Subarashi: A 16nm Sustainable SNN Processor with Cell-Composition-Based Radiation Resilience and Reconfigurable Array for Harsh Environments

Quan Cheng†, Haoyuan Li†, Masanori Hashimoto*
Department of Informatics, Kyoto University, Kyoto, Japan
†Equally Credited Authors (ECAs) *Email: hashimoto@i.kyoto-u.ac.jp

***Abstract*—Spiking neural network (SNN) processors are promising for edge AI in harsh environments due to their event-driven operation and improved algorithm-level robustness. However, conventional implementations remain vulnerable at the hardware level, particularly in control circuitry, where single-event upsets (SEU) and single-event transients (SET) can cause functional errors or even system failures. In addition, the diverse operator requirements of AI workloads often lead to suboptimal processing-element (PE) utilization. To address these challenges, this work presents Subarashi, a 16-nm SNN processor that jointly improves reliability and resource efficiency. To enhance soft-error resilience, Subarashi employs a cell-composition flow that selectively reconstructs sequential logic for SEU/SET mitigation in critical logic paths. To improve utilization across diverse operators and layer shapes, the processor further incorporates reconfigurable datapaths at both the PE and array levels, enabling efficient on-the-fly attention computation and workload-adaptive array shaping. Silicon measurements demonstrate that Subarashi achieves up to 101.67 TOPS/W peak energy efficiency and 50.83–68.70 TOPS/W across end-to-end NN evaluations. Under alpha irradiation, the design suppresses 97.6% of critical system failures with merely 9.68% area overhead, validating its effectiveness for reliable edge AI deployment in harsh environments.**

***Index Terms*—spiking neural network, cell composition, soft-error mitigation, reconfigurable architecture, radiation resilience**

I. Introduction

Recent advances in artificial intelligence (AI) have driven the rapid development of energy-efficient edge computing systems, particularly spiking neural network (SNN) processors, which offer inherent advantages in event-driven computation and low-power operation [1]–[5]. Compared with conventional neural network accelerators, spike-based systems are especially promising for deployment in radiation-prone harsh environments, where spike-based signaling has been shown to exhibit improved robustness [6], [7].

However, despite the intrinsic resilience of spike-based computation, practical SNN hardware implementations still face two critical challenges (Fig. 1 top). First, control circuitry within the processor remains vulnerable to single-event upsets (SEUs) and single-event transients (SETs), which can lead to functional errors or even system-level failures under radiation exposure. Second, due to the diverse operator requirements in AI workloads, processing elements (PEs) often suffer from underutilization, resulting in suboptimal energy efficiency and limited adaptability across different neural network models.

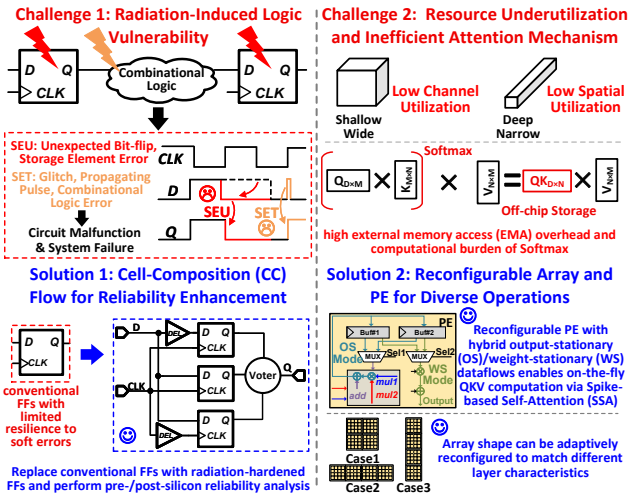


Fig. 1: Motivation and design overview of Subarashi.

To address these challenges, this work proposes **Subarashi**, a 16-nm **su**stainable SNN processor with cell-composition-**ba**sed **r**adiation resilience and reconfigurable **a**rray for har**sh** env**i**ronments. The proposed approach enables systematic reconstruction at both the control and datapath levels (Fig. 1 bottom). For sequential control logic (e.g., flip-flops (FFs)), the proposed cell-composition flow introduces structural redundancy to enhance robustness against SEU/SET-induced soft errors. For the PE array, two levels of reconfiguration are supported. First, intra-PE reconfiguration enables on-the-fly QKV computation, eliminating the need for explicit attention-cache storage. Second, the array shape can be adaptively reconfigured to match different layer characteristics, improving overall PE utilization. Together, these techniques improve both reliability and computational efficiency.

II. Proposed Design

A. System Architecture of Subarashi

Fig. 2 presents the architecture of Subarashi, a 16-nm SNN processor for reliable edge AI. The system integrates a full RISC-V sub-system with rich peripherals as the main controller, along with a digital PLL and on-chip memory banks. For reliability, only the I/DTCM features ECC protection,

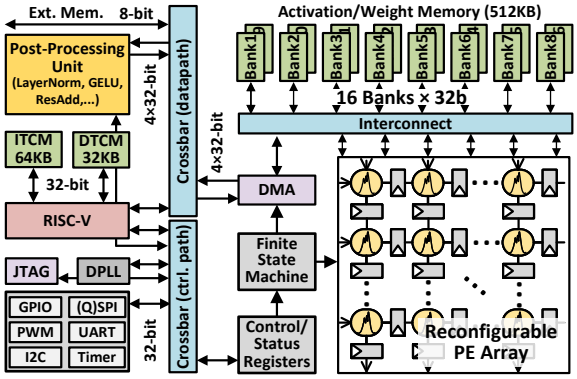


Fig. 2: System architecture of Subarashi.

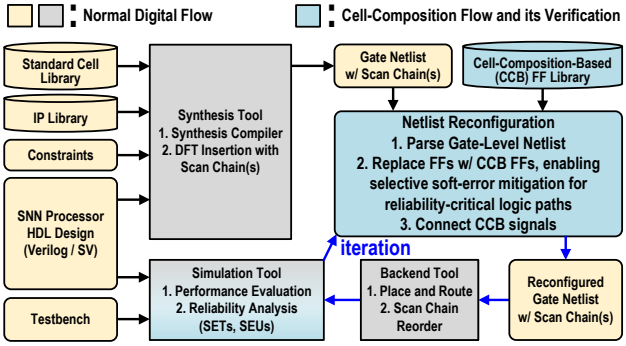
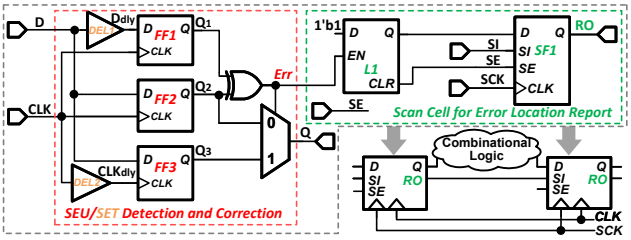


Fig. 3: Cell-composition flow for reliable Subarashi.

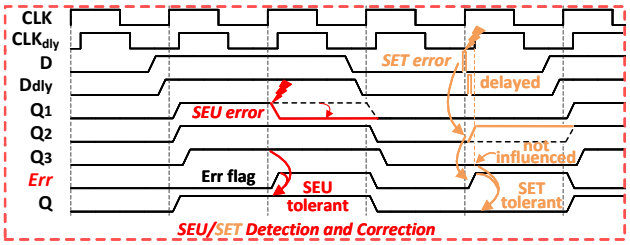
while the 512KB activation/weight memory pool remains unprotected but can be dynamically allocated between activations and weights based on application requirements. To enhance soft-error resilience, the entire digital logic is hardened using the proposed cell-composition (CC) flow, which reconstructs sequential elements to provide soft-error protection against SEU/SET. The processor features a spike-based reconfigurable PE array comprising 256 PEs, with each PE containing two multiply-accumulate (MAC) units for ping-pong operations. The datapath supports flexible dataflow configurations to accommodate diverse computational patterns, enabling adaptive mapping of different operators and improving PE utilization.

B. Cell-Composition Flow and Soft-Error Mitigation

Fig. 3 illustrates the CC flow for reliable Subarashi, which integrates soft-error resilience into conventional digital design flows at the netlist level with minimal disruption. After synthesis, the CC flow introduces a netlist-level reconstruction stage that parses the gate-level netlist and selectively replaces conventional FFs with CC-based (CCB) FFs. These CCB FFs enable soft-error mitigation against SEU/SET in targeting paths. Unlike traditional uniform hardening approaches, the proposed method adopts a scenario-driven replacement strategy. Pre-silicon reliability analysis techniques are employed to evaluate and rank circuit components based on their functional criticality. Components with higher error tolerance or lower reliability requirements (e.g., buffering FFs in AI accelerators) can be assigned lightweight configurations or left unprotected.



(a) Example of a CCB FF targeting SEU/SET Mitigation.



(b) Timing diagram illustrating SEU/SET detection and recovery.

Fig. 4: Cell-composition-based FF for SEU/SET mitigation.

TABLE I: Flip-Flops in Cell-Composition Flow.

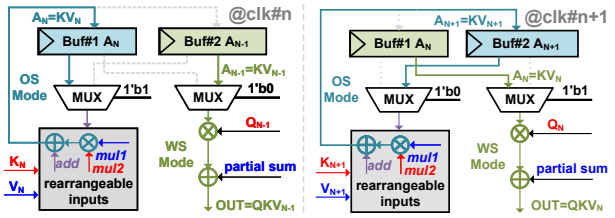
Flip-Flop Type	Configuration
SEU/SET-tolerant FF	Full configuration (as shown in Fig. 4)
SEU-tolerant FF	DEL1/2 (SET mitigation) are removed
Normal/Scan FF	SEU/SET detection and correction logic are removed

*Replacing conventional FFs with CCB FFs incurs an area overhead of 9.68%
*All CCB FFs are constructed by composing standard cell library elements.
*Users can further optimize area and power by designing application-specific FFs.

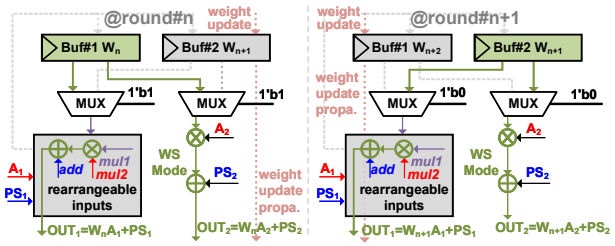
In contrast, critical logic paths can employ fully CCB FFs with enhanced error detection and recovery capability.

Fig. 4a illustrates an example of the proposed CCB FFs. It shows the most comprehensive FF variant that integrates key protection mechanisms, including detection and correction logic for SEU/SET (red) and scan cell (green) for error localization. The red part employs FF1–FF3 combined with XOR and MUX logic to protect against SEUs, while DEL1 and DEL2 associated with FF1 and FF3 introduce selective delays to clock and data paths for SET mitigation, where the delay values trade off SET filtering capability against timing overhead. The green part is used solely for error logging. It can be integrated into the scan chain to capture error information for reliability analysis. In practical deployments, however, scan cells can be omitted to reduce overhead. Besides, Fig. 4b shows the timing diagram of SEU and SET mitigation.

To support diverse application scenarios, Table I summarizes how different configurations of the CC flow can be instantiated to balance protection requirements and overhead. For instance, when only SEU tolerance is required, the delay elements for SET mitigation (DEL1/2) can be removed. For applications such as cache memories in AI accelerators, where algorithm-level redundancy can tolerate soft errors, SEU/SET protection logic may be omitted to minimize overhead. In contrast, for reliability-critical control paths, the full configuration with error detection and correction is retained to ensure robust operation. By selecting appropriate FF configurations based on pre-silicon reliability analysis, in which gate-level



(a) OS/WS hybrid dataflow enabling on-the-fly QKV computation.



(b) WS-only dataflow for general convolution and matrix operations.

Fig. 5: Reconfigurable dataflows in processing element.

fault injection (FI) is conducted across the RISC-V core, accelerator control logic, and computational datapaths, vulnerability hotspots can be identified to enable targeted reliability optimization. Scenario-driven hardening can be achieved. This approach enables system-level resilience improvements with minimized area, power, and timing overhead. Replacing conventional FFs with CCB FFs in Subarashi incurs an area overhead of 9.68% (excluding scan cells) for 26,164 FF replacements, demonstrating a scalable and cost-effective soft-error mitigation strategy.

C. Reconfigurable Array and PE for Diverse Operations

Fig. 5 illustrates the reconfigurable dataflows supported by the proposed PE, which is designed to natively support Spike-based Self-Attention (SSA). By exploiting the SSA property that the QKV operations can be performed in a chained manner without Softmax operation [3], [8], the PE directly maps the QKV computation flow onto hardware, enabling seamless QKV multiplication without intermediate cache storage. This avoids attention-cache overhead while improving the efficiency of data movement and computation. Specifically, a hybrid output-stationary/weight-stationary (OS/WS) data flow is introduced (Fig. 5a). Within each PE, the two local buffers operate in complementary modes at the same time. In one cycle, one buffer works in OS mode to compute and locally retain the intermediate KV result, while the other buffer works in WS mode to process Q and multiply it with the previously buffered KV data. In the next cycle, the two buffers swap their roles. By alternately exchanging the OS/WS functionalities between the two buffers in a ping-pong manner, the PE enables on-the-fly QKV computation without explicitly storing the attention cache in external memory. On the other hand, Fig. 5b shows the WS-only dataflow for conventional convolution and matrix operations. In this mode, a ping-pong buffering mechanism is also employed, where one buffer participates in computation while the other is updated with new weights.

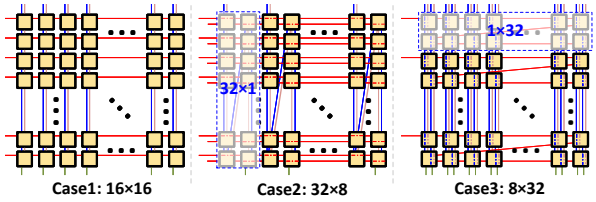


Fig. 6: Reconfigurable array for diverse workloads.

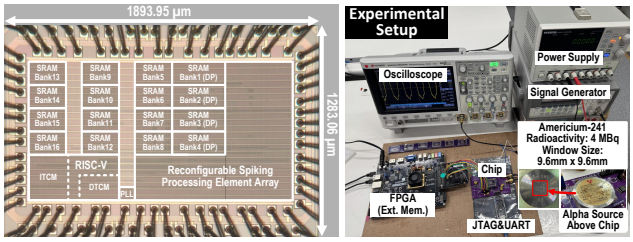


Fig. 7: Die photograph and test environment.

TABLE II: Event Statistics Across Different Experiments.

	Fluence(n/cm^2)	mSDC	cSDC	mDUE	cDUE
Baseline (FI exp.)	3.000e+12	24916	440	29843	2162
ECC-only (FI exp.)	3.000e+12	17142	512	35911	734
CCB+ECC (FI exp.)	3.000e+12	2379	1580	2047	38
CCB+ECC (Alpha exp.)	3.326e+12	3106	2329	3217	57

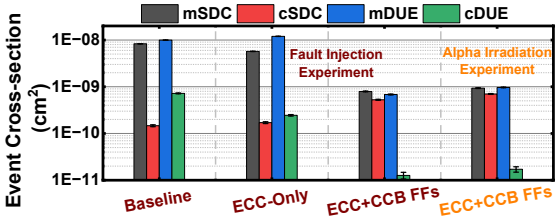


Fig. 8: Event cross-section across different experiments.

Once the current computation is completed, the updated buffer can immediately take over without additional data loading latency. This seamless alternation ensures continuous data supply to the PE, improving throughput and resource utilization.

The proposed design supports workload-adaptive array reconfiguration (Fig. 6). It supports three reconfigurable array modes, including 16×16, 32×8, and 8×32, to match different workload characteristics. Balanced workloads are mapped to the square configuration (16×16), while channel-dominated workloads benefit from wider arrays (8×32) to improve channel utilization. In contrast, spatially intensive or deep workloads are better supported by taller arrays (32×8), enhancing spatial utilization. By adaptively switching among these modes, the design improves overall PE utilization and computational efficiency across diverse workloads.

III. Fabricated Chip and Measurement Results

Fig. 7 shows the die photograph and experimental setup of the proposed Subarashi fabricated in a TSMC 16-nm FinFET CMOS process. An ^{241}Am alpha source with an activity of 4 MBq is used to evaluate its soft-error resilience.

To better assess the reliability, the soft error rates of memory banks and FFs are first measured as a reference. Under alpha

TABLE III: Performance Analysis on Different NN Models.

Model	DeiT-Small	ViT-Base	SDT-V2
Dataset	ImageNet-1k		
Accuracy [%]	74.89	75.04	76.24
Accuracy Loss [%]	5.01	3.20	4.06
Latency @ 1.1V [ms]	12.95	45.12	110.05
Task Performance @ 0.6V [TOPS/W]	59.20	68.70	50.83
External Memory Access Reduction [%]	20.57[a]	19.06[a]	NA[c]
Latency Reduction [%]	36.02[b]	31.43[b]	NA[c]

[a] Compared to an identical design w/o the proposed reconfigurable array and PE
[b] Compared to an identical model w/o SSA layer (i.e., w/ Softmax and QK^T storage)
[c] SDT-V2 is already an SSA-based NN, supporting on-the-fly QKV comp.

TABLE IV: Comparison with State-of-the-Art Designs.

	JSSC25 [2]	TCASAI25 [3]	VLSI24 [4]	CICC25 [5]	This work
Process	40nm	28nm	28nm	22nm	**16nm FinFET**
Archi.	SNN Accel.	Spiking ViT Accel.	BV-Transformer Accel.	16×RV32IM cores or Digital Accel.	**1×RV32IM core+ SNN Accel.**
Radiation Hardening	-	-	-	Adaptive DCLS[a] Boundary ReLU	**CCB FFs[b] w/ 9.68% area overhead**
Supply [V]	0.56-1.1	0.55-1.1	0.6-1.1	0.6-1.1	**0.6-1.1**
Frequency [MHz]	50-200	30-200	180-860	340-1225@Accel. 123-541@RISC-V	**342-990@Accel. 20-200@RISC-V**
Area [mm^2]	2.24	6.5	1.62	2.35	**Core Area: 1.36 Die Area: 2.43**
Power [mW]	2.84-43.58	9.04-222.65	51.9-960.3	3.73-118.05	**3.45-105.08**
Memory [KB]	192	640	66	608	**96 (I/DTCM w/ ECC) 512 (Memory Banks)**
Precision	INT1-8 (act.) INT8 (weight)	INT8	INT8	INT8	**8b[c]**
Power Efficiency [TOPS/W]	132.42[d] 42.8@Spikeformer	101.3[d] 57.7@ViT-Base	61.44[d] 25.01@DeiT-Small 29.66@Bert-Base	17.18[d] 14.72@SuperYOLO	**101.67[d] 59.20@DeiT-Small 68.70@ViT-Base 50.83@SDT-V2**

[a] Arch-dependent SEU resilience w/ partial coverage; [b] Arch-independent SEU/SET resilience w/ full coverage;
[c] 1b spikes with 8 timesteps; [d] Peak performance with full PE utilization.

irradiation at 0.7V, the cross-sections of the memory and FFs are 6.037×10^{-13} and 1.923×10^{-15} cm^2/bit, respectively. To further classify system-level reliability, four error categories are defined based on severity. A minor silent data corruption (mSDC) event refers to cases where intermediate computation errors occur but the final output remains correct. A critical SDC (cSDC) event corresponds to incorrect final outputs. A minor detected unrecoverable error (mDUE) denotes transient malfunctions, such as temporary stalls or non-responses in the computation array. In contrast, a critical DUE (cDUE) represents severe system failures, including RISC-V crashes.

Table II and Fig. 8 show the measured event statistics and cross-sections obtained from both gate-level FI and alpha irradiation experiments, using SD-Transformer V2 (SDT-V2) [8] as the benchmark. Three configurations are evaluated: 1) a baseline without protection, 2) ECC-only (i.e., I/DTCM protection), and 3) ECC combined with CCB FFs, where conventional FFs are replaced to mitigate SEU/SET effects. The results demonstrate that the ECC+CCB scheme significantly enhances system reliability. First, the cDUE cross-section is reduced by 42× (∼97.6% reduction) compared to the baseline. Second, the mDUE cross-section is reduced by approximately one order of magnitude with CCB FFs, confirming improved robustness of the computation array execution. In contrast, cSDC events remain relatively high, as ECC is not integrated in the activation/weight memory pool. Finally, a strong correlation coefficient (0.9866) between FI and irradiation results validates FI as a reliable proxy for reliability evaluation.

End-to-end NN executions are performed on multiple models (Table III). By leveraging the reconfigurable PE to enable on-the-fly QKV computation without explicit cache storage, and the reconfigurable array to adapt to diverse layer shapes, Subarashi achieves up to 36.02% latency reduction across different workloads. Table IV compares the proposed design with state-of-the-art works [2]–[5]. By leveraging the CC flow, this work provides full-chip SEU/SET resilience across all logic, which is not addressed in prior works, at an additional area overhead of 9.68%. Despite this area overhead, the impact on energy efficiency remains limited. In Subarashi, SET/SEU-hardened CCB FFs are primarily deployed in the low-frequency RISC-V domain, while SEU-only CCB FFs are used in the accelerator control logic, since SET-hardened CCB FFs reduce timing margin. For the main computational datapaths, such as the PEs, standard FFs are retained due to the inherent robustness of NNs. Thus, the proposed design still demonstrates compelling energy efficiency, achieving up to 101.67 TOPS/W peak energy efficiency and 1.19–4.67× energy efficiency across real end-to-end NNs.

IV. Conclusions

Subarashi is a 16-nm energy-efficient SNN processor designed for reliable edge AI in radiation-prone environments. The proposed CC flow systematically reconstructs sequential logic, enabling targeted soft-error mitigation against SEU/SET in critical control paths. In addition, the design incorporates reconfigurable datapath mechanisms at both the PE and array levels, enabling flexible computation patterns and adaptive array configurations to improve utilization across diverse workloads. Measurement results from the chip demonstrate 101.67 TOPS/W peak energy efficiency and 50.83–68.70 TOPS/W across end-to-end NN evaluations. Furthermore, the proposed design achieves 97.6% suppression of critical system failures under alpha irradiation at a cost of only 9.68% area overhead.

Acknowledgment

The authors would like to thank TSMC, Ltd. and TSMC Design Technology Japan, Inc. for the tape-out opportunity and the valuable support from Mr. Ayumu Yamada and Dr. Koji Nii.

References

[1] C. Frenkel and G. Indiveri, "ReckOn: A 28nm Sub-mm2 Task-Agnostic Spiking Recurrent Neural Network Processor Enabling On-Chip Learning over Second-Long Timescales," ISSCC, 2022.
[2] C. Fang et al., "An Energy-Efficient Unstructured Sparsity-Aware Deep SNN Accelerator With 3-D Computation Array," JSSC, 2025.
[3] X. Fang et al., "A Spiking Vision Transformer Accelerator with Energy-Efficient Attention Mechanism," TCASAI, 2025.
[4] C. Tang et al., "A 28nm 4.35TOPS/mm2 Transformer Accelerator with Basis-Vector-Based Ultra Storage Compression, Decomposed Computation and Unified LUT-Assisted Cores," Symp. VLSI Technology and Circuits, 2024.
[5] Q. Cheng et al., "A 22nm Resource-Frugal Hyper-Heterogeneous Multi-Modal System-on-Chip Towards In-Orbit Computing," CICC, 2025.
[6] J. Ding et al., "Neuromorphic computing paradigms enhance robustness through spiking neural networks," Nature Communications, 2025.
[7] Q. Cheng et al., "A Radiation-Hardened Neuromorphic Imager with Self-Healing Spiking Pixels and Unified Spiking Neural Network for Space Robotics," Symp. VLSI Technology and Circuits, 2025.
[8] M. Yao et al., "Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips," ICLR, 2024.