

A Deep Neural Network ~~based~~Based Level-1 Trigger for the Magnetic Monopole Search in the SCEPT Experiment

Changqing Ye, Yunhan Wang, Yifan Qiao, Zhe Cao, Beige Liu, Qing Lin, and Lei Zhao

Abstract—The Search for Cosmic Exotic Particles (SCEPT) experiment proposes a multilayer magnetic monopole (MM) detection array with a target exposure of $\Theta(10^4) \text{ m}^2 \cdot \text{year}$ ~~$O(10^4) \text{ m}^2 \cdot \text{yr}$~~ . Continuously recording the raw waveforms at this scale would generate approximately $O(10)$ EB of data per year, making full waveform storage impractical. A Level-1 (~~L-1~~) trigger is therefore required to reduce the stored-data volume while retaining rare ~~MM~~monopole signals. ~~MM~~Monopole waveforms vary with particle velocity and trajectory, so a single filter kernel cannot match all possible signals. A mismatch between the kernel and the signal waveform can reduce the detection probability. This work ~~therefore~~ develops a compact template-bank construction method for ~~MM~~monopole detection. Each incoming waveform is filtered with all kernels in the bank, and the best-matched kernel is identified from their responses. A multilayer perceptron (~~MLP~~) trigger is further developed to combine all template responses and improve signal detection. With a 31-sample window and a stored-data fraction of 10^{-3} , the ~~MLP~~multilayer perceptron increases the mean net acceptance P_{net} from 6.72% for the conventional template-bank trigger to 8.79%, a relative improvement of 30.8%. The network is then implemented on a Xilinx AC701 board equipped with a field-programmable gate array (~~FPGA~~) to process a 20-bit input waveform sampled at 1 MHz using 200 MHz trigger logic. At the target stored-data fraction, the fixed-point implementation reduces P_{net} by only $3/(3.2 \times 10^6)$ ~~relative to compared with~~ the floating-point reference. On-board tests reproduce the bit-accurate reference trigger outputs and sustain continuous 1 MHz processing without dropped input samples, verifying the functional correctness and real-time operation of the ~~FPGA-based L-1 hardware~~ trigger.

Index Terms—Field-programmable gate array (FPGA), Level-1 trigger, magnetic monopole detection, neural network, optimal filtering, template bank.

I. INTRODUCTION

MAGNETIC monopoles (MMs), first proposed by Dirac in 1931 [1], are hypothetical particles carrying iso-

lated magnetic charge. Their existence would explain electric-charge quantization through the Dirac quantization condition and is also predicted in various grand unified theories [2], [3], [4] [2]–[4]. Searches for MMs have used several detection strategies. Induction-based searches with superconducting coils look for the persistent current change induced by an MM passage [5], [6]. Other searches exploit the energy loss of MMs in detector media. The Monopole, Astrophysics and Cosmic Ray Observatory (MACRO) used scintillation counters, limited streamer tubes, and nuclear track detectors [7]. The IceCube Neutrino Observatory has searched for relativistic MMs through their Cherenkov-light signatures in ice [8]. These techniques have placed strong limits on the MM flux [9], [10], but their operating conditions or detector-scale requirements make further exposure increases challenging.

Room-temperature induction coils offer a scalable approach to large-area MM searches. Unlike superconducting systems, they do not require cryogenic cooling, which facilitates deployment in large-area, multilayer detector arrays. The Search for Cosmic Exotic Particles (SCEPT) experiment proposes a coincidence-detection scheme combining room-temperature induction coils with plastic scintillators. Its proposed multilayer array targets an exposure of $\Theta(10^4) \text{ m}^2 \cdot \text{year}$ ~~$O(10^4) \text{ m}^2 \cdot \text{yr}$~~ . Under the detector-performance assumptions of the proposal, this exposure would allow sensitivity comparable to existing MM-flux limits over a broad velocity range [11].

For a detector system of this scale, recording all waveform data continuously would overburden the bandwidth and storage capacity of the data-acquisition system. For an array designed to accumulate the target exposure over ten years, the annual raw-data volume is estimated to be $O(10)$ EB for the SCEPT design. Direct storage of all waveforms is therefore impractical, and a Level-1 (~~L-1~~) trigger is required to suppress noise-induced events while retaining potential signals for subsequent analysis.

Because MM events are extremely rare and a true event cannot be recovered once the front-end trigger rejects it, the ~~L-1~~L1 trigger must maintain high signal acceptance. In room-temperature induction-coil detection, coil thermal noise results in a low signal-to-noise ratio (SNR) for MM-induced signals. ~~For low-velocity MMs, a processed-signal SNR of approximately 4.5~~The SCEPT proposal requires a processed-signal power SNR of at least 4.5 for an MM crossing the coil center perpendicularly at 10^{-5} of the speed of light [11], and this value is adopted as the reference

This work was supported in part by the National Natural Science Foundation of China under Grant 12250011 and in part by the Frontier Scientific Research Program of Deep Space Exploration Laboratory under Grant 2022-QYKYJH-HXYF-013. (Corresponding authors: Zhe Cao, Qing Lin, and Lei Zhao.)

Changqing Ye and Beige Liu are with the Department of Modern Physics, University of Science and Technology of China, Hefei 230026, China.

Yunhan Wang is with the Institute of Advanced Technology, University of Science and Technology of China, Hefei 230088, China.

Yifan Qiao is with the School of Emerging Technology, University of Science and Technology of China, Hefei 230026, China.

Zhe Cao, Qing Lin, and Lei Zhao are with the Department of Modern Physics, University of Science and Technology of China, Hefei 230026, China, and also with the Deep Space Exploration Laboratory, Hefei 230022, China (e-mail: caozhe@ustc.edu.cn; qinglin@ustc.edu.cn; zlei@ustc.edu.cn).

signal level in this work. Such an SNR still leaves the signal vulnerable to noise fluctuations [44]. Maintaining high signal acceptance therefore requires a low trigger threshold, which produces a large number of noise-induced false triggers. The stored data are expected to be dominated by noise-induced false triggers rather than noise rather than by genuine signals. The data reduction is quantified by the stored-data fraction, defined as the number of saved waveform samples divided by the total number of raw input samples. A target stored-data fraction of 10^{-3} (≈ 10 PB/year ≈ 10 PB/yr for the SCEP design) is adopted as a manageable storage load. In addition, the L-L1 trigger should be implemented in front-end hardware to make trigger decisions in real time and thereby reduce the data-transfer load [42], [43] [12], [13].

The L-L1-trigger design for MM detection can draw on the approach used in gravitational-wave searches. Induction-based MM searches in SCEP and compact-binary gravitational-wave searches with the Laser Interferometer Gravitational-Wave Observatory (LIGO) [14], [15] share a common challenge: both seek weak transient signals in noise-dominated data streams, while the signal parameters are unknown before detection. For a known signal waveform and noise covariance, optimal filtering (OF) maximizes the output SNR, whereas a mismatch between the filter kernel and the signal causes an SNR loss [16]. A standard way to address unknown signal parameters is to filter each data window with a bank of candidate templates and use the template with the largest response magnitude [17]–[19].

Early template-bank constructions set the template spacing through a relative loss in SNR [20], [21]. Direct comparisons have shown that adding a more complex template family a bank built from a more complex template family does not necessarily improve detection performance at a fixed false-alarm rate [22]. Adding templates can increase the largest noise response. At a fixed false-positive probability (FPR), this can raise the trigger threshold and reduce the signal acceptance [23], because the additional templates can increase the largest noise response, which raises the trigger threshold and reduces the signal acceptance [23]. Template-bank construction must therefore satisfy the detection-probability coverage requirement while using as few templates as possible. In addition, retaining only the largest template response discards information in the remaining responses that may help distinguish signals from noise.

To address these limitations, this paper constructs a finite template bank using the absolute loss in detection probability as the coverage criterion. This criterion avoids unnecessary refinement of weakly detectable waveform regions. A multi-layer perceptron (MLP) trigger is then developed to combine all template responses at each sampling time, improving signal acceptance at the same stored-data fraction. The remainder of this paper is organized as follows. Section II describes the template bank, the MLP trigger, and the performance evaluation. Section III describes the streaming field-programmable gate array (FPGA) architecture for the selected network and its validation procedure. Section IV concludes the paper.

II. L-L1 TRIGGER ALGORITHM

A. Signal and Noise

In the SCEP induction-coil scheme, the MM signal waveform is set by the particle velocity and trajectory. Each trajectory can be parameterized in a coordinate system fixed to the induction coil. The coil plane is taken as the xy -plane, and the z axis is perpendicular to this plane. The physical parameters are denoted by

$$\gamma \equiv (v, \rho_0, \theta, \phi, q), \quad q \in \{-1, +1\}, \quad (1)$$

where v is the MM velocity, ρ_0 is the transverse distance from the coil center when the MM reaches the coil plane, and θ and ϕ are the polar and azimuthal angles of the directed incident trajectory. The parameter q denotes the magnetic-charge polarity, with $q = +1$ and $q = -1$ corresponding to magnetic charges of $+g_D$ and $-g_D$, respectively. Here, g_D is the unit Dirac magnetic charge [1], [9].

No well-established theoretical prior is available for the MM velocity. Accordingly, v is sampled uniformly on a logarithmic scale over $[10^{-5}, 10^{-1}]c$, where c is the speed of light. This choice prevents the sampled trajectories from being concentrated at high velocities. The dimensionless velocity $\beta \equiv v/c$ is used below. The incoming MMs are assumed to be isotropic, and their straight trajectories are sampled with a uniform flux through a sphere of radius 10 cm that encloses the coil. The transverse offset ρ_0 and the incident angles (θ, ϕ) are not sampled independently; they are determined by the geometrical relation between each isotropic trajectory and the induction coil. The magnetic-charge polarity q is sampled with equal probabilities for $+1$ and -1 . The signal parameter space of the trajectories parameterized in Eq. (1) is denoted by Γ .

The MM crossing induces a transient electromotive force $\mathcal{E}_s$ in the coil. The coil is read out by a low-noise amplifier with high input impedance connected in parallel across its terminals, as shown in Fig. 1, and the resulting circuit transfer function shapes the induced waveform. The signal and noise models use the same coil design and readout-circuit parameters as those specified in the SCEP proposal [11]. Together with the parallel capacitance C_d , the coil forms an LCR circuit with a resonance frequency of about 54.2 kHz. Owing to this resonance, the circuit-shaped signal oscillates, and the coil resistance damps the oscillation so that its amplitude falls below 1% of its initial value about 2.8 ms after the crossing, as shown in Fig. 2(a). The circuit-shaped signal waveform and the noise are shown in Fig. 2. Under room-temperature operation, the dominant noise contribution is the coil thermal noise $\mathcal{E}_{th}$, which can overwhelm the MM signal. Shaped by the same circuit, this random noise also oscillates near the resonance frequency, as shown in Fig. 2(b). The total noise is modeled as a zero-mean stationary Gaussian process. Non-stationary Nonstationary or non-Gaussian disturbances in a deployed detector are beyond the scope of this study.

B. Optimal Filter and Mismatch

Detecting an MM signal in a noisy waveform can be formulated as a statistical hypothesis-testing problem. Each input waveform window contains L samples. Under the null

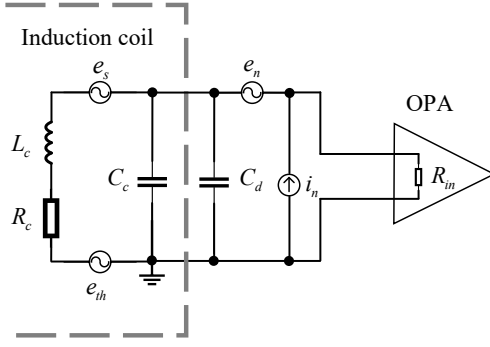


Fig. 1. Equivalent circuit of the induction-coil readout. The dashed box encloses the induction coil, represented by its inductance L_c , frequency-dependent resistance R_c , and capacitance C_c ; C_d is the parallel capacitance formed by the amplifier input capacitance and stray capacitances. The MM-induced electromotive force e_s and the coil thermal noise e_{th} act in series with the coil and are therefore shaped by the same transfer function; e_n and i_n are the voltage and current noise of the operational amplifier (OPA) with input resistance R_{in} .

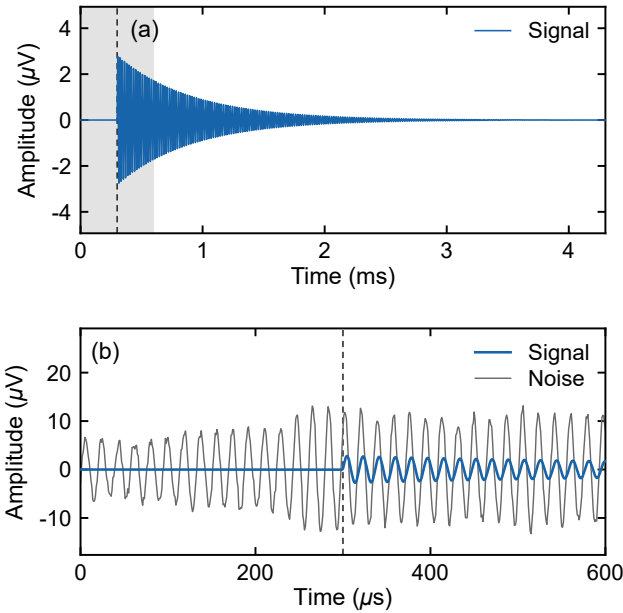


Fig. 2. Circuit-shaped MM signal and noise for an on-axis trajectory ($\rho_0 = \theta = \phi = 0$) with $\beta = 10^{-3}$. (a) Signal over 4.3 ms; the shaded interval is the window shown in (b). (b) Signal and a simulated realization of the total noise over 0–600 μ s. The blue curve shows the circuit-shaped MM signal, the gray curve shows the total noise (coil thermal noise plus readout-electronics contributions such as amplifier noise), and the dashed line marks the signal injection time in the analysis window. Vertical dashed lines mark the crossing time.

hypothesis, $\mathcal{H}_0$, the input waveform window contains only noise. Under the alternative hypothesis, $\mathcal{H}_1$, the window contains an MM signal plus noise. Expressed mathematically,

$$\begin{aligned} \mathcal{H}_0: \mathbf{x} &= \mathbf{n}, \\ \mathcal{H}_1: \mathbf{x} &= \mathbf{s}_\gamma + \mathbf{n}, \end{aligned} \quad (2)$$

where $\mathbf{x}$ is the input waveform window of L samples, $\mathbf{s}_\gamma$ is the circuit-shaped signal waveform for parameter γ , and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, C)$ states indicates that the noise vector $\mathbf{n}$ follows a multivariate Gaussian distribution with zero mean

and covariance matrix C , computed from the noise power spectral density. This is the sampled form of the Gaussian noise model introduced in Section II-A. For the hypotheses in Eq. (2) with a specified signal waveform, the Neyman–Pearson lemma [24] gives the likelihood-ratio test (LRT) as the optimal decision rule. In Gaussian noise, it reduces to testing the linear statistic [25]

$$\mathbf{s}_\gamma^\top C^{-1} \mathbf{x} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda. \quad (3)$$

Here λ is the decision threshold: when the linear statistic exceeds λ , the window is assigned to the signal hypothesis $\mathcal{H}_1$; otherwise it is assigned to the noise-only hypothesis $\mathcal{H}_0$. Equation (3) has the form of an OF statistic $\mathbf{h}_\gamma^\top \mathbf{x}$, where the OF kernel takes the form [26] [25], [26]

$$\mathbf{h}_\gamma = a_\gamma C^{-1} \mathbf{s}_\gamma, \quad (4)$$

where a_γ is a normalization coefficient chosen so that the root-mean-square noise after OF is unity.

In a streaming implementation, at each fixed sampling interval a new sample enters the trigger and the oldest sample leaves, forming the updated input window on which the OF is evaluated. The signal is centered within the kernel, and the filter response peaks when the signal in the current window aligns with the template. At this alignment, the output SNR is defined as the squared noise-free signal response divided by the mean squared noise response. With the normalization of the kernel in Eq. (4), this gives

$$\text{SNR}(\gamma; \gamma) = (\mathbf{h}_\gamma^\top \mathbf{s}_\gamma)^2. \quad (5)$$

The input waveform, OF kernel, and resulting filter output are shown in Fig. 3.

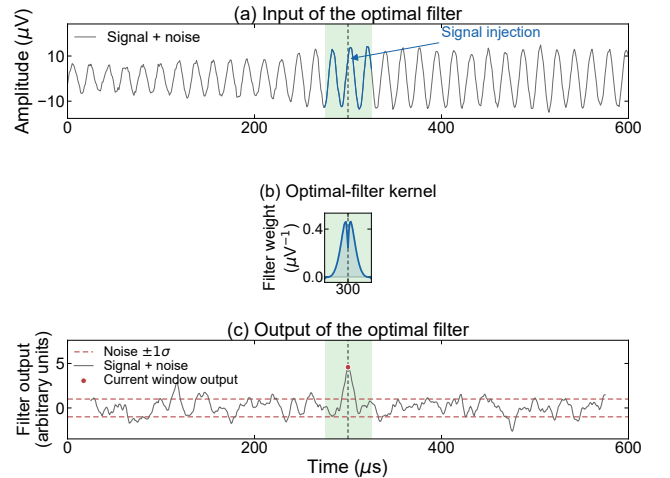


Fig. 3. Optimal filtering of a signal-plus-noise waveform. Top: the (a) circuit-shaped input waveform, with the signal injected at the window center marked by the dashed line. Middle: the (b) normalized OF kernel $\mathbf{h}_\gamma = a_\gamma C^{-1} \mathbf{s}_\gamma$ aligned with the signal. Bottom: the (c) OF output, where the matched response forms a peak near the injection time. The red dashed lines indicate the $\pm 1\sigma$ noise level after normalization, where σ is the noise standard deviation.

In practical detection, the true signal parameters γ are unknown. When the signal $\mathbf{s}_\gamma$ is filtered with the kernel $\mathbf{h}_\gamma$,

where η denotes the parameters of the reference signal used to construct the kernel, the resulting output SNR is denoted by $\text{SNR}(\gamma; \eta)$. It is largest for the matched case $\eta = \gamma$, where $\text{SNR}(\gamma; \gamma)$ equals the benchmark of Eq. (5).

The SNR loss caused by mismatch can be substantial. Filtering a $v=10^{-3}e\beta=10^{-3}$ signal with a $v=10^{-5}e\beta=10^{-5}$ template gives an power SNR loss of 55.6% relative to the matched case. The corresponding time-domain waveforms and filter-response comparison are shown in Fig. 4.

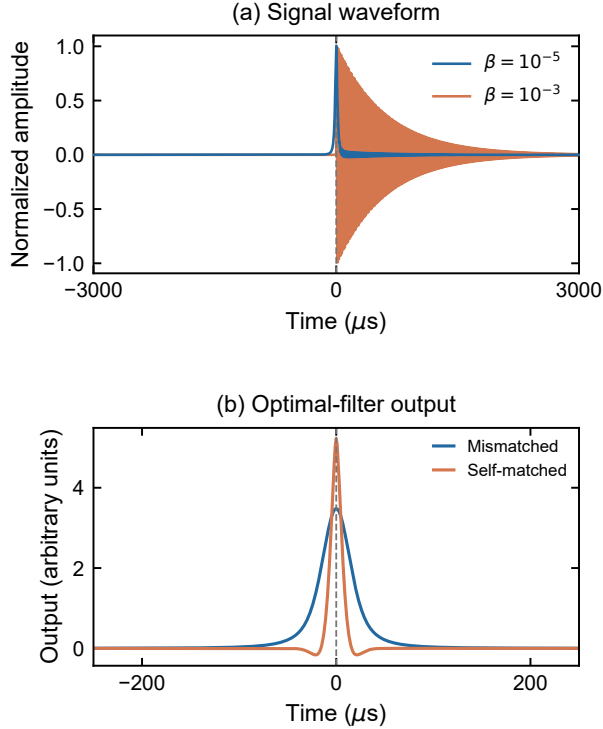


Fig. 4. Example of a large template mismatch for two on-axis trajectories with $\rho_0 = \theta = \phi = 0$. The upper-panel Panel (a) compares the time-domain waveforms for $v=10^{-5}e\beta=10^{-5}$ (blue) and $v=10^{-3}e\beta=10^{-3}$ (orange); the signals are normalized to facilitate the comparison of the two waveform shapes. The lower-panel Panel (b) shows that filtering the $v=10^{-3}e\beta=10^{-3}$ signal with the mismatched $v=10^{-5}e\beta=10^{-5}$ template (blue) yields a much smaller peak than filtering it with its self-matched template (orange), corresponding to an power SNR loss of 55.6%.

C. Template Bank

The preceding example shows that a filter kernel can incur a substantial SNR loss when its waveform does not match the signal. Because the true signal parameters are unknown and the parameter space Γ is continuous, a finite template bank $\mathcal{B}$ is constructed to cover the sampled signal distribution within this space. Each input window is filtered with every kernel in $\mathcal{B}$. Templates that differ only in magnetic-charge polarity have opposite kernels, so one kernel is retained for each waveform shape. The template-bank trigger uses the largest response magnitude,

$$\max_{\eta \in \mathcal{B}} |\mathbf{h}_{\eta}^{\top} \mathbf{x}| \stackrel{\mathcal{H}_1}{\geq} \lambda. \quad (6)$$

Although this bank statistic no longer retains the Neyman-Pearson optimality of the fixed-parameter LRT in (3), it is widely used as a practical approximation to that test, for example in LIGO searches [18], [19].

The bank construction has two objectives. First, the templates should cover the sampled signal distribution with a controlled loss in detection probability. Second, the number of templates should be kept small. In addition to increasing the filtering cost, adding templates can increase the probability of a noise-induced trigger under Eq. (6).

Coverage is quantified using the theoretical detection probability for a single normalized template response in Gaussian noise [25]. At a specified FPR false-positive rate (FPR), the detection probability is determined by the signal SNR, as shown in Fig. 5. For each signal, its own waveform is used as the ideal matched template and defines the reference detection probability. A signal is considered covered when, at every FPR from 10^{-6} to 0.1, the absolute loss in detection probability obtained with the best-matched template in $\mathcal{B}$ is less than 0.5 percentage points (pp). Here, pp denotes the absolute difference between probabilities expressed as percentages; a loss of 0.5 pp corresponds to a probability difference of 0.005, not a relative loss of 0.5%. Unlike the relative SNR-loss criterion commonly used in LIGO template-bank construction [20], [21], this absolute criterion relates coverage directly to detectability and prevents signals with self-matched detection probabilities close to the noise-only level from driving further refinement of the bank.

To satisfy this coverage criterion while limiting the number of templates, the bank is constructed iteratively using a farthest-first rule [27]. The construction set is generated using the sampling distribution of Section II-A. It provides a Monte Carlo representation of the continuous signal parameter space Γ under this distribution. At each step, the coverage of the current bank is evaluated over the construction set, and the trajectory with the largest current absolute loss in detection probability is added to the bank. The iteration stops when the loss is below 0.5 pp for every construction trajectory. By prioritizing the worst-covered trajectory, this rule avoids adding templates for waveforms that are already well represented.

The window length L affects the construction result. For $L = 31$ samples, the construction stops at 113 templates when all 10^6 isotropic physical trajectories in the construction set satisfy the 0.5-pp coverage bound. As shown in the upper panel of Fig. 6, all construction trajectories satisfy the coverage requirement, with the maximum loss below 0.50 pp, when the maximum loss over all 10^6 construction trajectories falls below 0.5 pp, as shown in Fig. 6(a). An independent audit is then performed using another 10^6 discrete trajectories drawn from the same sampling distribution. The empirical coverage rate exceeds 99.99%, with 15 trajectories falling outside the coverage bound; the largest observed loss is 0.57 pp, as shown in the lower panel of Fig. 6(b). These results demonstrate empirical coverage under the stated sampling distribution.

For comparison, a 601-template bank is constructed for $L = 31$ using the stochastic template-placement method employed in LIGO [28], applied to the same MM waveform model. This method controls the relative loss in matched-

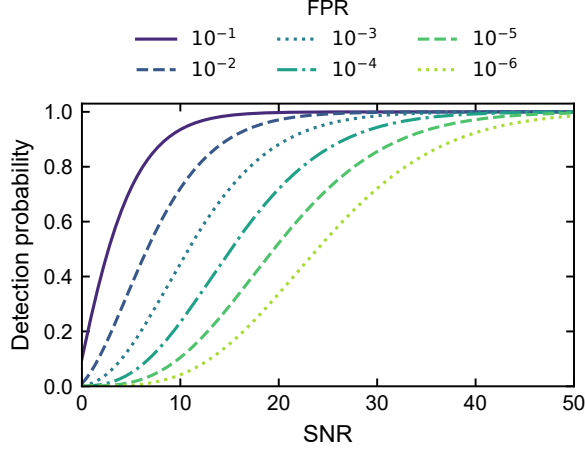


Fig. 5. Theoretical single-template detection probability for a normalized Gaussian response as a function of SNR. The curves correspond to the FPR values in the legend, and the SNR is the power SNR defined in (5).

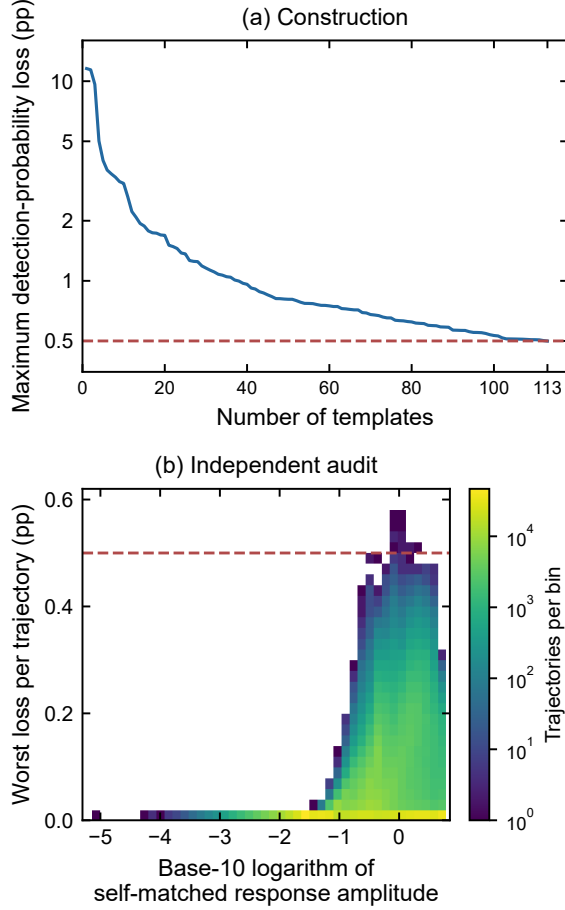


Fig. 6. Construction and independent audit of the 113-template bank. Upper: maximum (a) Maximum detection-probability loss over the 10^6 construction trajectories as templates are added. Lower: loss (b) Loss distribution for 10^6 independent audit trajectories. The dashed lines mark the 0.5-pp coverage bound. The lower horizontal axis (b) shows the base-10 logarithm of the self-matched response amplitude.

filter response amplitude by requiring a minimum normalized match of 0.97, corresponding to a maximum amplitude loss of 3%. The triggers using the 113-template and 601-template banks are referred to below as the compact and conventional template-bank triggers, respectively.

D. Streaming Trigger and Acceptance Evaluation

After the template bank has been constructed, its practical trigger performance must be evaluated on a continuous waveform stream. Previous gravitational-wave studies have likewise shown that template-space coverage alone does not determine overall detection performance [23]. At each sampling step, the current L -sample window is filtered with every template in the bank, and the largest absolute response in Eq. (6) is compared with the threshold. When the threshold is exceeded, the corresponding L raw samples are retained. In practice, an event often produces several consecutive above-threshold windows rather than a single isolated trigger. These consecutive windows are grouped into one event cluster. Because the windows are shifted sample by sample, a cluster spanning N_{trig} consecutive trigger windows retains

$$L_{\text{saved}} = L + N_{\text{trig}} - 1 \quad (7)$$

raw samples. If the saved intervals from neighboring clusters overlap, they are merged before storage, so that no raw sample is stored twice. The storage rule in Eq. (7), including merging of overlapping intervals, is illustrated in Fig. 7.

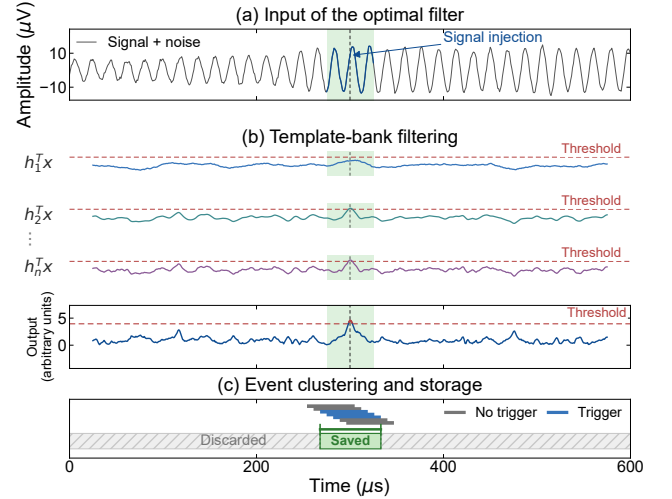


Fig. 7. Streaming evaluation of the template-bank trigger. Top: the (a) The signal-plus-noise waveform, with the injected signal highlighted around its center time. Middle: representative (b) Representative template responses and the maximum response magnitude, which is compared with the trigger threshold shown by the red dashed line. Bottom: consecutive (c) Consecutive above-threshold windows are grouped into an event cluster, and their overlapping intervals are merged into the saved waveform segment.

Under this storage rule, the acceptance of a template bank can be evaluated at a specified stored-data fraction. For each trigger method, the threshold is first calibrated on the negative set $\mathcal{D}_-$, taken as a simulated continuous background-only noise stream of 2.01×10^8 raw samples. The threshold is

chosen so that the stored-data fraction ε on $\mathcal{D}_-$ reaches the target value $\varepsilon_0 = 10^{-3}$.

After the threshold is fixed, the corresponding trigger is evaluated on the positive set $\mathcal{D}_+$. This set contains 3.2×10^6 signal-plus-noise records. Each record is paired with a noise-only record that contains the same noise realization without the injected signal. Because the MM crossing is asynchronous with the sampling clock, each injection is given a random sub-sample timing offset $\delta \in [-0.5, 0.5]$ samples, so that the evaluation reflects arbitrary sampling phases. A signal is counted as detected if its event cluster contains at least one trigger sufficiently close to the true crossing time. The allowed timing deviation accounts for the spread introduced mainly by noise fluctuations and the random sub-sample timing offset. The relevant quantity is the trigger offset

$$\Delta t = t_{\text{trig}} - t_0, \quad (8)$$

where t_{trig} denotes the center time of an above-threshold trigger window and t_0 is the true crossing time. The distribution of Δt defined in Eq. (8) is obtained from an independent timing-calibration set generated from the same signal distribution as $\mathcal{D}_+$. All above-threshold positions are included, and noise-only background is subtracted statistically. For each trigger method and target stored-data fraction, W_{99} is calibrated separately in each velocity bin using the frozen threshold. The symmetric interval $[-W_{99}, W_{99}]$ contains 99% of the corresponding background-subtracted distribution. Figure 8 shows the Δt distribution of the compact template-bank trigger at a stored-data fraction of 10^{-3} .

A signal-plus-noise record is classified as correctly triggered if at least one above-threshold position satisfies $|\Delta t| \leq W_{99}$; otherwise, it is classified as missed. The velocity-dependent W_{99} is used only in this offline evaluation; the online trigger does not use it. The net acceptance P_{net} is the fraction of correctly triggered signal-plus-noise records minus the fraction of paired noise-only records satisfying the same timing criterion. The mean P_{net} is evaluated over all records of $\mathcal{D}_+$ and thus averages over the sampled velocity distribution.

Using the calibrated threshold and the timing criterion above, Fig. 9 shows the velocity-dependent net acceptance P_{net} of the compact template-bank trigger and the conventional template-bank trigger at a stored-data fraction of 10^{-3} . For the compact template-bank trigger, P_{net} is lowest in the slowest velocity bin and increases from 0.96% to approximately 8.25%–9.27% over the intermediate and higher velocity bins. Because the trajectories are sampled through the 10-cm sphere, P_{net} also includes the geometric acceptance of the coil. Many of these trajectories pass outside the coil or cross only part of its turns, so only a small fraction of them produce detectable signals, and the values of P_{net} are relatively small. The compact template-bank trigger gives higher P_{net} than the conventional template-bank trigger in every velocity bin. Its mean P_{net} is 7.52%, compared with 6.72% for the conventional template-bank trigger, corresponding to an 11.9% relative improvement while using fewer templates. The 95% confidence-interval half-width of each mean P_{net} value is about 0.05 pp, well below the differences between the triggers.

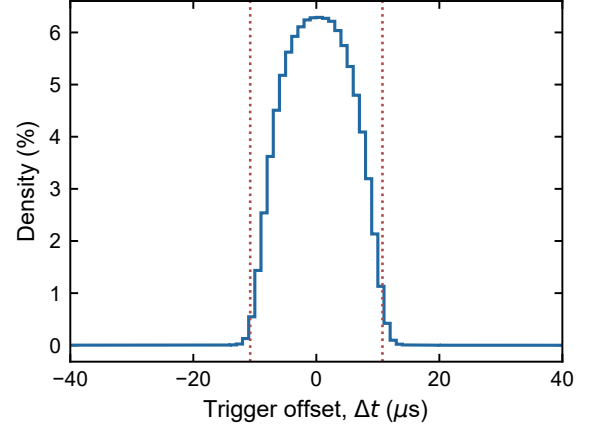


Fig. 8. Background-subtracted trigger-offset distribution in 1- μ s bins for the compact template-bank trigger at $\varepsilon_0 = 10^{-3}$. The dotted lines mark the pooled descriptive value $W_{99} = 10.75 \mu\text{s}$.

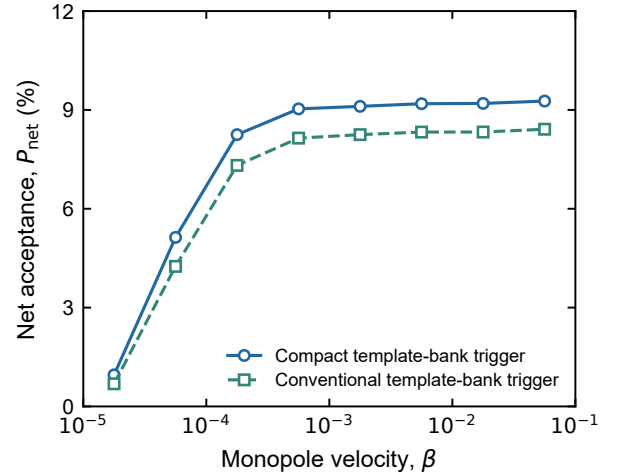


Fig. 9. Velocity-dependent net acceptance P_{net} of the compact template-bank trigger (113 templates) and the conventional template-bank trigger (601 templates) at a stored-data fraction of 10^{-3} . Both methods use separately calibrated thresholds and timing windows. The evaluation uses the same 3.2×10^6 signal-plus-noise records.

E. MLP Trigger

The compact template-bank trigger uses a practical approximation to the fixed-parameter LRT and therefore does not retain its Neyman-Pearson optimality. Although the template bank provides high empirical coverage under the stated sampling distribution, coverage alone does not guarantee the highest acceptance at a fixed stored-data fraction. Moreover, the compact template-bank trigger as noted in Section II-C, the compact template-bank trigger is not Neyman-Pearson optimal, and it retains only the largest response magnitude in Eq. (6), discarding the information carried by the remaining template responses. To address these limitations, the MLP trigger is trained to approximate a likelihood-ratio decision rule based on the vector of absolute

responses from all templates, thereby learning how to weight and combine all template outputs rather than relying only on their maximum [29].

At each sampling time, the absolute responses of all filters in the template bank form the input vector to the MLP trigger. The network maps this vector through two fully connected (FC) hidden layers with 16 and 8 units, respectively, followed by one output logit. Both hidden layers use the rectified linear unit (ReLU) activation [30], as illustrated in Fig. 10.

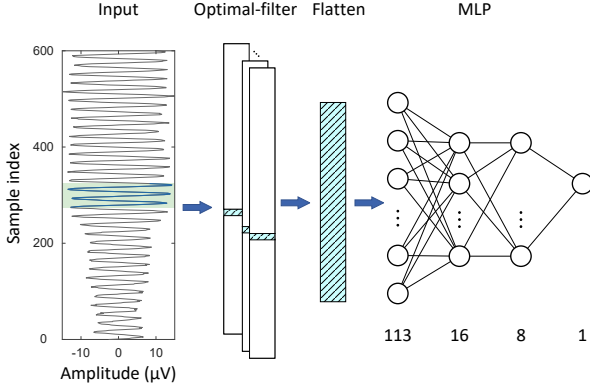


Fig. 10. Architecture of the MLP trigger. The absolute responses of all 113 templates form the input vector, which is processed by FC layers with 16 and 8 hidden units and one output logit. The hidden layers use ReLU activations.

Only the MLP weights and biases are trained, whereas the template bank is kept fixed. The training set contains 2×10^6 positive and 4×10^6 negative waveform windows, generated with the signal and noise models of Section II-A. A signal-injected window is labeled positive when its center lies within $3 \mu\text{s}$ of the true crossing time. Pure-noise windows are labeled negative, as are signal-injected windows whose centers are more than $10 \mu\text{s}$ from the true crossing time; signal-injected windows with offsets between 3 and $10 \mu\text{s}$ are excluded. An independent validation set of 4×10^5 positive and 8×10^5 negative windows, with the same signal distribution and labeling rules, is used to monitor overfitting and select the checkpoint. The networks are trained with ordinary binary cross-entropy. The network is trained in 32-bit floating point by minimizing the binary cross-entropy with the AdamW optimizer [31], using an initial learning rate of 10^{-3} , a weight decay of 10^{-4} , and a batch size of 1920 windows. The learning rate is halved whenever the validation loss has not improved for three epochs, and the training runs for 24 epochs.

Each frozen network is evaluated using the same streaming procedure as in Section II-D. Figure 11 compares the MLP trigger with both template-bank triggers for $L = 31$. The MLP trigger provides higher net acceptance P_{net} in every velocity bin, with increases of 0.42–1.46 pp over the compact template-bank trigger. Over the complete evaluation set, the mean P_{net} increases from 7.52% to 8.79%, corresponding to a relative improvement of 16.8%.

The waveform-window length creates a tradeoff in stream-level triggering. An above-threshold window causes its corresponding L -sample interval to be retained, so increasing L

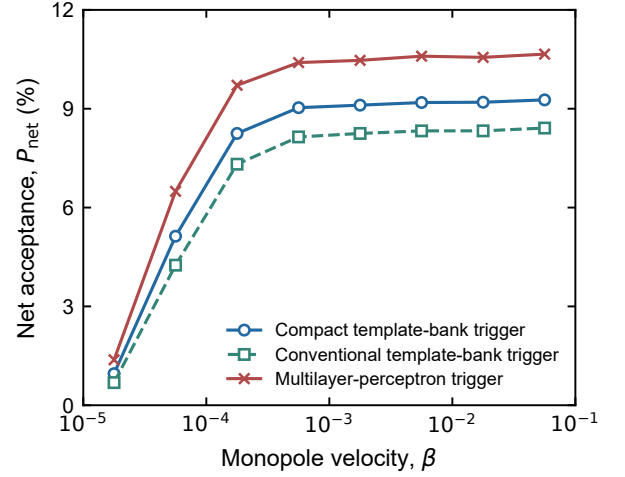


Fig. 11. Velocity-dependent net acceptance P_{net} of the MLP trigger, compact template-bank trigger (113 templates), and conventional template-bank trigger (601 templates) for $L = 31$ at a stored-data fraction of 10^{-3} . All three methods are evaluated using the same 3.2×10^6 signal-plus-noise records.

increases the amount of data saved per trigger. In contrast, an excessively short window captures less signal energy and makes signal-noise discrimination more difficult. An intermediate value of L may therefore provide the highest acceptance at the fixed stored-data fraction. The candidate lengths are $L \in \{11, 21, 31, 41, 51, 71, 101\}$ samples. To determine this length, MLP triggers are trained for seven window lengths between 11 and 101 samples, with three random seeds for each length. The frozen networks for all lengths are evaluated on the same 3.2×10^6 signal-plus-noise records, and the length giving the largest mean net acceptance P_{net} is selected. Figure 12 summarizes the length scan; for each length, the three networks give mean P_{net} values within about 0.01 pp. The $L = 31$ configuration gives the largest mean net acceptance P_{net} in the length-selection study and is selected for the subsequent FPGA implementation.

III. FPGA IMPLEMENTATION OF THE NEURAL-NETWORK TRIGGER

A. Overall Architecture

To process the continuous waveform stream in real time, the MLP trigger described in Section II-E is implemented on an FPGA the XC7A200T FPGA of a Xilinx AC701 board [32], [33]. The target input is a 20-bit waveform sampled at 1 MHz, while the trigger logic operates at 200 MHz on a Xilinx AC701 board equipped with an XC7A200T FPGA [32], [33].

Each input sample is assigned a sample index and sent to two paths. The waveform-cache path continuously writes the raw samples to a circular random-access memory (RAM). In the trigger path, a shift register maintains the current $L = 31$ -sample window. The 113 OF kernels are stored on chip and applied to this window. Their absolute responses at the current sampling time form the 113-dimensional input to the MLP trigger. The resulting output

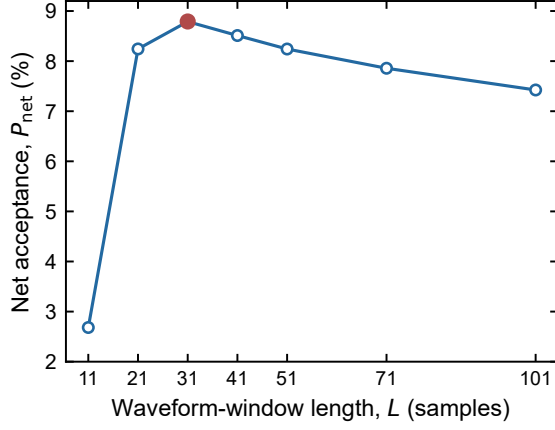


Fig. 12. Mean net acceptance P_{net} of the MLP trigger in the length-selection study as a function of the waveform-window length at a stored-data fraction of 10^{-3} . All lengths use the same 3.2×10^6 signal-plus-noise records. Open markers show the average over three training seeds. The highlighted marker identifies the selected configuration. The red marker shows the network implemented on the FPGA, $L = 31$.

logit is compared directly with a programmable threshold to produce the trigger decision. When the logit exceeds the threshold, the corresponding window start index is passed to the cluster-merging logic. Consecutive accepted windows are merged into one event cluster. A first-in, first-out (FIFO) queue stores the cluster boundaries and trigger metadata, and the waveform reader retrieves the corresponding raw samples from the circular RAM. The event formatter then combines the metadata and waveform samples into the record sent to the downstream data-acquisition system. Figure 13 shows the overall FPGA data flow.

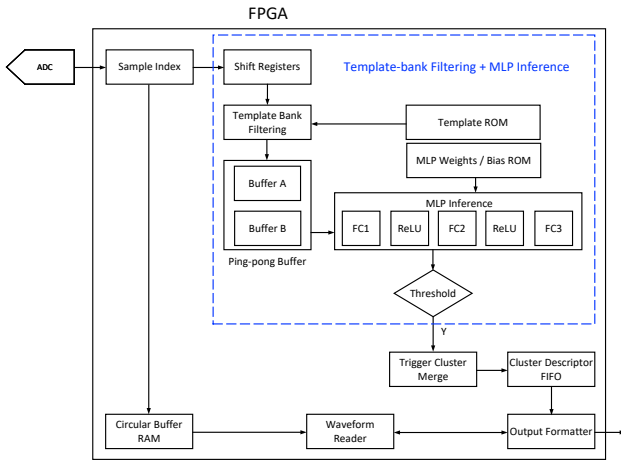


Fig. 13. Overall FPGA architecture for the streaming MLP trigger. ADC: analog-to-digital converter; FC: fully connected; ROM: read-only memory.

B. Template-Bank Matching Filtering and MLP Evaluation Inference

For each template η , the OF response is computed as

$$\mathbf{h}_{\eta}^{\top} \mathbf{x} = \sum_{i=0}^{L-1} h_{\eta,i} x_i, \quad (9)$$

where x_i is an input sample and $h_{\eta,i}$ is the corresponding discrete-time filter coefficient. In the fixed-point implementation, the input samples are signed 20-bit values, the filter coefficients are quantized to signed 18-bit integers, and the products are accumulated in 48 bits. With 113 templates and $L = 31$, evaluating Eq. (9) requires $3503 \times 31 = 3503$ coefficient multiplications for each new sample. The 113 signed responses are converted to their absolute values before entering the neural network and stored as signed 20-bit values with 12 fractional bits.

The 113-dimensional response vector is processed by FC layers with dimensions $113 \rightarrow 16 \rightarrow 8 \rightarrow 1$, with ReLU activations after the two hidden layers. The FC layers are evaluated with time-multiplexed multiply-accumulate (MAC) lanes. Instantiating one multiplier for every weight would be unnecessary at the 1 MHz input rate and would exceed the practical resource budget of the target FPGA. The design therefore uses parallel MAC lanes that are reused across templates and neurons, consistent with the MAC and pipelining capabilities of the 7-series digital signal processing (DSP) blocks [34]. Each neuron computes

$$y = \sum_k w_k u_k + b, \quad (10)$$

where u_k are the inputs of the neuron, w_k are the corresponding weights, and b is its bias. The inputs of the first layer are the absolute template responses, and the output of the last layer is the output logit. The weights are signed 18-bit integers, and the products are accumulated in 48 bits together with the biases, which are stored at the accumulator scale. The hidden activations use the same 20-bit format as the template responses, and the output logit has 32 bits with 18 fractional bits. Evaluating (10) for all neurons requires a further $113 \times 16 + 16 \times 8 + 8 = 1944$ multiplications per sample. Filter coefficients, network weights, and biases are stored in on-chip read-only memories (ROMs) and read in tiles [35]. Partial sums are accumulated for each output channel, after which the completed layer output is written to a buffer and passed to the next stage. The ReLU activations are implemented as comparisons with zero.

Because the input is sampled at 1 MHz and the trigger logic runs at 200 MHz, 200 clock cycles are available for processing each input sample. The computation is organized as a two-stage pipeline, with the template-bank filter followed by the MLP, and the multiplications of both stages are performed by DSP48E1 slices [34] that are reused over these clock cycles.

In the first stage, the template bank is evaluated one template per clock cycle. In each cycle, the 31 coefficients of one template are multiplied by the window samples in 31 parallel multipliers, which occupy 31 DSP48E1 slices. A five-level

pipelined adder tree sums the 31 products, and the absolute value of the sum is rounded, saturated, and written to a feature buffer. Each multiplier therefore performs 113 multiplications per sample. Including the delay of the adder tree and the output processing, the first stage has a latency of 127 clock cycles.

In the second stage, the MLP is evaluated by 16 multiply-accumulate (MAC) lanes, which occupy 16 DSP48E1 slices and are reused across the three FC layers. FC1, FC2, and FC3 take 113, 8, and 1 clock cycles, respectively. Including the delay of the memory reads, bias addition, rounding, and the ReLU activations, the second stage has a latency of 161 clock cycles.

Because both stages are shorter than the 200-cycle sample interval, ping-pong feature buffers allow the MLP to process the responses of one sample while the filter processes the next. The trigger therefore accepts one sample every 200 clock cycles, with a latency of 289 cycles (1.445 μ s) from the arrival of the last window sample to the trigger decision, including one cycle for the buffer handoff. In total, the two stages use $31 + 16 = 47$ DSP48E1 slices, whereas the remaining additions, such as the adder tree, are implemented in general logic. The internal filter and neural-network datapath is illustrated in Fig. 14.

C. On-Board Validation

The on-board test waveforms are represented by signed 20-bit samples and are replayed from a host computer over Gigabit Ethernet using SiTCP [36]. On the FPGA, the received samples are buffered in a FIFO and released to the trigger at a fixed interval of 200 clock cycles, emulating an analog-to-digital converter sampling at 1 MHz, and the trigger records are returned to the host over the same link. The same quantized inputs are supplied to the bit-accurate software reference, which reproduces the fixed-point template-bank and MLP datapaths. The floating-point model provides the performance reference.

The generated bitstream is successfully programmed on the AC701 board. The trigger-core resource utilization is summarized in Table I. The implementation meets the timing constraints at 200 MHz. The implementation meets the timing constraints at 200 MHz on the AC701 board, and its trigger-core resource utilization is summarized in Table I.

TABLE I
FPGA TRIGGER-CORE RESOURCE UTILIZATION. BRAM: BLOCK
RANDOM-ACCESS MEMORY; DSP: DIGITAL SIGNAL PROCESSING; LUT:
LOOK-UP TABLE.

Resource	Used	Utilization (%)
LUT	41,006	9.458
Register	15,357	12.770
BRAM tile	12	5.14
DSP48E1	7247	9.736

Figure 15 shows a representative signal-plus-noise record used in the on-board validation. All 27 above-threshold FPGA scores and their window indices agree exactly with the bit-accurate fixed-point reference. Relative to the floating-point output, the maximum absolute score difference is 4.12×10^{-4} .

The FPGA trigger was evaluated using the same $\mathcal{D}_-$ and $\mathcal{D}_+$ as in Section II-D, at a stored-data fraction of 10^{-3} . On-board tests confirmed continuous processing at 1 MHz without dropped samples. The FPGA trigger is evaluated with the same $\mathcal{D}_-$ and $\mathcal{D}_+$ as in Section II-D. All samples of both sets are replayed through the board without any dropped sample, using the threshold calibrated for a stored-data fraction of 10^{-3} . The above-threshold outputs and window indices, the event boundaries, and the saved waveforms returned by the board agree exactly with the bit-accurate fixed-point reference. Figure 16 compares the velocity-resolved net acceptance P_{net} of the floating-point reference and the FPGA fixed-point implementation. The fixed-point implementation gives a net acceptance lower than the floating-point reference by only $3/(3.2 \times 10^6)$. In absolute terms, and only 8 of the 3.2×10^6 signal-plus-noise records are classified differently. The overall mean values of P_{net} are 8.7911% for the floating-point reference and 8.7910% for the FPGA fixed-point configuration measured on the board. These tests verify the functional correctness and real-time operation of the FPGA trigger.

IV. CONCLUSIONS

A compact template-bank construction method has been developed for the SCEP L-H trigger using the absolute loss in detection probability as the coverage criterion. By adding the worst-covered waveform at each iteration, the method constructs a 113-template bank for $L = 31$ samples, with empirical coverage above 99.99% on an independent audit set. Compared with the conventional template-bank trigger, the compact template-bank trigger increases the mean P_{net} from 6.72% to 7.52% at the same stored-data fraction, an 11.9% relative improvement.

To exploit the information beyond the largest template response, an MLP trigger combines the absolute responses of all templates. A waveform-window length scan then identifies $L = 31$ as the configuration with the highest mean net acceptance. At a stored-data fraction of 10^{-3} , the MLP improves the mean P_{net} from 7.52% to 8.79%, corresponding to a relative increase of 16.8% over the compact template-bank trigger.

The resulting network is implemented on a Xilinx AC701 FPGA platform for real-time processing of 20-bit waveforms sampled at 1 MHz, with trigger logic operating at 200 MHz. The template-bank filter and the MLP form a two-stage pipeline that uses 47 DSP48E1 slices and has a latency of 1.445 μ s. On the board, all samples of $\mathcal{D}_-$ and $\mathcal{D}_+$ are processed without any dropped sample, and the returned outputs agree exactly with the bit-accurate fixed-point reference. Evaluation using the same $\mathcal{D}_-$ and $\mathcal{D}_+$ gives a reduction in P_{net} of only $3/(3.2 \times 10^6)$ relative to the floating-point reference. The fixed-point implementation reduces P_{net} by only $3/(3.2 \times 10^6)$ in absolute terms compared with the floating-point reference. Together, the acceptance comparison, on-board validation, and timing verification demonstrate the correctness and real-time operation of the FPGA-based L-H trigger.

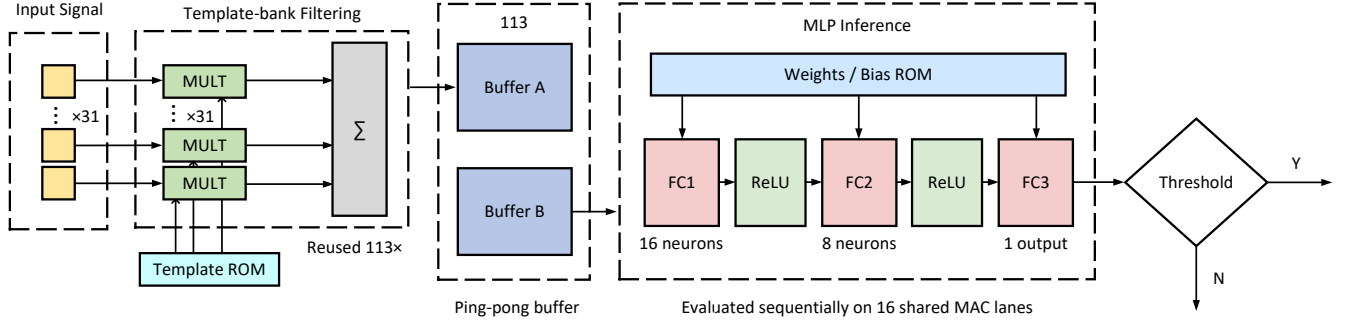


Fig. 14. Internal FPGA datapath for the streaming MLP trigger of the two-stage trigger pipeline. The input waveform is processed by the 113-template filter bank, converted to absolute responses, and passed through the $113 \rightarrow 16 \rightarrow 8 \rightarrow 1$ MLP network and a threshold comparison. In the template-bank filter, 31 multipliers and an adder tree evaluate one template per clock cycle and are reused for all 113 templates. The absolute responses are written to one half of a ping-pong buffer while the MLP reads the other half, and the two halves swap roles for every input sample. The three FC layers are evaluated sequentially on 16 shared MAC lanes, followed by the threshold comparison. MAC: multiply-accumulate; MULT: multiplier; ReLU denotes the ReLU activation.

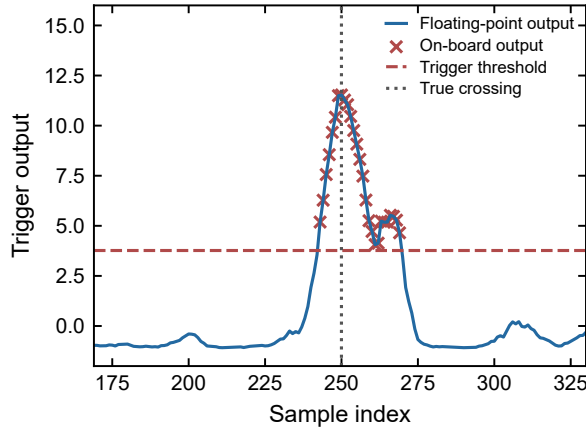


Fig. 15. On-board output validation for a representative $L = 31$ signal-plus-noise record. The floating-point MLP output is shown as a solid line, and the crosses mark the above-threshold FPGA outputs. The vertical dotted line indicates the true MM crossing time, and the horizontal dashed line indicates the FPGA threshold calibrated for a stored-data fraction of 10^{-3} .

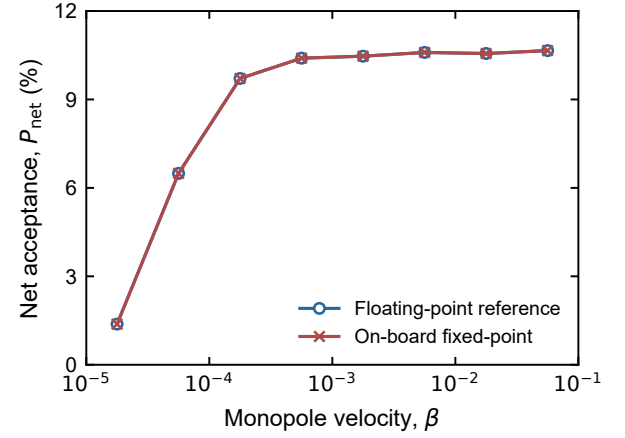


Fig. 16. Velocity-resolved net acceptance P_{net} comparison for the $L = 31$ MLP trigger at a stored-data fraction of 10^{-3} . The floating-point reference is compared with the selected FPGA fixed-point configuration, using 3.2×10^6 signal-plus-noise records.

ACKNOWLEDGEMENTS

This project is supported by grants from National Science Foundation of China (No. 12250011), and by the Frontier Scientific Research Program of Deep Space Exploration Laboratory under grant No. 2022-QYKYJH-HXYF-013.

REFERENCES

- [1] P. A. M. Dirac, “Quantised singularities in the electromagnetic field,” *Proc. R. Soc. Lond. A*, vol. 133, no. 821, pp. 60–72, 1931.
- [2] H. Georgi and S. L. Glashow, “Unity of all elementary-particle forces,” *Phys. Rev. Lett.*, vol. 32, no. 8, pp. 438–441, 1974.
- [3] J. Preskill, “Magnetic monopoles,” *Annu. Rev. Nucl. Part. Sci.*, vol. 34, no. 1, pp. 461–530, 1984.
- [4] G. ’t Hooft, “Magnetic monopoles in unified gauge theories,” *Nucl. Phys. B*, vol. 79, no. 2, pp. 276–284, 1974.
- [5] P. H. Eberhard, R. R. Ross, and J. D. Taylor, “Improvements of a magnetic monopole detector,” *Rev. Sci. Instrum.*, vol. 46, no. 4, pp. 362–366, 1975.
- [6] B. Cabrera, “First results from a superconductive detector for moving magnetic monopoles,” *Phys. Rev. Lett.*, vol. 48, no. 20, pp. 1378–1381, 1982.
- [7] M. Ambrosio *et al.*, “Final results of magnetic monopole searches with the MACRO experiment,” *Eur. Phys. J. C*, vol. 25, no. 4, pp. 511–522, 2002.
- [8] R. Abbasi *et al.*, “Search for relativistic magnetic monopoles with IceCube,” *Phys. Rev. D*, vol. 87, no. 2, p. 022001, 2013.
- [9] L. Patrizii and M. Spurio, “Status of searches for magnetic monopoles,” *Annu. Rev. Nucl. Part. Sci.*, vol. 65, no. 1, pp. 279–302, 2015.
- [10] L. Patrizii, Z. Sahnoun, and V. Togo, “Searches for cosmic magnetic monopoles: past, present and future,” *Philos. Trans. R. Soc. A*, vol. 377, no. 2161, p. 20180328, 2019.
- [11] C. Ye *et al.*, “Detector design proposal for a cosmic magnetic monopole search experiment,” *Phys. Rev. D*, vol. 111, no. 2, p. 023020, 2025.
- [12] D. S. Akerib *et al.*, “FPGA-based trigger system for the LUX dark matter experiment,” *Nucl. Instrum. Methods Phys. Res. A*, vol. 818, pp. 57–67, 2016.
- [13] J. Duarte *et al.*, “Fast inference of deep neural networks in FPGAs for particle physics,” *J. Instrum.*, vol. 13, no. 07, p. P07027, 2018.
- [14] J. Aasi *et al.*, “Advanced LIGO,” *Class. Quantum Grav.*, vol. 32, no. 7, p. 074001, 2015.
- [15] B. P. Abbott *et al.*, “Observation of gravitational waves from a binary black hole merger,” *Phys. Rev. Lett.*, vol. 116, no. 6, p. 061102, 2016.
- [16] G. L. Turin, “An introduction to matched filters,” *IRE Trans. Inf. Theory*, vol. 6, no. 3, pp. 311–329, 1960.
- [17] T. Cokelaer, “A template bank to search for gravitational waves from

- inspiralling compact binaries: II. Phenomenological model,” *Class. Quantum Grav.*, vol. 24, no. 24, pp. 6227–6242, 2007.
- [18] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, “FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries,” *Phys. Rev. D*, vol. 85, no. 12, p. 122006, 2012.
- [19] S. Babak *et al.*, “Searching for gravitational waves from binary coalescence,” *Phys. Rev. D*, vol. 87, no. 2, p. 024033, 2013.
- [20] B. J. Owen, “Search templates for gravitational waves from inspiraling binaries: Choice of template spacing,” *Phys. Rev. D*, vol. 53, no. 12, pp. 6749–6761, 1996.
- [21] B. J. Owen and B. S. Sathyaprakash, “Matched filtering of gravitational waves from inspiraling compact binaries: Computational cost and template placement,” *Phys. Rev. D*, vol. 60, no. 2, p. 022002, 1999.
- [22] C. Van Den Broeck *et al.*, “Template banks to search for compact binaries with spinning components in gravitational wave data,” *Phys. Rev. D*, vol. 80, no. 2, p. 024009, 2009.
- [23] R. P. Croce *et al.*, “Correlator bank detection of gravitational wave chirps—false-alarm probability, template density, and thresholds: Behind and beyond the minimal-match issue,” *Phys. Rev. D*, vol. 70, no. 12, p. 122001, 2004.
- [24] J. Neyman and E. S. Pearson, “On the problem of the most efficient tests of statistical hypotheses,” *Philos. Trans. R. Soc. Lond. A*, vol. 231, no. 694–706, pp. 289–337, 1933.
- [25] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1998.
- [26] S. Somalwar, H. Frisch, and J. Incandela, “Transient-response induction detectors for magnetic monopoles: First operation at 78 K,” *Phys. Rev. D*, vol. 37, no. 9, pp. 2403–2418, 1988.
- [27] T. F. Gonzalez, “Clustering to minimize the maximum intercluster distance,” *Theor. Comput. Sci.*, vol. 38, pp. 293–306, 1985.
- [28] I. W. Harry, B. Allen, and B. S. Sathyaprakash, “Stochastic template placement algorithm for gravitational wave data analysis,” *Phys. Rev. D*, vol. 80, no. 10, p. 104014, 2009.
- [29] K. Cranmer, J. Pavez, and G. Louppe, “Approximating likelihood ratios with calibrated discriminative classifiers,” 2015, arXiv:1506.02169.
- [30] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proc. 14th Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2011, pp. 315–323.
- [31] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- [32] *AC701 Evaluation Board for the Artix-7 FPGA User Guide (UG952)*, Xilinx, Inc., San Jose, CA, USA, 2019, v1.4.
- [33] *7 Series FPGAs Data Sheet: Overview (DS180)*, Xilinx, Inc., San Jose, CA, USA, 2020, v2.6.1.
- [34] *7 Series DSP48E1 Slice User Guide (UG479)*, Xilinx, Inc., San Jose, CA, USA, 2018, v1.10.
- [35] *7 Series FPGAs Memory Resources User Guide (UG473)*, Xilinx, Inc., San Jose, CA, USA, 2019, v1.14.
- [36] T. Uchida, “Hardware-based TCP processor for Gigabit Ethernet,” *IEEE Trans. Nucl. Sci.*, vol. 55, no. 3, pp. 1631–1637, 2008.