ZeroMQ Online ROOT-Output Storage and Express-Reconstruction System for the Belle II Experiment

Seokhee Park *et al.*

seokhee.park@kek.jp

KEK

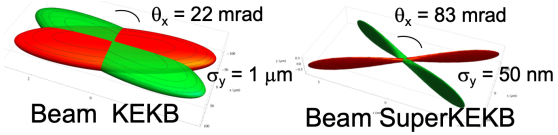
on behalf of the Belle II DAQ group

2022 August 1st

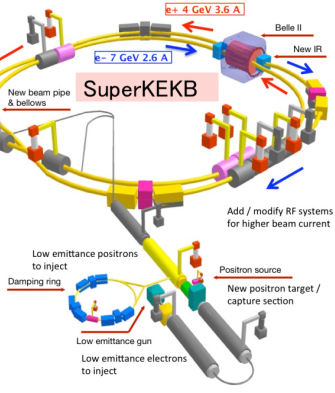
23rd Virtual IEEE Real Time Conference

SuperKEKB

- **Electron-positron collider with 7 GeV e^- and 4 GeV e^+**
  - Focused on $\Upsilon(nS)$, mainly $\Upsilon(4S)$
- **Aiming at 50 ab^{-1} of data (= 50× Belle) → Achieved 424 fb^{-1}**
- **Aiming at $6.5 \times 10^{35}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ of peak lumi (= 30× KEKB) → Achieved $4.7 \times 10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$**
  - corresponding to 30 kHz L1 trigger rate
  - 1/20 of beam size (nanobeam scheme)
  - 150% of beam current



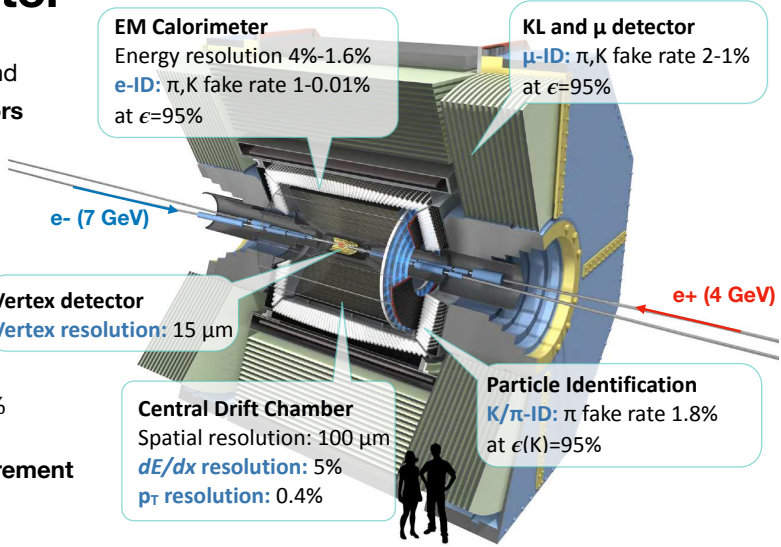
$$L = \frac{N_+ N_- n_\mathrm{b} f_0}{4\pi \sigma^*_{\mathrm{x,eff}} \sqrt{\varepsilon_\mathrm{y} \beta^*_\mathrm{y}}}$$



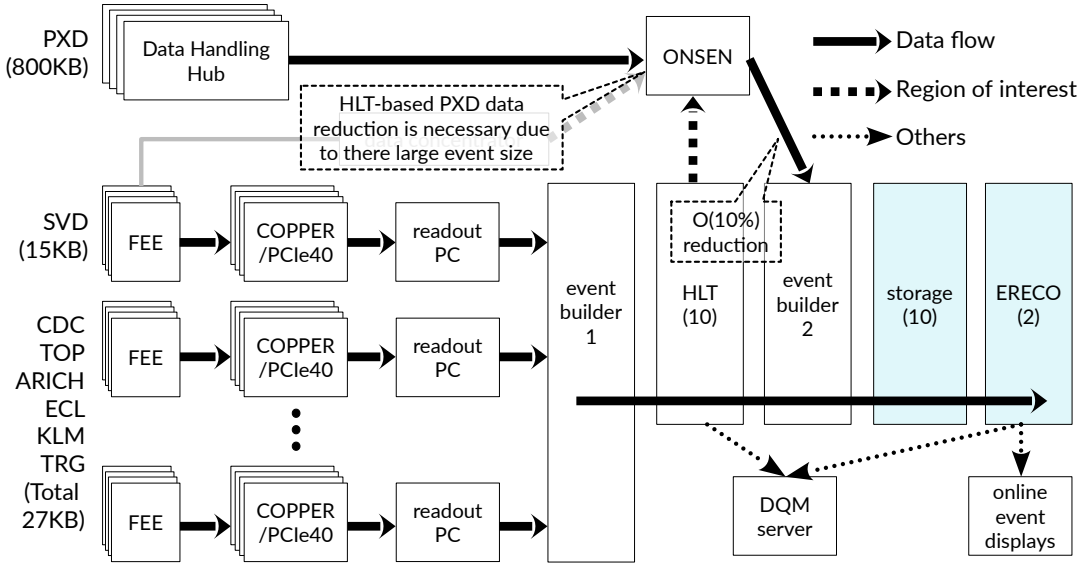
Belle II TDR, arXiv:1011.0352

Belle II detector

- Increased beam background
  - →**Upgraded sub-detectors and trigger**
- **βγ**=0.28 (vs 0.42 @KEKB)
  - →Reduced boost requires **improved vertex reconstruction:**
- Solid angle coverage >90%
  - →**High hermeticity for missing energy measurement**

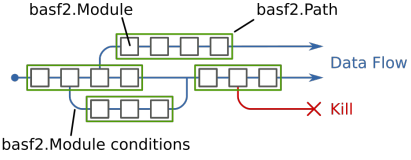


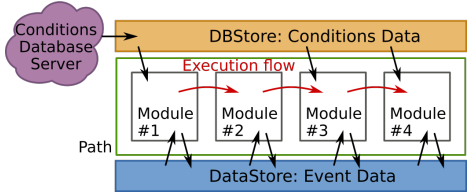
DAQ data flow



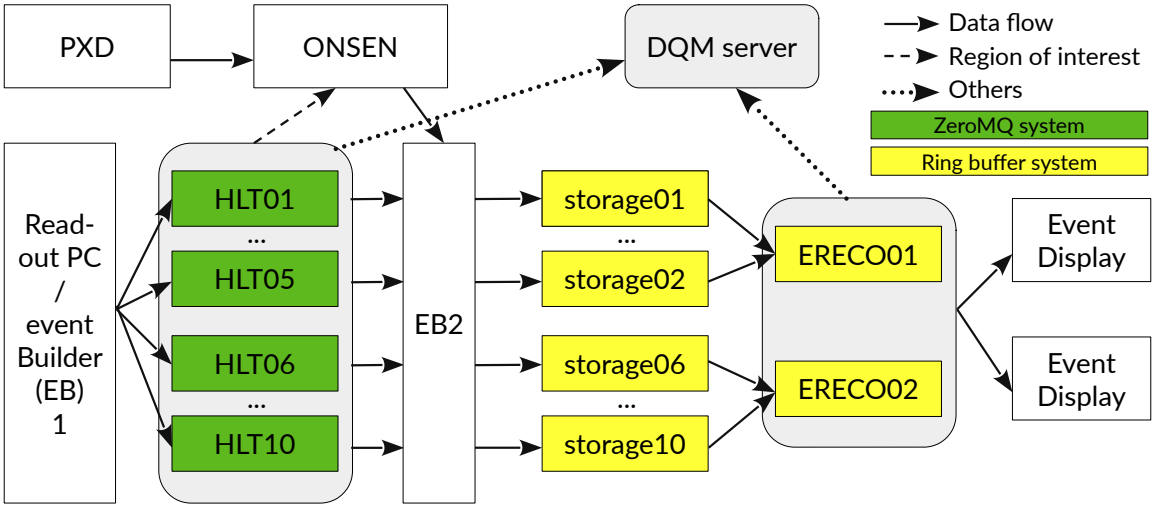
Introduction

- **Items to be shown**
  - Storage: online raw data storage, 32-48 threads CPU with three ~40 TB RAID units × 10
  - ERECO: Express-reconstruction system for online data quality monitoring (DQM), especially for vertex detectors and physics features
    - Till 2022: 2 ERECO consist of input, output (= control), and 8 worker nodes
    - ERECO has ~640 threads CPU → 10 times smaller than HLT (~6400 threads)
- **ZeroMQ library: embeddable networking library to give sockets that carry atomic messages across various transports like in-process, inter-process, TCP, and multicast**
- **BASF2: Belle II Analysis Software Framework**

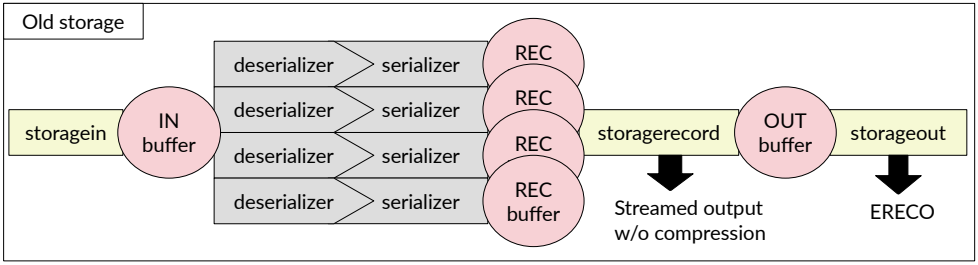




DAQ data flow: HLT to ERECO



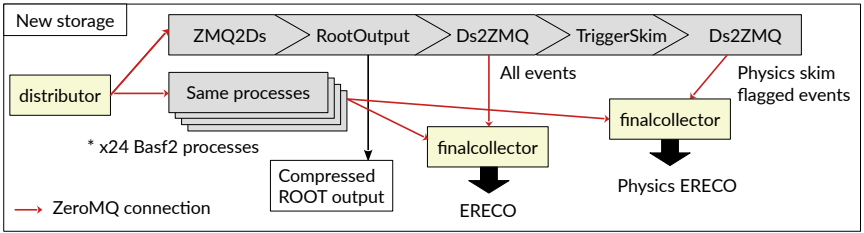
storage: Old



Old storage

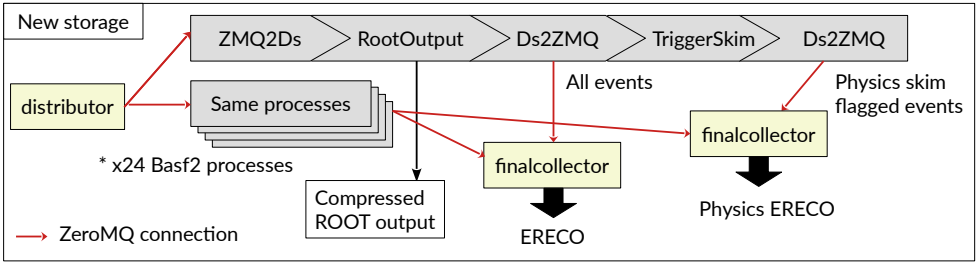
- Ring buffer + socket event distributor w/o HLT skim results
- Streamed output without compression, single output
- Pros: Small CPU usage for recording, no merging for reducing the number of output files, easy file salvage in case of troubles
- Cons: Large file size, additional ROOTization from the offline side

storage: New



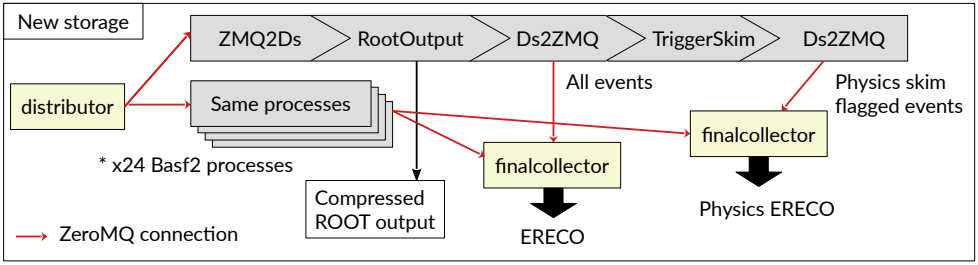
- **New storage**
  - ZeroMQ connections with HLT skim flags
  - Normal ROOT format with compression, multiple outputs
  - Events categorization by the HLT results for ERECO
  - Pros: Small file size, no additional offline processing
  - Cons: Large CPU usage for compression, requiring online side small-sized file merging, additional broken file salvage
- **With higher input rate, the pros of new storage is more important.**

ZeroMQ connections



- **Three types of input/output connections**
  - Load-balanced: 1 (input) / N (output) clients
  - Confirmed: N (input) / 1 (output) clients
  - Raw: 1 (input) / N (output) clients, for non-ZMQ applications

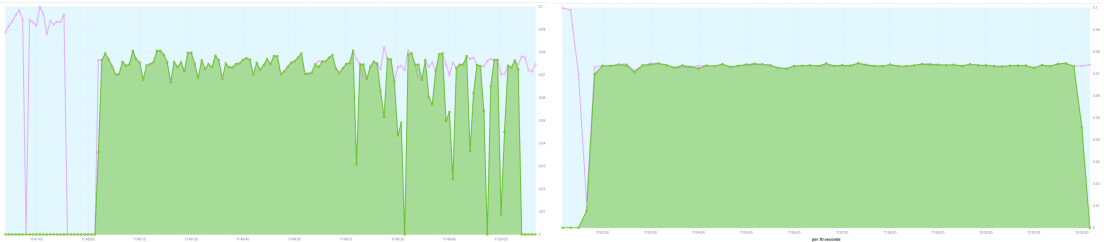
ZeroMQ connections



- **Combinations of the input/output connections create ZMQ applications**
  - distributor: Raw input + load-balanced output
  - finalcollector: Confirmed input + Raw output
  - ZMQ2Ds module: Load-balanced input to BASF2 DataStore
  - Ds2ZMQ module: Confirmed output from BASF2 DataStore

ROOT output: Performance test

- **We measured CPU consumption and disk usage using the real storage server.**
  - Compression algorithm: Zstandard
  - 1-proc can store 150 Hz events without event drop.
  - 24-proc can easily store the maximum rate of events from the Belle II detector.
  - Total disk I/O per second for Poisson trigger 3kHz is 93 MB/s = 327 GB/h.
    - Far away from the limit of the disk I/O

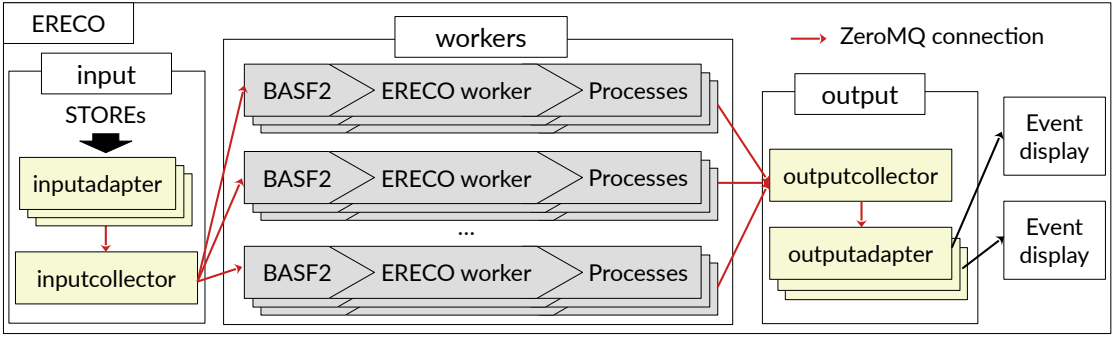


Test results of 200 (left) and 150 (right) Hz input rate. The pink line is the input rate, and the green colored region is the output rate.

ROOT output: After Processing Tool

- **After creating raw ROOT output files, the After Processing Tool performs additional treatment:**
  - Recovering incomplete output files caused by unknown errors
  - Merging small-size files
  - Checksum calculation
  - Making the final file list to be transferred
  - Updating the number of events / output files and "ready to be sent" flag into the run information DB
  - Getting the file transfer status and removing the completed files

ERECO overview

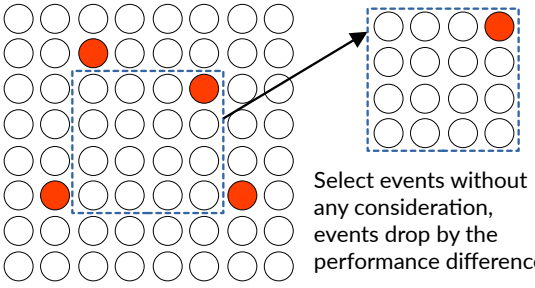


- **Functionality is the same with the current ring buffer + socket ERECO.**
  - However, the new ERECO gives better maintainability and stability.
    - No more shared memory-related issues
  - ERECO allows events to drop, unlike the HLT or storage.
  - From the HLT result based selection, dedicated ERECO for physics is possible.

HLT result based selection for ERECO

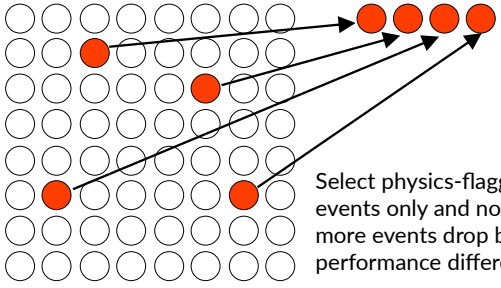
- **# of ERECO is smaller than HLT, therefore only a part of events can be processed.**
- **The less performance ERECO occurs random event selection caused by event drops.**
- **We want more statistics of physics features while keeping the random sampling.**
  - The random sampling is also important, especially for the pixel detector, since the pixel detector information is not in HLT.

< Random sampling >



Select events without any consideration, events drop by the performance difference

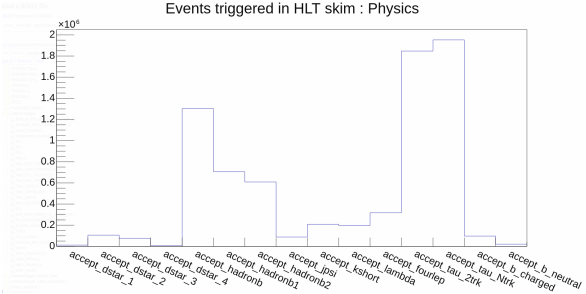
< Physics sampling >



Select physics-flagged events only and no more events drop by the performance difference

HLT result based selection for ERECO

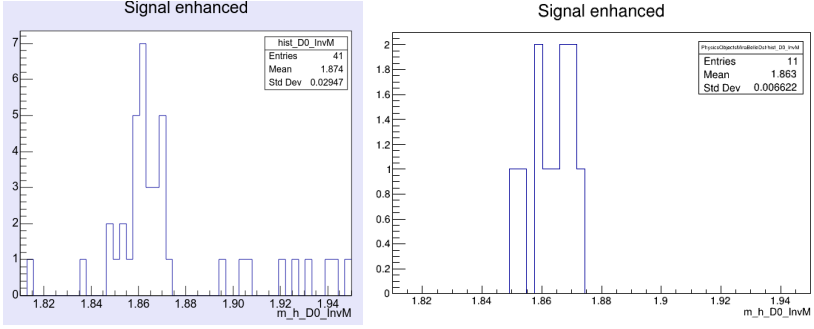
- **# of ERECO is smaller than HLT, therefore only a part of events can be processed.**
- **The less performance ERECO occurs random event selection caused by event drops.**
- **We want more statistics of physics features while keeping the random sampling.**
  - The random sampling is also important, especially for the pixel detector, since the pixel detector information is not in HLT.



The number of events for each physics skims from 4.7M events.

HLT result based selection for ERECO

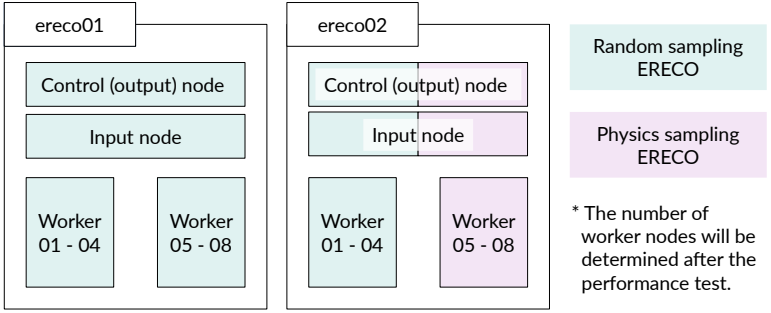
- **Simple ratio calculation for accept_dstar_1 events**
  - Random sampling: 41 D^0 from D^* events with 4.7M inputs → 8.7×10^{-6}
  - HLT result based selection: 11 D^0 from D^* events with 46K inputs → 2.4×10^{-4}
  - Roughly, over 25 times statistics for the physics flagged events



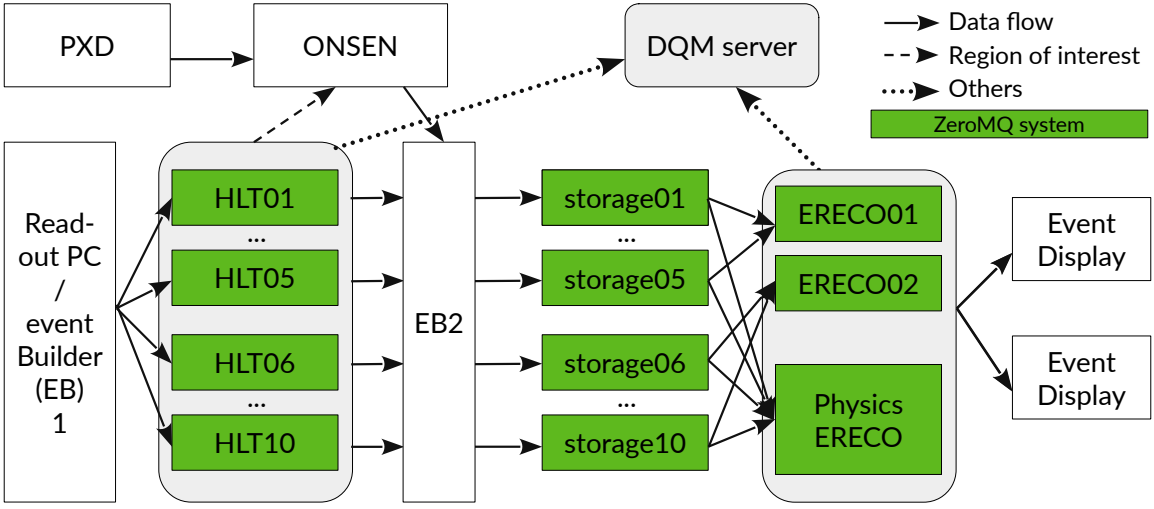
D^0 from D^* invariant mass histogram of 4.7M random sampling data (left) and 46K HLT result based sampling data (right).

Physics ERECO

- **The physics ERECO and one of normal ERECO share the same farm.**
  - Both ERECO share input and output (control) nodes.
  - A few worker nodes are dedicated to physics ERECO.
  - The number of physics ERECO worker nodes will be decided by the performance test and physics trigger menu.



Conclusion

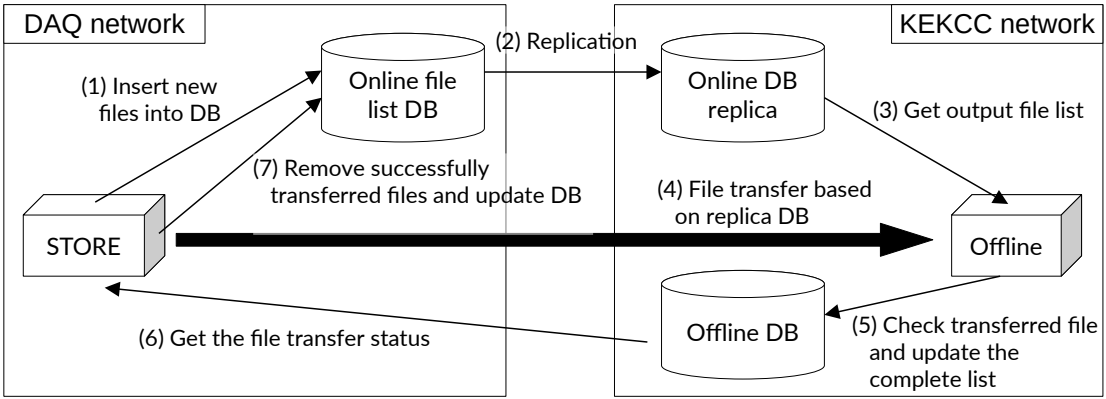

PXD
ONSEN
DQM server
Data flow
Region of interest
Others
ZeroMQ system
Read-out PC / event Builder (EB) 1
HLT01
...
HLT05
HLT06
...
HLT10
EB2
storage01
...
storage05
storage06
...
storage10
ERECO01
ERECO02
Physics ERECO
Event Display
Event Display

Conclusion

- **Belle II is a long shutdown period, and this is the chance to upgrade our storage and ERECO.**
- **Storage and ERECO will use the ZeroMQ framework, the same as HLT.**
  - Better maintainability and stability
- **storage will have new features:**
  - Direct ROOT output with compression
  - HLT result based sampling for ERECO
- **Dedicated physics ERECO will be used for more statistics of physics events from online data quality monitoring.**

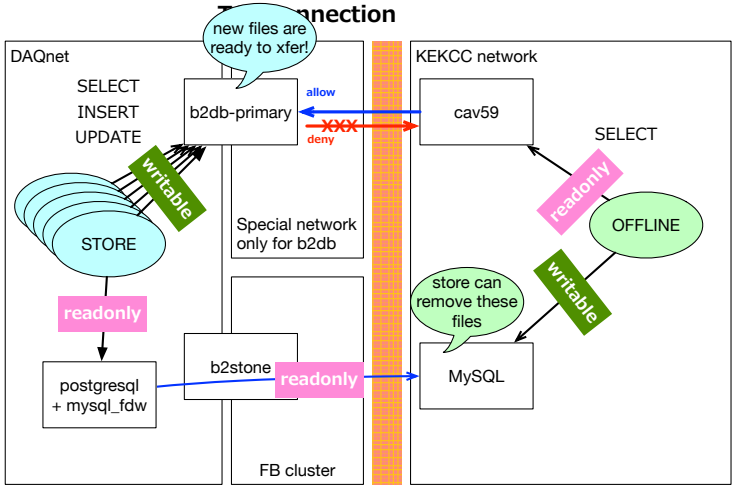
Backup

ROOT output: File list sharing b/w online and offline



- **DAQ network and KEKCC network are basically disconnected.**
  - Only the special connection is allowed for security reasons.
- **The overall design is still under discussion, including the online file list DB contents.**

File list sharing: Detail



accept_dstar_1 trigger rate