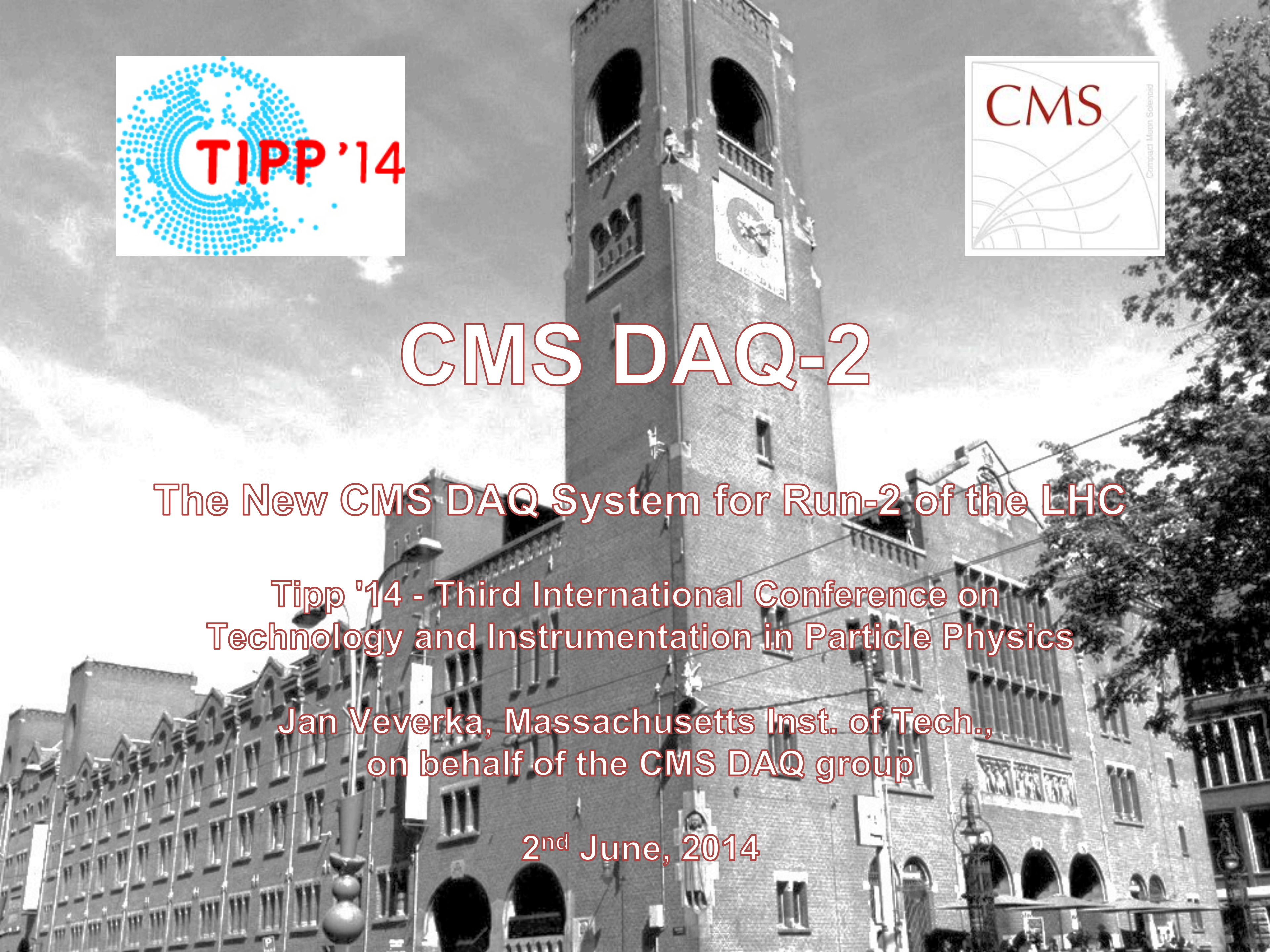


CMS DAQ-2

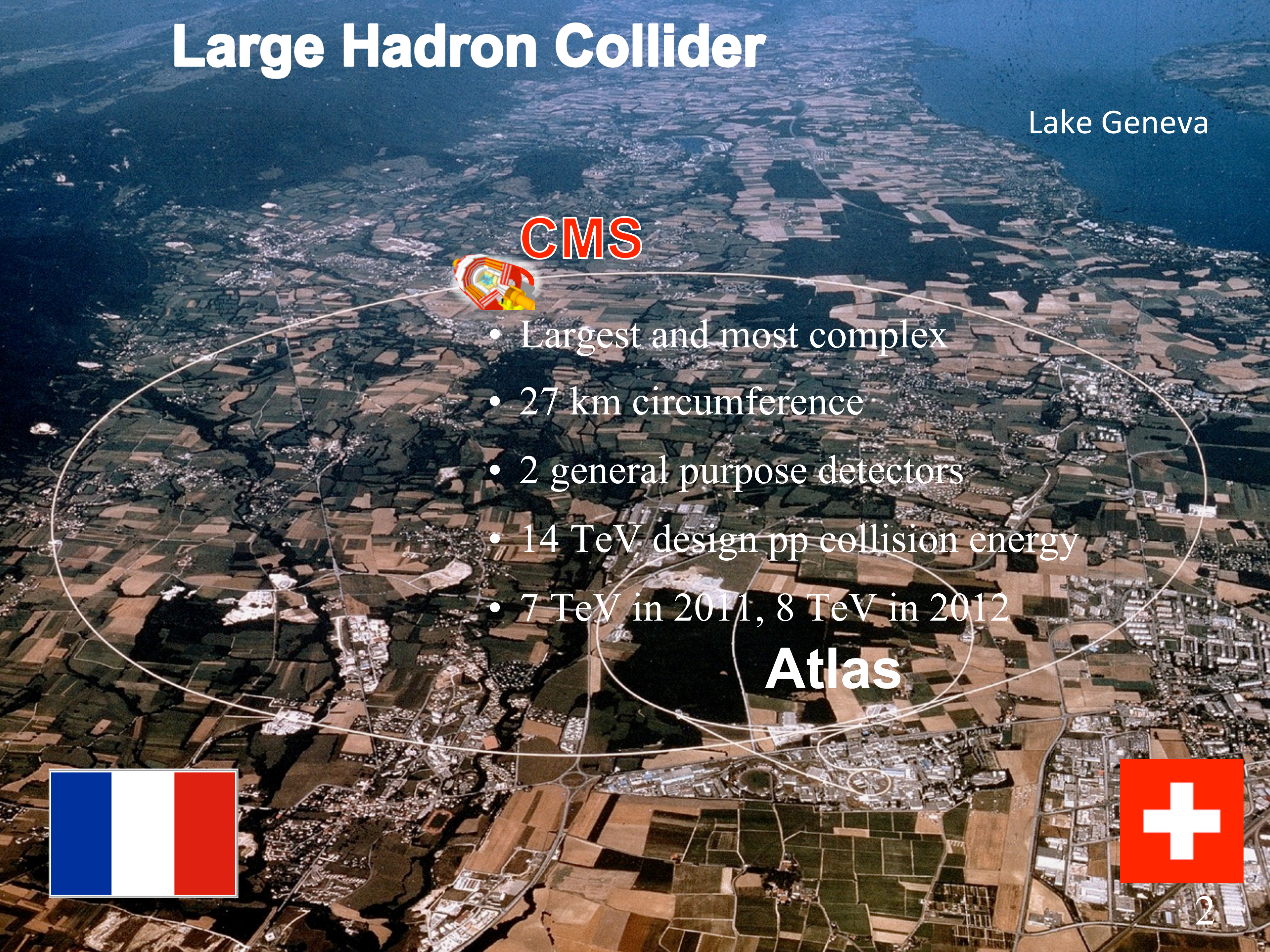
The New CMS DAQ System for Run-2 of the LHC

Tipp '14 - Third International Conference on
Technology and Instrumentation in Particle Physics

Jan Veverka, Massachusetts Inst. of Tech.,
on behalf of the CMS DAQ group

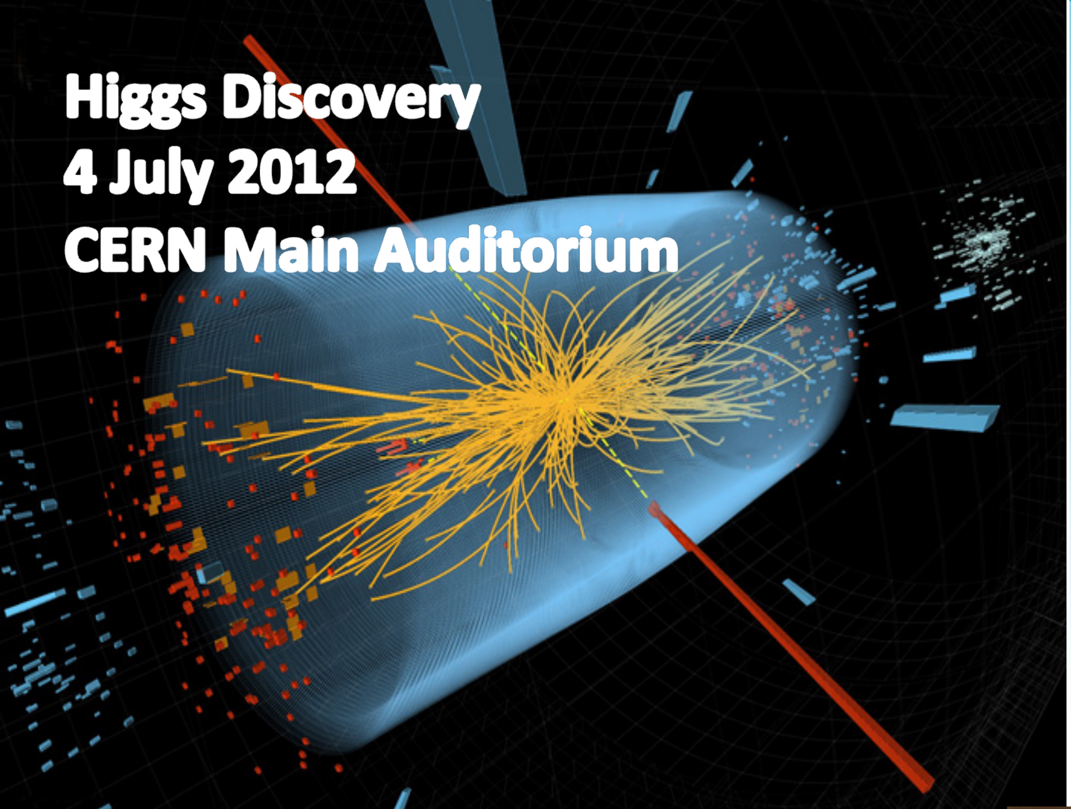
2nd June, 2014

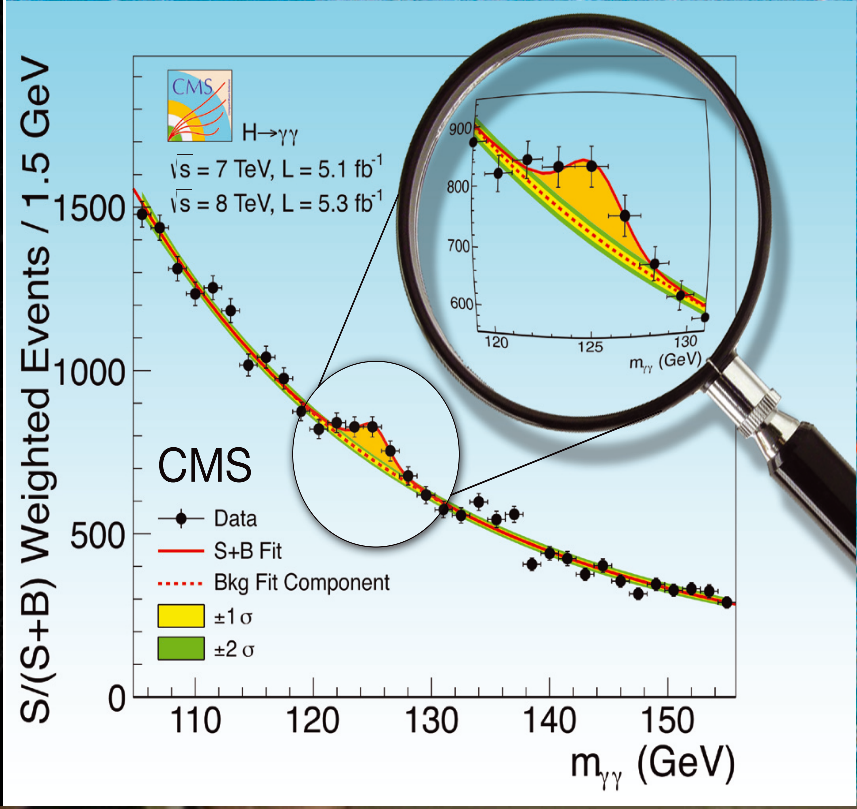
Large Hadron Collider

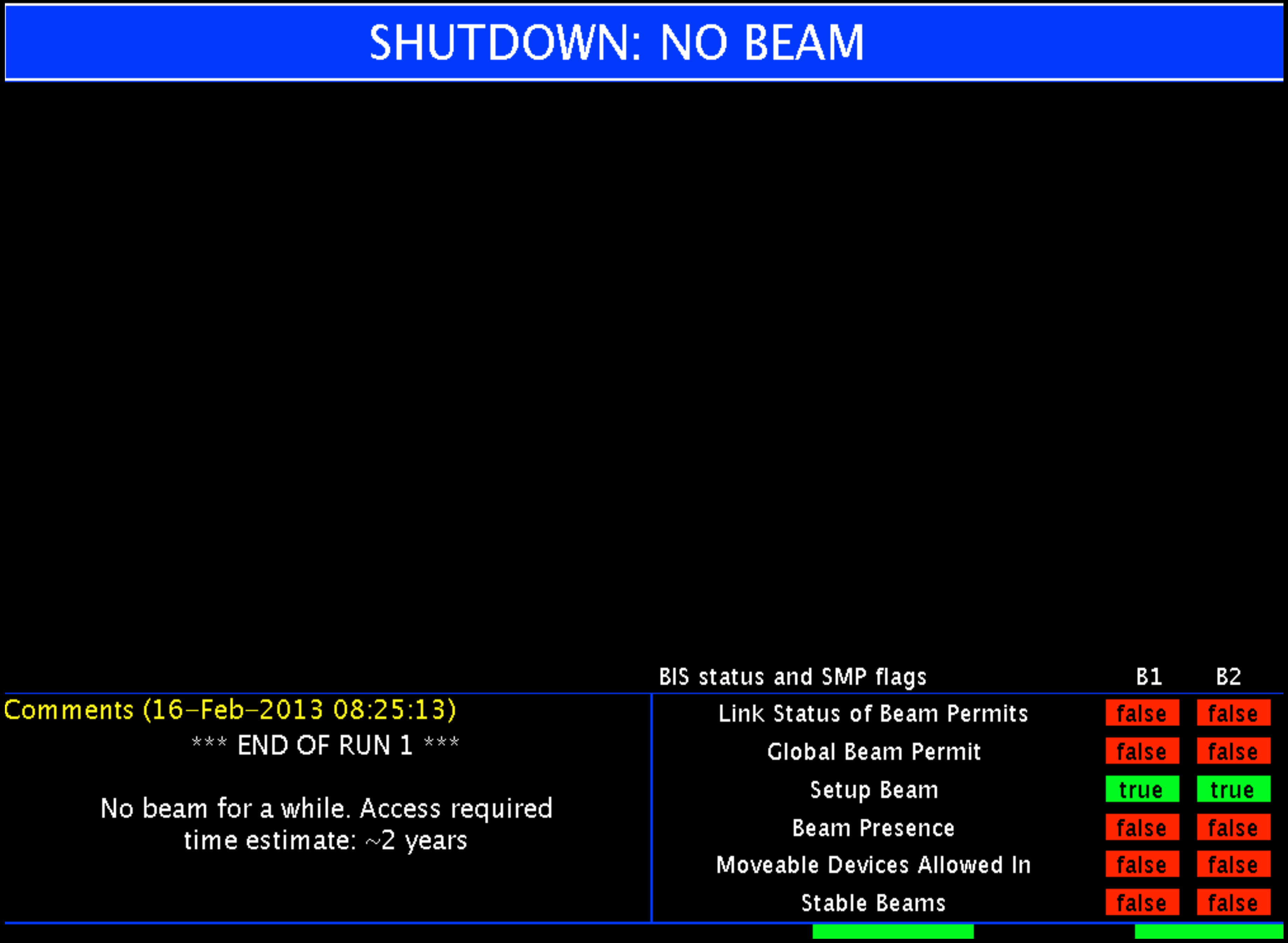


- Largest and most complex
- 27 km circumference
- 2 general purpose detectors
- 14 TeV design pp collision energy
- 7 TeV in 2011, 8 TeV in 2012

Higgs Discovery
4 July 2012
CERN Main Auditorium







SHUTDOWN: NO BEAM

Comments (16-Feb-2013 08:25:13)

*** END OF RUN 1 ***

No beam for a while. Access required
time estimate: ~2 years

BIS status and SMP flags	B1	B2
Link Status of Beam Permits	false	false
Global Beam Permit	false	false
Setup Beam	true	true
Beam Presence	false	false
Moveable Devices Allowed In	false	false
Stable Beams	false	false

Compact Muon Solenoid

2585 physiscists

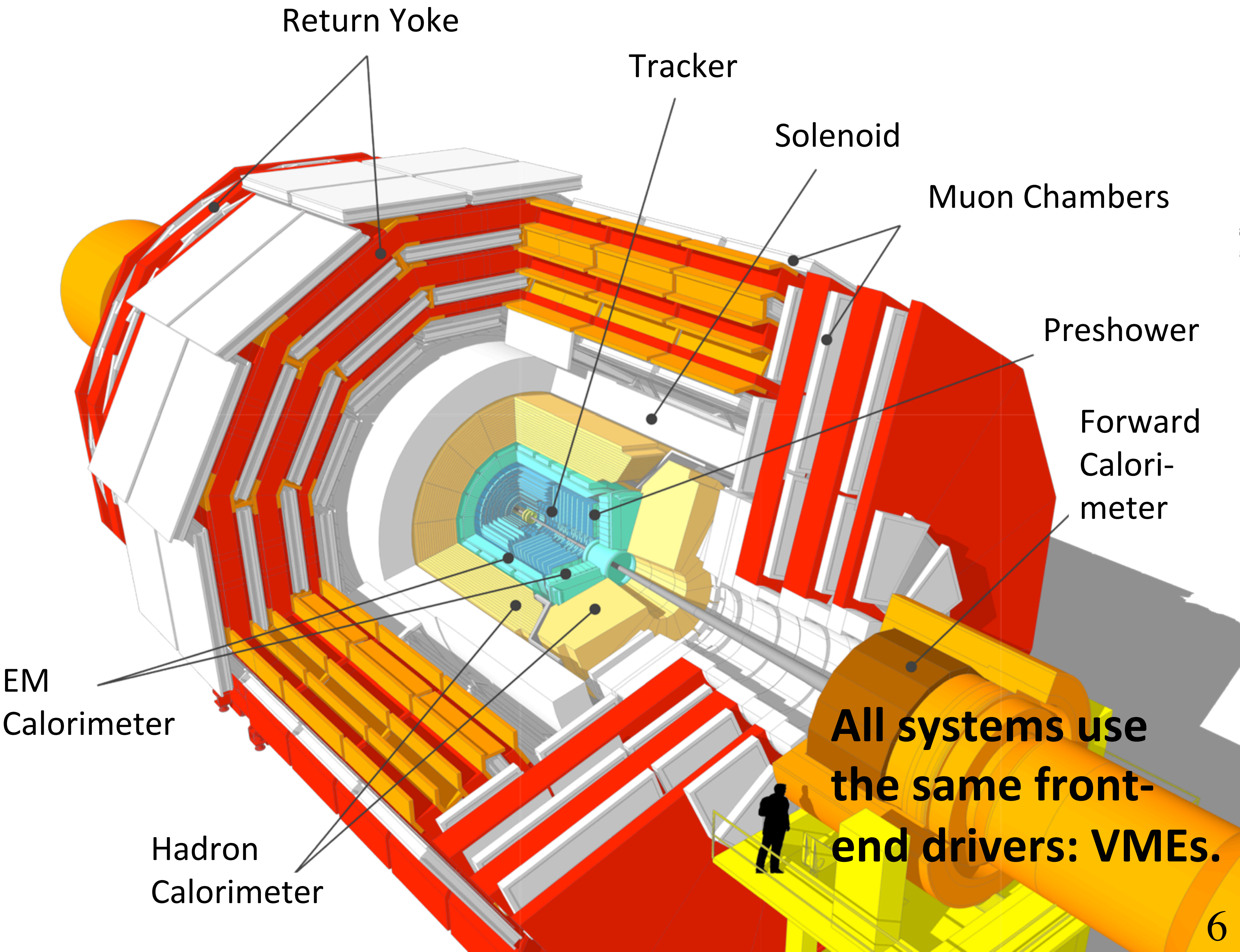
790 engineers

690 undergraduates

281 technicians

182 institutes

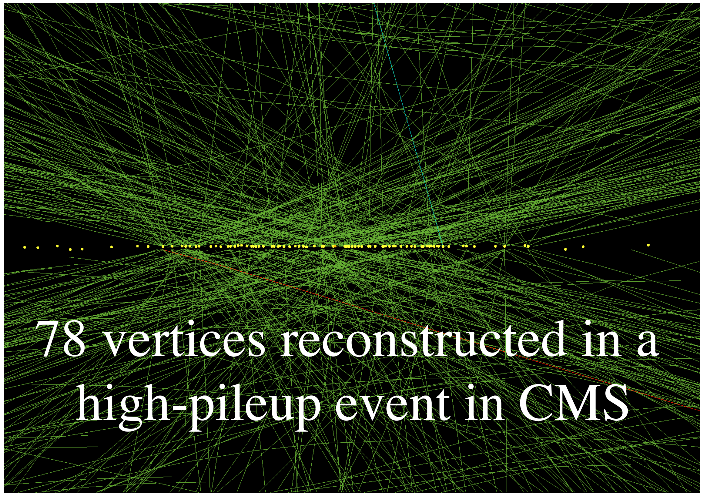
42 countries



All systems use the same front-end drivers: VMEs.

Run-2 Plans (2015-2018)

- 7-8 → 13 TeV pp energy
- 50 → 25 ns bunch spacing
- Up to 50 pileup
- Upgrading and adding several new CMS detector and on-line systems
  - Trigger Control and Distribution, Calorimeter Trigger (2014)
  - Hadron calorimeter readout electronics (2014/15)
  - Fully upgraded Level 1 Trigger (2016)
  - Pixel detector and readout electronics (2017)
- Event size 1 → 2 MB
- Support both *new* and legacy front-end drivers (*μTCA* and VME).



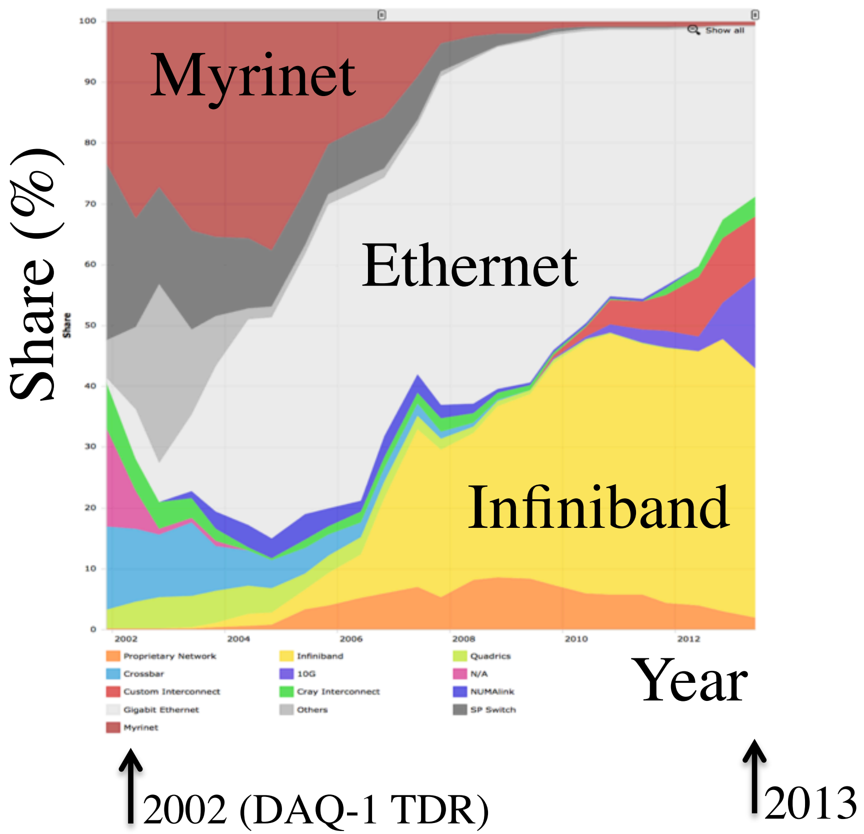
78 vertices reconstructed in a high-pileup event in CMS

Bunch Spacing (ns)	Beam Current (10^{11} e)	Emittance (μm)	Peak Lumi	Pileup
25	1.15	3.5	0.92	**21**
25	1.15	1.9	1.6	**43**
50	1.6	2.3	0.9-1.7	**40-76**
50	1.6	1.6	2.2	**106**

Why New DAQ?

- New requirements
- Ageing hardware
  - Most components at the end of life cycle
  - Run-1 NICs based on PCI-x
- New technologies
  - Myrinet widely used when DAQ-1 was designed
  - Ethernet and Infiniband dominate the Top-500 supercomputers

Top500.org by Interconnect Family



Myrinet

Ethernet

Infiniband

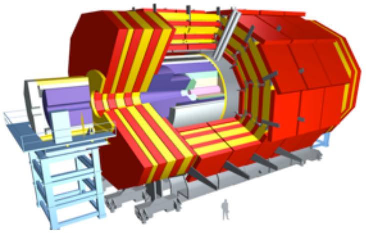
Share (%)

Year

2002 (DAQ-1 TDR)

2013

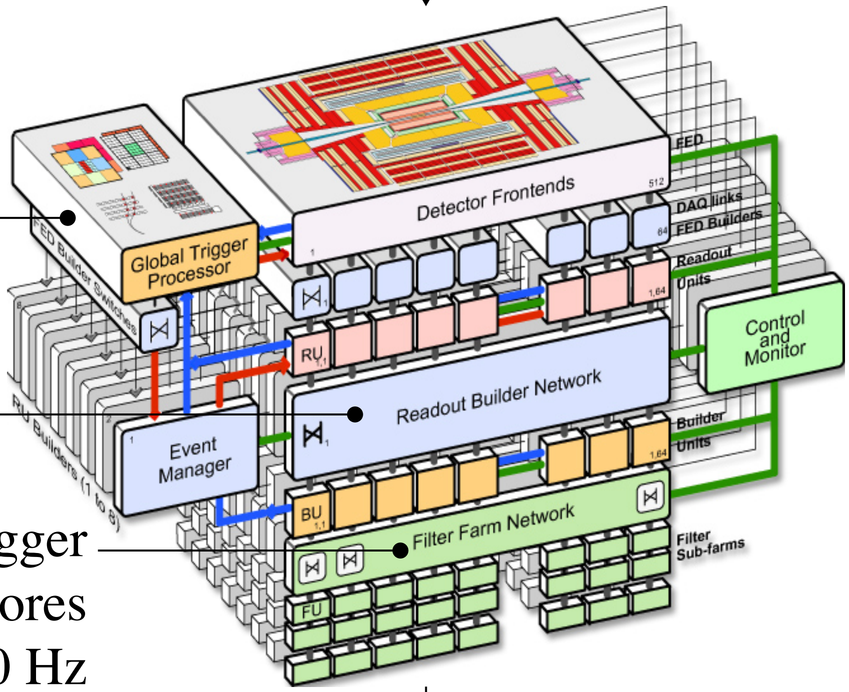
CMS DAQ-1 Overview



≤ 1 MB/event at ≤ 20 MHz

Level 1 Trigger
Custom electronics
Accept 100 kHz

2-Stage Event Builder
Myrinet, Gigabit Ethernet



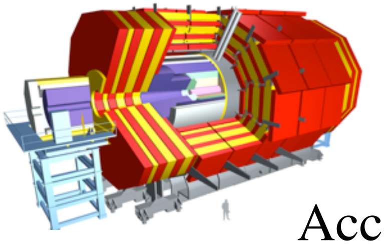
100 GB/s

High Level Trigger
13,000 Cores
Accept 500 Hz

≤ 1 GB/s

99.6 % Availability
(2010-2013 physics)

CMS DAQ-2 Overview

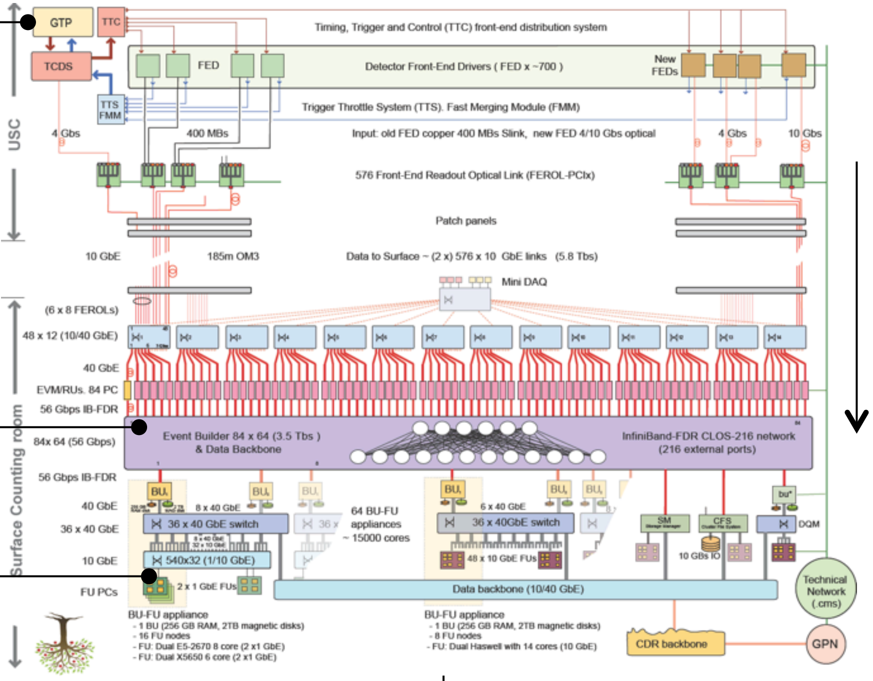


≤ 2 MB/event at ≤ 40 MHz

L1 Trigger
Accept 100 kHz

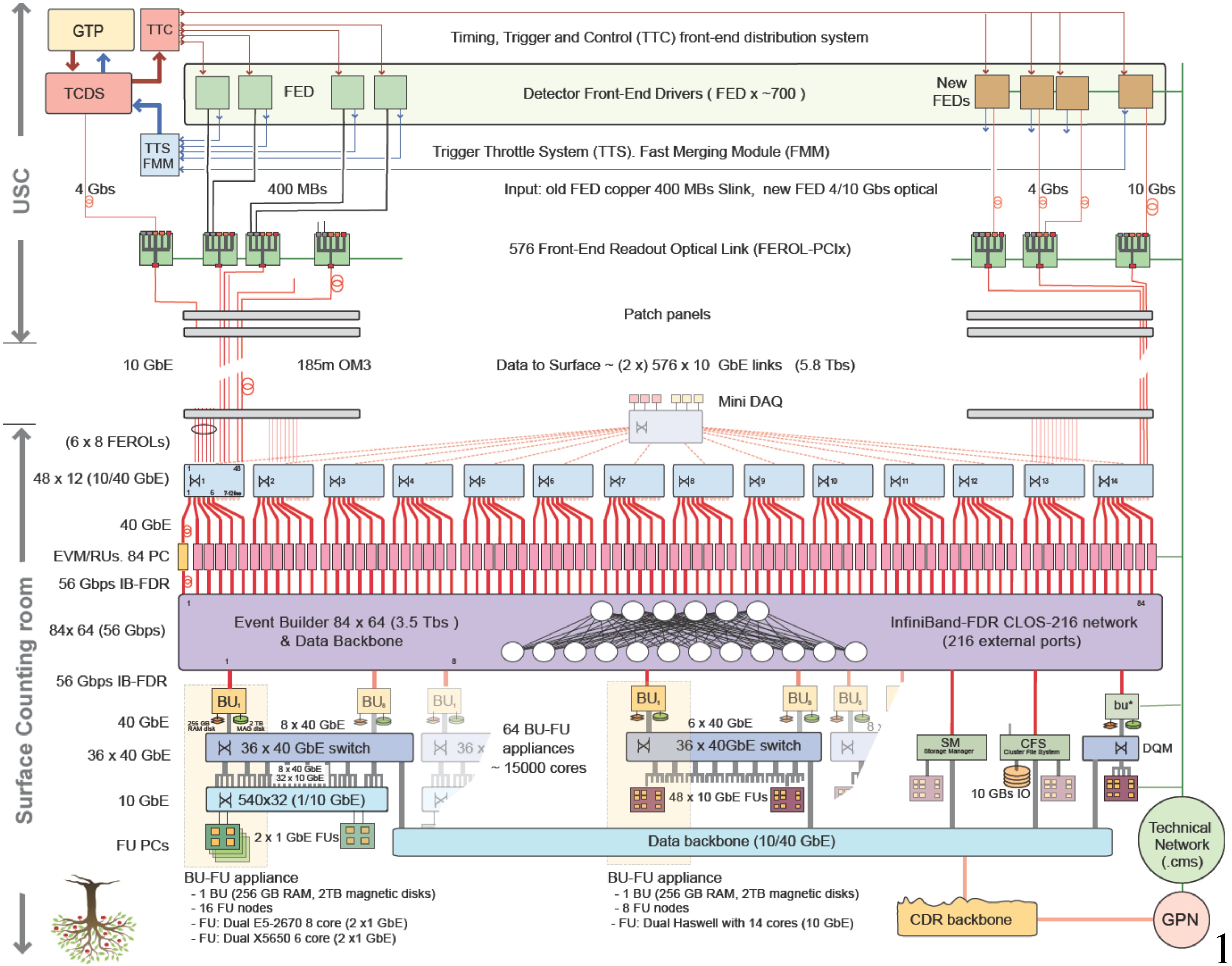
Single-Stage Event Builder
10/40 Gb/s Ethernet
56 Gb/s Infiniband

High Level Trigger
~15,000 Cores


GTP
TTC
TCDS
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
New FEDs
TTS FMM
Trigger Throttle System (TTS). Fast Merging Module (FMM)
USC
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)
Mini DAQ
(6 x 8 FEROLs)
48 x 12 (10/40 GbE)
40 GbE
EVM/RUs. 84 PC
56 Gbps IB-FDR
Surface Counting room
84x 64 (56 Gbps)
Event Builder 84 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
540x32 (1/10 GbE)
2 x 1 GbE FUs
64 BU-FU appliances ~ 15000 cores
6 x 40 GbE
36 x 40GbE switch
48 x 10 GbE FUs
SM
CFS
10 GBs IO
bu*
DQM
Data backbone (10/40 GbE)
Technical Network (.cms)
CDR backbone
GPN
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 14 cores (10 GbE)

200 GB/s

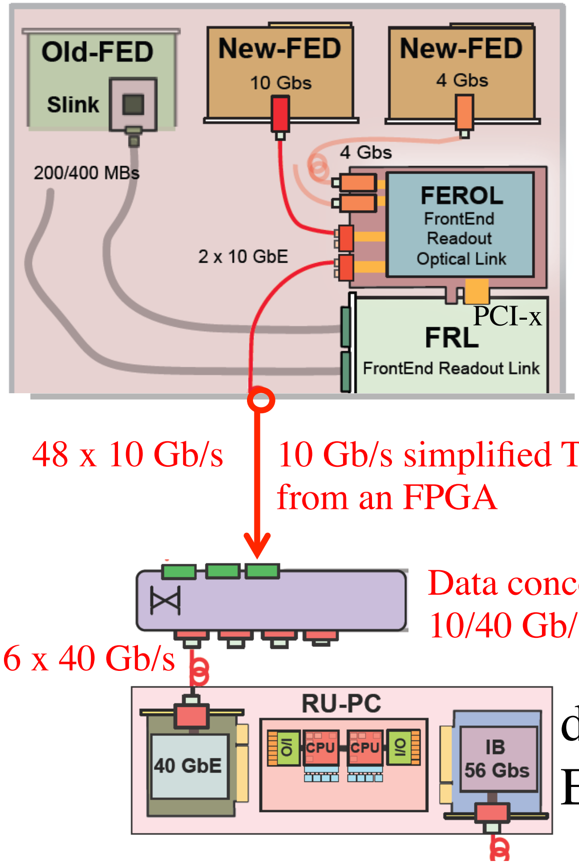
≤ 2 GB/s



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 14 cores (10 GbE)

Frontend-Optical Link & Data Concentrator



48 x 10 Gb/s

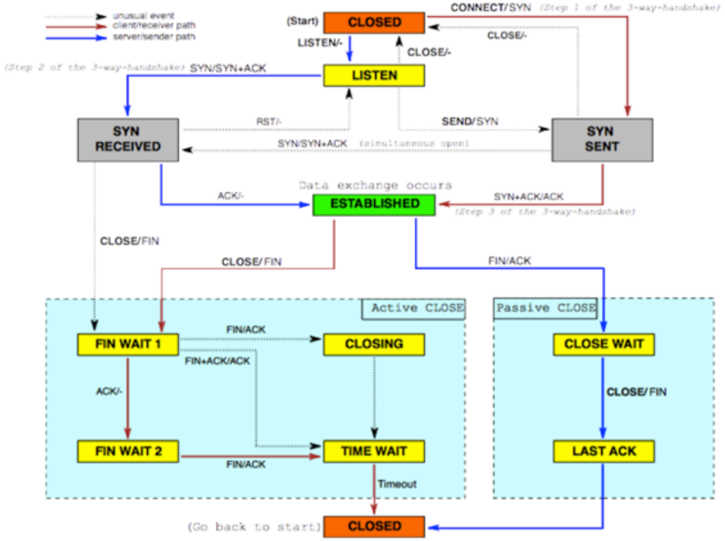
10 Gb/s simplified TCP/IP from an FPGA

Data concentration:
10/40 Gb/s Ethernet switch

6 x 40 Gb/s

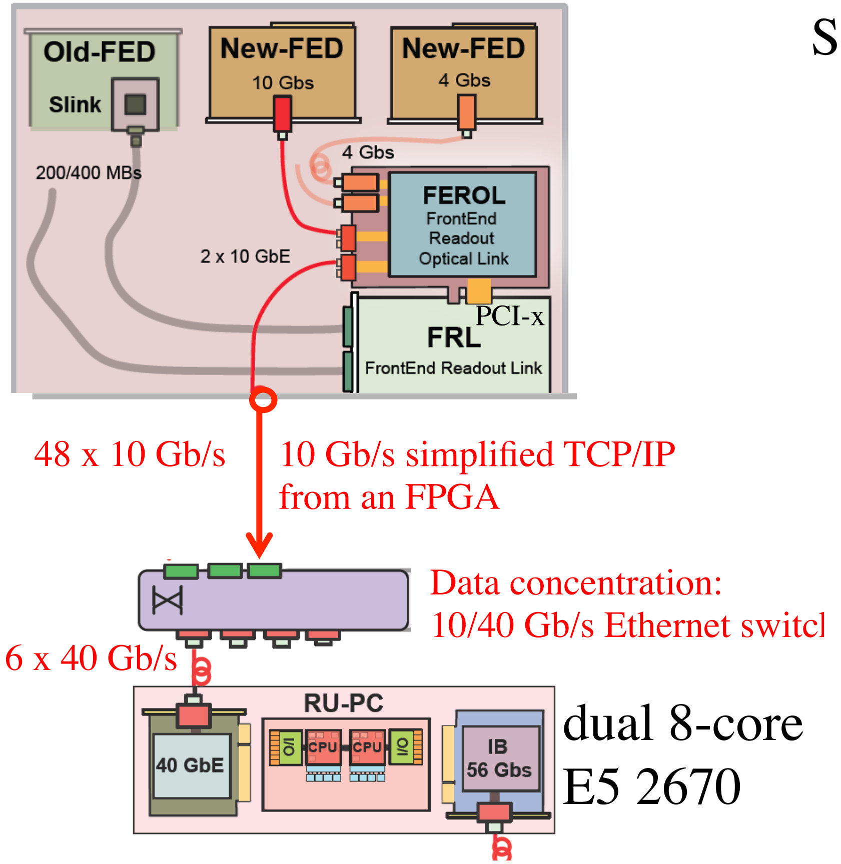
dual 8-core E5 2670

Full TCP/IP



Implementation in
an FPGA is challenging

Frontend-Optical Link & Data Concentrator



48 x 10 Gb/s

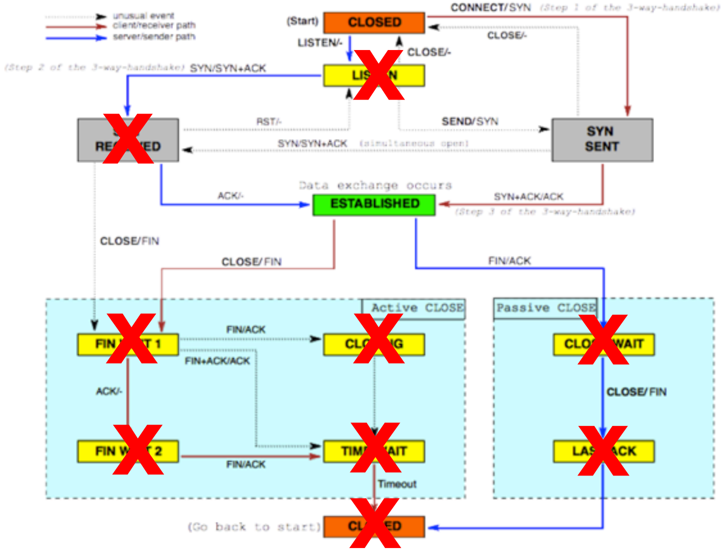
10 Gb/s simplified TCP/IP from an FPGA

Data concentration: 10/40 Gb/s Ethernet switcl

6 x 40 Gb/s

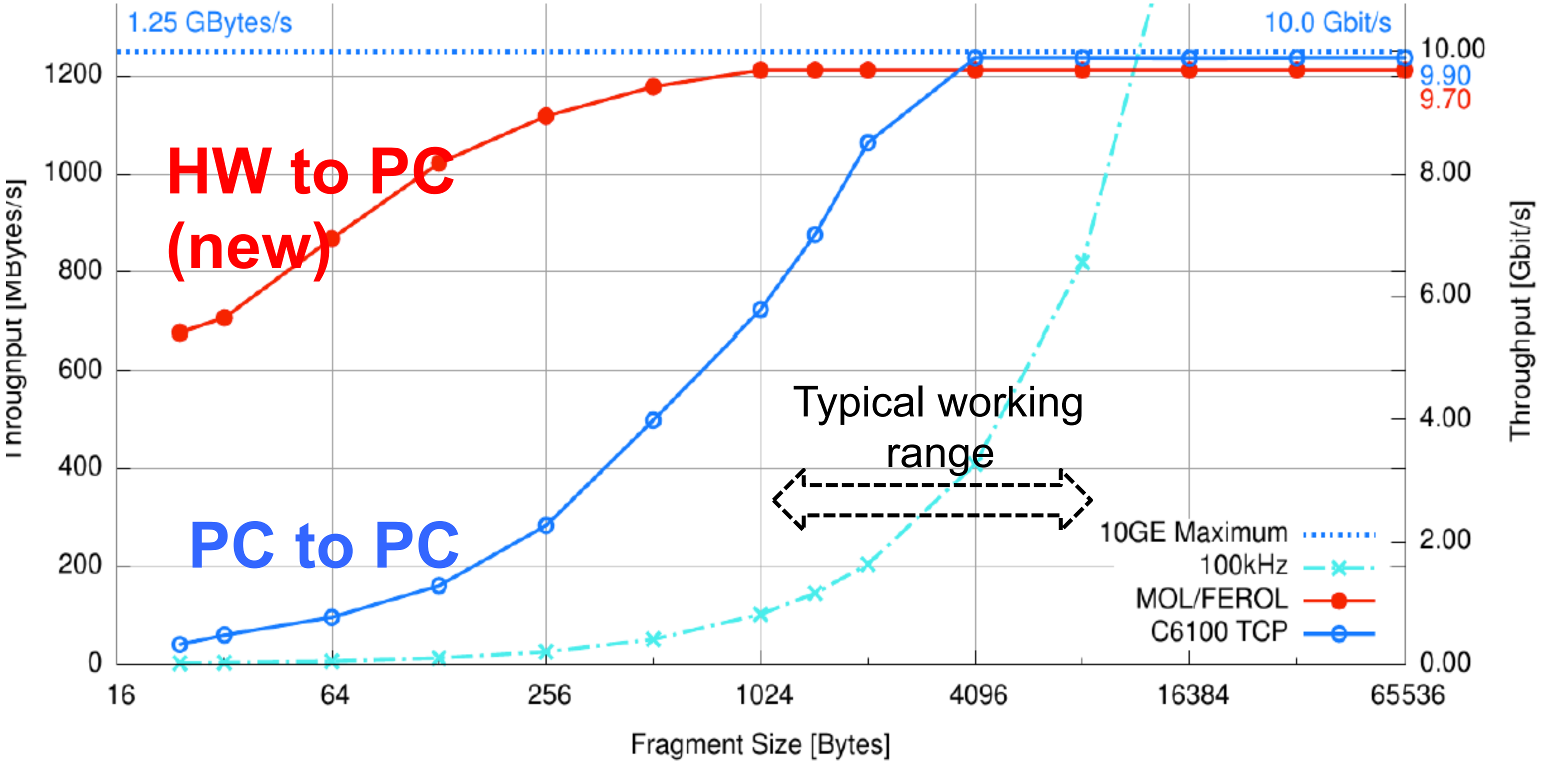
dual 8-core E5 2670

Simplified Unidirectional TCP/IP



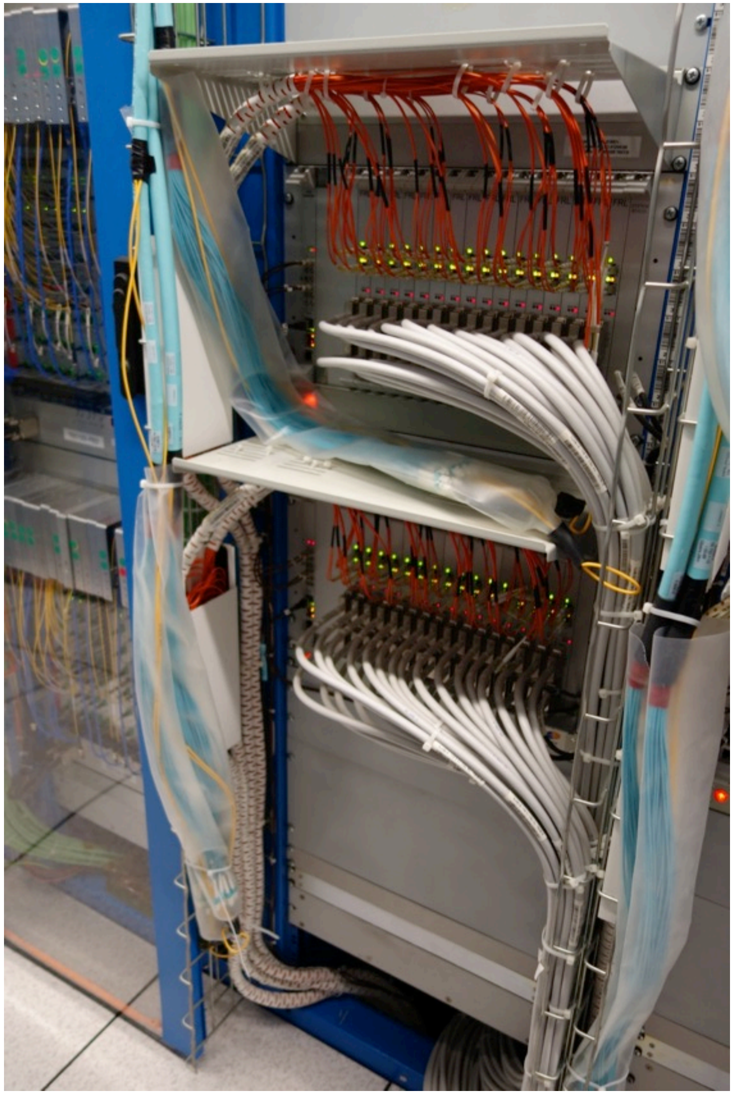
Only need 3 states ☺

10 Gb/s TCP/IP link from FPGA to PC



HW to PC: 9.7 Gb/s for fragments > 1 kB ☺

(receiving PC with 10 Gb/s NIC, performance tuned)

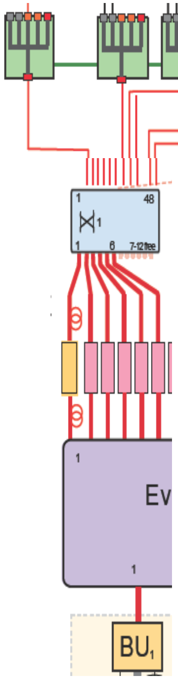


FRL/Myrinet

FRL/FEROL 10 Gb/s Ethernet

Switchover completed 2 weeks ago.

Data Concentrator Performance

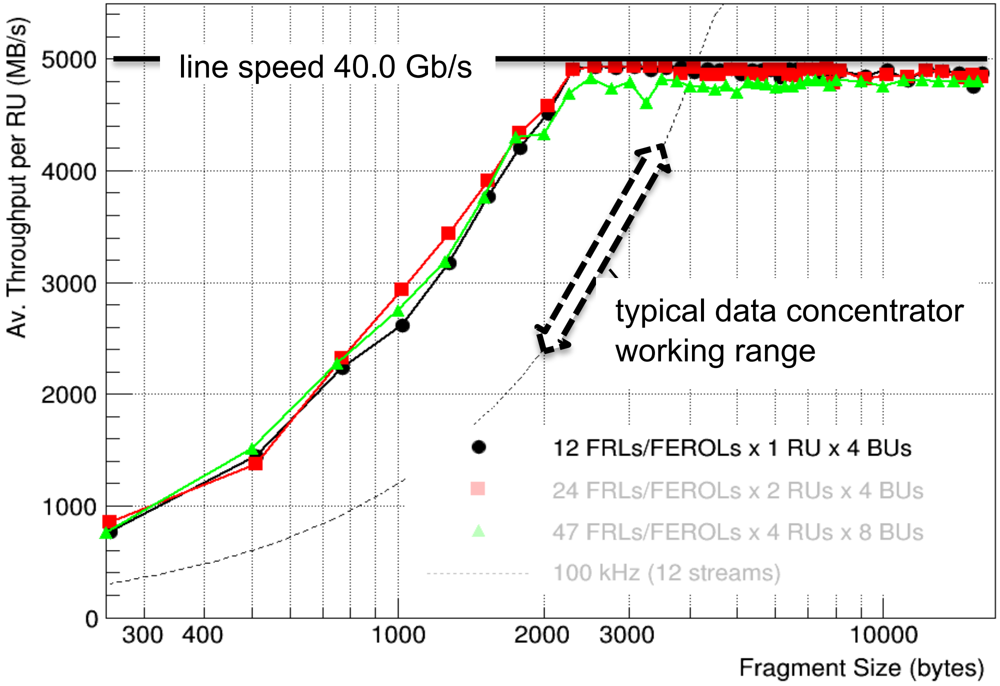


12 x FEROLs

10/40 Gb/s Ethernet switch
Mellanox MSX1024

1 x RU PC

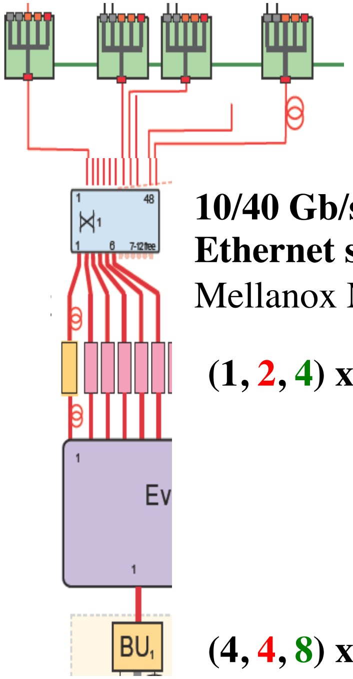
4 x BU PC



typical data concentrator working range

Performance meets requirements. ☺

Data Concentrator Performance

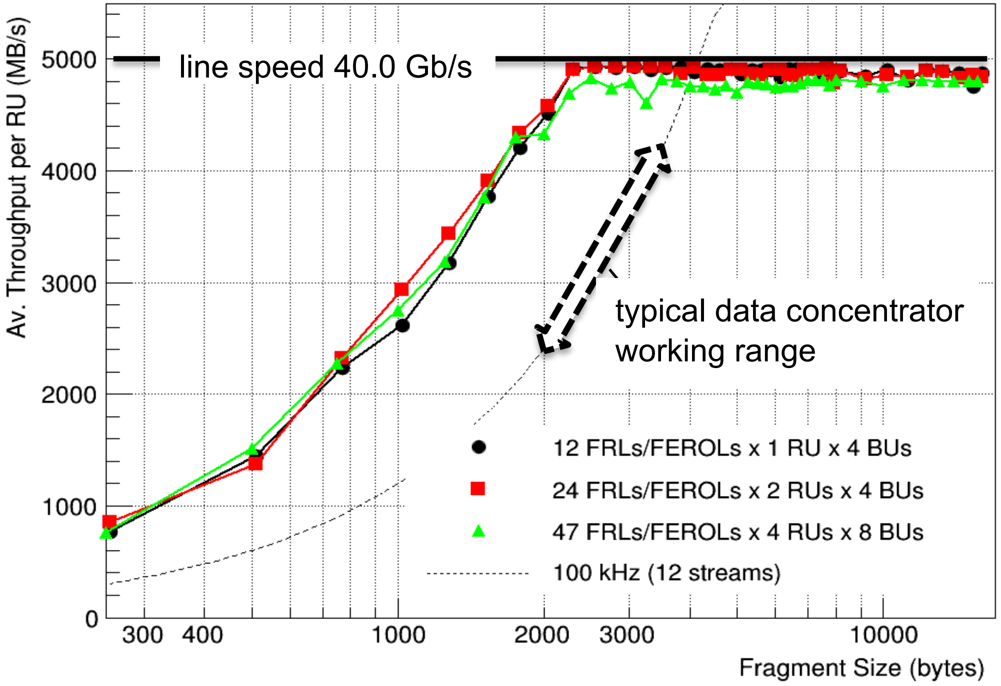


(12, 24, 47) x FEROLs

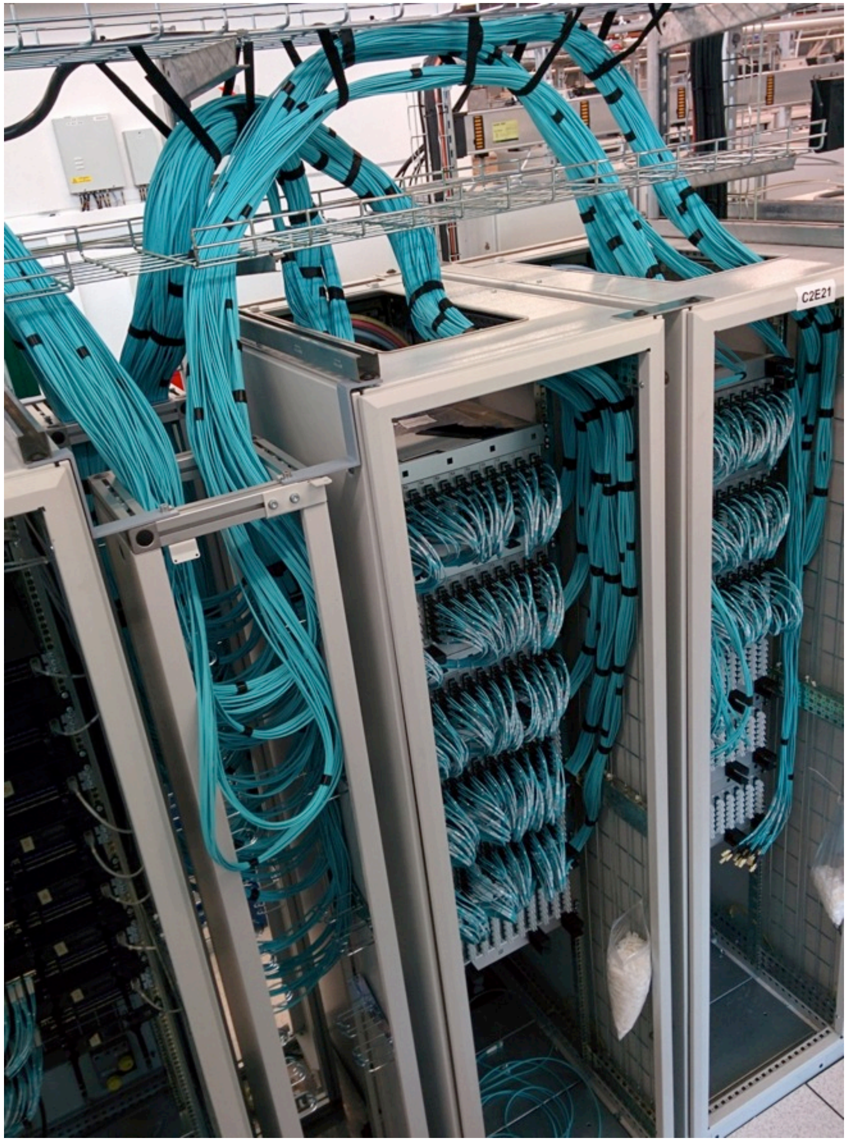
10/40 Gb/s Ethernet switch
Mellanox MSX1024

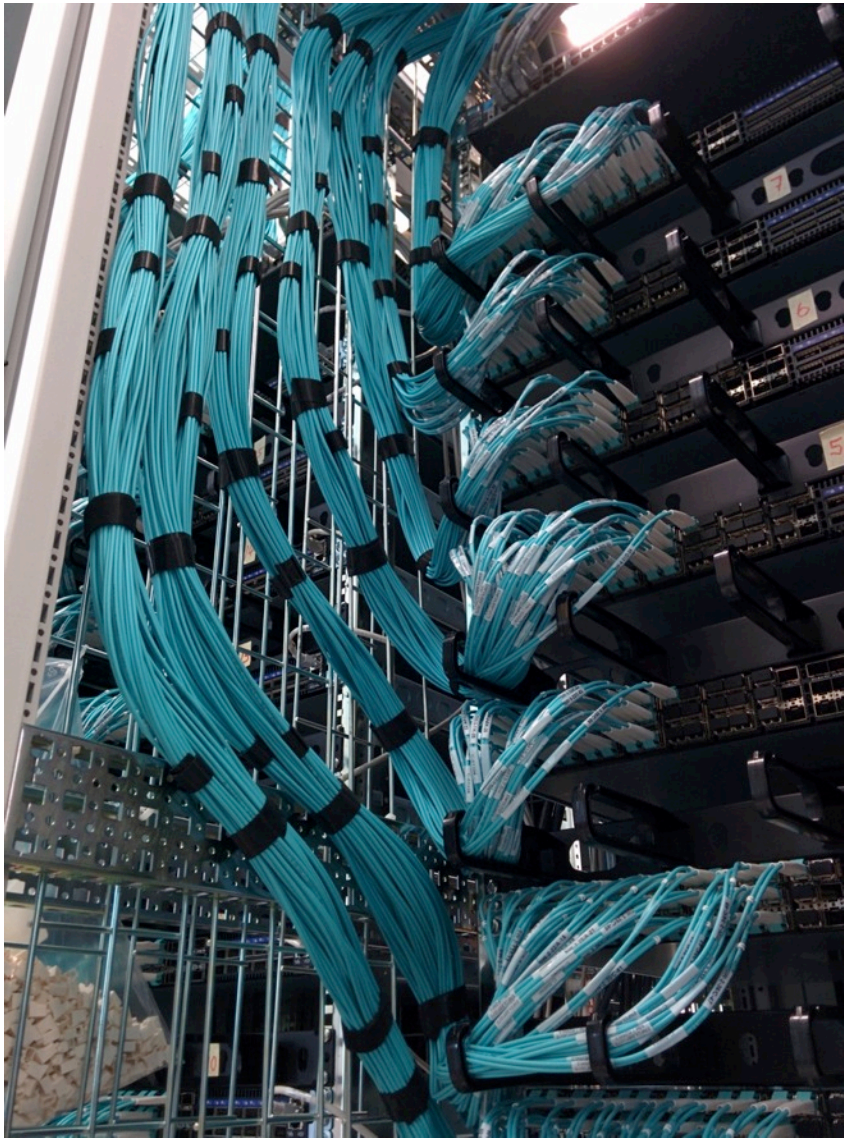
(1, 2, 4) x RU PC

(4, 4, 8) x BU PC



Performance meets requirements.
Scales from a single concentrator to a fully loaded switch. ☺

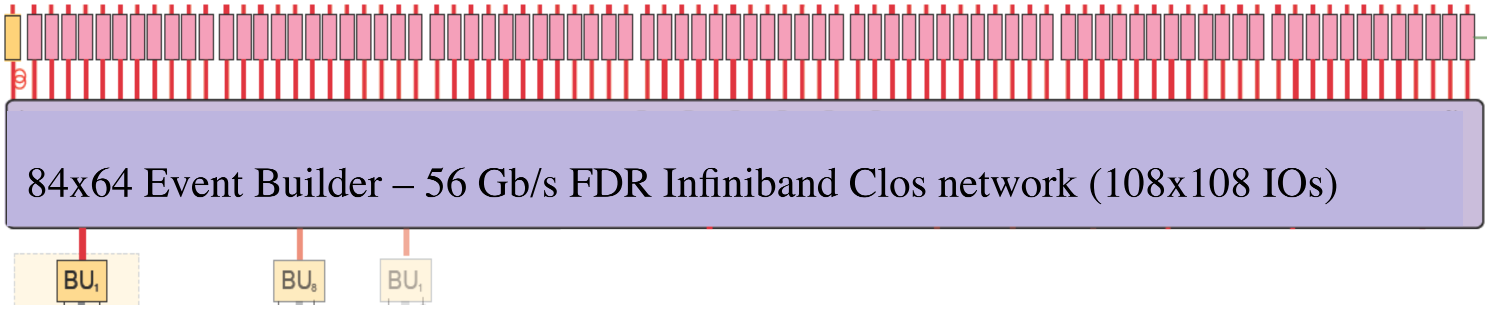




Data concentrator patch panels … and switches

See talk by Andrew Forrest

Core Event Builder

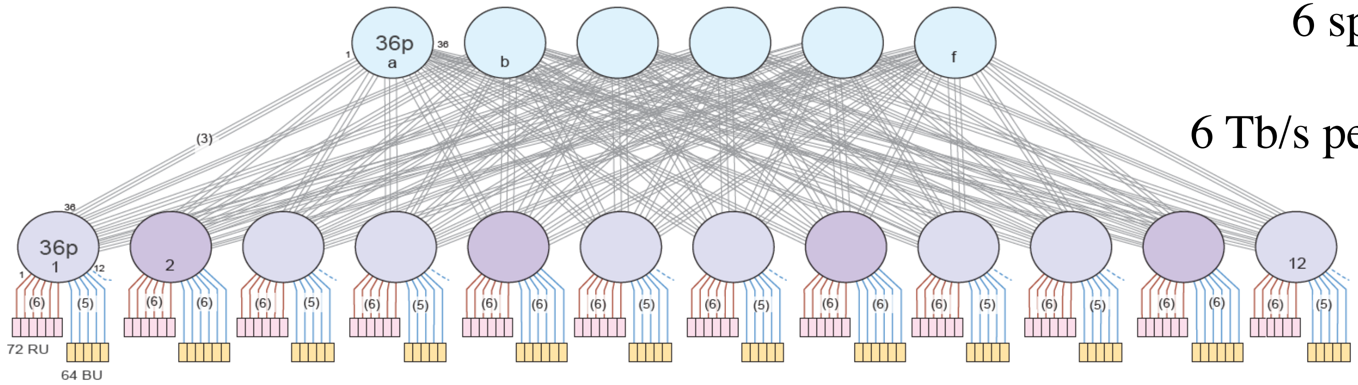

84x64 Event Builder – 56 Gb/s FDR Infiniband Clos network (108x108 IOs)
BU1
BU8
BU1

4 Tb/s in

3.5 Tb/s out

Infiniband

- reliable in hardware at link level (no heavy software stack needed)
- supports credit-based flow control
- switches do not need to buffer
- can construct large network from smaller switches



6 spine switches

6 Tb/s per direction

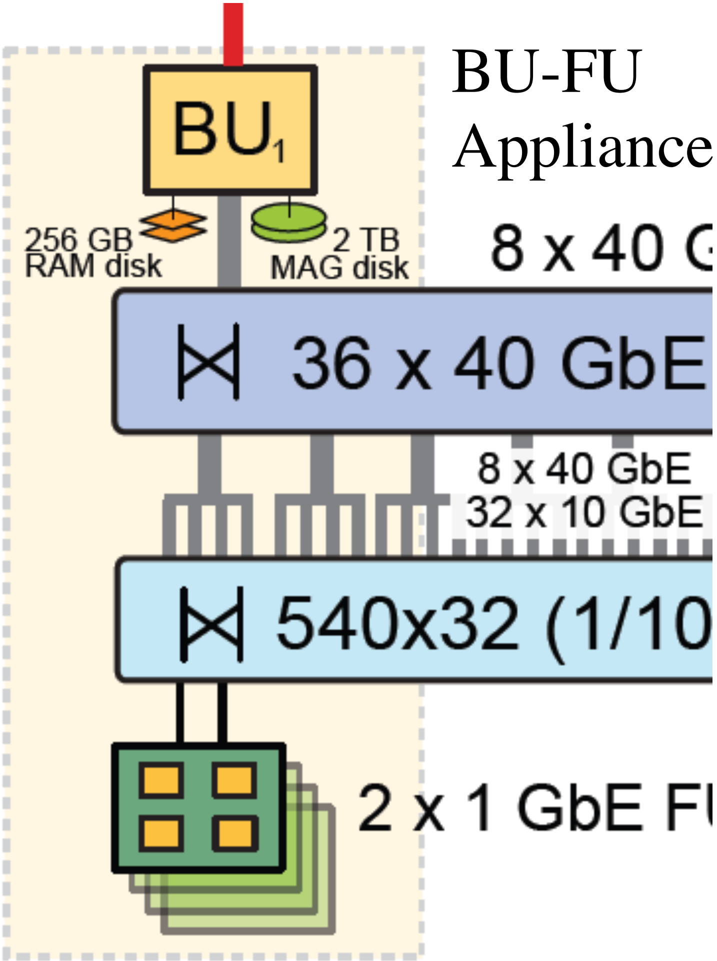
12 leaf switches

Inputs and outputs mixed on leafs to better utilize leaf-to-spine connections

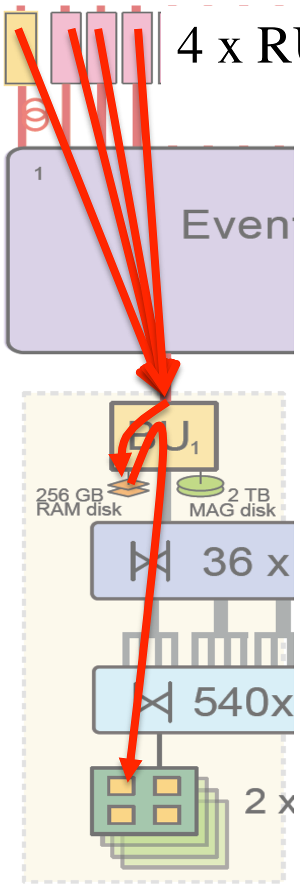
Infiniband CLOS network

File-based Filter Farm

- Goal: Fully factorize acquisition (XDAQ) and reconstruction (CMSSW) SW
  - Release cycle
  - Version of compiler and externals
  - Debugging
- Use files for the I/O (same as off-line)
- BU writes data to files on a RAM disk (256 GB/BU)
- 8-16 FUs mount it via NFS4 and run up to 2 SW processes per core reading the files.
- FU processes merge their outputs into a single file per FU and then write it back to a disk on the BU



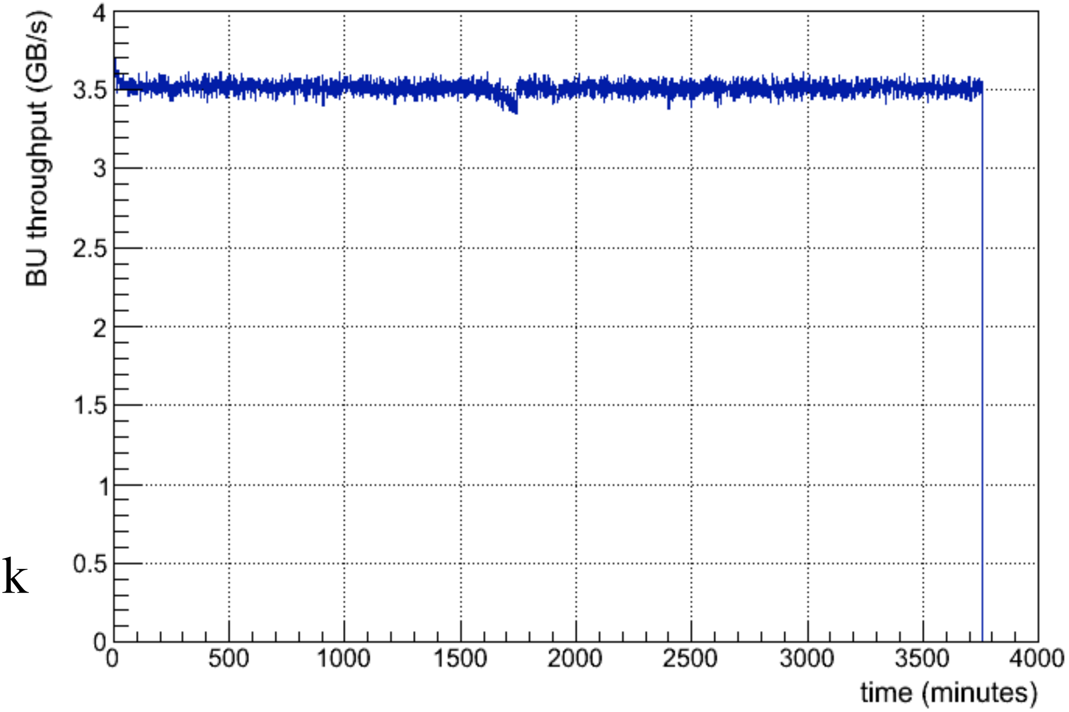
RU-BU-FU Performance



4 x RU

1 x BU
to/from
RAM disk

8 x FU
32 CMSSW
processes per FU



3.5 GB/s x 64 BUs: ~ 220 GB/s ☺

DAQ-2 High Level Trigger Farm

72x

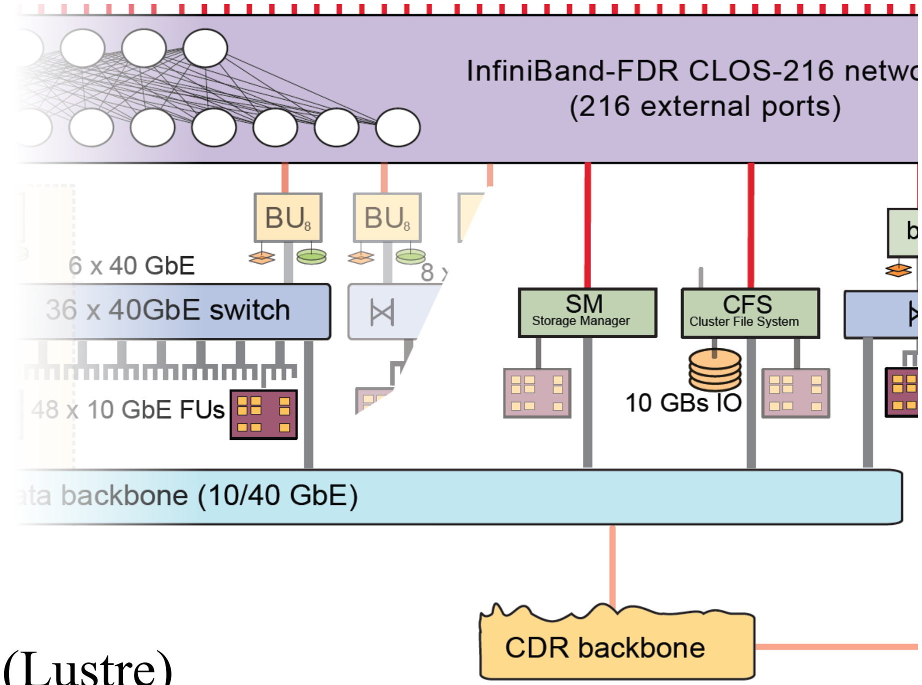
64x

	May 2011	*May 2012*	*Early 2015*
DAQ Version	DAQ-1	DAQ-1	DAQ-2
Model	Dell Power Edge c6100	Dell Power Edge c6220	To be decided
Form factor	4 motherboards in 2U box	4 motherboards in 2U box	
CPUs per mother-board	2 x **6-core** Intel Xeon 5650 **Westmere**, 2.66 GHz, hyper-threading, 24 GB RAM	2 x **8-core** Intel Xeon E5-2670 **Sandy Bridge**, 2.6 GHz, hyper-threading, 32 GB RAM	2 x **14-core** Intel **Haswell**
# Motherboards	288	256	256
# Cores	**3456**	**4096**	**7168**
Data link	2 x 1Gb/s	2 x 1Gb/s	1 x 10 Gb/s

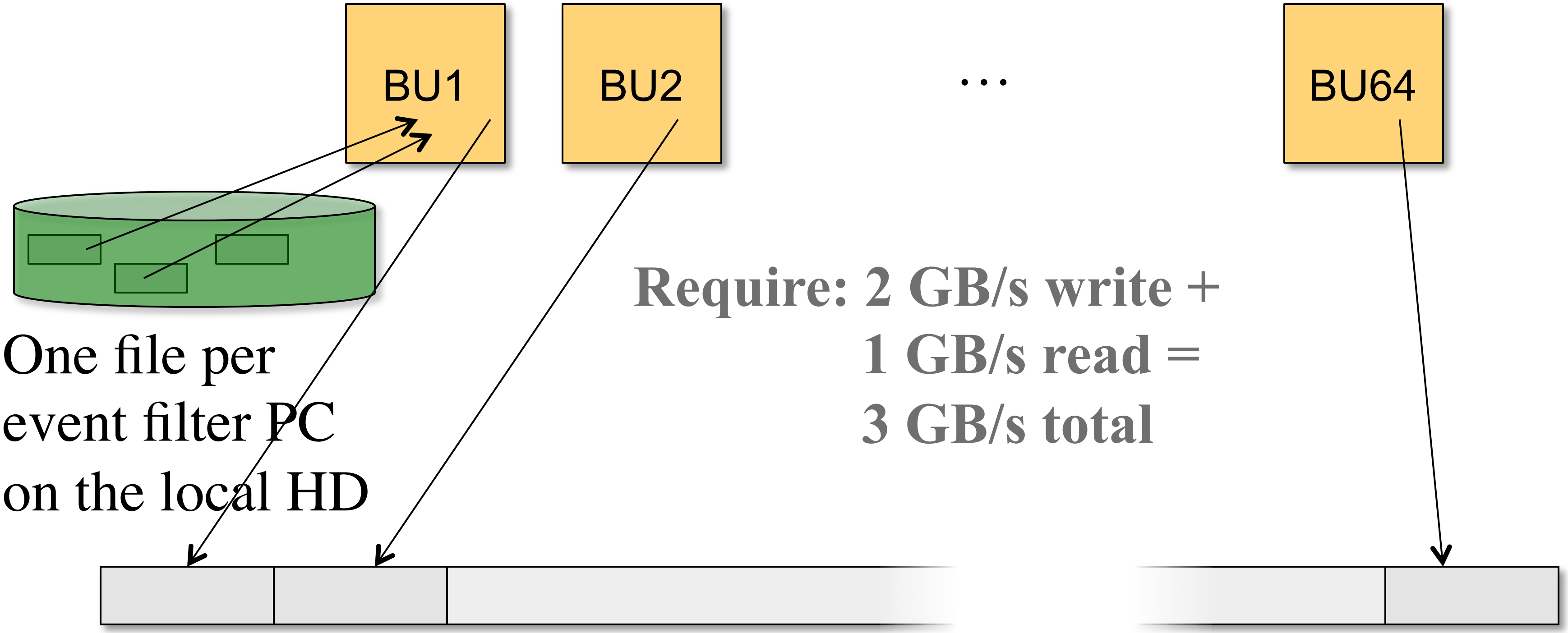
Total: ~ 15k cores on 800 motherboards (to be confirmed)

Merging and Storage

- File-Based Filer Farm produces output files
  - After merging on FU: 800 files x 10 streams scattered over 64 BUs every 23 seconds
  - To be merged to 1 file per stream 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
  - 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



Merging and Storage



One file per
event filter PC
on the local HD

Require: 2 GB/s write +
1 GB/s read =
3 GB/s total

Single output file in the cluster file system

Test system performance (NetApp) with 14 clients, 4 Object storage servers, Lustre 2.4 (1/2 scale): 4.8 GB/s write ☺

Summary

- CMS is installing a new DAQ system for Run-2 of the LHC
  - New optical SLINK-express readout link
  - 10 Gb/s TCP/IP from an FPGA
  - 10/40 Gb/s Ethernet data concentrator
  - 56 Gb/s FDR Infiniband core event builder
  - File-based high-level trigger (via 1/10/40 Gb/s Ethernet)
  - Cluster File System for storage
  - Throughput doubled to 200 GB/s
- Performance looks good. ☺
- Installation is advancing well. ☺
- Commissioning during the remainder of 2014.

References

- **CHEP 2013**, 20th International Conference on Computing in High Energy and Nuclear Physics 2013, Amsterdam, Netherlands, 14 - 18 Oct 2013
  - Andre Georg Holzner (UC San Diego) et al (CMS DAQ group), *The new CMS DAQ system for LHC operation after 2014 (DAQ2)*, 02 Nov 2013, CMS CR-2013/394, http://cds.cern.ch/record/1626828/
  - Petr Žejdl (CERN) et al (CMS DAQ group), *10Gbps TCP/IP streams from the FPGA for High Energy Physics*, 08 Nov 2013, CMS CR-2013/402, http://cds.cern.ch/record/1639563
- **RT 2014**, 19th Real-Time Conference, 26-30 May 2014, Osaka University, Nara (Japan)
  - Andrea Petrucci (CERN) et al (CMS DAQ group), *Achieving High Performance with TCP over 40GbE on NUMA architectures for CMS Data Acquisition*, CMS CR-2014/081
  - Hannes Sakulin (CERN) et al (CMS DAQ group), *The new CMS DAQ system for run 2 of the LHC*, CMS CR-2014/082
- **TIPP 2013** (coming up in 20 minutes) 3rd talk in the on-going session scheduled for 16:50),
  - Andrew Kevin Forrest (CERN) et al (CMS DAQ group), *Boosting Event Building Performance using Infiniband FDR for CMS Upgrade*, https://indico.cern.ch/event/192695/session/2/?slotId=0#20140602