

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} Takeo Higuchi^{1,2}

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

Fast Machine Learning for Science Conference, September 2025



Table of Contents

The Belle II Experiment

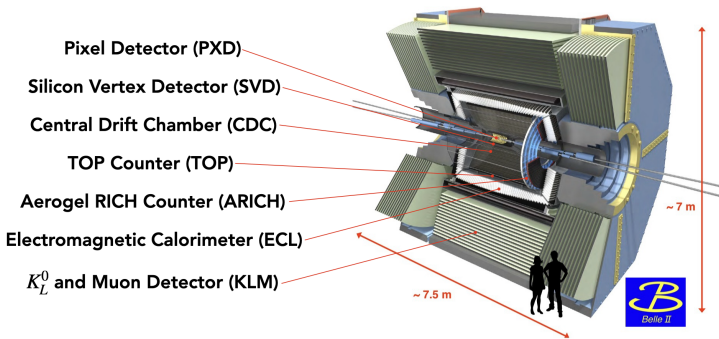
The Level-1 Trigger

Algorithm Design

On-Chip Deployment

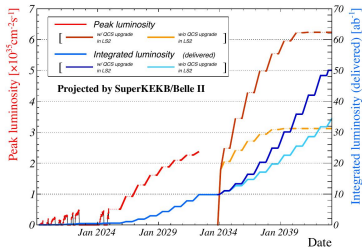
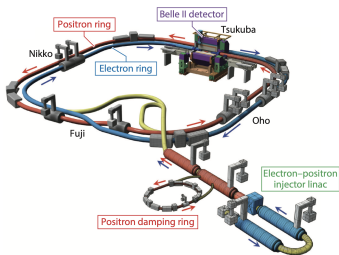
Belle II

Belle II is a luminosity frontier experiment at the SuperKEKB collider.



The first collisions occurred in April 2018; operation is expected to continue until at least ~2040.

SuperKEKB



Luminosity	Current	Target
Peak	$5.1 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$	$6 \times 10^{35} \text{ cm}^{-2} \text{ s}^{-1}$
Integrated	424 fb^{-1}	50 ab^{-1}

SuperKEKB is an asymmetric e^+e^- collider, operating at the $\Upsilon(4S)$ resonance to produce $B\bar{B}$ pairs. Collisions at $\Upsilon(4S)$ also generate a large number of τ leptons, enabling a robust τ physics program.

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
- Lorentz structure of charged-current weak interactions
 - Michel Parameters

Table of Contents

The Belle II Experiment

The Level-1 Trigger

Algorithm Design

On-Chip Deployment

Data Acquisition and Trigger Rate

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

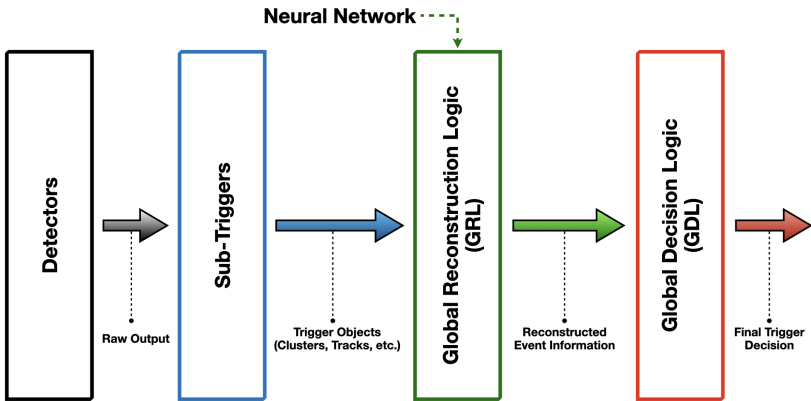
Process	Event rate
e^+e^- bunch collision	~ 200 MHz
Beam background	$> \sim 300$ kHz (2022)
Bhabha scattering	$> \sim 50$ kHz
Two photon processes	~ 10 kHz
$e^+e^- \rightarrow \gamma\gamma$	~ 2 kHz
$e^+e^- \rightarrow q\bar{q}$ ($q = u, d, s, c$)	~ 2 kHz
$e^+e^- \rightarrow Y(4S)$	~ 1 kHz
$e^+e^- \rightarrow \mu^+\mu^-$	~ 0.6 kHz
$e^+e^- \rightarrow \tau^+\tau^-$	~ 0.6 kHz
dark sector/new particle	???

physics target
 ~ 15 kHz

Item	Requirement	Present status
Trigger rate	< 30 kHz @ $6 \times 10^{35} \text{cm}^{-2}\text{s}^{-1}$	~ 8 kHz @ $4.7 \times 10^{34} \text{cm}^{-2}\text{s}^{-1}$
Latency	$4.4 \mu\text{s}$	$4.4 \mu\text{s}$
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

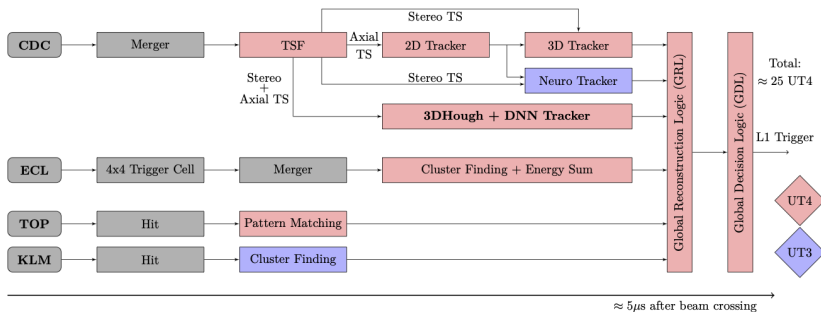
If the current trigger conditions are not modified, the data rate will quickly exceed ~ 30 kHz as SuperKEKB luminosity increases. Our goal is to limit the trigger rate to < 30 kHz while maintaining at least 95% efficiency on standard-model τ decays.

Level-1 Trigger Data Flow



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

Level-1 Trigger Data Flow



A detailed schematic of the Level-1 Trigger at Belle II

Input to Global Reconstruction Logic

Clustering is performed at the detector by the ECL sub-trigger.

(1,1)	(2,1)	(3,1)	(4,1)
(1,2)	(2,2) E1, T1	(3,2)	(4,2)
(1,3)	(2,3) E2, T2	(3,3) E3, T3	(4,3)
(1,4)	(2,4)	(3,4)	(4,4)

TC Hit

(1,1)	(2,1)	(3,1)	(4,1)
0	0	0	0
(1,2)	(2,2) 1	(3,2)	(4,2)
0	0	0	0
(1,3)	(2,3)	(3,3)	(4,3)
0	0	0	0
(1,4)	(2,4)	(3,4)	(4,4)
0	0	0	0

(a) ICN Output

(1,1)	(2,1)	(3,1)	(4,1)
(1,2)	(2,2)	(3,2)	(4,2)
(1,3)	(2,3)	(3,3)	(4,3)
(1,4)	(2,4)	(3,4)	(4,4)

(b) Find Cluster Position/Timing

(1,1)	(2,1)	(3,1)	(4,1)
(1,2)	(2,2)	(3,2)	(4,2)
(1,3)	(2,3)	(3,3)	(4,3)
(1,4)	(2,4)	(3,4)	(4,4)

(c) Estimate Cluster Energy

Parameter	Energy (E)	Azimuthal Angle (θ)	Polar Angle (ϕ)	Avg. Time (t)
Bits	12	7	8	8
LSB	5.5 MeV	1.40625°	1.40625°	1 ns

Information for the **six highest-energy clusters** in an event is sent to the Global Reconstruction Logic.

Table of Contents

The Belle II Experiment

The Level-1 Trigger

Algorithm Design

On-Chip Deployment

Signal Definition

We focus on reconstruction of the low-multiplicity standard model τ decays:

Leptonic Decays

$$\tau \rightarrow \mu \nu_\tau \bar{\nu}_\mu \mid \text{BR} = 17.82\% \mid (\text{"1-prong"})$$

$$\tau \rightarrow e \nu_\tau \bar{\nu}_e \mid \text{BR} = 17.39\% \mid (\text{"1-prong"})$$

Hadronic Decays

$$\tau \rightarrow \nu_\tau + \pi^\pm + n(\pi^0) \mid \text{BR} = 49.5\% \mid (\text{"1-prong"})$$

$$\tau \rightarrow \nu_\tau + 3\pi^\pm + n(\pi^0) \mid \text{BR} = 15.2\% \mid (\text{"3-prong"})$$

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

Dataset

Run Information:

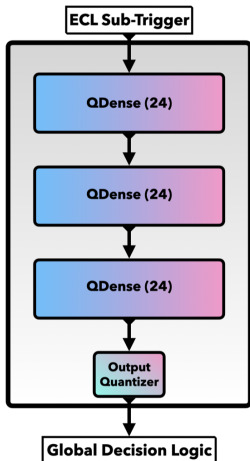
Integrated Luminosity: **36.1** pb^{-1}

Trigger Rate: **11.8** kHz

This study uses **real data** collected with a loose trigger condition to sample backgrounds. We use offline reconstruction to identify low-multiplicity standard model τ decays and define all other events as background.

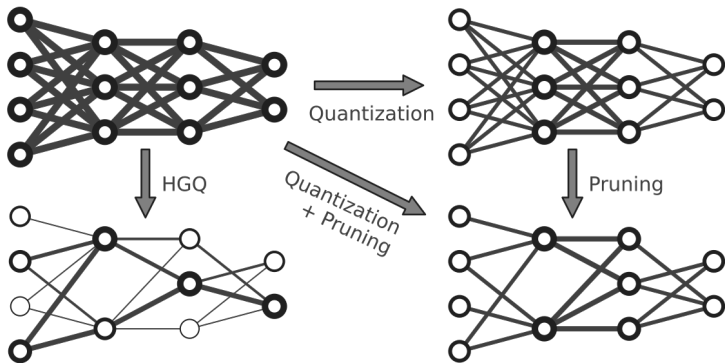
At Belle II, the dominant backgrounds come from beam interactions (**~ 300 kHz**) and Bhabha scattering (**~ 50 kHz**).

Model Architecture



- Feed-forward dense neural network implemented in Keras 3.0 / JAX
- Quantization-aware training with gradient-based bitwidth optimization (HGQ2)

High Granularity Quantization (HGQ2)



Sun et. al, *Gradient-based Automatic Mixed Precision Quantization for Neural Networks On-Chip*, arXiv:2405.00645 (2024)

Performance

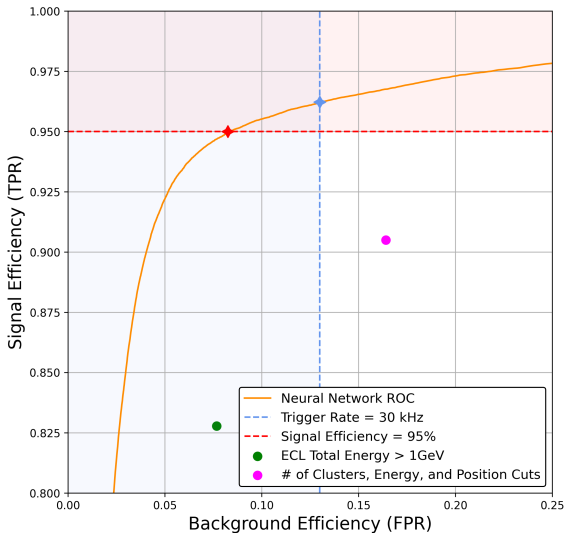


Table of Contents

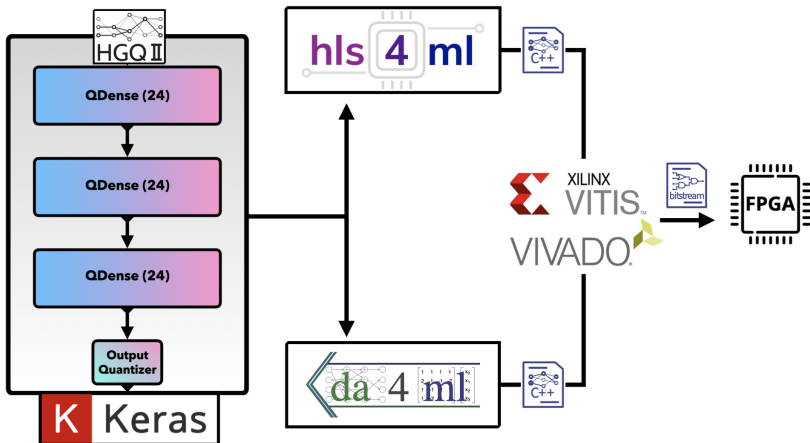
The Belle II Experiment

The Level-1 Trigger

Algorithm Design

On-Chip Deployment

High-Level Synthesis



Duarte et. al, *Fast inference of deep neural networks in FPGAs for particle physics*, JINST 13, no.07, P07027 (2018)

Sun et. al, *da4ml: Distributed Arithmetic for Real-time Neural Networks on FPGAs*, arxiv:2507.04535 (2025)

Latency & Utilization Estimates (hls4ml)

```
strategy = 'latency'
```

Latency	Maximum	Minimum
Cycles	12	12
Absolute	94 ns	94 ns

Resource	BRAM	DSP	FF	LUT	URAM
Total	0	205	11135	62599	0
Available	2842	672	891424	445712	0
Utilization	0%	30%	1%	14%	0%

Requirements:

Latency < 500 ns | DSP Utilization < 50% | LUT Utilization < 20%

Latency & Utilization Estimates (da4ml)

Latency	Maximum	Minimum
Cycles	9 (-3)	9 (-3)
Absolute	71 ns (-24 ns)	71 ns (-24 ns)

Resource	BRAM	DSP	FF	LUT	URAM
Total	0	0	7254	44849	0
Available	2842	672	891424	445712	0
Utilization	0%	0% (-30%)	0% (-1%)	10% (-4%)	0%

Requirements:

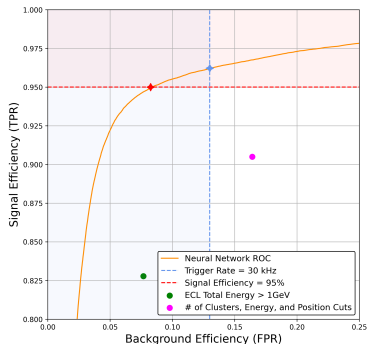
Latency < 500 ns | DSP Utilization < 50% | LUT Utilization < 20%

Firmware Validation

- Matched outputs in RTL behavioral simulation
- Install to AMD/Xilinx Virtex UltraScale FPGA on Belle II Universal Trigger Board 4 (UT4) for validation in cosmic run

Results & Conclusion

- New low-latency neural trigger logic for reconstruction of low-multiplicity standard-model τ decays
- Up to **50% reduction** in total trigger rate while maintaining **over 95%** signal efficiency at the target luminosity
- Firmware validated in RTL simulation and cosmic run



Implementation of the new logic is planned for the next Belle II physics run.

GRL Resource Utilization

Resource	LUT	LUTRAM	FF	BRAM	DSP	IO	GT	BUFG	MMCM
Total	141396	6071	150754	214	0	291	36	48	3
Available	445712	76800	891424	1421	672	702	64	960	16
Utilization	32%	8%	17%	15%	0%	41%	56%	5%	19%

Post-Synthesis

Resource	LUT	LUTRAM	FF	BRAM	DSP	IO	GT	BUFG	MMCM
Total	162537	6214	151612	512	0	291	44	54	4
Available	445712	76800	891424	1421	672	702	64	960	16
Utilization	36%	8%	17%	36%	0%	41%	69%	6%	25%

Post-Implementation

Requirements:

DSP Utilization < 50% | LUT Utilization < 50%