Upgrade of Online Storage and Express-Reconstruction system for the Belle II experiment

Seokhee Park *et al.*

seokhee.park@kek.jp

KEK

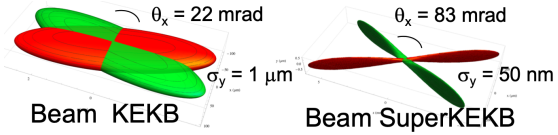
on behalf of the Belle II DAQ group

2023 May 11th

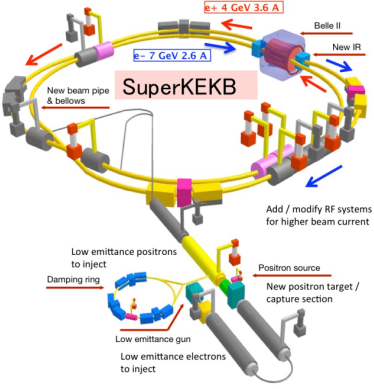
26th International Conference on Computing
in High Energy & Nuclear Physics

SuperKEKB

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



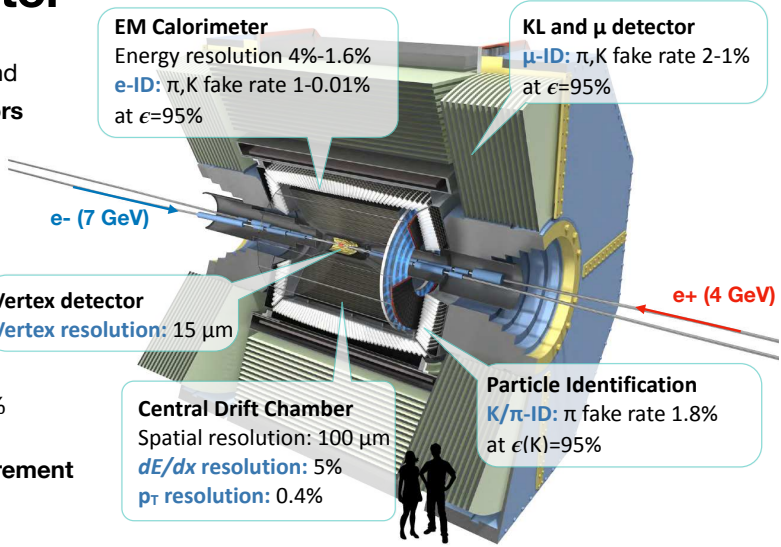
$$L = \frac{N_+ N_- n_\mathrm{b} f_0}{4\pi\sigma^*_{\mathrm{x,eff}}\sqrt{\varepsilon_\mathrm{y}\beta^*_\mathrm{y}}}$$



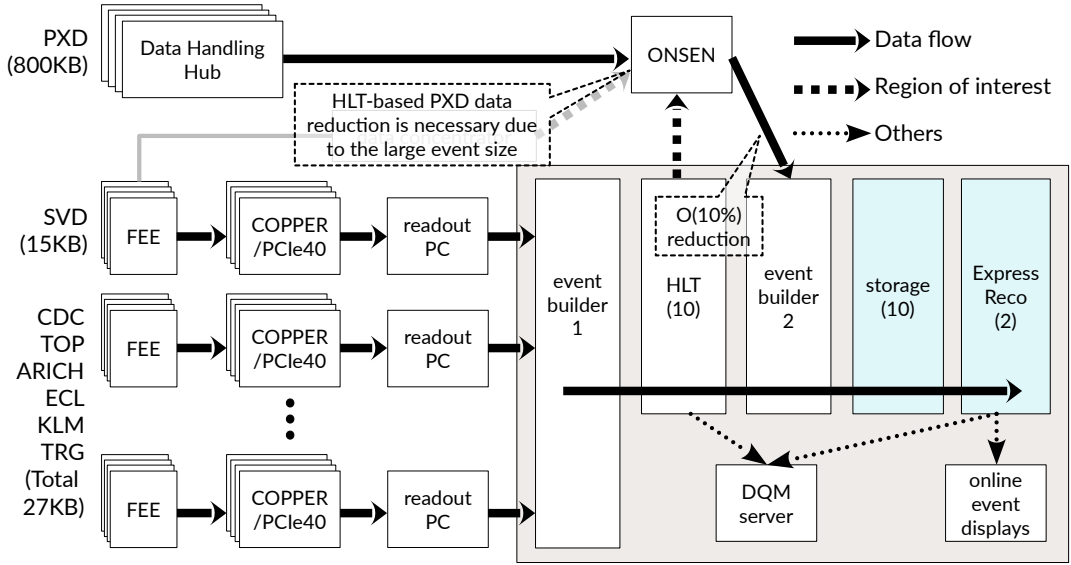
Belle II TDR, arXiv:1011.0352

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**

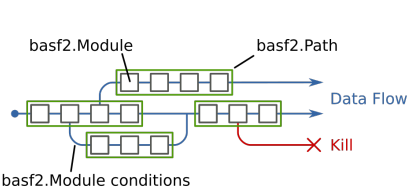


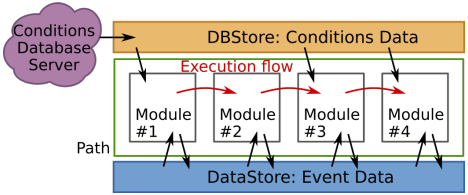
DAQ data flow



Introduction

- **Items to be shown**
  - Storage: online raw data storage, 32-48 thread 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 thread CPU → 10 times smaller than HLT (~6400 threads)
- **BASF2: Belle II Analysis Software Framework**





Why do we need to upgrade the DAQ backend?

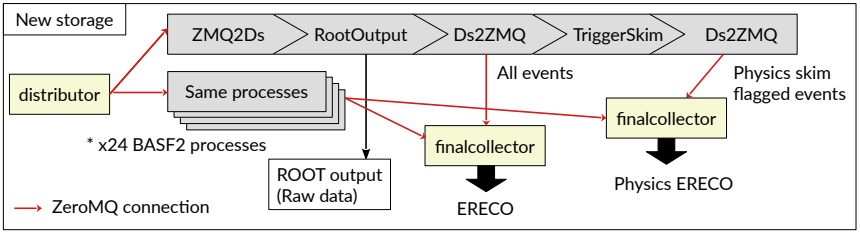
- **To unify design of DAQ backend components for maintainability**
  - HLT uses ZeroMQ (new), storage and ERECO use ring buffer for event distribution.

  → Implement ZeroMQ for storage and ERECO
- **To reduce bandwidth usage for file transfer and offline computing resources**
  - Online storage stores the generated raw data in internally developed format without compression, and offline and grid side store in ROOT format.

  → Directly store raw data in ROOT format
- **To increase statistics of data quality monitoring for physics-tagged events**
  - # of ERECO is smaller than HLT, therefore only a fraction of events can be processed.
  - Events are randomly dropped without any condition due to lower performanc of ERECO.
  - We want more statistics of physics features while keeping the random sampling.

  → Add selection mechanism in STORE and dedicated physics ERECO

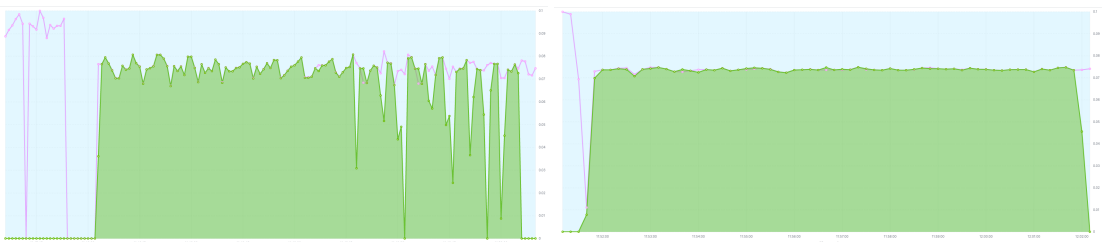
The new online storage



- **New storage**
  - ZeroMQ connections with HLT skim flags
  - Standard ROOT format with compression, multiple outputs
  - Events categorization by the HLT results for ERECO
  - Pros: Small file size, no additional offline processing, all available disks to be usable
  - 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.**

ROOT output: Performance test

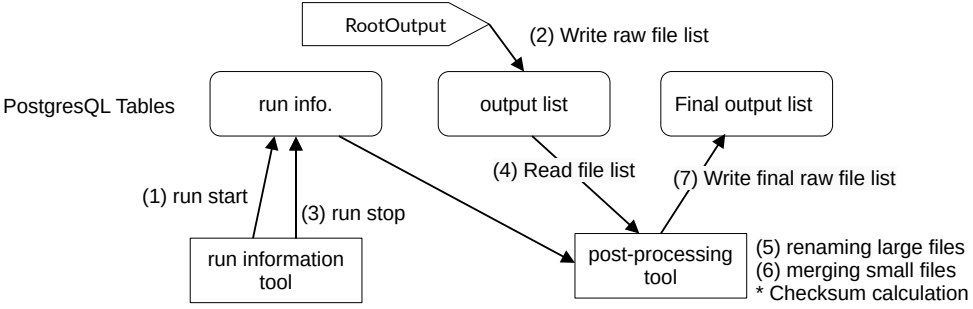
- **We measured CPU consumption and disk usage using the real storage server.**
  - Compression algorithm: Zstandard
  - 1-process can store 150 Hz events without event drop.
  - 24-process can easily store the maximum rate of events from the Belle II detector.
  - Total disk I/O per second for dummy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



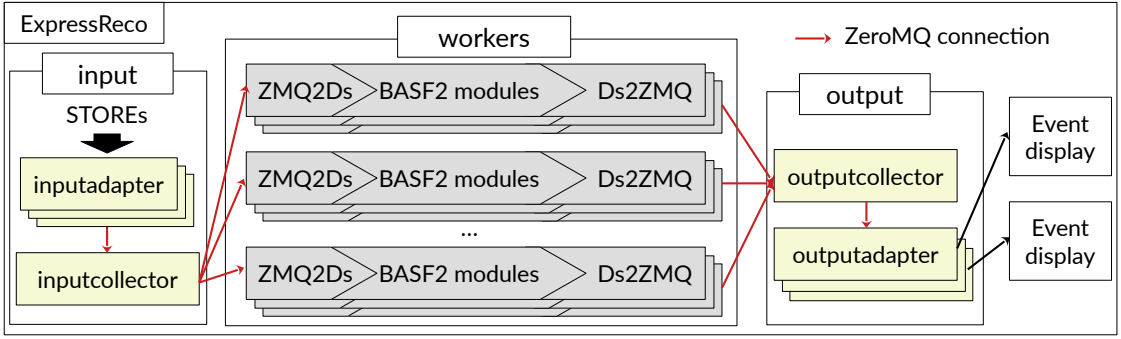
Test results of 200 (left) and 150 (right) Hz input rate. The pink line is the input rate, and the green colored region is the output rate.

Post-Processing Tool

- **After creating raw files, the post-processing Tool performs additional treatment:**
  - Merging small-size files / renaming large-size files for consistency
  - Checksum calculation
  - Making the final file list to be transferred
  - Getting the file transfer status and removing the completed files



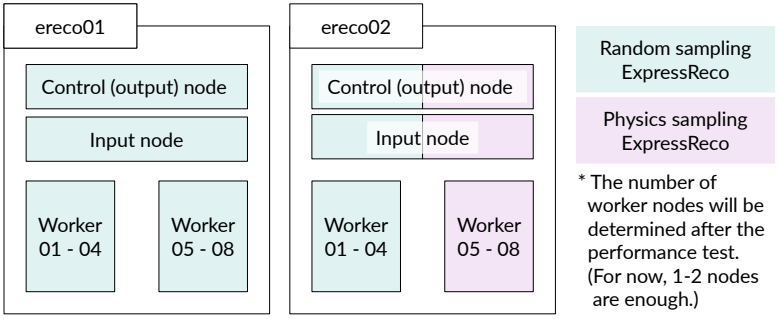
ERECO overview (Express-reconstruction system for DQM)



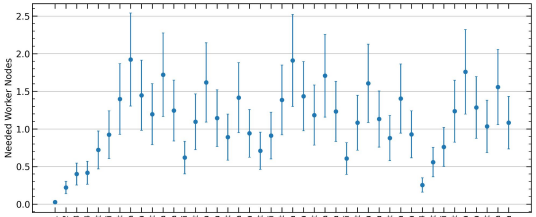
- **Functionality is the same with the current ring buffer + socket ERECO.**
  - However, the new ERECO gives better maintainability and stability.
    - No shared memory-related issues, low delay of data quality monitoring histogram update
  - ERECO allows events to drop, unlike the HLT or storage.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

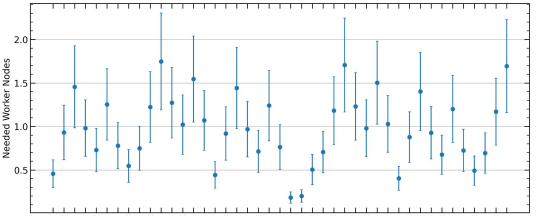
Physics ERECO

- **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.

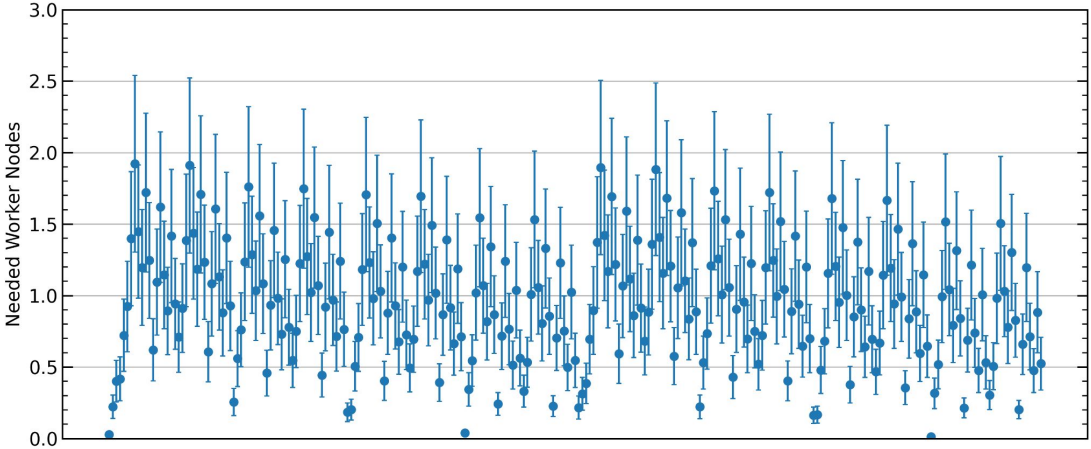


The number of workers for Physics ERECO



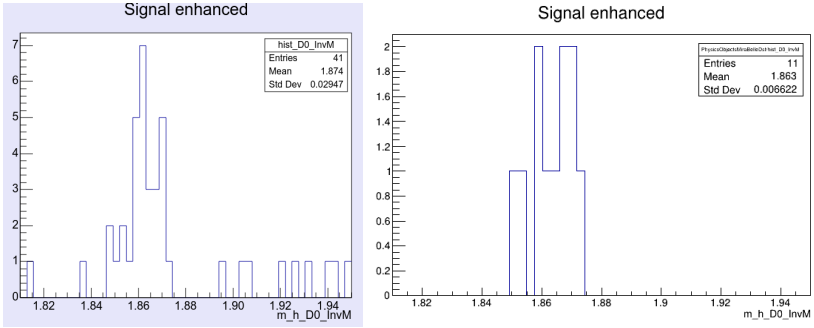


The number of workers for Physics ERECO



Impact of physics ERECO

- **Simple ratio calculation for one of HLT physics flags**
  - 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



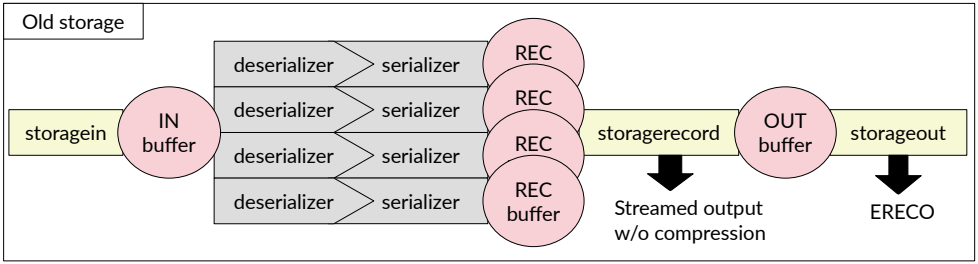
D^0 from D^* invariant mass histogram of 4.7M random sampling data (left) and 46K HLT result based sampling data (right).

Conclusion

- **Belle II is in a long shutdown period, and we are testing the new system.**
- **Storage and ERECO will use the ZeroMQ framework, the same as HLT.**
  - Better maintainability and stability
  - Resolve ring buffer related issues, like delayed data quality monitring histograms.
- **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.**
- **File transfer mechanism also will be updated based on DB table file listing and xrootd.**

Backup

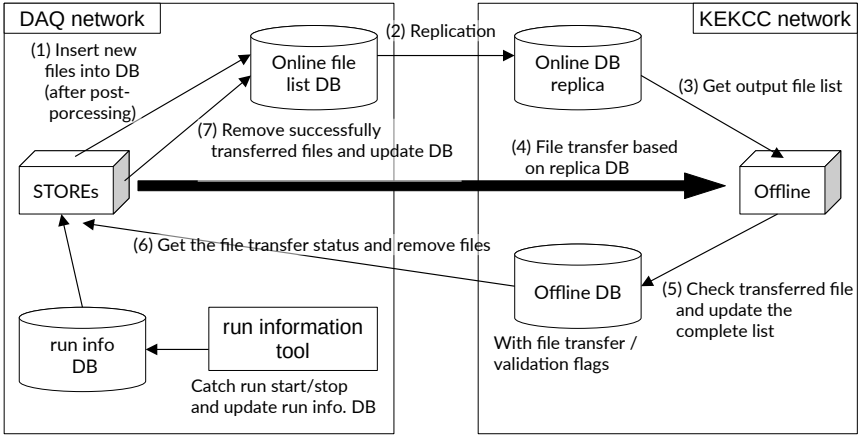
storage: Old



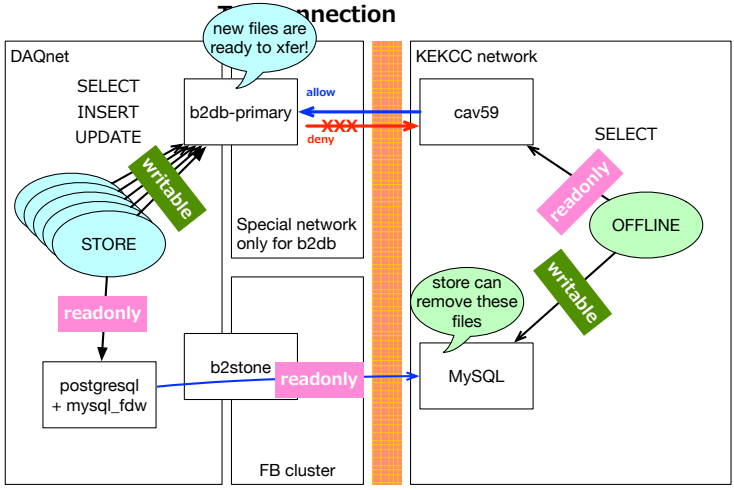
- **Old storage**
  - Ring buffer + socket event distributor w/o HLT skim results
  - Streamed output without compression, single output
  - Pros: Small CPU usage for recording, no merging for reducing the number of output files, easy file salvage in case of troubles
  - Cons: Large file size, additional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

File transfer

- **We plan to update file transfer mechanism completly for ROOT raw output.**
  - File listing: text-file based → database table based
  - File transfer: rsync → xrootd



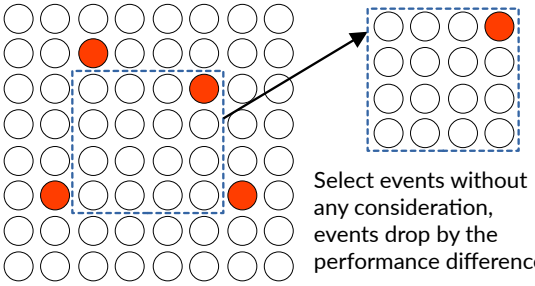
File list sharing: Detail



HLT result based selection for ERECO

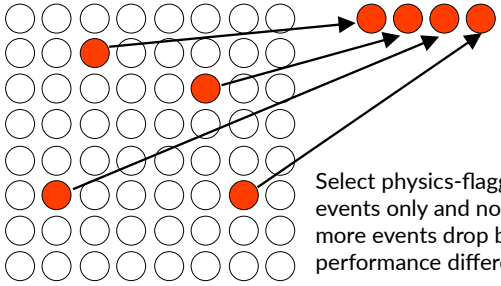
- **# of ERECO is smaller than HLT, therefore only a part of events can be processed.**
- **The less performance ERECO occurs random event selection caused by event drops.**
- **We want more statistics of physics features while keeping the random sampling.**
  - The random sampling is also important, 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