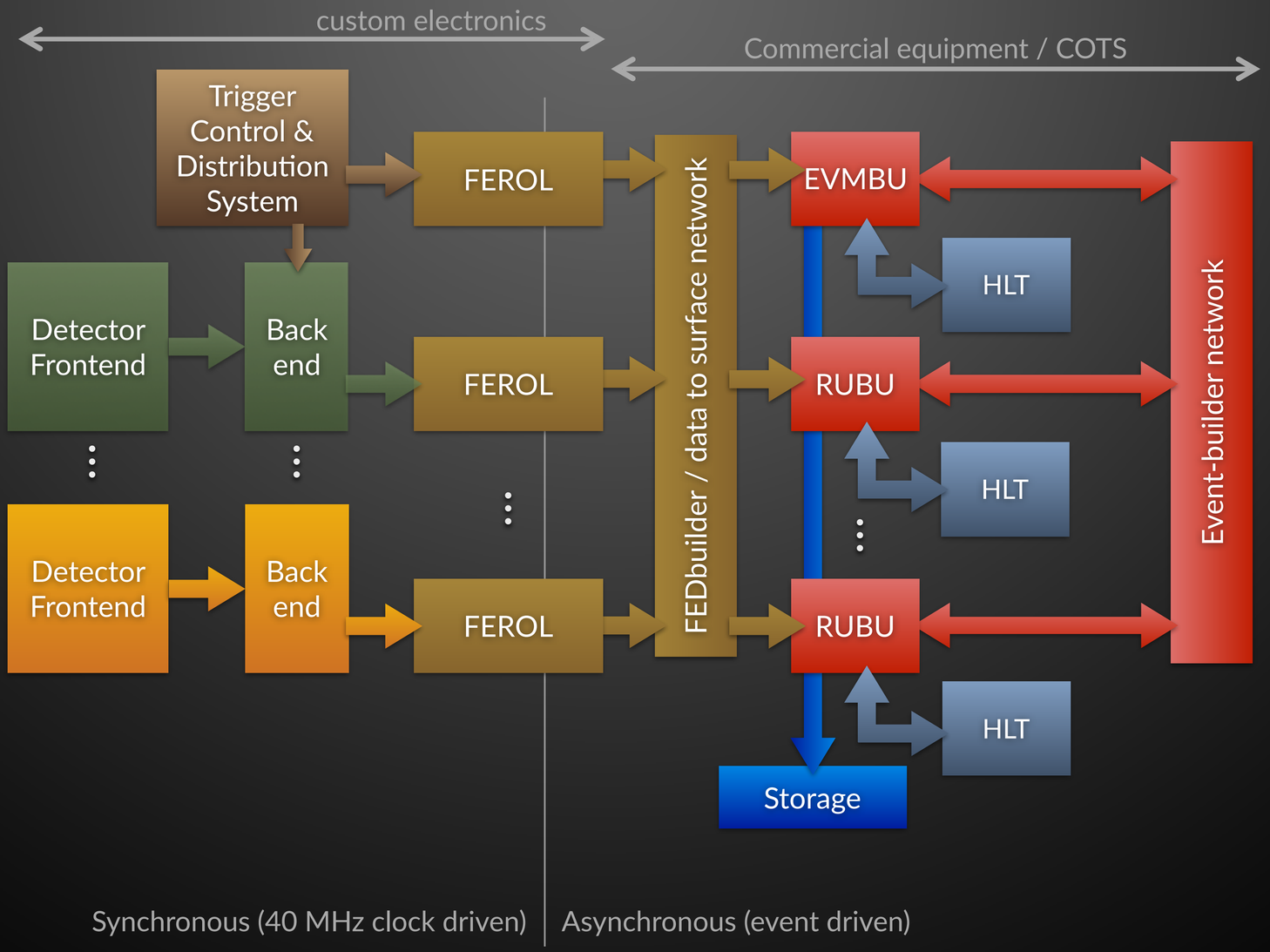
CMS Event-Builder Performance on State-of-the-Art Hardware

Remigius K Mommsen
Fermilab

on behalf of the CMS DAQ Group

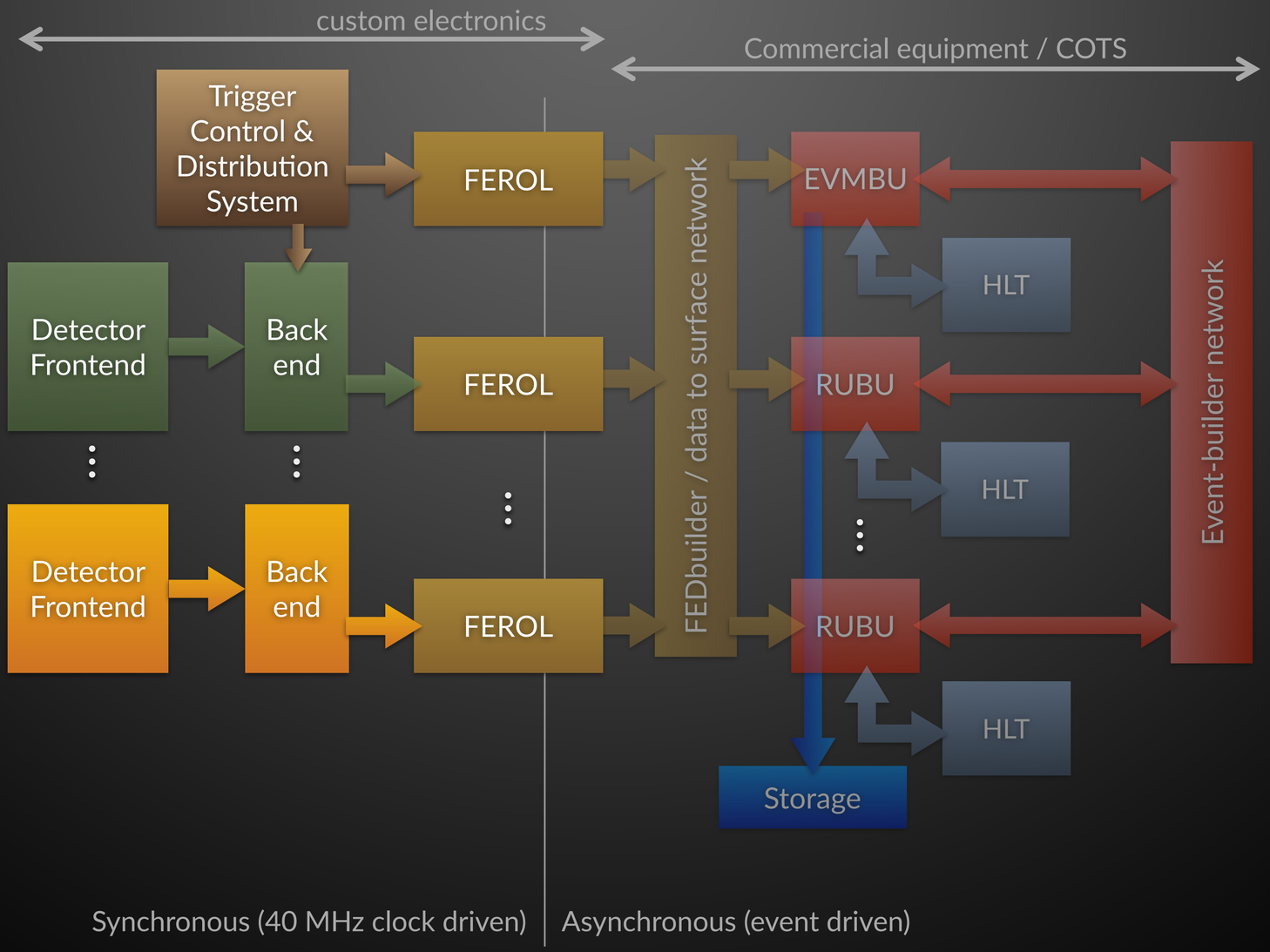
Fermilab

Overview of DAQ for LHC Run 3 (2021-2024)



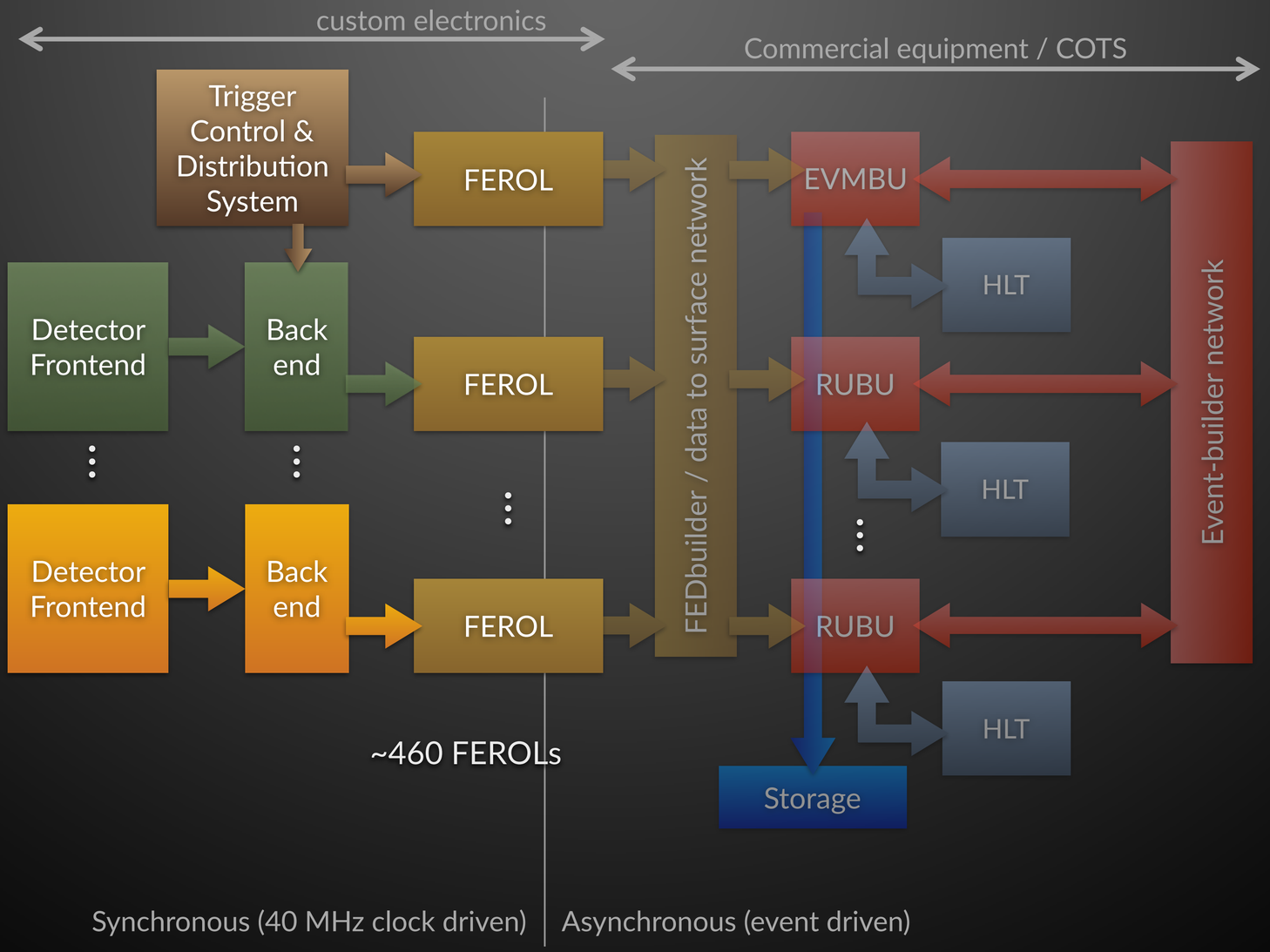
Overview of DAQ for LHC Run 3 (2021-2024)

~750 backends
0.1 - 9 kB fragments
@ 100 kHz L1 trigger rate



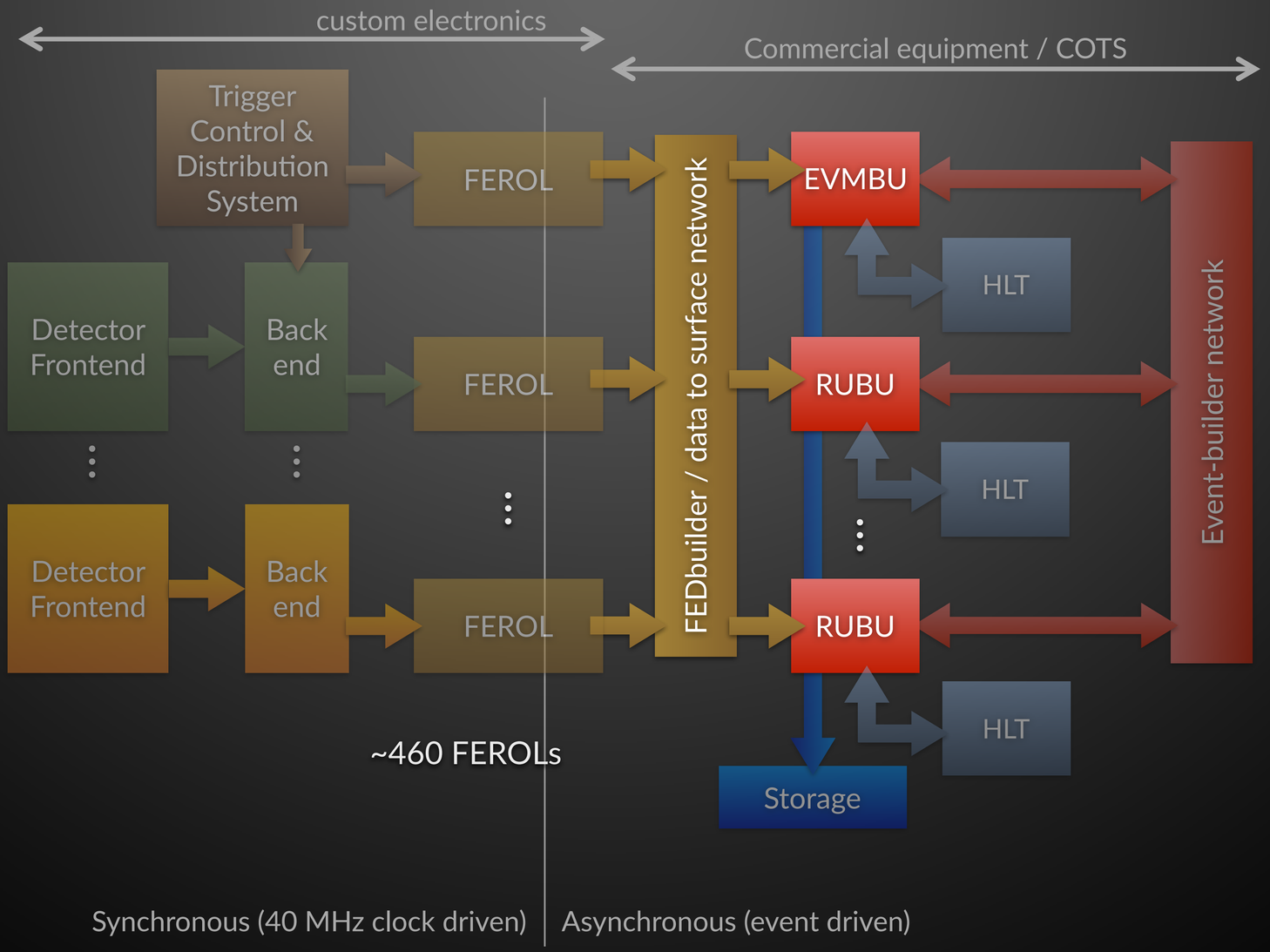
Overview of DAQ for LHC Run 3 (2021-2024)

~750 backends
0.1 - 9 kB fragments
@ 100 kHz L1 trigger rate



Overview of DAQ for LHC Run 3 (2021-2024)

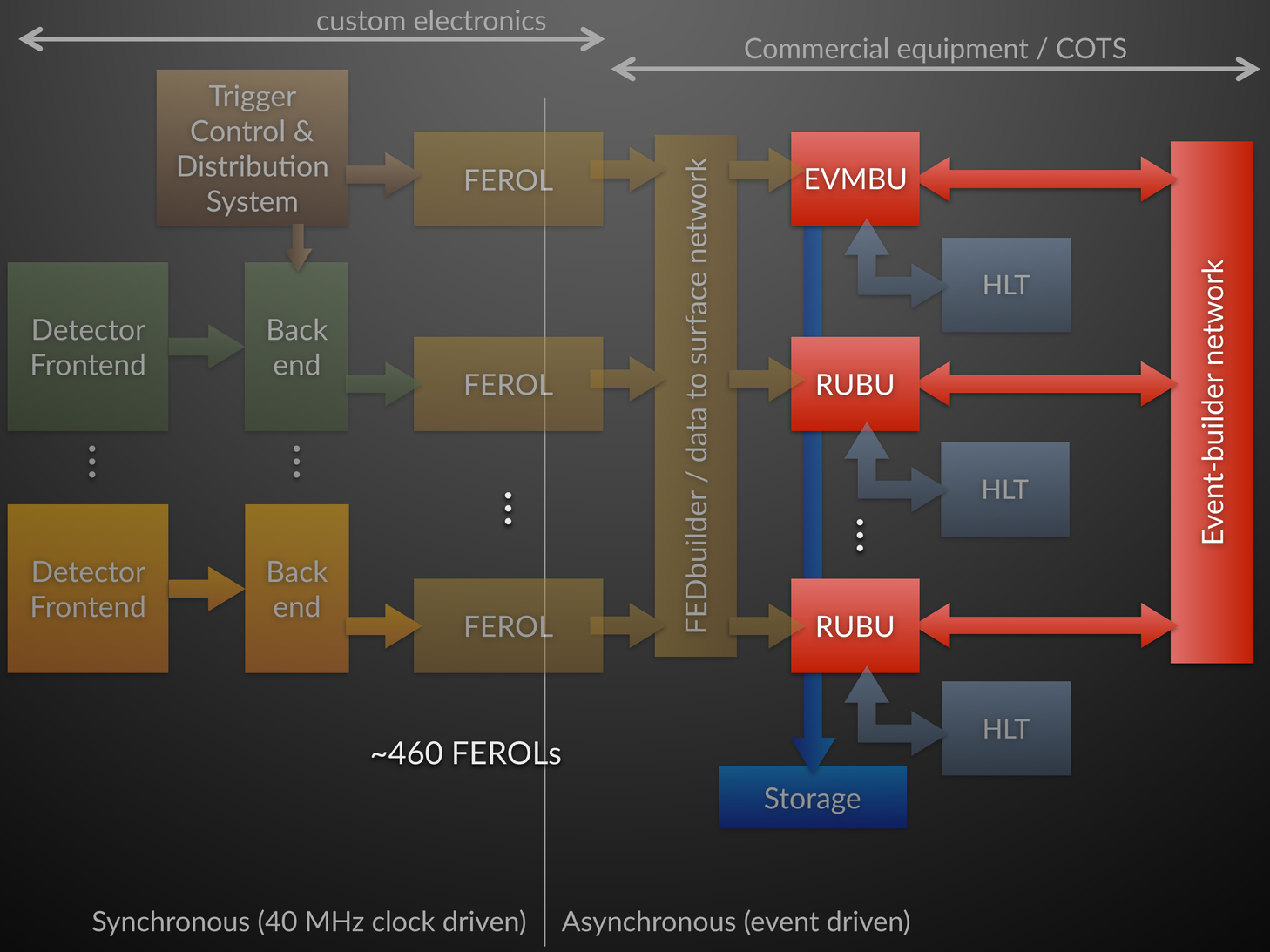
~750 backends
0.1 - 9 kB fragments
@ 100 kHz L1 trigger rate



30-50 RUBU nodes

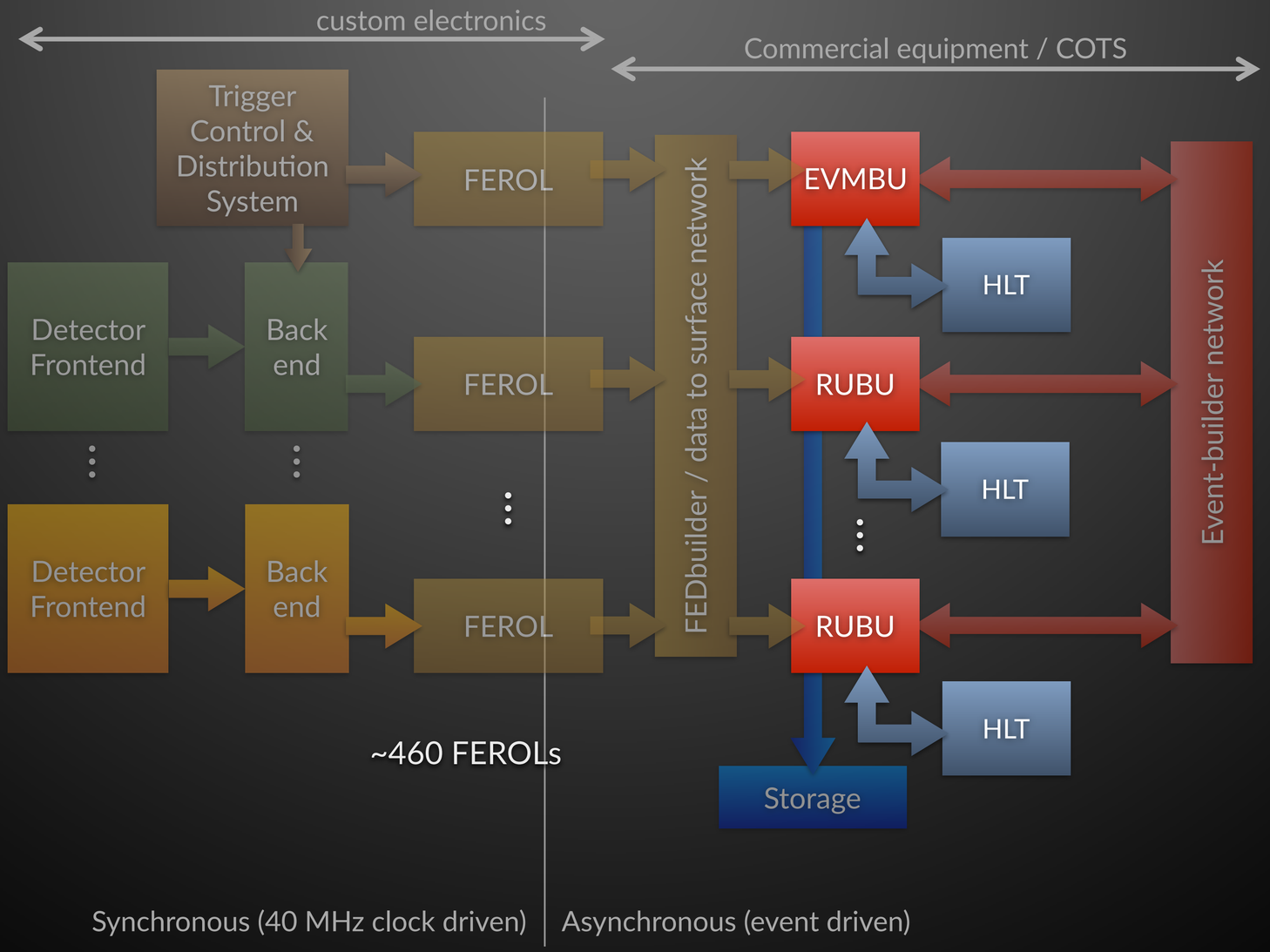
Overview of DAQ for LHC Run 3 (2021-2024)

~750 backends
0.1 - 9 kB fragments
@ 100 kHz L1 trigger rate



30-50 RUBU nodes

Overview of DAQ for LHC Run 3 (2021-2024)

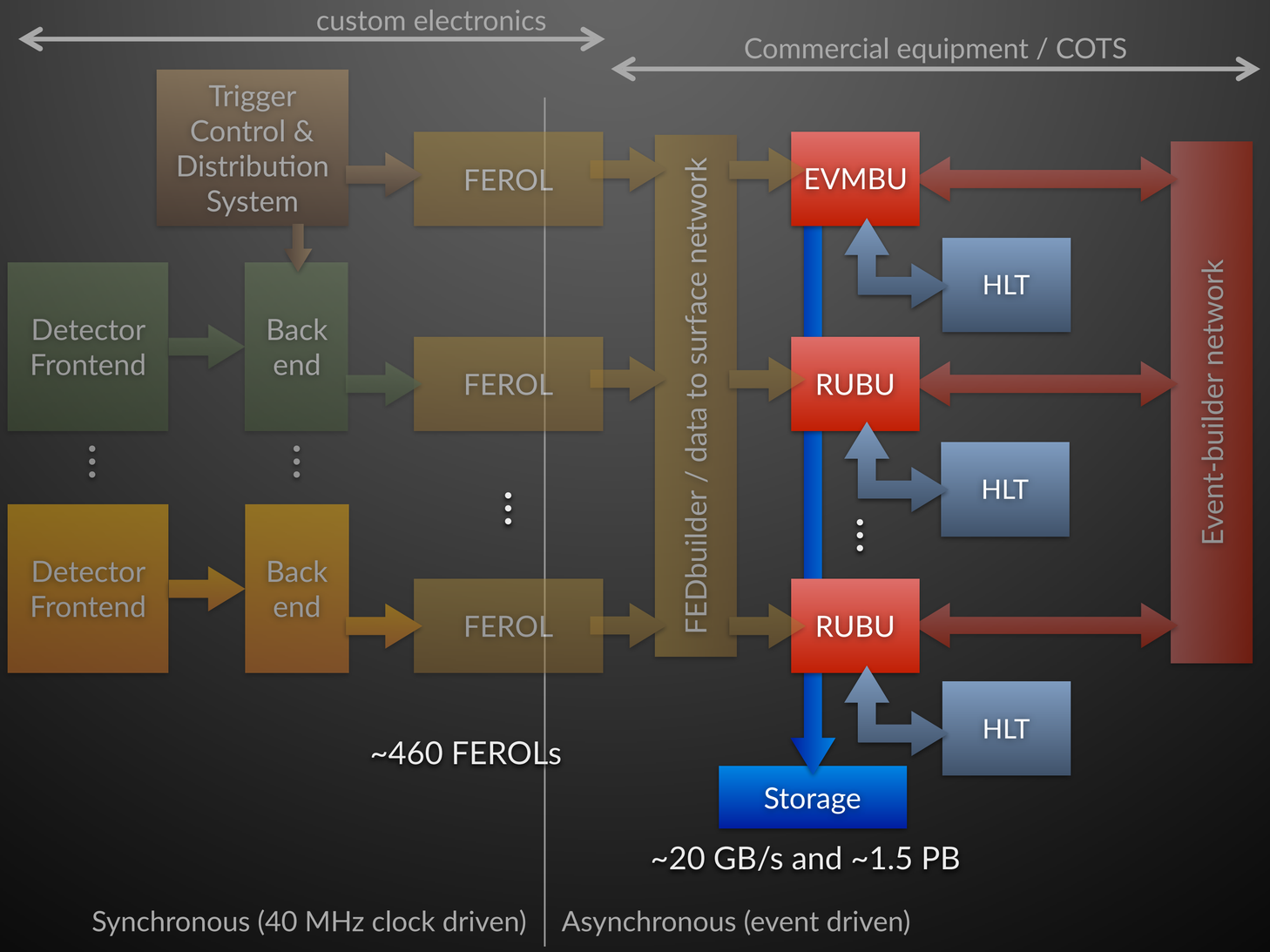


~750 backends
0.1 - 9 kB fragments
@ 100 kHz L1 trigger rate

30-50 RUBU nodes

~1100 HLT nodes
~32k physical cores
Nvidia T4 GPUs

Overview of DAQ for LHC Run 3 (2021-2024)

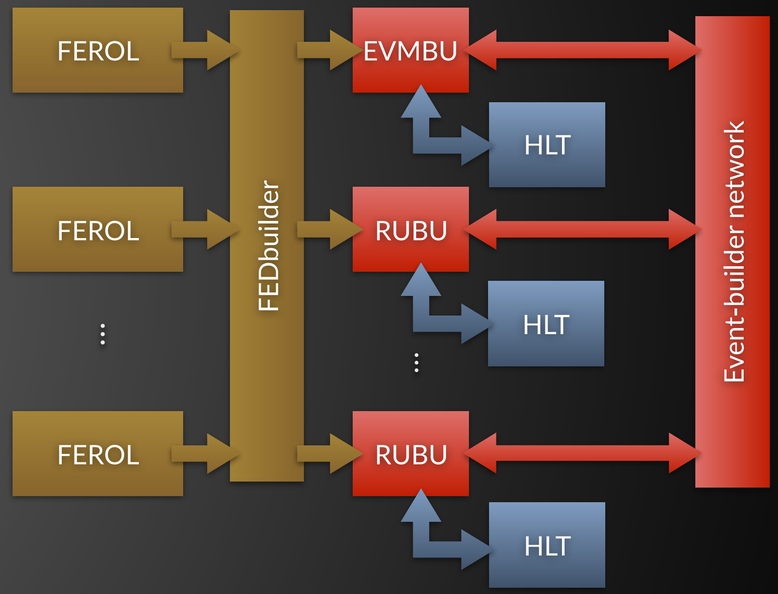


~750 backends
0.1 - 9 kB fragments
@ 100 kHz L1 trigger rate

30-50 RUBU nodes

~1100 HLT nodes
~32k physical cores
Nvidia T4 GPUs

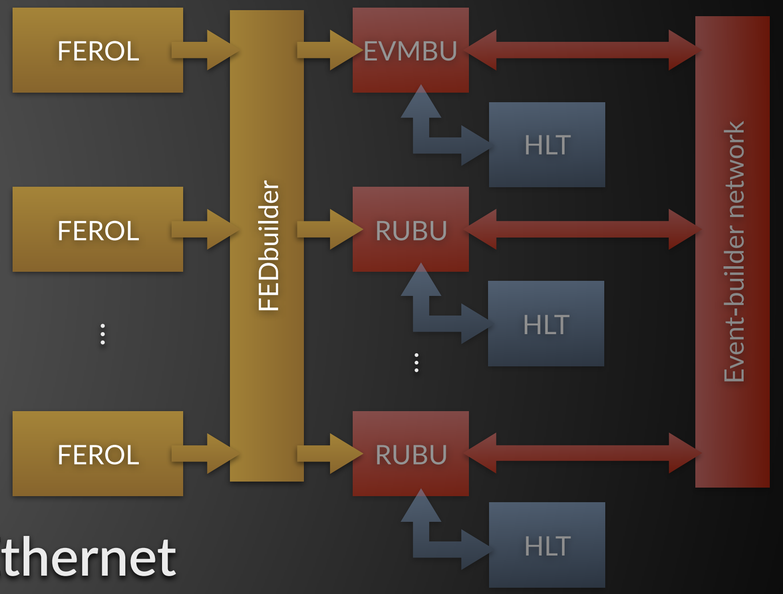
Network Technologies



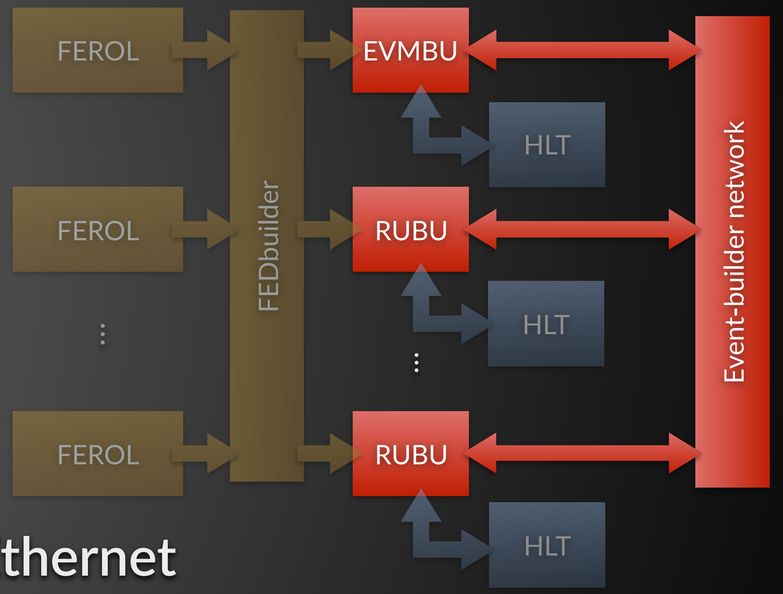
Network Technologies

FEDbuilder

- Data transferred from underground to surface over 200 m
- FEROLs use a simplified TCP/IP protocol to send data
- FEROL output streams aggregated from 10 Gbps into 100 Gbps Ethernet



Network Technologies



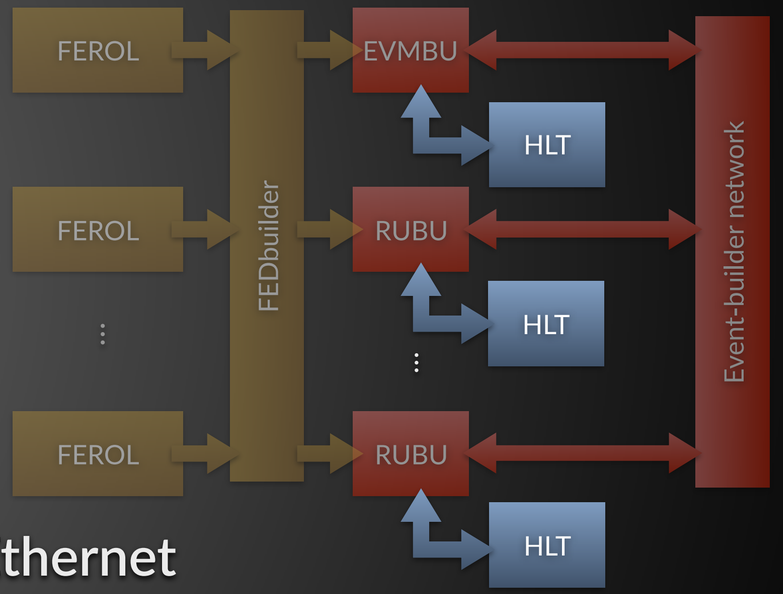
FEDbuilder

- Data transferred from underground to surface over 200 m
- FEROLs use a simplified TCP/IP protocol to send data
- FEROL output streams aggregated from 10 Gbps into 100 Gbps Ethernet

Event-builder

- Interconnects all RUBU nodes
- Used to build complete events
- 100 Gbps Infiniband (EDR) or Ethernet using RoCE (RDMA over Converged Ethernet)

Network Technologies



FEDbuilder

- Data transferred from underground to surface over 200 m
- FEROLs use a simplified TCP/IP protocol to send data
- FEROL output streams aggregated from 10 Gbps into 100 Gbps Ethernet

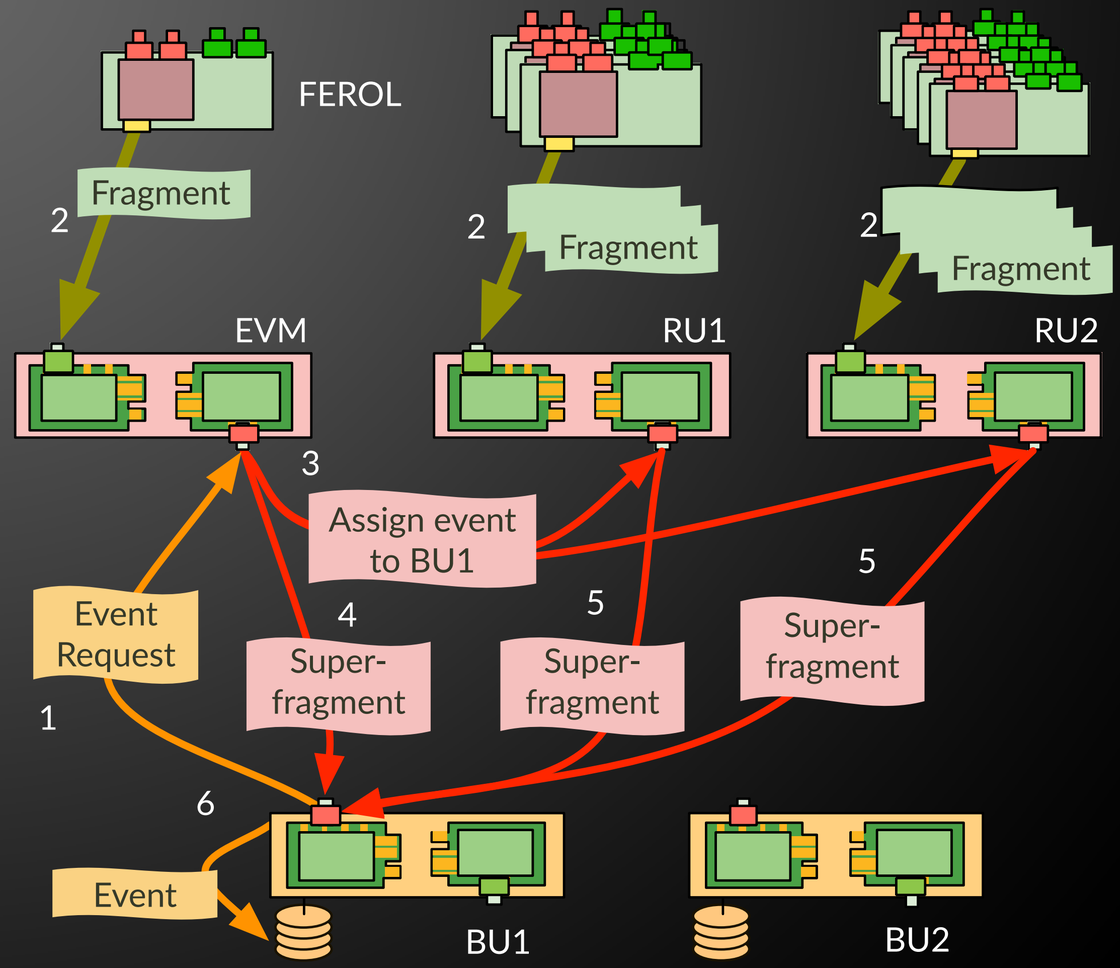
Event-builder

- Interconnects all RUBU nodes
- Used to build complete events
- 100 Gbps Infiniband (EDR) or Ethernet using RoCE (RDMA over Converged Ethernet)

High-Level Trigger (HLT)

- Interconnects RUBU node to a set of HLT nodes running the event selection
- Events read and written back using NFS
- 100 Gbps fanned out to 10 Gbps Ethernet network

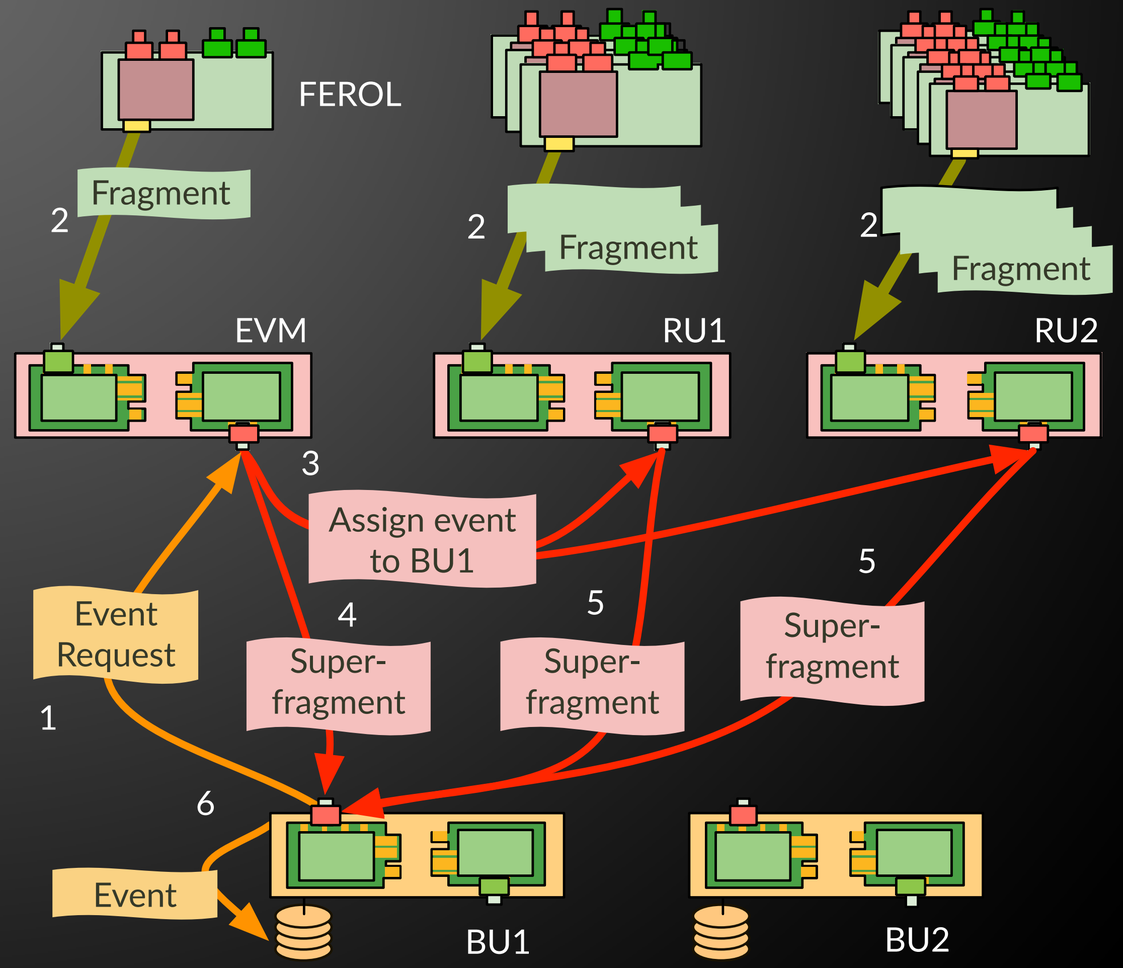
Event-Builder Protocol



Event-Builder Protocol

Event Manager — EVM

- Orchestrates the event building
- Receives the master fragment from TCDS



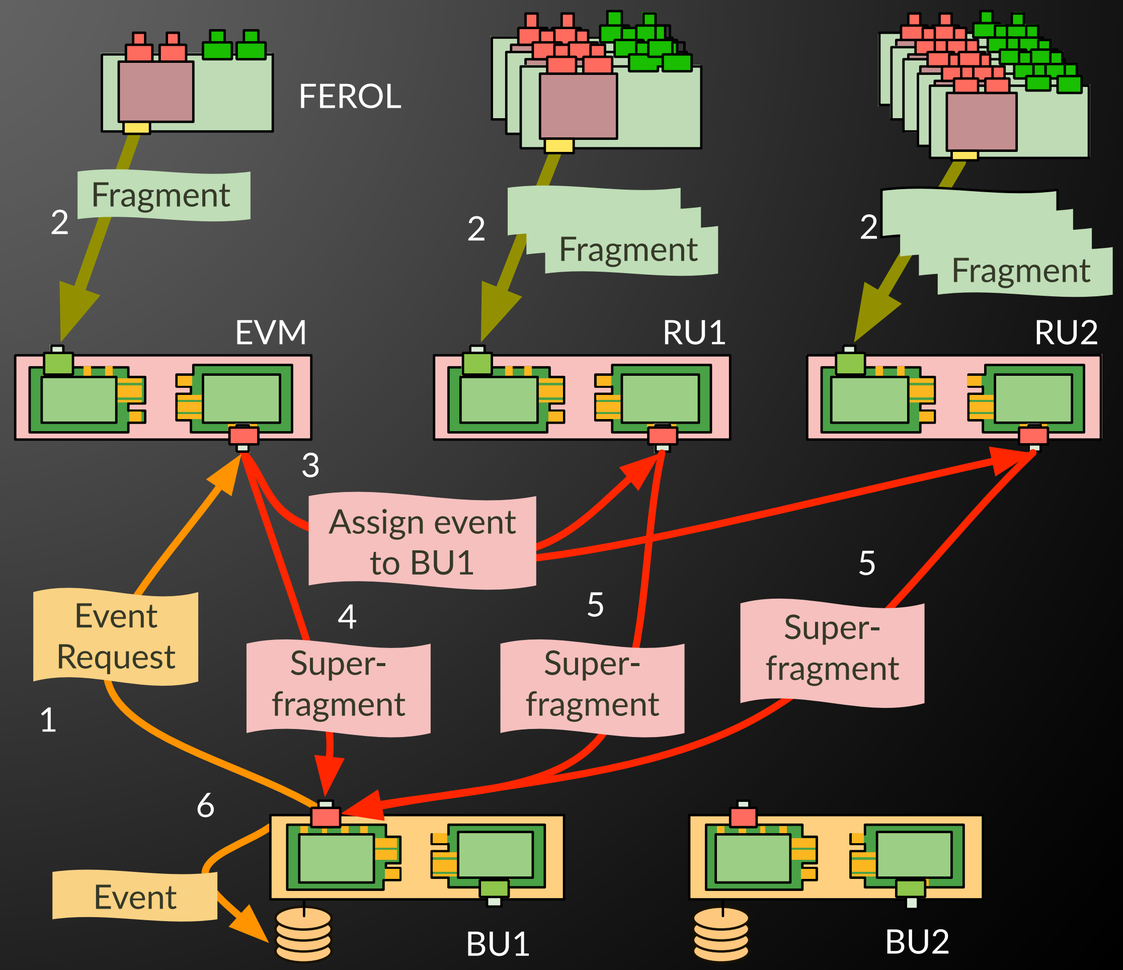
Event-Builder Protocol

Event Manager — EVM

- Orchestrates the event building
- Receives the master fragment from TCDS

Readout Unit — RU

- Receive TCP/IP streams from FEROLs
  - Number of streams depend on fragment sizes
- Checks data integrity (CRC) and sequence
- Buffers fragments and combines them into larger messages (super fragments)



Event-Builder Protocol

Event Manager — EVM

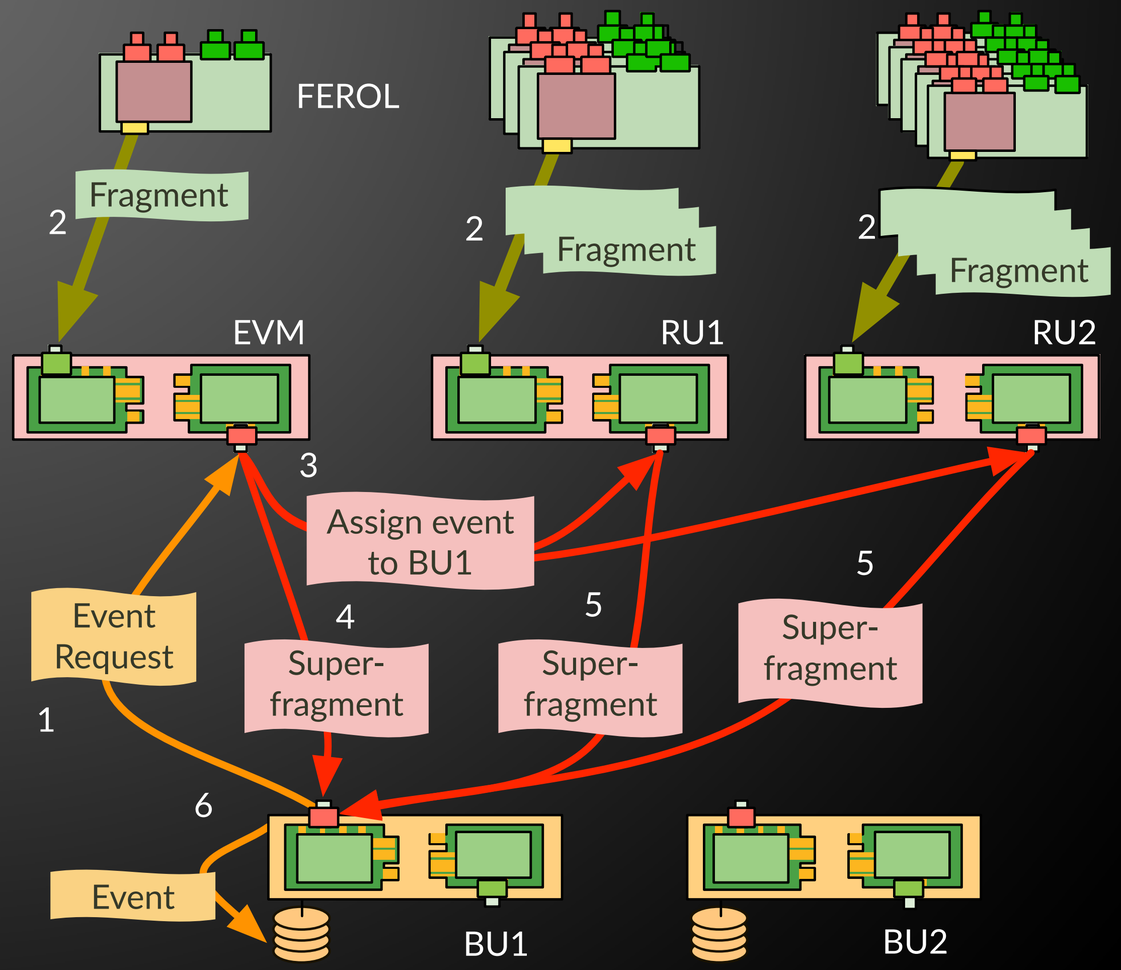
- Orchestrates the event building
- Receives the master fragment from TCDS

Readout Unit — RU

- Receive TCP/IP streams from FEROLs
  - Number of streams depend on fragment sizes
- Checks data integrity (CRC) and sequence
- Buffers fragments and combines them into larger messages (super fragments)

Builder Unit — BU

- Builds complete events
- Checks consistency of event
- Writes events to files on local RAM disk



DAQ3 Testbed (daq3val)

Readout/Builder Unit (RUBU)

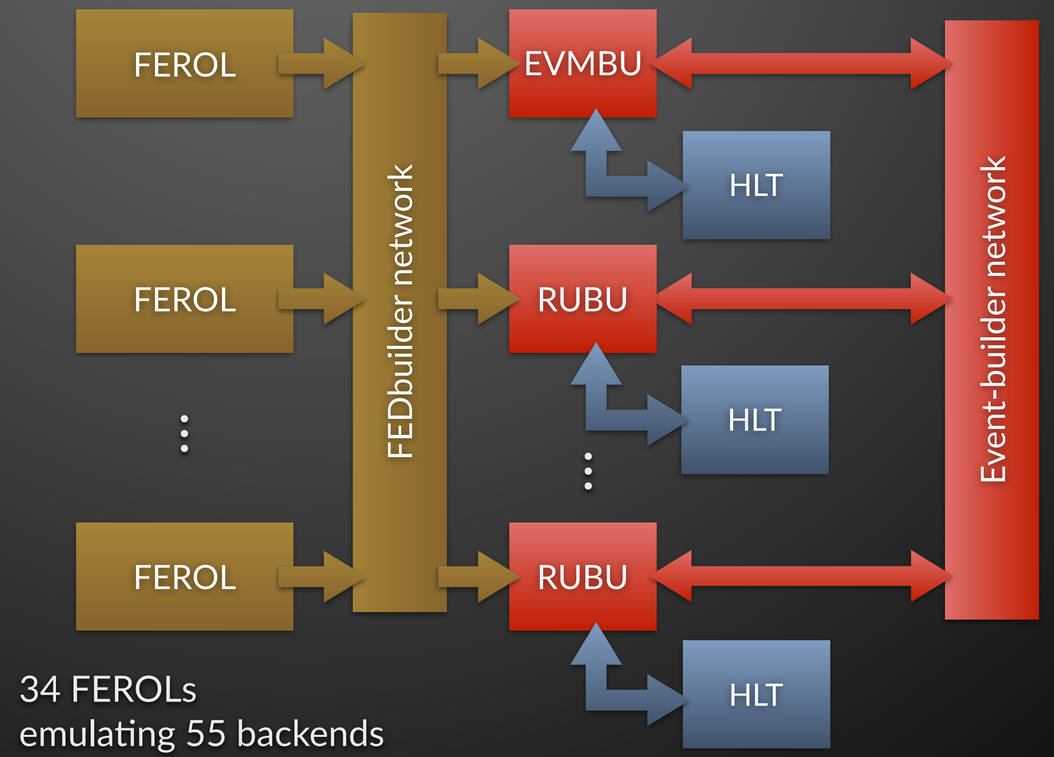
- Dell R740 dual skylake & 288 GB RAM
  - 14 RUBUs with 16-cores @ 2.1 GHz
  - 2 RUBUs with 14-cores @ 2.6 GHz
- 1 Dell R6515 with single AMD 7502P (EPYC ROME) @ 2.5 GHz & 251 GB RAM
- ConnectX-5 NICs (dual and single port)

Ethernet networks

- Juniper QFX10002-72Q
  - 288 x 10 Gbps or 24 x 100 Gbps
  - 16 GB buffer
- Juniper QFX5200-32C
  - 32 x 100 Gbps
  - 16 MB buffer

Infiniband EDR

- Mellanox SB7800
- 36 x 100 Gbps



34 FEROLs
emulating 55 backends

16 RUBU nodes

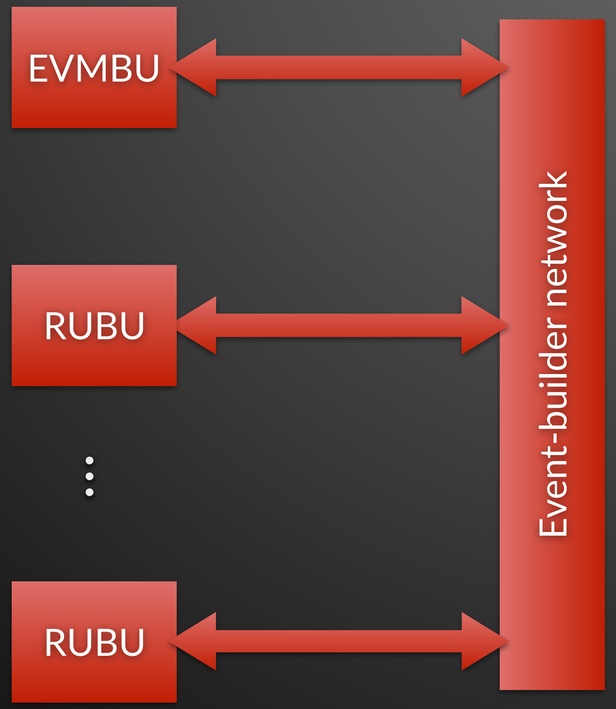
80 HLT nodes
2560 physical cores
(w/o GPUs)

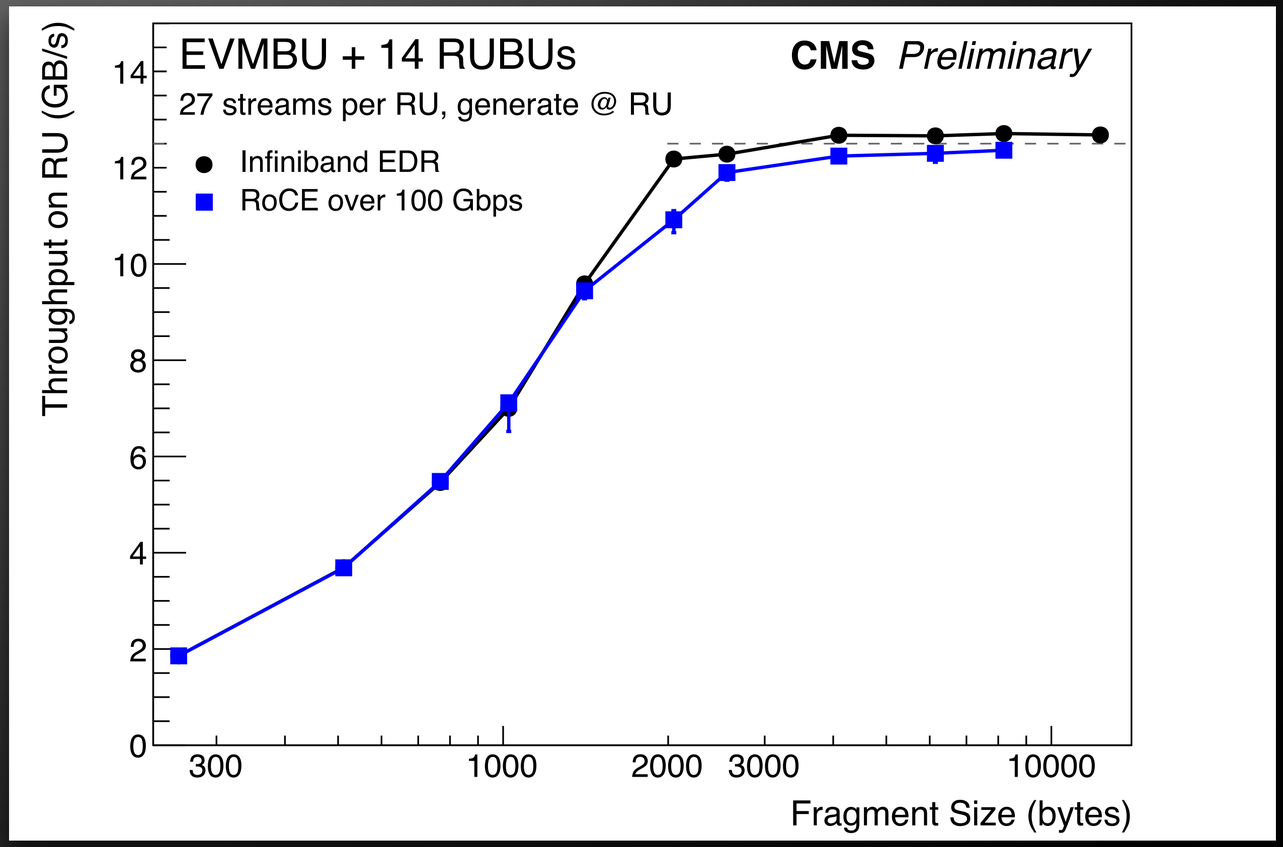
no Storage

All to All

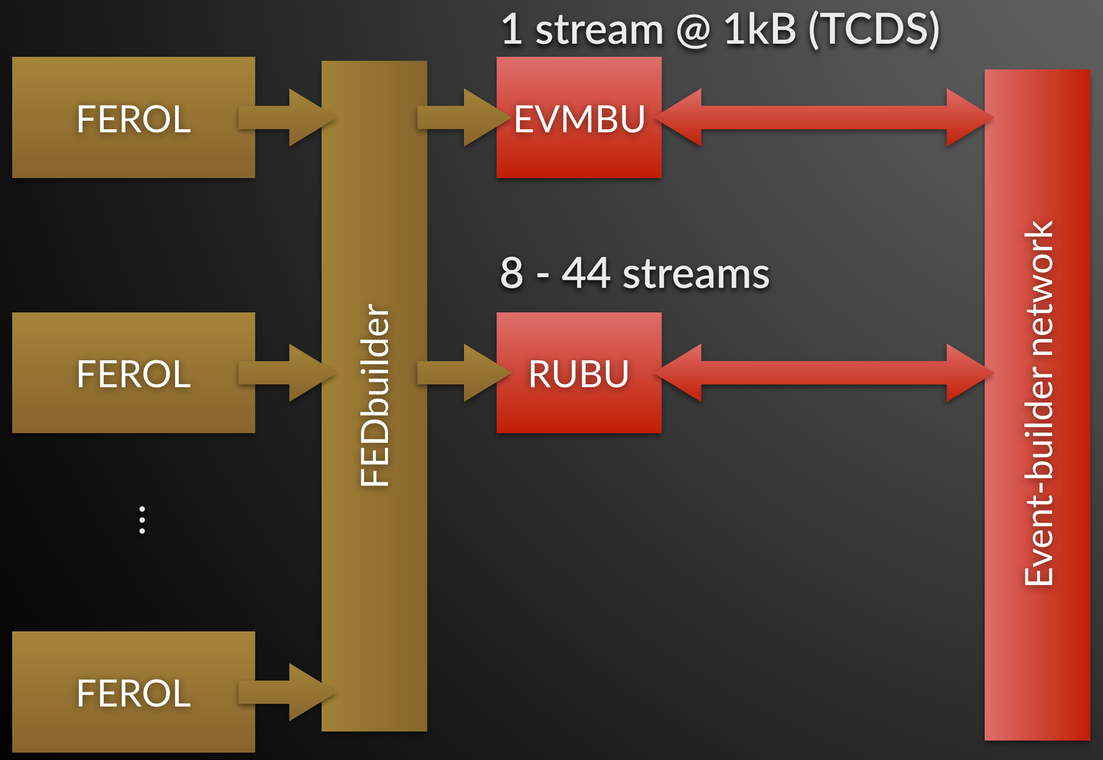
Emulate 1 fragment with 1kB (TCDS)

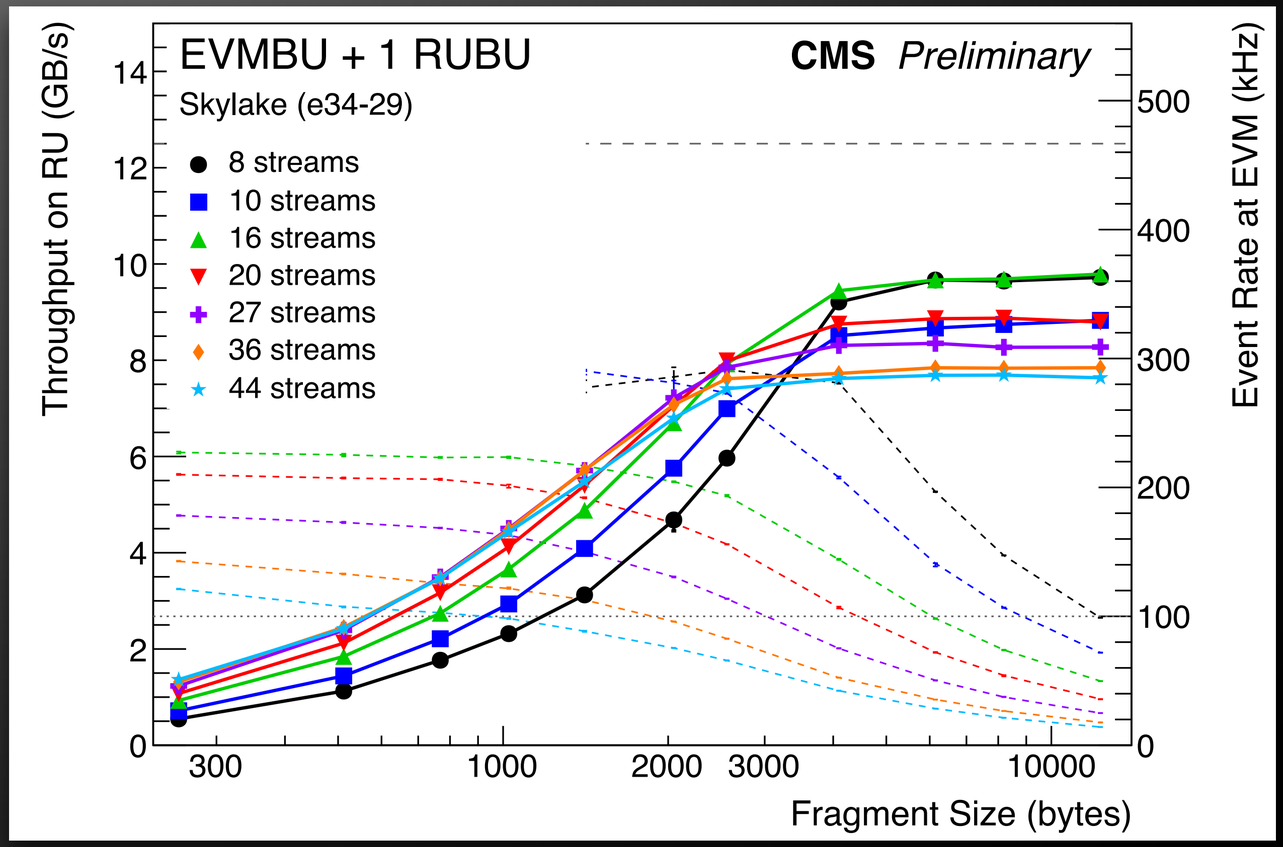
Emulate 27 fragments on 14 RUs



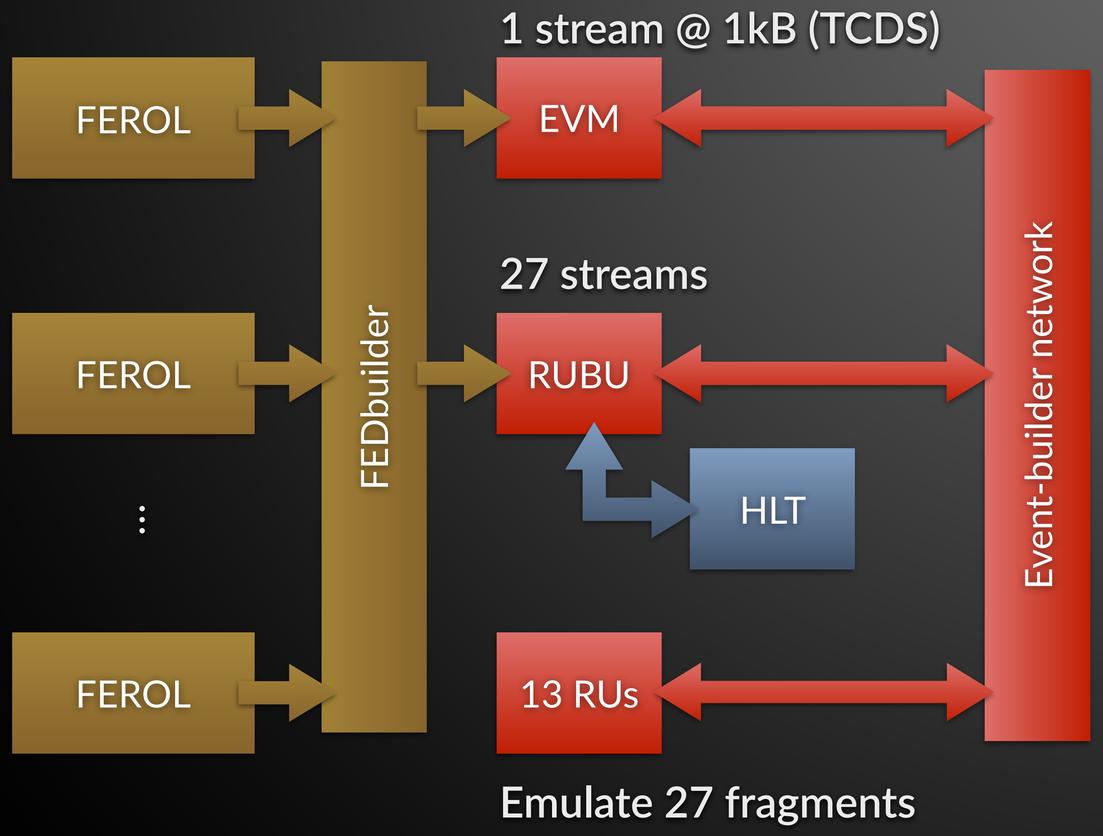


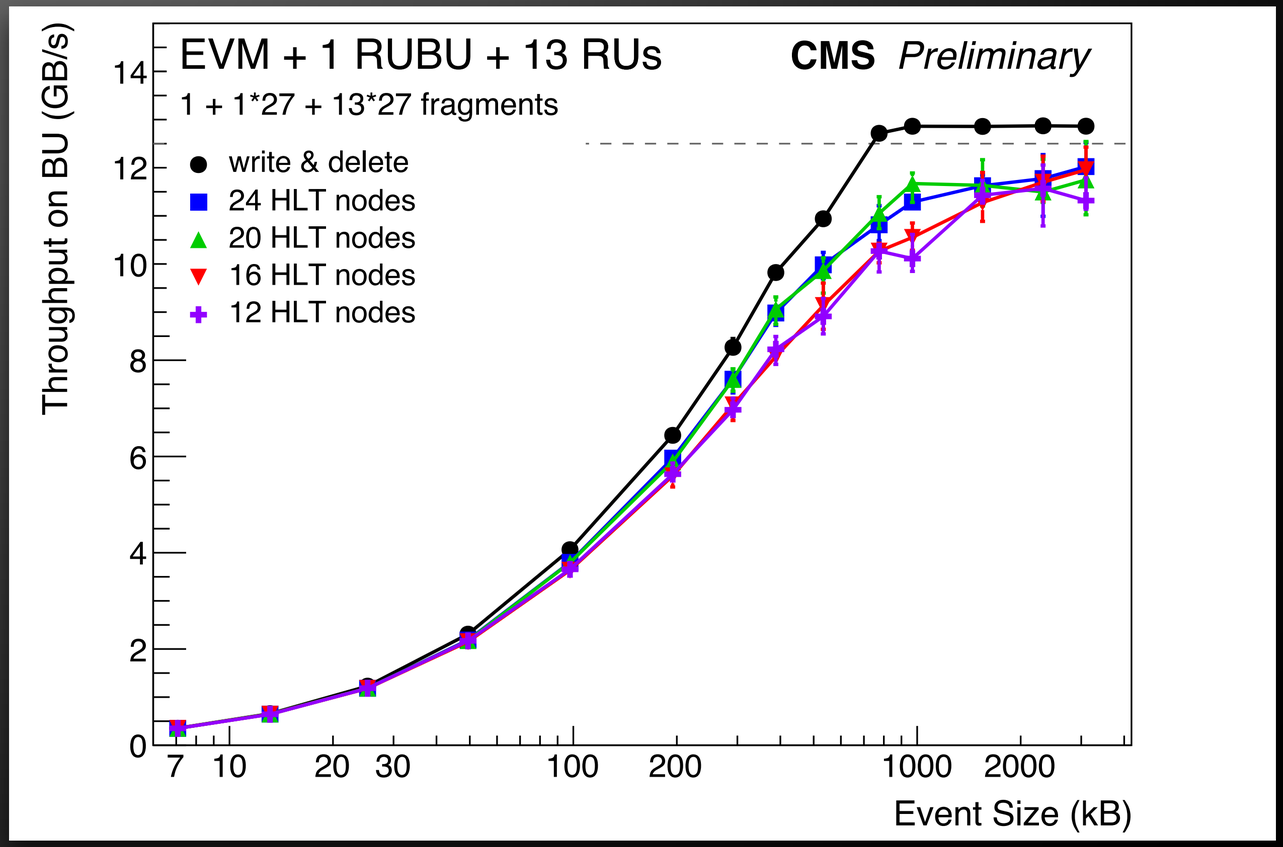
Scaling Number of Input Streams





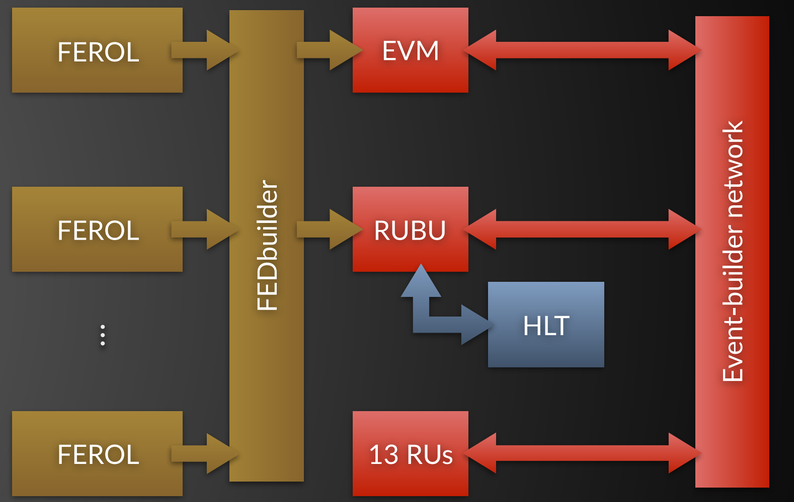
Performance of One RUBU Node





A HLT node runs 6 HLT processes with 4 threads each

Throughput of Disk Writing



Events are written to files on a RAM disk

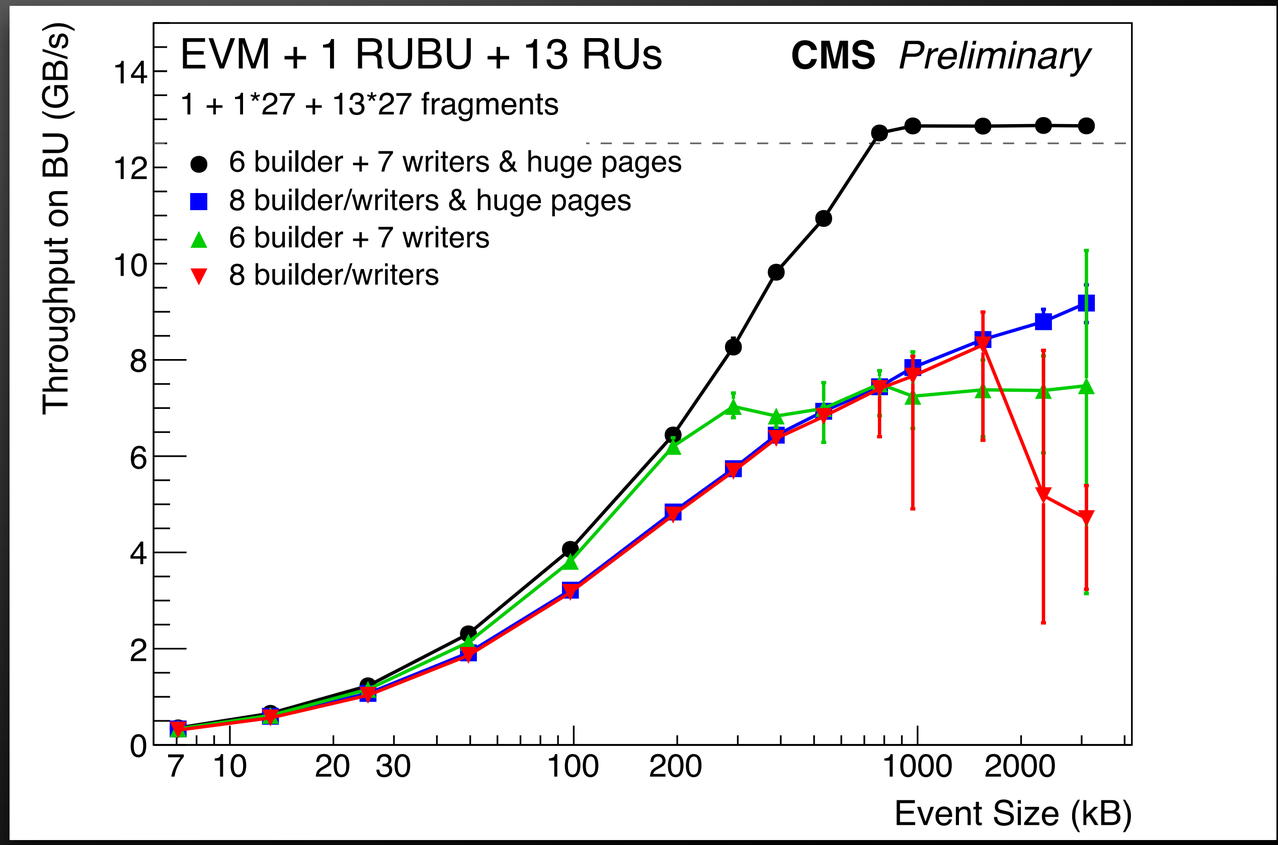
- Several threads build and write events concurrently into separate files
- Each file contains 100 events

Throughput is limited by size of memory pages

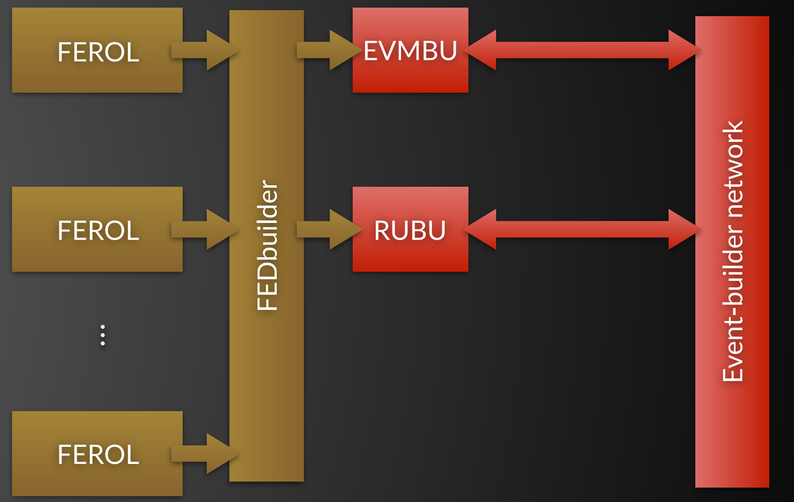
- Move to kernel 4 which supports huge memory pages for RAM disk

Further improvement by splitting event building and disk writing task into separate threads

- 6 builder threads
- 7 writer threads



Cache Misses

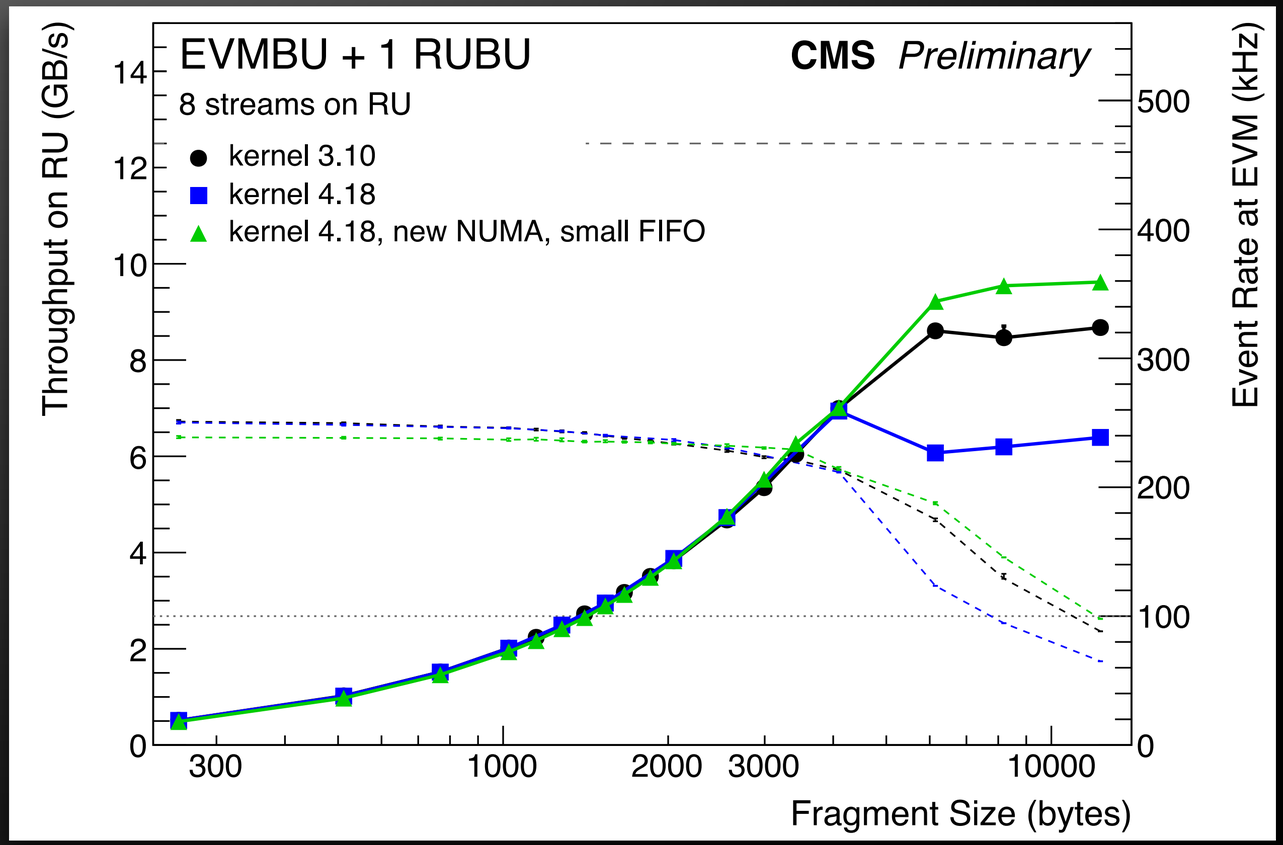


Observed large performance drop when moving from kernel 3.10 to kernel 4.18

- TCP buffer handling in kernel has changed
- Event fragments no longer in cache when re-calculating CRC in input stage

Improved performance

- Tuning of NUMA settings to assure locality
- Reduced FIFO length between input and fragment handling code



AMD Architecture

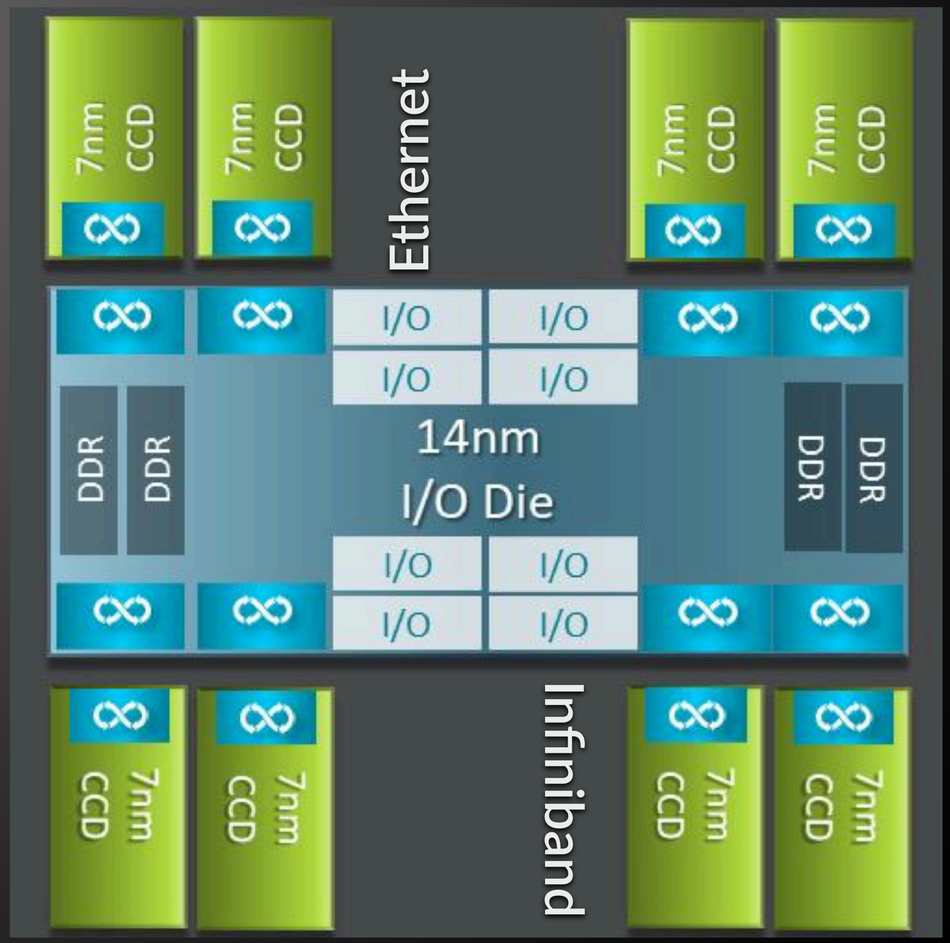
2nd generation EPYC architecture (“Rome”)

- Multiple dies interconnected by AMD Infinity Fabric
- I/O die with 128 PCIe 4.0 lanes
- 8 memory channels @ 3.2 GHz

Single socket CPU provides up to 64 cores

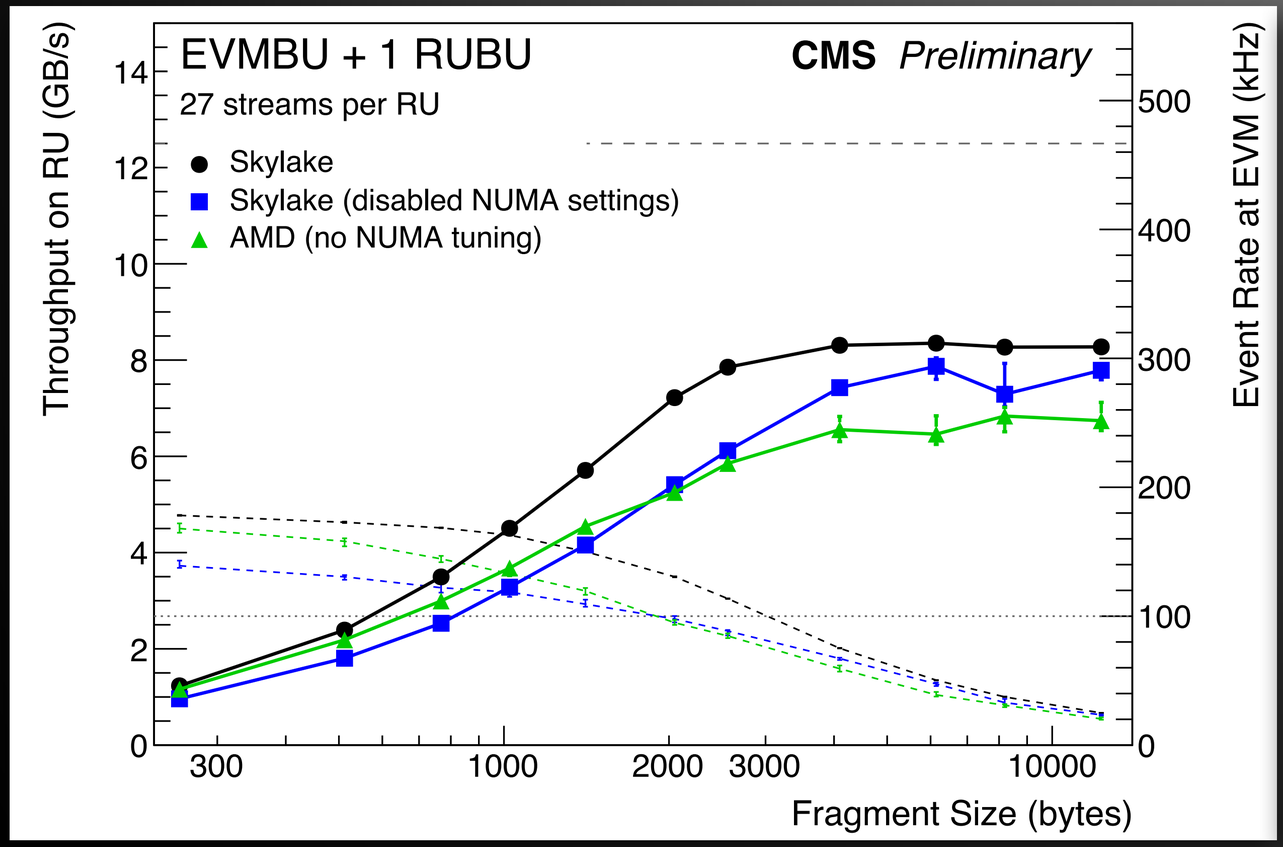
- No need to use dual CPUs as for Intel
- Less worry about NUMA as all I/O on single socket
- Still need to deal with NUMA inside chip

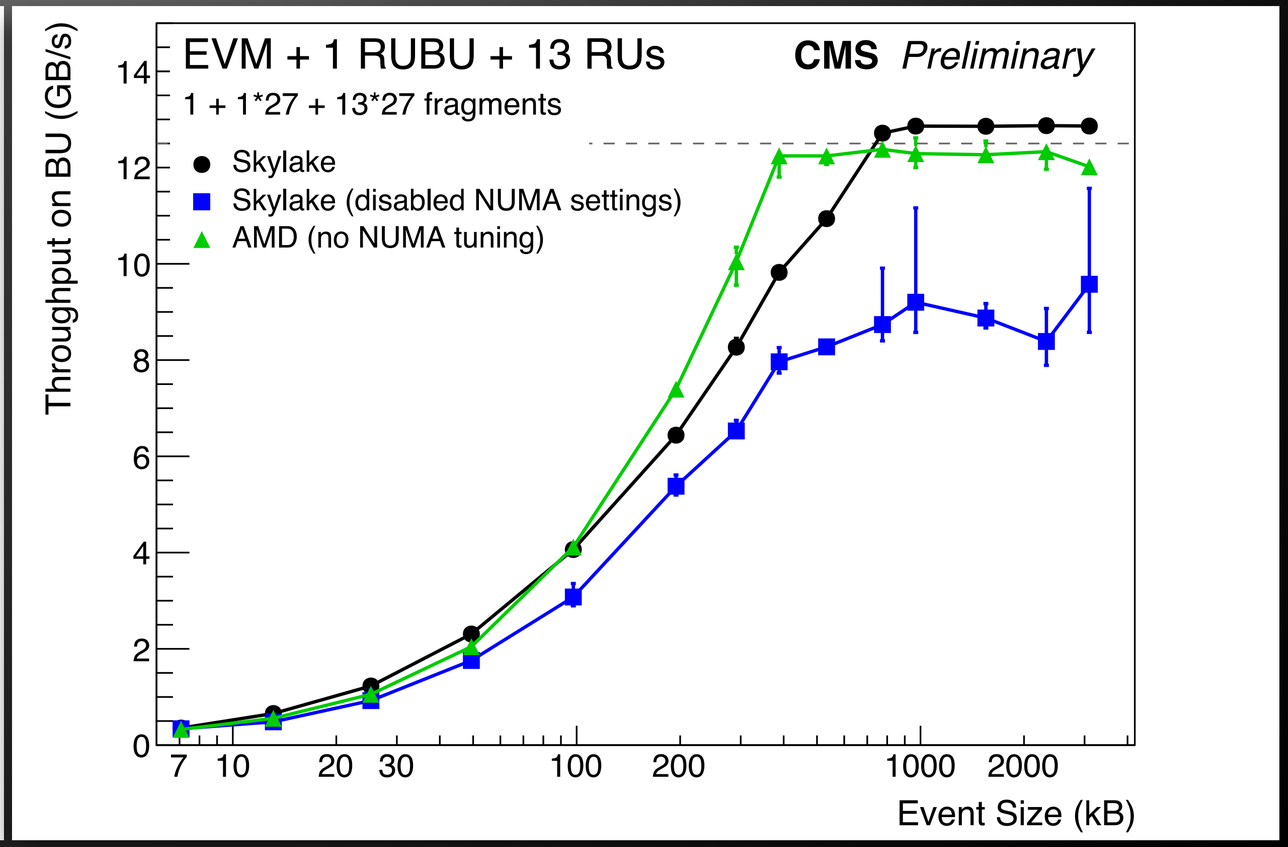
CCD = 4 CPU cores & caches:
64 kB L1, 512 kB L2 & 16 MB L3

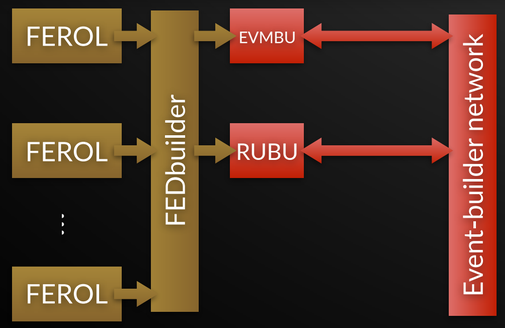


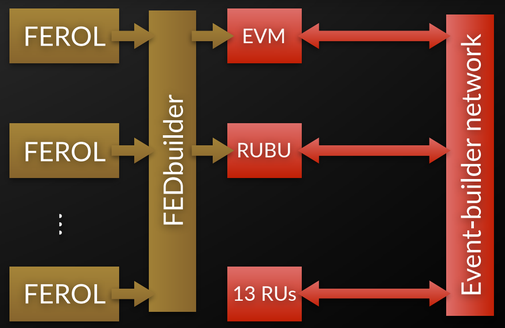
https://www.amd.com/system/files/documents/TIRIAS-White-Paper-AMD-Infinity-Architecture.pdf

AMD vs Intel Skylake









Summary

Data acquisitions system for LHC run 3 will be based on recent hardware

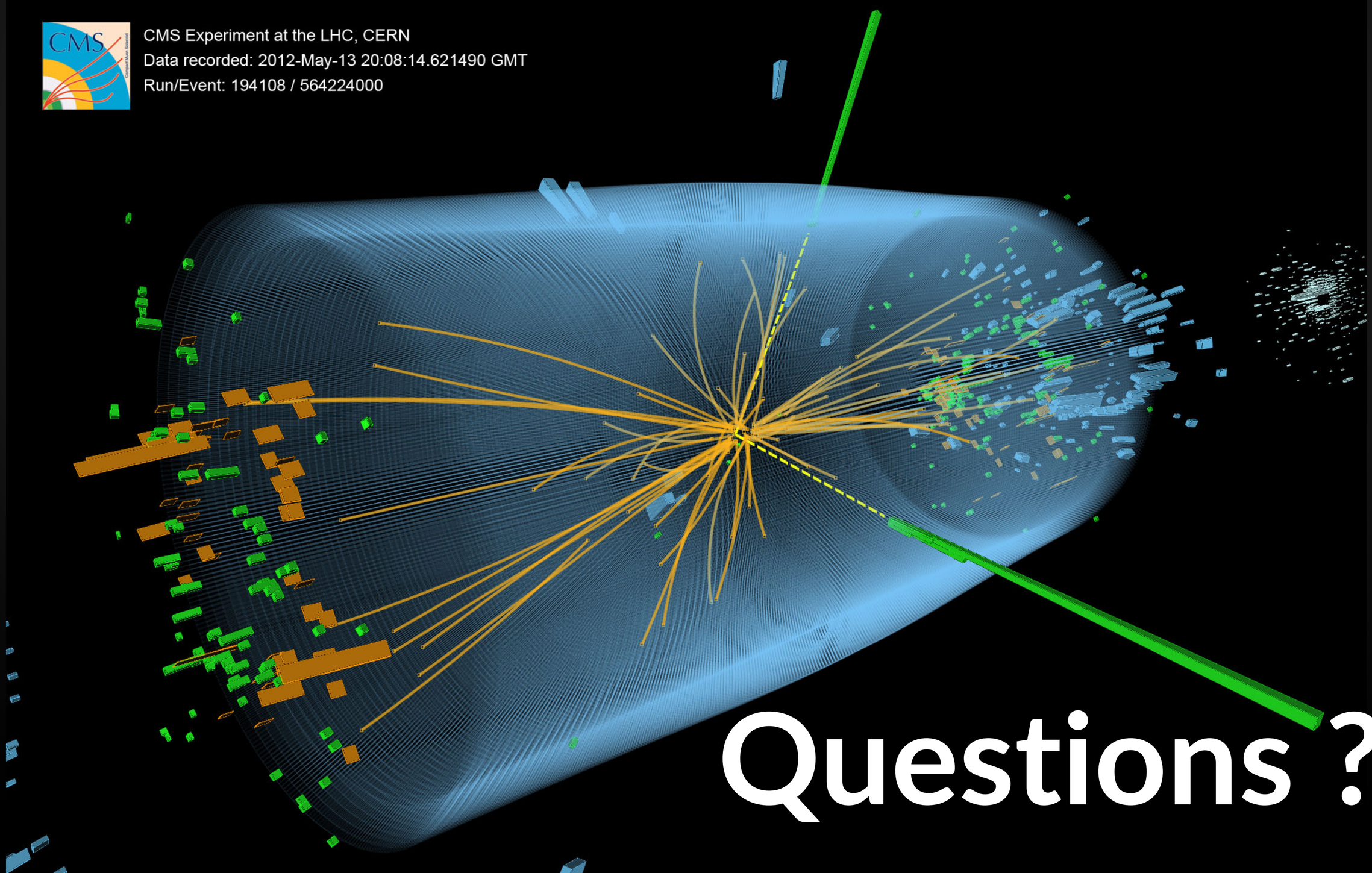
- 100 Gbps Ethernet Networks with TCP/IP and RoCE
- RoCE offers similar performance at lower cost than Infiniband

High-end servers with 30-40 cores can handle TCP/IP, event-builder traffic and NFS

- Allows to reduce the DAQ system by a factor ~4 compared to the run 2 system
- Throughput limited by software overhead in handling fragment rate from TCP streams
- AMD single-socket EPYC ROME is an interesting alternative, but requires more studies

Kernel 4.18 required to meet performance goals

- Huge memory pages for RAMdisk are a game changer
- Some interesting surprises encountered along the way


CMS Experiment at the LHC, CERN
Data recorded: 2012-May-13 20:08:14.621490 GMT
Run/Event: 194108 / 564224000
Questions ?