Low-Latency On-Chip τ Event Selection with Machine Learning for the Belle II Level-1 Trigger

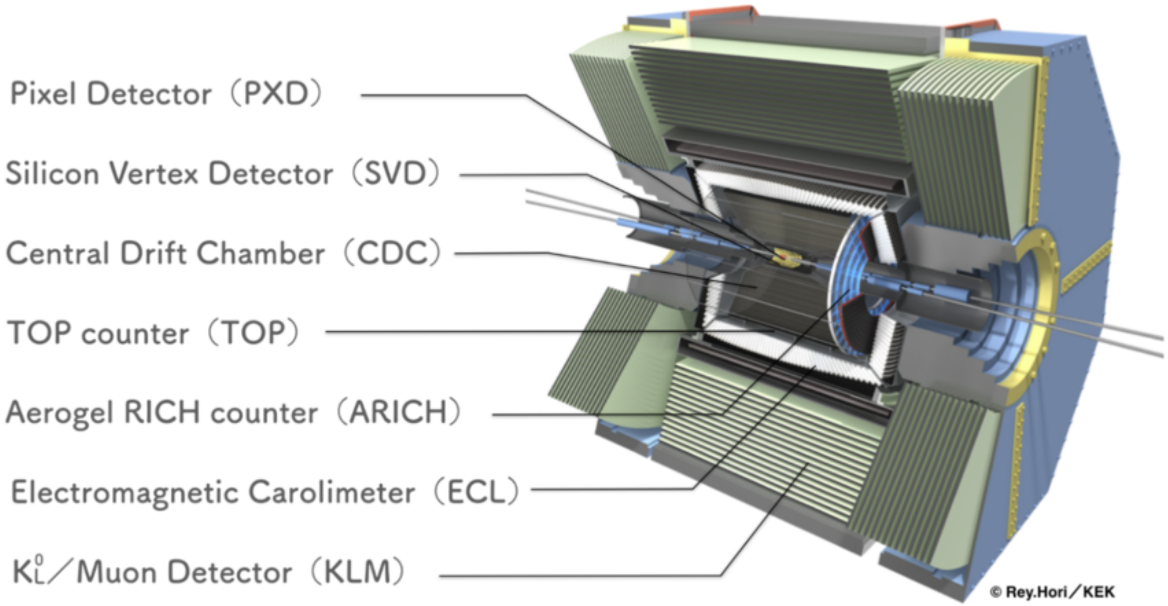
Deven Misra[1,2], Taichiro Koga[3,4], Yu Nakazawa[3,4], and Takeo Higuchi[1,2]

[1]*The University of Tokyo*, [2]*Kavli IPMU*, [3]*KEK*, [4]*SOKENDAI*

May 20, 2025

Belle II

Belle II is the second-generation flavor physics experiment (“B-Factory”) located at the SuperKEKB accelerator in Tsukuba, Japan.

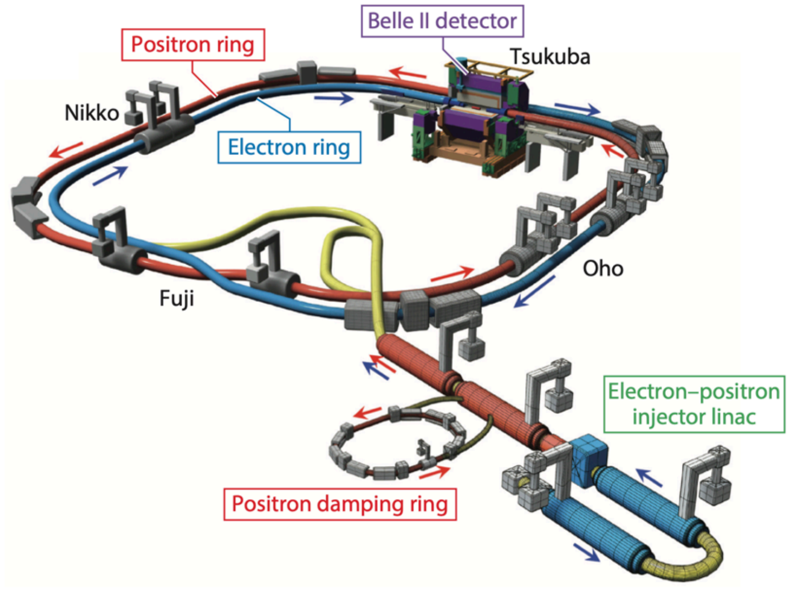


The Belle II Detector

Data taking began in early 2018, and operation is expected to continue until at least ~2040.

SuperKEKB

- Asymmetric e^+e^- collider
- Center-of-mass energy corresponding to the $\Upsilon(4S)$ resonance to produce $B\bar{B}$ pairs
  - → $E_{e^+} = 4\ \mathrm{GeV}$
  - → $E_{e^-} = 7\ \mathrm{GeV}$
- SuperKEKB utilizes a unique nano-beam collision scheme to achieve extremely high luminosities



The SuperKEKB Accelerator

Peak Luminosity: $5.1 \times 10^{34}\ \mathrm{cm^{-2}s^{-1}}$ (World Record, December 2024)

Target Luminosity: $6 \times 10^{35}\ \mathrm{cm^{-2}s^{-1}}$

Tau Physics at Belle II

Probes of New Physics:

- Tests of lepton flavor universality with leptonic τ decays
- Searches for CP violation in the lepton sector
- Constraints on lepton flavor violation in τ decays

Precision Measurements:

- Measurement of τ lepton properties
  → Mass, Lifetime, EDM/MDM
- Lorentz structure of charged-current weak interactions
  → Michel parameters

Tau Event Reconstruction

Our focus is on the reconstruction of low-multiplicity τ decays.

Leptonic Decays

- $\tau \to \mu\, \nu_\tau\, \overline{\nu}_\mu \quad \mathrm{BR} = 17.82\%$ ("1-prong")
- $\tau \to e\, \nu_\tau\, \overline{\nu}_e \quad \mathrm{BR} = 17.39\%$ ("1-prong")

Hadronic Decays

- $\tau \to \nu_\tau + \pi^\pm + n(\pi^0) \quad \mathrm{BR} = 49.5\%$ ("1-prong")
- $\tau \to \nu_\tau + 3\pi^\pm + n(\pi^0) \quad \mathrm{BR} = 15.2\%$ ("3-prong")

In principle, any decay mode which can be identified with offline reconstruction can also be included.

Data Acquisition and Trigger Rate

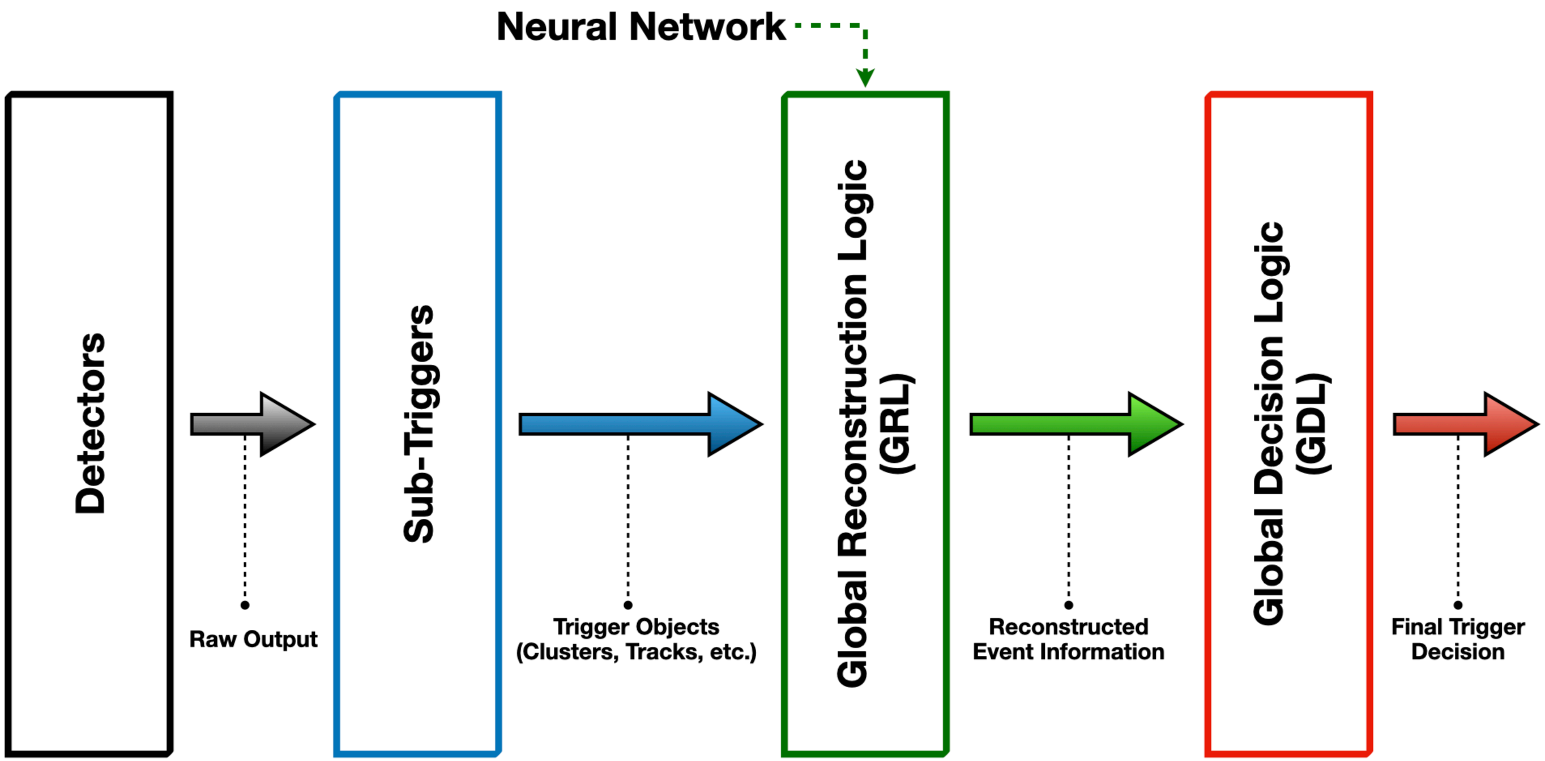
The Data Acquisition System (DAQ) is limited to a rate of ~30 kHz.

Process	Event rate	
e^+e^- bunch collision	~200 MHz	
Beam background	> ~300 kHz (2022)	
Bhabha scattering	> ~50 kHz	
Two photon processes	~10 kHz	
$e^+e^- \to \gamma\gamma$	~2 kHz	physics target
$e^+e^- \to q\bar{q}$ $(q = udsc)$	~2 kHz	
$e^+e^- \to \Upsilon(4S)$	~1 kHz	
$e^+e^- \to \mu^+\mu^-$	~0.6 kHz	~ 15 kHz
$e^+e^- \to \tau^+\tau^-$	~0.6 kHz	
dark sector/new particle	???	

Item	Requirement	Present status
Trigger rate	< 30 kHz @ $6 \times 10^{35}\mathrm{cm}^{-2}\mathrm{s}^{-1}$	~8 kHz @ $4.7 \times 10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ → reducible by increasing prescale
Latency	$4.4\ \mu s$	$4.4\ \mu s$
Event timing resolution	10 ns	~8 ns
Efficiency	> 99% for $B\bar{B}$ pair	> 99% for $B\bar{B}$ pair > 95% for $\tau^+\tau^-$ pair + low multiplicity triggers for dark sector and new physics

As luminosity increases, online trigger efficiency will need to increase to compensate. In particular, the current τ and dark sector trigger lines will require upgrades to function at the target luminosity.

Level-1 Trigger



This will be the first implementation of neural-network based real-time event selection in the Belle II trigger system.

ECL Sub-Trigger (ECLTRG)

Clustering is performed at the detector by the ECL sub-trigger. Information for the six highest-energy clusters in an event is sent to the GRL:

Energy

$\{E_1, \ldots, E_6\}$ (12 bits, LSB = 5.5 MeV)

Azimuthal Angle

$\{\theta_1, \ldots, \theta_6\}$ (7 bits, LSB = 1.6025°)

Polar Angle

$\{\phi_1, \ldots, \phi_6\}$ (8 bits, LSB = 1.6025°)

Average Cluster Time

$\{t_1, \ldots, t_6\}$ (8 bits, LSB = 1 ns)

Model Structure

Our model is a feed-forward dense neural network with three hidden layers. We utilize mixed-precision quantization-aware training to ensure the model fits within available FPGA resources.

Pre-Processing

- Typical inputs from ECLTRG are not normally distributed
- We observe significant performance gains with feature standardization
- Since our model is trained on scaled data, the same scaling needs to be implemented on-edge before input to neural network

Inputs

- $\left\{ E, \theta, \phi \right\}_{i=1}^{6}$

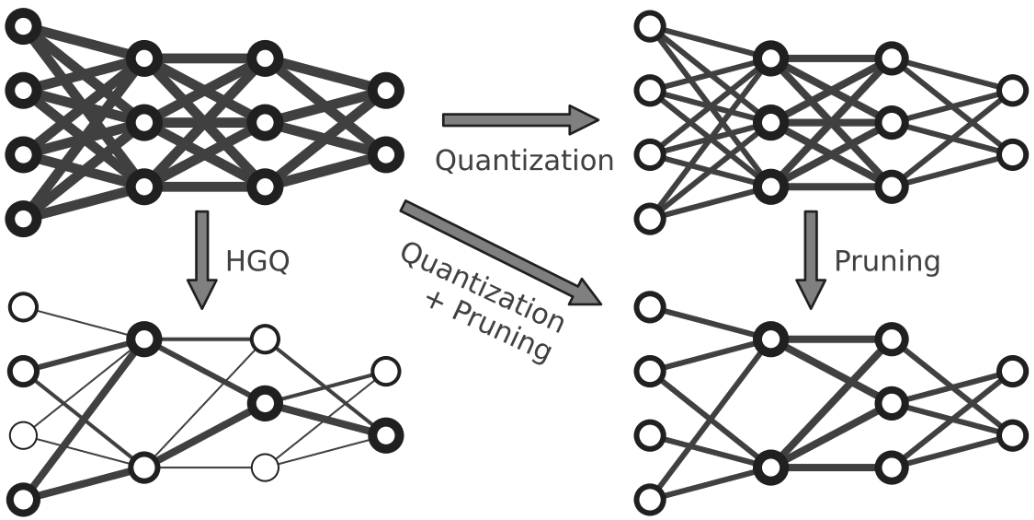
Hidden

- 64 Nodes per Layer
- Dense + ReLU Activation

Outputs

- $y \in (-5, 5)$ [Logits]

High-Granularity Quantization



Features

- Heterogeneous quantization-aware training with differentiable bitwidth
- Bit-accurate conversion to firmware with **high-level synthesis**
- Support for multiple backends (PyTorch, TensorFlow, JAX)
    - → TPU acceleration for TensorFlow and JAX

Dataset

Run Information

- Integrated Luminosity: $\mathbf{36.1\ pb}^{-1}$
- Trigger Rate: $\mathbf{11.8\ kHz}$
- Signal:Background $\approx \mathbf{1:200}$

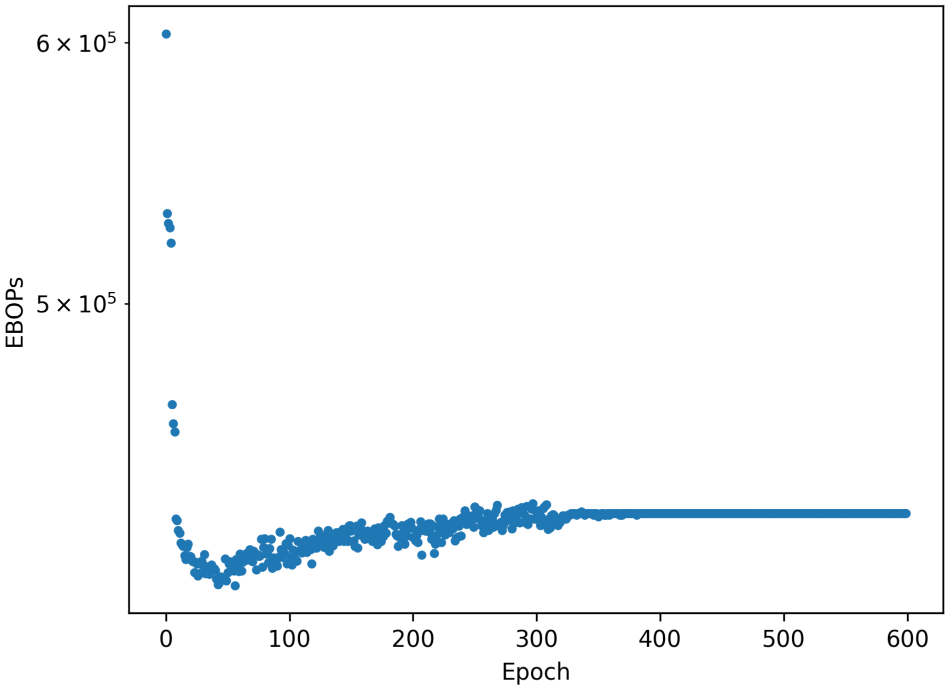
This study uses real data collected with a loose trigger condition to sample backgrounds. We use preselection cuts from offline reconstruction to identify τ decays. All other events are taken to be background, and thus we include both continuum and beam backgrounds.

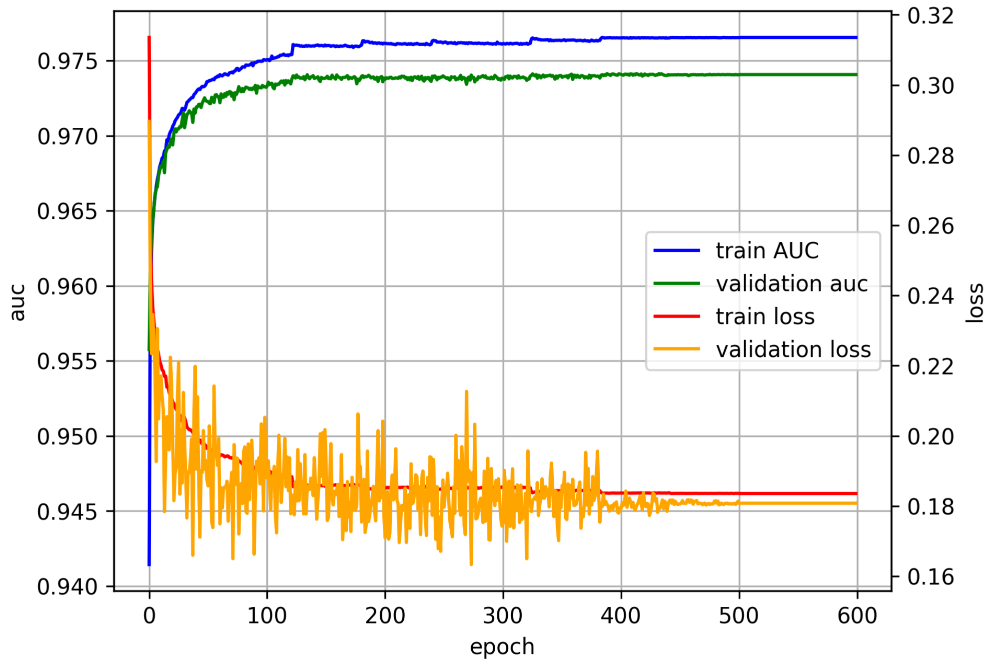
For the training and evaluation of the neural network, we split the data with

$$\mathbf{Train : Test = 0.8 : 0.2.}$$

Training

Since the parameter bitwidths are differentiable, resource utilization and latency can be optimized as part of the training cycle.

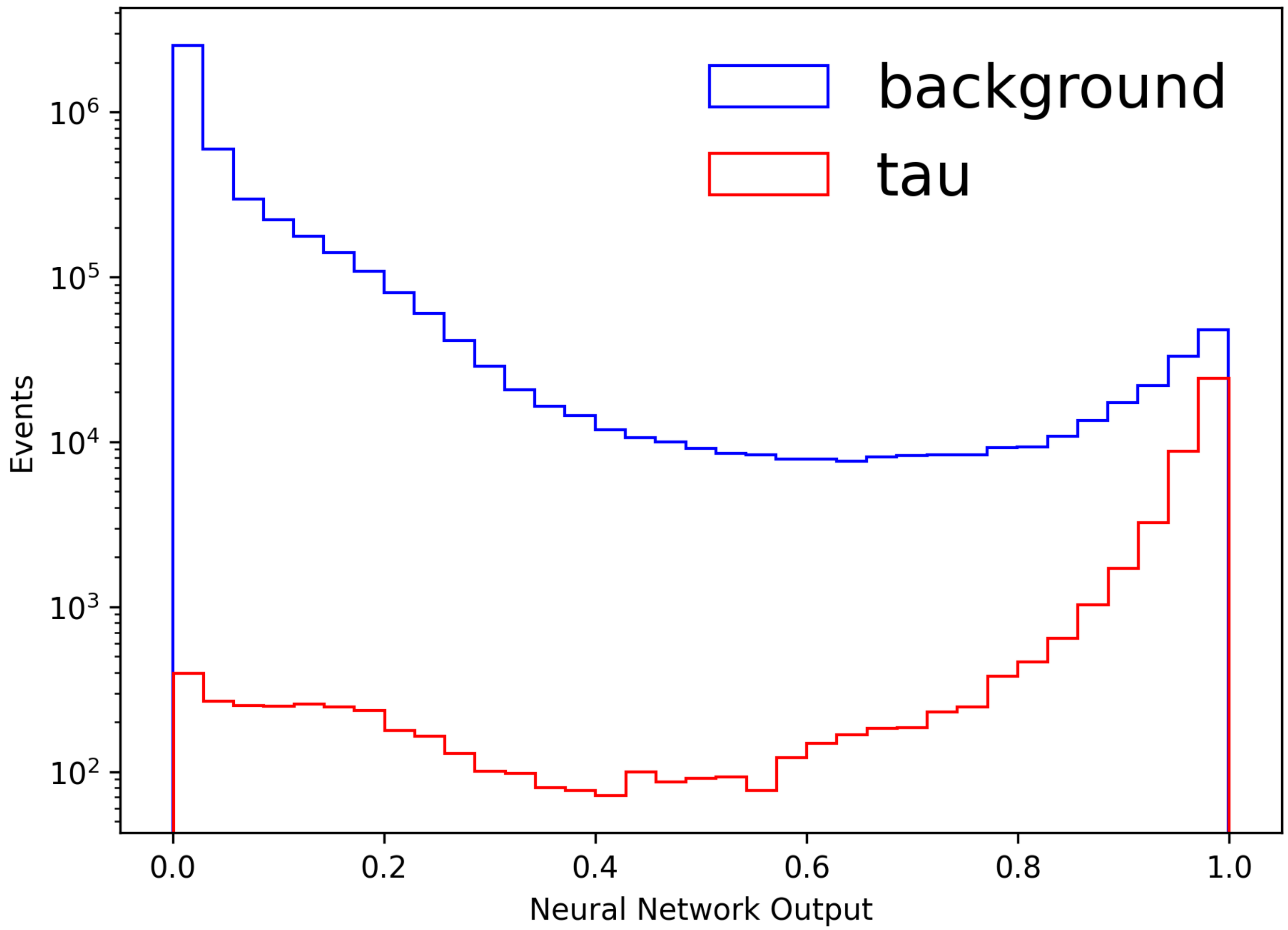




$$\mathrm{EBOPs} \approx \#\mathrm{LUTs} + 55\#\mathrm{DSPs}$$

"Effective Bit Operations"

Evaluation



Visualizing separation of τ signal from background

Signal & Background Efficiency

Signal Efficiency

$$\frac{\#\left[(\mathbf{isTauSignal} == \mathbf{True}) \&\& (\mathbf{Network\ Output} > \mathbf{Threshold})\right]}{\mathbf{isTauSignal} == \mathbf{True}}$$

"True Positive Rate (TPR)"

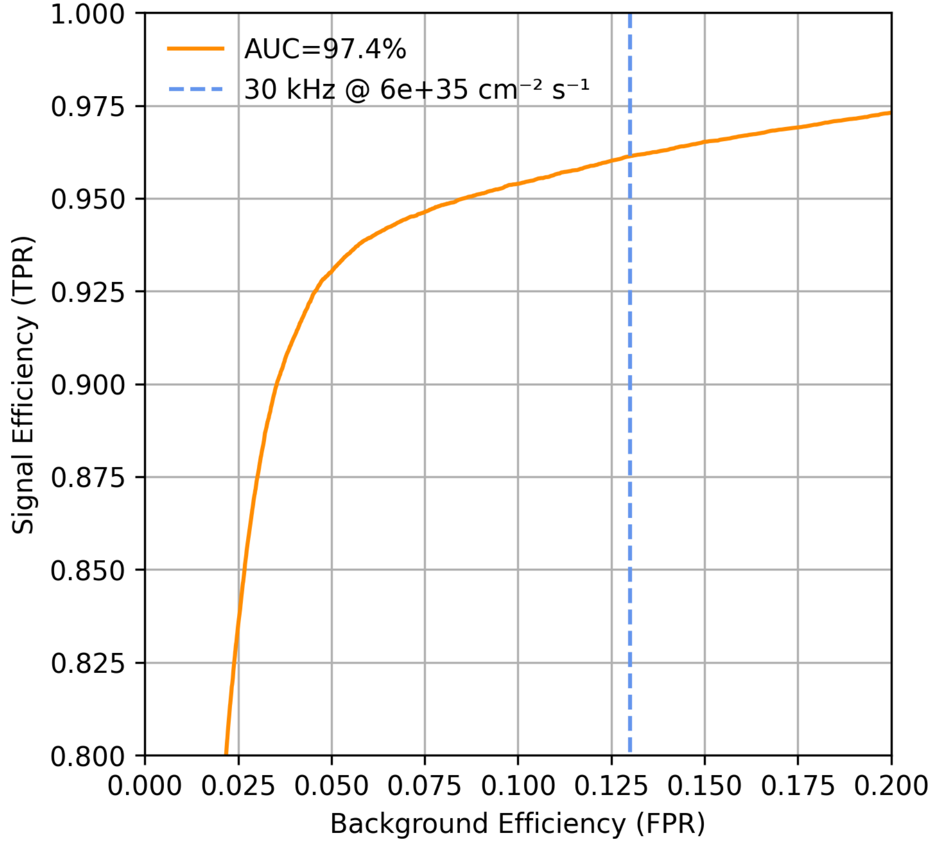
Background Efficiency

$$\frac{\#\left[(\mathbf{isTauSignal} == \mathbf{False}) \&\& (\mathbf{Network\ Output} > \mathbf{Threshold})\right]}{\mathbf{isTauSignal} == \mathbf{False}}$$

"False Positive Rate (FPR)"

Performance

- We use the ($\boldsymbol{\tau}$) Signal Efficiency and Background Efficiency to plot the Receiver Operating Characteristic (ROC) curve.
- The total area under the ROC curve (AUC) is generally (though not always!) a good estimate of overall classification performance.
- At target luminosity, the projected signal efficiency for a trigger rate of $\mathbf{30\ kHz}$ is $\sim \mathbf{96\%}$.



Latency & Utilization Estimates

Latency

```
|  Latency (cycles) |  Latency (absolute) |
|   min   |   max   |    min   |    max   |
+---------+---------+----------+----------+
|       21|       21| 0.105 us | 0.105 us |
```

Required

Latency $\leq \mathbf{500\ ns}$

Resources

```
|      Name       | BRAM_18K| DSP48E|   FF    |   LUT   | URAM|
+-----------------+---------+-------+---------+---------+-----+
|Total            |        0|     72|    57678|   233548|    0|
+-----------------+---------+-------+---------+---------+-----+
|Available        |     2842|    672|   891424|   445712|    0|
+-----------------+---------+-------+---------+---------+-----+
|Utilization (%)  |        0|     10|        6|       52|    0|
```

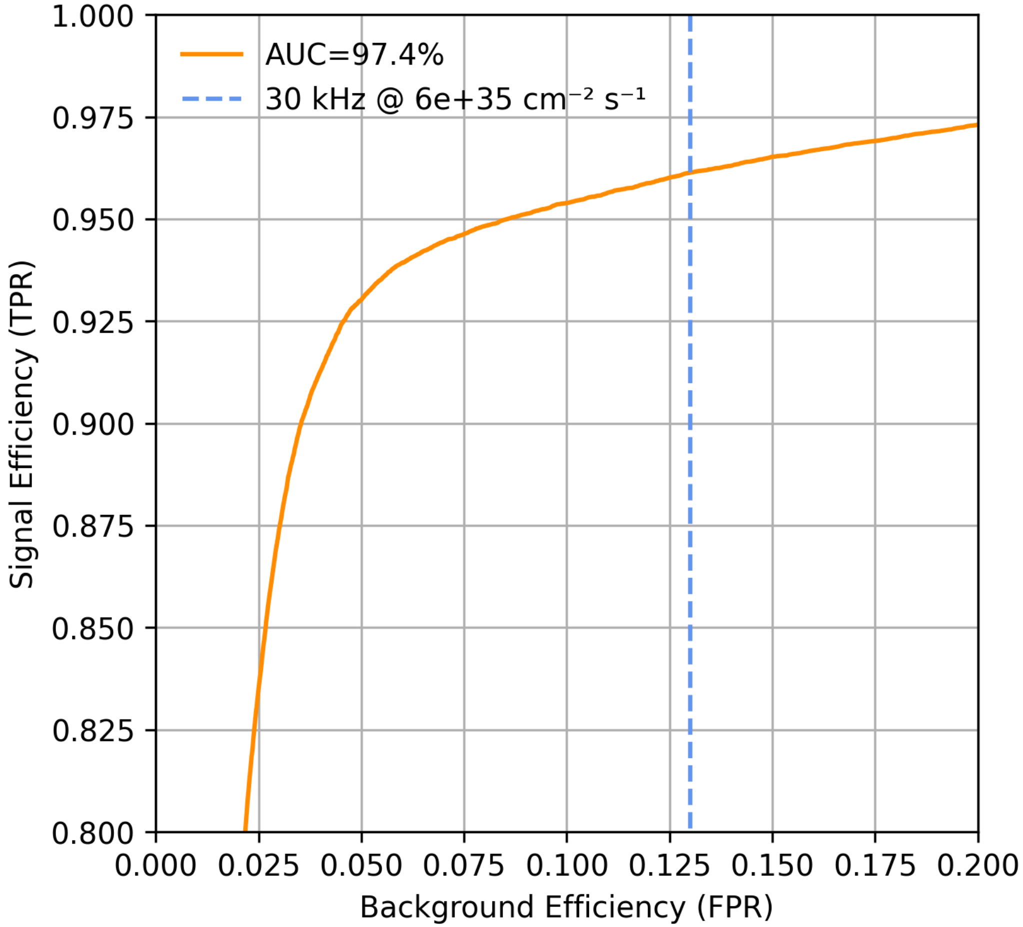
Required

Utilization $\sim \mathbf{50\%}$

- No practical benefit to latency/resource reduction beyond minimal constraints
- Resource utilzation can be reduced to $\sim \mathcal{O}(\mathbf{1\%})$ with latency $\sim \mathcal{O}(\mathbf{10\ ns})$ if necessary with small impact on efficiency ($\sim \mathbf{1-5\%}$)

Conclusion

- Development of new low-latency neural trigger logic with projected signal efficiency of $\sim \mathbf{96\%}$ for a trigger rate of $\mathbf{30\ kHz}$ at SuperKEKB target luminosity.
- Model size can be increased for reconstruction of additional decay modes.

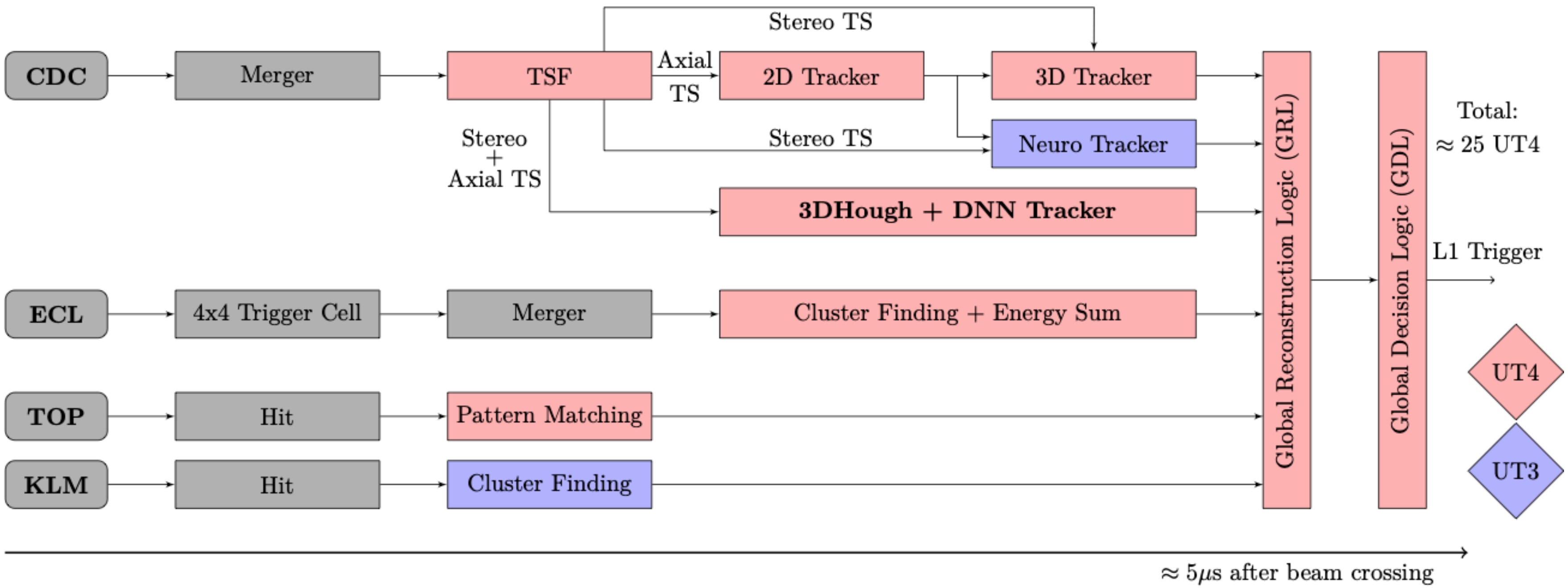


Implementation of the new trigger logic is planned in advance of the upcoming 2025 Belle II physics run.

Q & A

Backup

Level-1 Trigger at Belle II

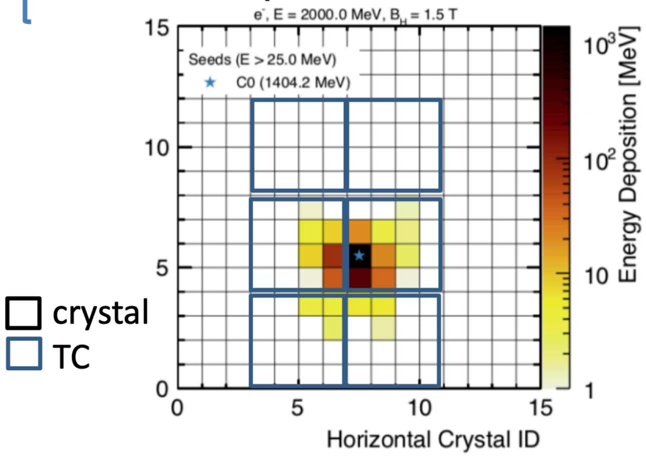


Sub-trigger outputs from ECL, CDC, KLM, and TOP are sent to the Global Decicision Logic (GDL), which performs the final online trigger decision.

ECL Sub-Trigger (ECLTRG)

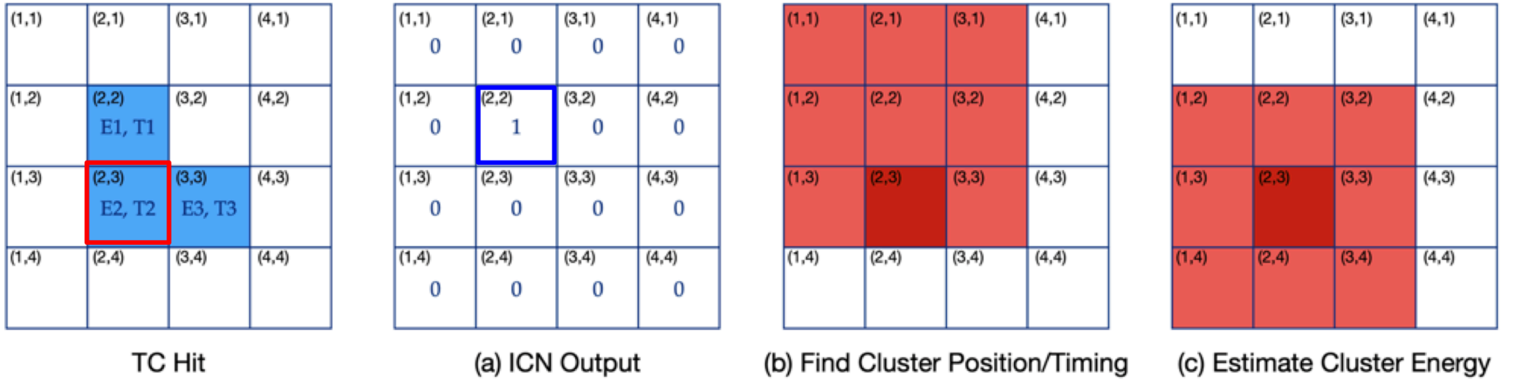
ECL readout does not utilize the full granularity of the crystal segmentation.

-Trigger Cell(TC) is minimum unit
-8736 CsI -> 546TC
-shower shape information is lost



ECL Trigger Cells

ECL Sub-Trigger (ECLTRG)



ECL Cluster Finding and Energy Estimation

CDC Sub-Trigger (CDCTRG)

1. Track Charge
   - $\{Q_1, \ldots, Q_4\}$ (2 bits)
2. Track Curvature
   - $\{\omega_1, \ldots, \omega_4\}$ (7 bits)
3. Angular Position
   - $\{\phi_1, \ldots, \phi_4\}$ (7 bits)

KLM Sub-Trigger (KLMTRG)

1. BKLM Trigger (1 bit) & EKLM Trigger (1 bit)
   - $\mathbf{1}$ if $\geq \mathbf{6}$ scintillator/RPC hits, else $\mathbf{0}$
2. BKLM Trigger Mask (16 bits)
3. EKLM Trigger Mask (8 bits)