CIM-Based Parallel Fully FFNN Surface Code High-Level Decoder for Quantum Error Correction

Hao Wang[1*], Erjia Xiao[1*], Songhuan He[2], Zhongyi Ni[1], Lingfeng Zhang[1], Xiaokun Zhan[3], Yifei Cui[2], Jinguo Liu[1], Cheng Wang[2+], Zhongrui Wang[4+], Renjing Xu[1+]
[1]Hong Kong University of Science and Technology (Guangzhou), [2]University of Electronic Science and Technology of China, [3]Harbin Institute of Technology, [4]Southern University of Science and Technology
wangch87@uestc.edu.cn; wangzr@sustech.edu.cn; renjingxu@hkust-gz.edu.cn

***Abstract*—In all types of surface code decoders, fully neural network-based high-level decoders offer decoding thresholds that surpass decoder-Minimum Weight Perfect Matching (MWPM), and exhibit strong scalability, making them one of the ideal solutions for addressing surface code challenges. However, current fully neural network-based high-level decoders can only operate serially and do not meet the current latency requirements (below 440 ns). To address these challenges, we first propose a parallel fully feedforward neural network (FFNN) high-level surface code decoder, and comprehensively measure its decoding performance on a computing-in-memory (CIM) hardware simulation platform. With the currently available hardware specifications, our work achieves a decoding threshold of 14.22%, and achieves high pseudo-thresholds of 10.4%, 11.3%, 12%, and 11.6% with decoding latencies of 197.03 ns, 234.87 ns, 243.73 ns, and 251.65 ns for distances of 3, 5, 7 and 9, respectively.**

***Index Terms*—Surface code, quantum error correction, decoder, compute in memory**

I. Introduction

Quantum computing holds significant promise as a solution to complex problems that classical computers struggle to solve. Unfortunately, qubits are extremely sensitive to their environment, making them prone to decoherence, which can lead to the loss of stored information [1]. To counter this, quantum error correction (QEC) is essential. QEC involves two critical steps: encoding and decoding. Among all error correction codes, surface code [2] stands out due to it being easy to implement on physical platforms.

Several decoders, such as Union Find (UF) [3], Look-Up Table (LUT) [4], Neural Network (NN) [5]–[8], MWPM [9], and modified MWPM [10] [11], have been employed for surface code decoding. Among all decoders, NN-based decoders have garnered considerable attention due to their potential to surpass MWPM in terms of decoding performance and scalability. NN-based decoders can be categorized into two types: low-level decoders (containing a single neural network module) and high-level decoders (comprising a simple decoder and a classifier module, as shown in fig.1) [1]. Compared to low-level decoders, the two modules in high-level decoders exhibit lower computational complexity and operate in parallel, thereby demonstrating superior decoding thresholds and reduced decoding latency. However, current NN–based approaches have not yet demonstrated acceptable decoding latencies (below 440 ns) on hardware platforms.

* indicates co-first authors, and + indicates the corresponding author. This project was supported by Guangdong Basic and Applied Basic Research Foundation (No. 23201910240002532).

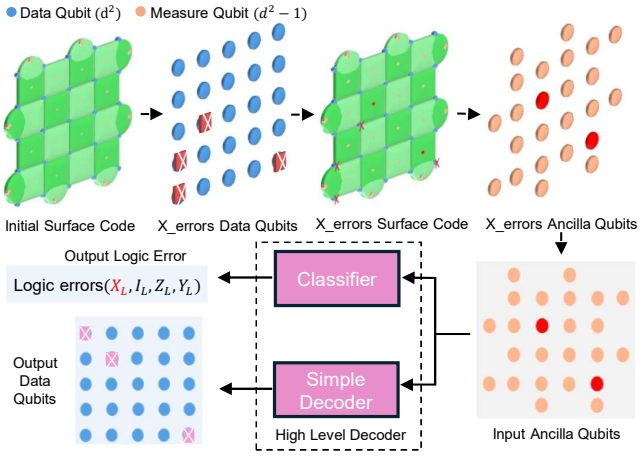

Fig. 1. Decoding distance-5 surface codes using high-level decoders.

To address this challenge, we propose a novel fully-parallel FFNN high-level decoder. Our decoder achieves high decoding thresholds (14.22%) on small-distance tasks under depolarizing noise models. Moreover, using the MNSIM 2.0 [12] simulator, we first demonstrate quantum surface code decoding on code distances 3 to 9 using a computing-in-memory (CIM) hardware architecture, achieving below-440ns latency on available hardware specifications.

II. Decoder Design

A. Surface code and high-level decoder

As shown in fig. 1, the surface code is a fault-tolerant encoding constructed from $d \times d$ data qubits and d^2-1 measurement qubits (ancilla qubits). Since directly measuring the data qubits would cause quantum collapse and result in the loss of stored information, ancilla qubits are employed to build quantum circuits for information readout. The **error syndrome** is the combination of quantum information obtained from ancilla qubits. Following the fig. 1, the error syndrome generated by the current data qubits. The high-level decoders then concurrently feed the error syndrome into both the classifier and the simple decoder, resulting in the final decoding outcome. This outcome includes predictions for the output data qubits as well as for the logical error.

978-3-9826741-0-0/DATE25/© 2025 EDAA

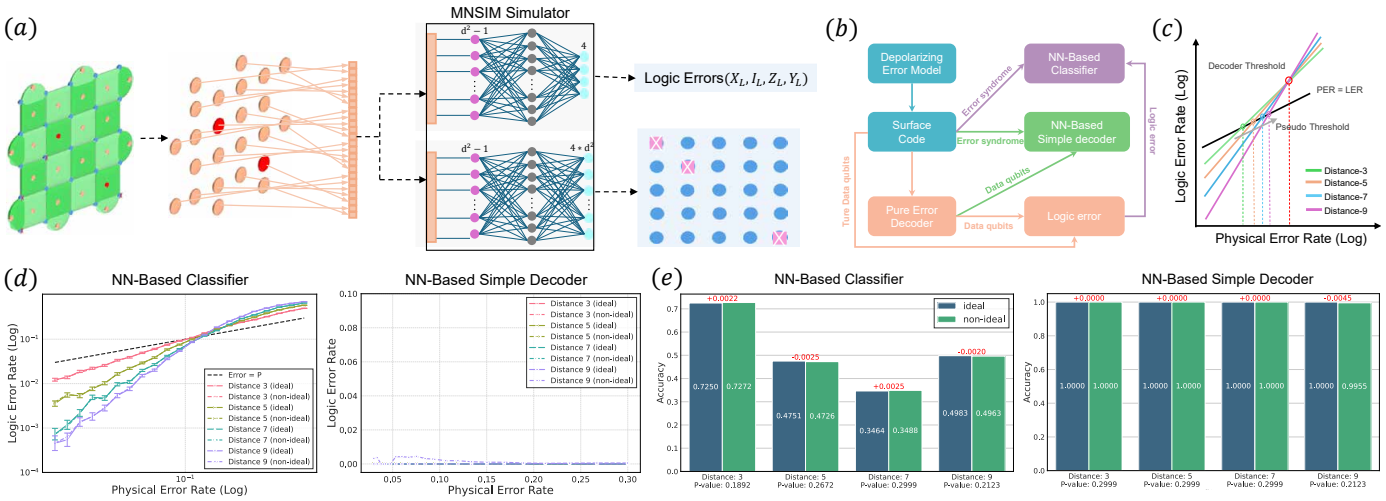

Fig. 2. (a) Our fully FFNN high-level decoders; (b) Simulation experiment implementation framework; (c) Diagram illustrating the decoding threshold and pseudo-threshold definitions; (d) Experimental results of high-level decoders; (e) Impact of non-idealities on the experiments

B. *Parallel fully NN-Based high-level decoder*

Inspired by [8], we first experimentally demonstrate that two-layer FFNN can achieve decoding performance comparable to previous methods. Moreover, since the implementation is entirely based on neural networks, it facilitates hardware acceleration using the CIM platform. We employed MNSIM 2.0 to measure the comprehensive hardware performance. The network parameters are presented in fig.2 (a).

III. EXPERIMENTS

A. *Experimental preparation*

In our experiment, we use the depolarizing noise model to generate the dataset, as detailed in [1] and [8]. The procedure for generating the dataset is described in [1]. According to the proposed guidelines, the minimum training set sizes for surface code decoding with distances of 3, 5, 7, and 9 should be 256, 2×10^5, 3×10^6, and 2×10^7, respectively. Open-source code is provided in [8]. We conducted the corresponding model training according to the rules outlined in fig.2 (b). Meanwhile, fig.2 (c) illustrates how the decoding threshold and pseudo-threshold were measured, with the decoding time defined as the maximum execution time of the classifier and simple decoder on the CIM platform.

B. *Experimental results and conclusion*

The experimental results for the decoding threshold are presented in fig.2 (d). Our decoder achieves a threshold of up to 14.22%, and the proposed simple decoder produces prediction outcomes that are approximately 100% consistent with the pure error decoder (PED) utilized in [8], thereby validating our approach. Hardware performance measured on MNSIM 2.0 is detailed in Tables I and II, with decoding latencies of 197.03 ns, 234.87 ns, 243.73 ns, and 251.65 ns for code distances of 3, 5, 7, and 9 respectively. In Table II, since [8] only includes hardware measurements for the classifier, we measured the latency of the PED component using open-source code at 250 MHz FPGA, and the results are recorded. Additionally, simulations performed under default hardware non-ideality parameters are depicted in fig.2 (e), indicating that non-idealities impact the experimental performance by no more than approximately 0.5%.

TABLE I
DECODER HARDWARE PERFORMANCE(NVM-BASED)

Distance	Pth	NN-Based Classifier				NN-Based Simple Decoder			
		Latency(ns)	Area(mm^2)	Power(W)	Energy(nJ)	Latency(ns)	Area(mm^2)	Power(W)	Energy(nJ)
3	10.40%	197.03	72.92	0.16	31.18	196.11	55.55	0.11	20.92
5	11.30%	234.87	98.97	0.75	177.06	203.93	81.61	0.22	45.79
7	12%	243.73	197.95	1.89	461.63	215.37	133.72	0.54	116.74
9	11.60%	251.65	296.93	2.94	740.82	239.65	479.34	4.23	1018.12

TABLE II
COMPARISON RESULTS(DISTANCE = 9)

Methods	Dth	Latency	Area	Power	Platform	Noise Model	Temperature
AFS [3]	2.60%	150ns(d=11)	-	-	-	flip and measurement error	300k
MWPM [10]	2.90%	-	-	-	CPU	Circuit-level	300k
QECool [10]	1%	400ns	183.45mm^2	400.32μW	SFQ	Circuit-level	4k
Fully_RNN [6]	0.10%	-	-	-	CIM	Circuit-level	300k
QECool [10]	6%	400ns	183.45mm^2	400.32μW	SFQ	Depolarizing	4k
AQEC [11]	5%	19.8ns	329.46mm^2	3.88mW	SFQ	Depolarizing	4k
UF+NN [7]	16.20%	>1ms	-	-	FPGA	Depolarizing	300k
LUT+NN [1]	>12.45%	31.34ms	-	-	CPU	Depolarizing	300k
PED+NN [8]	>12.49%	1.936μs	-	-	CPU	Depolarizing	300k
PED+NN [8]	>12.49%	5.312μs	-	-	FPGA	Depolarizing	300k
Ours(Parallel_FFNN)	**14.22%**	**251.65ns**	**479.34mm^2**	**6.99W**	**CIM**	Depolarizing	300k

REFERENCES

[1] S. Varsamopoulos and et al., "Comparing neural network based decoders for the surface code," *IEEE Transactions on Computers*, 2019.

[2] A. G. Fowler and et al., "Surface codes: Towards practical large-scale quantum computation," *Physical Review A*, 2012.

[3] P. Das and et al., "Afs: Accurate, fast, and scalable error-decoding for fault-tolerant quantum computers," in *HPCA*, 2022.

[4] P. Das and el at., "Lilliput: a lightweight low-latency lookup-table decoder for near-term quantum error correction," in *ASPLOS*, 2022.

[5] P. Baireuther and el at., "Machine-learning-assisted correction of correlated qubit errors in a topological code," *Quantum*, 2018.

[6] F. Marcotte and el at., "A cryogenic memristive neural decoder for fault-tolerant quantum error correction," *arXiv preprint*, 2023.

[7] K. Meinerz and el at., "Scalable neural decoder for topological surface codes," *Physical Review Letters*, 2022.

[8] O. el at., "Neural-network decoders for quantum error correction using surface codes: A space exploration of the hardware cost-performance tradeoffs," *IEEE Transactions on Quantum Engineering*, 2022.

[9] "Suppressing quantum errors by scaling a surface code logical qubit," *Nature*, 2023.

[10] Y. Ueno and el at., "Qecool: On-line quantum error correction with a superconducting decoder for surface code," in *DAC*, 2021.

[11] A. Holmes and el at., "Nisq+: Boosting quantum computing power by approximating quantum error correction," in *ISCA*, 2020.

[12] Z. Zhu and el at., "Mnsim 2.0: A behavior-level modeling tool for processing-in-memory architectures," *TCAD*, 2023.