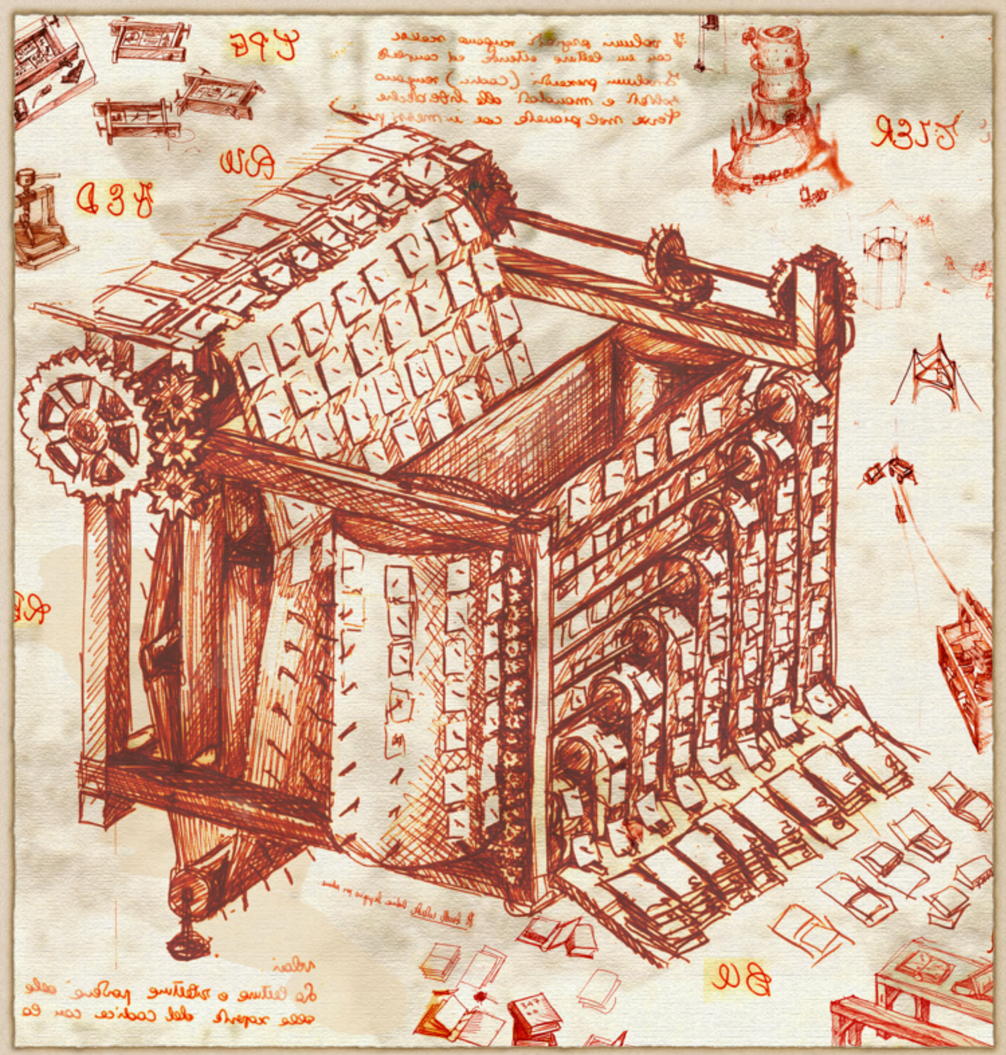
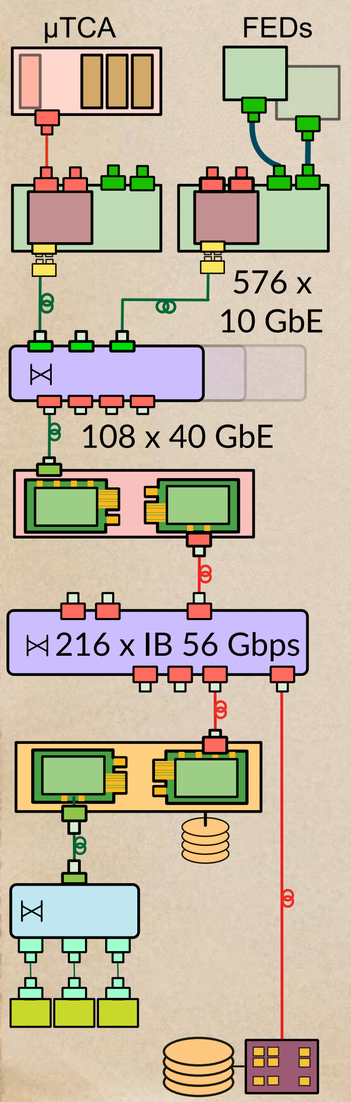
The CMS DAQ System for Run-2

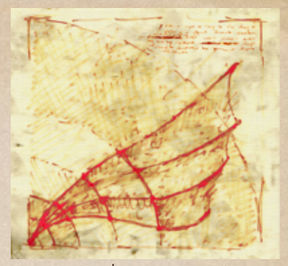
Remigius K Mommsen
Fermilab

on behalf of the CMS DAQ group



CMS Data Acquisition System





Detector front-end (custom electronics)
- ~700 front-end drivers (FEDs)
- 0.1 - 8 kB fragments at 100 kHz (1.2 MB event size)

Front-End Readout Optical Link (FEROL)
- Custom protocol from FEDs
- Optical 10 GbE TCP/IP

Data Concentrator switches
- Data to Surface over ~200m
- Aggregate into 40 GbE links

Up to 108 Readout Units (RUs)
- Combine FEROL fragments into super-fragment
- Buffer fragments

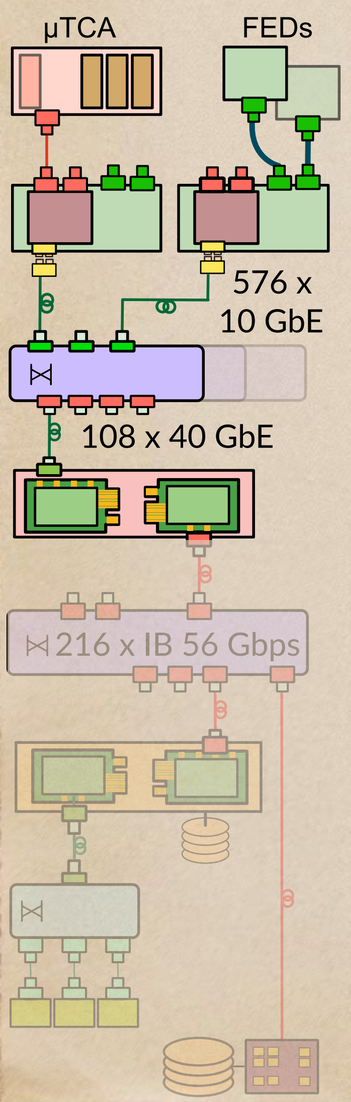
Event Builder switch
- Infiniband FDR 56 Gbps CLOS network

73 Builder Units (BUs)
- Event building & temporary recording to RAM disk

Filter Units (FUs)
~22k cores in ~940 boxes
- Run HLT selection using files from RAM disk
- Select O(1%) of the events for permanent storage

Storage and Transfer System
350 TB Lustre file system
- Merge output files from filter unit
- Transfer files to tier 0 or online consumers at pt.5

Data Concentrator



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

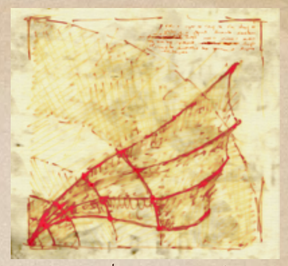
- Simplified TCP protocol over 10 GbE
- 1-18 FEDs merged into 40 Gbit Ethernet at switch level
- Fat-tree architecture interconnects any FEROL to any RU at full bandwidth

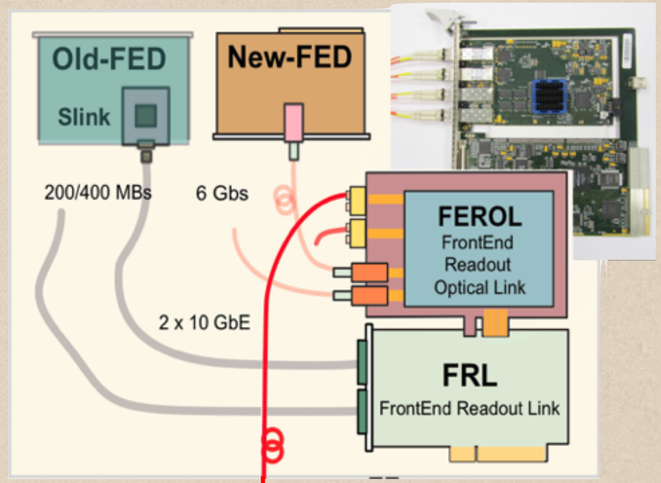
Each FED has one TCP stream

- Readout Unit (RU) splits stream into FED fragments
- Checks FED fragments for consistency and buffers them

FEROL40 under development

- µTCA standard (without legacy FRL board)
- 4x10 Gbps optical input and 40 GbE output



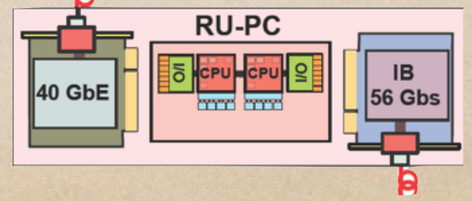


48 x 10 Gb/s

10 Gb/s simplified TCP/IP from an FPGA

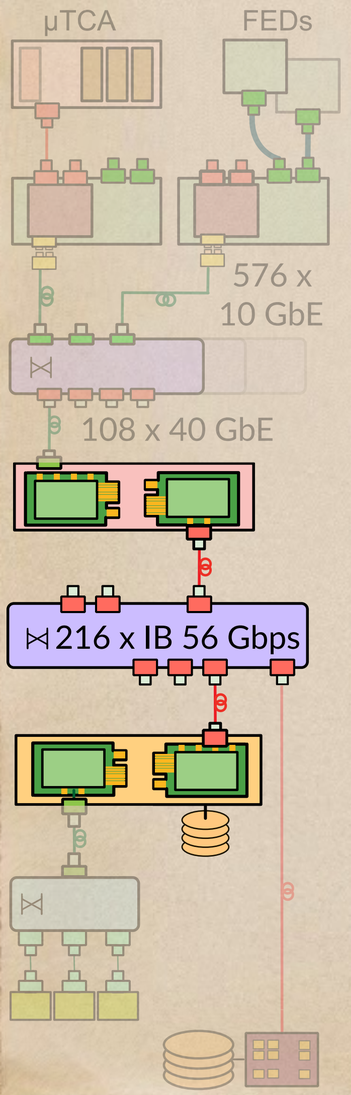
Data concentration:
10/40 Gb/s Ethernet switch

6 x 40 Gb/s



dual 8-core
E5 2670

Event Builder

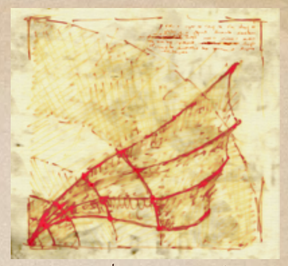


InfiniBand – most cost-effective solution

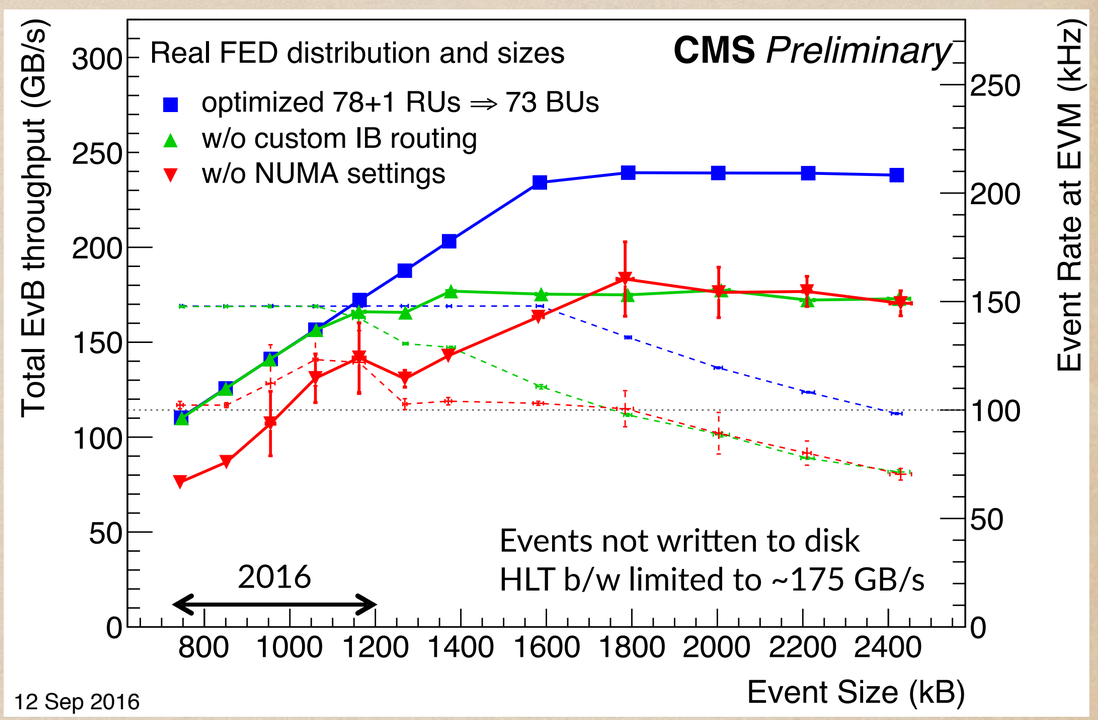
- Reliability in hardware at link level (no heavy software stack)
- Credit-based flow control (switches do not need to buffer)
- Easy to construct a large network from smaller switches

Event Builder Performance

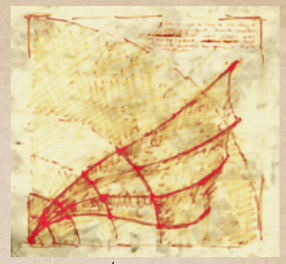
- Avoid high rate of small messages
- Avoid copying data
- Parallelize the work
- Bind to CPU cores and memory (NUMA)
- Tune Linux TCP stack for maximum performance
- Use custom IB routing taking into account the event-building traffic pattern

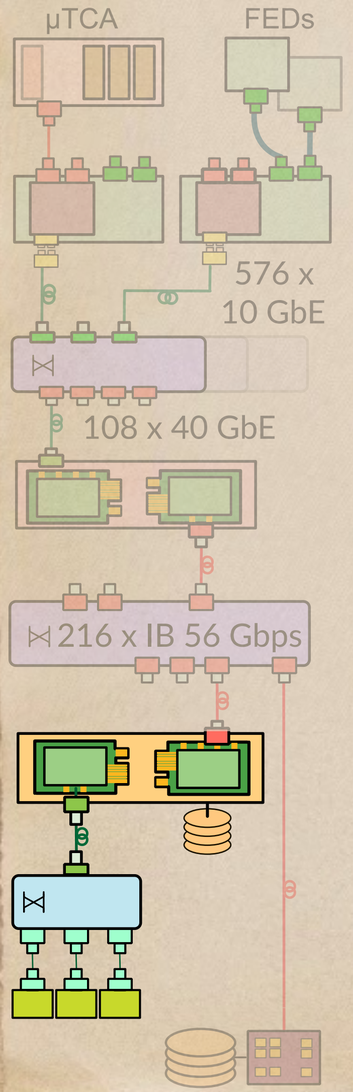


More information on performance of the CMS Event Builder on poster #116



File-Based Filter Farm (F³)





Each builder unit has 12 or 16 filter units

- Static mapping depending on machine generation
- Filter units mount RAM disk on BU via NFSv4
- Filter units pick next available file to process

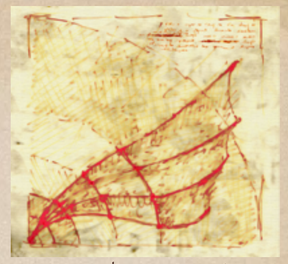
HLT selection uses standard CMSSW jobs

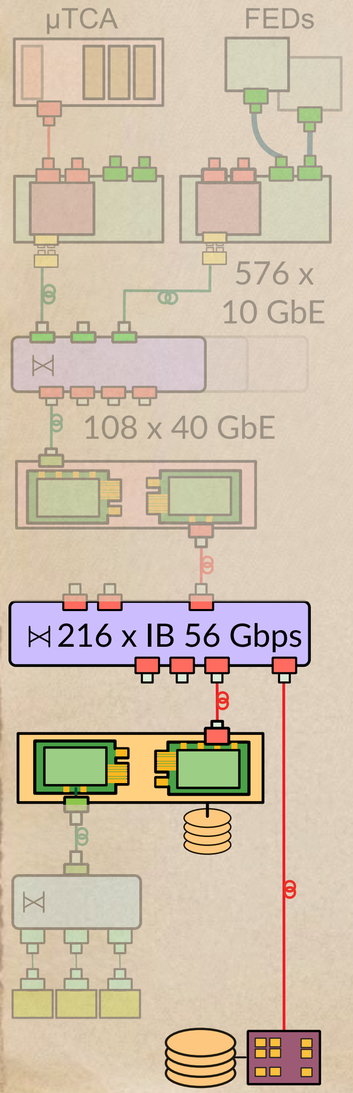
- Standalone process independent from online data-acquisition framework
- DAQ specific plug-ins for file discovery & monitoring
- Each filter unit runs several CMSSW instances
- Each CMSSW instance uses 4 threads
- New processes are started for each run
- Selected events are written to local files
- Files are copied back to output disk on the BU
- Processes exit once the last file of the run has been processed

	Dell C6220	Megware S2600KP	Action S2600KP
CPU (2x)	E5-2670 (sandy bridge)	E5-2680v3 (haswell)	E5-2680v4 (broadwell)
Cores	16	24	28
RAM	32 GB	64 GB	64 GB
HS06/ node	350	538	659
#nodes	256	360	324
#cores	4096	8640	9072

Total: ~22k cores on 940 motherboards with ~500 kHS06

Storage and Transfer System





File-Based Filter Farm produces output files

- 940 FU nodes create their own files
  - One file for each of the ~25 different output and monitoring streams
  - A new file for each luminosity section (~23s)
- To be merged into 1 file per stream and luminosity section in a central place

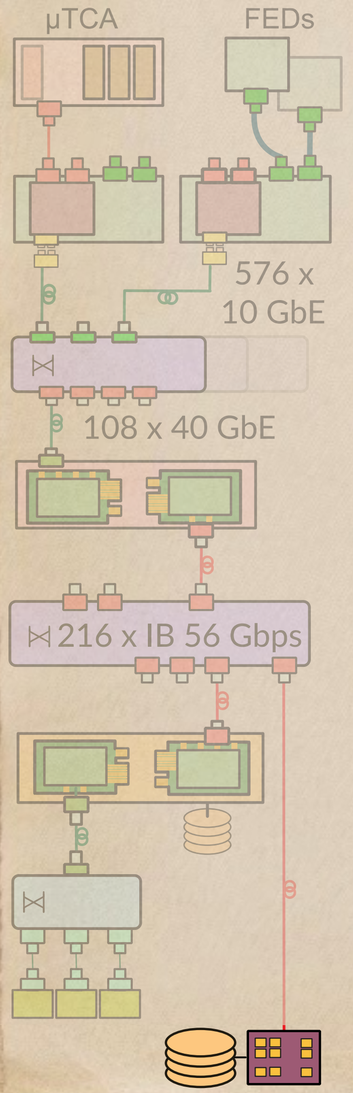
Files merged into a global file system (Lustre) on a storage system with 350 TB

- Merger process on BU reads data from the local output disk
- Event-data files are concurrently written into a single file on the global file system
- Monitoring data (histograms or scaler data) are aggregated first per BU and then on the global file system

Transfer system distributes the merged files

- Transferred to tier 0 for offline processing
- Copied to local consumers at pt.5 for data-quality monitoring, event display & fast calibration
- Monitoring data (HLT rates and event counts) are inserted into DB

Lustre Filesystem



Lustre

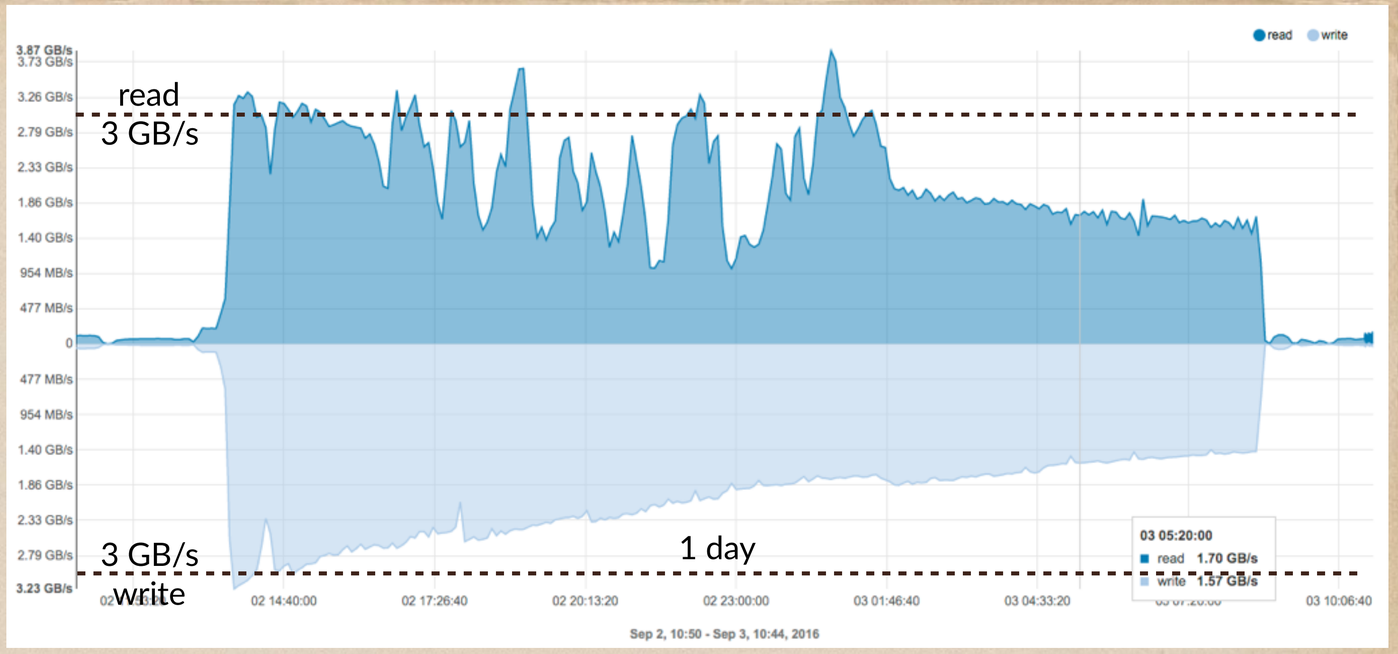
- 1 Metadata Service (MDS)
- 2 Object Storage Services (OSS)
- Added 3rd OSS yielding 50% more throughput

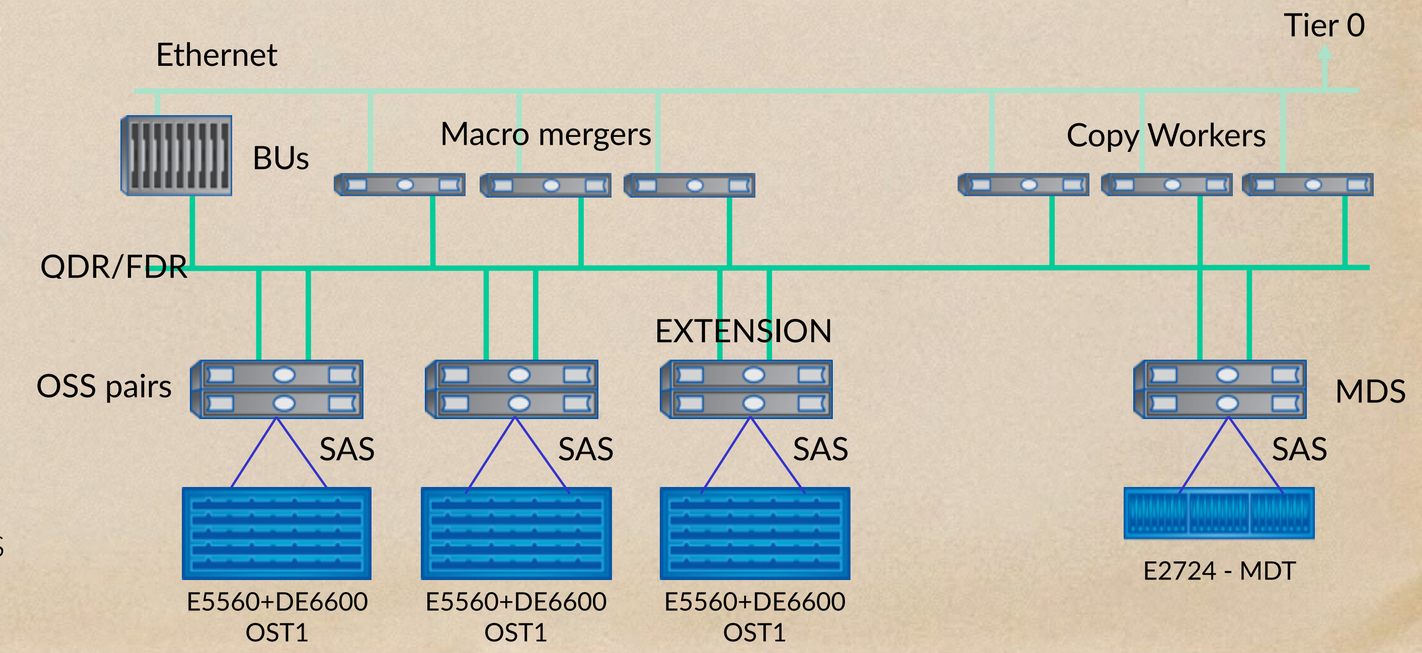
NetApp E-Series

- 1 TB for Metadata (MDS/MDT)
- 240 TB raw space per OSS
- RAID 6 systems
- Fully redundant
- Connected over IB and 40 GbE

Experience

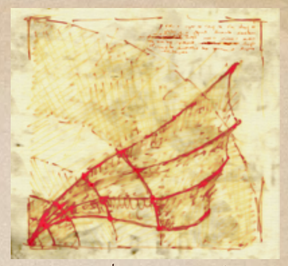
- Careful tuning to get full performance
- Sensible to network instabilities
- No data loss

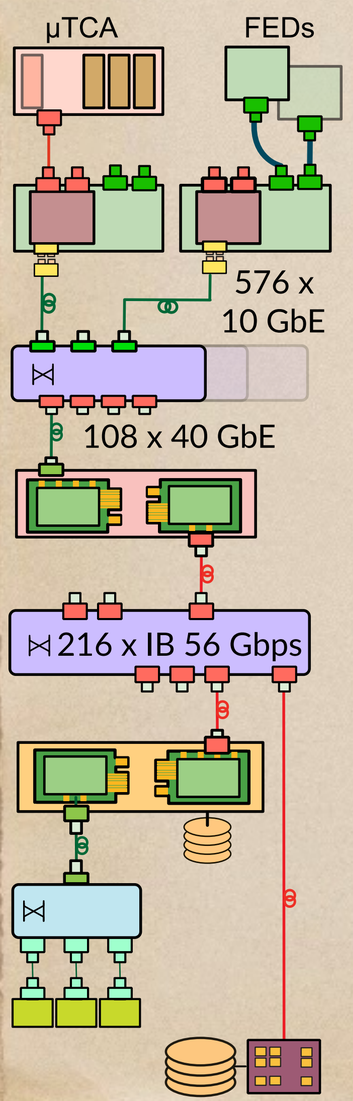




Configuration & Control

More information in RCMS talk #264





Data-flow applications based on XDAQ C++ framework

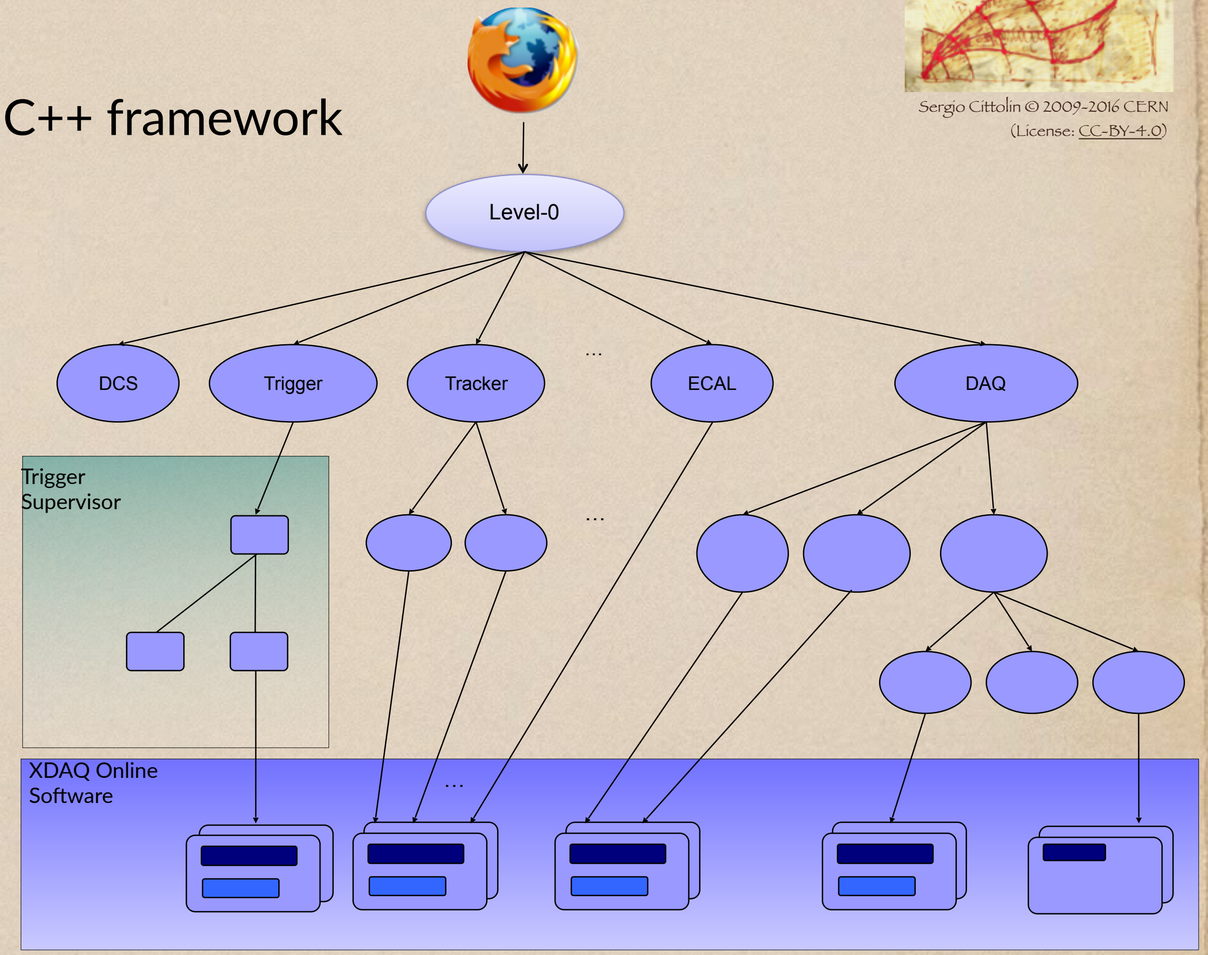
- Reusable building blocks for
  - Hardware access
  - Transport protocols
  - Services
- Dynamic configuration based on XML
- Controlled and browsable with HTTP/SOAP

Run-Control & Monitoring System

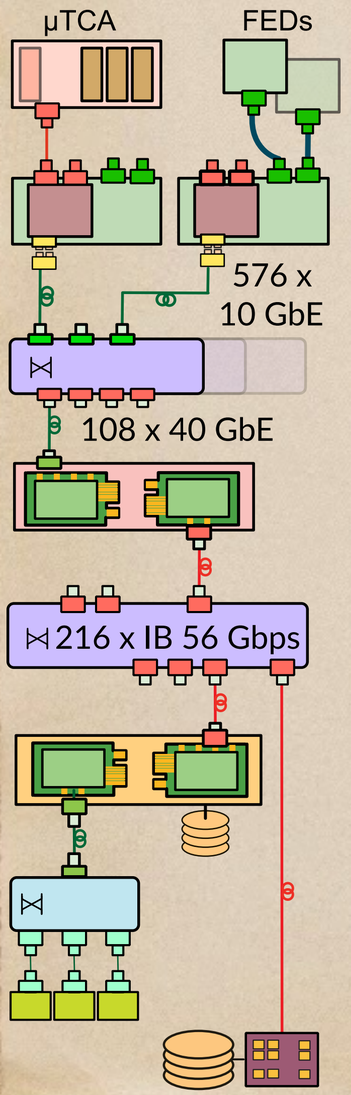
- Hierarchical control structure
- Java code running as Tomcat servlets
- React on state machine events
  - Commands from parents
  - Errors from children

File-based filter farm

- Daemons running asynchronously to run boundaries
- Driven by appearance of directories or files



Monitoring & Error reporting



XDAQ services in each application

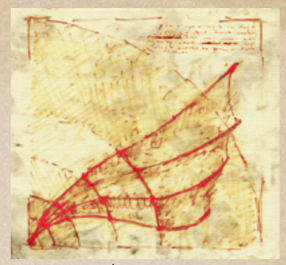
- Periodically publish monitoring data
- Central logging facilities
- Error reporting

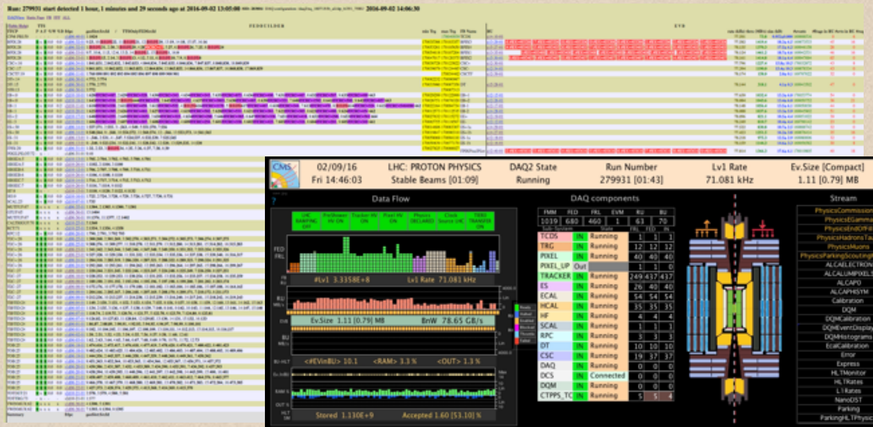
Services to centrally access the data

- Monitoring tools aggregate the data
- Display information for shifters and experts
- Expert system is being commissioned

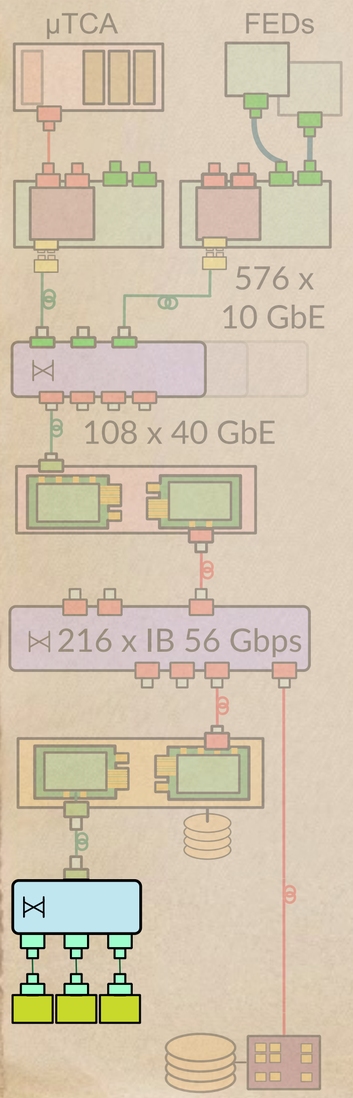
File-based filter farm uses Elastic Search

- Near real-time indexing of *O*(40000) JSON files / s
- Instantaneously querying and displays
- Investigating feasibility to migrate monitoring of all DAQ applications to JSON & Elastic Search





Online Cloud



HLT computing power similar to all CMS tier 1 sites combined

- Profit from CPU power outside of physics-data taking for offline computing workflows

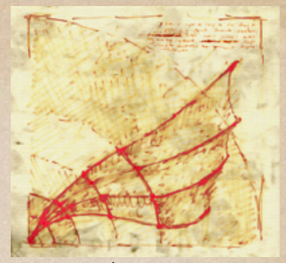
Cloud overlay acting as tier 2 site

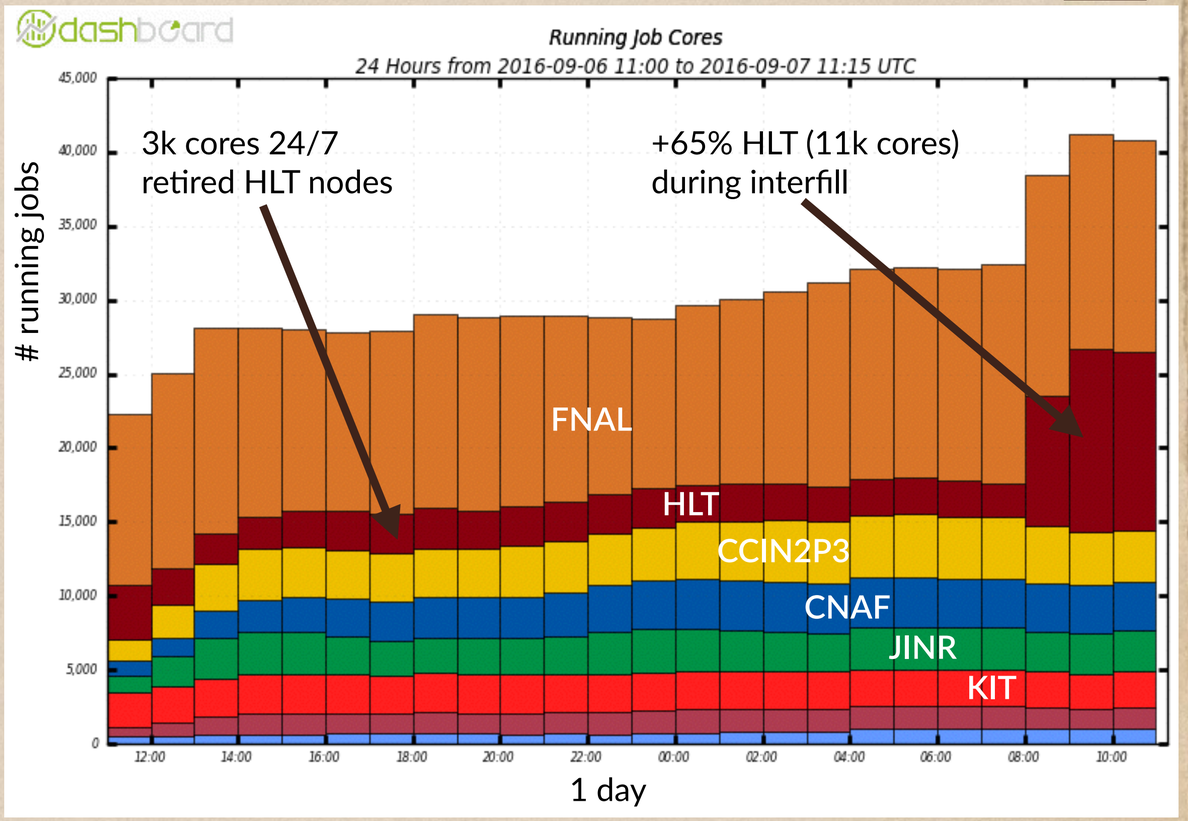
- Virtual machines using OpenStack Grizzly
- No local data storage (retrieved from CERN)
- Started and stopped based on LHC beam states (257 days in 2016)
- Retired HLT nodes permanently available for cloud
- Average 10k cores across year

Challenges

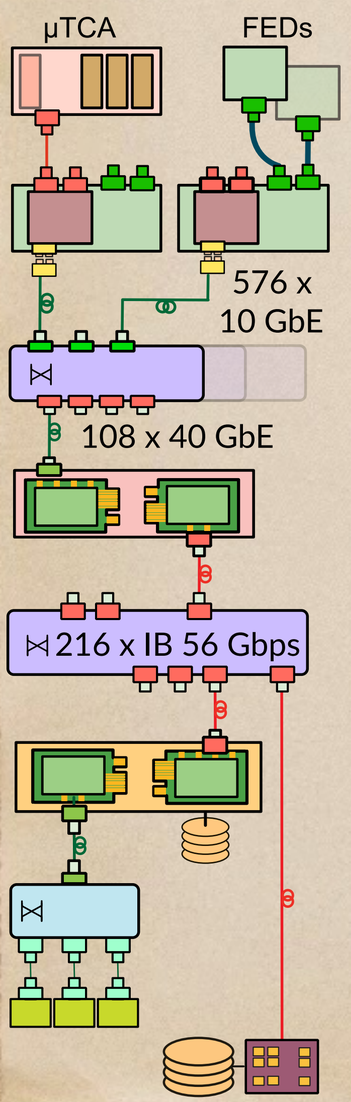
- Quickly start 800-1000 of virtual machines simultaneously
- Avoid process timeouts when hibernating VM images for several hours during data taking

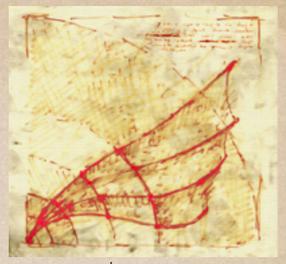
More information on CMS online cloud in talk #412





Summary





CMS DAQ system for run 2 fully commissioned

- Fulfills functional and performance requirements
- Ready to integrate new/upgraded sub-detectors in 2017
- Main challenge will be the new FEROL40 hardware

Extensive tuning to achieve full performance

- Throughput per data concentrator is ~4 GB/s
- Total event building bandwidth ~230 GB/s
- Bandwidth to HLT ~175 GB/s
- HLT output bandwidth to storage ~4.5 GB/s

Focus shifts on monitoring and automation

- Improved monitoring with Elastic Search
- New expert system is being commissioned
- Further automation of routine tasks and error recovery

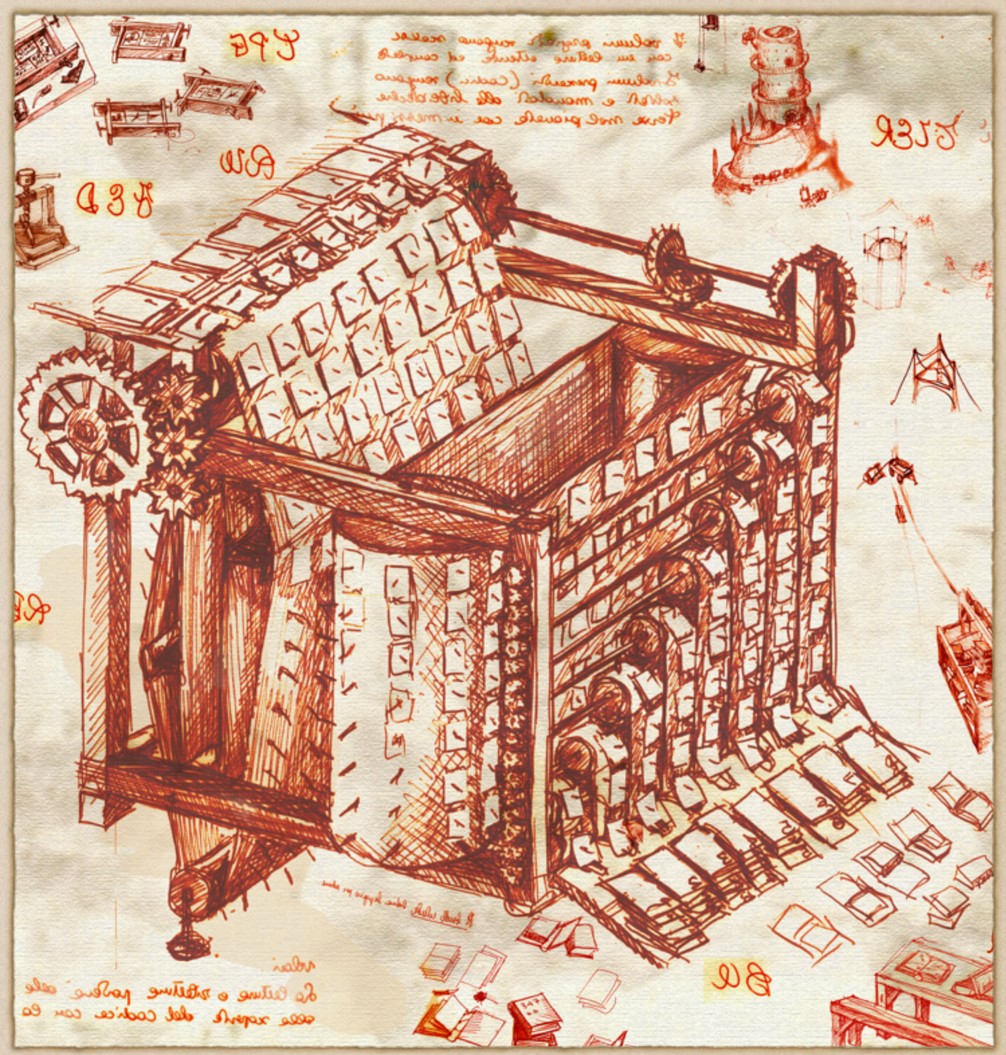
Questions?

Other talks about CMS DAQ @ CHEP

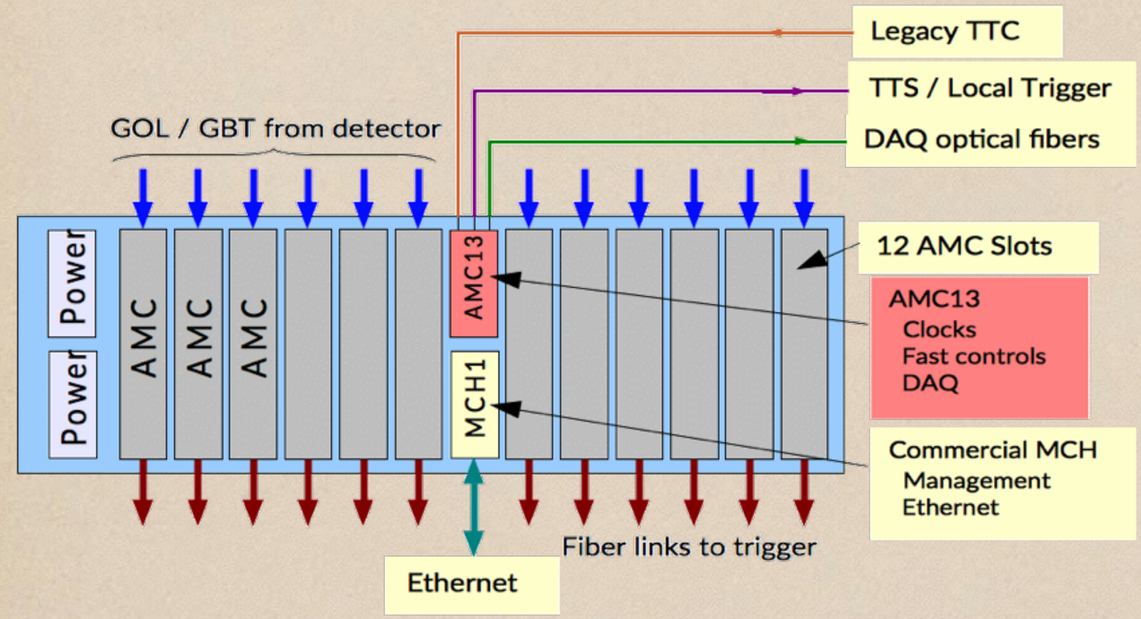
- New operator assistance features in the CMS Run Control System (Hannes Sakulin, #264)
- The CMS Data Acquisition – Architectures for the Phase-2 Upgrade (Emilio Meschi, #299)
- Dynamic resource provisioning of the CMS online cluster using a cloud overlay (Marc Dobson, #412)

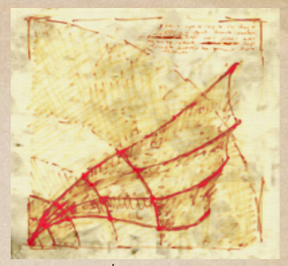
Posters

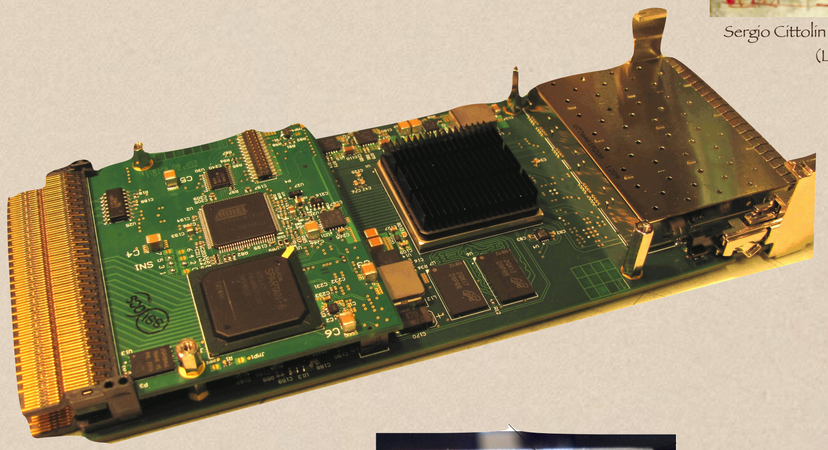
- Evolution, design, management and support for the CMS Online computing cluster (Marc Dobson, #506)
- Performance of the CMS Event Builder (Remi Mommsen, #116)

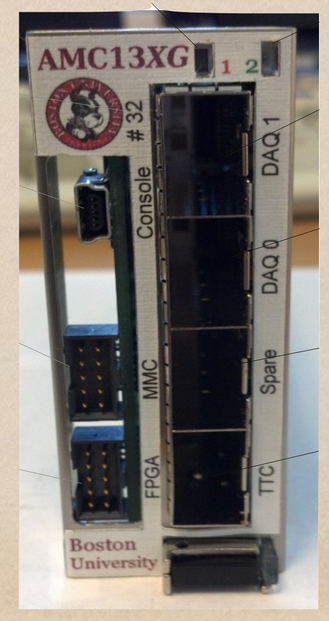


AMC13



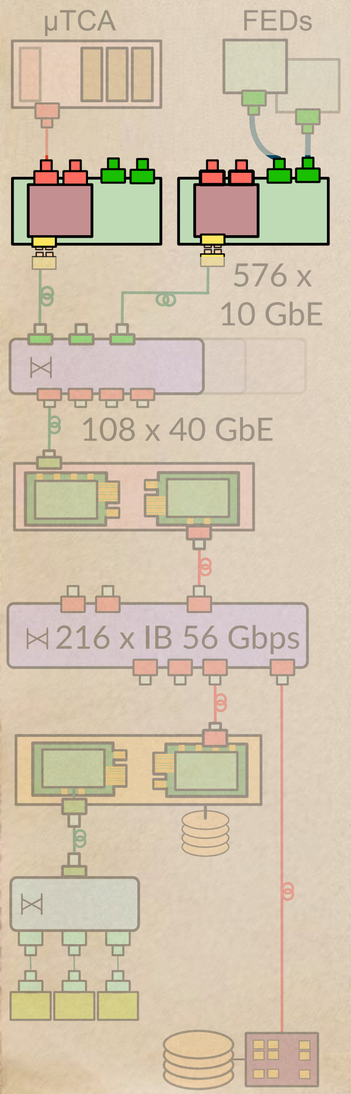


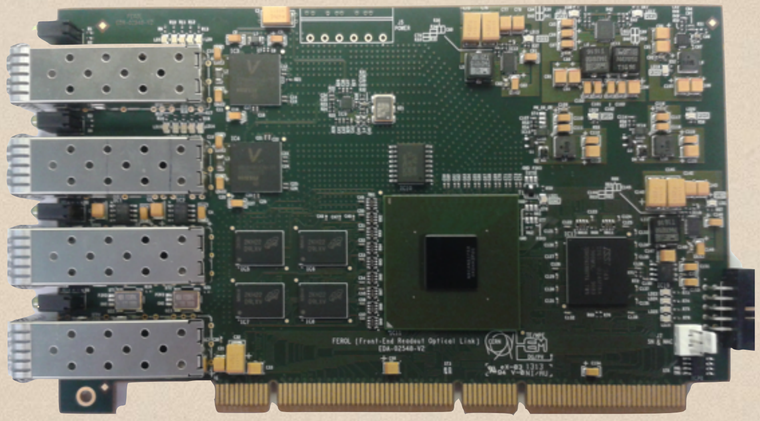


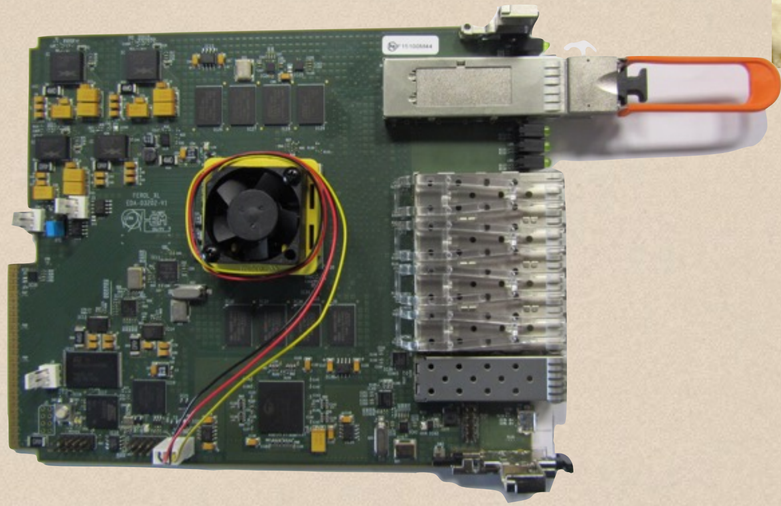


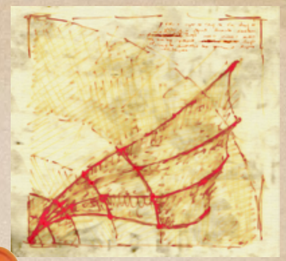
- It is not an MCH! It is a 13th AMC in MCH-2 slot
- It distributes LHC clock / timing / controls to AMCs
- It collects DAQ data from AMCs
- It provides standard interface to CMS subdetectors:
  - CMS DAQ via 1-3 optical fibers at 10 Gb/s (64/66b encoded)
  - TTC via 1300nm fiber @ 160Mb/sec biphase mark code

Front-End Readout Optical Link (FEROL)



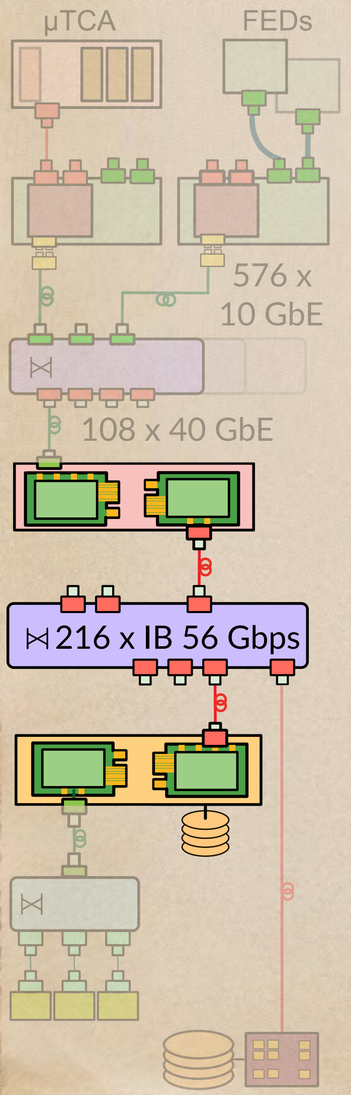






	FEROL	FEROL40
FPGA	Altera Arria II GX	Altera Arria V GZ
QDR Memory	16 MB	32 MB
DDR Memory	512 MB DDR2	2x 1GB DDR3
Input (SLINKXpress)	2x optical 6 Gbit/s or 1x optical 10 Gbit/s	4x optical 10 Gbit/s
DAQ interface (Ethernet)	1x optical 10 Gbit/s	4x10 Gbit/s or 40 Gbit/s

Event Builder Performance



Avoid high rate of small messages

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

Avoid copying data

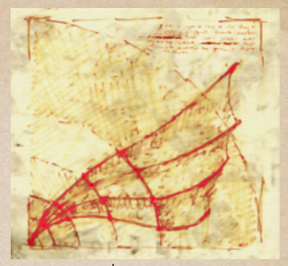
- Operate on pointers to data in receiving buffers
- Copy data directly into RDMA buffers of IB NICs
- Stay in kernel space when writing data

Parallelize the work

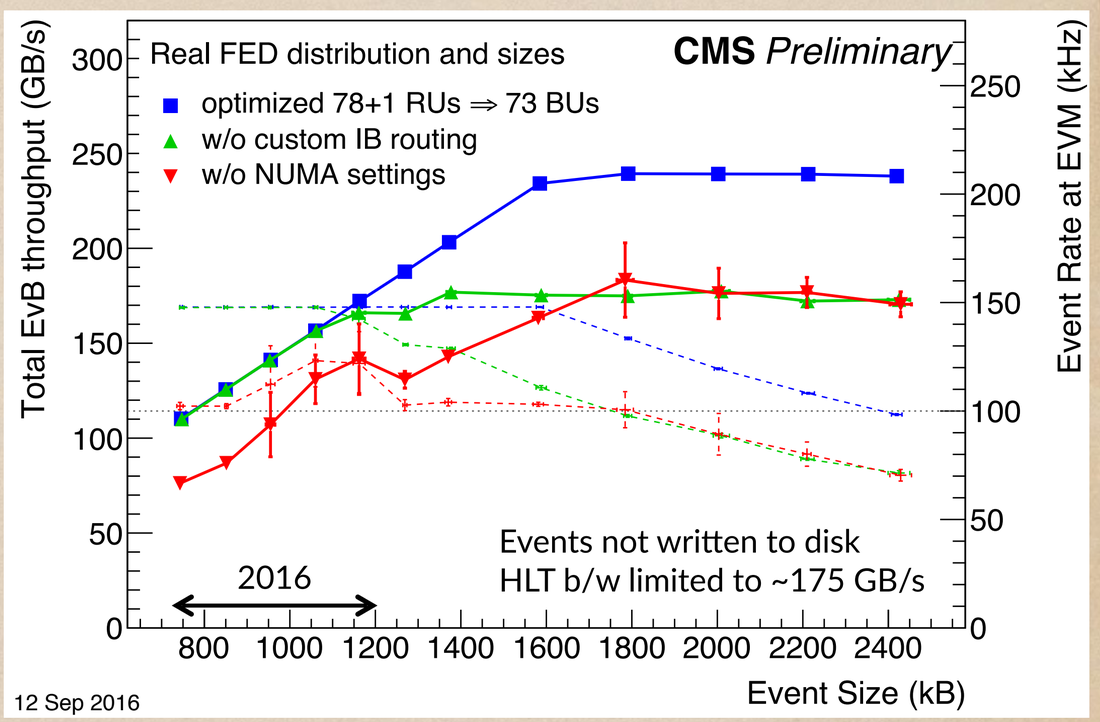
- Multiple threads parallelize event handling
- Write events concurrently into multiple files

Bind to CPU cores and memory (NUMA)

- Bind threads & memory structures to cores
- Restrict interrupts from NICs to certain cores
- Tune Linux TCP stack for maximum performance

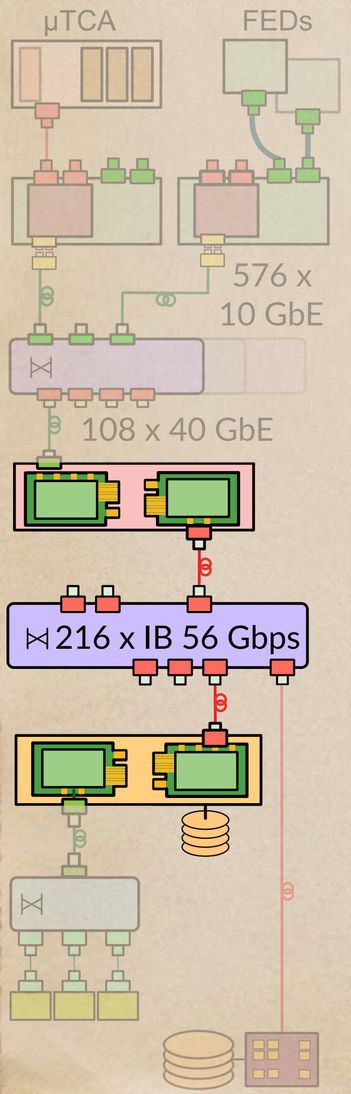


More information on performance of the CMS Event Builder on poster #116



Use custom IB routing taking into account the event-building traffic pattern

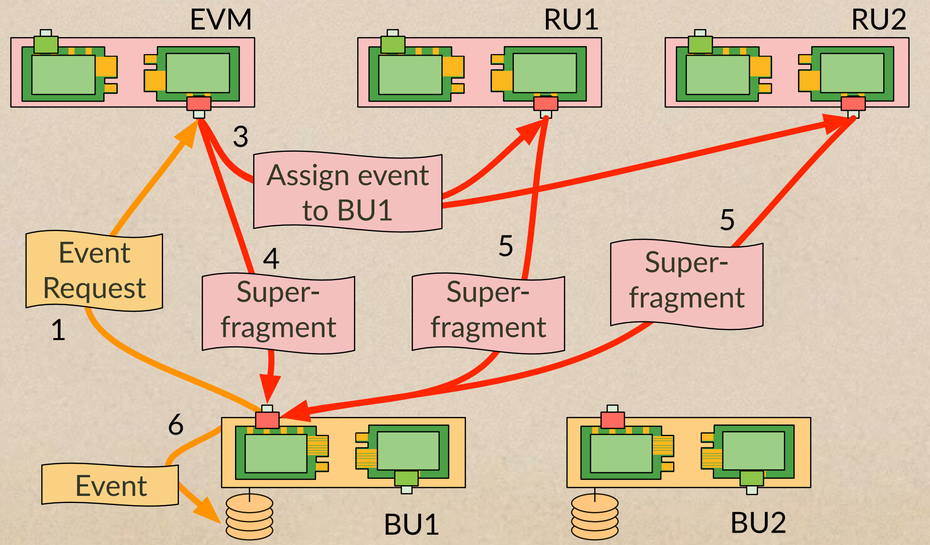
Event Builder

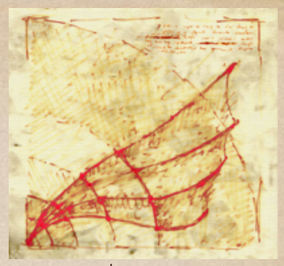


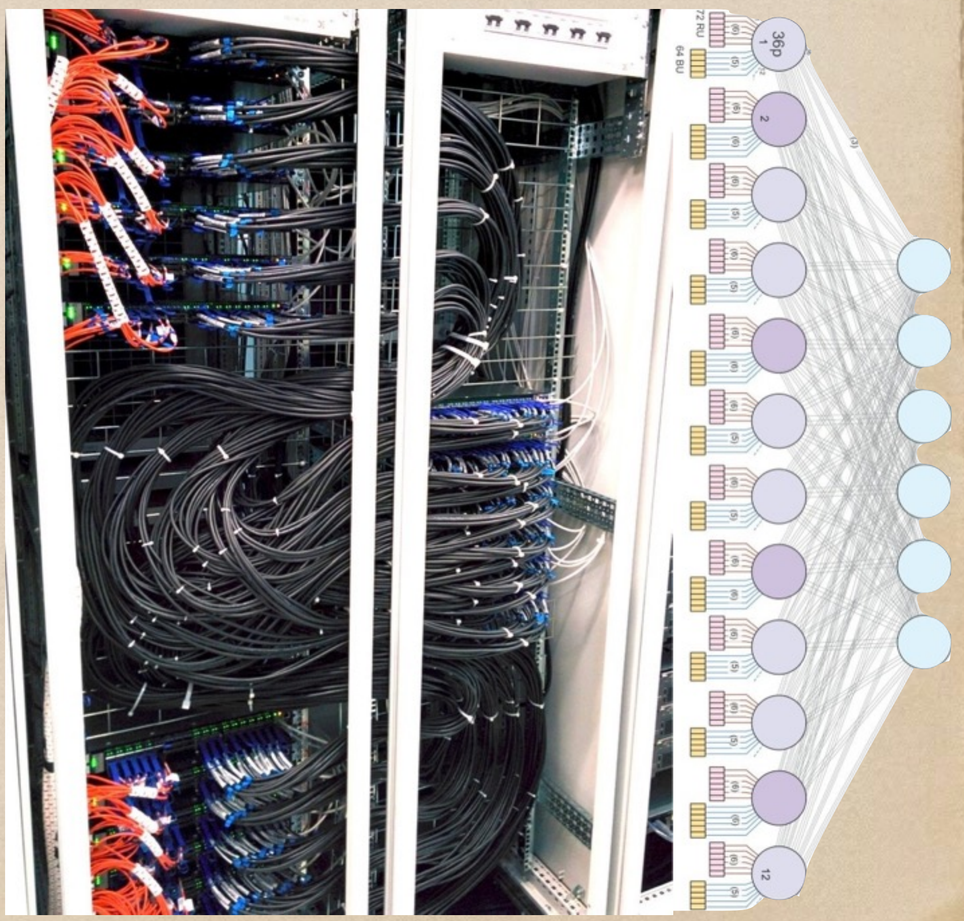
InfiniBand – most cost-effective solution

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

Event Builder protocol

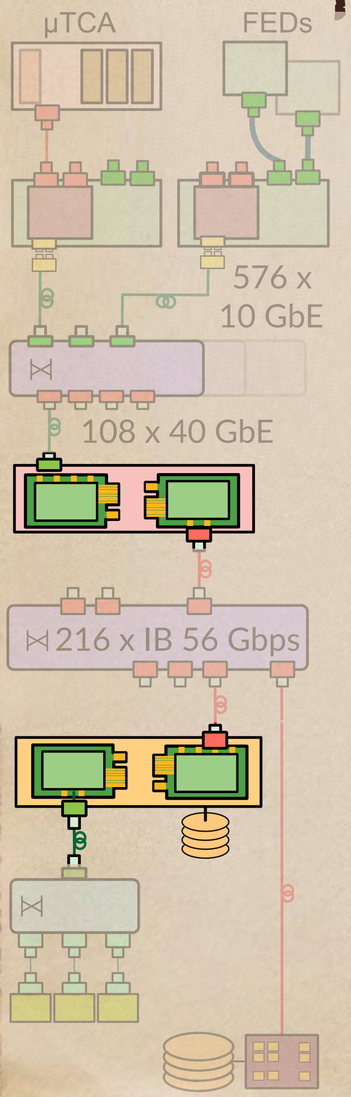






Infiniband CLOS network

Computers



Readout Unit (RU)

- Dell PowerEdge R620
- Dual 8 core Xeon CPU E5-2670 0 @ 2.60GHz
- 32 GB of memory

Builder Unit (BU)

- Dell PowerEdge R720
- Dual 8 core Xeon CPU E5-2670 0 @ 2.60GHz
- 32+256GB of memory (240 GB for Ramdisk on CPU 1)