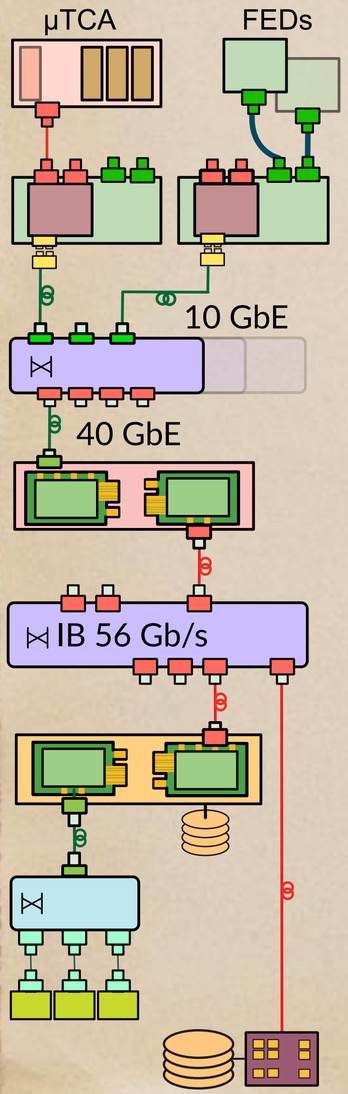
The CMS Event-Builder System for LHC Run 3 (2021-23)

Remigius K Mommsen
Fermilab

on behalf of the CMS DAQ group

CMS Data Acquisition System



Detector front-end (custom electronics)
- ~740 front-end drivers (FEDs)
- 0.1 - 8 kB fragments at 100 kHz (1.4 MB event size)

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

Data Concentrator switches
- Data to Surface over ~200m
- Aggregate into 40 Gb/s Ethernet links

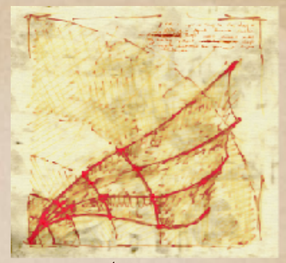
Up to 108 Readout Units (RUs)
86 RUs actively used
- Combine FEROL fragments into super-fragment
- Buffer fragments

Event Builder switch
- Infiniband FDR 56 Gb/s CLOS network

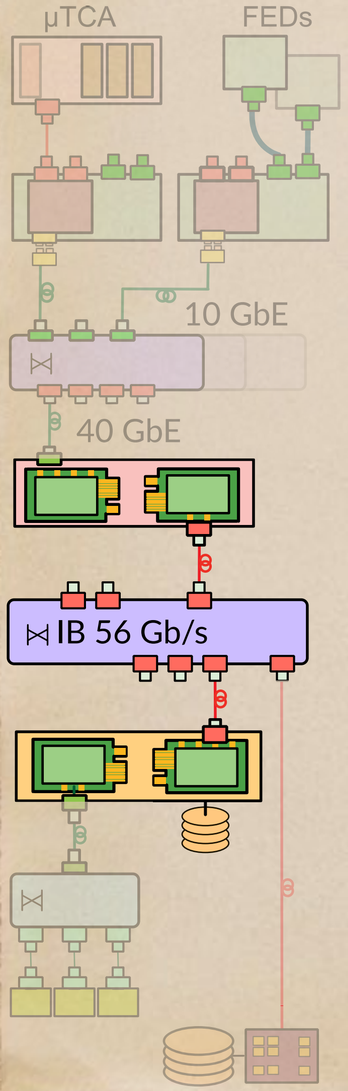
73 Builder Units (BUs)
- Event building & temporary recording to RAM disk

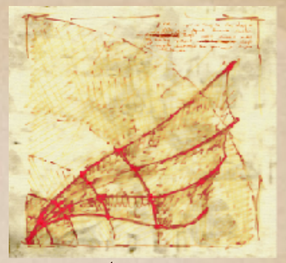
Filter Units (FUs)
~32k physical cores in ~1100 boxes
- Run HLT selection using files from RAM disk
- Select O(1%) of the events for permanent storage

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



Event-Builder Network





Infiniband CLOS network

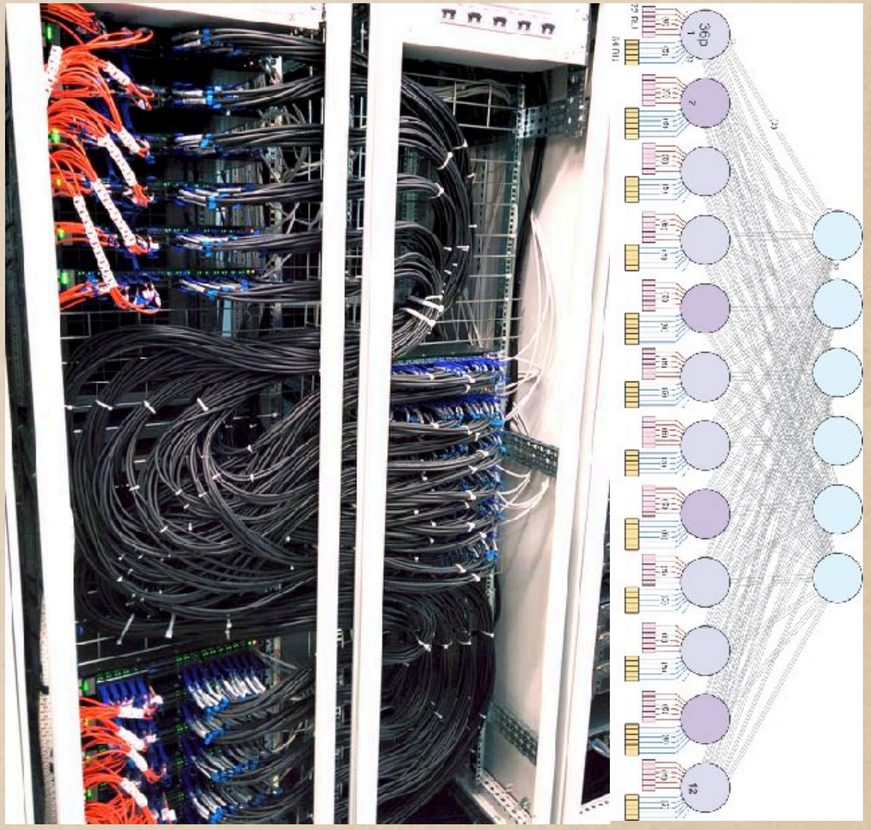
- 12 leaf and 6 spine switches (Mellanox SX6036)

Interconnects readout (RU) and builder units (BU)

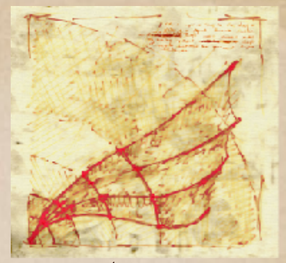
- RUs receive data from backends over Ethernet
- RUs push data to BUs upon receiving a readout message
- BUs build events from fragments from all RUs
- BUs write complete events to file for HLT processing

290 GB/s throughput of event-builder traffic

- ~66% of the line speed due to head-of-line blocking
- Optimized routing takes into account unidirectional traffic
- All processes/threads are pinned to CPU cores and memory



The DAQ System for LHC Run 3



Requirements do not change

- Collect data from ~800 sources at 100 kHz L1 trigger rate
- Transport the data from underground over ~200 meters to the online computing farm
- Build complete events of ~1.5 MB size
- Store events to a file system where they can be accesses by the HLT processes

Why replace the current DAQ system?

- Commercial equipment reaches end-of-life after 5 years
- Keep abreast with technological evolution

Custom electronics (FEROLs) will stay

- Data-concentrator network still using Ethernet

Explore solutions for next generation DAQ system

- Combine readout and builder units into a single I/O processor (folded architecture)
- Investigate the integration co-processors (GPUs or FPGAs) into pre-process events before handing them to the HLT
- Partial event acquisition at bunch-crossing rate of 40 MHz

Interconnects

DAQ1 2002 — 100 GB/s handled by 640 RUs & 720 BUFUs

- 2-rail 2 Gb/s Myrinet
- Multi-rail 1 Gb/s Ethernet

DAQ2 2014 — 200 GB/s handled by 90 RUs & 70 BUs

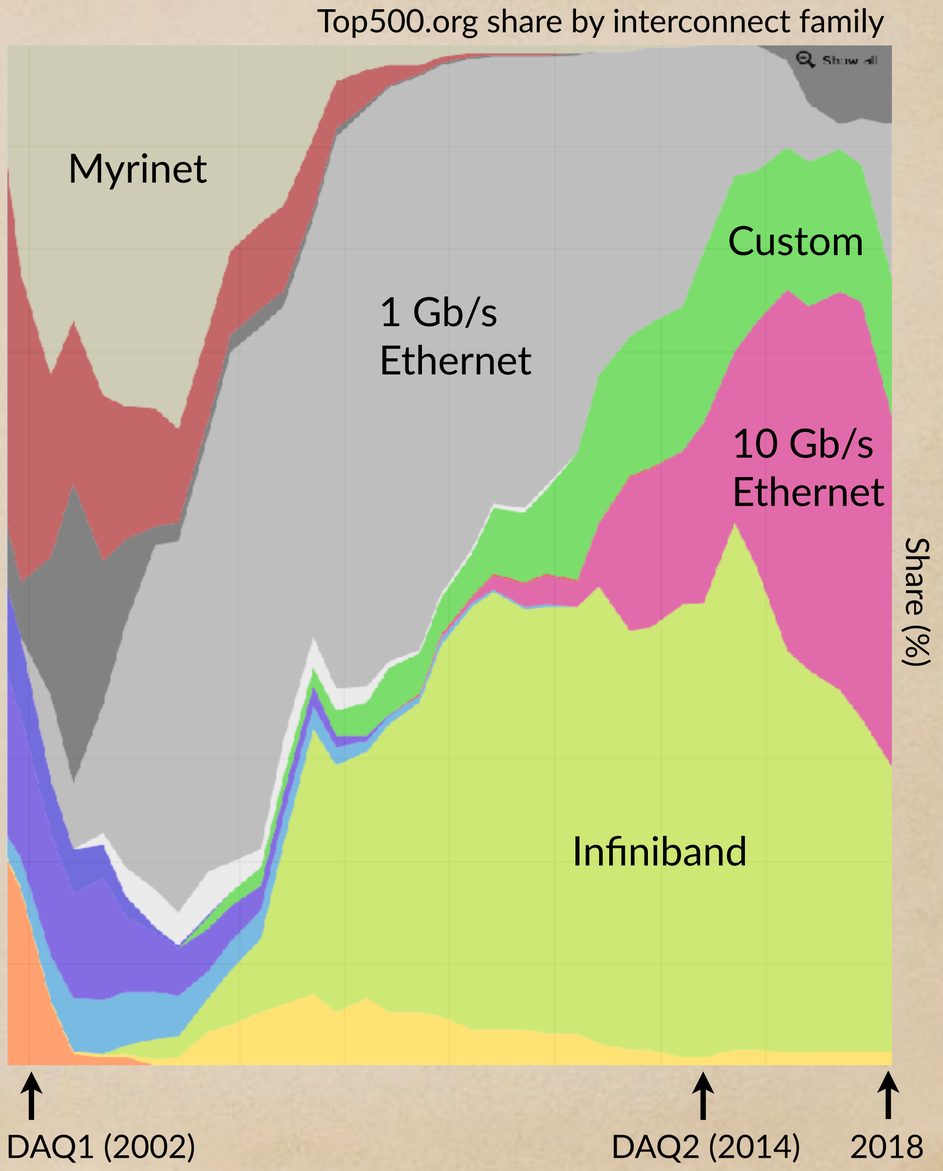
- 10 and 40 Gb/s Ethernet
- Infiniband FDR 56 Gb/s

DAQ3 2020 — 200 GB/s handled by 50-100 RUBUs

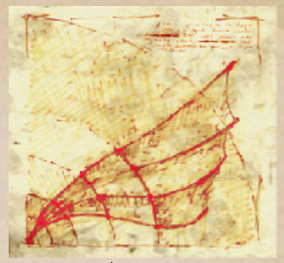
- 10/40/100 Gb/s Ethernet
- Infiniband EDR 100 Gb/s or HDR 200 Gb/s

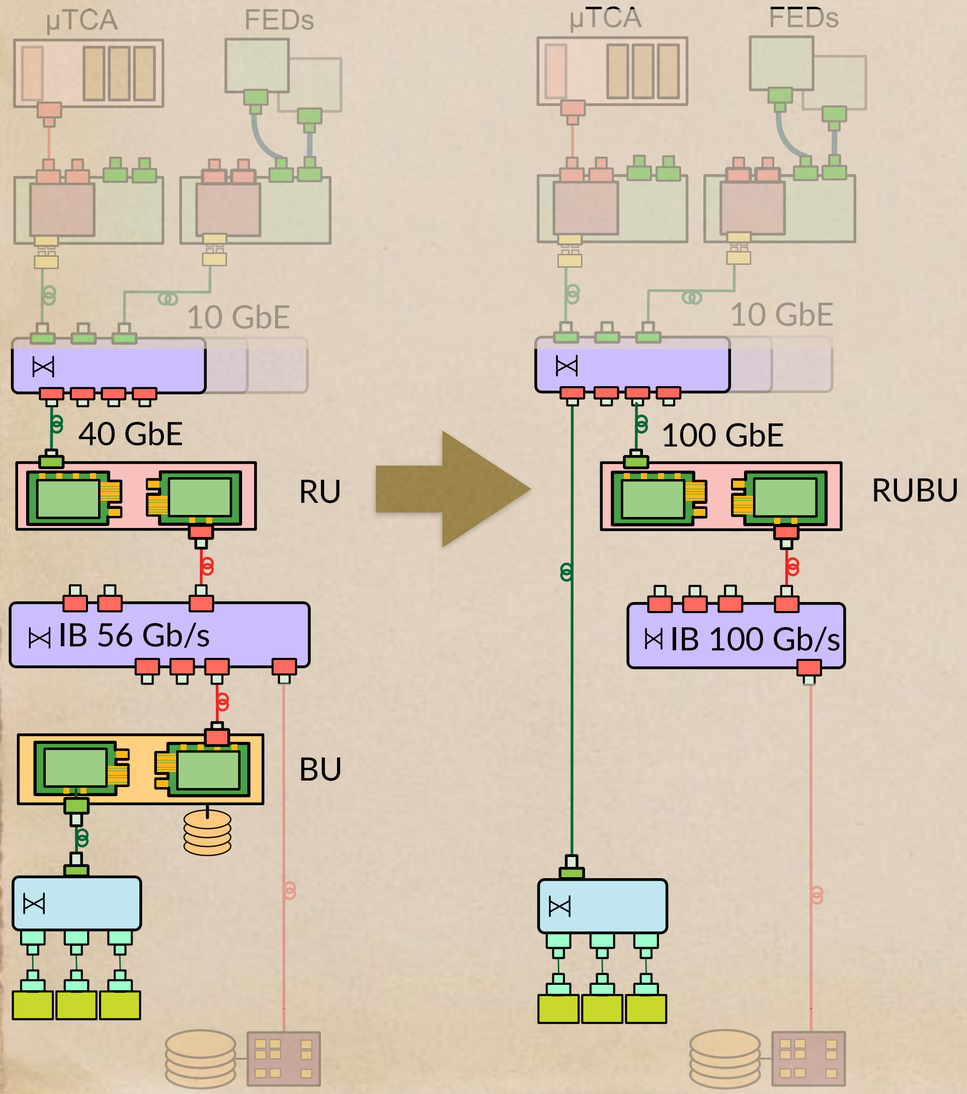
DAQ4 2025 — 5.5 TB/s handled by ~500 RUBUs (HL-LHC)

- 100/200 Gb/s Ethernet
- Whatever technology will be the most cost effective in terms of cost per port vs number of ports/machines needed



Folded Event-Builder Architecture





Combine readout and builder units into a single machine

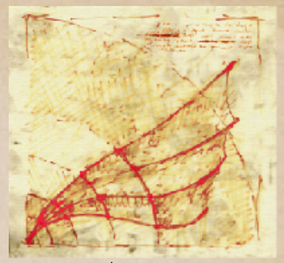
Less hardware to purchase and maintain

- Needs about half the number of machines and switch ports
- Exploit bi-directional links
- Traffic balancing becomes more challenging
- Required for HL-LHC (2025) to save costs

Demanding I/O performance and memory performance

- Process 100 Gb/s TCP/IP stream
- Distribute event fragments to other builder units at 100 kHz
- Build complete events at 1-2 kHz
- Write events to files
- Possibly hand events to co-processors

Feasibility Test

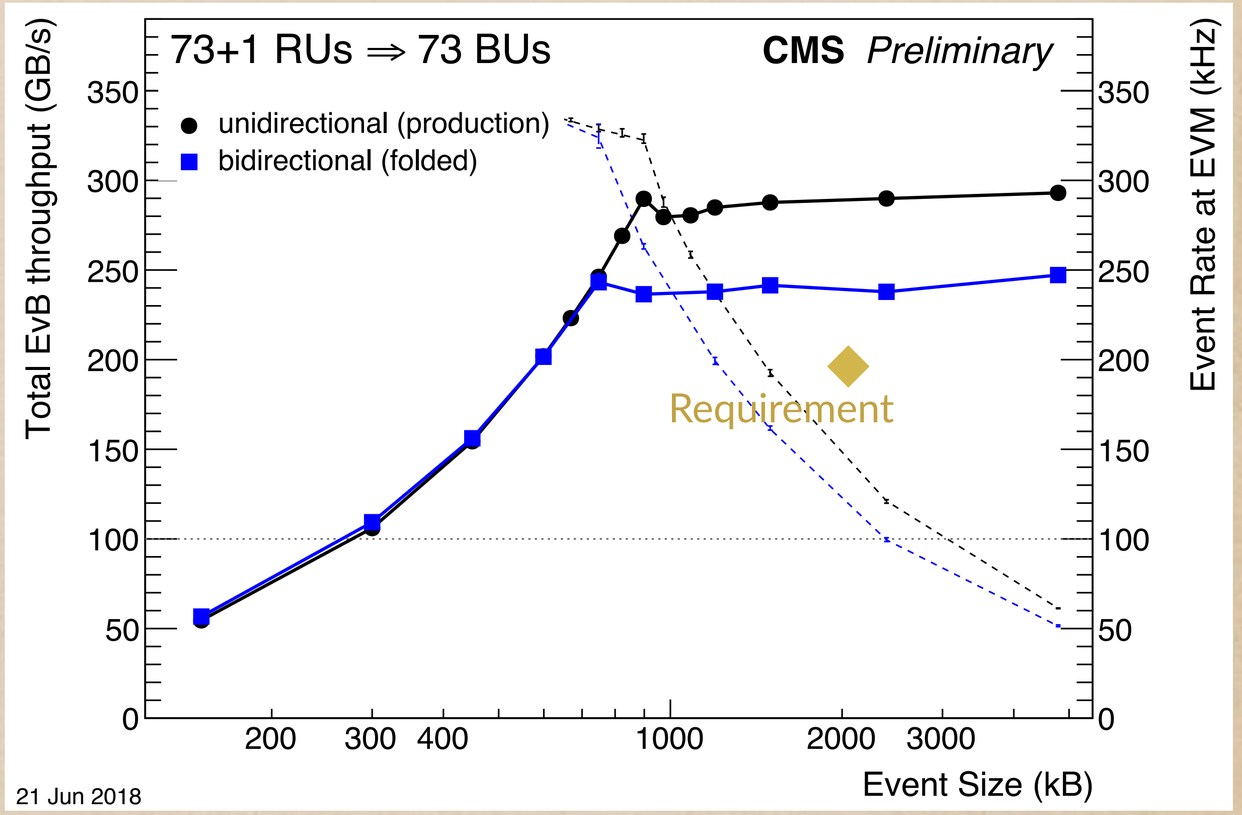


Run BU on the RU node

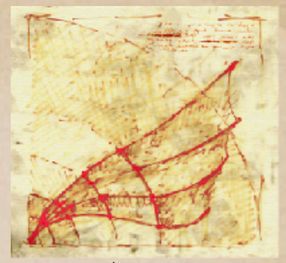
- Unchanged s/w stack
- Re-tuned NUMA settings
- Events generated on the RU, i.e. no TCP/IP streams
- Events not written to disk

Compare performance using production system

- Dell R620/R720 dual 8-core sandy bridge @ 2.6 GHz
- Mellanox SX6036 CLOS network
- Plateau throughput reduced by ~15%
- Could still build 2 MB at 100 kHz L1 trigger rate
- Distribution of events to HLT farm limited to ~220 GB/s



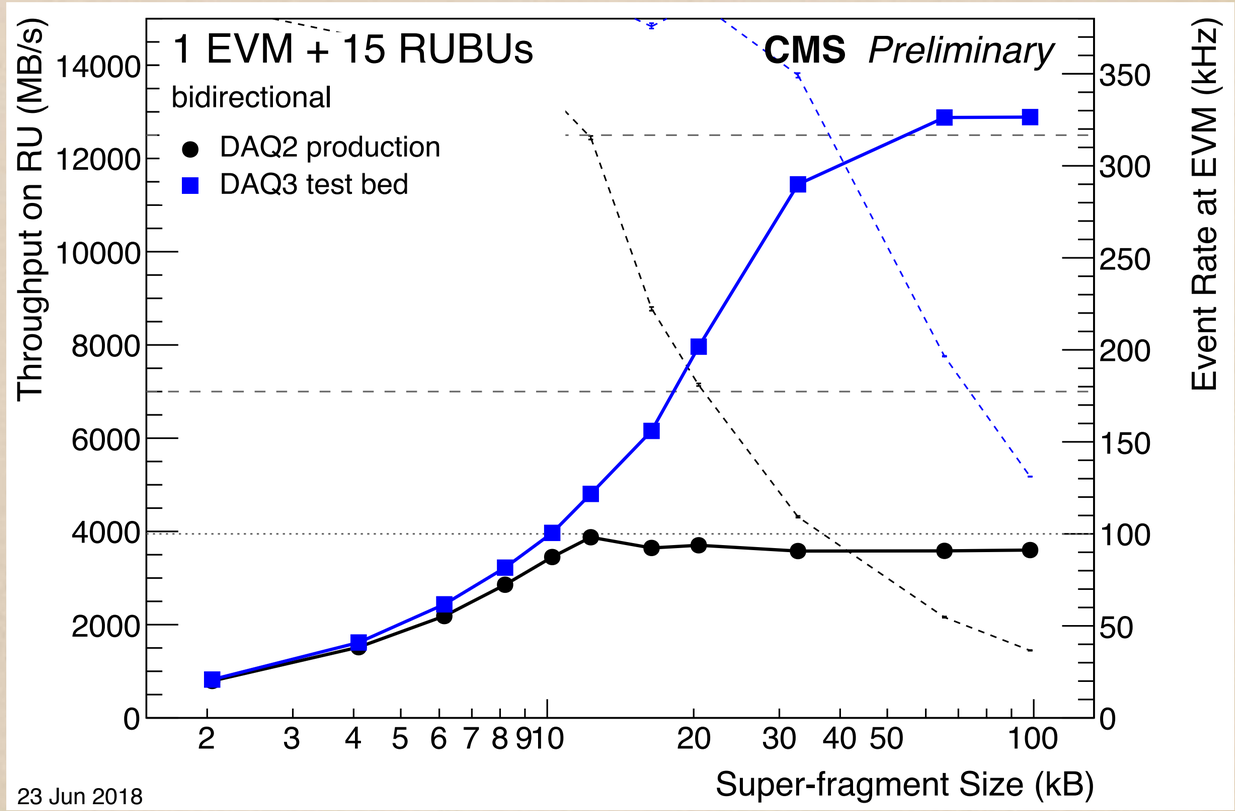
First Look at Latest Hardware



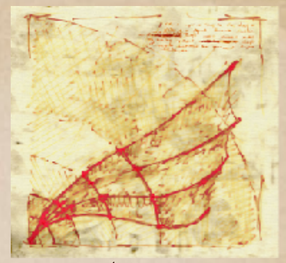
Testbed with 16 nodes

- Dell R740 dual 16-cores skylake @ 2.1/2.6 GHz
- Single Mellanox MSB7800 EDR switch (100 Gb/s)

Up to 3 times higher throughput for folded architecture compared to DAQ2 hardware



Event Building with MPI



Can we profit from MPI for the event building?

- MPI gather and alltoall methods attractive
- Do not support variable fragment sizes
- Padding to max. fragment size wastes a lot of bandwidth

Implement MPI application based on Isend/Irecv

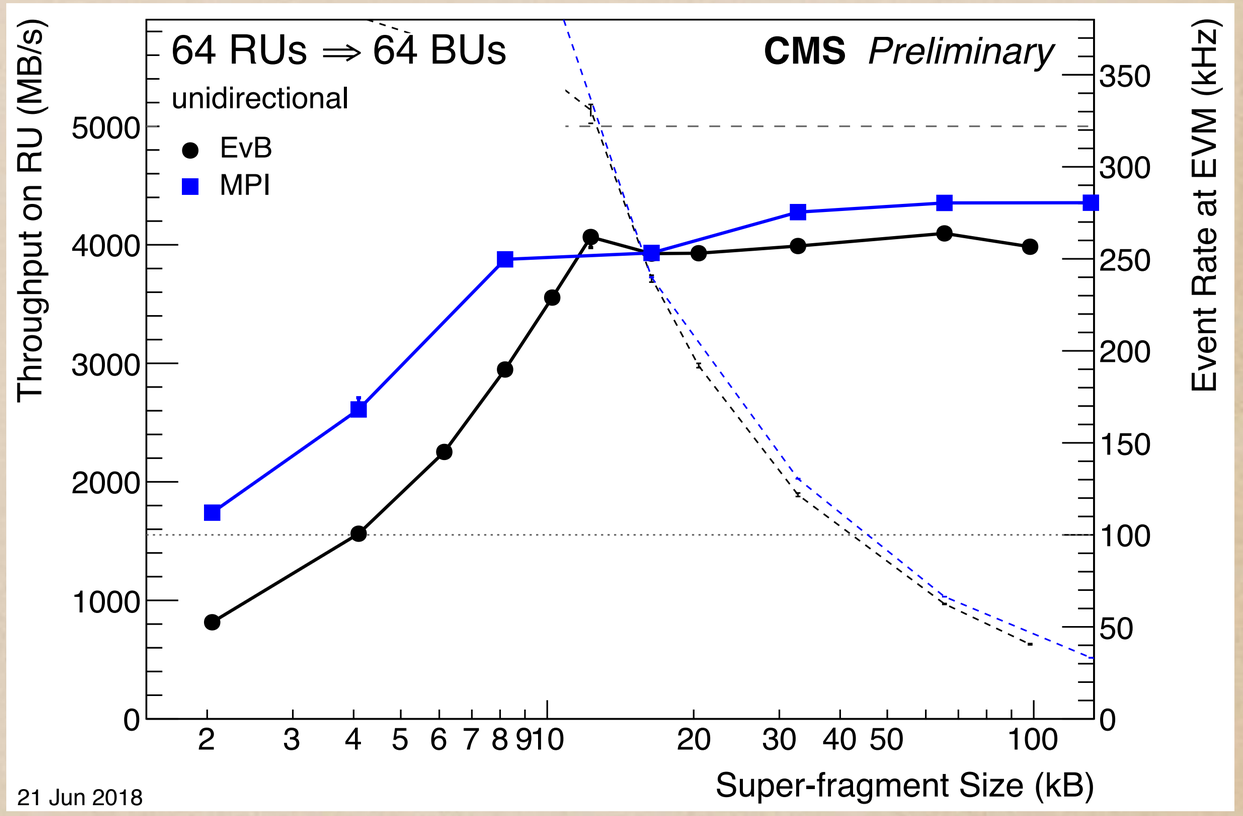
- Non-blocking and no external synchronization
- Backpressure through limited buffers
- Simple streaming of event-fragments

Comparison with full event-builder software

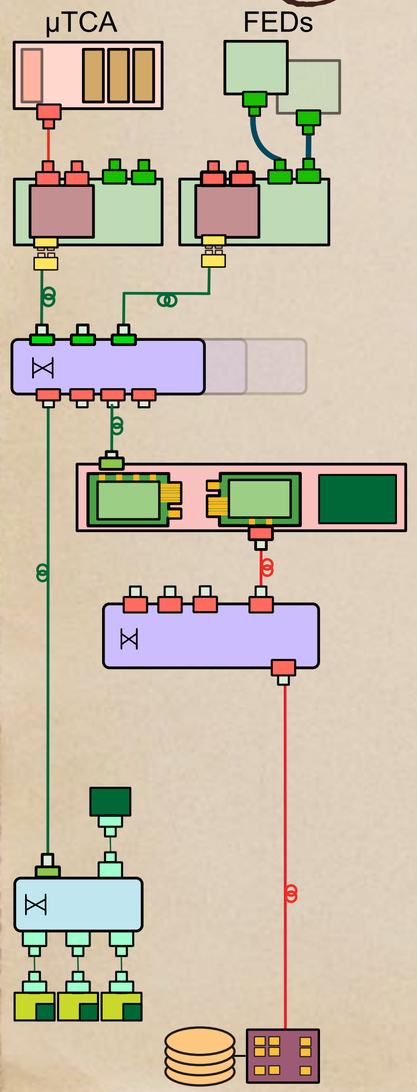
- Better throughput for small fragments
- No gain in plateau throughput

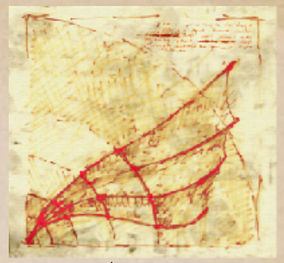
No benefit compared to verbs-based implementation

- Tuning of MPI parameters to achieve good performance
- MPI might become useful to integrate GPUs



Integration of Co-Processors





10-20 fold increase in CPU power of HLT event selection for HL-LHC

- Might need to integrate GPUs and/or FPGAs into the event selection

Equipping each HLT node with a GPU not cost effective

- GPU most likely cannot be fully loaded
- Host machines and GPUs have different life cycles

Offload specific parts of the event selection to GPU farm during the HLT processing

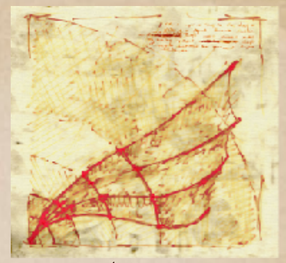
- Requires mechanisms and network to transfer data to and from a GPU farm
- Adds latency for data transport
- Careful tuning needed to avoid that HLT CPU stalls on network transfers

Preprocess all L1 accepted events before handing them to the HLT

- Avoids any latency during HLT processing
- Could be effective for specific tasks on a subset of the event data
- Would waste many GPU cycles if result is only used for a small number of events

DAQ3 could be a good testbed to investigate technical solutions

Summary



Commercial components of current DAQ system reach end-of-life

- New system needs to be built in 2020 for LHC run 3 (2021-23)
- Requirements do not change compared to today's system

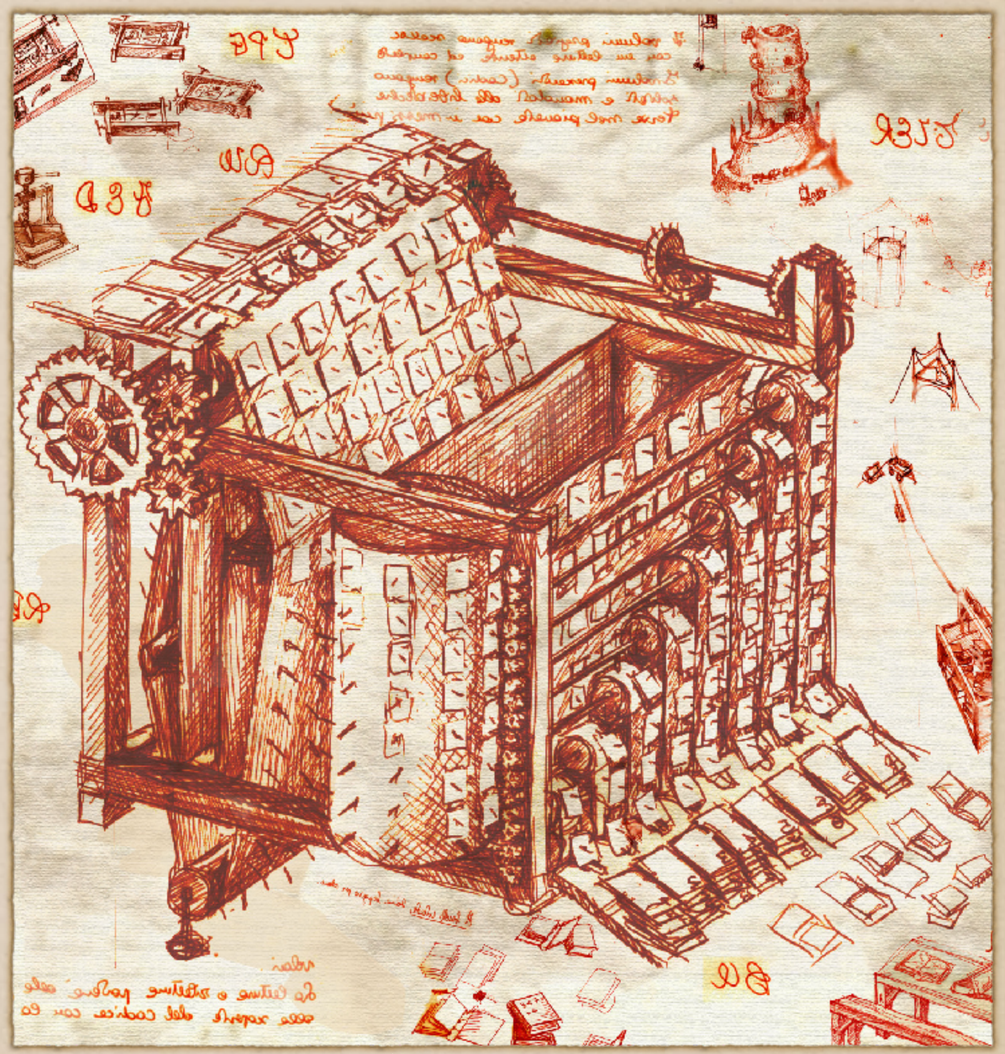
DAQ3 is a stepping stone to much more performant DAQ system needed for HL-LHC (phase 2)

- Folded architecture to better exploit hardware capabilities
- Investigate techniques to integrate GPU/FPGAs into the event selection
- Explore the possibility of a parasitic DAQ system to acquire partial events at a bunch crossing rate of 40 MHz

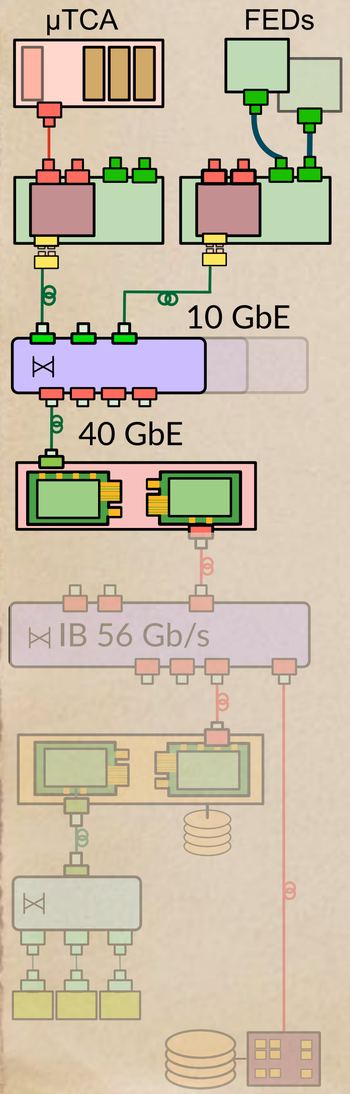
A folded architecture seems to be feasible with today's hardware

- Handling TCP/IP streams at 100 Gb/s will be a challenge
- Need a more efficient distribution of events to the HLT farm

MPI for the event building does not offer a clear benefit to the current verbs-based implementation



Data Concentrator



Front-End Readout Optical Link (FEROL)

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

New version (FEROL40)

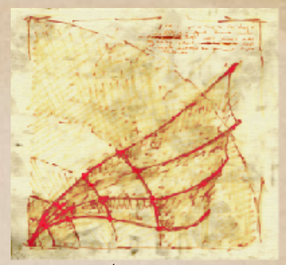
- µTCA standard (without legacy FRL board)
- 4x10 Gbps optical input and 40 GbE output

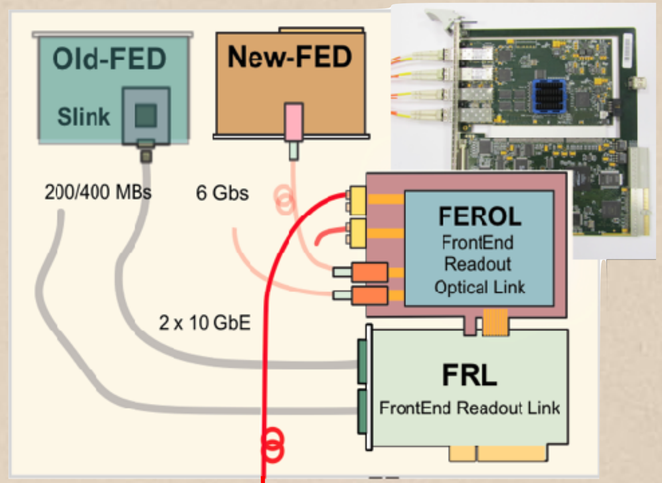
Data to surface

- 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



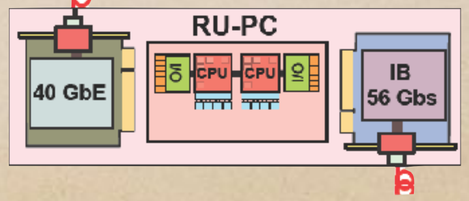


48 x 10 Gb/s

10 Gb/s simplified TCP/IP from an FPGA

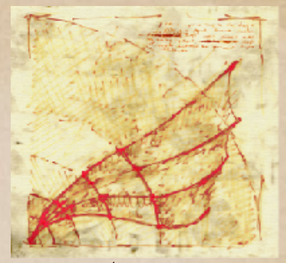
Data concentration:
10/40 Gb/s Ethernet switch

6 x 40 Gb/s



dual 8-core
E5 2670

Partial Event Acquisition at 40 MHz



At HL-LHC most detectors provide data for each bunch crossing at 40 MHz

- All tracks down to Pt>2GeV, |η|<2.4
- Triggerless streaming of e.g. calorimeters
- L1 objects used on L1 trigger

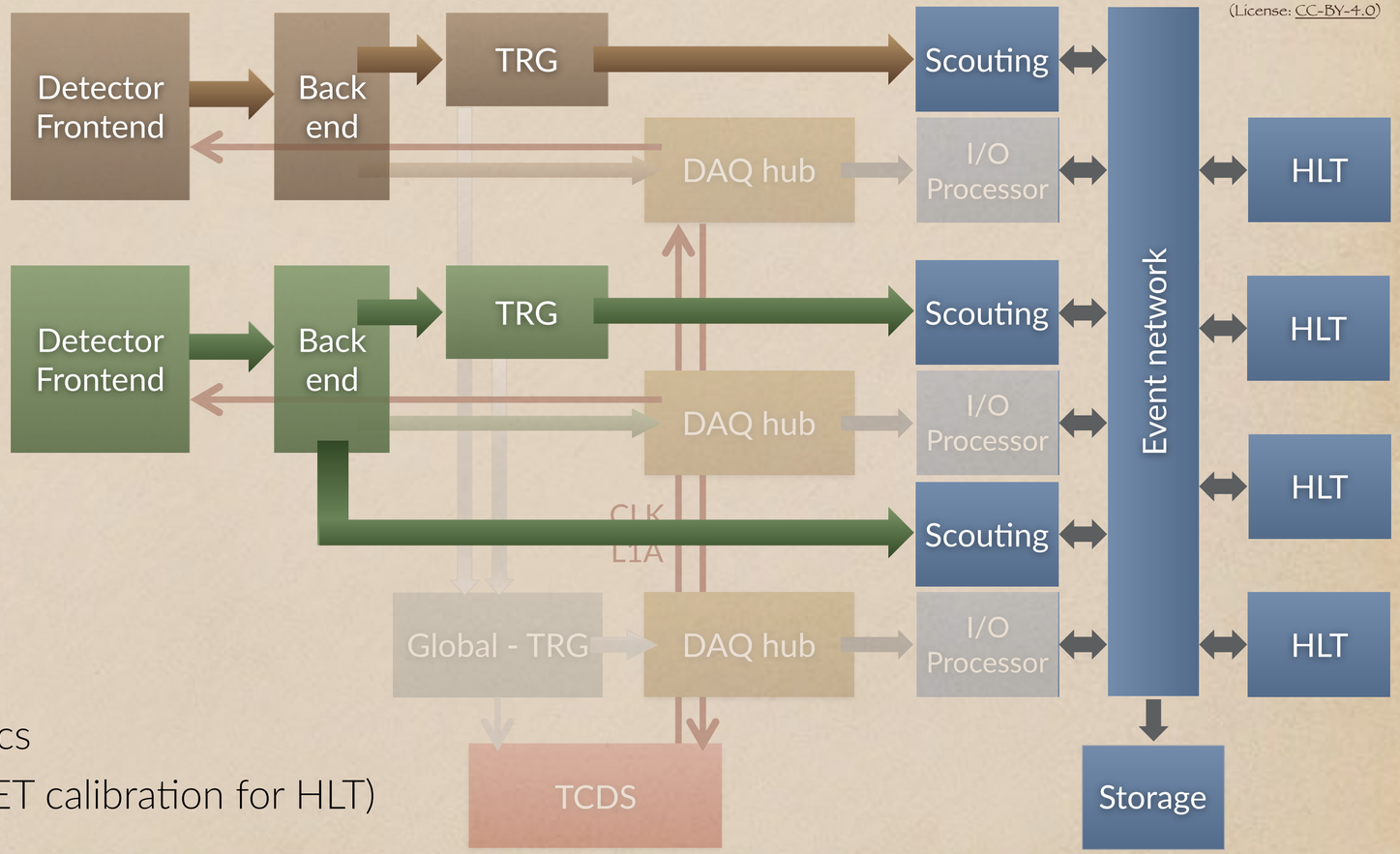
Partial acquisition at 40 MHz

- Parasitic DAQ' system w/o backpressure
- Data with reduced accuracy and/or content

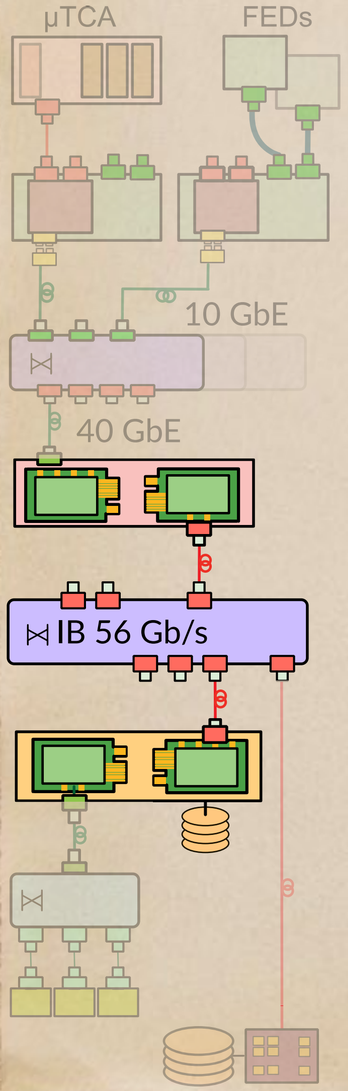
High-statistics real-time data analysis

- 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)

Study feasibility during run 3



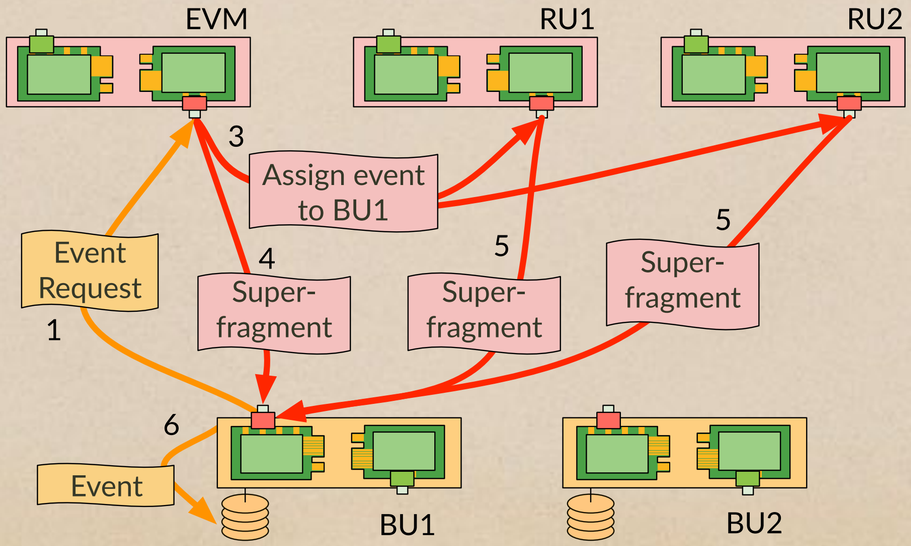
Event Builder

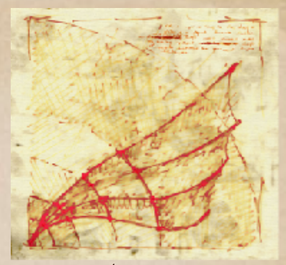


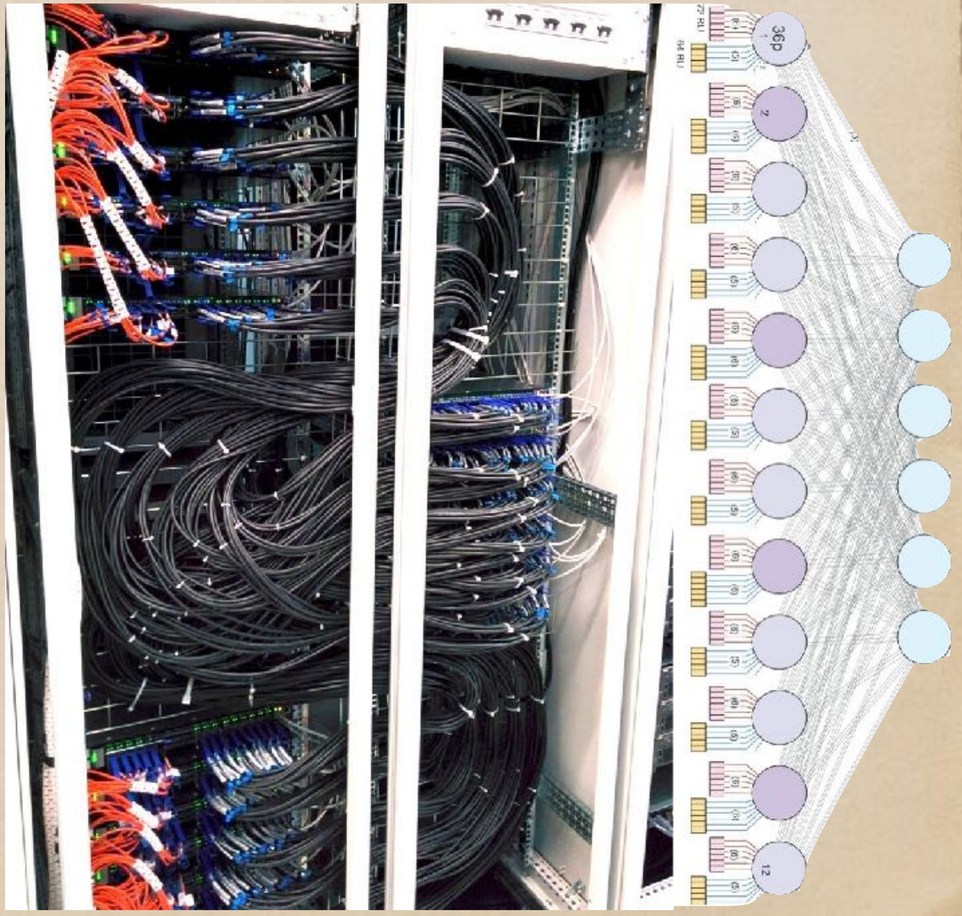
InfiniBand – most cost-effective solution

- 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