Evolution of the energy efficiency of LHCb's real-time processing

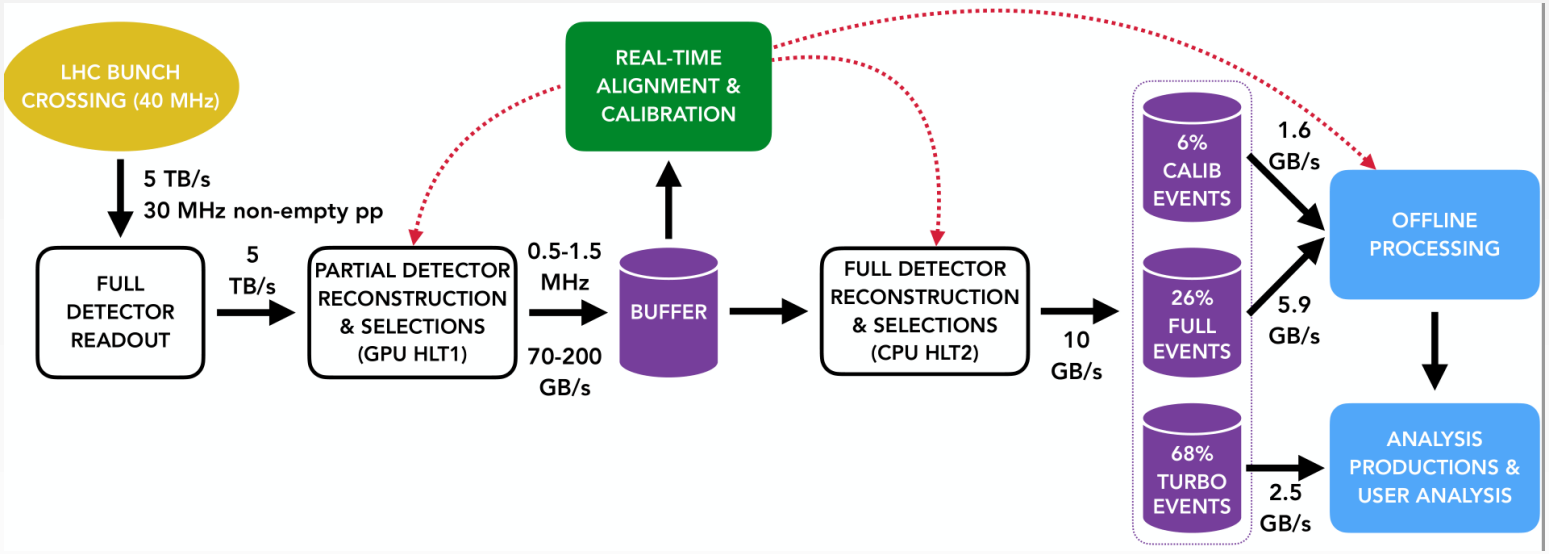
Tommaso Colombo, Vladimir Gligorov, Arthur Hennequin,
Niko Neufeld, Rainer Schwemmer

vChep 2021

Motivation

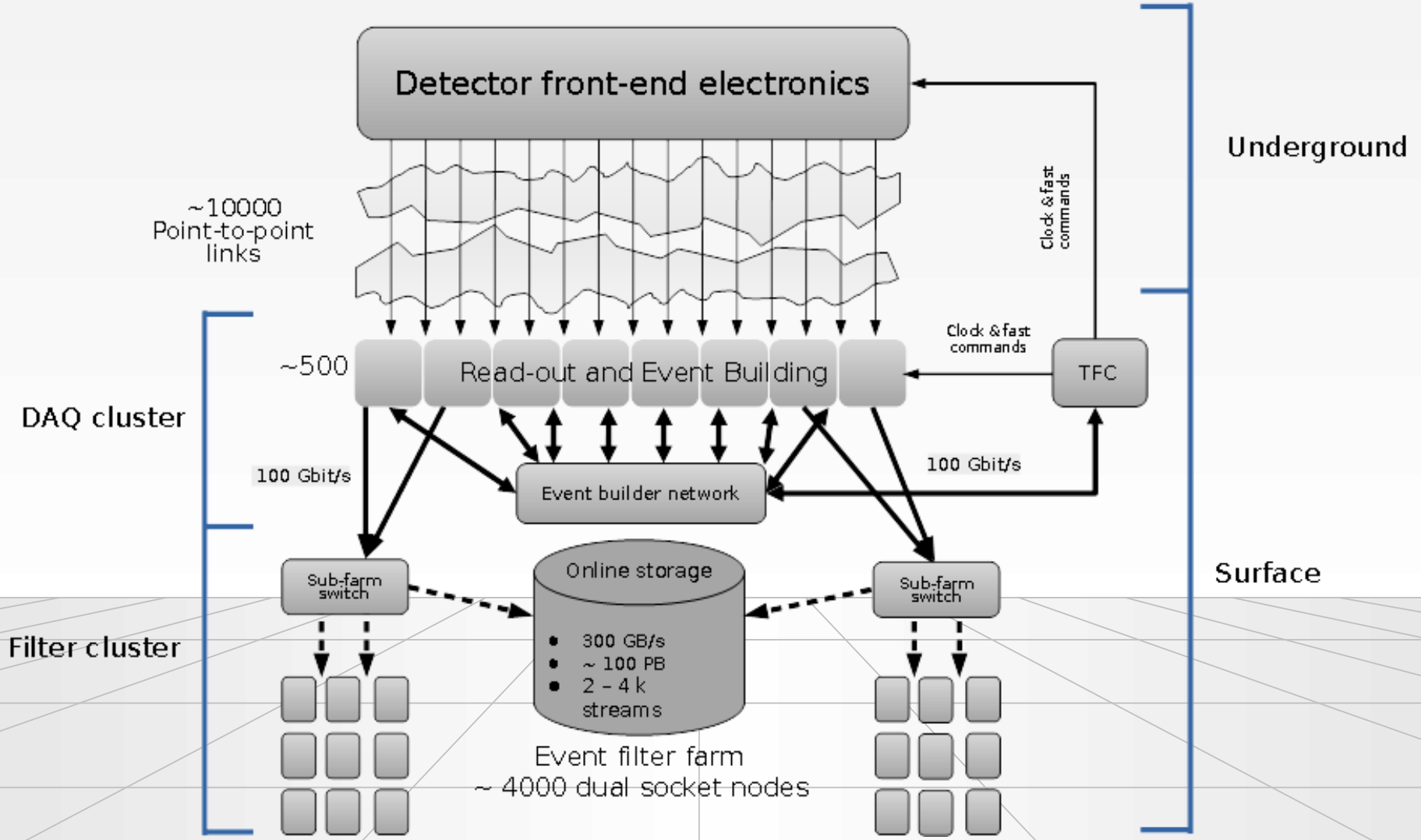
- Power efficiency is a current goal for all major data center providers
  - Global Warming / Agenda 2030
  - Subsequent rise in energy costs
  - At 20 Cent per kWh, a server costs as much in electricity over 3 years as it costs to buy it
- Data Center cooling overheads are reaching their optimization limits
  - LHCb Data Center cooling overheads: approx. 5-6%
- Future efficiency goals must be achieved with
  - Better software efficiency
  - More efficient compute hardware
- LHCb: Change from CPU to GPU based RT processing
  - GPU power budget not as straight forward as CPU
  - What is the impact on energy budget

LHCb Upgrade Dataflow

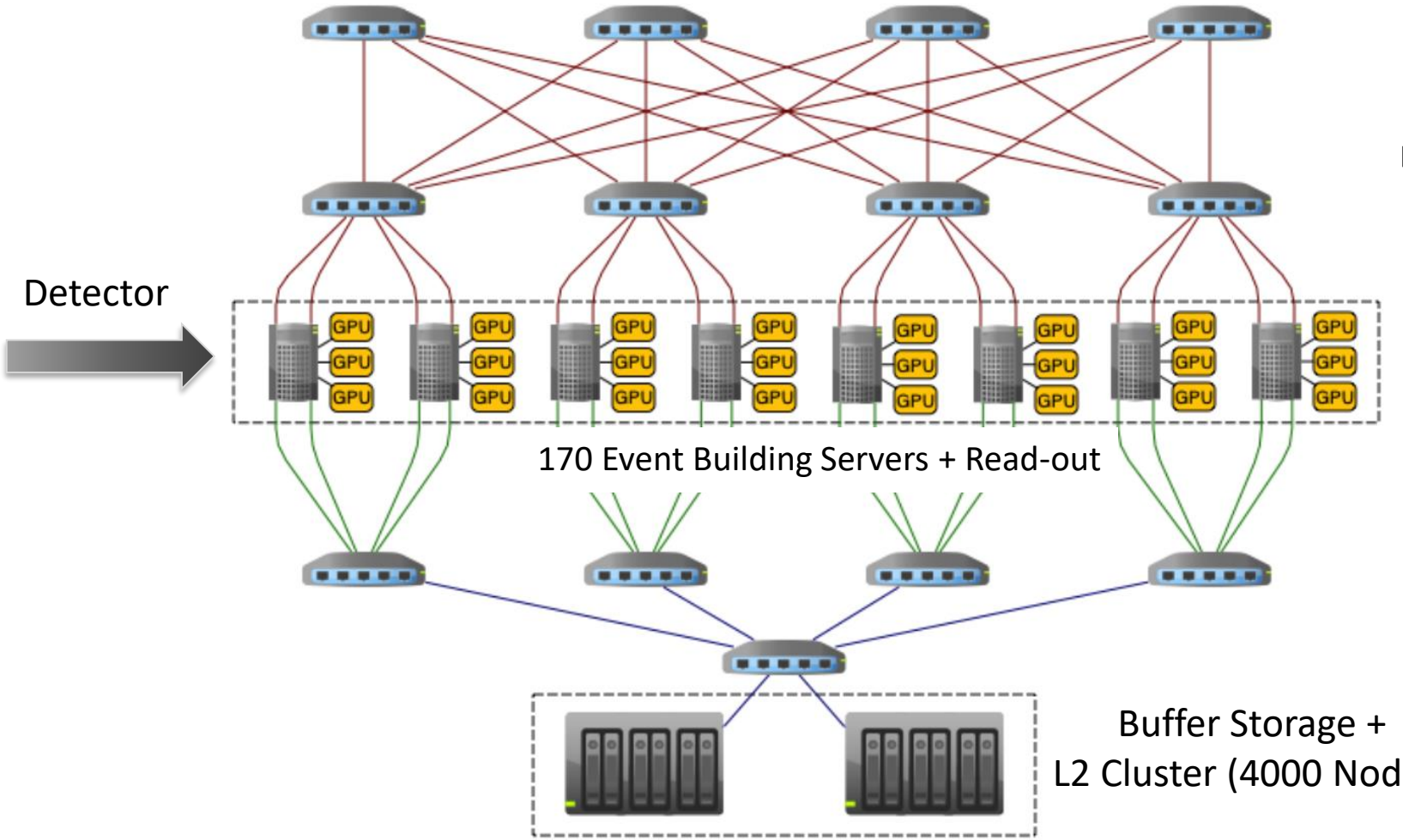


- Full detector read-out of every collision
- Track based L1 GPU trigger @ 4-5 TB/s
- Full reconstruction based L2 software trigger
- Peak input rate of up to 5 TB/s
- Up to 10 GB/s filtered output

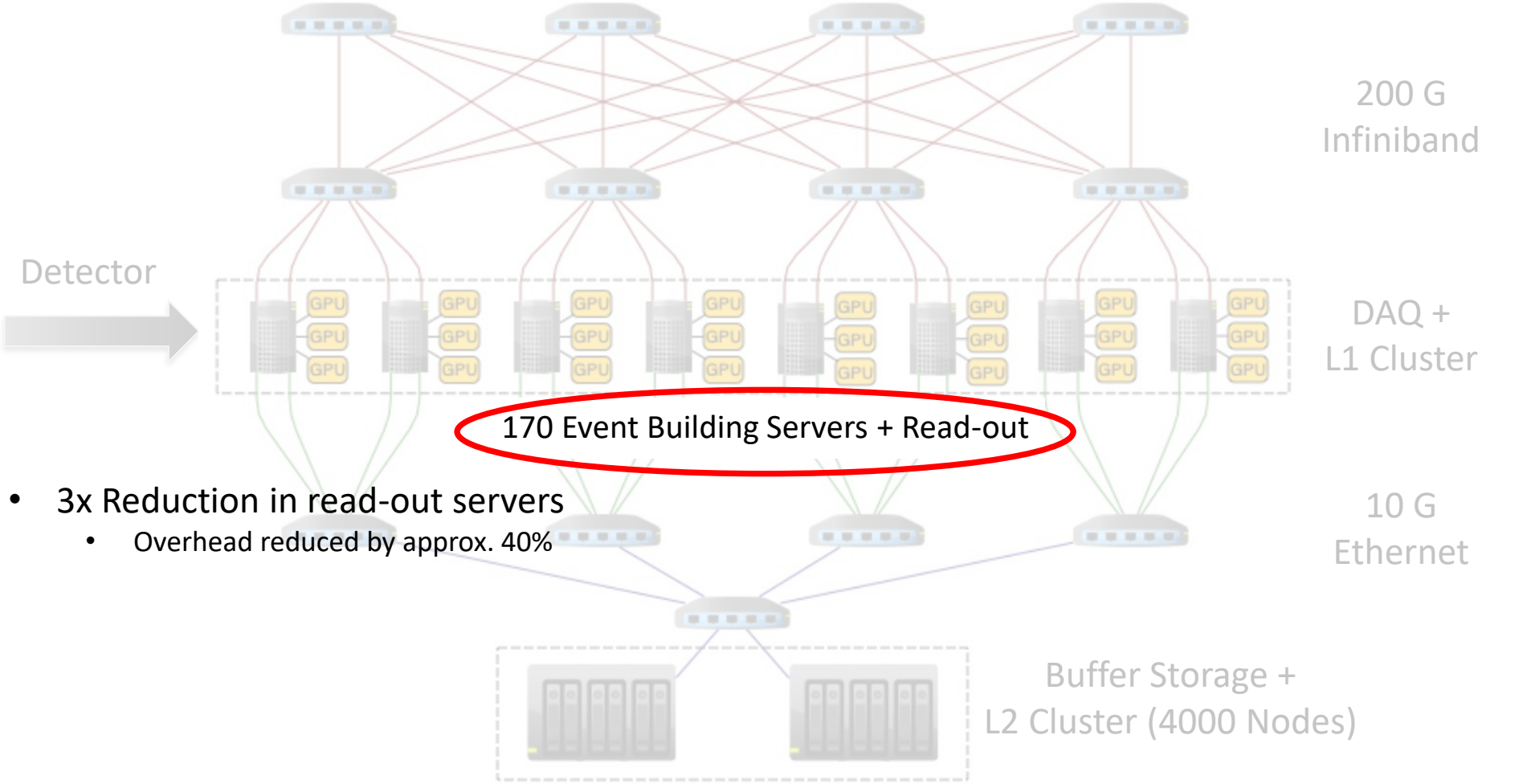
LHCb Upgrade DAQ Architecture – how it started



LHCb Upgrade DAQ Architecture - current

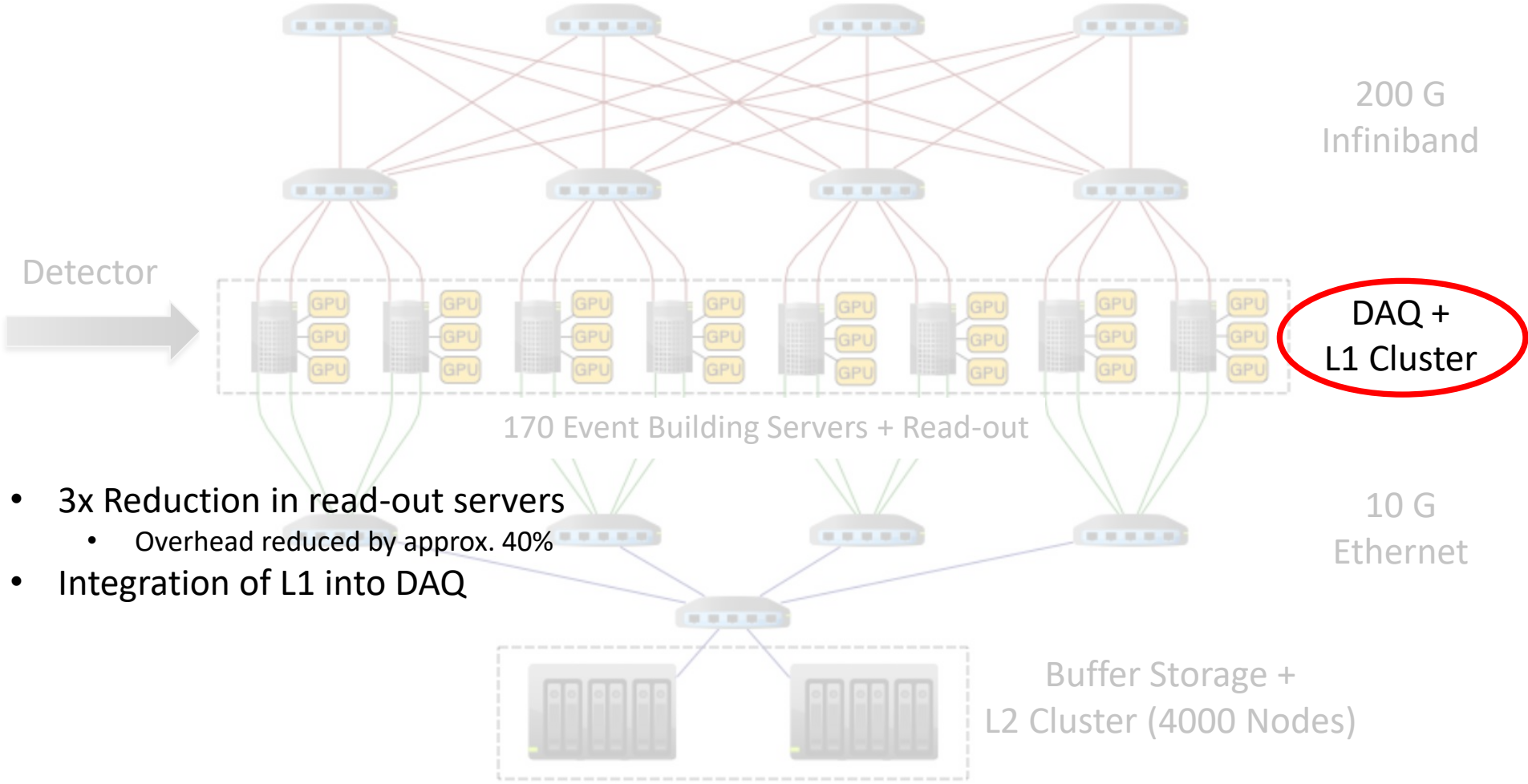


LHCb Upgrade DAQ Architecture 2/2



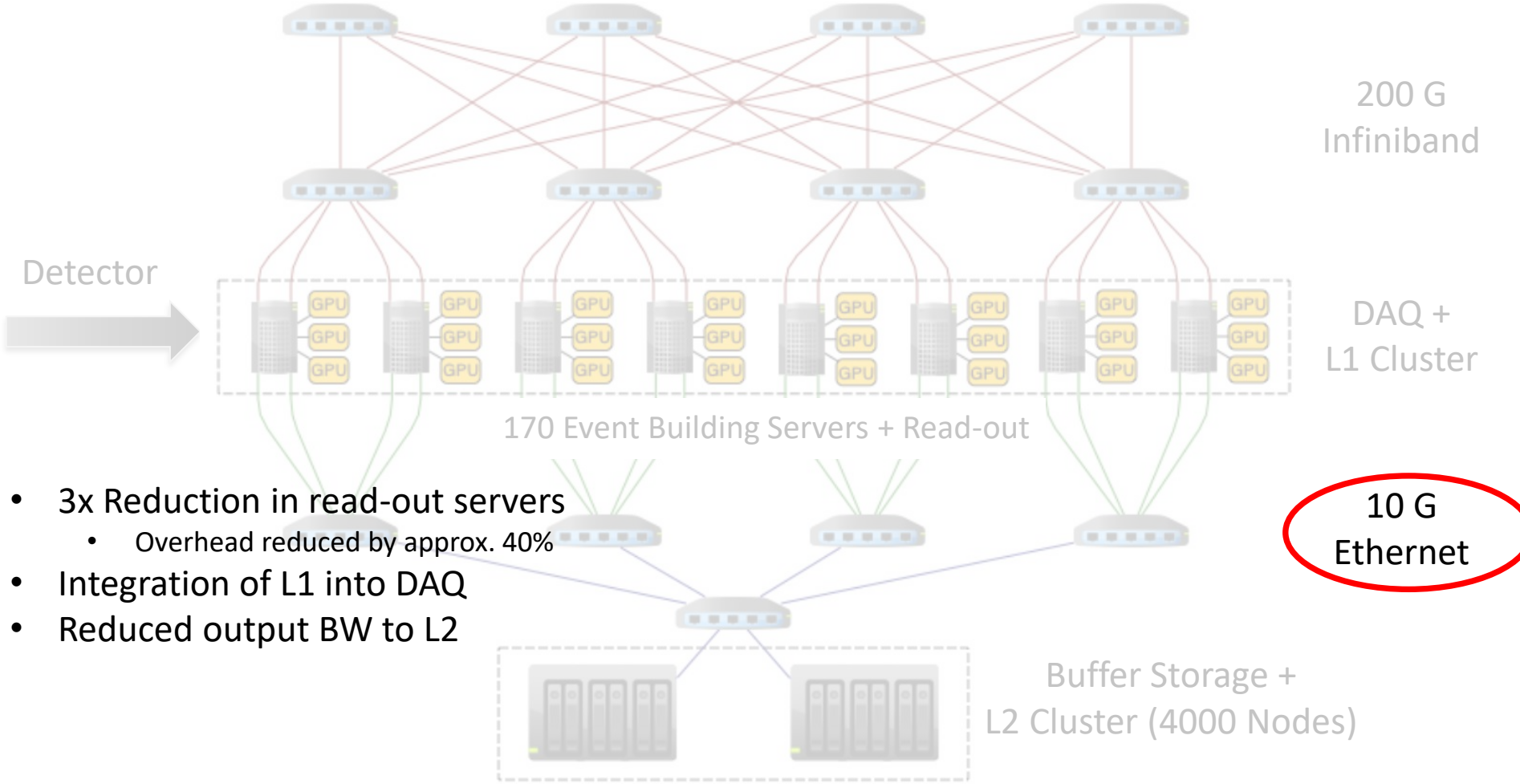
- 3x Reduction in read-out servers
  - Overhead reduced by approx. 40%

LHCb Upgrade DAQ Architecture 2/2



- 3x Reduction in read-out servers
  - Overhead reduced by approx. 40%
- Integration of L1 into DAQ

LHCb Upgrade DAQ Architecture 2/2

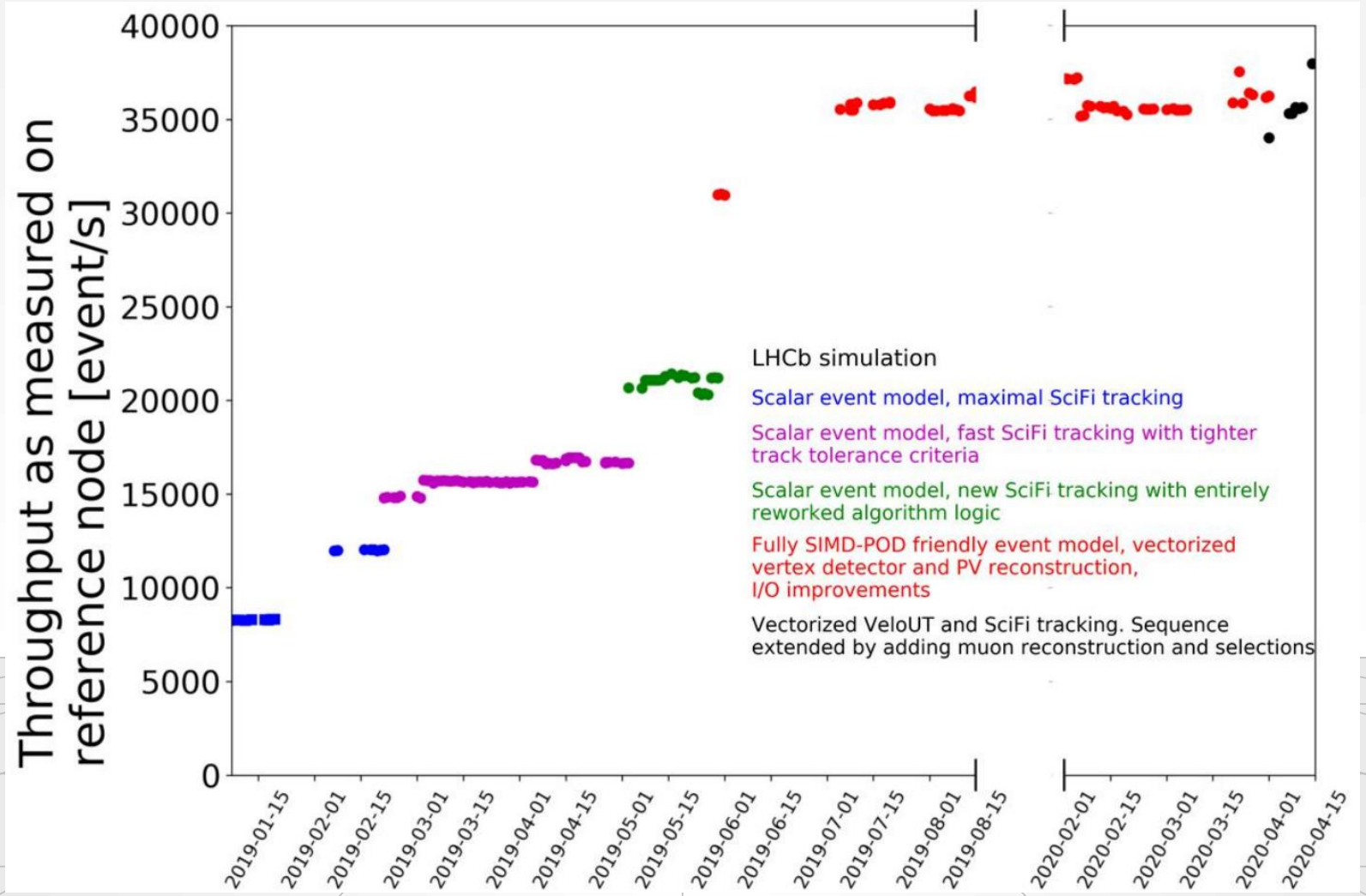


- 3x Reduction in read-out servers
  - Overhead reduced by approx. 40%
- Integration of L1 into DAQ
- Reduced output BW to L2

Main drivers behind development

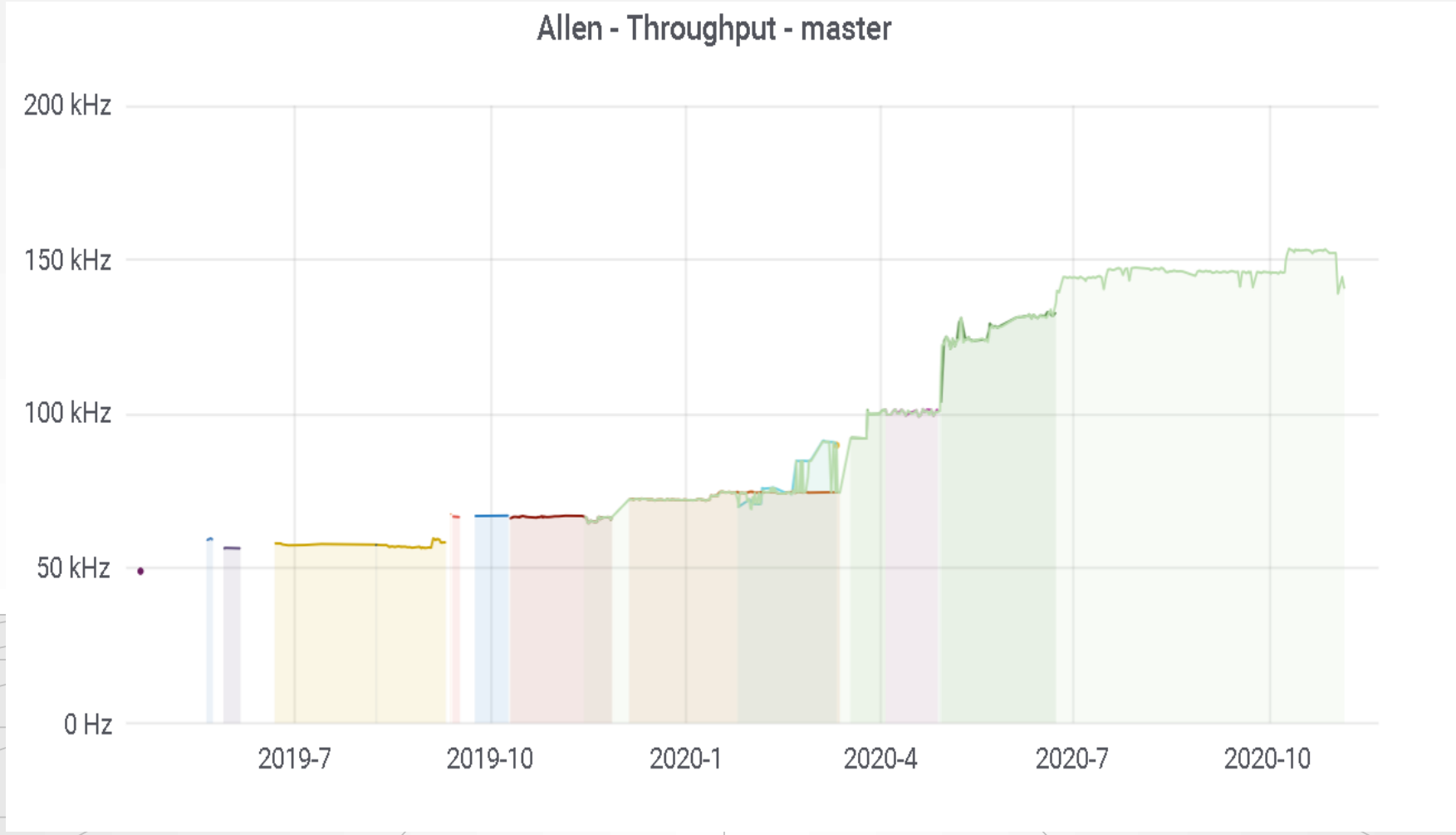
- Improvements in Host IO bandwidth
  - PCIe Gen 3 → PCIe Gen 4
  - 2 PCIe x16 slots per socket → 4 PCIe x16 per socket
  - 100 G Ethernet → 200 G Infiniband
- Improvements in Memory capacity
  - 0.5 TB Ram / machine is relatively easy today
  - Can buffer 5 TB/s approx. 15s @ 170 servers
  - Previous, FPGA based L1 trigger: O(us)
- Massive improvements in trigger software
  - Optimizations in Physics Algorithms
  - Rework toward SIMD friendly data structures
  - Single Threaded → Multi threaded

Trigger Software Improvements: CPU

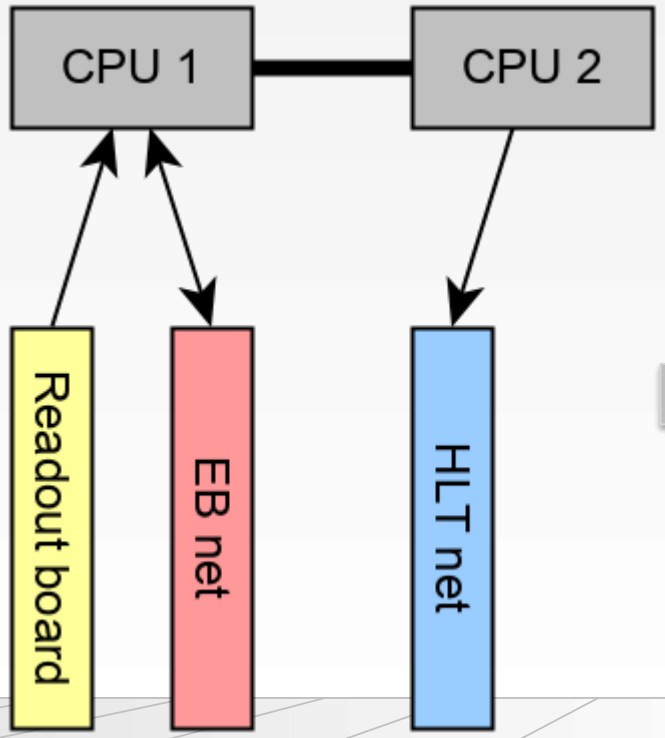


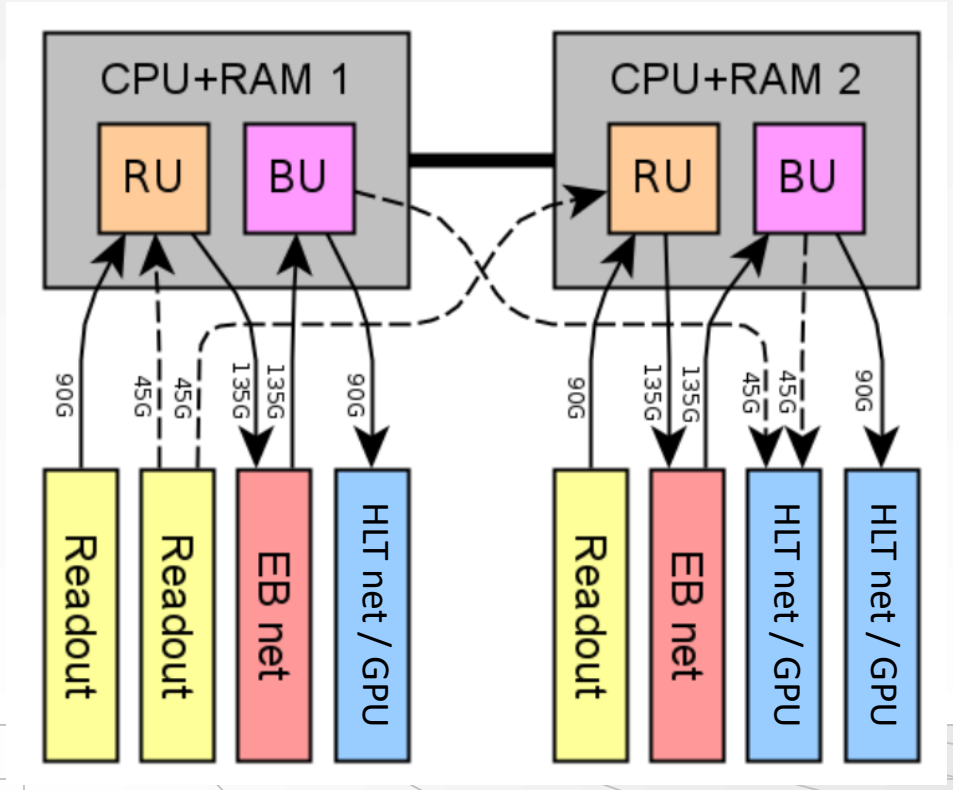
[https://cds.cern.ch/record/2715210]

Trigger Software Improvements: GPU



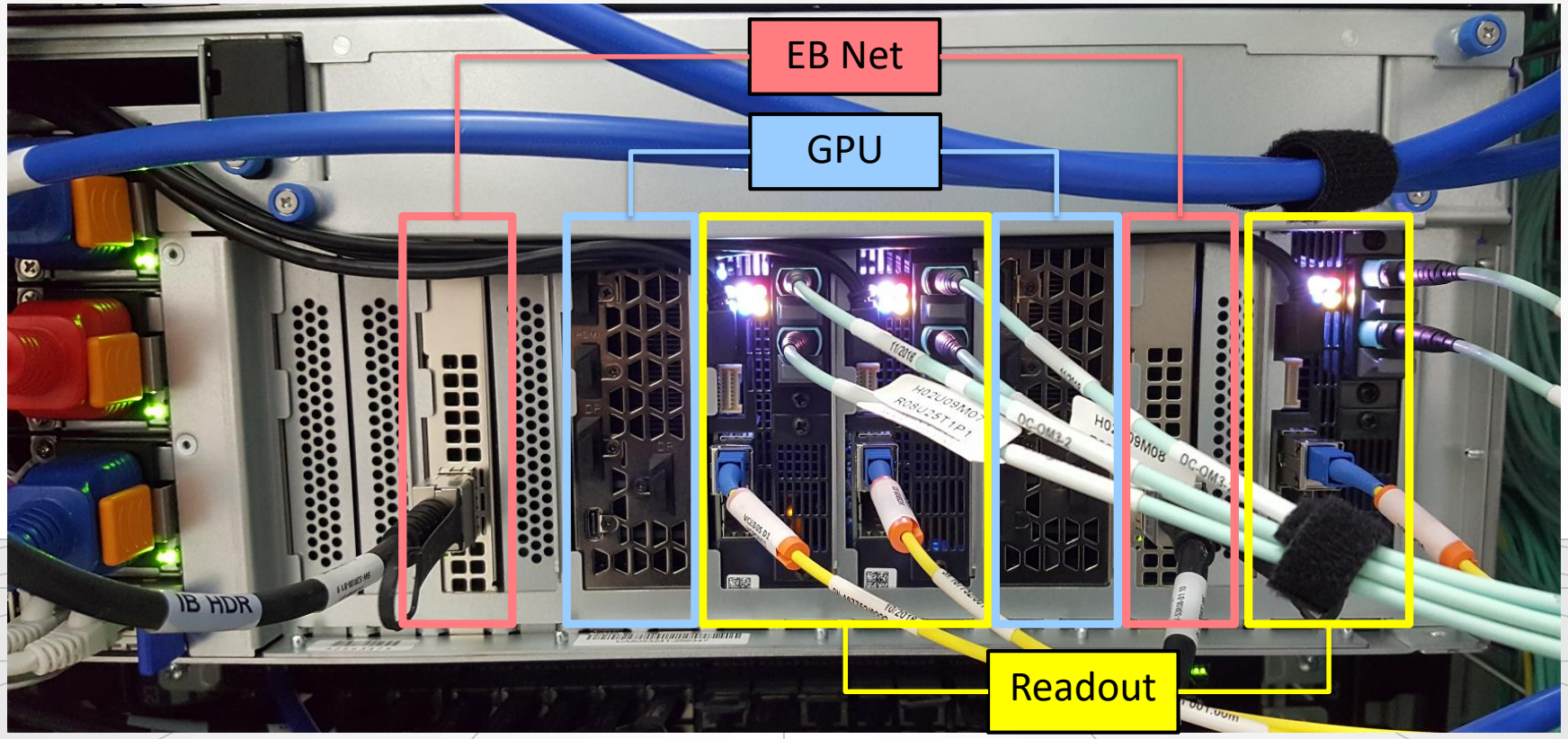
Host IO Improvements





- Move from Intel Xeon to AMD Ryzen
- Double PCIe BW
- Double PCIe slots
- Use On Board 10G NIC for output in case of GPUs / Compute Accelerators

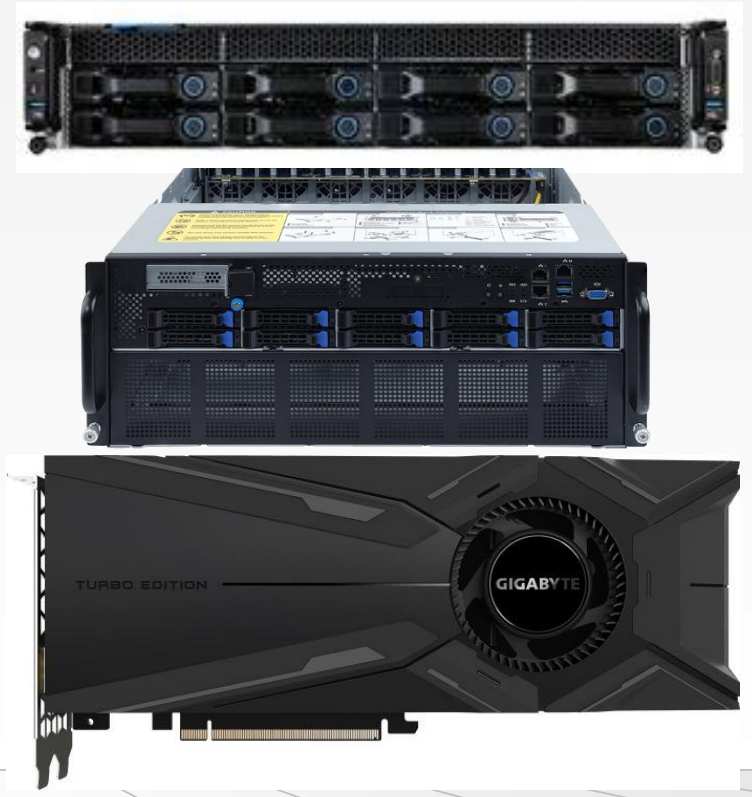
Host IO - Physical



COMPUTING POWER EFFICIENCY

Hardware used

- Quanta DA0S2SMBCE0
  - Classic 2U four node server
  - 2 x Xeon 2630-v4 per node
  - Shared cooling + PSU
- Gigabyte G482-Z5
  - 2 x AMD EPYC 7502
  - Cooling tuned for CPU only
- Gigabyte G482-Z5
  - 2 x AMD EPYC 7502
  - Various number of GPUs, network and DAQ cards
  - GPUs: GV-N208TTURBO-11GC-rev-10 (RTX280-TI)
- Measurements done with BMC instrumentation



Test Configurations

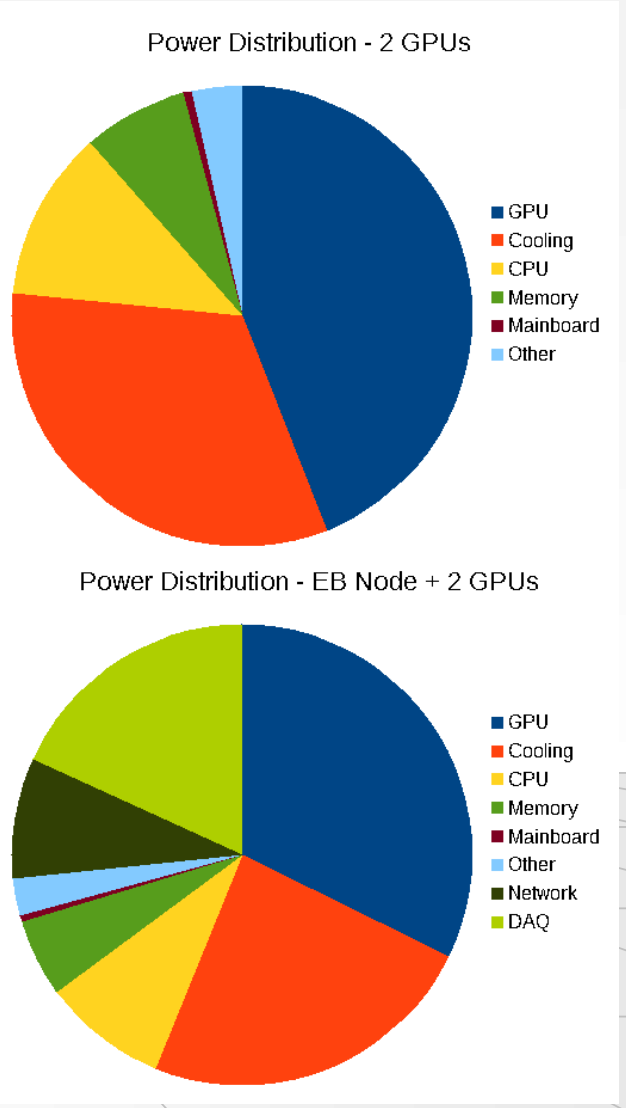
- Pure CPU Trigger
  - Illustration of software improvements
- GPU Based trigger in dedicated hosts
  - Architecture similar to pure CPU but with GPU acceleration
- DAQ Integrated GPU Trigger
  - Current Baseline
- Pure GPU Machine
  - Not relevant for our form of data processing
  - Good estimate for upper limit

DAQ integrated Power efficiency accounting

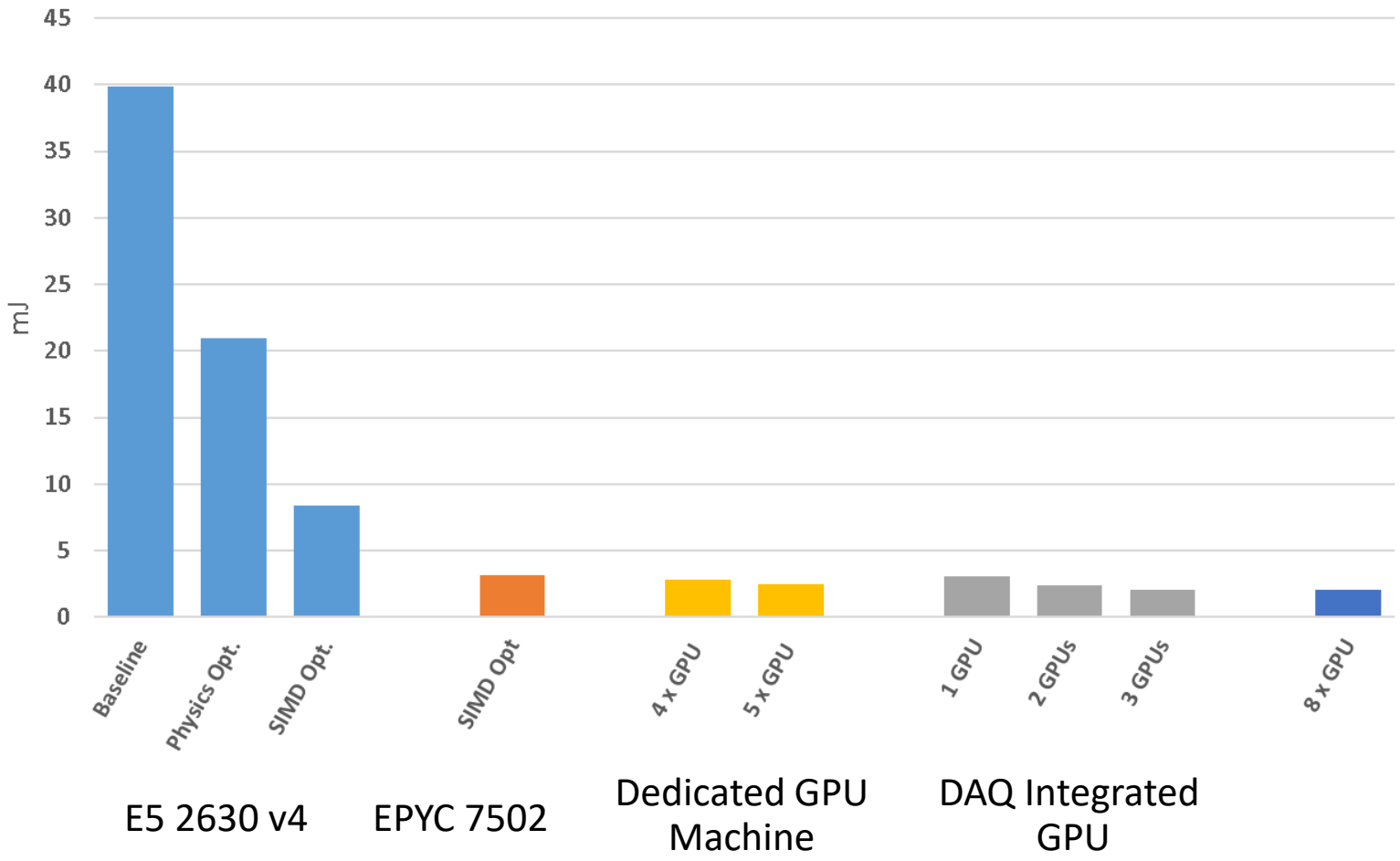
- GPUs replace output network cards
  - Subtract output network card power consumption
- Host needed for DAQ activity even without GPUs
  - Subtract Host DAQ activity power consumption
  - Mostly memory, network and DAQ cards
- Accounting of shared cooling (Network/DAQ/GPU/CPU)
  - DAQ board cooling needs almost full server cooling resources
  - Slightly unfair toward GPUs (in our specific case)
  - Distribute cooling power over number of slots

Power Distribution

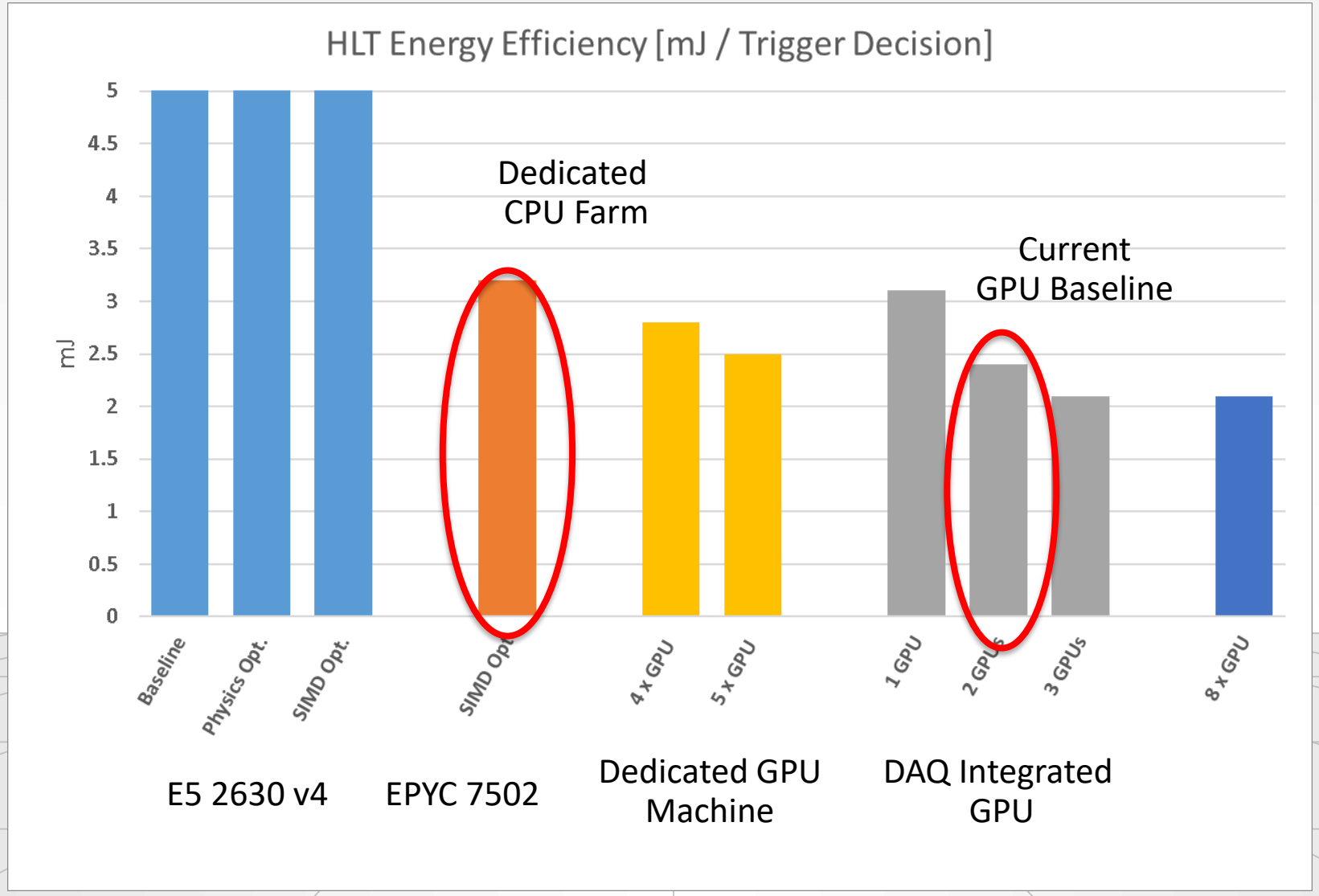
- DAQ configuration vs plain, 2 GPU configuration
  - Machine is made for 8 GPUs
  - Large portion of cooling air bypasses the 2 GPUs
- Cooling is a surprisingly large amount
  - approx. 400W! at full fan speed
- Adding DAQ and network
  - Fans serve additional purpose
  - Fans need to work slightly more
  - Power efficiency still increases per slot



Results


HLT Energy Efficiency [mJ / Trigger Decision]
mJ
45
40
35
30
25
20
15
10
5
0
Baseline
Physics Opt.
SIMD Opt.
SIMD Opt
4 x GPU
5 x GPU
1 GPU
2 GPUs
3 GPUs
8 x GPU
E5 2630 v4
EPYC 7502
Dedicated GPU Machine
DAQ Integrated GPU

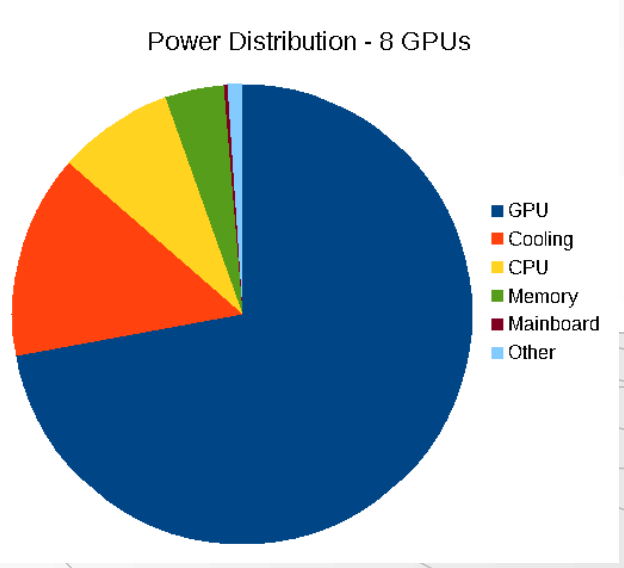
Results - Zoomed In



Power efficiency discussion

- By far largest contribution is software optimization
- Relying on just CPU hardware advances would have been tight in both cost and power infrastructure

- Naïve efficiency from POV of GPU TDP: 1.6 mJ
  - Additional 1.3 x over best configuration



Conclusions

- LHCb upgrade RT strategy change CPU→GPU
- GPUs are between 3-33% more efficient
  - Our particular software
  - Depending on full architecture
  - Comparable physics but radically different software architectures
- Developers, developers, developers
  - GPU computing still relatively new (for us) → better optimization more likely
  - Despite decades of CPU experience, a lot of gains on the table

→ Hire more software engineers

Thank you for your attention