Storage in CMS DAQ

Openlab Technical Workshop, 1 Mar 2017, CERN

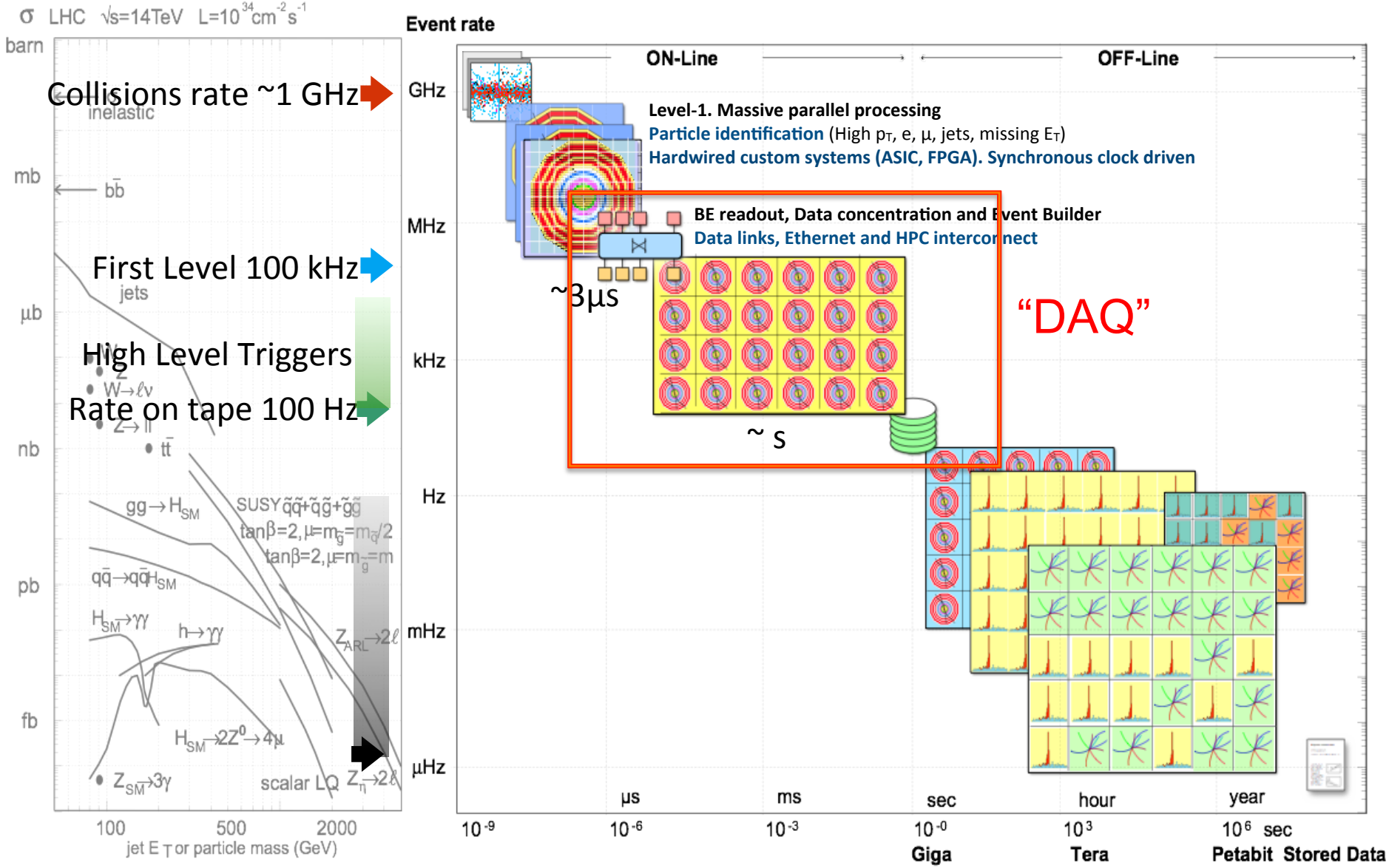
Prepared by Frans Meijers – CERN PH-CMD

CMS DAQ team

Outline:

- Introduction
- Storage in current DAQ and baseline for HL-LHC
- Alternative approaches for DAQ at HL-LHC

CMS Online/Offline computing model



Timeline and CMS DAQ parameters

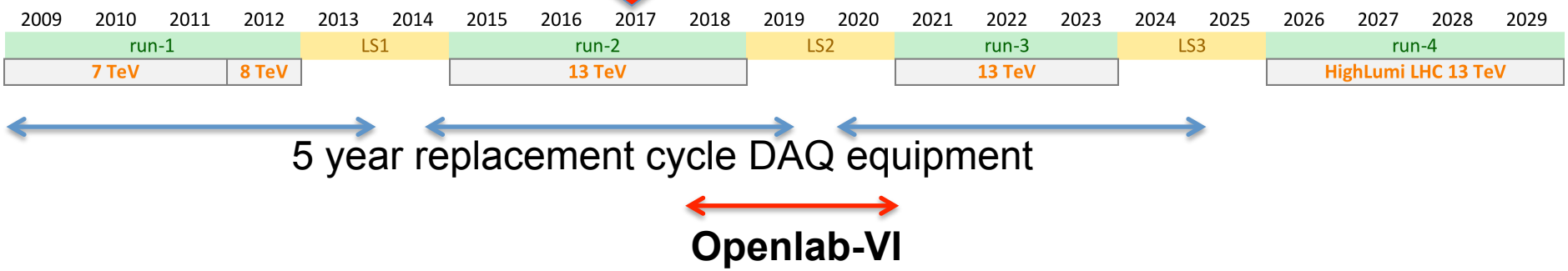
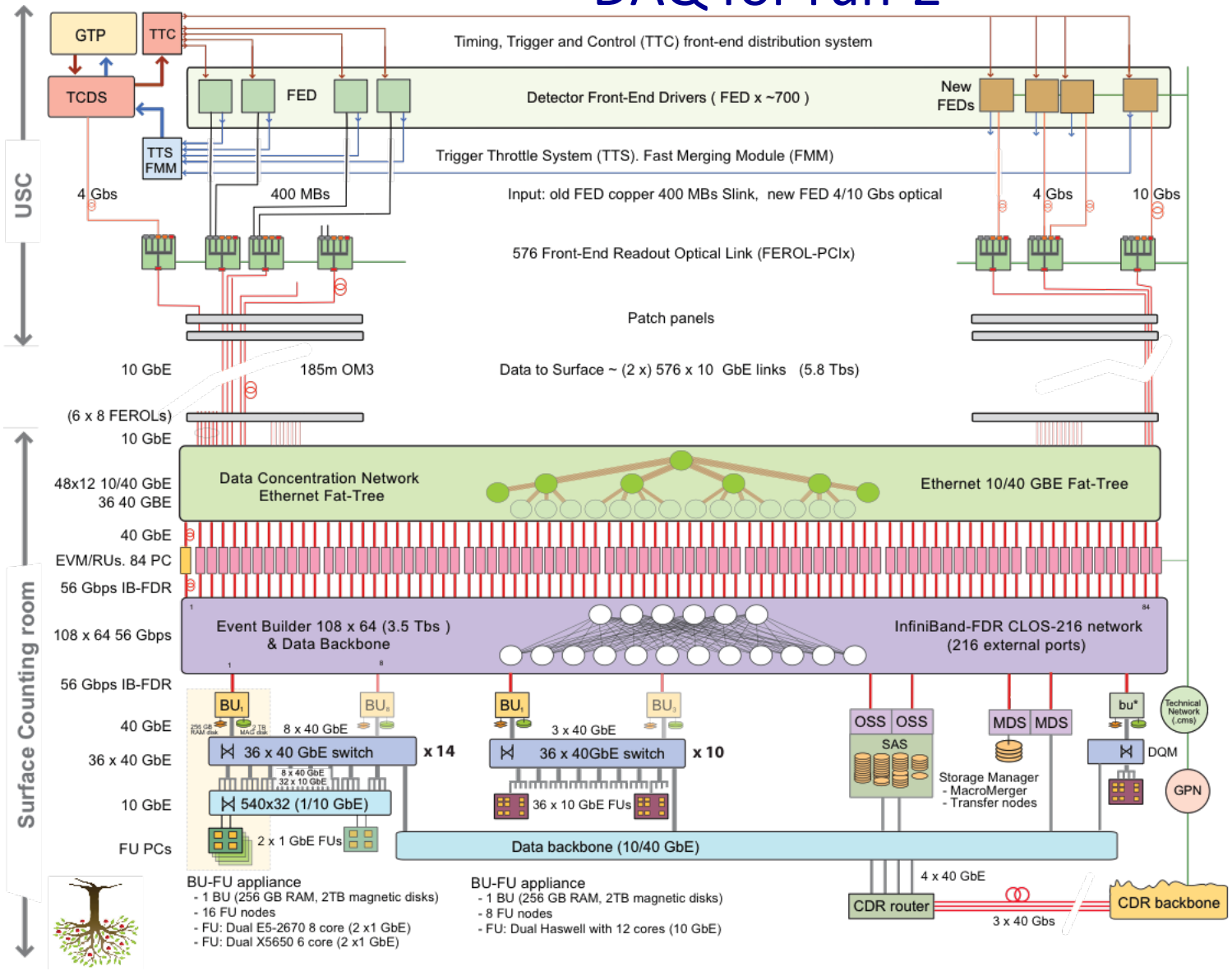
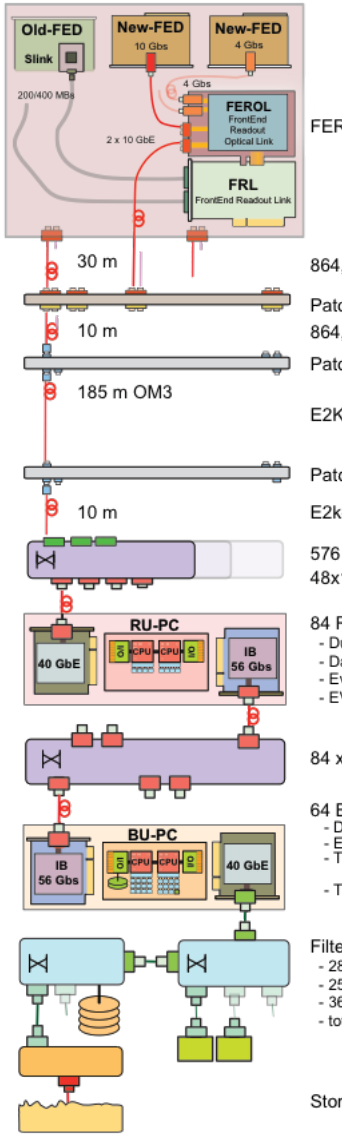


Table 7.1: DAQ/HLT system parameters.

	LHC Run-I	LHC Phase-I upgr.	HL-LHC Phase-II upgr.	
Energy	7-8 TeV	13 TeV	13 TeV	
Peak Pile Up (Av./crossing)	35	50	140	200
Level-1 accept rate (maximum)	100 kHz	100 kHz	500 kHz	750 kHz
Event size (design value)	1 MB	1.5 MB	4.5 MB	5.0 MB
HLT accept rate	1 kHz	1 kHz	5 kHz	7.5 kHz
HLT computing power	0.21 MHS06	0.42 MHS06	5.0 MHS06	11 MHS06
Storage throughput (design value)	2 GB/s	3 GB/s	27 GB/s	42 GB/s

DAQ for run-2


USC
Surface Counting room
GTP
TTC
TCDS
TTS FMM
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
New FEDs
Trigger Throttle System (TTS). Fast Merging Module (FMM)
4 Gbs
400 MBs
Input: old FED copper 400 MBs Slink, new FED 4/10 Gbs optical
4 Gbs
10 Gbs
576 Front-End Readout Optical Link (FEROL-PCIx)
Patch panels
10 GbE
185m OM3
Data to Surface ~ (2 x) 576 x 10 GbE links (5.8 Tbs)
(6 x 8 FEROLs)
10 GbE
48x12 10/40 GbE
36 40 GBE
Data Concentration Network
Ethernet Fat-Tree
Ethernet 10/40 GBE Fat-Tree
40 GbE
EVM/RUs. 84 PC
56 Gbps IB-FDR
108 x 64 56 Gbps
Event Builder 108 x 64 (3.5 Tbs)
& Data Backbone
InfiniBand-FDR CLOS-216 network
(216 external ports)
56 Gbps IB-FDR
40 GbE
36 x 40 GbE
10 GbE
FU PCs
BU
8 x 40 GbE
36 x 40 GbE switch
x 14
540x32 (1/10 GbE)
2 x 1 GbE FUs
3 x 40 GbE
36 x 40GbE switch
x 10
36 x 10 GbE FUs
OSS
SAS
Storage Manager
- MacroMerger
- Transfer nodes
MDS
bu*
DQM
Technical Network (.cms)
GPN
Data backbone (10/40 GbE)
4 x 40 GbE
CDR router
3 x 40 Gbs
CDR backbone
BU-FU appliance
- 1 BU (256 GB RAM, 2TB magnetic disks)
- 16 FU nodes
- FU: Dual E5-2670 8 core (2 x1 GbE)
- FU: Dual X5650 6 core (2 x1 GbE)
BU-FU appliance
- 1 BU (256 GB RAM, 2TB magnetic disks)
- 8 FU nodes
- FU: Dual Haswell with 12 cores (10 GbE)

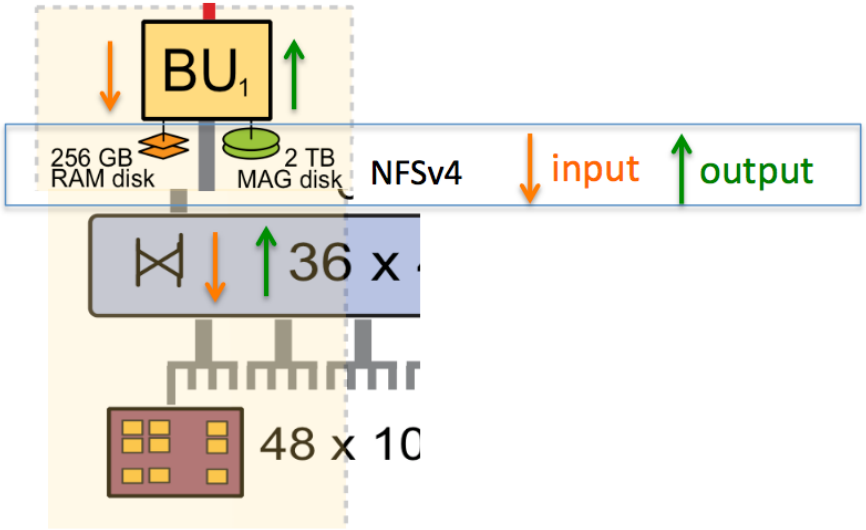

Old-FED
Slink
New-FED
10 Gbs
New-FED
4 Gbs
200/400 MBs
4 Gbs
FEROL
FrontEnd Readout Optical Link
2 x 10 GbE
FRL
FrontEnd Readout Link
30 m
10 m
185 m OM3
10 m
RU-PC
40 GbE
IB 56 Gbs
BU-PC
IB 56 Gbs
40 GbE

Storage for IT infrastructure

- Online cluster:
  - Usage: repositories, user home directories, VM images, ..
  - Current size: ~ 250 TB
  - Requirements:
    - HA
    - “turn key solution ”, minimal operation needs
  - Current
    - NetApp NAS filer
- Evolution:
  - Requirements met with standard commercial solution

Event Data Storage for DAQ in DAQ2 (step-1)

- The output of the EVB = input for High Level Trigger uses temporary storage on file system
  - allows the DAQ and HLT systems to be independent and to use the HLT software in the same way as for the offline processing.
  - 100 kHz, 1.5 MB evt size, yields **150 GB/s aggregate** with ~75 servers

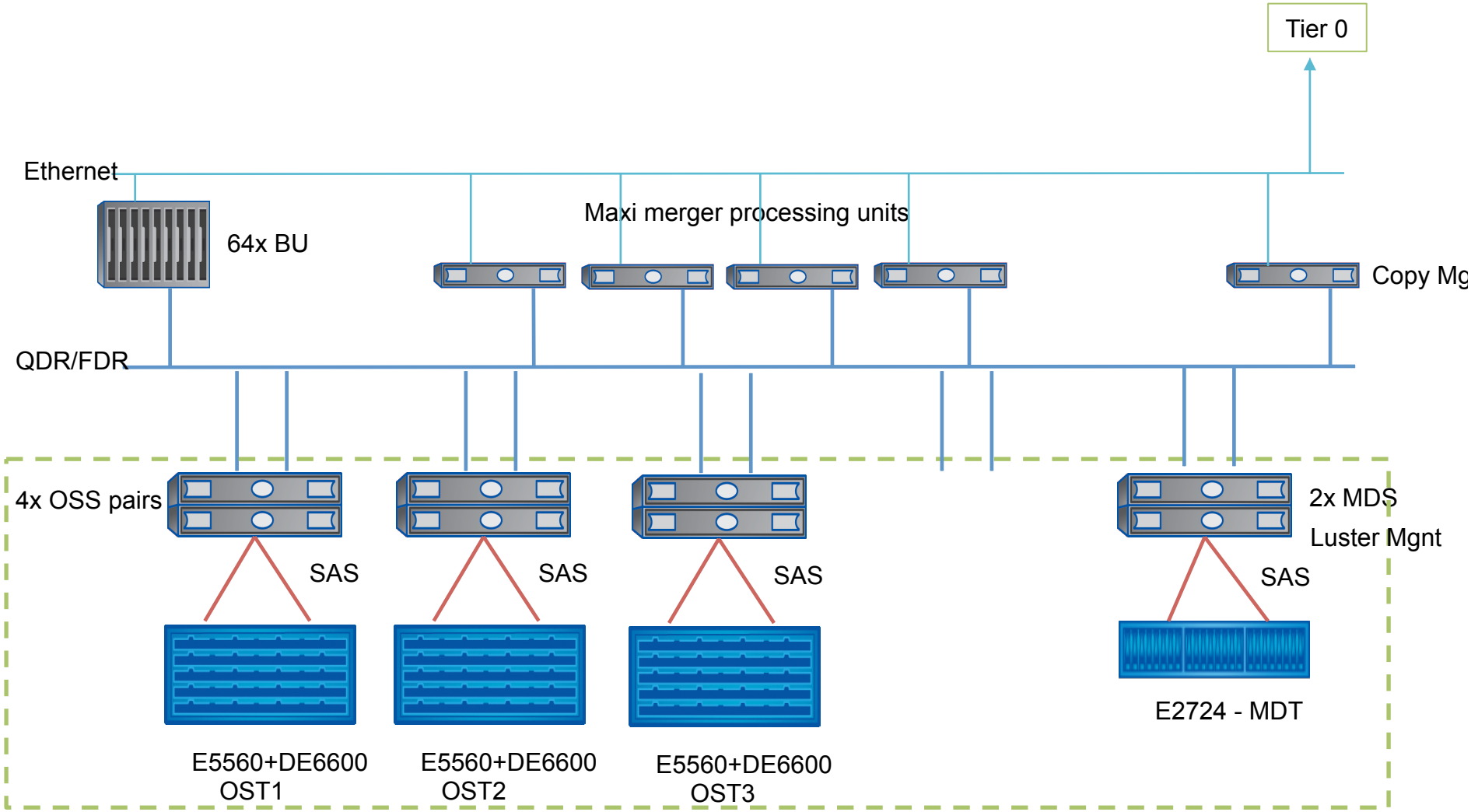


  - Per server
    - Requirements (2 GB/s W+R, few minutes buffer) met with RAM disk technology
    - SSD technology is improving both in terms of I/O throughput, as well as endurance.

Event Data Storage for DAQ in DAQ2 (step-2)

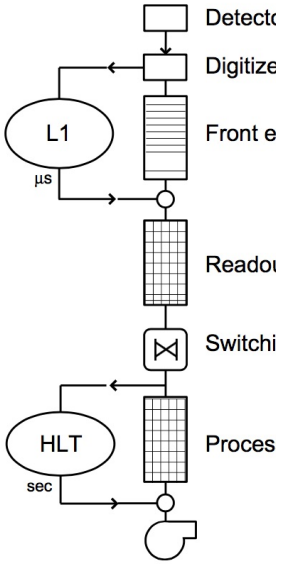
- The output of the EVB = input for High Level Trigger uses temporary storage on file system
- Output of HLT stored in (~20) “streams” per “Lumi-section”
  - Function: Merging, storage, monitoring and transfer to EOS
  - Storage system with Global File System
  - HA
  - Current implementation
    - Storage system and Lustre FS, interconnected via IB (and Ethernet)
      - ~100 clients, 360 disks over 3x2 MDS
    - Application level throughput ~6 GB/s, sequential write ~16 GB/s
  - Real GFS; Exploit multiple writers to same file
    - Reduces BW requirements by factor 2

E5560/DE6600 Lustre



Lustre File System

CMS L1 / DAQ / HLT

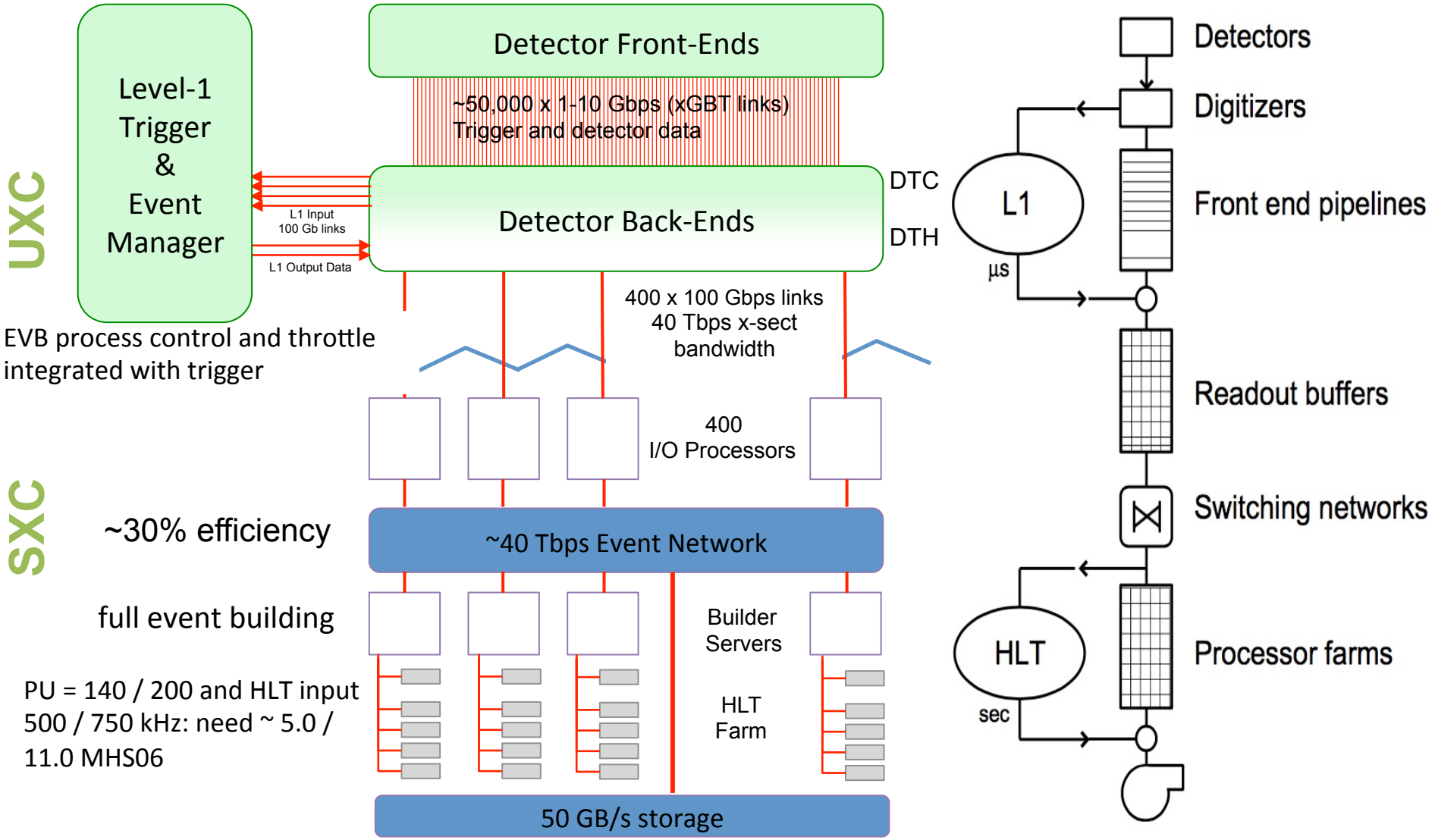


- Same two-level architecture as current system
  - L1 hardware trigger: 40 MHz clock driven, custom electronics
  - High Level Trigger (HLT): event driven, COTS computing nodes

Table 7.1: DAQ/HLT system parameters.

	LHC Run-I	LHC Phase-I upgr.	HL-LHC Phase-II upgr.	
Energy	7-8 TeV	13 TeV	13 TeV	
Peak Pile Up (Av./crossing)	35	50	140	200
Level-1 accept rate (maximum)	100 kHz	100 kHz	500 kHz	750 kHz
Event size (design value)	1 MB	1.5 MB	4.5 MB	5.0 MB
HLT accept rate	1 kHz	1 kHz	5 kHz	7.5 kHz
HLT computing power	0.21 MHS06	0.42 MHS06	5.0 MHS06	11 MHS06
Storage throughput (design value)	2 GB/s	3 GB/s	27 GB/s	42 GB/s

L1 / DAQ / HLT: Baseline



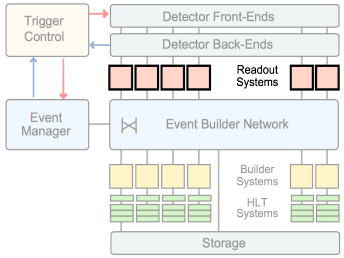
Baseline: Challenges

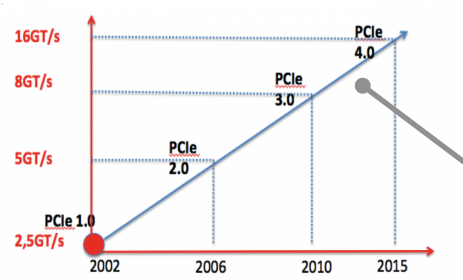
- Baseline CMS DAQ architecture for Run4 feasible with readily available technology
- Main Challenge: Limited Budget
- Directions:
  - Data to Surface: efficient concentration and transition to asynchronous/ reliable protocol
  - Event Builder: Reduce size and complexity
    - I/O processors
    - Choice of Network
  - Size/cost of HLT farm
    - Understand actual evolution of hardware
    - Use of heterogeneous architectures
    - Evolution of reconstruction software

Cursory conclusion

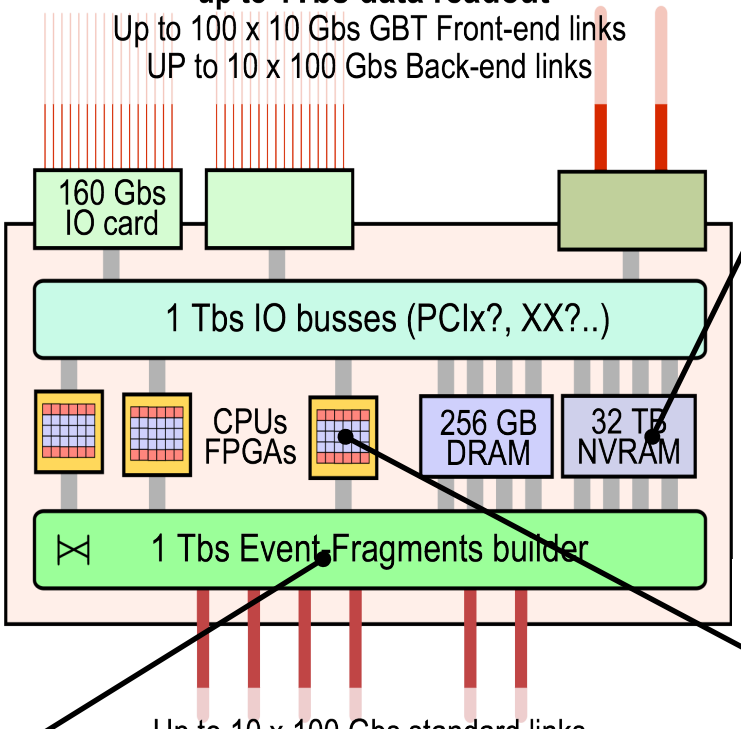
- Reduction of complexity and cost may result from tracking and late adoption of maturing technologies
  - Multi - 100 Gb links, on chip fabric
  - High bandwidth and large NV memory I/O servers
- Design options and exploratory work
  - Event Network: Non-building
  - HLT: programming styles more suitable for truly distributed systems with large NV memory
  - HLT: coprocessors and offload engines

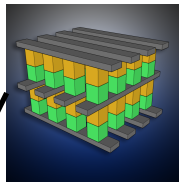
I/O, Buffer and Process (extrapolation)





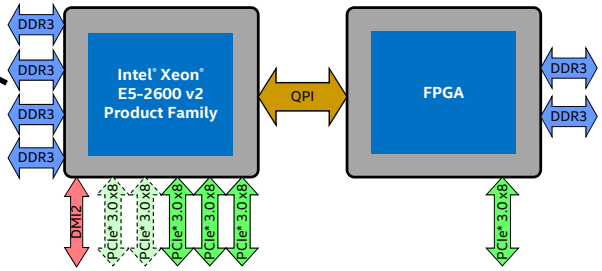
up to 1Tbs data readout
Up to 100 x 10 Gbs GBT Front-end links
UP to 10 x 100 Gbs Back-end links



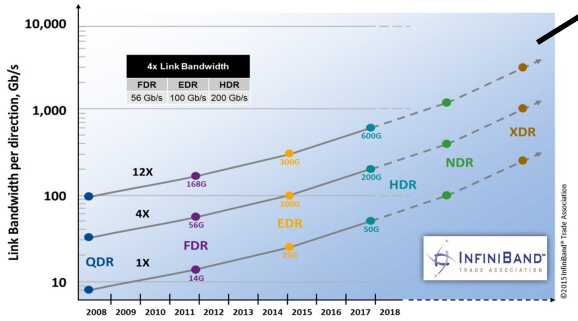


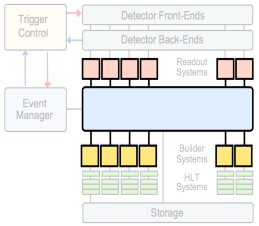
Readout Server
PC, ATCA crates etc..

CPU power for:
- Link protocols
- Detector hit builder
- Event fragment builder
- Large latency buffer
- Detetcor data monitor

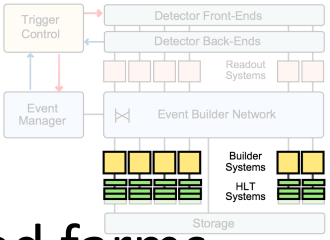


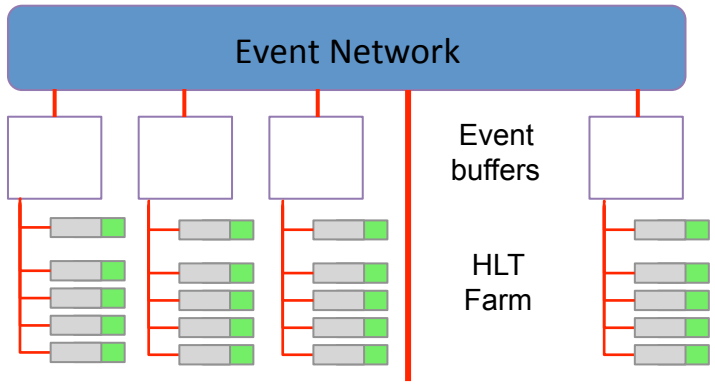
Up to 10 x 100 Gbs standard links
up to 1 Tbs Event-Fragments output

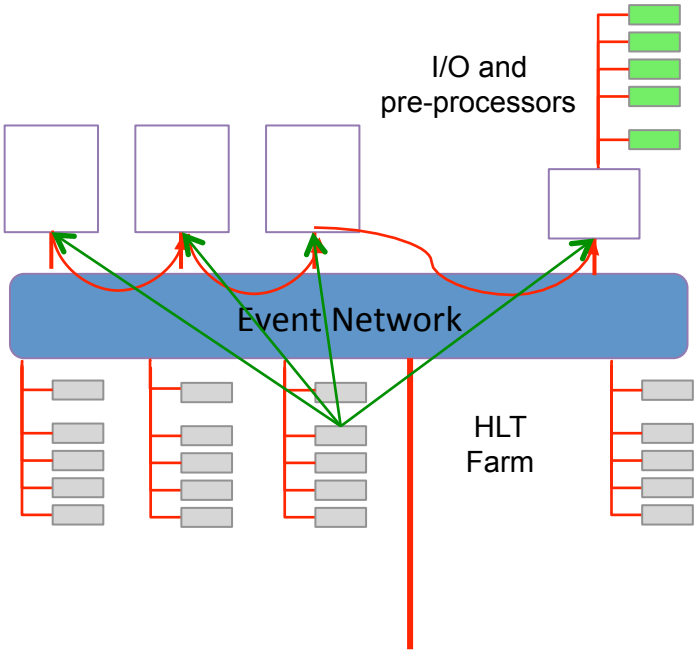




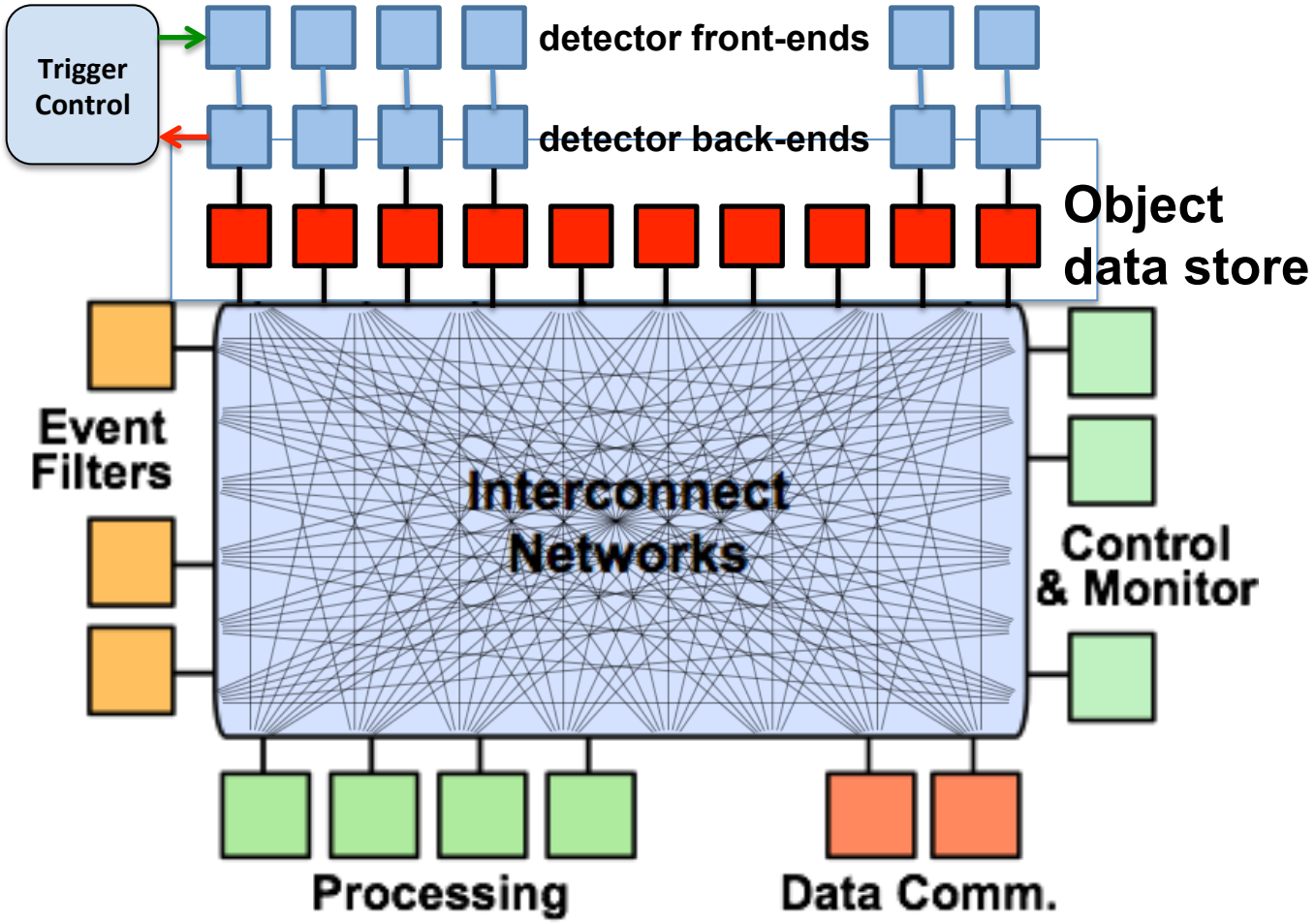
HLT: harness efficient CPU power







- Coprocessors and offload farms
  - CUDA, OpenCL…
  - Preemptive local reconstruction with ad-hoc software/hardware
- Truly distributed local processing (exploiting high-performance fabric)
- FPGA-assisted regional reconstruction
- Early classification – real-time indices
- Container / query programming style leveraging large NV memory


Trigger
Control
detector front-ends
detector back-ends
Object
data store
Event
Filters
Interconnect
Networks
Control
& Monitor
Processing
Data Comm.

- 5000 GB/s

Hardware key-value (object) datastore

- Use of up-and-coming **object datastore** "standards" in DAQ
  - in a first phase, implementation of parts of the existing CMS DAQ architecture using the open storage API (but not necessarily the hardware).
    - For example, replace parts of the current EVB protocol, by implementing a "readout unit" using a logical open storage device and complete event building as a lookup/ read from a "cluster" of RU devices. This paves the way to selective build directly from the HLT application
  - if high-throughput devices become available, it is imaginable to implement the protocol at the level of the common detector interface ("FED"). This could allow building a "virtual" pipeline at the output of the L1 trigger, thus potentially enabling differed and selective processing of L1-accepted events.

EXTRA MATERIAL