LHC Event Building Systems

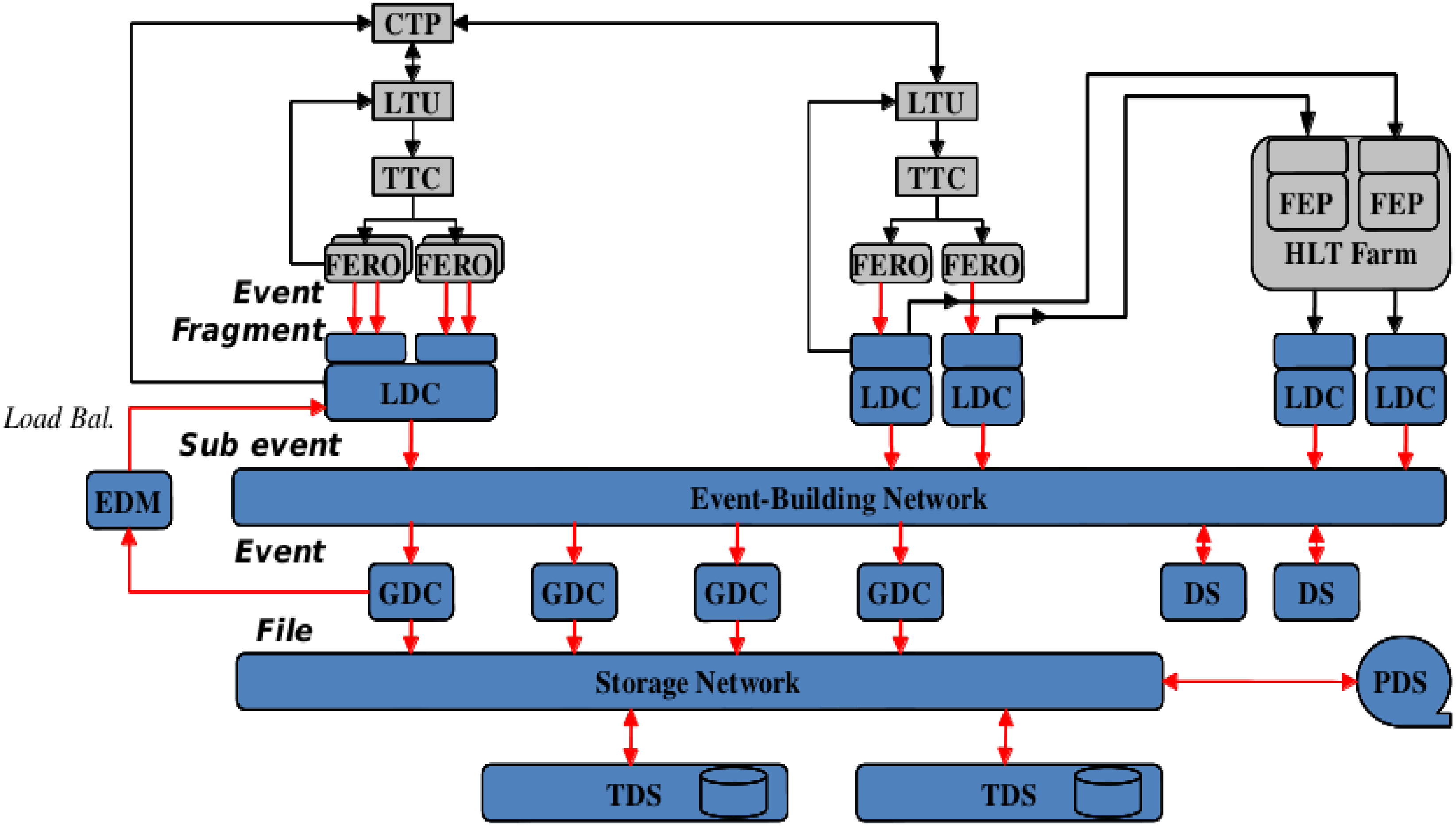
DAQ@LHC, March 12th 2013

W.Vandelli CERN/PH-ATD

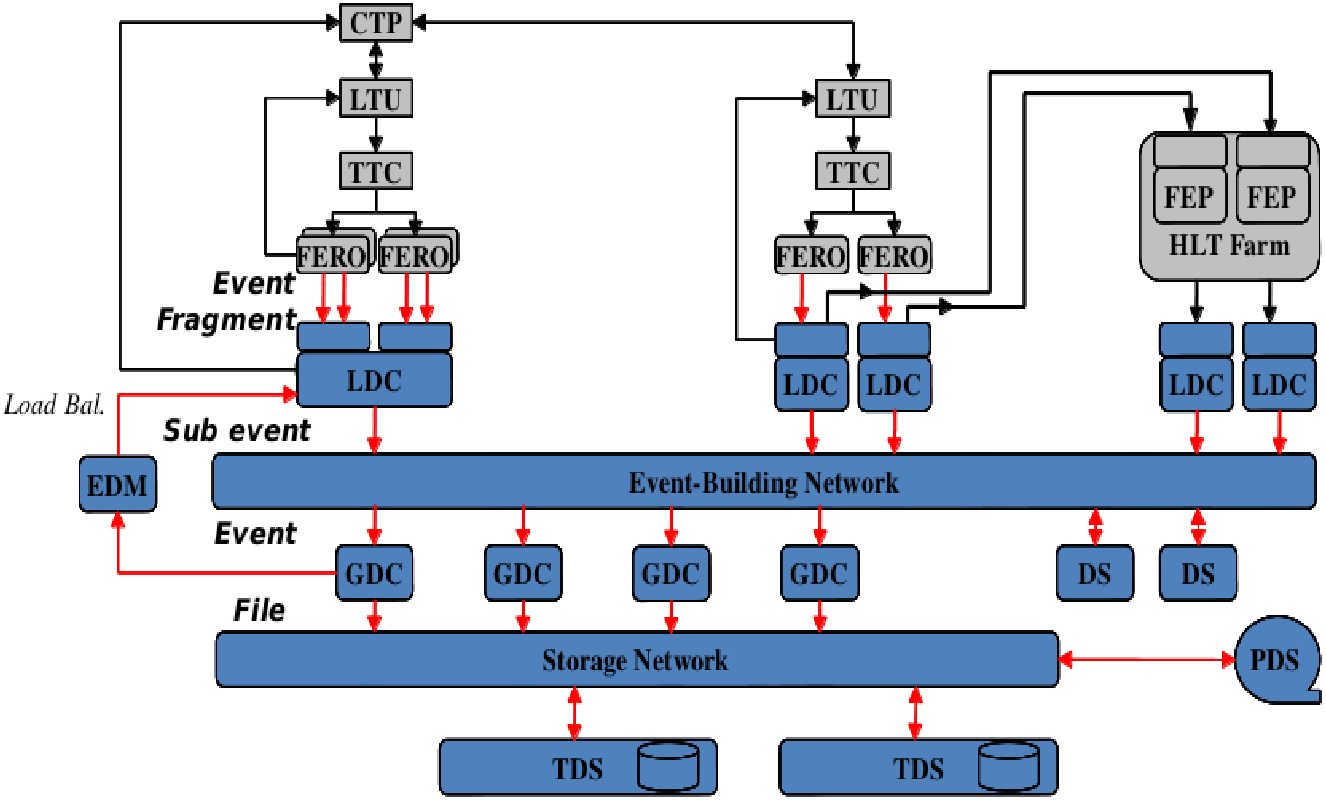
Outline

- ➔ The Fantastic Four
  - Architecture & Implementation
- ➔ Scaling & Performance
- ➔ Fault tolerance & Heterogeneity
- ➔ Conclusion

ALICE

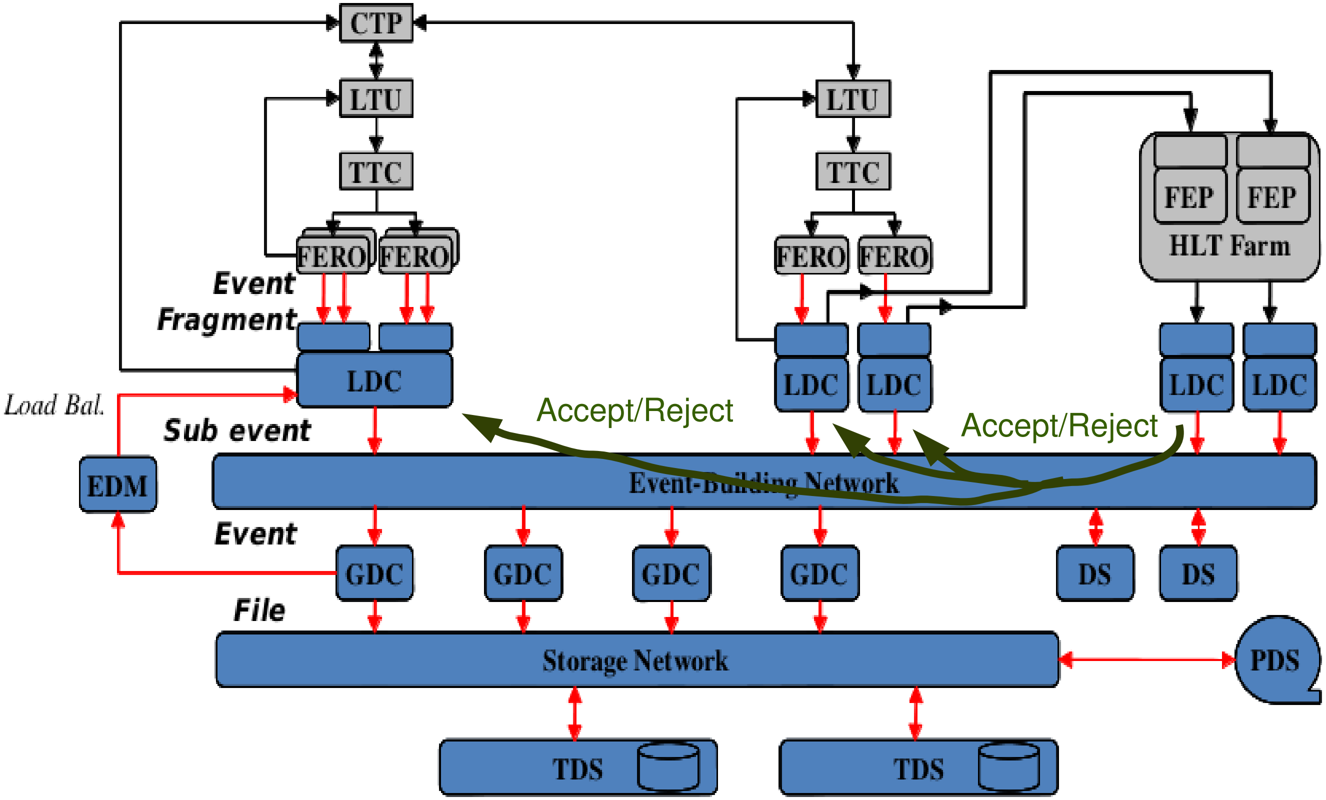

CTP
LTU
TTC
FERO
FERO
Event
Fragment
LDC
Load Bal.
Sub event
EDM
LTU
TTC
FERO
FERO
LDC
LDC
FEP
FEP
HLT Farm
LDC
LDC
Event-Building Network
Event
GDC
GDC
GDC
GDC
DS
DS
File
Storage Network
PDS
TDS
TDS

ALICE explained



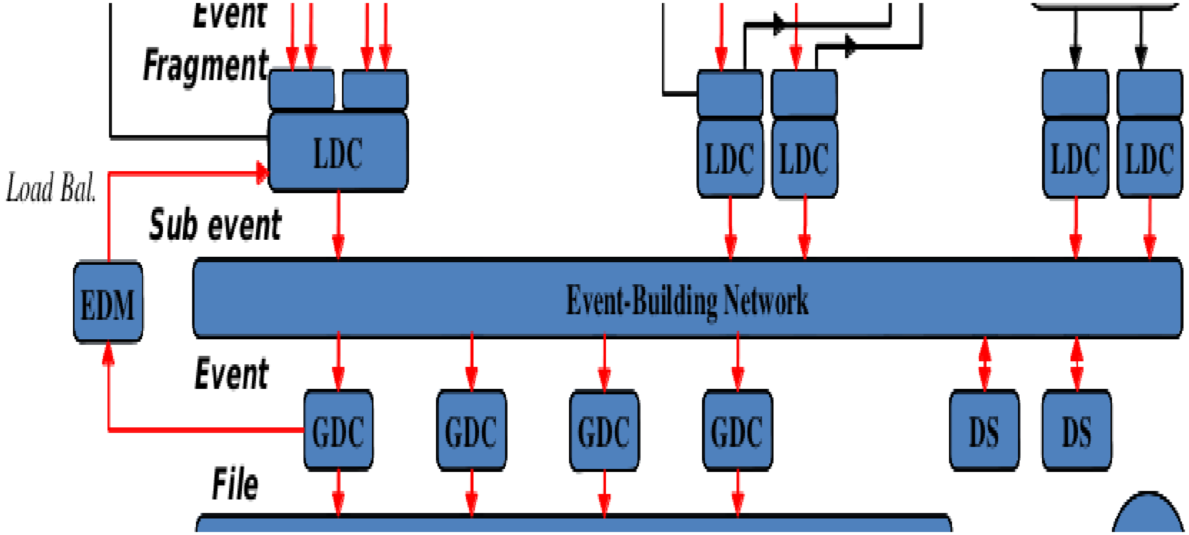
- ➔ How does the HLT work?

ALICE explained



- ➔ How does the HLT work?
- ➔ **HLT decision propagated to LDCs**
  - via EB network
  - large buffering using PC memory
- ➔ HLT also performs (TPC) data reduction
  - for most events HLT LDC fragments replace detector LDC fragments
  - driven by event type tag
  - **event-by-event GDCs know involved LDCs**

ALICE Event Builder



➔ Push protocol using TCP/IP

- LDC are PCs housing custom cards

➔ EDM not used

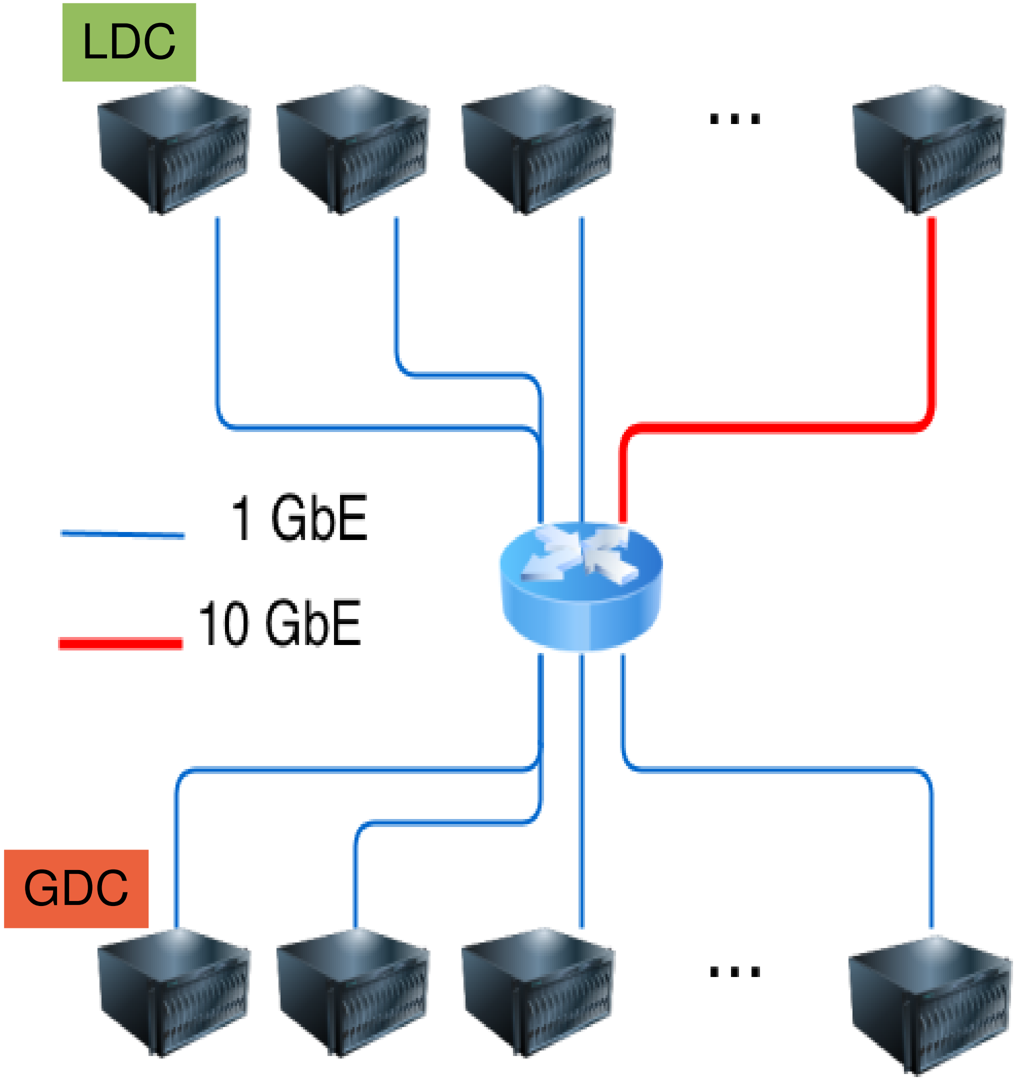
- LDCs event ID-based round-robin over independent streams

➔ Full event content depends on event type and HLT decision

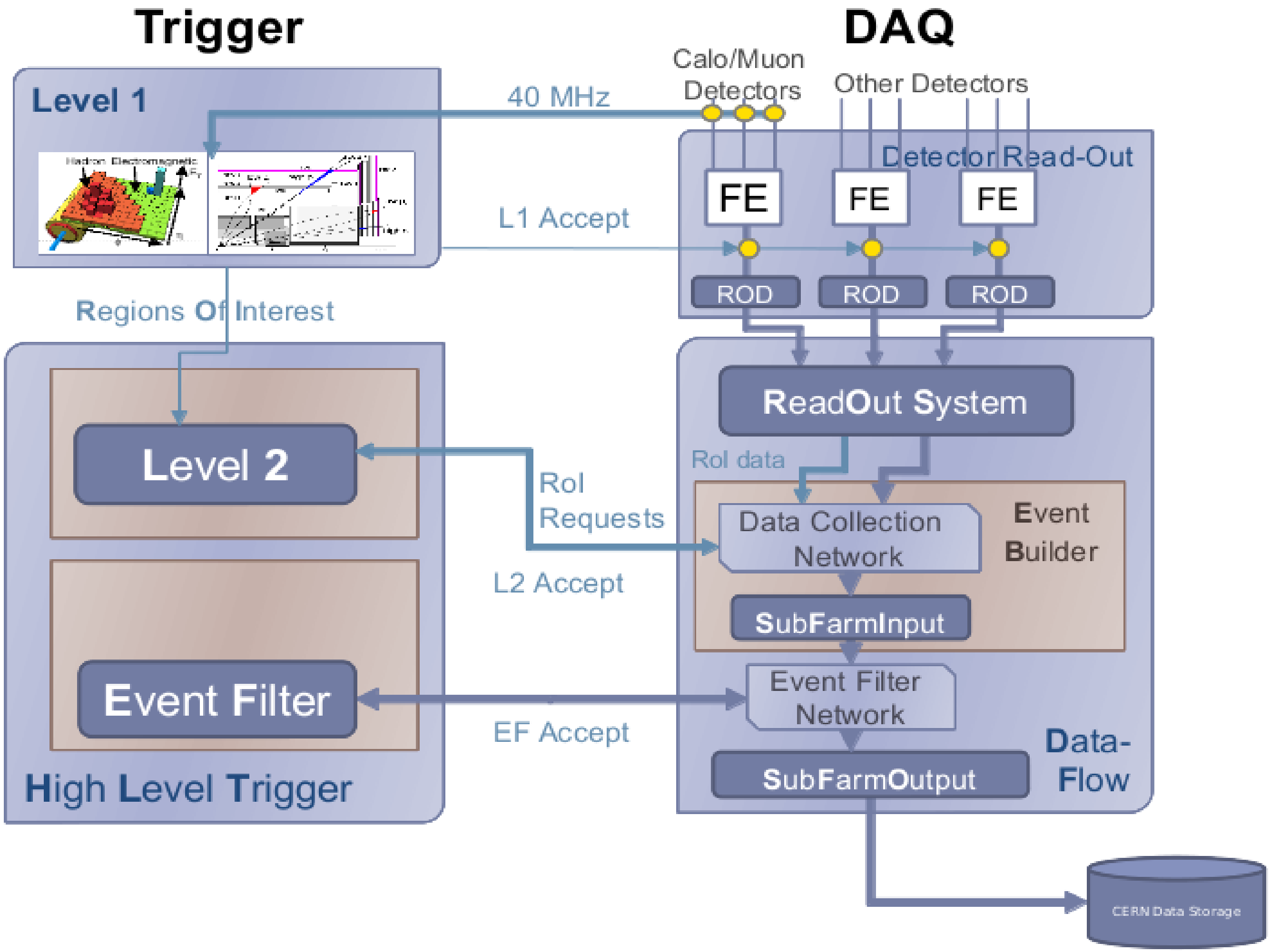
➔ Full events handed over to local streaming/objectification/writing tasks via shared memory

ALICE Design Parameters

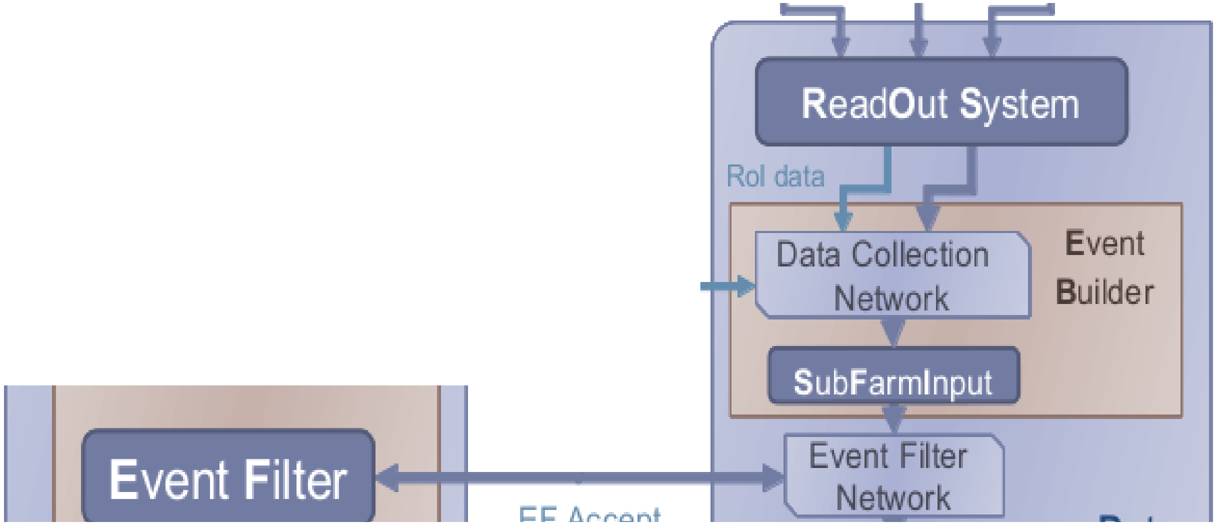
- ➔ 185 data sources (LDC)
  - variable fragment size
    - detector dependent
  - 1GbE/10GbE
- ➔ 85 builder units (GDC)
  - 1GbE
- ➔ EB rate & bandwidth
  - HI Central
    - 39MB/event@40Hz = 1.5 GB/s
  - HI Dimuon
    - 250kB/event@1kHz = 0.25 GB/s
  - pp
    - 500kB/event@1kHz = 0.5 GB/s



ATLAS


Trigger
DAQ
Level 1
40 MHz
Calo/Muon Detectors
Other Detectors
Detector Read-Out
FE
FE
FE
L1 Accept
ROD
ROD
ROD
Regions Of Interest
ReadOut System
Level 2
RoI data
RoI Requests
Data Collection Network
Event Builder
L2 Accept
SubFarmInput
Event Filter Network
Event Filter
EF Accept
Data-Flow
SubFarmOutput
High Level Trigger
CERN Data Storage

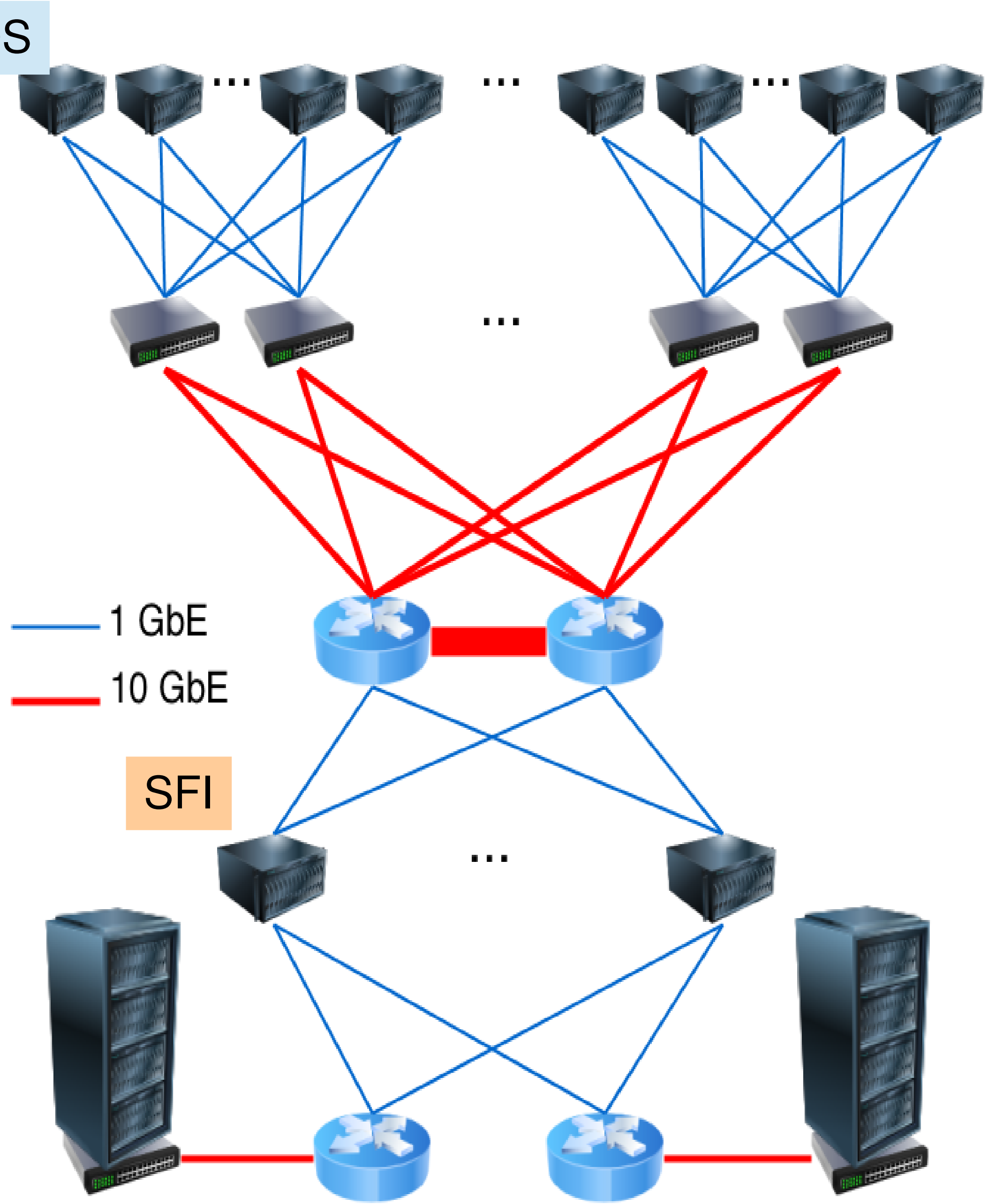
ATLAS Event Builder



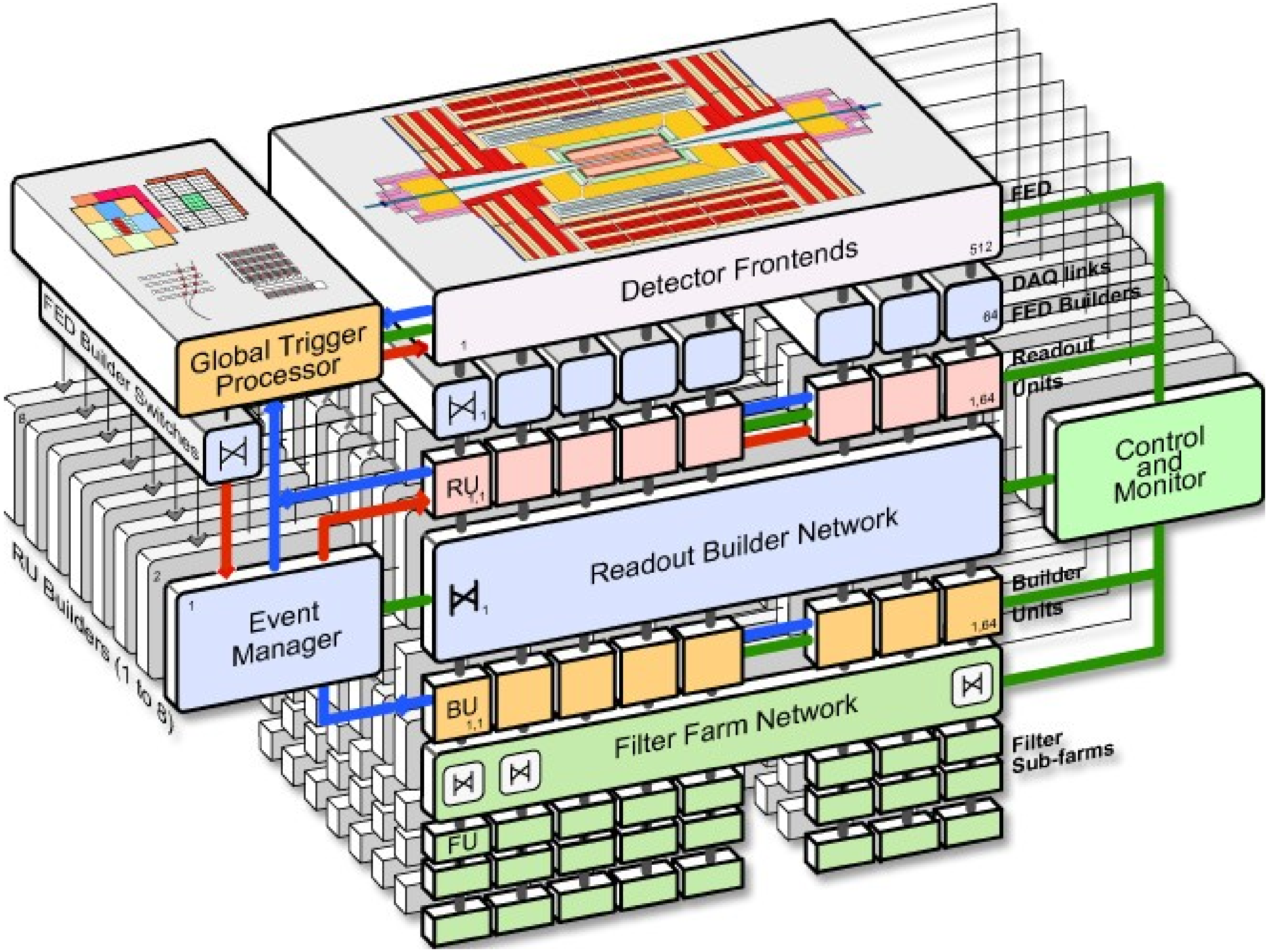
- ➔ Pull protocol using TCP over Ethernet
  - UDP possible, never used in production
- ➔ Event building manager
  - Educated round-robin based on occupancy and XON/XOFF messages
- ➔ Traffic shaping
- ➔ Full events handed to HLT farm using TCP connections over a second network

ATLAS Design Parameters

- ➔ 150 data sources (ROS)
  - Average fragment size ~10 kB
    - detector dependent
  - 2x 1GbE
  - 10 GbE up-links - redundancy
- ➔ 100 builder units (SFI)
  - 1GbE
  - later dual builder units
- ➔ ~1000 data destinations (HLT)
- ➔ EB rate 3.5 kHz
- ➔ EB bandwidth 5.25 GB/s
- ➔ ~1.5MB/event

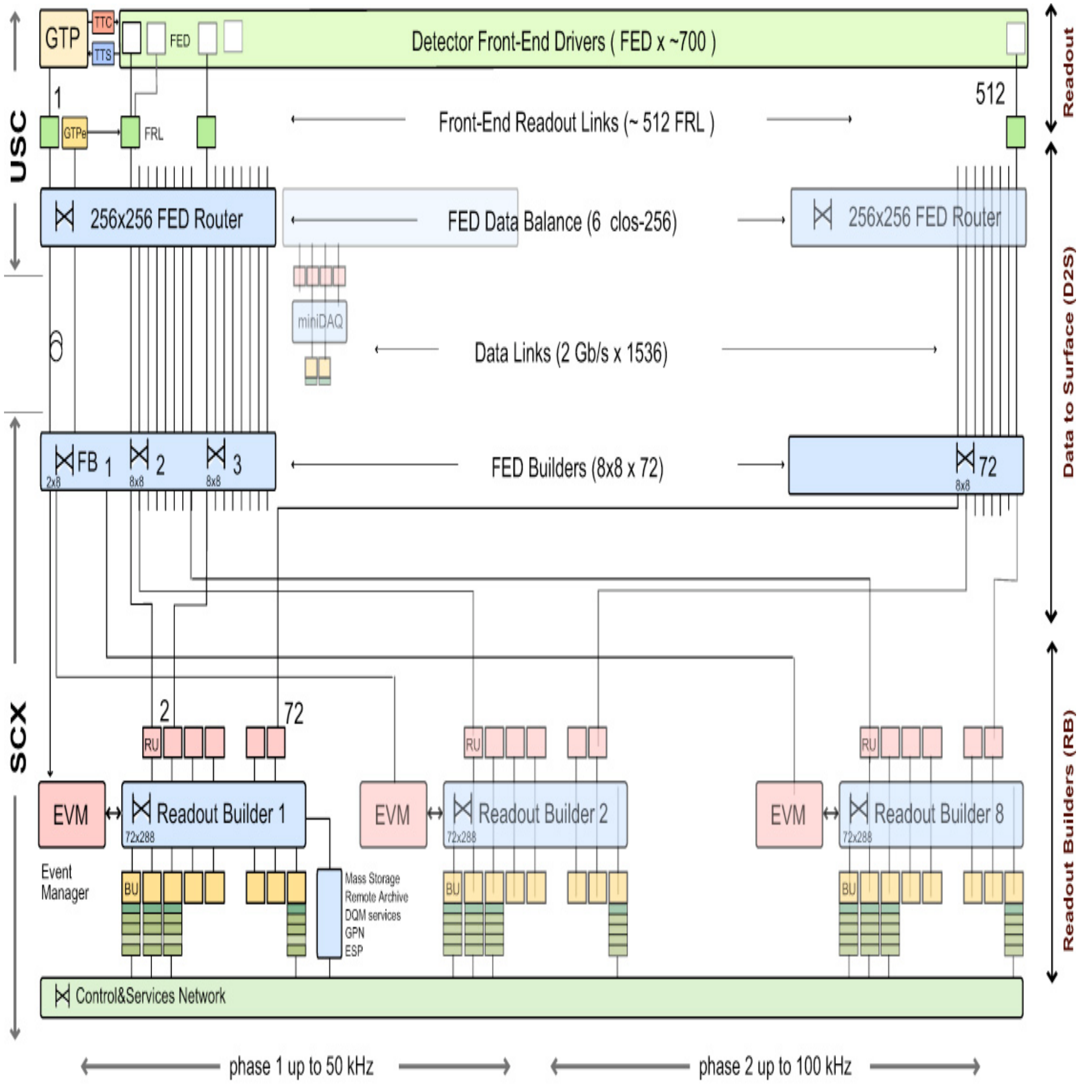


CMS


Detector Frontends
512
1
FED
DAQ links
FED Builders
64
Readout
Units
1,64
Global Trigger
Processor
FED Builder Switches
RU
1,1
Readout Builder Network
Control
and
Monitor
Event
Manager
1
2
8
RU Builders (1 to 8)
Builder
Units
1,64
BU
1,1
Filter Farm Network
Filter
Sub-farms
FU

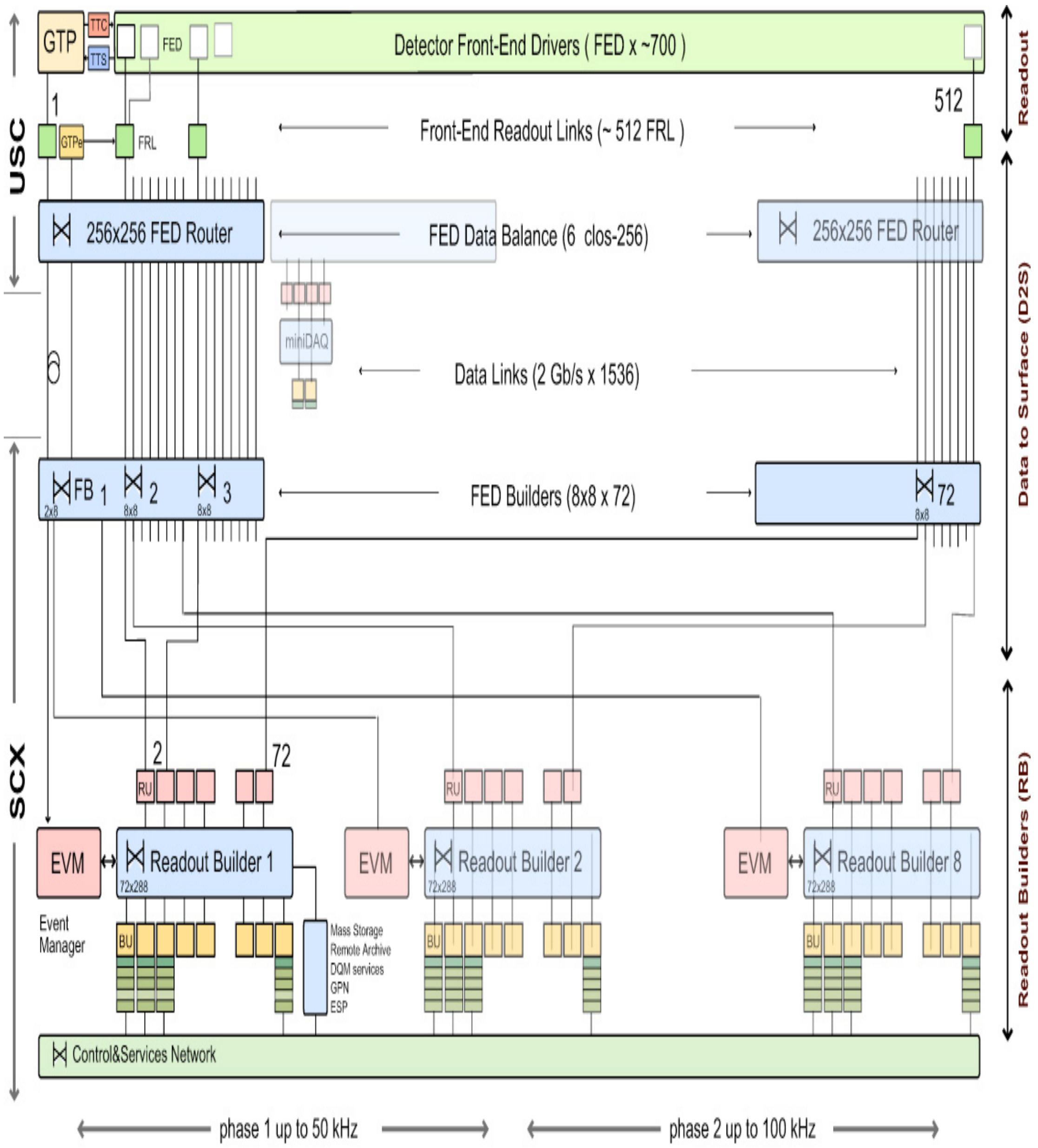
CMS Event Builder

- ➔ Push-pull protocol over a Myrinet + Ethernet networks
  - sliced HLT farm
- ➔ 1st step
  - programmable NICs perform
    - statically-weighted round-robin
    - super-fragment assembly
- ➔ 2nd step
  - TCP/IP
  - dynamic distribution by event manager per slice
- ➔ Event building distributed over the HLT farm
  - full events locally assigned for processing

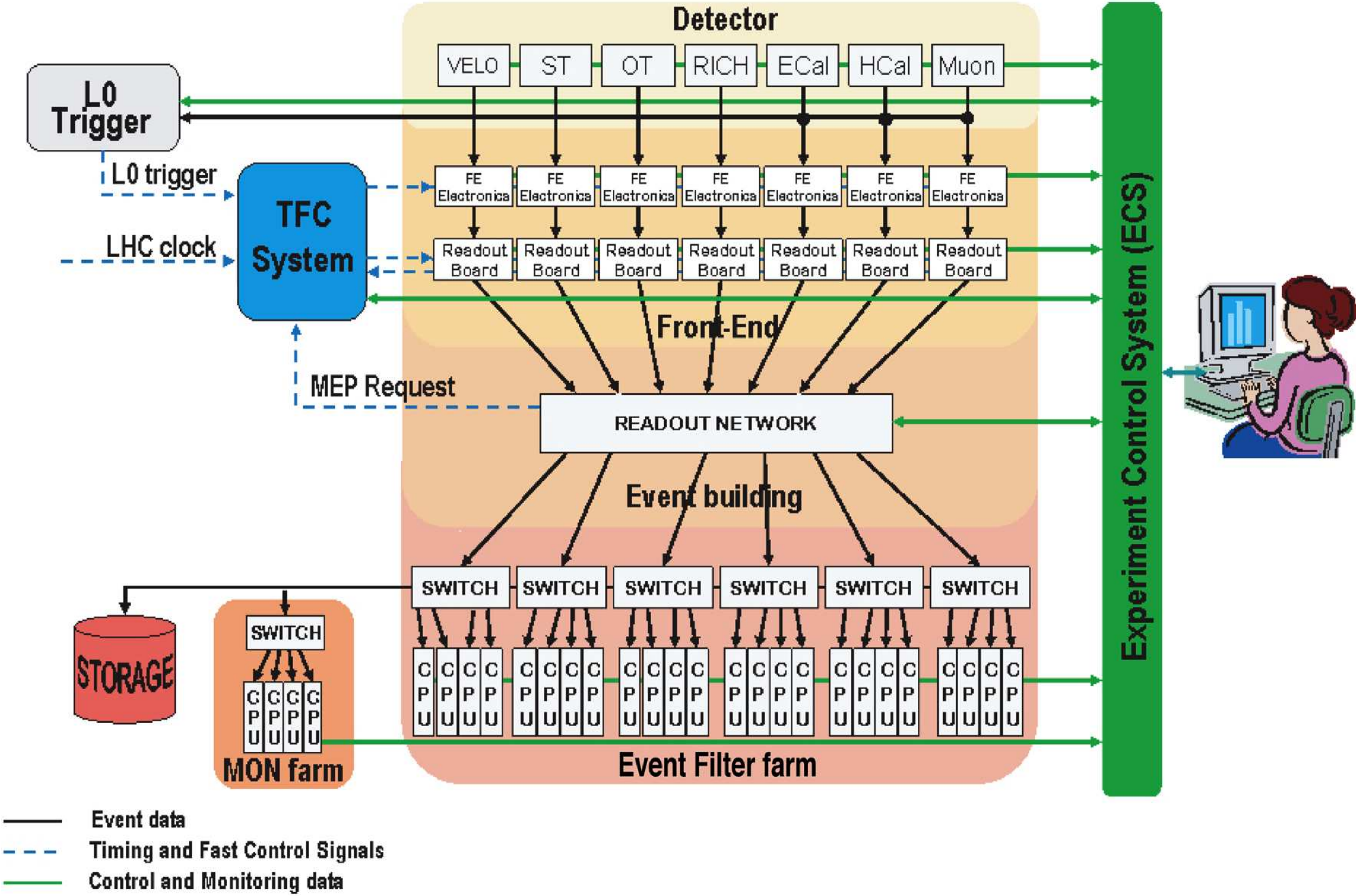


CMS Design Parameters

- ➔ 512 data sources (FRL)
  - Average fragment size ~ 2kB
  - 2x 2Gb Myrinet - redundancy
- ➔ 8x64 = 512 intermediate builder units (RU)
  - Average fragment size ~ 16kB
  - 2x 2Gb Myrinet + 3x 1GbE
- ➔ 8x90 = 720 builder units (BU)
  - 1GbE
- ➔ EB rate 100 kHz
- ➔ EB bandwidth 100 GB/s
  - ~1MB/event



LHCb


Detector
VELO
ST
OT
RICH
ECal
HCal
Muon
L0 Trigger
L0 trigger
LHC clock
TFC System
FE Electronics
Readout Board
Front-End
MEP Request
READOUT NETWORK
Event building
SWITCH
CPU
STORAGE
MON farm
Event Filter farm
Experiment Control System (ECS)
Event data
Timing and Fast Control Signals
Control and Monitoring data

LHCb Event Builder

➔ Push protocol using UDP over Ethernet

- Readoud Boards are FPGA-based custom-cards

➔ Multi-event fragment (MEP)

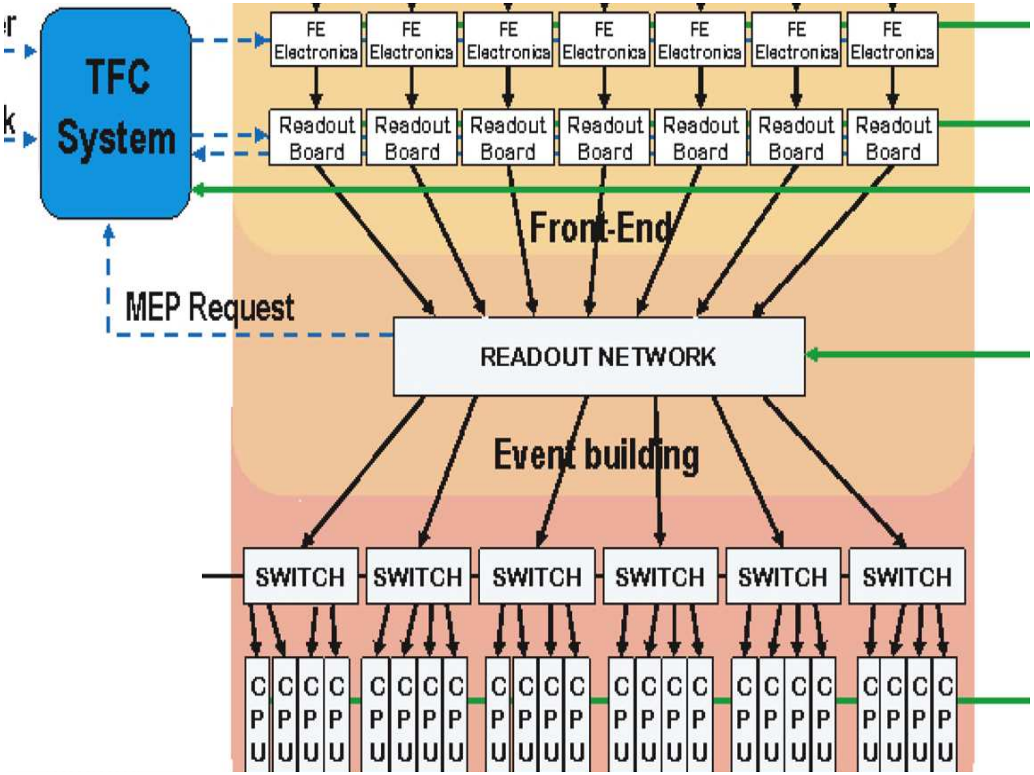
- reduce overhead due to small event size

➔ Educated round-robin

- Builder availability distributed over spare TTC channel

➔ Event building distributed over the HLT farm

- full events locally assigned for processing



LHCb Design Parameters

→ 313 data sources (TELL1)

- Average fragment size ~100 B
- Multi-fragment size chosen to fit jumbo frames
- up to 4 GbE

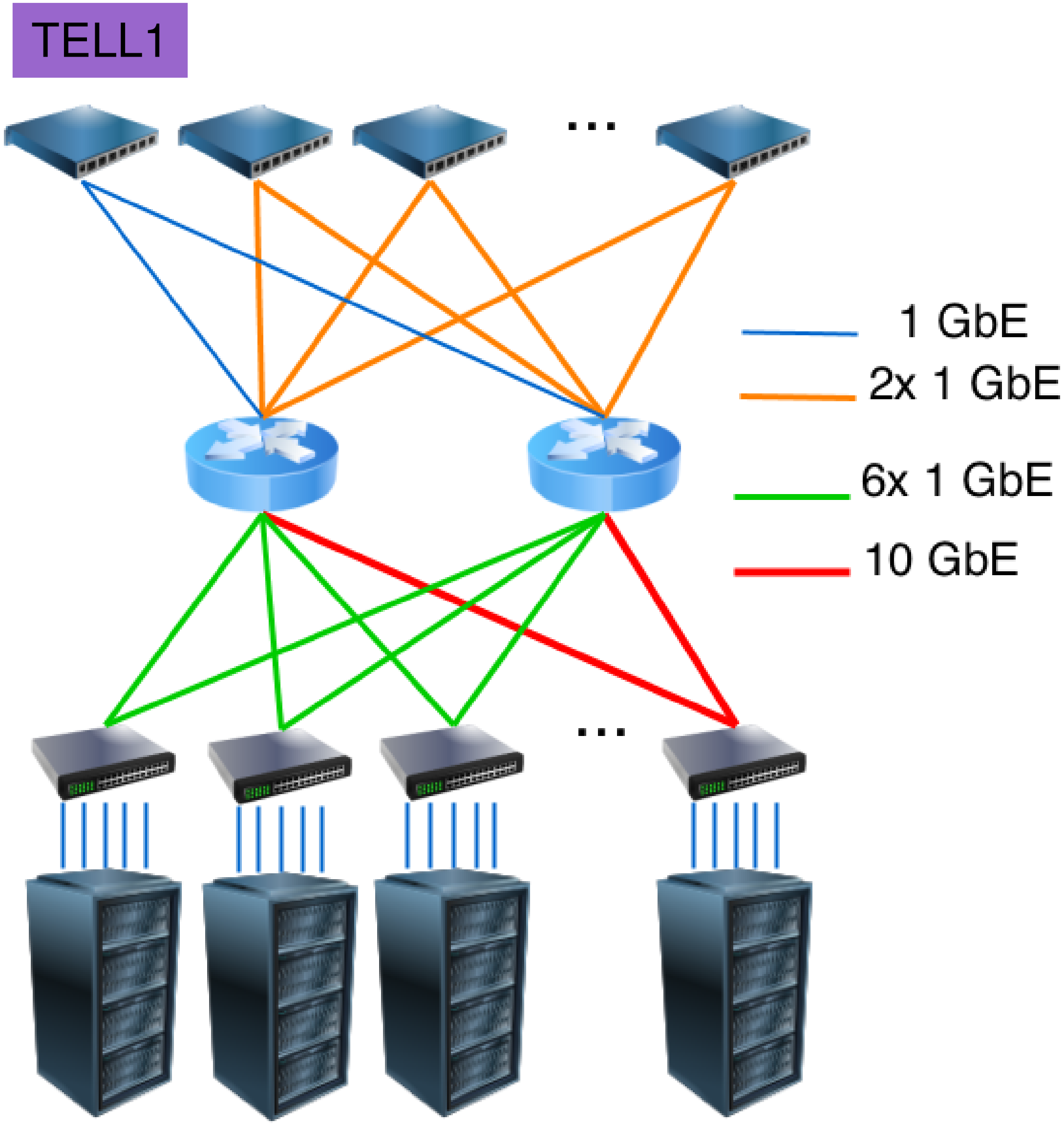
→ 1500 builder units (HLT)

- 1GbE

→ EB rate 1 MHz

→ EB bandwidth 40 GB/s

- ~35 kB/event



Architecture & Requirement Summary

	ALICE	ATLAS	CMS		LHCb
Protocol	Push TCP/IP	Pull TCP/IP (UDP)	Push Myrinet	Pull TCP/IP	Push UDP
Ev. assignment	Static	Dynamic	Static	Dynamic	Dynamic
Topology	Concentrated	Concentrated	Distributed		Distributed
Full Event Destination	Local (Storage)	Remote TCP/IP	Local (HLT)		Local (HLT)
Rate (kHz)	1	3.5	100		1000
BW (GB/s)	2	5.25	100		40
Data Sources	185	150	512	(8x) 64	313
Builder Units	85	100	8x64	HLT Farm	HLT Farm

Scaling and ultimate performance

➔ Scaling

- ALICE: variable operating conditions (data size, event rates, HLT operation) → built with large margins
  - EB scales up to 7 GB/s
- ATLAS: horizontal scaling possible, not implemented. HW & SW improvement wrt design margins
- CMS: scaling not needed. Free parameter event size → 50ns operation supported with design margins
- LHCb: scales with HLT farm size

	Design EB size		Final EB size		Design BW (GB/s)	Peak BW (GB/s)
ALICE	-		85		2	2 (2011)
ATLAS	100		(2x) 48		5.25	10 (2012)
CMS	8x 64	720	8x 64	1264	100	100
LHCb	1500		1500		40	60 (2012)

Fault tolerance

➔ ALICE

- builder unit crash stops data-flow
- resilient to missing fragment
- incomplete events recorded

➔ ATLAS

- resilient to both missing fragments and builder crashes
- incomplete events are preserved and processed

➔ CMS

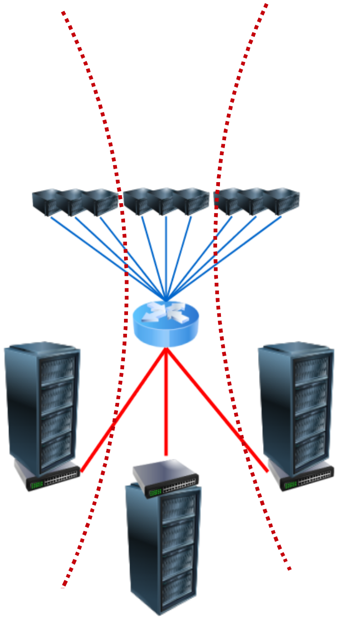
- resilient to full builder crashes, intermediated builder failure stops system
- missing fragment stops data-flow → recovery mechanism

➔ LHCb

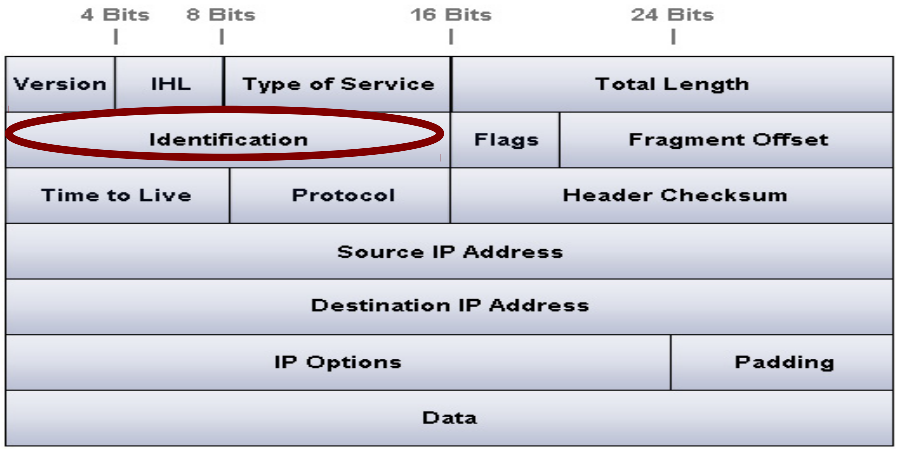
- resilient to both missing fragments and builder crashes
- incomplete events are dropped

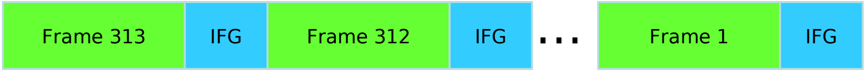
Handling heterogeneity

- HLT farms composed by heterogeneous hardware → Event builder serve events matching computing performance
- LHCb
  - one network, distributed event building → heterogeneity implicitly handled in event assignment
- CMS
  - sliced network → 8 parallel distributed event builders
  - heterogeneity within a slice is ok
  - slices have to be matched in term of computing power
- ATLAS
  - concentrated event builder distributing events over a flat network
  - logical slicing → less network connections, simpler recoveries
    - initially per rack → computing HW matching needed
    - later pseudo-random

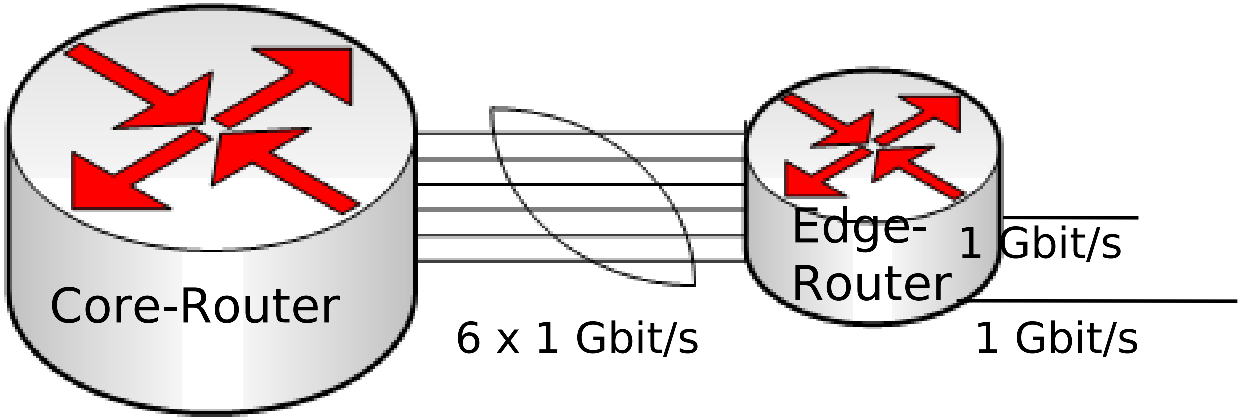


Squeezing the HW

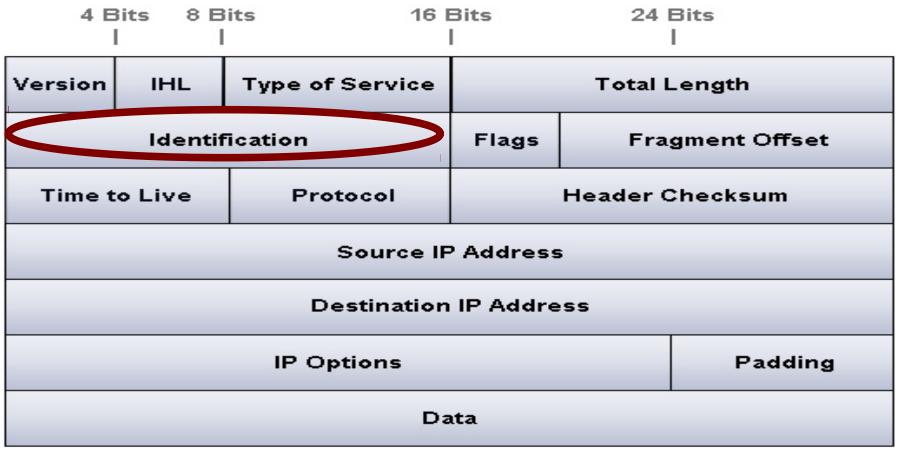




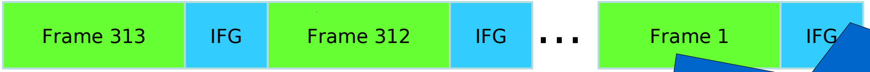
- ➔ Push protocol and UDP → **Minimize packet drops**
  - high performance edge routers
- ➔ Balancing policy for aggregate links using **event ID** as hash
  - frames from the same event are serialized → prevent over-commit
- ➔ Inter-frame gap size tuning
  - correct mismatches between transmitter and receiver frequencies (125 MHz±0.01%)



Squeezing the HW



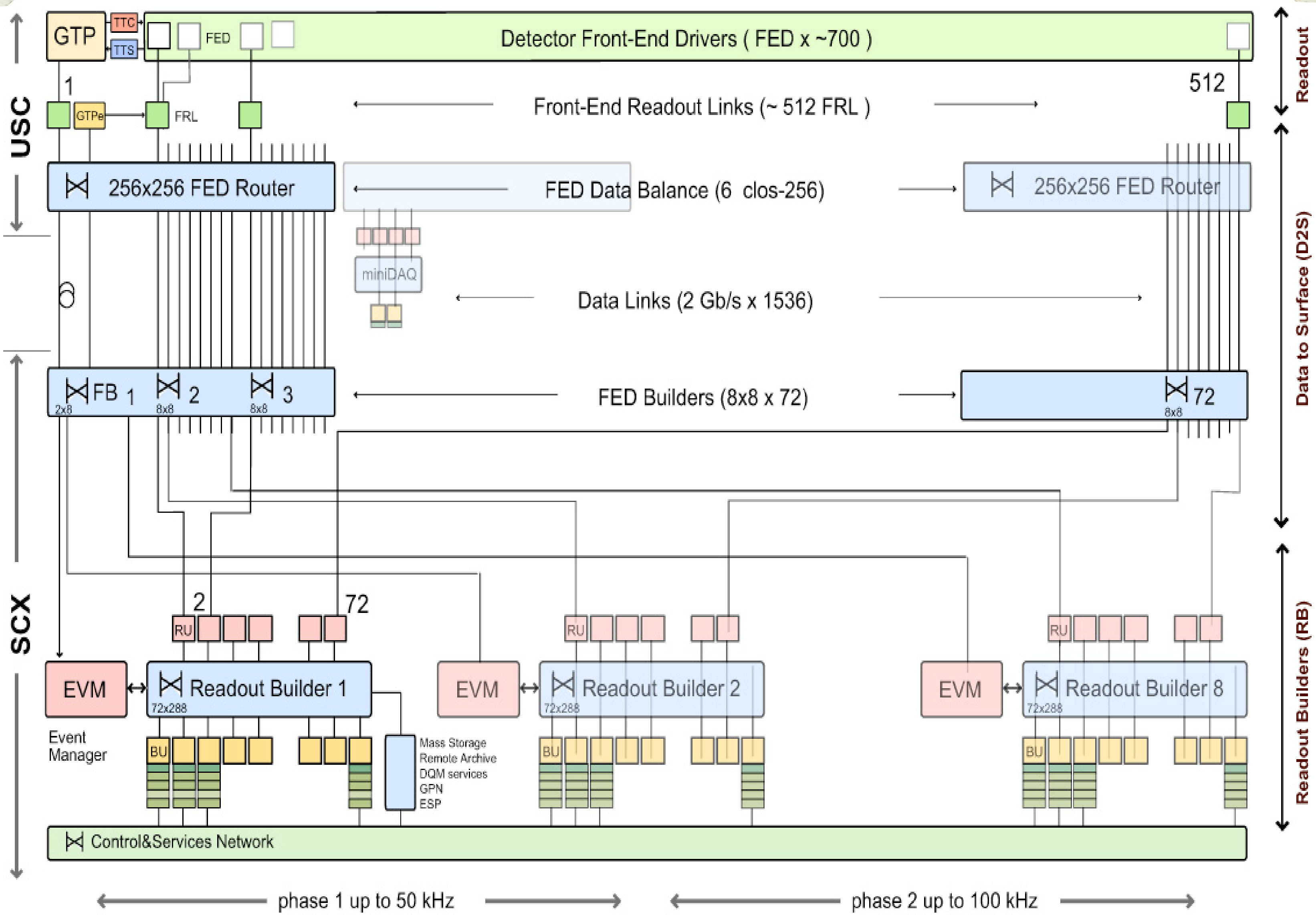
- ➔ Push protocol and UDP → **Minimize packet drops**
  - high performance edge routers
- ➔ Balancing policy for aggregate links using **event ID** as hash
  - frames from the same event are serialized → prevent over-commit
- ➔ Inter-frame gap size tuning
  - smatches between
  - receiver
  - z±0.01%)

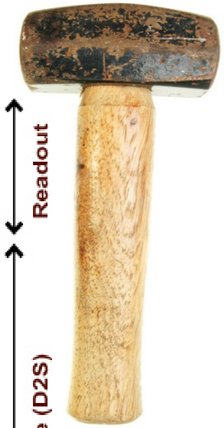


Core-Router
6 x 1 Gbit/s
1 G it/s

60 GB/s and ~0.1Hz incomplete events on a relatively small network
BUT
fine tuning, long-term stability concerns, support from HW manufacturers, non-standard options

Alternatively take the hammer





Conclusions

- ➔ Unsurprisingly, LHC EB systems based on large network infrastructures
  - network tuning and monitoring fundamental for system operation
- ➔ Combinations of few basic principles
  - distributed vs concentrated
  - push vs pull
  - static vs dynamic assignment
- ➔ Designs largely driven by
  - overall DAQ architecture
  - **resource availability**
- ➔ Different views (cultures) on hardware and data fault tolerance
  - quantitative comparison?
- ➔ In fact …
  - *"Upgrade: HLT frameworks & Event Building"*

Thanks to

- Roberto Divia'
- Niko Neufeld
- Andrea Petrucci