Commissioning and Early Experience of the New Online Storage and Express-Reconstruction System for the Belle II Experiment

Seokhee Park *et al.*

seokhee.park@kek.jp

KEK

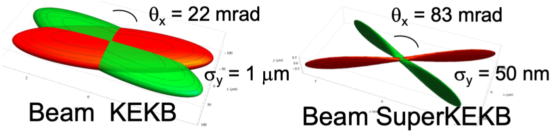
on behalf of the Belle II DAQ group

2024 April 25th, Vietnam Quy Nhon

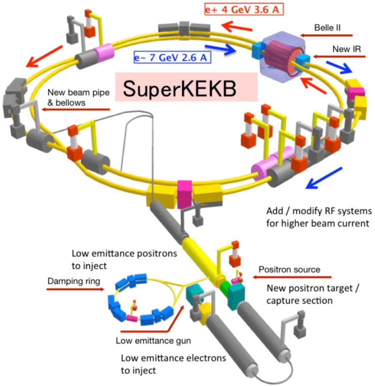
25th IEEE ~~Virtual~~ Real Time Conference

SuperKEKB

- **Electron-positron collider with $7\ \text{GeV}\ e^-$ and $4\ \text{GeV}\ e^+$**
  - Focused on $\Upsilon(nS)$, mainly $\Upsilon(4S)$
- **Aiming at $50\ \text{ab}^{-1}$ of data ($= 50\times$ Belle) → Achieved $456\ \text{fb}^{-1}$**
- **Aiming at $6.5 \times 10^{35}\text{cm}^{-2}\text{s}^{-1}$ of peak lumi ($= 30\times$ KEKB) → Achieved $4.71 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$**
  - corresponding to 30 kHz L1 trigger rate
  - 1/20 of beam size (nanobeam scheme)
  - 150% of beam current



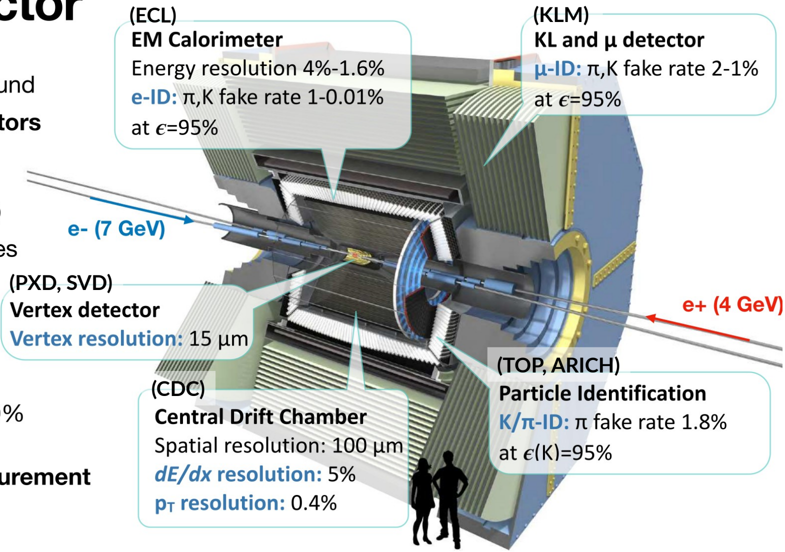
$$L = \frac{N_+ N_- n_b f_0}{4\pi \sigma^*_{x,\text{eff}} \sqrt{\varepsilon_y \beta^*_y}}$$



Belle II TDR, arXiv:1011.0352

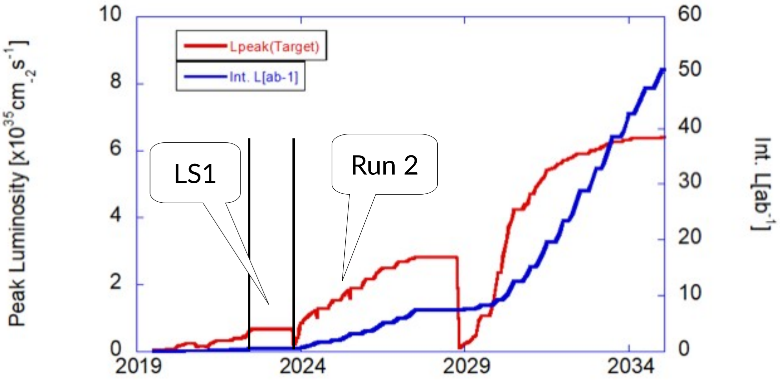
Belle II detector

- 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**

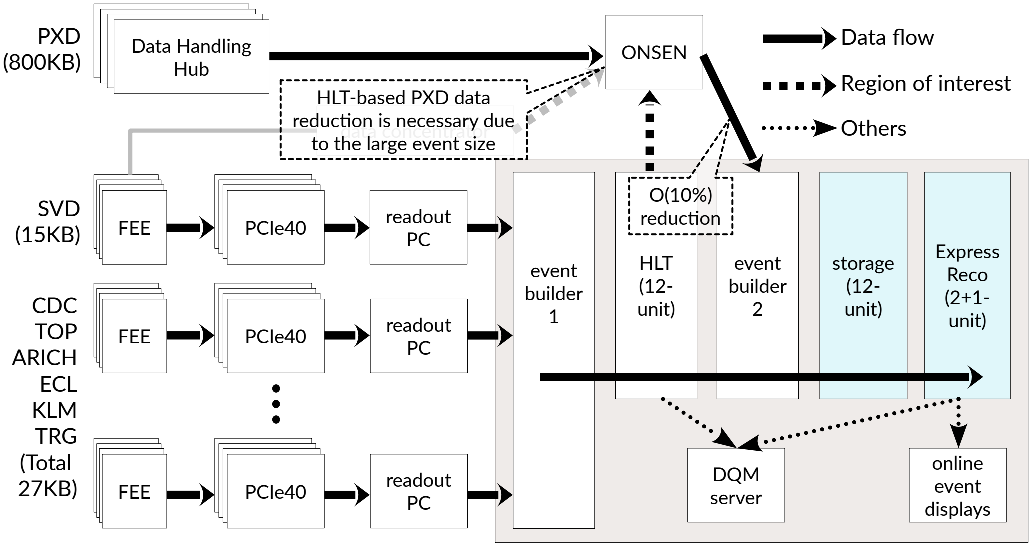


Run 2 operation

- **Run 1 (-2022) → Long shutdown 1 (2022-2024) → Run 2 (Feb. 2024-)**
  - LS1: PXD, TOP PMT, DAQ readout, and **Online storage / Express-Reconstruction upgrade**



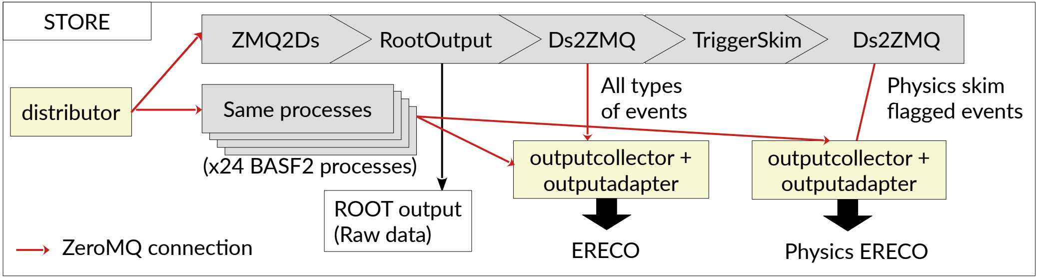
DAQ data flow



Introduction

- **Motivations of the upgrade**
  - Unify the framework across the HLT, online storage (STORE), and express-reconstruction (ERECO) for better maintainability and stability
  - Record output files to ROOT format to reduce the file transfer bandwidth and offline computing resource usage
  - Provide ERECO only for physics-tagged events for higher statistics of the monitoring
- **Hardwares**
  - STORE ($\times 12$): 32-48 threads CPU with three ~40 TB RAID6 units (HDD)
  - ERECO: Express-reconstruction system for online data quality monitoring (DQM), especially for vertex detectors and physics features
    - Two types of ERECO: random sampling (normal) and physics sampling (physics)
    - Normal ERECO ($\times 2$): input, output (= control), and 8 worker nodes (~160-core per unit)
    - Physics ERECO ($\times 1$): (input, output: normal ERECO shared,) 2 worker nodes (96-core in total)

Key updates: Online storage



- **Data distribution using the ring buffer + TCP/IP socket → ZeroMQ connections**
- **Single SROOT (home-made format) → Standard ROOT format with compression**
  - Multiprocessing to achieve the online compression and multiple output files at the same time
- **(New) 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**

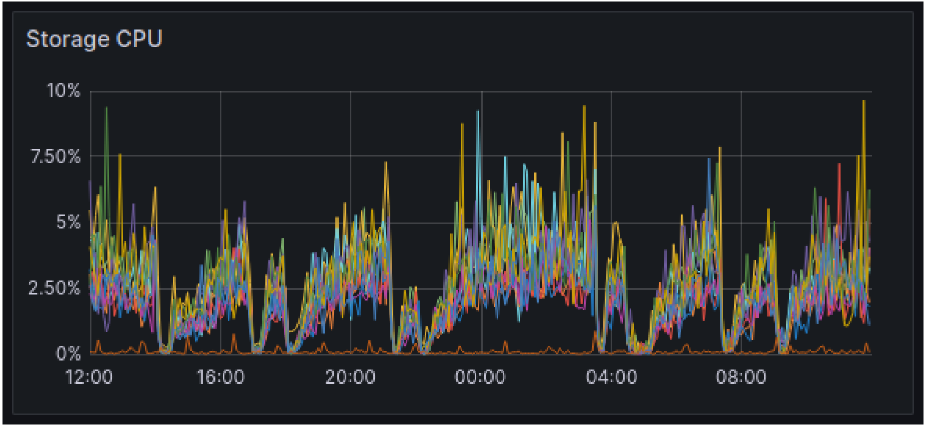
Write cache disks

- **During the test, we faced some troubles on creating and closing ROOT files.**
  - It's because 8 files are trying to be created at the same time in an HDD array.
- **To solve the issue, write cache disks are installed.**
  - 2TB SATA SSD per RAID disk → 6TB buffer space
  - Once a file is correctly closed, the file immediately moved to the corresponding RAID disk.
    - Since the buffer disk is large enough, we can use it as temporary space in case of RAID disk issue.
  - The buffer space also prevents performance degradation of output writing which can be caused by reading files simultaneously
- **All the SSDs are hot swappable and monitored by zabbix smartmon.**

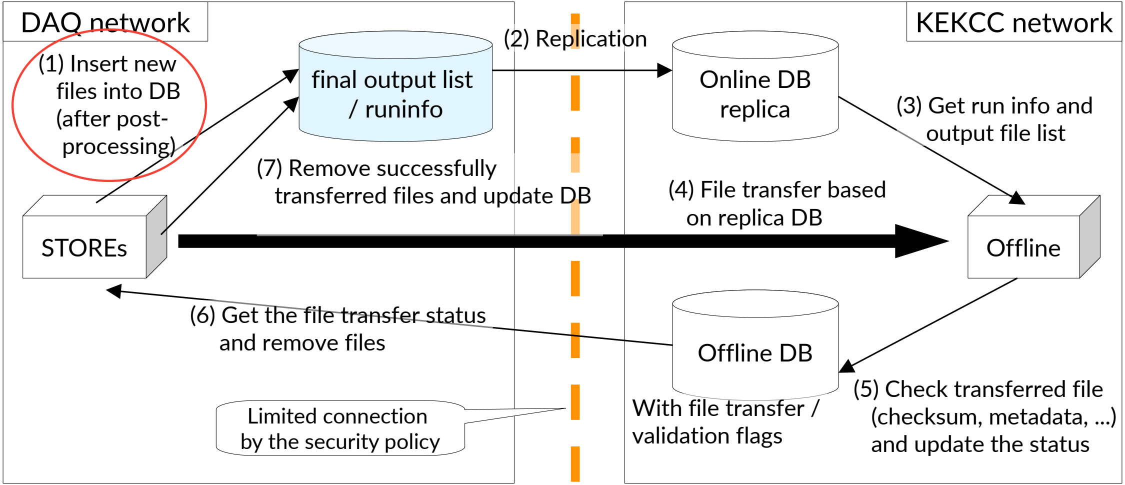
```
/dev/sdh1        1.9T  1.1G  1.9T   1% /buffer/rawdata/disk03
/dev/sdf1        1.9T  1.1G  1.9T   1% /buffer/rawdata/disk01
/dev/sdg1        1.9T  1.1G  1.9T   1% /buffer/rawdata/disk02
/dev/sda1         33T  2.5T   31T   8% /rawdata/disk02
/dev/sdb1         33T  2.7T   31T   9% /rawdata/disk01
/dev/sdc1         33T  2.6T   31T   8% /rawdata/disk03
```

CPU usage in the operation

- **We are now early phase of the run 2 operation, so the input rate is not so high.**
  - Roughly, 15-20% of maximum design
  - Even though, the CPU usage is very acceptable level.

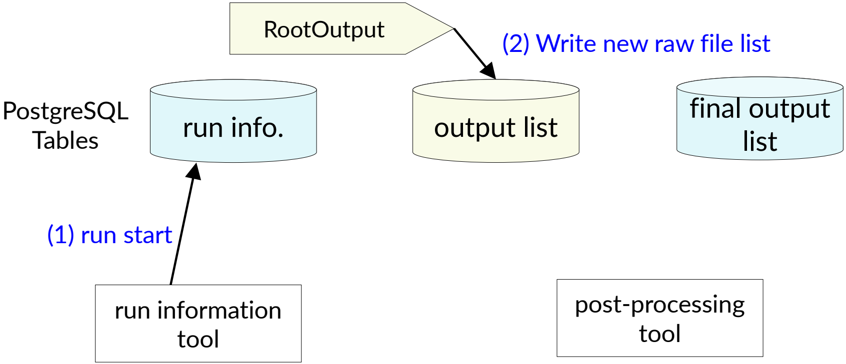


File transfer to offline computing site



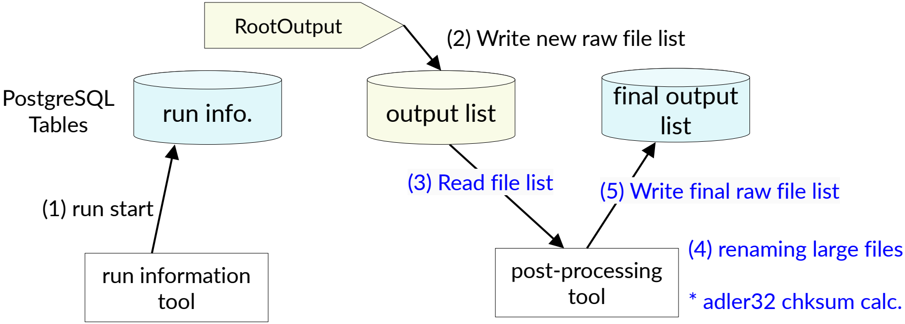
- **In run 1, the file transfer is performed based on the text file, so we should improve this.**

Post-processing and database-based file listing



- **Beginning of the run**
  1. run info table: New run is recorded
  2. output list table: New files are listed by RootOutput modules

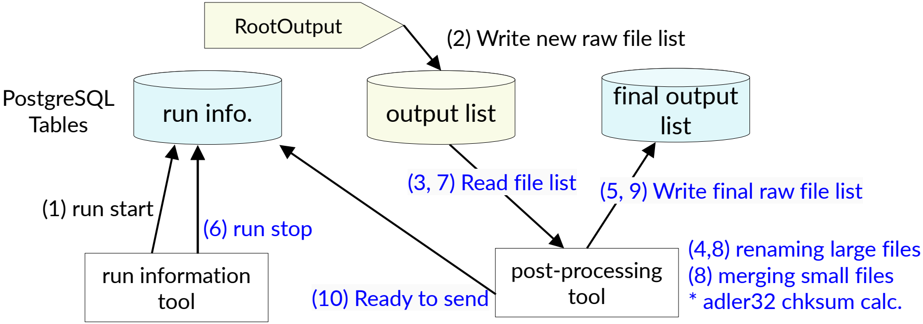
Post-processing and database-based file listing



■ Middle of the run

1. output list table: once file is reached at the size limit, close the files and update "closed" flag
2. Once file closing is confirmed, move files from buffer disk to RAID disk
3. final output list table: rename, calculate checksum, and update the entry

Post-processing and database-based file listing

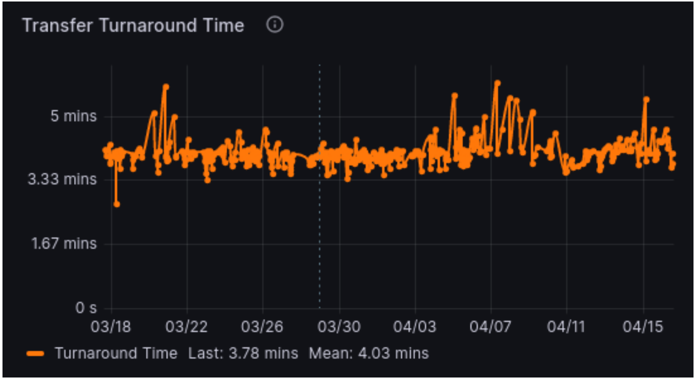


■ After the run end

1. run info table: flag the run end
2. output list table: close the files and update a "closed" flag
3. final output list table: rename or merge files, calculate checksum, and update the entry
4. Once everything is ready, set "ready to send" flag

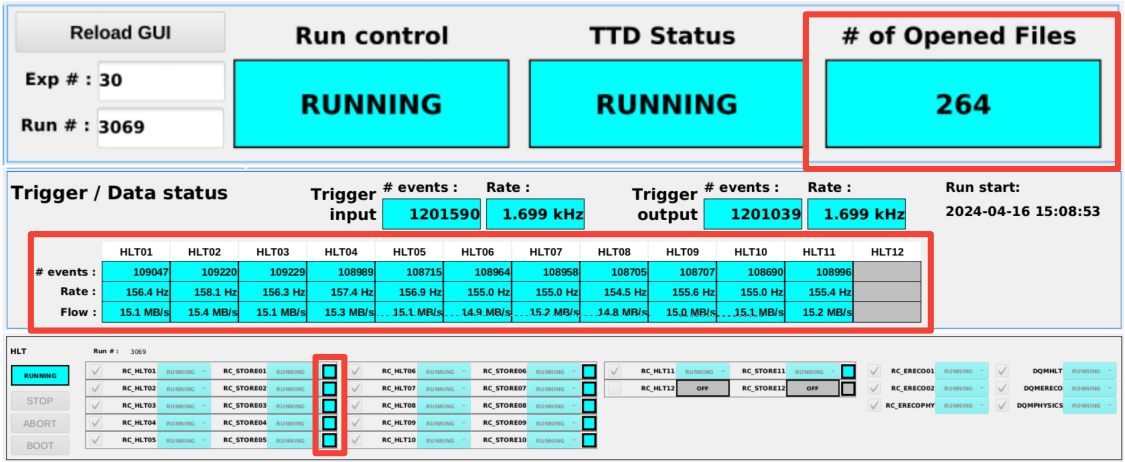
File transfer to offline computing site

- **The file transfers are done almost within 5 minutes**
  - Done by xrootd (run 1: rsync)
  - Much faster than the previous text-based file listing & transfer
  - No additional format conversion and compression is needed from the offline computing site

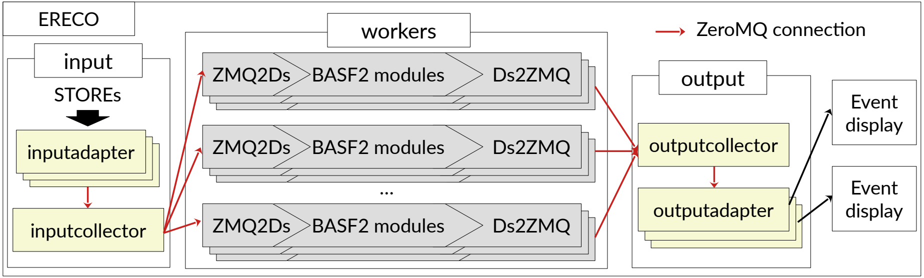


Monitoring

- **Run control GUI provides useful information and CR shift can check the STORE status**
  - The color becomes red or orange if the state is wrong



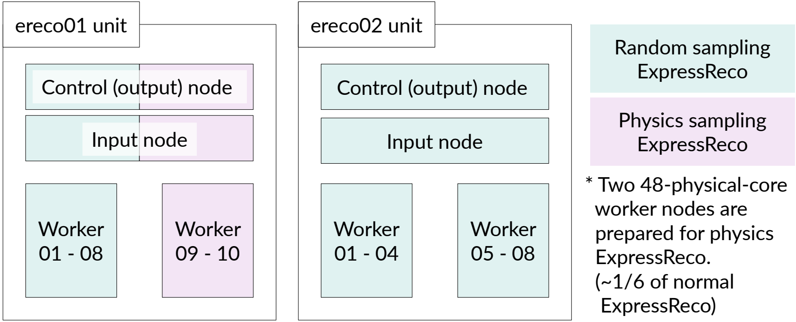
Key updates: Express-reconstruction system



- **Data distribution using the ring buffer + TCP/IP socket → ZeroMQ connections**
- **Better maintainability and stability**
  - The operation is very stable.
  - Some old bugs in the previous system are gone
    - Slow DQM histogram update, run number mixing, silence crash, shard memory issue, ...
- **DQM and online event display for physics-tagged events**

Physics ERECO

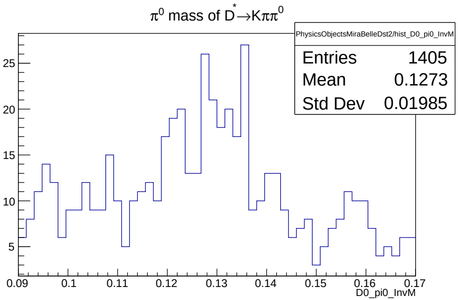
- **The ERECO performance is $\mathcal{O}(10\%)$ of HLT → many events are randomly discarded.**
  - Prepare dedicated ERECO only for physics-tagged event for more statistics of DQM
- **The physics ERECO and one of normal ERECO share the same farm.**
  - Both ERECO share input and output (control) nodes.
  - Two worker nodes (~100-core) are prepared only for physics ERECO.



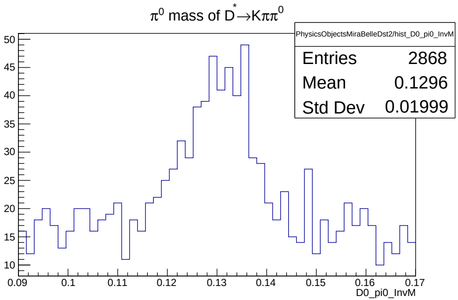
* Two 48-physical-core worker nodes are prepared for physics ExpressReco. (~1/6 of normal ExpressReco)

Physics ERECO DQM

- **The trigger lines for physics ERECO is now studied**
  - In the early phase of run 2, input rate is small, so try to include as many as possible trigger lines
  - The statistical enhancement is depending on the input rate and trigger line selection
- **Both normal and physics ERECO DQM files are stored**
  - Even with the low lumi, # event of physics info in the physics ERECO are double of normal one



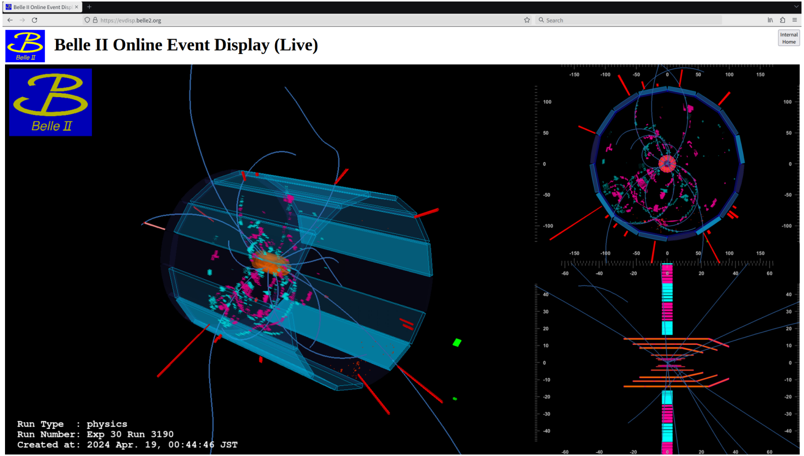
m_{π^0} from the $D^* \to D(K\pi)\pi^0$ in normal ERECO.



m_{π^0} from the $D^* \to D(K\pi)\pi^0$ in physics ERECO.

Online event display

- **Public online event display is now running with the physics ERECO output**
  - We can provide only physics live events (available on **https://evdisp.belle2.org**)



Conclusion

- **During the long shutdown period, we decided to upgrade our systems to**
  - Unify the structure for better maintainability and stability
  - Use standard ROOT format to save bandwidth and offline computing resource usage
  - More statistics of physics objects in the data quality monitoring histograms
- **Belle II is now on the early phase of Run 2, and new online storage and express-reconstruction systems are successfuly running.**
- **Several practical solutions are implemented for the online storage operation.**
  - Performances are in the acceptiable range.
- **The new express-reconstruction system is stably running.**
  - Physics express-reconstruction unit provides DQM histograms and online event display data.

Backup

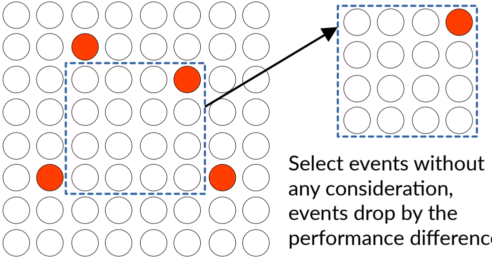
Post-processing and database-based file listing

- **After the new output files are placed in the RAID disks, further processing is necessary.**
  - Renaming large enough files
  - Mreging small files
  - Checksum calculation
  - Making the final file list to be sent
- **For the file listing, three PostgreSQL tables are used.**
  - run info table: recording run information, exp/run number, run type, global flags, ...
  - output list table: file list before the post processing, recorded by the RootOutput module
  - final output list table: file list after the post processing, used for the online-offline file transfer

HLT result based selection for ERECO

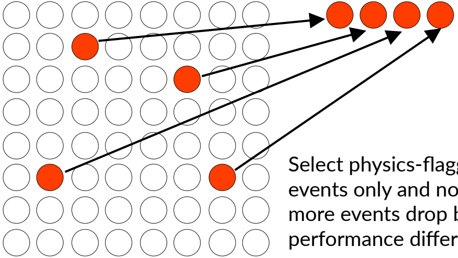
- **# of ERECO is smaller than HLT, therefore only a part of events can be processed.**
- **The less performance ERECO occurs random event selection caused by event drops.**
- **We want more statistics of physics features while keeping the random sampling.**
  - The random sampling is also important,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

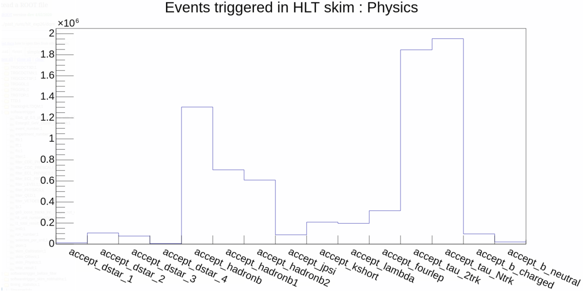
< Physics sampling >



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.

accept_dstar_1 trigger rate