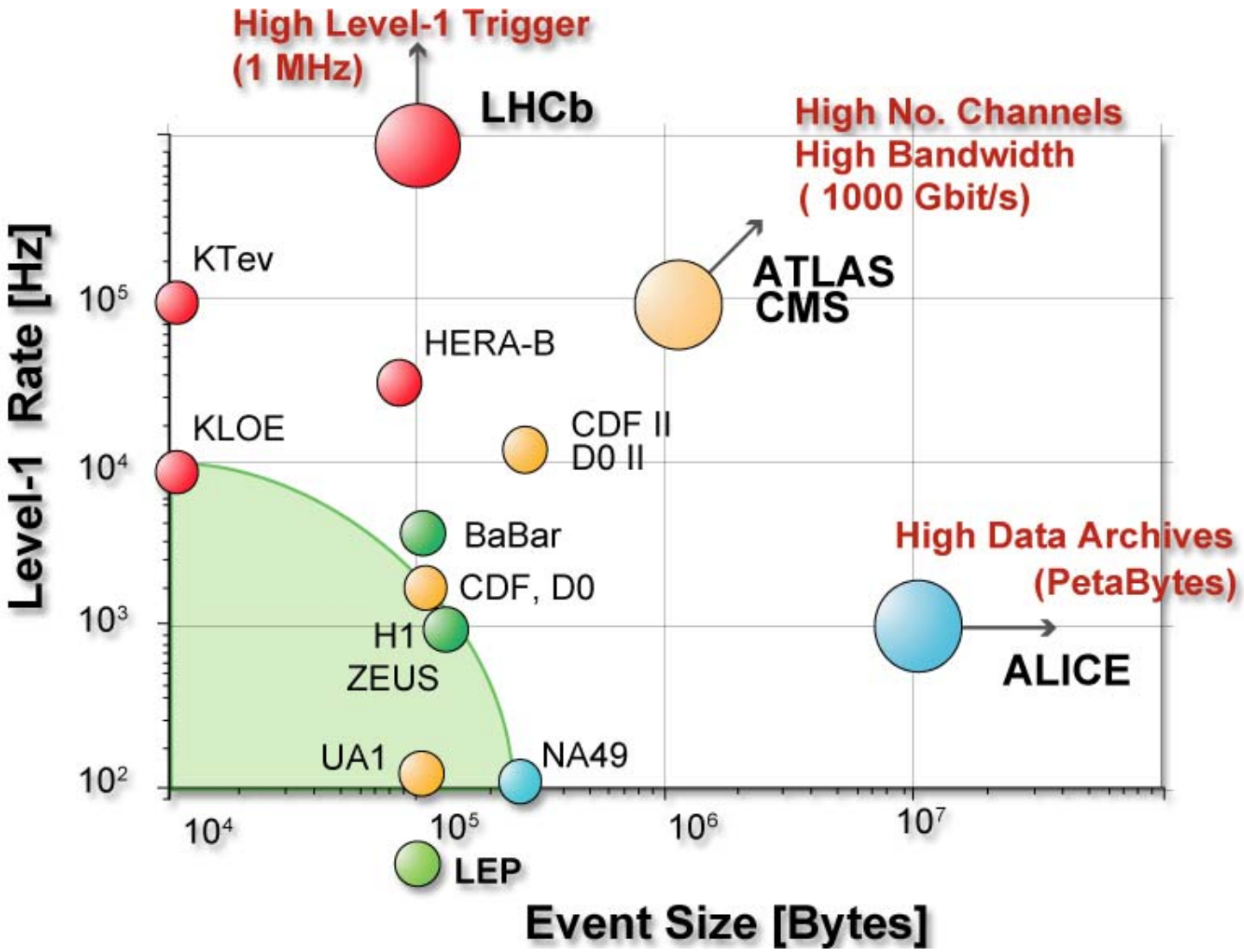
Detector	Channels	Control	Ev. Data
Pixel	60000000	1 GB	50 (kB)
Tracker	10000000	1 GB	650
Preshower	145000	10 MB	50
ECAL	85000	10 MB	100
HCAL	14000	100 kB	50
Muon DT	200000	10 MB	10
Muon RPC	200000	10 MB	5
Muon CSC	400000	10 MB	90
Trigger		1 GB	16

Event size	**1 Mbyte**
Max LV1 Trigger	**100 kHz**
Online rejection	**99.999%**
System dead time	**~ %**

LHC trigger and DAQ summary

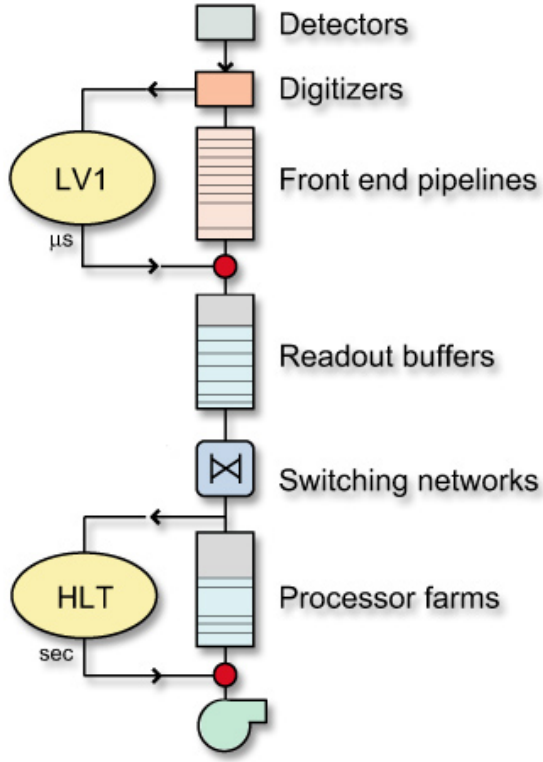
	No.Levels Trigger	First Level Rate (Hz)	Event Size (Byte)	Readout Bandw.(GB/s)	Archive MB/s (Event/s)
ATLAS 	3	10^5 LV-2 10^3	10^6	10	100 (10^2)
CMS 	2	10^5	10^6	100	100 (10^2)
LHCb 	3	LV-0 10^6 LV-1 $4\ 10^4$	$2x10^5$	4	40 ($2x10^2$)
ALICE 	4	Pp-Pp 500 p-p 10^3	$5x10^7$ $2x10^6$	5	1250 (10^2) 200 (10^2)

Trigger and data acquisition trends



CMS DAQ structure: 2 physical triggers

40 MHz
Clock driven
Custom processors



Level-1 Trigger
Custom design

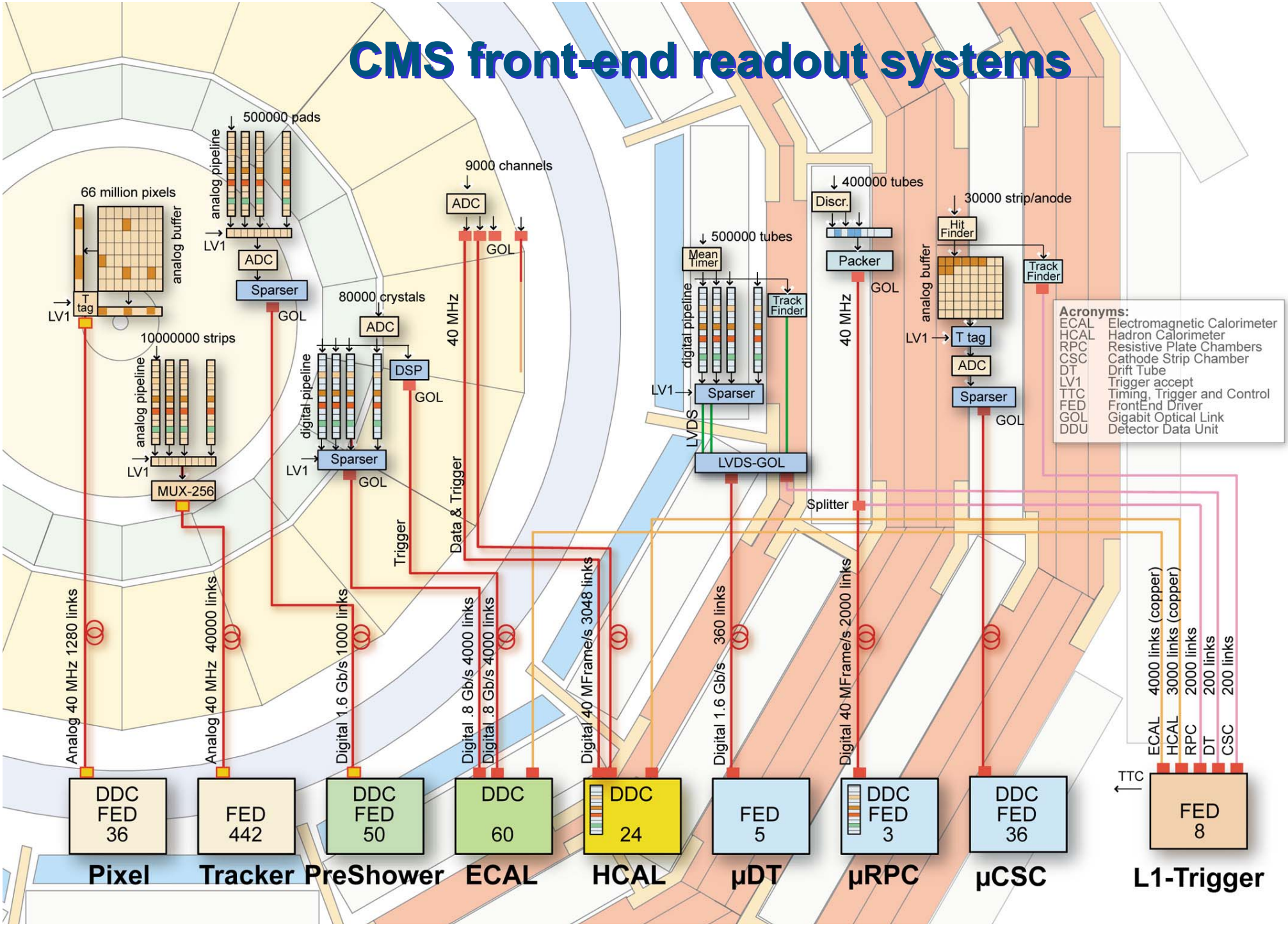
100 kHz
Event driven
PC network

High-Level Trigger
Industry products

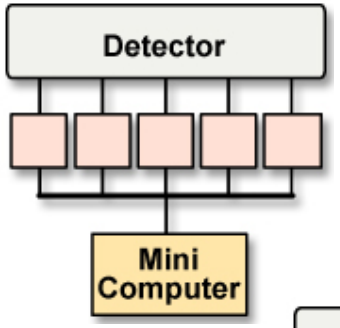
Level-1 output / HLT input	**100 kHz**
Network bandwidth	**1 Terabit/s**
HLT output	**10^2 Hz**

Invest in data transportation and CPU

CMS front-end readout systems

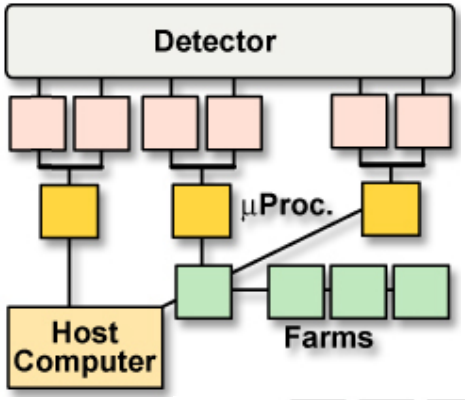


Evolution of DAQ technologies and structures



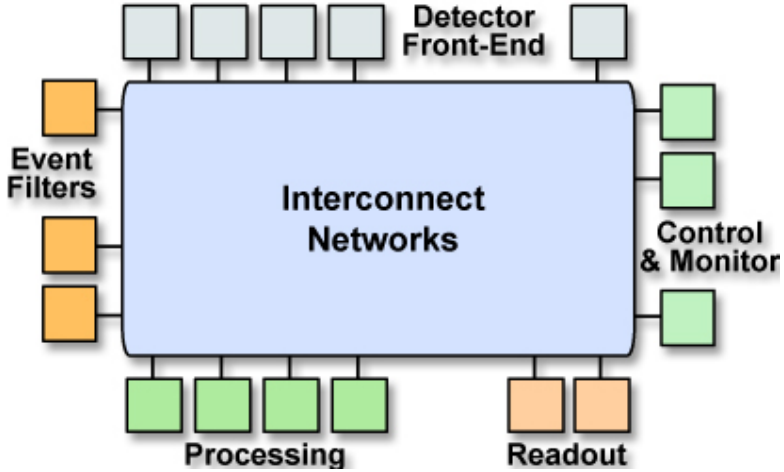
1970-80: Minicomputers

Readout custom design
First standard: CAMAC
• **kByte/s**



1980-90: Microprocessors

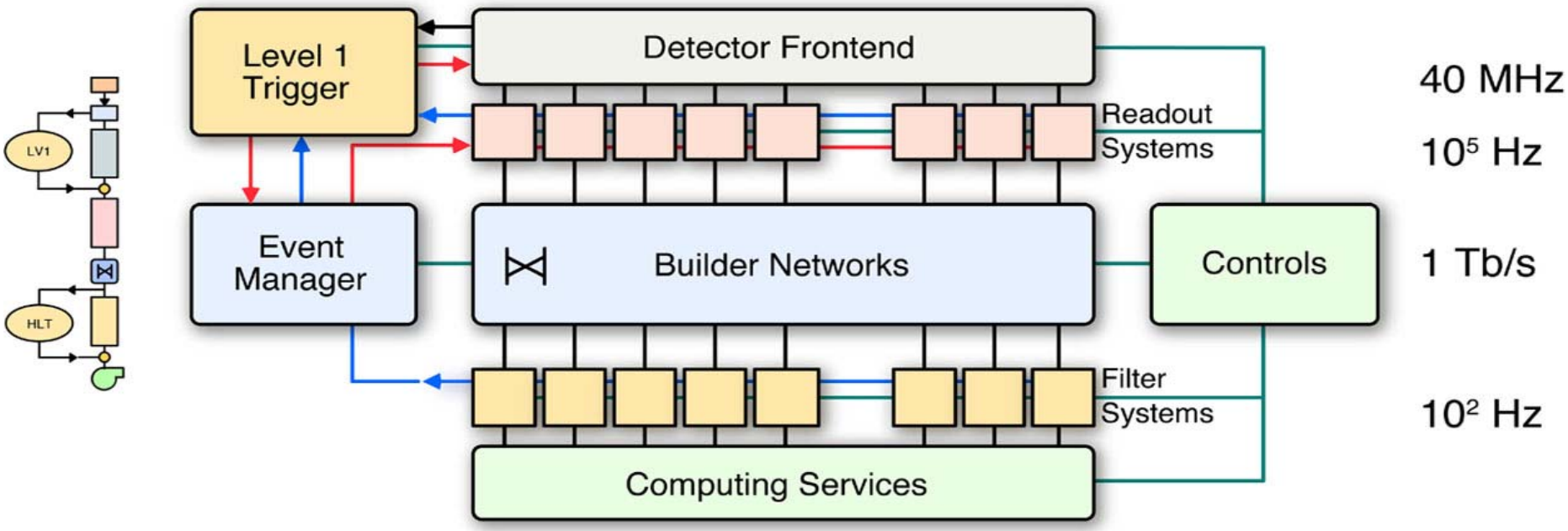
HEP standards (Fastbus)
Embedded CPU, Industry standards (VME)
• **MByte/s**



2000-xx: Networks

IT commodities, PC, Clusters
Internet, Web, etc.
• **GByte/s**

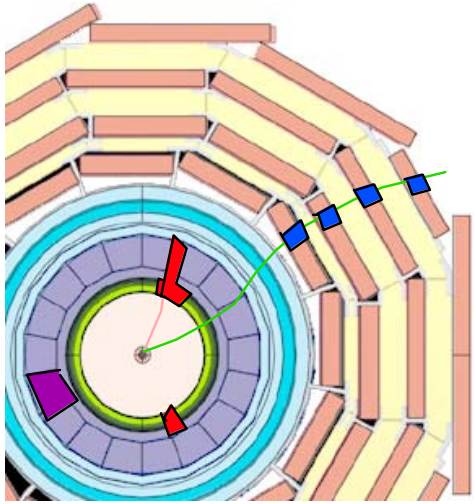
DAQ baseline structure



Collision rate	40 MHz
Level-1 Maximum trigger rate	**100 kHz**
Average event size	**≈ 1 Mbyte**
Event Flow Control	≈ 10^6 Mssg/s

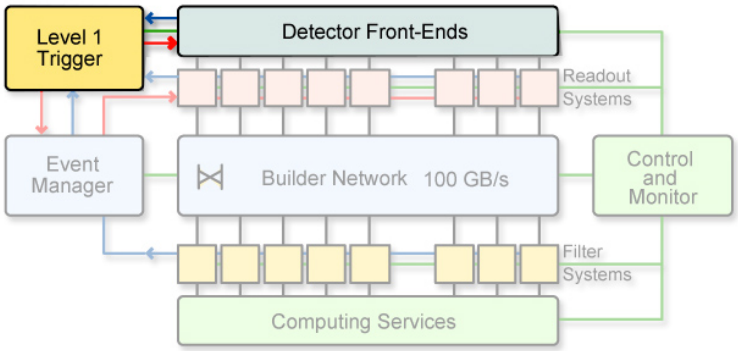
No. of In-Out units	**512**
Readout network bandwidth	**≈ 1 Terabit/s**
Event filter computing power	**≈ 10^6 SI95**
Data production	≈ Tbyte/day
No. of PC motherboards	≈ Thousands

Two trigger levels

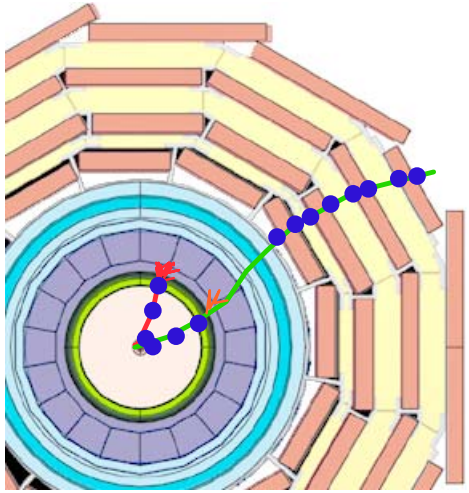


Level-1: Specialized processors
40 MHz synchronous

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

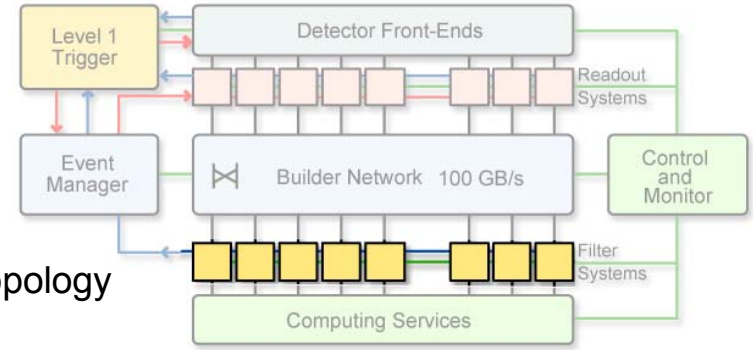


99.99 % rejected 0.01 Accepted

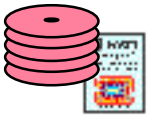


High trigger levels: CPU farms
100 kHz asynchronous farms

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



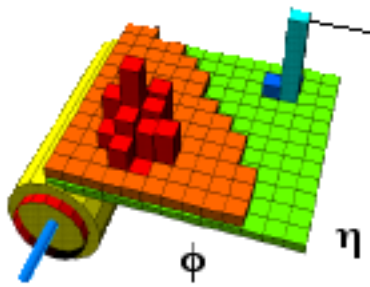
100-1000 Hz. Mass storage
Reconstruction and analysis.



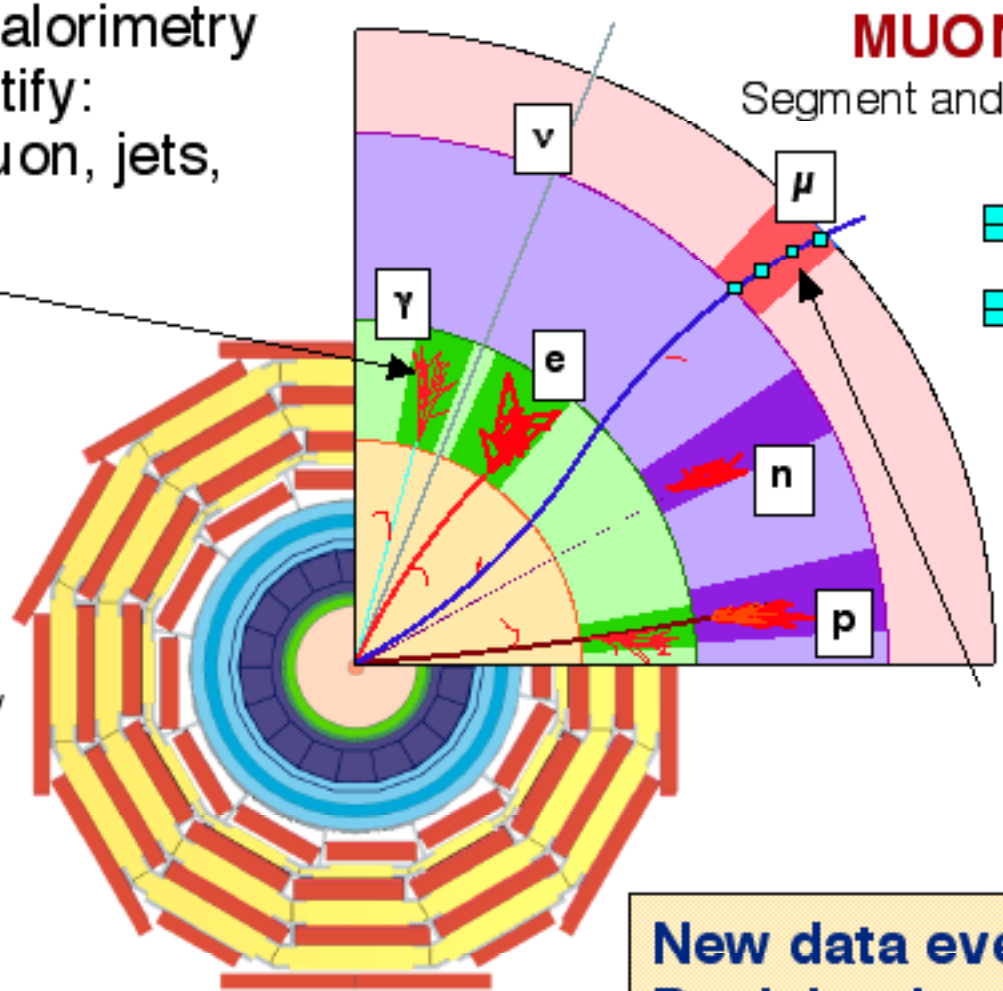
99.9 % rejected 0.1 Accepted

Level-1. Particle identification

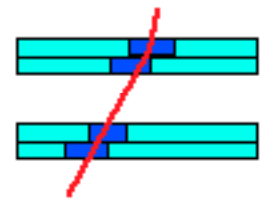
Use prompt data (calorimetry and muons) to identify:
High p_t electron, muon, jets, missing E_T



CALORIMETERs
Cluster finding and energy deposition evaluation

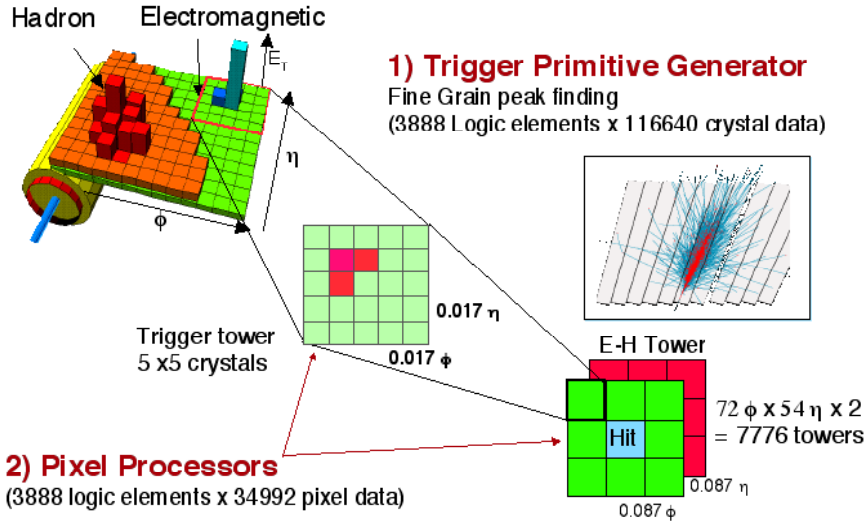


MUON System
Segment and track finding



New data every 25 ns
Decision latency ~ μs

Level-1 trigger systems



Trigger Primitive Generator

Fine grain — Flag Max of (, , , ,)

Pixel Processor

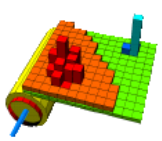
E_t cut — + Max ( ) > Threshold

Longitudinal cut (H/E) — AND / < 0.05

Neighbors longitudinal cut — AND / < 2 GeV

AND

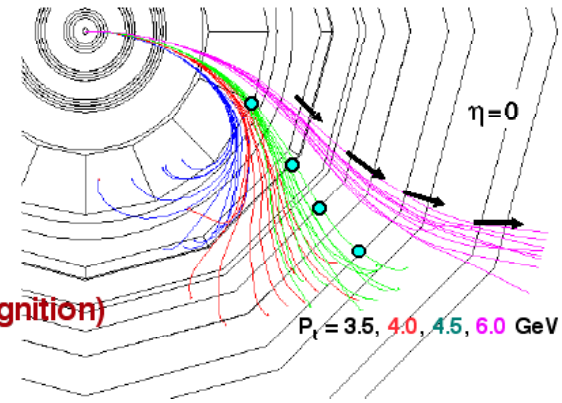
One of (, , ,)



ISOLATED ELECTRON

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

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



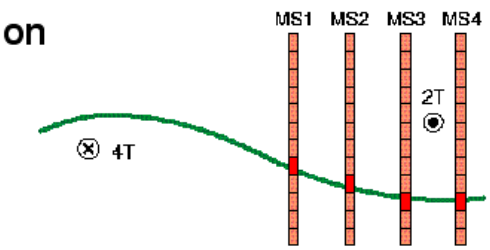
Detectors:

RPC (pattern recognition)

DT(track segment)

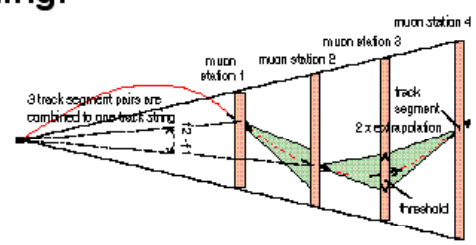
RPC pattern recognition

- Pattern catalog
- Fast logic

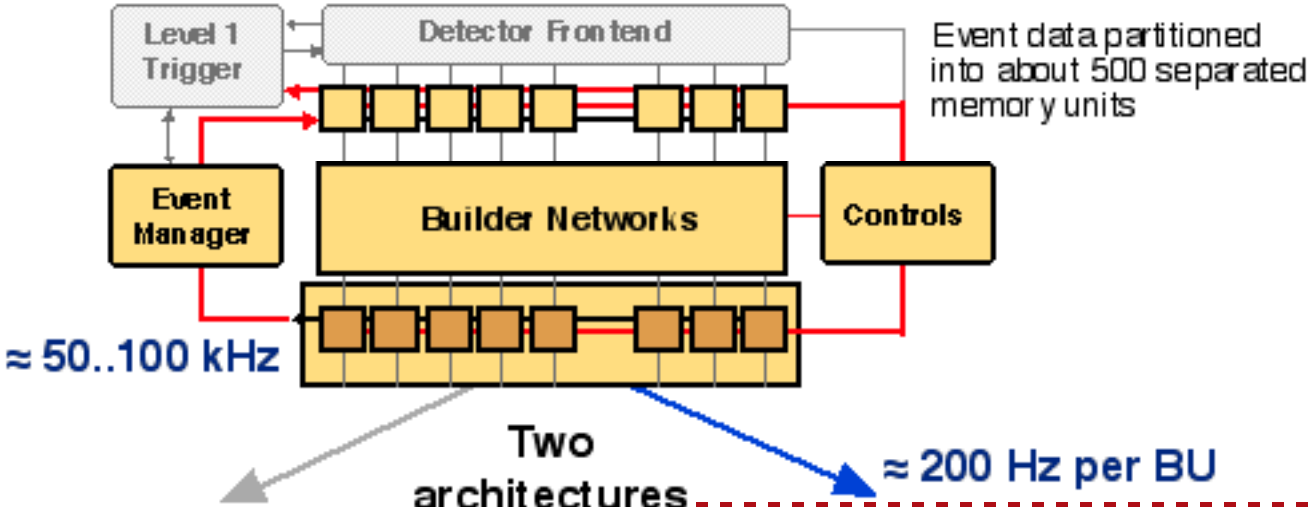


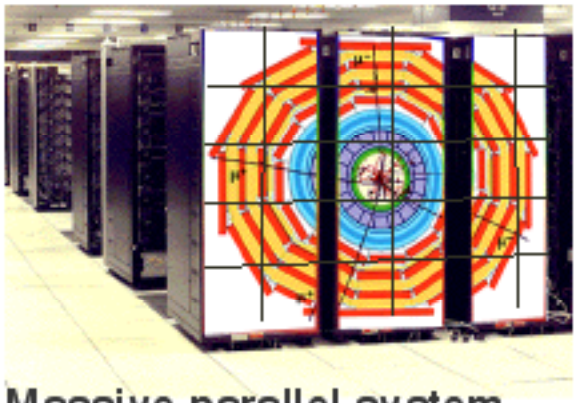
DT and CSC track finding:

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



High Level Trigger: Event Filter farm

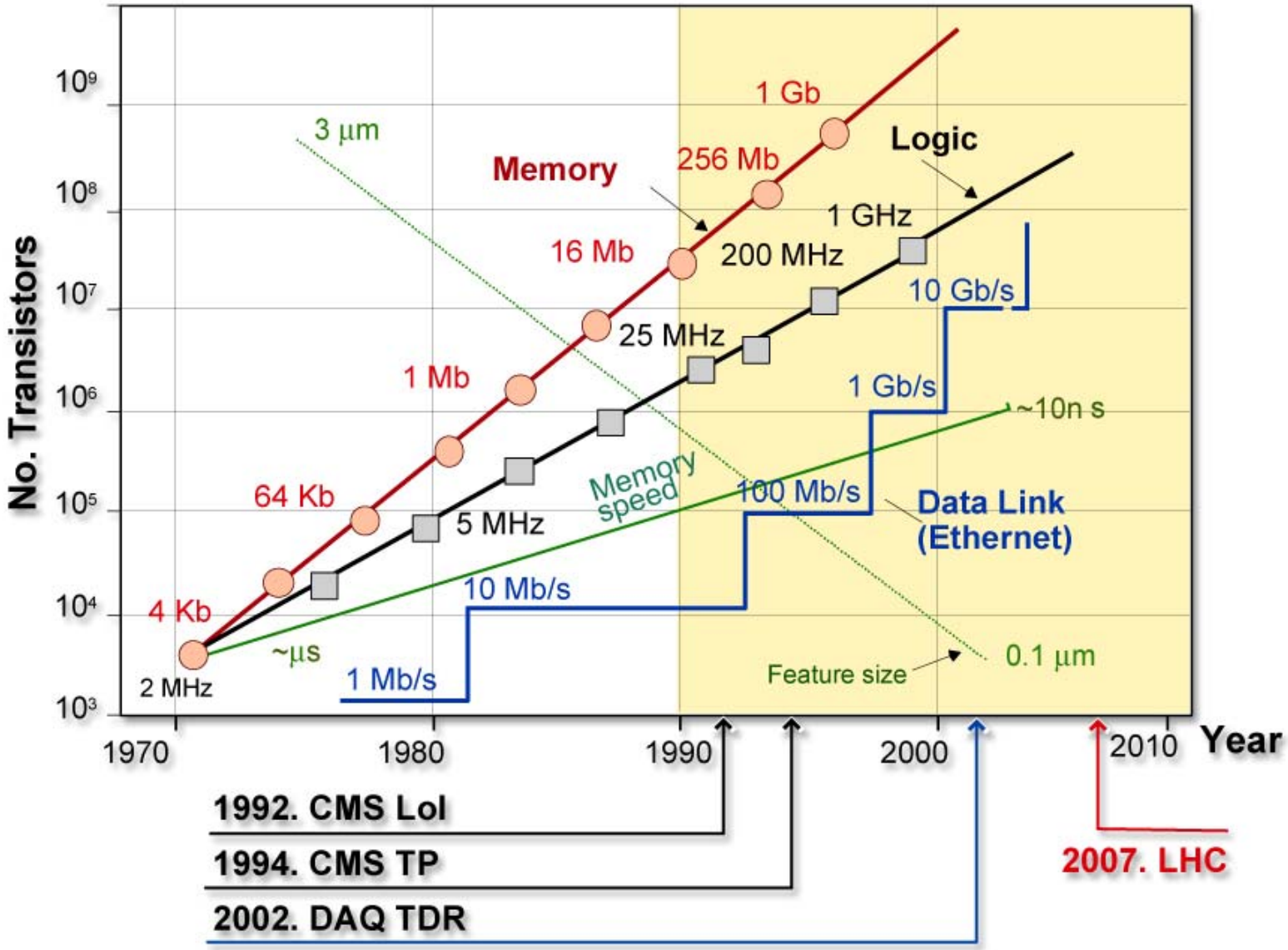




Massive parallel system
ONE event, ALL processors
- Low latency
- Complex I/O
- Parallel programming

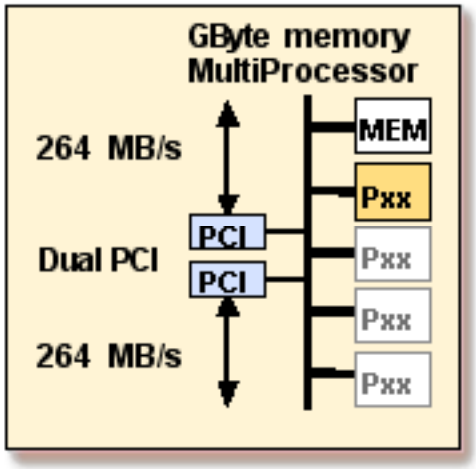
Farm of processors
ONE event, ONE processor
- High latency (larger buffers)
- Simpler I/O
- Sequential programming

Technology trends



Readout and computing IO structures

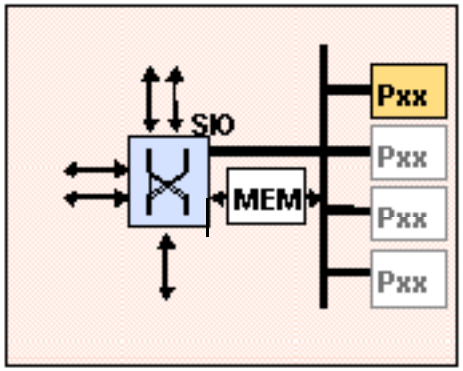
1990' PCI



IO and Processing systems :Commercial PCs
Operating systems : Unix(Linux),
Interfaces standards :PC IO systems (e.g. PCI)

Desktop/Server
current architecture
Peripheral IO **bus PCI:**
33/66 MHz x 32/64 bit
100/200/400 MB/s

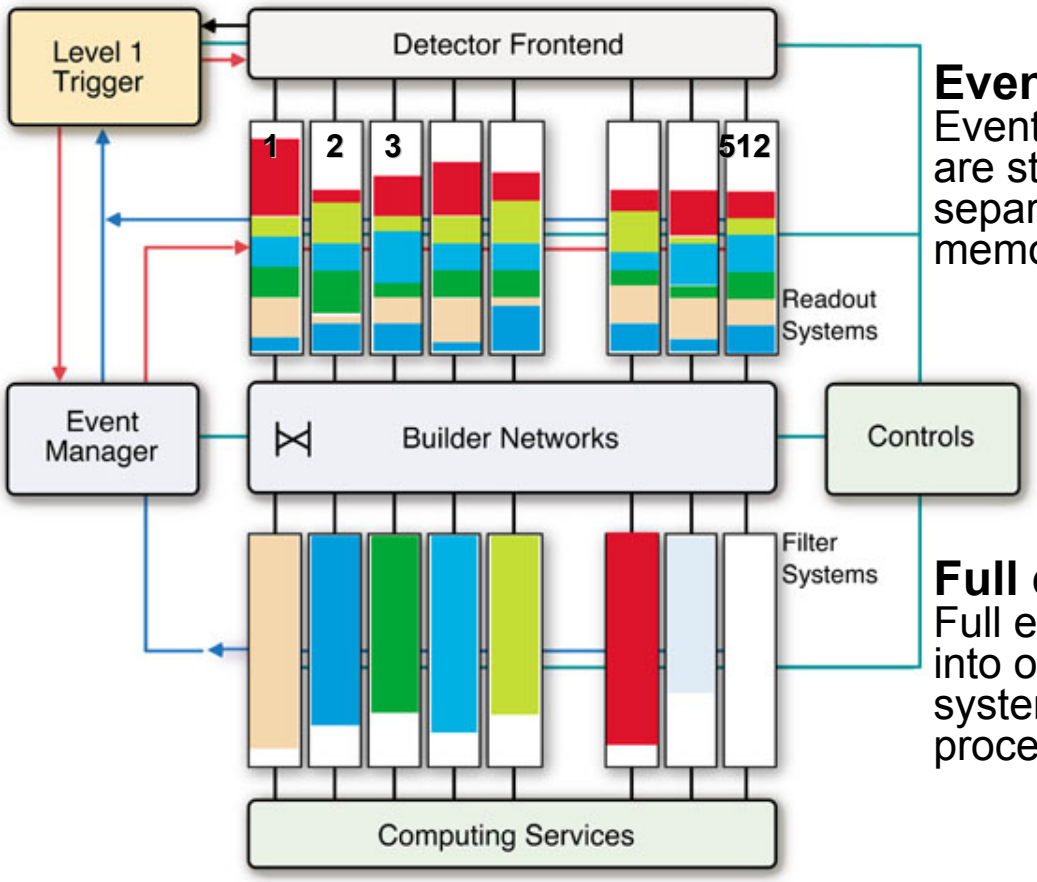
200X: PCI-X …



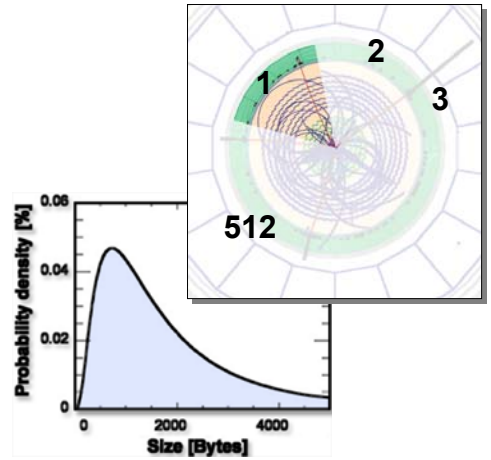
2002 PC mother boards:
- 2 GHz dual processors
- 4 PCI-X ports at 1GB/s
- 3 GB/s memory bandwidth
- Suitable for all DAQ readout applications

Building the event (EVB)

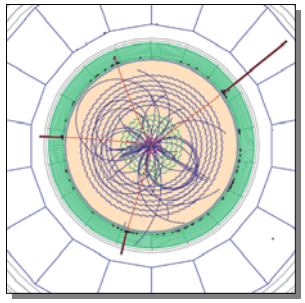
Event builder :
Physical system interconnecting data sources with data destinations. It has to move each event data fragments into a same destination



Event fragments :
Event data fragments are stored in separated physical memory systems



Full events :
Full event data are stored into one physical memory system associated to a processing unit

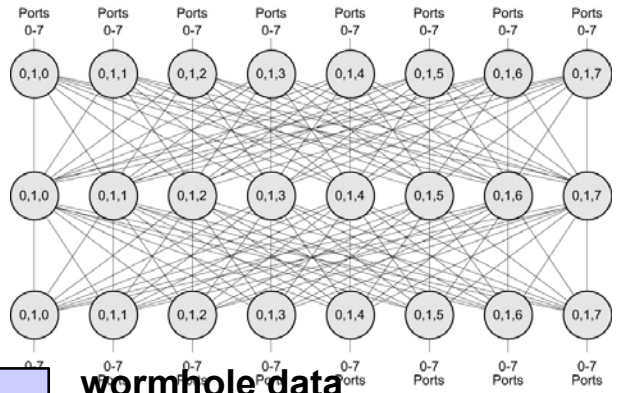


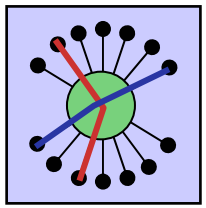
512 Data sources for 1 Mbyte events
~1000s HTL processing nodes

EVB and switch technologies

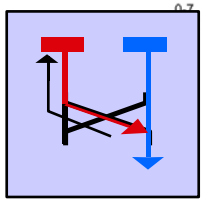
Myricom Myrinet 2000

- Switch: **Clos-128 x 2.5** Gb/s port
- **NIC**: M3S-PCI64B-2 (**LANai9**)
- **Custom Firmware**





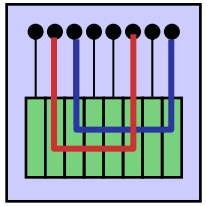
Implementation :
16x16 port X-bar capable of channeling data between any two ports.



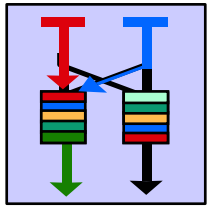
wormhole data transport with flow control at all stages

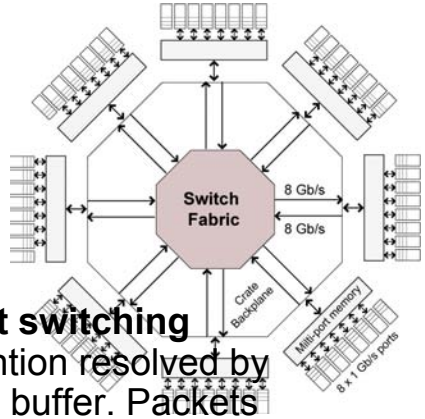
Gigabit Ethernet

- Switch: Foundry **FastIron 64 x 1.2** Gb/s port
- **NIC**: **Alteon** (running standard firmware)



Implementation:
Multi-port memory system of R/W access bandwidth greater than the sum of all port speeds

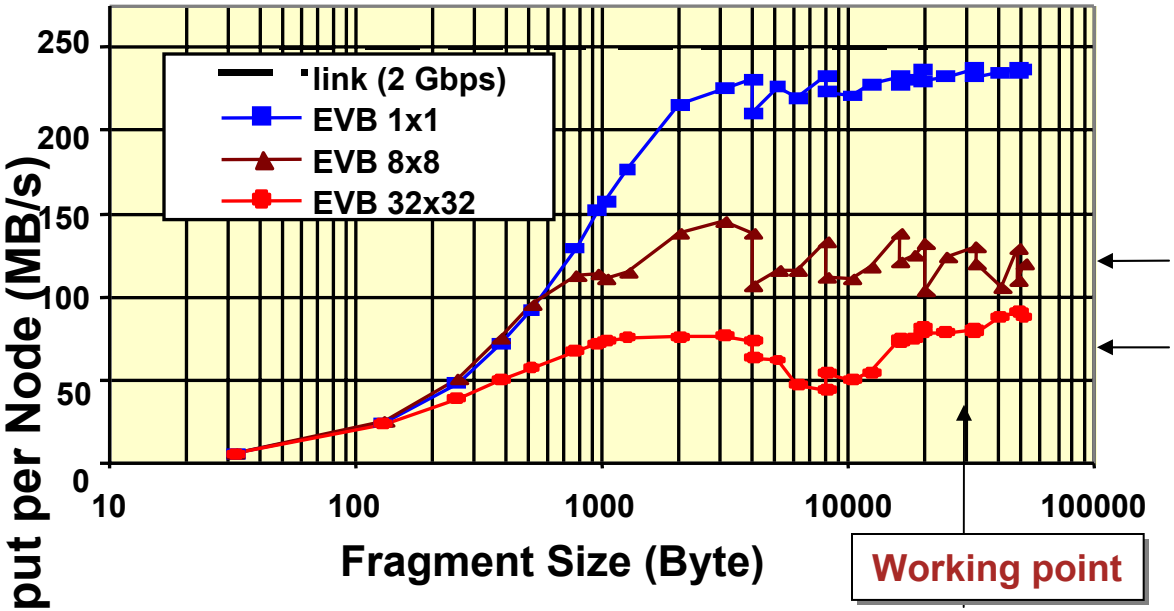




Packet switching
Contention resolved by Output buffer. Packets can be lost.

Infiniband • 2.5 Gb/s demo product. Tests ongoing with a small 2x2 setup

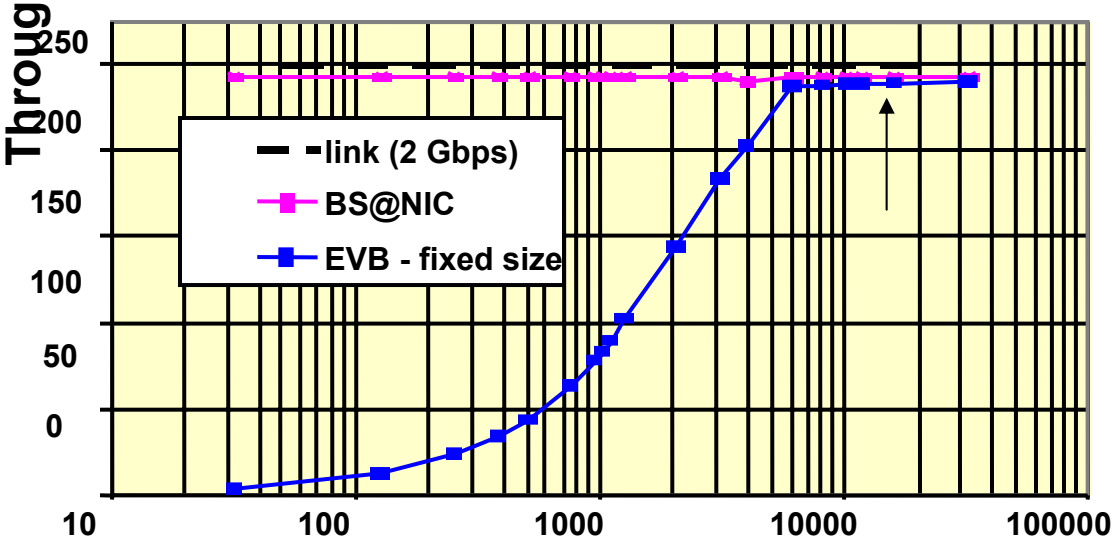
32x32 Myrinet EVB protocols results



Random traffic

8x8: single stage:
max. **utilization: ≈ 50%**

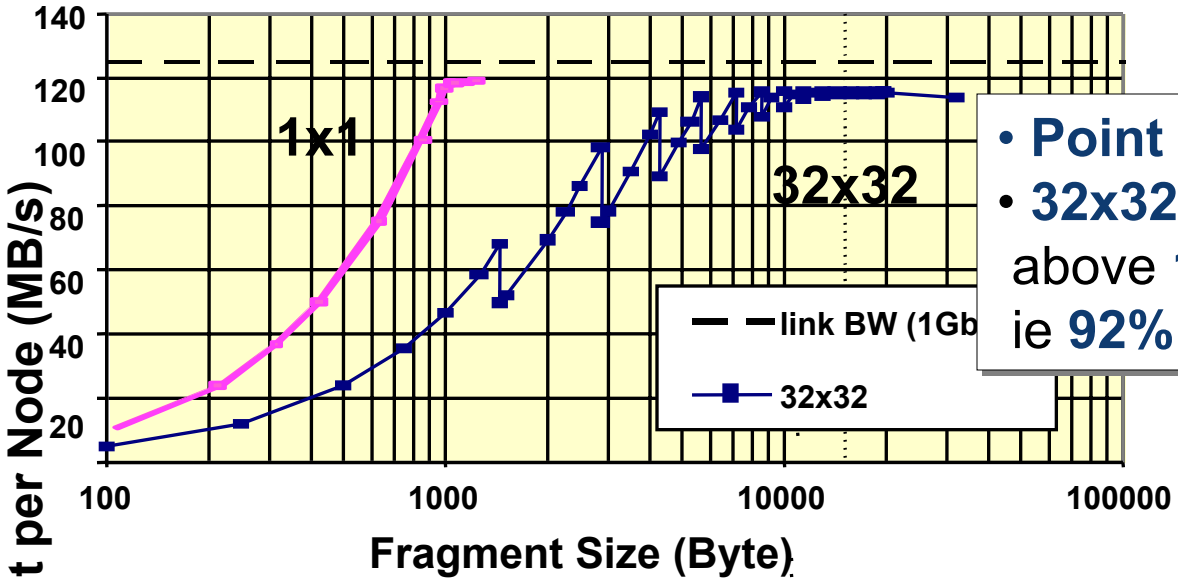
32x32: two stage network
max. **utilization ≈ 30%**



Barrel shifter

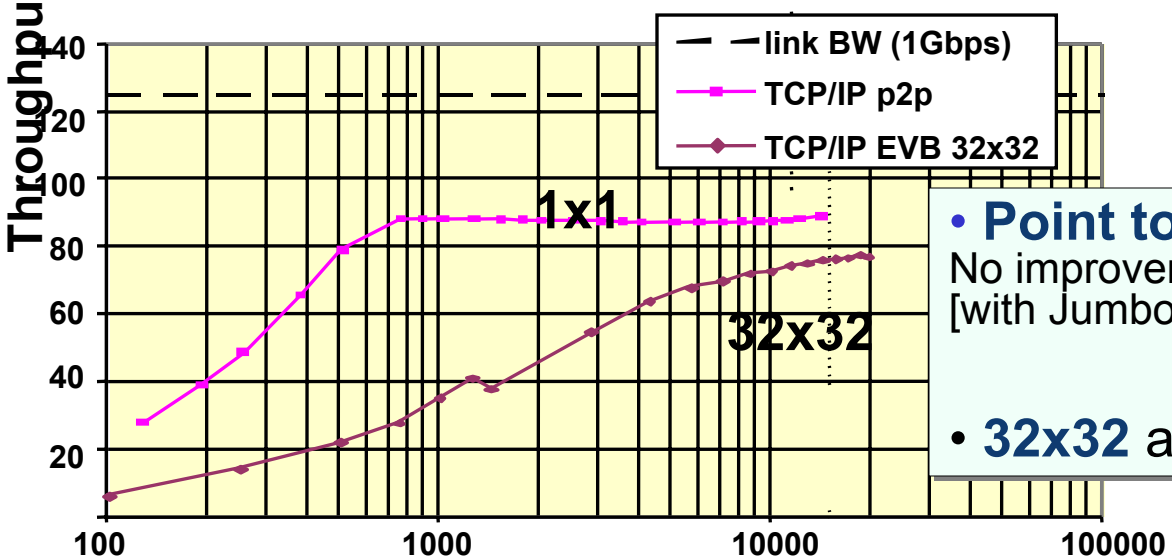
- Fixed size event fragments
  below 4k: Fragment < BS carrier
  above 4k: Fragment > BS carrier
- Throughput at **234 MB/s**
  = **94% of link Bandwidth**

32x32 GbE EVB protocols results



Raw packets (L2 frames)

- **Point to point** asymptotically to 125 MB/s
- **32x32** saw tooth due to MTU above **10k**: plateau of 115 MB/s ie **92%** of link speed (1Gbps)



TCP/IP

- **Point to point** At 800 B: 88 MB/s (**70%**). No improvement if host CPU733 -> 1000 MHz [with Jumbo frames 110 MB/s (88%)]
- **32x32** at **16 kB** 75 MB/s (**60%** load)

EVB demonstrators summary

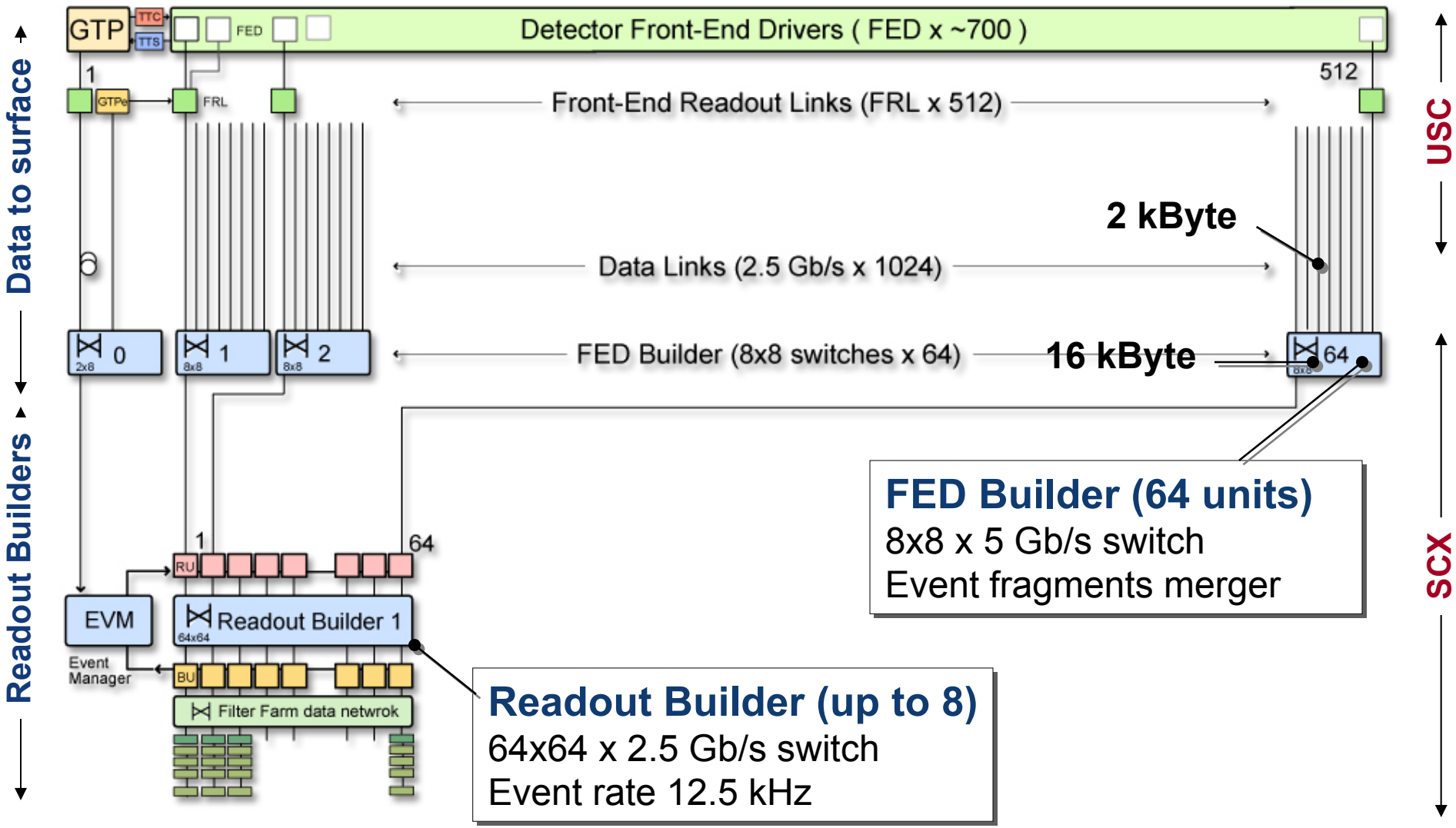
	Myrinet 2000	GbE raw packet	GbE TCP/IP
Test bench	32x32	32x32	32x32
Port speed	2.5 Gbit/s	1.2 Gbit/s	1.2 Gbit/s
Random traffic	30-50%	50%, 92% (*)	30%, 60% (*)
Barrel switch	94%	-	-
CPU load	Low	High	High
1 Tbit/s EVB	512x512	1024x1024	2048x2048
No. switches	8 128-Clos	16 256-port	32 256-port

Industry standards →

← Proprietary standards

(*) with fragment sizes larger than 16kB

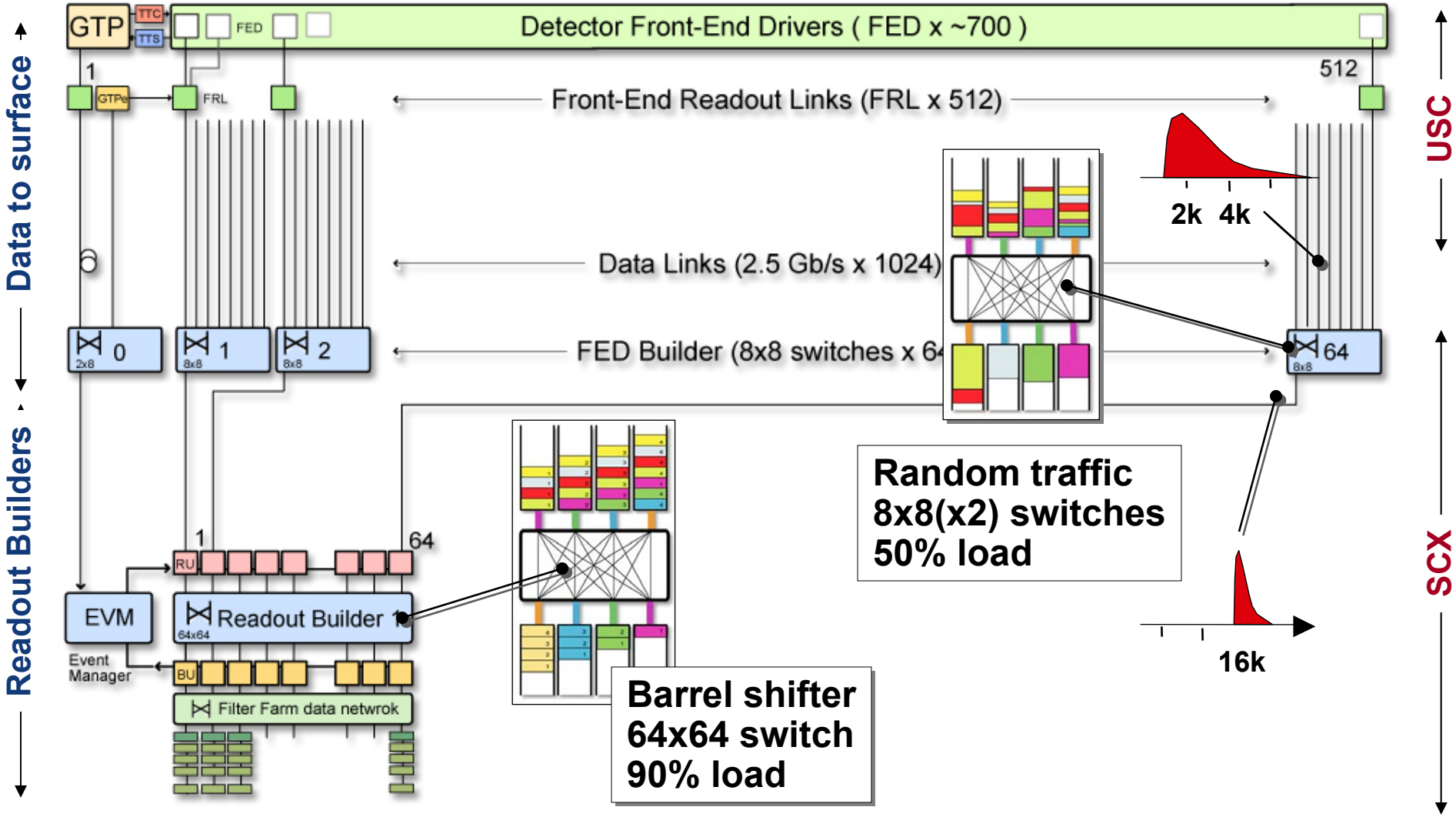
2 stages: Data to surface & Readout Builder



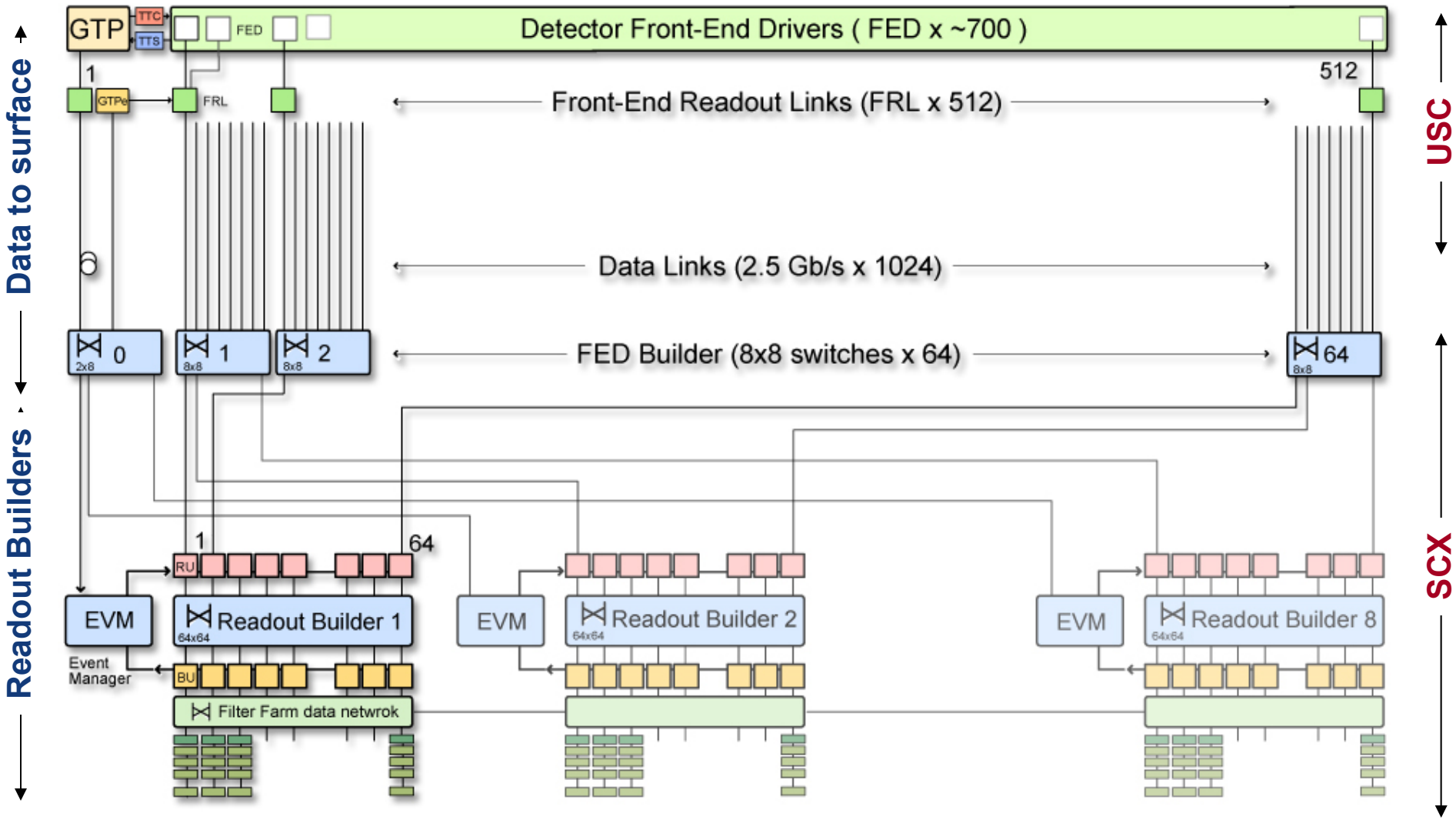
FED Builder (64 units)
8x8 x 5 Gb/s switch
Event fragments merger

Readout Builder (up to 8)
64x64 x 2.5 Gb/s switch
Event rate 12.5 kHz

Builders protocols (e.g. Myrinet)

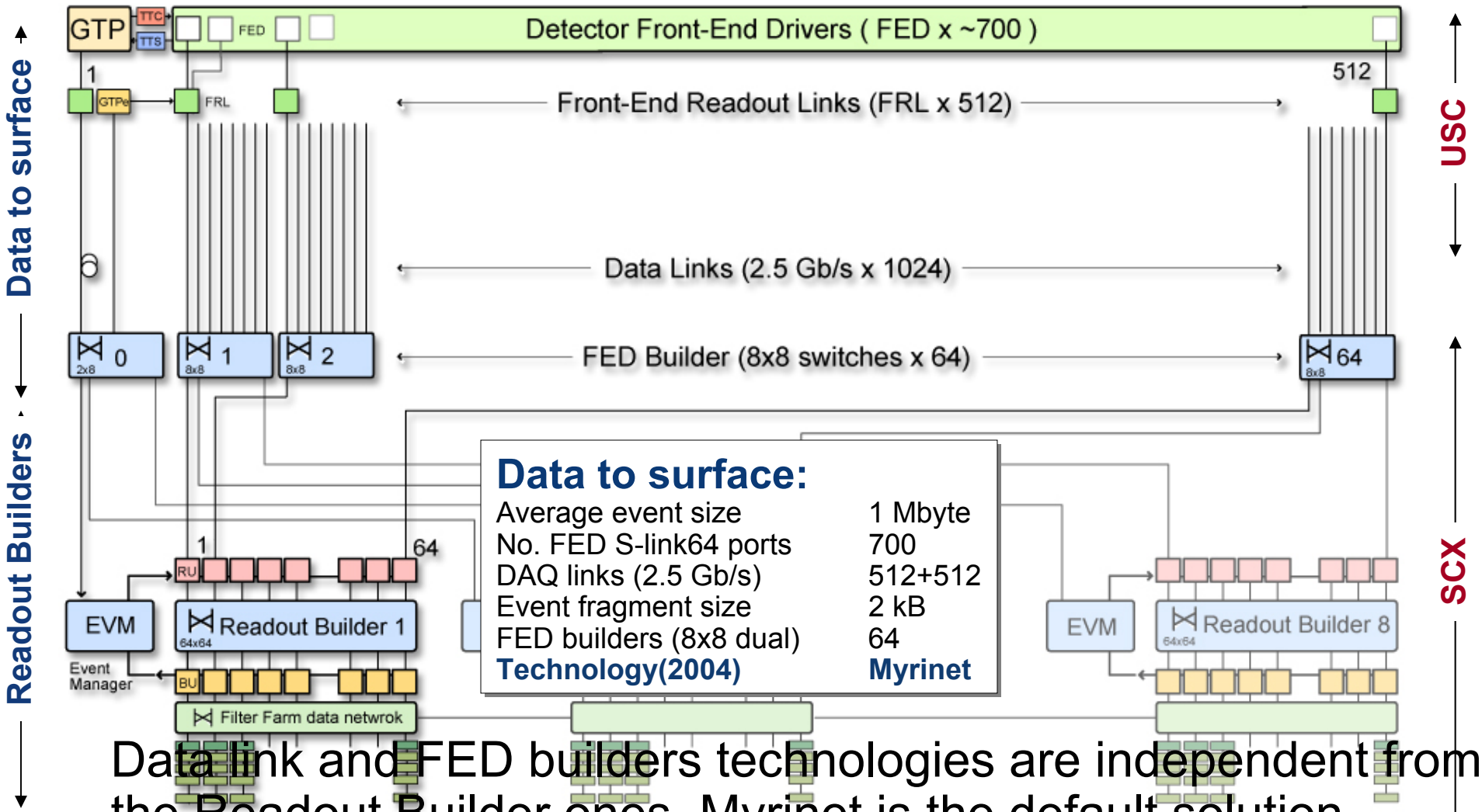


DAQ staging : 1 to 8 RBs = 100 kHz



12.5 kHz **+12.5 kHz** **+12.5 kHz**

I) Data to surface (D2S)

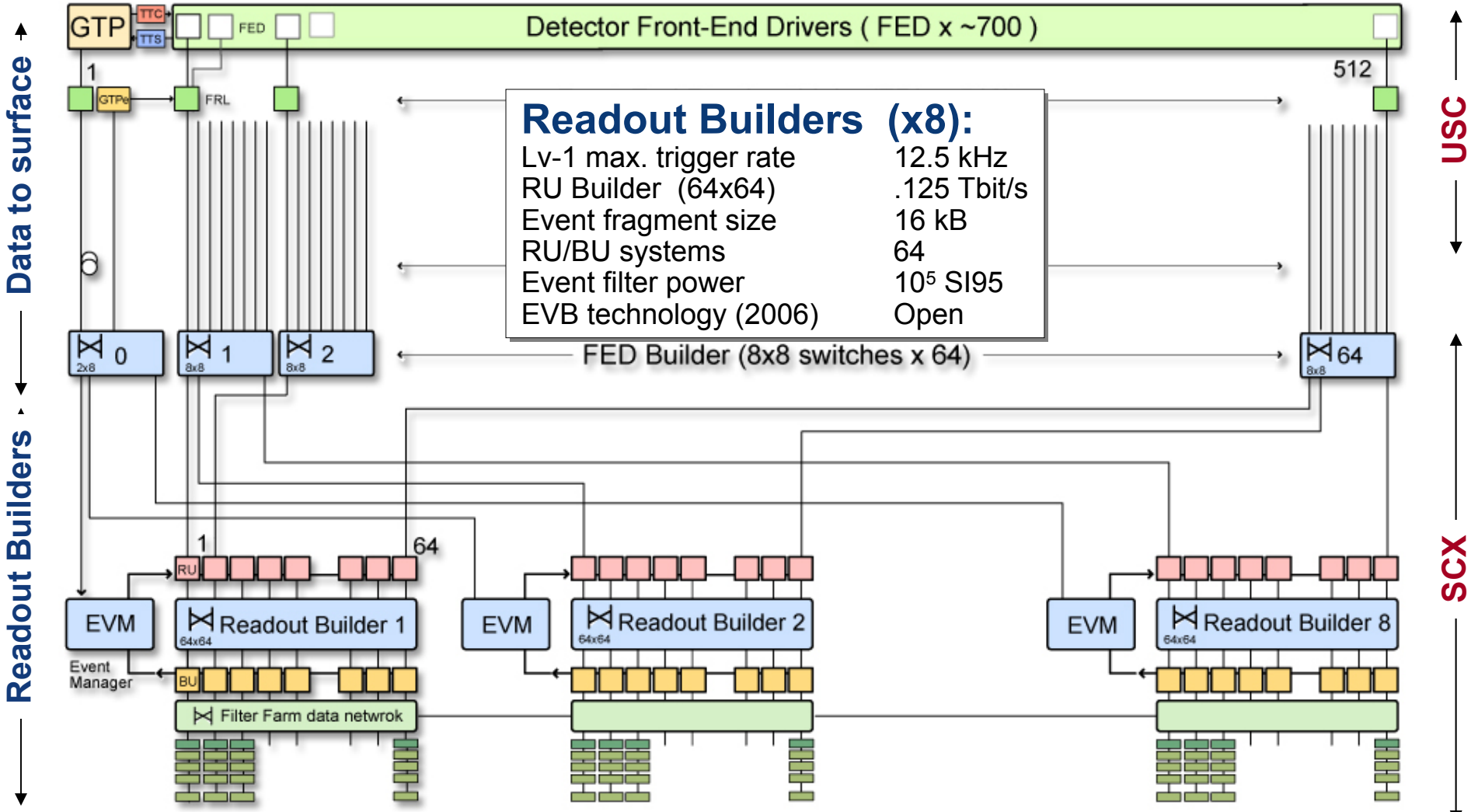


Data to surface:

Average event size	1 Mbyte
No. FED S-link64 ports	700
DAQ links (2.5 Gb/s)	512+512
Event fragment size	2 kB
FED builders (8x8 dual)	64
Technology(2004)	**Myrinet**

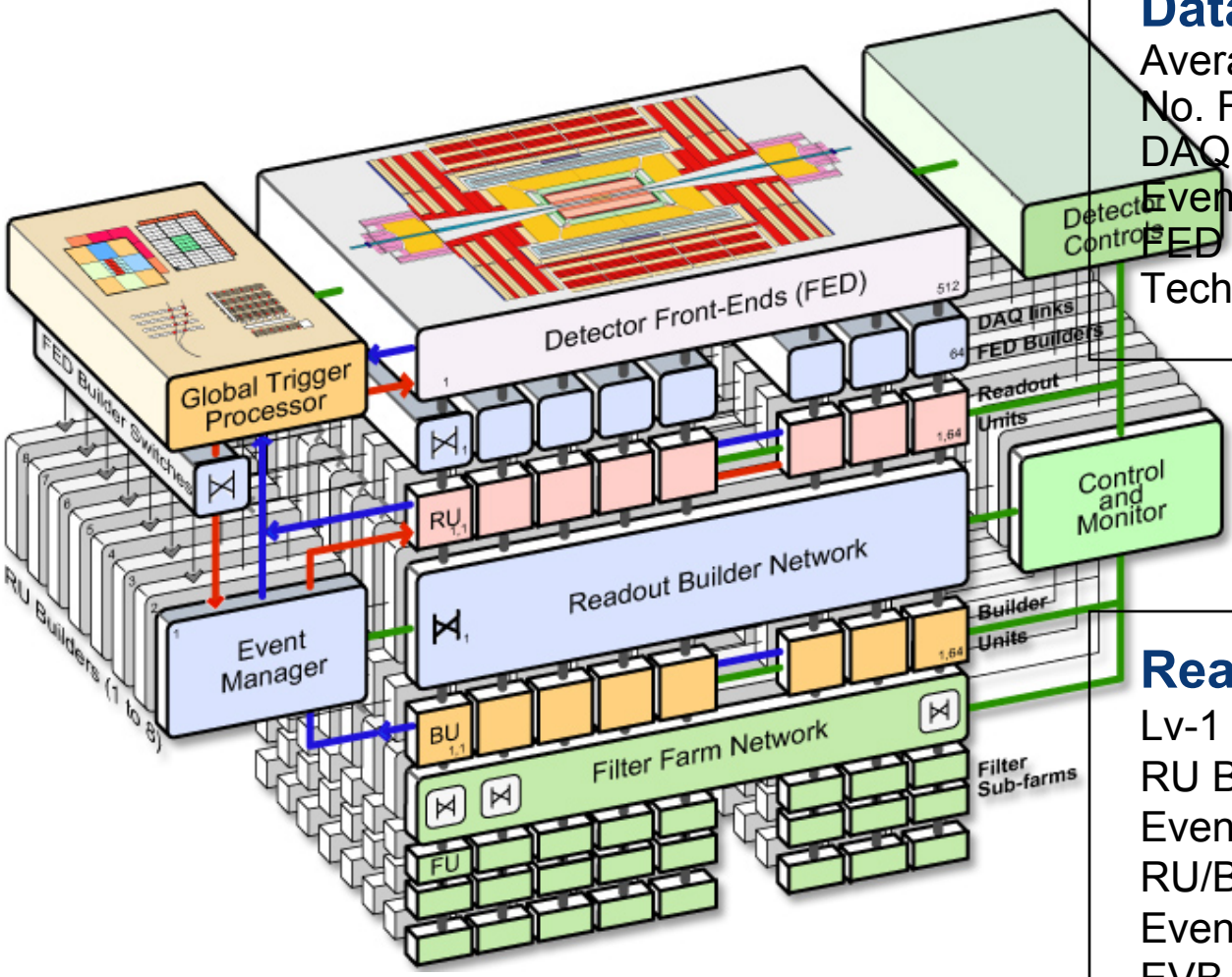
Data link and FED builders technologies are independent from the Readout Builder ones. Myrinet is the default solution. Decision must be taken in 2004. Installation in 2005

II) Readout Builders (RB)



Readout builders technologies are independent from the D2S one. Decision will be taken later in 2006

8-fold DAQ 3-D design



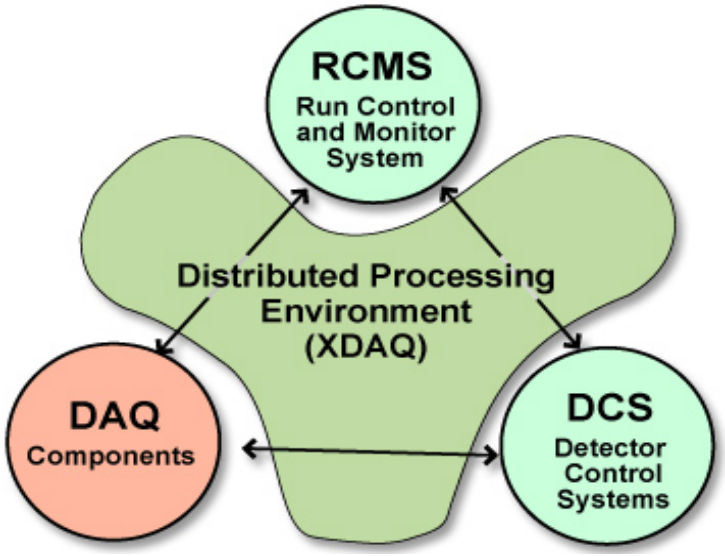
Data to surface:

Average event size	1 Mbyte
No. FED S-link64 ports	700
DAQ links (2.5 Gb/s)	512+512
Event fragment size	2 kB
FED builders (8x8 dual)	64
Technology(2004)	Myrinet

Readout Builders (x8):

Lv-1 max. trigger rate	12.5 kHz
RU Builder (64x64)	.125 Tbit/s
Event fragment size	16 kB
RU/BU systems	64
Event filter power	10^5 SI95
EVB technology (2006)	Open

On-line software : framework and subsystems

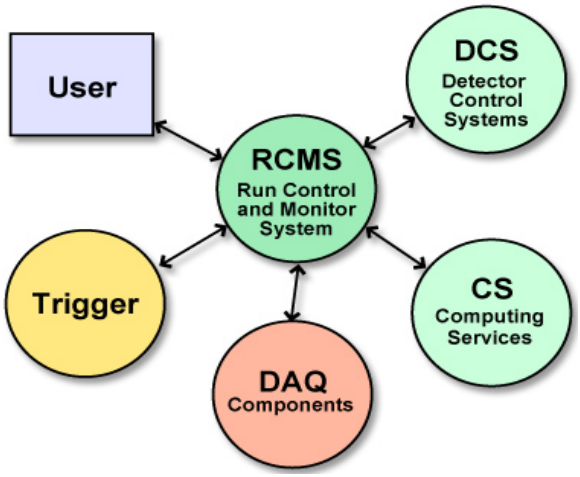


Online software architecture :

- **Cross-platform DAQ framework: XDAQ**
- **Data acquisition components**
- **Run Control and Monitor System (RCMS)**
- **Detector Control System (DCS)**

The RCMS, DCS and data acquisition components interoperate through a distributed processing environment called XDAQ (cross-platform DAQ framework)

Configuration, operation and monitoring

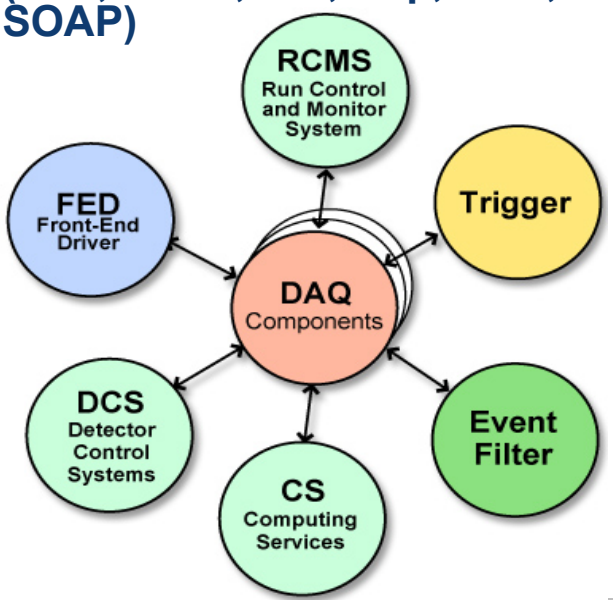


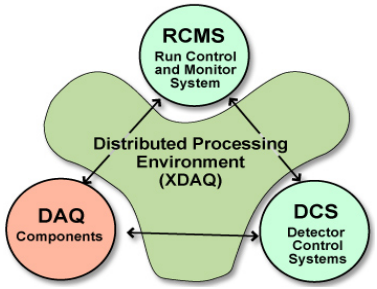
RCMS: Run Control and Monitoring System

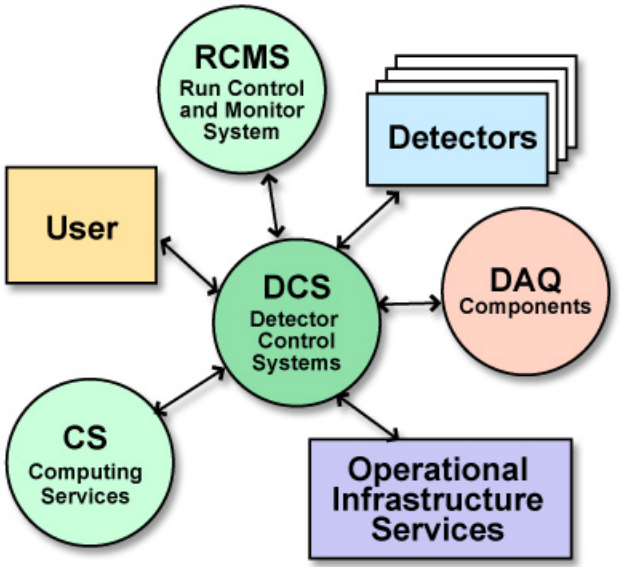
Based on open protocols, web services and emerging e-tools tools **(JAVA, http, XML, MySQL, ….)**

XDAQ:on-line framework and DAQ components:

- Services and tools for local and remote inter-process communication, configuration and control and data storage
Components to build data acquisition systems (RU,BU,EVM,..)
(C++, JAVA, I2O, http, XML, SOAP)





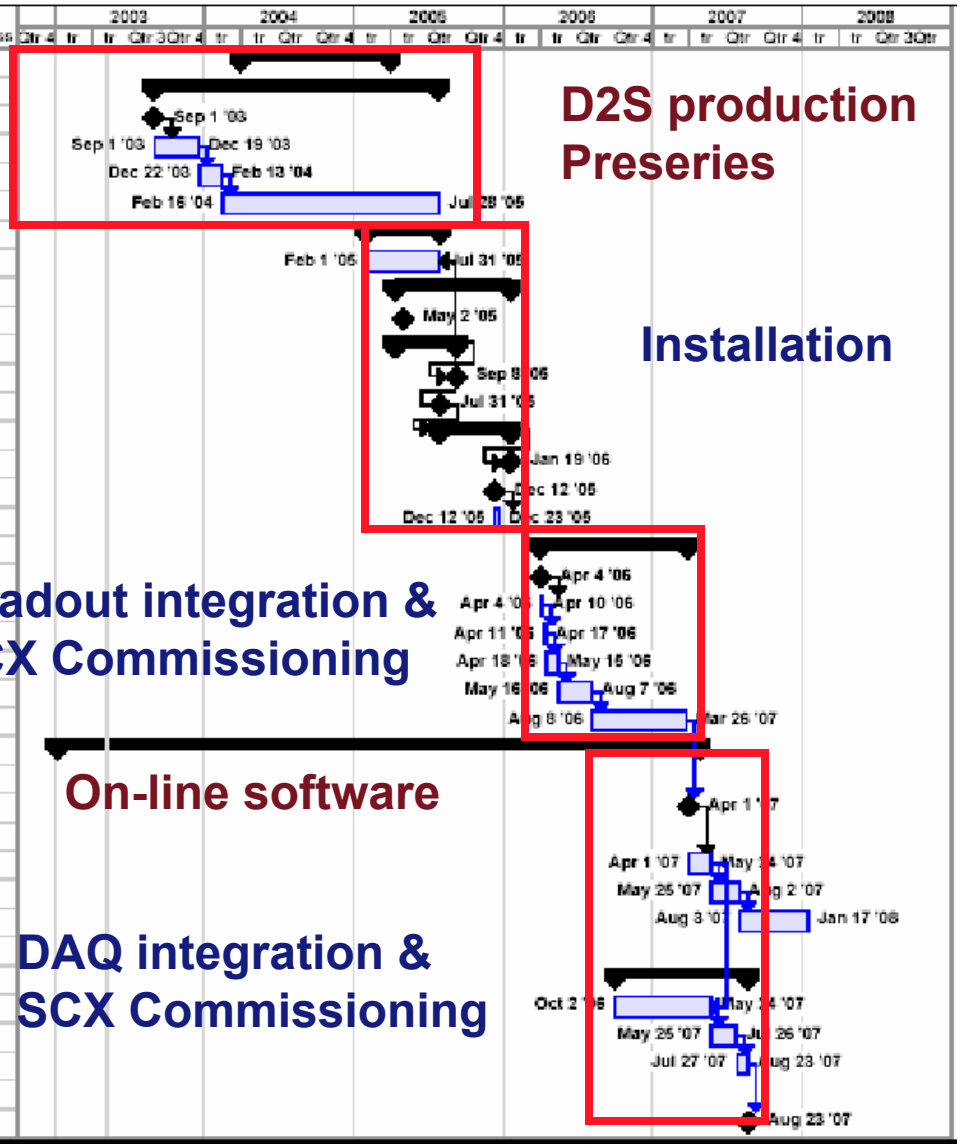


DCS: Detector Control Systems

Based on industry supported hardware and software
(PLC, field buses, PVSS and JCOP tools)

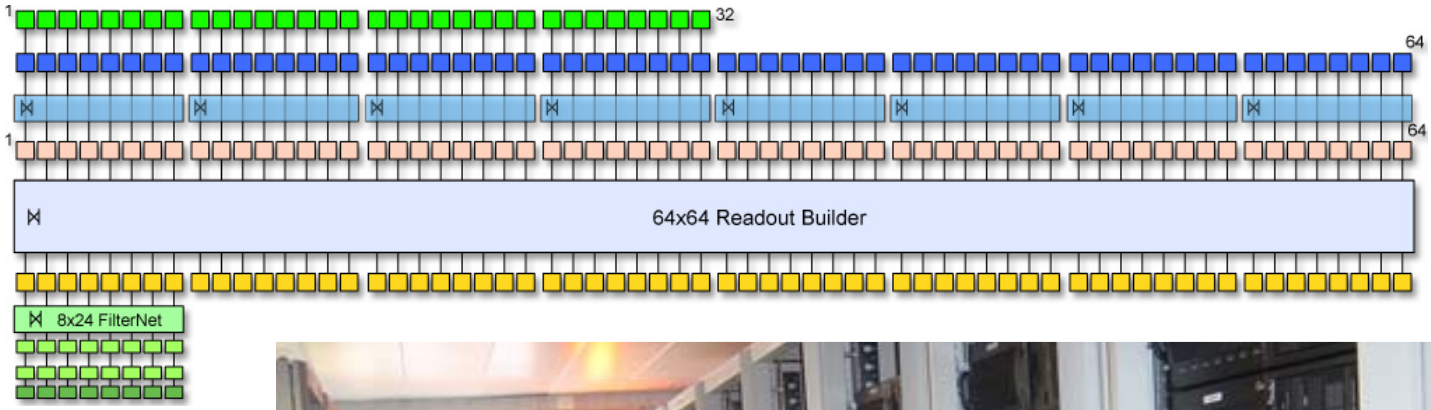
DAQ raw schedule

ID	Task Name	Duration	Start	Finish	Predecess
1	**1 D2S full production**	**260 days**	**Apr 1 '04**	**Mar 30 '05**	
29	**2 Preserie**	**600 days**	**Sep 1 '03**	**Jul 28 '05**	
30	2.1 Approval/decision	0 days	Sep 1 '03	Sep 1 '03	
31	2.2 Procurement	80 days	Sep 1 '03	Dec 19 '03	30
32	2.3 Install	40 days	Dec 22 '03	Feb 13 '04	31
33	2.4 Debug/exploitation	350 days	Feb 16 '04	Jul 28 '05	32
34	**3 First readout builder (12.5kHz)**	**130 days**	**Feb 1 '05**	**Jul 31 '05**	
35	3.1 Procurement (1 rb)	130 days	Feb 1 '05	Jul 31 '05	44SF
36	**4 Installation**	205 days	**Apr 12 '05**	**Jan 19 '06**	
37	4.1 USC55 ready for crate install	0 days	May 2 '05	May 2 '05	
38	**4.2 DAQ install in USC55**	**110 days**	**Apr 12 '05**	**Sep 8 '05**	
43	4.3 DAQ install in USC done: wait for FRL/FED int	0 days	Sep 8 '05	Sep 8 '05	38
44	4.4 SCX ready for DAQ install	0 days	Jul 31 '05	Jul 31 '05	
45	**4.5 DAQ install/internal integration in SCX**	**125 days**	**Jul 31 '05**	**Jan 19 '06**	44
52	4.6 DAQ install in SCX done	0 days	Jan 19 '06	Jan 19 '06	45
53	4.7 Fibers from USC to SCX installed	0 days	Dec 12 '05	Dec 12 '05	
54	4.8 Communication test USC/SCX	10 days	Dec 12 '05	Dec 23 '05	53,40
55	**5 Integration/commisionning**	**255 days**	**Apr 4 '06**	**Mar 26 '07**	
56	5.1 Trigger ready for DAQ integration	0 days	Apr 4 '06	Apr 4 '06	
57	5.2 GTP->EVM communication	5 days	Apr 4 '06	Apr 10 '06	56
58	5.3 GTP readout via FRLs	5 days	Apr 11 '06	Apr 17 '06	57
59	5.4 Trigger detectors readout via FRLs	20 days	Apr 18 '06	May 15 '06	58
60	5.5 Other detector readout readout via FRLs	60 days	May 16 '06	Aug 7 '06	59
61	5.6 General TRIDAS/Detector integration	165 days	Aug 8 '06	Mar 26 '07	60
62	**6 On-line software**	**1127 days**	**Jan 7 '03**	**Apr 27 '07**	
71					
72	7 CMS Ready to take data (12.5kHz)	0 days	Apr 1 '07	Apr 1 '07	61
73					
74	8 LHC pilot run (tentative)	40 days	Apr 1 '07	May 24 '07	72
75	9 LHC shutdown (tentative)	50 days	May 25 '07	Aug 2 '07	74
76	10 LHC physics run (tentative)	120 days	Aug 3 '07	Jan 17 '08	75
77					
78	**11 DAQ 50 kHz**	**235 days**	**Oct 2 '06**	**Aug 23 '07**	
79	11.1 Procurement (3rbs)	170 days	Oct 2 '06	May 24 '07	74FF
80	11.2 Install	45 days	May 25 '07	Jul 26 '07	79
81	11.3 DAQ 50 kHz (4 rbs) integ/debug	20 days	Jul 27 '07	Aug 23 '07	80
82					
83	12 DAQ 50kHz running	0 days	Aug 23 '07	Aug 23 '07	81



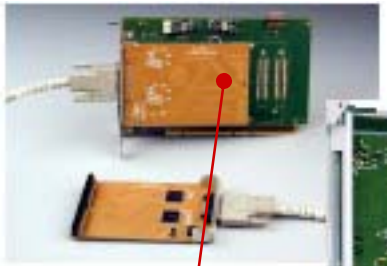
2004 P5 green barrack: Pre-series RB

-32 GII-FED emulators
-64 FRLs
-13 Water cooled racks
-93 PC dual-CPU
-D2S Myrinet equipment
-Readout Builder Myrinet
-16 PC Filter Farm

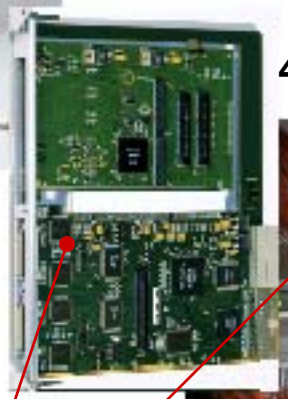


Preseries integration programme

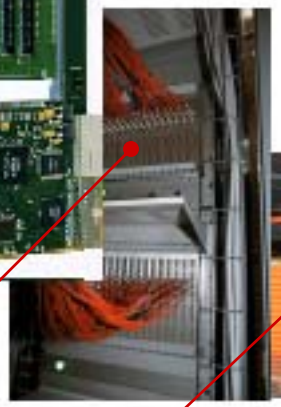
32 GIII FED emulators



64 FRLs



4 FRL crates

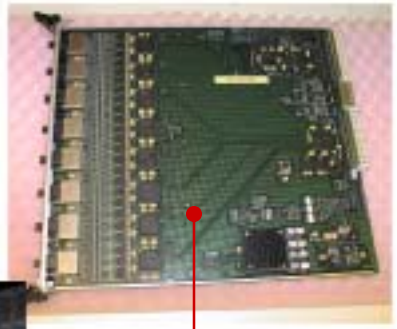


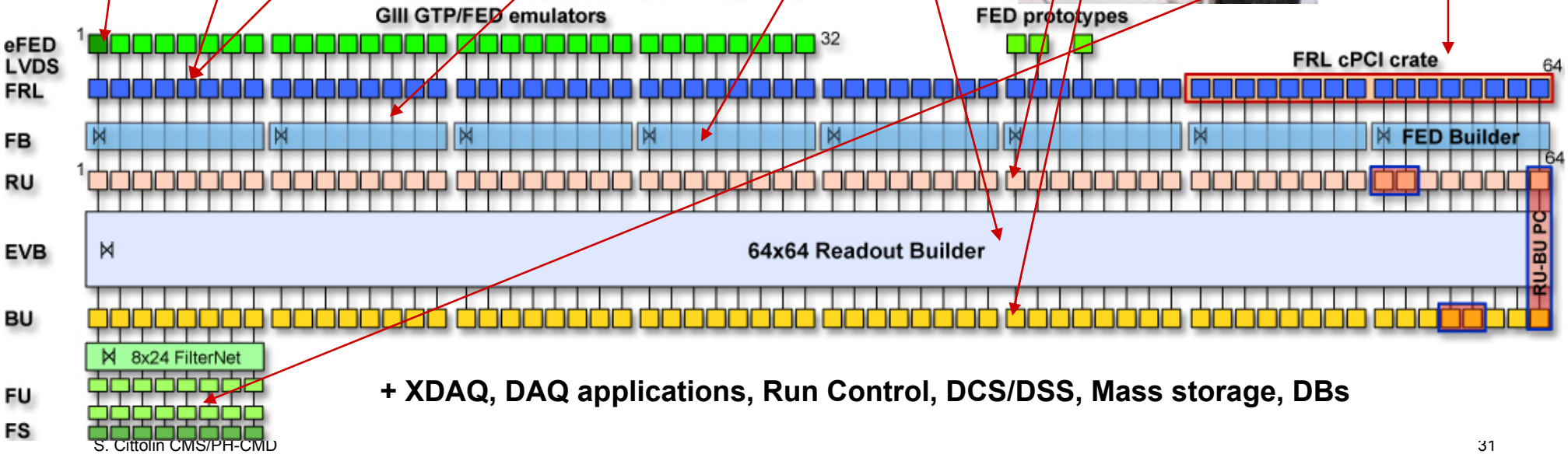
200m fibers
Data links

8 8x8 FED builders

64x64 RU builders
16 FUs Filter SF

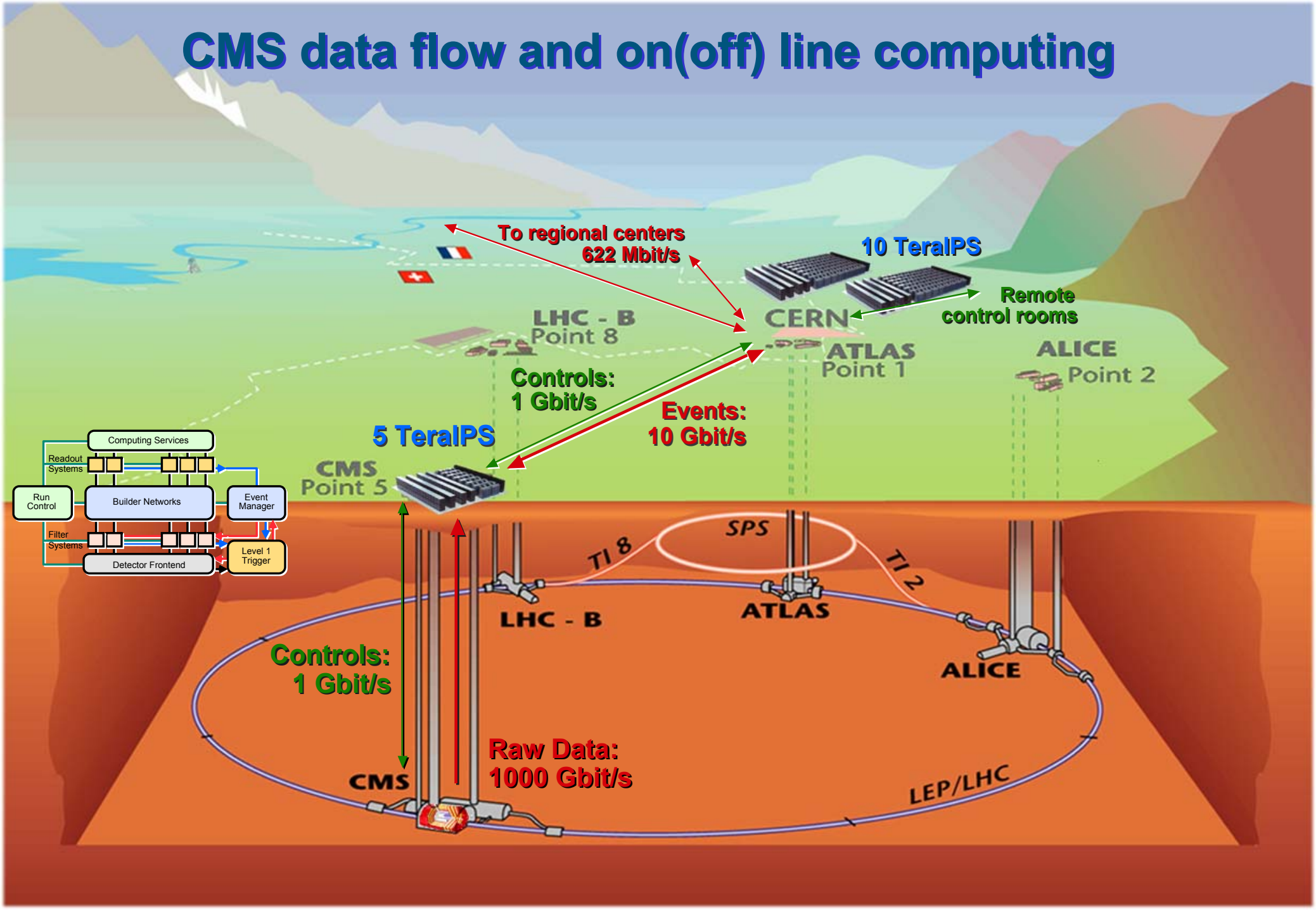
Up to 32 detector FED



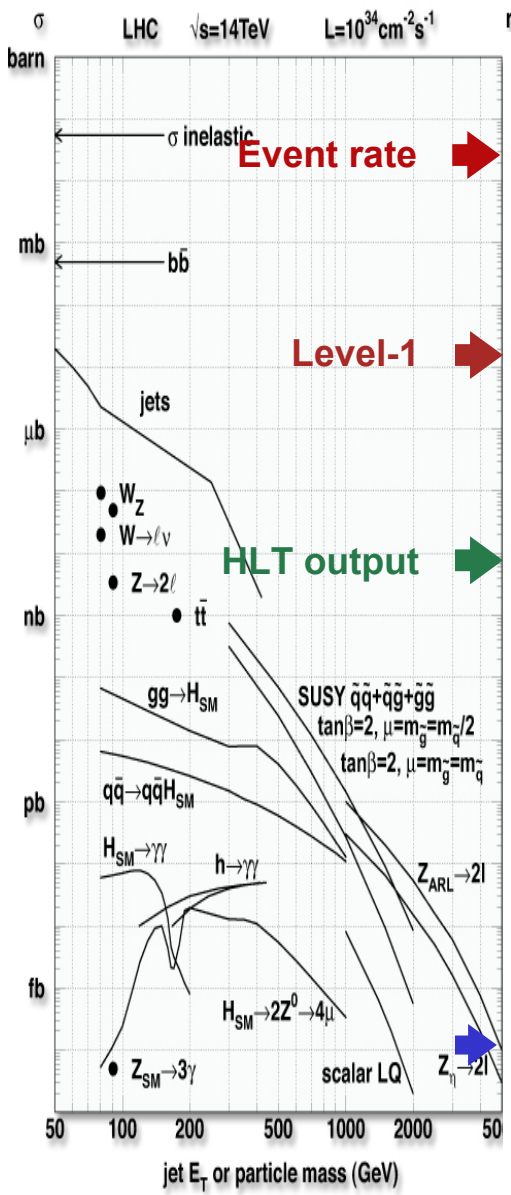


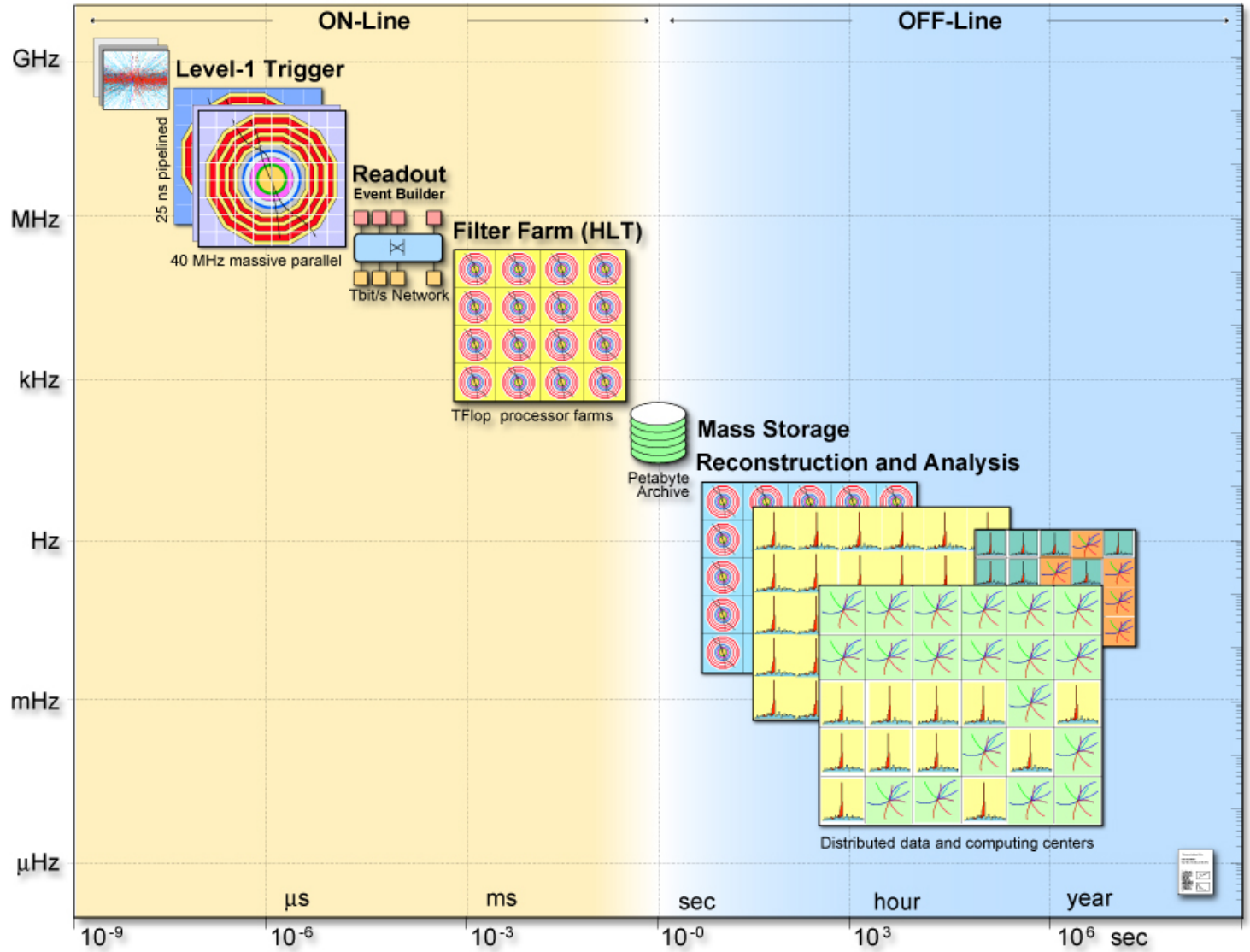
+ XDAQ, DAQ applications, Run Control, DCS/DSS, Mass storage, DBs

CMS data flow and on(off) line computing



DAQ data flow and computing model





Summary design principles

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

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

- **Rely on hardware and software industry standards**
  - Custom/**standards** (PCI, Ethernet, C++, JAVA, http, XML,..)

Conclusions

The CMS design fulfils the major requirements:

- ✓**100 KHz level-1 readout**
- ✓**Event builder:**
  a scalable structure that can go up to 1 Terabit/s
- ✓**High-Level Trigger by fully programmable processors**

•This **design** should be considered **complete,** but not final.
e.g. switches procured in 2008-09 can be different from those of the startup system

•It is a **system** that is **expected to change** with time, accelerator and experiment conditions. And it has been designed to do so

•It is conceived to provide the **maximum possible flexibility** to execute a physics selection on-line