High Level Trigger of Belle II experiment

GDR-InF Annual workshop 2022

Valerio Bertacchi, Karim Trabelsi, Vidya Vobbilisetti

2 November 2022

Belle II experiment at SuperKEKB collider

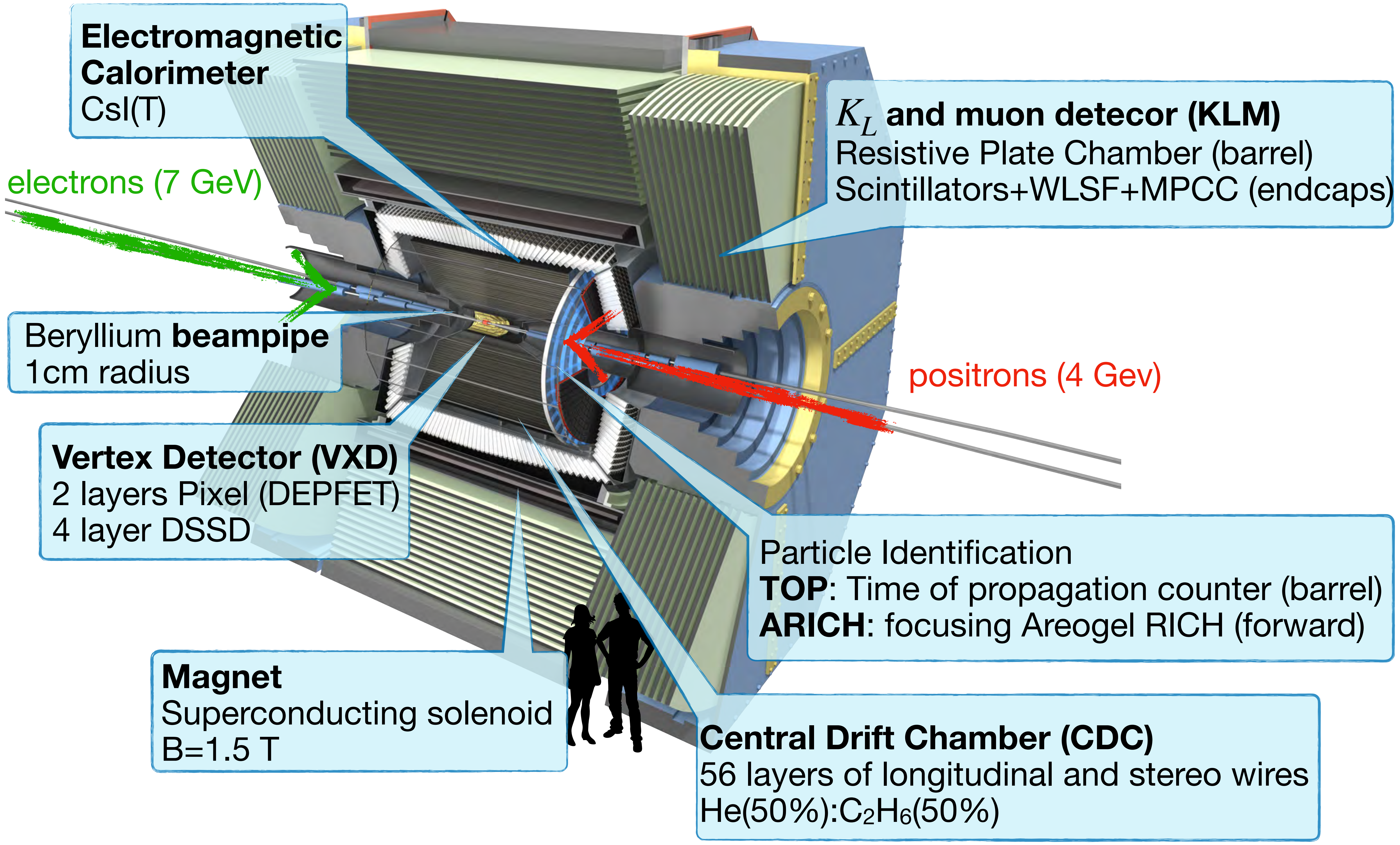
SuperKEKB

- Target peak luminosity: $6 \cdot 10^{35}\ \mathrm{cm^{-2}s^{-1}}$ (x 30 of KEKB)
- Target integrated luminosity: $50\ \mathrm{ab^{-1}}$ (x 70 Belle at $\Upsilon(4S)$)

Current Status

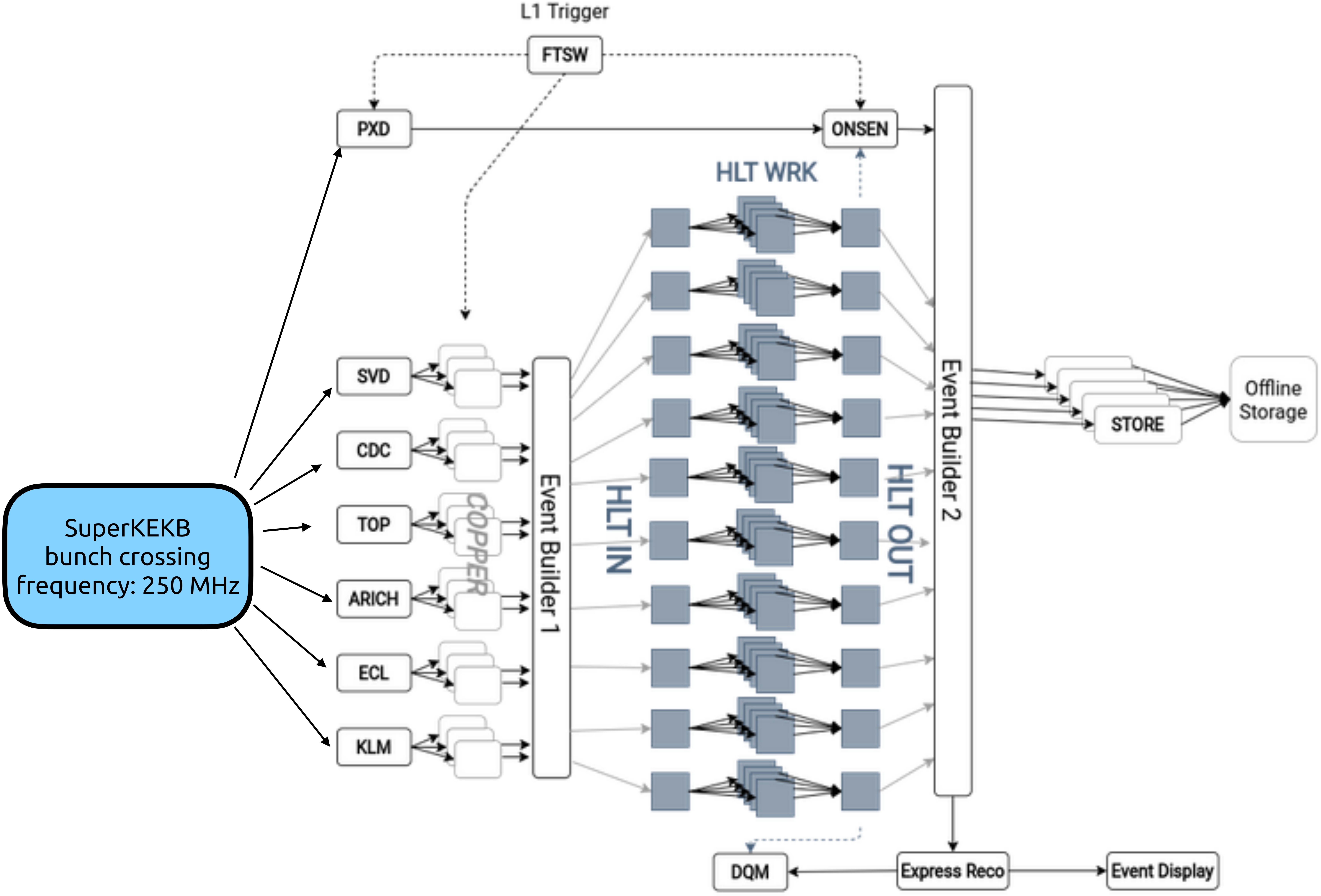
- complete detector data taking started in 2019
- Current peak luminosity $4.7 \cdot 10^{34}\ \mathrm{cm^{-2}s^{-1}}$ (reached the 22/06/2022)
- current integrated luminosity: $\sim 424\ \mathrm{fb^{-1}}$ (~Babar~0.5 Belle)
- Long Shutdown 1 (LS1) started in July

Belle II

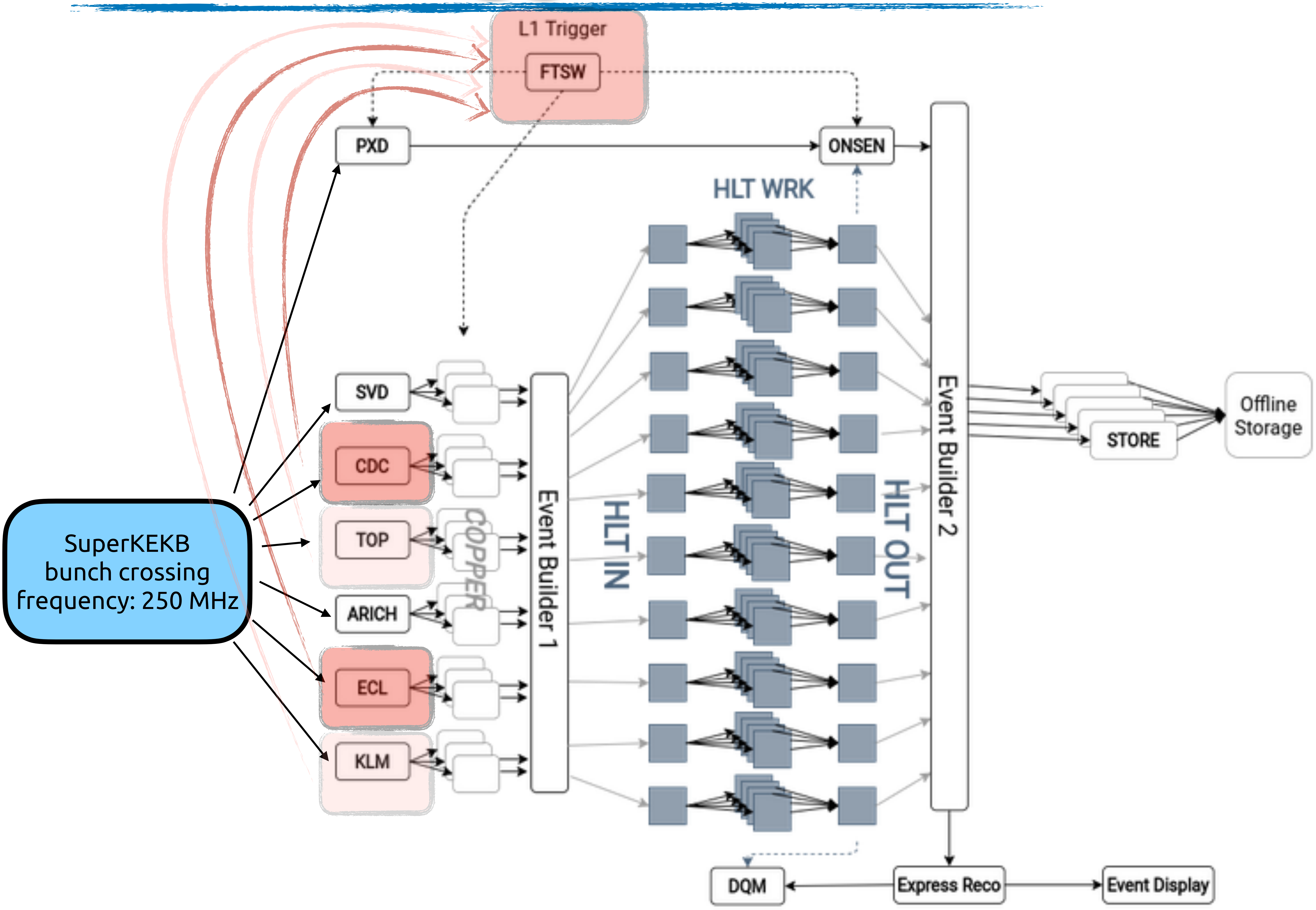


[Belle II Technical Design Report, arXiv:1011.0352]

Belle II trigger dataflow



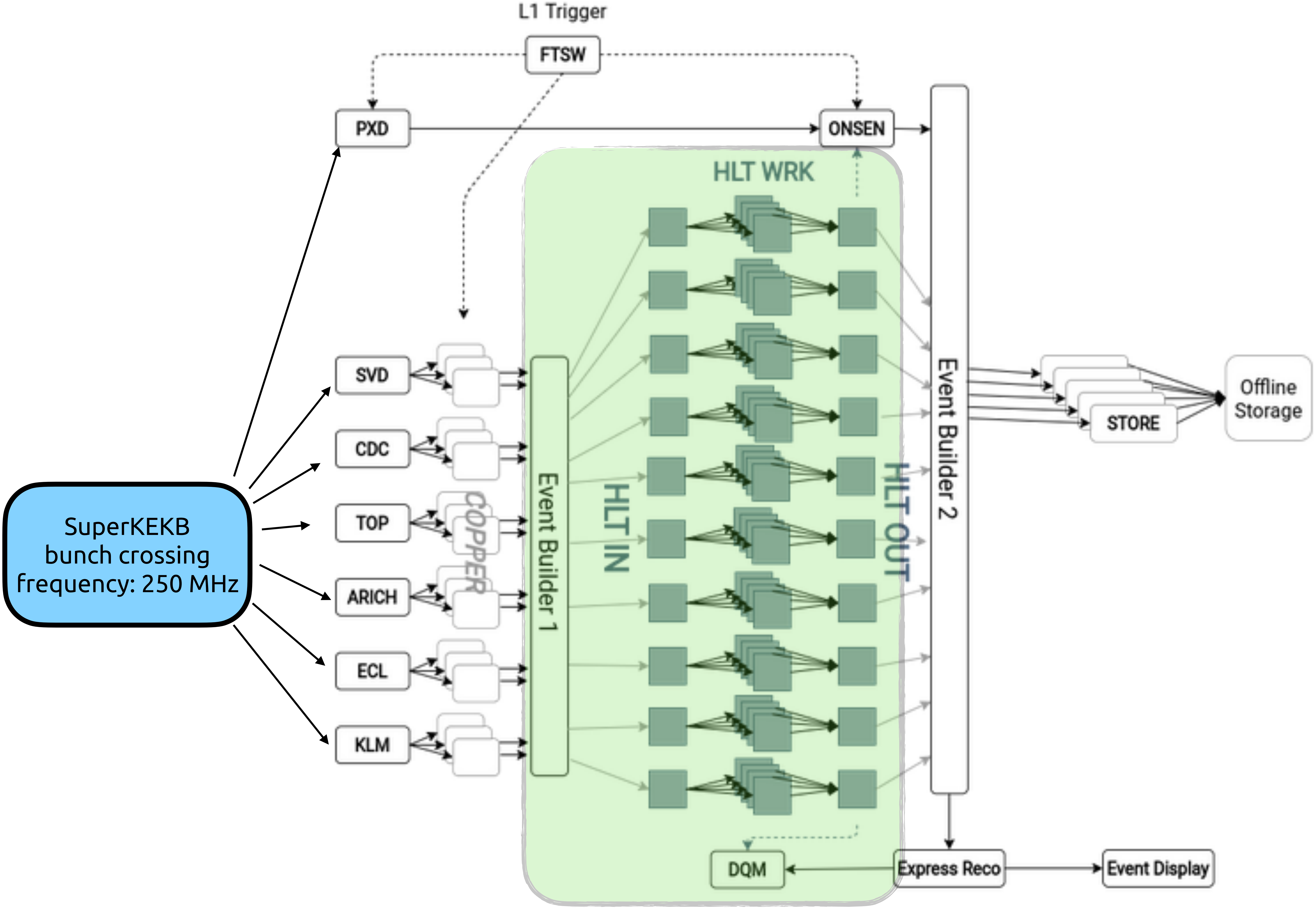
Belle II trigger dataflow: Level 1 trigger



L1 Trigger

- Purpose: **suppress the background** rate, retaining ~100% of $b\bar{b}$ events with high efficiency also for $c\bar{c}$ and $\tau^+\tau^-$
- Output rate:
  - Now: about **10 kHz**
  - Expected at target luminosity: 30 kHz
- latency: few μs
- Strategy:
  - processing on **FPGA**,
  - using OR of different, **orthogonal**, trigger lines (**CDC**, **ECL**)$\Rightarrow$ conservative approach

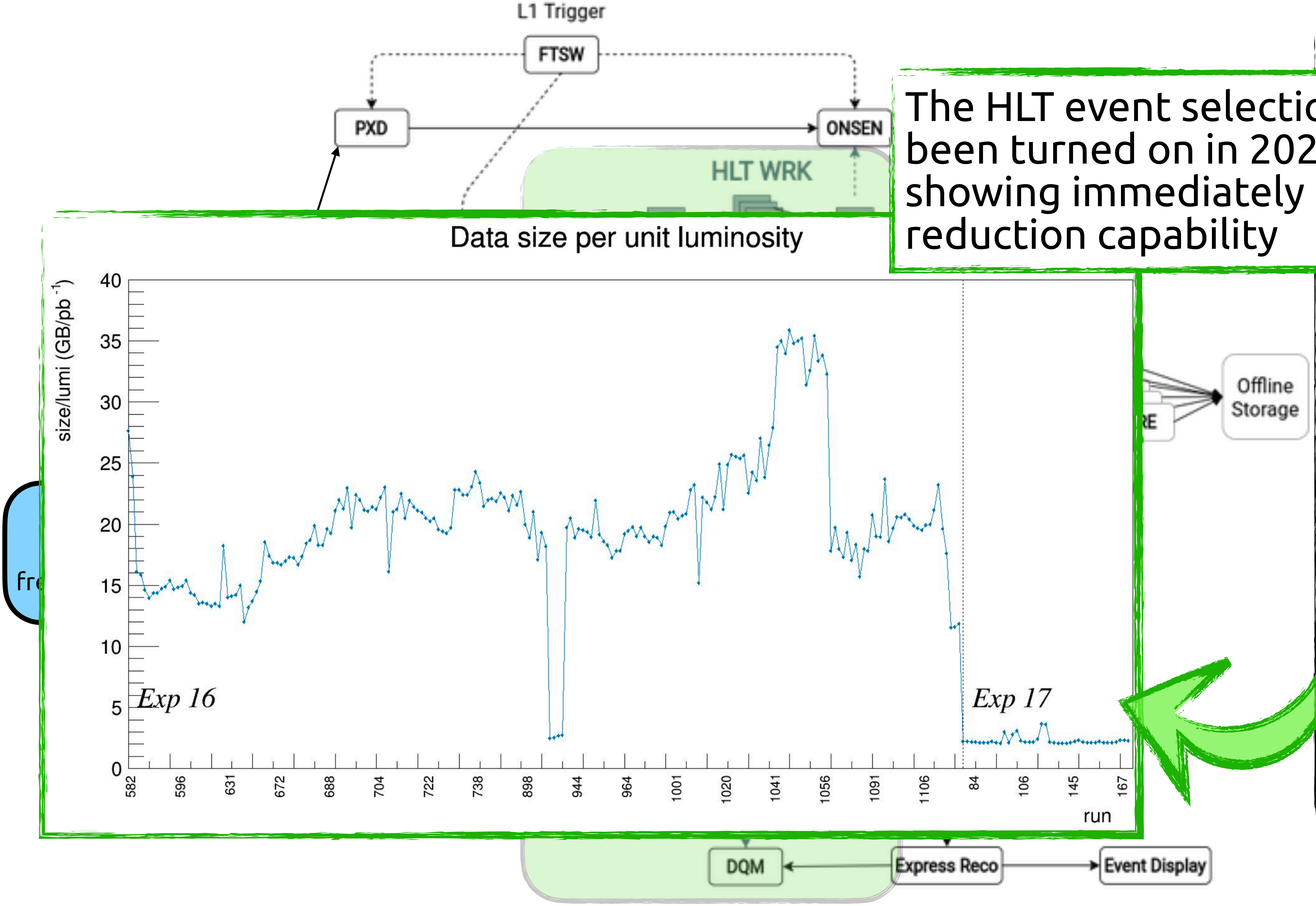
Belle II trigger dataflow: HLT



HLT

- Purpose:
  - reduce the trigger rate to a **storable rate**
  - run **DQM**
  - produce the **ROIs** for the PXD
  - assign the skim flag
- Output rate: ($\varepsilon \simeq 10 - 20\,\%$)
  - Now: about **2 kHz**
  - Expected at target luminosity: 6 kHz
- **Processing time: 300 ms**
- **Budget time** (N_{proc}/L1 rate)**: 400 ms**
- Strategy: **fast reconstruction** on CPU
- hardware:
  - Now: 10 units, about 500 cores per unit--> **2 x 4800 processors**
  - After LS1: +3 units (to sustain 20 kHz input rate)

Belle II trigger dataflow: HLT



The HLT event selection has been turned on in 2021, showing immediately its data reduction capability

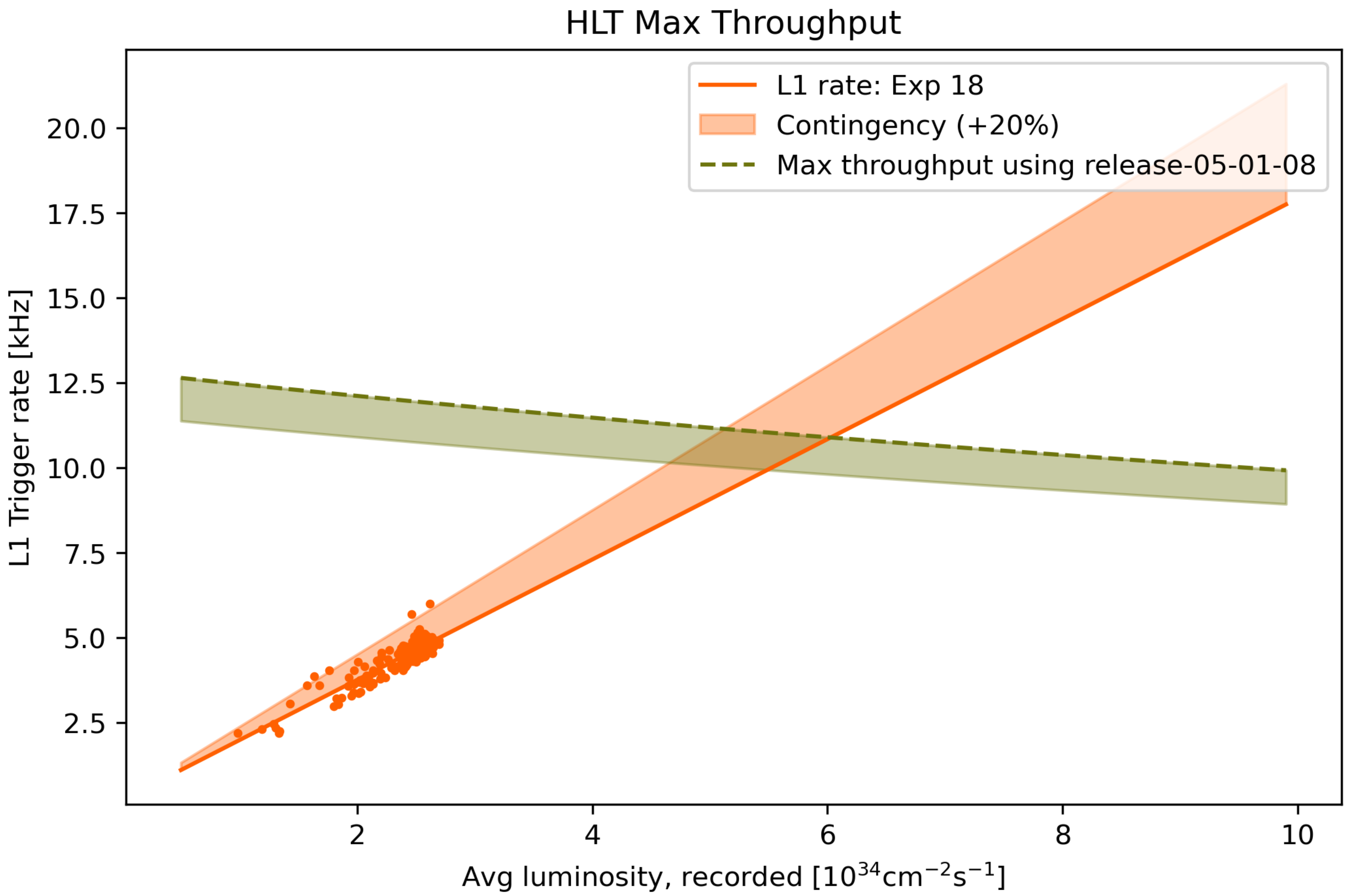
HLT

- Purpose:
  - …e trigger rate to a **storable**
  - …he **ROIs** for the PXD
  - assign the skim flag
- Output rate: ($\varepsilon \simeq 10-20\,\%$)
  - Now: about **2 kHz**
  - Expected at target luminosity: 6 kHz
- **Processing time: 300 ms**
- **Budget time** (N_{proc}/L1 rate)**: 400 ms**
- Strategy: **fast reconstruction** on CPU
- hardware:
  - Now: 10 units, about 500 cores per unit--> **2 x 4800 processors**
  - After LS1: +3 units (to sustain 20 kHz input rate)

HLT limits (exp 18 ~ 2021 data taking)

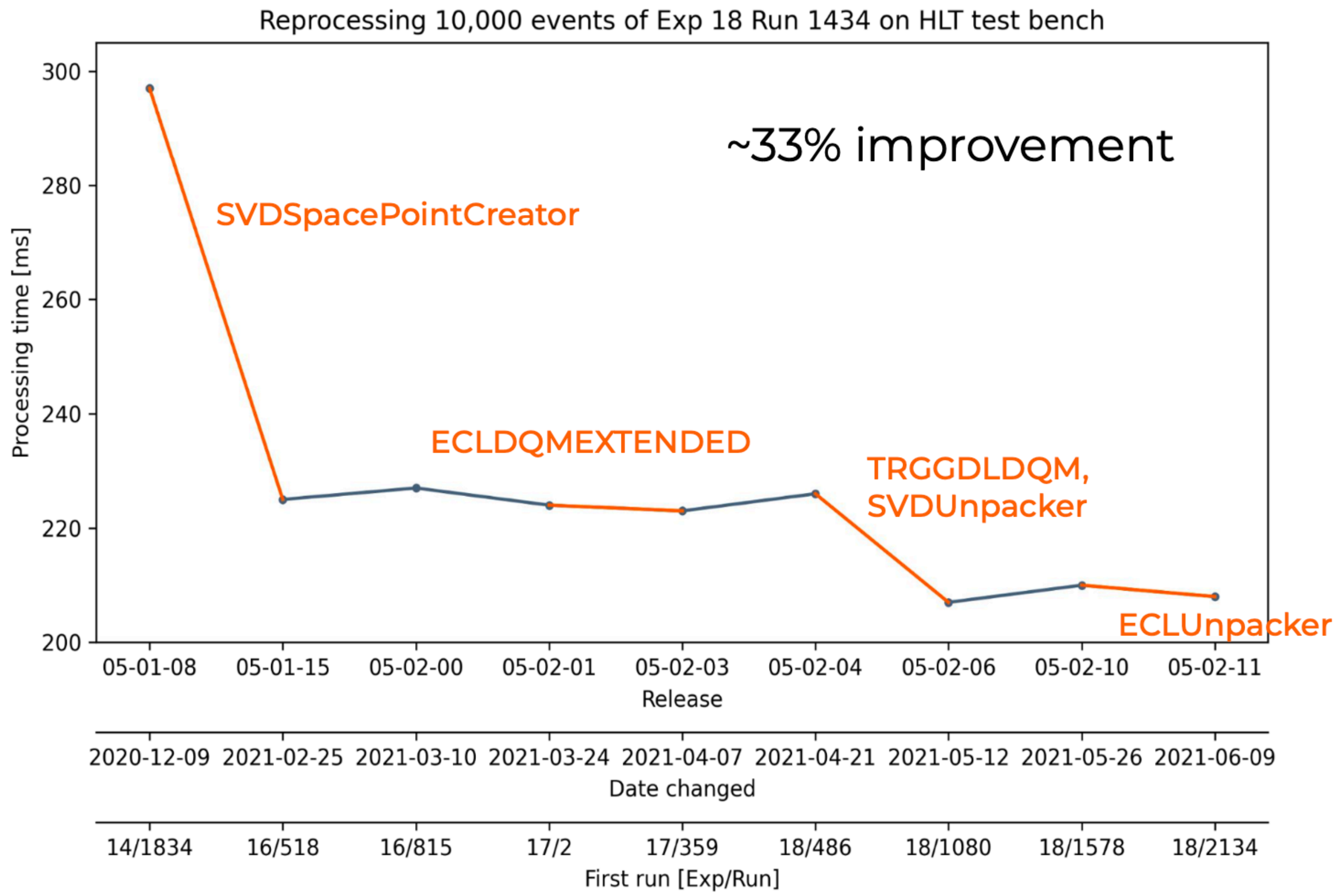
Throughput = $N_{\text{processes}}$/process time

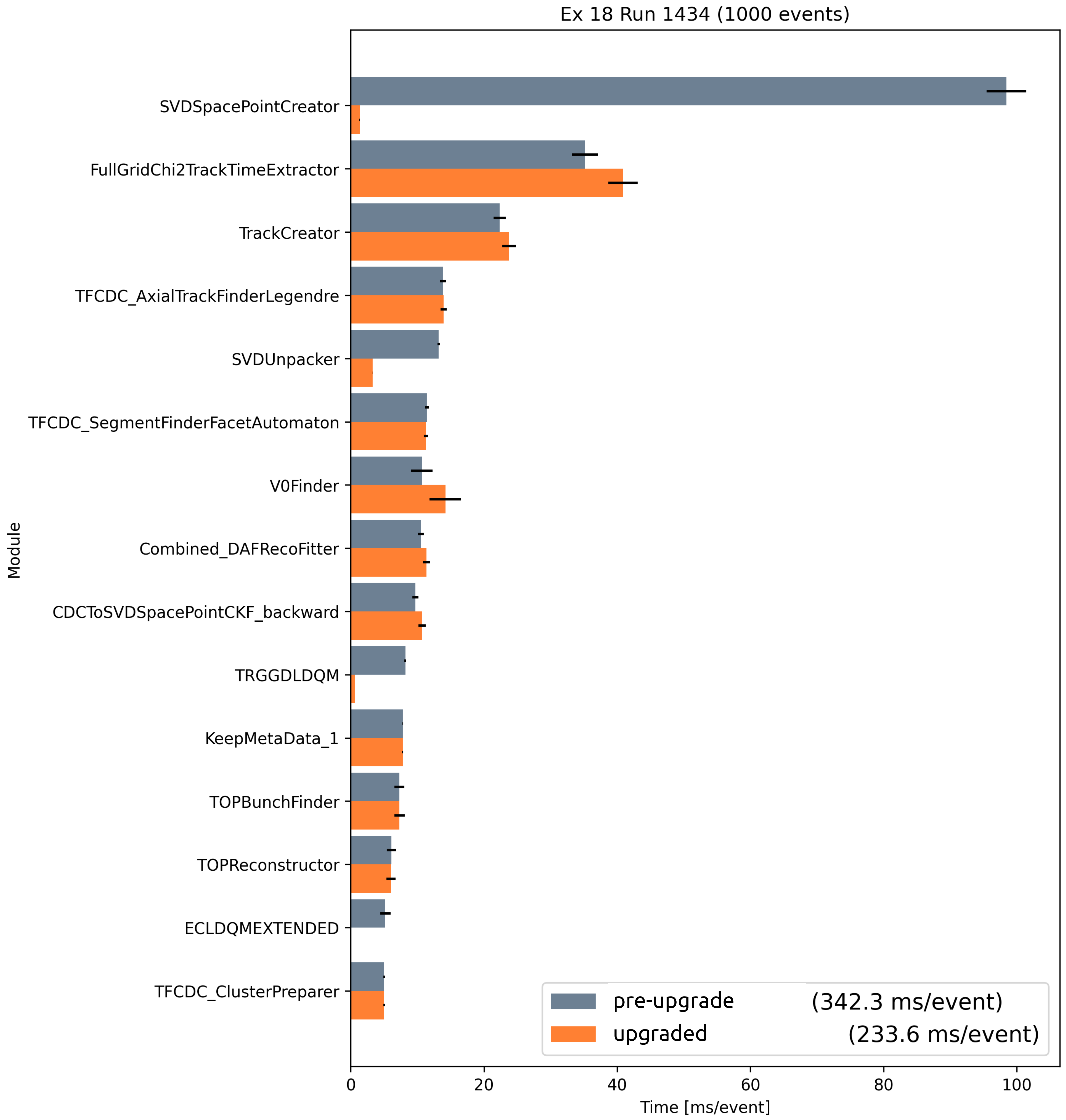
- L1 output (HLT input) **increase with luminosity** given the increased event rate
- Throughput decrease with luminosity given the **increasing complexity of the events** (higher background) which requires longer processing time
- In 2021 ($\mathscr{L} = 2 \cdot 10^{34}\text{cm}^{-2}\text{s}^{-1}$) Belle II realised that the conditions are not sustainable to reach the LS1
- Optimization of HLT is needed to increase the throughput (**decrease the processing time**)



First optimization before the Long Shutdown 1

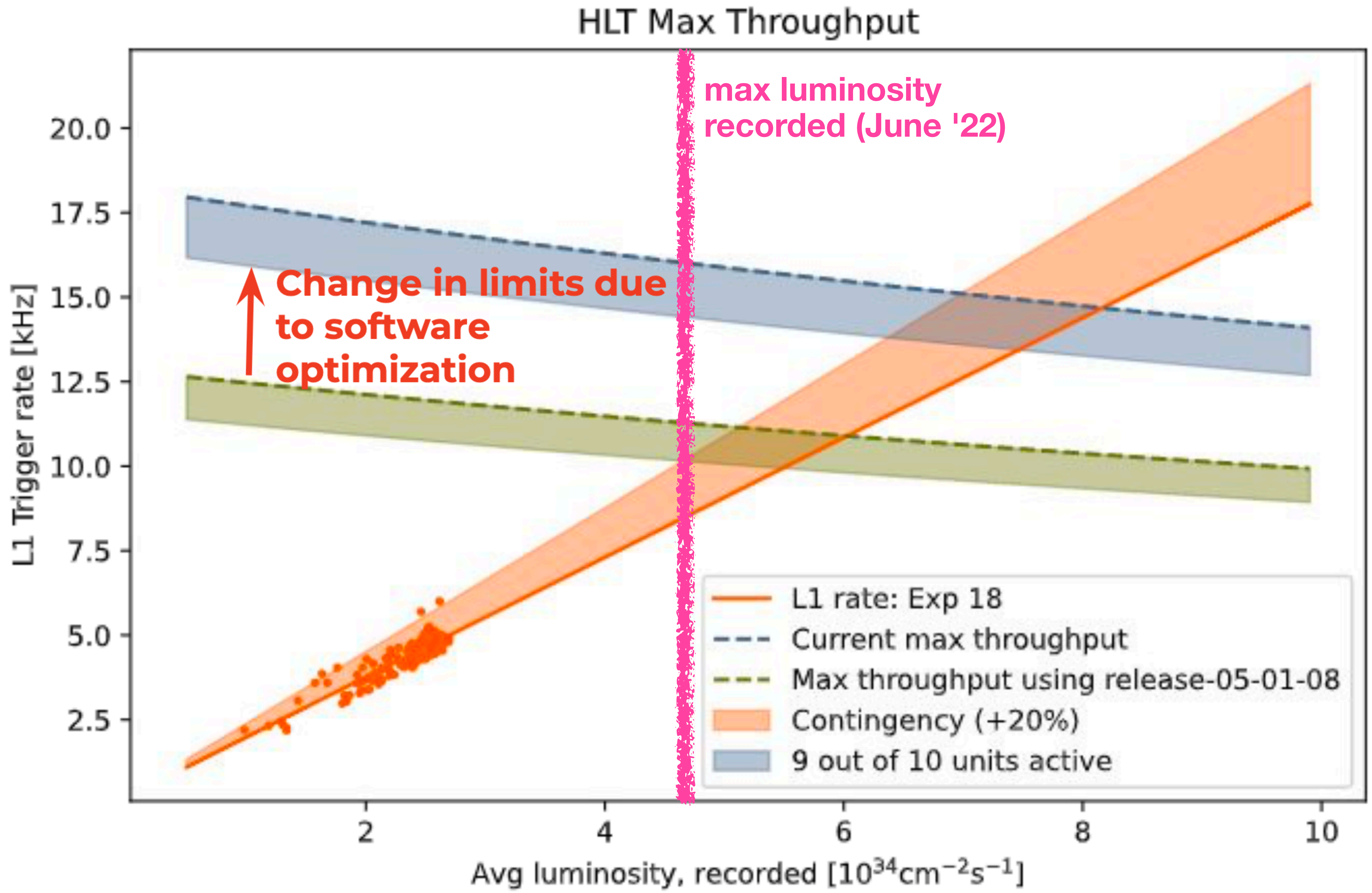
- Strategy: optimize the code, **producing identical results**
- Constraint:
  - implemented during data taking to **survive until LS1**
  - the HLT decision, the DQM, the skim **cannot change** to keep consistent data taking
- Optimized the ROOT object management





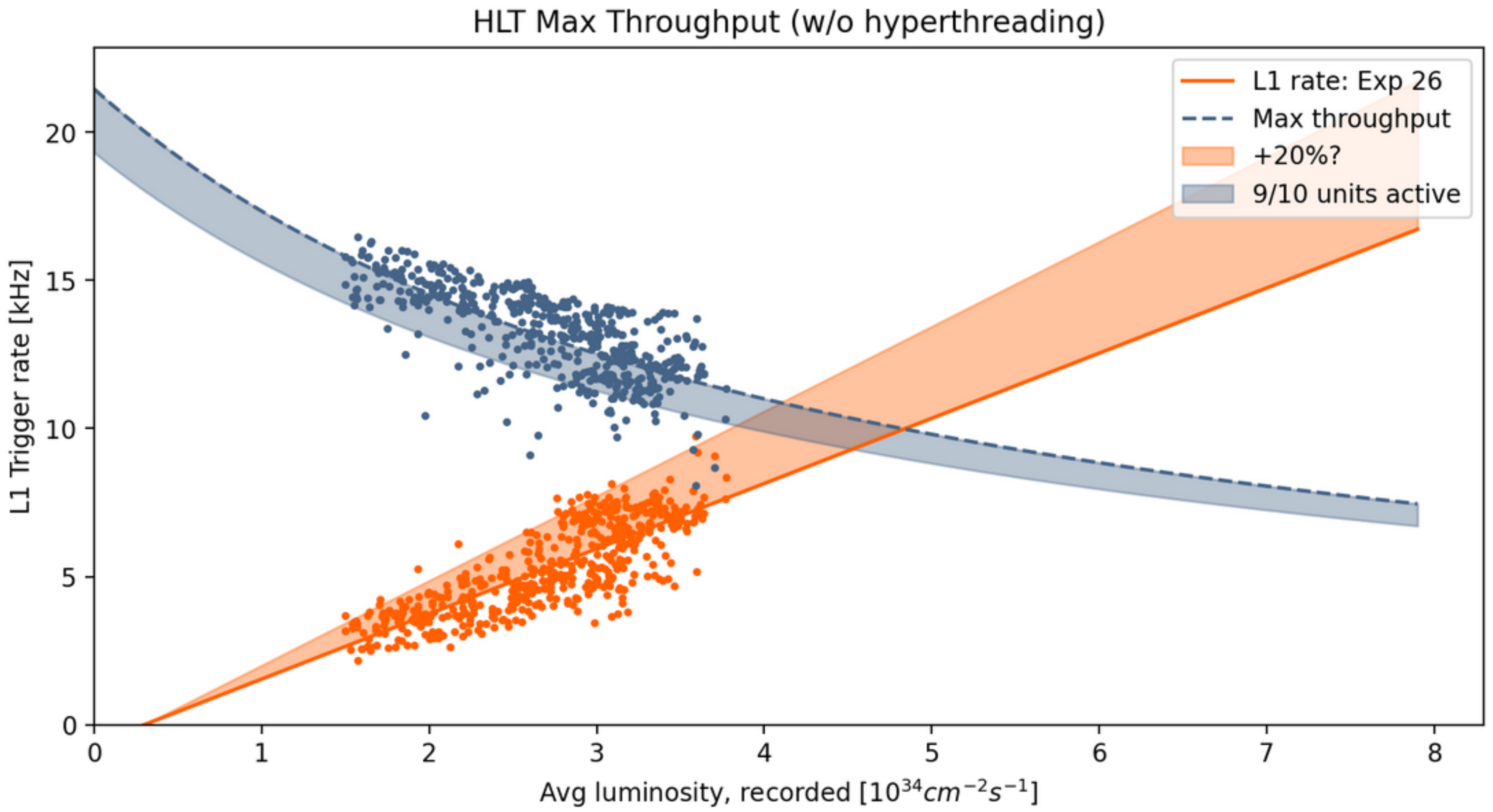
First optimization impact

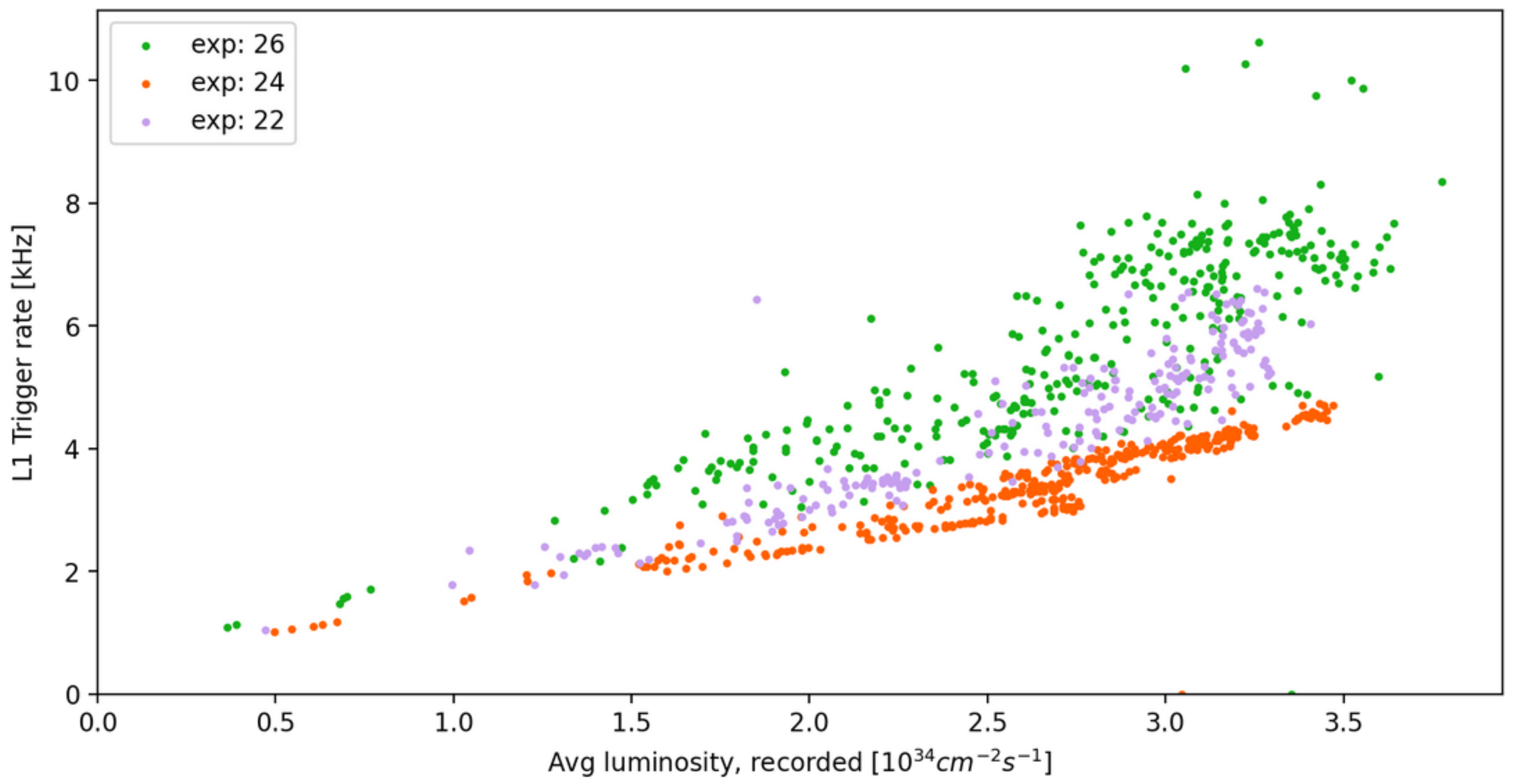
- Thanks to this optimization work we will survived until LS1!



HLT current limits

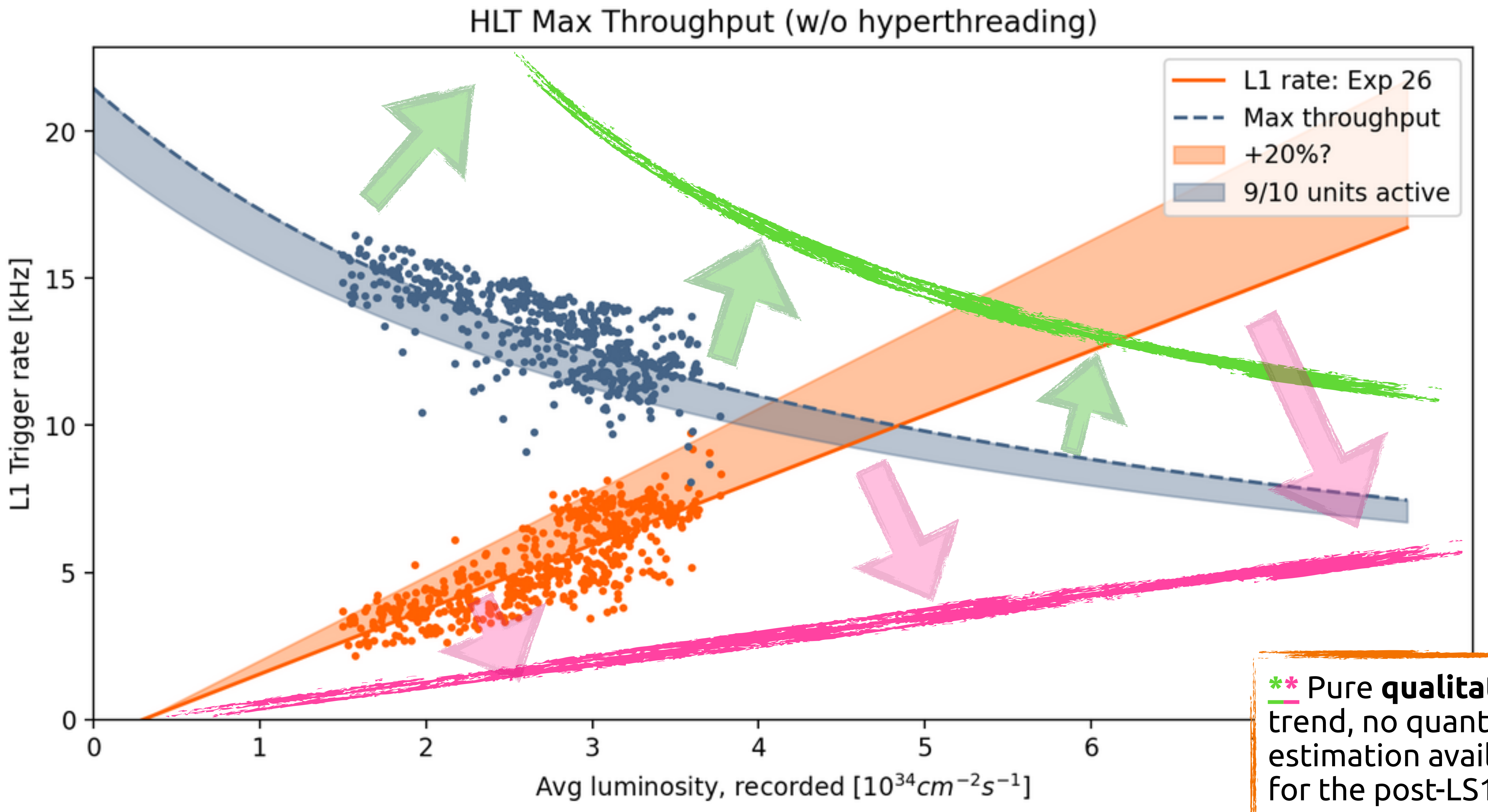
- But we arrived very close and the trend is not promising…



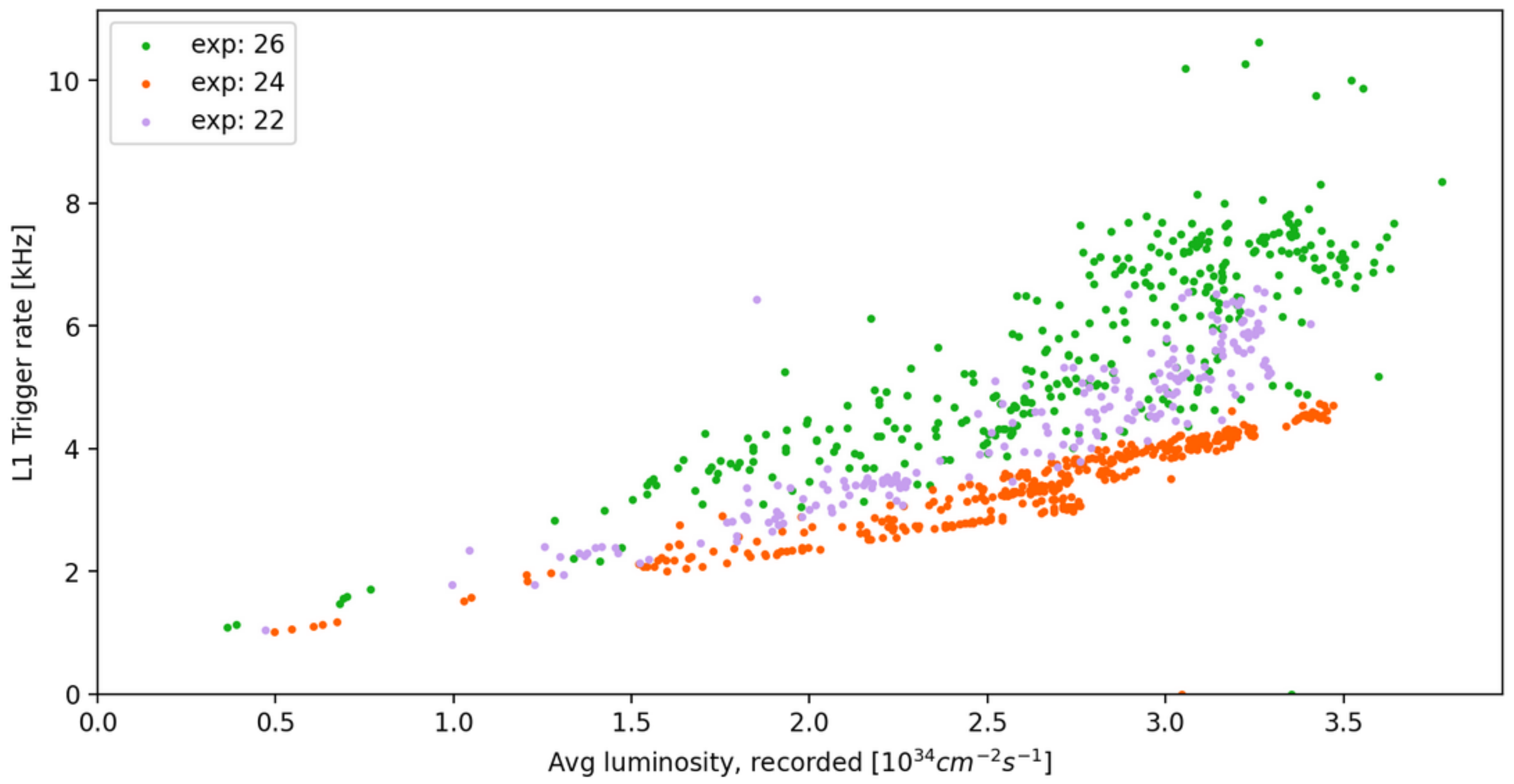


HLT current limits

- But we arrived very close and the trend is not promising...
- After LS1 we will have:
  - **new collimators** (reduce input rate overall and bkg in $b\bar{b}$ events) $\Rightarrow$ Reduce L1 output
  - 3 **additional HLT units** $\Rightarrow$ increase Throughput
- However further HLT optimization is needed:
  - as **safety factor**
  - to reduce the computing burden
  - indirect impact on **MC production** and **data reprocessing**

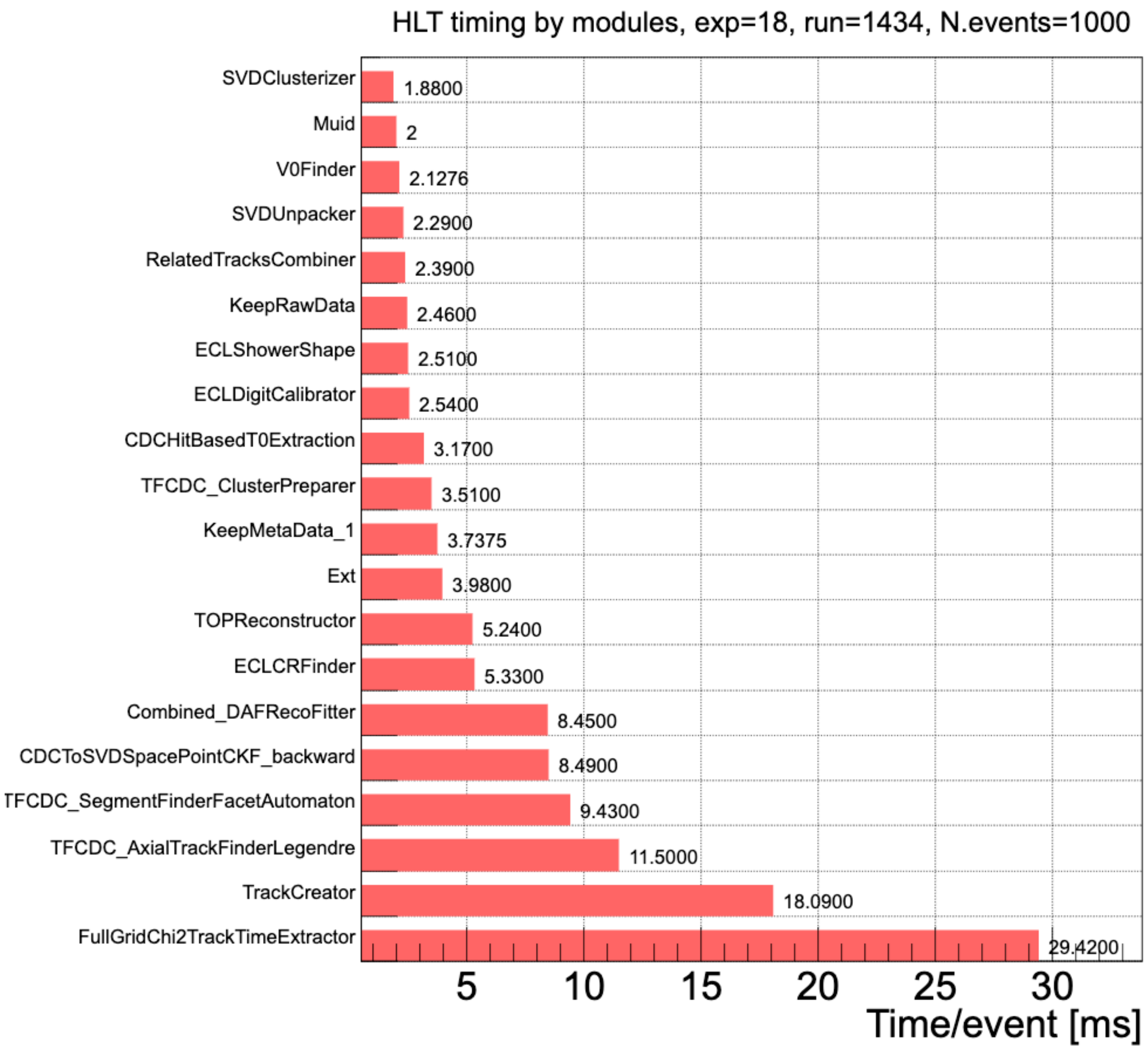


** Pure **qualitative** trend, no quantitative estimation available for the post-LS1



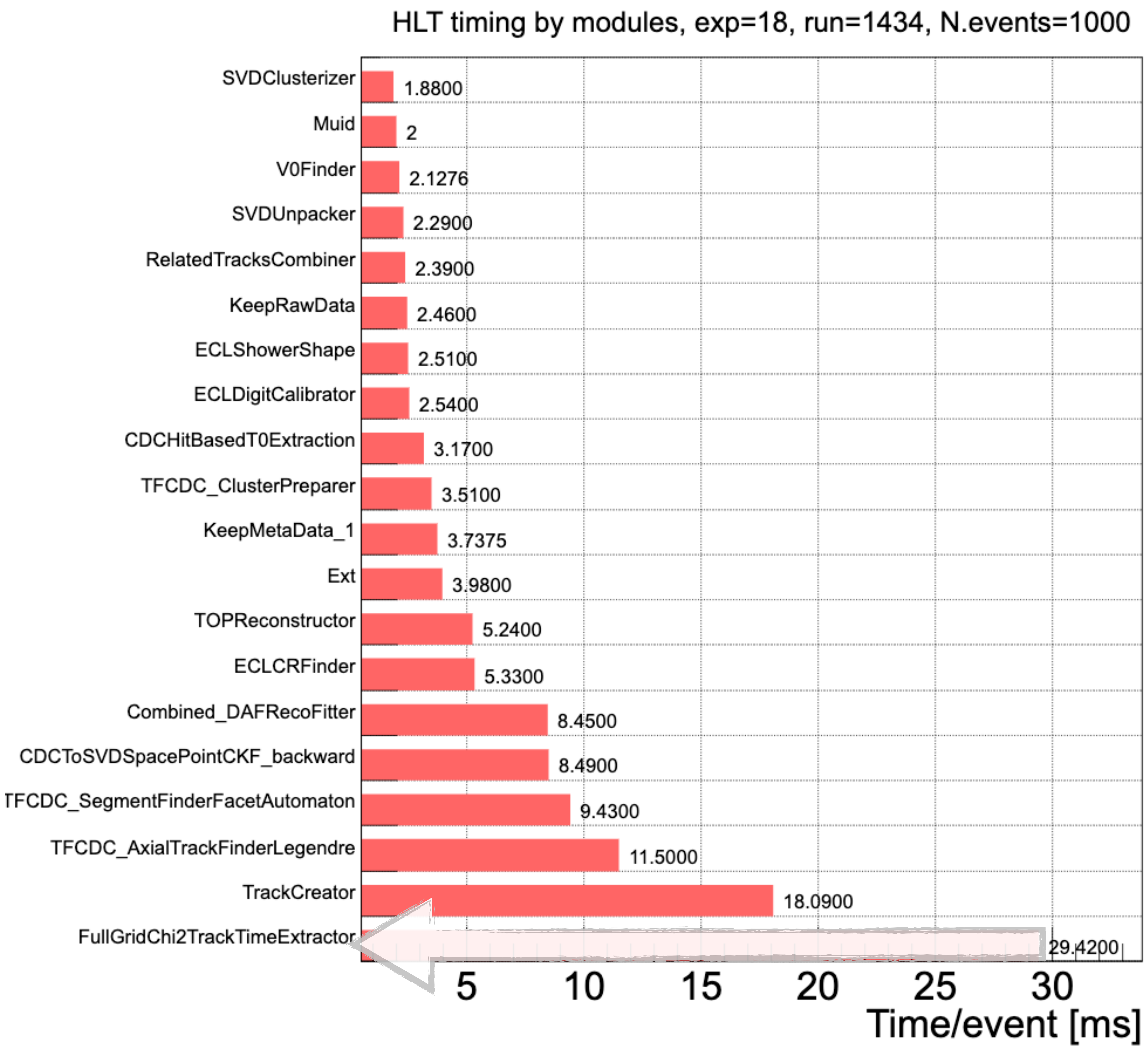
Further optimization is needed

- Strategy: **modify the reconstruction** strategies, allowing also *small degradation*, to save processing time



Further optimization is needed

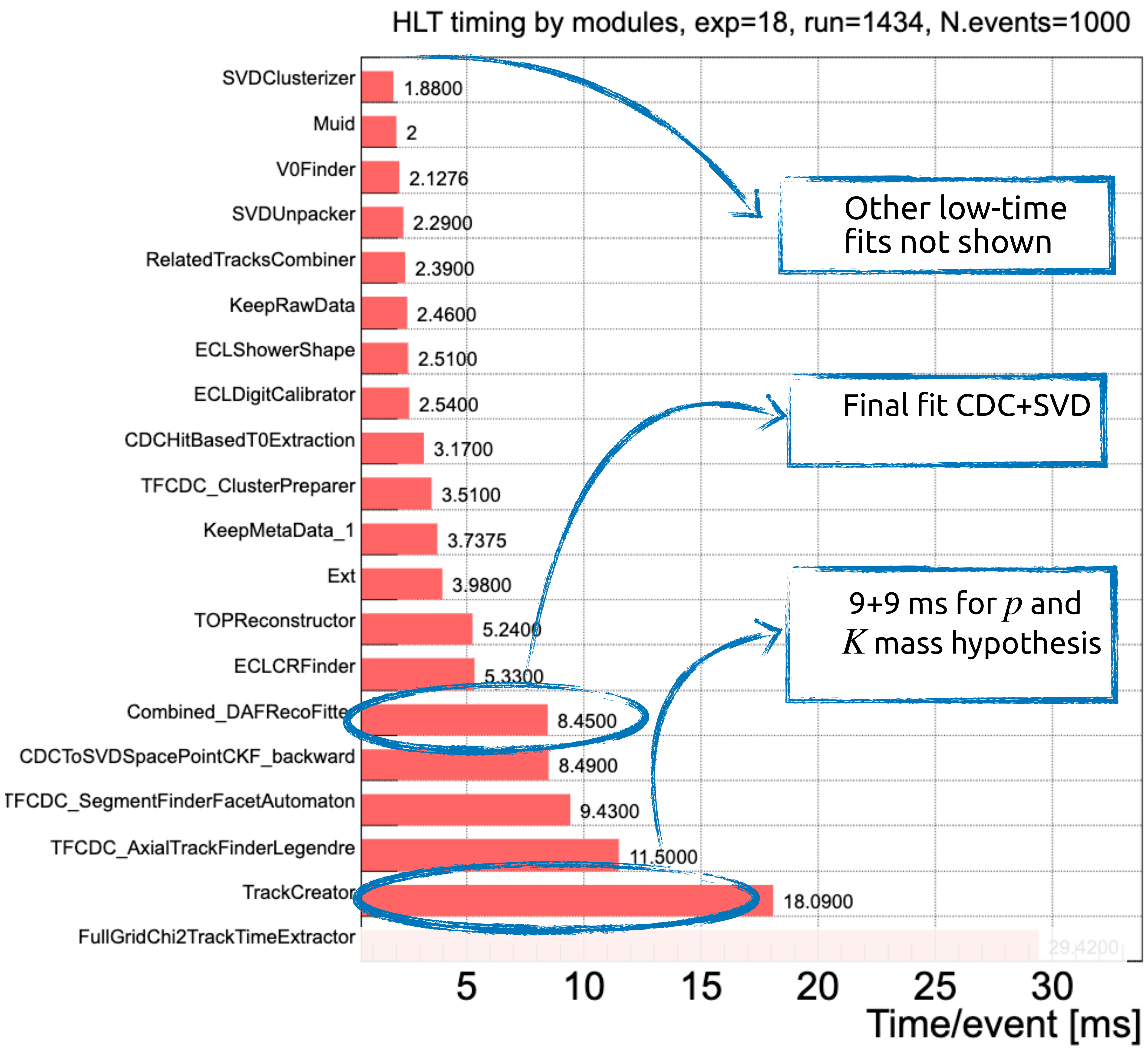
- Strategy: **modify the reconstruction** strategies, allowing also *small degradation*, to save processing time
- First achieved result: CDC Event Time estimation has been **replaced with SVD Event Time** estimation ⇒ 2000 times faster [see backup]
- Next step: reducing tracking processing time **(track fitting)**



Track Fitter calls

- The fitter is **called ~5 times per track**, using a Deterministic Annealing Filter (**DAF**)
- With the current configuration the DAF takes **15 ms/track** for each call

The DAF its optimization has a **radical impact** on reconstruction CPU time (and tracking performance)



DAF

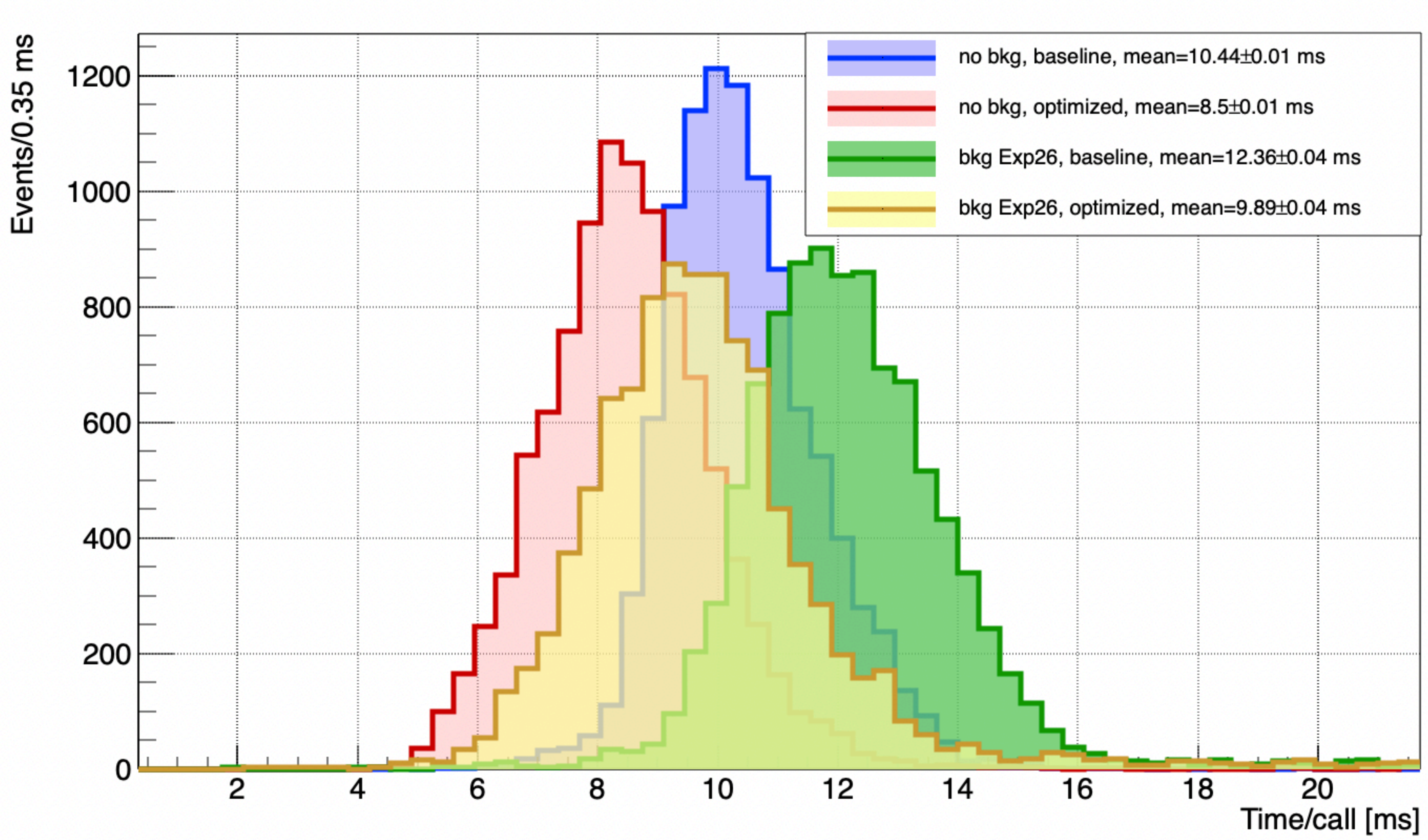
- For each call of the fitter the DAF (*Deterministic Annealing Filter*) is called
- The purpose of the DAF is to remove from the fit the **outlier hits** to improve the fit accuracy
- Method:
  - The DAF is **assigning weights** (in the range [0,1]) to each hit, accordingly to the residuals between the measurement and the Kalman Filter prediction.
  - The fit is r**epeated multiples times lowering an annealing temperature**
  - A **convergence criterion** is defined, based on the variation of the weights and the p-value of the fit (see next slide)
- Status: the **DAF has been never optimized**, and in the current configuration the convergence is not tuned ⇒ **extremely time consuming**

DAF demonstrative optimization

Test condition:
- CDC only tracking
- Muon Gun
- pT=1 GeV
- $\theta = 70^{\circ}$

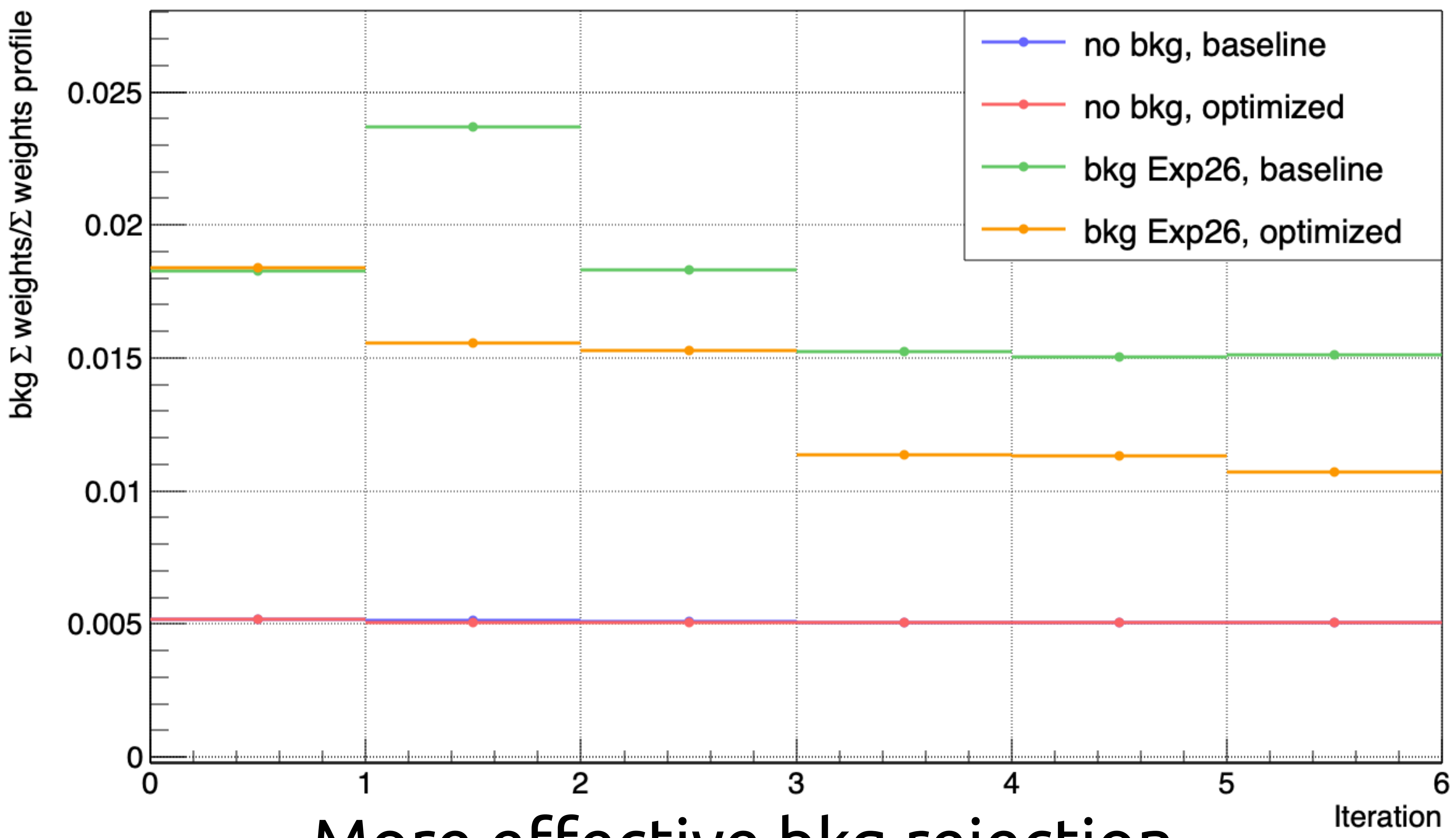
Changed some hyperparameters of the DAF *[see backup]*:

- to obtain reasonable **convergence behaviour** (use the iteration range, use mainly primary convergence criterion, exploit more wisely the p-value)
- having the CPU **time figure of merit**



Better timing performance

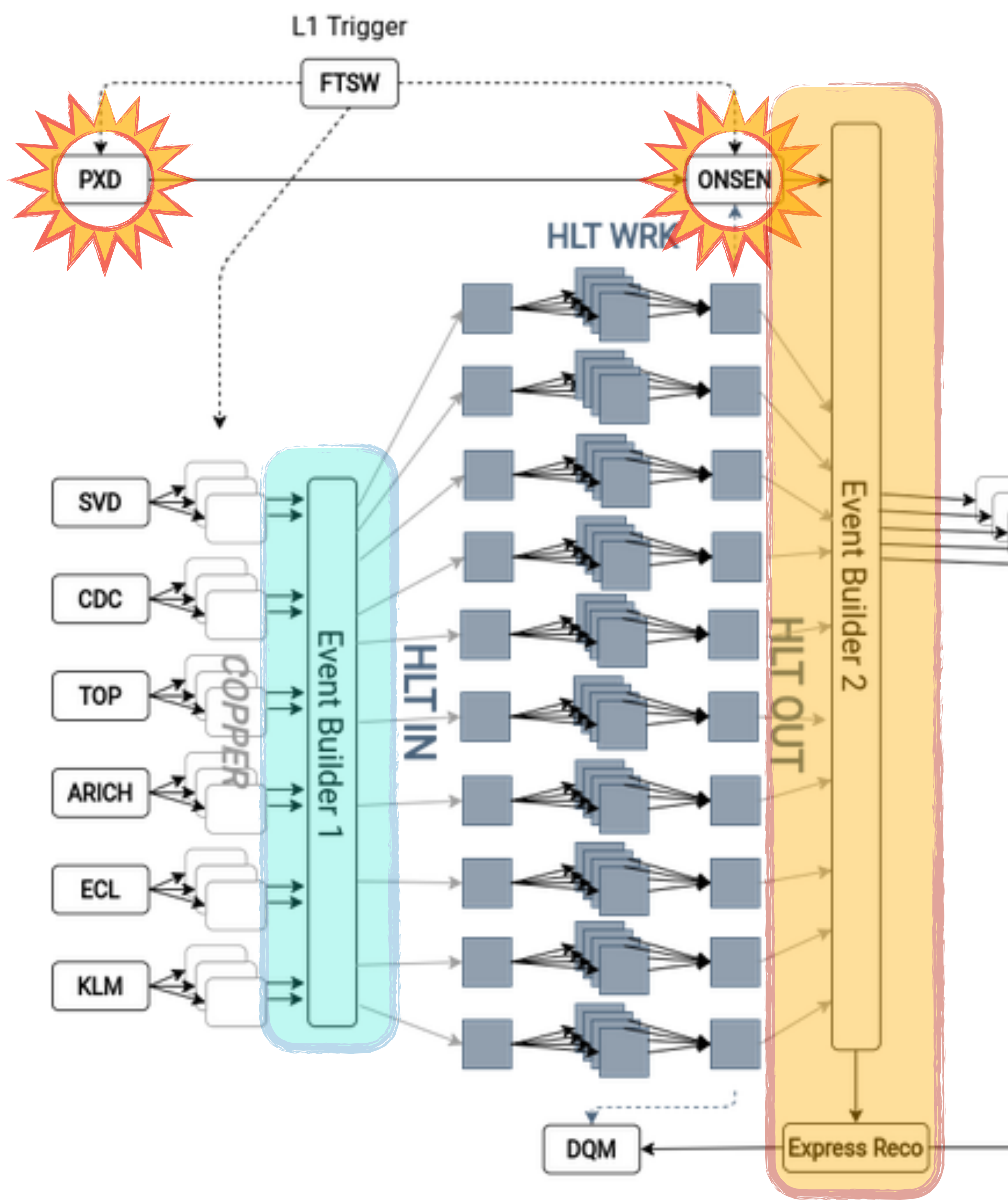
Un-purity of the tracks



More effective bkg rejection (improving in high-bkg scenario)

The future of the HLT

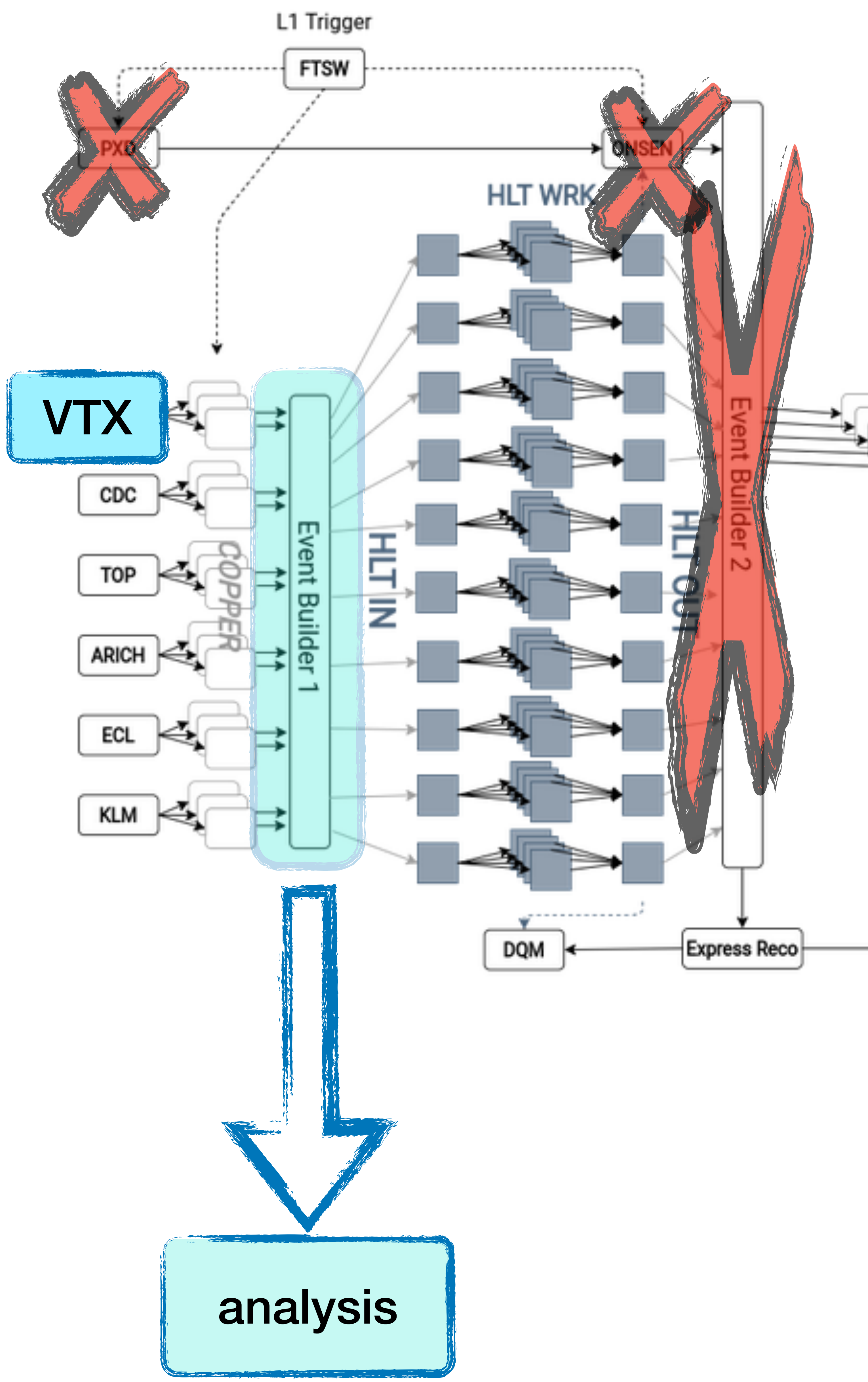
- With current HLT scheme we already have a **full online reconstruct** with 2 missing elements:
  - **calibrations** with most updated condition: built and then applied using run-by-run information
  - **PXD**: too slow to be read or used online *[see backup]*



The future of the HLT

- With current HLT scheme we already have a **full online reconstruct** with 2 missing elements:
  - **calibrations** with most updated condition: built and then applied using run-by-run information
  - **PXD**: too slow to be read or used online *[see backup]*
- Can we change the scheme to use the HLT reconstruction as final one?
  - calibrations: with a **more stable detector (accelerator)** we can use previous run information to **calibrate online**
  - PXD: after the LS2 (2026-2027) we plan to replace the PXD+SVD with the **VTX: six-layer CMOS pixel detector,** which can be fully used at HLT level

- Advantages: **faster** and **tidier** dataflow, **online final** reconstruction



Conclusions

- The HLT of Belle II is powerful tool to obtain close-to-final online reconstruction
- Given the increasing background the **HLT need constant optimization** to fulfil the timing constraints
  - Large room for improvement in the **track fitting** step
- Thanks to the HLT, with an upgrade of the Belle II vertex detector, Belle II can obtain a r**eady-to-analysis online reconstruction** for free

NB: dotted = my "prediction" only

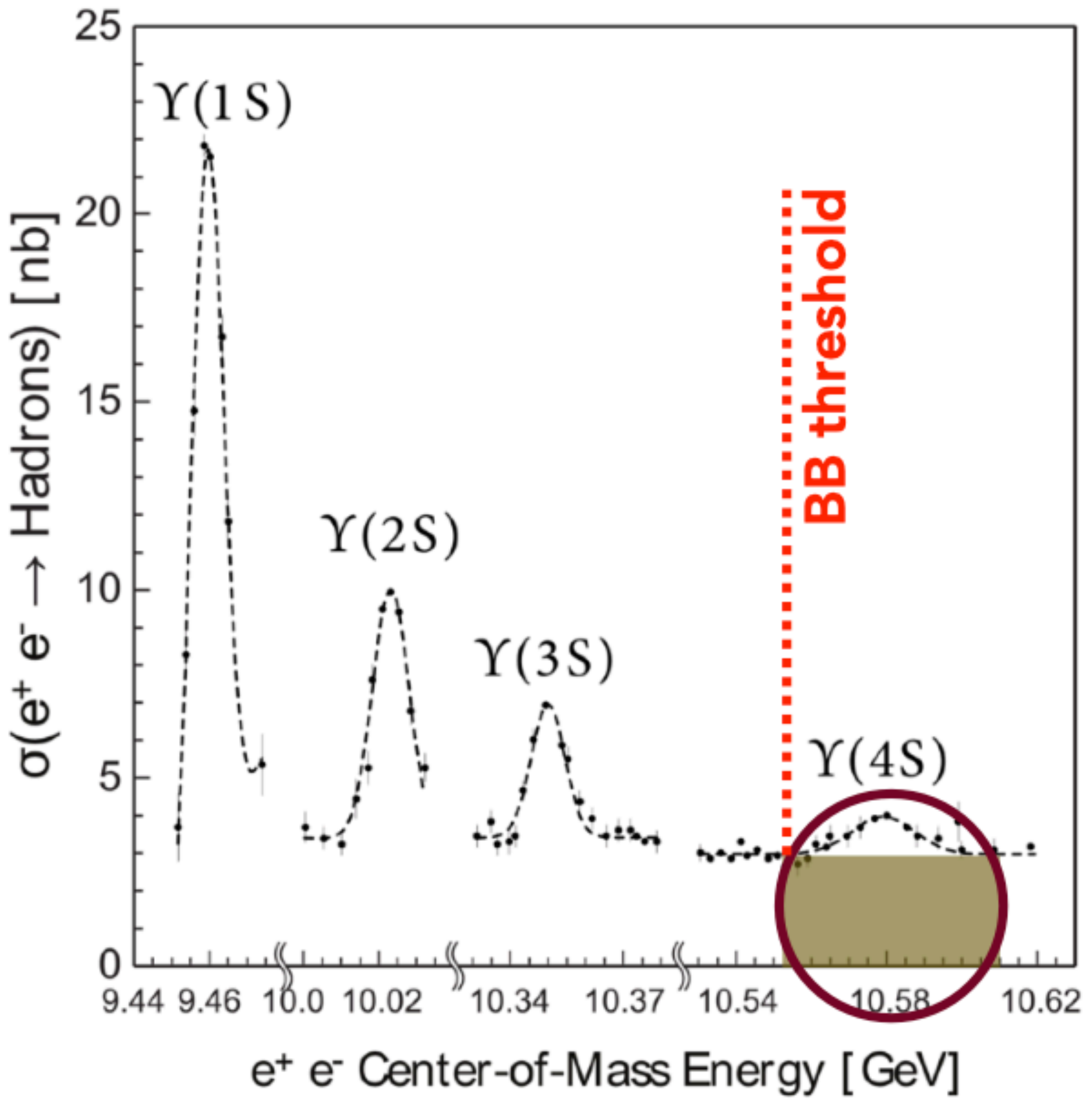
BACKUP SLIDES

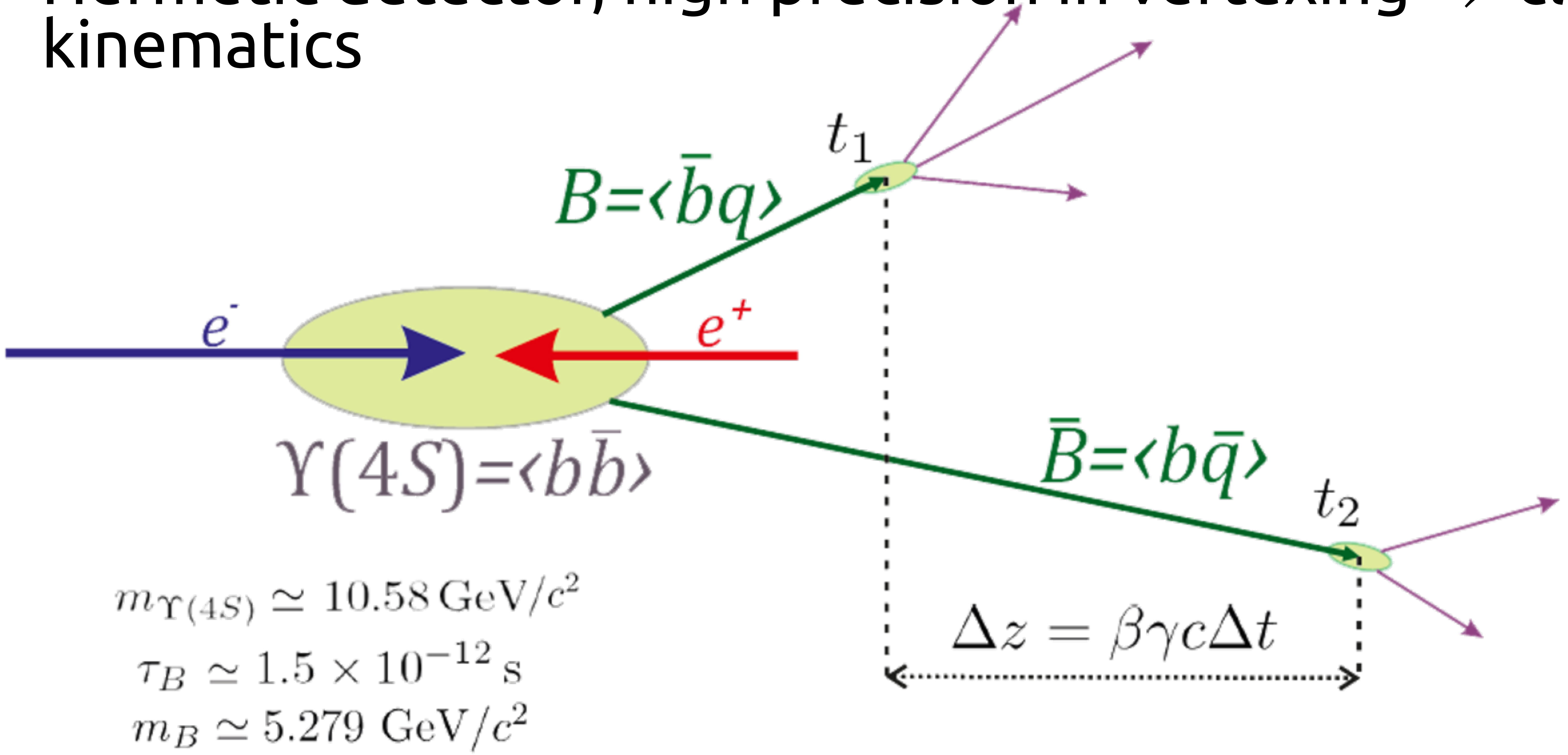
Outline

- SuperKEKB and Belle II experiment
- High Level Trigger (HLT) structure and data flow
- Limitations and first upgrade
- Further upgrade: Tracking optimization
- Far future: a tidier and faster HLT for the target luminosity

B-Factory idea

- Asymmetric collider e^+e^-, $E_{cm} = m(\Upsilon(4S)) = 10.58\ \text{GeV}$ $\Rightarrow$ coherent $B\overline{B}$ pairs
- Boost of center-of-mass ($\beta\gamma = 0.28$)$\Rightarrow$ measure of Δz
- High luminosity $\Rightarrow$ precision measurements
- Hermetic detector, high precision in vertexing $\Rightarrow$ closed kinematics



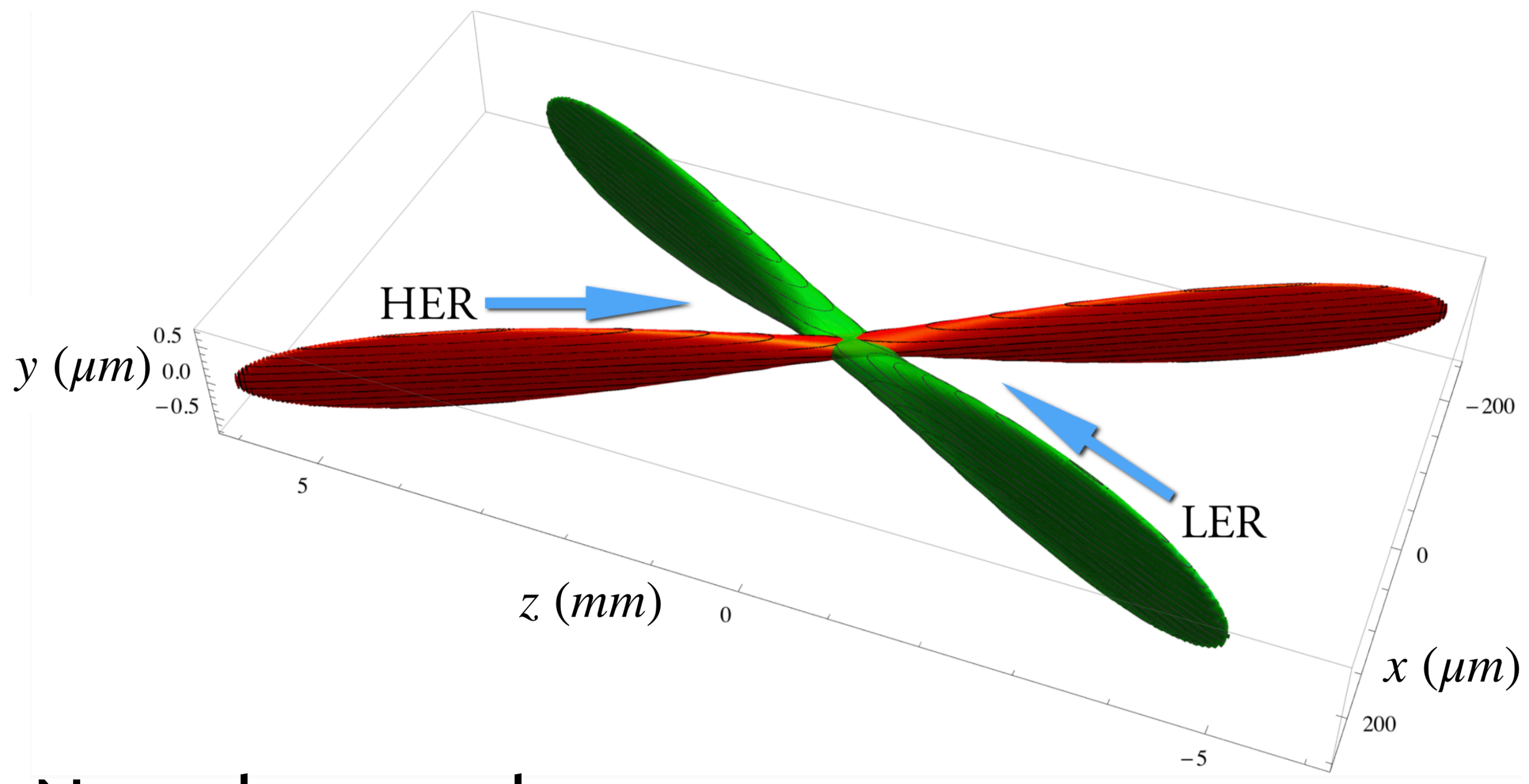


$e^+e^- \to$	Cross section [nb]
$\Upsilon(4S)$	1.05 ± 0.10
$c\bar{c}$	1.30
$s\bar{s}$	0.38
$u\bar{u}$	1.61
$d\bar{d}$	0.40
$\tau^+\tau^-(\gamma)$	0.919
$\mu^+\mu^-(\gamma)$	1.148
$e^+e^-(\gamma)$	300 ± 3

Belle II experiment at SuperKEKB collider

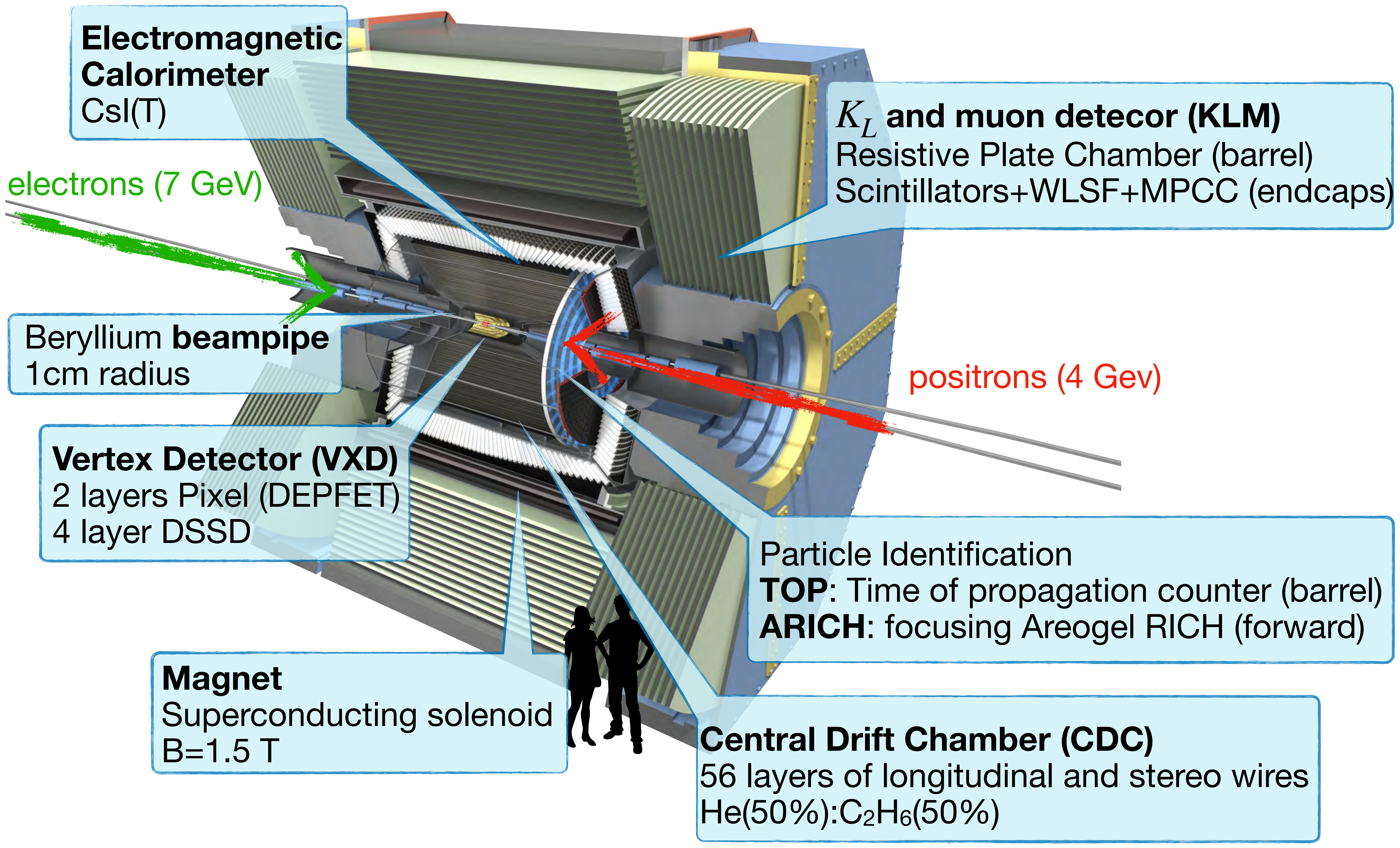
SuperKEKB

- Successor of KEKB (1999-2010, KEK, Japan)
- Target peak luminosity: $6 \cdot 10^{35}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$ (x 30 of KEKB)
- Target integrated luminosity: $50\ \mathrm{ab}^{-1}$ (x 70 Belle at $\Upsilon(4S)$)



Nano-beam scheme:
250 μm (Z) × 10 μm (X) × 50 nm (Y)

Belle II



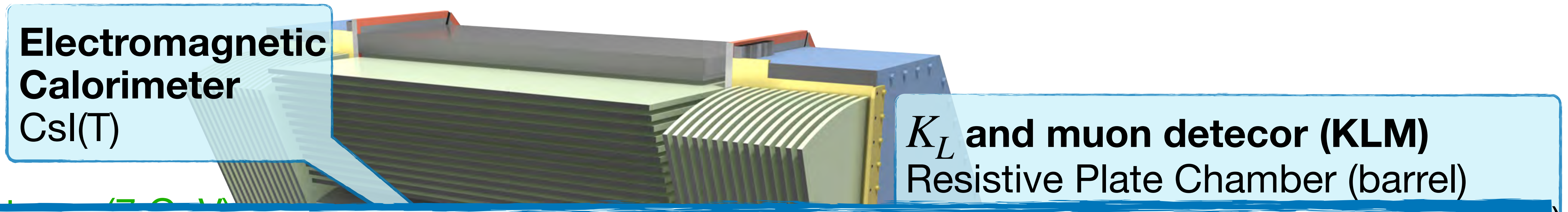
[Belle II Technical Design Report, arXiv:1011.0352]

Belle II experiment at SuperKEKB collider

SuperKEKB

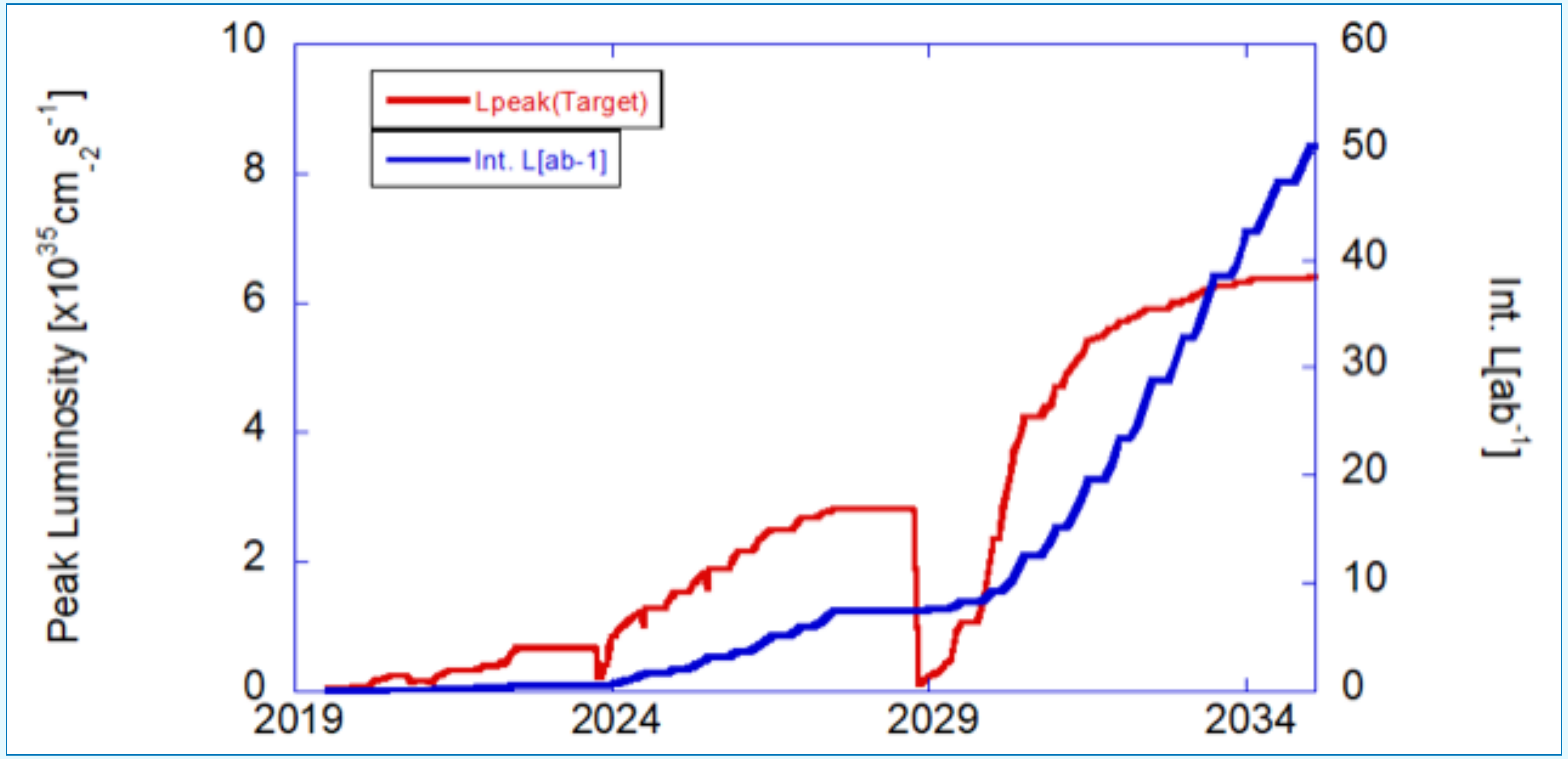
- Successor of KEKB (1999-2010, KEK, Japan)

Belle II

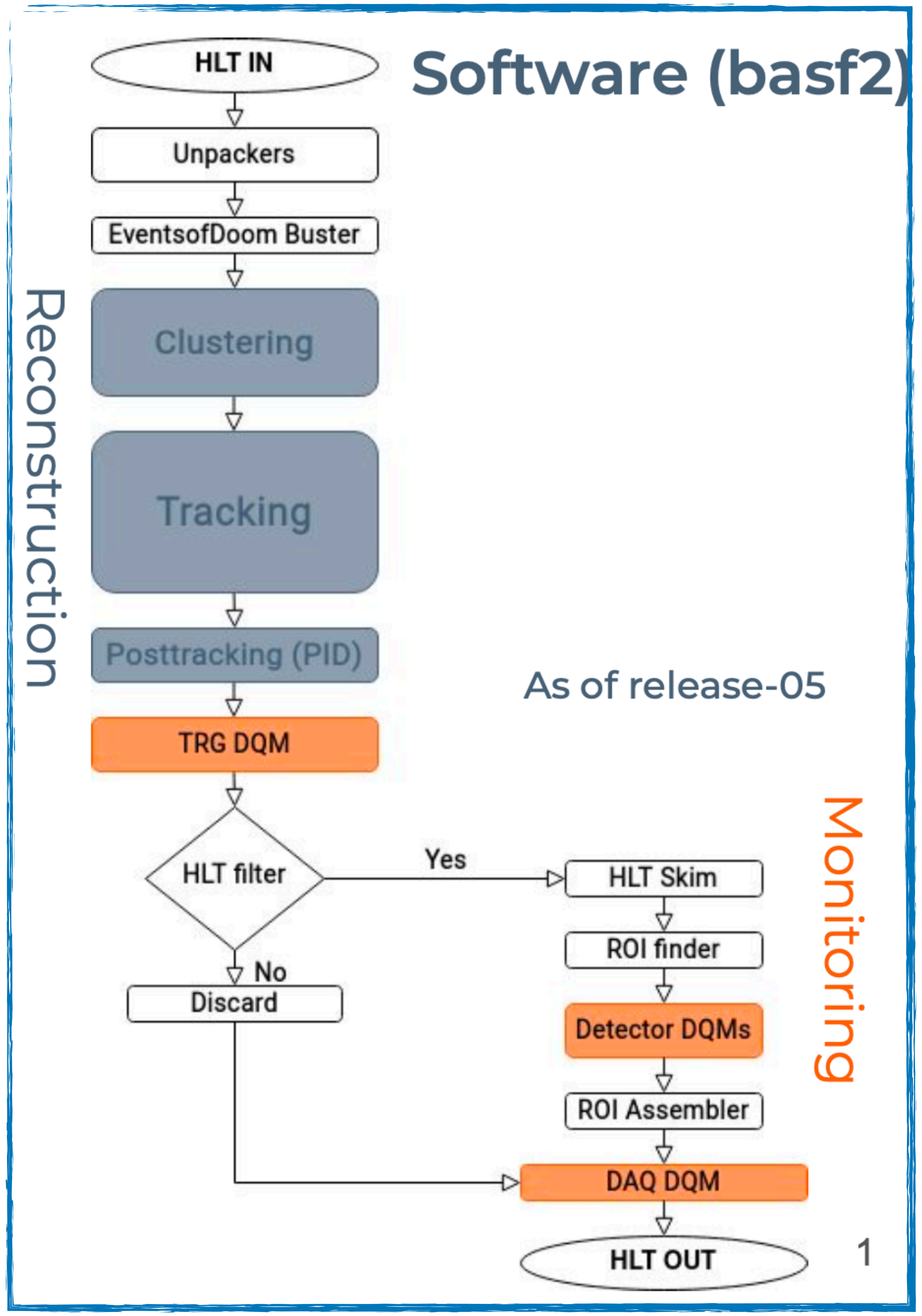


Current Status

- complete detector data taking started in 2019
- Current peak luminosity $4.7 \cdot 10^{34}\ \mathrm{cm^{-2}s^{-1}}$ (reached the 22/06/2022)
- current integrated luminosity: $\sim 424\ \mathrm{fb}^{-1}$ (~Babar~0.5 Belle)
- Long Shutdown 1 (LS1) started in July for several upgrades (beam pipe, pixel, TOP PMT)



HLT software



[by Vidya Vobbilisetti]

SVD Time

- How
  average of cluster time of all cluster associated to tracks in the event

- When
  estimation performed after clustering, within SVD track finding

- Performance
  efficiency=99.8% (higher than CDC)
  resolution=1ns (as CDC)
  Time consumption=0.015 ms/event (2000 times better than CDC)

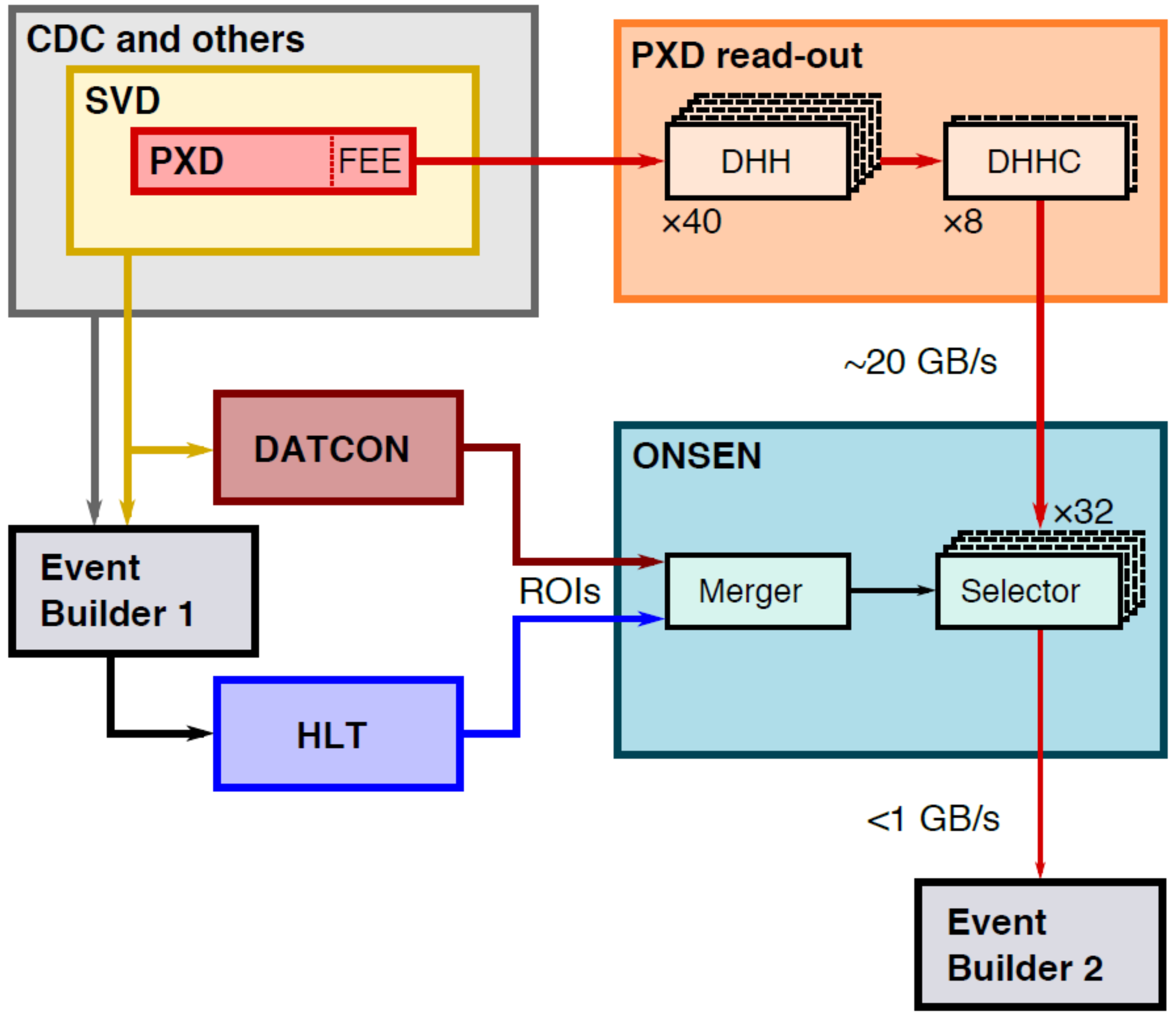
PXD and ROI

L1 trigger numbers

- input rate: 250 MHz (4ns)
- latency: few μs
- Output rate: 10-30 kHz

- Redout time of all PXD sensors: 20 $\mu s \Rightarrow$ too slow for L1
- full PXD output rate: 20 GB/s (with zero-suppression applied) $\Rightarrow$ too big for the bandwidth

- PXD saved on the ONSEN: FPGA system to collect and temporary store PXD data
- HLT takes the decision and cut events in the ONSEN: **x3 data reduction**
- HLT evaluates ROIs (Region Of Interest) on the PXD layers, using CDC+SVD tracks: **x10 data reduction**
- Event builder 2 merging HLT and PXD data



Demonstrative optimization

Disclaimer: *simply a good setup after few days of tests, not really optimized!*

Adopted criteria: *obtain reasonable* ***convergence behaviour*** *(use the iteration range, use mainly primary convergence criterion, exploit more wisely the p-value), with the CPU* ***time figure of merit***

Parameter	BASELINE VALUE	NEW VALUE
Δw	0.001	0.1
Δp	1	0.001
Prob cut	0.001	0.001
Max failed hits	5	5
(Tmax, Tmin, Niter)	(100, 0.1, 5)	(2, 0.01, 5)
Min Iterations	Niter (5)	1
Min iteration for pval check	MinIter (5)	MinIter (1)
Max iterations	Niter+4 (9)	Niter+4 (9)

Optimization approach:

- increase Δw threshold ⇒ **allow convergence**
- decrease Δp threshold ⇒ **avoid automatic convergence**
- Change min-max iterations ⇒ **reduce average number of iterations**
- Charge annealing scheme (lowering initial T) ⇒ **avoid discarding weights** ⇒ "speed-up" the convergence

Convergence behaviour after optimization:

- Convergence spread between iteration 2 and 6 (peak at 4)
- convergence by pval about 10% of the time

DAF general features

- For each call of the fitter the DAF (*Deterministic Annealing Filter*) is called
- The purpose of the DAF is to remove from the fit the **outlier hits** to improve the fit accuracy
  - removing of beam bkg hits
  - removing of hits from other tracks
  - removing of δ-rays
  - removing/fix of wrong L/R CDC hit assigment
- The DAF is **assigning weights** (in the range [0,1]) to each hit, accordingly to the residuals between the measurement and the Kalman Filter prediction.
- The fit is r**epeated multiples times lowering a temperature** parameter T
  - high T --> softer assignment , weights tend to move to 0.5
  - low T --> harder assignment, weights tend to be 1 or 0
- A **convergence criterion** is defined, based on the variation of the weights and the p-value of the fit (see next slide)

DAF convergence criterion

1. if $\max_{i \in \text{track}} (|w_{j-1} - w_j|) < \Delta w = 10^{-3}$, where i=hits, j=iterations

2. if $j > N_{\min} = 5$ and $|p_{j-1} - p_j| < \Delta p = 1$, where p=p-value of the fit

3. if $j > N_{\max} = 9$

Additional parameters which regulate convergence:

- The **annealing temperature** is lowered from $T_{max} = 100$ to $T_{min} = 0.1$ in N_{min} steps (T is constant in the iterations $[N_{min}, N_{max}]$)
- a probability cut $P = 10^{-3}$ regulate a damping factor of the weights, to force them to be 0 if their value is below a threshold.

DAF behaviour

1. $\max_{i \in \text{track}} (|w_{j-1} - w_j|) < \Delta w = 10^{-3}$
2. $j > N_{\min} = 5$ and $|p_{j-1} - p_j| < \Delta p = 1$
3. $j > N_{\max} = 9$

- The criterion 1 (weights) is never satisfied
- The criterion 2 is immediately satisfied as soon as checked
- The **DAF always* run 5 iterations**
- * = sometimes (<0.1%) the pvalue==0, in that case additional iterations are run

There are room for optimize the DAF iterations!

- NB: the **CPU time is ~proportional** to number of iterations of the DAF
- Some examples (single mass hypothesis):
  - $N_{min} = 2$: 6.5 ms (7.7 ms) w/o SVD (with SVD)
  - $N_{min} = 5$: 10.4 ms (15.5 ms) w/o SVD (with SVD)

DAF convergence visualization

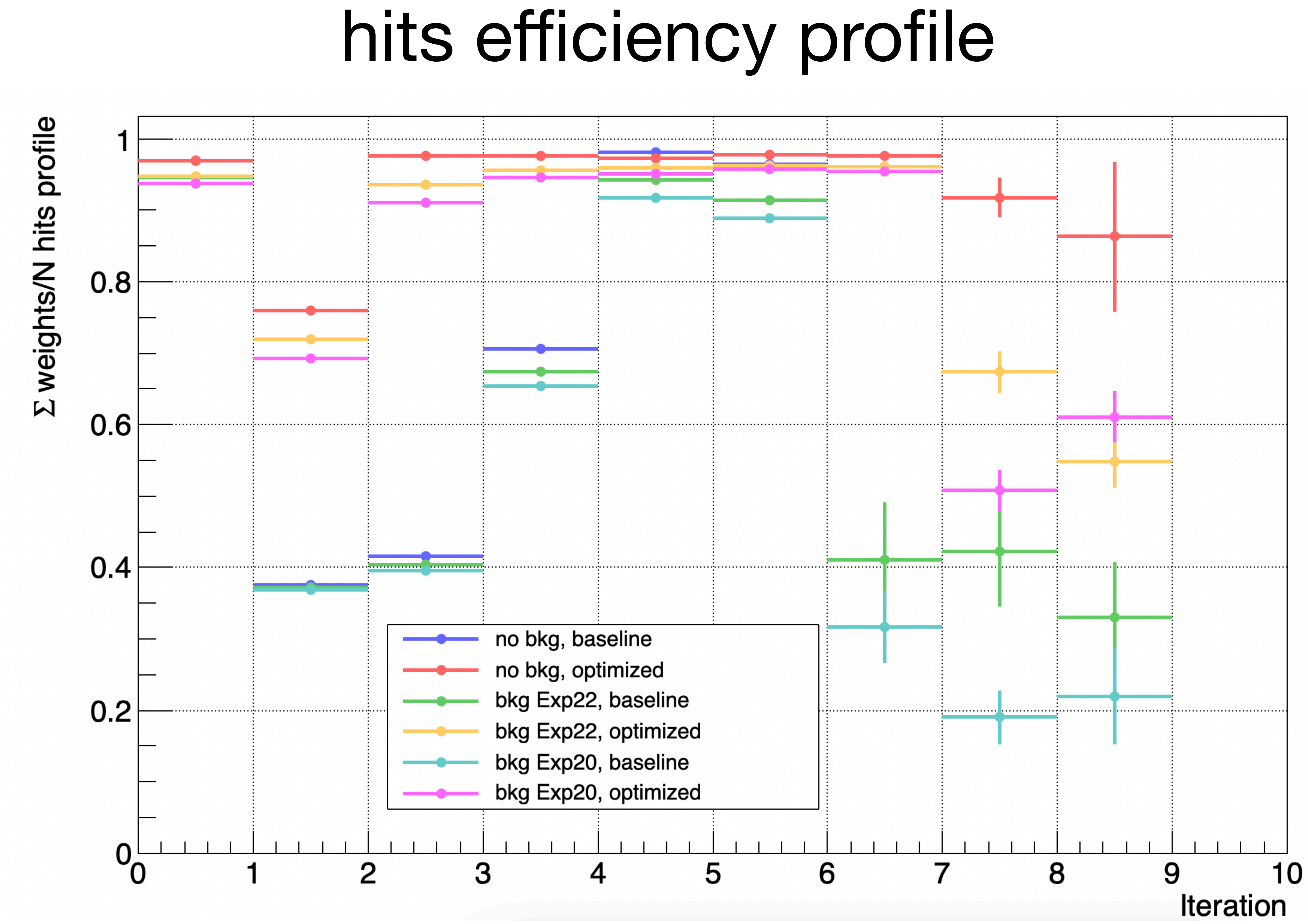
Test condition:
- CDC only tracking
- Muon Gun
- pT=1 GeV
- $\theta = 70^{\circ}$

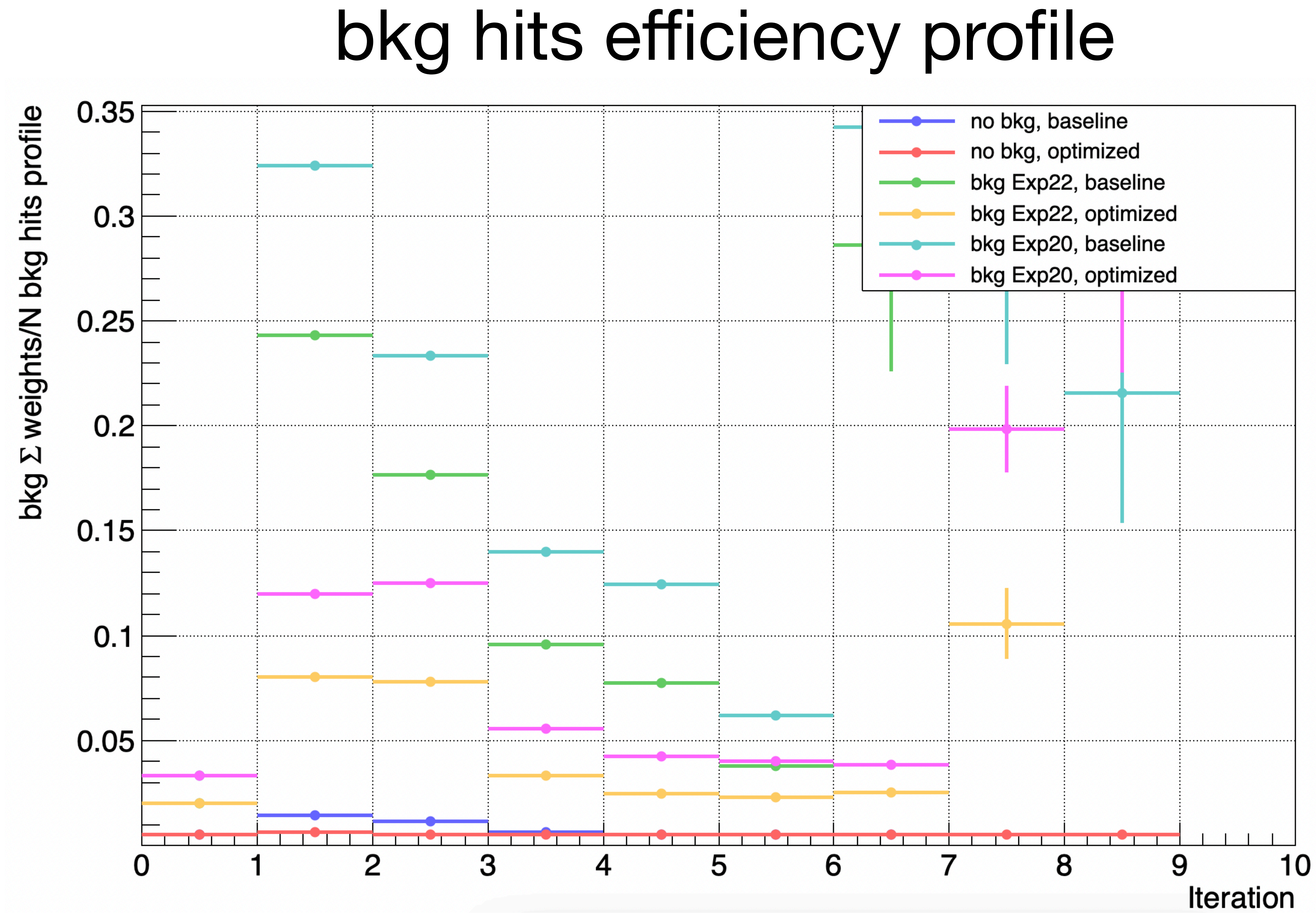
The Faster convergence is visible from the weight evolution

Optimization performance (DAF 1/4)

Test condition:

- CDC only tracking
- Muon Gun
- pT=1 GeV
- $\theta = 70^{\circ}$



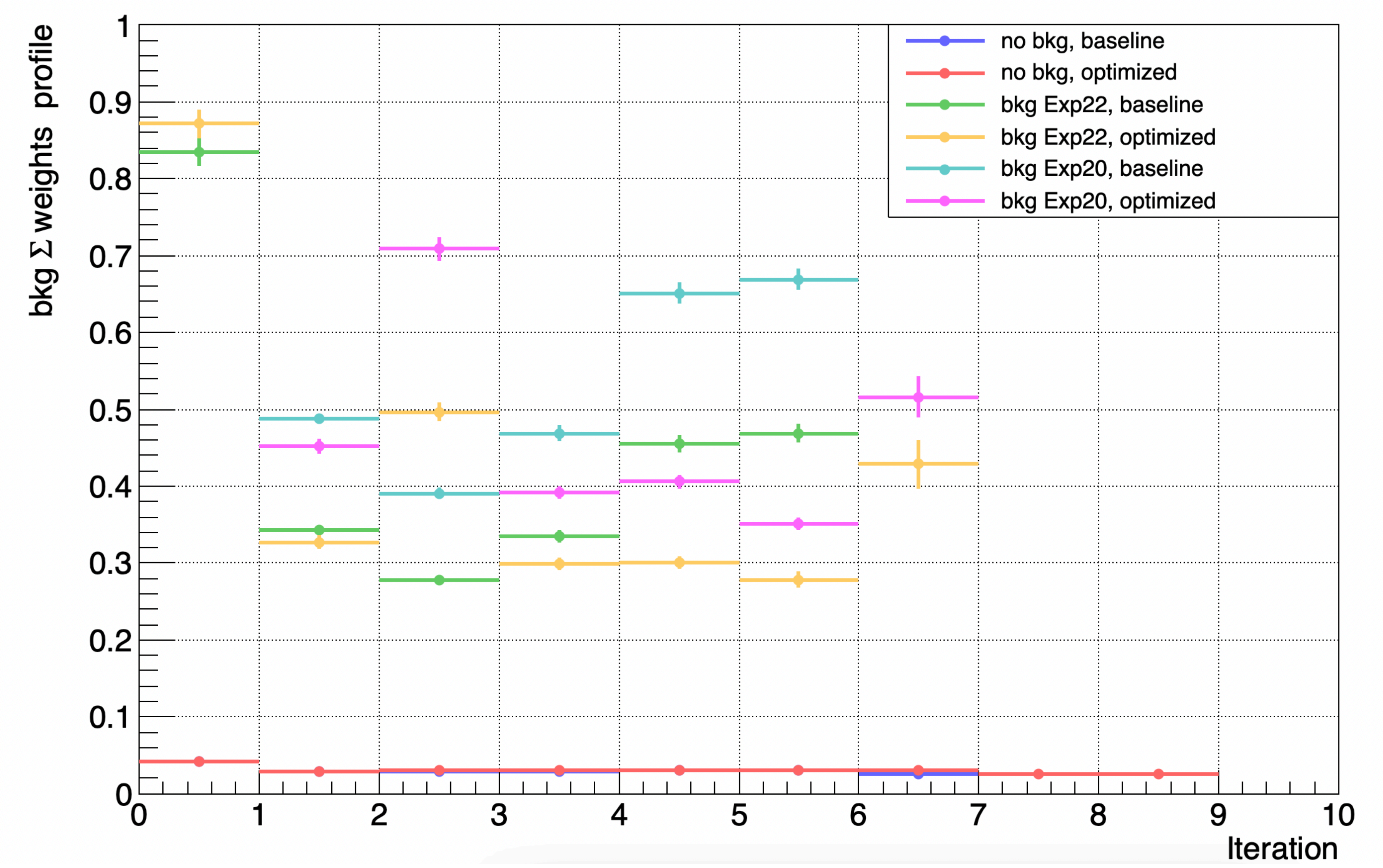


At the end of the day, about the same bkg rejection performance

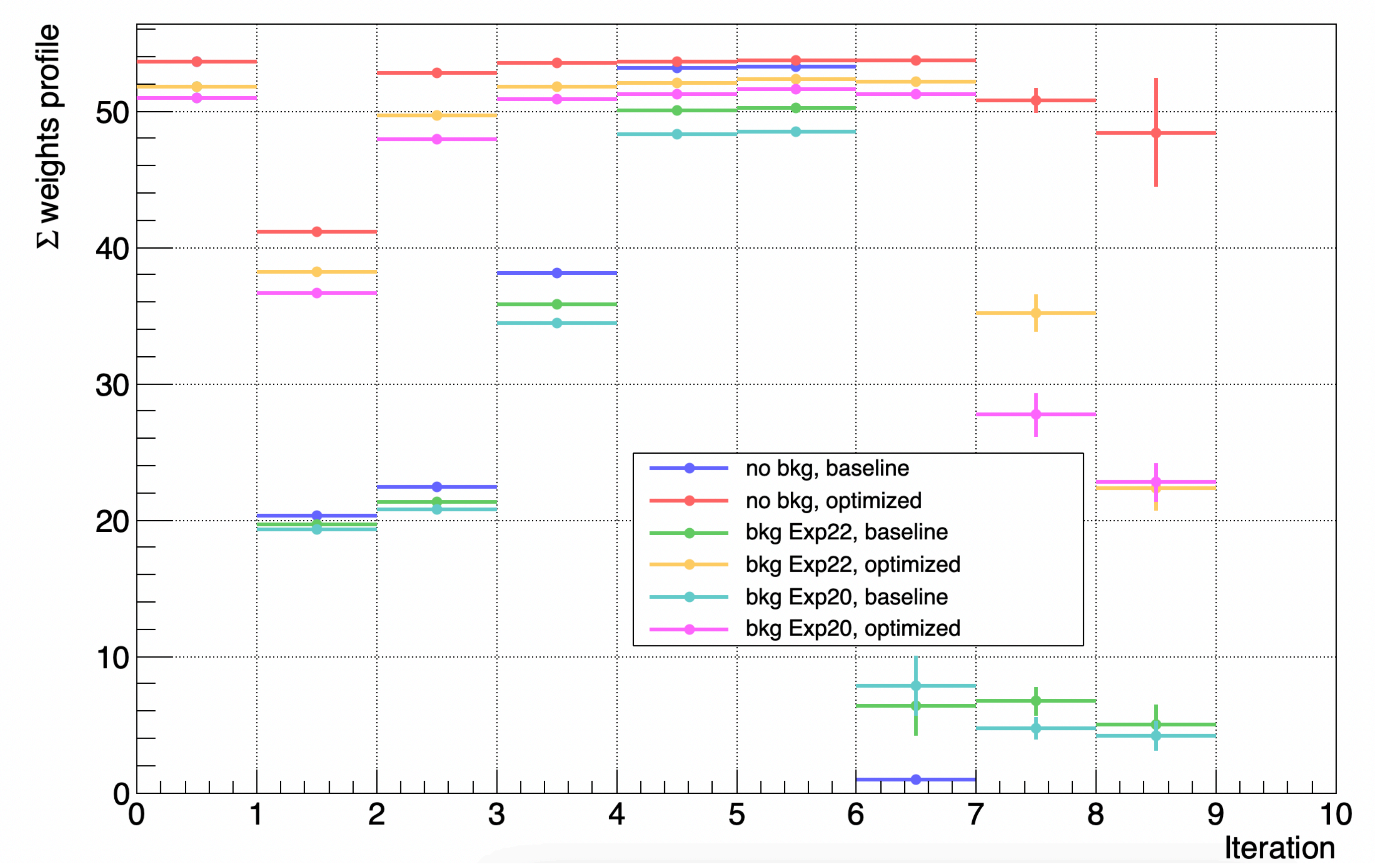
Clear (good) effect of reduced Temperature

Optimization performance (DAF 2/4)

<u>Test condition:</u>
- CDC only tracking
- Muon Gun
- pT=1 GeV
- $\theta = 70^{\circ}$

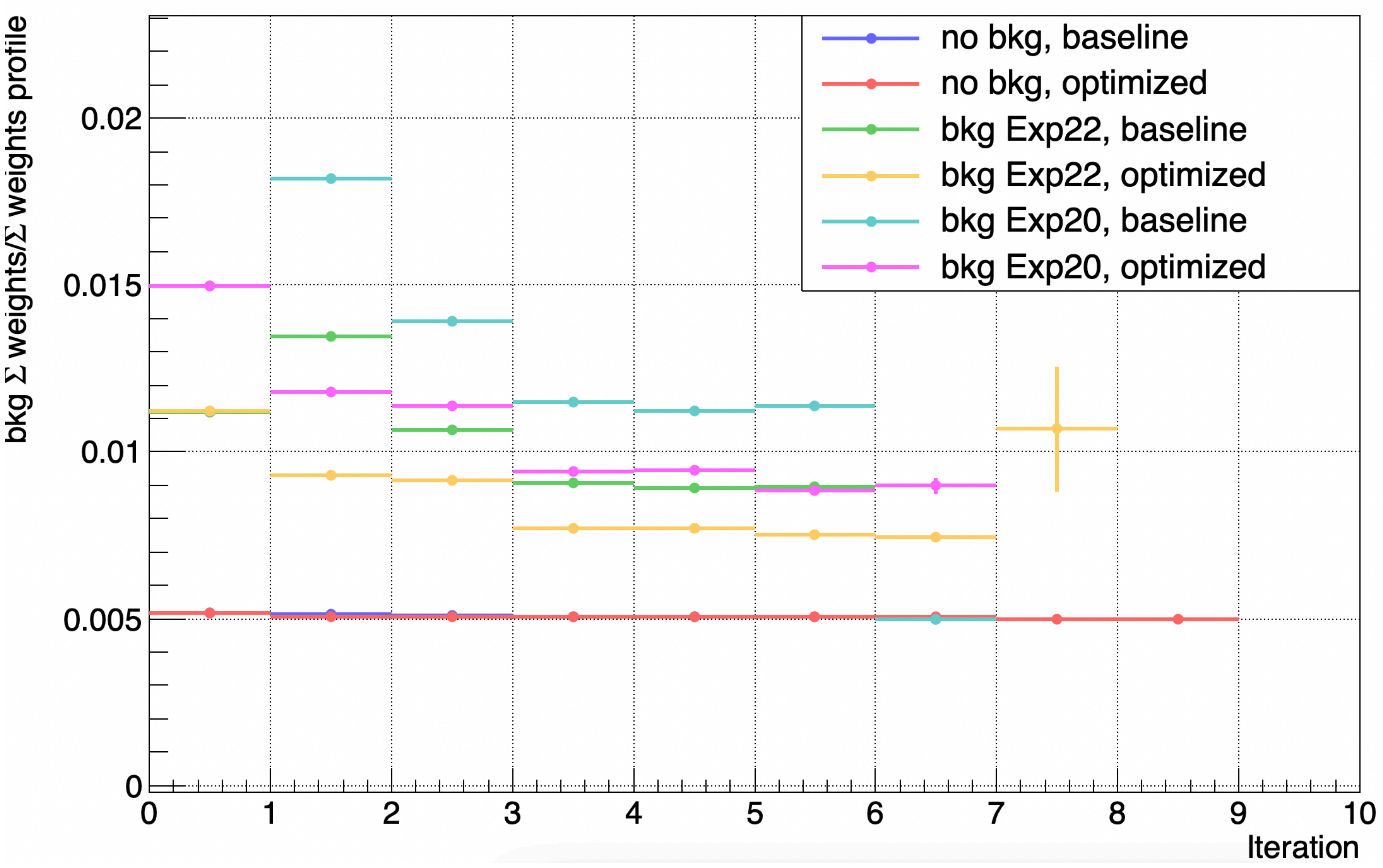


—



=

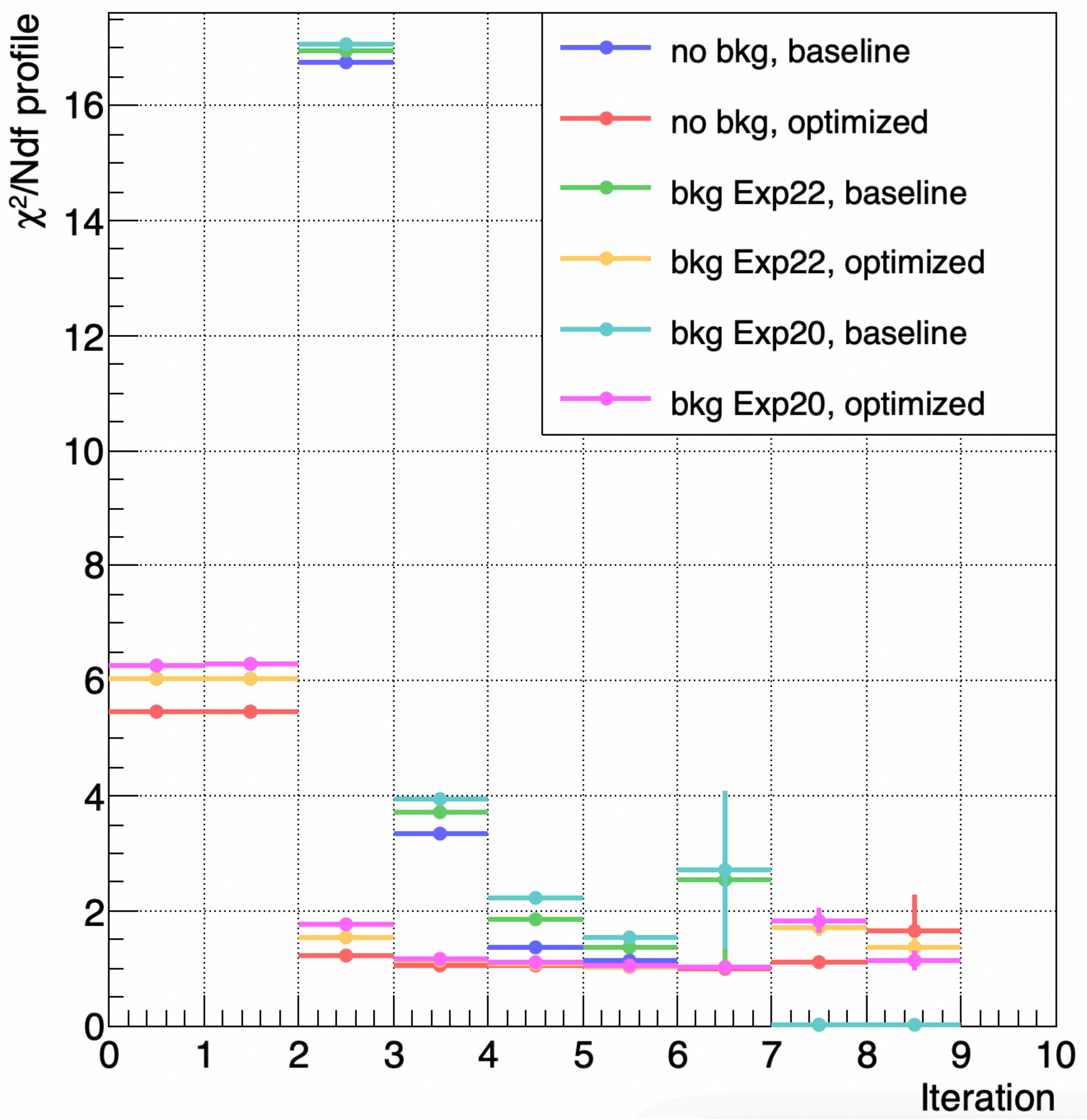
un-purity

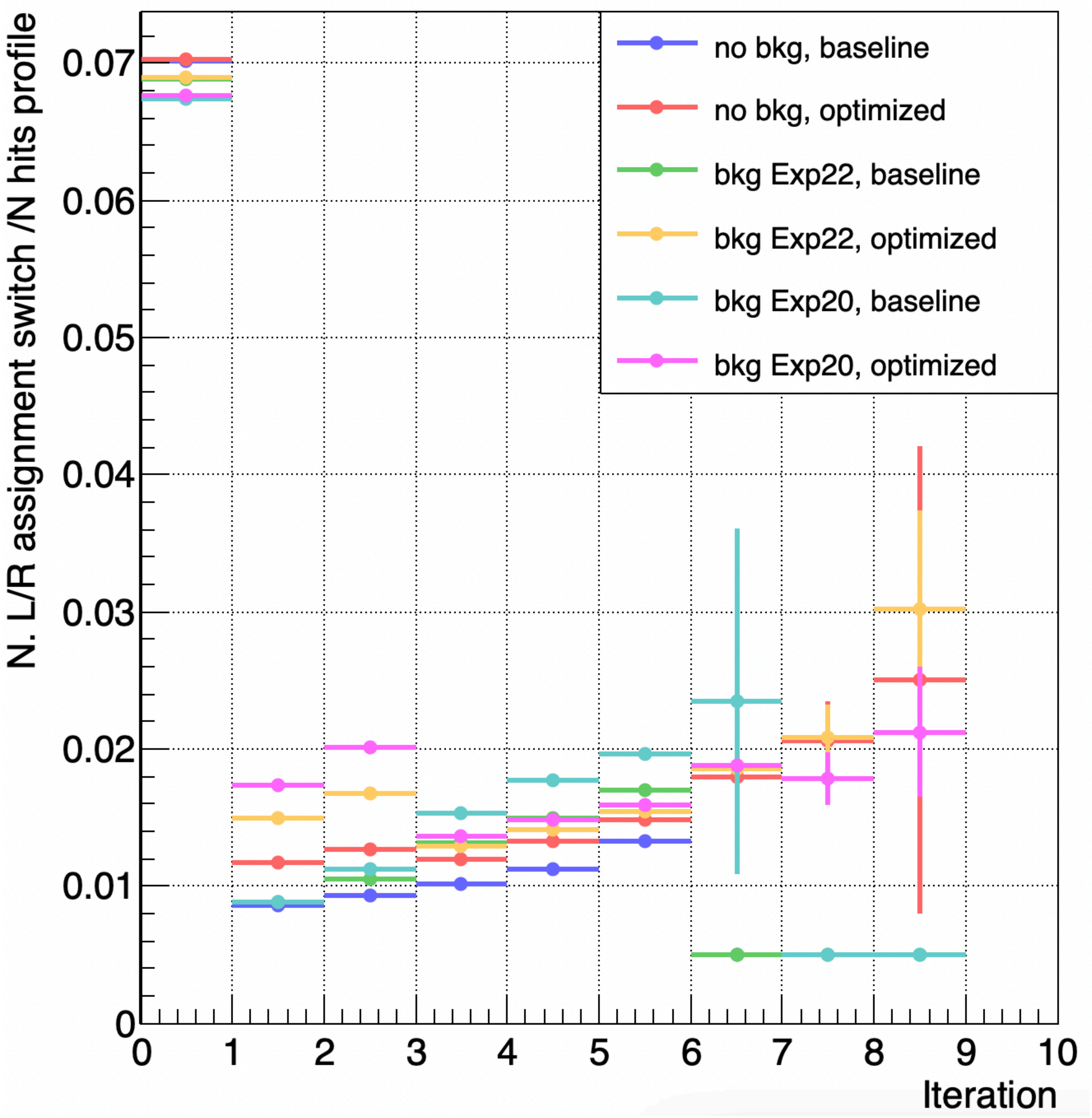


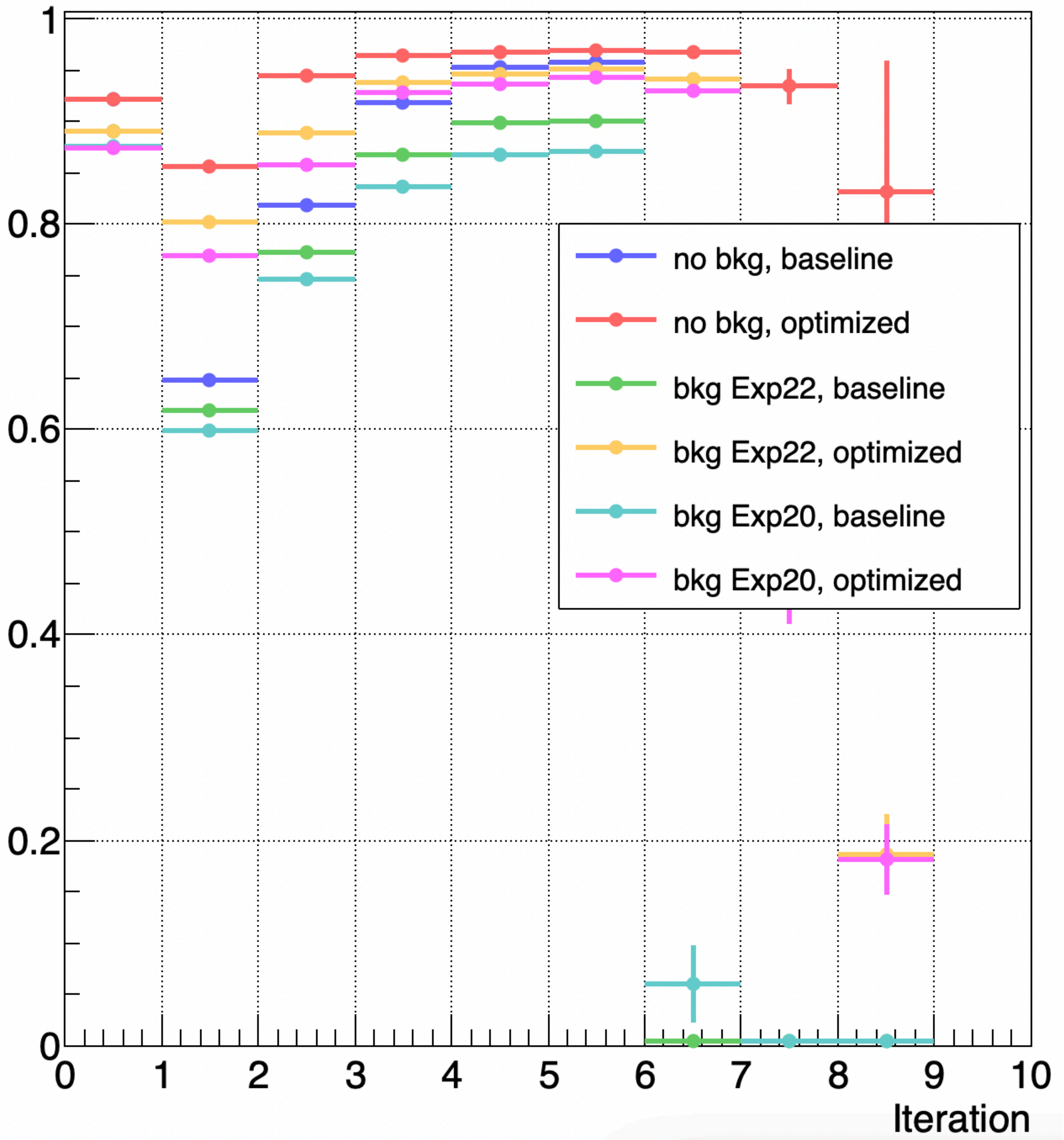
Optimization performance (DAF 3/4)

<u>Test condition:</u>

- CDC only tracking
- Muon Gun
- pT=1 GeV
- $\theta = 70^{\circ}$



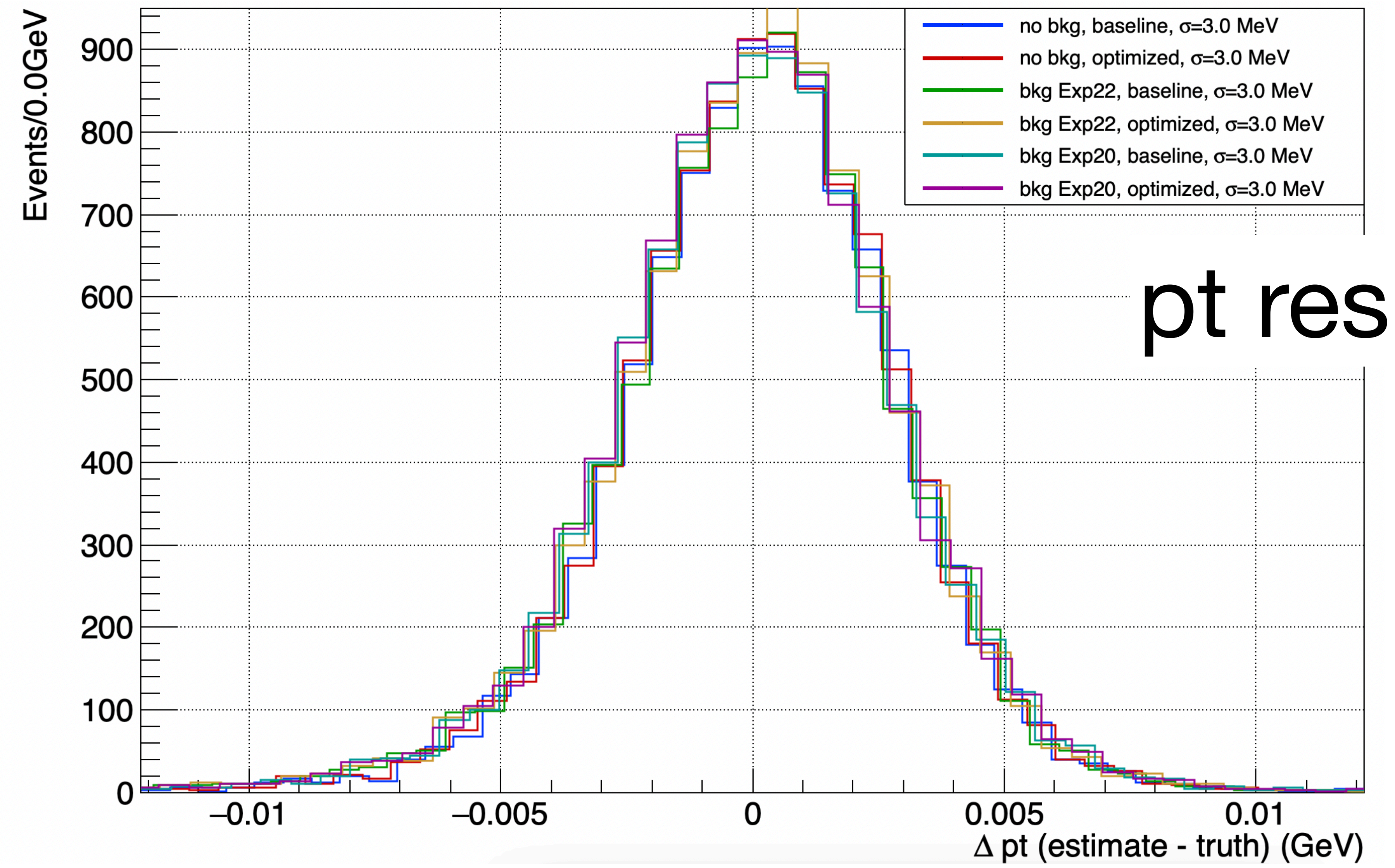


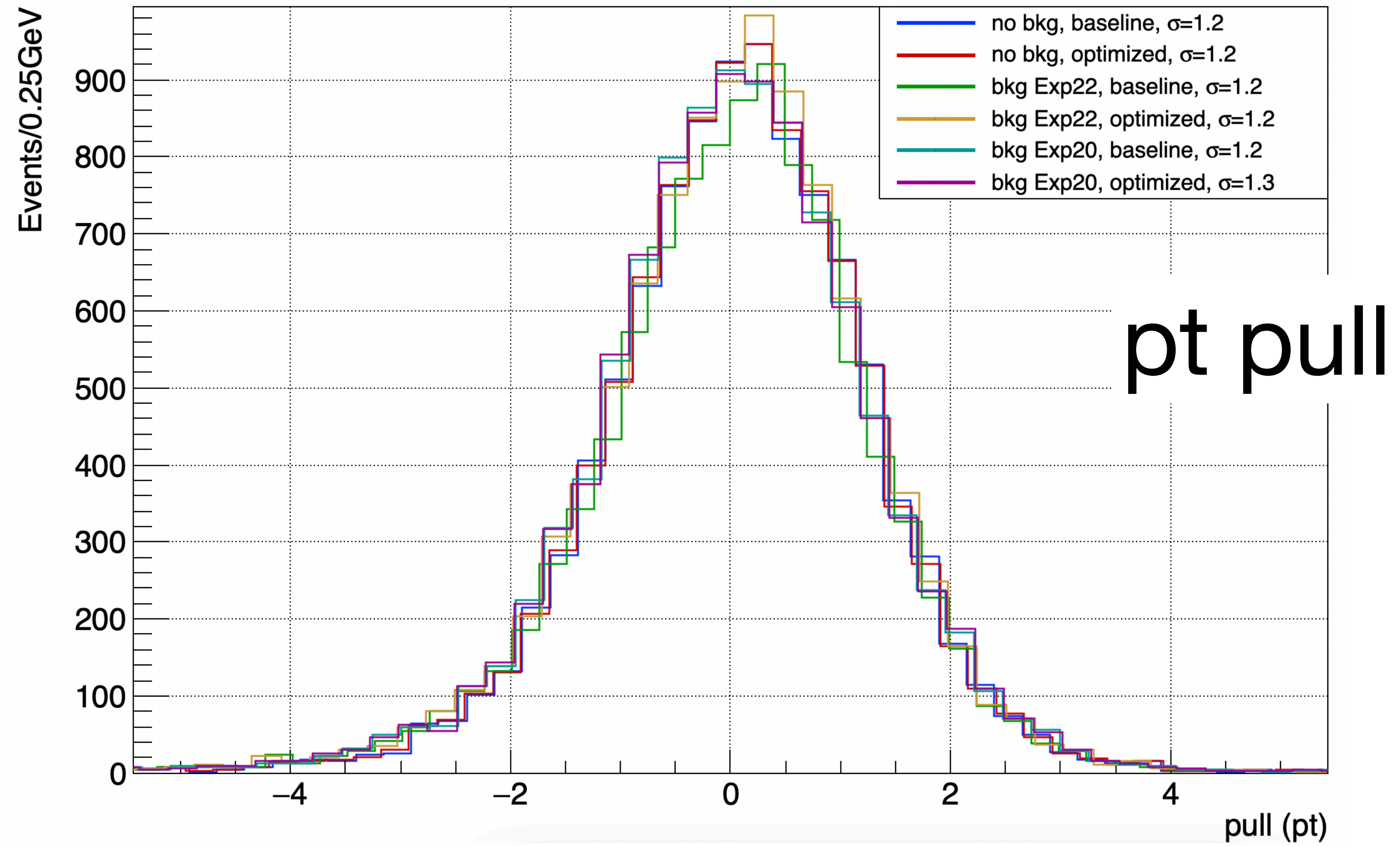


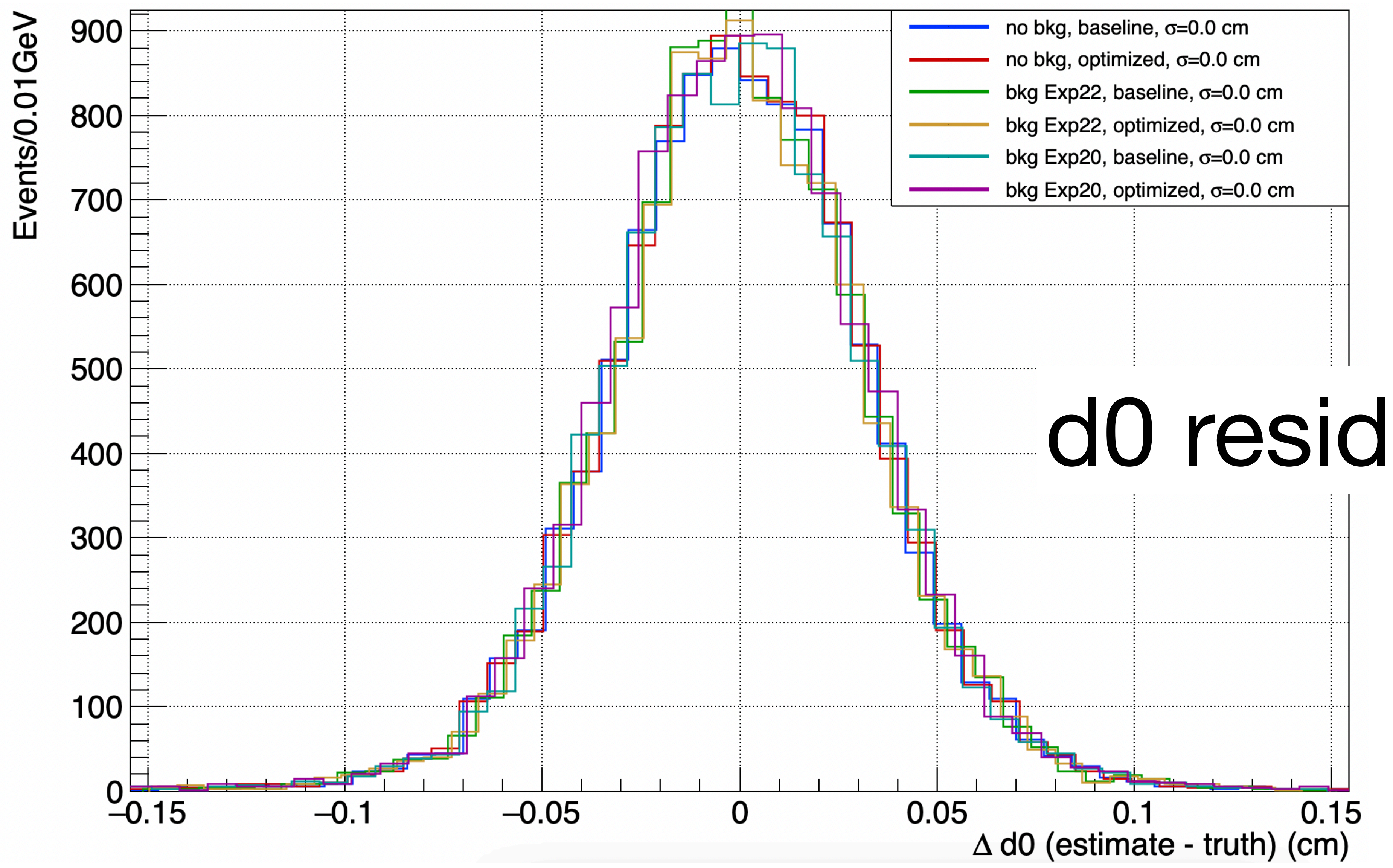
At the end of the day, about the same L/R performance

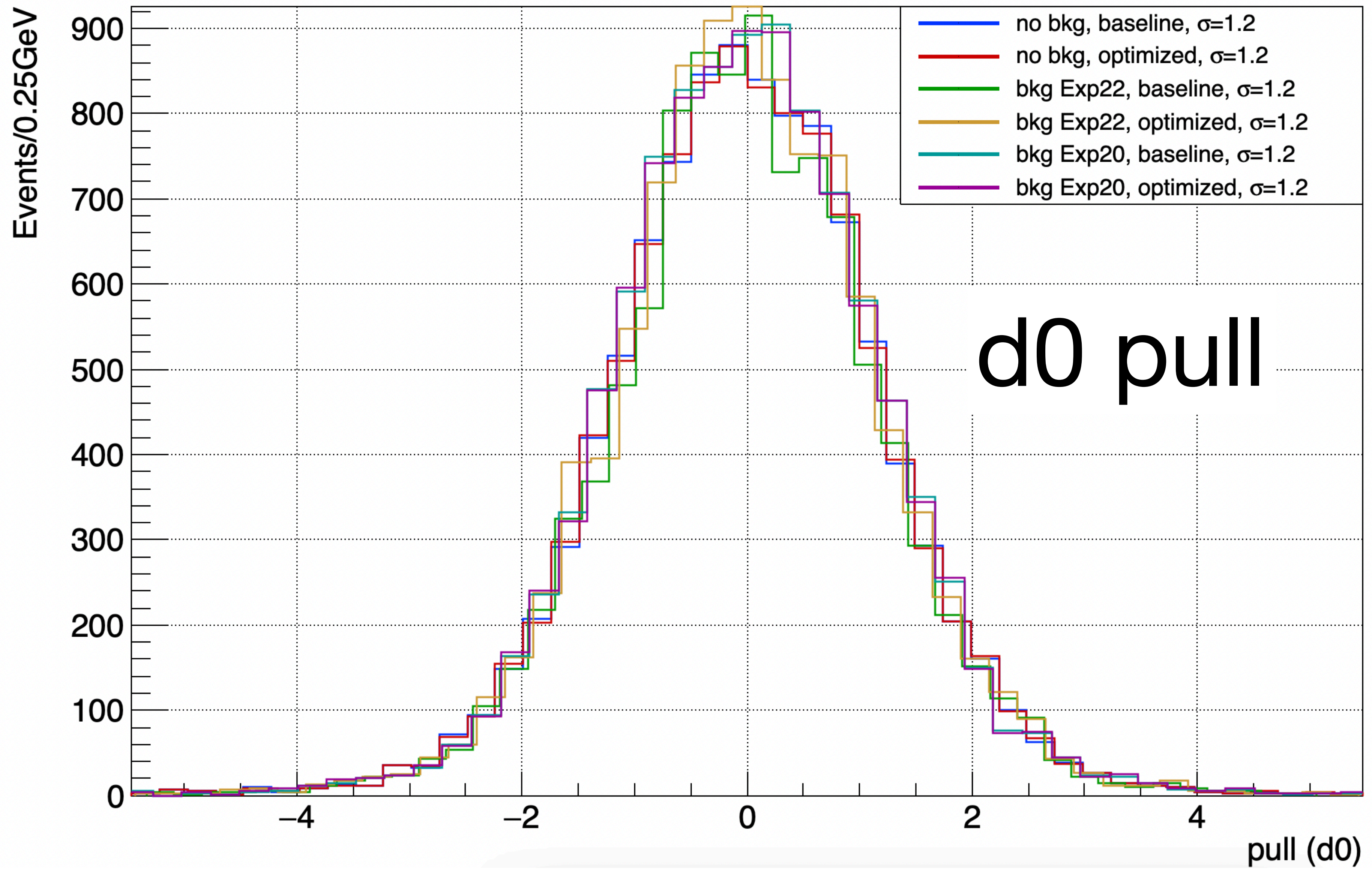
Optimization performace (tracking 1/2)

Test condition:
- CDC only tracking
- Muon Gun
- pT=1 GeV



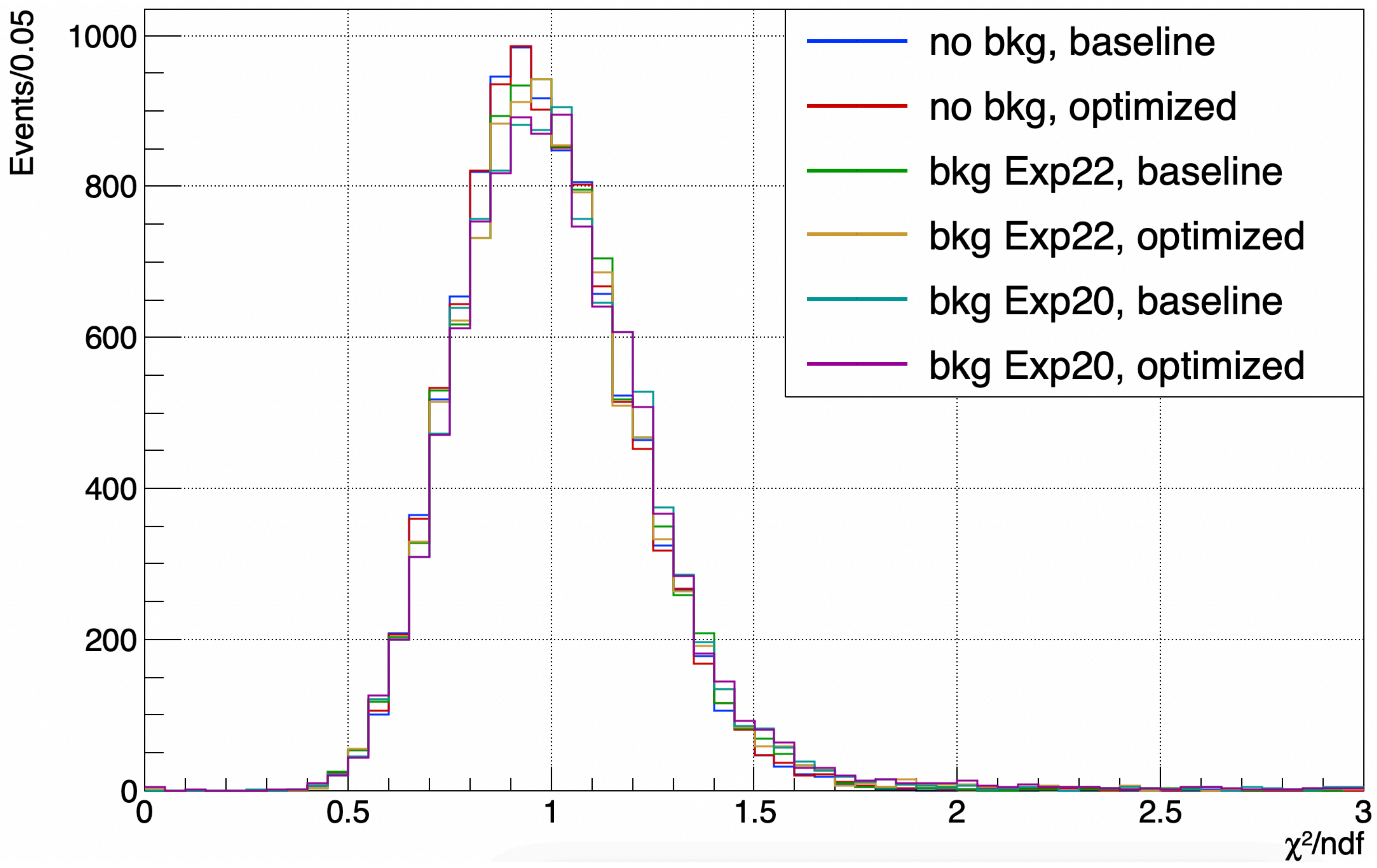


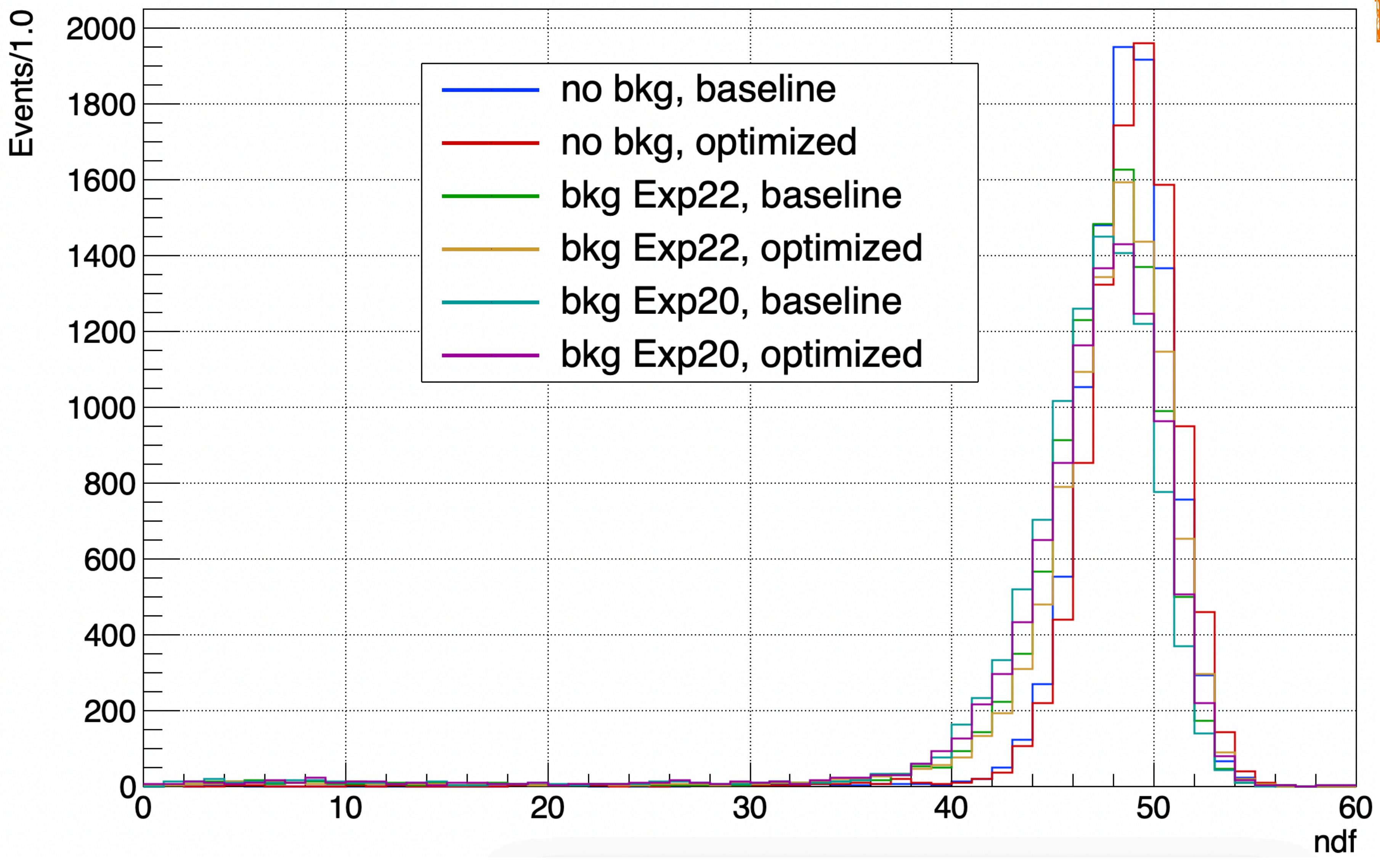


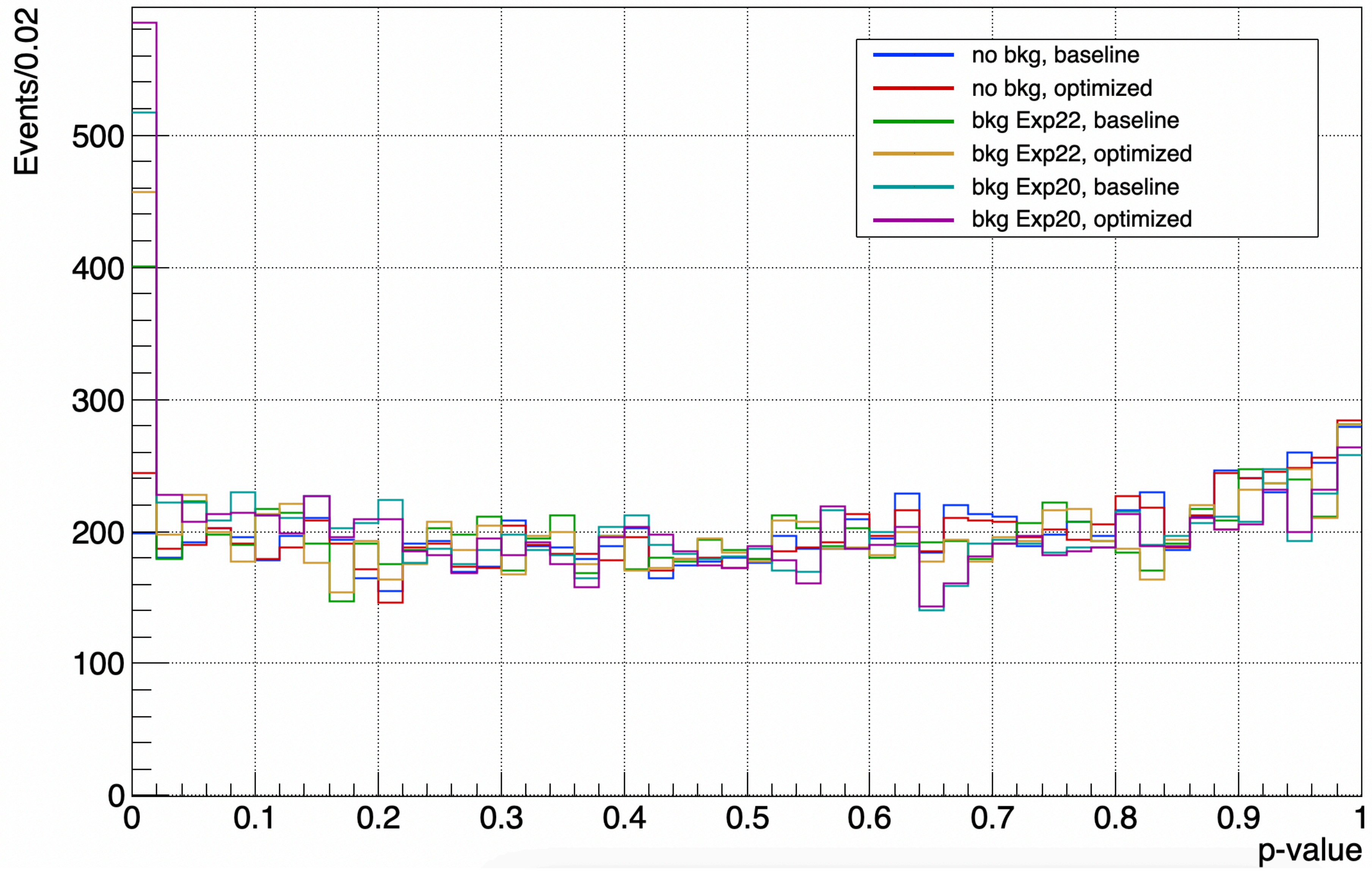


Optimization performace (tracking 2/2)

Test condition:
- CDC only tracking
- Muon Gun
- pT=1 GeV
- $\theta = 70^{\circ}$







Timing performance is different condition

	Combined_DAFRecoFitter (ms/ev)
DAF (no bkg)	10.5
DAF (exp20 bkg)	10.3
fit without DAF (no bkg)	2.2
fit without DAF (exp20 bkg)	2.3