DAQ FOR HIGHER L1 RATE + TRIGGERLESS DAQ

S. Yamada (KEK)

DISCLAIMER

- This talk is not about a concrete plan but just an idea to facilitate discussion about future high-rate TRG and DAQ. So, please do not too much worry about the contents in the talk, where feasibility or necessity was not extensively studied.

CONTENTS

- Current throughput limitation in Belle II DAQ
- Triggerless DAQ

DATA FLOW ESTIMATION (SCALING 2021B RUN'S EVENT SIZE)

Higher trigger rate : Let's say 100kHz
(No specific reason to choose 100kHz)

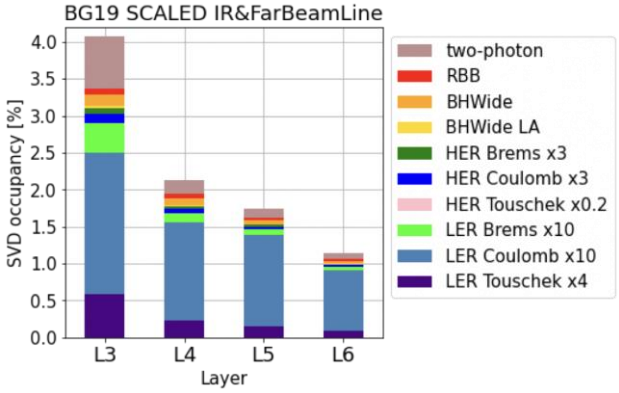
	# of ROPC (COPPER)	dataflow of the current system at 30kHz by just scaling data on May18.2021 [MB/s]	100kHz (3.3 x 30kHz) [MB/s]	# of ROPC (PCIe40)	100kHz /PCIe40 [MB/s]†
SVD	9	1026	**3420**	5	684
CDC	9	613	**2043**	7	292
TOP	3	208	**693**	2	347
ARICH	6	375	**1250**	2	625
ECL	10	601	**2003**	3	668
KLM	3	45	**148**	1	148
TRG	3	137(9COPPERs)	**457**	1	457
Total	43	3005	**10015**	21	

† Note : Since DAQ overhead size is different between COPPER and PCIe40 system, we should convert the event-size for the PCIe40 system, but I will use COPPER value in this talk.

DATA FLOW ESTIMATION (SCALING 2021RUN'S EVENT SIZE + SVD ESTIMATION)

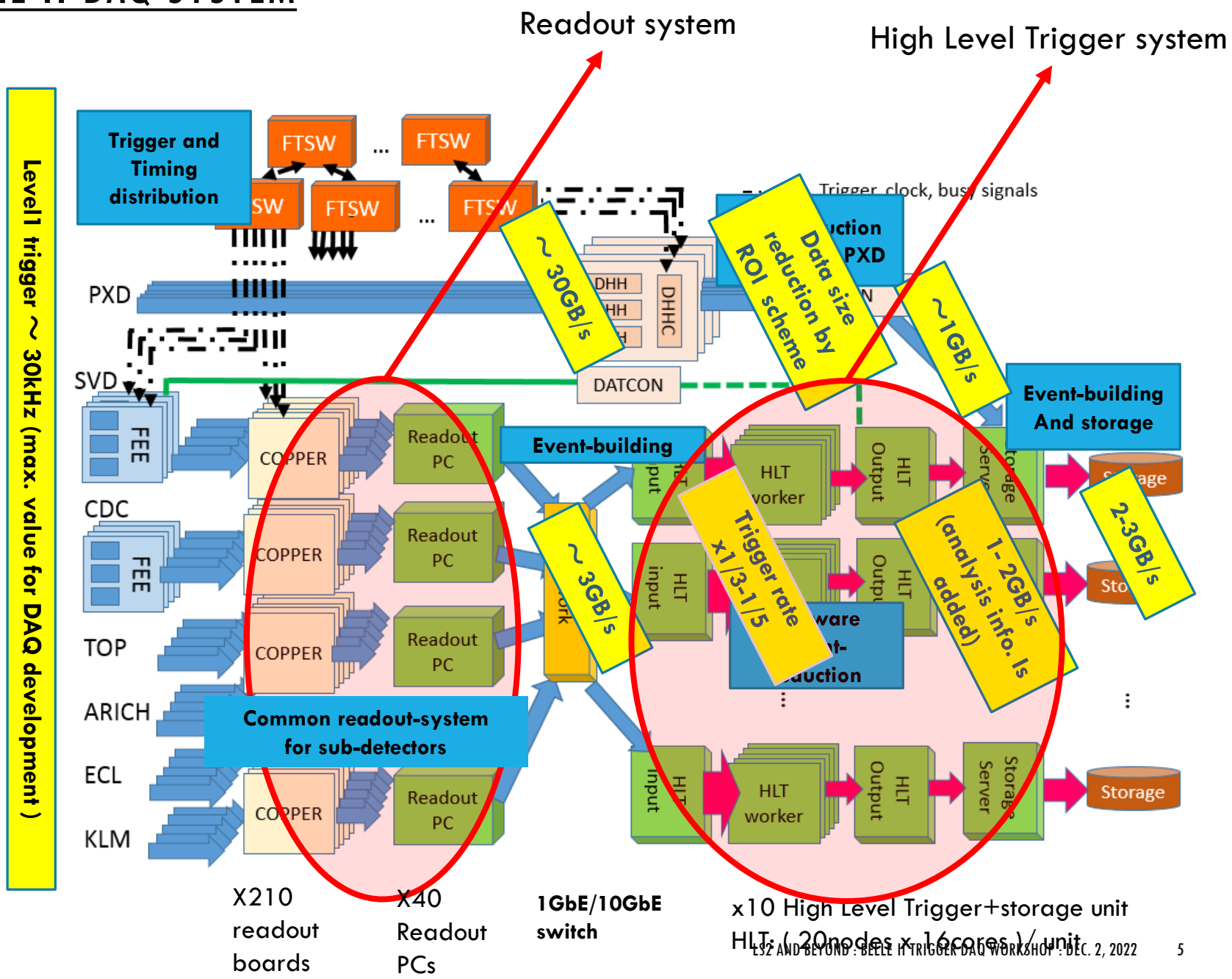
According to Katuro-san's e-mail "SVD PCIe40 打ち合わせ議事録"

	# of ROPC (COPPER)	dataflow of the current system at 30kHz by just scaling data on May18.2021 [MB/s]	100kHz (3.3 x 30kHz) [MB/s]	# of ROPC (PCIe40)	100kHz /PCIe40 [MB/s]†
SVD	9	3640	**12133**	5	2427
CDC	9	613	**2043**	7	292
TOP	3	208	**693**	2	347
ARICH	6	375	**1250**	2	625
ECL	10	601	**2003**	3	668
KLM	3	44.5	**148**	1	148
TRG	3	137	**457**	1	457
Total	43	5619	**18728**	21	



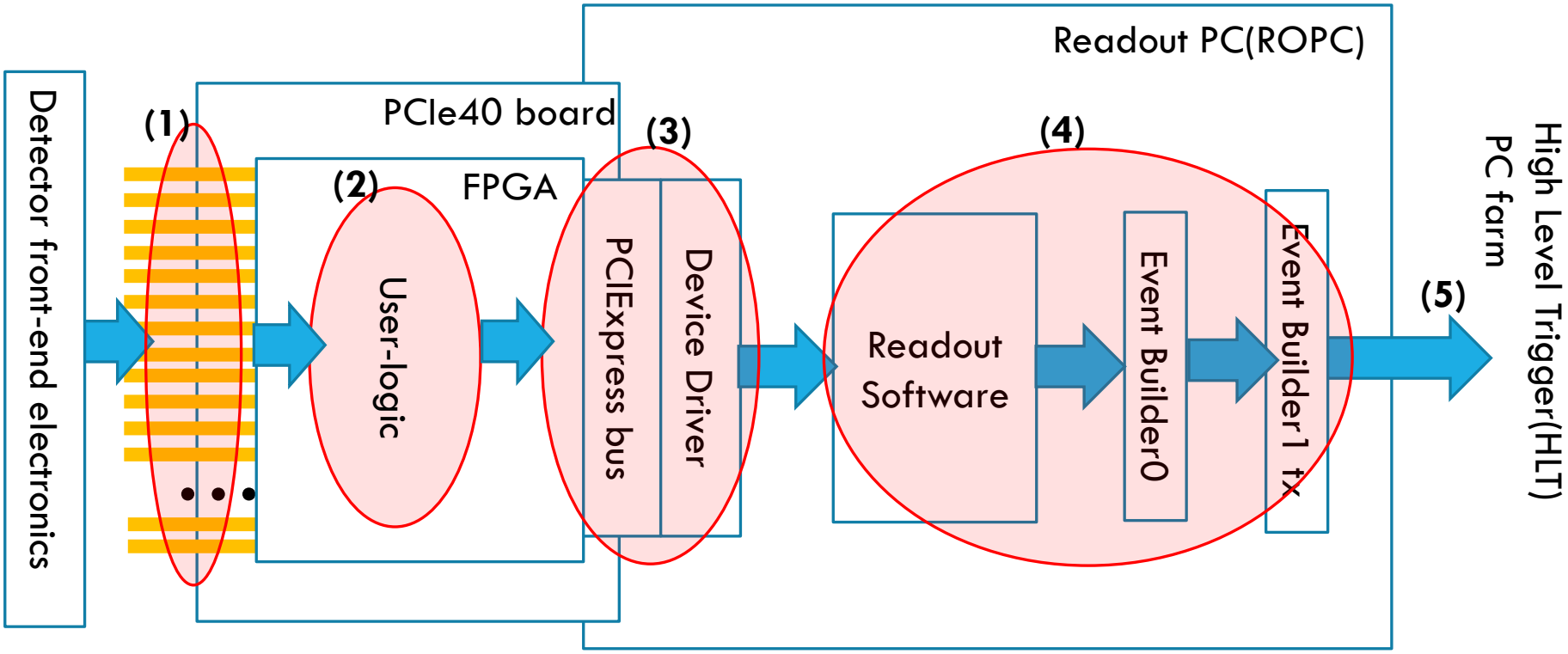
- According to recent SVD group's estimation, event size will be 3.6 times larger than 2021b run.
- **Same things could happen for other sub-detectors ?**

BELLE II DAQ SYSTEM



Readout system

POSSIBLE BOTTLENECKS IN THE NEW READOUT SYSTEM



(1) Belle2link -> No change (Line-rate 2.54Gbps)
  (1) Actual max. rate could be increased by belle2link upgrade work
(2) Processing data : formatting, event-building, data-check
  -> In recent development, event-building moves to ROPC
(3) DMA transfer via PCIExpress
(4) DAQ sotftware on readout PC
(5) Network bandwidth from ROPC to HLT -> 25GbE

(3) DMA transfer via PCIExpress

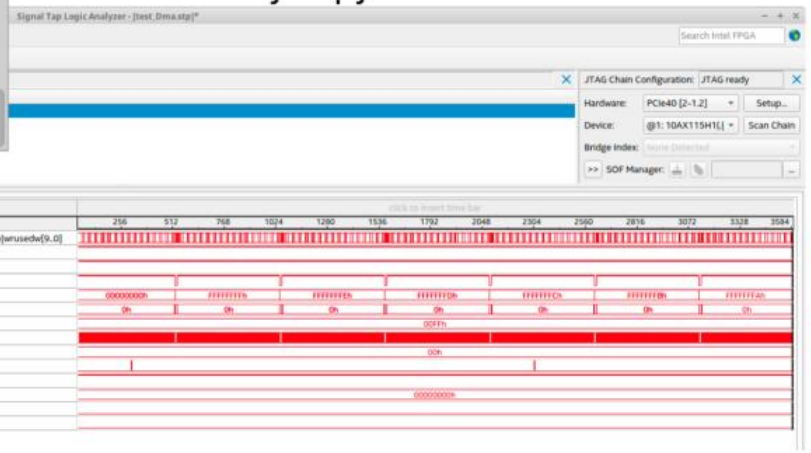
: Large than estimated data-rate at 100kHz

: Marginal : Not enough

(P. Robbe, D. Charlet)

Test 1: Maximum speed with back-pressure

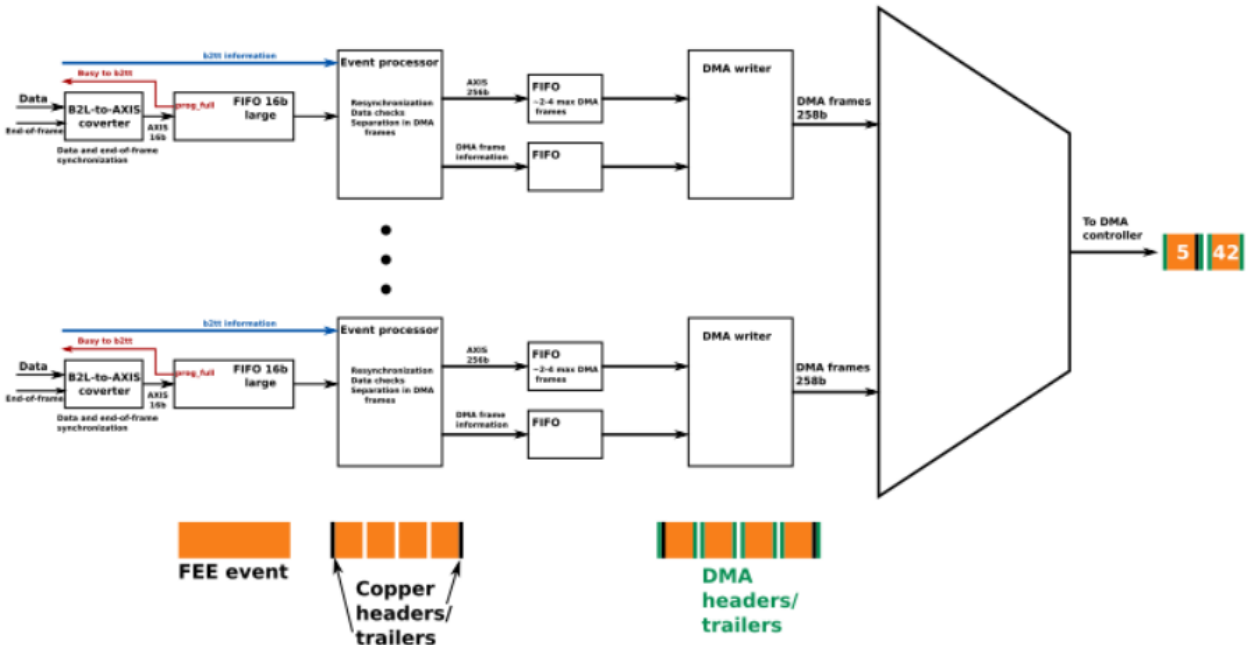
- Pulse trigger rate : 470 kHz (times 8 kBytes)
- Transfer data rate : ~~39 Gbits/s~~ 3.9GB/s
- 10 % of event with back-pressure active



	100kHz /PCIe40 [MB/s]†
SVD	2427
CDC	292
TOP	347
ARICH	625
ECL	668
KLM	148
TRG	457
All PCIe40	**18728**

(2)+(3)+(4) : PCIe40 firmware + software on ROPC

- PCIe40 firmware block diagram for software assisted event-building



- Performance of the system measured and bottlenecks identified
  - **2.4 GB/s** sustained data rate with CRC calculation
  - **4 GB/s** sustained data rate without CRC calculation

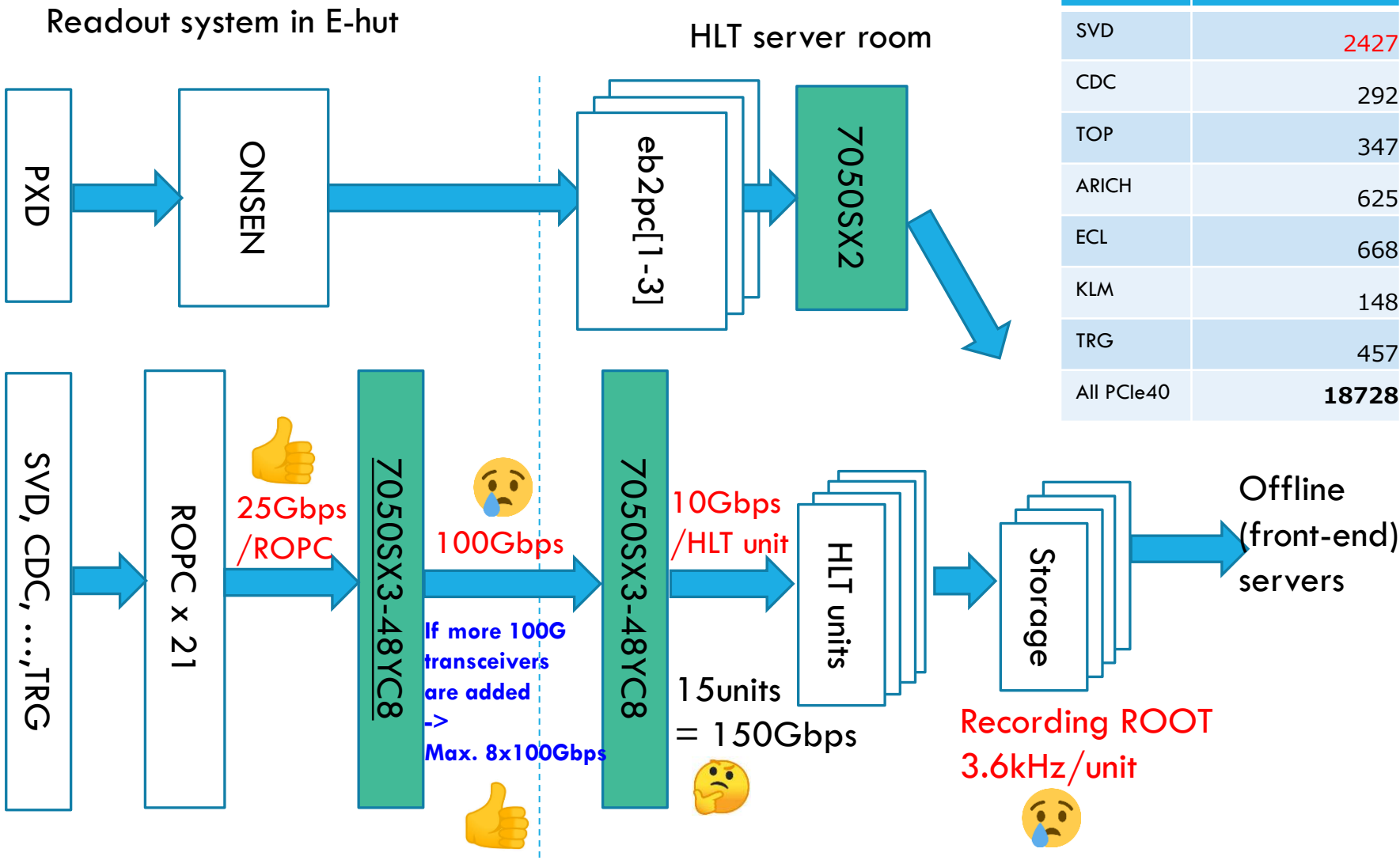
	100kHz /PCIe40 [MB/s]†
SVD	2427
CDC	292
TOP	347
ARICH	625
ECL	668
KLM	148
TRG	457
All PCIe40	**18728**

Note : This is the test result at test bench. Need to be checked in global Belle II DAQ.

Network and HLT

Network bandwidth

	100kHz /PCIe40 [MB/s]†
SVD	2427
CDC	292
TOP	347
ARICH	625
ECL	668
KLM	148
TRG	457
All PCIe40	**18728**



High-level Trigger

Number of physical cores after 2021 HLT Reinforcement

HLT01:
16 cores * 9 + 20 * (2+2) + 28 * 2 + 36 * 2 + 40 * 3 = 472 cores
(replaced 11 of 16 core servers with new ones).
HLT02-05
20 cores * 16 + 36 cores * 2 + 40 cores * 2 = 472 cores
HLT06-09
28 cores * 12 + 36 cores * 2 + 40 cores * 2 = 488 cores
HLT10
28 cores * 12 + 36 cores * 2 + 40 cores * 2 = 488 cores

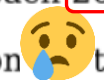
4800 cores

- The reinforcement achieved 75% of the design number of cores(6400).
- At the same time, the operating system has been upgraded to CentOS7.

Summary

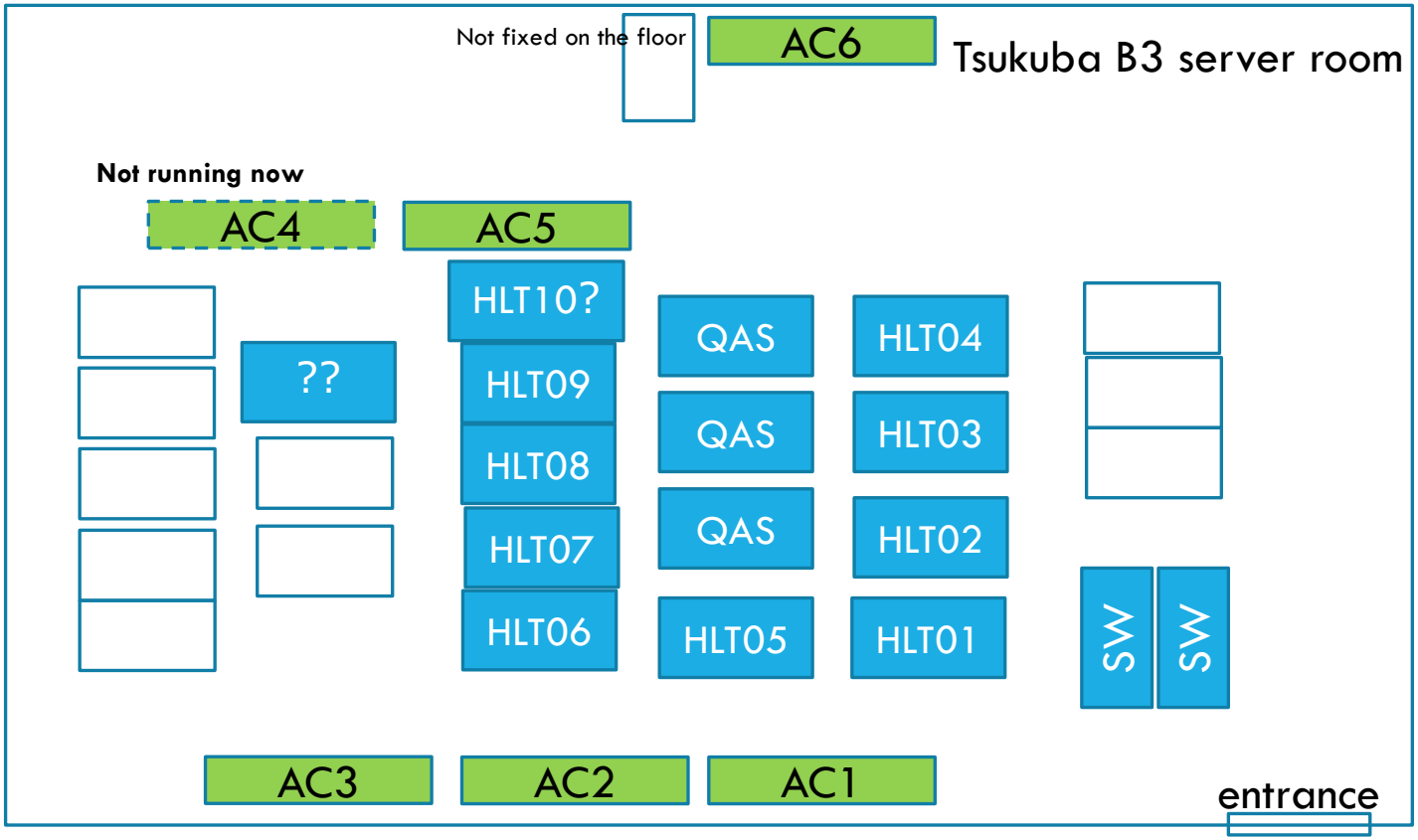
- Optimization during data taking (release-05) allowed us to survive until LS1 (13 kHz)
- with 3 HLT units more + release-08 (event time+ track fitting improvements)

⇒ should be able to reach 20 kHz (Luminosity ~ $10^{35}/cm^2/s$)

- need to carefully mon  tracking performance at higher luminosity
- need to make sure software development keeps CPU budget under control
- no clear path beyond 20 kHz... unless significant improvements in pattern recognition

Room for more HLT units in B3 server room (1)

- # of Vacant fixed racks : 10
  - (According to Itoh-san, max. # is 15 and increase servers in a unit)
  - Maybe more room to place racks in the room

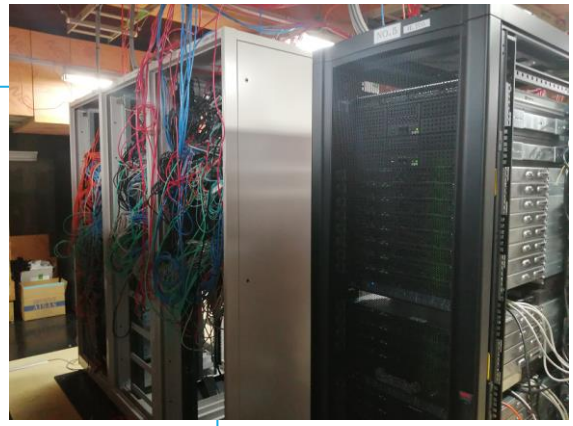


QAS : Quick analysis servers

AC : Air conditioner

SW : network switches

Room for more HLT units in B3 server room (2)



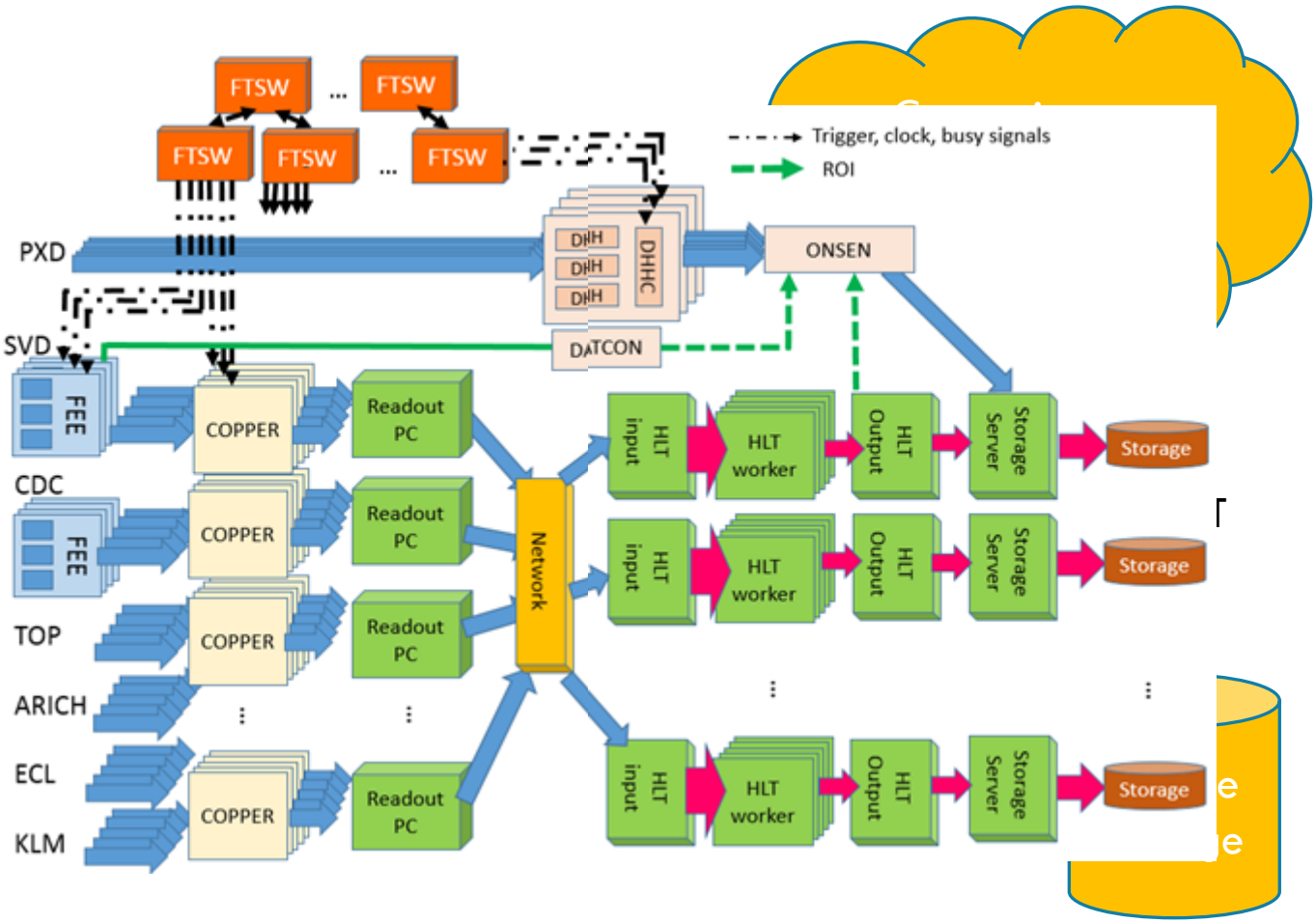
- Cooling in the room
  - Currently 6 Air-conditioner units
    - Power-consumption : 10kW/unit†
  - Need to add more units ?
- Power-supply in the room
  - Power consumption : ～10kW/unit †
  - ExpressReco+QAS : ～10kW
  - HLT 15units operation : 15x10 + 10 = 160kW
  - Need to add additional power supply to the room ?
- Cost of servers
  - 1 HLT unit ～ 20M JPY

† Itoh-san's e-mail, "Re: [str-jp:442] 電気代節約作戦"

- Hardware trouble of such a large amount of servers could be an issue in coming years.

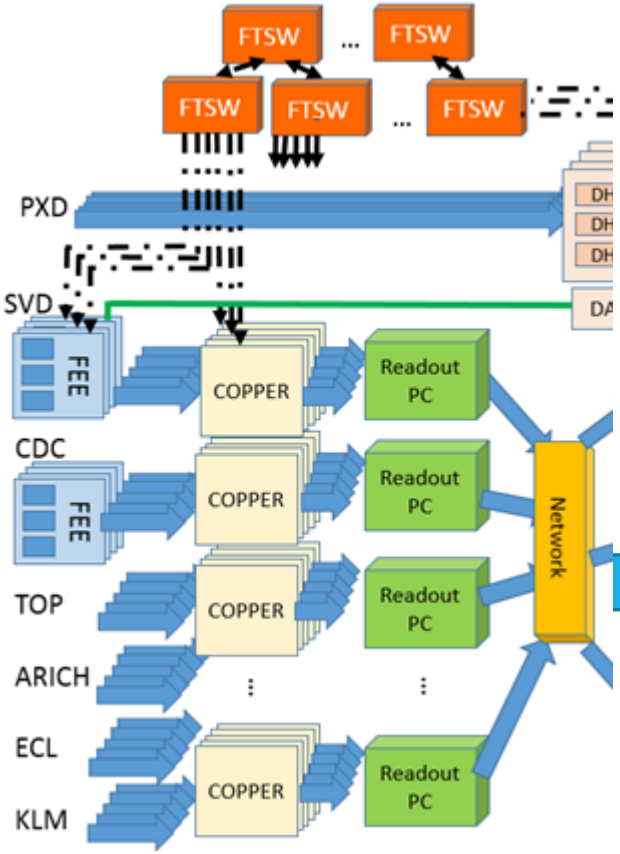
IF “high-level trigger” computing is outsourced… (1)

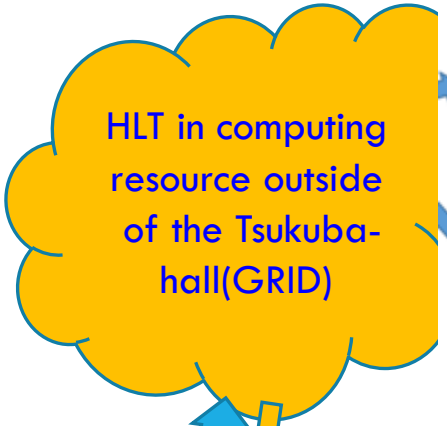
- Store all L1 trigger events
  - Reconstruction done by PC farm somewhere
  - Selected events are permanently stored

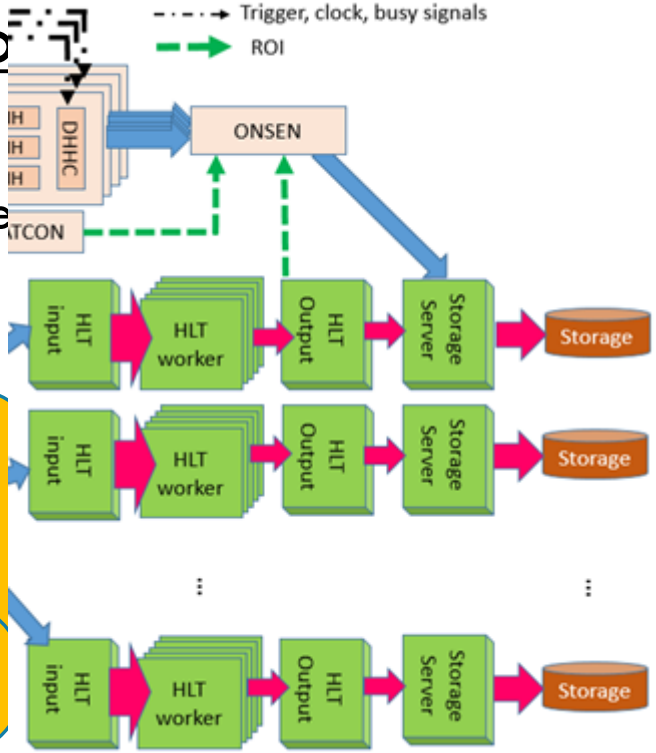


IF “high-level trigger” computing is outso

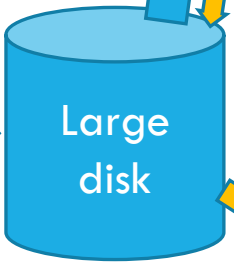
- Store all L1 trigger events
  - Reconstruction done by PC farm somewhere
  - Selected events are permanently stored







Selected events by HLT



IF “high-level trigger” computing is outsourced… (2)

If we assume;
- HLT CPU : Intel Xeon E5-2650 v4
- HS06/core for the CPU : ～11

Then, HLT CPU power : 6400cores x 0.011HS06 = **70.4 kHS06** ?

- HLT CPU power is still comparably large compared with the current Belle II computing resources.
  - Currently, it seems difficult to secure computing resource for HLT.

From computing coordinator’s report at the 42nd B2GM

c.f. Pledged resources for 2022 JFY

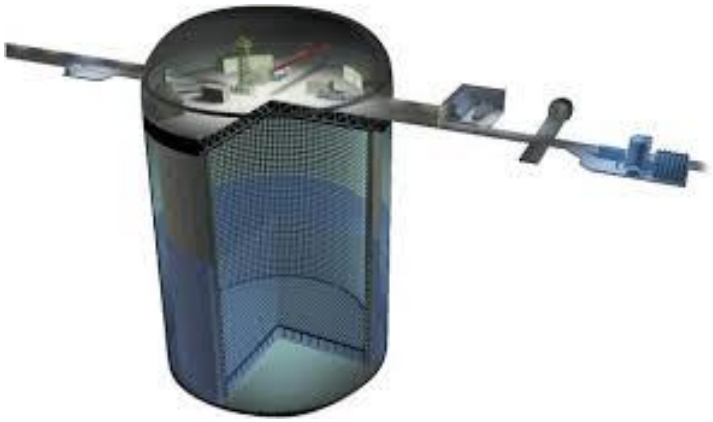
Year	Pledged 2022
Total tape (PB)	8.8
Total Disk (PB)	16.5
Total CPU (kHEPSpec)	385

Other things to be considered…
- Large network bandwidth to/from KEKCC
- Reconstruction for ROI calculation is at least necessary for PXD data reduction
  - Only limited HLT for this in Tsukuba-hall ?
- Online data-quality monitor by reconstructing pre-scaled events.

What is Triggerless DAQ ?

Neutrino experiment

- Triggerless DAQ in Super-Kamiokande was deployed in 2008
  - No hardware trigger but online software event-selection is applied.
- Data : PMT dark rate + cosmic ray interaction and decays in the water tank
- Software trigger for such as delayed trigger can be easily implemented.
  - e.g. Detect neutron captured gamma a few microseconds after an event
    - In Belle II, particles fly away from the detector in nanoseconds…



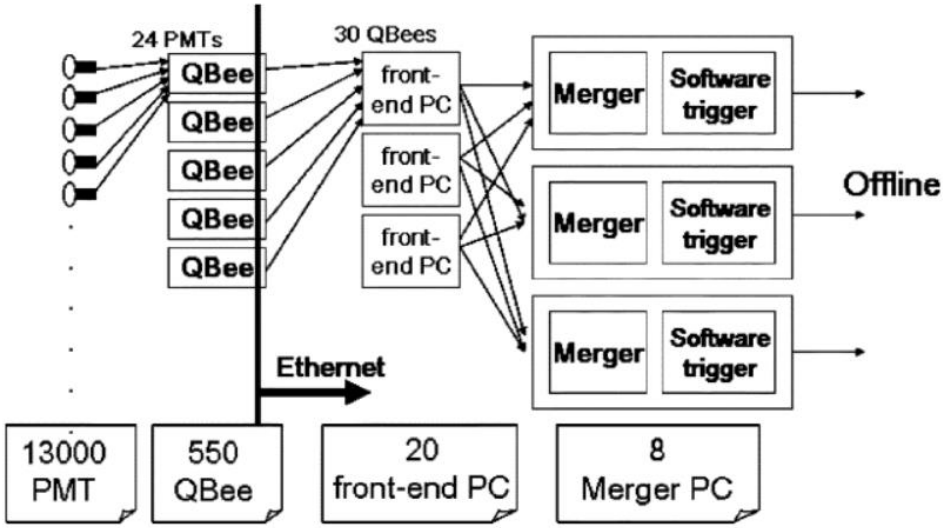
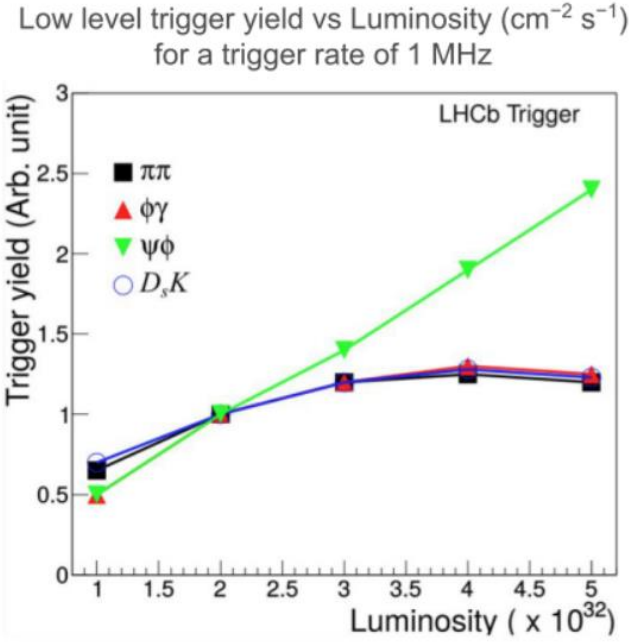
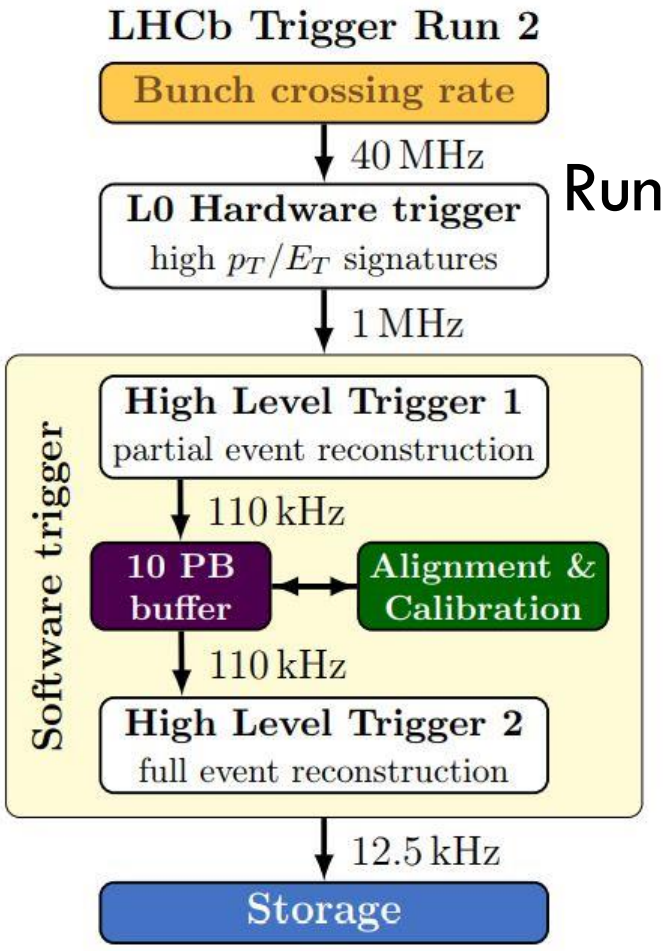


Fig. 2. Schematic view of the new data acquisition system. In the left part, newly developed front-end electronics boards connected with PMTs are shown. In the right part, components of the new online system are shown. Those PCs are connected via Ethernet.

LHCb

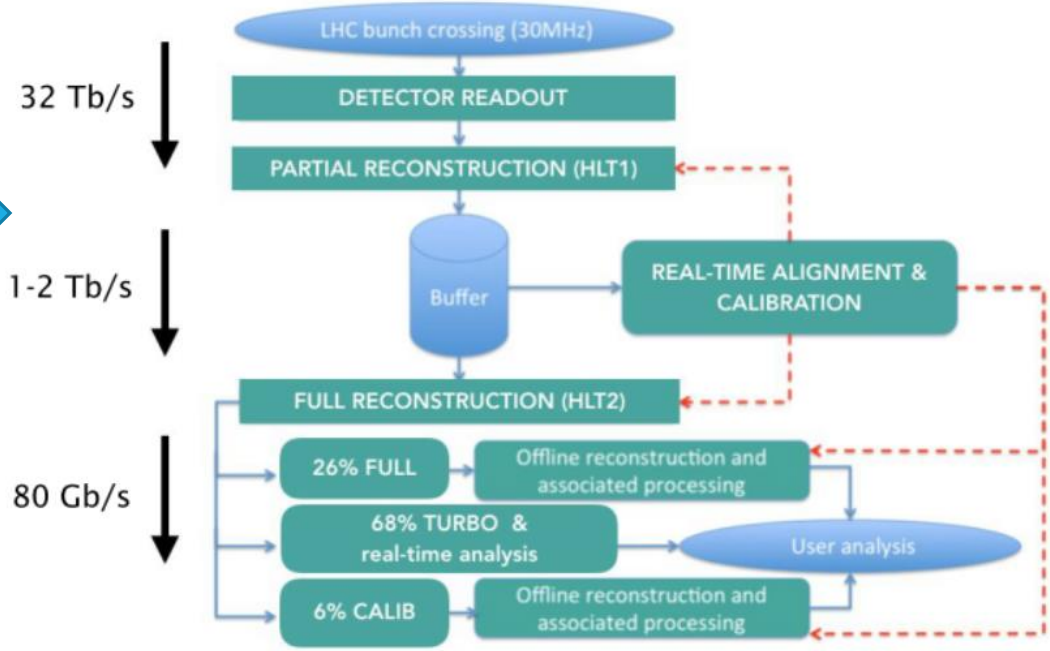
- Hardware (L0) trigger : based on calorimeter and muon system information(high pt/Et)
- Better efficiency can be achieved by using all detector information





Run2

Run3



LHCb : From Nico-san's slides

https://indico.cern.ch/event/1037447/

Triggerless read-out: Why?

- No trigger is 100% efficient, hardware triggers always have to compromise
- A trigger-less front-end is much *less complex* and *more robust*
  - No buffering, no selection logic,
- Selection in "software" / using compute infrastructure from the data-centre world has a lot of advantages:
  - Scalability
  - Cost: costs in electronics are driven (down) by scale
  - **Cost-efficiency**: 1 USD spent on an ASIC trigger board will be "active" only during the operation of the trigger, 1 USD spent on a CPU or GPU can be "active" round the clock
  - Flexibility: new hard and software can be integrated without impacting the detector or operations
  - Operations: compute for filtering does not need to be "on-prem", *not operated by experiment, not even "owned"*

Triggerless read-out: Why not?

- More data from the front-end
  - 0-suppression / compression / clever encoding required
  - more links → more power, more cost, more material
- Cultural shift:
  - from electronics to software engineers
  - from hardware projects to software commitments: What does your funding agency say?

In Belle II case…

Belle II case …

https://indico.cern.ch/event/1037447/

Triggerless read-out: Why?

- No trigger is 100% efficient, hardware triggers always have to compro
- A trigger-less front-end is much *less complex* and *more robust*
  - No buffering, no selection logic,
- Selection in "software" / using compute infrastructure from the data-centre has a lot of advantages:
  - Scalability
  - Cost: costs in electronics are driven (down) by scale
  - **Cost-efficiency**: 1 USD spent on an ASIC trigger board will be "active during the operation of the trigger, 1 USD spent on a CPU or GPU ca "active" round the clock
  - Flexibility: new hard and software can be integrated without impacting detector or operations
  - Operations: compute for filtering does not need to be "on-prem", *not operated by experiment, not even "owned"*

BB pair : 100% efficiency
This part (physics motivation) should be more studied.
(low-multiplicity event ?
Displaced vertex ₮?

This part is true even if hardware trigger is still available ?

N. Neufeld CERN - Triggerless readout 22

Triggerless read-out: Why not?

- More data from the front-end
  - 0-suppression / compression / clever encoding required
  - more links → more power, more cost, more material
- Cultural shift:
  - from electronics to software engineers
  - from hardware projects to software commitments: What does your funding agency say?

Large upgrades of FEE will be necessary.

More HLT computing resource, network bandwidth,
New readout board ?

TRIGGER EFFICIENCY IN BELLE II

L1 Trigger Menu for Low Multiplicity Physics BELLE2-NOTE-PH-2015-011

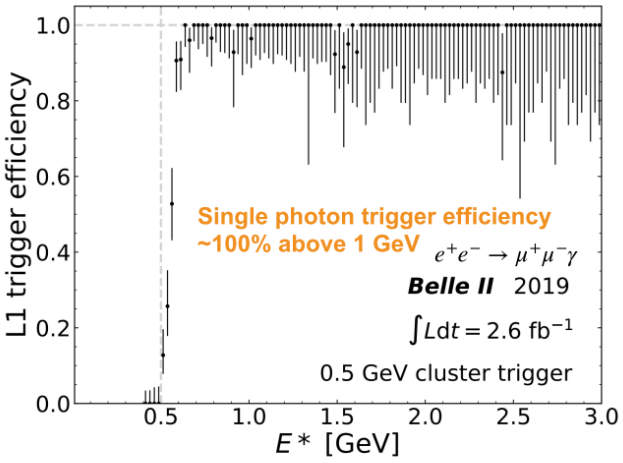
TABLE VIII: Efficiencies and Cross section after triggers

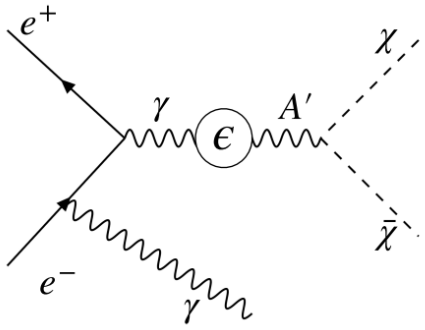
	Processes	T1:2trk	T2:1trk1mu	T3:1mu	T4:1trk1c	T1:bbc	T2:3g	T3:3t	Combine
	$B^0\bar{B}^0$	-	96.5	50.0	82.9	44.8	93.4	99.4	> 99.9
	B^+B^-	-	96.5	51.7	84.1	46.2	92.6	99.5	> 99.9
	ccbar	-	96.8	65.9	89.4	52.1	84.8	98.0	> 99.9
	uds	-	96.5	68.0	89.1	50.0	81.1	97.2	> 99.9

- Already very high for hadronic events in Belle II
- What kind of events can benefit from triggerless DAQ/software trigger ?
  - Low multiplicity event ?
  - Displaced vertex ?

Single Photon Search

- Search for massive Dark Photon, A', which mixes with Standard Model photon.
- Detector signature is a single initial-state radiation photon.

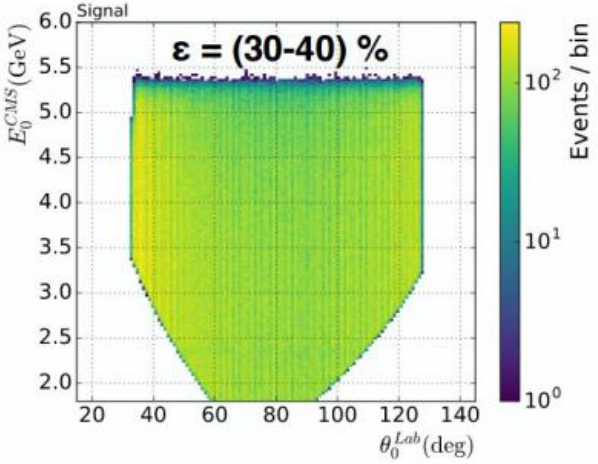




- Single photon trigger is crucial:
  - ➡ Maintaining acceptable rate challenging due to beam-induced backgrounds

Discriminant variables:

E_{CMS} vs. polar angle of "single photon"



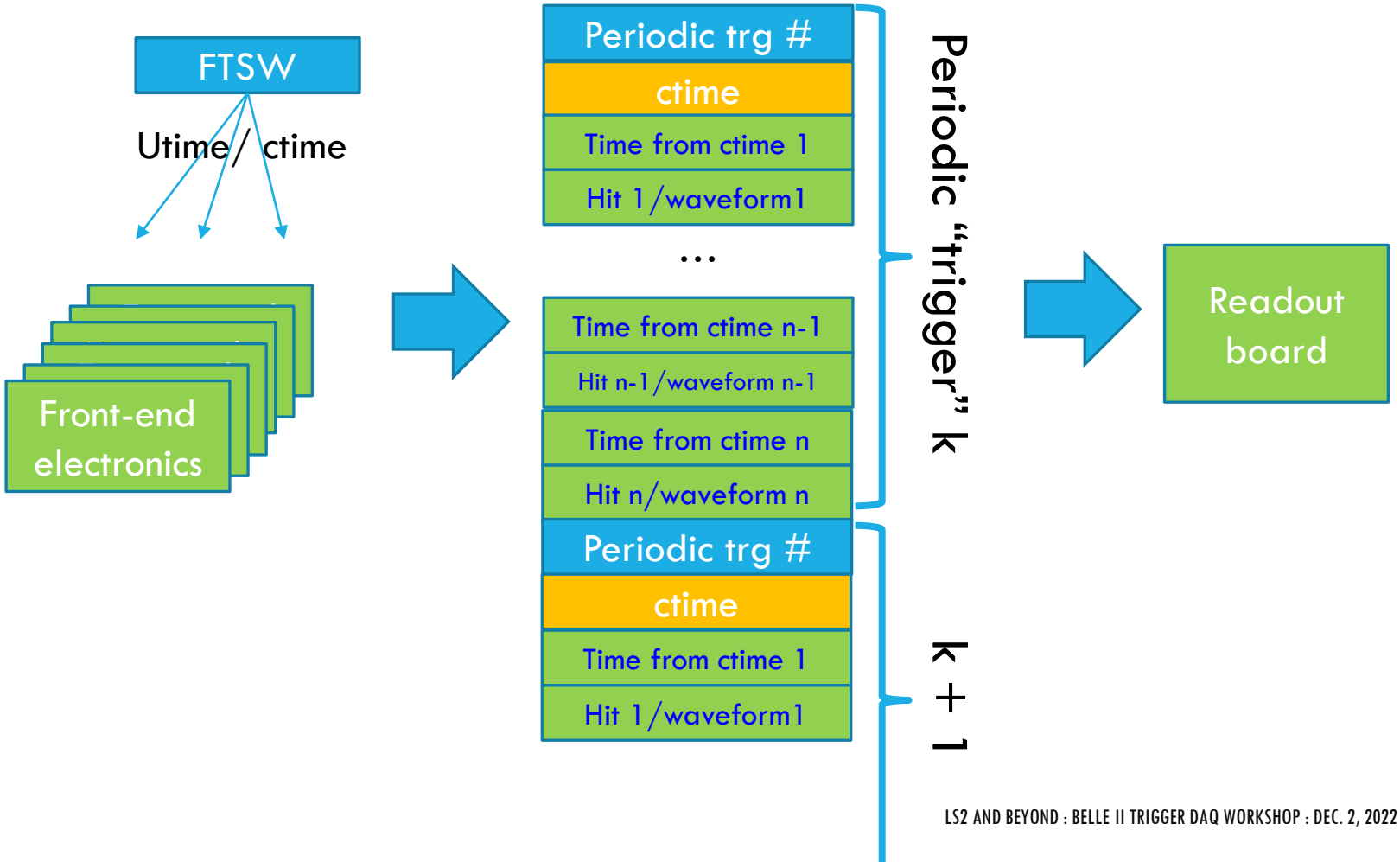
Signal signature:

peak in E_{CMS} (horizontal band)

But… If TRG experts thinks that they can reduce trigger rate more easily by software trigger with event data, that could be another motivation.

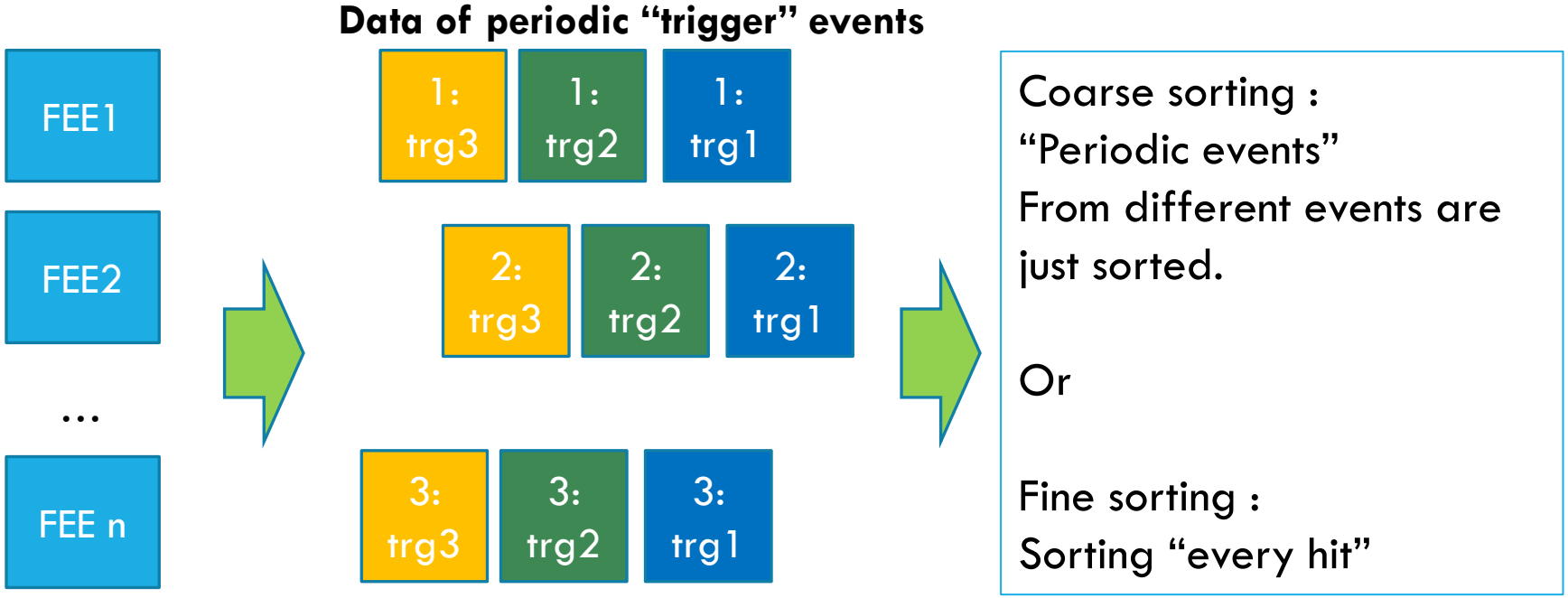
What needs to be done in triggerless DAQ (1)

- Data format
  - No event #
    - Instead time counter from FTSW will be used (timestamp)
  - It would be good to have coarse time-stamp and fine timestamp



What needs to be done in triggerless DAQ (2)

- Time sorting
  - Instead of event-building, data from different FEEs needs to be sorted for software trigger.

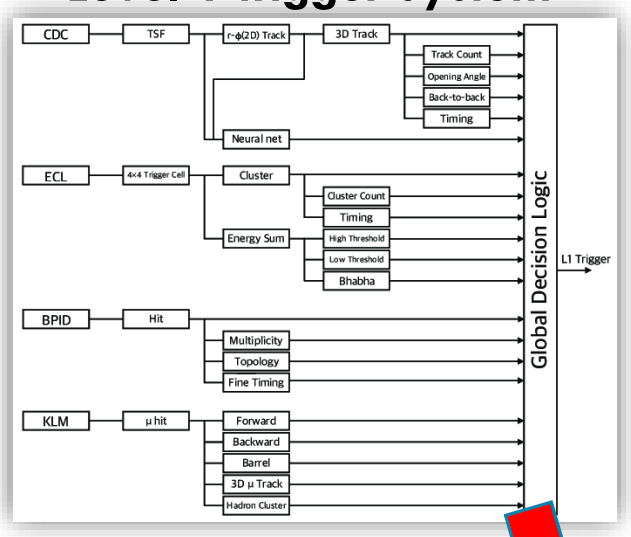


Probably, this part is also resource-consuming.
It will be useful to use FPGA or GPU for the sorting.

It could affect the load of software trigger.

Trigger system then moves to PC farm…

Level 1 trigger system



- Hardware trigger system itself might stay as a throttling system for a while.

PXD → ONSEN

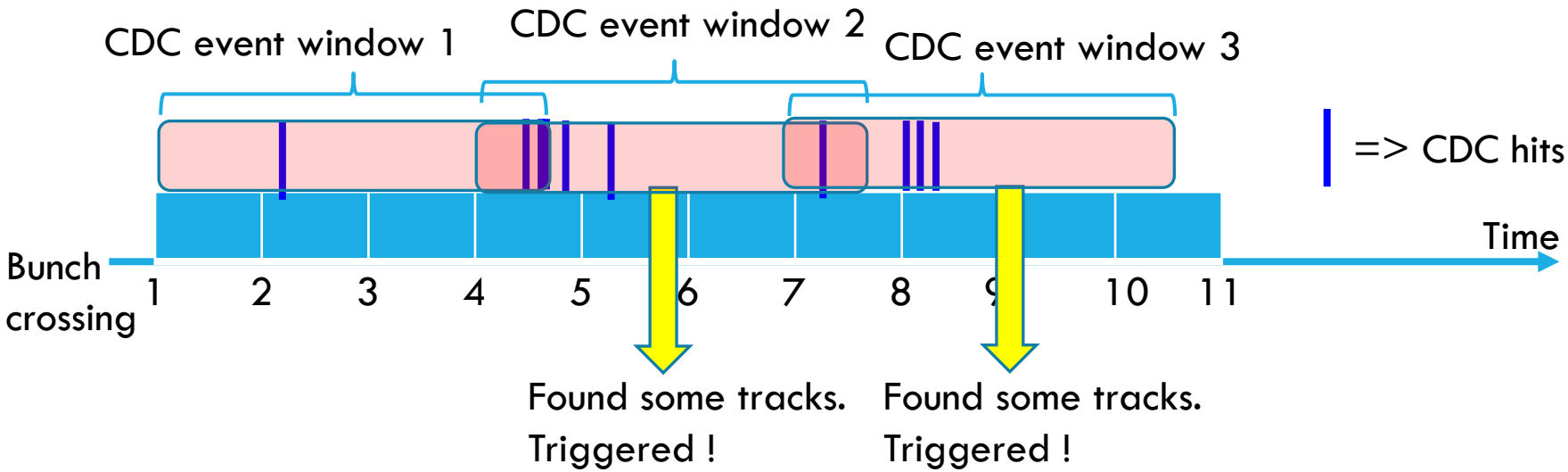
Detector FEE → Readout system → **"Level 1" trigger by software** → High-level Trigger (by offline reconstruction software) → Storage

(Small effect to software gr.)

TRG group's PC farm

What needs to be done in triggerless DAQ (3)

- Online event selection
- How to set trigger-window is a complicaed part
- Check every bunch crossing ? :
  - LHCb max. 40MHz(=25ns)
  - Belle II 254MHz(=4ns)
    - Smaller than timing resolution of some sub-detectors
- Set a certain trigger-window which has some overlap



About this session…

Feasibility of higher rate or ultimately trigger less(or streaming) readout rely on not only online/offline system but also front-end electronics capability.

- I understand that the current FEEs of many sub-detectors are not capable of triggerless or streaming readout. Anyway, inputs from sub-detector DAQ experts about the performance of their FEEs is important for higher trigger rate operation.

- Acceleration by hardware (GPU, FPGA) could cost-effectively improve the situation ? -> Zhou-san's talk

SUMMARY

- For higher trigger rate (e.g. 100kHz) in future, we estimate the current limitation in the throughput of DAQ.
  - Readout system and network bandwidth to HLT is not so bad.
  - More CPU power is needed for reconstruction software in HLT
  - Off-loading HLT computation to Belle II GRID computing resource for HLT is currently unrealistic.
  - Adding HLT units :
    - Need to consider cost, space, power-consumption, cooling
- Triggerless DAQ
  - Recently, HEP experiments start to adopt the scheme.
  - Upgrades of FEEs for many sub-systems will be necessary.
  - Motivation for Belle II physics (improve efficiency etc.) needs to be considered more.
  - Software trigger with all data could help TRG experts' efforts to reduce L1 trigger rate before HLT.

Backup

Advantage of steaming readout (towards EIC DAQ)

https://www.bnl.gov/srv2019/

The next logical step is to eliminate the hardware trigger altogether, and replace the trigger decision with a data selection realized in software, with the following advantages:

- Since all data is already in the digital domain, latency constraints on the selection algorithm are seriously loosened compared to a hardware trigger.
- The software algorithm can access all detector information, allowing us to better suppress noise and be more efficient.
- A streaming readout is, in principle, less complex. Many problems are moved from hardware to software, where better tooling and more expert knowledge is available and more people can contribute. Other problems are reduced due to less hardware being required, or due to the removal of the event building bottleneck.
- The architecture furthers the convergence of online and offline analysis, leading to better data quality control during data taking and shorter analysis cycles.
- Streaming readout allows the efficient readout of detectors operating on longer timescales like TPCs at high event rates without incurring excessive dead-time, and simplifies the read-out of high-channel count, high-rate detectors as it can be scaled up easily.
- For bandwidth-to-disk-limited experiments (e.g. Run 3 of LHC), a streaming readout combined with online analysis allow to drastically reduce the amount of data stored for each event by pre-processing the raw data to extract features like clusters, or even fully analyze the raw data and store only analysis level data structures. This maximizes the amount of physics that can be extracted from the experiment.

For EIC, the adoption of streaming readout has significant advantages:
Current rate predictions indicate that, contrary to LHC Run 3, all raw data can be saved to disk after a first-level zero-suppression, maximizing the physics impact of EIC by allowing for data mining in a completely unbiased data set.

CBM experiment(1)

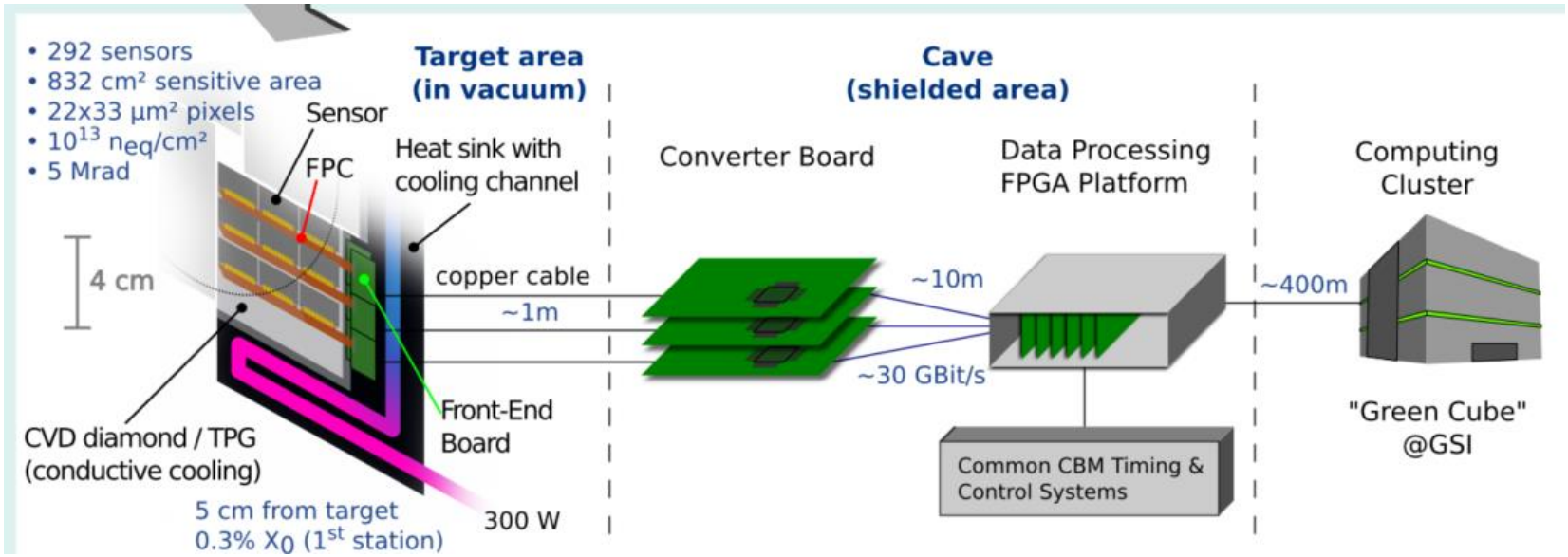
The Compressed Baryonic Matter (CBM) experiment is designed to explore the phase diagram of nuclear matter in the region of moderate temperatures and high net baryon densities. Due to the high expected reaction rates and particle multiplicities, CBM has to run free-streaming (triggerless) and thus requires extensive online data pre-processing by a computing cluster.

Event rate : up to 10MHz
Raw data rate : 1 TB/s

Sorting by FPGA

Online reconstruction+ event selection

Readout chain



CBM experiment(2)

Fluctuation of particle rates

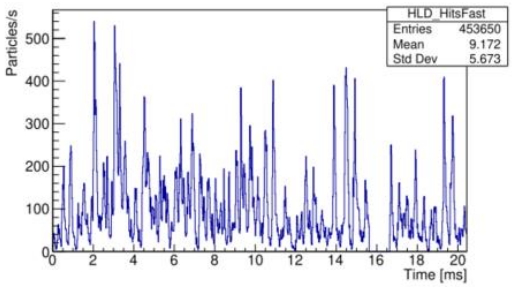


Figure 5: Micro-spill time structure of the particle rates.

Throttling system will be implemented.

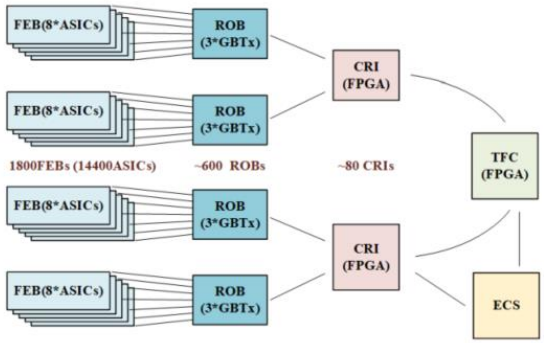


Figure 1: Hierarchy of the readout tree of the STS subsystem

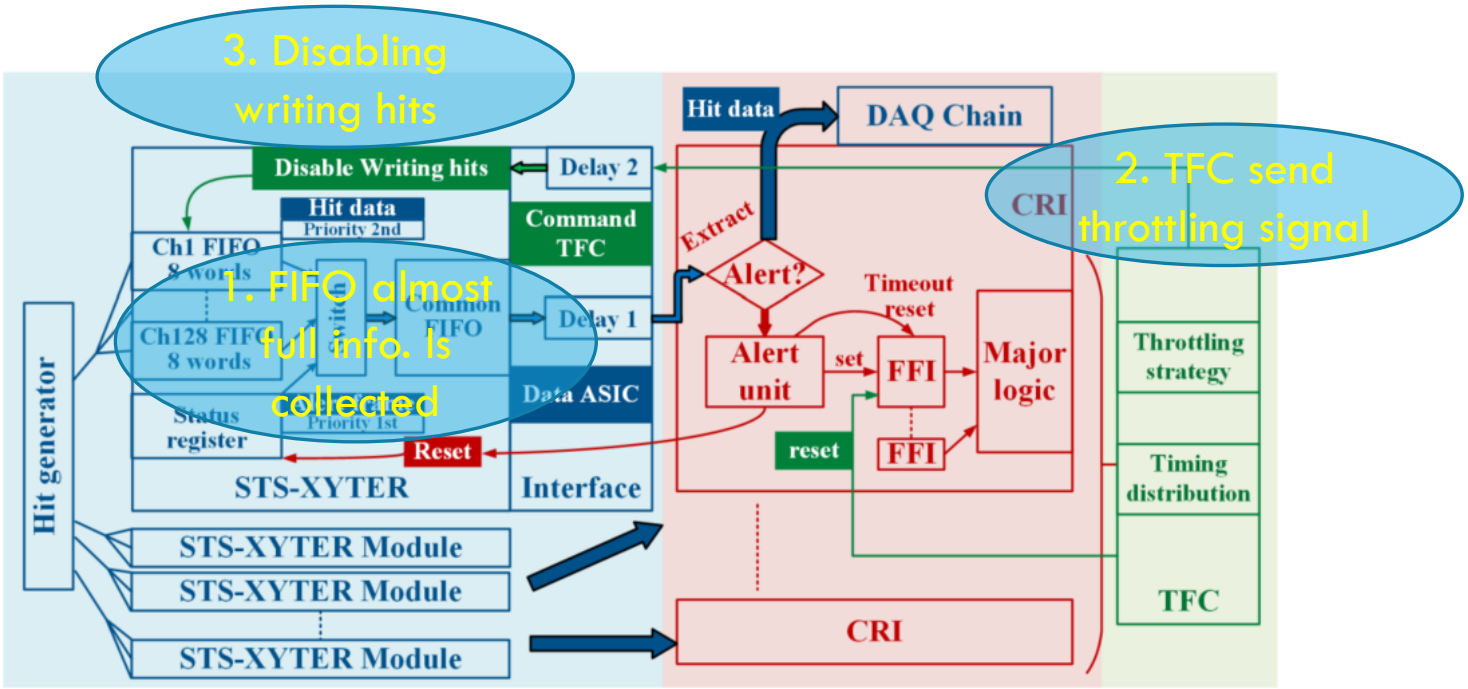


Figure 2: Hardware functional diagram