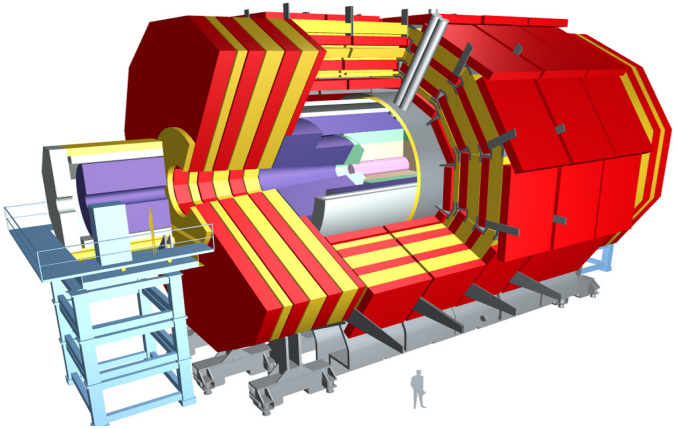
The Run-2 DAQ System of CMS

ALICE, ATLAS, CMS & LHCb Second Joint Workshop on
DAQ@LHC, Chateau de Bossey, Switzerland, 12th April, 2016

Srećko Morović, CERN/EP-CMD
On behalf of the CMS DAQ group

Introduction: Run-1 DAQ

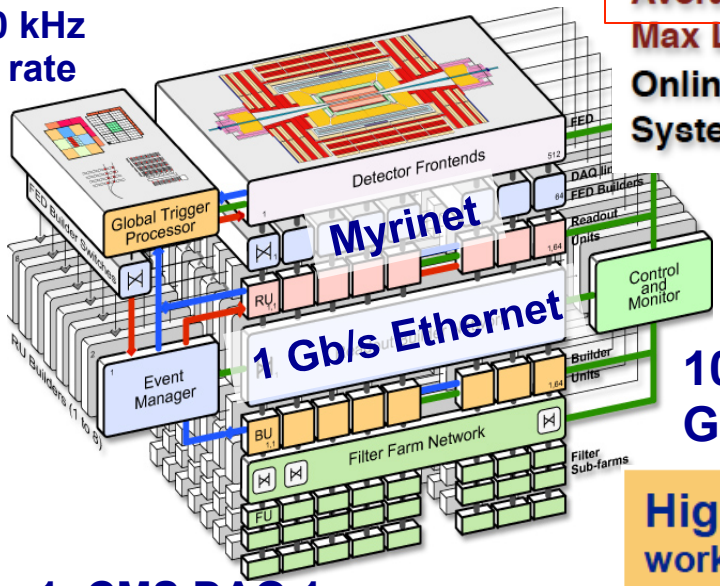


Detector	Channels	Control	Ev. Data
Pixel	60000000	1 GB	50 (kB)
Tracker	10000000	1 GB	650
Preshower	145000	10 MB	50
ECAL	85000	10 MB	100
HCAL	14000	100 kB	50
Muon DT	200000	10 MB	10
Muon RPC	200000	10 MB	5
Muon CSC	400000	10 MB	90
Trigger		1 GB	16

Only 2 trigger levels in CMS

Level-1 Trigger accepting 100 kHz
Custom electronics

100 kHz L1 rate



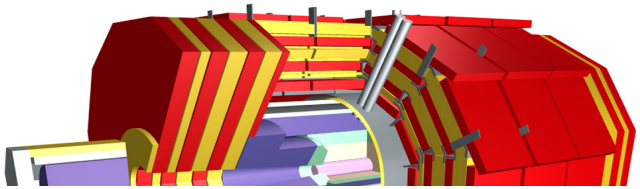
100 GB/s

Run-1: CMS DAQ 1

Average Event size	1 Mbyte
Max LV1 Trigger	100 kHz
Online rejection	99.999%
System dead time	~ %

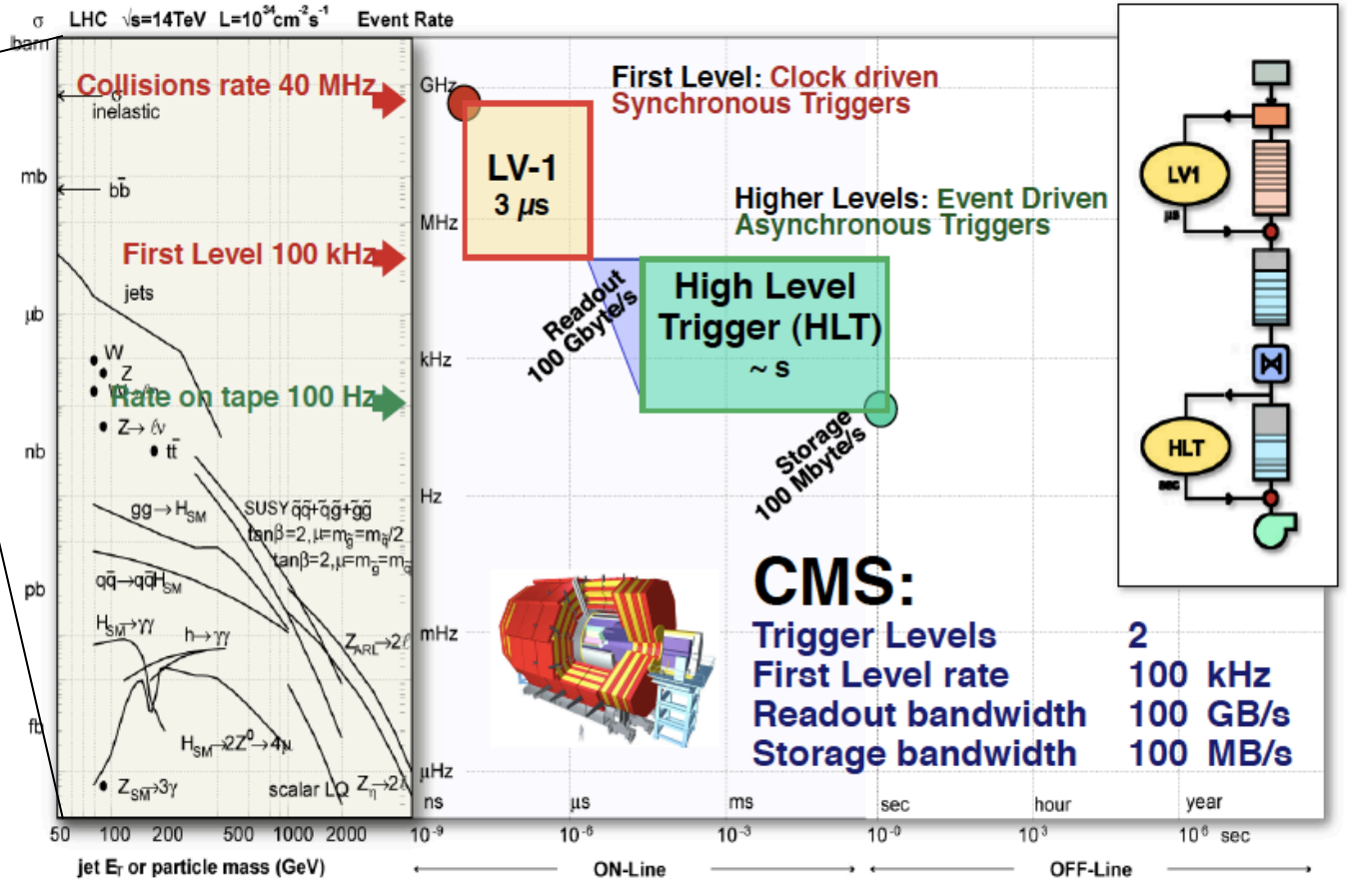
High-Level Trigger
working on full events
13000 cores
~500 Hz accept rate

Introduction:
Run-1 DAQ



Only 2 trigger levels in CMS

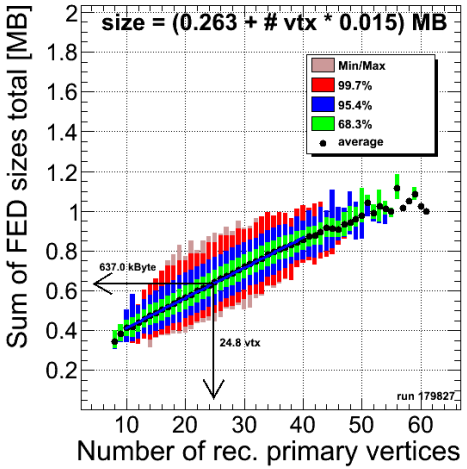
Level-1 Trigger accepting 100 kHz
Custom electronics



~500 Hz accept rate

CMS DAQ for LHC Run-2

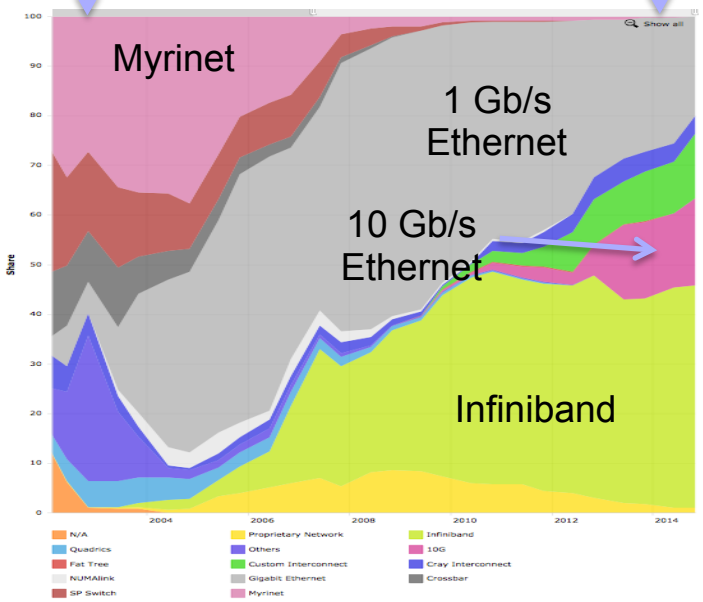
- Requirements in Run-2
  - 100 kHz level 1 trigger rate (unchanged)
  - 1 MB event size → 2 MB (large margin)
    - Increase in pileup
    - Detector upgrades
  - Accommodate legacy and new uTCA-based detector readouts
    - 1-2 kB (old) or 8 kB (new) fragments
- Other reasons to upgrade:
- Ageing hardware
  - 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



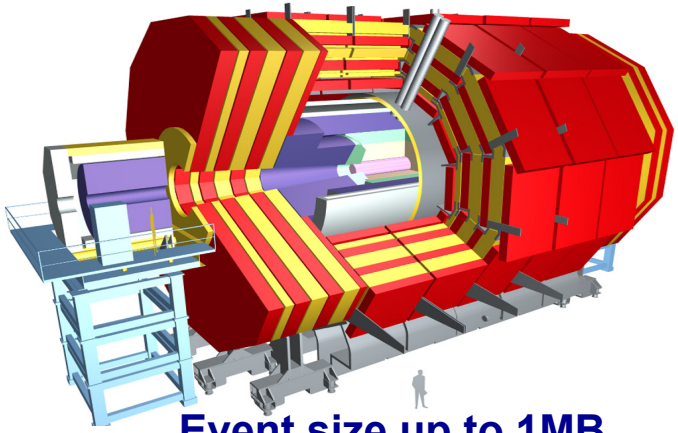
CMS event size

(Run-1 subsystems) linear increase with pile-up

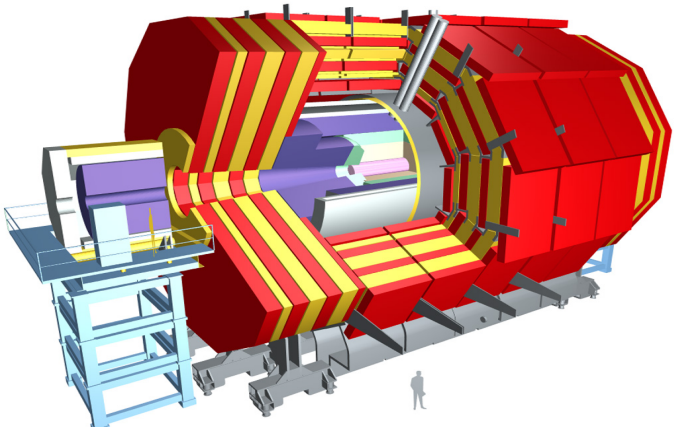
DAQ1 TDR (2002) 2014



Top500.org share by Interconnect family

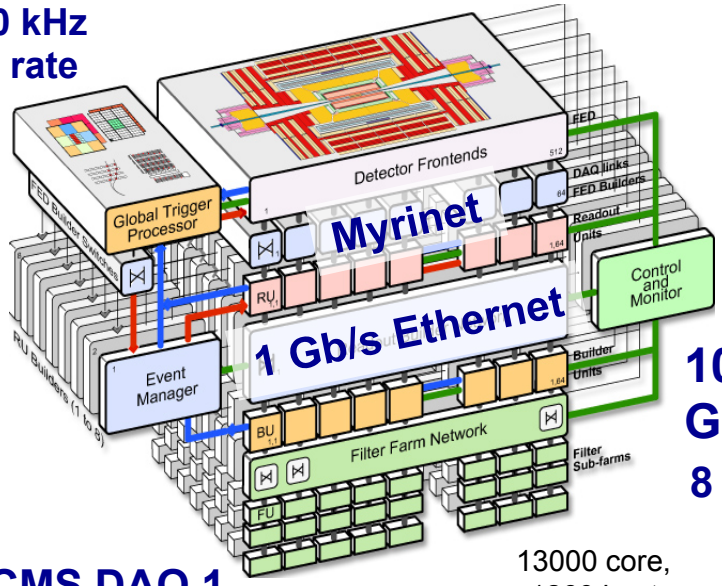


Event size up to 1MB



Event size up to 2MB

100 kHz
L1 rate



100 GB/s

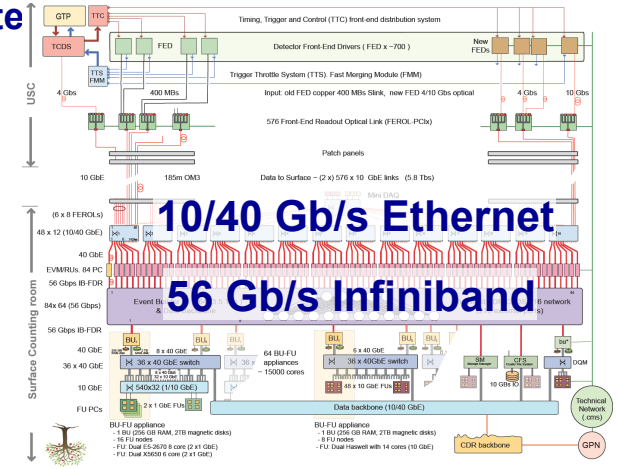
8 slices

CMS DAQ 1

13000 core,
1260 host
filter farm

max. 1.2 GB/s to storage

100 kHz
L1 rate



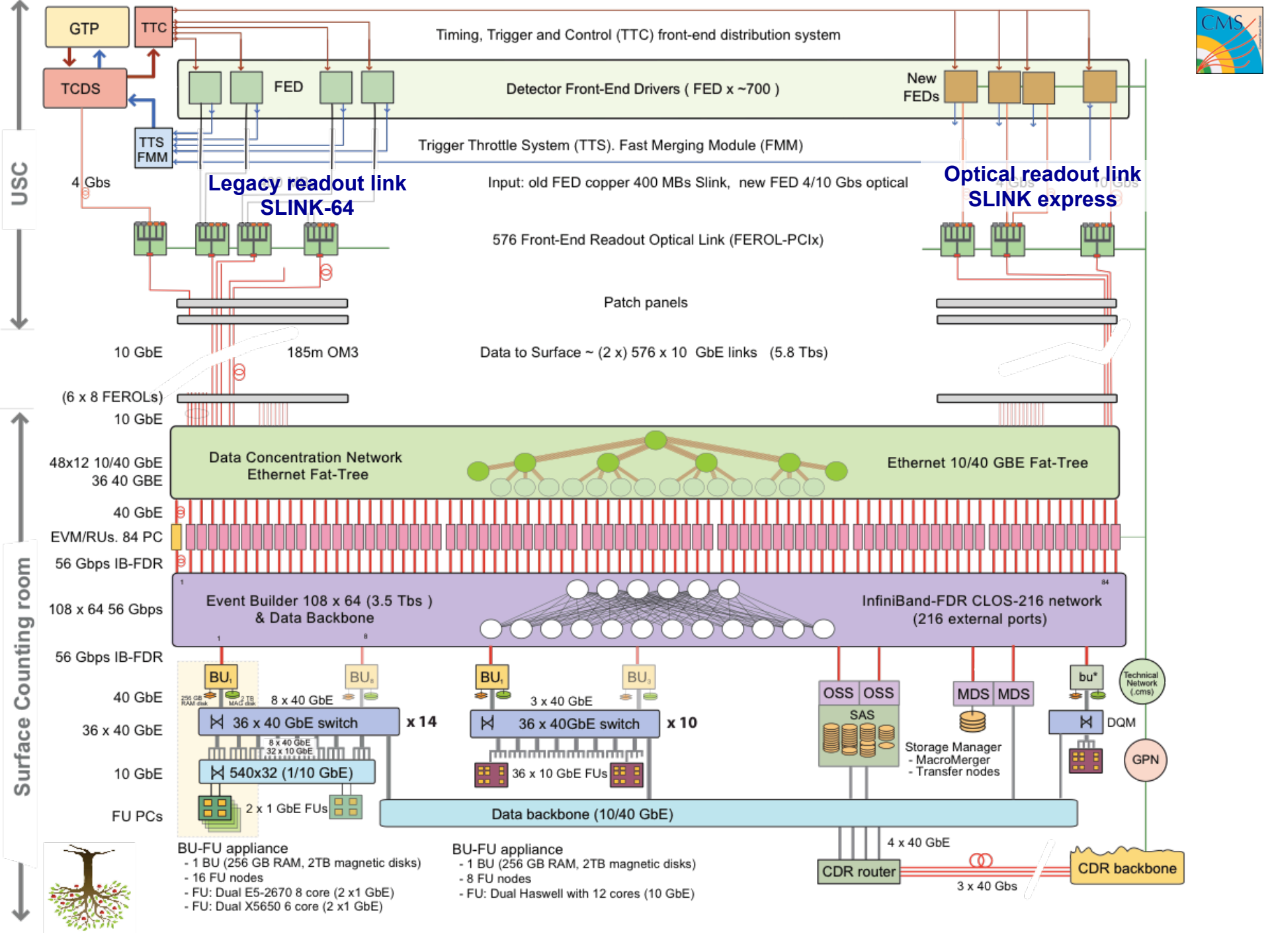
~200 GB/s

1 slice

CMS DAQ 2

16000+ core,
900 host
filter farm

~ 3-6 GB/s to storage


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
Legacy readout link
SLINK-64
Input: old FED copper 400 MBs Slink, new FED 4/10 Gbs optical
Optical readout link
SLINK express
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
8 x 40 GbE
32 x 10 GbE
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
MDS
Storage Manager
- MacroMerger
- Transfer nodes
bu*
DQM
Technical Network (.cms)
GPN
Data backbone (10/40 GbE)
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)
4 x 40 GbE
CDR router
3 x 40 Gbs
CDR backbone

New or upgraded detectors in CMS

- Several detectors / online-systems upgraded to cope with higher luminosity
- Increase of event size
- New readout electronics based on µTCA

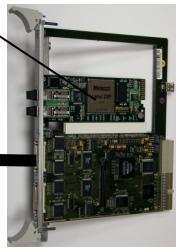
- 2014: New Trigger Control and Distribution System
- 2014: Stage-1 calorimeter trigger upgrade
- 2014/15: new HCAL readout electronics
- 2016: Full trigger upgrade
- 2017: New pixel detector and readout electronics

SLINK sender mezzanine
Plugged onto VME electronics

Fragment size 1..4 kB

SLINK copper cable 400 MB/s

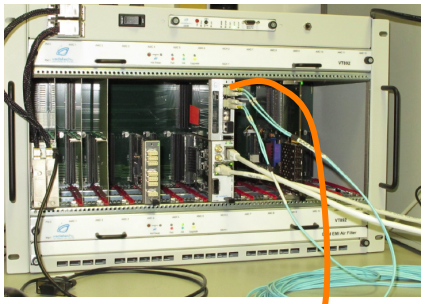
Myrinet NIC



Frontend-Readout Link

640 Legacy Links: SLINK-64

(600 after Pixel upgrade)

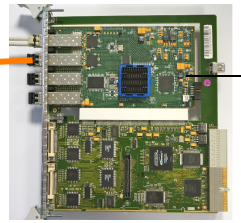


SLINK express sender =IP-core

µTCA electronics

Fragment size 2..8 kB

Optical SLINK-express
5 Gb/s - 10 Gb/s (2016)
- 4 kB packets
- retransmit

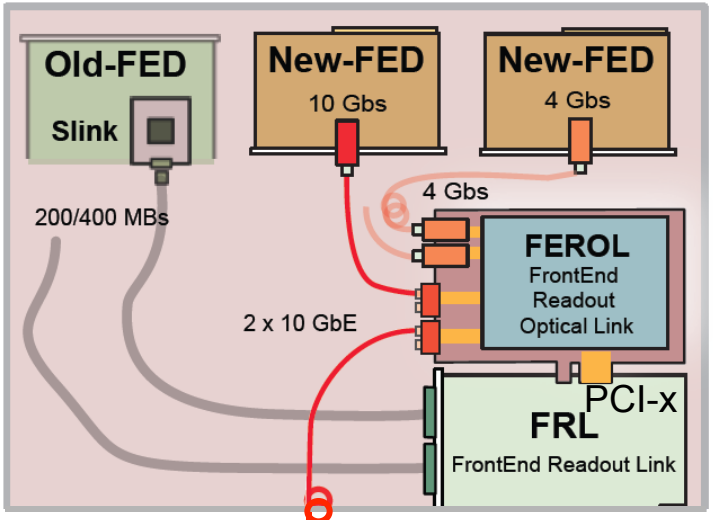


Frontend-Readout Optical Link

+ 50 new Links: SLINK-express

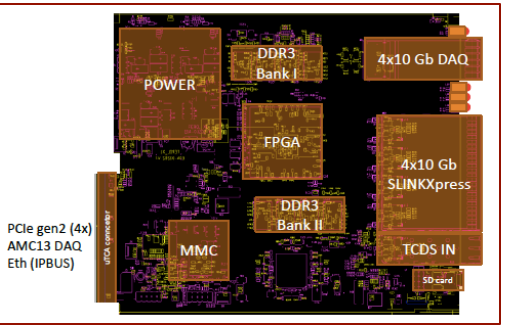
(170 after Pixel upgrade)

Frontend-Optical Link & Data Concentrator



10 Gb/s throughput in FEROL
- 2 x 5 Gb/s or 1x10 Gb/s new FEDs
- Legacy 200/400 MB/s FRL cards

early 2017:
FEROL-40
- 4x 10 Gb/s inputs
- 4x 10 Gb/s outputs

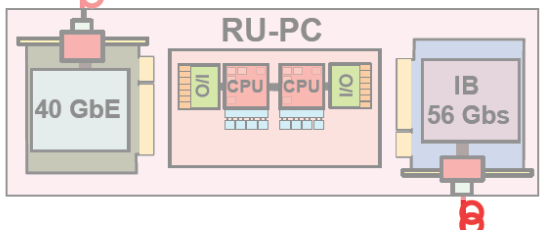


10 Gb/s simplified TCP/IP from an FPGA

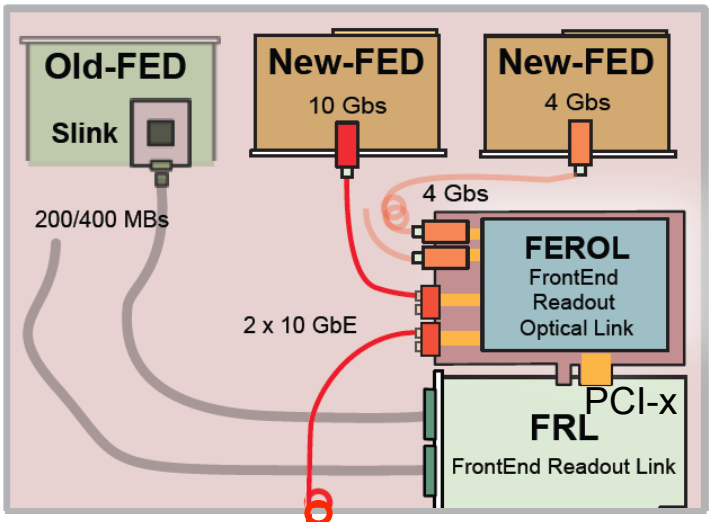
48 x 10 Gb/s in

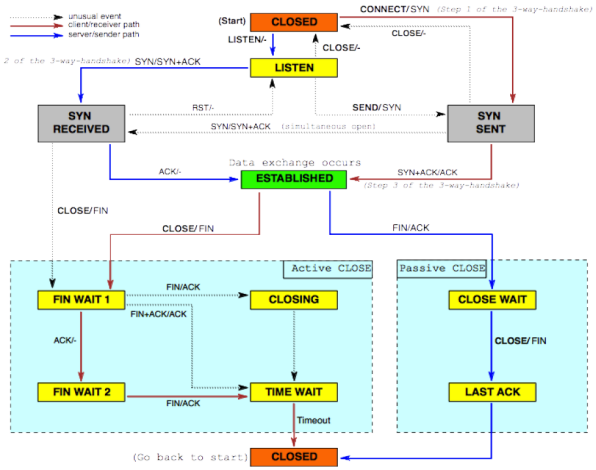
Data concentration:
10/40 Gb/s Mellanox
Ethernet switch

6 x 40 Gb/s out



Frontend-Optical Link & Data Concentrator





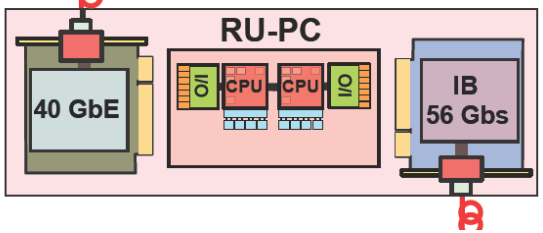
TCP/IP in principle difficult to implement in an FPGA …

10 Gb/s simplified TCP/IP from an FPGA

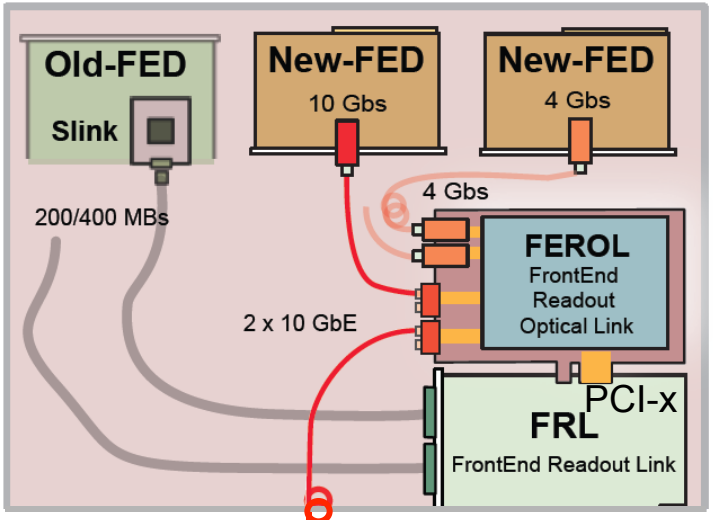
48 x 10 Gb/s in

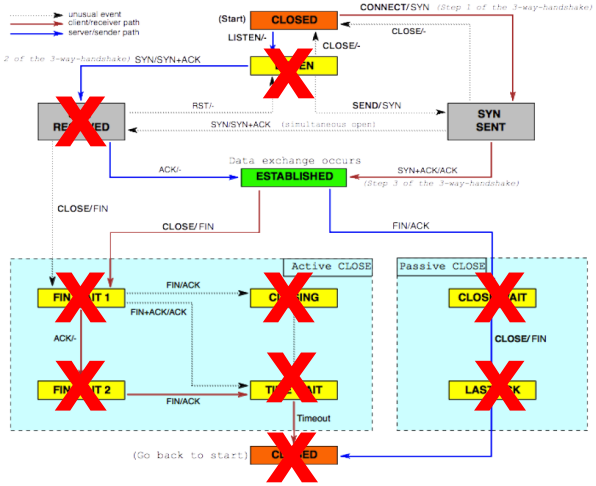
Data concentration: 10/40 Gb/s Mellanox Ethernet switch

6 x 40 Gb/s out



Frontend-Optical Link & Data Concentrator



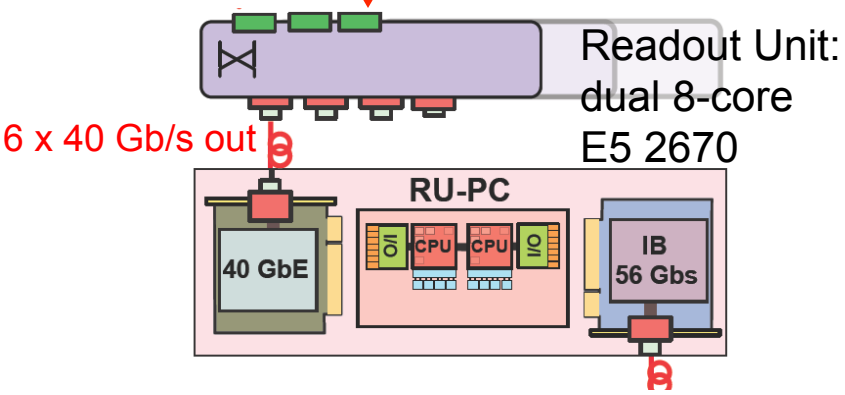


Simplified unidirectional TCP/IP only needs 3 states

10 Gb/s simplified TCP/IP from an FPGA

48 x 10 Gb/s in

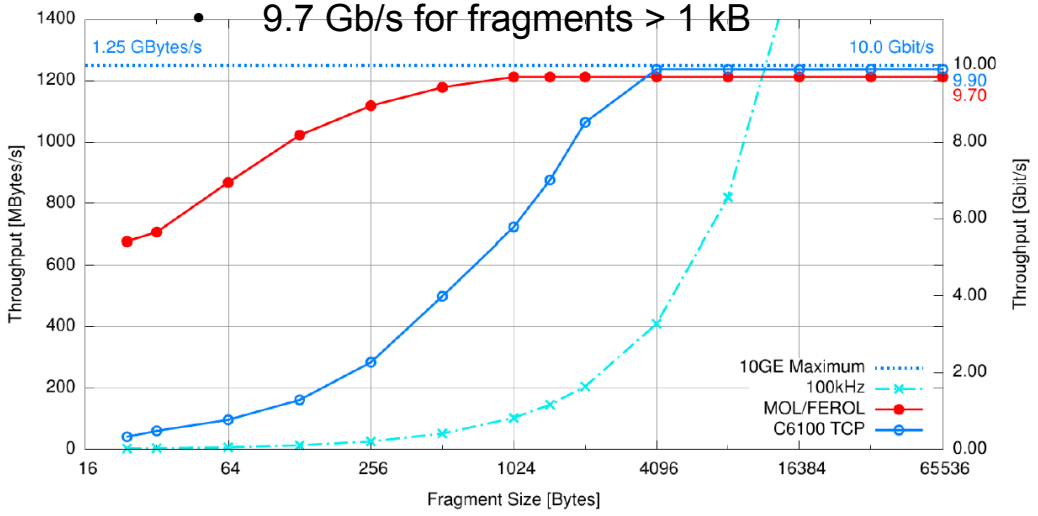
6 x 40 Gb/s out



Readout Unit: dual 8-core E5 2670

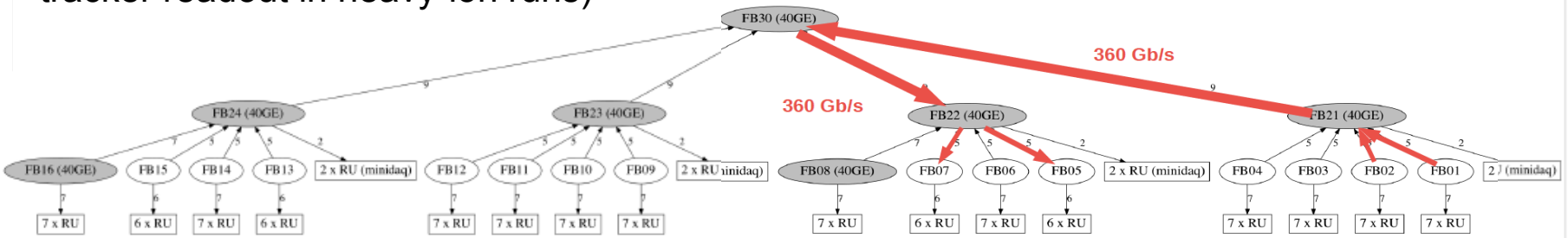
Point2point TCP FEROL FPGA → RU PC

- 9.7 Gb/s for fragments > 1 kB



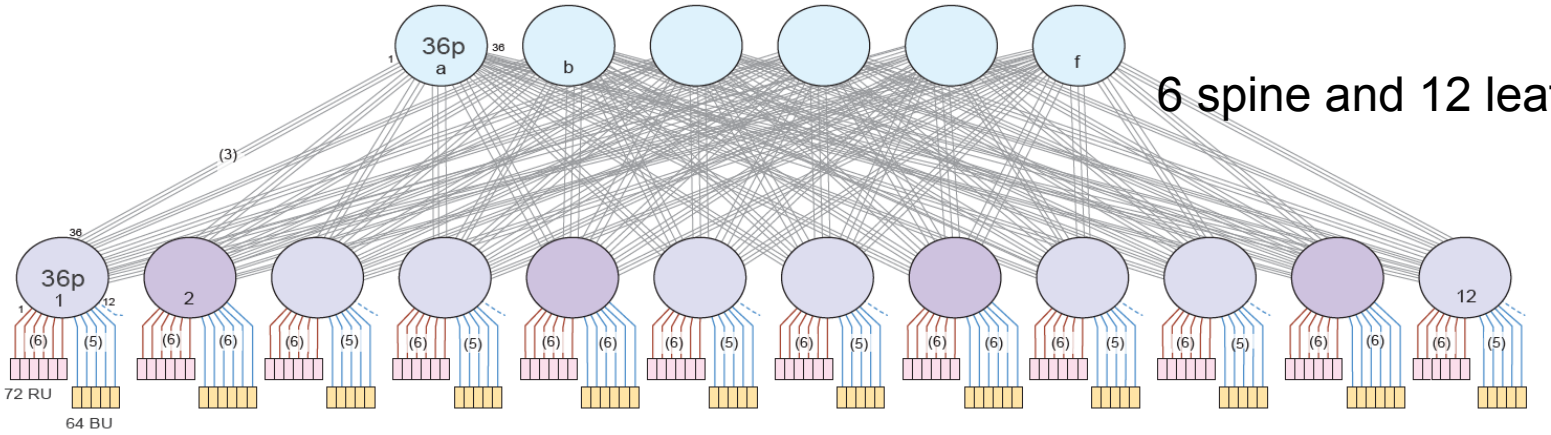
Data concentrator and event building networks

- Fat-Tree data concentrator setup:
- 16 x 10/40 GbE leaf switches + middle and top spine switches
- 108 Readout-Units (including spares)
- Allows FEROL – RU traffic routing throughout data concentrator network (needed for tracker readout in heavy-ion runs)



Infiniband

- 108 x 72 Event Builder – 56 Gb/s FDR Infiniband Clos network (108x108 I/Os)



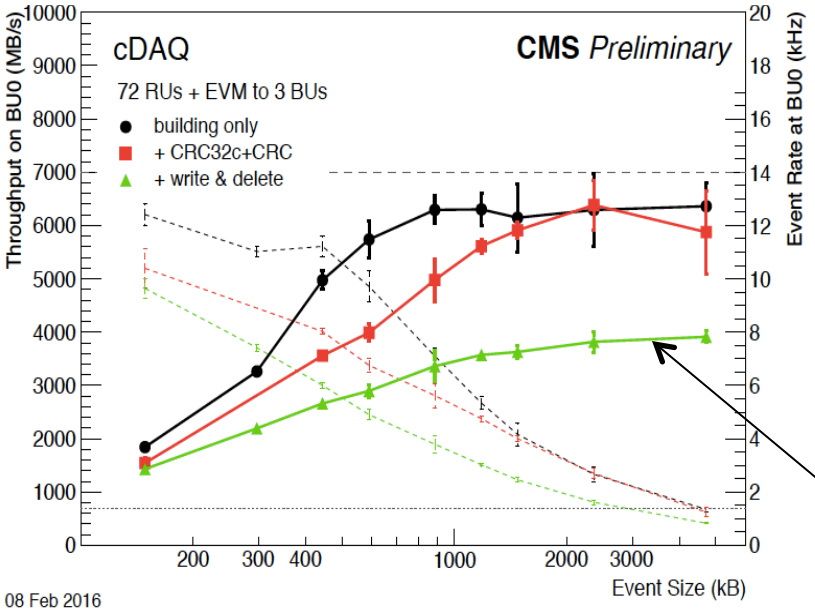
6 spine and 12 leaf switches

For details on Ethernet and Infiniband experience see P. Zejdl and A. Holzner talk

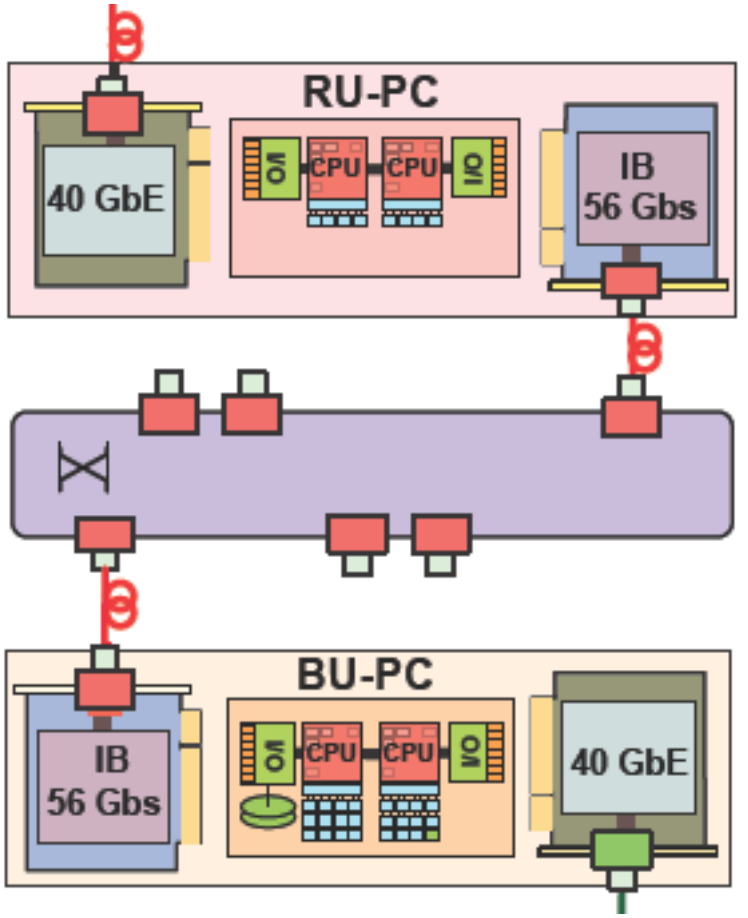
Event Builder

- Software based on Linux TCP stack (RU 40 GbE) and Infiniband Verbs
- On both ends :
  - Multiple threads for data reception/ writing
- Performance: sensitive to scaling effects of IB network, NUMA-specific software tuning

Builder Unit (BU)-side performance:



08 Feb 2016



256 GB RAM mounted as ramdisk
(EvB output buffer)

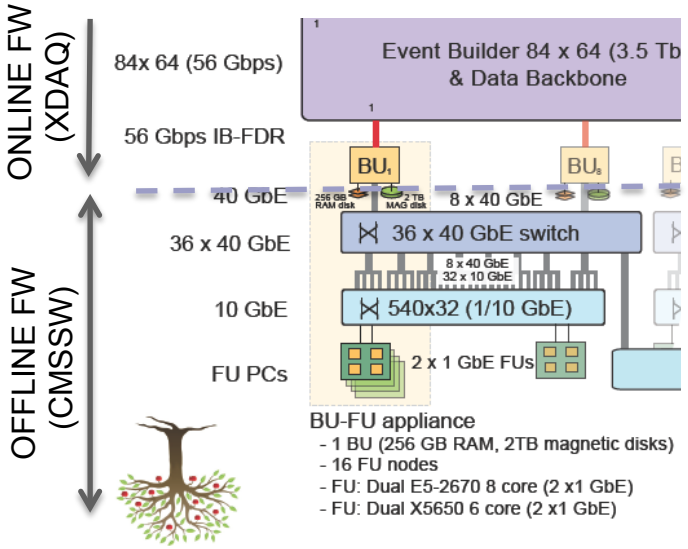
~3 GB/s building+writeout rate on BU (>1MB events)

For details on Event Building architecture and performance see R. Mommsen talk

Filter Farm and High Level Trigger

- **New approach in Run-2: File-based Filter Farm**
  - Input, output of event and non-event data, monitoring and logging through **files**
  - **Network filesystem** used as transport (and resource arbitration) protocol
- Reduced coupling between DAQ and HLT
  - Use of standard offline software (CMSSW), with DAQ-specific code as modules

  → no mixing with Online (XDAQ) framework as in Run-1
    - separate release cycles, simplified development, maintenance and debugging
  - Data-driven execution, decoupled from upstream DAQ state

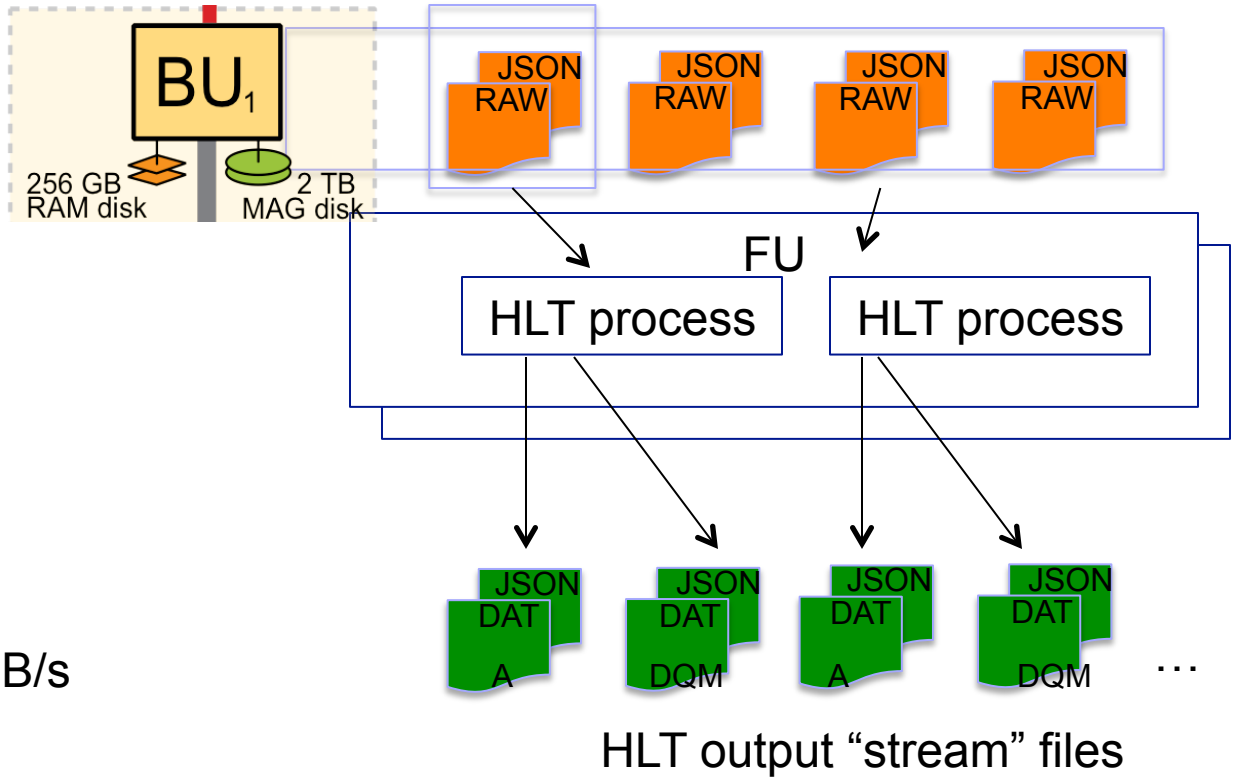


- Filter Farm consists of 62 – 72 **HLT appliances**
  - 12 – 18 Filter-Units (FU) mount the 256 GB RAM disk of each BU via network (**NFSv4**)
  - Connected through:
    - 40 GbE (BU) to 10 GbE network (new FU nodes) or 40/10 to 1 GbE network (legacy Run-1 FU nodes)

Filter Farm data flow

- Data read from RAM disk and processed/filtered by HLT jobs running in FUs

Every data file accompanied by a **metadata file in JSON format** (description of input or output file)



HLT output "stream" files

Approx ~30 streams (in collision HLT 'menu')
Physics, Calibration, DQM, Event Display, Monitoring …

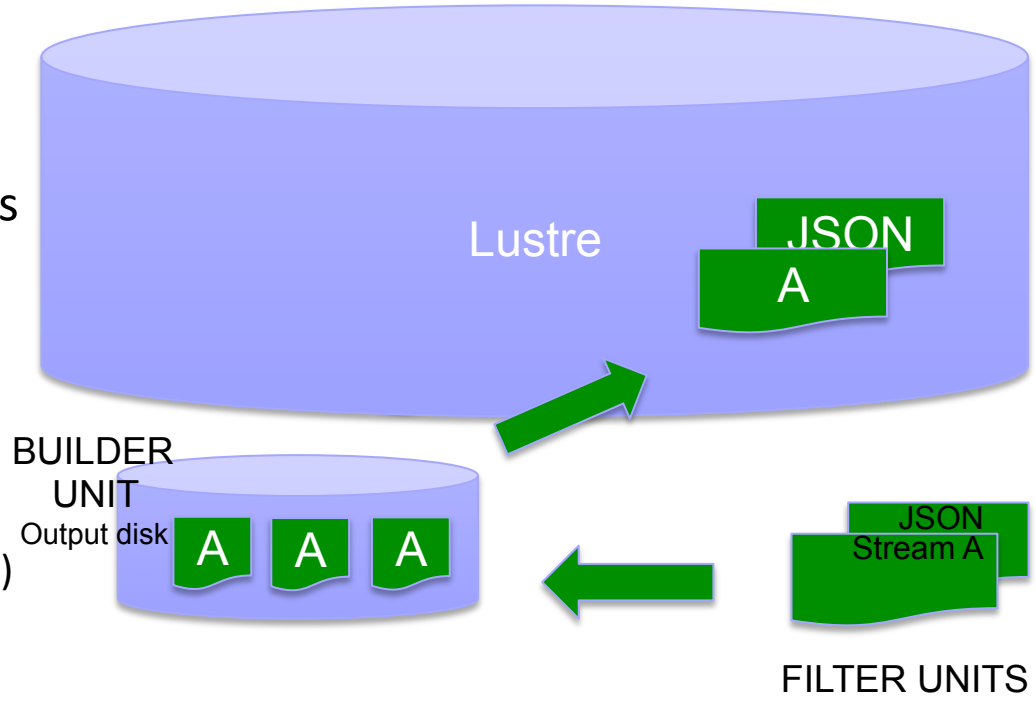
Max. input rate per BU:
2kHz, ~ 2.2 – 2.6 GB/s
Max. output rate to BU: 150 MB/s
(into 4x disk RAID0 array)

Merging and storage

- HLT output scattered on ~900 nodes
- 3 levels of data merging

aggregate data into a global filesystem (Lustre FS):

1. micro-merging
   - Per-stream HLT output (JSON/data) concatenated in FU and copied to BU output disk



2. mini-merging: services running on BUs
   - Merge step 1 data and metadata to a location in Lustre FS
   - data single per-stream output file in Lustre (done in parallel on BUs)
3. macro-merger - running on dedicated merger nodes
   - completes event count bookkeeping

HLT Farm Hardware in Run-2

HLT farm 2015:

Installed	product	vendor	cpu type	#cores/ box	network	#boxes	#cpus	kHS
2011 *	C6100 (Westmere)	Dell	X5650	12	1 GbE x2	288	3456	59
2012	C6220 (Sandy Bridge)	Dell	E5-2670	16	1 GbE x2	256	4096	90
2015	S2600xp (Haswell)	Megware	E5-2680v3	24	10 GbE	360	8640	194
total						904	16190	342

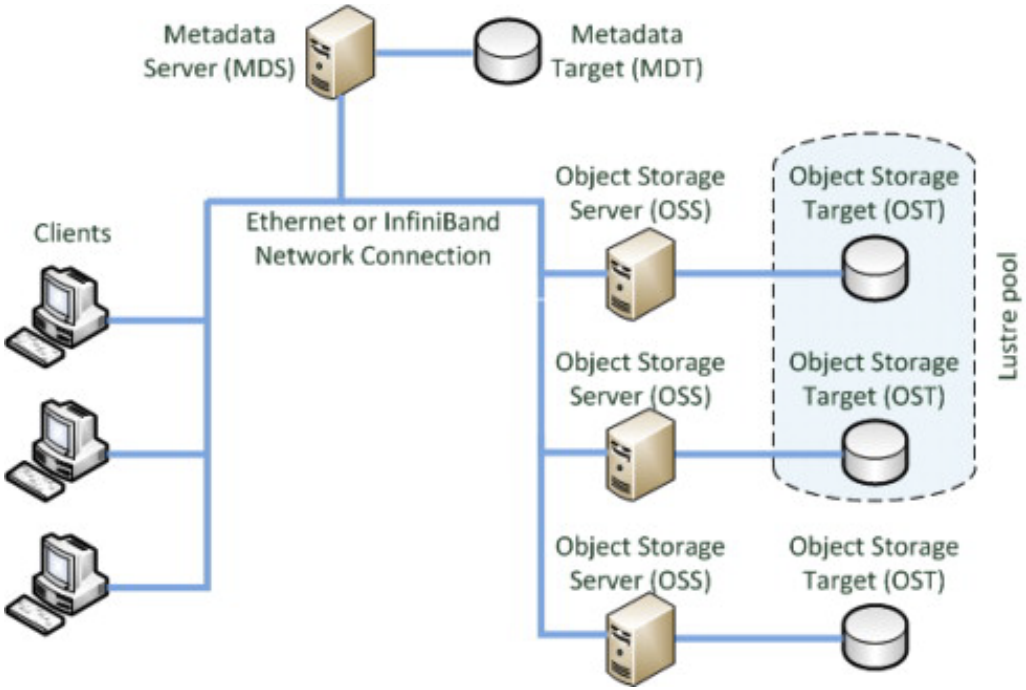
HLT farm 2016:

Installed	product	vendor	cpu type	#cores/ box	network	#boxes	#cpus	kHS
2012	C6220 (Sandy Bridge)	Dell	E5-2670	16	1 GbE x2	256	4096	90
2015	S2600xp (Haswell)	Megware	E5-2680v3	24	10 GbE	360	8640	194
total	S2600xp (Broadwell)	Action	E5-2680v4	28	10 GbE	324	9072	195
total						940	21808	479

HLT processing capacity 2016 / 2015 (kHS) = 1.4

* C6110 nodes will be kept in 2016 for running Cloud

Storage: Lustre FS



Hardware:

- 6 Dell R720 servers: 2 x MDT nodes, 4 x OSS nodes (with OST + disk shelves)
- 350 TB usable storage space in RAID6 volumes

- Mounted on each Builder Unit (Infiniband), merger nodes and transfer system nodes (40Gb/s Ethernet)

Production performance in 2015:

- 4 GB/s write (merging) + 2 GB/s read (Heavy-Ion runs)

For more details on Lustre deployment see L. Darlea talk

Transfer system

Services with access to merger output on Lustre

- Merged data copied to Tier0
  or to consumers (e.g DQM, Event display)

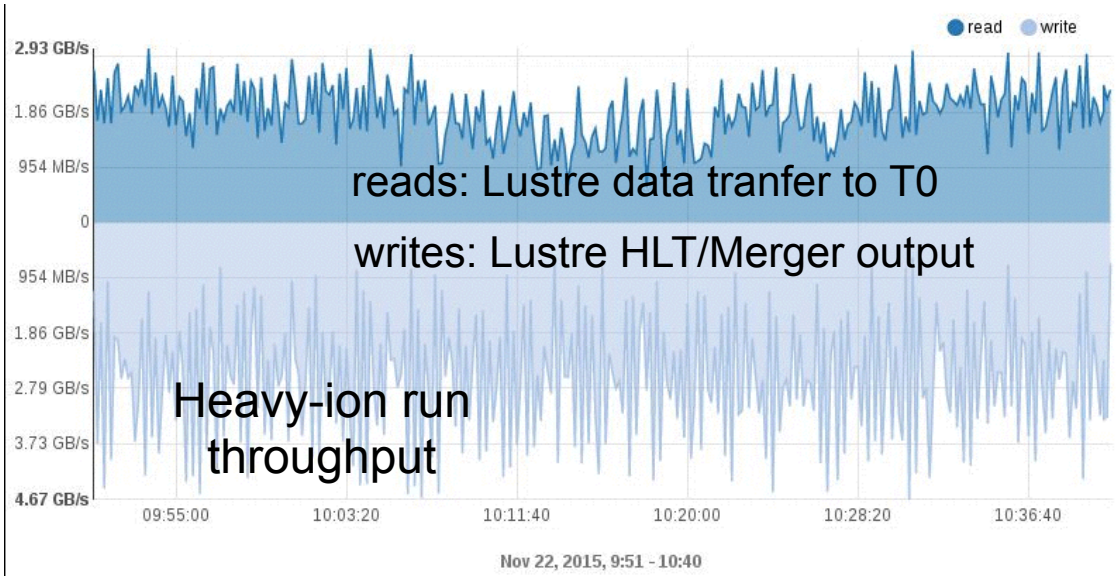
Tier0 transfers:

- 4x40 GB/s CMS CDR uplink to IT
- Xrdcp copy to EOS
  - Multi-threaded copying from up to 3 nodes
- Injection of run event accounting/bookkeeping into DB for Tier0

Parts of transfer system inherited from Run-1 →

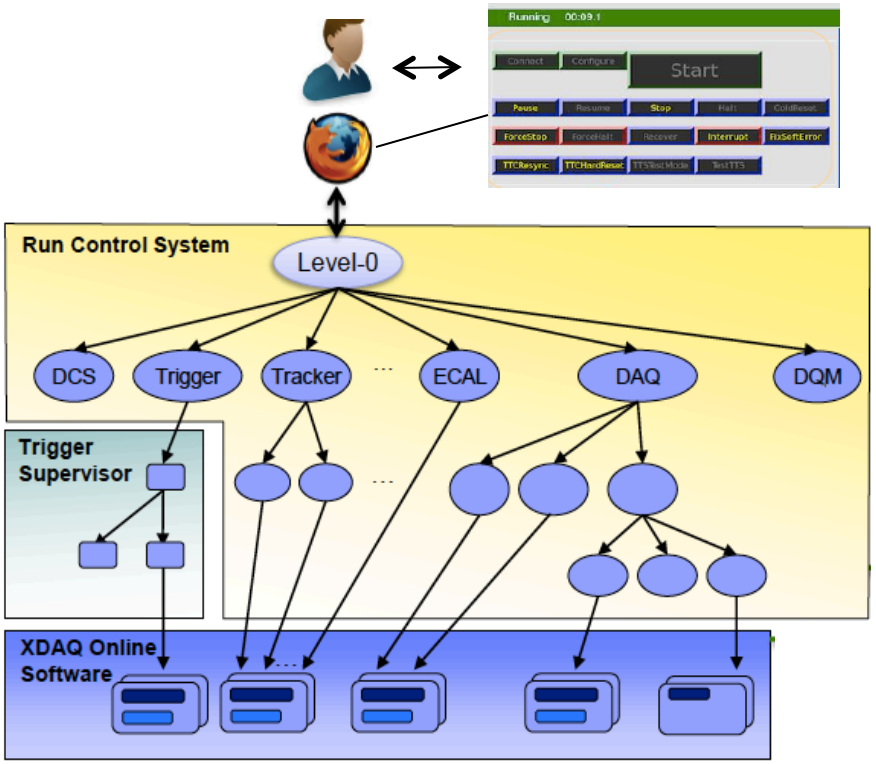
Transfer monitoring page:

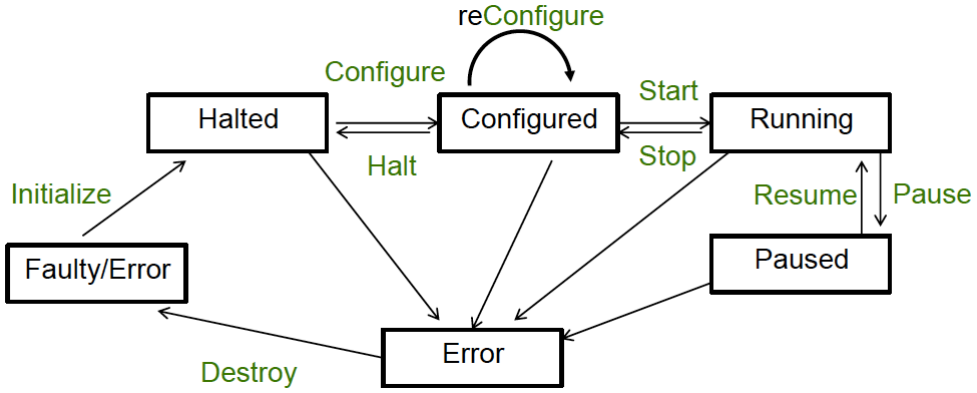
Size [GB]	Files	Evnts	Disk MB/s	Tier0 MB/s	Open	Close	Inject	Transfr	Check	Repack	Delet
58.79	1634	6816991	2351.28	369.06	0	1634	1624	190	190	0	
.4	15	430	2.03	.52	0	15	15	2	2	2	
12.55	10170	42307039	2292.09	328.28	0	10170	10170	1322	1322	0	
48.45	1266	10650462	10.11	.53	0	1266	1266	211	211	211	20
99.08	1387	1873774	214.99	28.52	0	1387	1387	263	263	252	25
47.07	1502	9539146	8.09	.52	0	1502	1502	254	254	254	25
00.87	619	11372329	169.48	1.81	0	619	619	103	103	103	10
44.62	14101	56427100	2263.48	246.31	0	14101	14101	1802	1802	1691	171
.24	16	33212	1.9	.15	0	16	16	3	3	3	
3.19	72	757662	12.8	.48	0	72	72	12	12	12	1
73.91	132	6596042	359.77	79.73	0	132	132	22	22	22	2
6.43	393	386717	4.26	.52	0	393	393	67	67	67	6
11.89	426	7882646	133.11	2.26	0	426	426	71	71	71	7
89.19	14225	54802741	2206.21	300.1	0	14225	14225	1832	1832	1820	183
.01	6	2280	.85	.01	0	6	6	1	1	1	
02.68	4008	17584788	6.48	.51	0	4008	4008	692	692	692	69
67.77	564	9980374	126.53	2.13	0	564	564	94	94	94	9
.44	1107	408751	.01	.01	0	1107	1107	1107	1107	1107	110
0	14	2916	.01	.01	0	14	14	14	14	14	1
0	10	2301	.01	.01	0	10	10	10	10	10	1
0	16	3742	.01	.01	0	16	16	16	16	16	1
.03	125	30423	.01	.01	0	125	125	125	125	125	12
0	13	3069	.01	.01	0	13	13	13	13	13	1
0	6	1248	.02	.01	0	6	6	6	6	6	
0	6	1340	.01	.01	0	6	6	6	6	6	
04.77	9632	36831588	1461.86	106.39	0	9632	9632	3935	3934	1206	120
.02	12	13	.13	0	0	12	12	12	12	12	1
.05	55	44022	.04	.04	0	55	55	55	55	55	5
60.17	32	72922	684.64	.51	0	32	32	32	32	30	2



Control structure

- DAQ (and other CMS subsystems) instantiation, configuration, states and transitions defined/controlled through synchronous state machine hierararchy - Run Control and Monitoring System (RCMS) system





DAQ data readout, concentrator and event builder software controlled by RCMS structure

- RCMS retrieves system configuration from DB – contains definition of hardware, applications & parameters, interconnects, etc.
- Blacklist DB – special database defining which hosts should not be used (mostly used to exclude broken hardware)

Control structure: Filter Farm, mergers, transfers

- HLT, merging and transfer system are permanently running system services
  - Not integrated into run control structure
- Services await external notification:
  e.g. new run in F^3 ← creation of new run directory by Event Builder in ramdisk

- All relevant information propagated through RAM disk from BU, such as:
  - HLT menu (configuration), which FU host should participate in the run (blacklist), transfer destination

- Feedback returned to EvB through ramdisk
  - State of resources, latency (EvB can apply backpressure or raise alerts in monitoring)

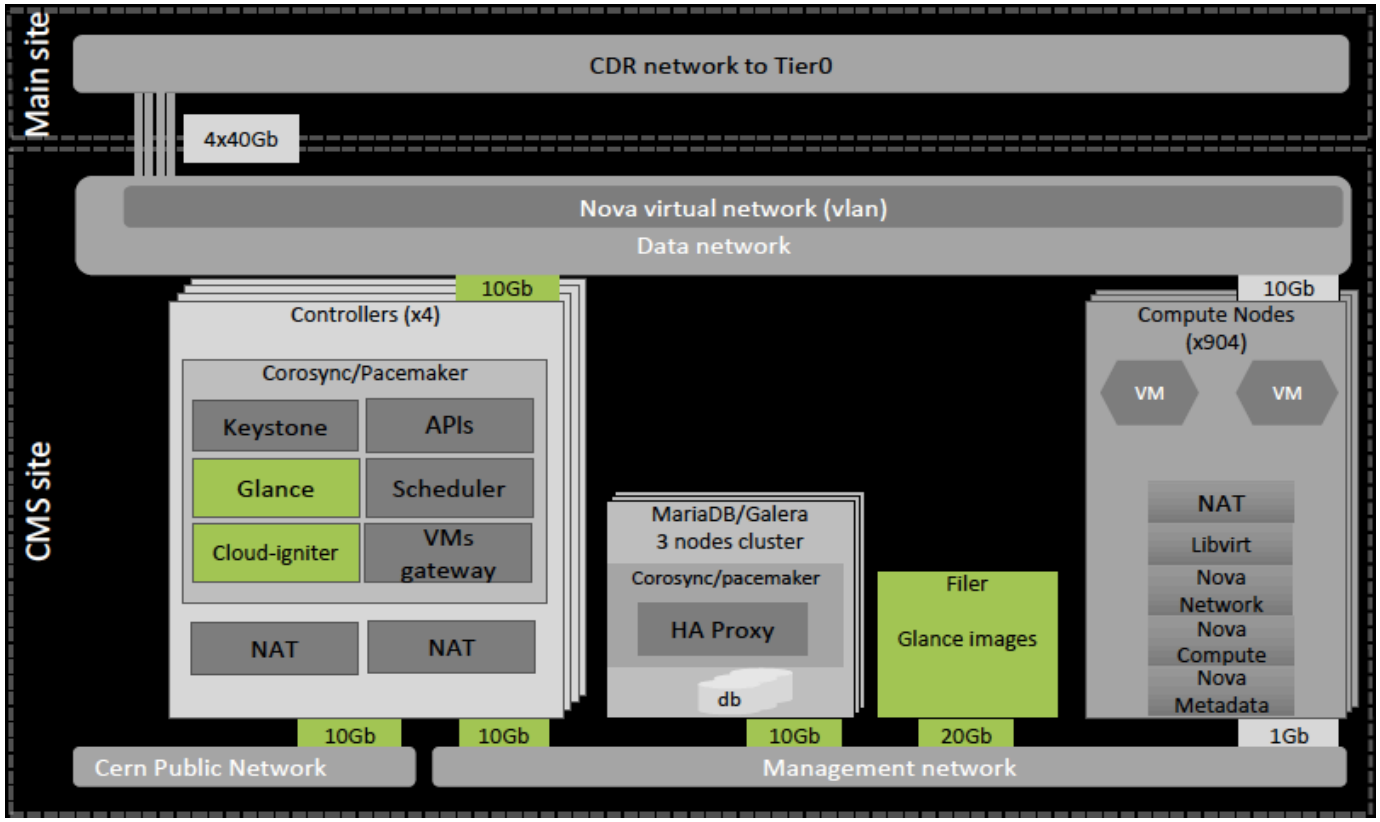
- Asynchronous execution
  - run in Filter Farm lasts until data is processed
  - Following runs are queued and executed as resource are releases

Filter Farm Cloud architecture

- HLT farm is under-utilized in technical stops and LHC interfills periods

Farm size in 2015 (HEP-SPEC06)	HLT	Tier0	All CMS Tier1 sites
CMS	350K (~500k in 2016)	300K	~300K (>500k in 2016)

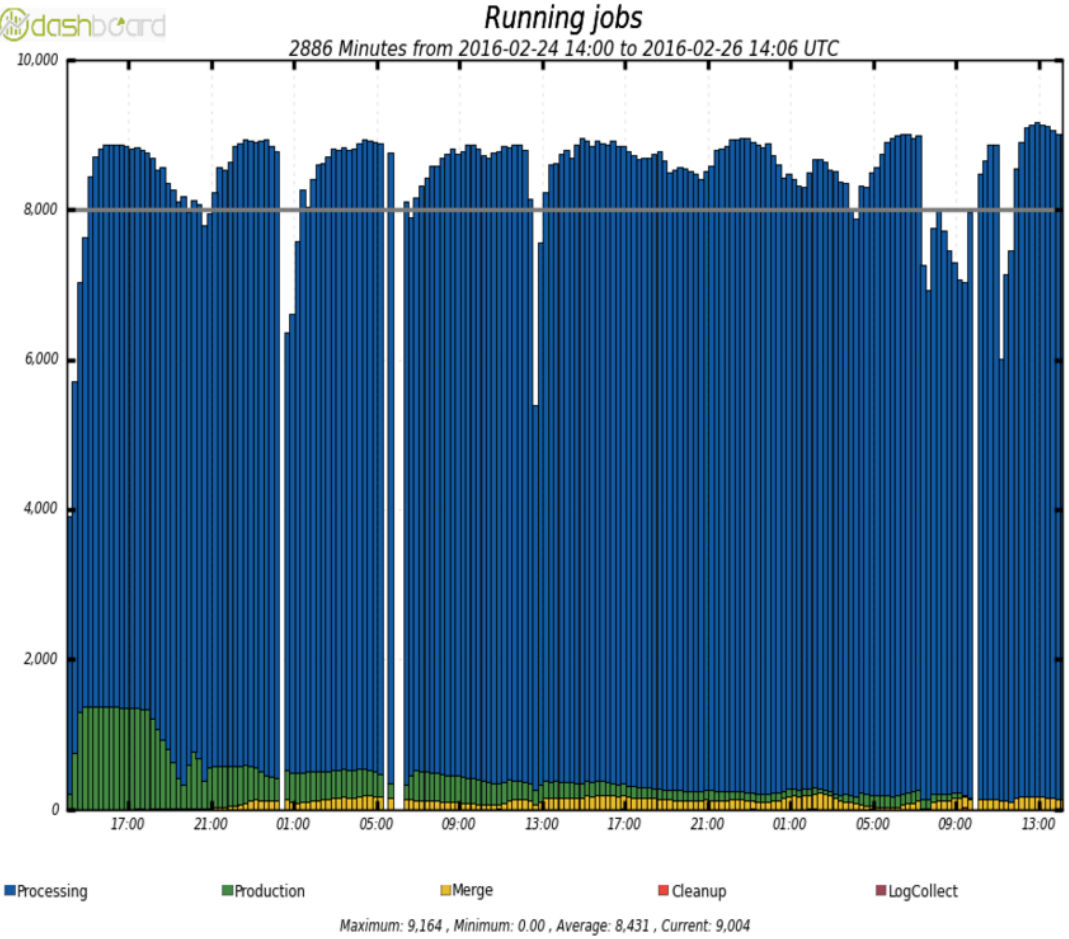
- Added ability to run WLCG grid jobs in FUs (switch off HLT) - T2_CH_CERN_HLT
  - jobs encapsulated in VMs on FUs (Cloud)
  - OpenStack based infrastructure



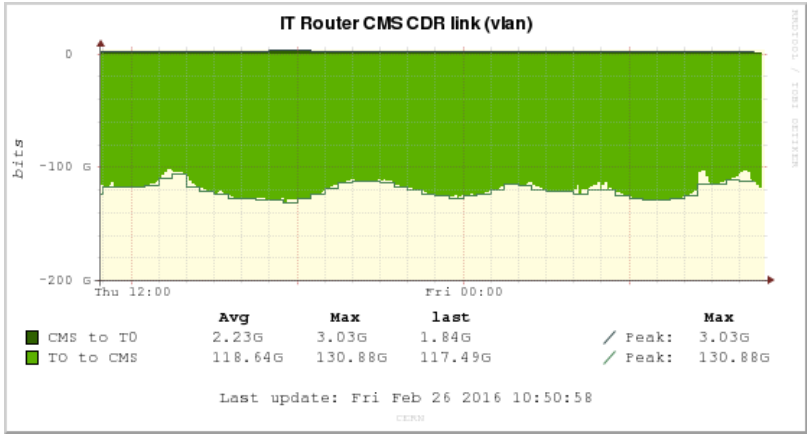
- 40GbE data network (using F^3 network)
- VMs interfaced to VLANs for connections outside of CMS technical network

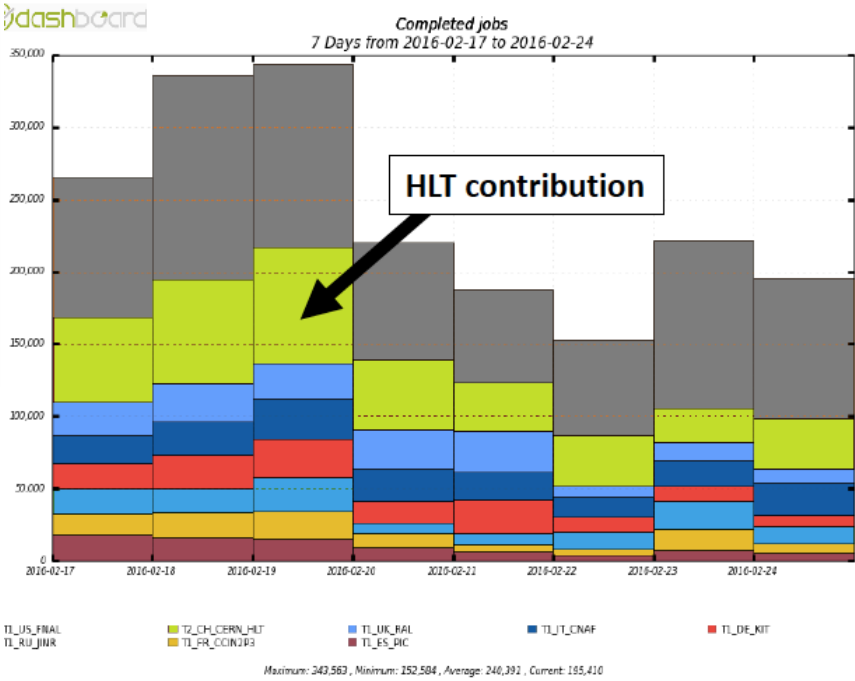
Jobs running in the cloud

with ~70% of Filter Farm:



Network intensive jobs can run in the Farm





Opportunistically running Cloud during interfills

- Challenges:
  - Unpredictable and short periods of inactivity (3 to 6 hours)

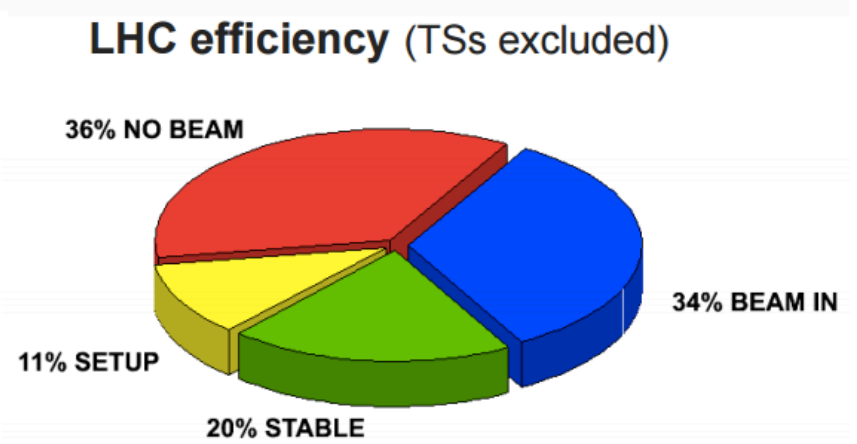
  Jobs could be terminated on a short notice
  → Duration suitable for only a subset of CMS jobs

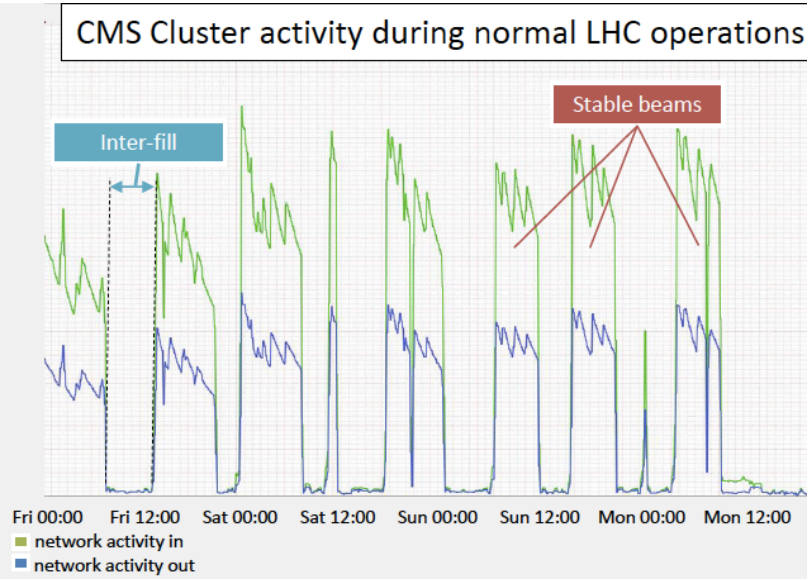
  - Fast switchover and release of resources between Cloud and HLT implemented

  - VM image distribution – large traffic volume (initially lasted hours)

  → Deployment to FUs optimized to 8 minutes using Squid proxies, compression and transfer over 10 GbE connections

- Fully automatized cloud start / stop for 2016:
  → based on parsing LHC states (started at beam dump, stopped at next flat top)





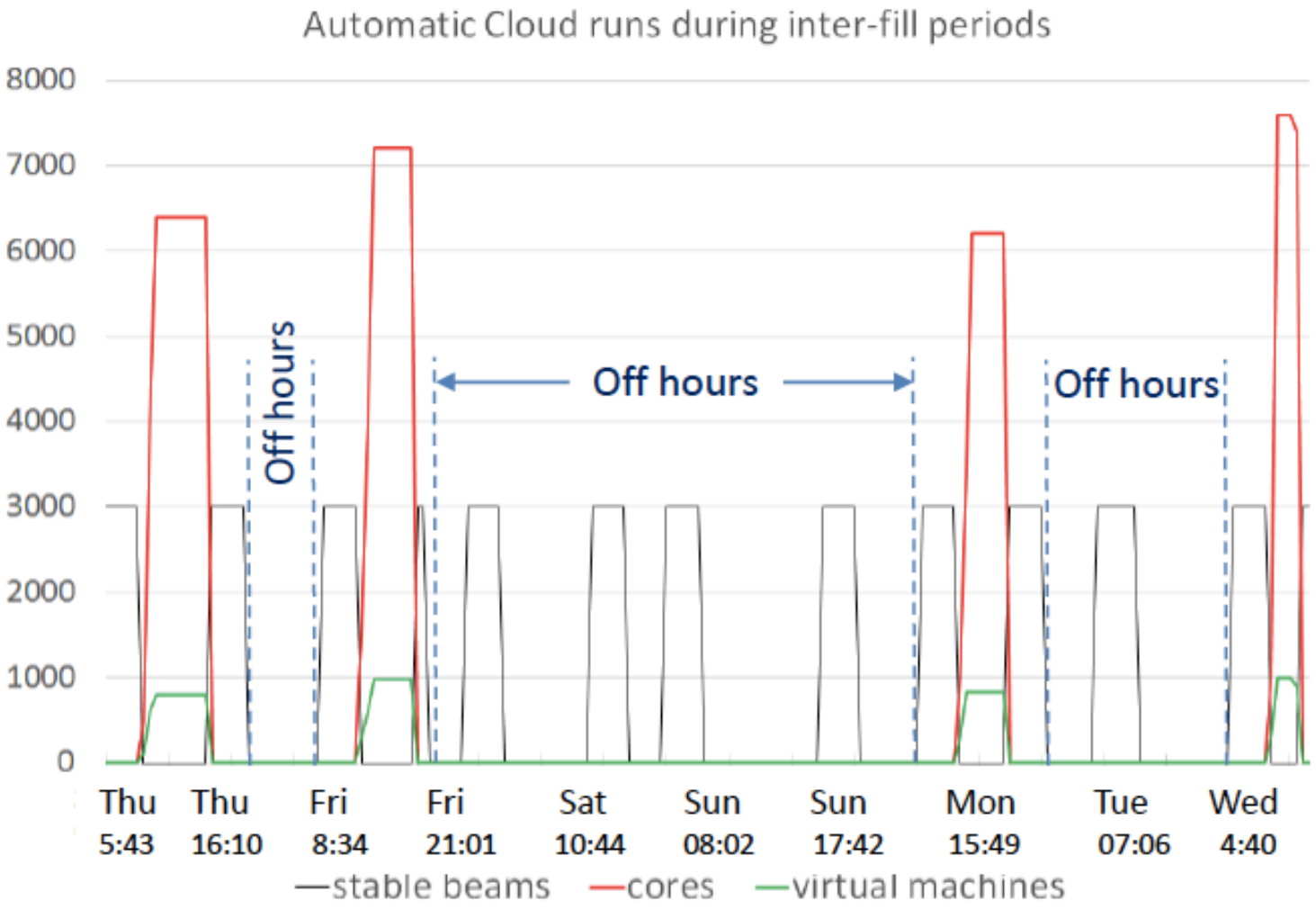
Summary

- A new DAQ system for Run-2 has been commissioned and used in production
  - 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)
  - Cloud capability integrated in the Filter Farm
  - Lustre File System for storage
  - Throughput doubled to 200 GB/s
- Performance required for Run-2 within DAQ2 capacity
  - Sufficient for proton-proton and Heavy-Ion runs in 2015
  - Higher readout capacity for detector upgrades planned (FEROL40)

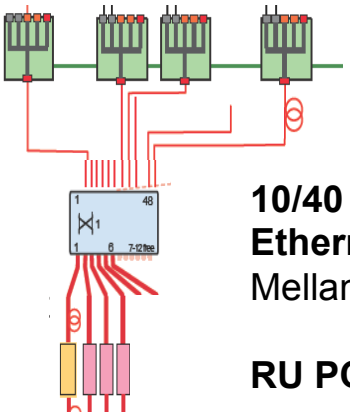
Backup

Cloud proof-of-concept running in interfills

- Tested in 2015



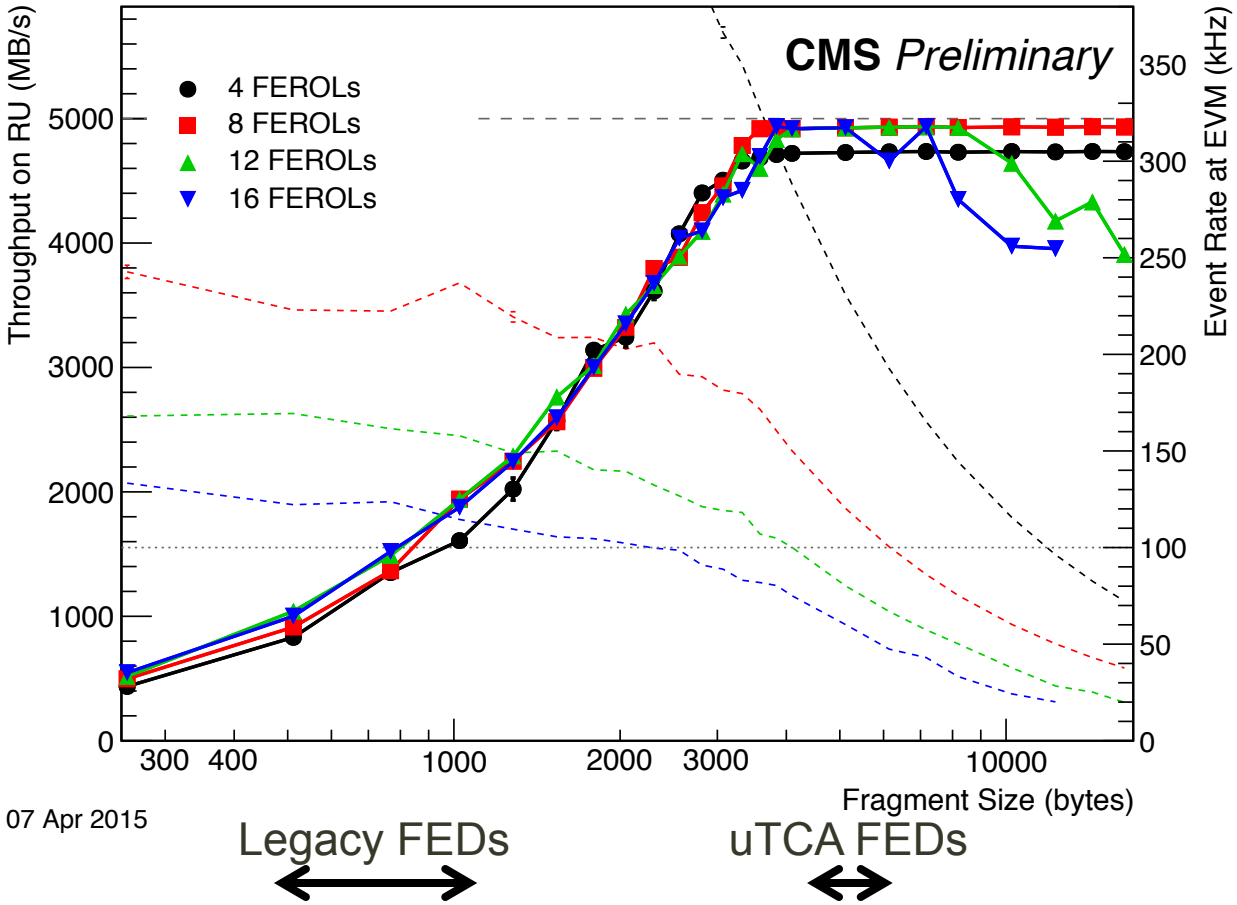
Data concentrator scaling



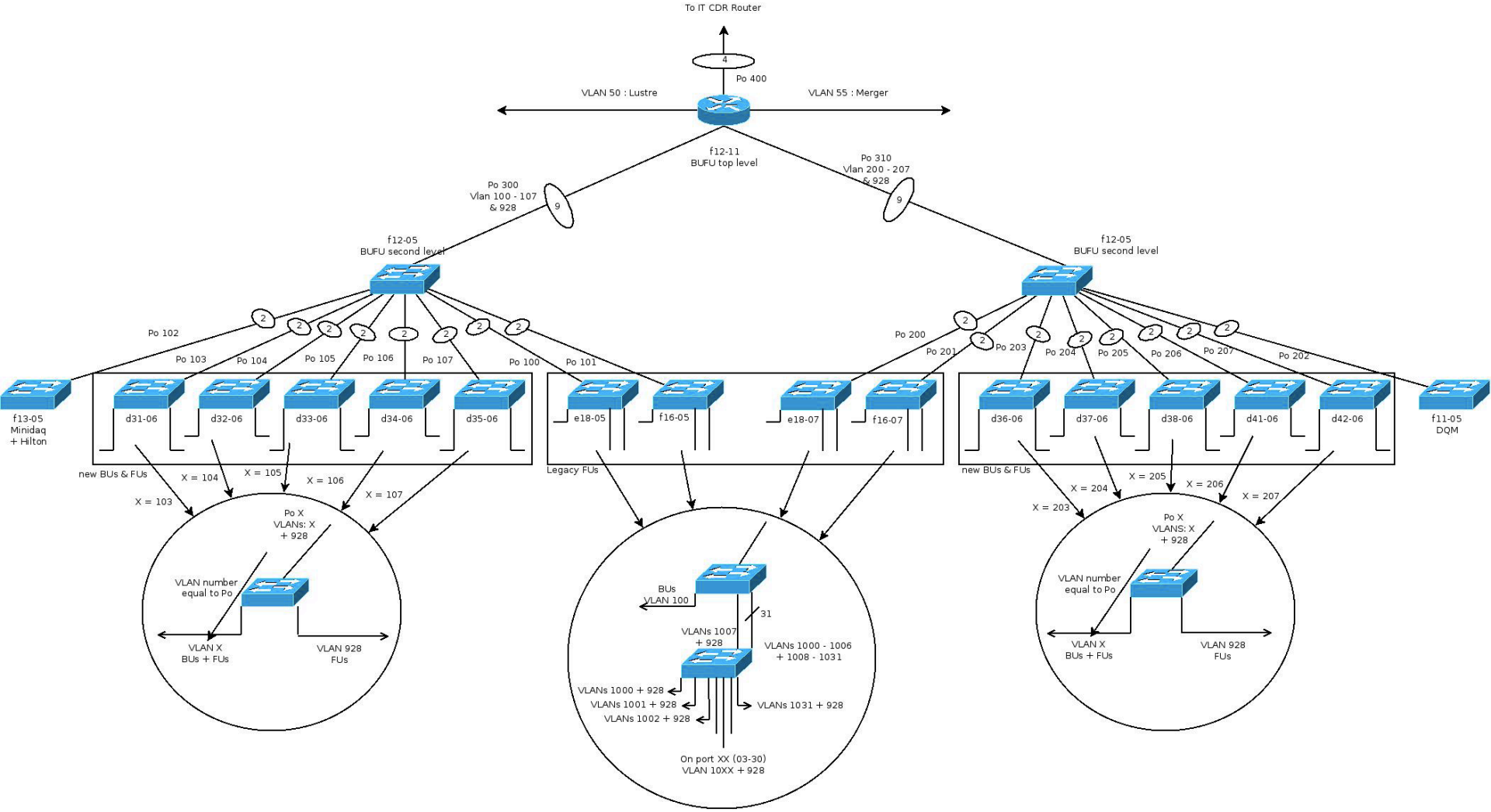
FEROLs 1..16

10/40 Gb/s
Ethernet switch
Mellanox MSX1024

RU PC



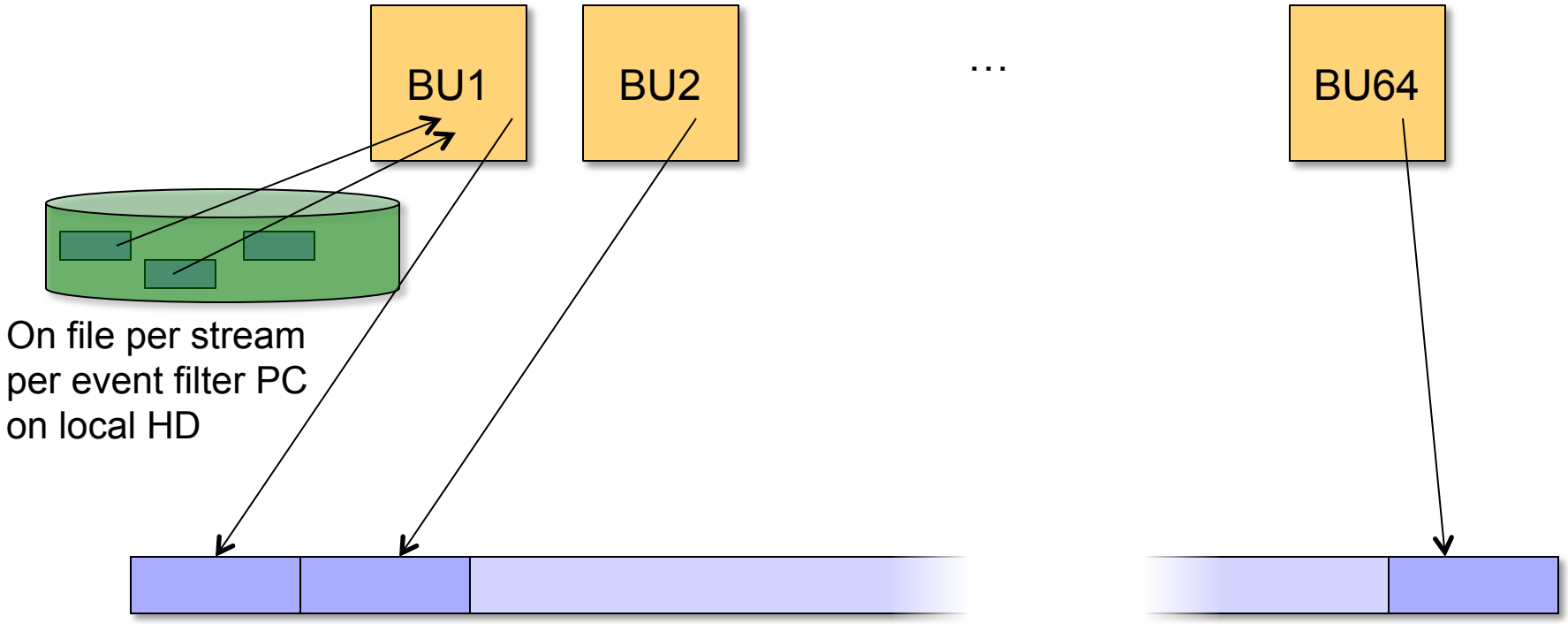
Filter Farm / merger data network



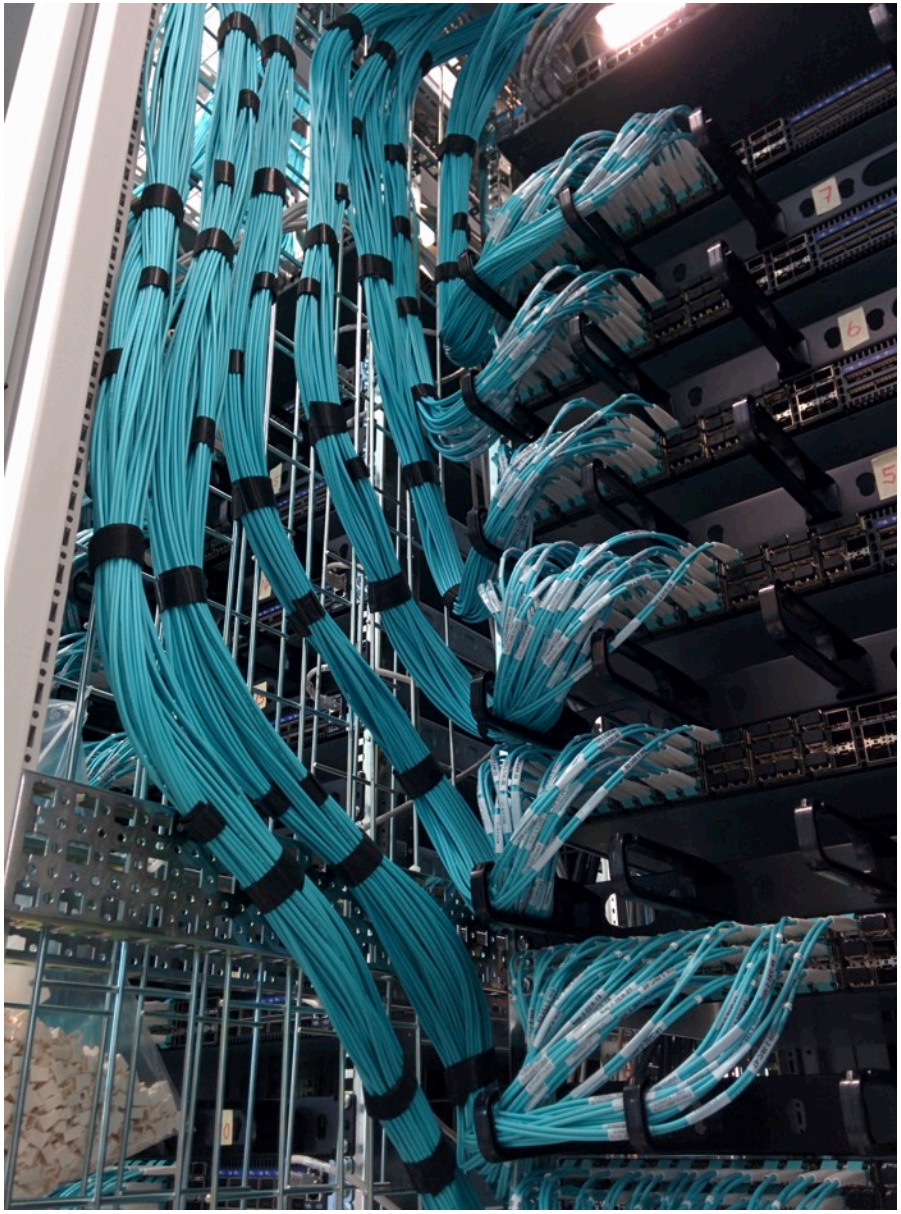
Merging algorithm

For large streams – single-copy algorithm

- BU merger service appends output to a single global file in Lustre (chunks written to in parallel from BUs)

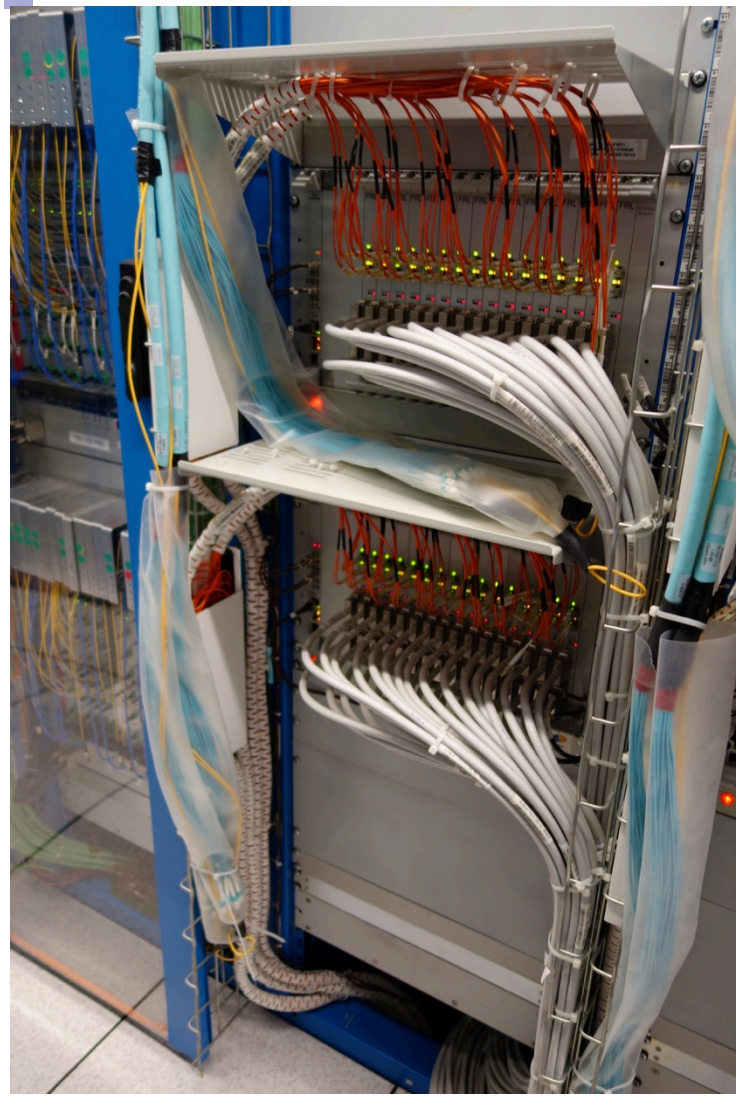


Single output file in the cluster file system (per output stream)



Data concentrator patch panels **and switches**

Infiniband Clos network

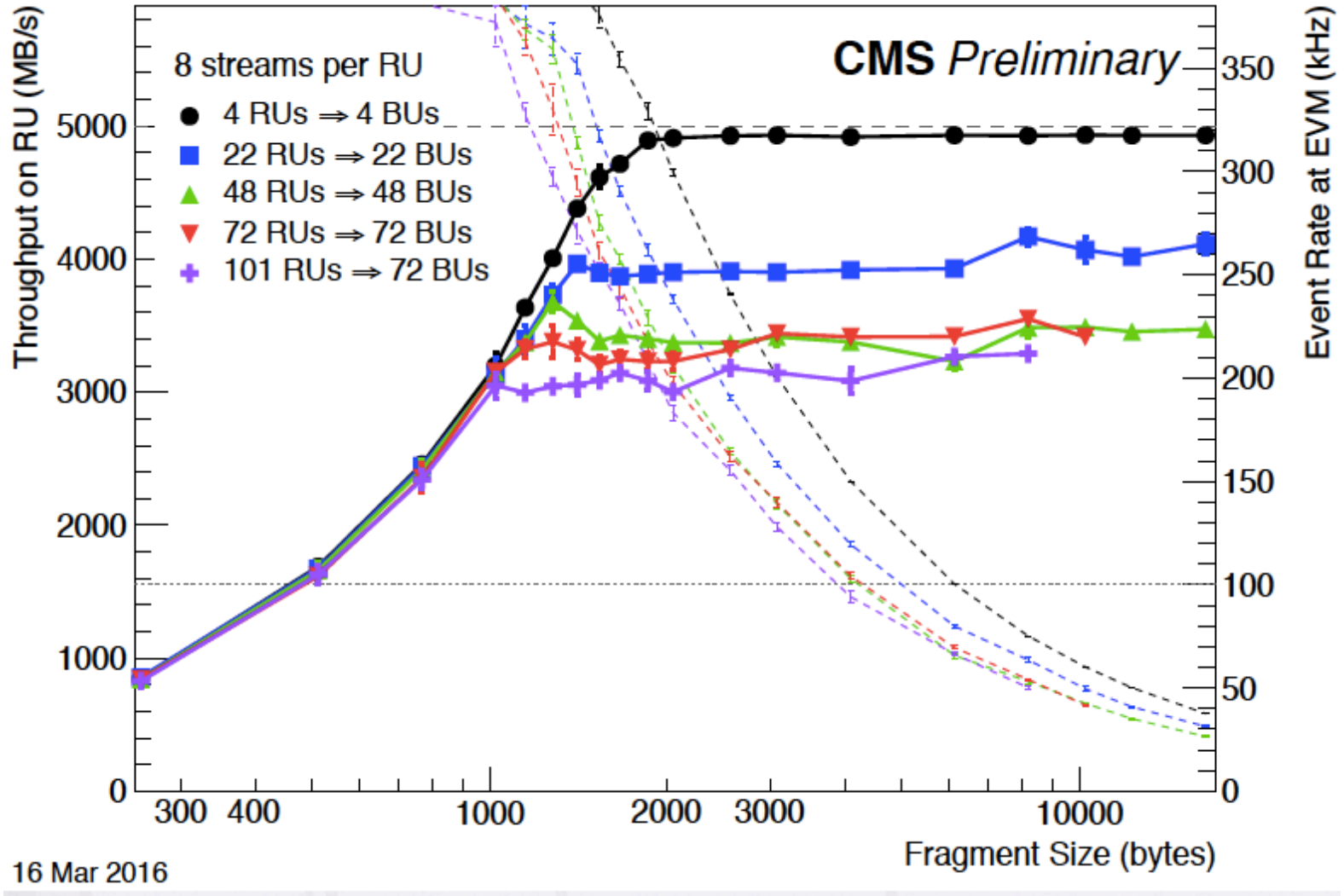


FRL/Myrinet

FRL/FEROL 10 Gb/s Ethernet

Switchover in 2014

Performance of EVB network (readout-unit)



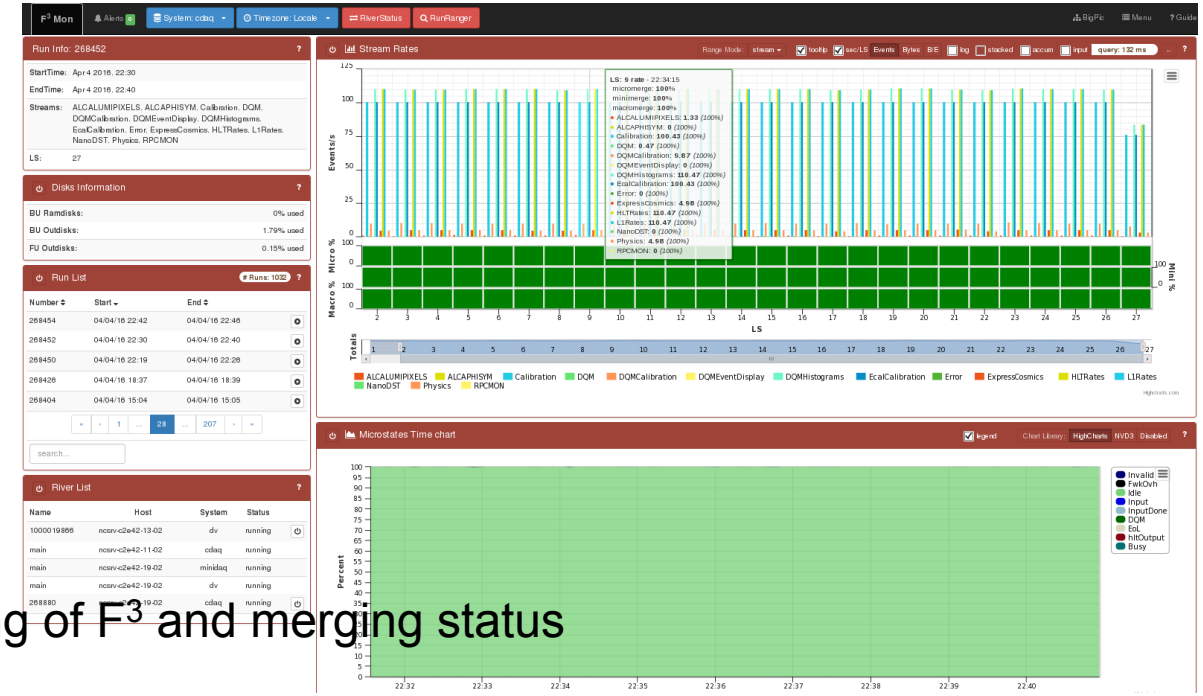
Streaming of data from *n* to *n* nodes (no event building)

HLT and merger monitoring

Elasticsearch (NoSQL DB) based monitoring

- near real-time latency (~second)
- Injection of all JSON metadata files used in Filter Farm (and merging system)

Used to build several web tools



F^3Mon UI - monitoring of F^3 and merging status

Elasticsearch also used for monitoring of other DAQ components (XDAQ monitoring)

For more details see talk by D. Simelevicius