Developments for CMS DAQ and possible collaboration with Openlab

Openlab Technical Workshop, 5-6 Nov 2015, CERN

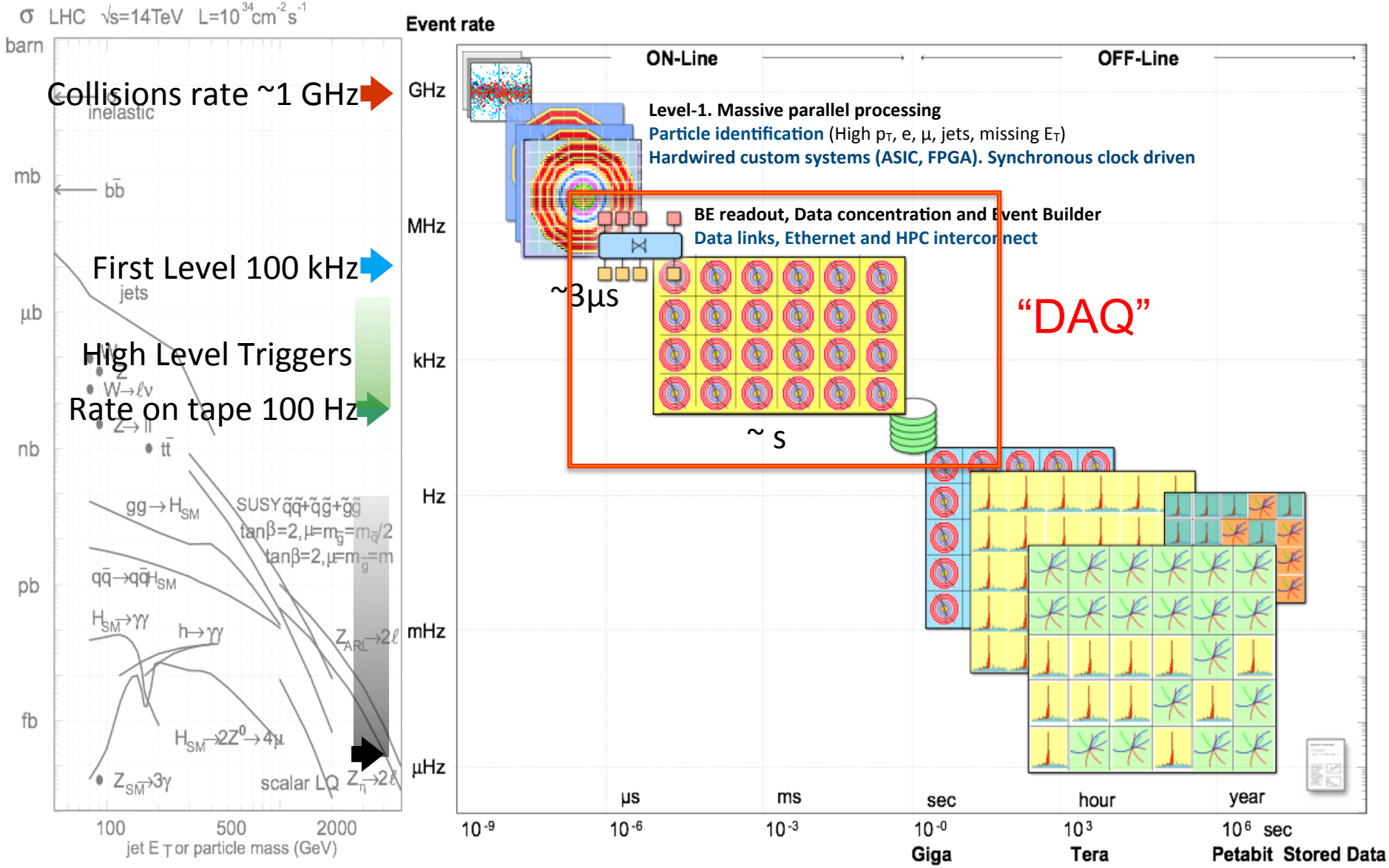
Prepared by Frans Meijers and Emilio Meschi – CERN PH-CMD

CMS DAQ team

Outline:

- Introduction
- R&D areas for CMS DAQ
- Possible projects with Openlab

CMS Online/Offline computing model



Timeline and CMS DAQ parameters

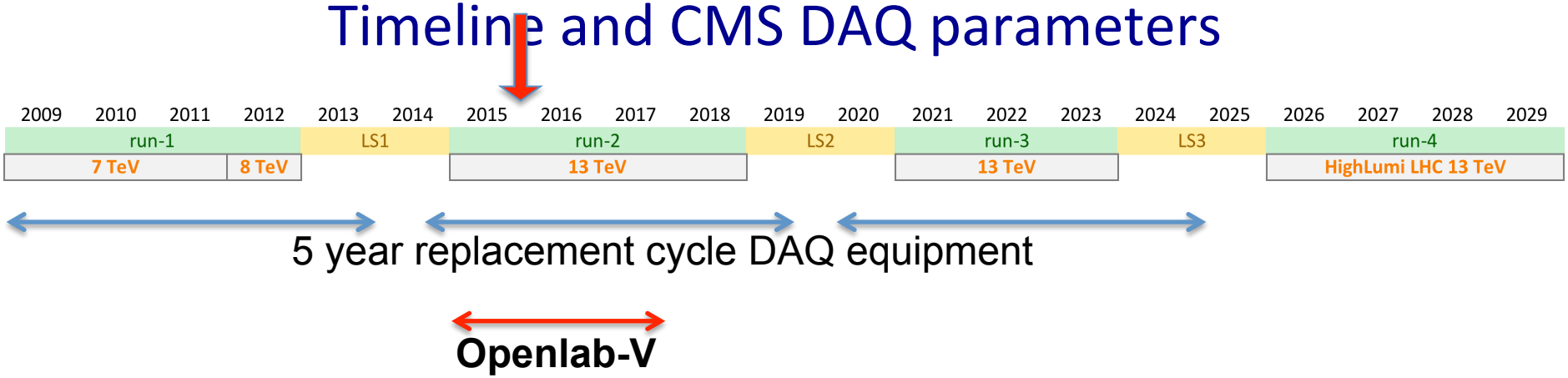
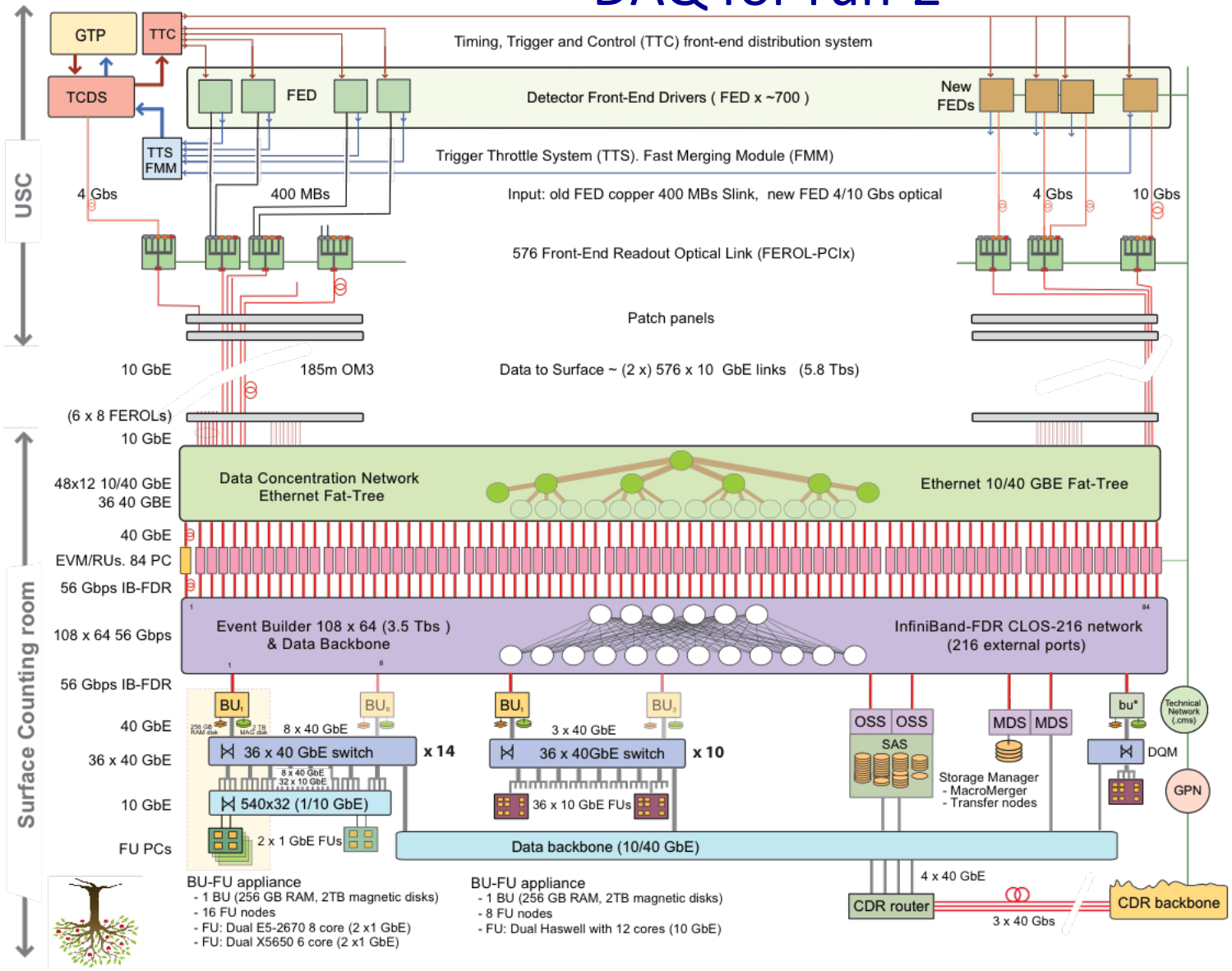
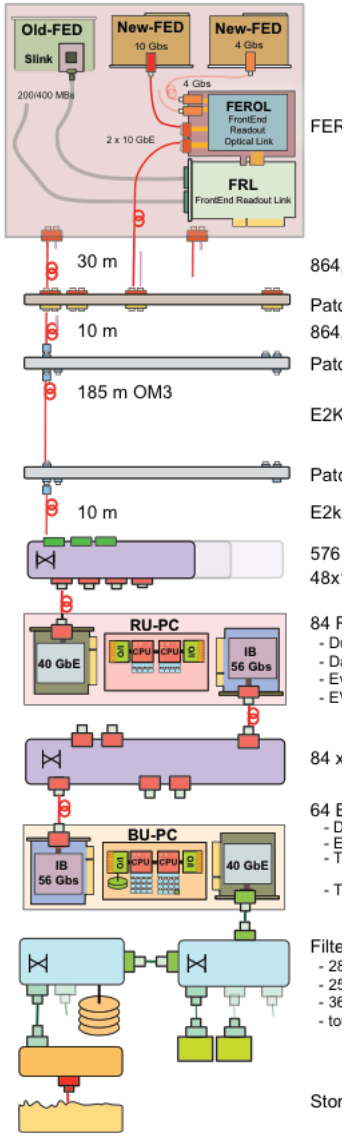


Table 7.1: DAQ/HLT system parameters.

	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

Areas of R&D for CMS online

- FPGA and Links and SerDes
- Networking for event building and event distribution
- Processor nodes for DAQ
- Processor nodes for HLT
- Storage for DAQ
- Software

FPGA and Links and SerDes

- R&D includes the
  - evaluation of new products with higher I/O bandwidth
  - study of custom protocols to transmit event data
  - feasibility of implementing standard protocols in an FPGA in order to interface directly to standard commercial switching networks and computing nodes (both Ethernet and HPC interconnects).
  - evaluation of high level programming languages (in order to simplify network protocol implementations and testing)

Networking for event building and event distribution

- The two main technologies of interest are Ethernet and HPC fabric interconnects.
- The CMS DAQ2 system:
  - For data concentration 10/40 Gbps Ethernet,
  - implementation of a reduced TCP/IP in FPGA for a reliable transport between custom electronics and commercial computing hardware.
  - The HPC Interconnect Infiniband (4xFDR 56 Gbps) is used for the event builder network.
- Evaluation of new products. One of the main objectives study
  - Throughput for the event building traffic. The event building traffic does not resemble the typical traffic in a data center and is also not typical for a HPC cluster which is mainly concerned with low latency.
  - effective use of the bi-directional links in network, considering that event building traffic is essentially uni-directional.

Processor nodes for DAQ

- EVB processor nodes connected to switching networks
  - input/output and buffering of event data and the control of this process
  - Now dual-socket x86 based servers, 40 PCIe-Gen3 lanes and up to 12 physical cores per socket and a NUMA (Non Uniform Memory Access)
  - trend towards integration of components in to the same Xeon die (CPU itself, memory controllers, PCI interface, GPU), which might lead to a further incorporation of the network interface in the future.
  - Experience has shown that extensive effort in software development and testing is required to fully exploit the full potential of the underlying hardware
- R&D includes study of I/O performance, NUMA, multi-core, use of co-processors for partial event building and effective utilization of very high bandwidth network interfaces.

Processor nodes for HLT

- The HLT code is common with the offline software.
- Platforms
  - General purpose servers based on dual-CPU (x86) configurations were the main platform for both HLT and offline processing in the past decade. This trend will continue.
  - However, there are also developments towards specialized co-processors and vectorization,
  - Specialized hardware and architectures are likely to be deployed first in controlled environments, such as the HLT farm, where the hardware can be controlled and specified.
- R&D includes the integration of HLT processor nodes based on alternative architectures.

Storage for DAQ

- 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 MB evt size, yields 100 GB/s aggregate with ~50 servers
  - 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.
- Output of HLT stored in (~20) “streams” per “Lumi-section”
  - Storage system with Global File System (capable of ~10 GB/s)
- R&D evaluation of cost-effective storage and access methods

Software

- R&D includes areas of
  - network technologies and protocol stacks,
  - web technologies,
  - databases,
  - cluster file systems,
  - expert systems and online failure prediction methods,
  - data analytics tools.

Examples of possible DAQ projects with Openlab

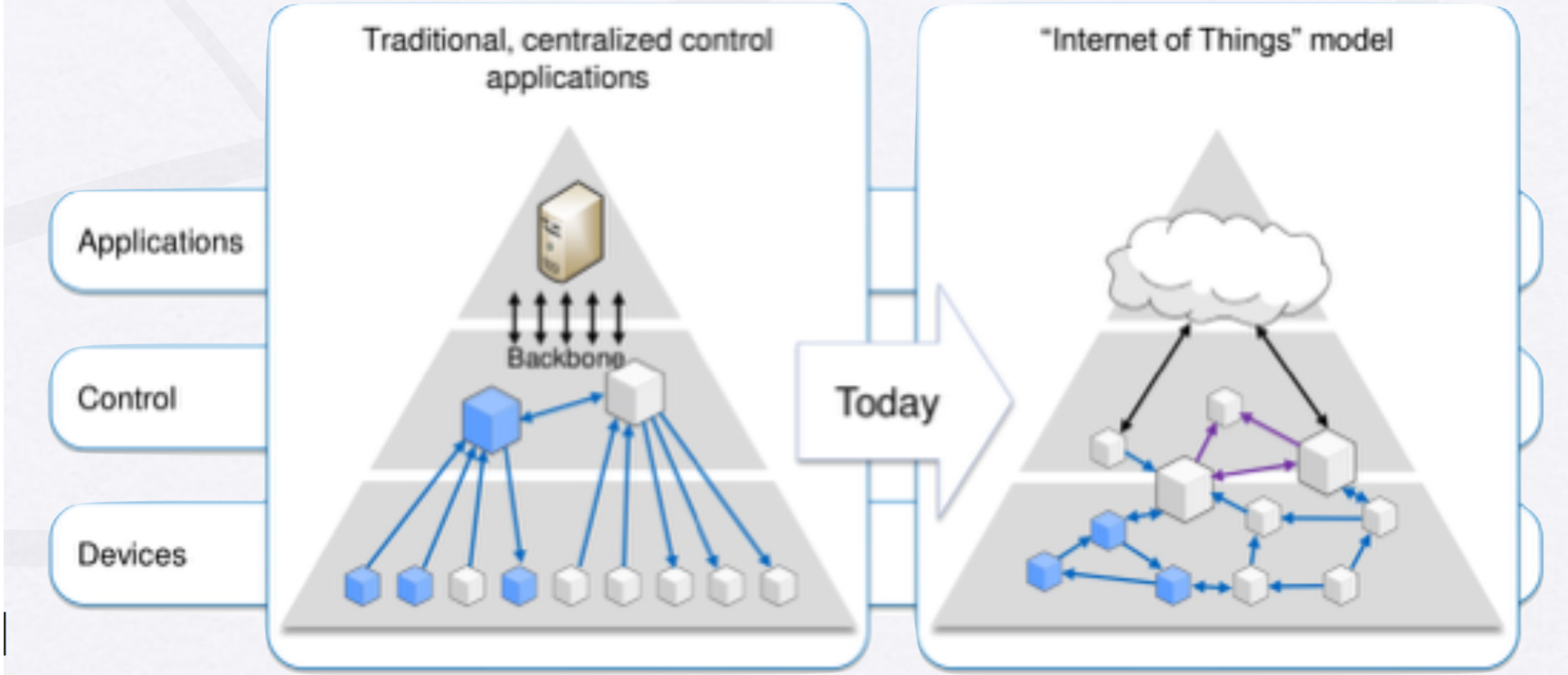
- Interconnect for Event Building
- NoSQL and search engines/analytics
- Hardware key-value (object) datastore

Interconnect for Event Building

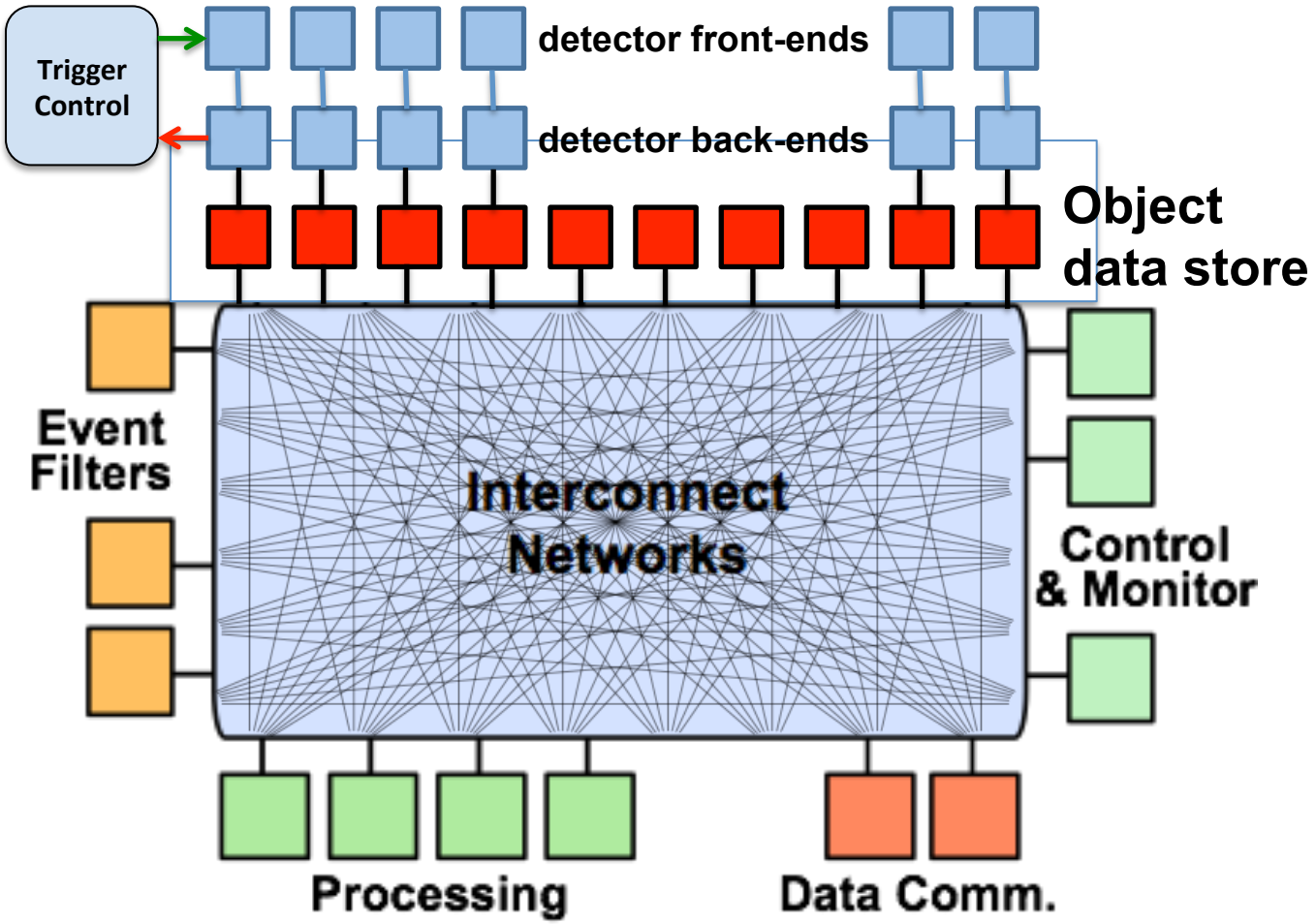
- Evaluate Intel Omni-path.
  - First generation product with 100 GBbps port bandwidth provides PCIe3.0 interface cards and 48 port switches.
- In particular
  - evaluate performance and functionality for event builder application on a single switch setup.
  - Using the VERBS API interface this could be immediately integrated in the existing CMS event builder application stack
  - Scaling of effective throughput with multi-stage switch network
- For openlab we would profit from:
  - test hardware (network switches and interface cards), either at CERN or at CMS site and / or remote access to large system
  - access to behaviour simulation model

NoSQL and search engines/analytics

- Investigate applications of real-time indexing and NoSQL datastore (elastic search, apache solr, paired with stream-processing services)
  - monitoring of all DAQ applications (EVB, etc)
  - monitoring the execution of the HLT algorithms:
    - large amounts of metadata are generated as by-product of the HLT execution
    - Online monitoring of system and info about the execution and the data flow
    - means to investigate and debug problems by analysing accumulated historical data.
    - correlating the HLT metrics with experiment and LHC in an integrated way
  - event categorisation for data scouting:
    - HLT algorithms feed summary information to such a service to enable fast identification of interesting feature, as well as providing event, file and dataset level cataloging. Advantages could be flexible feature identification (features do not have to be pre-defined) and immediate access to relevant data via a hierarchical catalog.
- For openlab we could profit from
  - involvement of developer to implement our specific features into the services



Source: http://db.in.tum.de/teaching/ws1314/industrialIoT


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

Hardware key-value (object) datastore

- Use of up-and-coming **object datastore** "standards" (such as seagate openstorage) 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.
- For openlab we could profit from
  - access to developer tools / expertise to kickstart development (e.g. implementing the "standard" on live RAM), as well as test hardware at a later stage (e.g. hi-performance SSD-based openstore devices/systems).