Vegapunk: Accurate and Fast Decoding for Quantum LDPC Codes with Online Hierarchical Algorithm and Sparse Accelerator

Kaiwen Zhou[1], Liqiang Lu[1] *, Debin Xiang[1], Chenning Tao[1], Anbang Wu[2], Jingwen Leng[2], Fangxin Liu[2], Mingshuai Chen[1], Jianwei Yin[1] *

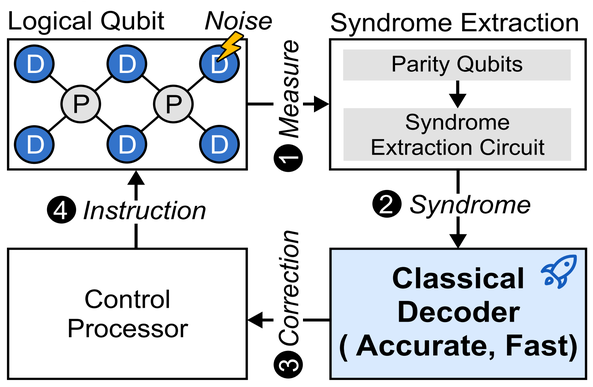
[1]Zhejiang University, [2]Shanghai Jiao Tong University

Email: kaiwenzhou@zju.edu.cn

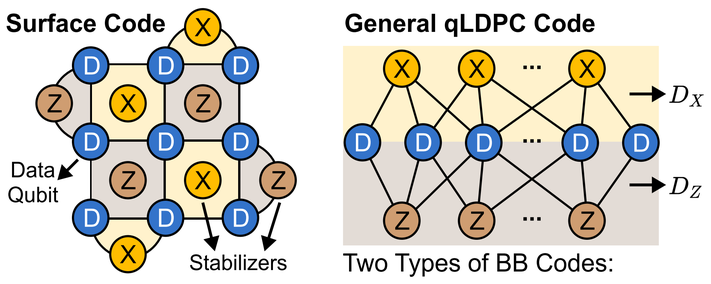
MICRO'25
Oct. 20th
Session 4C

Introduction

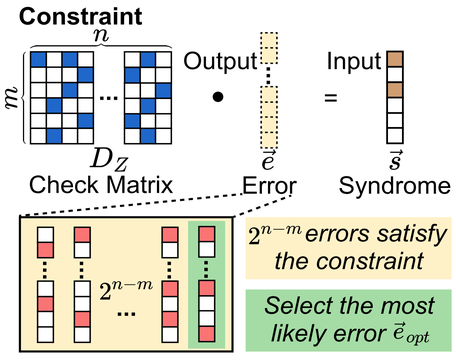
- Quantum Error Correction (QEC) is a cyclical process: first, parity qubits interact with data qubits to detect errors and generate the syndrome. This syndrome is then sent to a classical decoder to calculate a correction plan, which is executed by the control processor to fix the error.



- Quantum Low-Density Parity-Check (qLDPC) codes are a promising path towards scalable fault-tolerant quantum computation, offering constant-rate encoding and high error thresholds.

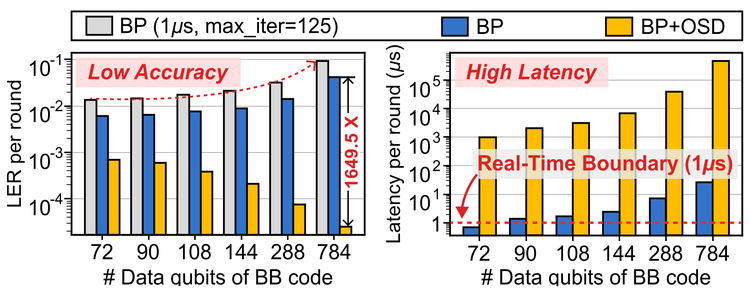


- This image illustrates the core challenge of decoding: the same syndrome can be caused by a vast number of different error patterns. The decoder's job is to find the one most likely error out of all these possibilities.



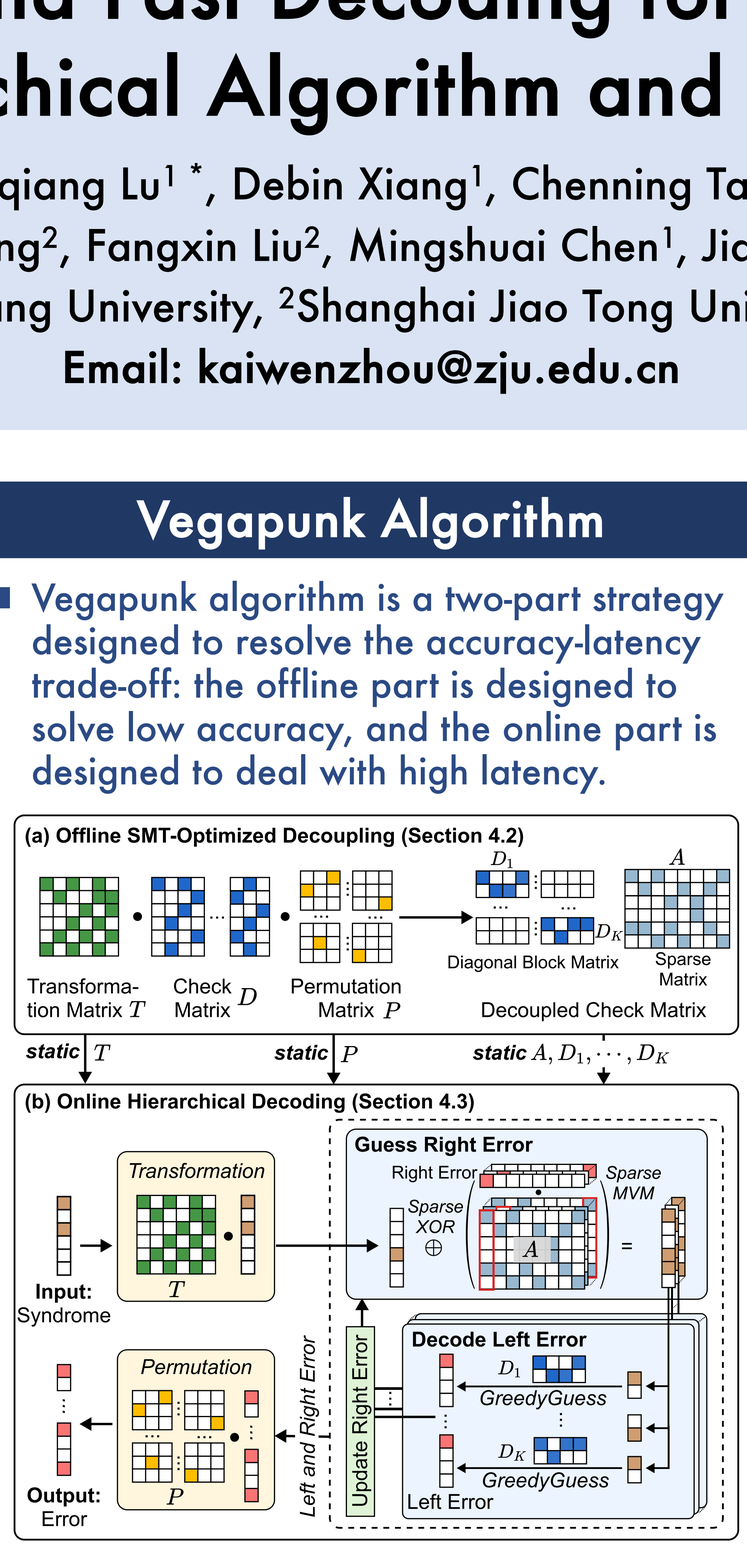
Motivation

- Existing qLDPC decoders force a trade-off between accuracy and speed:
  - **Low Accuracy**: BP is fast and parallelizable, but suffers from low accuracy due to quantum degeneracy, where multiple error patterns can produce the same syndrome.
  - **High Latency**: BP+OSD uses Ordered Statistics Decoding (OSD) as a post-processing step. However, OSD involves costly operations, like linear system solving, failing to satisfy the real-time requirement.



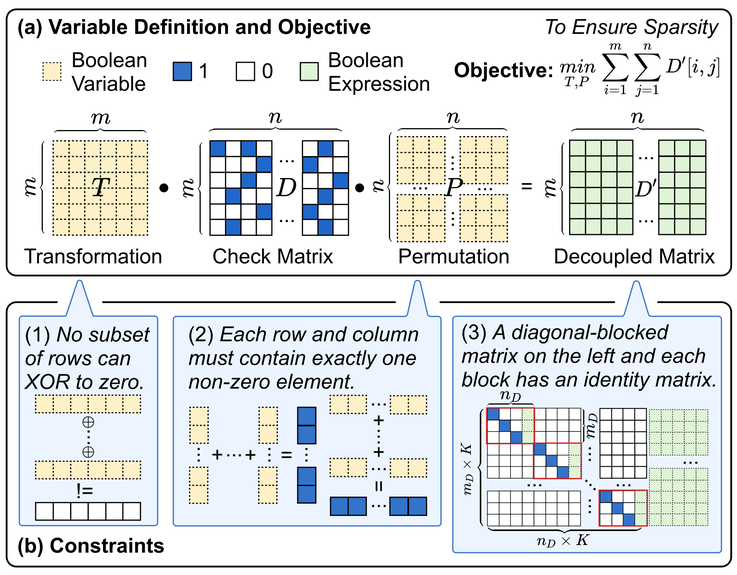
Vegapunk Algorithm

- Vegapunk algorithm is a two-part strategy designed to resolve the accuracy-latency trade-off: the offline part is designed to solve low accuracy, and the online part is designed to deal with high latency.



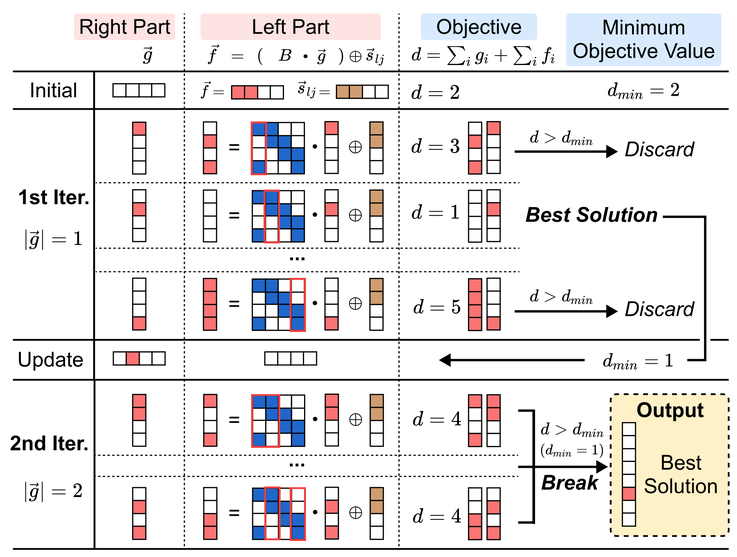
Offline SMT-Based Decoupling

- The core idea is to transform the original check matrix into a sparser, partially diagonal-blocked matrix. We reformulate this process as an SMT problem and use an SMT solver to obtain the matrix T and P.



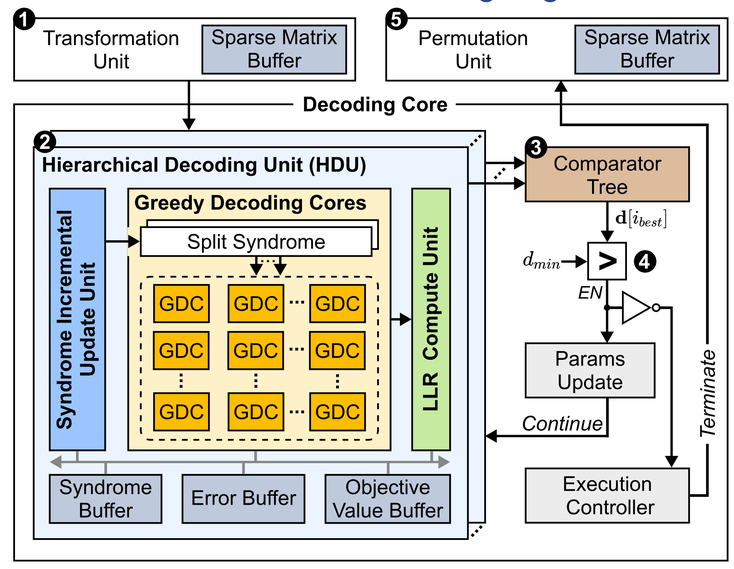
Online Hierarchical Decoding

- Our online hierarchical decoding algorithm utilizes pre-computed matrices and applies a greedy strategy for fast qLDPC decoding.

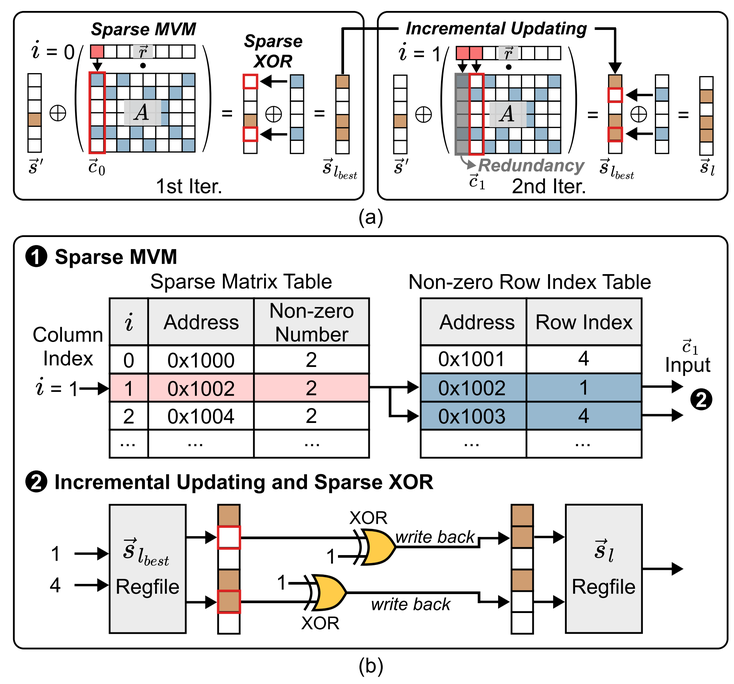


Vegapunk Accelerator

- To meet the stringent sub-microsecond latency requirement, we designed a custom FPGA-based accelerator based on our online hierarchical decoding algorithm.



- We design the Syndrome Incremental Update Unit, which efficiently computes residual syndromes by exploiting sparsity and reusing results from previous iterations, thereby avoiding redundant calculations.



Evaluation

- We develop Vegapunk using Xilinx HLS C++ and implement it with Vitis 2023.1. Vegapunk achieves high-accuracy decoding within the real-time latency budget.