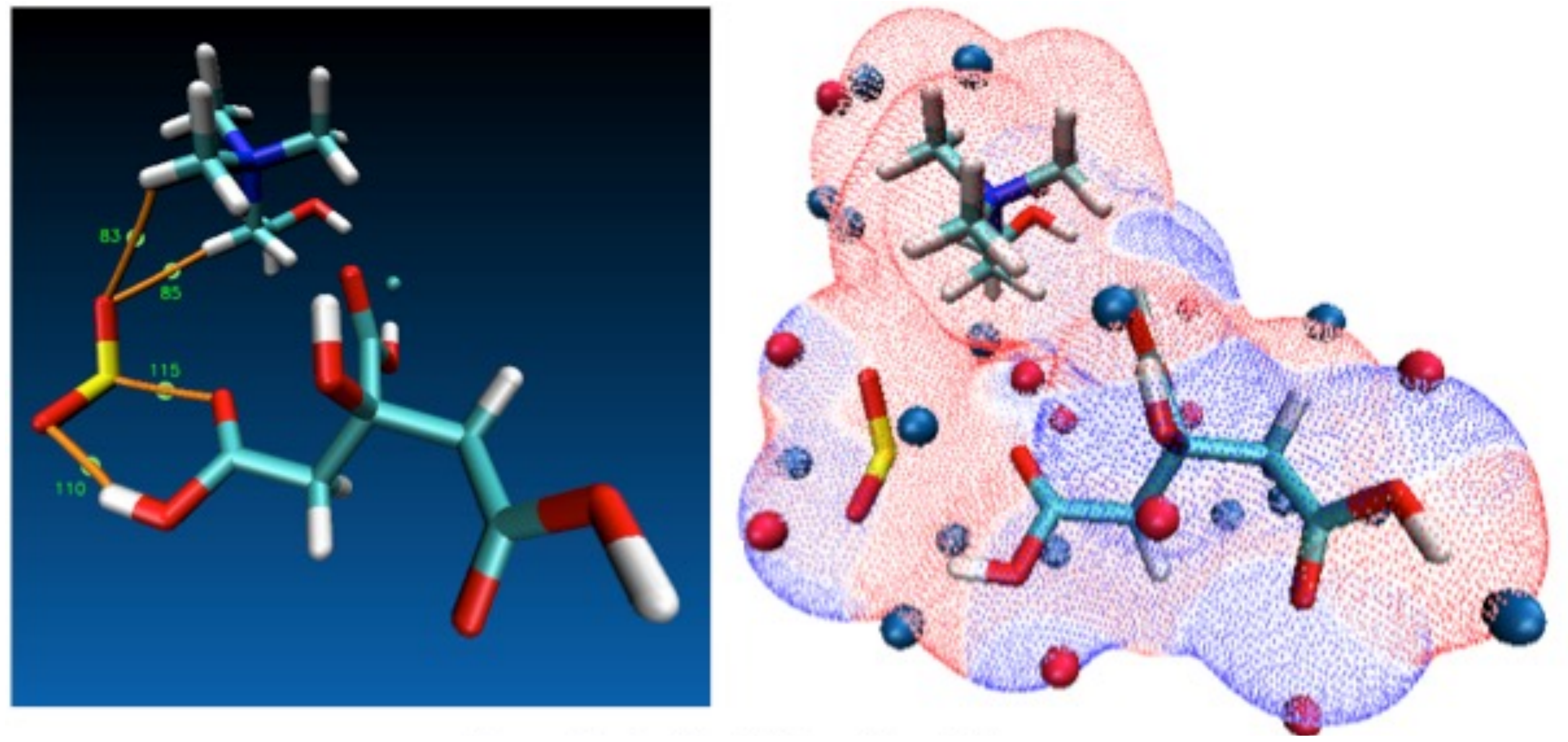
FASTML Workshop @ ICCAD 2023

TT-QEC: Transferable Transformer for Quantum Error Correction Code Decoding

Hanrui Wang[1], Pengyu Liu[2], Kevin Shao[1], Dantong Li[3], Jiaqi Gu[4],
David Z. Pan[3], Yongshan Ding[3], Song Han[1]
[1]MIT [2]CMU [3]Yale University [2]ASU [2]UT Austin

Quantum Computing has Ubiquitous Potential Applications

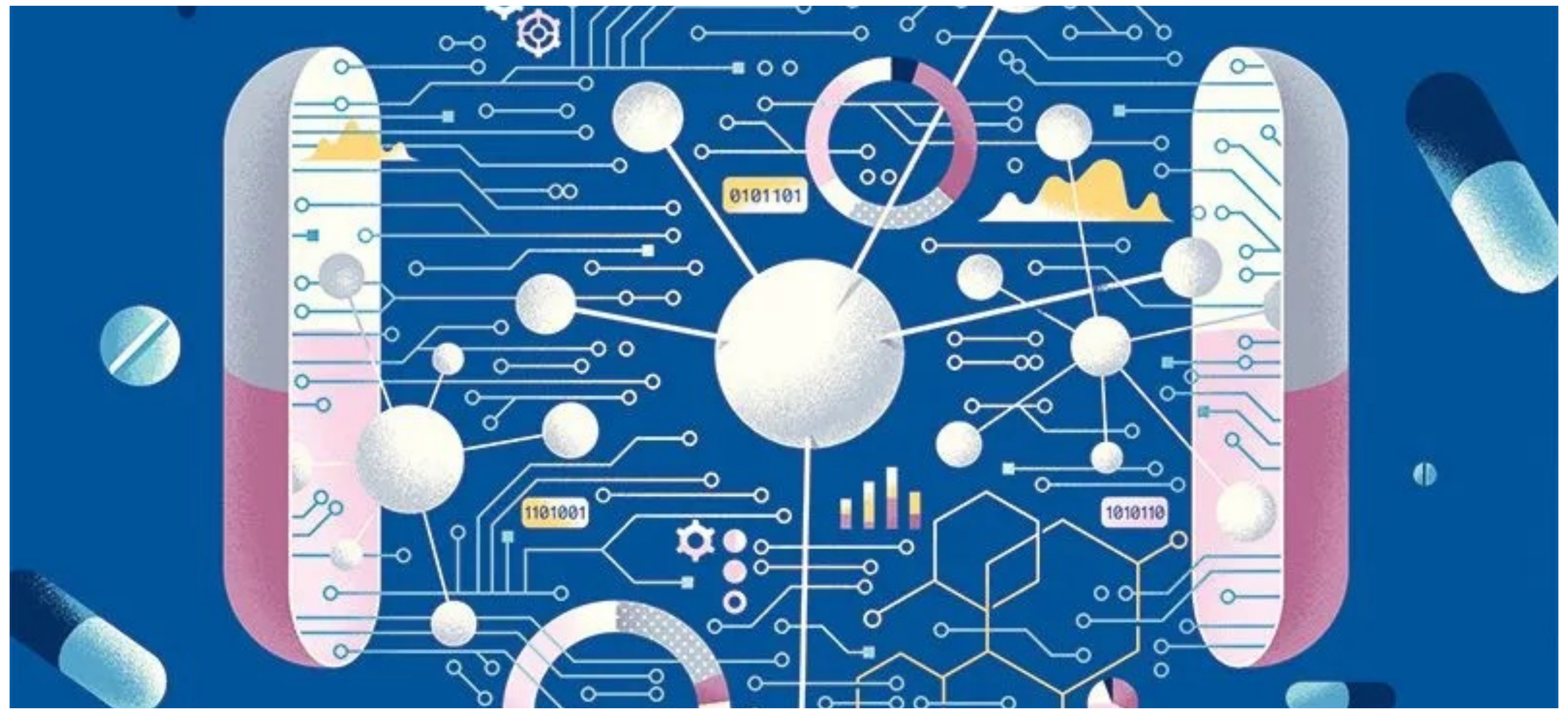
Cryptograph



Chemistry

Optimization

Machine Learning



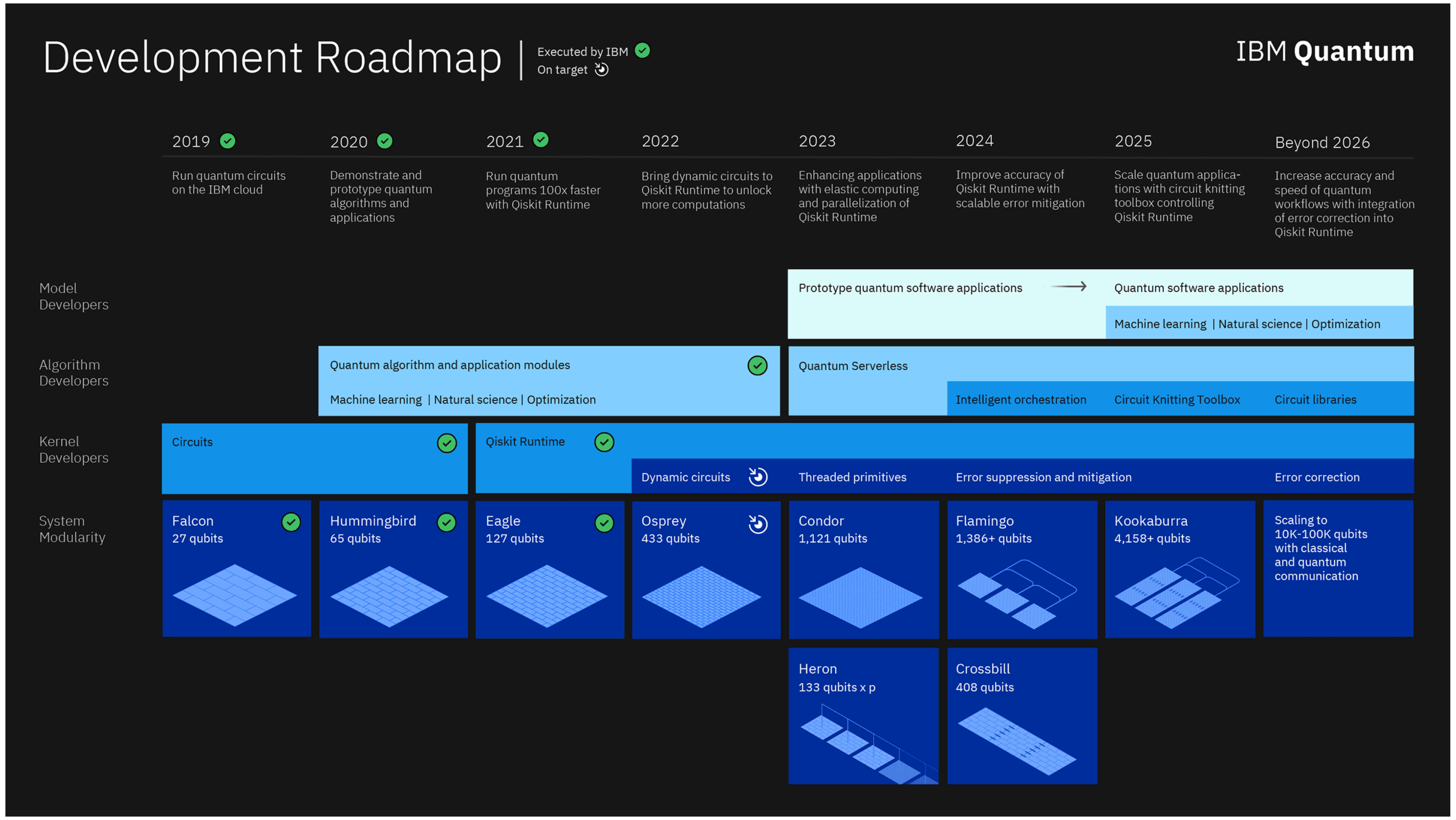
Pharmaceutical

Climate

Torch
Quantum

MIT HAN LAB

Practical Quantum Computing is Getting Real



Development Roadmap | Executed by IBM ✓ On target

IBM **Quantum**

2019 ✓ Run quantum circuits on the IBM cloud

2020 ✓ Demonstrate and prototype quantum algorithms and applications

2021 ✓ Run quantum programs 100x faster with Qiskit Runtime

2022 Bring dynamic circuits to Qiskit Runtime to unlock more computations

2023 Enhancing applications with elastic computing and parallelization of Qiskit Runtime

2024 Improve accuracy of Qiskit Runtime with scalable error mitigation

2025 Scale quantum applications with circuit knitting toolbox controlling Qiskit Runtime

Beyond 2026 Increase accuracy and speed of quantum workflows with integration of error correction into Qiskit Runtime

Model Developers

- Prototype quantum software applications → Quantum software applications
- Machine learning | Natural science | Optimization

Algorithm Developers

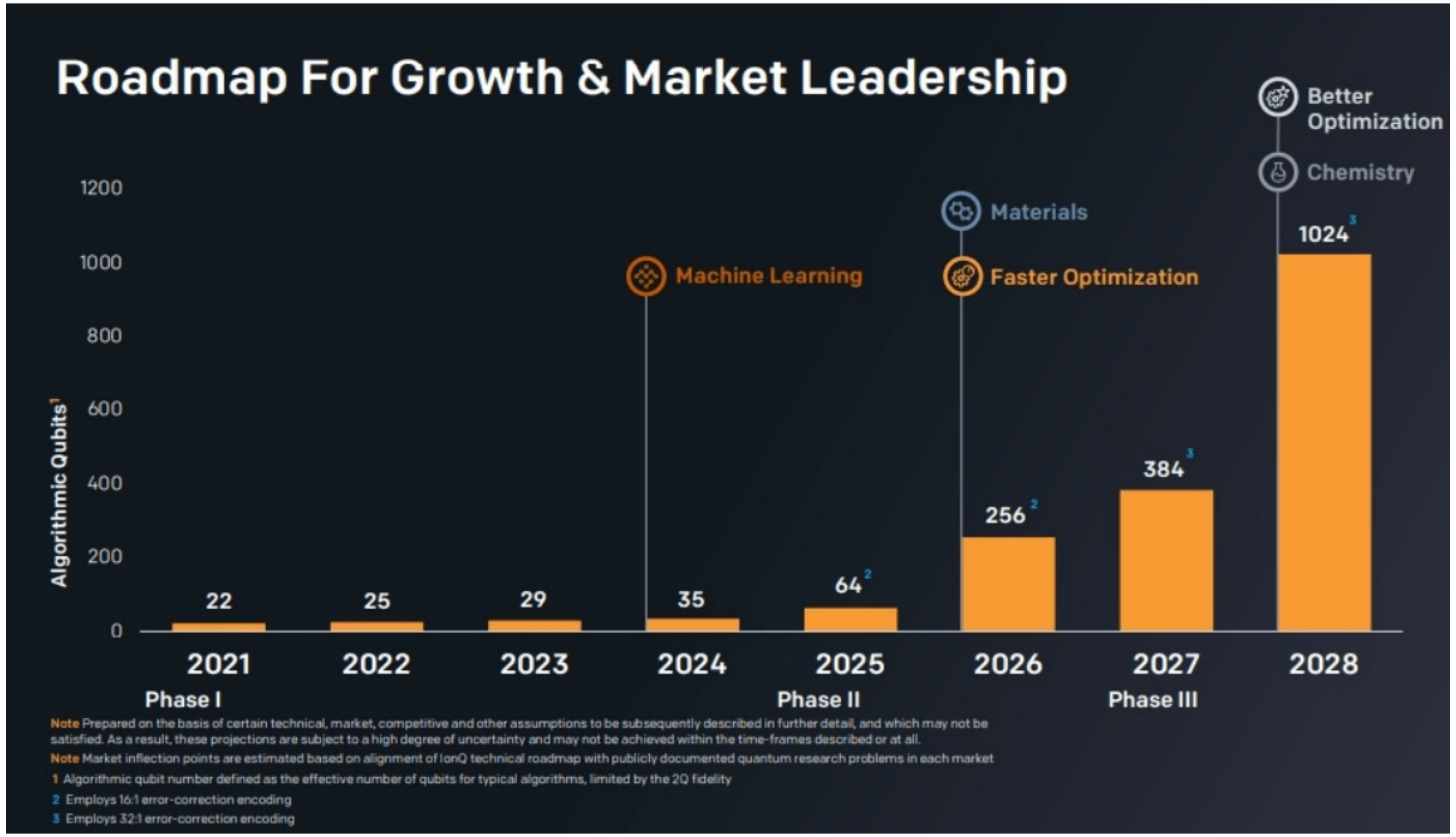
- Quantum algorithm and application modules ✓
- Machine learning | Natural science | Optimization
- Quantum Serverless
- Intelligent orchestration
- Circuit Knitting Toolbox
- Circuit libraries

Kernel Developers

- Circuits ✓
- Qiskit Runtime ✓
- Dynamic circuits
- Threaded primitives
- Error suppression and mitigation
- Error correction

System Modularity

- Falcon ✓ 27 qubits
- Hummingbird ✓ 65 qubits
- Eagle ✓ 127 qubits
- Osprey 433 qubits
- Condor 1,121 qubits
- Flamingo 1,386+ qubits
- Kookaburra 4,158+ qubits
- Scaling to 10K-100K qubits with classical and quantum communication
- Heron 133 qubits x p
- Crossbill 408 qubits



Roadmap For Growth & Market Leadership

Note Prepared on the basis of certain technical, market, competitive and other assumptions to be subsequently described in further detail, and which may not be satisfied. As a result, these projections are subject to a high degree of uncertainty and may not be achieved within the time-frames described or at all.

Note Market inflection points are estimated based on alignment of IonQ technical roadmap with publicly documented quantum research problems in each market

[1] Algorithmic qubit number defined as the effective number of qubits for typical algorithms, limited by the 2Q fidelity

[2] Employs 16:1 error-correction encoding

[3] Employs 32:1 error-correction encoding

IBM Superconducting Roadmap

IonQ Trapped-Ion Roadmap

Torch Quantum

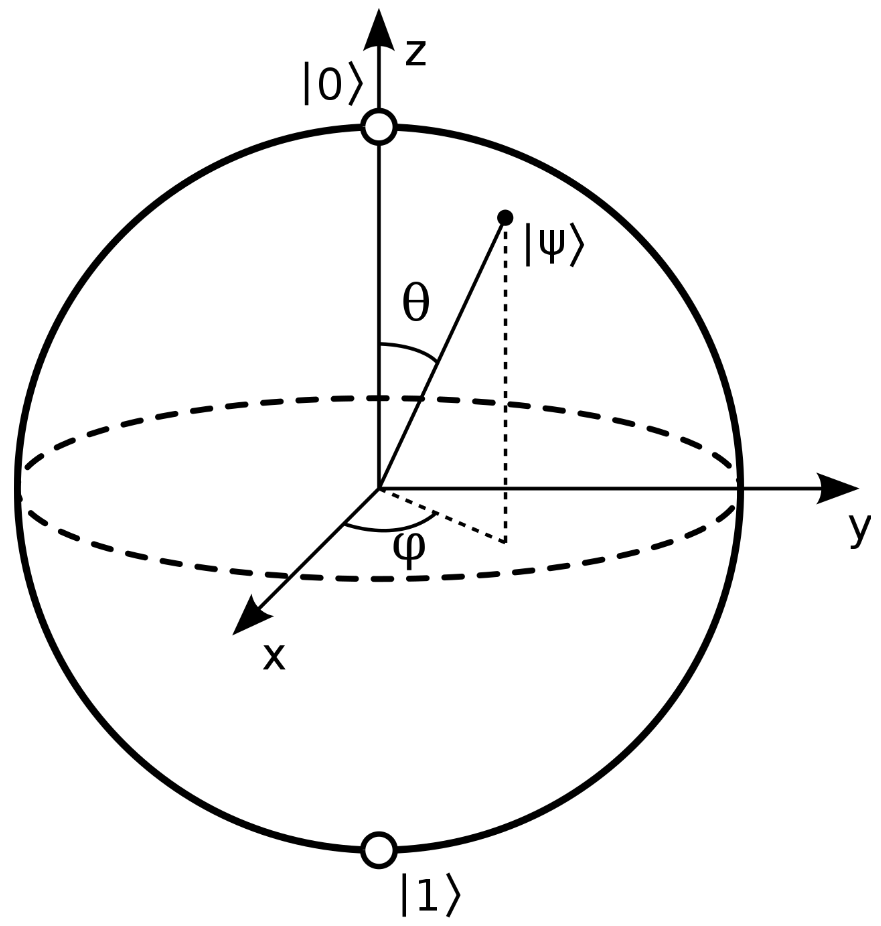
MIT HAN LAB

Quantum Computing Basics

- Quantum Bit (Qubit)
- Statevector: contains 2^n complex numbers for n qubit system
- The square sum of magnitude of 2^n numbers are 1
- 1 qubit:

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} \qquad a_0, a_1 \in \mathbb{C}$$

$$|a_0|^2 + |a_1|^2 = 1$$



- 2 qubits:

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} \qquad a_0, a_1, a_2, a_3 \in \mathbb{C}$$

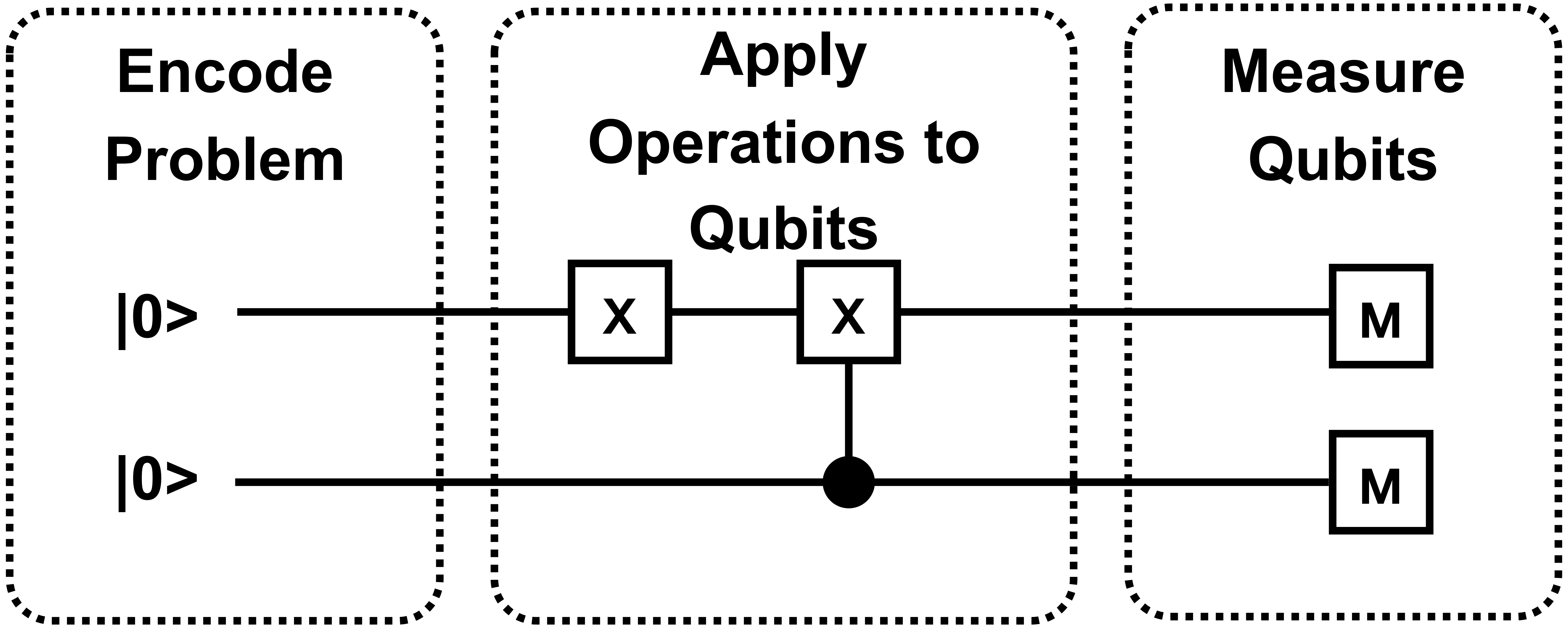
$$|a_0|^2 + |a_1|^2 + |a_2|^2 + |a_3|^2 = 1$$

Quantum Computing Basics

- Quantum Operations (Gates)
- Quantum algorithms apply gates to qubit to manipulate the quantum states

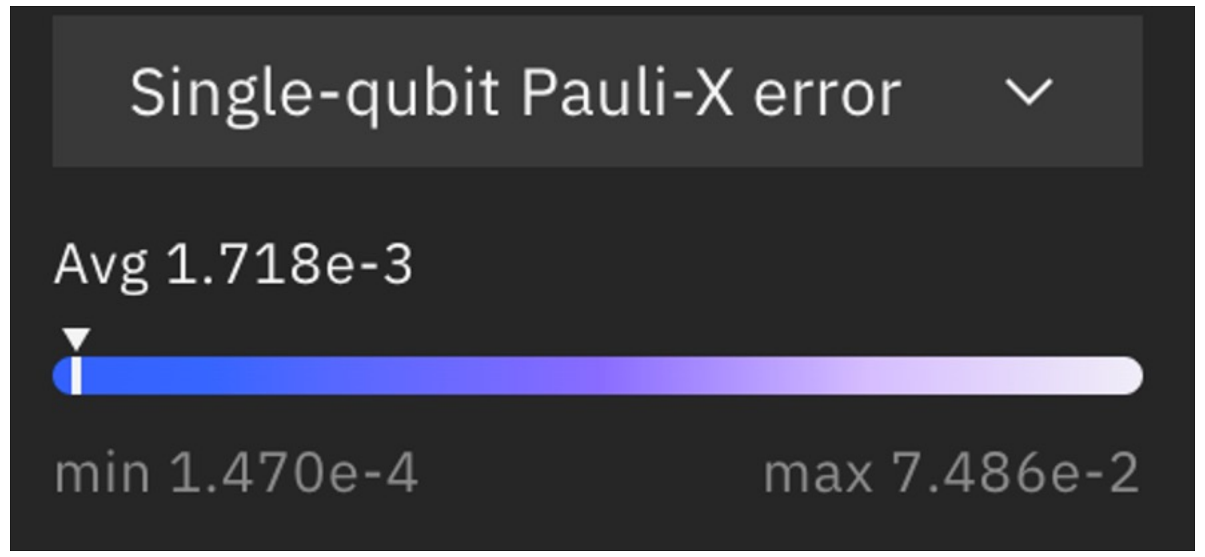
$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$CRX(\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\frac{\theta}{2} & -i\sin\frac{\theta}{2} \\ 0 & 0 & -i\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{bmatrix}$$



Quantum Computing Challenges

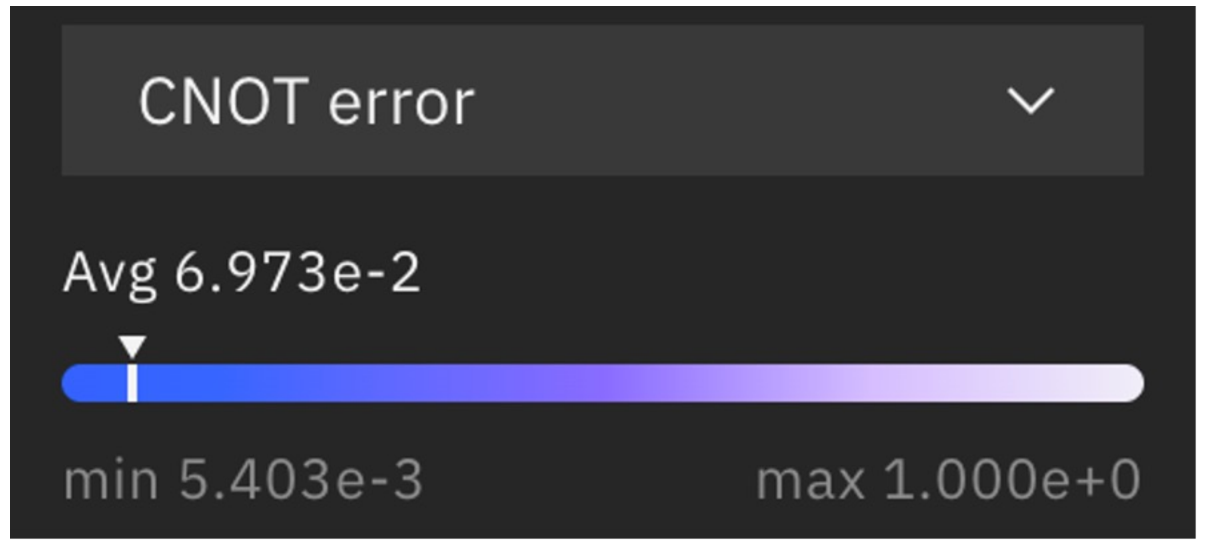
- Reliability
  - 1Q gate error rate ~10^{-3}
  - 2Q gate error rate ~10^{-2}



Single-qubit Pauli-X error

Avg 1.718e-3

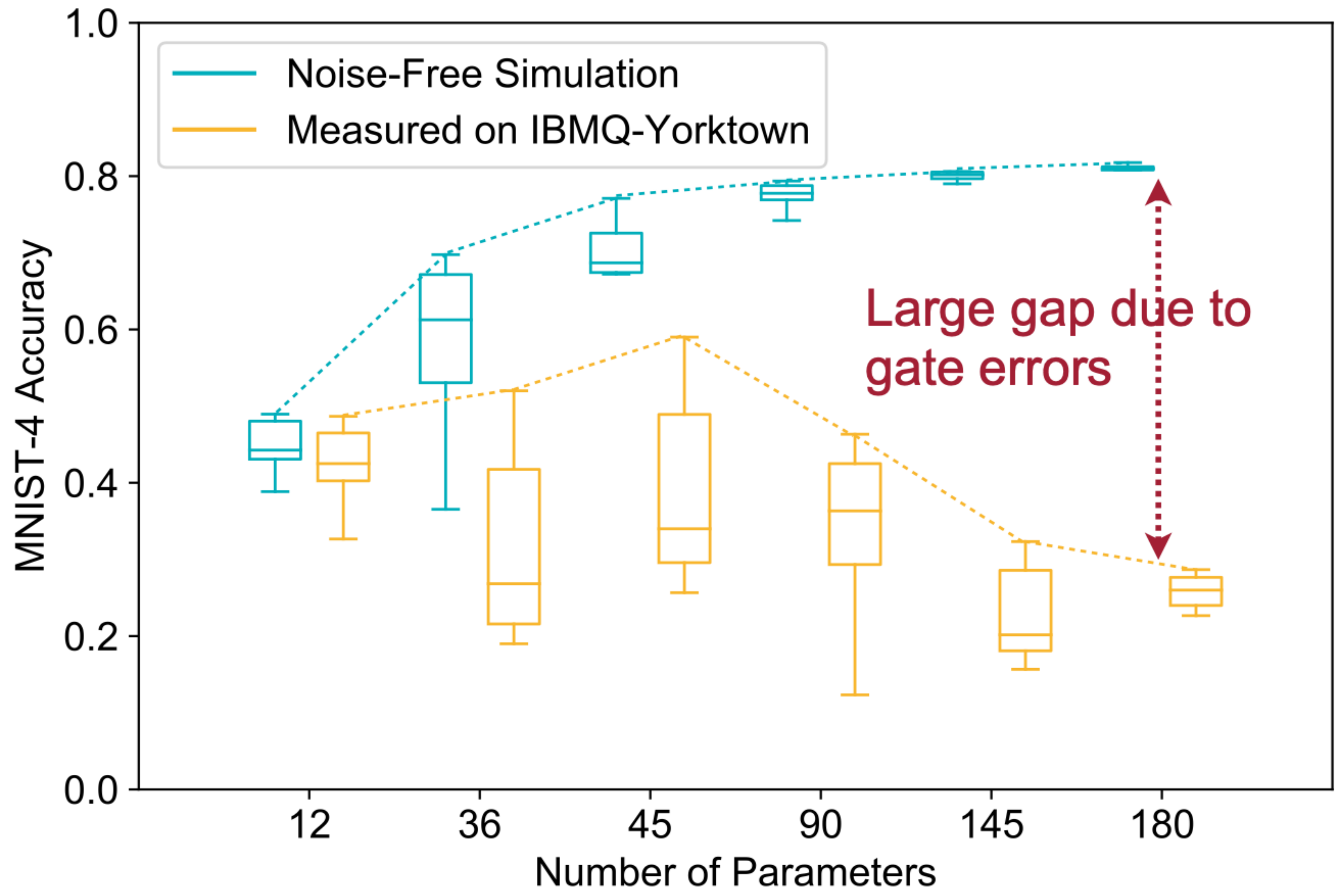
min 1.470e-4 max 7.486e-2



CNOT error

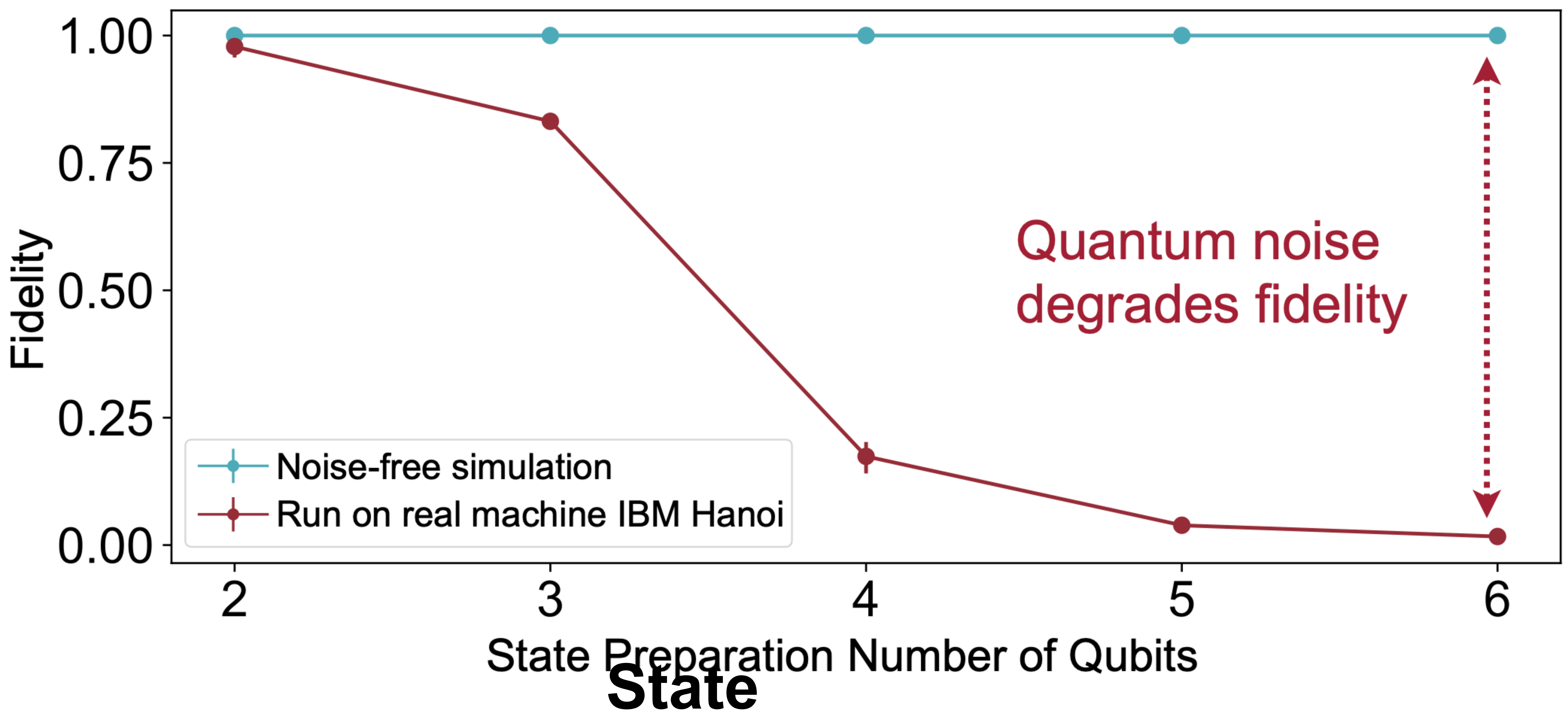
Avg 6.973e-2

min 5.403e-3 max 1.000e+0



Large gap due to gate errors

Quantum Classifier



Quantum noise degrades fidelity

State Preparation

Error Correction

- Trade redundancy for reliability
- In the classical case

$$0\ 0\ 0\ 0\ 0 \xrightarrow{\text{Noise}} 0\ 0\ 1\ 1\ 0 \xrightarrow[\text{Voting}]{\text{Majority}} 0\ 0\ 0\ 0\ 0$$

Quantum Error Correction

- Difference
  - Both bit flip (0 to 1) and phase flip error (1 to -1)
  - We need two dimensions of the error correction
    - One checks X dim and one checks Z dim

Quantum Error Correction

- Difference
  - Cannot directly measure the quantum information
    - The redundancy is on the **basis** of information, not information itself
  - q_0 = a_0 |0> + a_1 |1>
  - |0> = $[1\ 0]^T$
  - |1> = $[0\ 1]^T$

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$$

$$a_0, a_1 \in \mathbb{C}$$

$$|a_0|^2 + |a_1|^2 = 1$$

  - Due to non-clone theorem, cannot get q_1 that is exactly the same as q_0
  - However, we can get a_0 |00> + a_1 |11>
    - For more qubits a_0 |00000> + a_1 |11111>

Quantum Error Correction

- When bit flip error occurs
  - a_0 |10000> + a_1 |01111>
- When phase flip error occurs
  - a_0 |00000> + a_1 |-11111> = a_0 |00000> - a_1 |11111>
- How to know where is the error?
  - Check the qubit **parity**

a_0 |00000>

Parity = 0

a_0 |10000>

Parity = 1

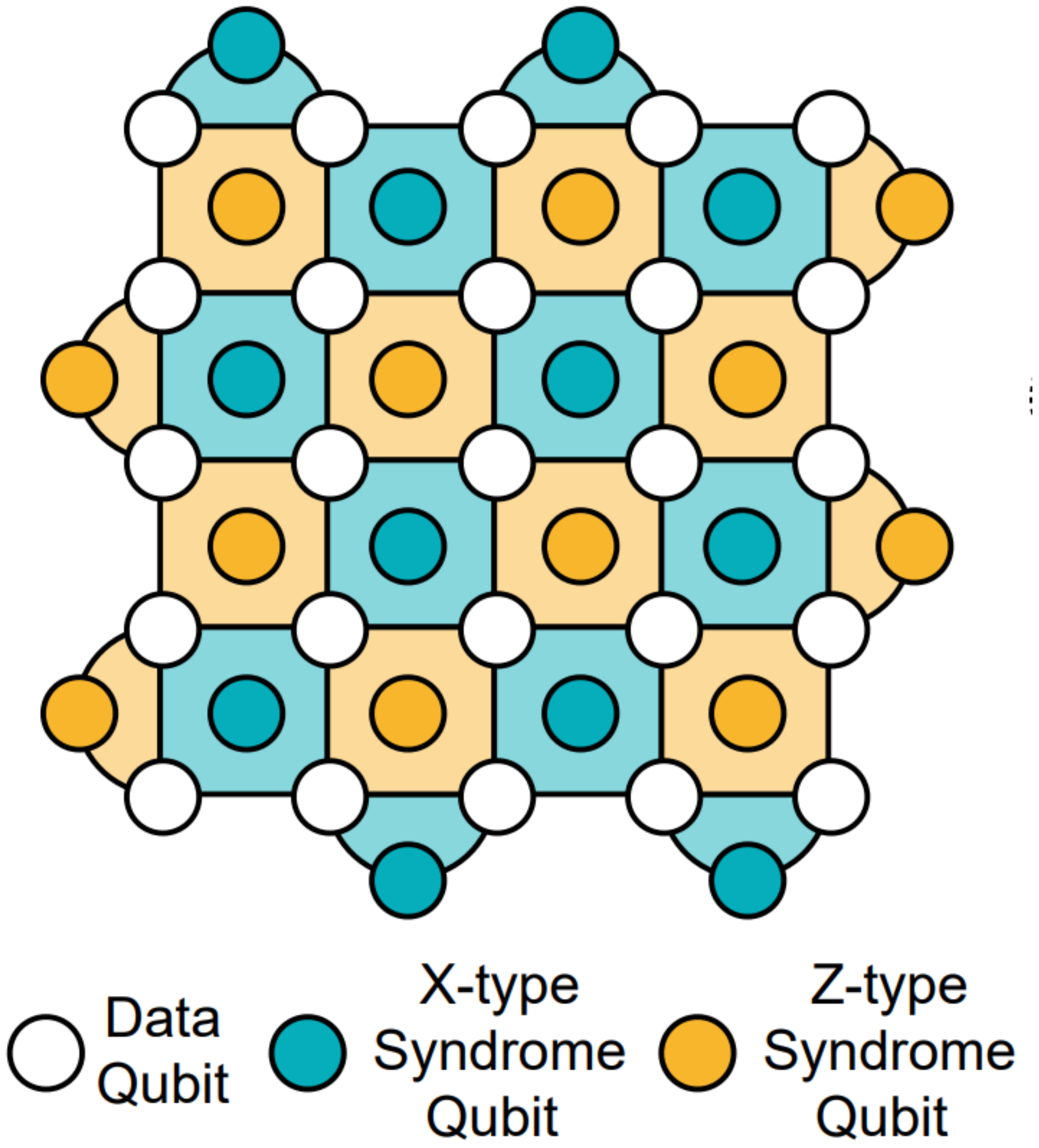
a_0 |01000>

Parity = 1

- **We use some qubits to store information and some obtain parity (syndromes)**

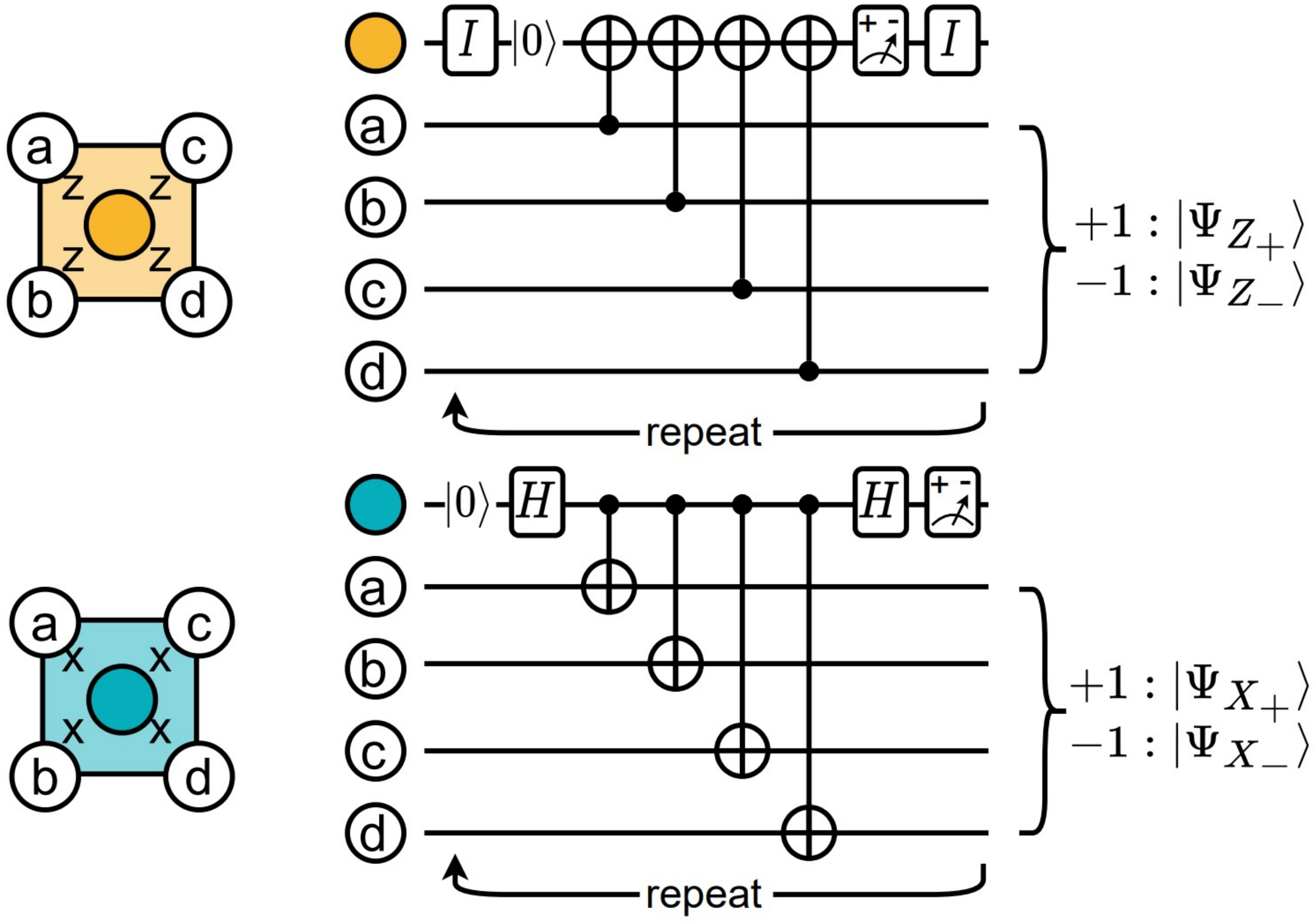
Surface Code

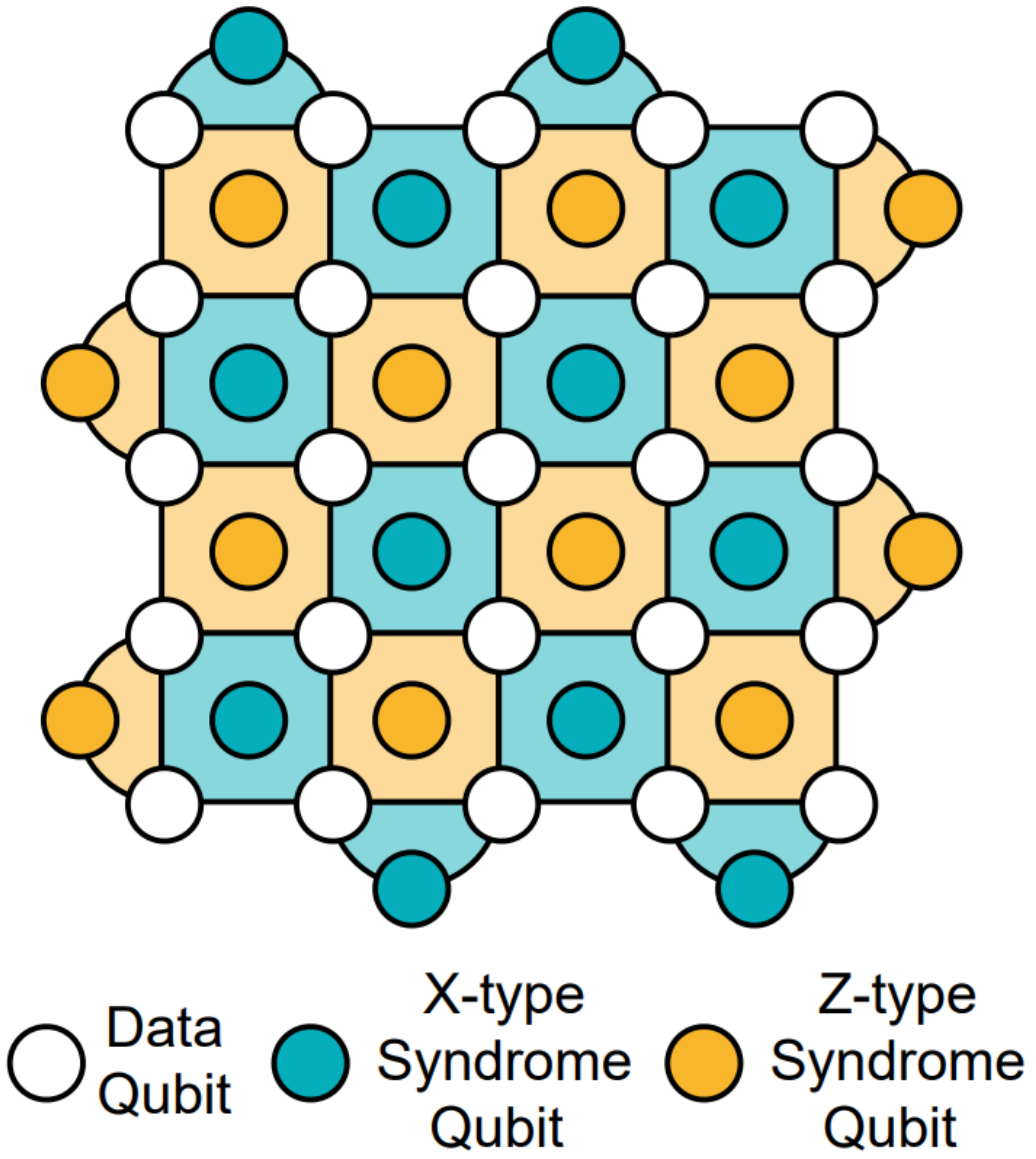
- Data qubits (white) store information distributedly
- a_0 |0000000000000000000000000> +
  a_1 |1111111111111111111111111>
- Syndrome Qubits (Green): check bit flip parity
- Syndrome Qubits (Yellow): check phase flip parity



Syndrome Extraction

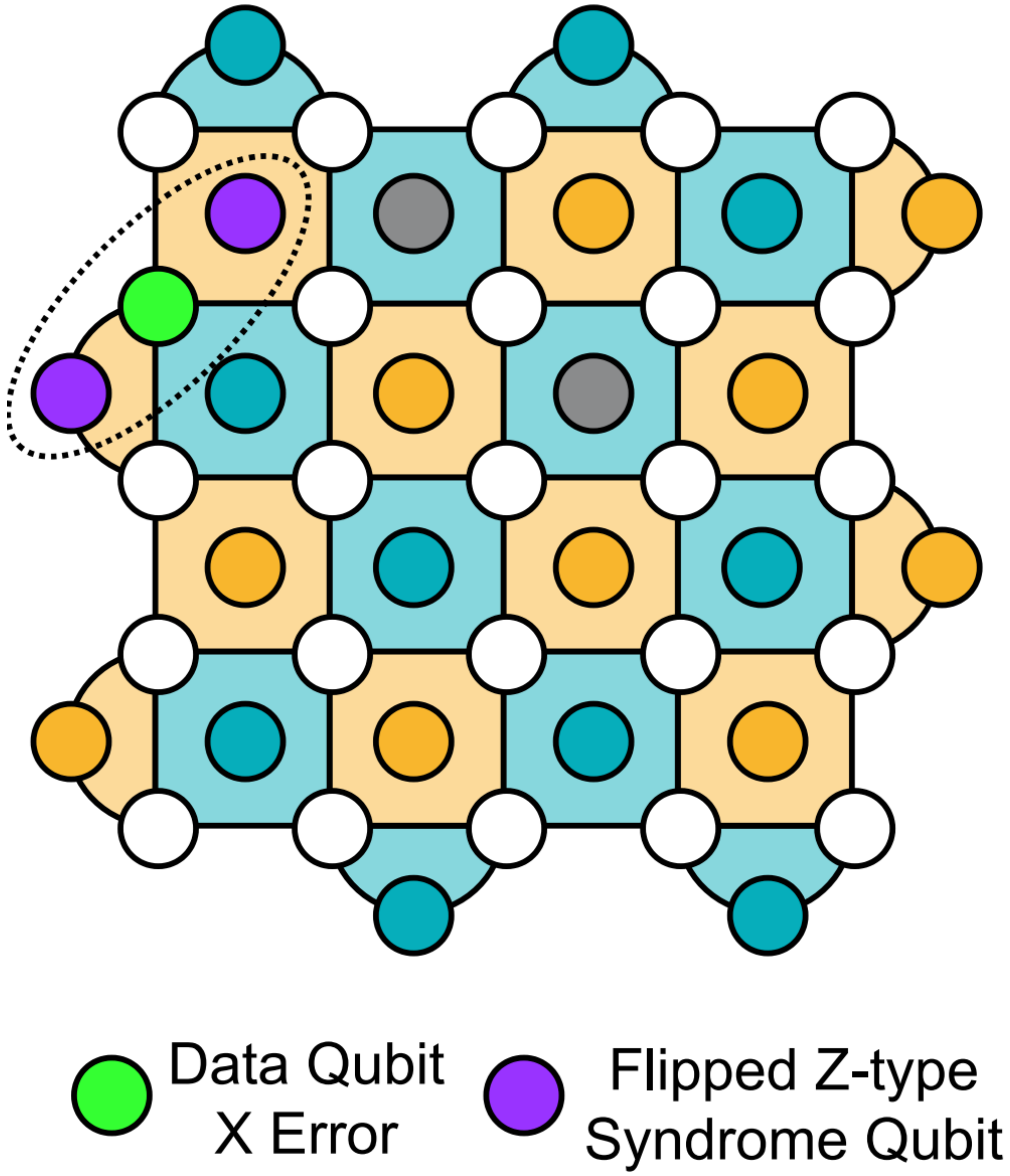
- Syndrome Extraction process





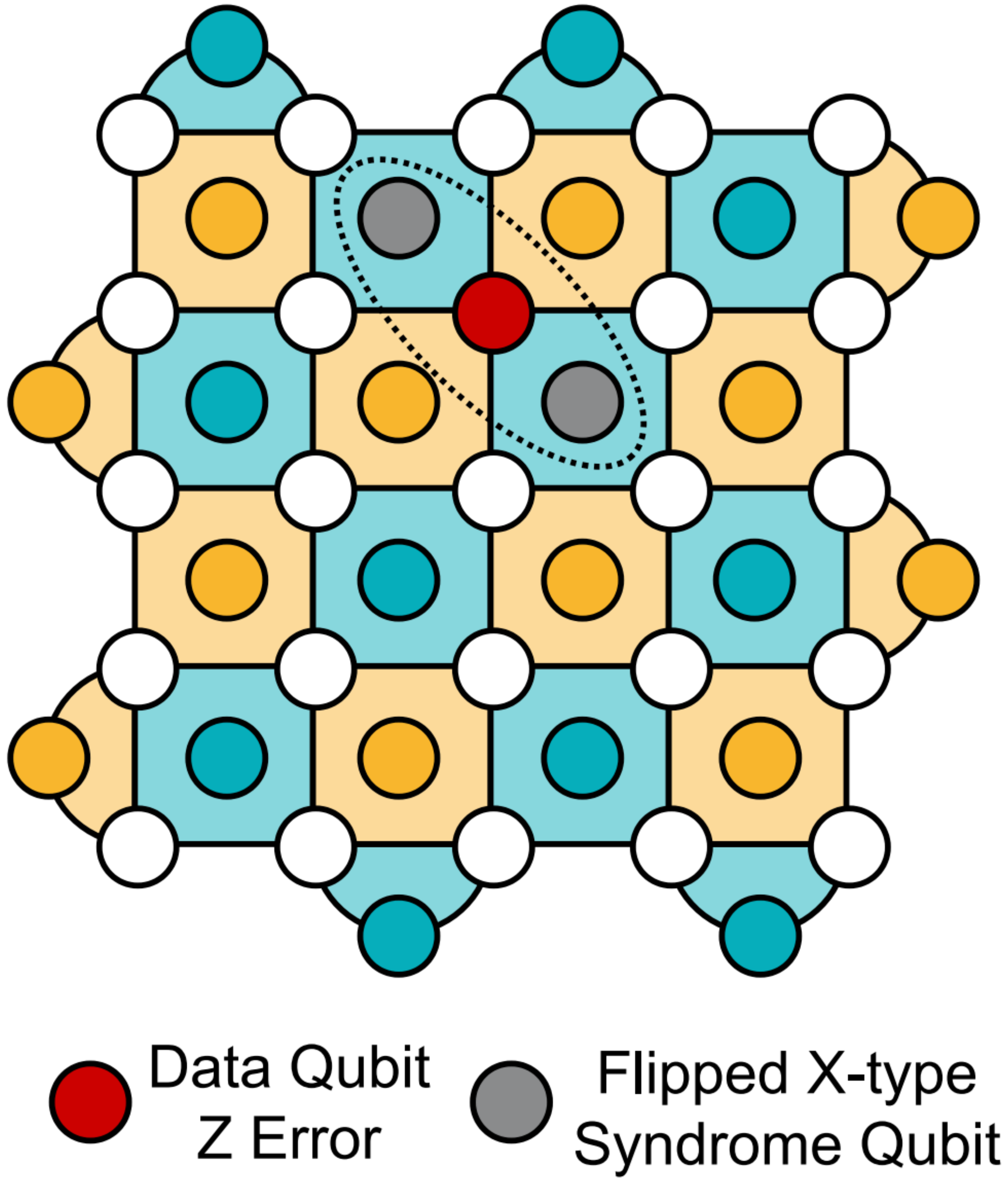
Surface Code

- Example of indicating X error from Z



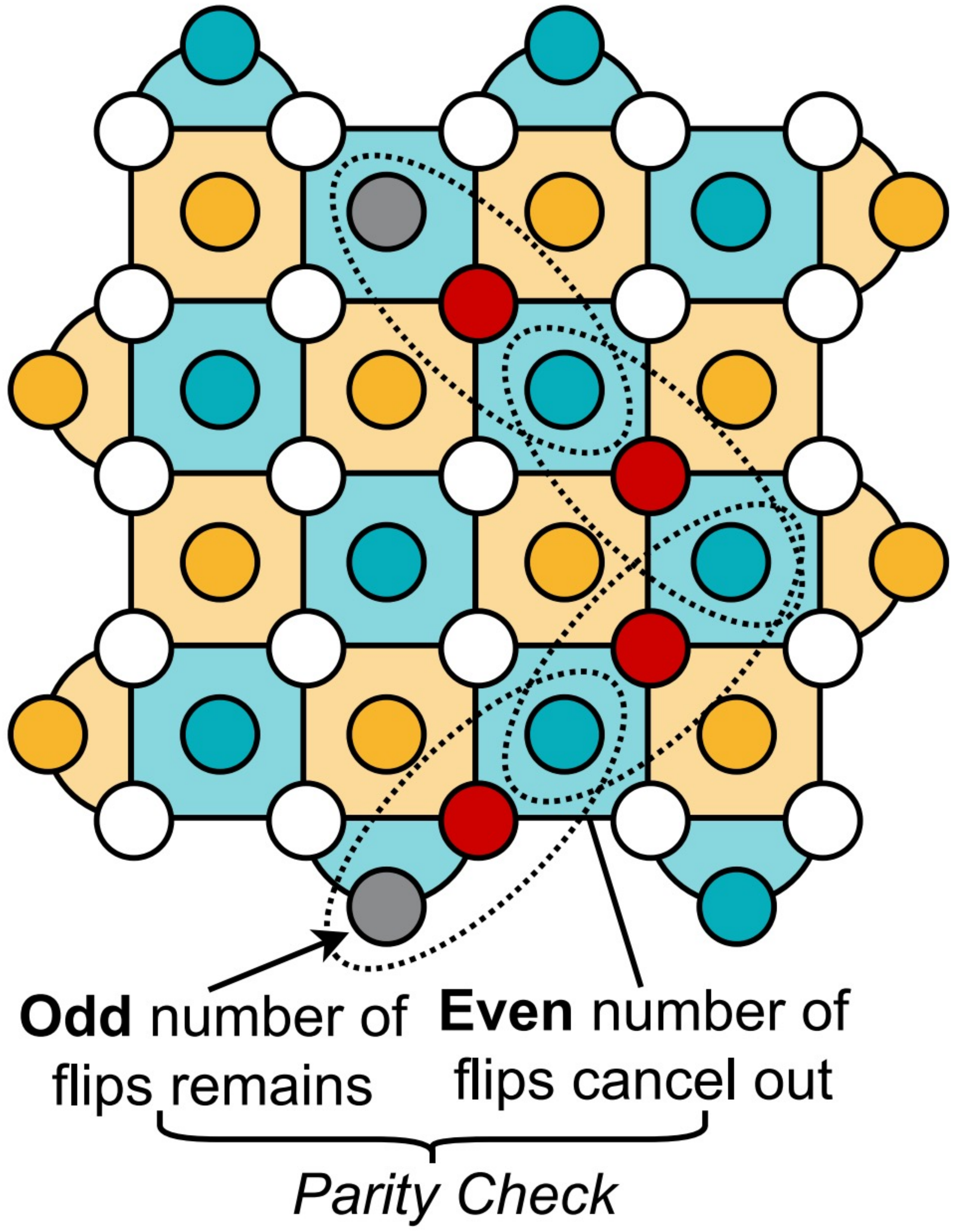
Surface Code

- Example of indicating X error from Z



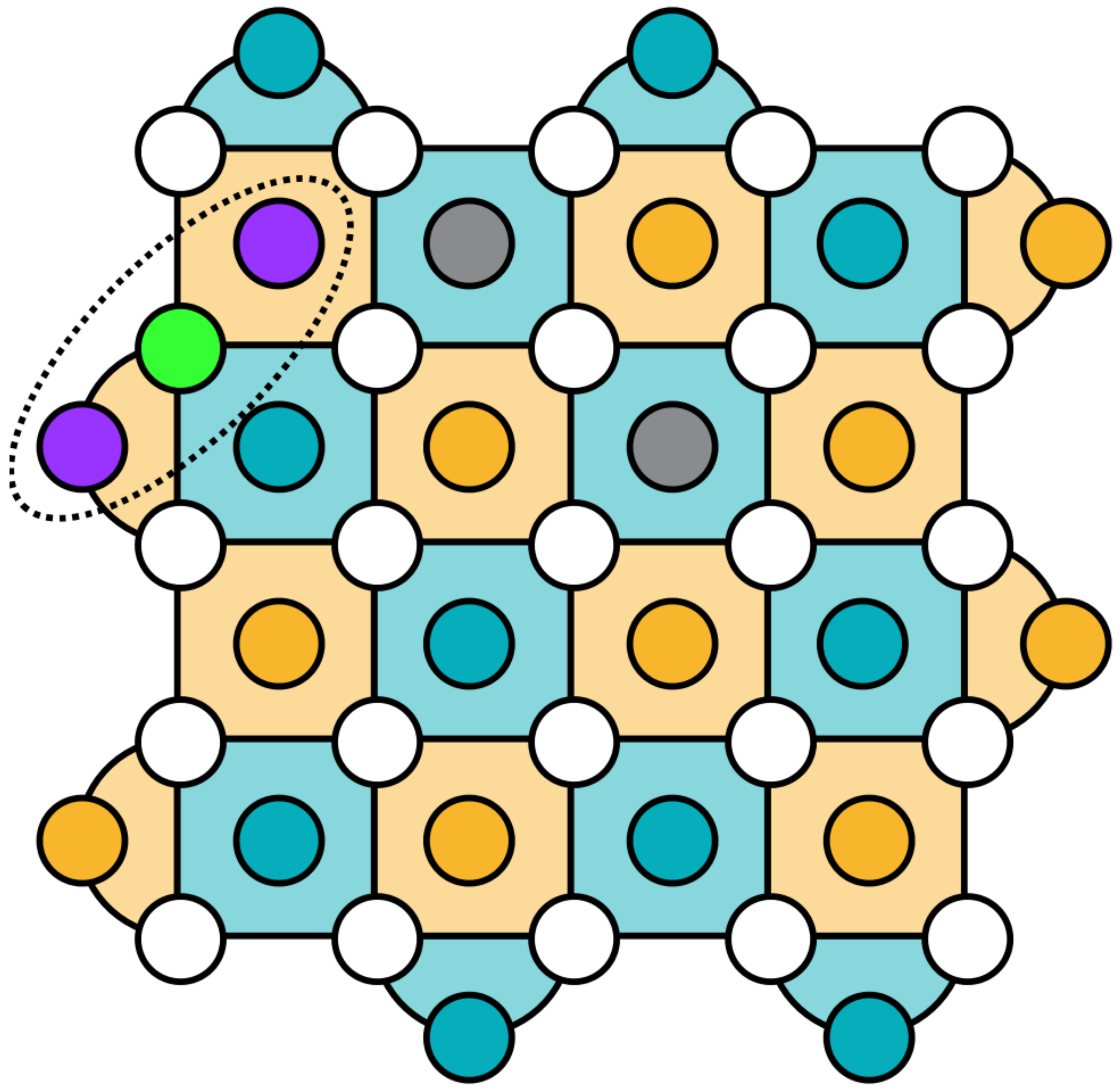
Challenge of QEC

- Complicated syndromes

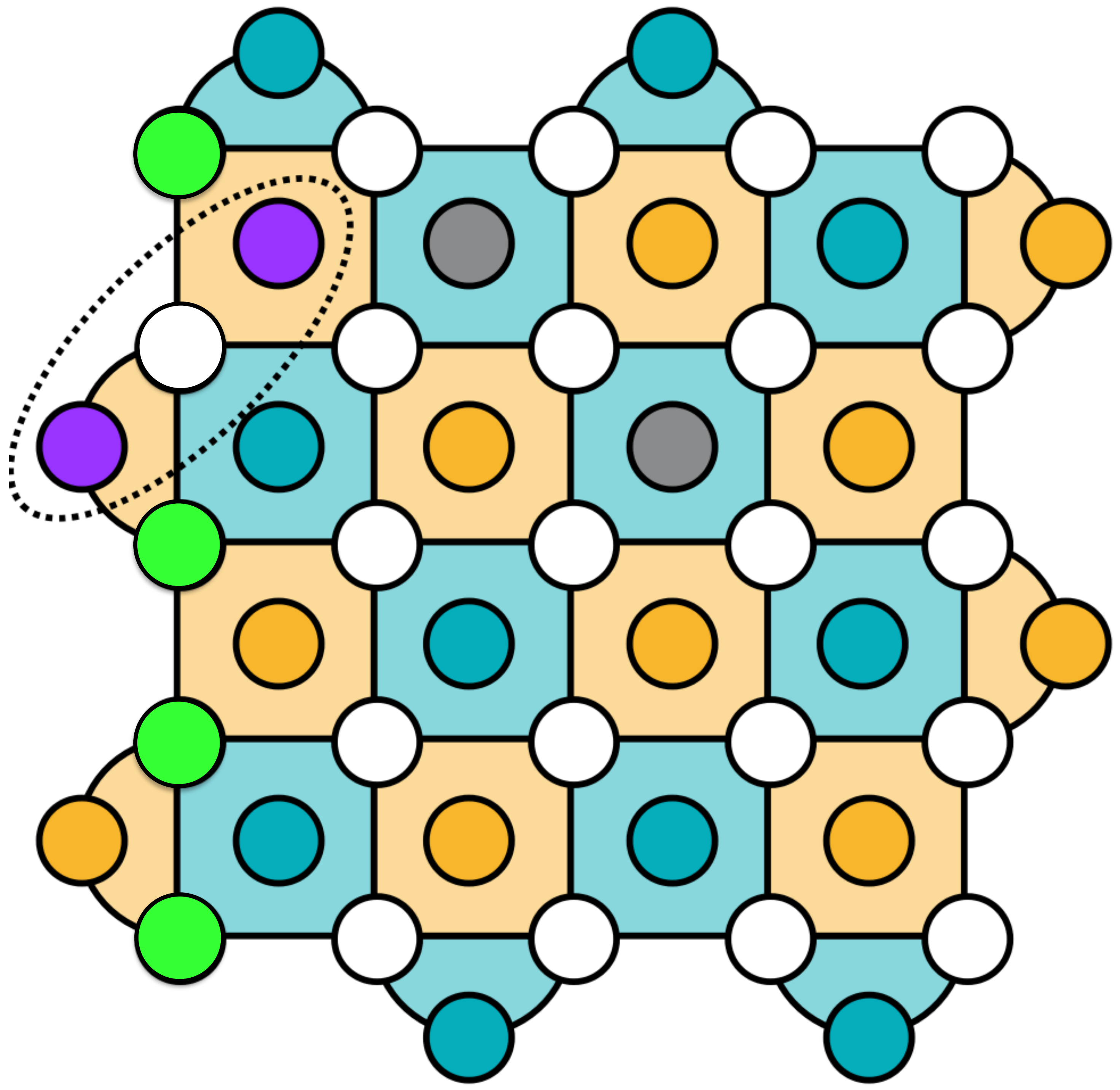

Odd number of flips remains
Even number of flips cancel out
Parity Check

Challenge of QEC

- Degeneracy



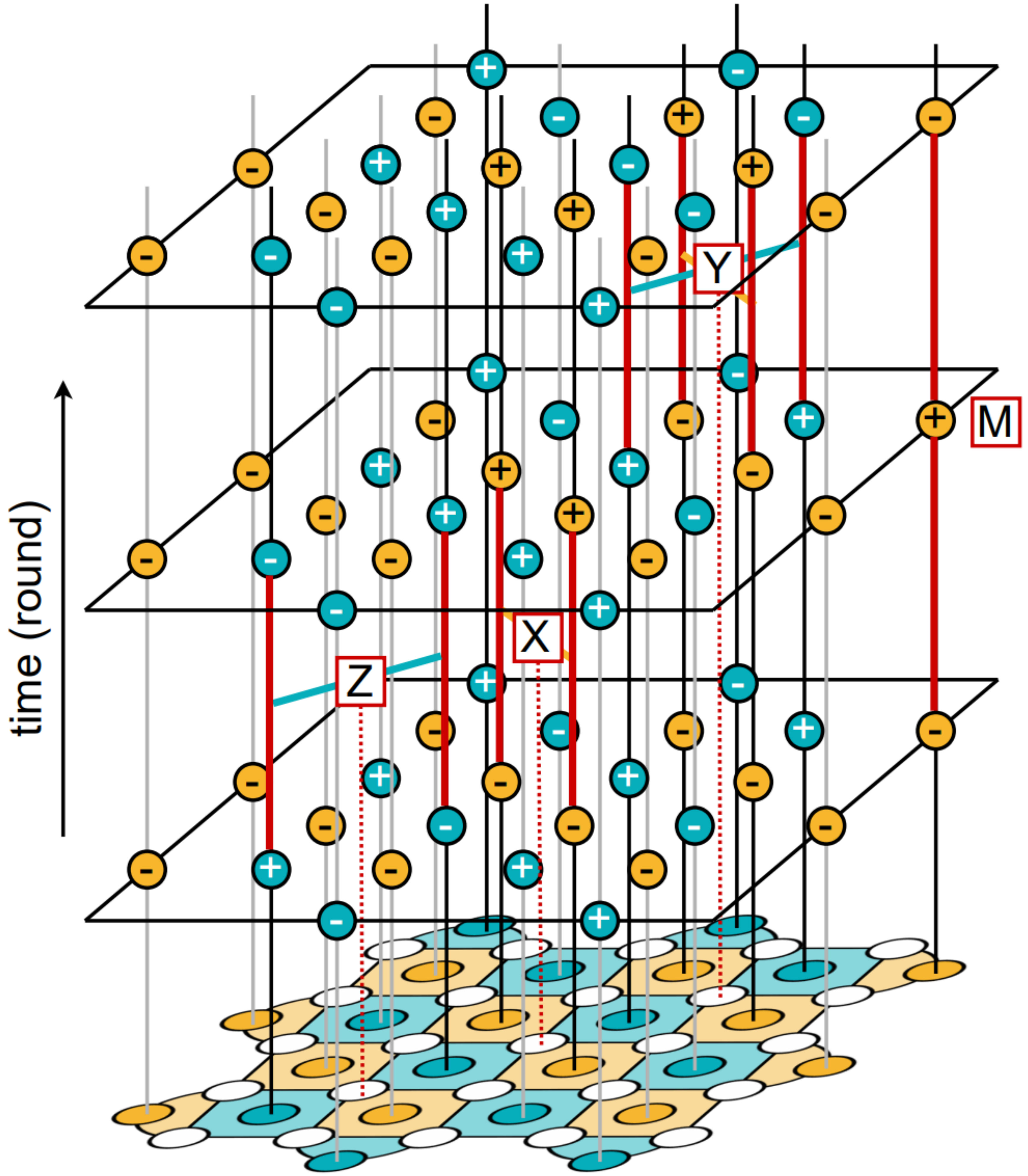
● Data Qubit X Error ● Flipped Z-type Syndrome Qubit



● Data Qubit X Error ● Flipped Z-type Syndrome Qubit

Challenge of QEC

- Complicated graph for practical case
- Multiple rounds of extraction
- Extraction itself may contains errors

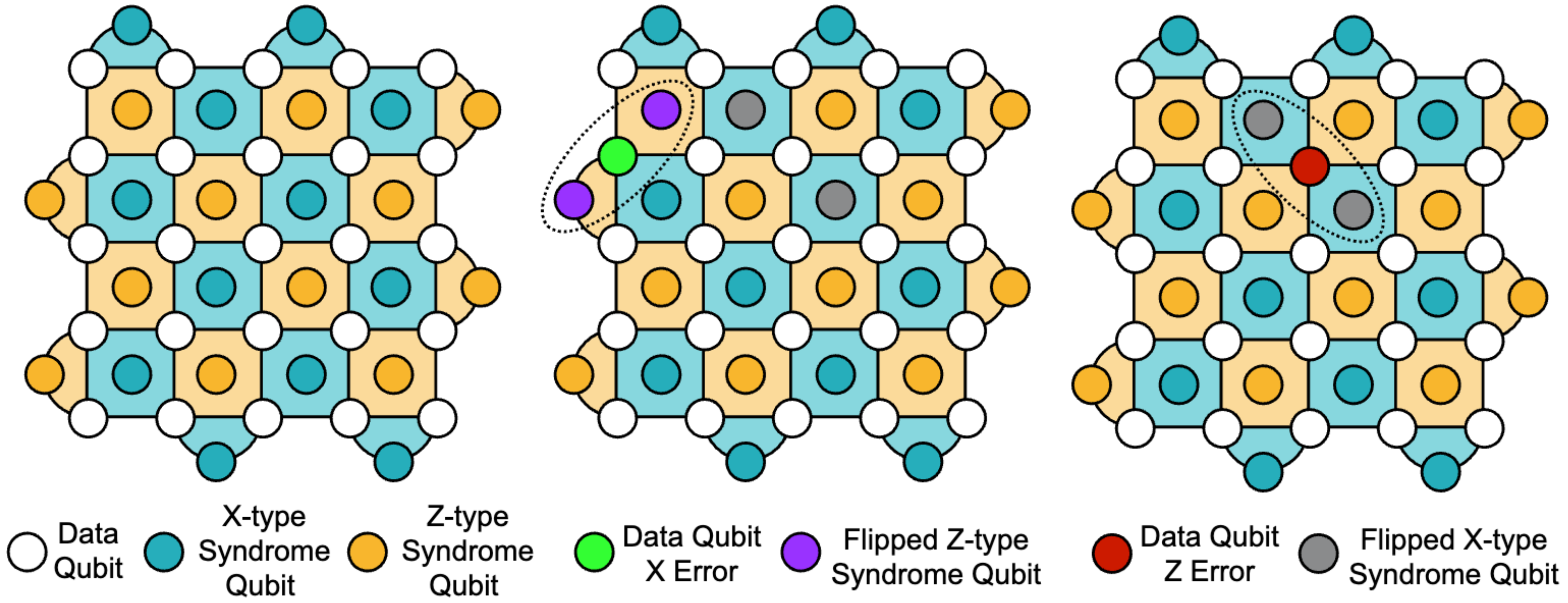


The iterative correction process

- Quantum Error Correction

Error correction code running on quantum

Decoder running on classical

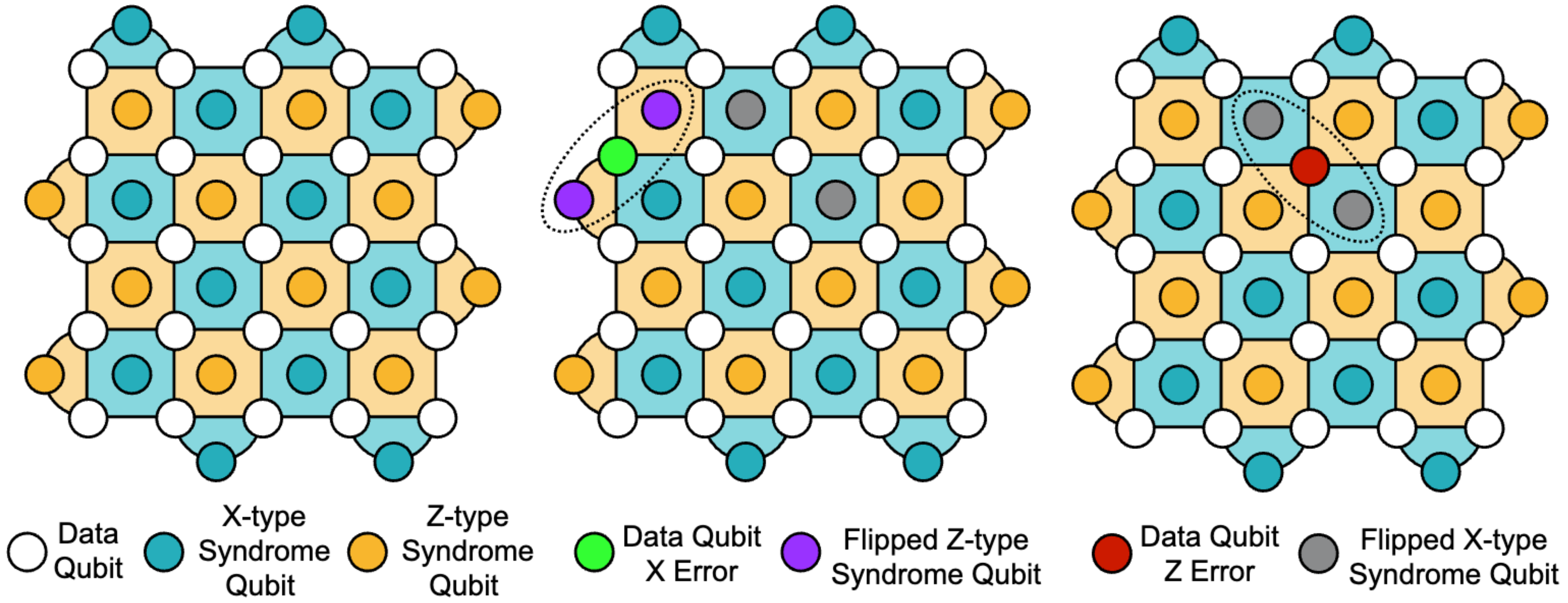


The iterative correction process

- Quantum Error Correction

Error correction code running on quantum

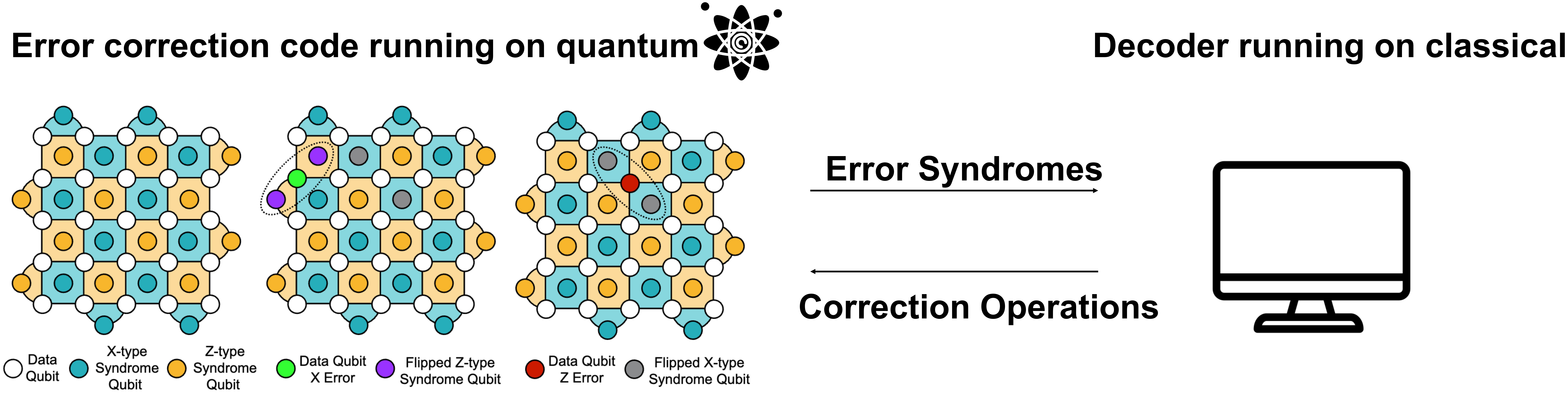
Decoder running on classical



Error Syndromes →

The iterative correction process

- Quantum Error Correction



Error correction code running on quantum

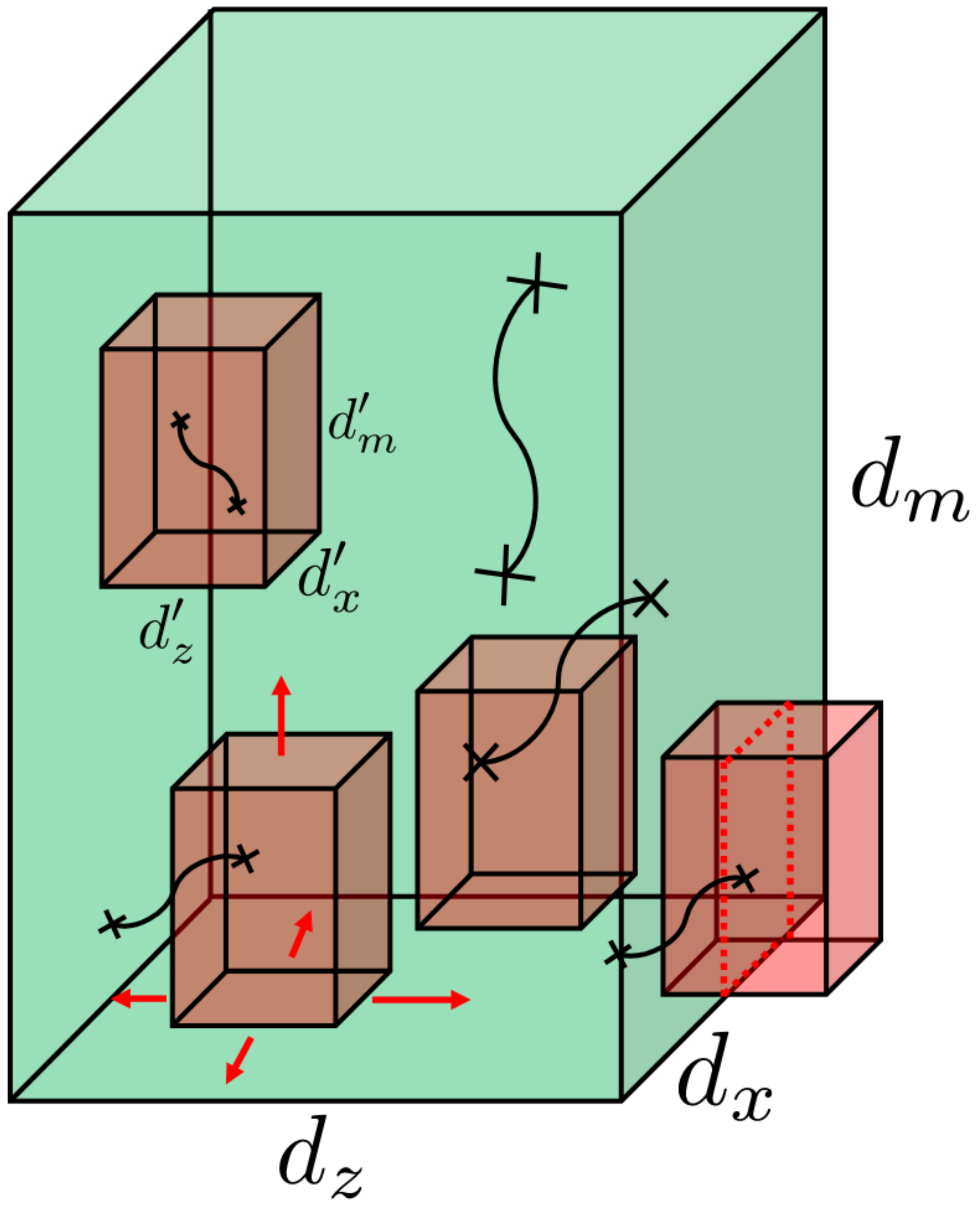
Decoder running on classical

Error Syndromes

Correction Operations

ML-based Decoders

- Reduced decoding time
- Adaptable to various noise models
- Easy for retraining for performance optimization
- Different models has been proposed
  - MLP, 3D convolution, Graph Neural Networks



Krastanov, Stefan, and Liang Jiang. "Deep neural network probabilistic decoder for stabilizer codes." *Scientific reports* 7.1 (2017): 11003.
Varsamopoulos, Savvas, Ben Criger, and Koen Bertels. "Decoding small surface codes with feedforward neural networks." Quantum Science and Technology 3.1 (2017): 015004.
Chamberland, Christopher, et al. "Techniques for combining fast local decoders with global decoders under circuit-level noise." arXiv preprint arXiv:2208.01178 (2022).

Challenges of ML Decoder

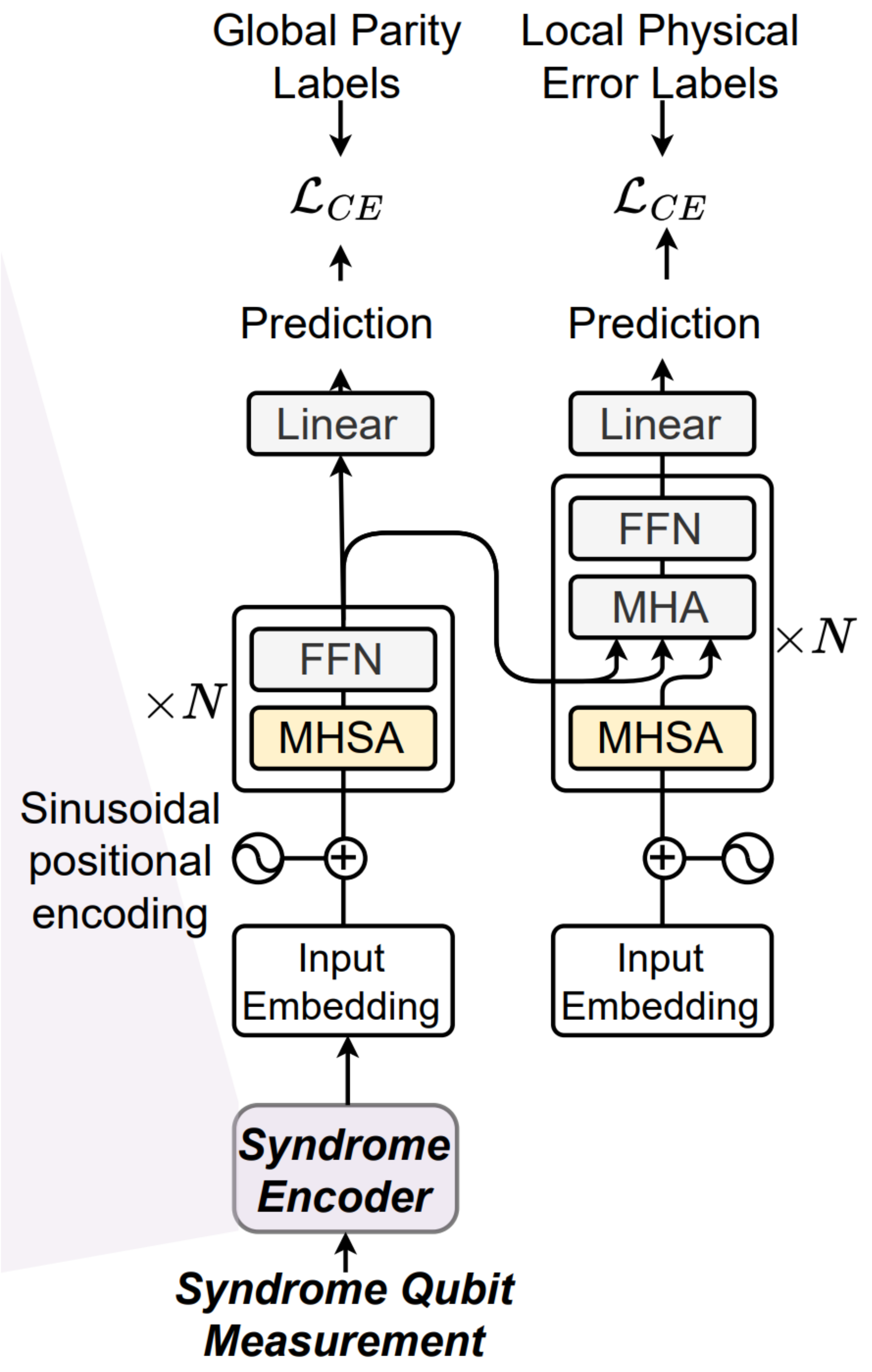
- Large training cost for different distance of QEC codes
- Efficiency and speed of the ML model

Proposed Transformer Based QEC

- Easy transfer learning between different code distance with **Transformer model**
- Specific **hardware accelerator** for Transformer ML

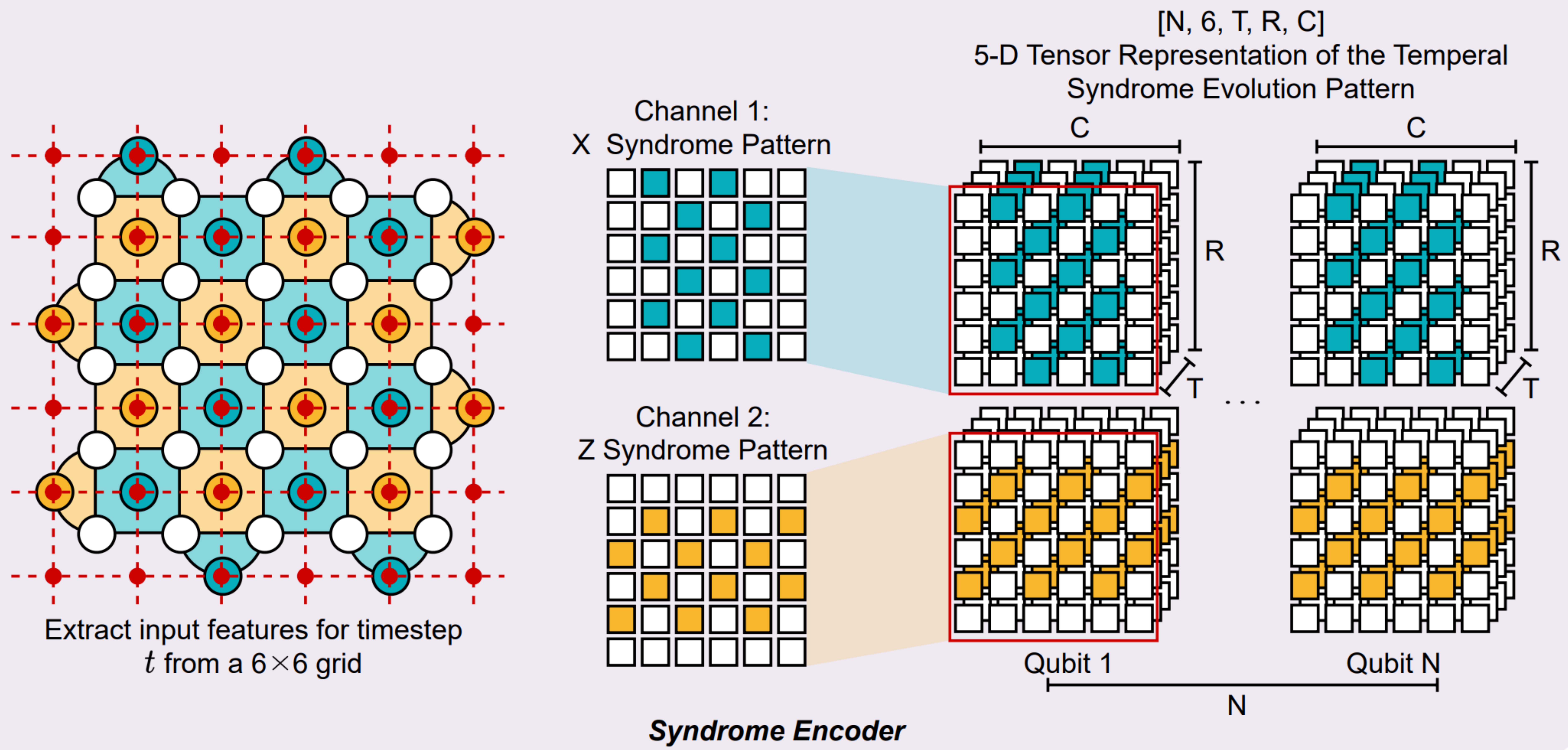
Model Architecture

- Transformer-Encoder to process the input syndromes
- Transformer-Decoder to predict the error on each of the qubit



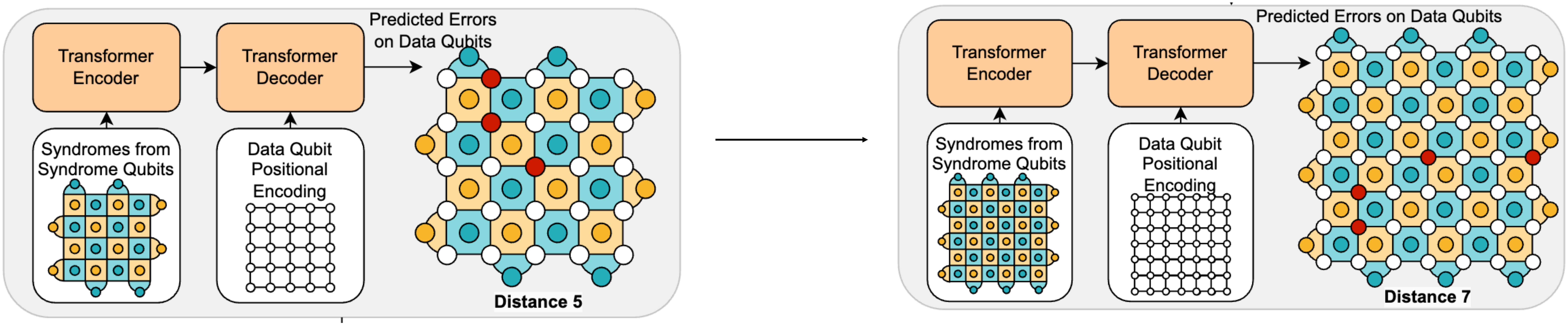
Input Features

- Features contains the locations and the binary syndrome value
- 3D positional positional encoding


[N, 6, T, R, C]
5-D Tensor Representation of the Temperal Syndrome Evolution Pattern
Channel 1:
X Syndrome Pattern
Channel 2:
Z Syndrome Pattern
C
R
T
C
R
T
. . .
Qubit 1
Qubit N
N
Extract input features for timestep t from a 6×6 grid
Syndrome Encoder

Transformer based QEC decoder

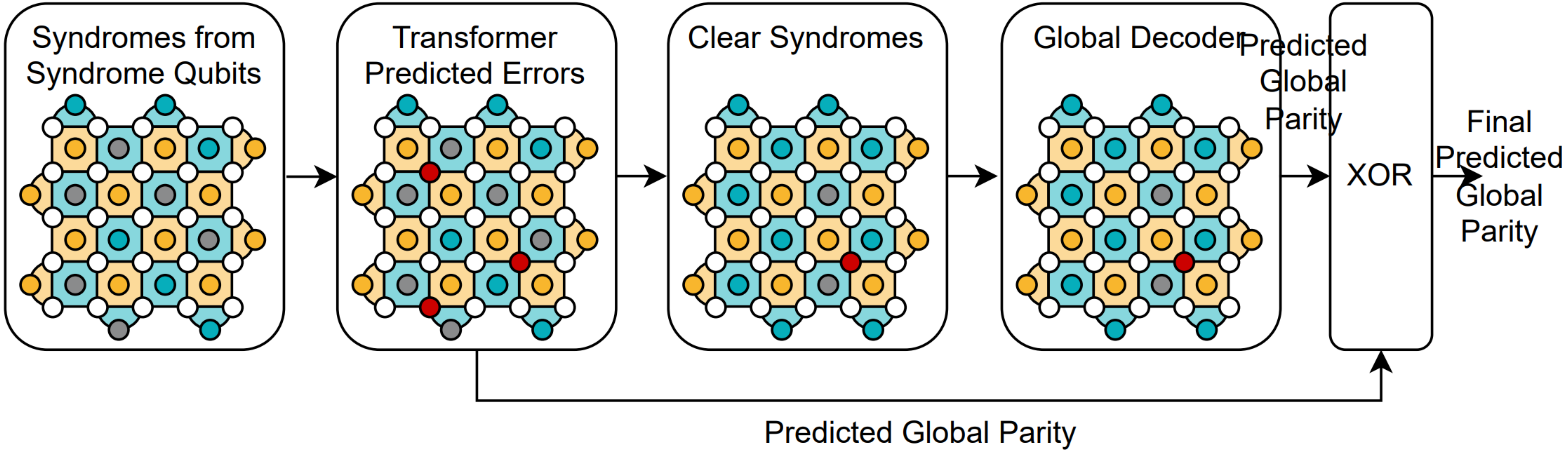
- Transfer learning to other code distances


Transformer Encoder
Transformer Decoder
Predicted Errors on Data Qubits
Syndromes from Syndrome Qubits
Data Qubit Positional Encoding
Distance 5
Transformer Encoder
Transformer Decoder
Predicted Errors on Data Qubits
Syndromes from Syndrome Qubits
Data Qubit Positional Encoding
Distance 7

H. Wang, et al. “TransformerQEC: Transferable Transformer for Quantum Error Correction Code Decoding." *FASTML@ICCAD 2023*

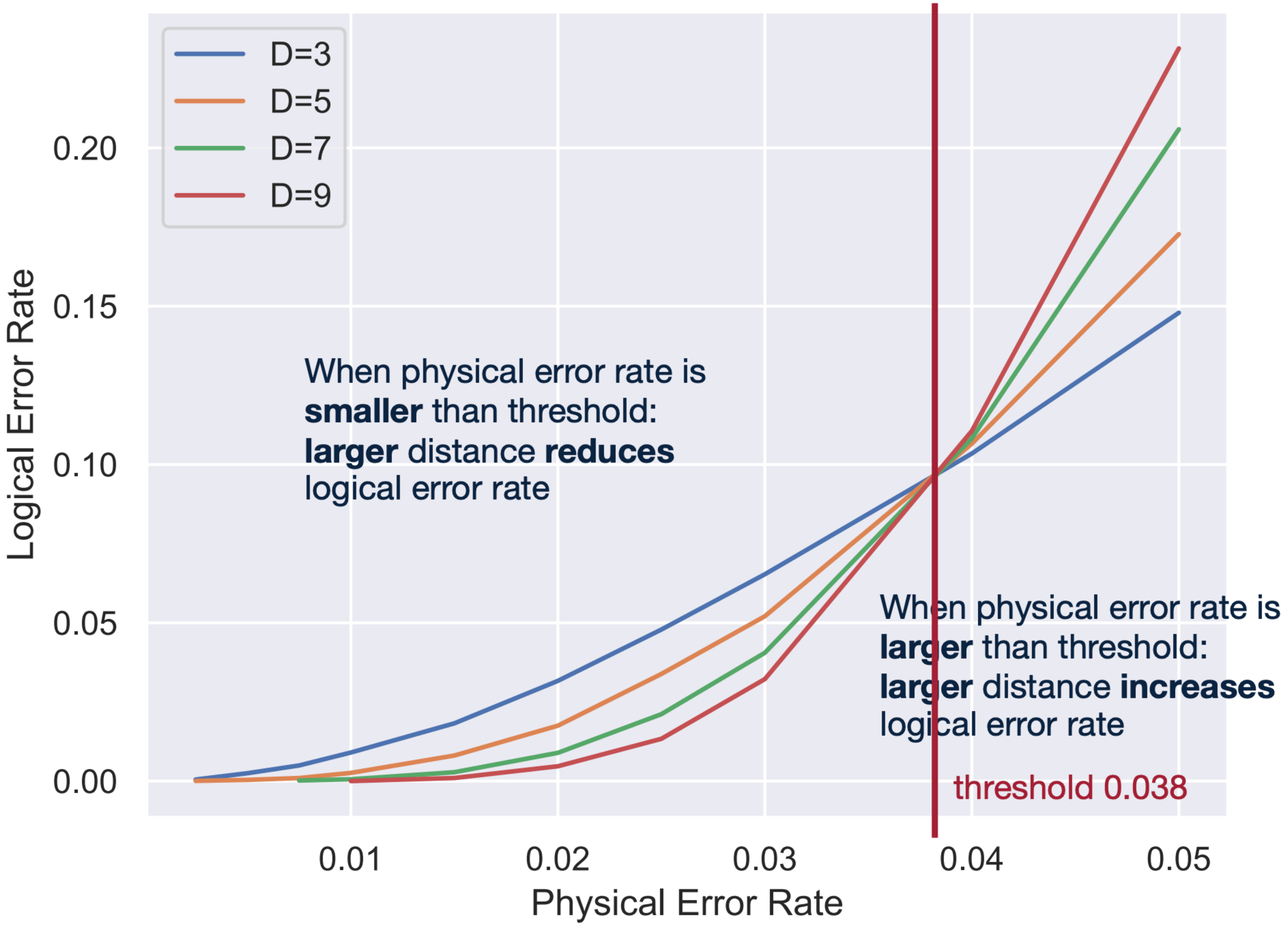
Whole Pipeline

- One final layer of global decoder for the


Syndromes from Syndrome Qubits
Transformer Predicted Errors
Clear Syndromes
Global Decoder
Predicted Global Parity
XOR
Final Predicted Global Parity
Predicted Global Parity

Accuracy Results

- Low logical error rate with QEC



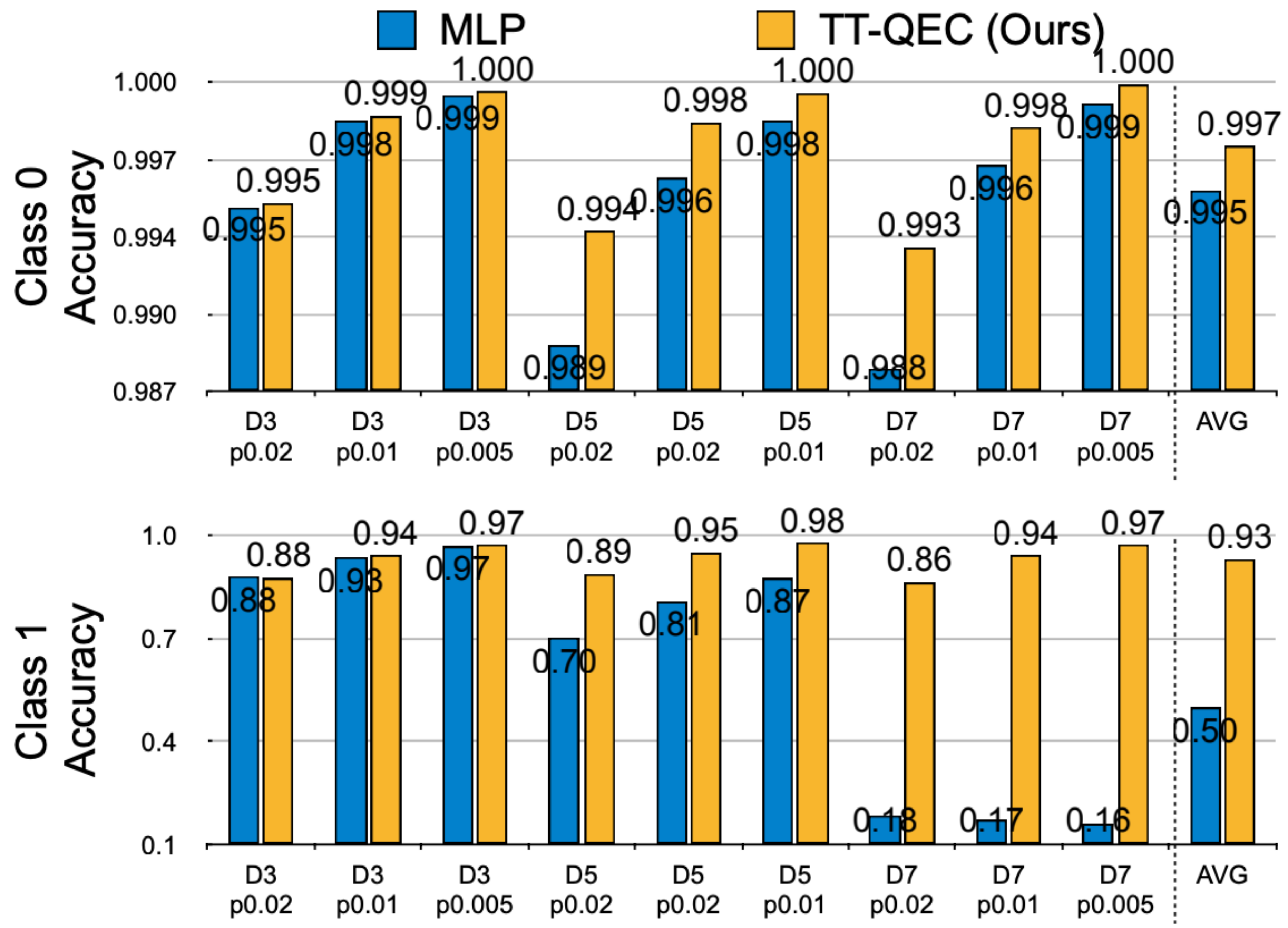
When physical error rate is **smaller** than threshold: **larger** distance **reduces** logical error rate

When physical error rate is **larger** than threshold: **larger** distance **increases** logical error rate

Accuracy Results

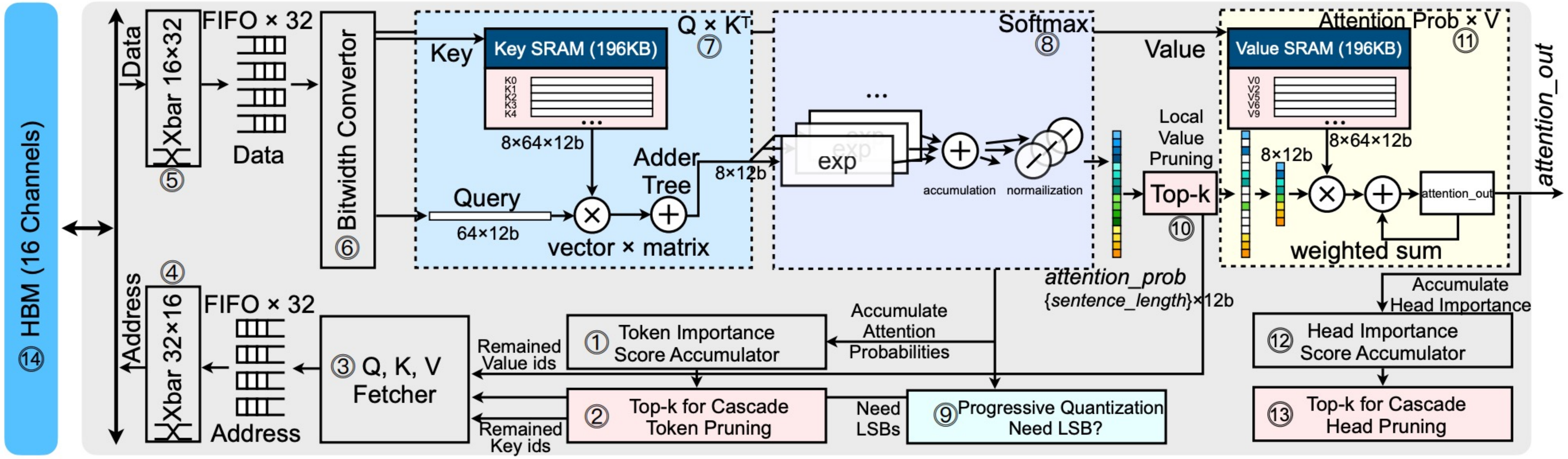
- Lower logical error rate than baseline methods

		Logical Error Rate ↓			
Distance	Phys. Err. Rate	UF	MWPM	MLP	TT-QEC
3	0.0500	0.16745	0.14063	0.14794	**0.13005**
	0.0100	0.01039	0.00800	0.00903	**0.00784**
5	0.0500	0.24120	0.17279	0.20888	**0.17232**
	0.0100	0.00406	0.00268	0.00443	**0.00254**
7	0.0500	0.29813	0.20178	0.28454	**0.20590**
	0.0100	0.00113	0.00064	0.00197	**0.00059**
9	0.0500	0.35250	0.23161	0.32770	**0.23144**
	0.0100	0.00028	0.00002	0.00017	**0.00001**

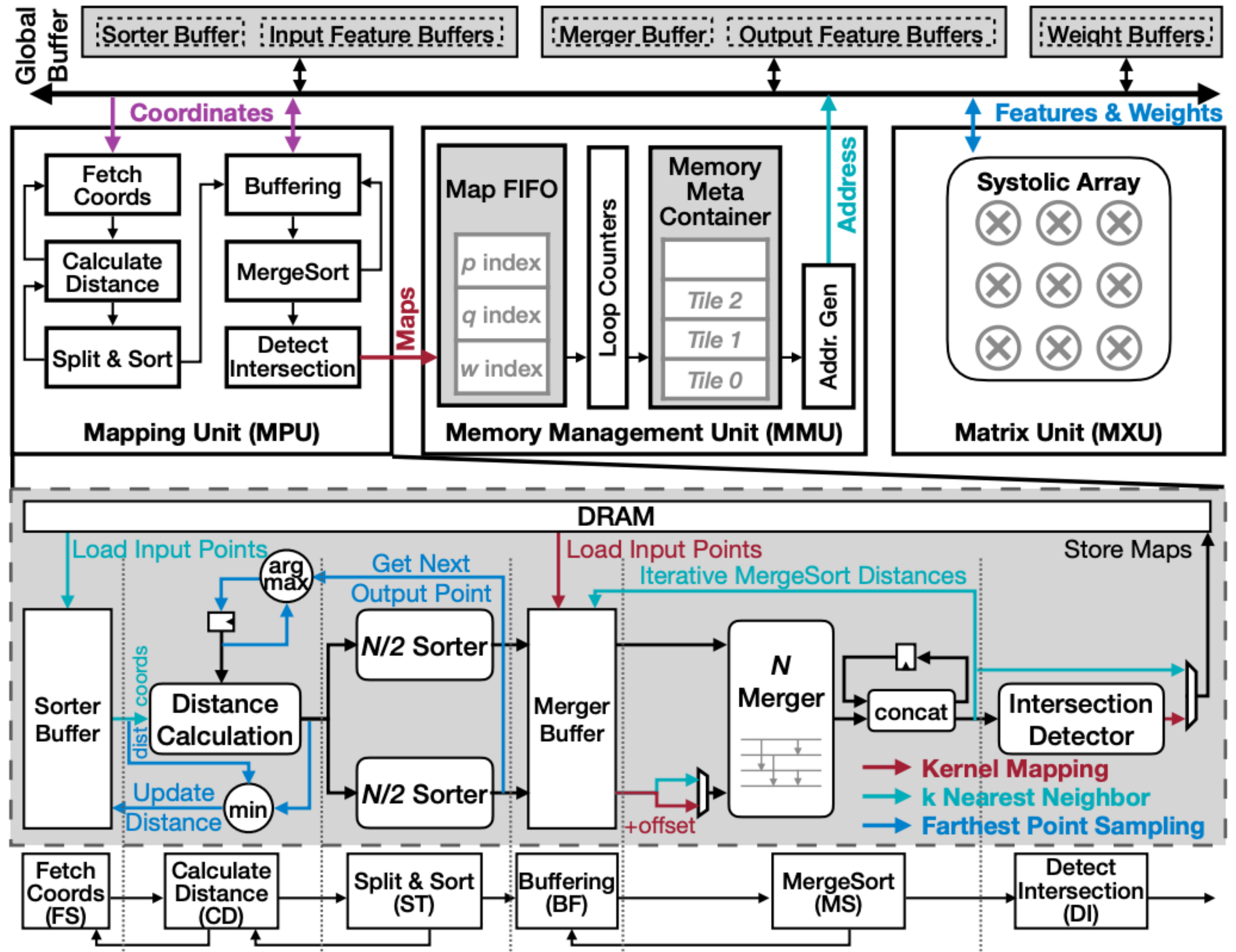


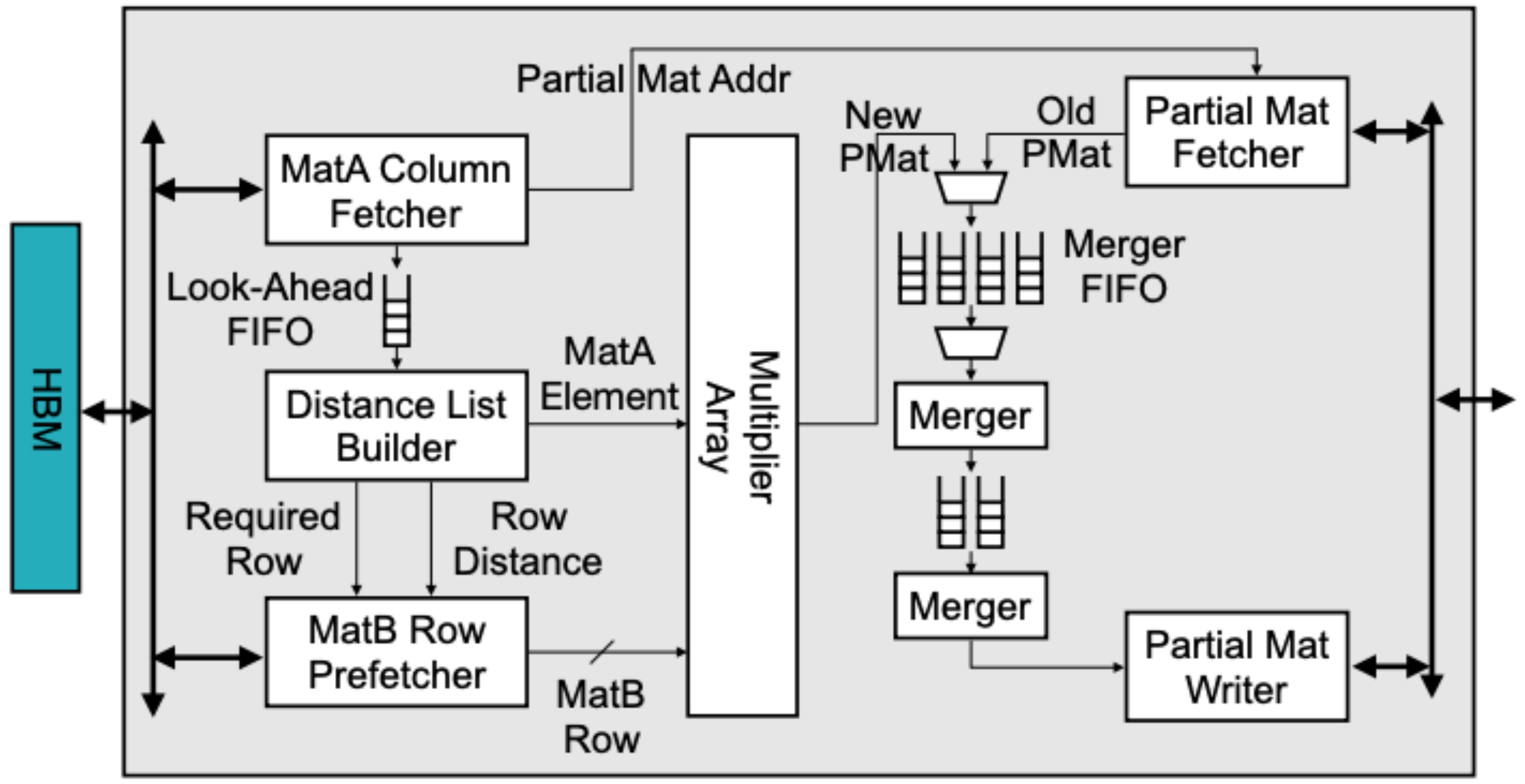
How to further improve the efficiency of ML for Quantum Science?

Classical Accelerator Support



SpAtten for Transformer Acceleration [HPCA'21]



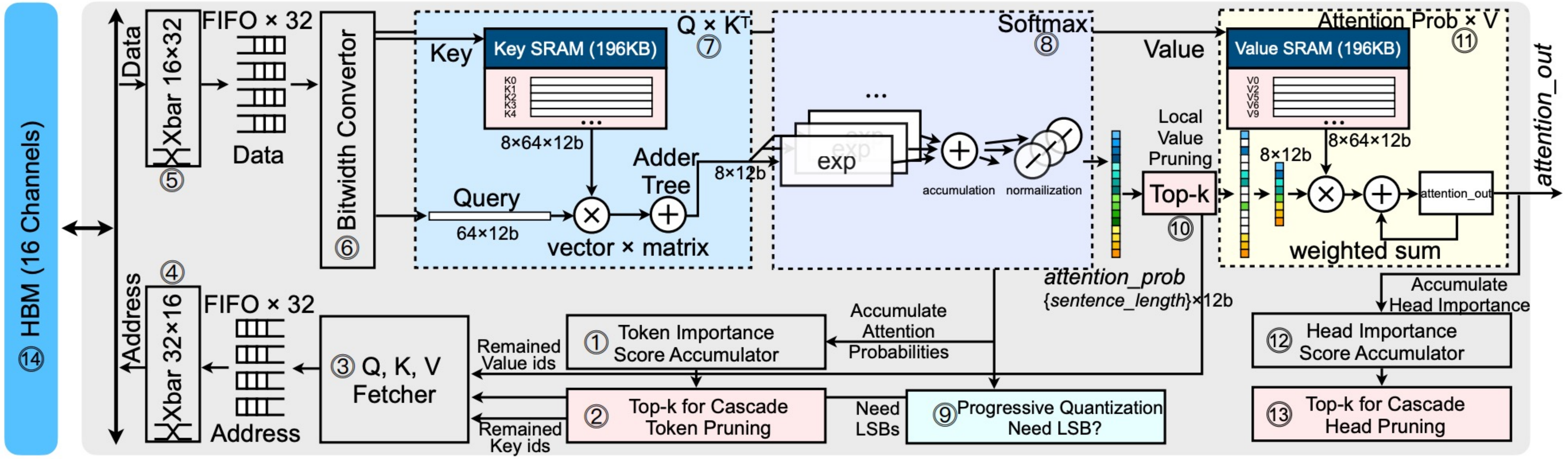


SpArch for sparse matrix multiplication [HPCA'20]

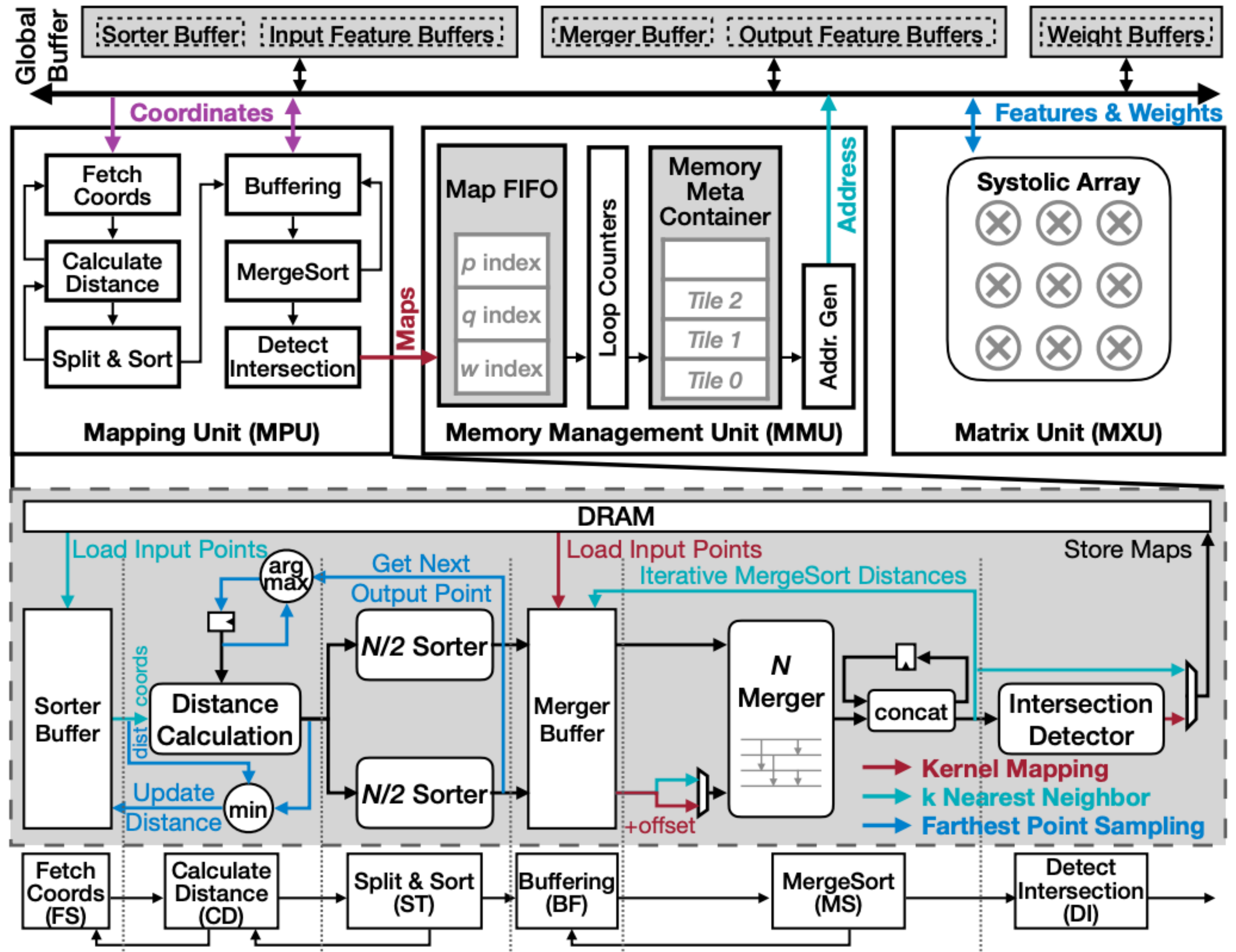
PointAcc for 3D Conv Acceleration [MICRO'21]

Hanrui Wang, Zhekai Zhang, and Song Han. "Spatten: Efficient sparse attention architecture with cascade token and head pruning." *HPCA* 2021.
Zhang, Zhekai*, Hanrui Wang* (co-first). "Sparch: Efficient architecture for sparse matrix multiplication." *HPCA*. 2020.
Yujun Lin, Zhekai Zhang, Haotian Tang, Hanrui Wang, Song Han "Pointacc: Efficient point cloud accelerator." *MICRO*, 2021.

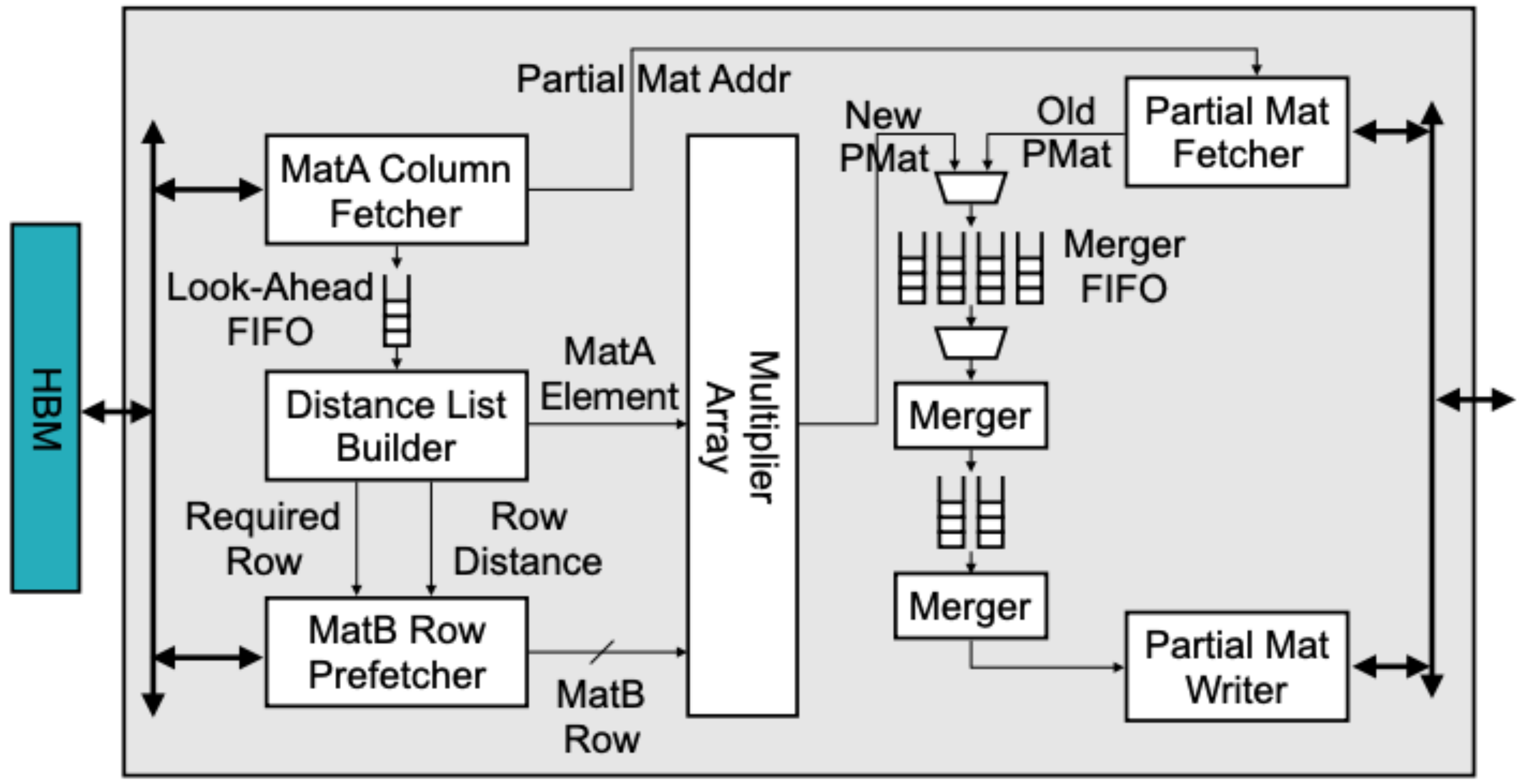
Classical Accelerator Support



SpAtten for Transformer Acceleration [HPCA’21]



PointAcc for 3D Conv Acceleration [MICRO’21]



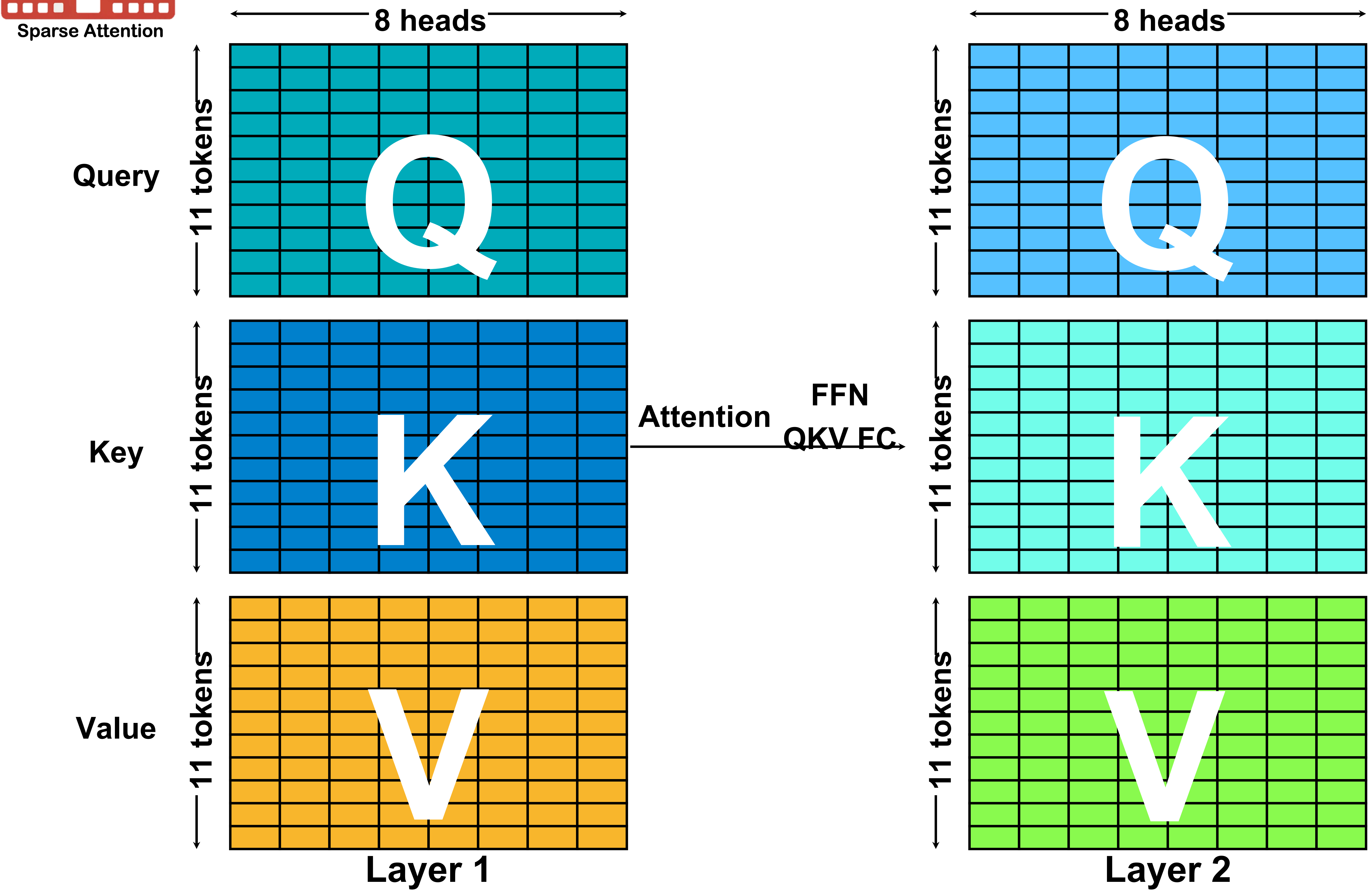
SpArch for sparse matrix multiplication [HPCA’20]

Hanrui Wang, Zhekai Zhang, and Song Han. "Spatten: Efficient sparse attention architecture with cascade token and head pruning." *HPCA* 2021.
Zhang, Zhekai*, Hanrui Wang* (co-first). "Sparch: Efficient architecture for sparse matrix multiplication." *HPCA*. 2020.
Yujun Lin, Zhekai Zhang, Haotian Tang, Hanrui Wang, Song Han "Pointacc: Efficient point cloud accelerator." *MICRO*, 2021.

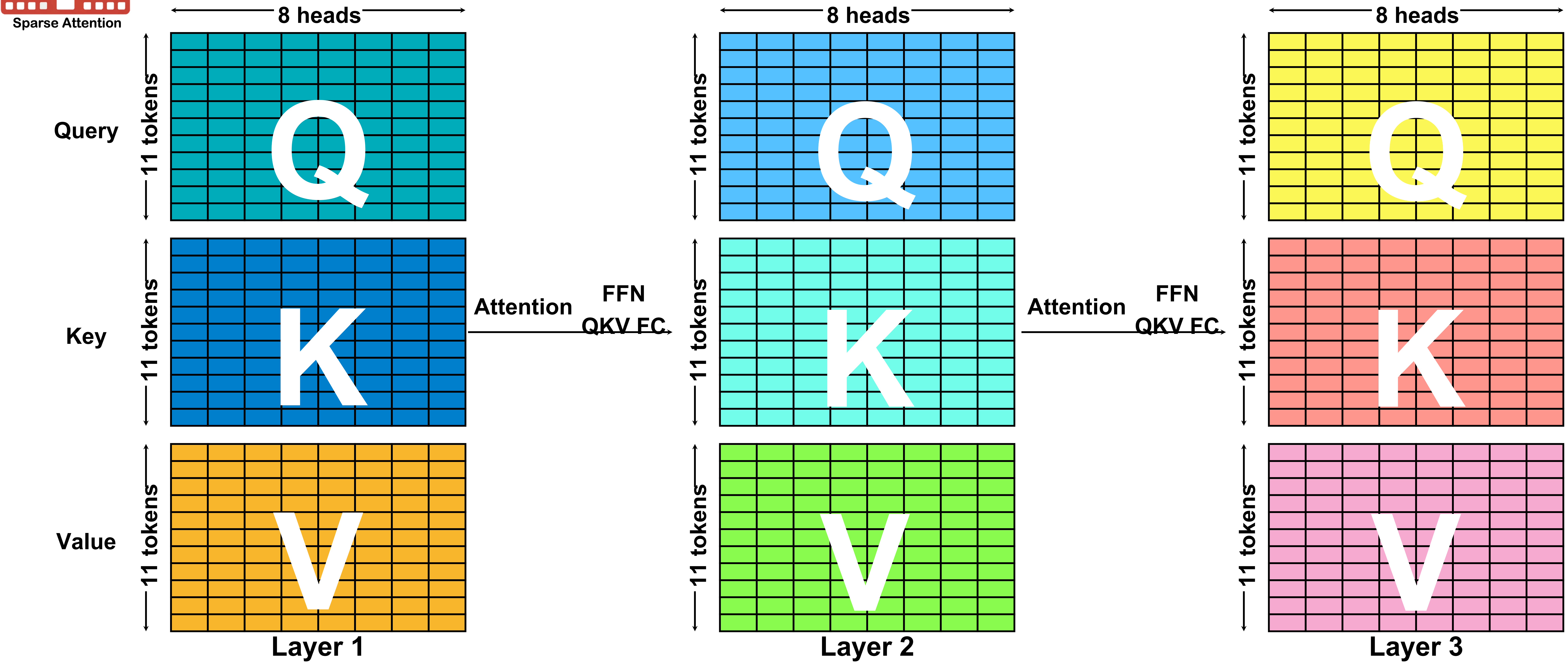
Cascade Token/Head Pruning

8 heads
Head 8
Query
Key
Value
11 tokens
11 tokens
11 tokens
Q * K^T
11
11
Attention Score
Row-wise Softmax
11
11
Attention Probability
8 heads
Attention Probability * V
Compute next layer's Query, Key and Value

Cascade Token/Head Pruning

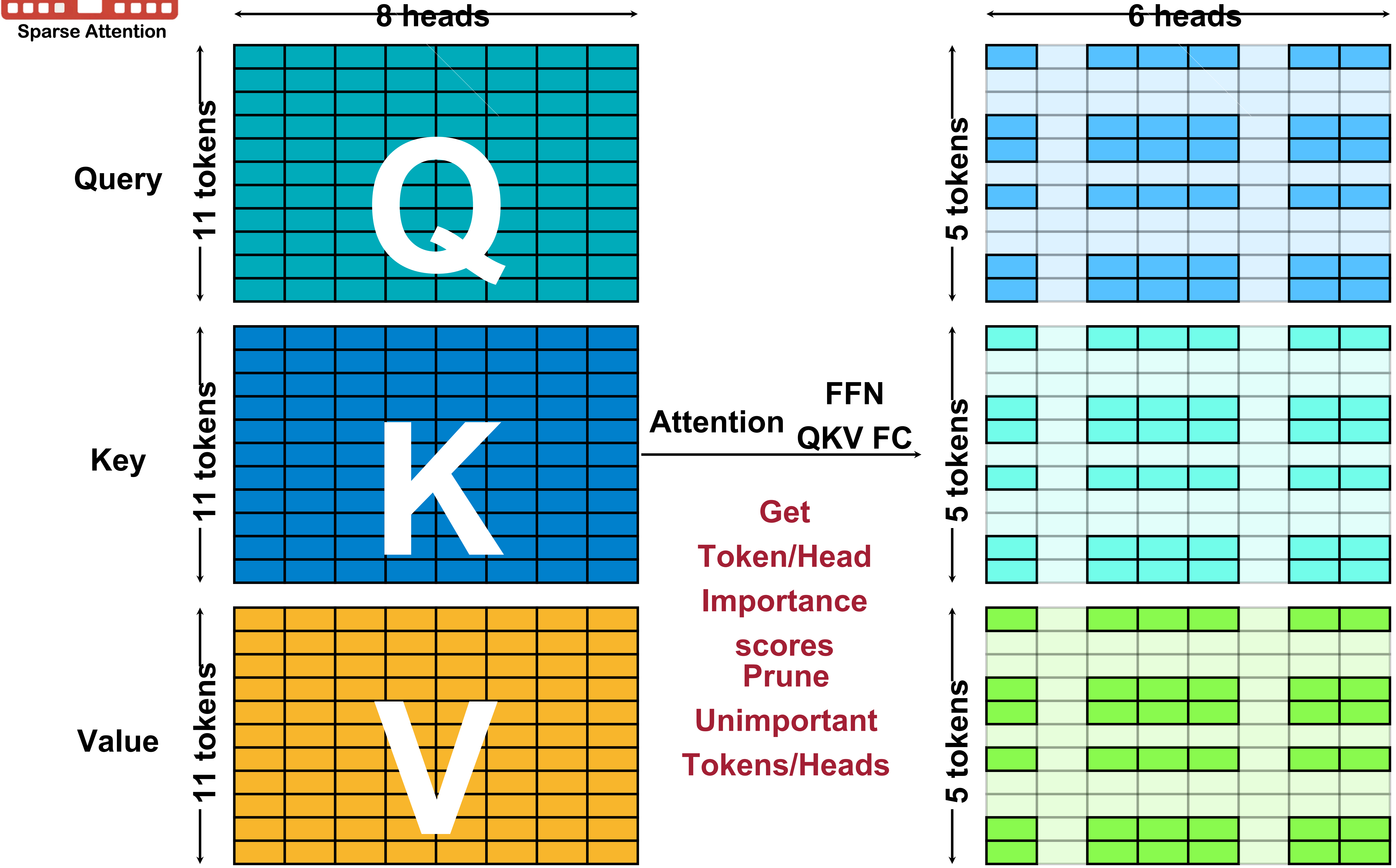


Cascade Token/Head Pruning



- Not all tokens/heads are created equal
- Find unimportant tokens and heads in front layers
- Remove them in latter layers

Cascade Token/Head Pruning



Get Token/Head Importance scores
Prune Unimportant Tokens/Heads

11 tokens, 8 heads

5 tokens, 6 heads

Cascade Token/Head Pruning

8 heads

Query

11 tokens

Q

Key

11 tokens

K

Value

11 tokens

V

11 tokens, 8 heads

Attention

FFN

QKV FC

Get Token/Head Importance scores

Prune Unimportant Tokens/Heads

6 heads

5 tokens

5 tokens

5 tokens

5 tokens, 6 heads

Attention

FFN

QKV FC

Get Token/Head Importance scores

Prune Unimportant Tokens/Heads

4 heads

2 tokens

2 tokens

2 tokens

2 tokens, 4 heads

Attention

FFN

Classifier

✅

- Pruned tokens/heads will never be used in all following layers: “Cascade”

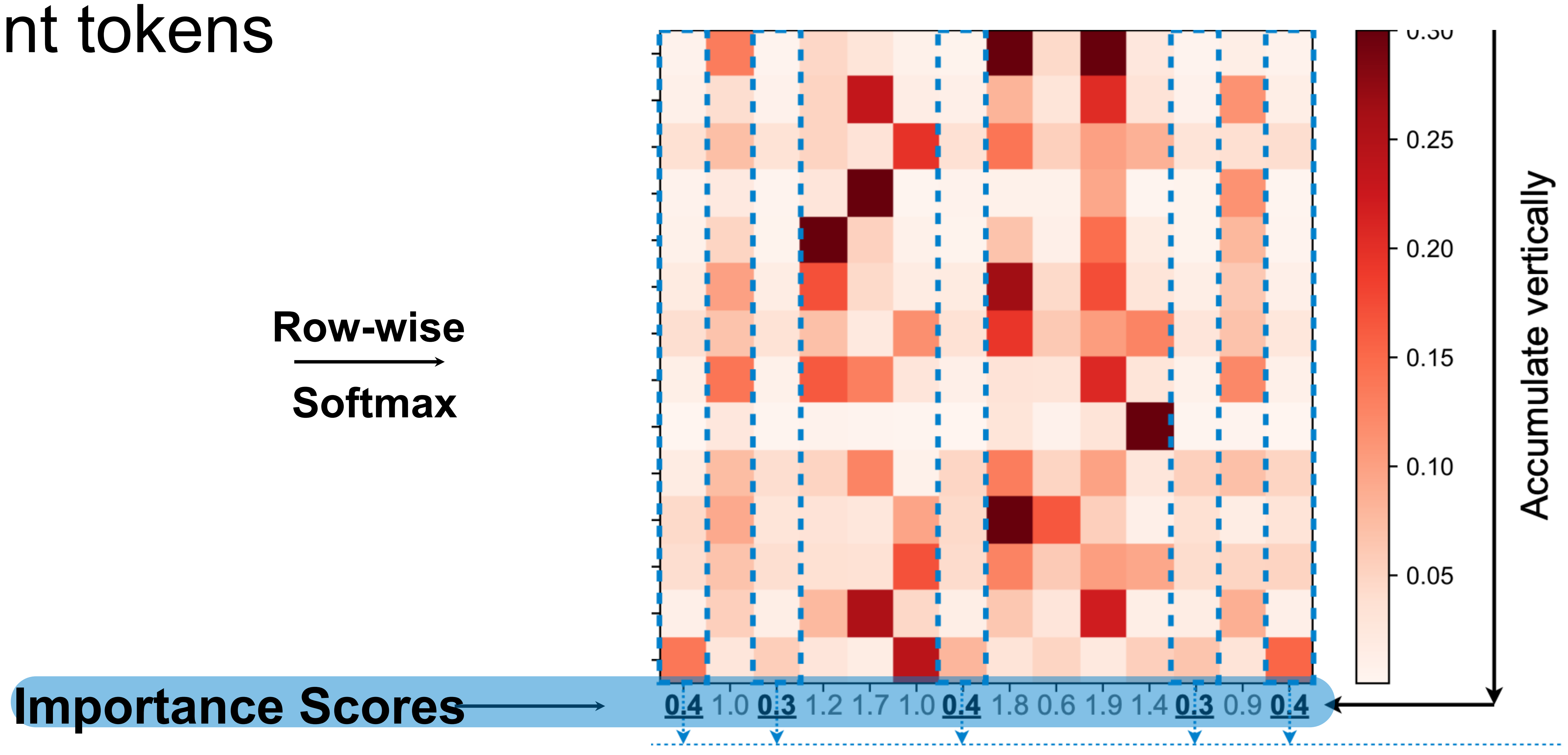
Cascade Token/Head Pruning

8 heads
6 heads
4 heads
Query
Key
Value
11 tokens
11 tokens
11 tokens
Q
K
V
5 tokens
5 tokens
5 tokens
2 tokens
2 tokens
2 tokens
Attention
FFN
QKV FC
Get Token/Head Importance scores
Prune Unimportant Tokens/Heads
Attention
FFN
QKV FC
Get Token/Head Importance scores
Prune Unimportant Tokens/Heads
Attention
FFN
Classifier
✅
11 tokens, 8 heads
5 tokens, 6 heads
2 tokens, 4 heads

- Pruned tokens/heads will never be used in all following layers: “Cascade”

Find Unimportant Tokens with Attention Probabilities

- If one column in attention probability is **small**: the token is **unimportant** to all other tokens
- Maintain an **importance score** for each token
- **Accumulate** attention probs to the importance scores
- **Top-k** scores indicate top-k important tokens



Row-wise Softmax

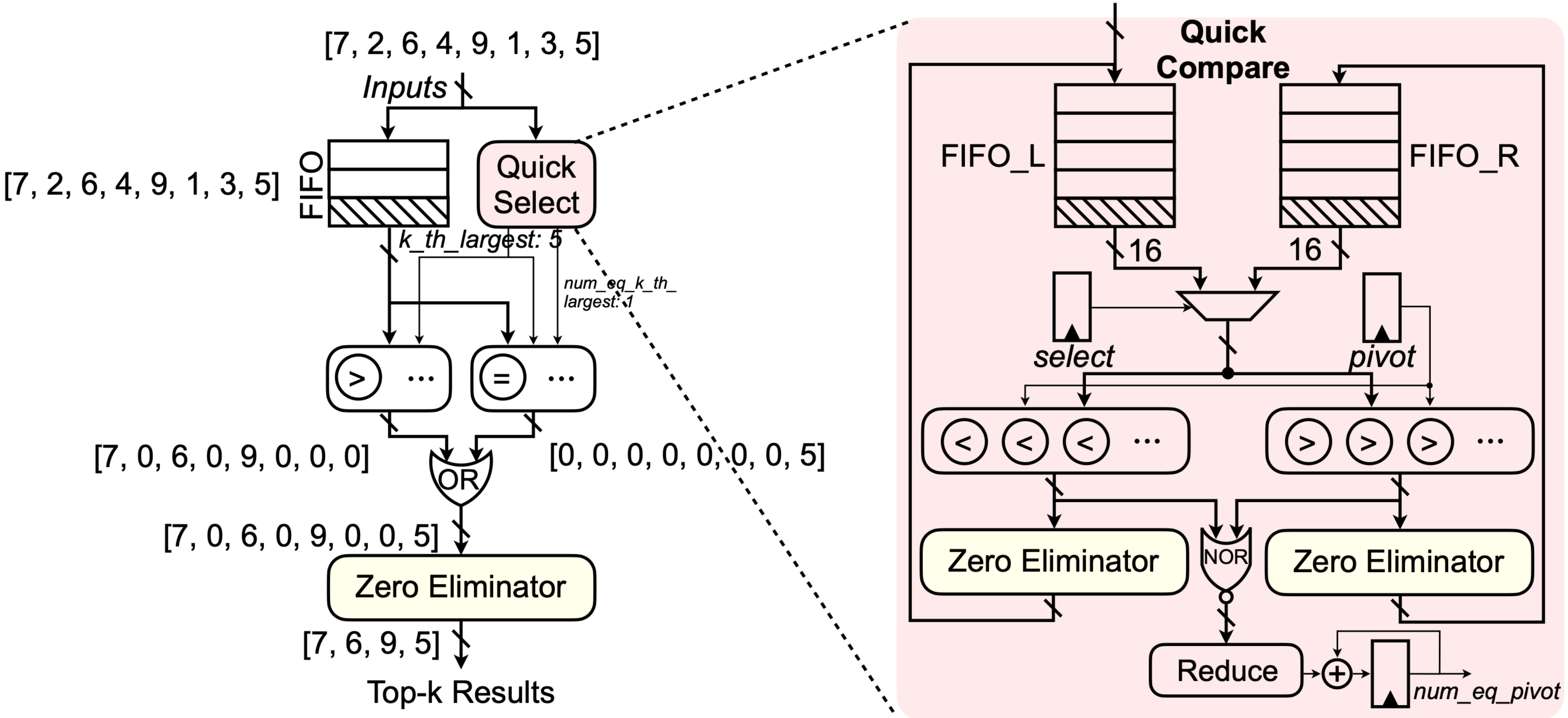
Importance Scores

Small score tokens can be pruned away

Attention Probability

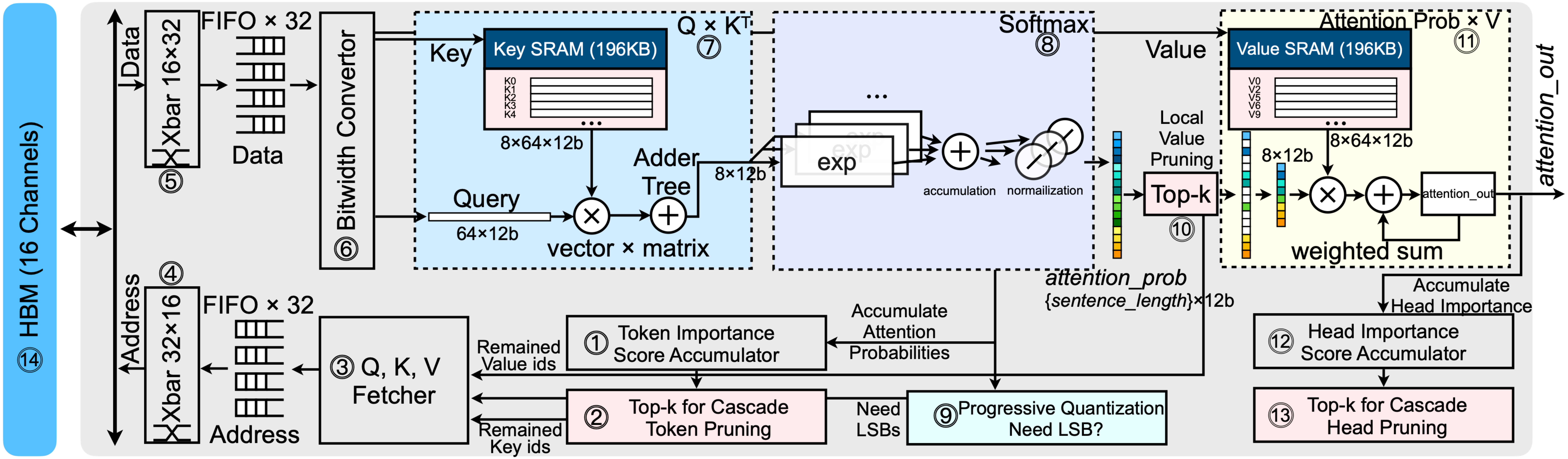
Quick Select

- Top-k Engine has **high-parallelism**
  - 16 ‘<’ comparators and 16 ‘>’ comparators in Quick Select
  - Compare the elements with pivot **in parallel**



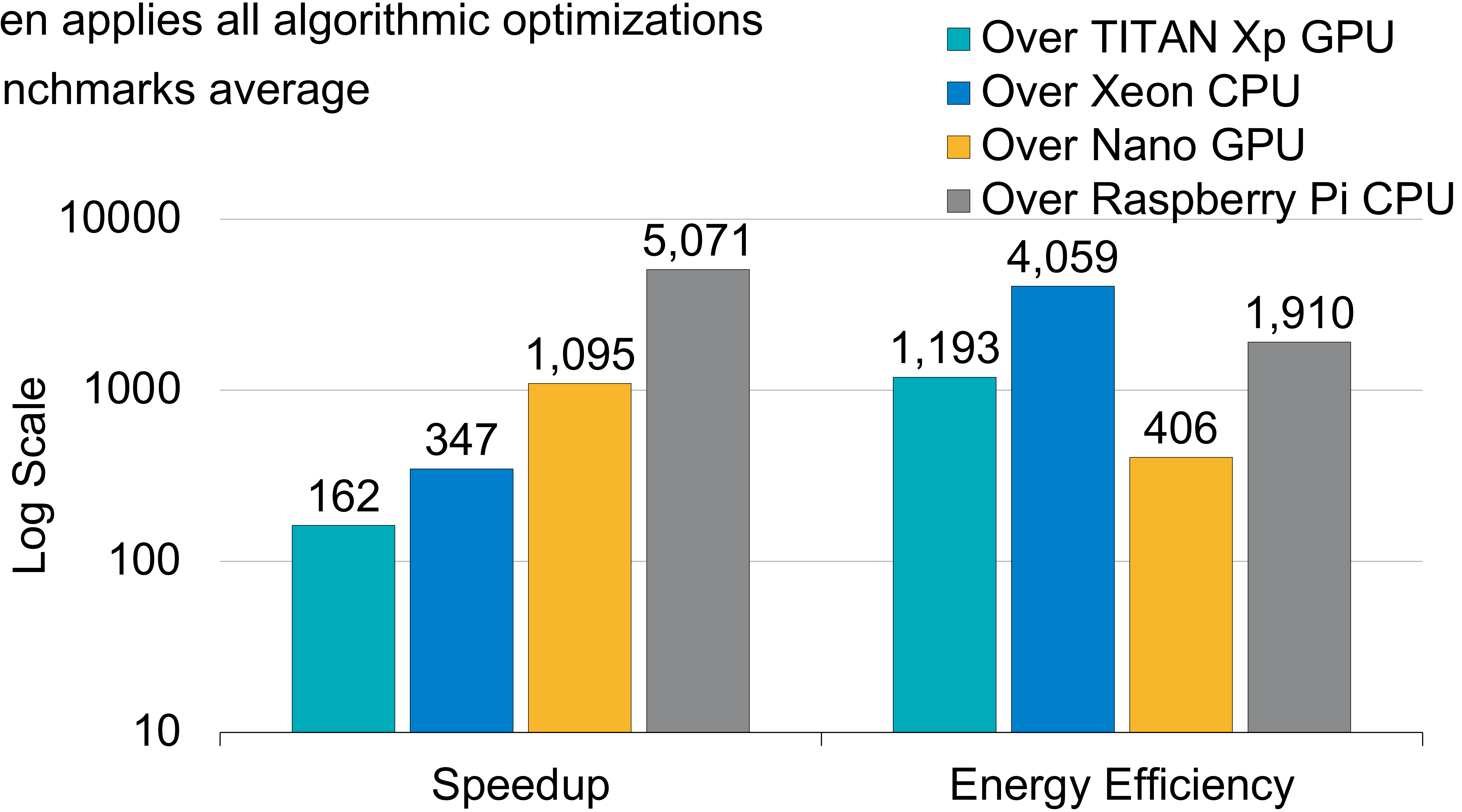
Dedicated Accelerator

- Pipelined architecture to improve the throughput



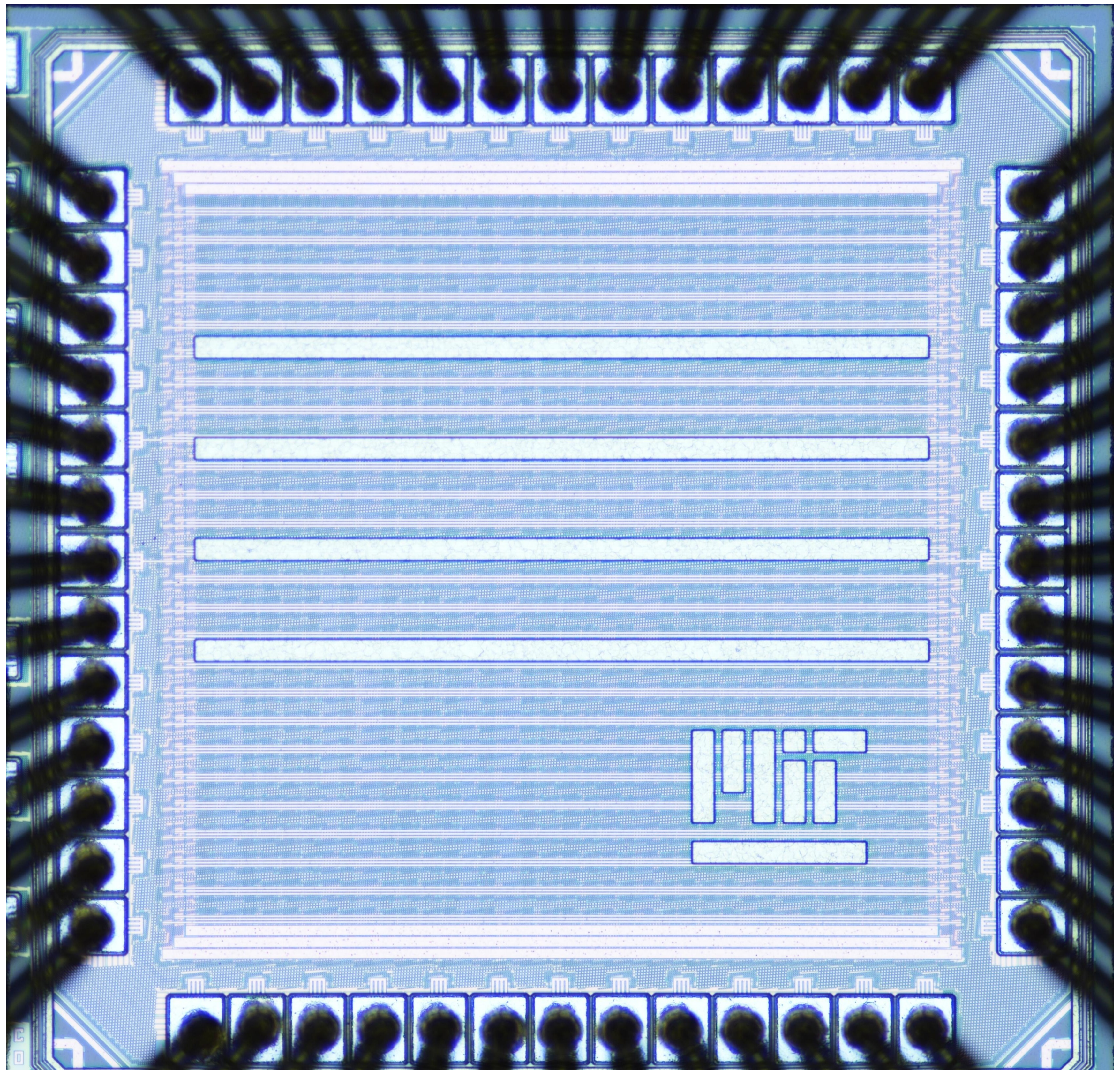
Performance Comparisons

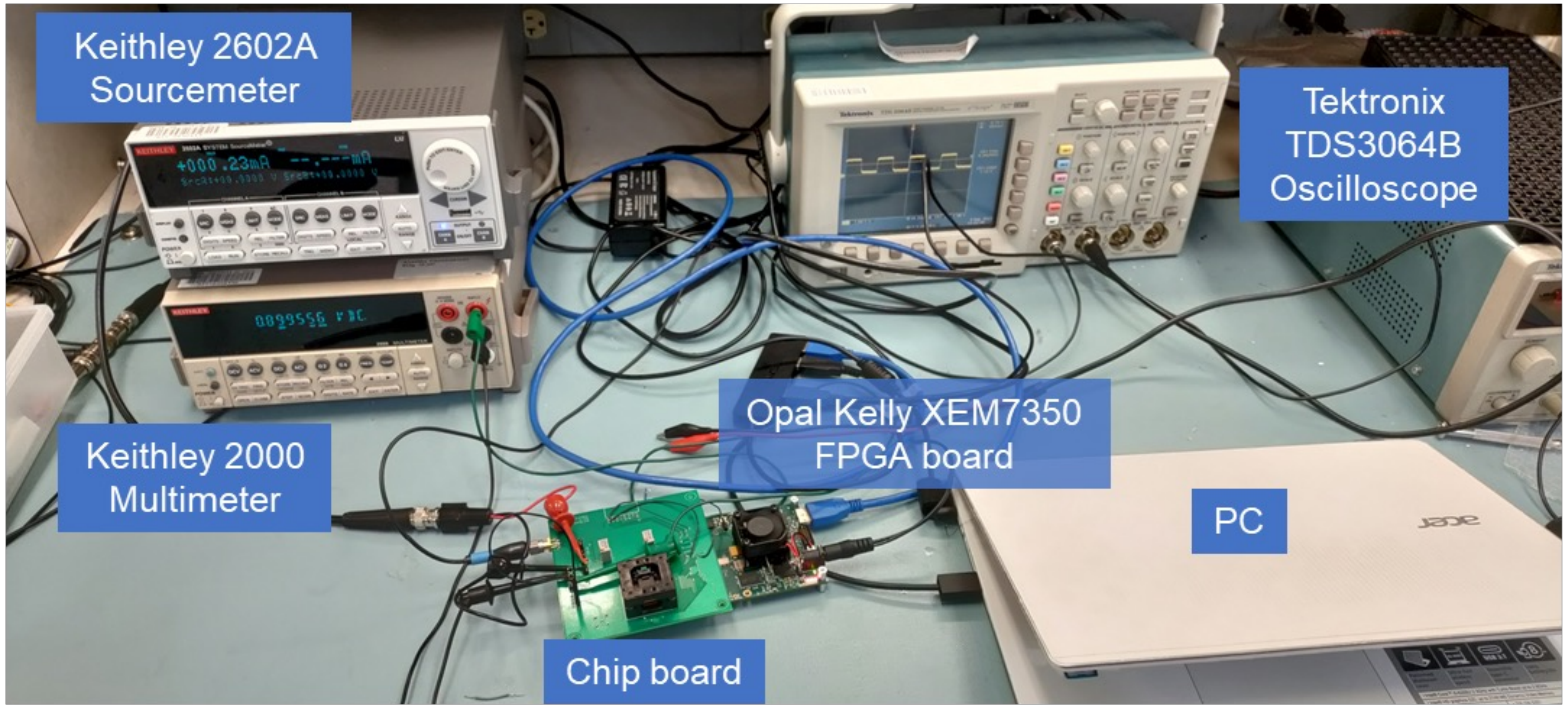
- Over general-purpose CPUs/GPUs on attention layers
  - SpAtten applies all algorithmic optimizations
  - 30 benchmarks average



SpAtten Transformer Accelerator & Chip Tape-out

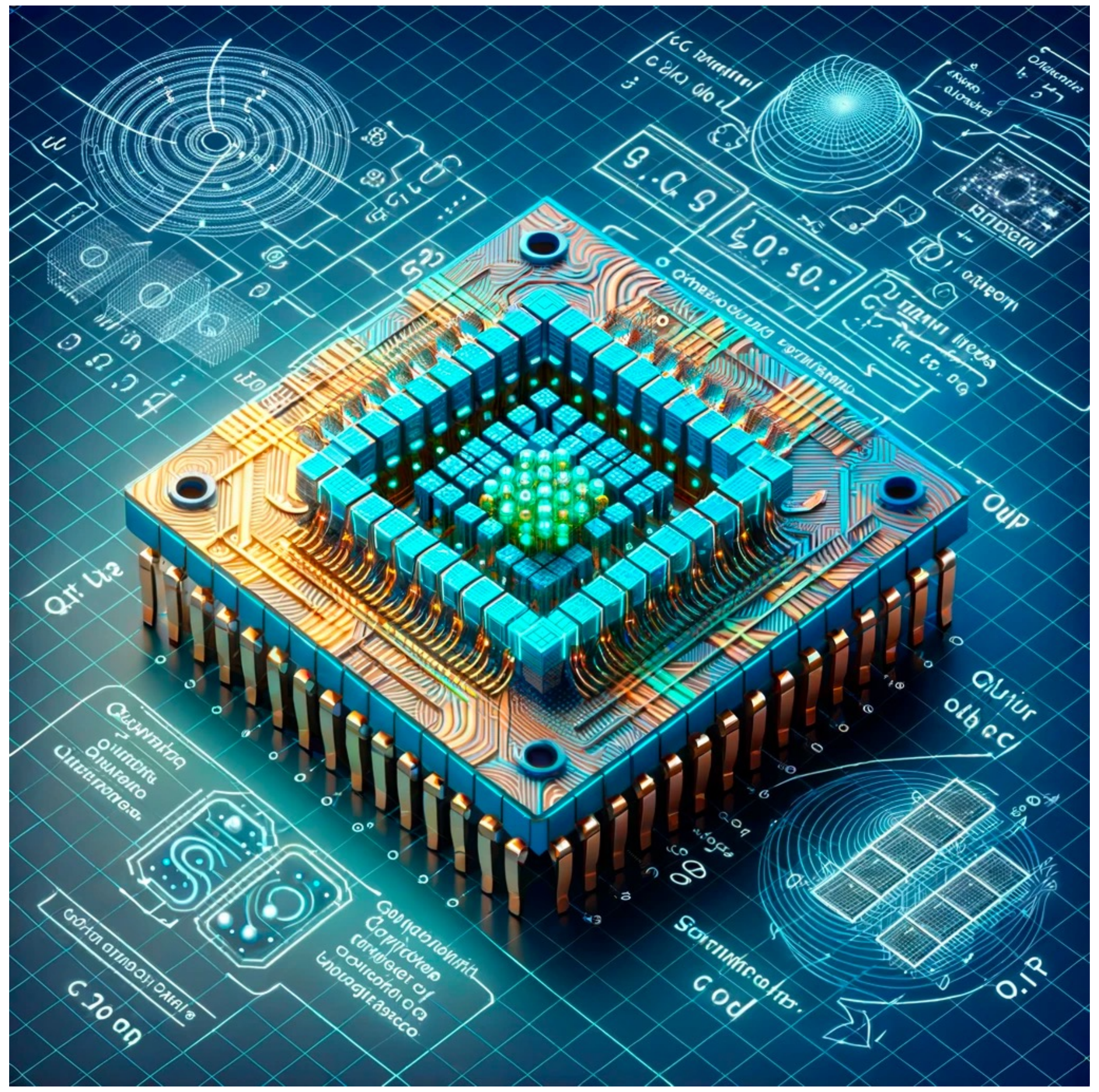
- Transformer accelerator leverages attention sparsity for better efficiency
- Achieve 0.6ms latency, 1.6uJ energy for one round of correction



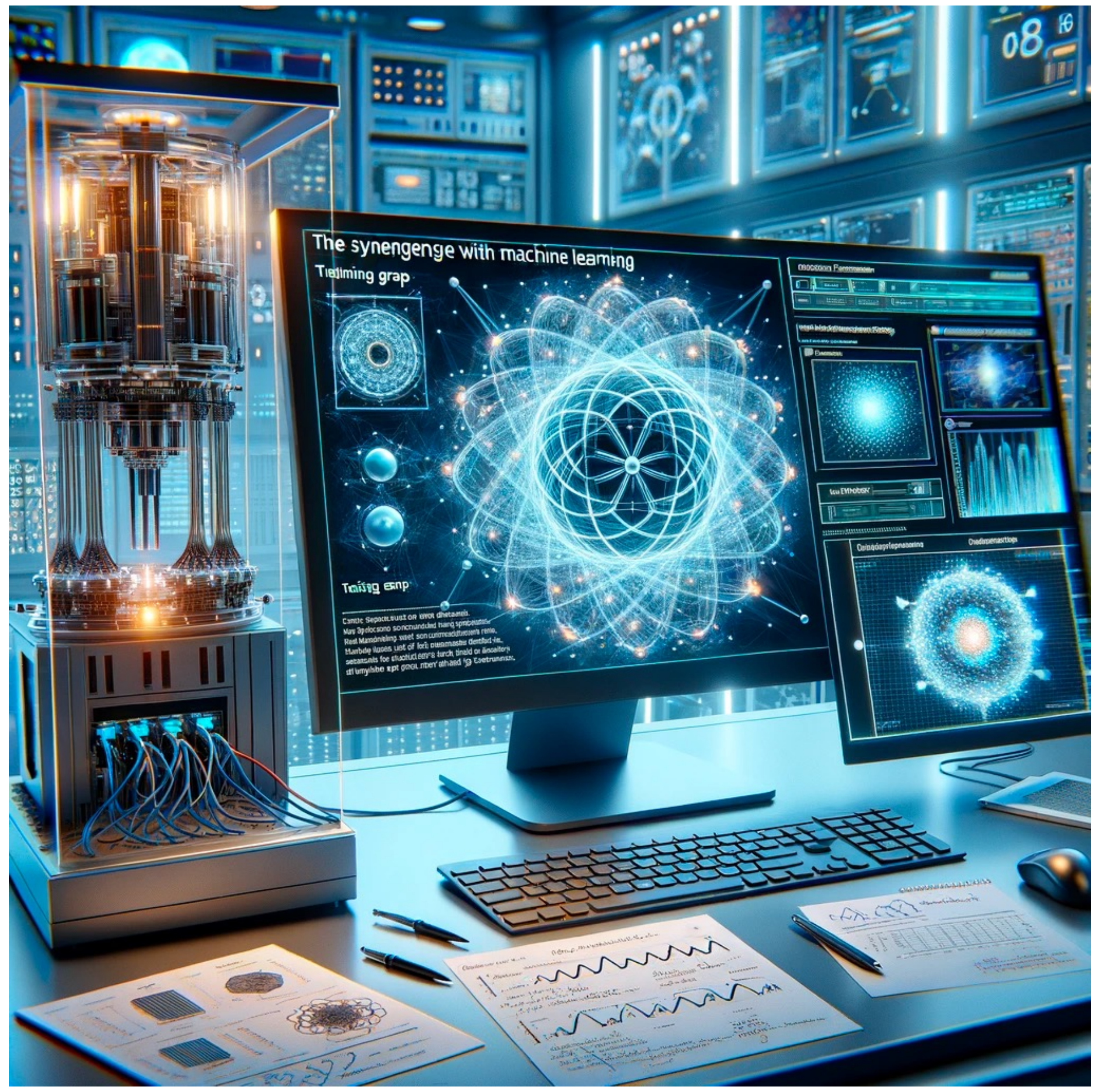

Keithley 2602A Sourcemeter
Tektronix TDS3064B Oscilloscope
Opal Kelly XEM7350 FPGA board
Keithley 2000 Multimeter
PC
Chip board

H. Wang, et al. "SpAtten: Efficient Sparse Attention Architecture with Cascade Token and Head Pruning." HPCA 2021

Future Research



Compilation stack and hardware accelerator for fault tolerant quantum computing



Efficient machine learning algorithms and systems for quantum information science

Thank You!

Hanrui Wang is on academic job market this year, please reach out for any opportunities.