Storage

Seokhee Park

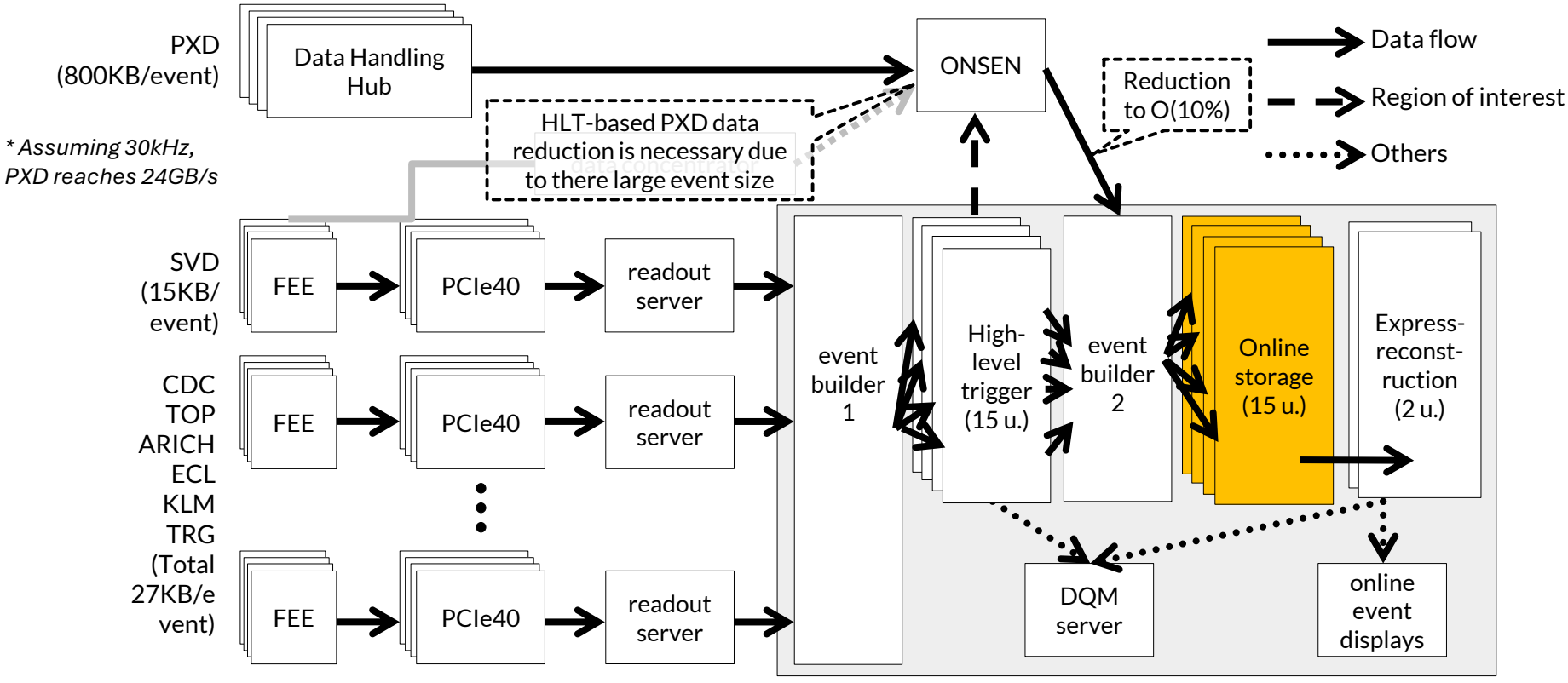
KEK, IPNS

on behalf of the Belle II DAQ group

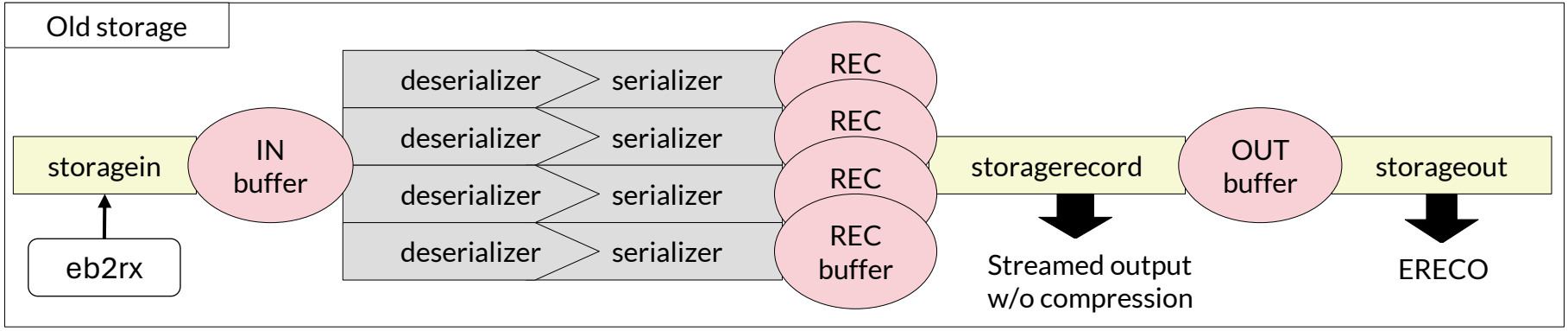
Introduction

- What is online storage (STORE)?
  - A kind of temporary storage, store HLT and EB2 output into the ROOT files and send them to KEKCC.
- The online storage (STORE) is installed in the each HLT rack
  - A single STORE consists of a 3U main server and 3 disk enclosures.
  - Intel dual CPU configuration
  - Each disk enclosure has 12 HDD (4TB). One disk is dedicated as a hot spare; 11 disks construct one RAID-6 volume. → 36TB per volume → ~100TB per STORE
  - To support multiple file writing at once, three 2TB SATA SSDs are installed.
  - Many network connections: hltout, eb2, ereco, QAS, KEKCC, DAQNET, B2NSM

Introduction

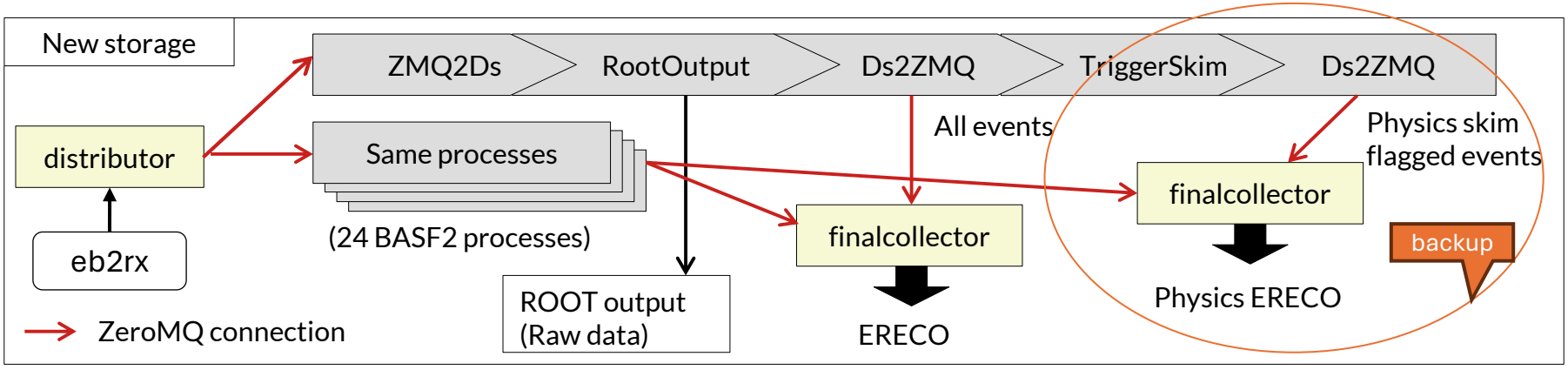


Before direct ROOT recording



- We were using “home-made” SROOT format
  - Streamed output without compression, very simple structure.
    → Small CPU consumption, large file size
  - But compression and format conversion are necessary from offline center.
  - Single disk is used (only one output file) and rotating disks
- Event distribution via ring buffer (rfarm framework)
  - HLT was already move to ZMQ.
  - Occasional SHM-related troubles.

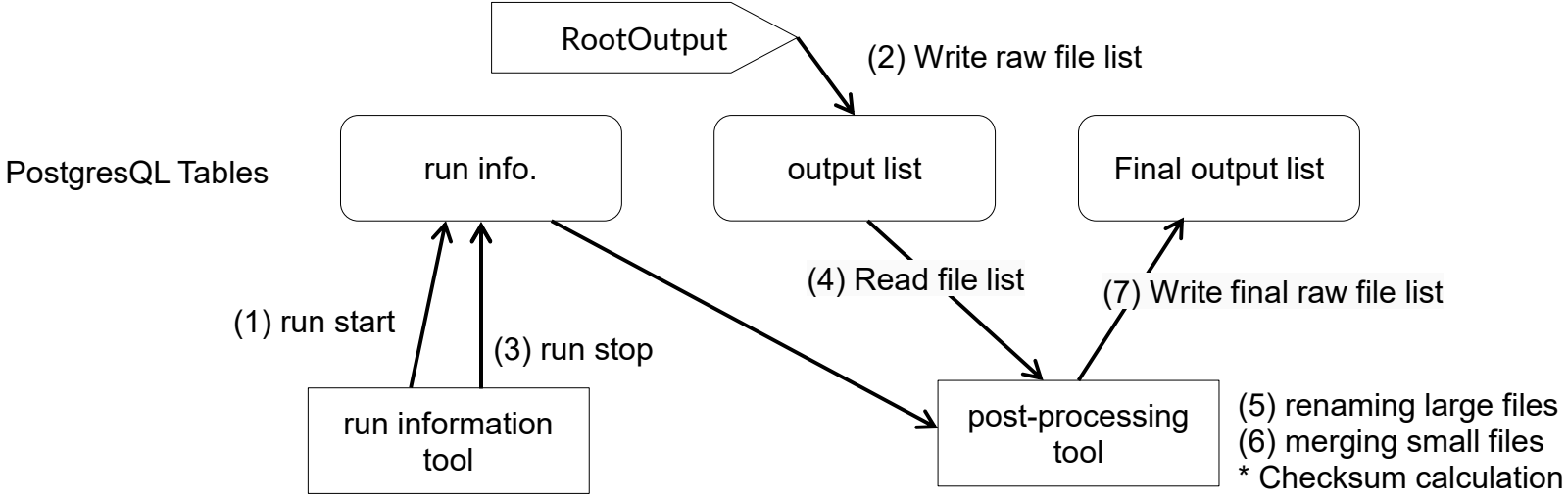
Upgrade to direct ROOT recording (LS1)



- Record "ROOT" directly with the compression (Zstd, level = 5)
  - Compression consumes high CPU power → multiple CPU output in one time.
  - At the time, total $N_{\mathrm{proc.}} = 24$ (limit from STORE01).
  - No format conversion is needed from KEKCC. Only book-keeping is enough.
- Event distribution via ZMQ (hbsaf2), same framework with HLT
- At the last of the run, many small files can be created
  - Need additional merging before sending file to KEKCC.

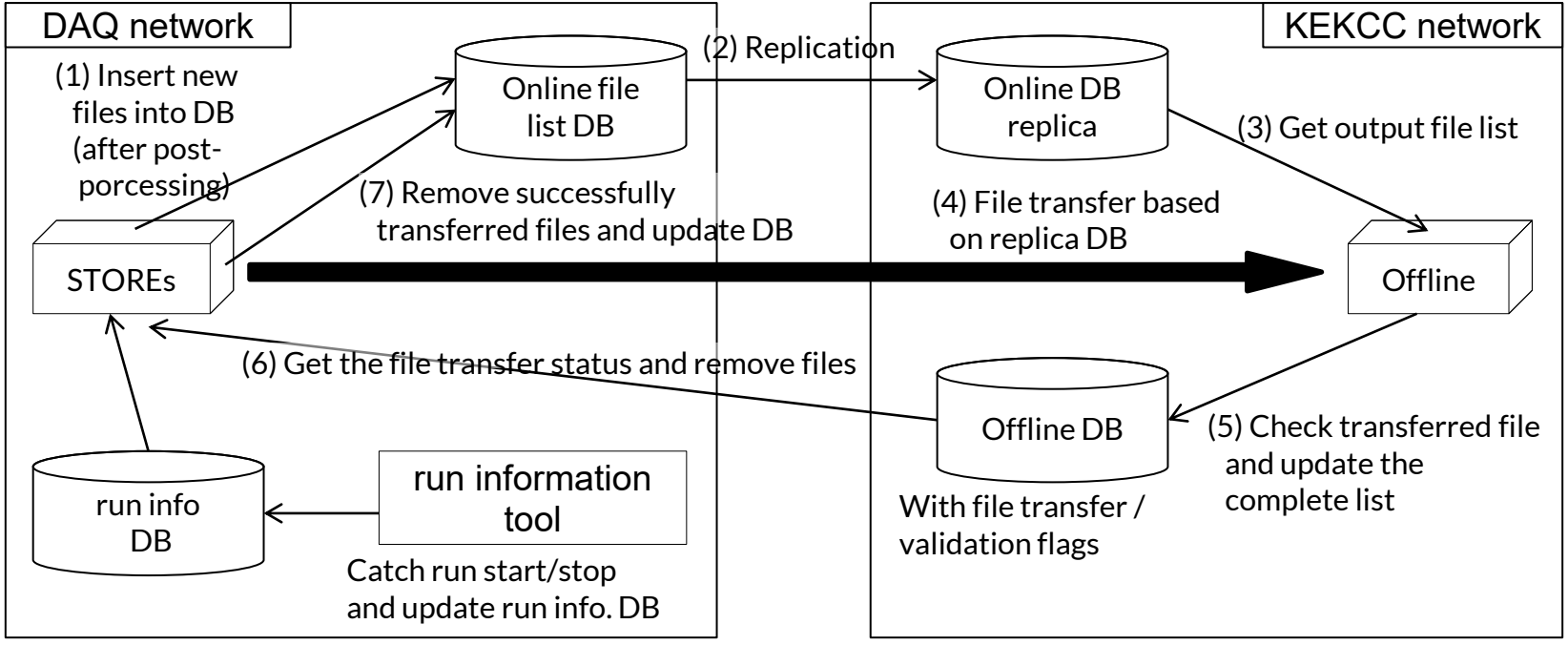
Upgrade to direct ROOT recording (LS1)

- After creating raw files, the "post-processing tool" works:
  - 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



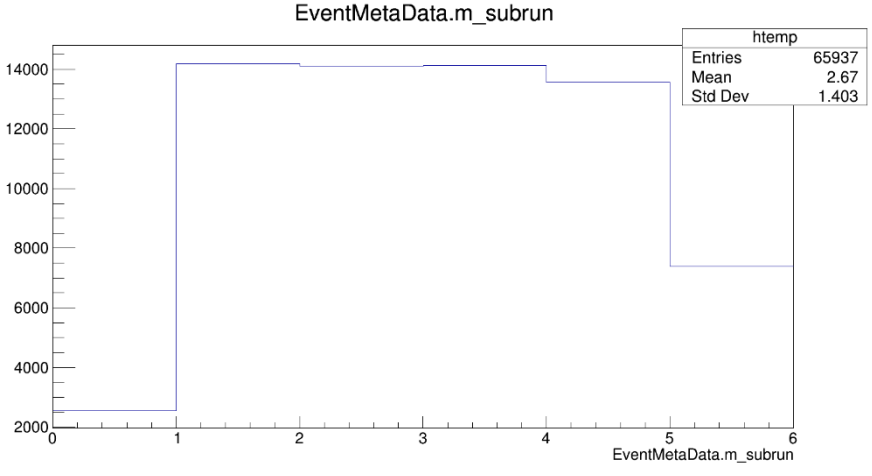
Upgrade to direct ROOT recording (LS1)

- I haven't mention, but file book-keeping method has been updated.
- With SROOT STORE, the file list is managed by a text file, "list_send"
- New file transfer handshake with databases and FileMetaData



Operation status

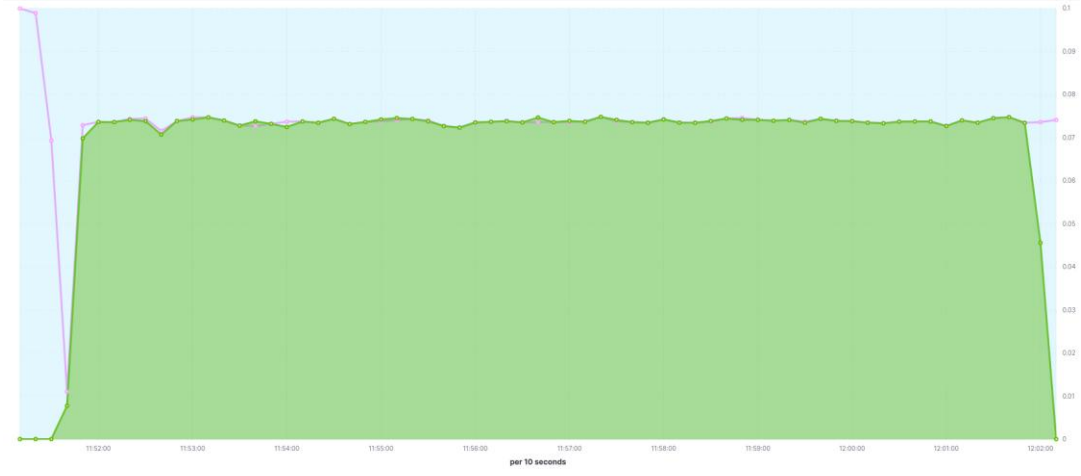
- Very stable operation after fixing two major issues in early stage:
  - Run stop without file closing → fixed by file closing at stopping and signal handling
    - Matt prepared a tool to recover the files which are not correctly closed, no FileMetaData.
  - Continuous file overwriting with old run's event → adding run number checker
- After that, no updates since it just works.
- Sub-run number is well recorded in EventMetaData without any update.
- One concern: the post-processing tool (python) code is very dirty; we should clean up at some point.
- New STORE01 and 15 setup is done by Ansible easily (almost…).



- OS "minor" version is different server by server (Rocky 9.2 – 9.6…)

Sustainability of STORE

- Early performance test using STORE01 (retired, slowest STORE) ensure more than ~100 MB/s output rate (after compression)
  - Compression ratio is ~ 50%, so if the event size is 100kB, 2kHz data can be handled without HLT filtering.
  - Cannot touch SATA3 bandwidth.
  - We have also an option to lower the compression level to reduce the CPU load.
- RAM consumptions are always small, not a blocker.



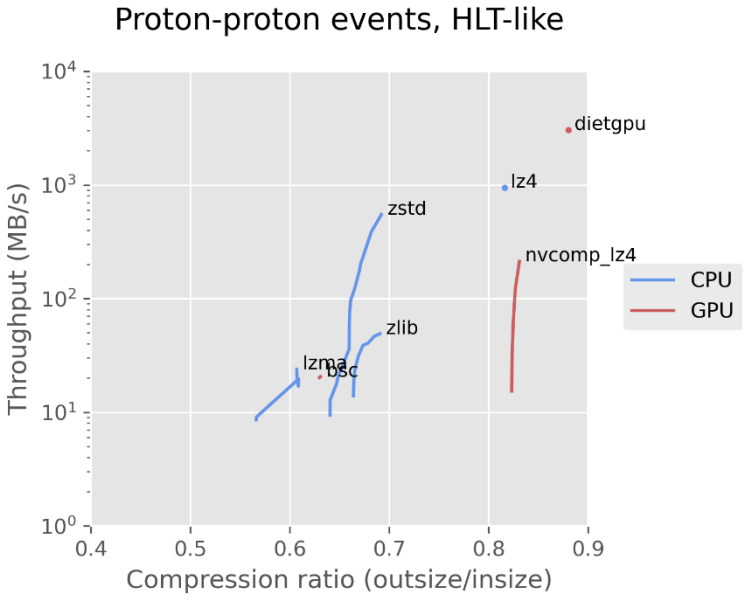
Sustainability of STORE

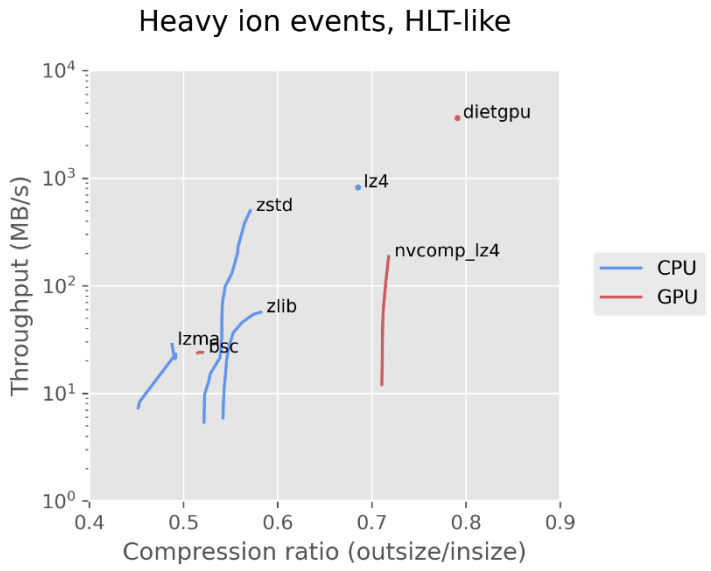
- First point: performance
  - After introducing VTX, what will happen in the data size?
  - (With VTX, we may not need EB2, different discussion point.)
- Currently, 5 types of CPU is used (always Xeon dual-CPU):
  - E5-2650v3 (STORE02-03), E5-2650v4 (STORE04-07), Silver 4116 (STORE08-10, 13), Silver 4214R (STORE11-12), and Silver 4310 (STORE01, 14-15)
  - Now there is no blocker, we can increase the process number to 36 = 50% more
- From the SPECrate(2017), the CPU scores are similar except Silver 4310.
  - If we want to upgrade effectively, change all the server at once.
- We should also consider the data size from STORE.

host	store01	store02	store04	store08	store11
CPU	Silver 4310	E5-2650 v3	E5-2650 v4	Silver 4116	Silver 4214R
# of physical core	24	20	24	24	24
# of logical thread	48	40	48	48	48
Int SPECrate(2017)	174	95	111	110	129
SPECrate/core	7.25	4.75	4.63	4.58	5.38

Sustainability of STORE

- By seeing old study (2022) from CERN, GPU algorithms consumes similar level of throughput, but worse compression ratio.
  - https://stefanrua.github.io/gpucomp/acat2022.html
  - Sorry for missing the latest improvement.





Sustainability of STORE

- Second point: ARECA RAID controller
- Our STOREs fully rely on ARECA RAID controller. All of the STOREs using ARC-1883LP except STORE12 (using ARC-1884IXL)
- While upgrading RHEL7 to RHEL9, ARECA driver support has been dropped from RHEL official repository.
  - While upgrading RHEL9, we installed DKMS module manually, provided by ARECA.
- The support is not continued (or very slow). The latest build is RHEL 9.4. (The latest version is RHEL 9.6.)

OpenSUSE 11	1.20.0X.15	2010/08/02
RHEL 9.4	1.51.0X.16	2023/12/26
This driver also supports Rocky Linux 9.4/AlmaLinux 9.4.		
RHEL 9.3	1.51.0X.15	2023/10/12
This driver also supports Rocky Linux 9.3/AlmaLinux 9.3.		

Sustainability of STORE

- Already the driver source code cannot be attached by default.
- While installing STORE01 and STORE15, I have used Rocky 9.6 and the DKMS build was failed.
- In this time, I can easily modify the source code by Googling.
  - Should not work with the next major OS upgrade.

```
[psh@store01 arcmsr-1.51.00.16]$ diff --color arcmsr.c.orig arcmsr.c
1282c1282,1286
<                nvec = pci_alloc_irq_vectors(pdev, 1, 1, PCI_IRQ_LEGACY);
---
>                 unsigned int flags = PCI_IRQ_MSIX | PCI_IRQ_MSI;
> #ifdef PCI_IRQ_LEGACY
>                 flags |= PCI_IRQ_LEGACY;
> #endif
>                  nvec = pci_alloc_irq_vectors(pdev, 1, 1, flags);
```

- **Q. can we just upgrade RAID card without changing disk enclosures?**
  - Should we go to ARECA? Can we use different manufacturer?
  - If none of solution works and we faced OS upgrade, what can we do?

Summary

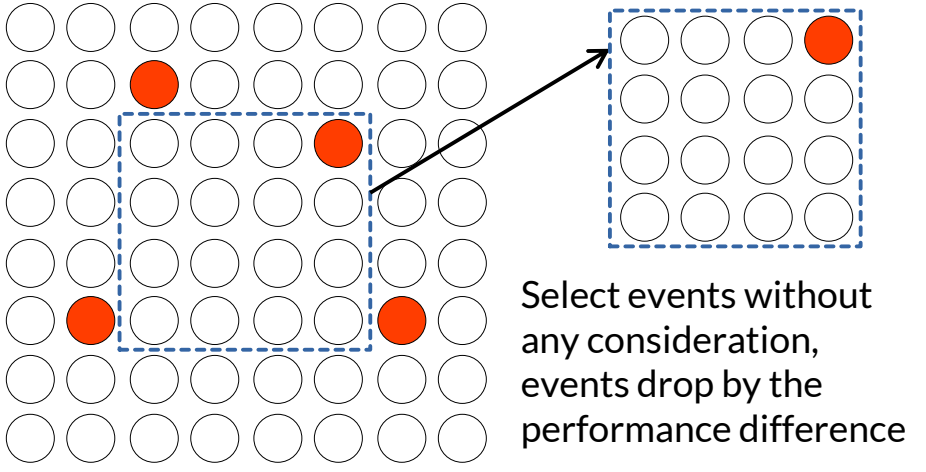
- Basic STORE introduction
  - Pros and Cons of SROOT Storage
  - How the direct ROOT recording upgrade is done during LS1
- A bit detail of post-processing tools and online-offline file transfer handshake in the view of online side.
- Operation status: basically, not many changes, STORE is very stable.
  - A couple of minor concerns
- After upgrading STORE01, the STORE performance can be increased ~50% more, by increasing the number of basf2 process.
- Must keep watching the impact of VTX upgrade in terms of data size.
- A future support of ARECA RAID controller is another point for sustainability.

Backup

Physics ERECO

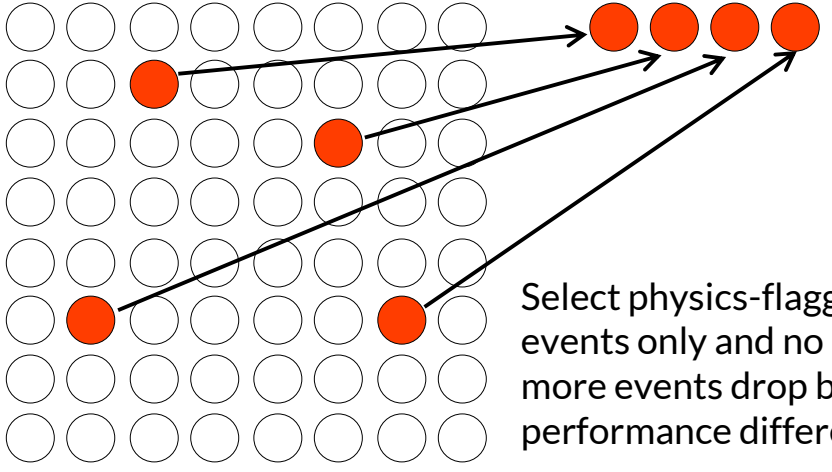
- # of ERECO is smaller than HLT → Not all events can be processed.
  - Events drops randomly.
- We wanted more statistics of physics features with the random sampling.
  - The random sampled DQMs are also important especially for VXD performance.

< Random sampling >



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

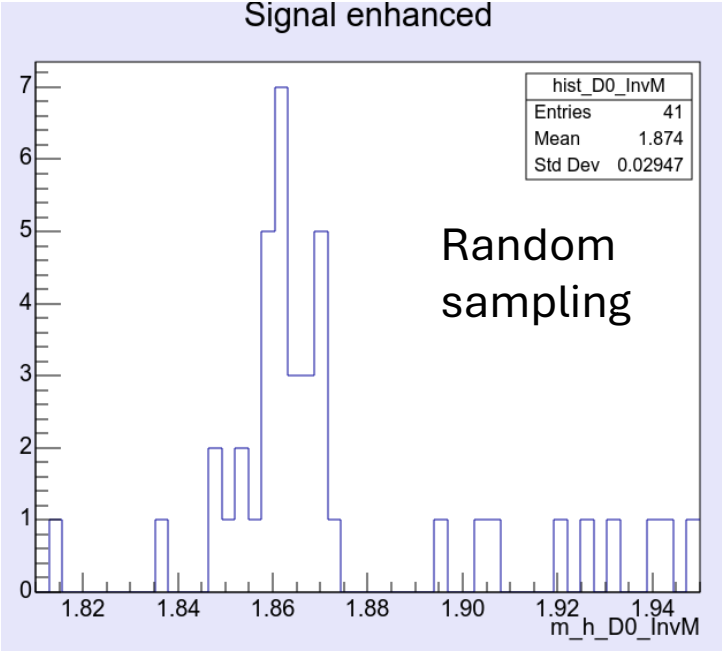
< Physics sampling >

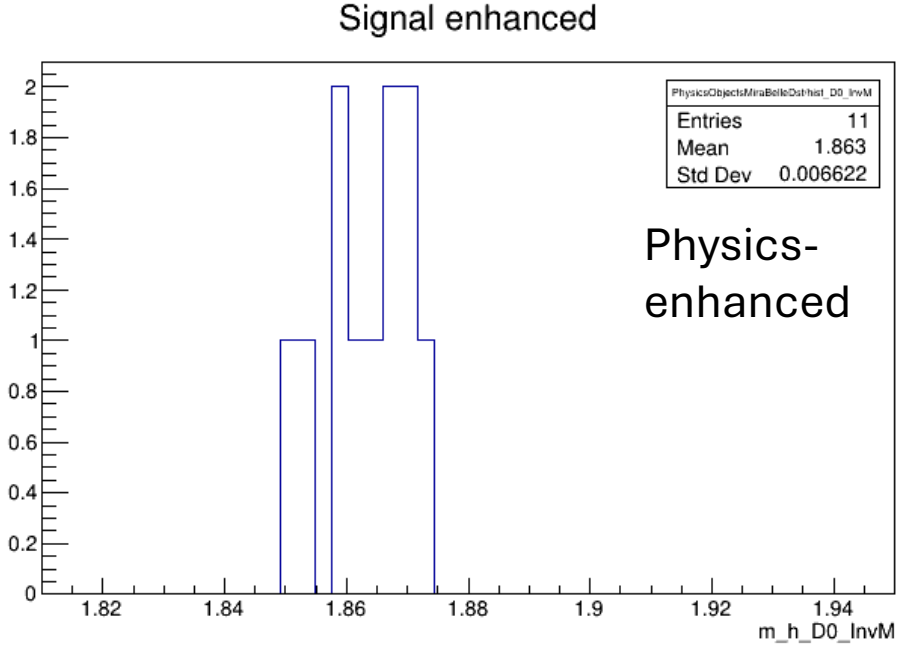


Select physics-flagged events only and no more events drop by the performance difference

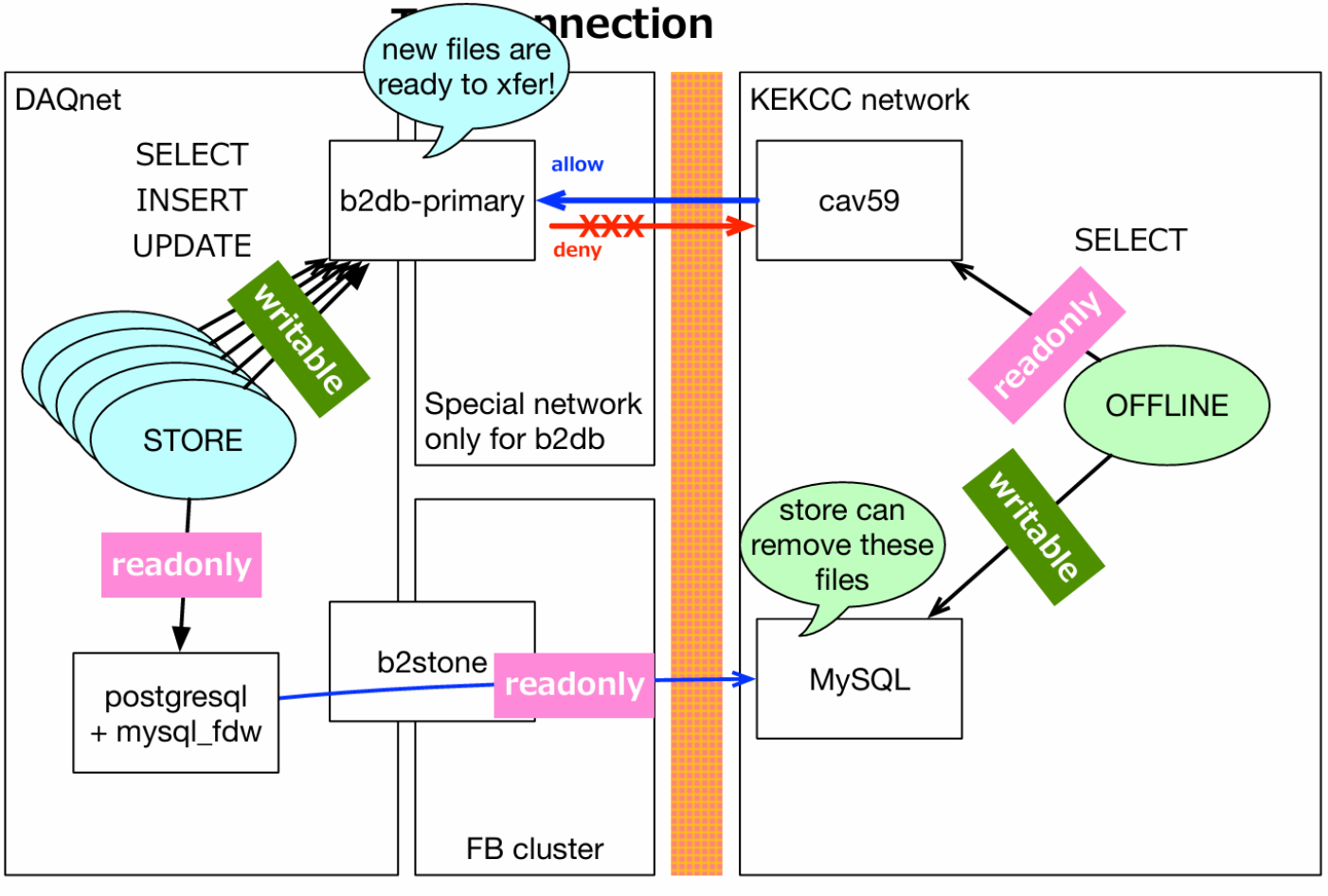
Physics ERECO

- 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}
  - Physics-enhanced: 11 D^0 from D^* events with 46k inputs → 2.4×10^{-4}
  - Over 25 times statistics of the physics flagged events from physics ERECO.





DB replication detail



15 STOREs Spec.

host	store01	store02	store03	store04	store05	store06	store07	store08	store09	store10	store11	store12	store13	store14	store15
CPU	Silver 4310	E5-2650 v3	E5-2650 v3	E5-2650 v4	E5-2650 v4	E5-2650 v4	E5-2650 v4	Silver 4116	Silver 4116	Silver 4116	Silver 4214R	Silver 4214R	Silver 4116	Silver 4310	Silver 4310
# of physical core	24	20	20	24	24	24	24	24	24	24	24	24	24	24	24
Int SPECrate(2017)	174	95	95	111	111	111	111	110	110	110	129	129	110	174	174
SPECrate/core	7.25	4.75	4.75	4.63	4.63	4.63	4.63	4.58	4.58	4.58	5.38	5.38	4.58	7.25	7.25
RAM (G)	96	64	64	32	32	32	64	96	96	96	96	96	96	96	96
RAID card	ARC-1883LP	ARC-1882LP	ARC-1883LP	ARC-1883LP	ARC-1883LP	ARC-1883LP	ARC-1883LP	ARC-1883LP	ARC-1883LP	ARC-1883LP	ARC-1883LP	ARC-1884IXL-[8|12]	ARC-1883LP	ARC-1883LP	ARC-1883LP