Overview of CMS DAQ Upgrades

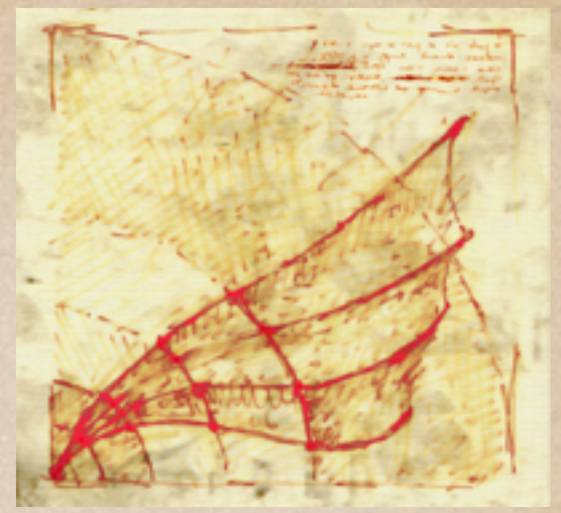
Remigius K Mommsen

Fermilab

on behalf of the CMS DAQ group

Sergio Cittolin © 2009-2016 CERN (License: CC-BY-4.0)

Contents



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

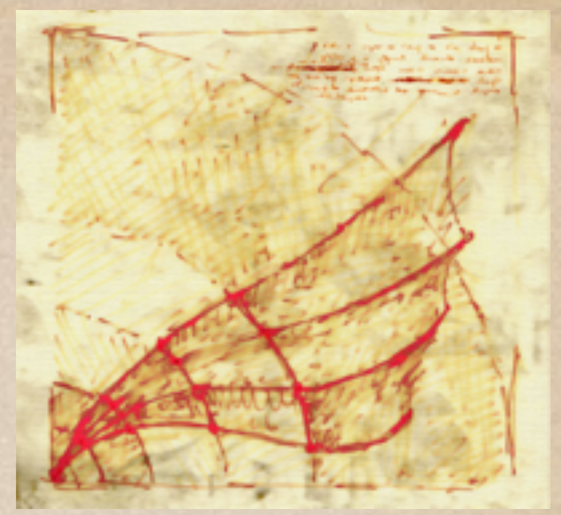
The current DAQ system for run 2

DAQ for after LS2 (run 3)

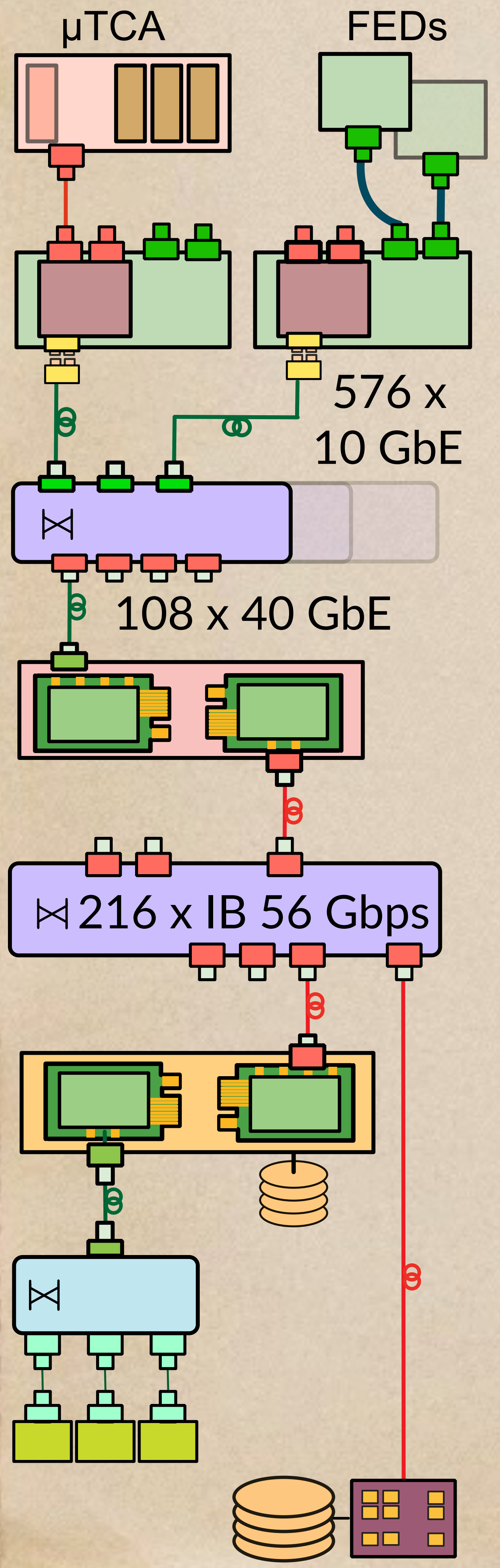
Ideas for Phase II

- TCDS-DAQ hub
- High performance event-builder node
- Event-builder architectures
- High-level trigger challenges
- Data scouting at 40 MHz

CMS Data Acquisition System



Sergio Cittolin © 2009-2016 CERN (License: CC-BY-4.0)



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

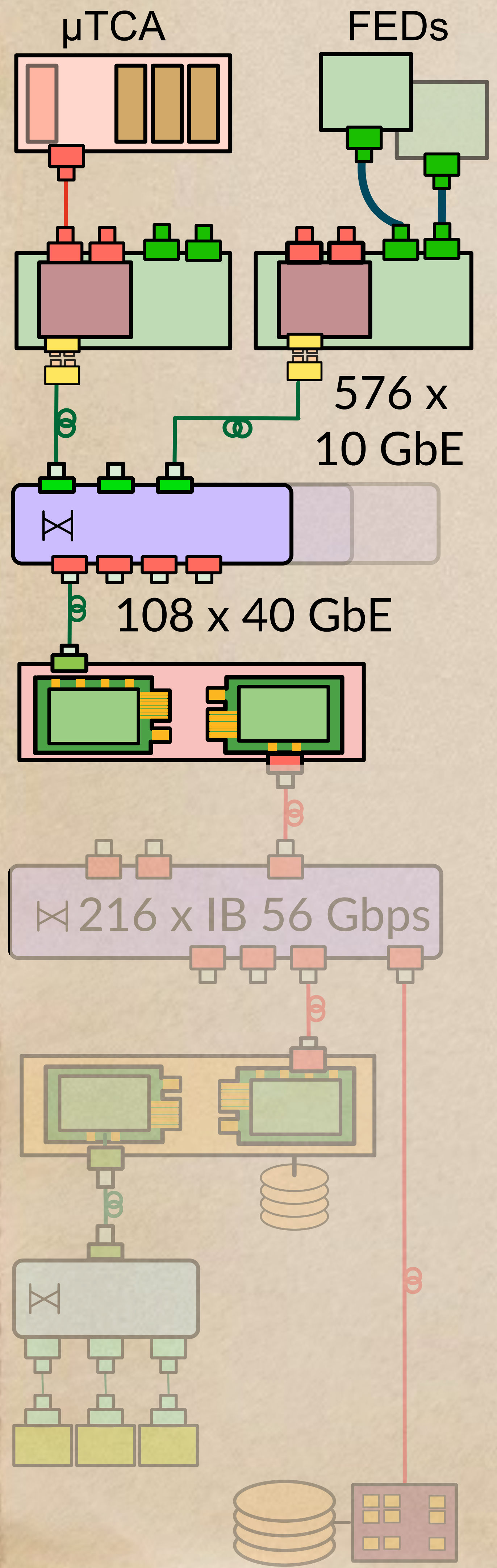
- Run HLT selection using files from RAM disk
- Select O(1%) of the events for permanent storage

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

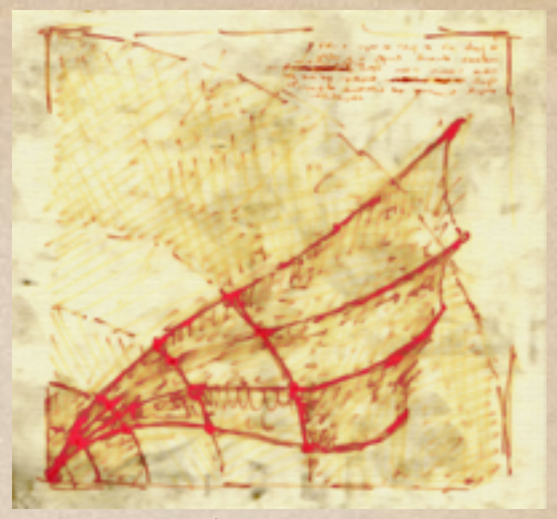
- Simplified TCP protocol over 10 GbE
- 1-18 FEDs merged into 40 Gbit Ethernet at switch level
- Fat-tree architecture interconnects any FEROL to any RU at full bandwidth

Each FED has one TCP stream

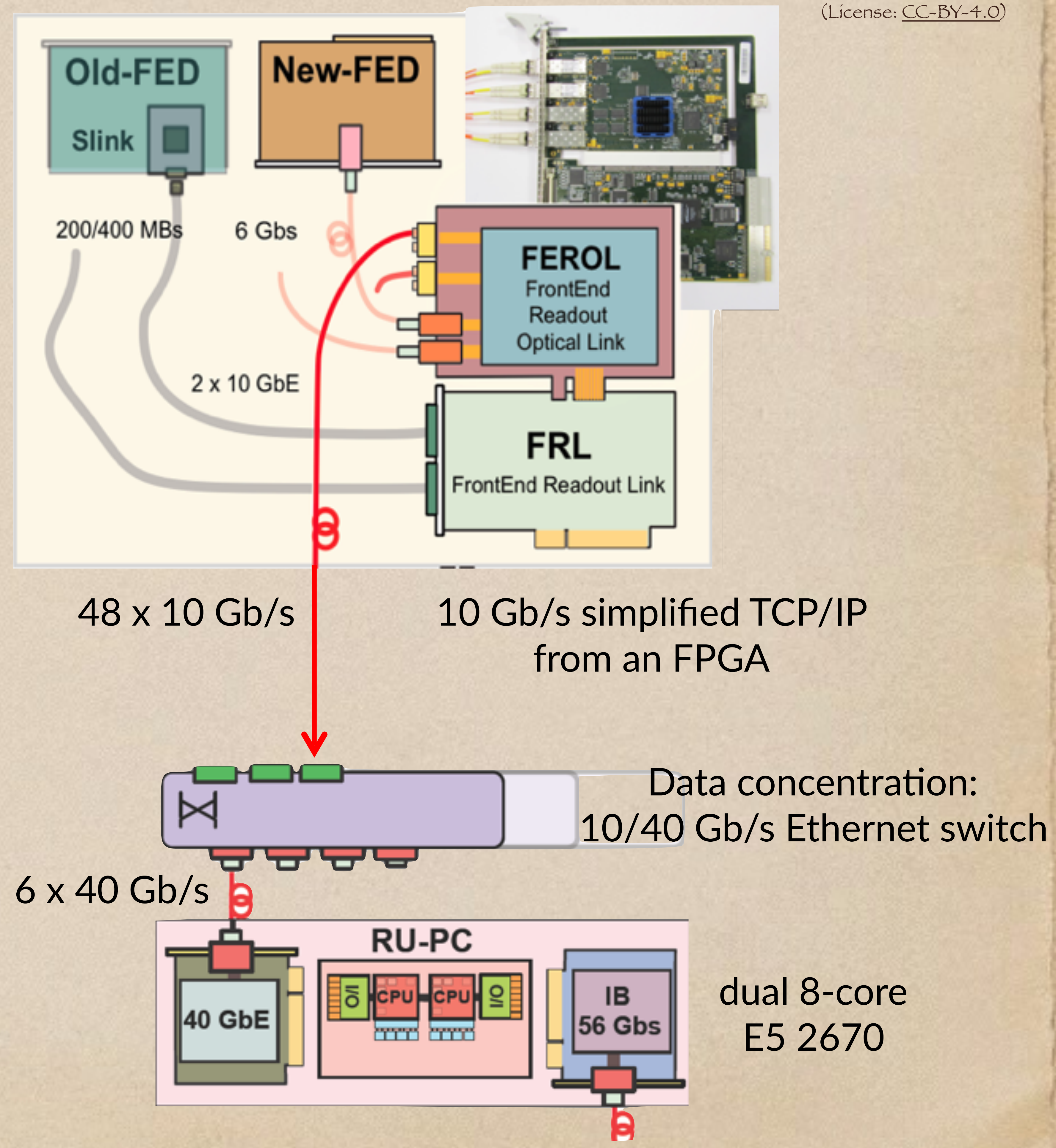
- Readout Unit (RU) splits stream into FED fragments
- Checks FED fragments for consistency and buffers them

FEROL40 under development

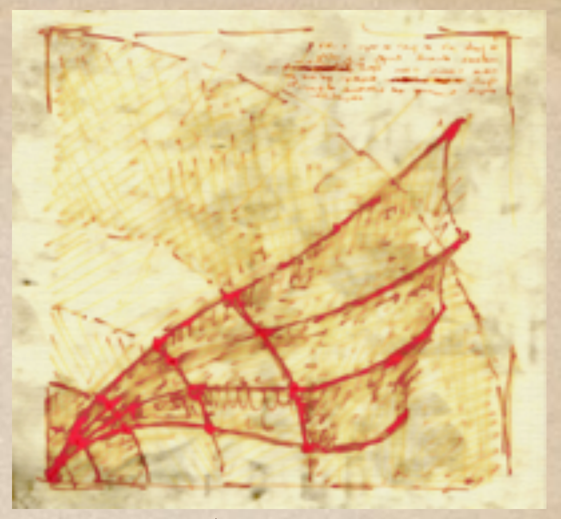
- µTCA standard (without legacy FRL board)
- 4x10 Gbps optical input and 40 GbE output



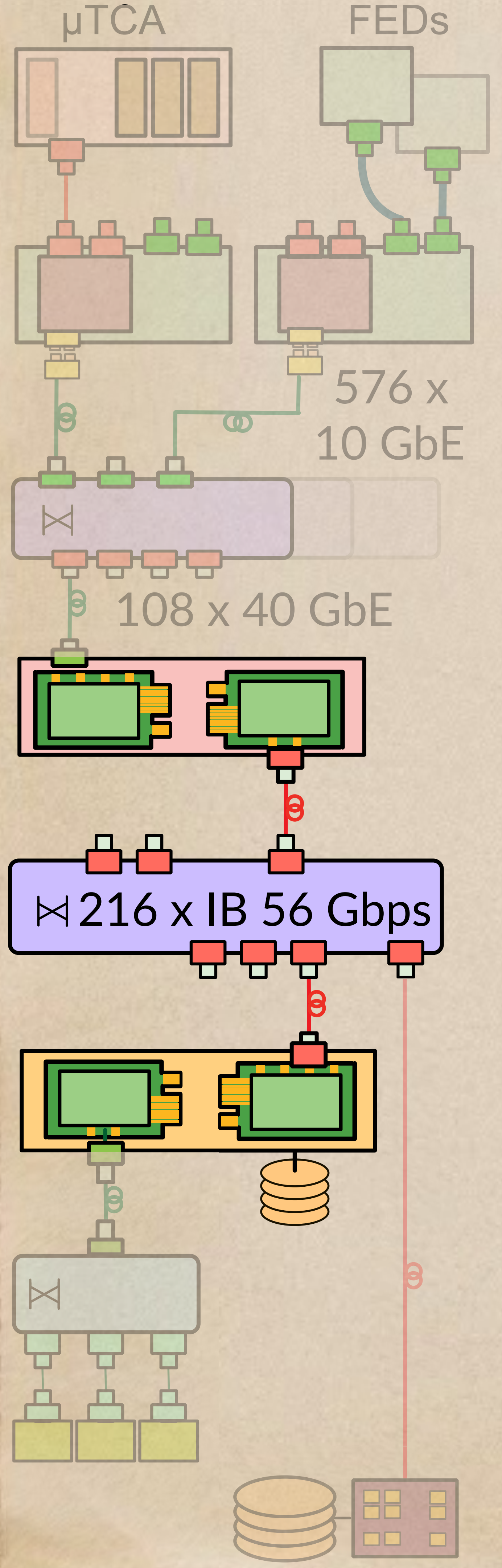
Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)



Event Builder



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

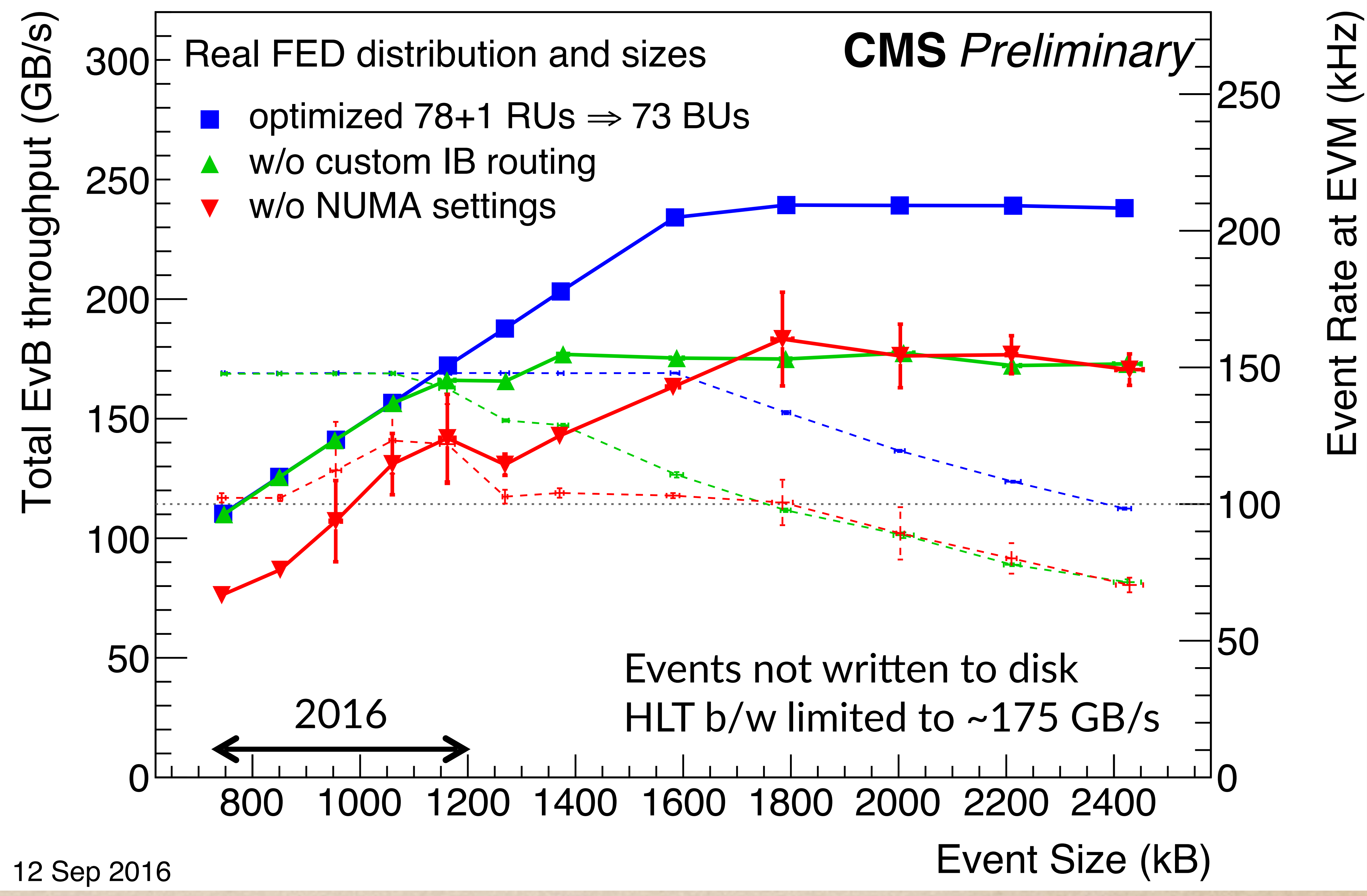


InfiniBand – most cost-effective solution

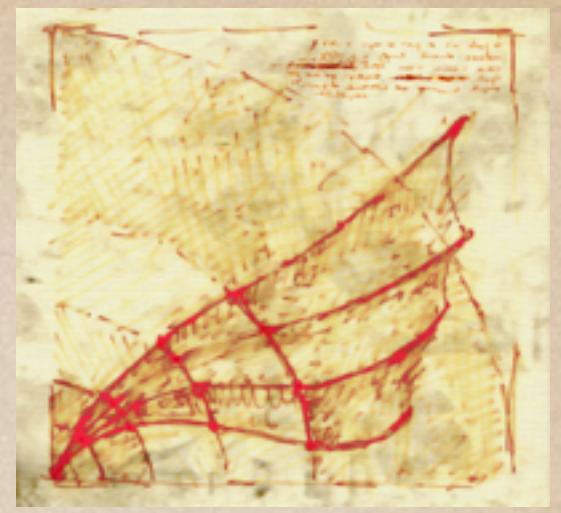
- Reliability in hardware at link level (no heavy software stack)
- Credit-based flow control (switches do not need to buffer)
- Easy to construct a large network from smaller switches

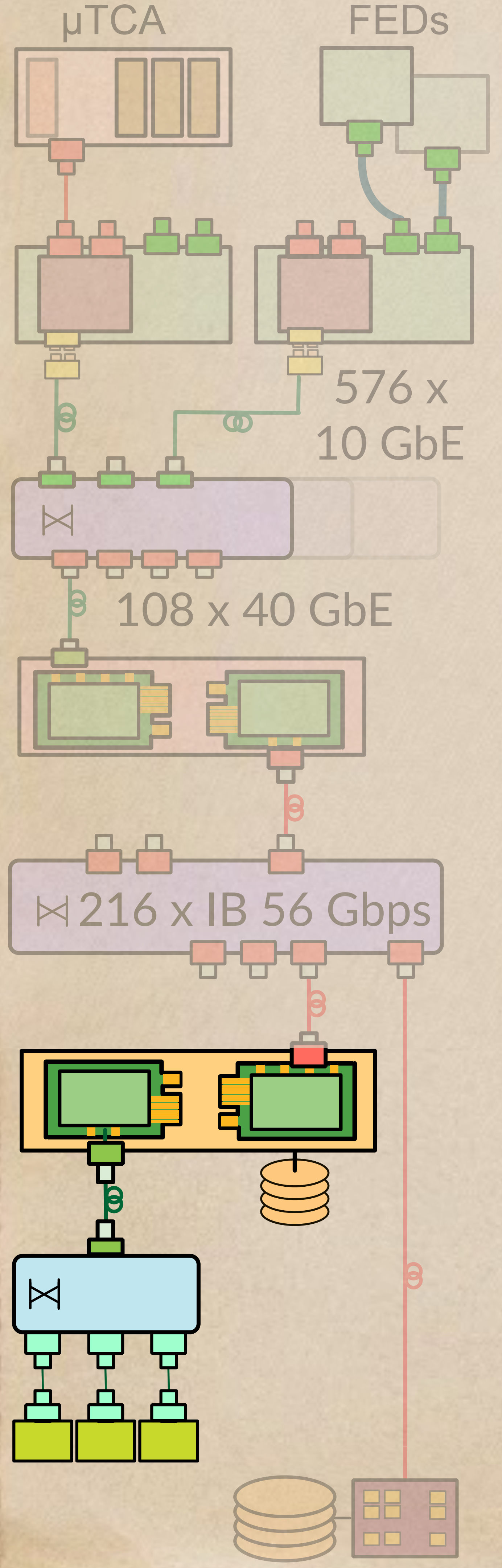
Event Builder Performance

- 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



File-Based Filter Farm (F^3)


Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)



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

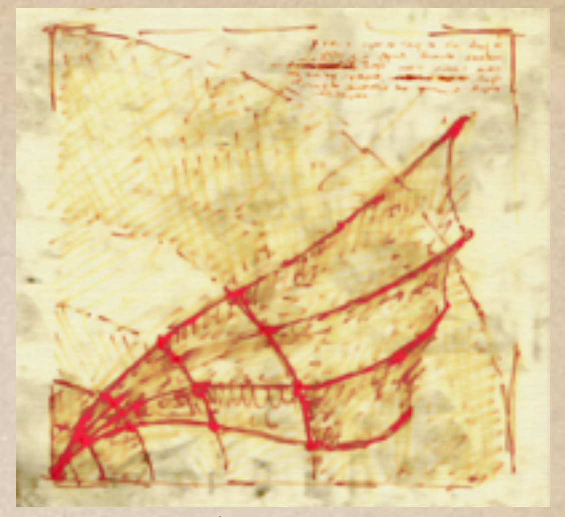
HLT selection uses standard CMSSW jobs

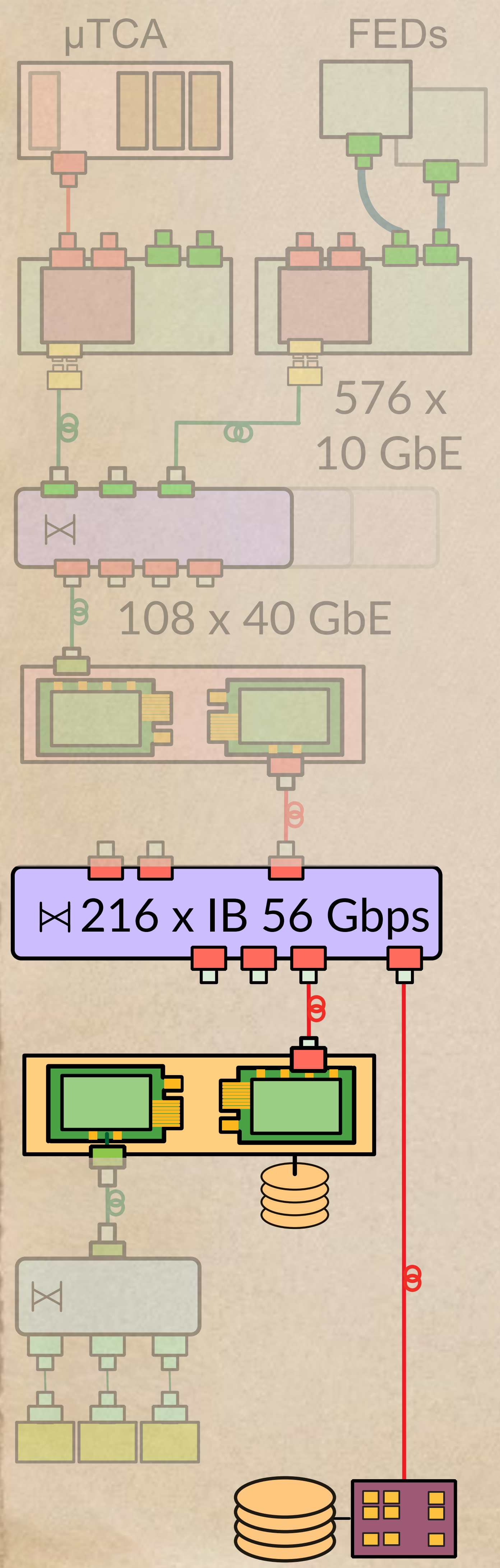
- Standalone process independent from online data-acquisition framework
- DAQ specific plug-ins for file discovery & monitoring
- Each filter unit runs several CMSSW instances
- Each CMSSW instance uses 4 threads
- New processes are started for each run
- Selected events are written to local files
- 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


Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)



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

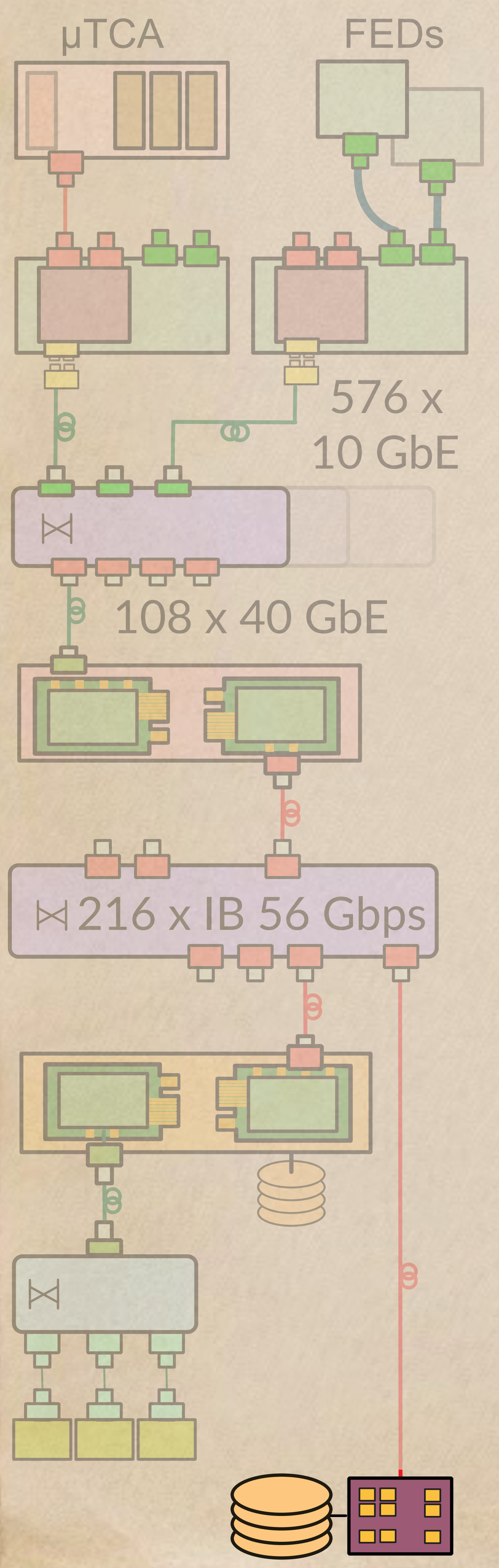
Files merged into a global file system (Lustre) on a storage system with 350 TB

- 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

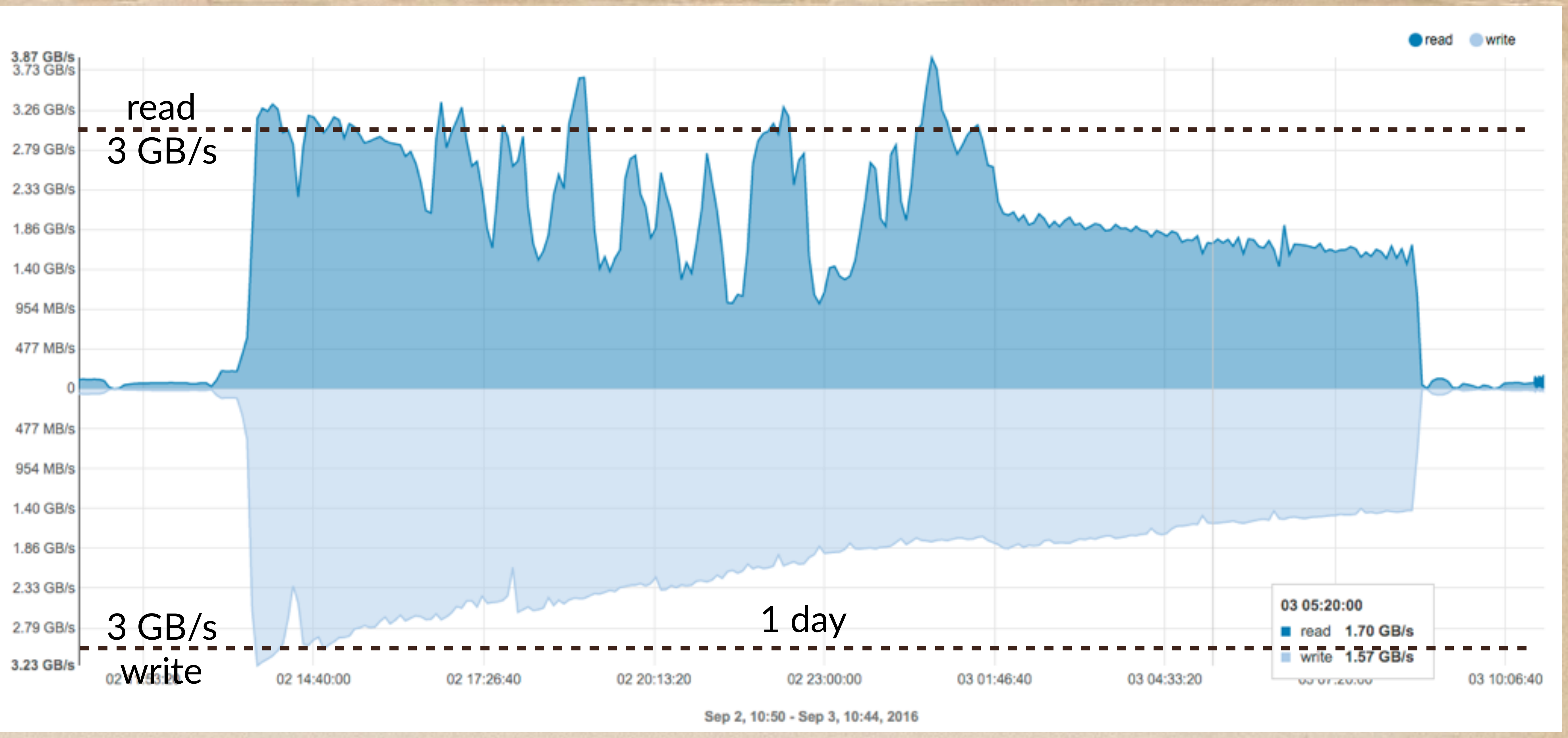
- 1 Metadata Service (MDS)
- 2 Object Storage Services (OSS)
- Added 3rd OSS yielding 50% more throughput

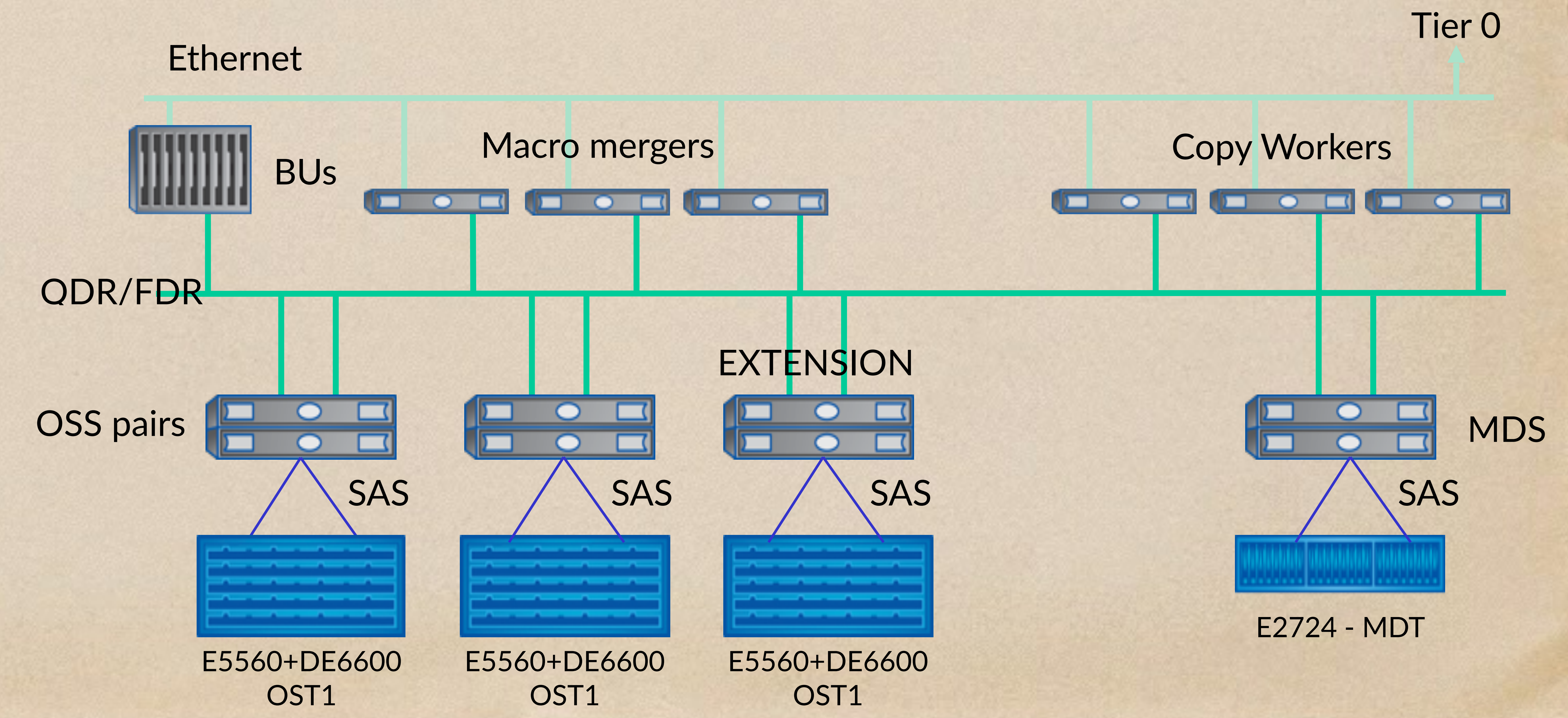
NetApp E-Series

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

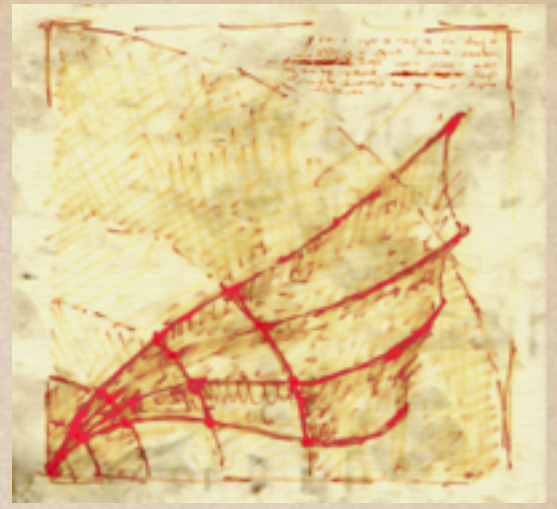
Experience

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

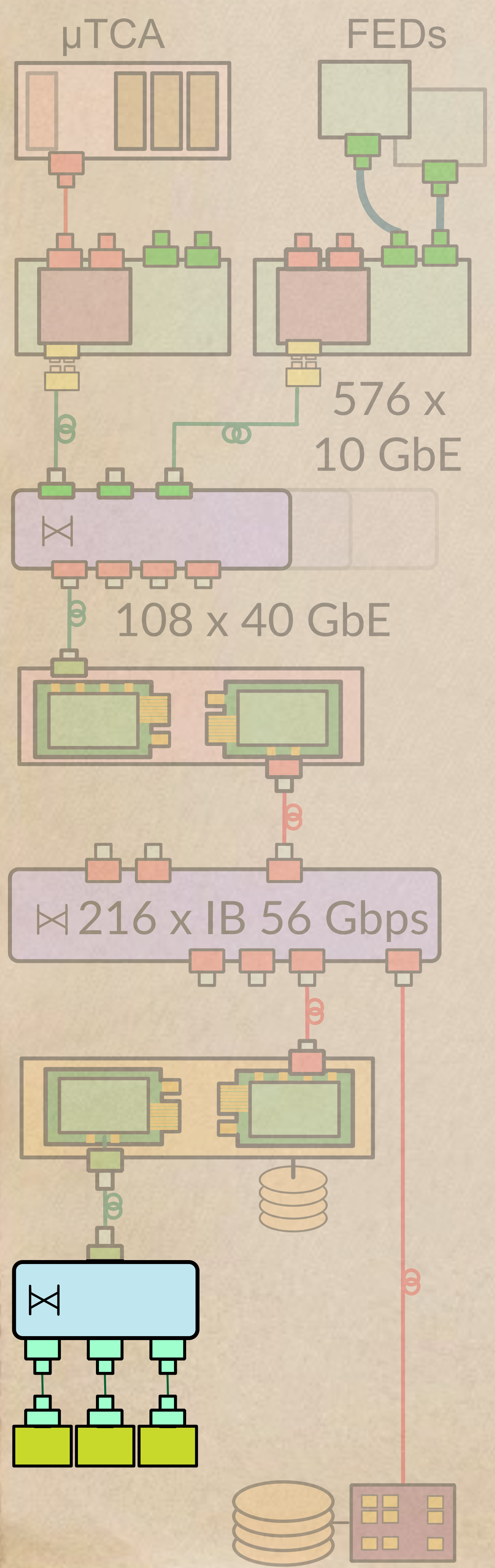




Online Cloud



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

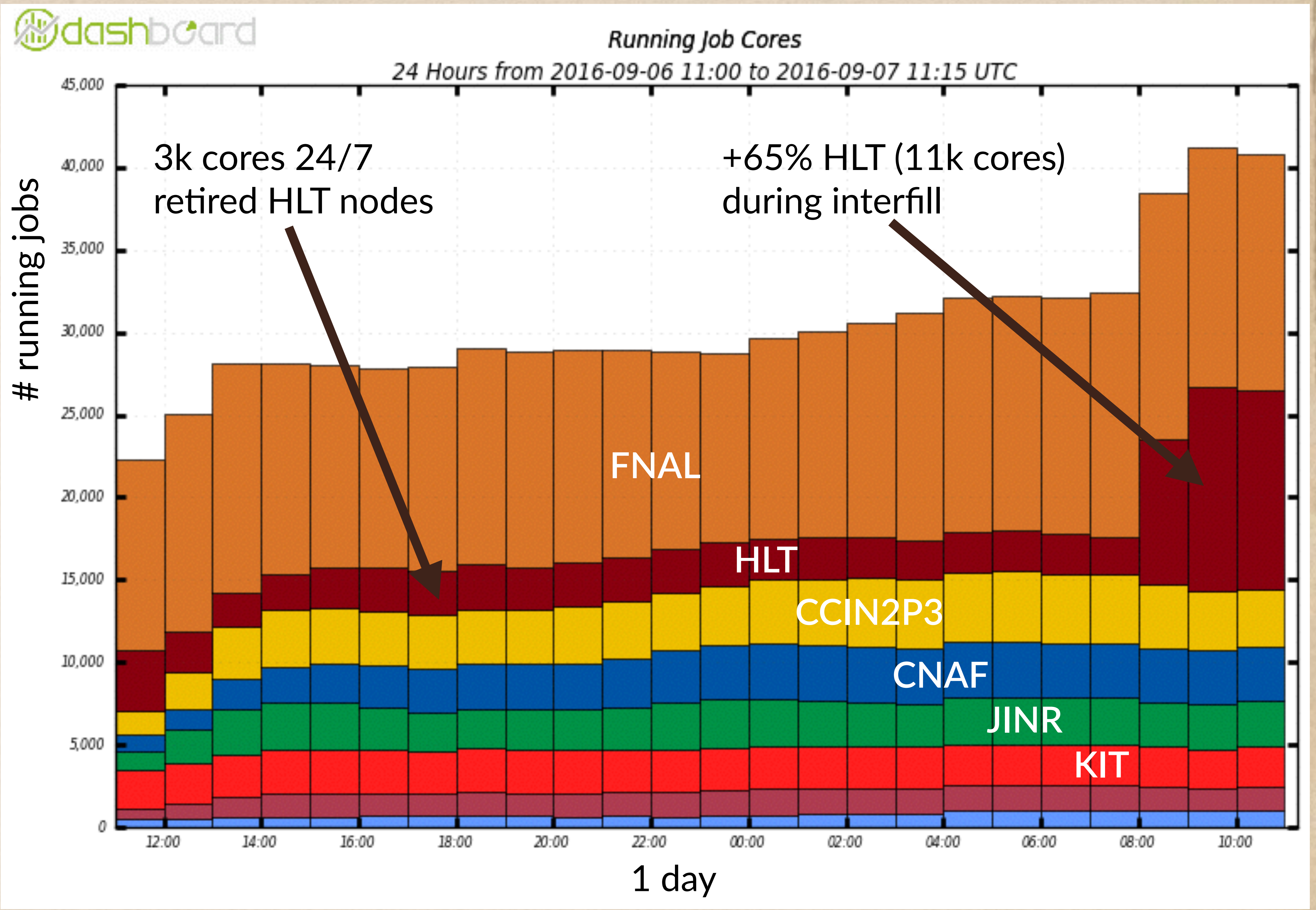


HLT computing power similar to all CMS tier 1 sites combined

- 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

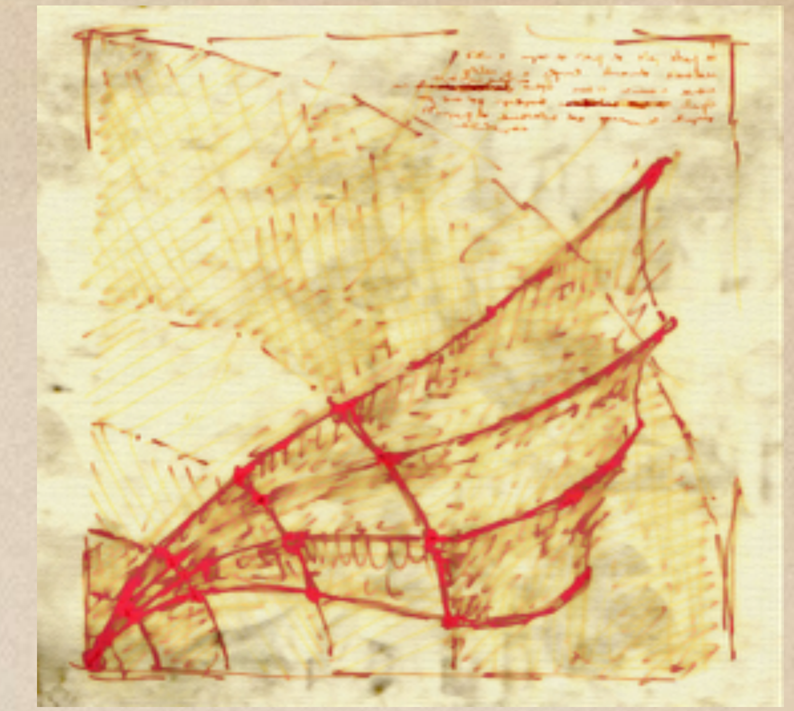


3k cores 24/7 retired HLT nodes

+65% HLT (11k cores) during interfill

Challenges

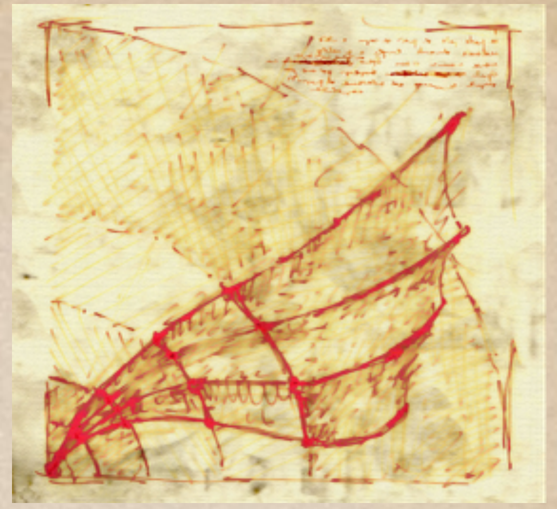
- Quickly start 800-1000 of virtual machines simultaneously
- Avoid process timeouts when hibernating VM images for several hours during data taking



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

Plans for CMS DAQ

No Radical Change for DAQ3



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

Requirements about the same as today

- Some increase in event size due to pileup & upgraded detectors

DAQ2 h/w will be at end-of-life at end of run 2 in 2019

- Need to replace computers and network infrastructure

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

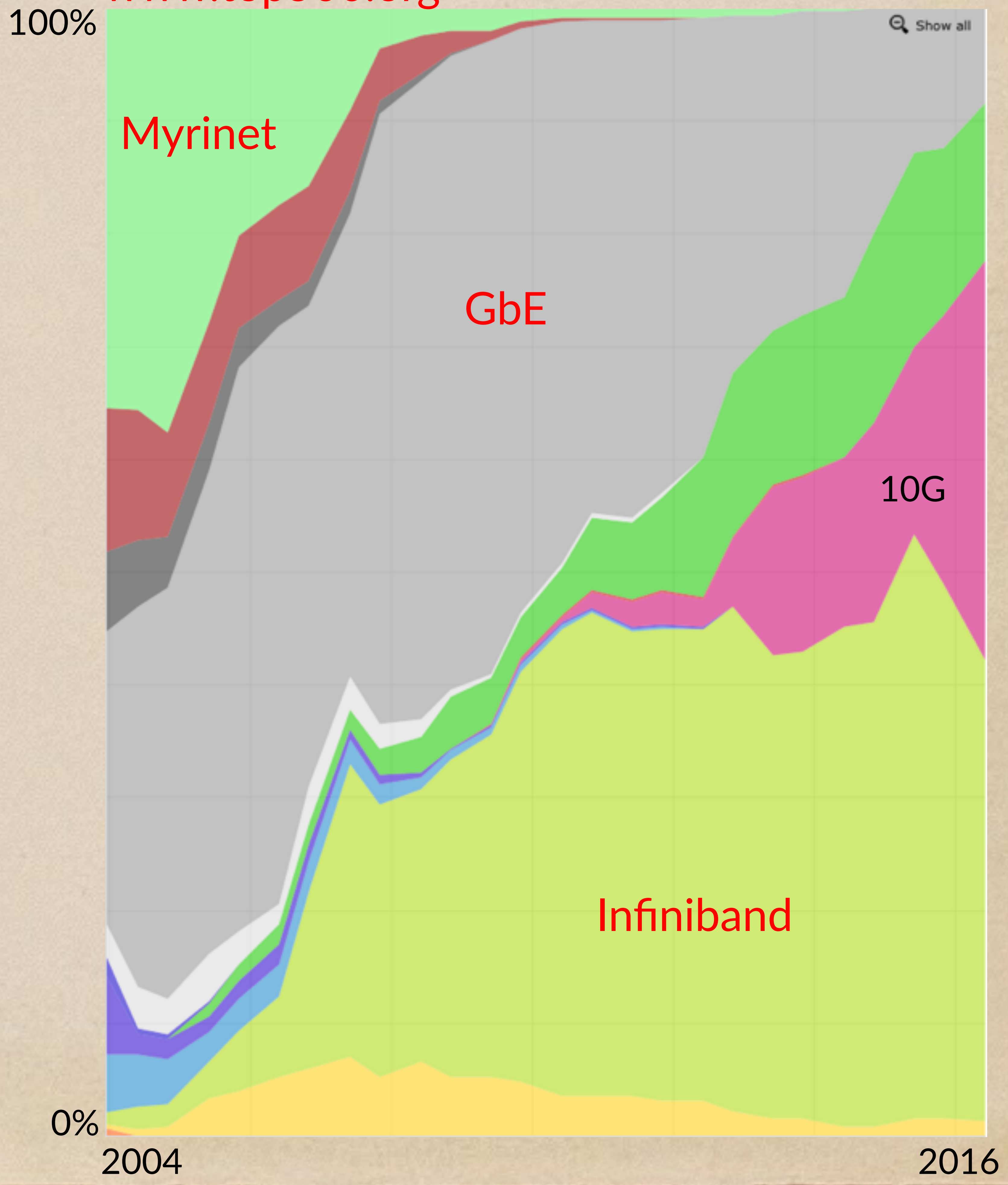
- More systems will switch to μTCA readouts
- Next generation FEROL is still using TCP/IP
- Data-concentrator network will stay on Ethernet

Need to re-evaluate event-builder network

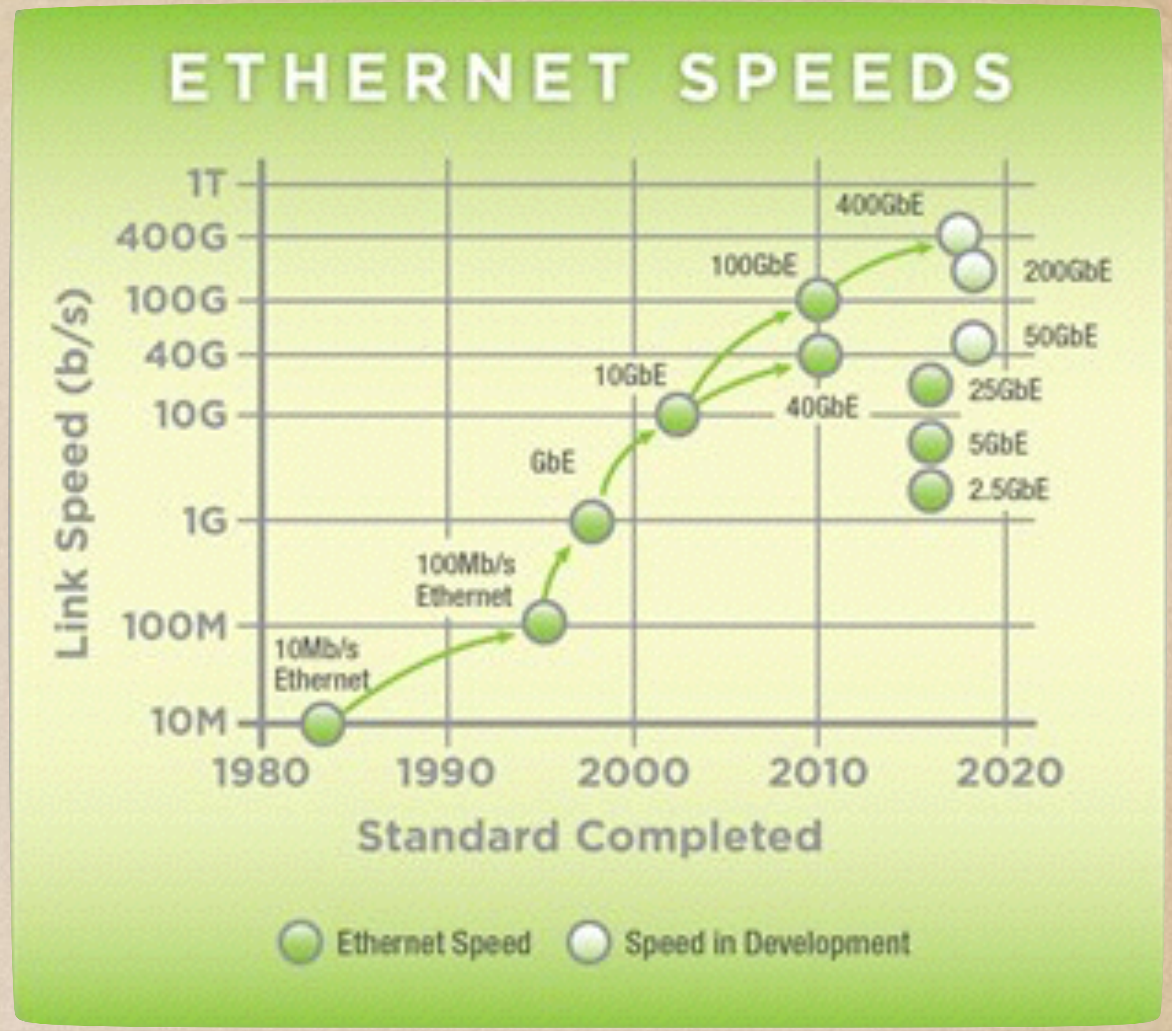
- Will Infiniband still be the most cost effective solution?
- Unlikely that there's a technology which allows to shrink the DAQ system substantially (as for DAQ2)

Take into account lessons to be learned during run 2

www.top500.org



Networking for Event-Builder

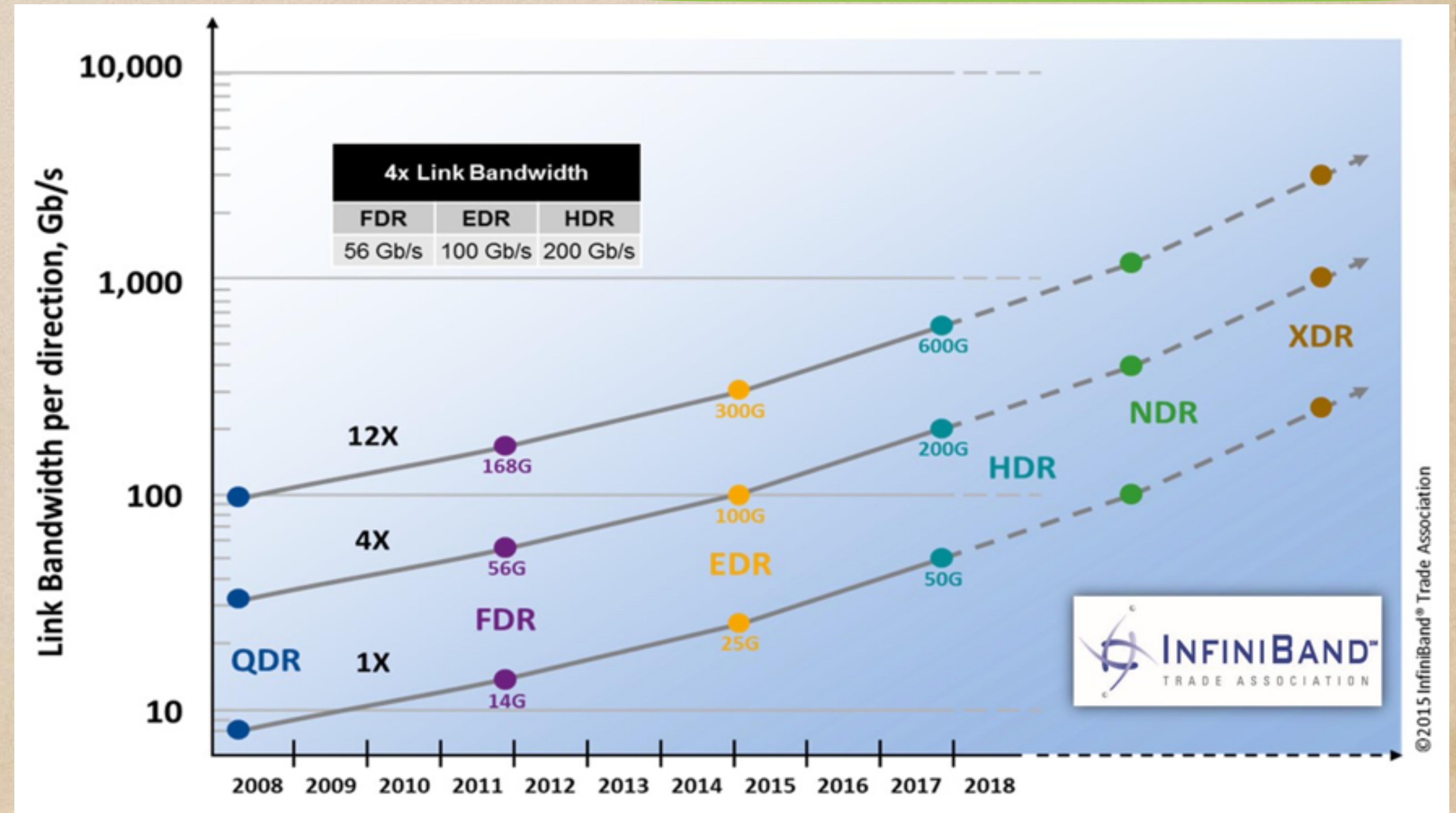


Ethernet

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

High-Performance Computing (HPC) Fabrics

- Low-latency, reliable
- Infiniband
  - 4xFDR 56 Gbps and 4xEDR 100 Gbps available
  - 4xHDR 200 Gbps (2017-18)
  - Offload network processing to NIC/switches
- Intel Omni-Path 100 Gbps
  - Integration of fabric port onto the CPU socket (onload)
  - Tight integration with specialized processors (Xeon Phi)

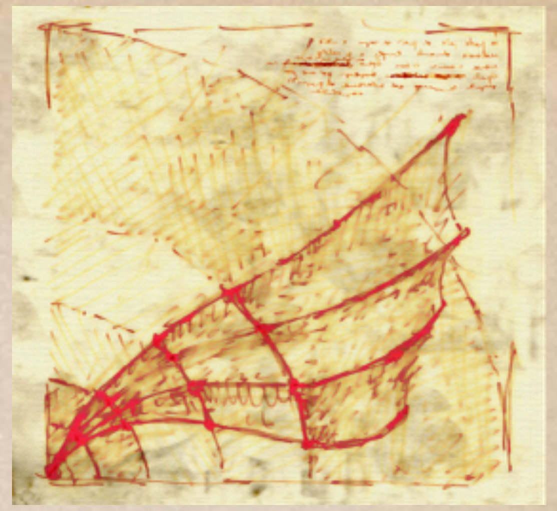


4x Link Bandwidth

FDR	EDR	HDR
56 Gb/s	100 Gb/s	200 Gb/s

Both technologies have switches with ~50 Tbp

Requirements for Phase II



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

Requirements on DAQ increase by factor ~25

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

Same two-level trigger architecture as current system

- L1 hardware trigger: 40 MHz clock driven, custom electronics
- High Level Trigger (HLT): event driven, COTS computing nodes

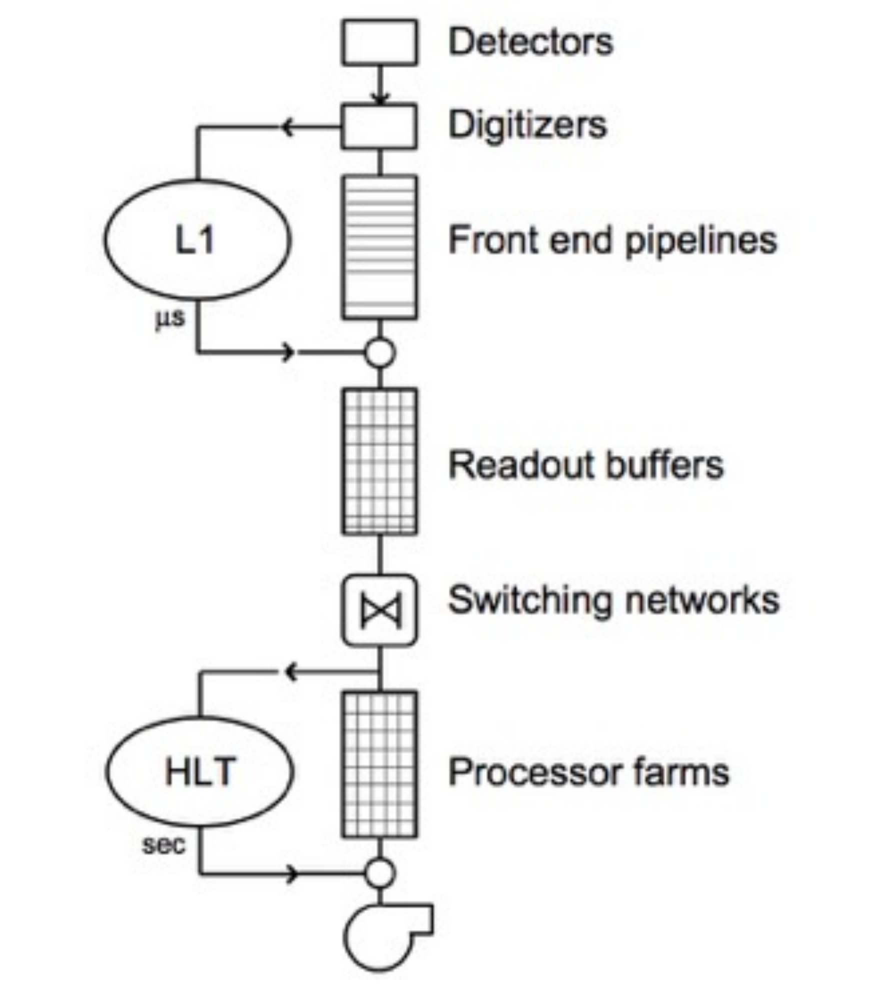
DAQ readout, network, and storage

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

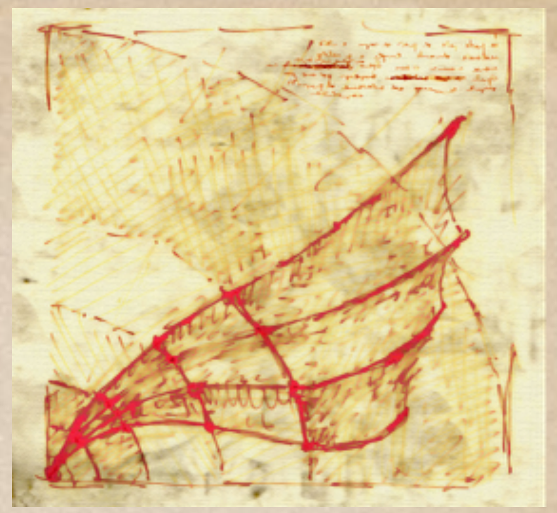
High Level Trigger with similar reduction factor as present (1/100)

- Large uncertainty from technological progress and physics requirements
- Possibly more cost effective CPU/GPU/co-processor architectures in a decade

	Run 2	HL-LHC (Phase-II)	
Peak Pile Up	**50**	**140**	**200**
Level-1 rate (kHz)	100	500	750
Event size (MB)	1.5	4.5	5.0
HLT accept rate	1	5	7.5
HLT power (MHS06)	0.5	5	11
Storage throughput (GB/s)	3	27	42



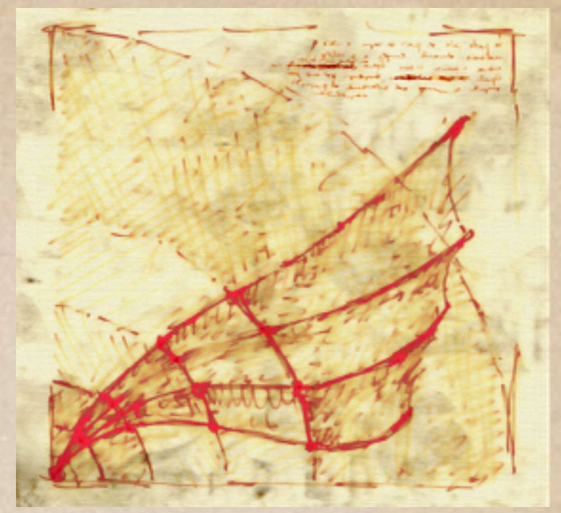
CMS Phase-II Detector Readout



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

Sub-det	# links on-/ off-detector	Type (Gbps)	use	Data reduction	Event size (MByte)	#DAQ links (100 Gbps)
Tracker-outer	13 k 2 k	GBT (4 G) GBT (9 G)	DAQ + Trig 20% + 80%	On-det	0.5 – 0.6	100
Tracker-pixel	1 k	lpGBT (9 G)	DAQ	On-det	0.7 – 1.0	200
ECAL-barrel	12 k	GBT (3 G)	streaming	Off-det	1.2	200
HCAL	2 k	GBT (3 G)	streaming	Off-det	0.2	40
HGCAL	9 k	lpGBT(9 G)	streaming?	On-det?	1.2	200
Muons DT	6 k	GBT (3 G)	streaming	Off-det	0.1	20
Muons CSC	1 k	GBT (3 G)	DAQ+Trig 50%+50%	Off-det	0.1	20
Trigger						20
EVB					***4.2-4.6***	***800***

CMS DAQ Concept



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

Sub-detector specific readout

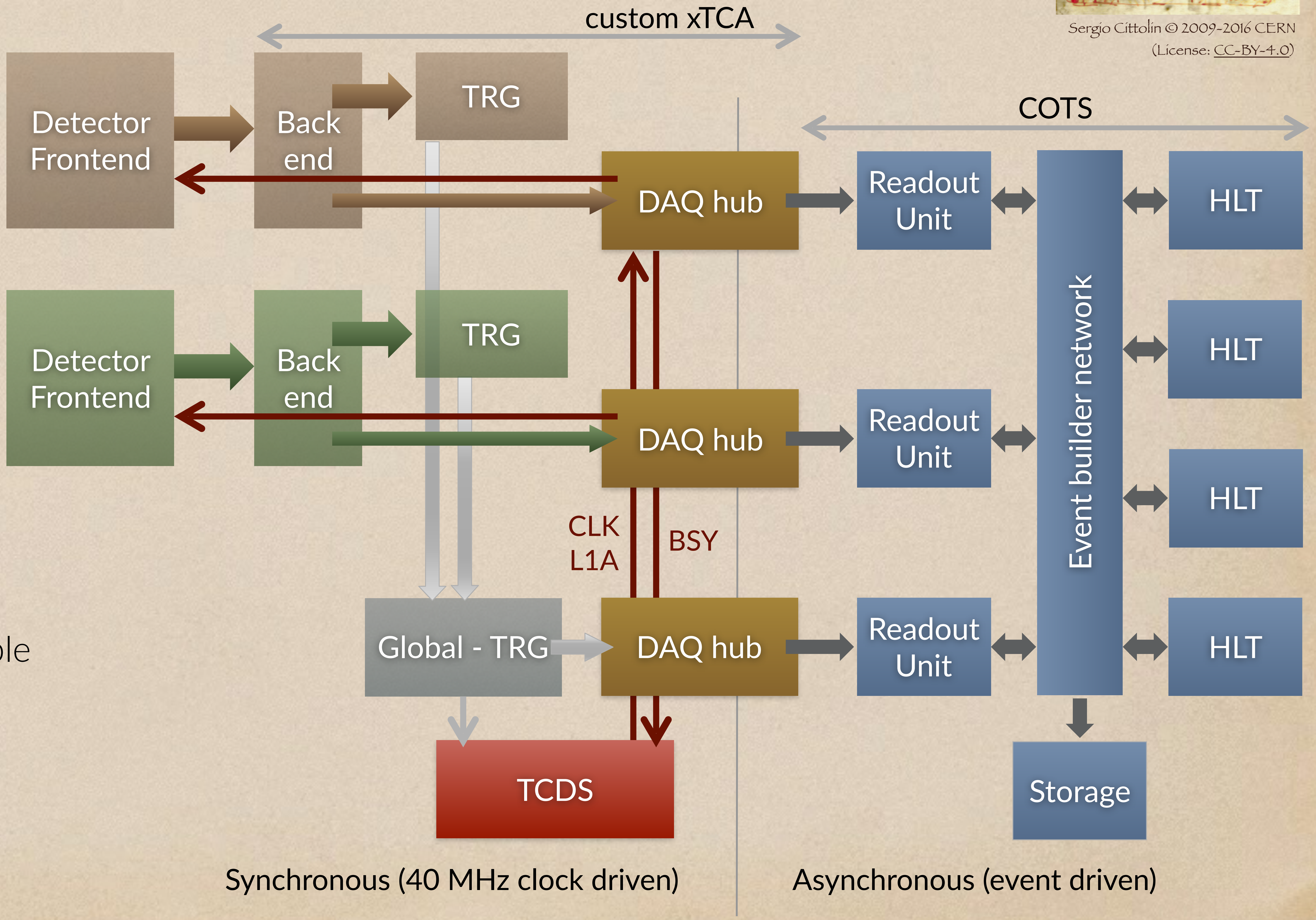
- Frontend to off-detector electronics
- Connection to L1 trigger (if any)

Interface to central DAQ system

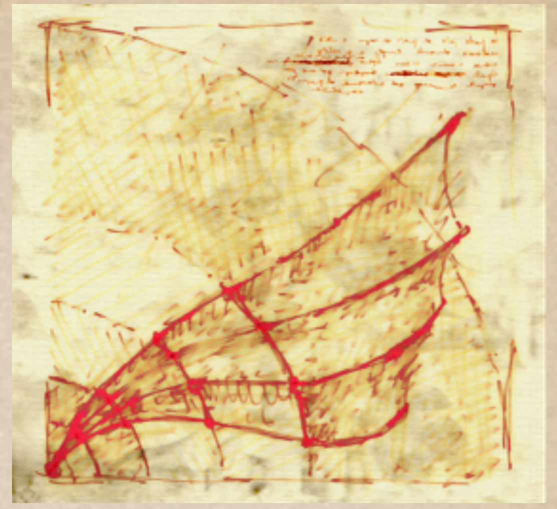
- Timing & Control Distribution System (TCDS)
  - Timing, Trigger & Control (TTC)
  - Trigger Throttling System (TTS)
- Common protocol for data readout

Central DAQ system

- FPGA based read-out board (data to surface)
  - Ethernet: Layer 2 ok, but transmission unreliable
  - (reduced) TCP/IP possible in FPGA, but needs memory for buffering
  - HPC fabric is difficult on FPGA
- Event building, HLT & storage using commercial hardware



New TCDS-DAQ Hub



rgio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

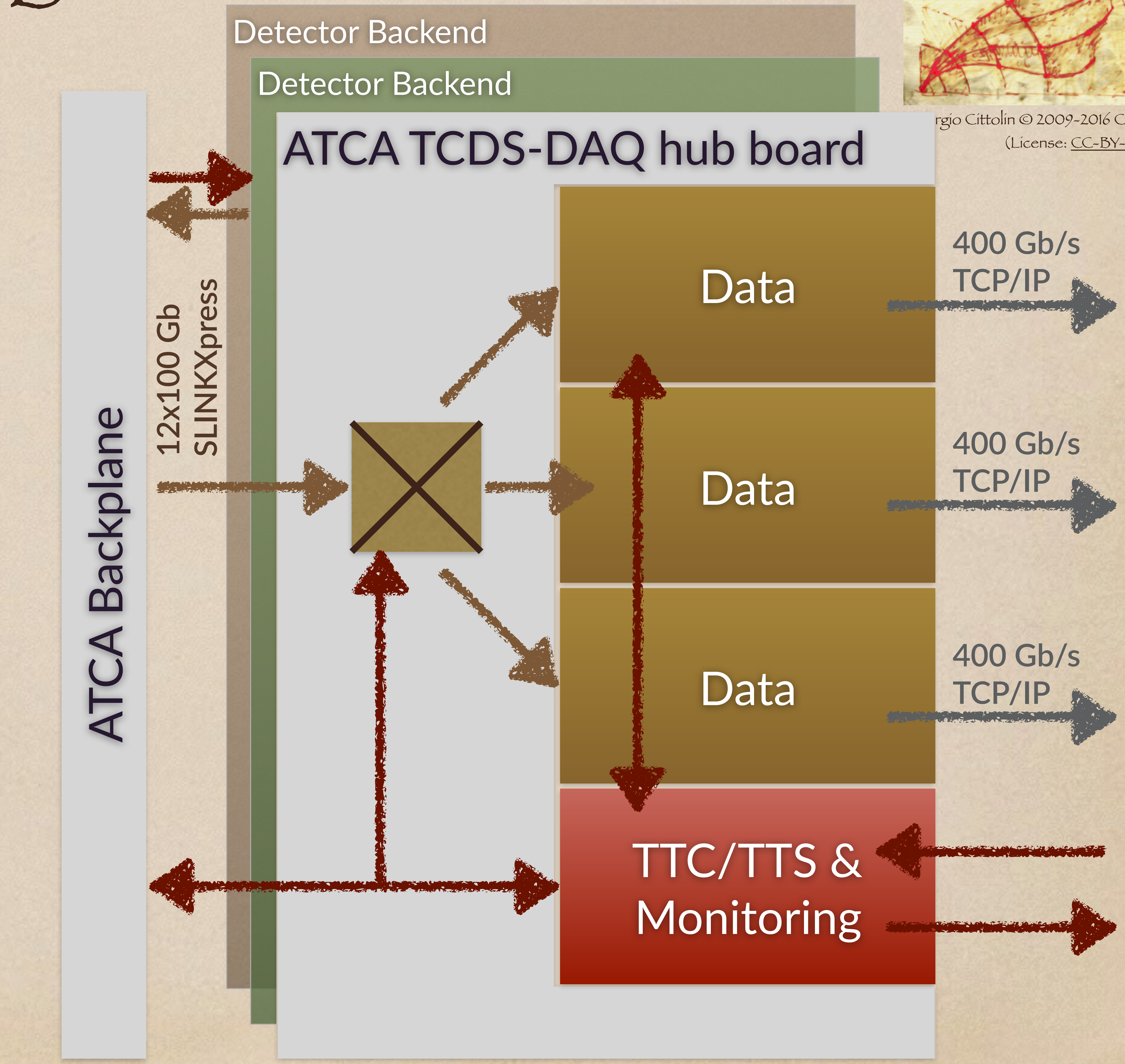
Common interface across detectors

- Between synchronous clock-driven system and asynchronous event driven
- Between custom and COTS networking/computing

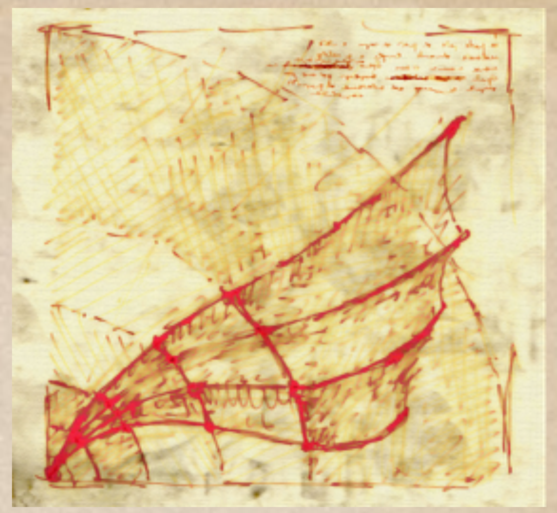
Functionality of the current AMC13 & FEROL

- Timing, trigger & control (TTC) & trigger throttling (TTS)
- Data aggregation from leaf cards in crate
  - Backplane fat pipes
  - Dedicated fibre connections (front panel or back adapter)
- Convert to commercial network & protocol
  - Most likely Ethernet with 25 Gbps or 50 Gbps lanes
  - TCP/IP would need fast buffer memory
- Emulation / data generation for testing purposes
- Improved monitoring, e.g. congestion in BE vs DAQ bp
- Full software stack for standalone & cDAQ mode

Full local DAQ system



Planning for TCDS-DAQ Hub



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

First prototype

- Develop the ATCA carrier board
- Reuse existing µTCA h/w as mezzanine boards
    - FEROL40 (DW mezzanine): 4 fat pipes & 40 GbE
    - AMC13 T1+T2 (SW,FH mezzanine)
- FPGA to act as a concentrator/router between streams from leaf cards and mezzanines
    - Both backplane & dedicated links
    - Timing & control via backplane (fat pipe?)
    - Foresee 8 fat pipes to mezzanines
    - Take FPGA which supports 25 Gbps lanes?
- MMC and system-on-a-module
- Firmware & software stack

Target for first version: mid 2018

- DAQ demonstrator
    - At least 2 leaf cards emulating sub detector data
    - COTS network & readout unit PC receiving data
- Aligned with Tracker DTC 1st prototype

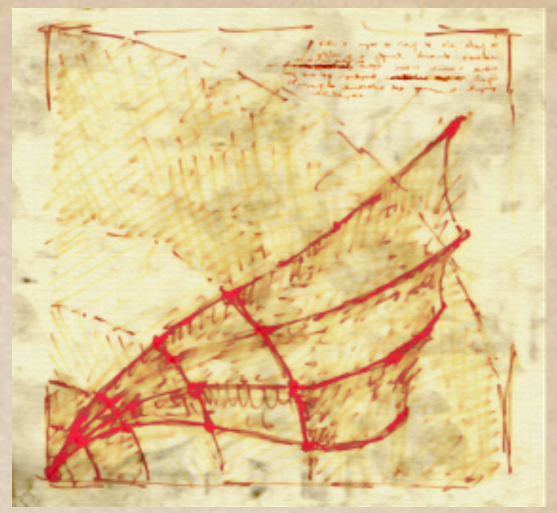
Version 2

- Keep existing TCDS-DAQ-Hub v1
- Construct FEROL100 mezzanine
    - Input 8x10 Gbps lanes
    - Output 4x25 Gbps lanes for 100 Gb Ethernet

Version 3

- ATCA card without mezzanines

Readout Unit



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

Commercial PC on surface

- Receiver from synchronous system
  Commercial NIC (Ethernet L2), or custom card
- Data concentrator and temporary buffer
- Protocol converter to event-builder network
  (Ethernet or HPC interconnect)

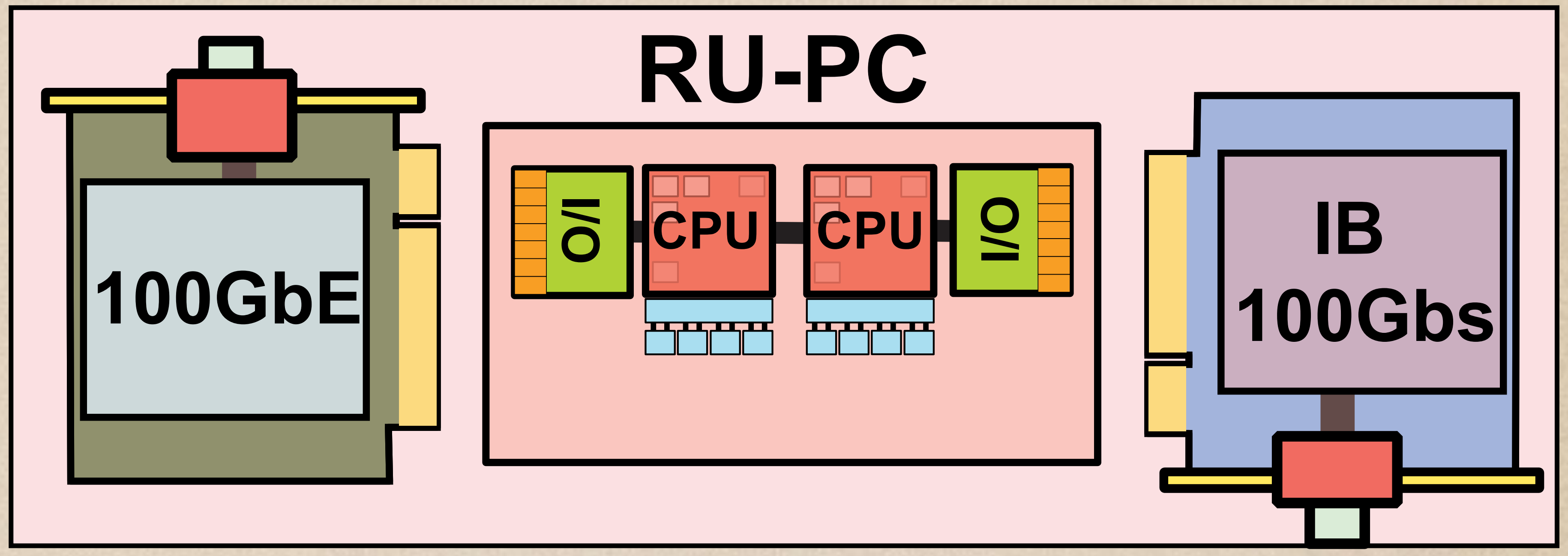
CPU power for

- Link protocols
- Event-fragment checks
- Super-fragment building

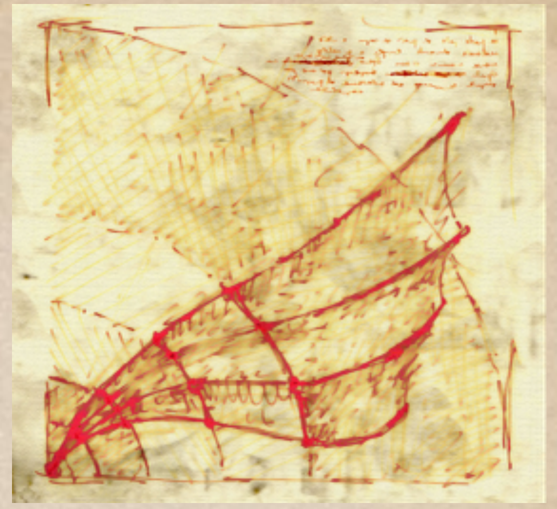
Memory to buffer O(100) event fragments

Would need O(1000) machines with 100 Gbps I/O

- Could be done today, albeit prohibitively expensive



Industry Trends — CPU



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

Hybrid CPU-FPGA system

- FPGA offload with OpenCL
- Prototypes for categorization and search (e.g. Google)

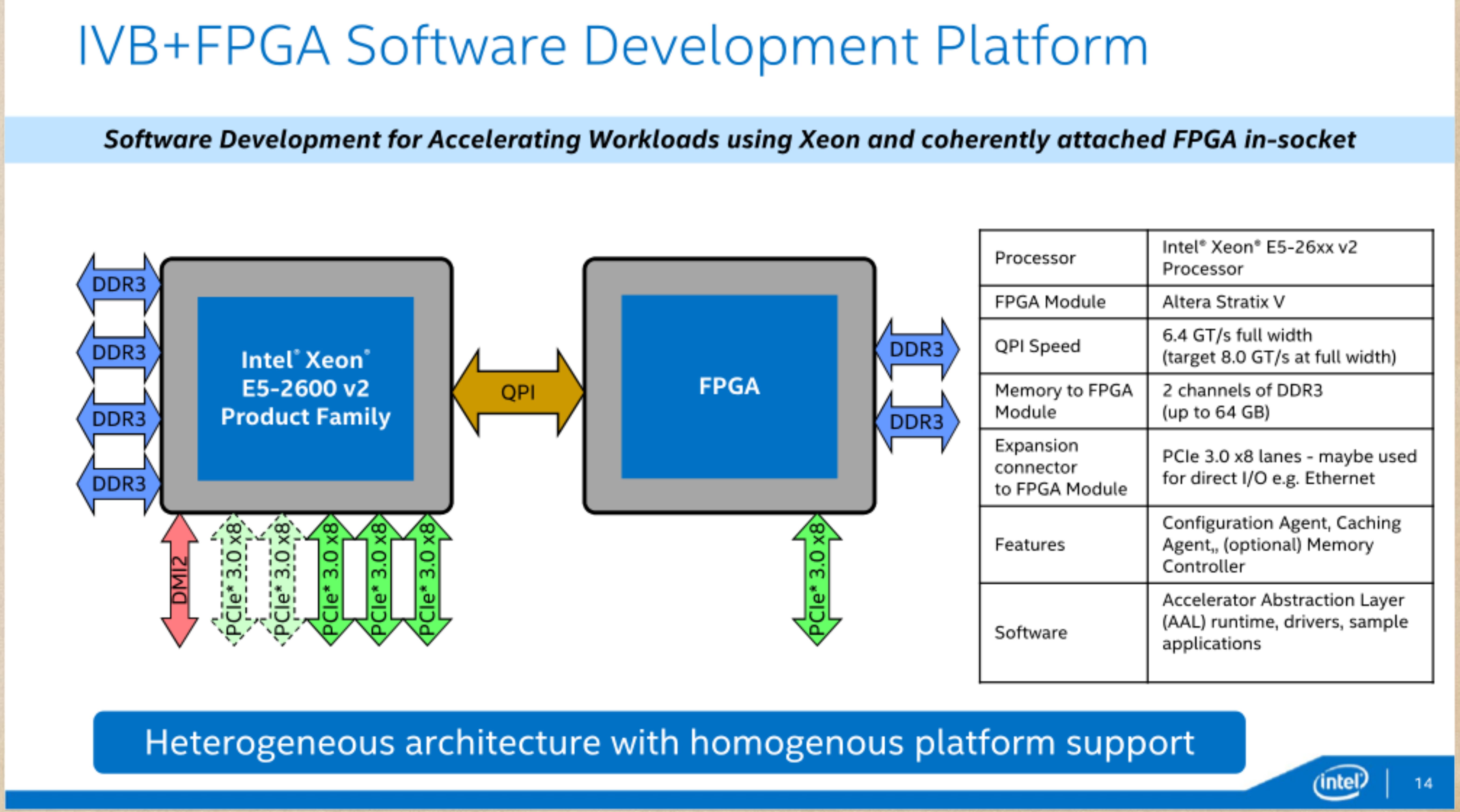
On-chip fabric

- Intel OmniPath
- Reliable and scalable 100 Gbps interconnects

Specialized many-core processors (Xeon Phi, GPUs, …)

- Many cores with wide vector units
- Network fabric will be integrated in the future

Gen4 PCIe (~16 GT/s/lane)



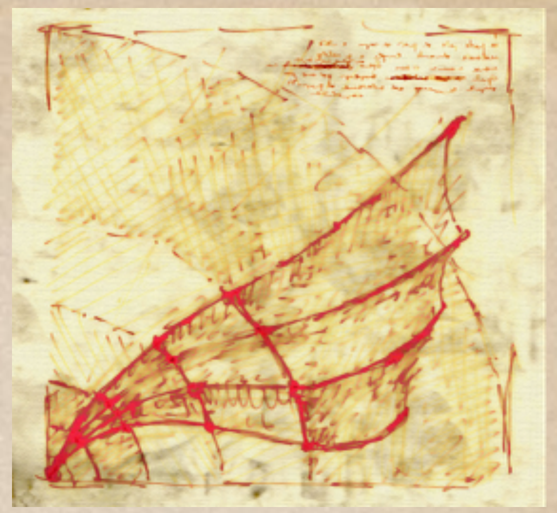
IVB+FPGA Software Development Platform

Software Development for Accelerating Workloads using Xeon and coherently attached FPGA in-socket

Processor	Intel® Xeon® E5-26xx v2 Processor
FPGA Module	Altera Stratix V
QPI Speed	6.4 GT/s full width (target 8.0 GT/s at full width)
Memory to FPGA Module	2 channels of DDR3 (up to 64 GB)
Expansion connector to FPGA Module	PCIe 3.0 x8 lanes - maybe used for direct I/O e.g. Ethernet
Features	Configuration Agent, Caching Agent,, (optional) Memory Controller
Software	Accelerator Abstraction Layer (AAL) runtime, drivers, sample applications

Heterogeneous architecture with homogenous platform support

Industry Trends — Memory



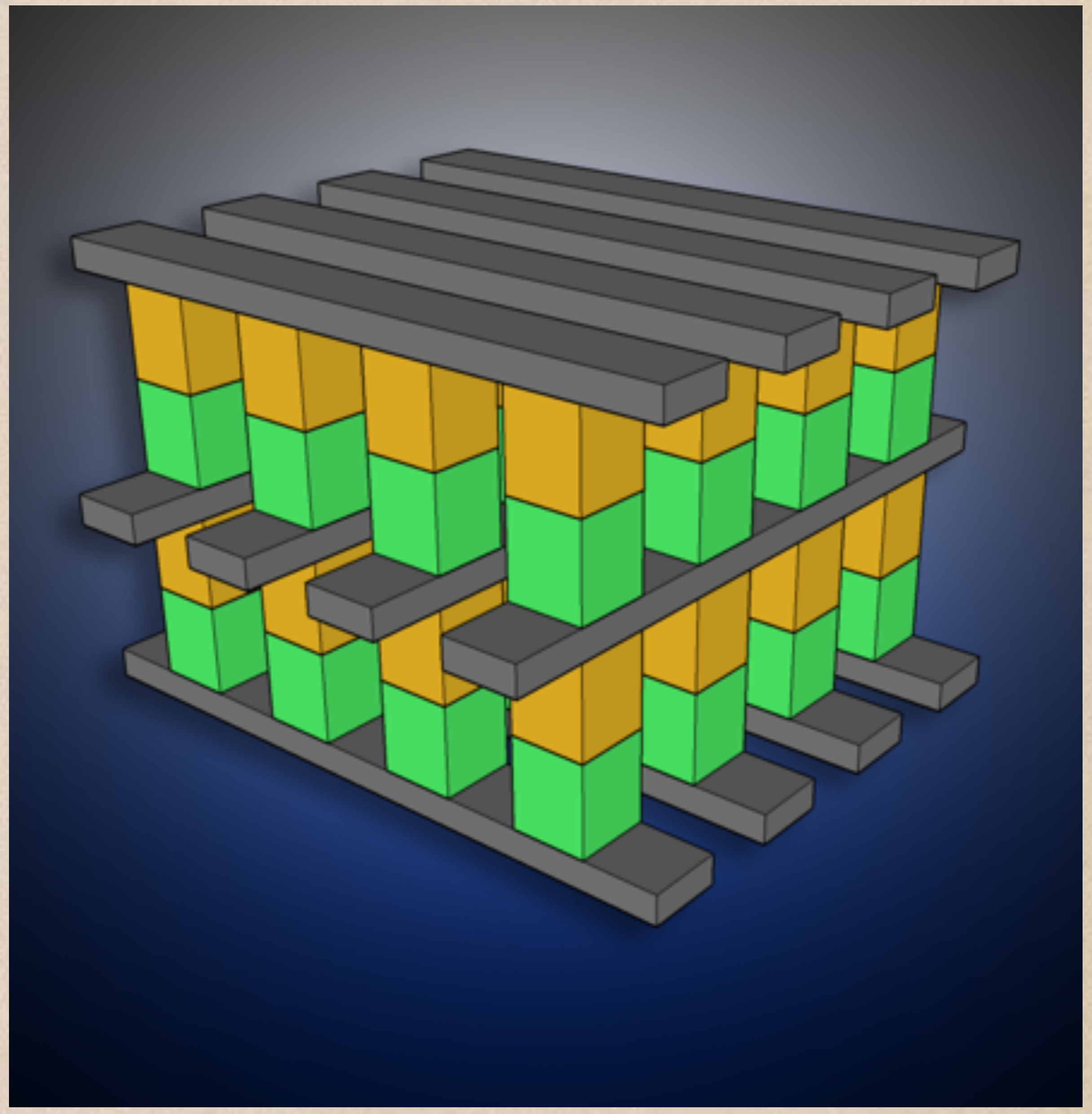
Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

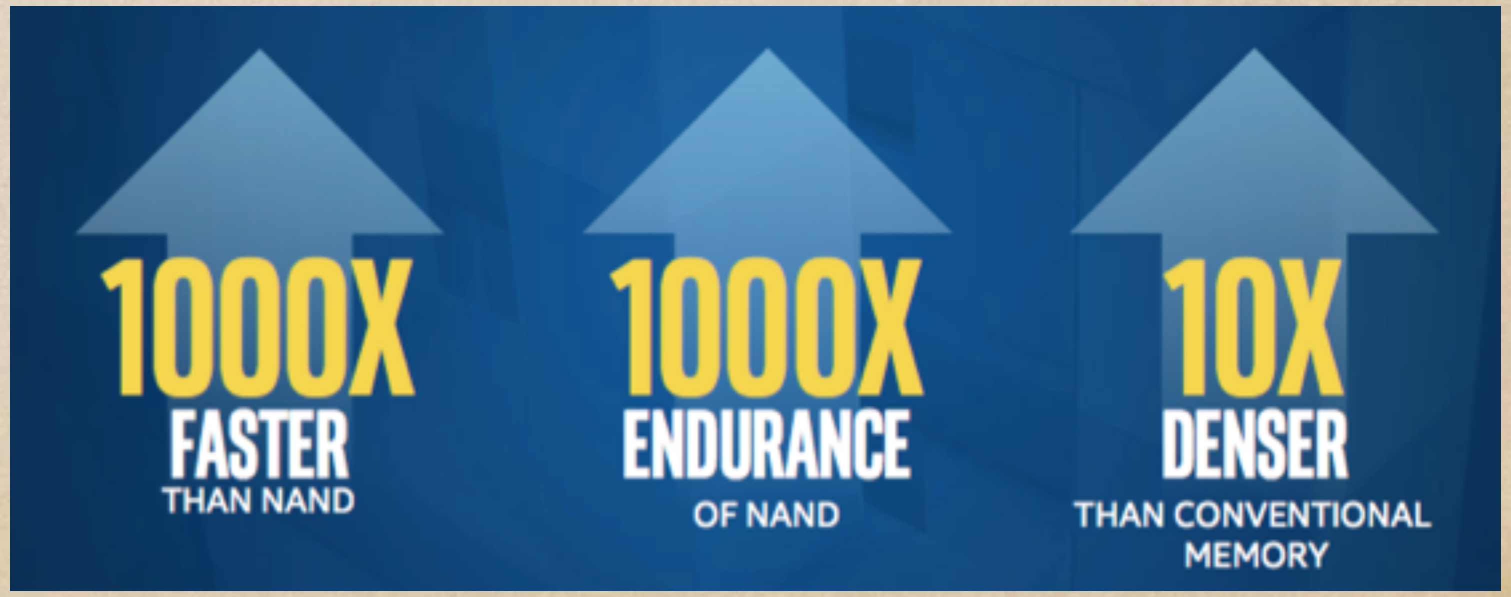
3D XPoint memory (Intel & Micron)

- Non-volatile memory
- Faster than NAND SSD
- Denser than DRAM

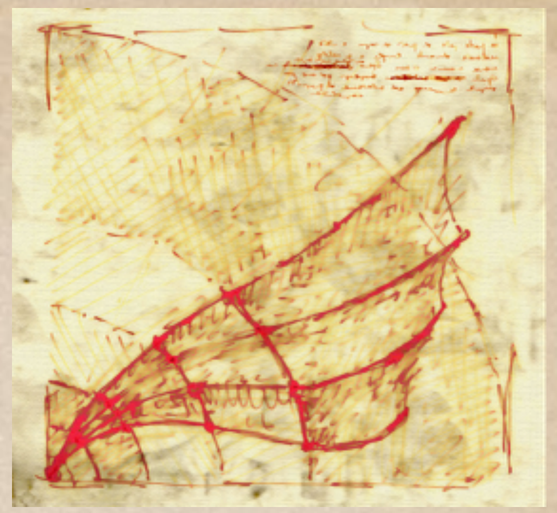
Allows unprecedented addressable storage capacity with low-latency

- Data can be buffered for much longer times than today
- No risk of data loss due to power failures





Readout Processor



Commercial PC on surface

- Receiver from synchronous system
  Commercial NIC (Ethernet L2), or custom card
- Data concentrator and temporary buffer
- Protocol converter to event-builder network
  (Ethernet or HPC interconnect)

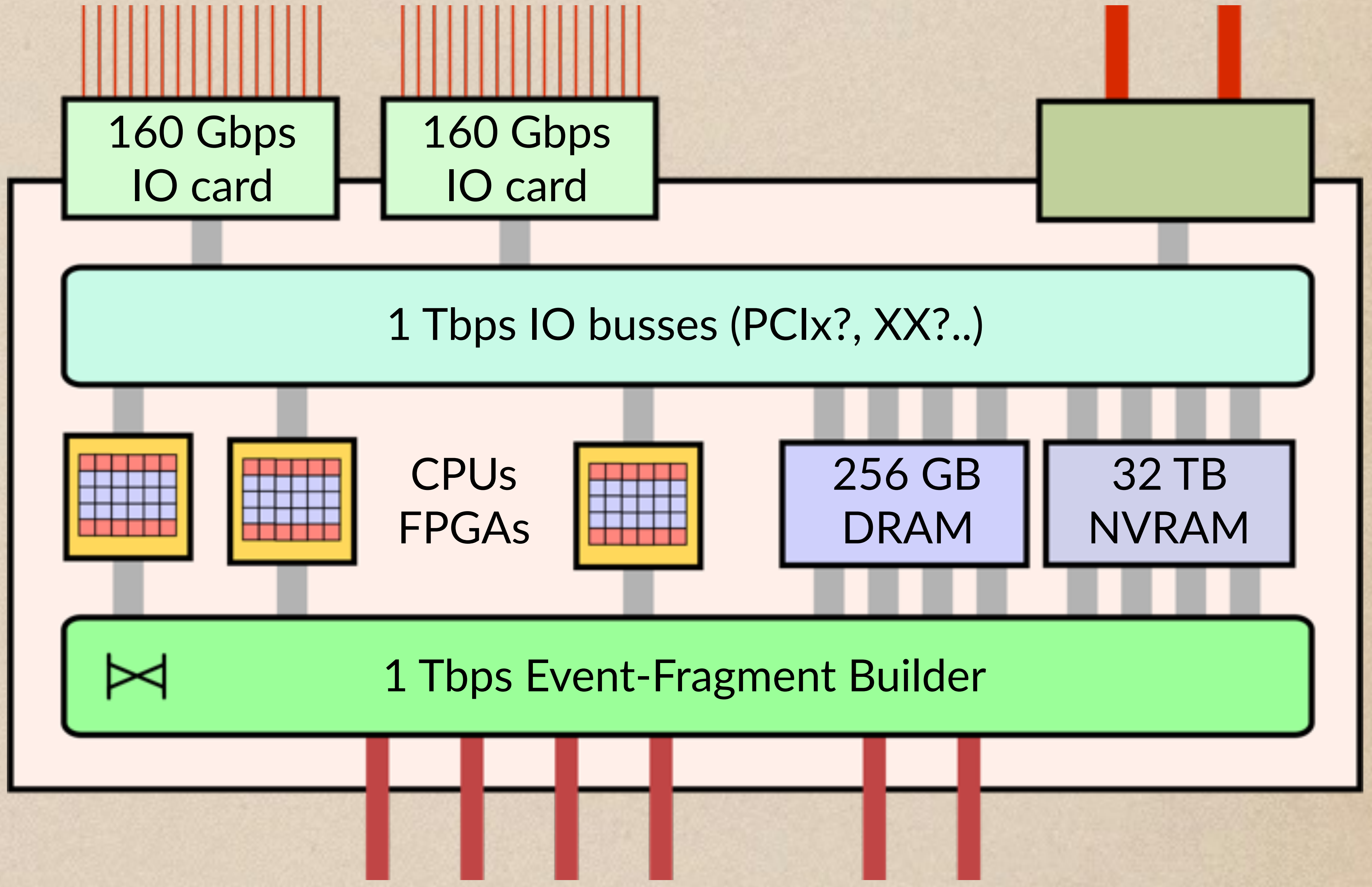
CPU power for

- Link protocols
- Event-fragment checks
- Super-fragment building
- Pattern recognition, preprocessing or fast calibration
- Event classification
- Detector data monitoring

Large, non-volatile memory

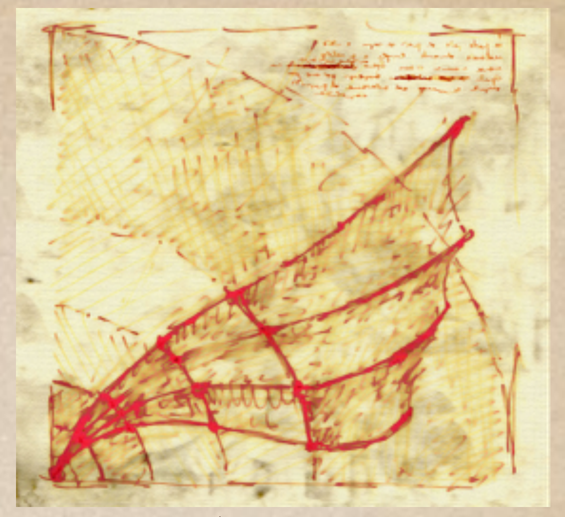
- Detector calibration before HLT
- Store data until needed for HLT selection

up to 1 Tbps data readout

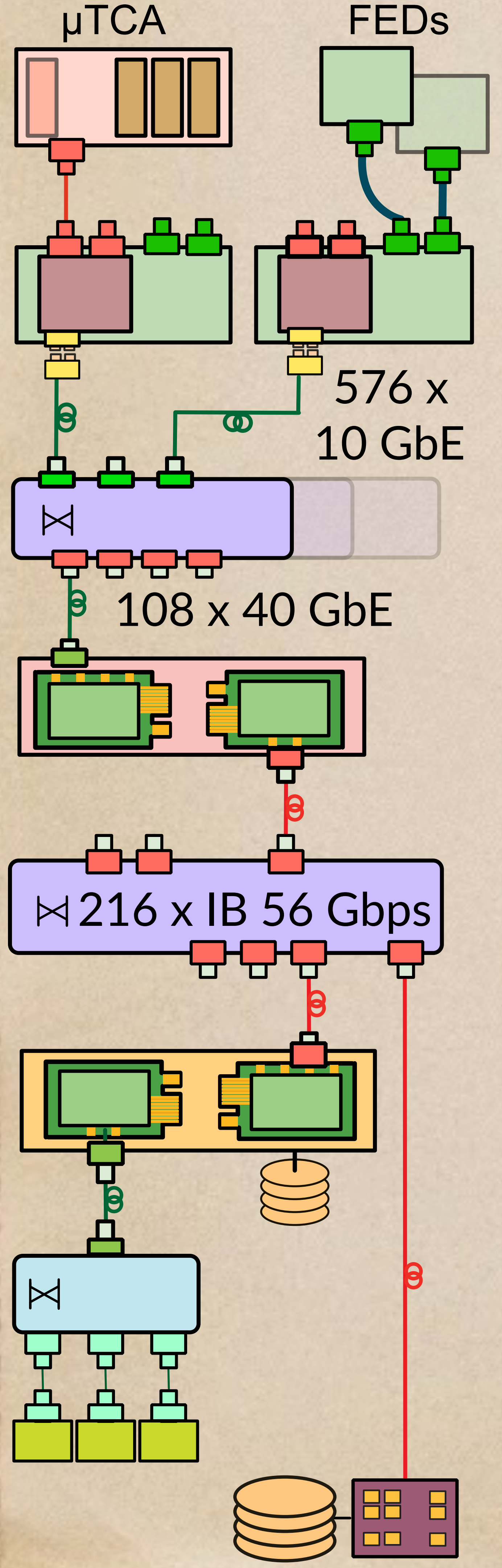


up to 1 Tbps event-fragment output

Alternative Event-Builder Architectures



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)



Folded event builder

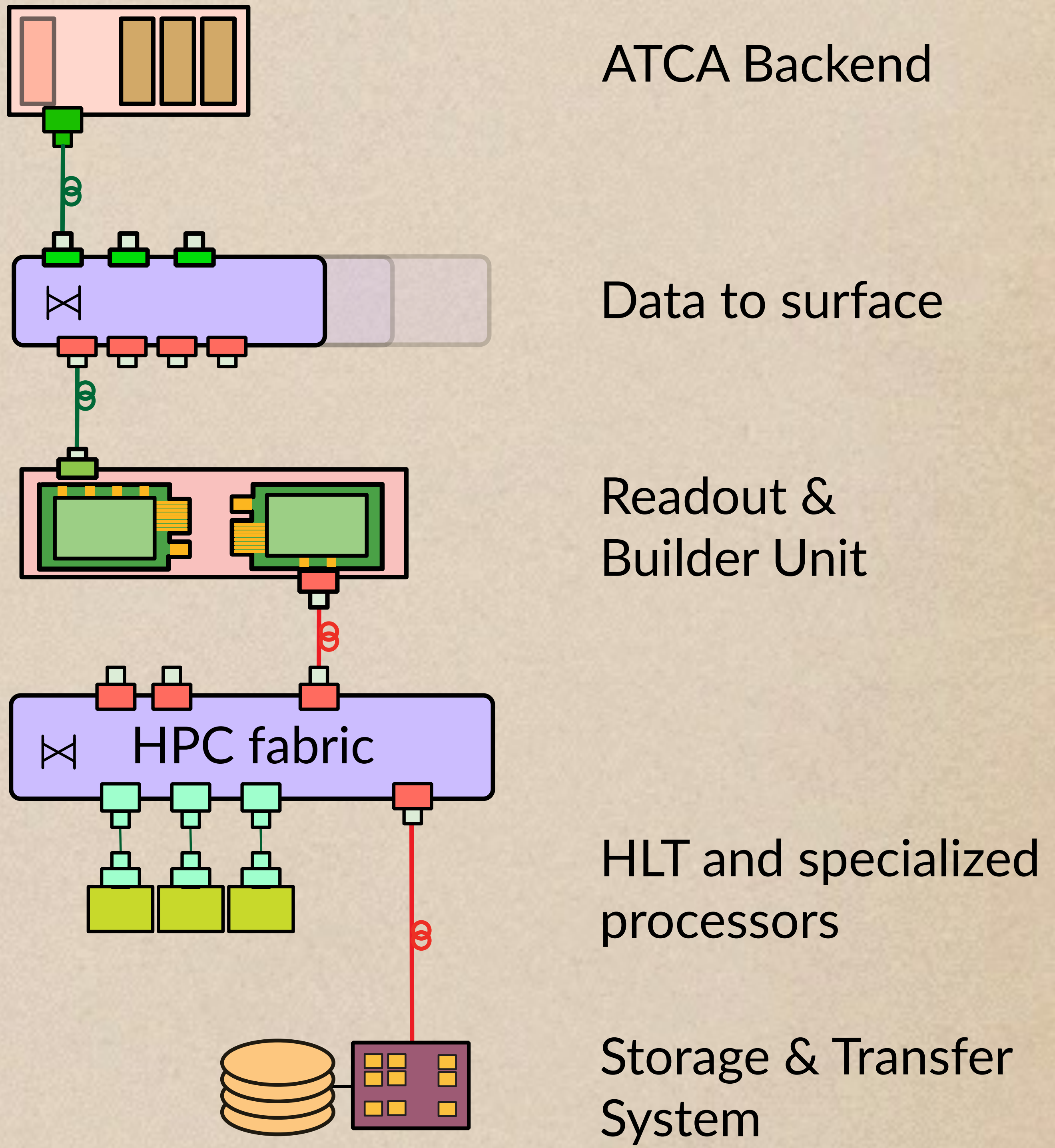
- Same event-builder node used for detector read-out and full event building
- Exploit bi-directional links
- Traffic balancing becomes more challenging
- Higher demand on I/O and memory performance

Data federations & event building on demand

- Each readout unit holds data from pre-defined set of detector
- Data is accessed directly from HLT process through HPC fabric
  - Complete events are built during or after HLT processing
  - Requires support from s/w framework to deal with partial events and longer latencies to access sub-detector data

Co-processor farms

- Specialized processors for feature extraction might be beneficial
- Access data from readout unit and provide information to HLT nodes



High-Level Trigger

Guestimate based on current detector, PU dependence, assuming current algorithms scale

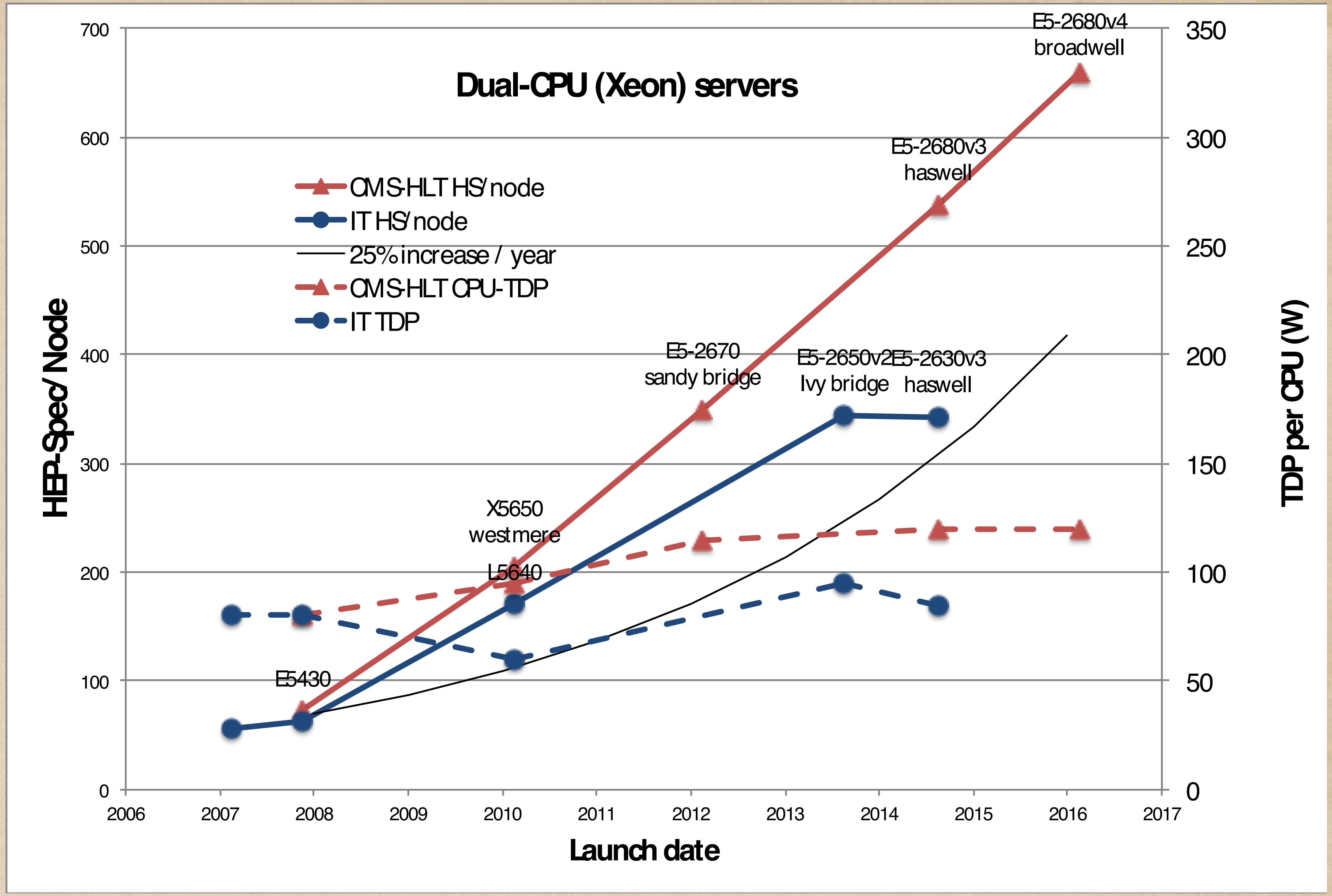
- Possible gain with using L1 Track trigger
- PU = 140 / 200, L1 Rate 500 / 750 kHz:
  ~ 5.0 / 11.0 MHS06
- Compare LHCb: need 3.3 MHS06 in 2021

Existing data center at pt.5 supports 1 MW

- Upgrading in place is difficult & expensive

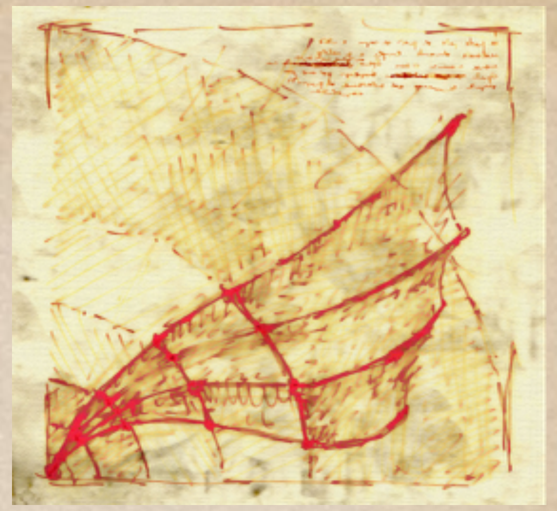
New data center might be required

- New building at pt.5
- Remote data center at Prévessin for all LHC exp.
  - ~4 TB/s over ~10 km
  - Costs for data links might be prohibitive



Assumed perf. increase of server	Exponential 25% / year [WLCG 2014]	Exponential 12.5% / year	Linear 73 HS / year
11 years	12	3.7	2.2
#servers in Q1-27	1431	4562	7860
Total power Q1-27	0.5 MW	1.6 MW	3 MW

Possible Mitigations for HLT



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

Co-processor farms

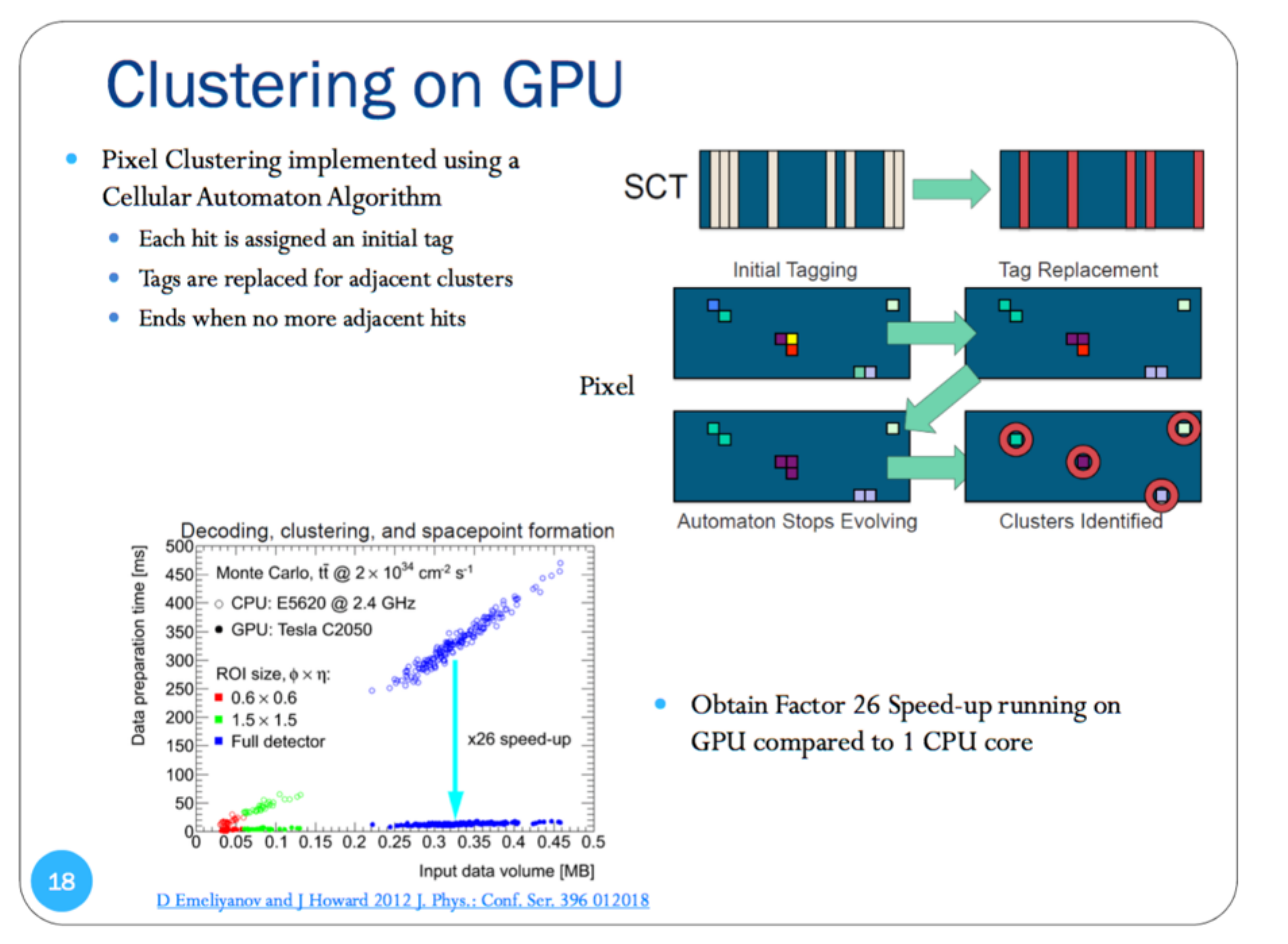
- GPU or Xeon Phi for specialized tasks
- Need to study data movement overhead

Hybrid CPU-FPGA system

- Use OpenCL to make code portable to normal CPUs

Truly distributed local processing

- Exploiting high-performance fabric
- Container & query programming style leveraging large non-volatile memory



Clustering on GPU

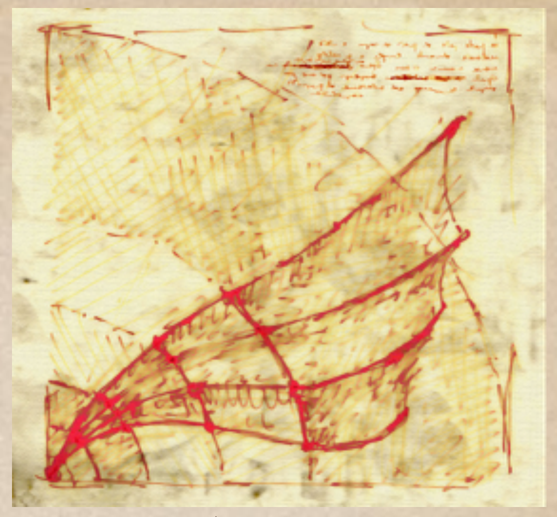
- Pixel Clustering implemented using a Cellular Automaton Algorithm
  - Each hit is assigned an initial tag
  - Tags are replaced for adjacent clusters
  - Ends when no more adjacent hits
- Obtain Factor 26 Speed-up running on GPU compared to 1 CPU core

SCT
Initial Tagging
Tag Replacement
Pixel
Automaton Stops Evolving
Clusters Identified
Decoding, clustering, and spacepoint formation
Data preparation time [ms]
500
450
400
350
300
250
200
150
100
50
0
Monte Carlo, $t\bar{t}$ @ 2 × 10^{34} cm^{-2} s^{-1}
○ CPU: E5620 @ 2.4 GHz
● GPU: Tesla C2050
ROI size, $\phi \times \eta$:
0.6 × 0.6
1.5 × 1.5
Full detector
x26 speed-up
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Input data volume [MB]

D Emeliyanov and J Howard 2012 J. Phys.: Conf. Ser. 396 012018

18

Scouting for New Physics



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

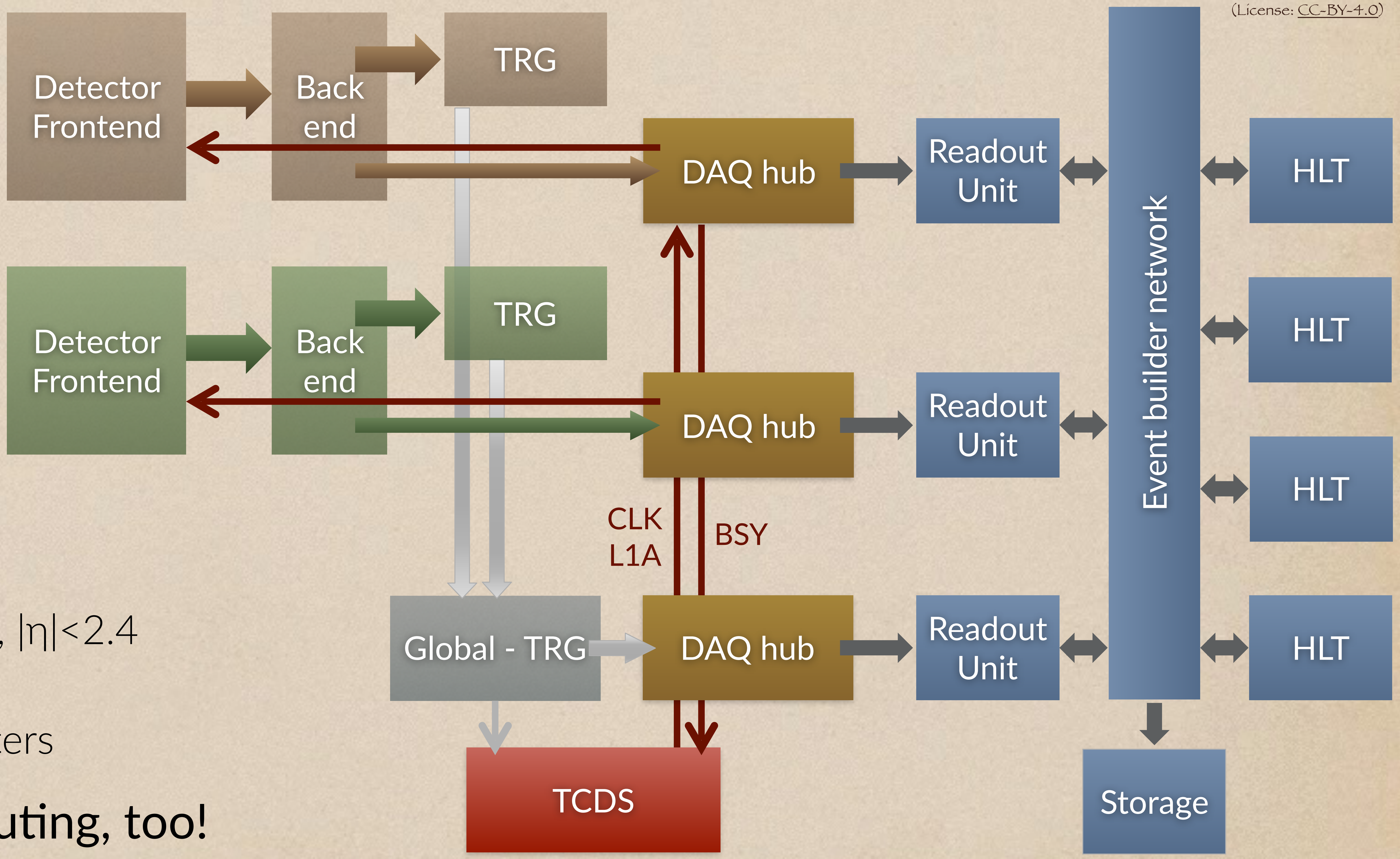
Current scouting technique on HLT

- Signatures with acceptable rates at L1, but without substantial rate reduction at HLT
- Simple analysis on HLT using HLT objects with reduced accuracy (e.g. CaloJets)
- Store minimal information to find interesting features offline

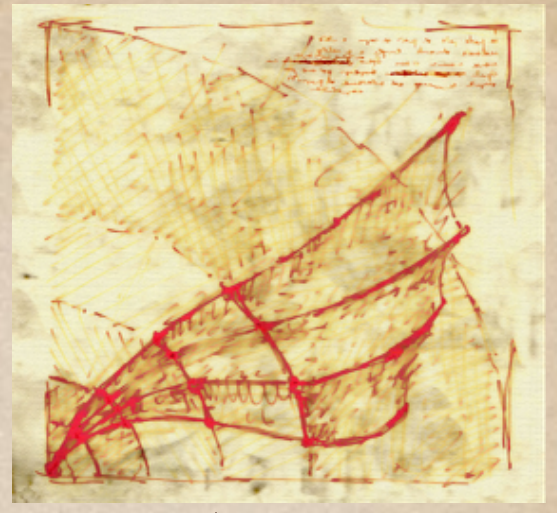
In phase II most detectors provide data for each bunch crossing at 40 MHz

- Track trigger data, all tracks down to Pt>2GeV, |η|<2.4 with high efficiency
- Triggerless streaming of data e.g. for calorimeters

Trigger information could be used for scouting, too!



Scouting at 40 MHz (Emilio Meschi)



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

Extend scouting technique to L1

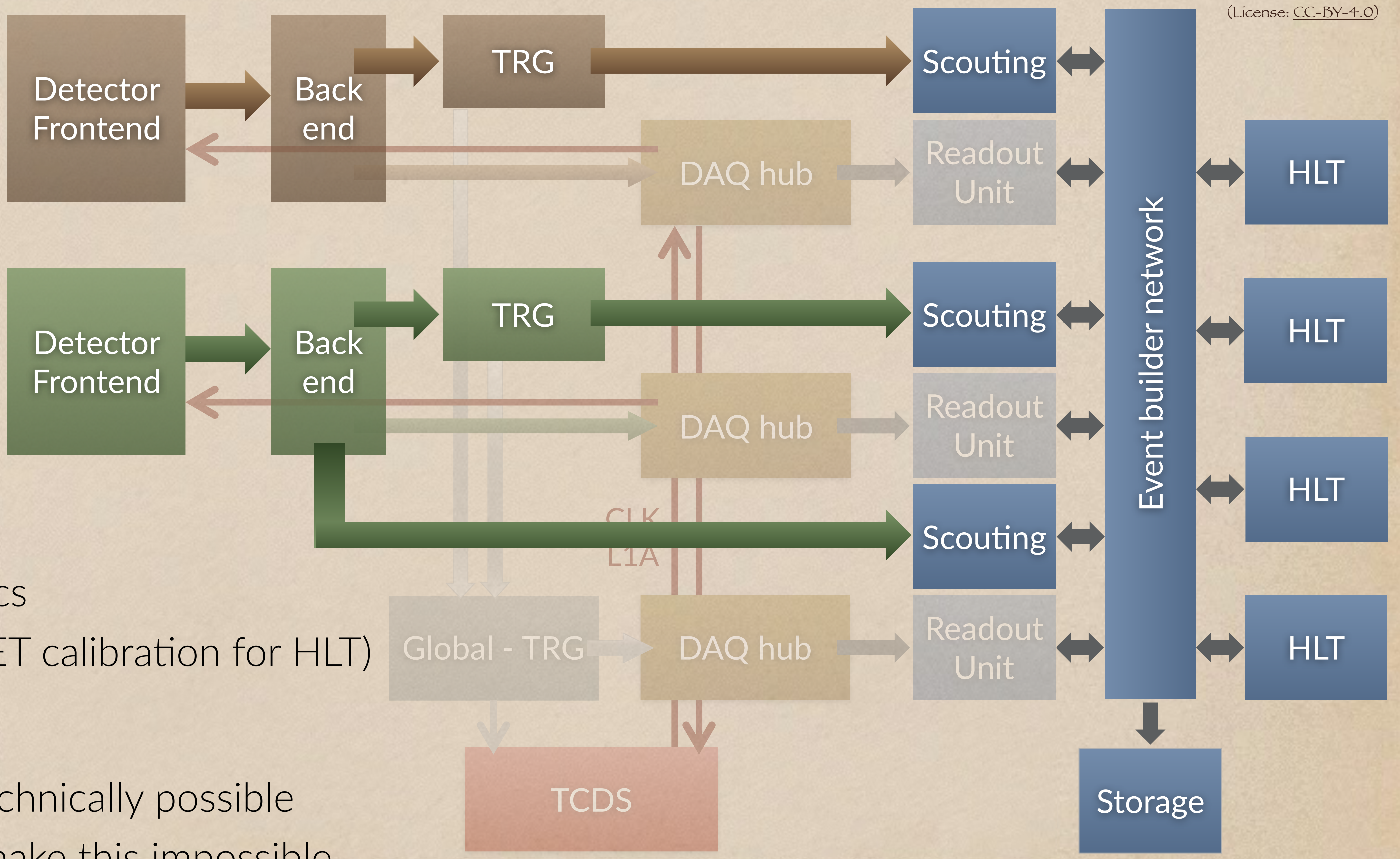
- Collect data available at 40 MHz
  - Parasitic DAQ' system w/o backpressure
  - Data with reduced accuracy and/or content
- Look for signatures with too high L1 rate
  - Dedicated scouting processors
  - Analyze with spare HLT power

High-statistics real-time data analysis

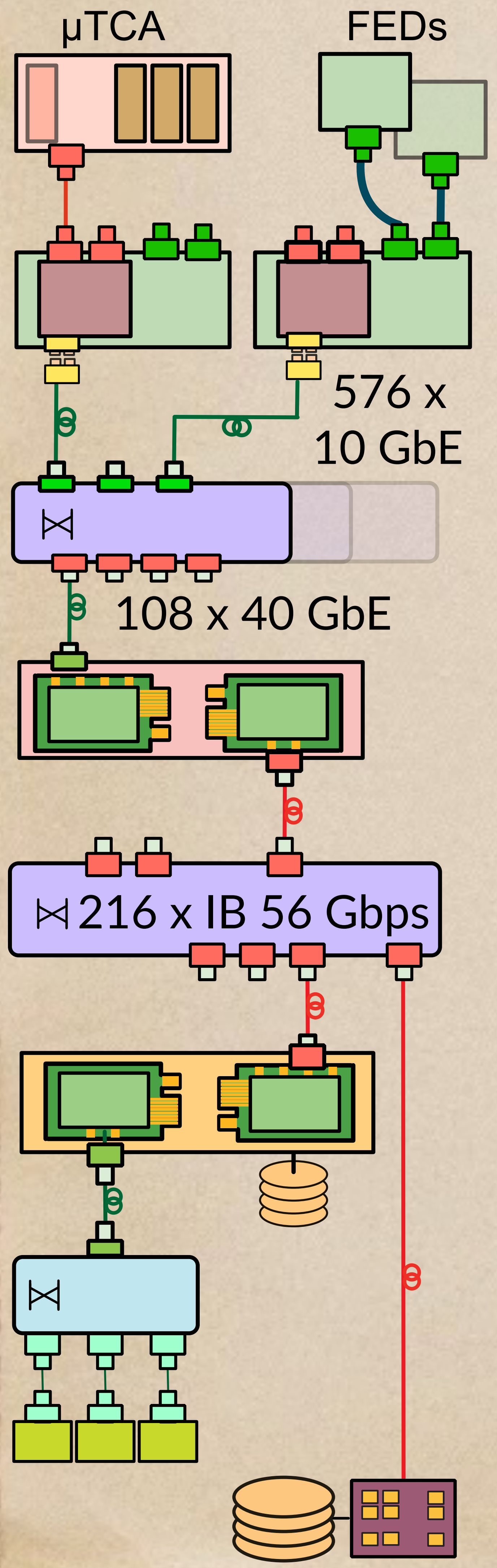
- Understand physics limited by L1
- Rapidly attain and monitor best calibration of (e.g. L1) quantities that require high statistics
- Calibrations in real time (e.g. fast, accurate MET calibration for HLT)

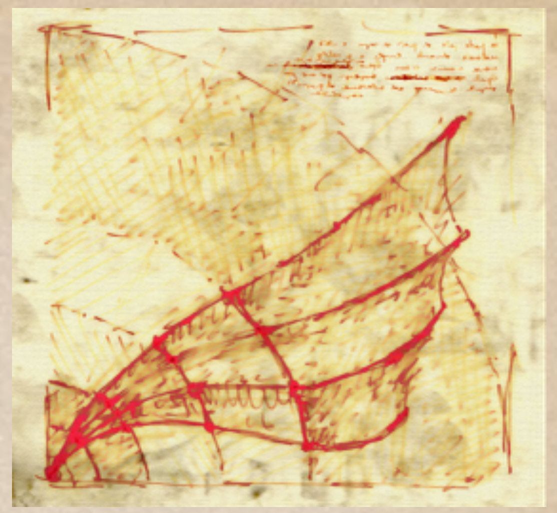
Opportunity for phase II

- Understand where splitting 40 MHz data is technically possible
- Avoid taking decisions on hardware that will make this impossible
- We do not need to build anything for day 1



Summary





Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

CMS DAQ system for run 2 fully commissioned

- Fulfills functional and performance requirements
- Extensive tuning is needed to take advantage of state-of-the-art technologies
- Ready to integrate new/upgraded sub-detectors in 2017
- Main challenge will be the new FEROL40 hardware

No radical changes for run 3

- Evaluate event-building network technology

Ideas being developed for phase II

- Common DAQ-TCDS hub based on ATCA
  - Moves interface to central DAQ into the backend crate
- High performance event-builder node with non-volatile memory
  - Pre-processing and online calibration
  - On-demand event building or remote data access
- Big uncertainty on high-level trigger CPU needs
- 40 MHz scouting for new physics

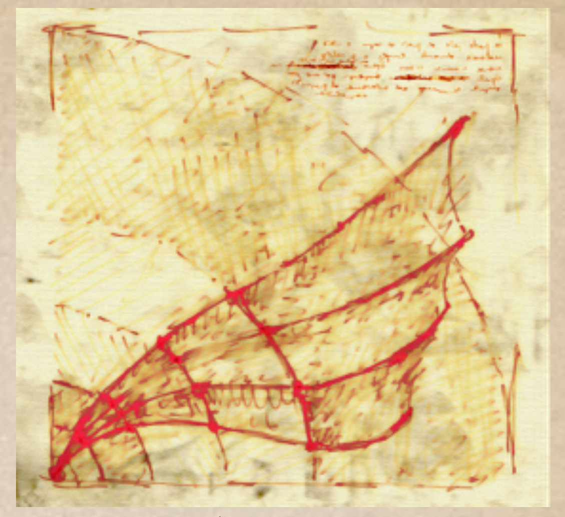
Questions?

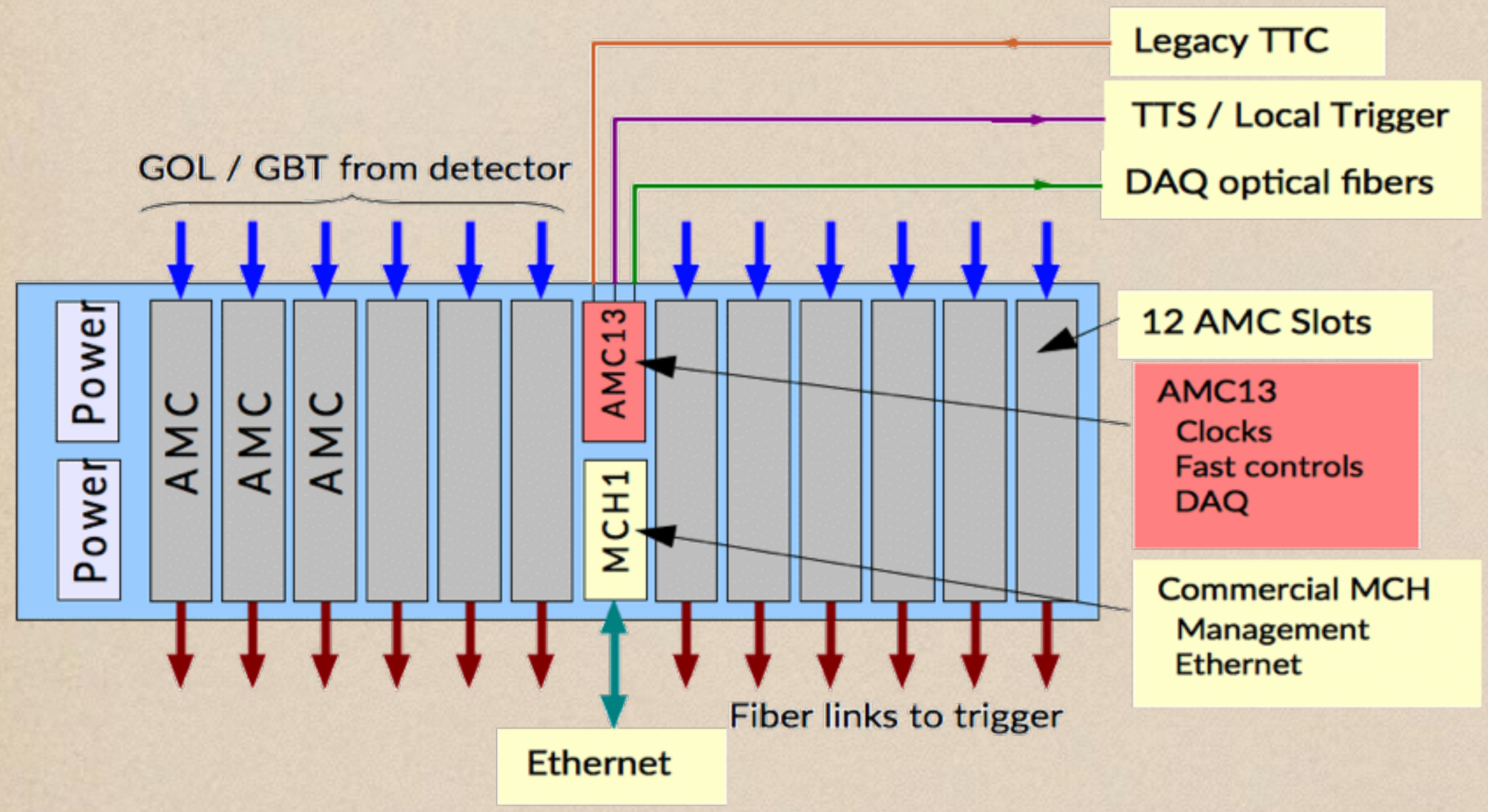
Sergio Cittolin © 2009-2016 CERN (License: CC-BY-4.0)

← ↑

Sergio Cittolin © 2009-2016 CERN (License: CC-BY-4.0)

AMC13


Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)


Legacy TTC

TTS / Local Trigger

DAQ optical fibers

GOL / GBT from detector

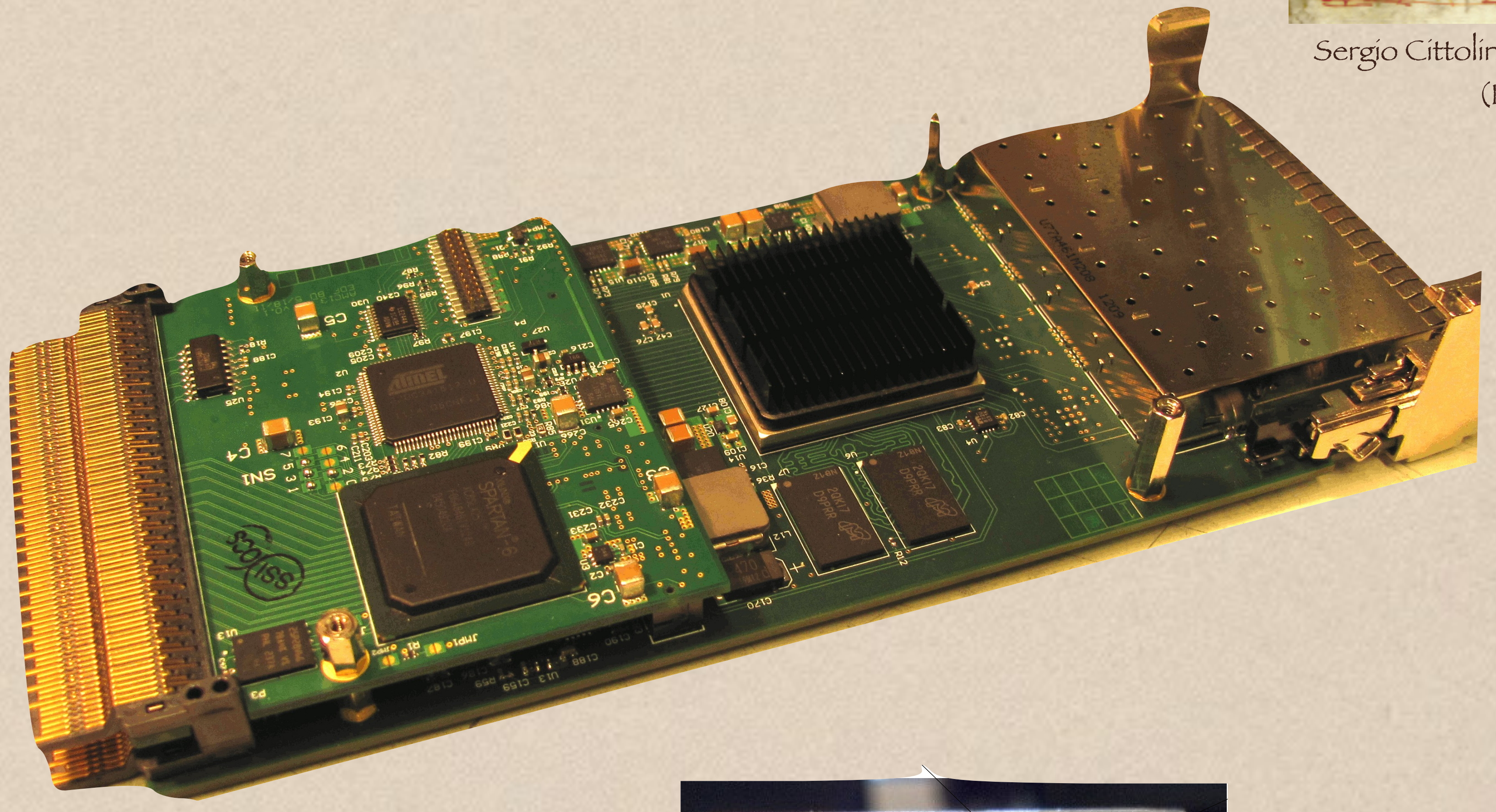
12 AMC Slots

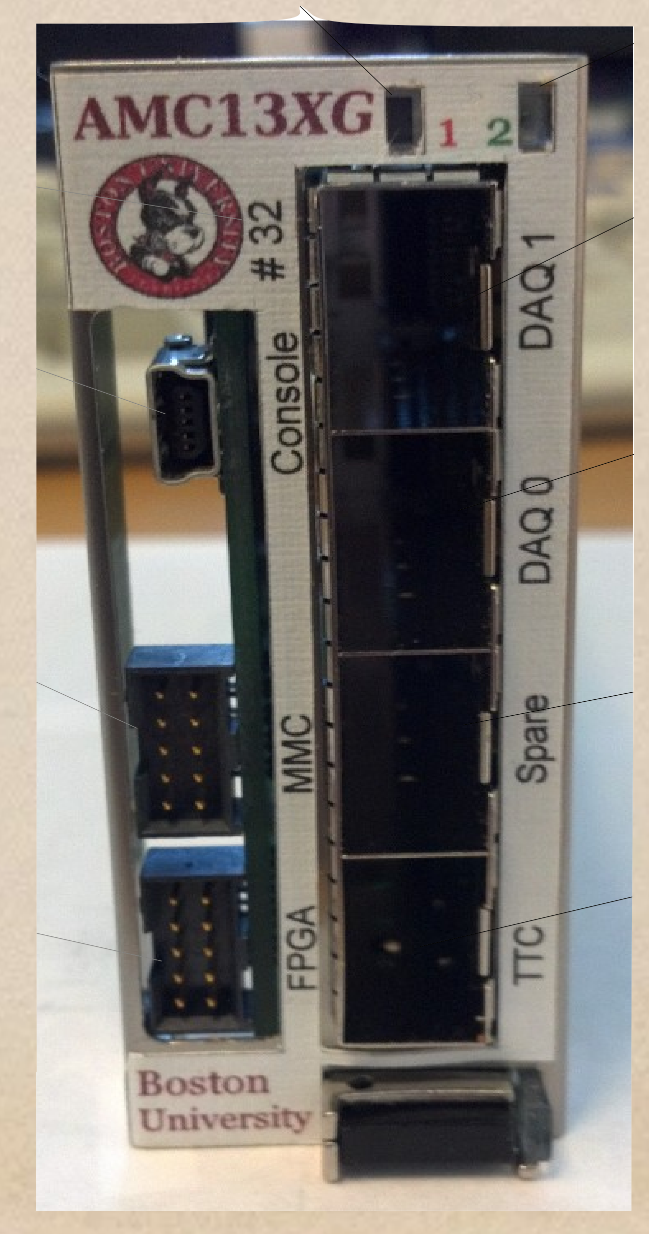
AMC13
Clocks
Fast controls
DAQ

Commercial MCH
Management
Ethernet

Fiber links to trigger

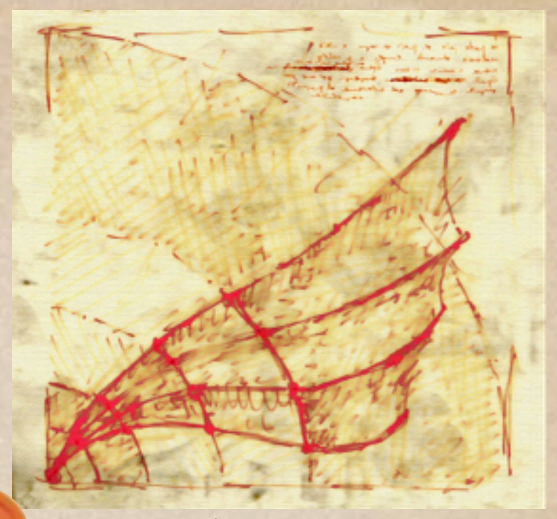
Ethernet



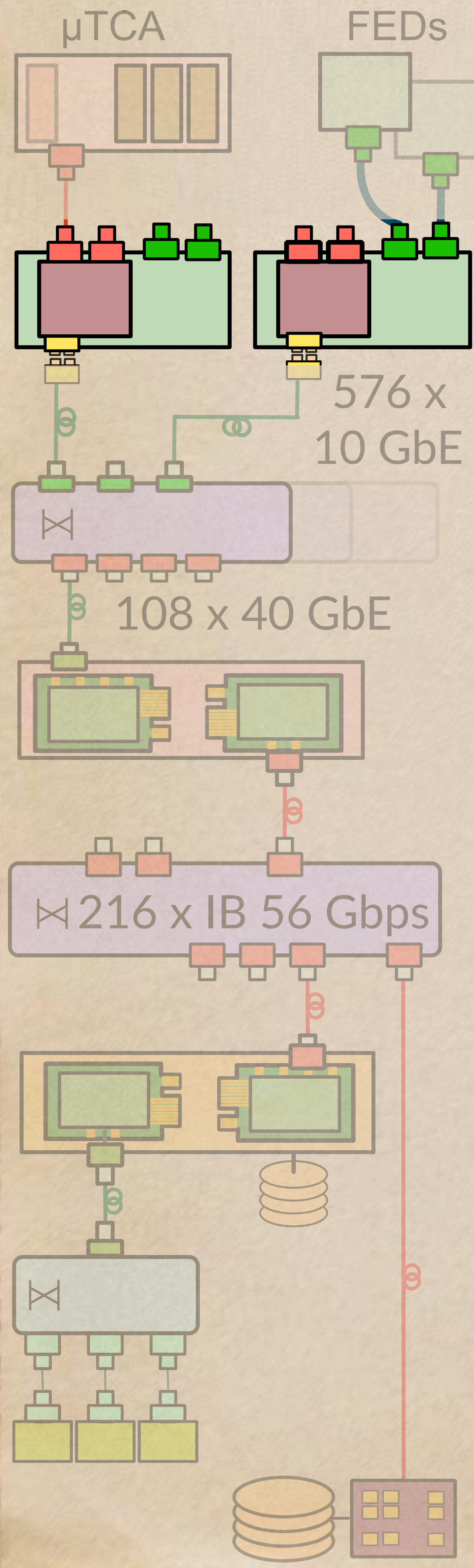


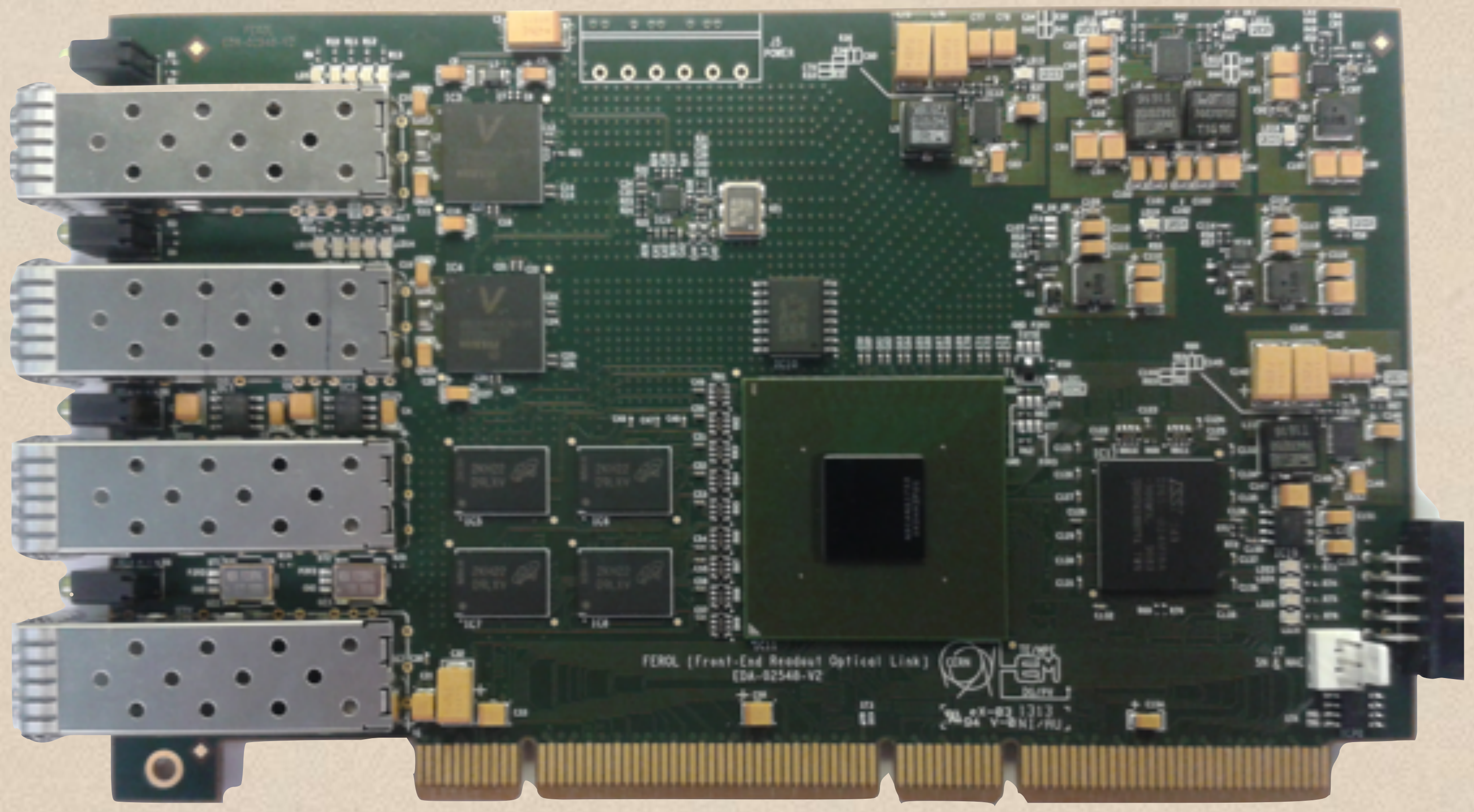
- It is not an MCH! It is a 13th AMC in MCH-2 slot
- It distributes LHC clock / timing / controls to AMCs
- It collects DAQ data from AMCs
- It provides standard interface to CMS subdetectors:
  - CMS DAQ via 1-3 optical fibers at 10 Gb/s (64/66b encoded)
  - TTC via 1300nm fiber @ 160Mb/sec biphase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

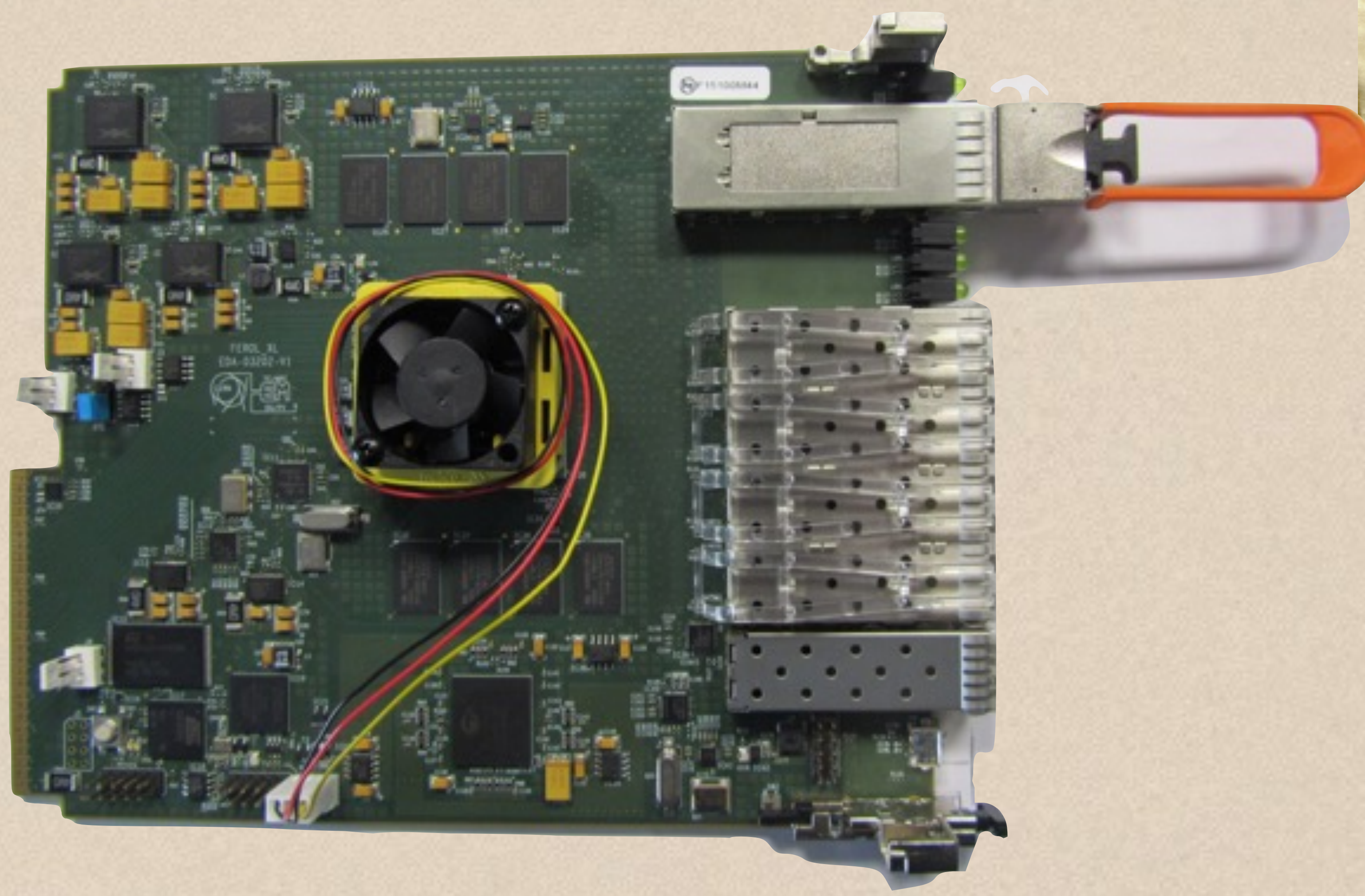
Front-End Readout Optical Link (FEROL)



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

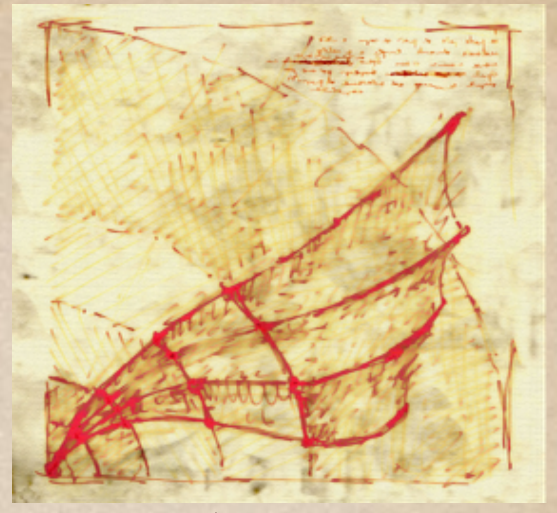




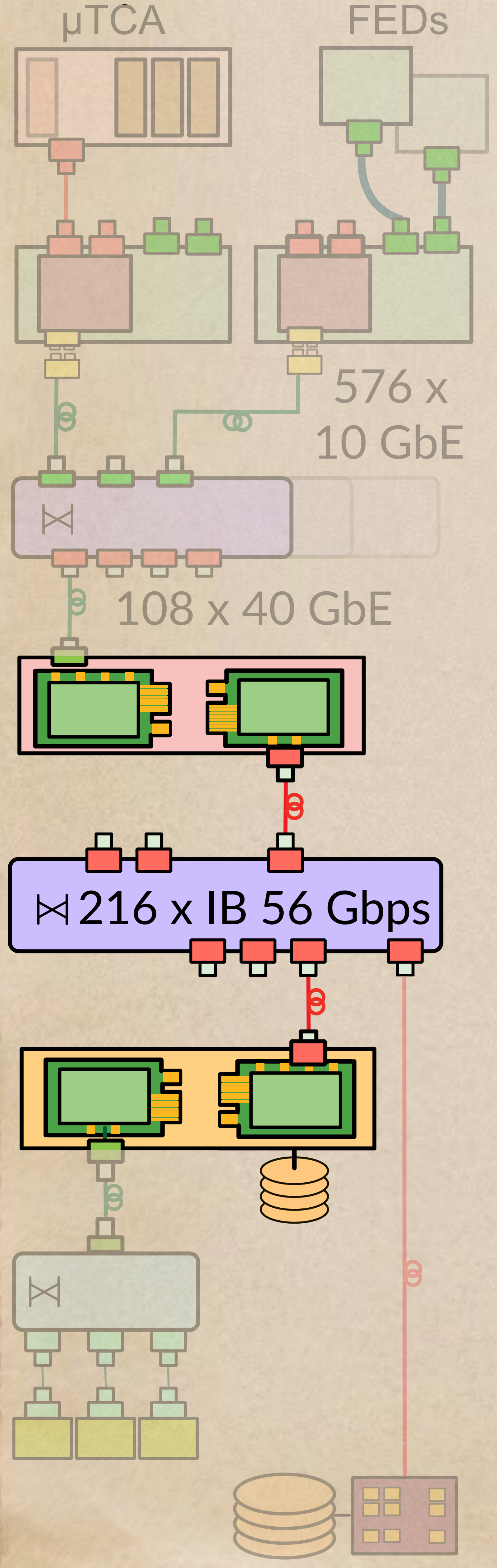


	FEROL	FEROL40
FPGA	Altera Arria II GX	Altera Arria V GZ
QDR Memory	16 MB	32 MB
DDR Memory	512 MB DDR2	2x 1GB DDR3
Input (SLINKXpress)	2x optical 6 Gbit/s or 1x optical 10 Gbit/s	4x optical 10 Gbit/s
DAQ interface (Ethernet)	1x optical 10 Gbit/s	4x10 Gbit/s or 40 Gbit/s

Event Builder Performance



Sergio Cittolin © 2009-2016 CERN (License: CC-BY-4.0)



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

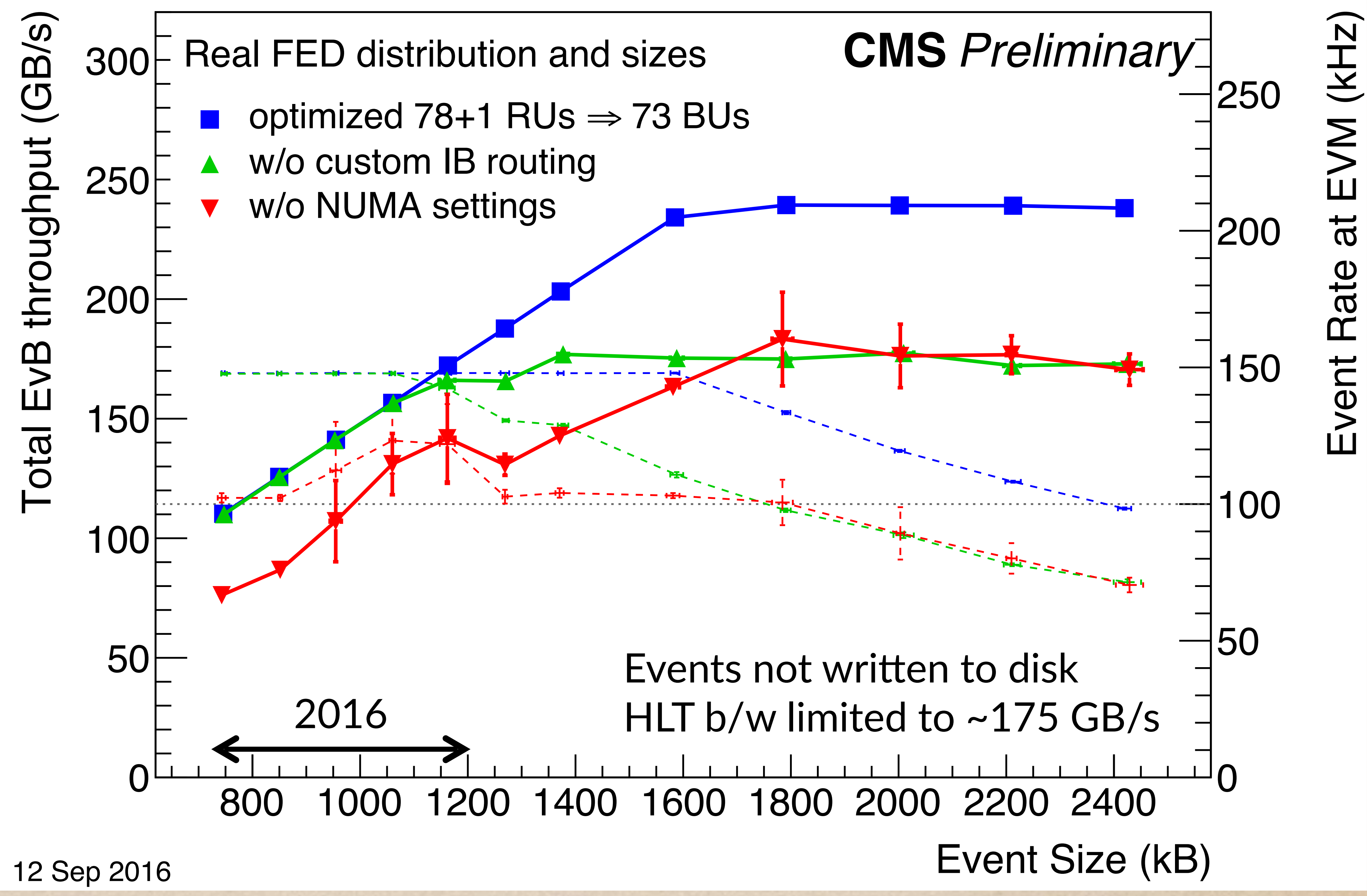
Parallelize the work

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

Bind to CPU cores and memory (NUMA)

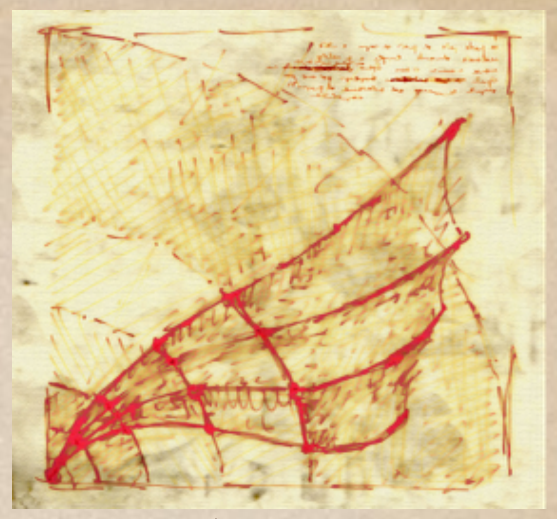
- Bind threads & memory structures to cores
- Restrict interrupts from NICs to certain cores
- Tune Linux TCP stack for maximum performance

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

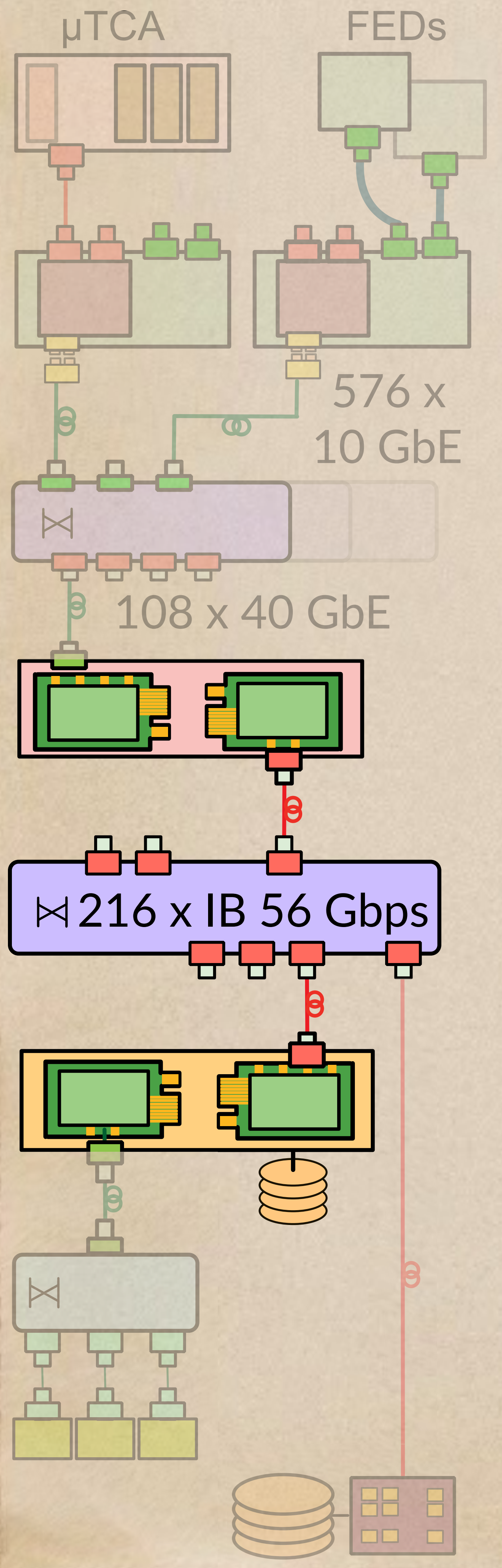


Use custom IB routing taking into account the event-building traffic pattern

Event Builder



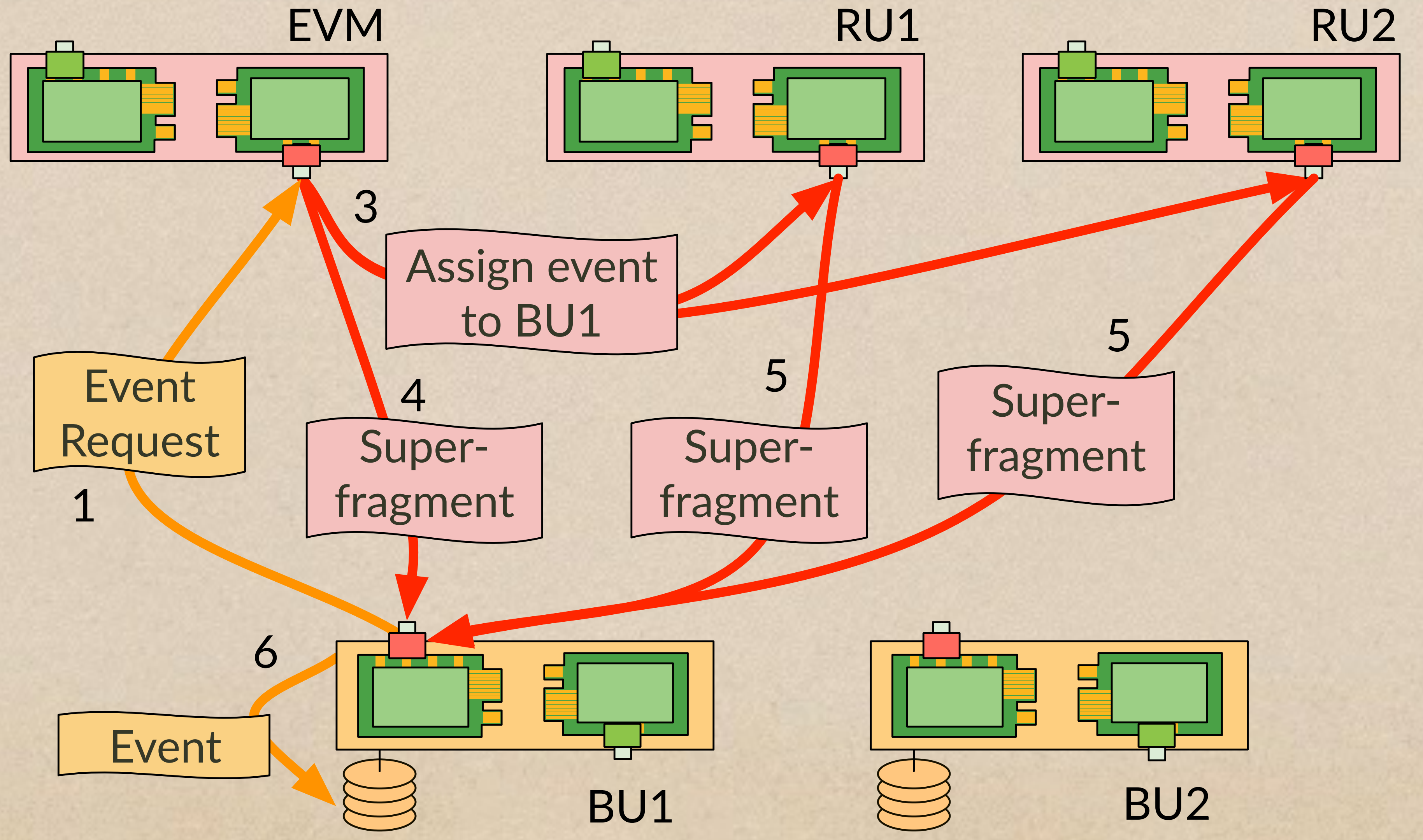
Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)

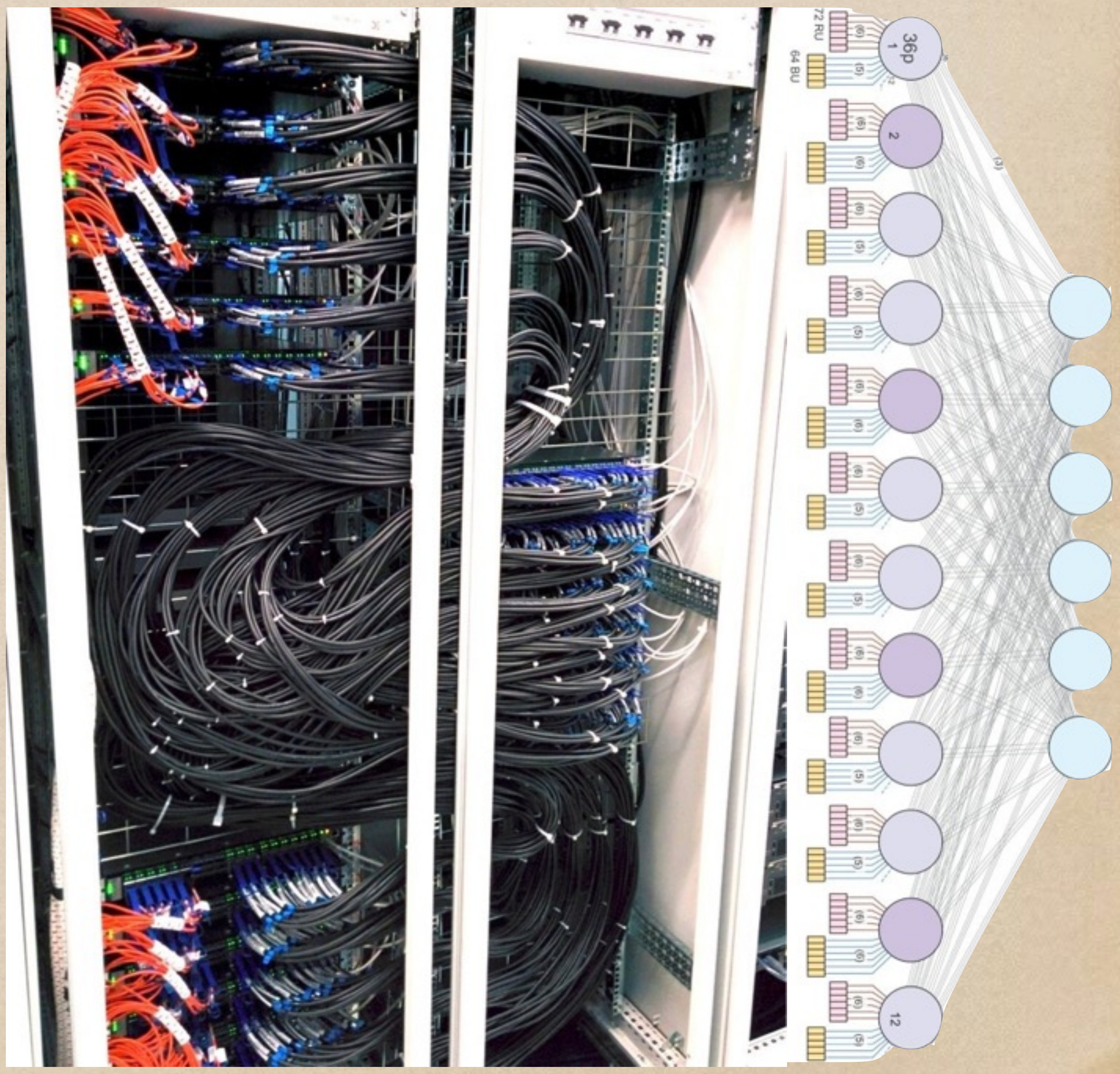


InfiniBand – most cost-effective solution

- 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

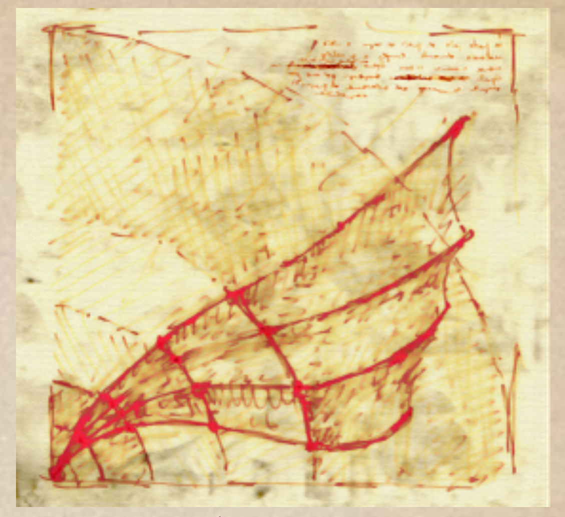
Event Builder protocol



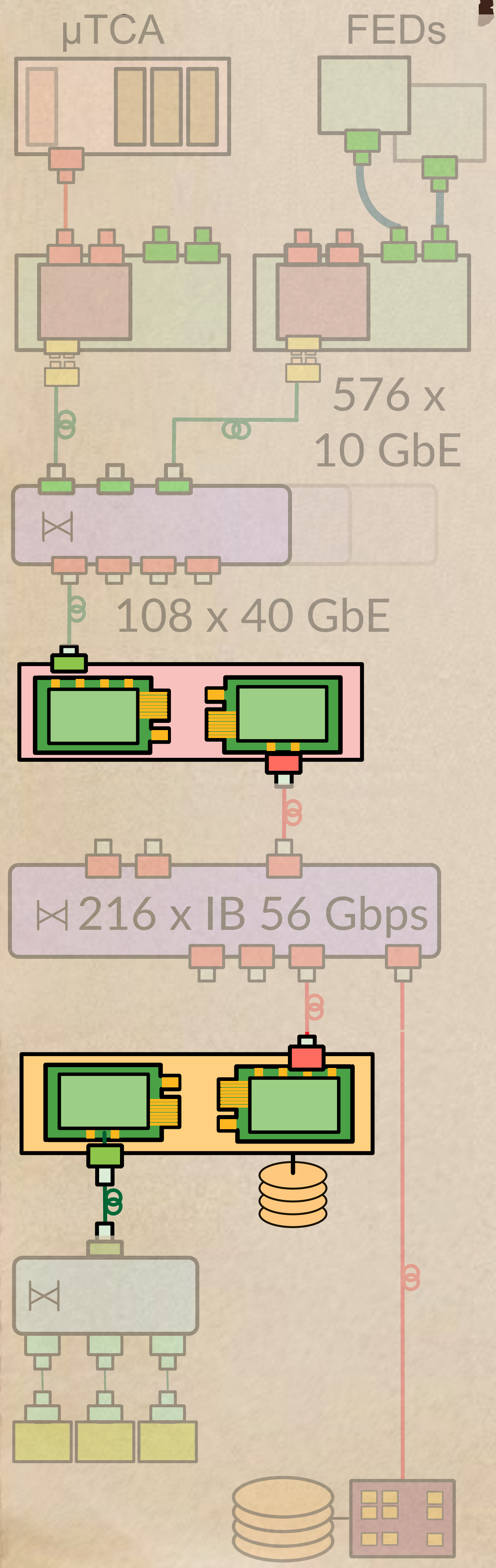


Infiniband CLOS network

Computers



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)



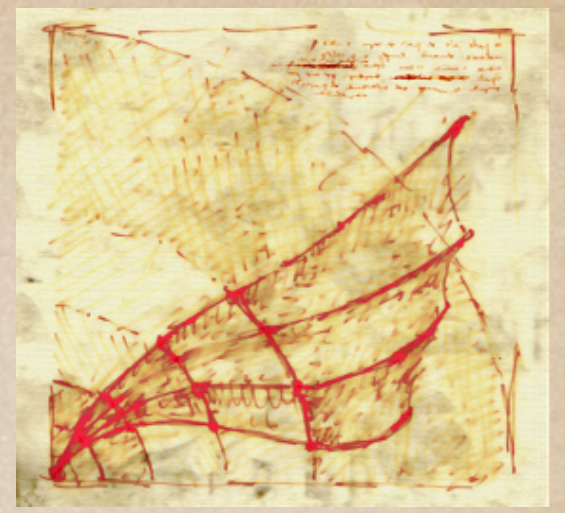
Readout Unit (RU)

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

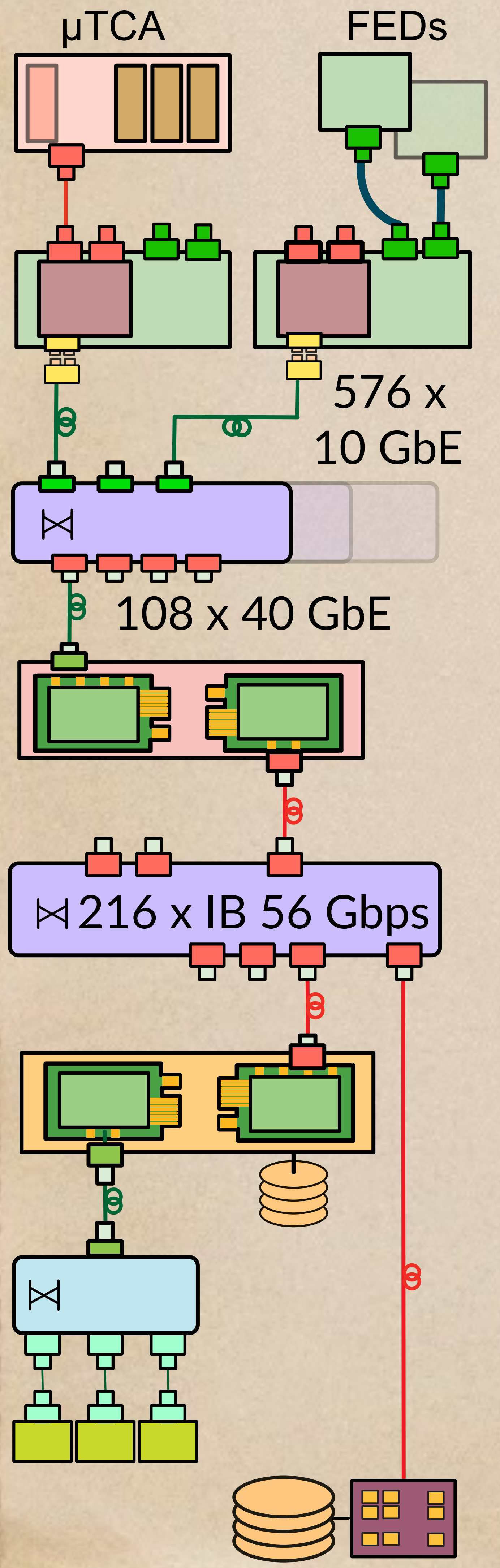
Builder Unit (BU)

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

Configuration & Control



Sergio Cittolin © 2009-2016 CERN
(License: CC-BY-4.0)



Data-flow applications based on XDAQ C++ framework

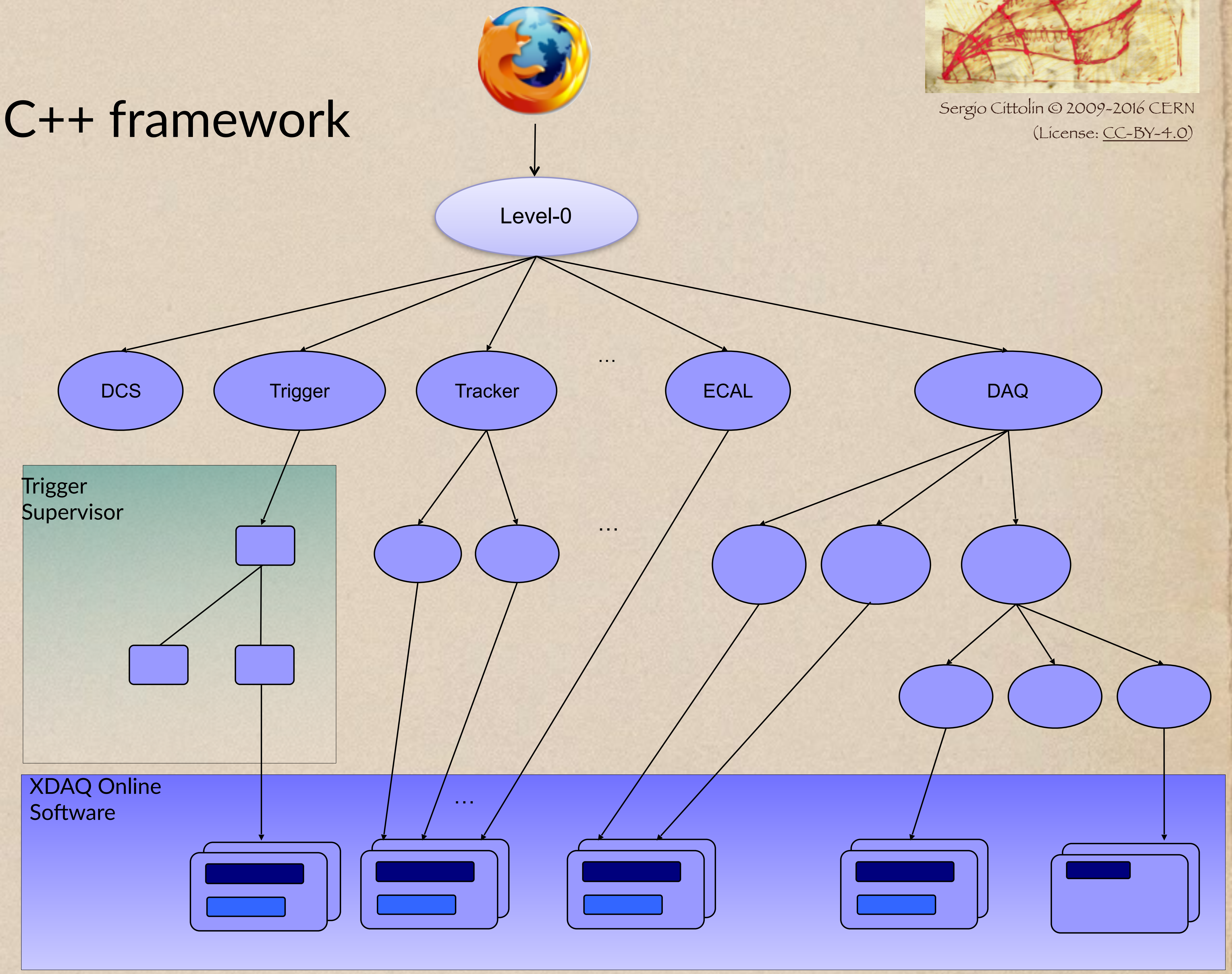
- Reusable building blocks for
  - Hardware access
  - Transport protocols
  - Services
- Dynamic configuration based on XML
- Controlled and browsable with HTTP/SOAP

Run-Control & Monitoring System

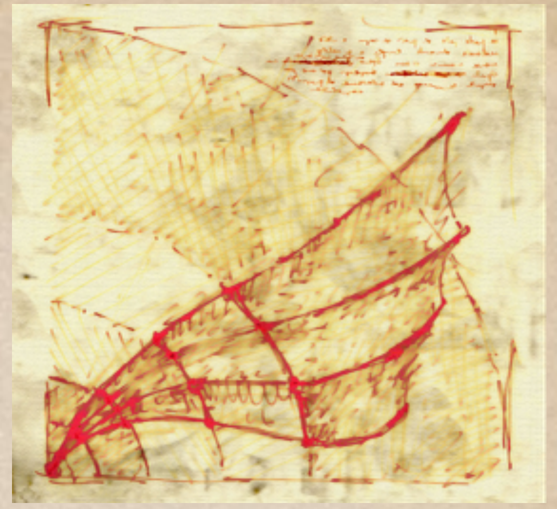
- Hierarchical control structure
- Java code running as Tomcat servlets
- React on state machine events
  - Commands from parents
  - Errors from children

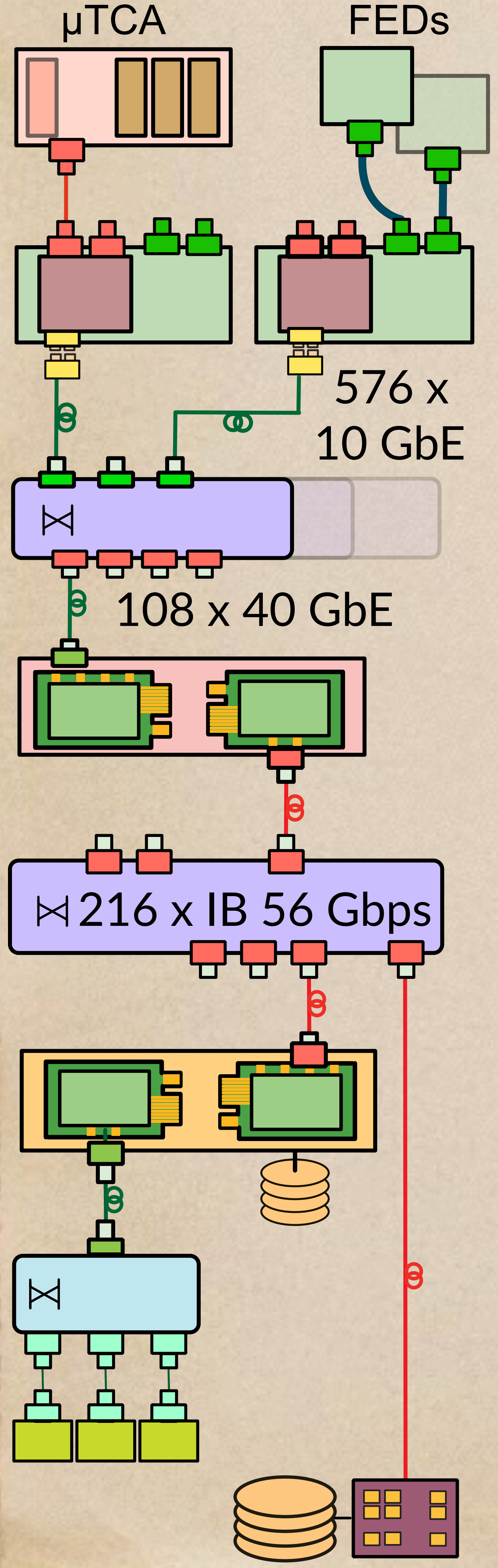
File-based filter farm

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



Monitoring & Error reporting





XDAQ services in each application

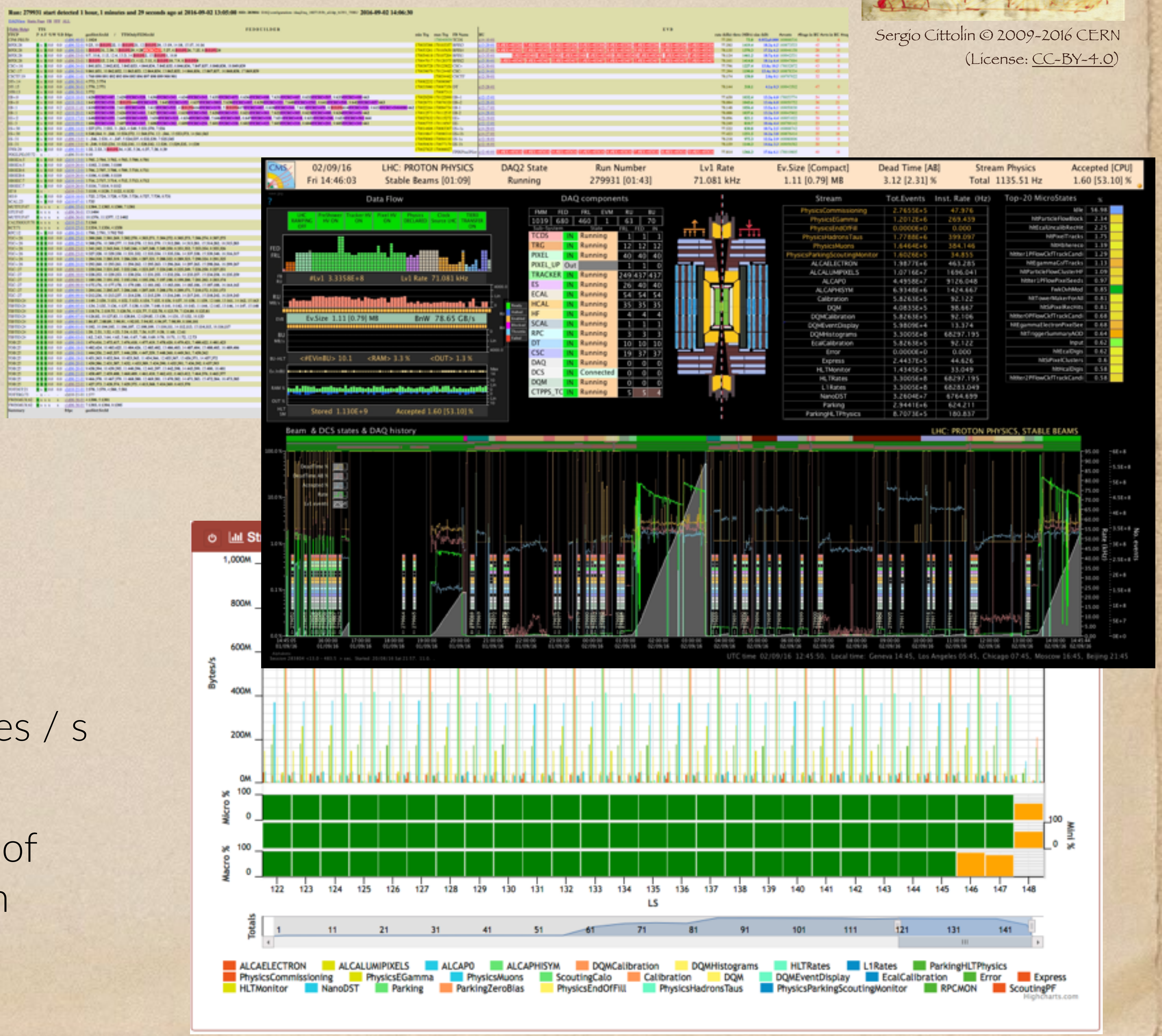
- Periodically publish monitoring data
- Central logging facilities
- Error reporting

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


02/09/16
Fri 14:46:03
LHC: PROTON PHYSICS
Stable Beams [01:09]
DAQ2 State
Running
Run Number
279931 [01:43]
Lv1 Rate
71.081 kHz
Ev.Size [Compact]
1.11 [0.79] MB
Dead Time [AB]
3.12 [2.31] %
Stream Physics
Total 1135.51 Hz
Accepted [CPU]
1.60 [53.10] %
Data Flow
DAQ components
Stream
Top-20 MicroStates
Beam & DCS states & DAQ history
LHC: PROTON PHYSICS, STABLE BEAMS
1,000M
800M
600M
400M
200M
0M
Bytes/s
Micro %
Macro %
Mini %
LS
Totals
ALCAELECTRON
ALCALUMIPIXELS
ALCAP0
ALCAPHISYM
DQMCalibration
DQMHistograms
HLTRates
L1Rates
ParkingHLTPhysics
PhysicsCommissioning
PhysicsEGamma
PhysicsMuons
ScoutingCalo
Calibration
DQM
DQMEventDisplay
EcalCalibration
Error
Express
HLTMonitor
NanoDST
Parking
ParkingZeroBias
PhysicsEndOfFill
PhysicsHadronsTaus
PhysicsParkingScoutingMonitor
RPCMON
ScoutingPF