CMS DAQ
Today and the Future

Remigius K Mommsen
Fermilab

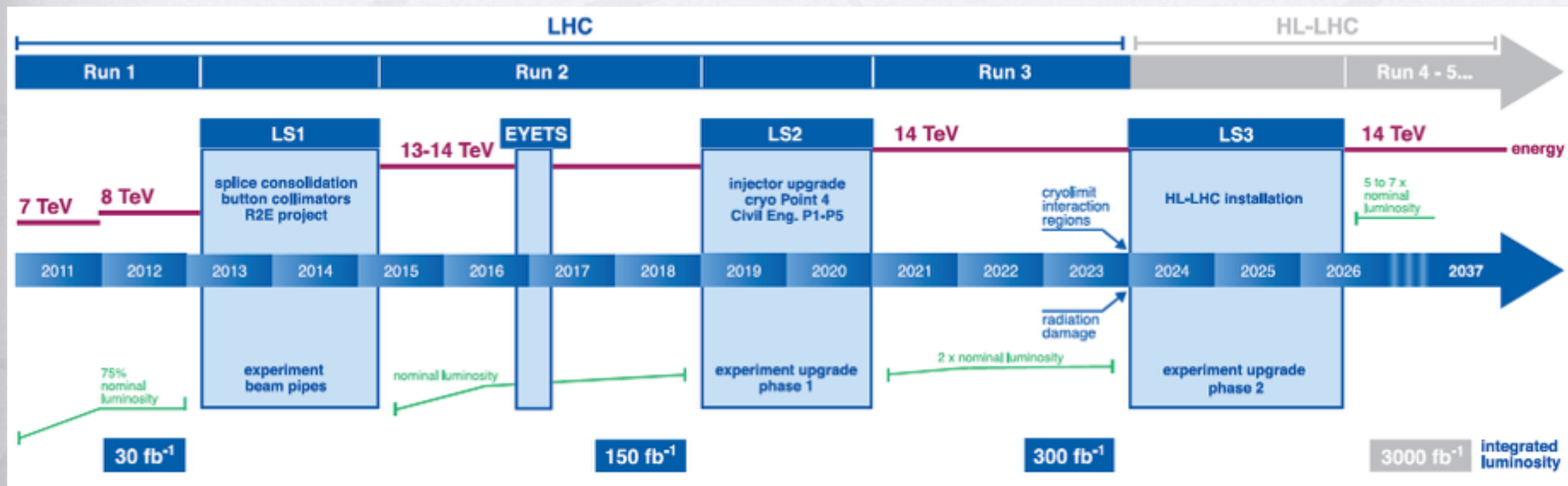
Overview

A new DAQ for run II

Plans for the near future

Ideas about DAQ3 (2019)

Data-acquisition on the time-scale of CMS phase 2 (2025)



CMS DAQ for LHC Run II

Requirements

- 100 kHz level 1 trigger rate (unchanged)
- Event size might double to 2 MB
  - Increase in pileup
  - New detectors
- Accommodate legacy and new µTCA-based detector readouts

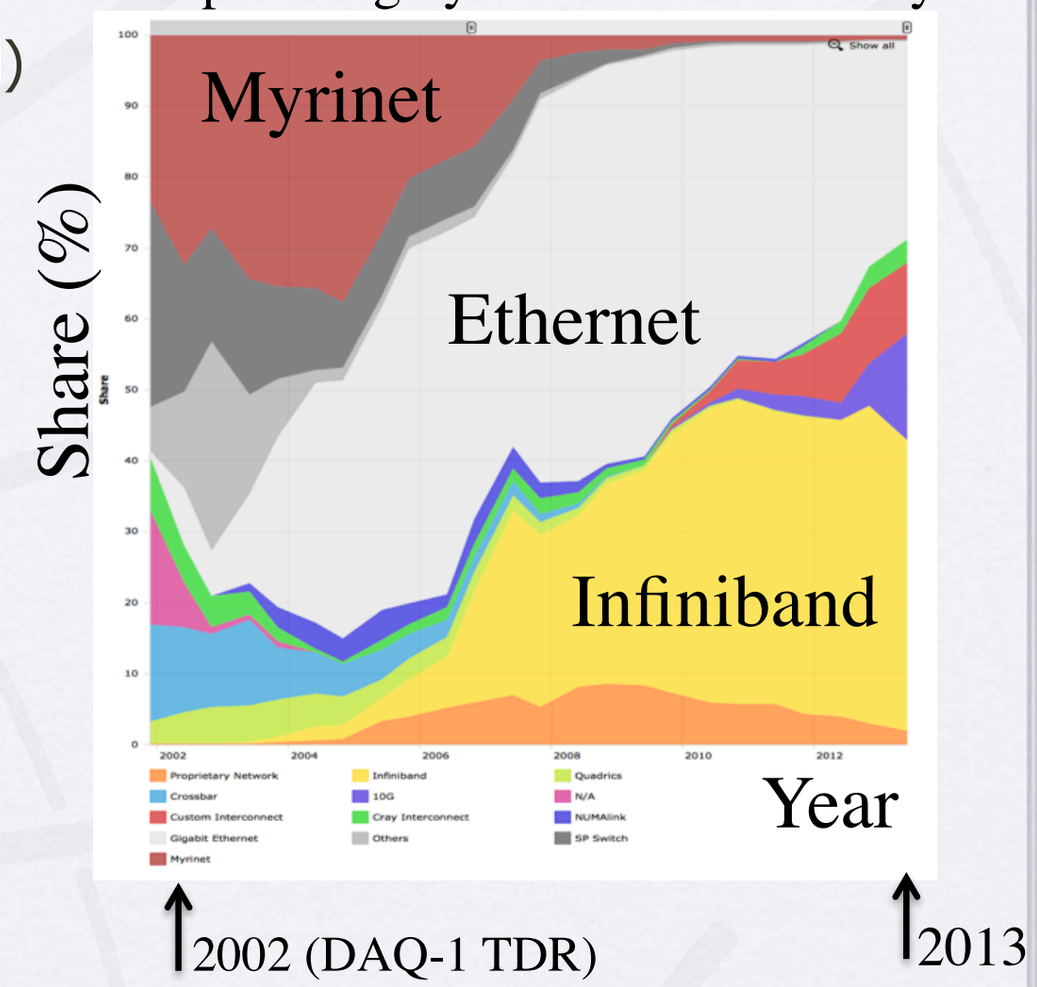
Aging hardware

- Most components reached end-of-life cycle

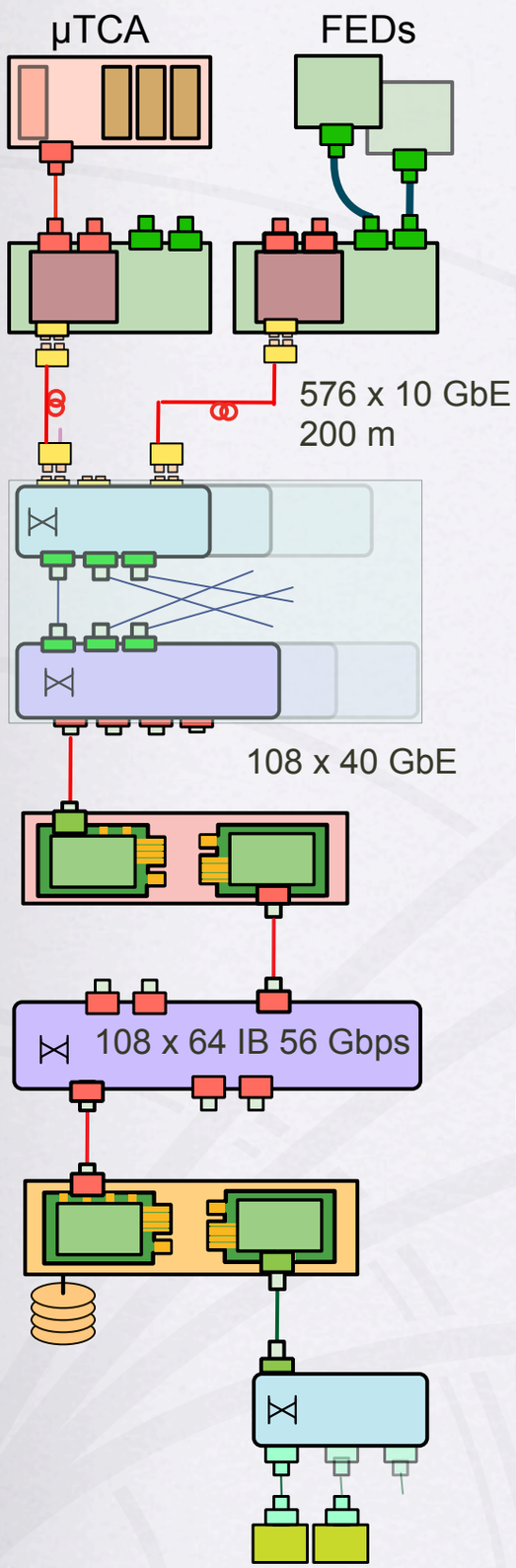
New technologies

- Myrinet widely used when DAQ-1 was designed
- Ethernet and Infiniband dominate the top-500 supercomputers today

Top500.org by Interconnect Family



CMS Event Builder



Detector front-end (custom electronics)

- ~700 front-end drivers (FEDs) with ~2kB/fragment at 100 kHz

Front-End Readout Optical Link (FEROL)

- Optical 10 GbE TCP/IP

Data Concentrator switches

- Data to Surface
- Aggregate into 40 GbE links

Up to 108 Readout Units (RUs)

- Combine FEROL fragments into super-fragment

Event Builder switch

- Infiniband FDR 56 Gbps CLOS network

Up to 64 Builder Units (BUs)

- Event building
- Temporary recording to RAM disk

Filter Units (FUs) (~16k cores in ~900 boxes)

- Run HLT selection using files from RAM disk
- Select O(1%) of the events for permanent storage

10 GbE Replacing Myrinet

Front-End Readout Optical Link (FEROL)

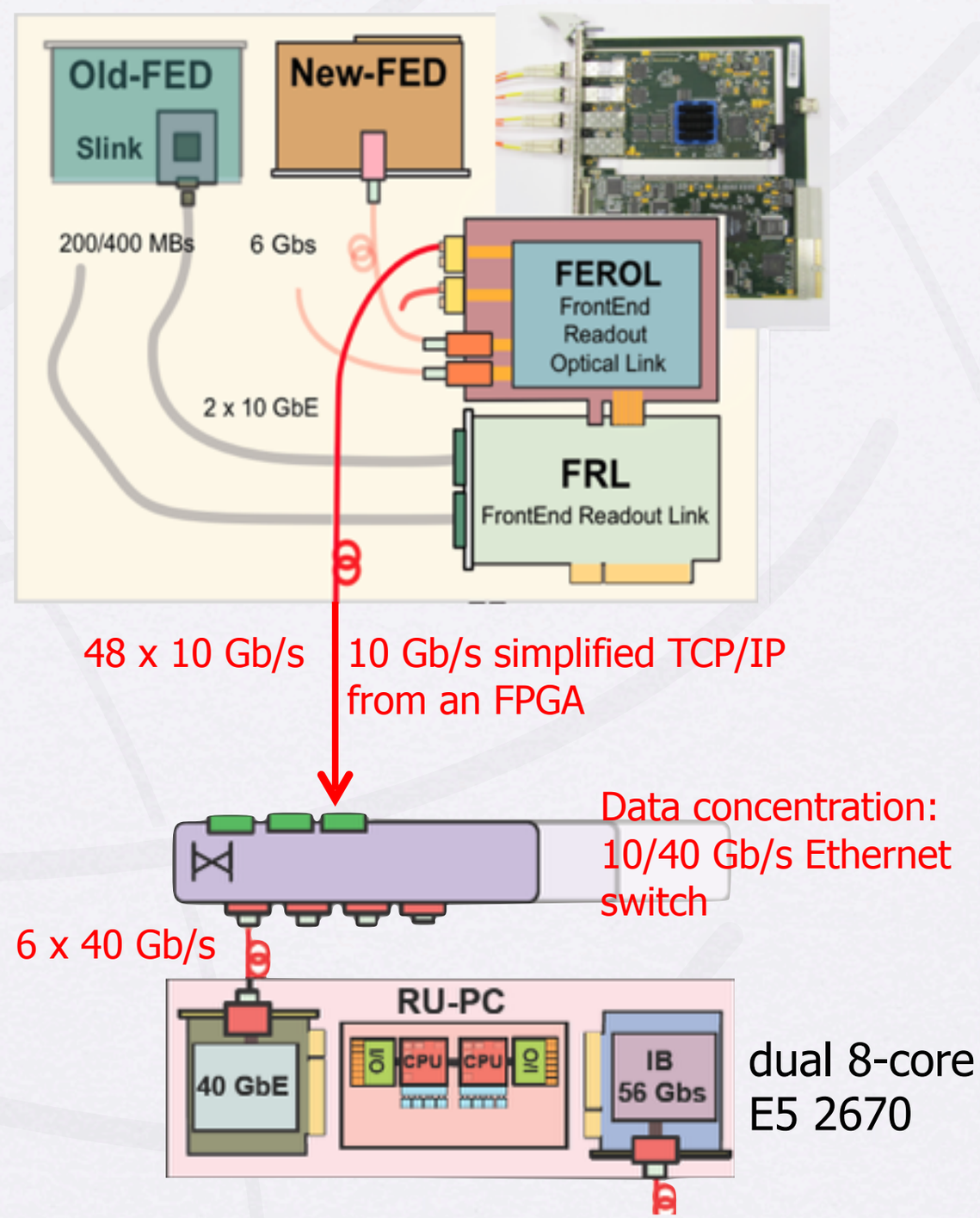
- Legacy input via Slink / FRL
- Optical up to 10 Gb/s from new µTCA crate via AMC13

Data to surface

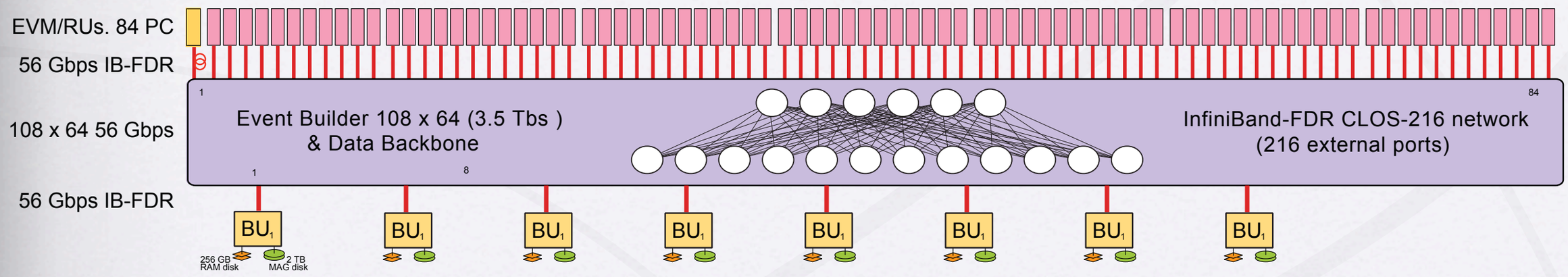
- Simplified TCP protocol over 10 GbE
- New fibers from USC to SCX
- 4-16 FEROL merged into 40 Gbit Ethernet at switch level

Each FED has one TCP stream

- Readout Unit (RU) builds super-fragments



Event Builder



Single-sliced event builder (DAQ1: 8 separate slices)

- 63 RUs and 62 BUs (DAQ1: 480 RUs and ~1000 BUs)
- Required re-engineered event-builder protocol and software
- Heavily multi-threaded & templated code to meet performance goal
- Optimized for NUMA (non-uniform memory architecture)

InfiniBand – most cost-effective solution

- Reliability in hardware at link level (no heavy software stack)
- Credit-based flow control (switches do not need to buffer)
- Can construct a large network from smaller switches

HLT farm organized in sub-farms connected to one BU

- DAQ1: each HLT node is also doing event building

How to Achieve Performance

Avoid high rate of small messages

- Request multiple events at the same time
- Pack data of multiple events into one message

Avoid copying data

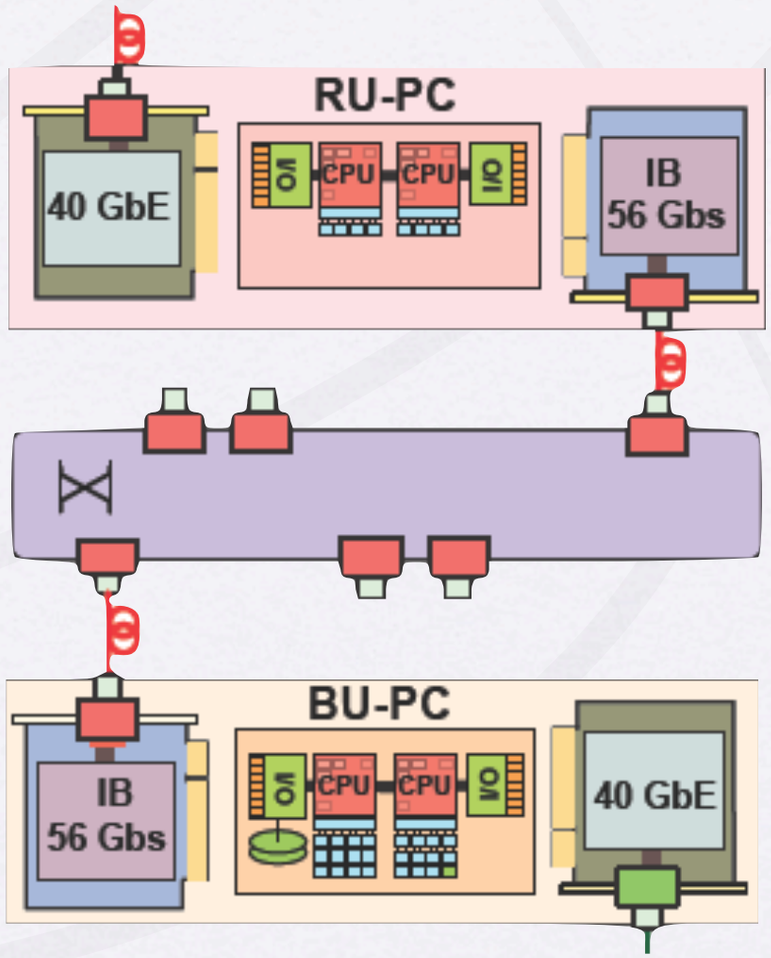
- Operate on pointers to data in receiving buffers
- Copy data directly into RDMA buffers of Infiniband NICs
- Stay in kernel space when writing data

Parallelize the work

- Use multiple threads for data transmission and event handling
- Write events concurrently into multiple files

Bind everything to CPU cores and memory (NUMA)

- Bind each thread to a specific core
- Allocate memory structures on pre-defined CPU
- Restrict interrupts from NICs to certain cores
- Tune Linux TCP stack for maximum performance



File-based Filter Farm

Decouple CMSSW from XDAQ world

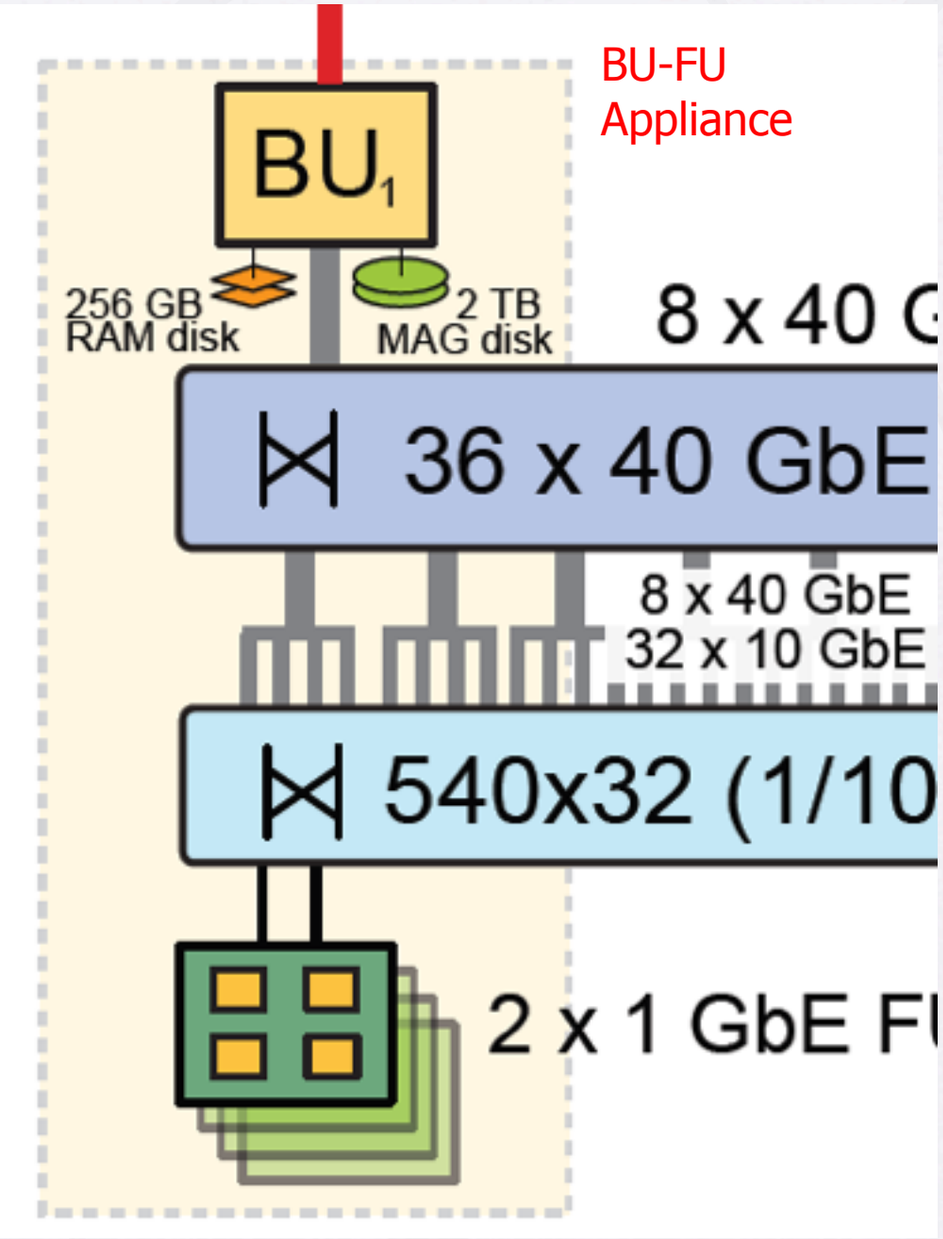
- Different frameworks with separate state machines, services, etc
- Different release cycles, compilers, etc
- Convoluted debugging

CMSSW input/output is file based

- Write the raw data to RAM disk on BU (256 GB/BU)
- Run standard CMSSW jobs (with cmsRun) with output to disk

8-16 FUs mount RAM disk via NFS4

- Each FU runs multiple CMSSW instances
- Discover new files on BU RAM disk
- Output merged into a single file per FU
- Copied it back to output disk on the BU



HLT farm, DAQ2

May 2011
72x

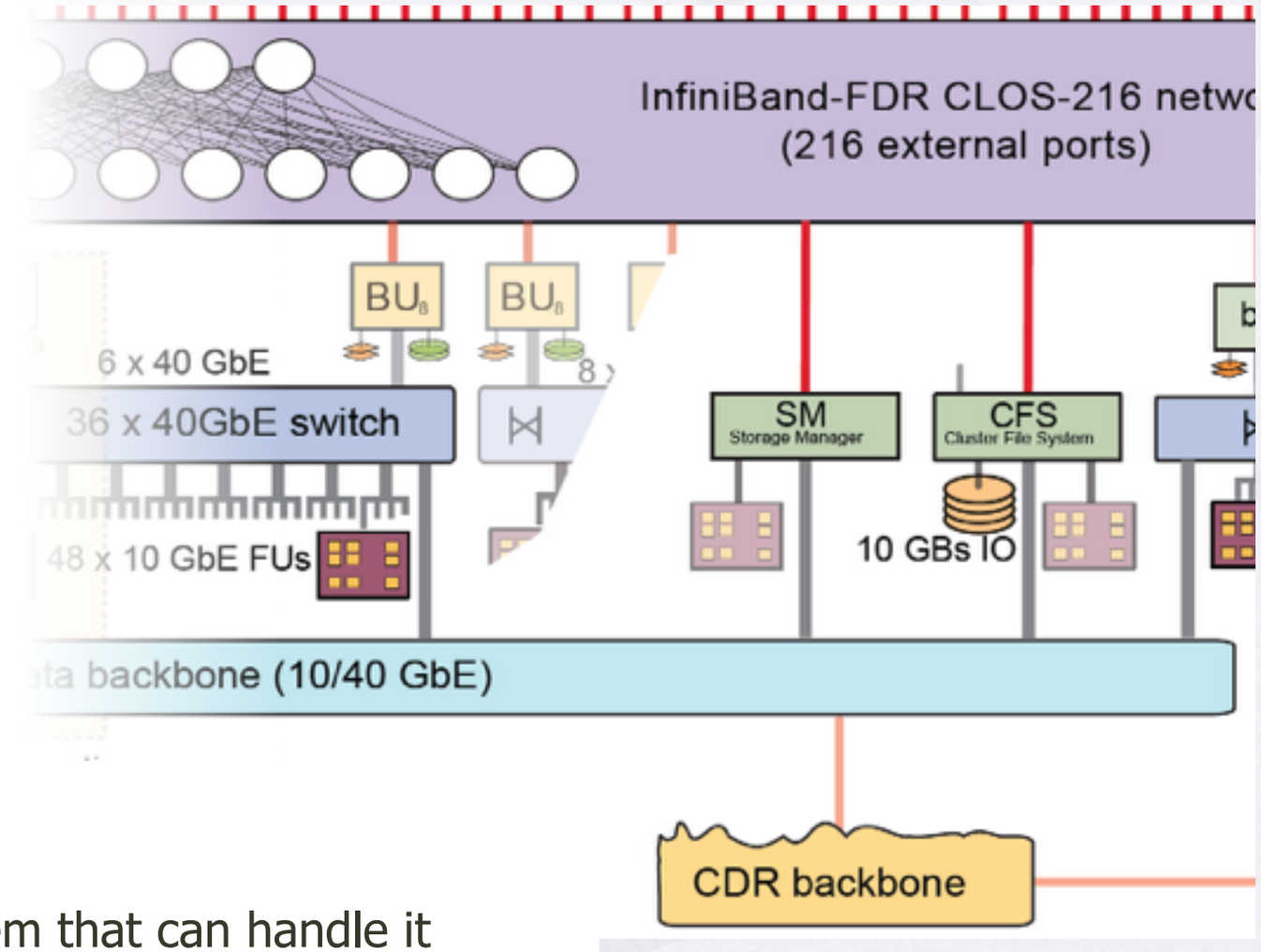
May 2012
64x

2015
90x

	2011 extension of DAQ-1 Dell Power Edge c6100	2012 extension of DAQ-1 Dell Power Edge c6220	HLT PC 2015 Megware S2600KP
Form factor	4 motherboards in 2U box	4 motherboards in 2U box	4 motherboards in 2U box
CPUs per mother-board	2x 6-core Intel Xeon 5650 **Westmere**, 2.66 GHz, hyper-threading, 24 GB RAM	2x 8-core Intel Xeon E5-2670 **Sandy Bridge**, 2.6 GHz, hyper threading, 32 GB RAM	2x 12-core Intel Xeon E5-2680v3 **Haswell**, 2.6 GHz, hyper threading, 64 GB RAM
#boxes	72 (=288 motherboards)	64 (=256 motherboards)	90 (=360 motherboards)
#cores	3456	4096	8640
Data link	2x 1Gb/s	2x 1Gb/s	1x 10 Gb/s, 1x 1Gb/s
Rel. perf. per node	0.6	1.0	1.66

Total: ~ 16 k cores on 900 motherboards

Merging and Storage



File-Based Filter Farm produces output files

- After merging on FU: 800 files x 10 streams scattered over 62 BUs every lumi section (23s)
- To be merged into 1 file per stream and lumi section in a central place

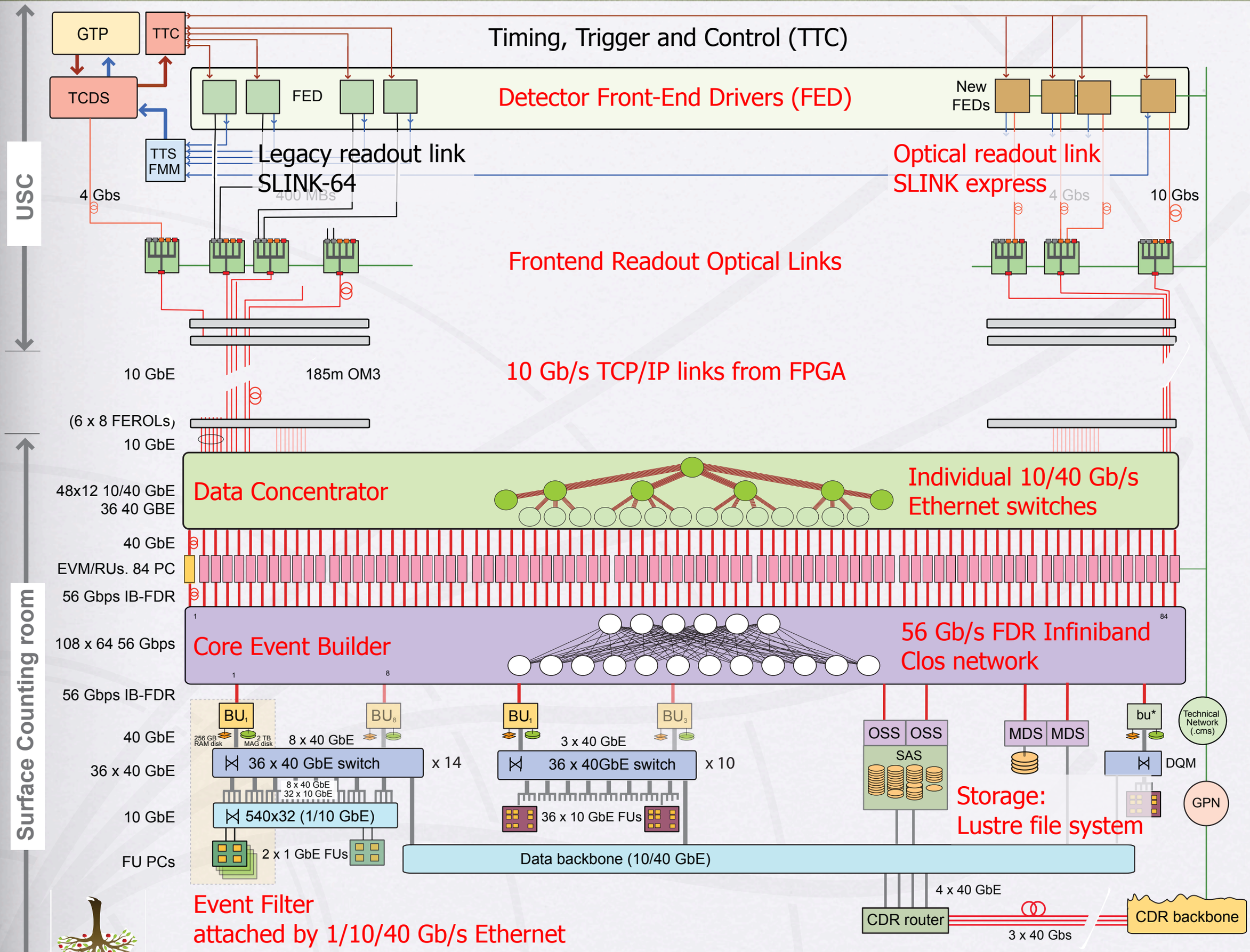
Merging can be done by a file system

- Just need to find a file system that can handle it

Solution: Global File System (Lustre) on a Storage System with 350 TB

- Merger process on BU reads data from all FUs in appliance
- Data are written directly from the BUs to a single output file in the global file system

Merged data is then transferred to tier 0 or to consumers at pt.5


GTP
TTC
Timing, Trigger and Control (TTC)
TCDS
FED
Detector Front-End Drivers (FED)
New FEDs
TTS FMM
Legacy readout link
SLINK-64
Optical readout link
SLINK express
USC
4 Gbs
400 MBs
4 Gbs
10 Gbs
Frontend Readout Optical Links
10 GbE
185m OM3
10 Gb/s TCP/IP links from FPGA
(6 x 8 FEROLs)
10 GbE
48x12 10/40 GbE
36 40 GBE
Data Concentrator
Individual 10/40 Gb/s Ethernet switches
40 GbE
EVM/RUs. 84 PC
56 Gbps IB-FDR
108 x 64 56 Gbps
Core Event Builder
56 Gb/s FDR Infiniband Clos network
Surface Counting room
56 Gbps IB-FDR
40 GbE
36 x 40 GbE
10 GbE
FU PCs
BU1
BU8
BU1
BU3
256 GB RAM disk
2 TB MAG disk
8 x 40 GbE
3 x 40 GbE
36 x 40 GbE switch
x 14
36 x 40GbE switch
x 10
8 x 40 GbE
32 x 10 GbE
540x32 (1/10 GbE)
36 x 10 GbE FUs
2 x 1 GbE FUs
Data backbone (10/40 GbE)
OSS
OSS
SAS
MDS
MDS
bu*
Technical Network (.cms)
DQM
GPN
Storage:
Lustre file system
4 x 40 GbE
CDR router
3 x 40 Gbs
CDR backbone
Event Filter
attached by 1/10/40 Gb/s Ethernet

Near Future

Integration of μTCA Readouts

Integrated into DAQ

- TCDS with 1 μTCA FED - in the readout since Sep 2014
- HCAL HF with 3 μTCA FEDs - working fine since May 2015

Infrastructure installed and commissioning on the way

- HCAL HB/HE with 9 μTCA FEDs
- L1 upgrade adds another 28 μTCA FEDs
- Pixel test blade with 1 μTCA

YETS 2016-17

- New pixel with 112 μTCA FEDs @ 10Gb/s
- Developing next generation FEROL based on μTCA
  - 4x 10Gb/s slink express in, 1x 40 Gb/s TCP/IP out
  - Not depending on FRL anymore

Work for the Next Years

Focus during LS 1 on commissioning the core DAQ system

- Many not-so-urgent issues pushed aside
- Will be worked on during run 2

Software improvements

- Improve fault tolerance and error reporting
- Performance improvements to reach ultimate DAQ 2 goal
- Develop automatic testing facilities in DAQ FM

Monitoring

- Basic tools adapted from run 1
- Plan to use elastic search for the whole DAQ (beyond FFF)
  - Some work already started
  - Build on expertise already gained to improve legacy monitoring
  - Allows much more post-mortem analysis
- New expert system for shifters & experts
- Revamp the error reporting for DAQ and CMS

Plans for CMS DAQ

No Radical Change for DAQ3

DAQ2 h/w will be at end-of-life in 2019

- Need to replace computers and network infrastructure

FEROLs will stay (unless there's a major disaster)

- More systems will switch to μTCA readouts
- Next generation FEROLs will still use TCP/IP
- Data-concentrator network will stay on Ethernet

Need to re-evaluate event-builder network

- Will Infiniband still be the most cost effective solution?
- Unlikely that there's a technology which allows to shrink the DAQ system substantially (as for DAQ2)

Take into account lessons to be learned during run 2

Networking for Event-Builder

Ethernet

- Not a reliable network in switched environment
- Speed
  - 40 GbE exists on switch and NIC since ~2012
  - 100 GbE exists but still very expensive
  - 400 Gbps defined

High-Performance Computing (HPC) Fabric interconnect

- Low-latency, reliable
- Infiniband 4xFDR 56 Gbps and 4xEDR 100 Gbps available
- New fabric interconnect forthcoming ..
  - 128 Gbps (2017-18), 200 Gbps (after 2020)
  - Integration of fabric port onto the CPU socket

Both technologies have switches with ~50 Tbps

Phase II (2025)

	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

Requirements on DAQ increase by factor ~25

- Feasible from technical point of view
- Likely obtainable within reasonable budget

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

Main cost driver will be the HLT farm

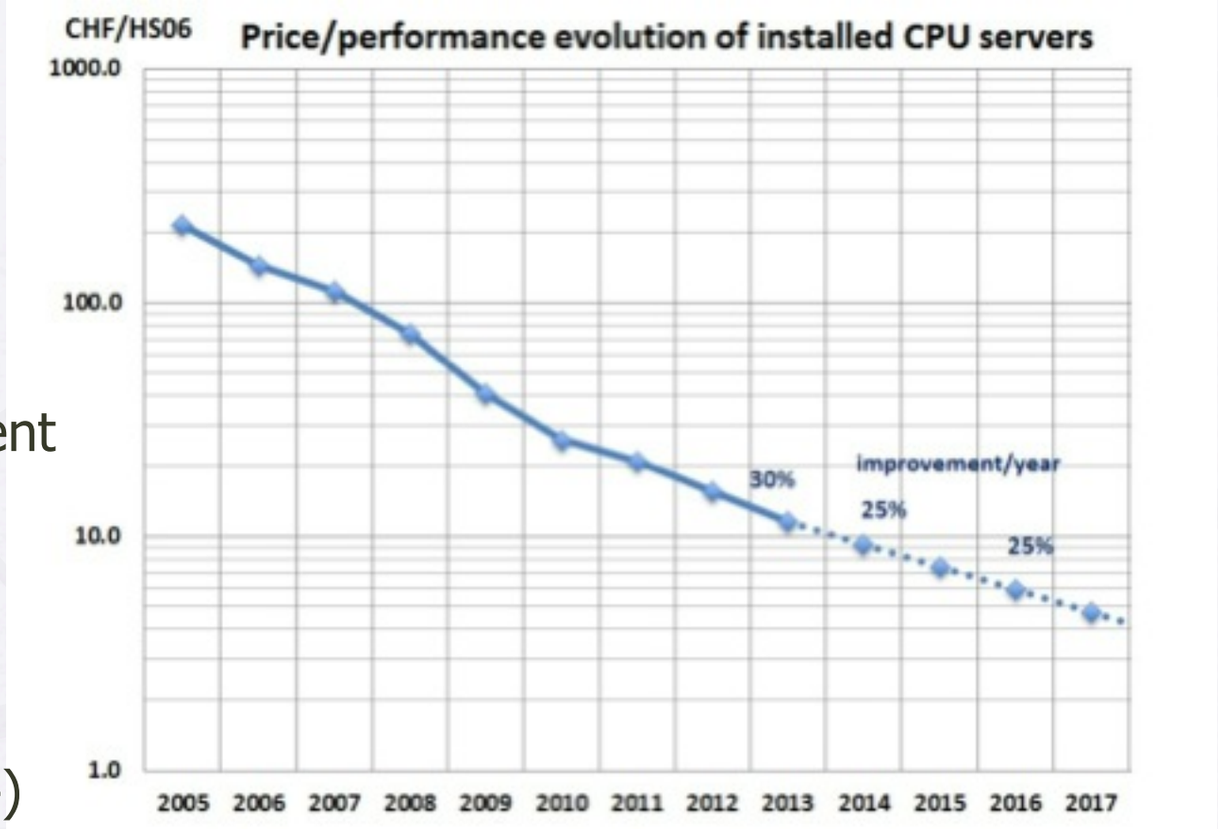
- Large uncertainty from technological progress and physics requirements

Costs Estimates

Extrapolation from existing system assuming technology evolution

DAQ readout, network, and storage

- Built with COTS computing, networking and storage equipment
- Assume factor of 10-20 performance improvement over 10 years at fixed costs
  - Observed in last decade from DAQ1 (~2007) to DAQ2 (2014)



High Level Trigger with similar reduction factor as present (1/100)

- CPU time per event largely unknown
  - Phase-II detector and L1/HLT menu
  - Influence of level-1 track trigger
    - Less time spend on track finding, but harder to reject events
  - Improvements due to code and selection optimization
- Possibly more cost effective CPU/GPU architectures in a decade

Staging Options

DAQ/HLT system can relatively easily be de-scoped / staged

- DAQ needs full connectivity from the start
  - Throughput can be adjusted
- HLT processing capacity can be adjusted according to needs
  - Easily expandable at a later stage
  - Delaying CPU purchased yields more CPU power at same cost
  - Consider trade-off between HLT / offline

Personal Ideas

Rethinking the DAQ?

Predicting affordable technologies over >10 years is hard

- There will be CPUs with many, many cores
- Different specialized cores/processors will be available
- Access to main memory will still be relatively slow
- Accessing memory on remote machines will become faster
  - Intel® Omni-Path Architecture will move interconnects into the CPU

A big challenge for high-throughput computing

- Minimize time needed to get data into and out of the boxes
- Optimize data structures for easy access from CPU/GPU

Distributed computing/storage/search will be the standard

- DAQ has different usage pattern: write once, read once, delete
- Industry focuses on latency which is less important for DAQ

A chance to rethink how a data-acquisition system works?

Getting Data of the Detector

Ideally, data transferred at bunch-crossing rate from the detector

- Otherwise large fraction of b/w is used for trigger data

Trigger-less DAQ challenging in LHC environment

- High radiation environment affects on-detector electronics
- Power dissipation limits complexity of readout chips
- Would need high-speed, low-power, very radiation-hard readout links
  - Optical or (highly directional) wireless links integrated in the readout-ASICs
  - Maybe some SiliconPhotonics or Internet-of-Things technology might help (if it's accidentally rad-hard)

Need careful definition of "raw" data

- Zero-suppression
- Pre-processing (e.g. tracklet)
- Can the same data be used for trigger and offline reconstruction?
  - Avoids that data needs to be transferred twice
  - Requires "offline" quality pre-processing on the detector

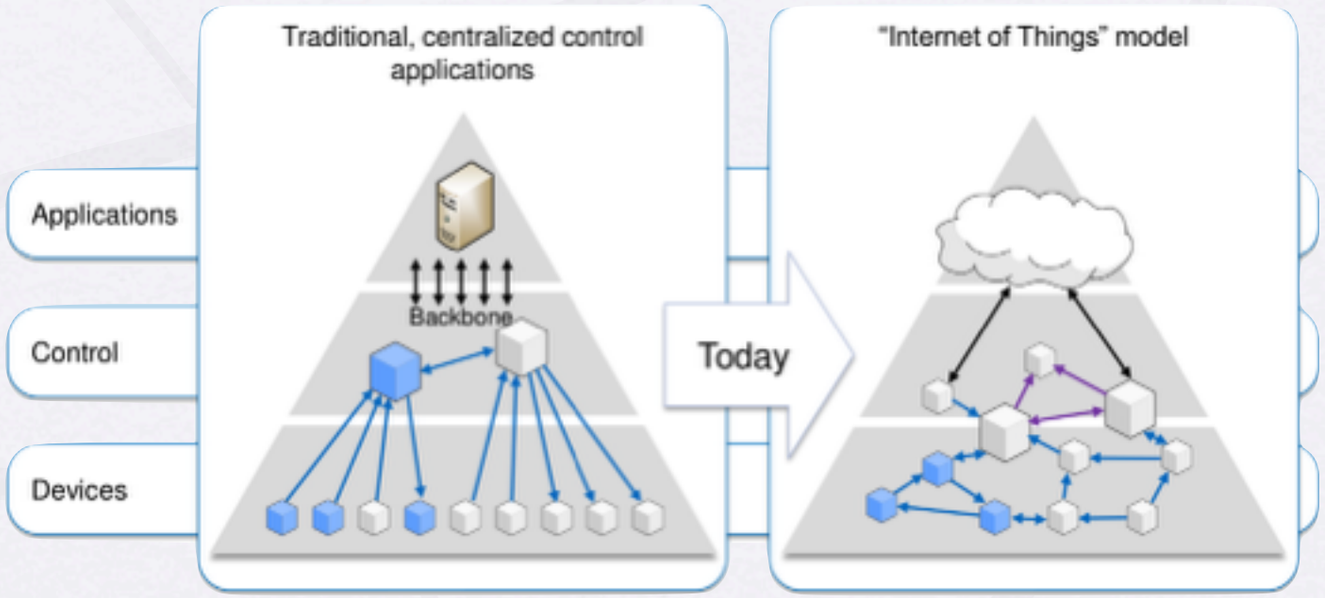
A New World?

Traditional event building uses hardware inefficiently

- Needs a lot of resources to transport data which is mostly unused
  - Used only for L1 trigger
  - Not processed by HLT and then discarded
- Network b/w is used only in one direction

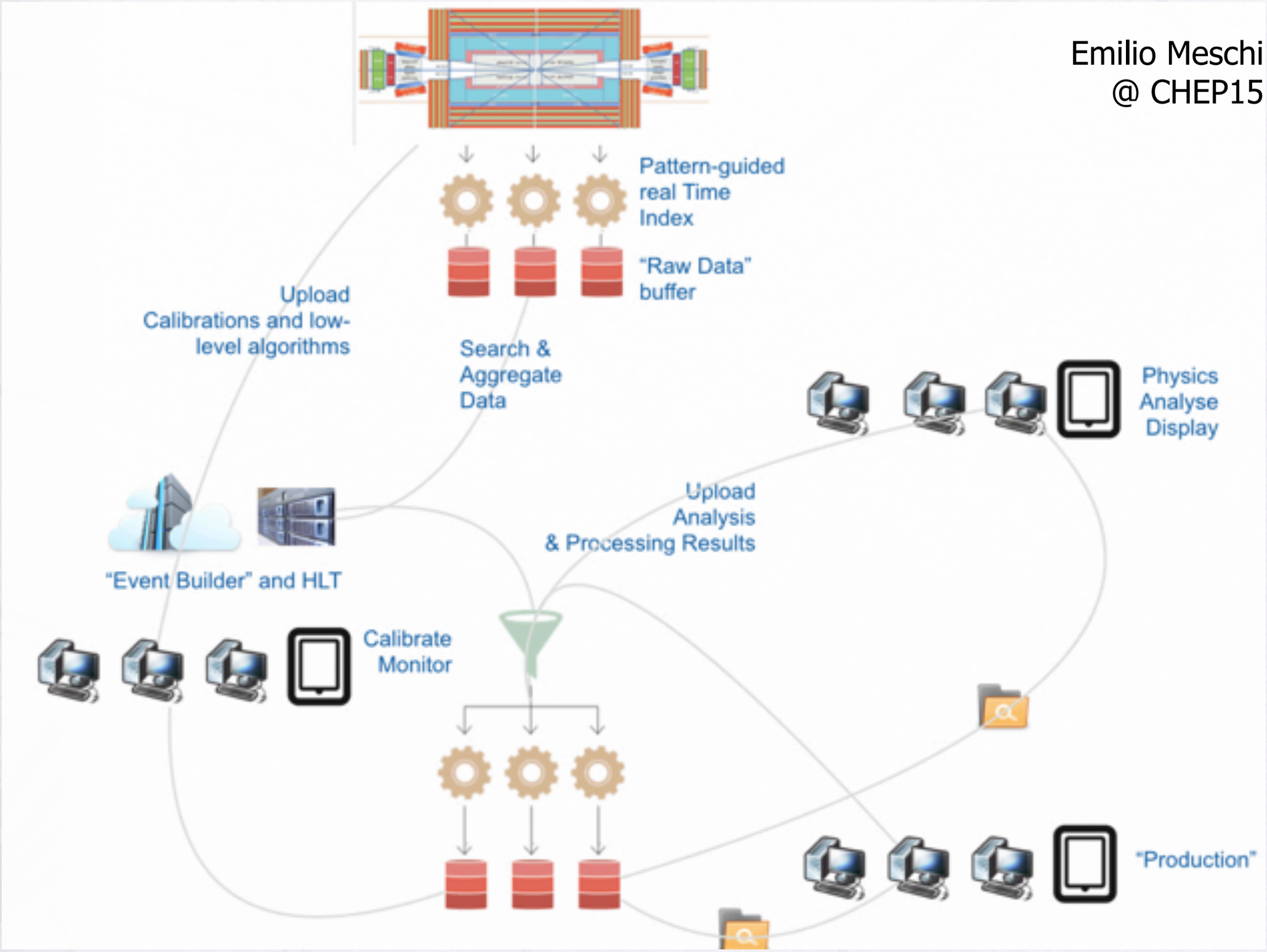
Think more of a mesh (or Internet of Things)

- Leave data as close as possible to the detector
- Pre-process it locally
  - Specialized processors (custom or commercial)
  - Generic CPUs
- Access it remotely
  - Event-building on demand
- Continuous calibrations with feedback to processors
  - Allows near offline-quality selection to reduce the event rate
  - Blurs boundary between online and offline reconstruction



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

Search for Your Data?



Summary

CMS has a complete new DAQ system for LHC run 2

- State-of-the-art technology
- Order of magnitude smaller than old DAQ system
- Achieves run 1 performance
- More work needed to use full potential
- New sub-system readouts are being integrated

The changes for run 3 (2019) will be less radical

- But still a lot of planning and work will be needed

Open field for phase 2 DAQ (2025)

- Evolution of current DAQ as baseline
- Toying with radical new ideas
- Opportunities for R&D on modest budget
  - High-speed, low-mass, rad-hard readout links (mostly unidirectional)
  - Data-reduction schemes on- or near-detector w/o affecting physics
  - Tie event building and selection into a mesh

Endnotes

Acknowledgments

- The CMS DAQ group
- S. Cittolin, S. Erhan, F. Meijers, E. Meschi, N. Neufeld, H. Sakulin, P. Zejdl for their input and slides

References

- Technical Proposal for the Phase-II Upgrade of the CMS Detector (CERN-LHCC-2015-010)
- CMS Phase II Upgrade Scope Document (CERN-LHCC-2015-019)
- The New CMS DAQ System for Run 2 of the LHC (CMS-CR-2014-082)
- 10 Gbps TCP/IP streams from the FPGA for the CMS DAQ Eventbuilder Network (CMS-CR-2013-416)