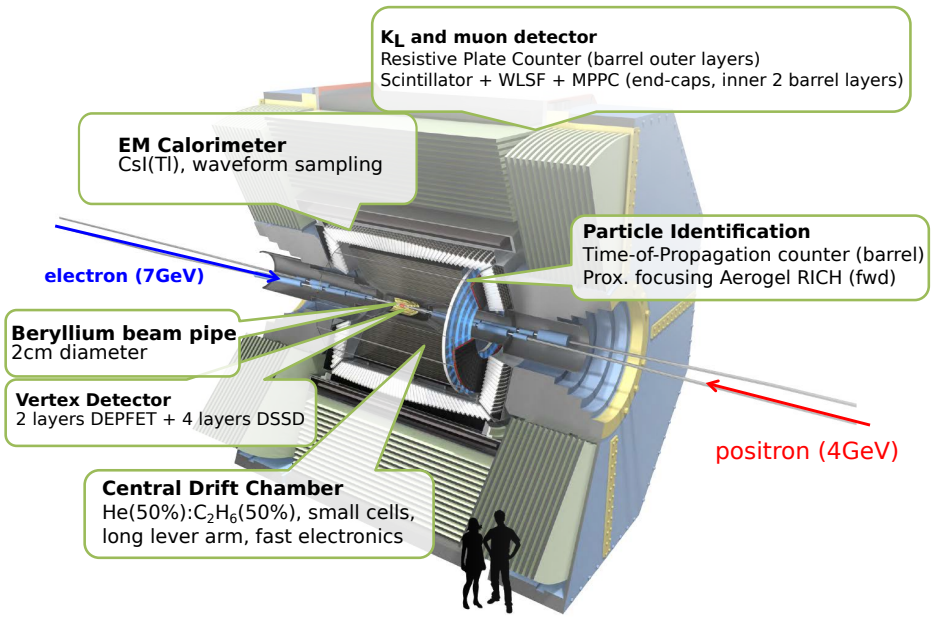
Software-assisted Event Builder for the Belle II Experiment

Dmytro Levit on behalf of the Belle II DAQ group

Institute of Particle and Nuclear Studies, KEK

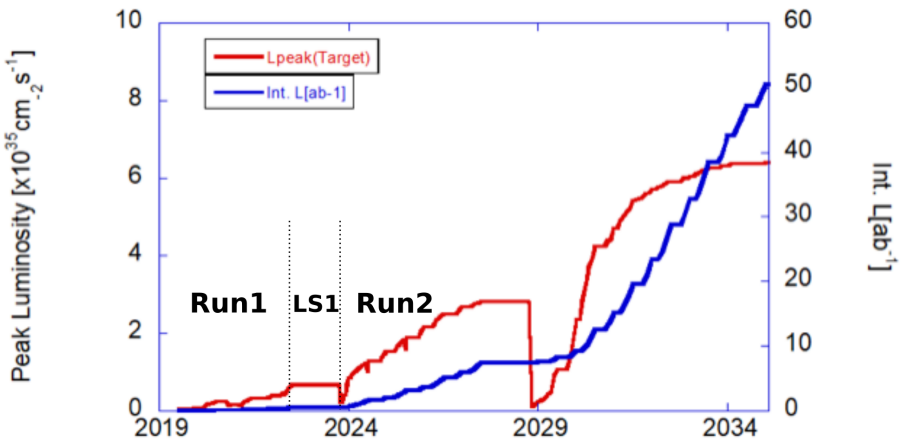
April 22, 2024

Introduction



- Study of CP violation
- Search for physics beyond the Standard Model
- Goal: **50x** Belle data sample

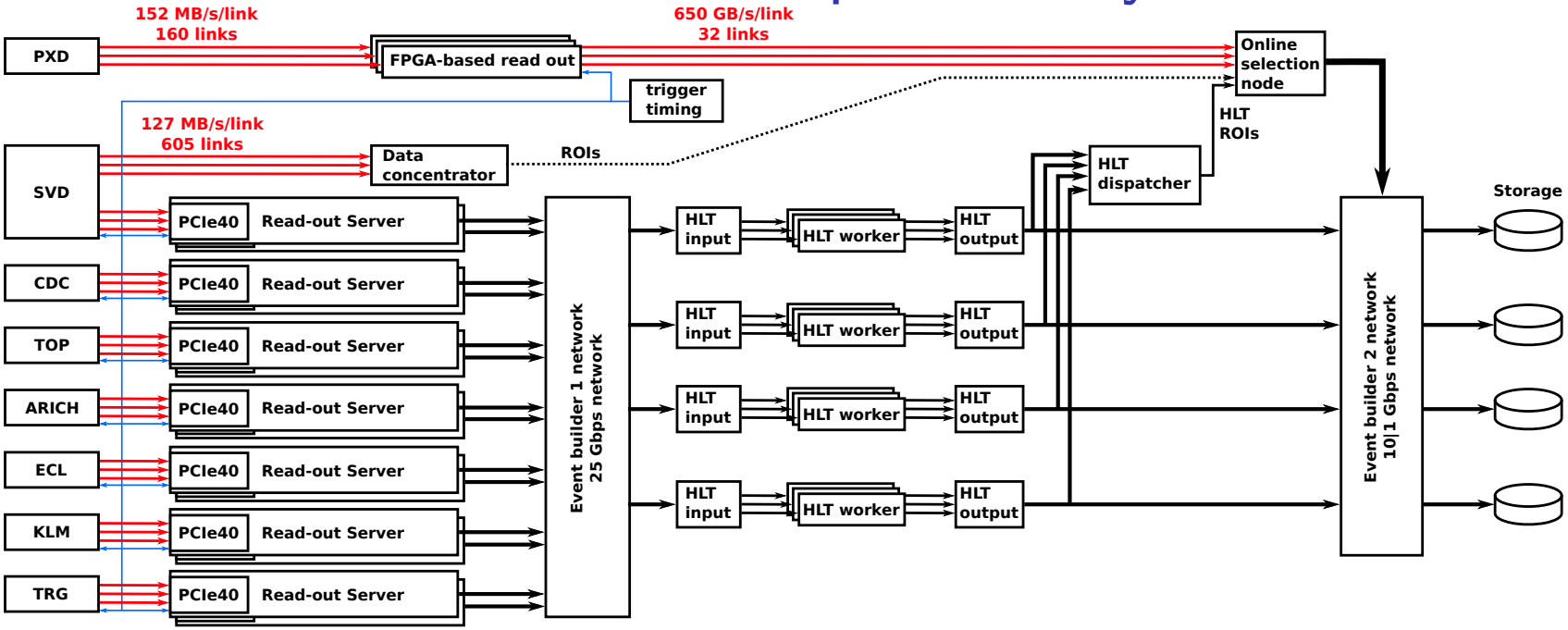
Introduction



- Study of CP violation
- Search for physics beyond the Standard Model
- Goal: **50x** Belle data sample
- Run 1: 2019–2022
- Long shutdown 1: 2022–2024
  - New pixel detector
  - TOP PMT upgrade
  - **DAQ read-out upgrade**
- Run 2: February 2024–

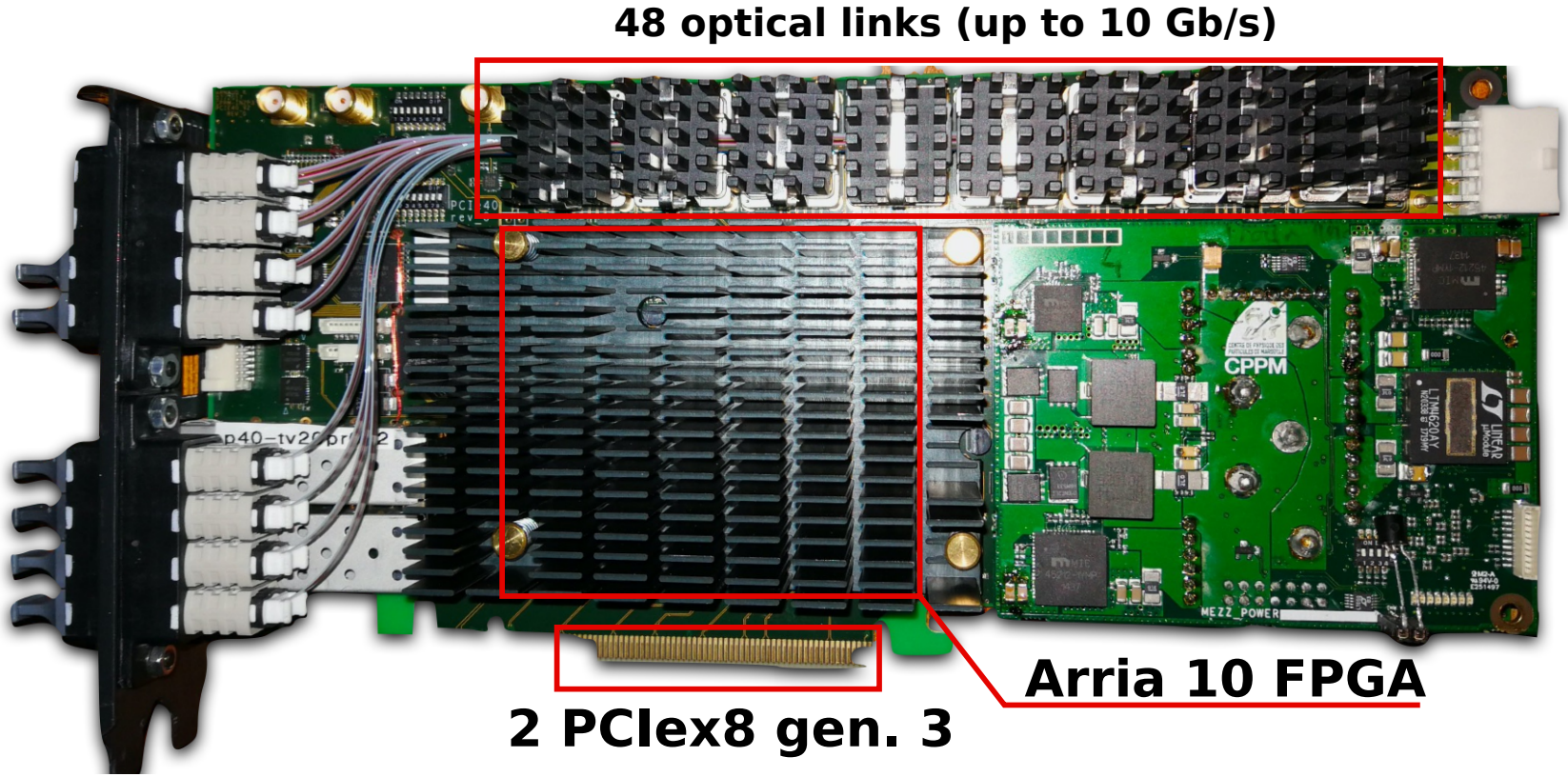
DAQ Layout and Hardware

Layout of the Belle II Data Acquisition System



- Trigger rate: **30 kHz**
- Event size: **1.5 – 700 kB** at full luminosity
- Large data rate asymmetry
  - PXD vs. rest of Belle II: **20 GB/s** vs. **2 GB/s**
  - Separate read-out paths
- Two-stage event builder
- Online event reconstruction and data filtering
- Unified read-out system for 6 subsystems

PCIe40 Hardware



- Designed for LHCb and ALICE experiments (CROC)
  - no external memories
- Read-out system upgrade from COPPER system
  - 2021: TOP and KLM
  - 2022: ARICH
  - 2024: all subsystems

Event Builder Algorithms

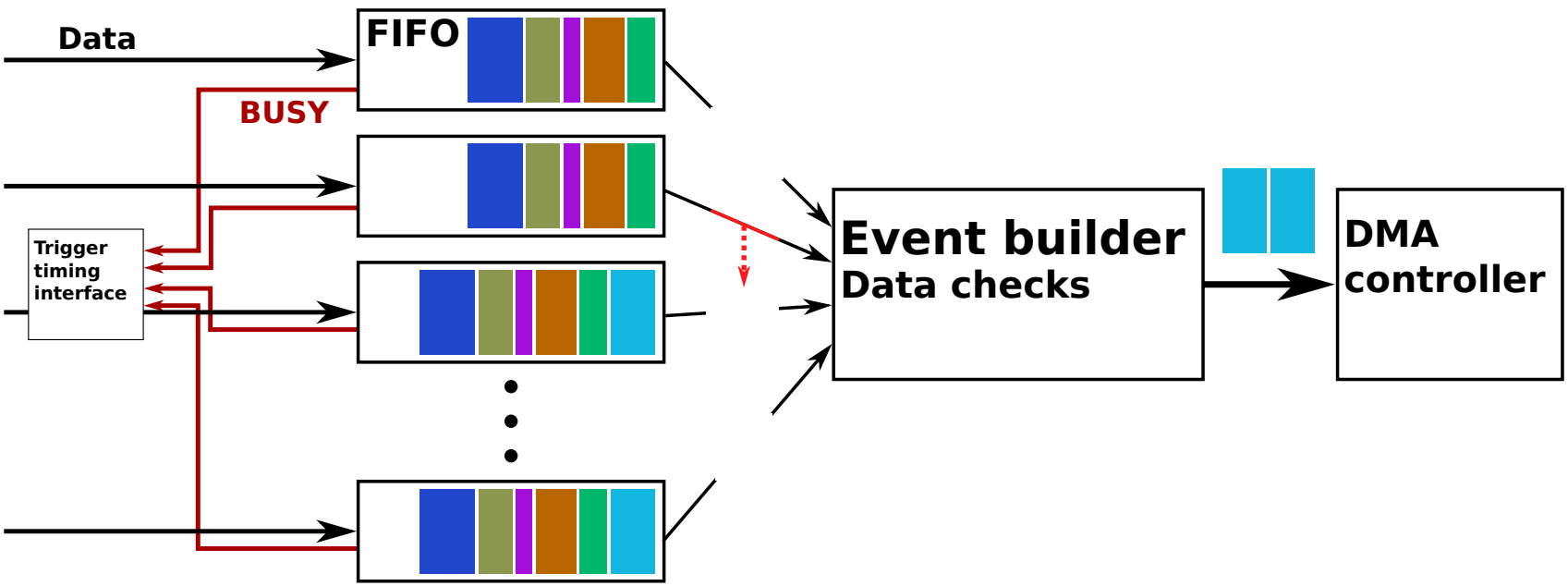
System Requirements

- Designed trigger rate: **30 kHz**
  - limitation of the APV25 used in SVD
  - estimated maximum data flow (SVD)/read-out PC: **720 MB/s**
- Event size: up to a few kB/channels
- 600 detector channels
- Event processing limited by PXD buffer size 128 GB
  - 5 s at maximum occupancy

  $\Rightarrow$ fast event builder
- Maximum throughput per board
  - Belle2link protocol throughput: **127 MB/s** (254 MB/s)
  - 32 channels: **4 GB/s** (8 GB/s)
  - 42 channels: **5.3 GB/s** (10.6 GB/s)

Event Builder in Firmware

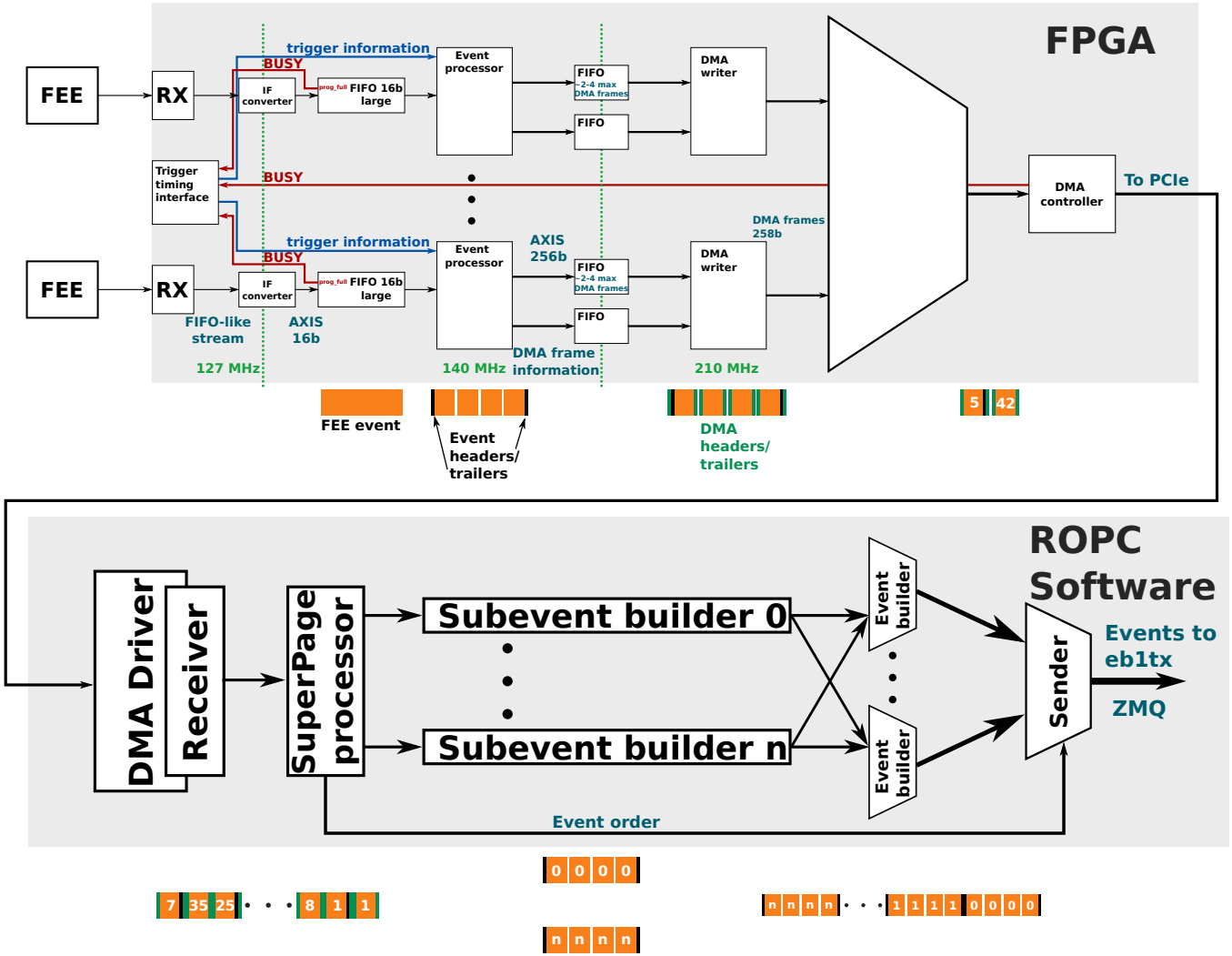
- Data buffering in internal memory
- Event builder
  - round-robin FIFO read-out
  - data consistency checks
- Data read-out by DMA controller



Performance and Limitations of the Firmware-based Event Builder

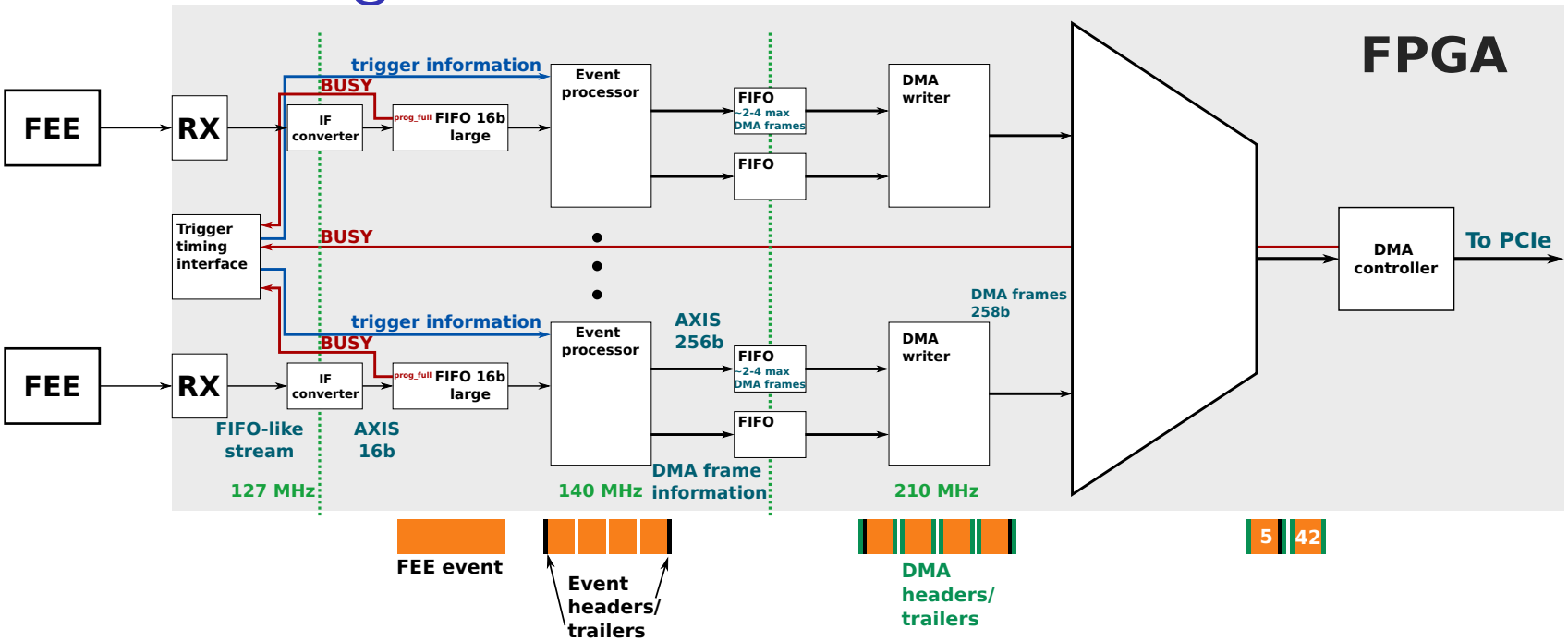
- System in operation in 2021-2022 run with 3 subdetectors
  - trigger rates: up to **15 kHz**
  - data rates: $<$ **100 MB/s/ROPC**
- Dead time in DAQ
  - limited FIFO size: 64 or 128 kB/channel
  - round-robin event builder
  - large event sizes
  - large delay spread
- Buffer overflow
  - late backpressure to trigger distribution
  - multiple events in FEE read-out chain
- Performance limited to 600 MB/s by software
  - inefficient CRC calculation
  - 1 GB/s without CRC calculation

Software-Assisted Event Builder



- Independent event processing in firmware
- Data buffering in the read-out PC RAM
- Event building in software

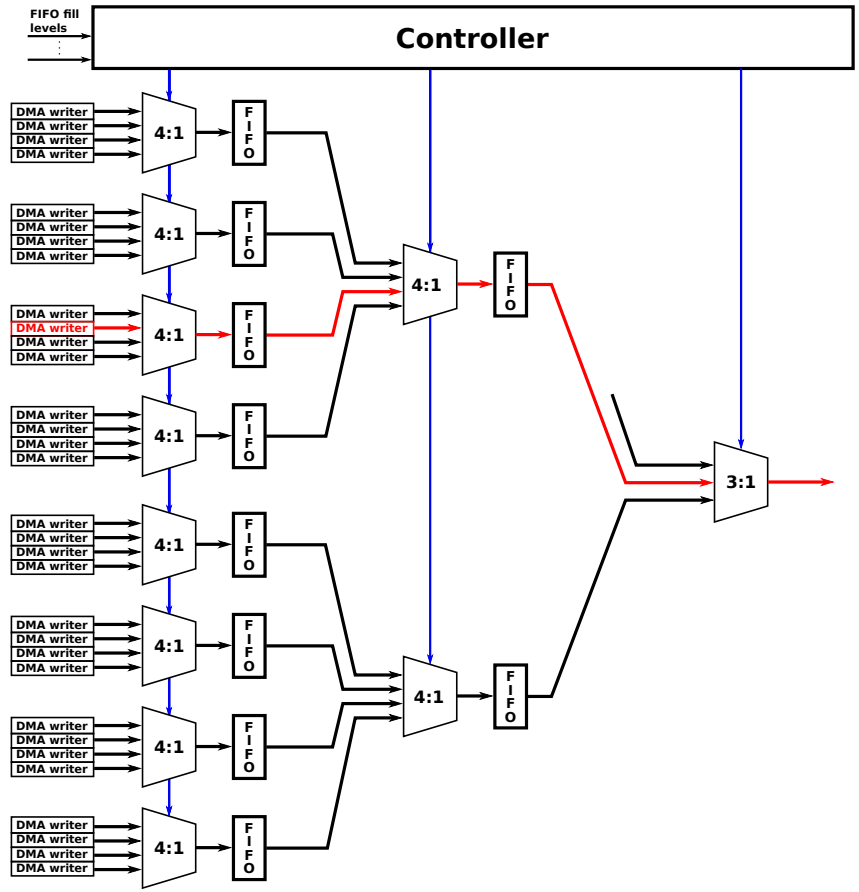
Data Processing in Firmware



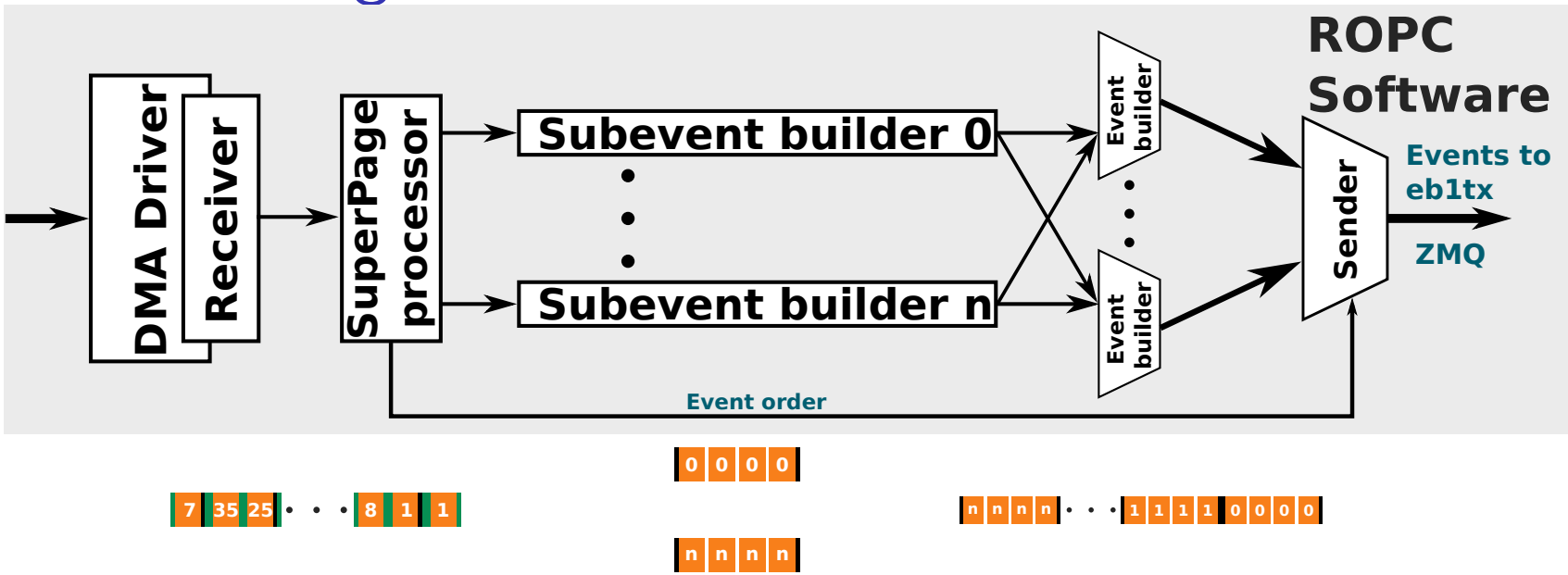
- Synchronization of trigger and data streams
- Data consistency check
- Event separation into fragments
- Routing of fragments to DMA controller
  - throughput limit **6.7 GB/s**
- Flow control based on DMA descriptor FIFO fill level

Multiplexer

- Intelligent link readout scheduling
  - highest FIFO fill level
  - data availability
- Reevaluation on the frame-by-frame basis

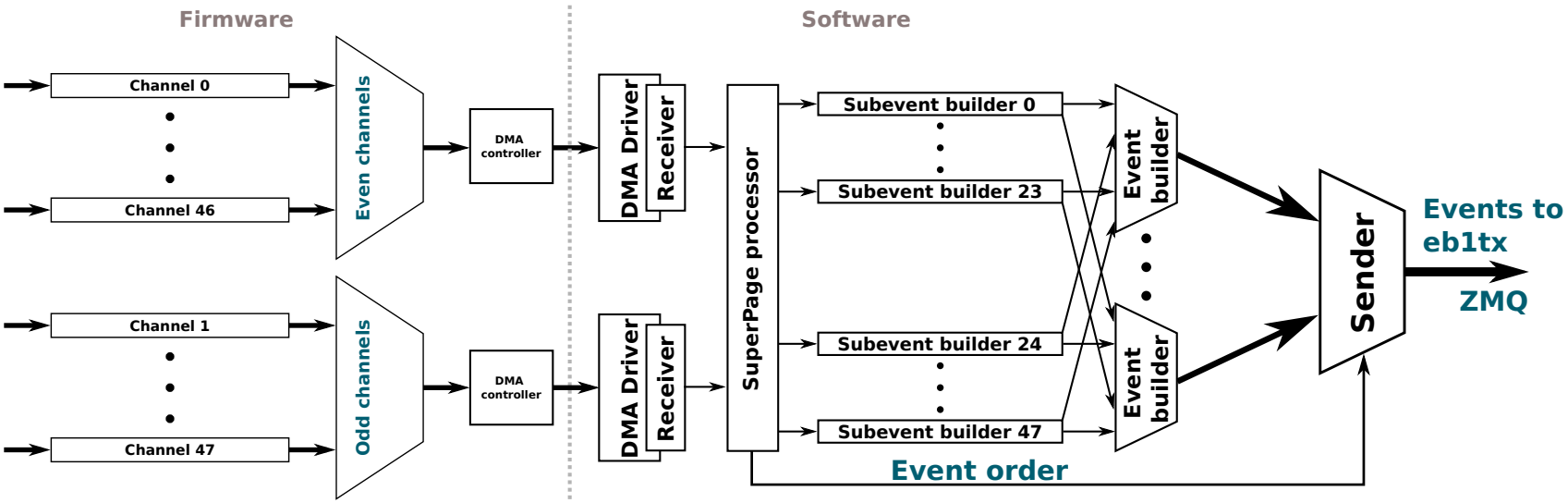


Data Processing in Software



- Pulling events from driver
  - 1 CPU core for receiver + driver
- Subevent reassembly
  - synchronization and data format checks
- Multiple event mergers
  - CRC calculation
- Online channel masking
  - replace faulty event with a placeholder event
  - persistent until the next run

Double PCIe Interface

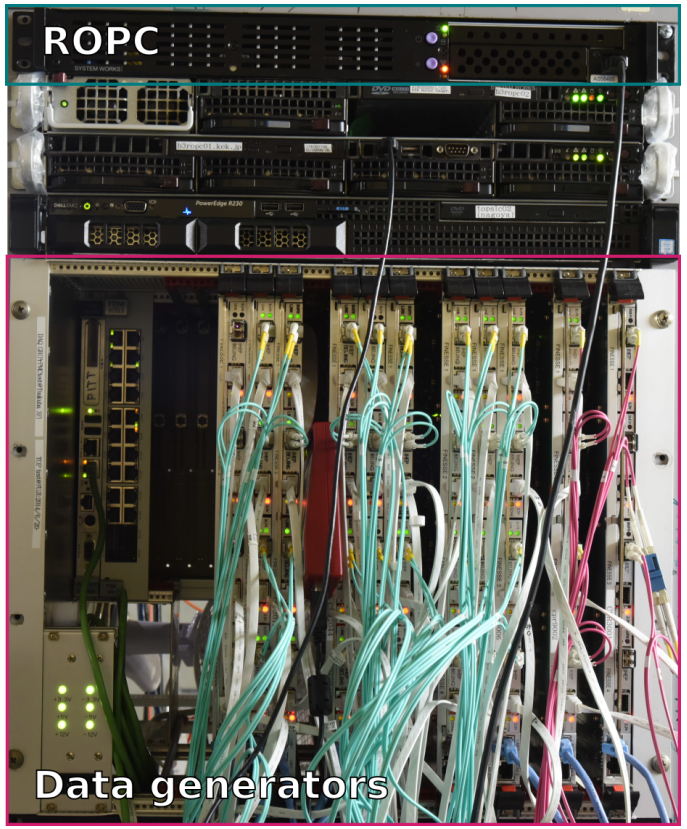


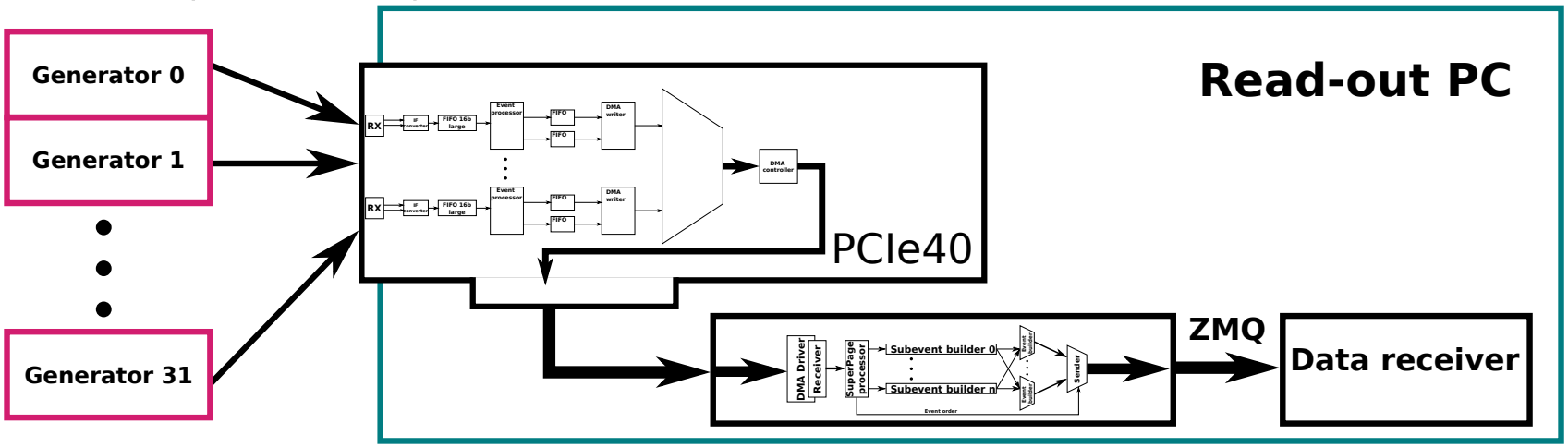
- Duplication of DMA controllers and receiver threads
- **Almost no change** to software or data processing logic
- Load balancing through separation of even and odd channels

System Performance

Test Setup

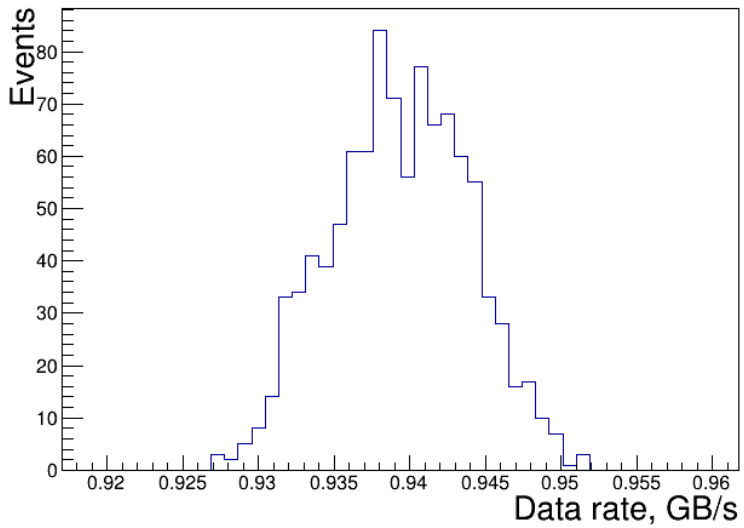
- 32 hardware data generators
  - programmable event size distribution
- Unthrottled Belle2Link: **254 MB/s**
- **Double** DMA controller firmware
- Discard data in receiver
- ROPC configuration:
  - Intel(R) Xeon(R) Silver 4214R CPU (12 cores)
  - 48 GB memory, 6 channels (2400 MHz)





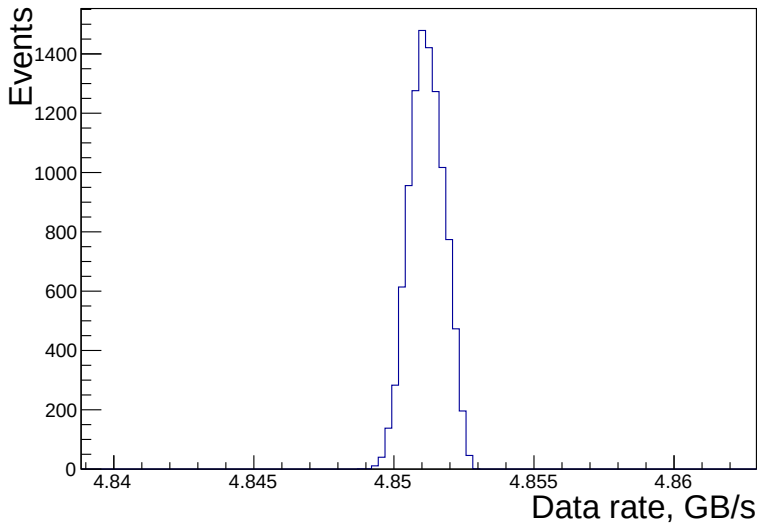
Performance Measurements

- **Sarwate CRC** calculation in event builder
  - **0.94 GB/s**
  - 147.000 frames/s



Performance Measurements

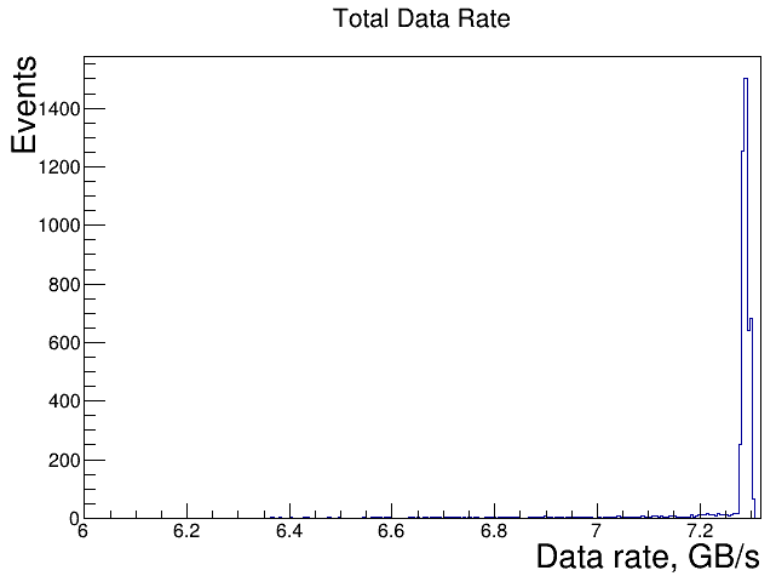
- **Sarwate CRC** calculation in event builder
  - **0.94 GB/s**
  - 147.000 frames/s
- **Slice-by-16*** in event builder, **single** DMA
  - **4.85 GB/s**
  - 1.000.000 frames/s
  - limited by interface to DMA controller



*DOI: 10.1109/TC.2008.85

Performance Measurements

- **Sarwate CRC** calculation in event builder
  - **0.94 GB/s**
  - 147.000 frames/s
- **Slice-by-16*** in event builder, **single** DMA
  - **4.85 GB/s**
  - 1.000.000 frames/s
  - limited by interface to DMA controller
- **Slice-by-16*** in event builder, **double** DMA
  - **7.2 GB/s**
  - 1.000.000 frames/s
  - limited by ZMQ interface



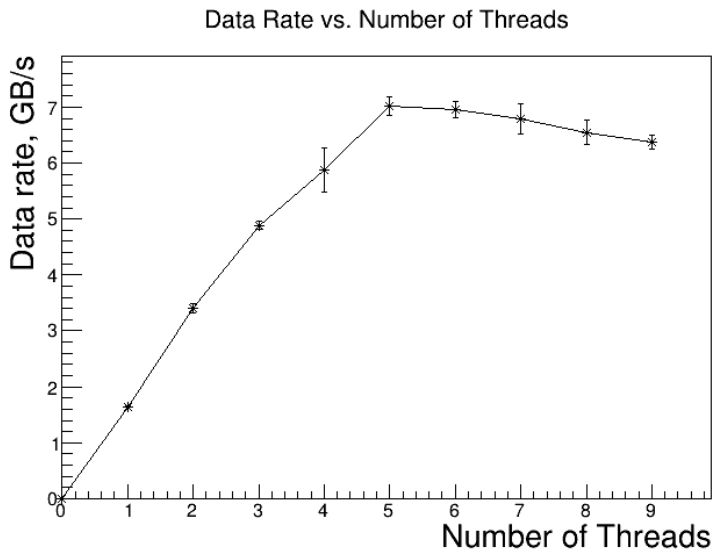
*DOI: 10.1109/TC.2008.85

Performance Bottlenecks

- CRC algorithm
  - solved by using Slice-by-16

Performance Bottlenecks

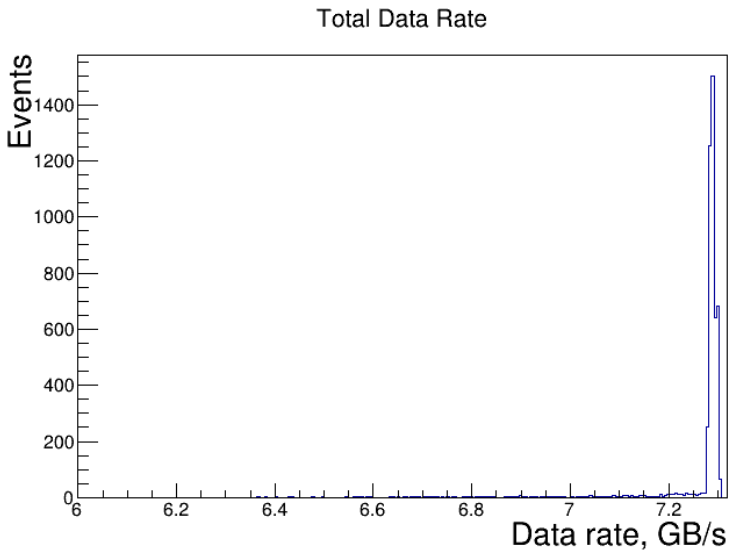
- CRC algorithm
  - solved by using Slice-by-16
- Performance of the event builder thread
  - $\sim$**1.52 GB/s/thread**
  - scale number of threads to expected performance



Performance Bottlenecks

- CRC algorithm
  - solved by using Slice-by-16
- Performance of the event builder thread
  - $\sim$**1.52 GB/s/thread**
  - scale number of threads to expected performance
- Interface to DMA controller
  - double number of PCIe interfaces
  - double PCIe performance: **7.2 GB/s**
  - tail towards low data rates

    $\Rightarrow$ ZMQ limitation



Performance Bottlenecks

- CRC algorithm
  - solved by using Slice-by-16
- Performance of the event builder thread
  - $\sim$**1.52 GB/s/thread**
  - scale number of threads to expected performance
- Interface to DMA controller
  - double number of PCIe interfaces
  - double PCIe performance: **7.2 GB/s**
  - tail towards low data rates

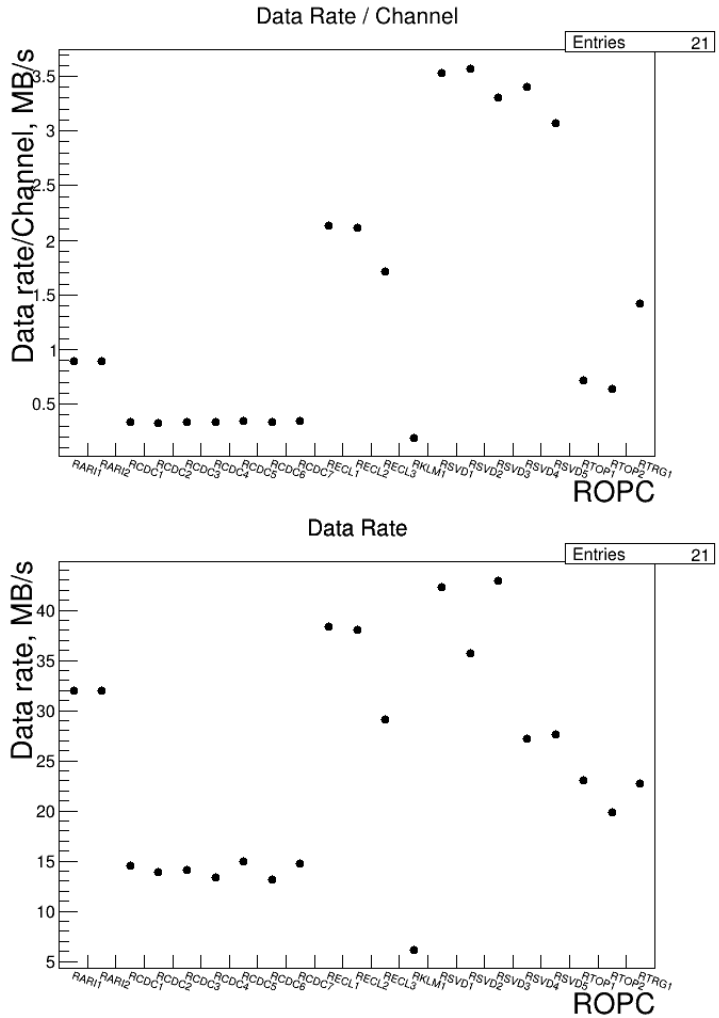
    $\Rightarrow$ ZMQ limitation
- Memory bandwidth
  - use faster memories

Operation of the System in Run 2

- 21 systems, 600 detector links
- **Single** DMA controller firmware
- Data taking with up to **5.5 kHz** trigger rate
  - scales with luminosity
- Stable operation at 30 kHz artificial trigger rate without beams
  - 18.000.000 frames/s

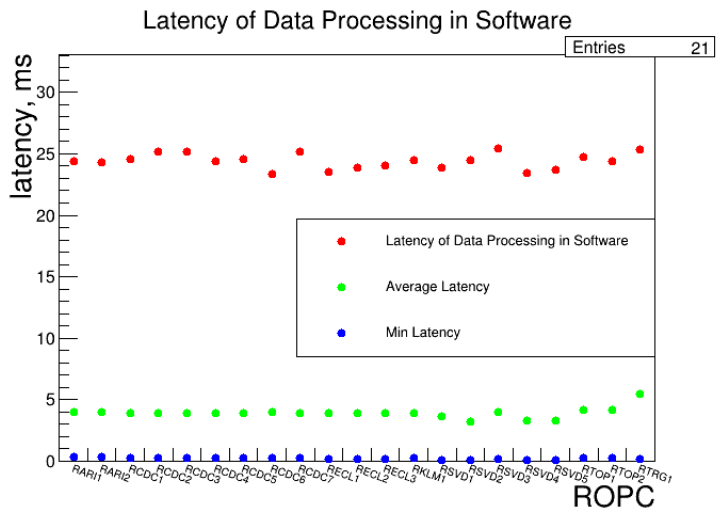
Operation of the System in Run 2

- 21 systems, 600 detector links
- **Single** DMA controller firmware
- Data taking with up to **5.5 kHz** trigger rate
  - scales with luminosity
- Stable operation at 30 kHz artificial trigger rate without beams
  - 18.000.000 frames/s
- Data rates per channel at $3.27x10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$
  - **200 kB/s – 3.5 MB/s**



Operation of the System in Run 2

- 21 systems, 600 detector links
- **Single** DMA controller firmware
- Data taking with up to **5.5 kHz** trigger rate
  - scales with luminosity
- Stable operation at 30 kHz artificial trigger rate without beams
  - 18.000.000 frames/s
- Data rates per channel at $3.27x10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$
  - **200 kB/s – 3.5 MB/s**
- Average event processing time **in software**: 4 ms
  - independent of subsystem (event size)



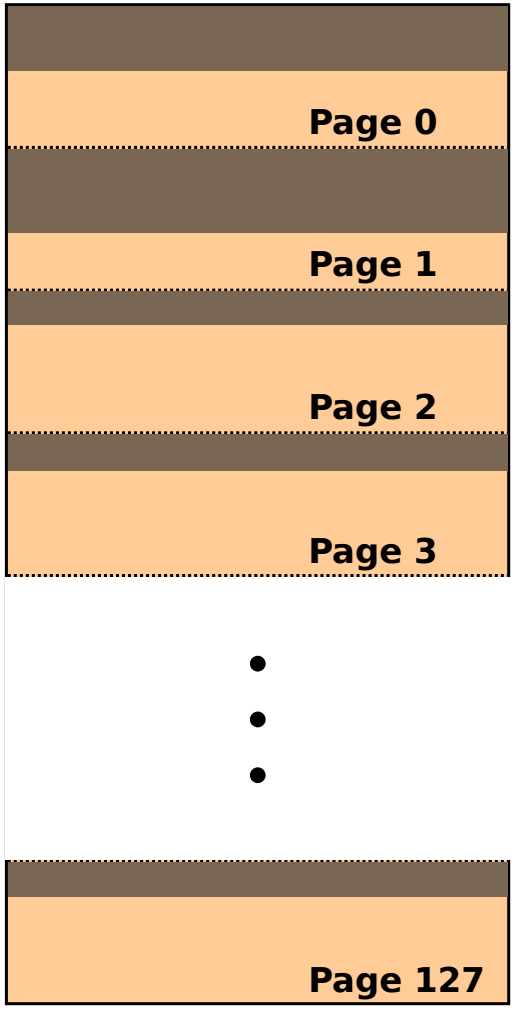
Summary

- New read-out system for Belle II experiment
  - combined data processing in firmware and software
- System characterized and performance measured
- Throughput of **7.2 GB/s** measured on the testbench
  - much higher than expected by Belle II (**720 MB/s**)
  - headroom for future detector upgrades
- Stable operation in run 2 at a fraction of the peak performance

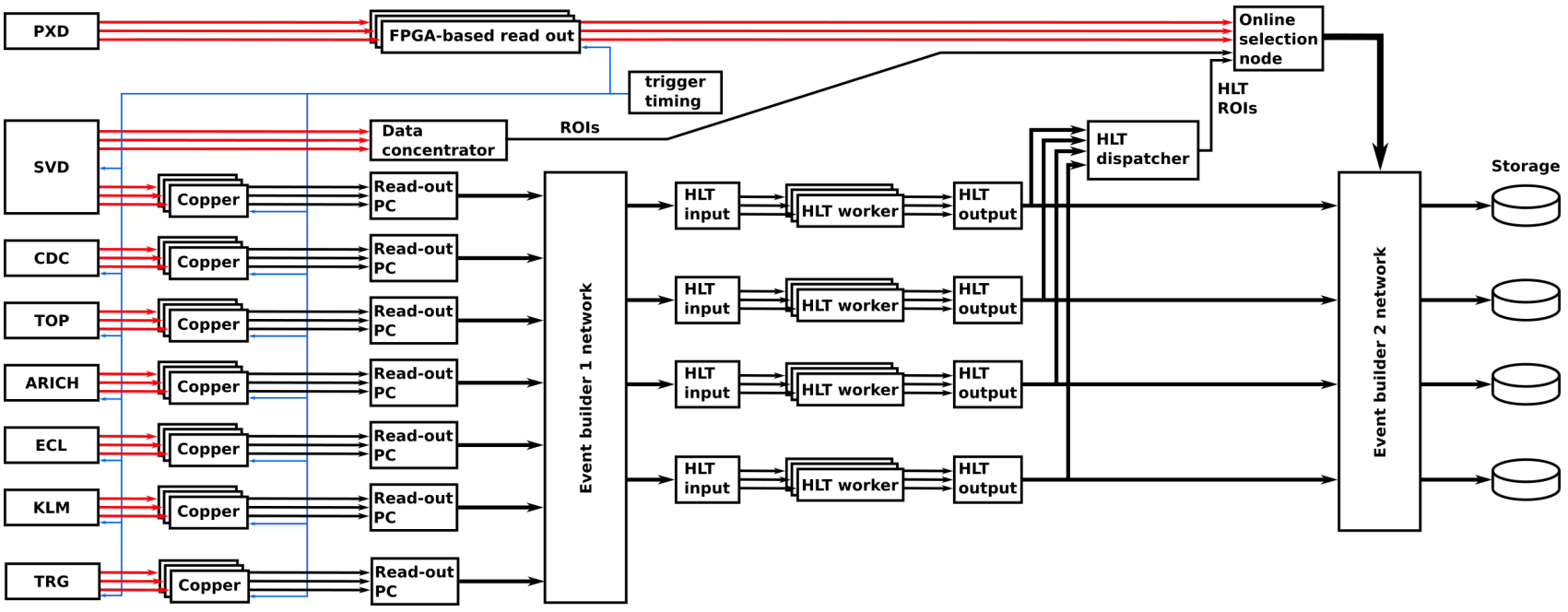
Backup slides

DMA Controller and Memory Organization

- Memory organization:
  - page-wise (8 kB) transactions: min. 1 page/event
  - superpage: 128 pages
- Ring buffer: 16 superpages in PC memory
- DMA operation:
  - superpage descriptors maintained by driver
  - page addresses calculated by DMA controller

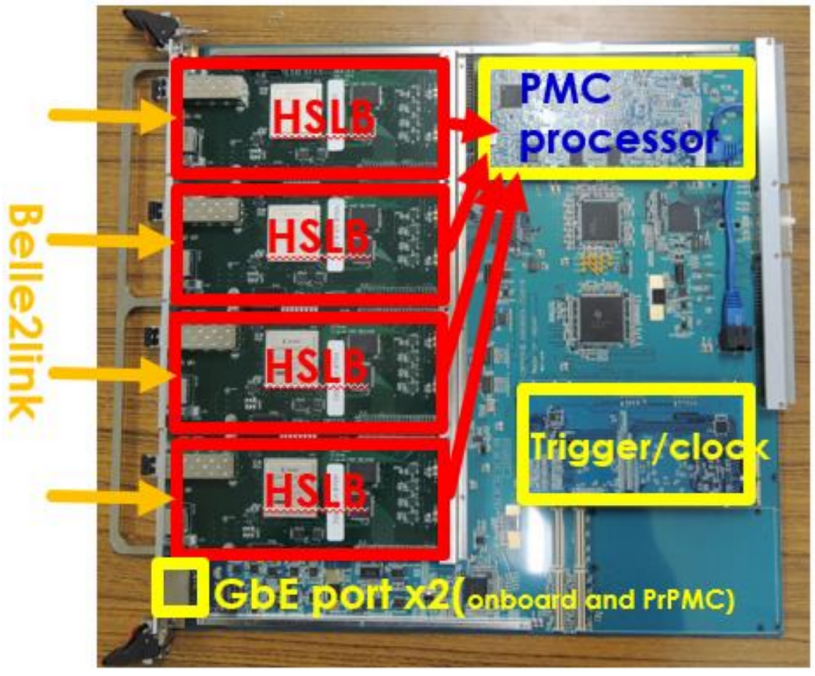


COPPER Read-out System



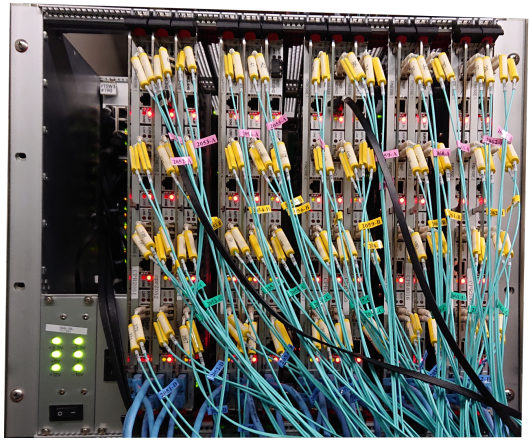
- COPPER: COmmon Pipelined Platform for Electronics Readout
- 600 read-out cards (HSLB), 150 COPPER modules

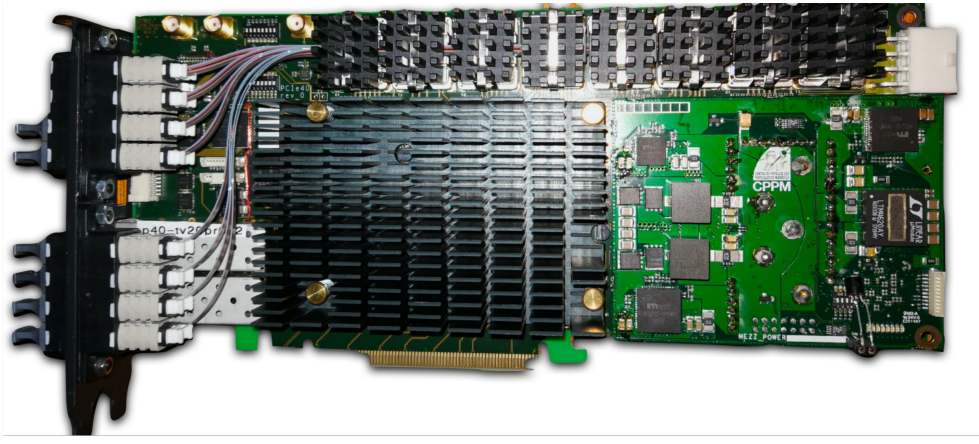
COPPER System and Its Limitations



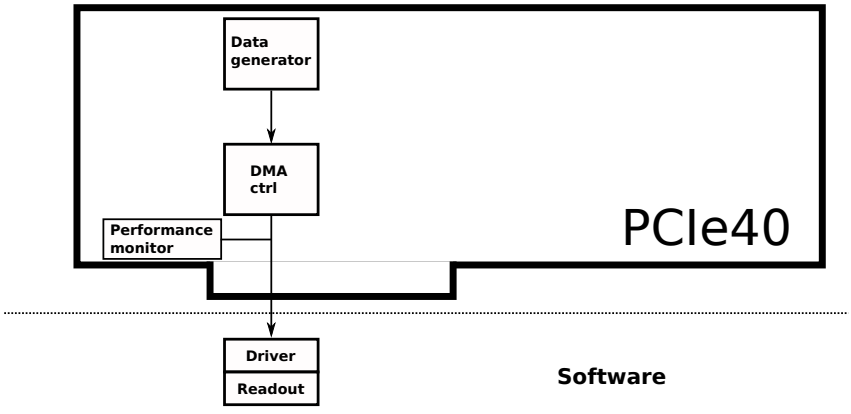
- High-speed link boards with Virtex-5 FPGA
- Event builder in Atom processor
  - 60 % CPU load at 30 kHz
- Output over 1 Gb/s Ethernet to read-out PC
- Deprecated hardware
  - Expensive maintenance

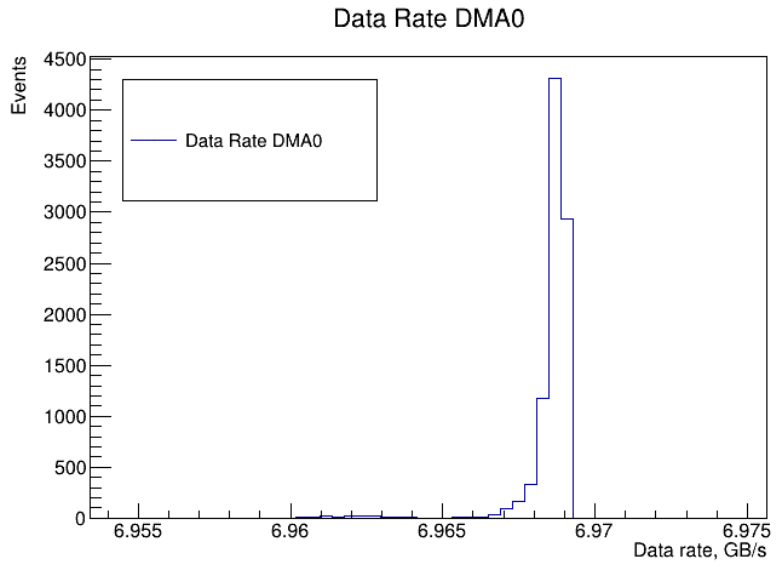
COPPER and PCIe40 Comparison





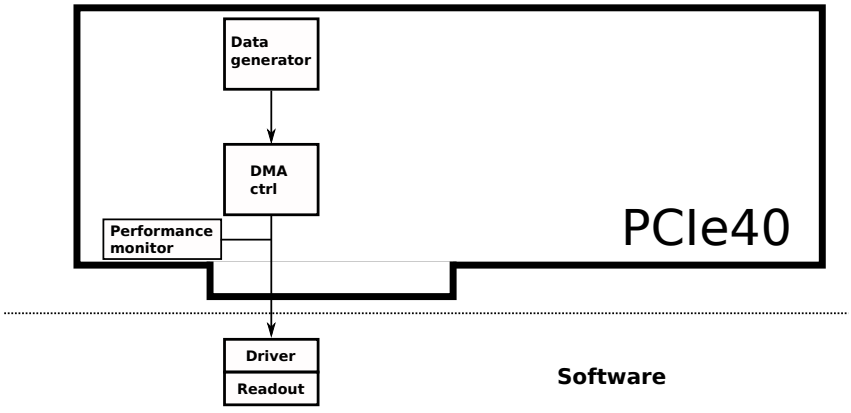
DMA Performance

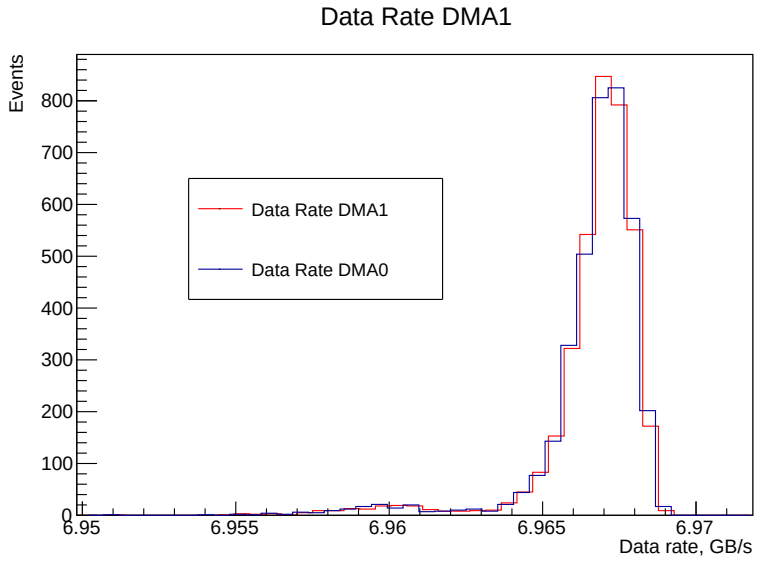




- Setup
  - data generation in firmware at **250 MHz**
  - discard data in software without processing
- Single DMA: **6.965 GB/s**
- Double DMA: **14 GB/s**

DMA Performance





- Setup
  - data generation in firmware at **250 MHz**
  - discard data in software without processing
- Single DMA: **6.965 GB/s**
- Double DMA: **14 GB/s**