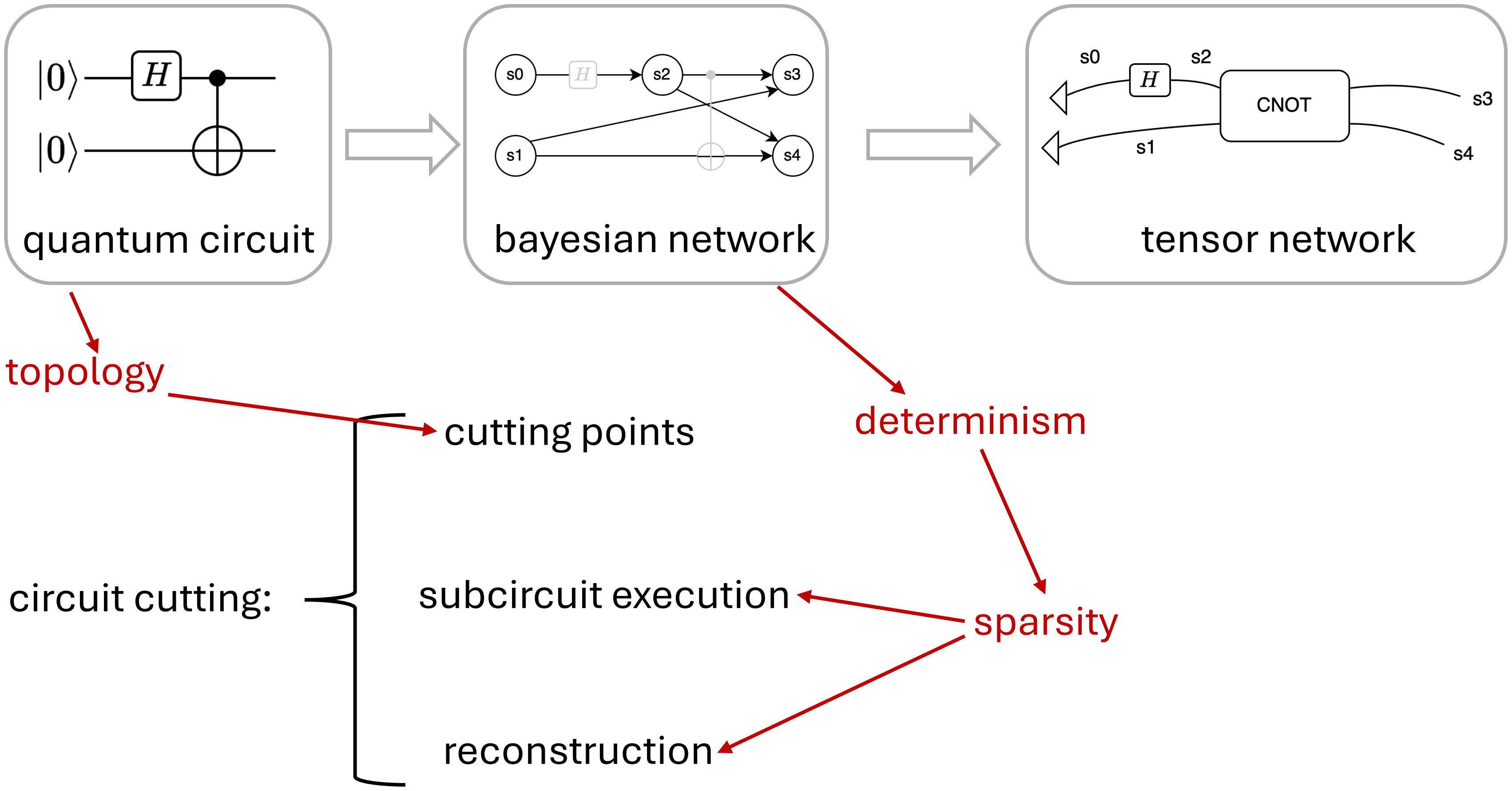
What is the Cost of Quantum Circuit Cutting?

Impact of topology, determinism, and sparsity

Zirui Li

09/25/2024

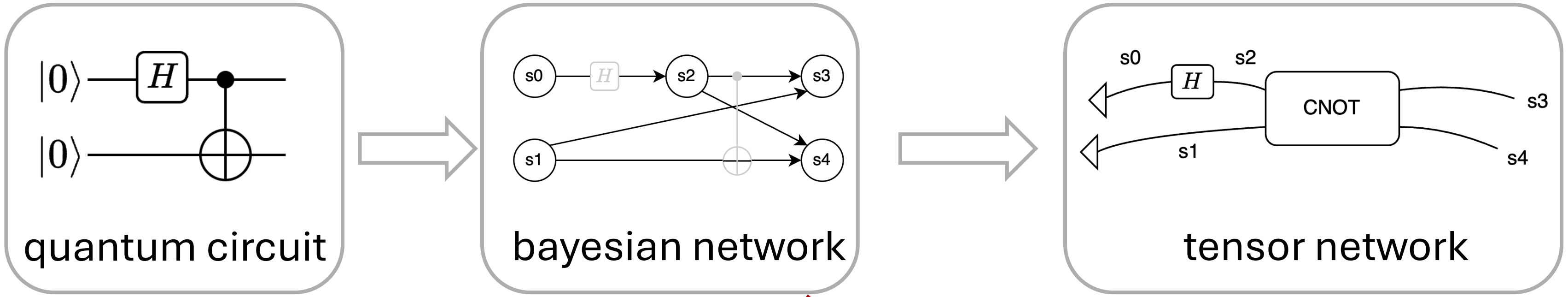


|0⟩
|0⟩
H
quantum circuit
s0
H
s2
s3
s1
s4
bayesian network
s0
H
s2
CNOT
s3
s1
s4
tensor network
topology
determinism
cutting points
circuit cutting:
subcircuit execution
sparsity
reconstruction

Outline

1. Quantum computing basic knowledge.

2. The flow:



topology

determinism

circuit cutting:

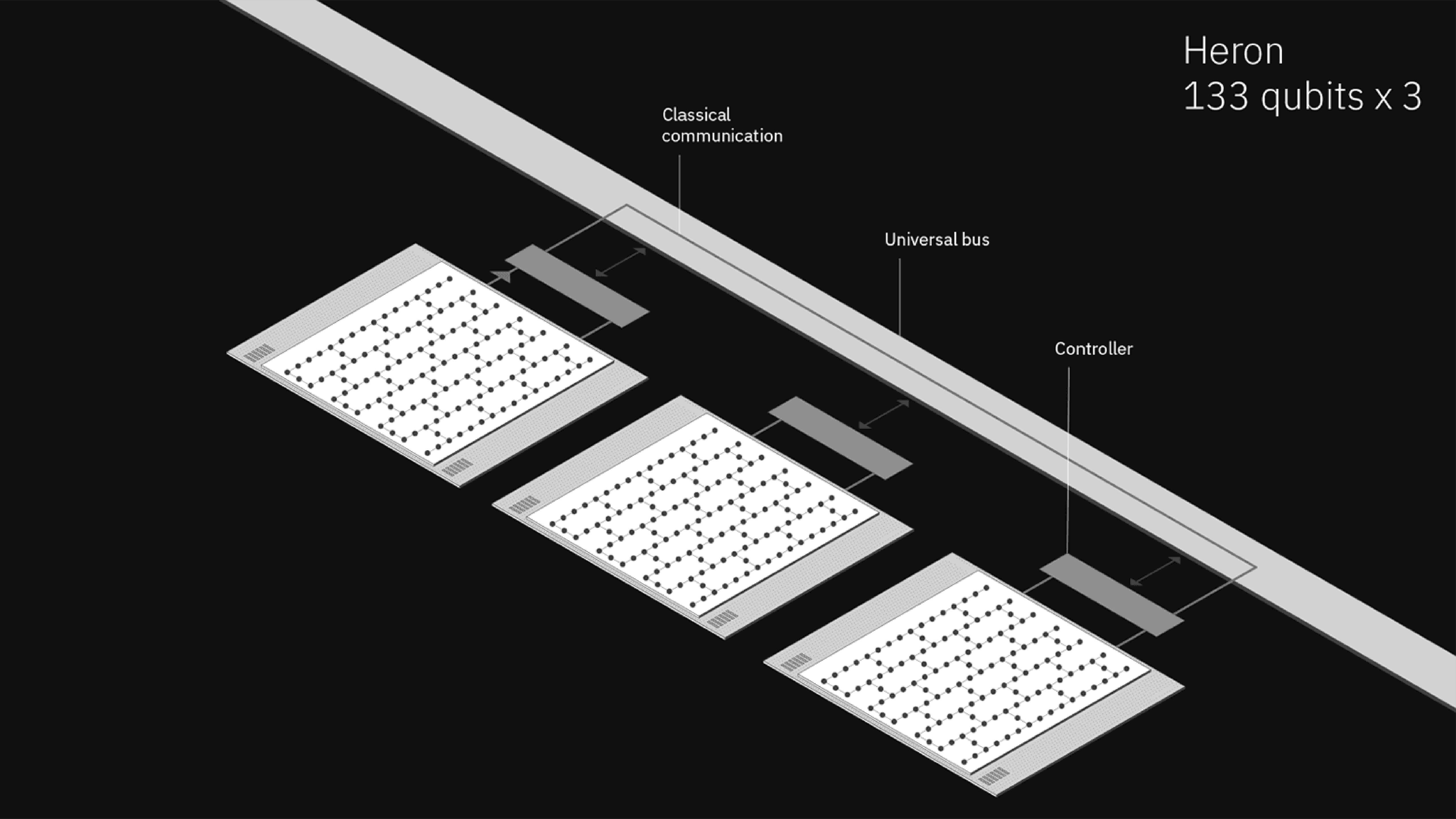
- cutting points
- subcircuit execution
- reconstruction

sparsity

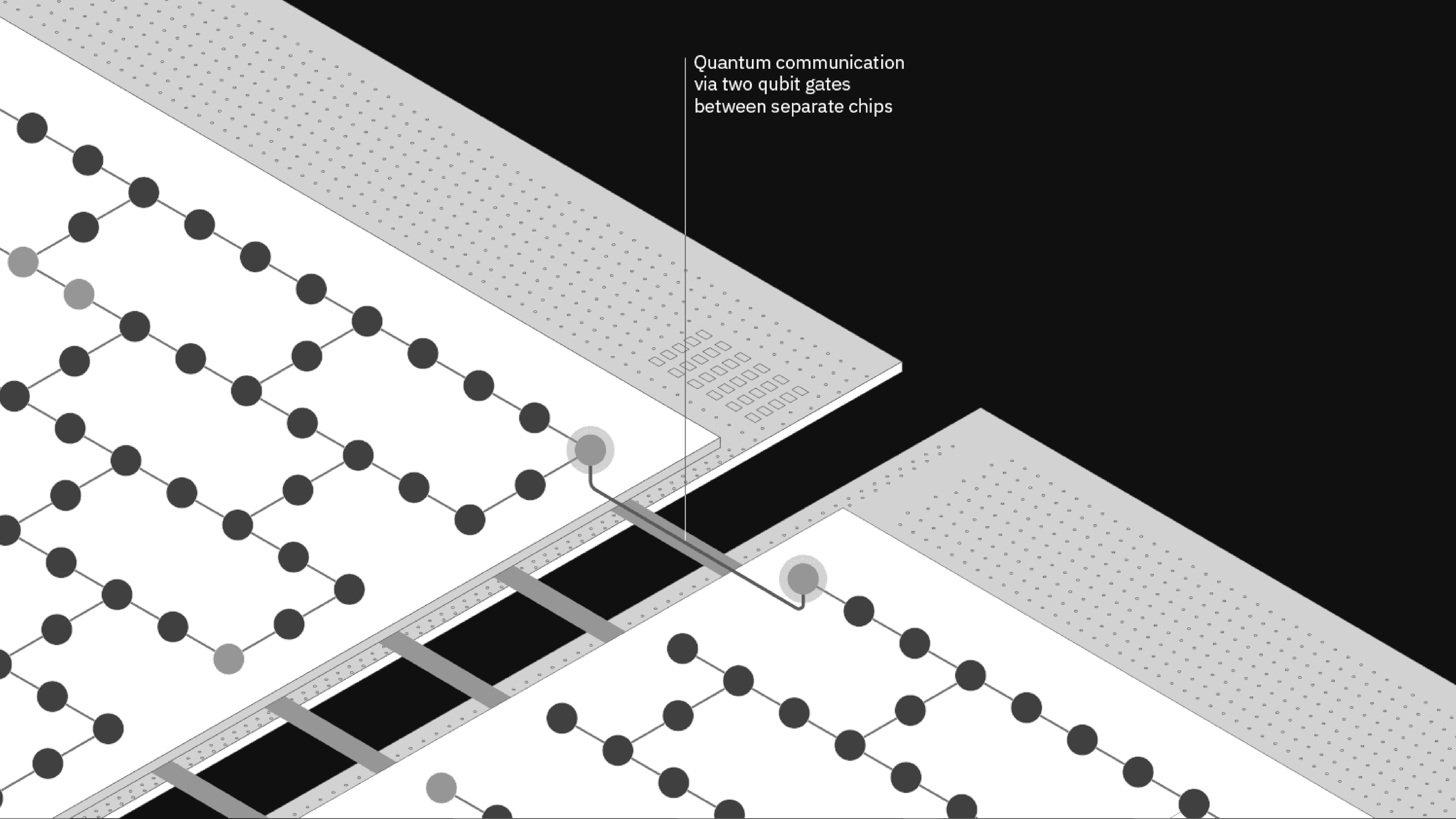
3. Optimized circuit cutting: sparse tensor preparation and contraction.

Three Ways to Scale Quantum Computers

- Circuit cutting/knitting
- Multi-chip processor
- Quantum communication to link multi-chip processors

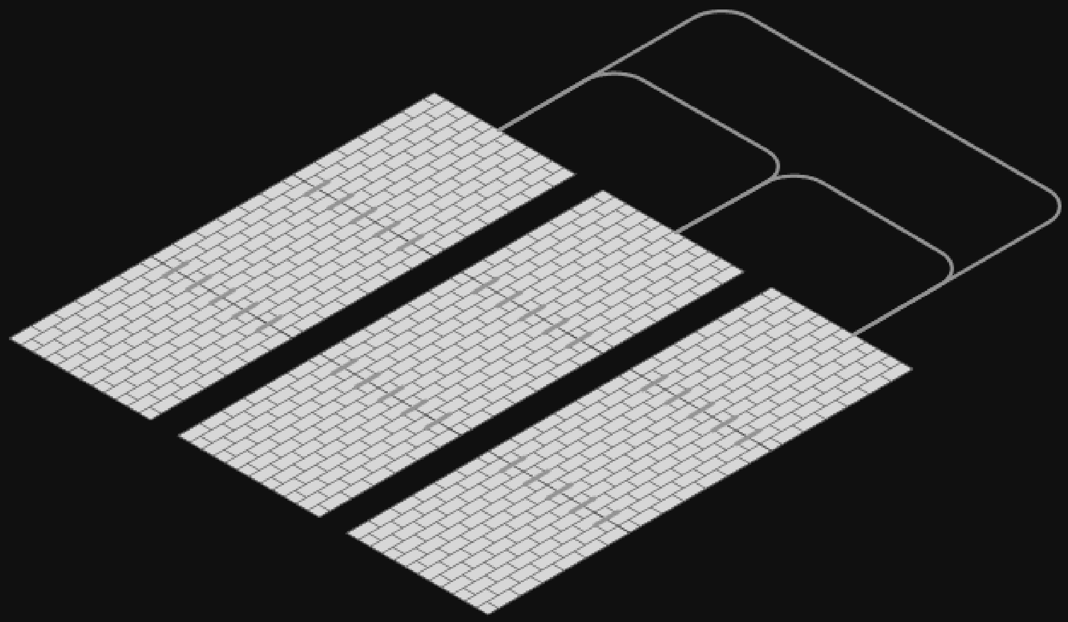


Heron
133 qubits x 3
Classical communication
Universal bus
Controller


Quantum communication
via two qubit gates
between separate chips

2025
Quantum parallelization of
multi-chip quantum processors

Kookaburra
4,158+ qubits



Three Ways to Scale Quantum Computers

- Circuit cutting/knitting
- Multi-chip processors
- Quantum communication to link multi-chip processors

Intro to Quantum Computing

- quantum state representation: $\begin{cases} \text{state vector} \\ \\ \text{density matrix} \end{cases}$

Intro to Quantum Computing

- State vector can represent a quantum state.
- State vector is in a complex inner product space.
- Bra-ket notation for state vector:
  - $|\psi\rangle$ (called ket psi) is a coloum vector.
  - $\langle\phi|$ (called bra phi) is a row vector.
  - $\langle\psi|$ is the conjugate transpose of $|\psi\rangle$
  - $\langle\phi|\psi\rangle$ is the complex inner product of the two vectors.

Intro to Quantum Computing

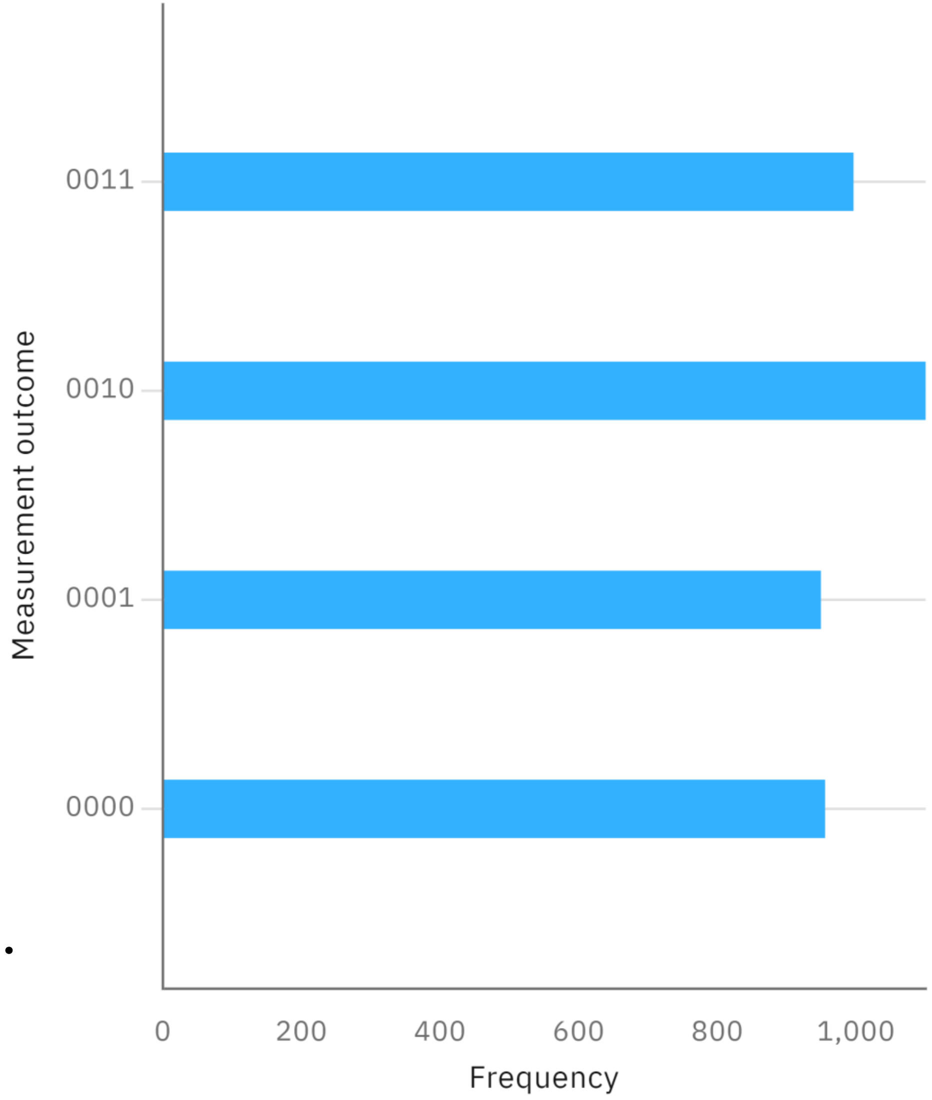
- Quantum State:
  - 1-qubit state: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$
  - $|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ forms an orthonormal bases in $\mathbb{C}^2$.
  - Other choices of bases:
    - $|+\rangle = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, |-\rangle = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$
    - $|i\rangle = \begin{bmatrix} 1/\sqrt{2} \\ i/\sqrt{2} \end{bmatrix}, |-i\rangle = \begin{bmatrix} 1/\sqrt{2} \\ -i/\sqrt{2} \end{bmatrix}, \langle i|-i\rangle = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-i}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ -i/\sqrt{2} \end{bmatrix} = 0$

Intro to Quantum Computing

- Quantum State:
  - 1-qubit state: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$
  - $|\alpha|^2 + |\beta|^2 = 1$
  - When you observe $|\psi\rangle$ in $|0\rangle, |1\rangle$ bases, $|\psi\rangle$ will collapse to any of the two basis.
  - You will observe $|0\rangle$ with probability $|\alpha|^2$, observe $|1\rangle$ with probability $|\beta|^2$.
  - In real quantum computers, multiple shots will be performed to have a guess of $|\alpha|^2$ and $|\beta|^2$.

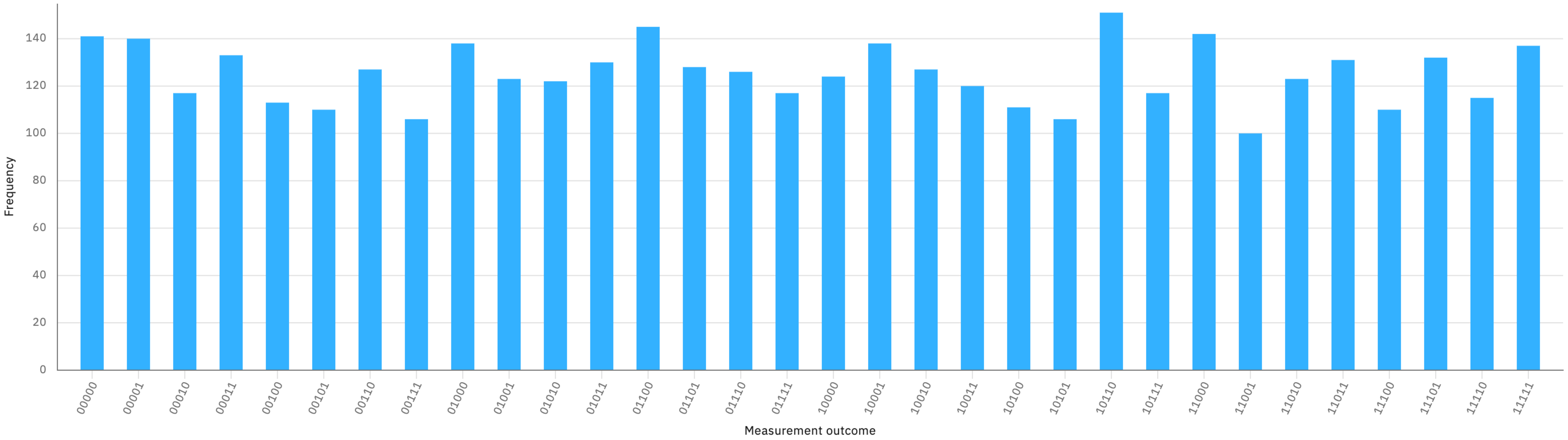
Intro to Quantum Computing

- Quantum State:
  - $|00\rangle$ is the tensor product of $|0\rangle$ and $|0\rangle$
  - $|00\rangle = \begin{bmatrix}1\\0\end{bmatrix} \otimes \begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}$
  - 2-qubit state: $|\psi\rangle = \alpha_0|00\rangle + \alpha_1|01\rangle + \alpha_2|10\rangle + \alpha_3|11\rangle$
  - $|\alpha_0|^2 + |\alpha_1|^2 + |\alpha_2|^2 + |\alpha_3|^2 = 1$
  - For example, $\frac{1}{2}|00\rangle + \frac{1}{2}|01\rangle + \frac{1}{2}|10\rangle + \frac{1}{2}|11\rangle$ taking 4000 shots, see the right Fig.



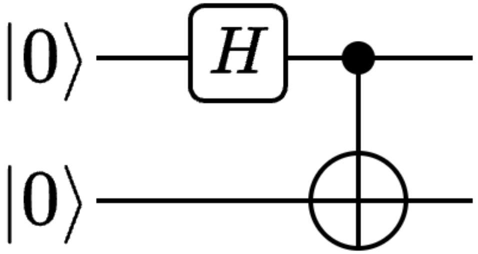
Intro to Quantum Computing

- Quantum State:
  - N-qubit state: $|\psi\rangle = \alpha_0|00\ldots00\rangle + \alpha_1|00\ldots01\rangle + \cdots \alpha_{2^n-1}|11\ldots11\rangle$
  - $\sum_{i=0}^{2^n-1} \alpha_i = 1$
  - For example, 5-qubit state after 4000 shots.



Intro to Quantum Computing

- Unitary Operation:
  - U is a unitary matrix. $|\psi^*\rangle \leftarrow U|\psi\rangle$
- For example, to create a bell state $\frac{1}{\sqrt{2}}|00\rangle + \frac{1}{\sqrt{2}}|11\rangle$:
  - 1. initial quantum state: $|00\rangle = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$
  - 2. apply Hadamard gate to qubit zero:
    - Hadamard gate: $H = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
    - the state after Hadamard gate: $\mathrm{I} \otimes H|00\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = |0\rangle \otimes \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$
  - 2. apply CNOT gate from qubit zero to qubit one:
    - CNOT gate: $CNOT = \begin{bmatrix} 1 & & & \\ & & & 1 \\ & & 1 & \\ & 1 & & \end{bmatrix}$
    - the state after CNOT gate: $CNOT(|0\rangle \otimes \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)) = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}}|00\rangle + \frac{1}{\sqrt{2}}|11\rangle$



$CNOT(I \otimes H)|00\rangle$

Intro to Quantum Computing

- State vector simulator:
  - exponential time and memory cost *w.r.t.* #qubits.
  - 128GB memory needed to simulate 34-qubit, then twice the memory needed for each extra qubit.

💰 Pricing

Check out our pricing page for the full info.

- **34 qubit CPU** (state-vector): FREE
- **36 qubit GPU** (state-vector): $3/hour/gpu
  - We will be using 1 GPU for up to 32 qubits, then twice as many for each extra qubit:
  - 2 GPUs for 33 qubits
  - 4 GPUs for 34 qubits
  - 8 GPUs for 35 qubits
  - 16 GPUs for 36 qubits
  - Minimum charge is $0.20 and we charge in 1-second increments
- **Quantum** (IQM Garnet): $0.3 + $0.00145 / shot. So 1000 shots (default) will be $1.75.

Credit: BlueQubit
https://app.bluequbit.io/docs#a_b

Intro to Quantum Computing

- quantum state representation: $\begin{cases} \text{state vector} \\ \\ \text{density matrix} \end{cases}$

Intro to Quantum Computing

- The expectation value of a state $|\psi\rangle$ on an observable $\hat{O}$.
$$\langle\psi|\hat{O}|\psi\rangle$$
- In quantum chemistry: minimizing $\langle\psi|\hat{O}|\psi\rangle$ $\Rightarrow$calculating the ground state energy of a molecule.

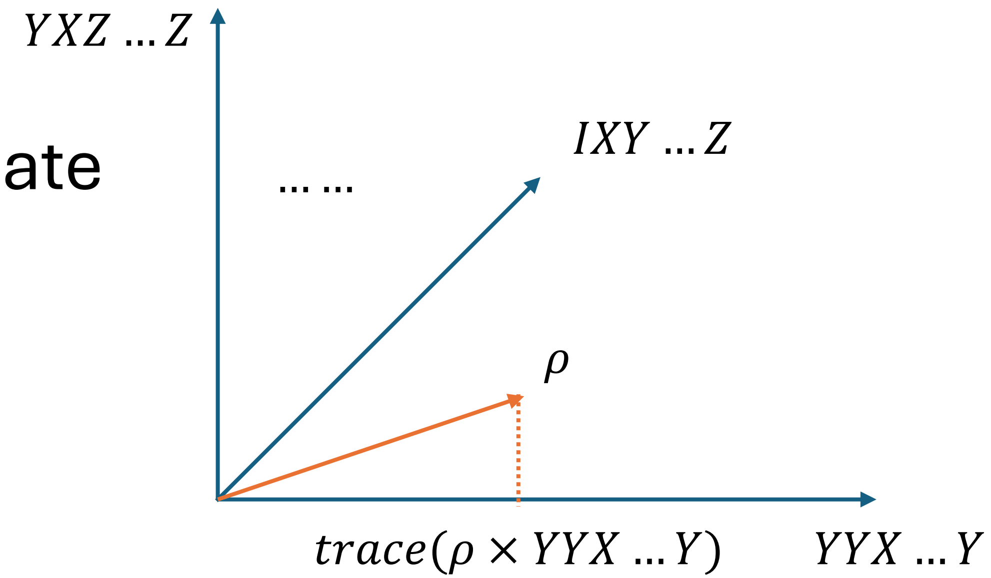
- Density matrix: $\rho = |\psi\rangle\langle\psi|$, then $\langle\psi|\hat{O}|\psi\rangle = trace(\rho\hat{O})$.
- $|\psi\rangle \in \mathbb{C}^{2^n}$; $\rho, \hat{O} \in \mathbb{C}^{2^n \times 2^n}$; $\rho, \hat{O}$ are Hermitian matrices.
- $trace(\rho\hat{O})$ is taking the inner product of the two matrices.

Intro to Quantum Computing

- Pauli matrices:

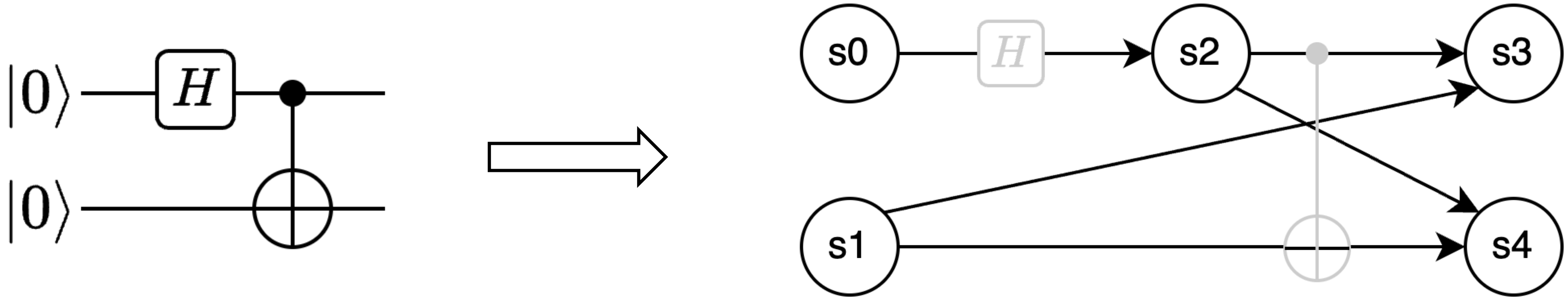
$$I = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}, X = \begin{bmatrix} & 1 \\ 1 & \end{bmatrix}, Y = \begin{bmatrix} & -i \\ i & \end{bmatrix}, Z = \begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$$

- I, X, Y, Z forms a bases in $\mathbb{C}^{2\times 2}$.
- The tensor product of them forms a bases in $\mathbb{C}^{2^n \times 2^n}$.
- E.g. $YYXY := Y \otimes Y \otimes X \otimes Y$
- 4^n Pauli strings/bases for n-qubit state
- Quantum state tomography:
  - Measure the expval on each basis.



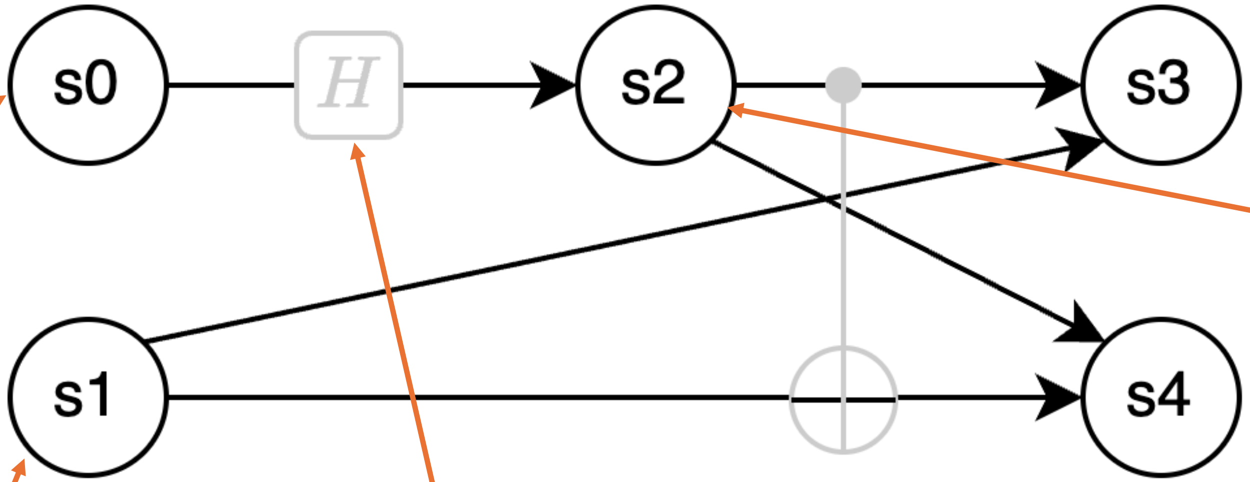
Quantum Circuit to Bayesian Network

- The Bell state: $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle) = CNOT(I \otimes H)|00\rangle$ in density matrix representation calculate by hand :
  - The initial state $\rho_0 = |00\rangle\langle 00| = \frac{1}{4}II + \frac{1}{4}IZ + \frac{1}{4}ZI + \frac{1}{4}ZZ.$
  - After Hadamard $\rho_1 = H\rho_0 H^\dagger = \frac{1}{4}II + \frac{1}{4}IX + \frac{1}{4}ZI + \frac{1}{4}ZX.$
  - After CNOT $\rho_2 = CNOT\rho_1 CNOT^\dagger = \frac{1}{4}II + \frac{1}{4}XX - \frac{1}{4}YY + \frac{1}{4}ZZ.$



s0	w
I	0.5
X	0
Y	0
Z	0.5

Tensor 1



s1	w
I	0.5
X	0
Y	0
Z	0.5

Tensor 2

s0	s2	w
I	I	1
X	Z	1
Z	X	1
Y	Y	-1
otherwise		0

Tensor 3

s2	w
I	0.5
X	0.5
Y	0
Z	0

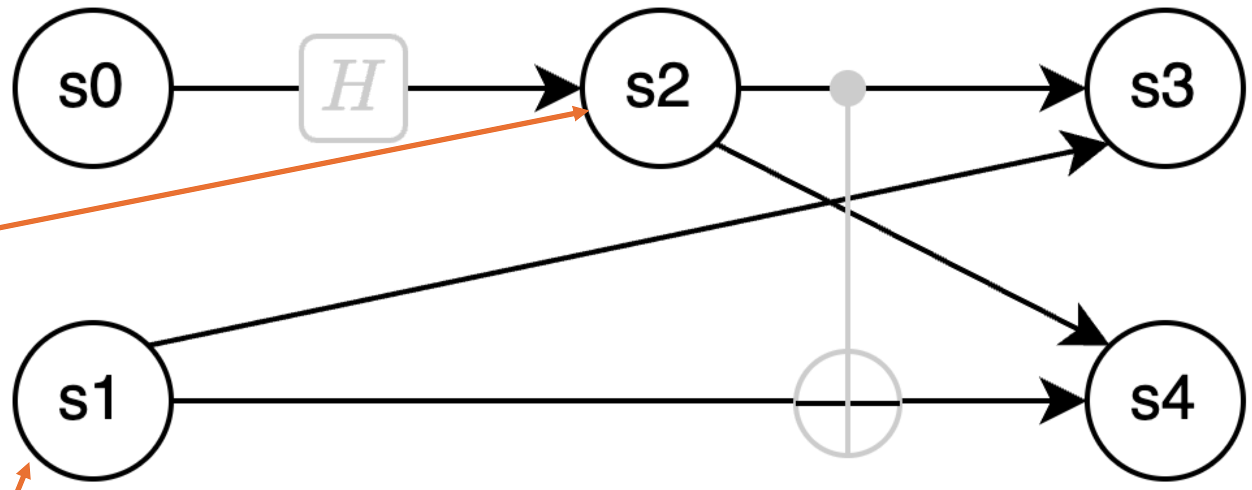
Tensor 4

Contract tensor 1 and tensor 3, get tensor 4.

This tensor's size is 16 but only 4 entries have non-zero weights.

s2	w
I	0.5
X	0.5
Y	0
Z	0

Tensor 4



s1	w
I	0.5
X	0
Y	0
Z	0.5

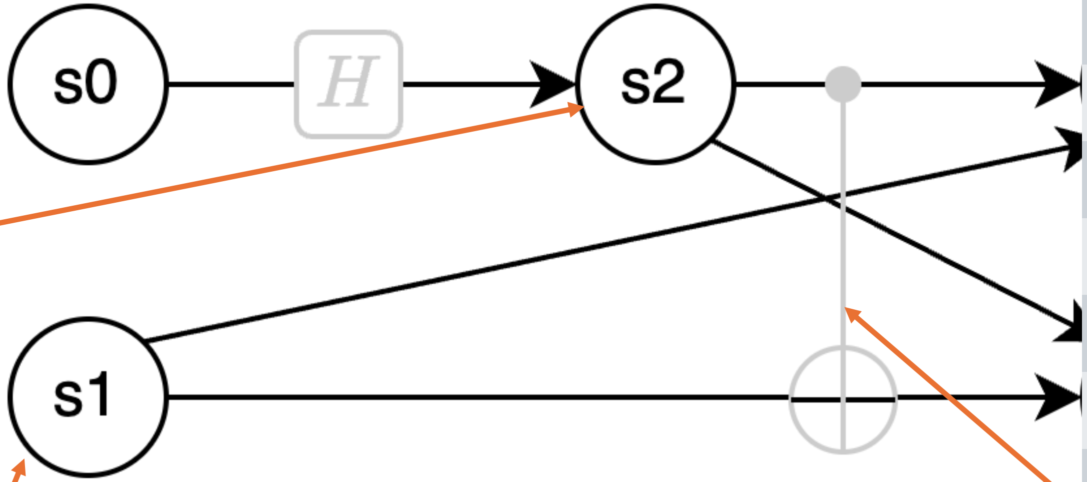
Tensor 2

s2	w
I	0.5
X	0.5
Y	0
Z	0

Tensor 4

s1	w
I	0.5
X	0
Y	0
Z	0.5

Tensor 2

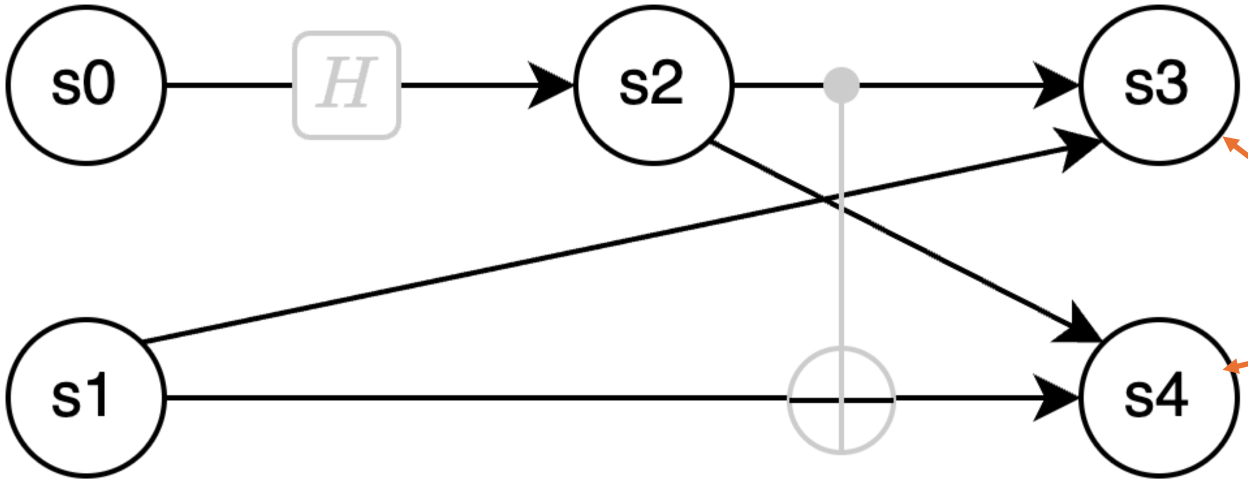


Next page: contract tensor 4, 2 and 5, get tensor 6.

s2	s1	s3	s4	w
I	I	I	I	1
I	X	I	X	1
I	Y	Z	Y	1
I	Z	Z	Z	1
X	I	X	X	1
X	X	X	I	1
X	Y	Y	Z	1
X	Z	Y	Y	-1
Y	I	Y	X	1
Y	X	Y	I	1
Y	Y	X	Z	-1
Y	Z	X	Y	1
Z	I	Z	I	1
Z	X	Z	X	1
Z	Y	I	Y	1
Z	Z	I	Z	1
otherwise				0

Tensor 5

This tensor's size is 256 but only 16 entries have non-zero weights.



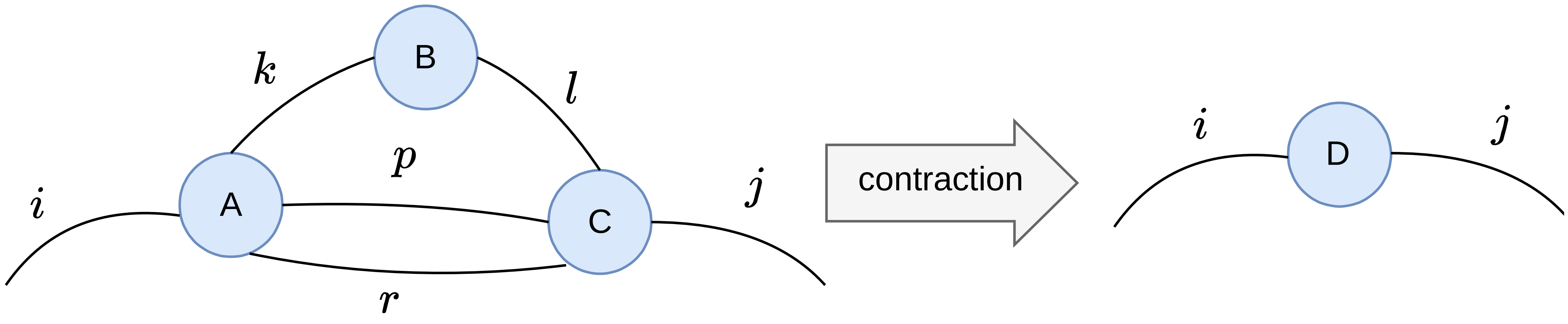
s3	s4	w
I	I	0.25
X	X	0.25
Y	Y	-0.25
Z	Z	0.25
otherwise		0

Tensor 6

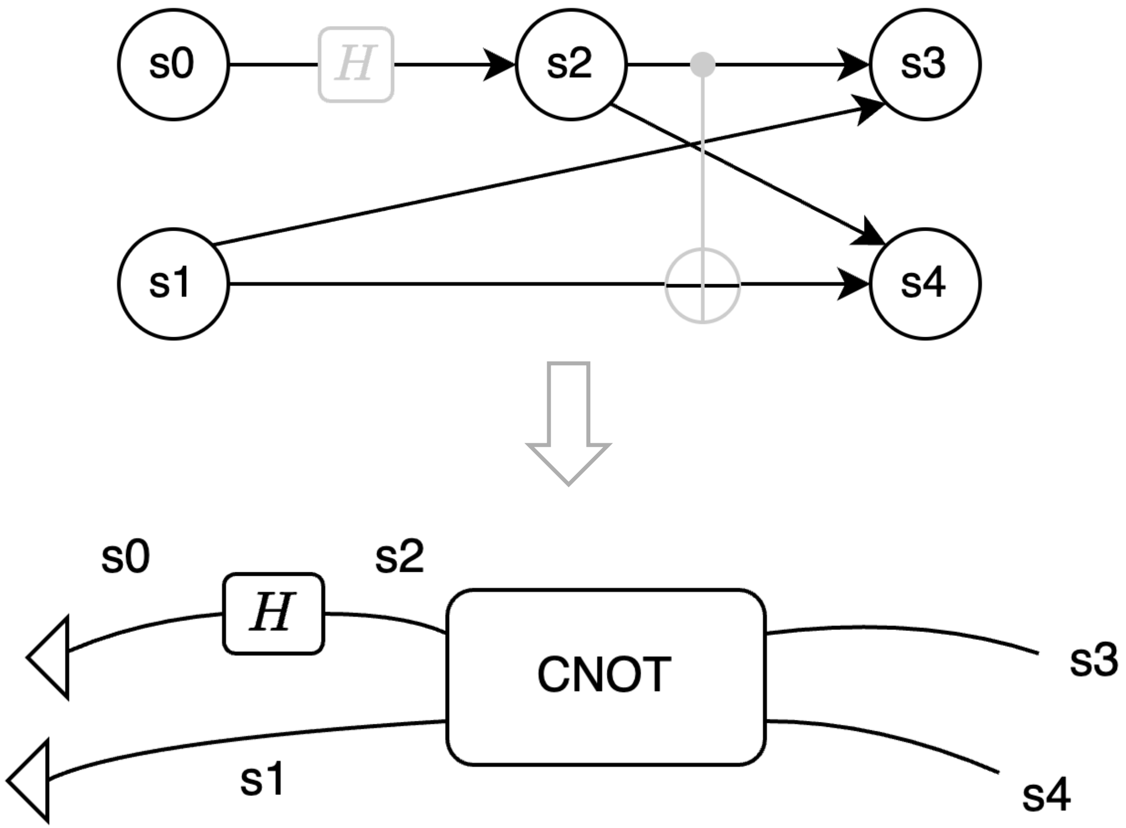
Tensor 6 matches the hand-calculated $\frac{1}{4}II + \frac{1}{4}XX - \frac{1}{4}YY + \frac{1}{4}ZZ$.

Tensor Contraction

- $D_{i,j} = \sum_{k,l,p,r} A_{i,k,p,r} B_{k,l} C_{l,p,r,j}$



Bayesian Network to Tensor Network



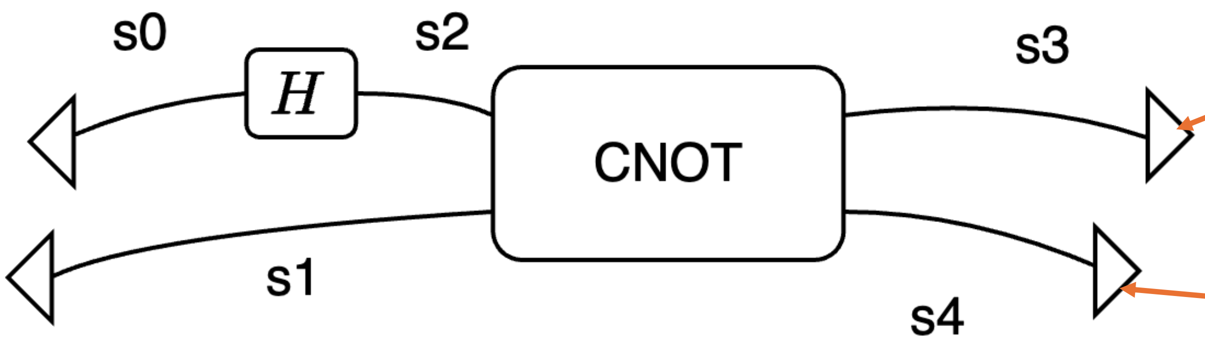
After contraction:

s3	s4	w
I	I	0.25
X	X	0.25
Y	Y	-0.25
Z	Z	0.25
otherwise		0

Tensor 6

Expectation Value on an Observable

- If we want to know the expectation value on observable YY.
- $\hat{O} = YY$, the bell state is $\rho = \frac{1}{4}II + \frac{1}{4}XX - \frac{1}{4}YY + \frac{1}{4}ZZ$.
- The expectation value is $trace\left(\rho\hat{O}\right) = -\frac{1}{4}trace(II) = -1$.



s3	w
I	0
X	0
Y	1
Z	0

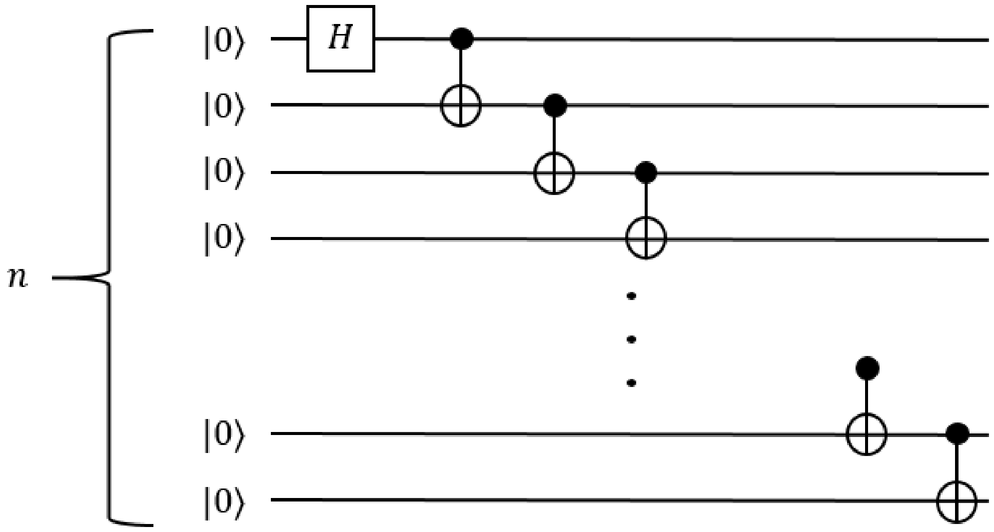
Tensor 7

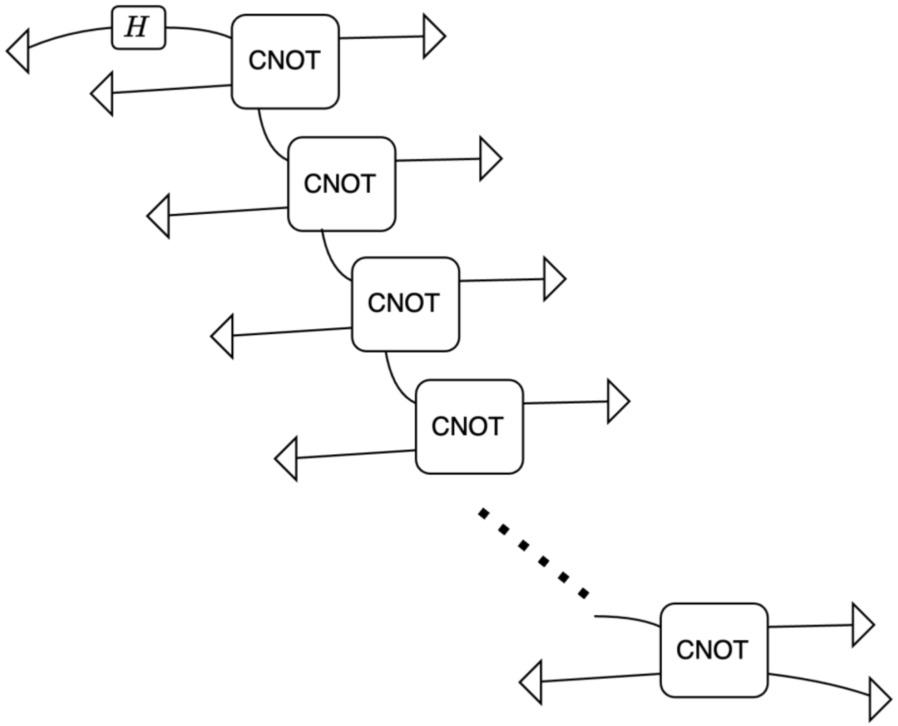
s4	w
I	0
X	0
Y	1
Z	0

Tensor 8

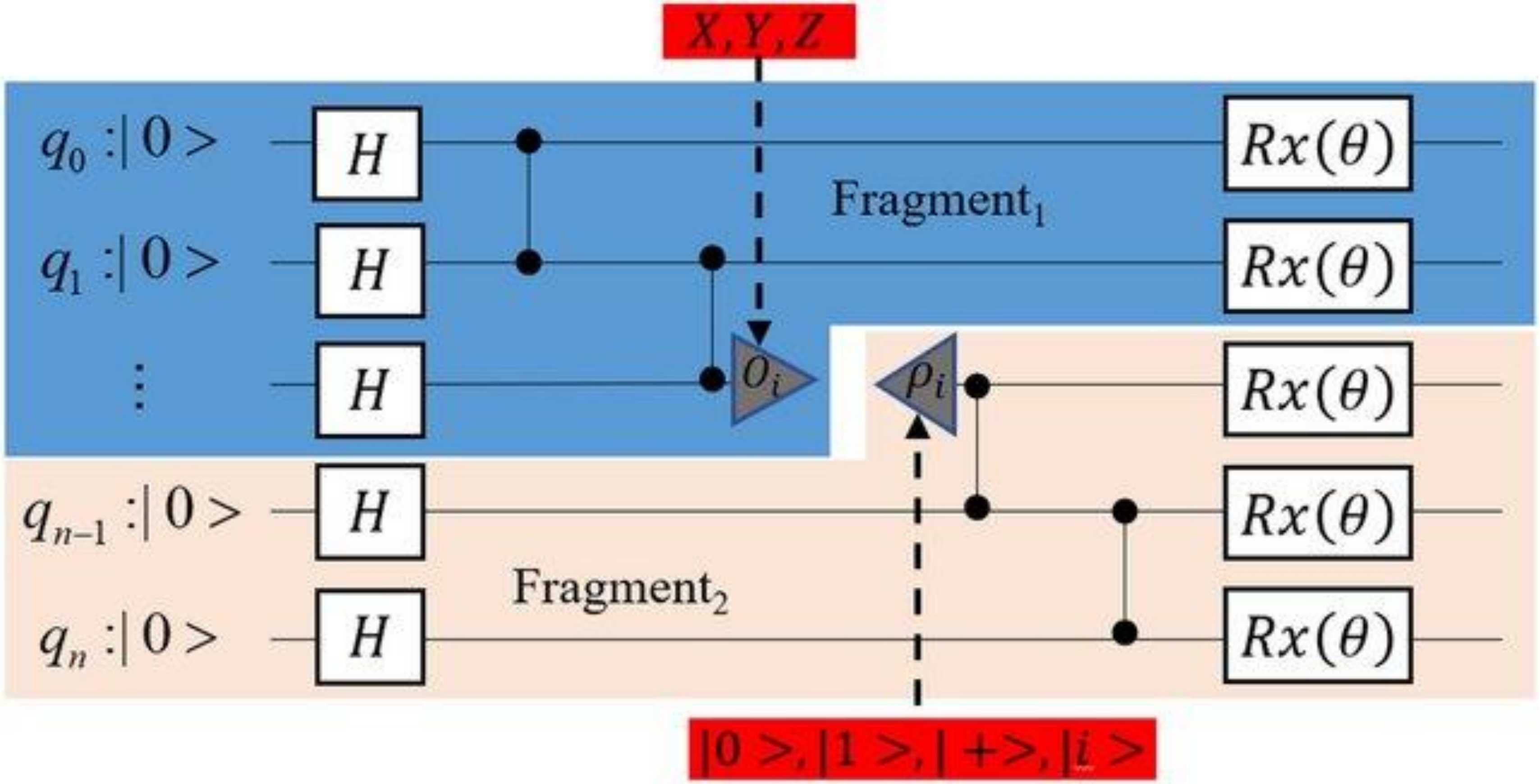
#Qubits vs Treewidth

- Suppose we want to know the expval of GHZ state on an observable.





Circuit Cutting



Credit: Lian, Hang & Xu, Jinchen & Zhu, Yu & Fan, Zhiqiang & Liu, Yi & Shan, Zheng. (2023).
Fast reconstruction algorithm based on HMC sampling. Scientific Reports. 13. 10.1038/s41598-023-45133-z.

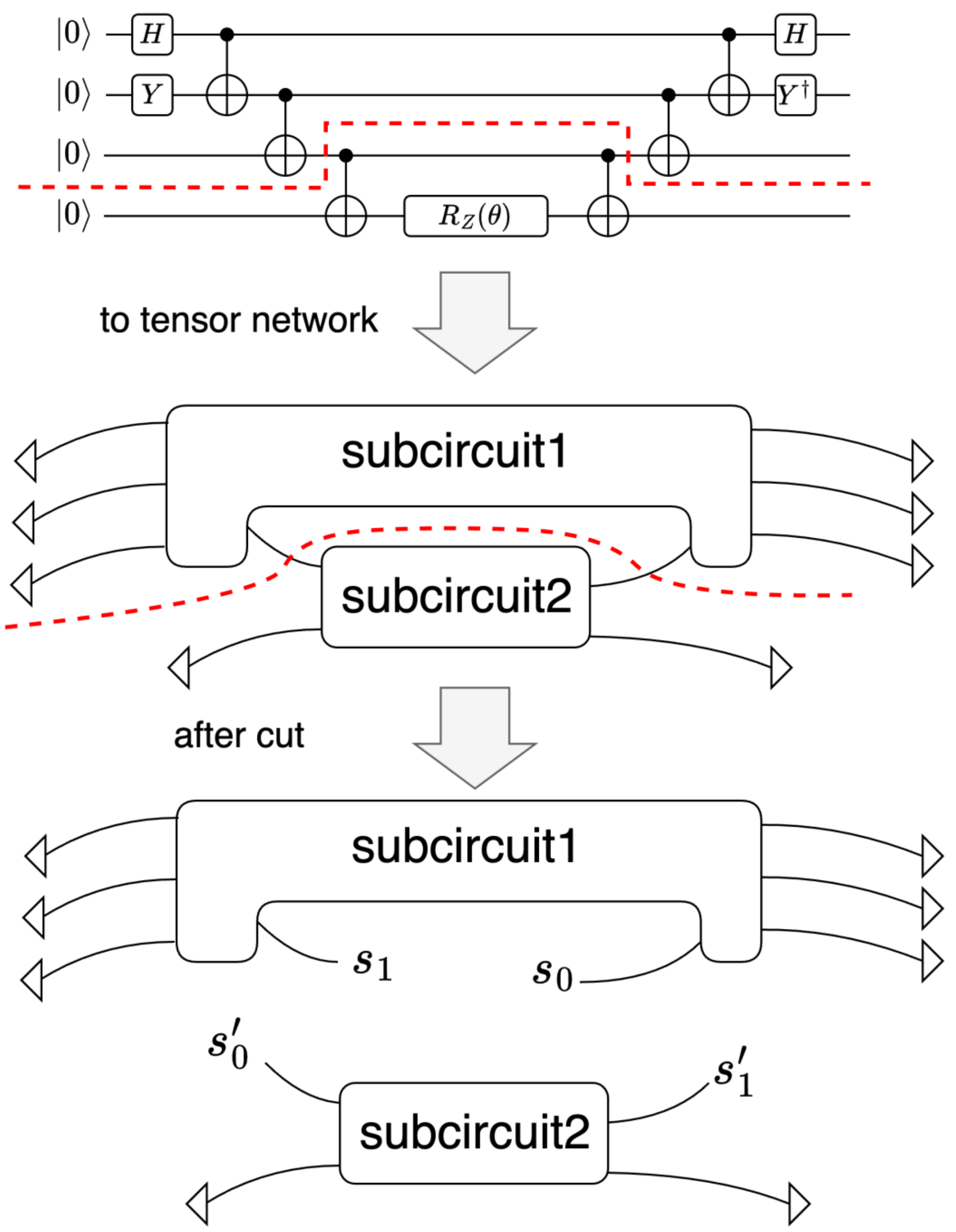
Example

- Subcircuit 1 has 2 open edges;
- Subcircuit 2 has 2 open edges;
- Run 16 different settings of

each subcircuit to fill in the two tensors.

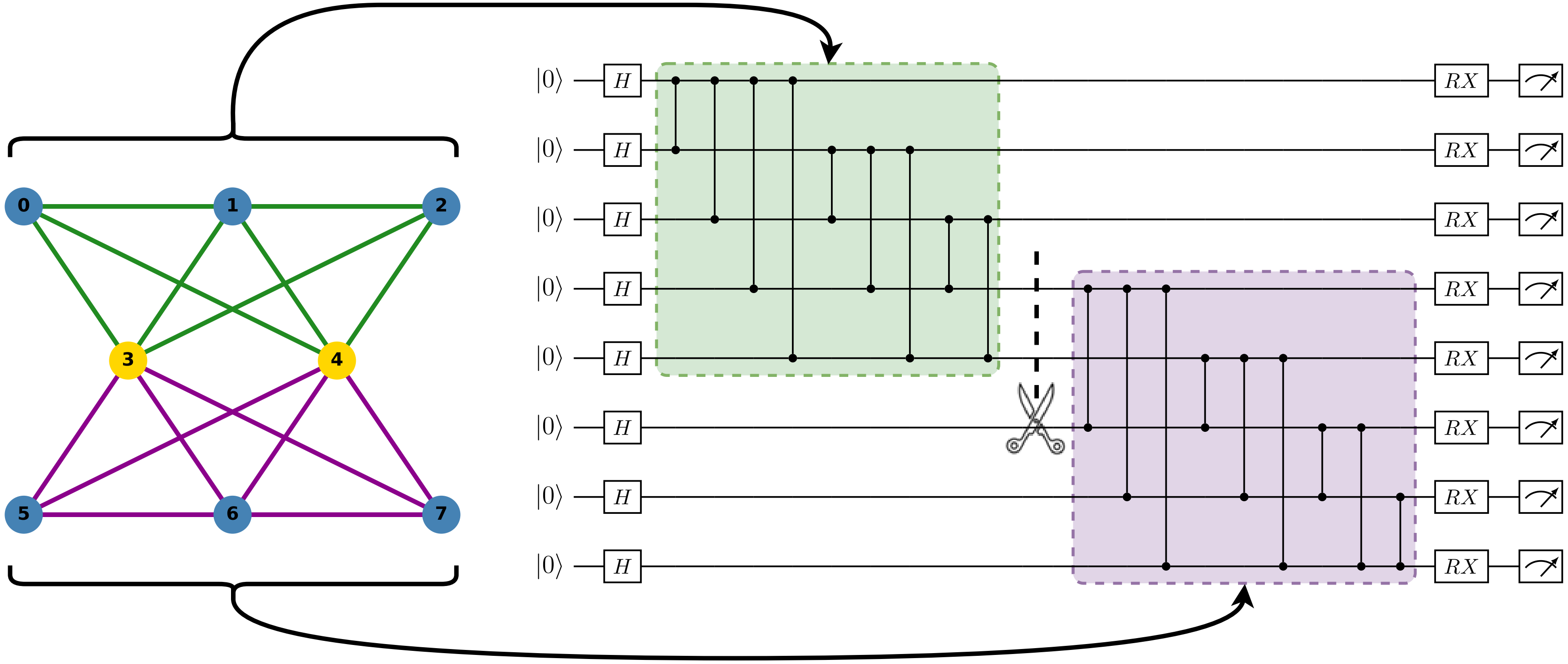
s0	s1	w
I	I	?
I	X	?
I	Y	?
I	Z	?
…	…	…

s'0	s'1	w
I	I	?
I	X	?
I	Y	?
I	Z	?
…	…	…



Impact of Topology

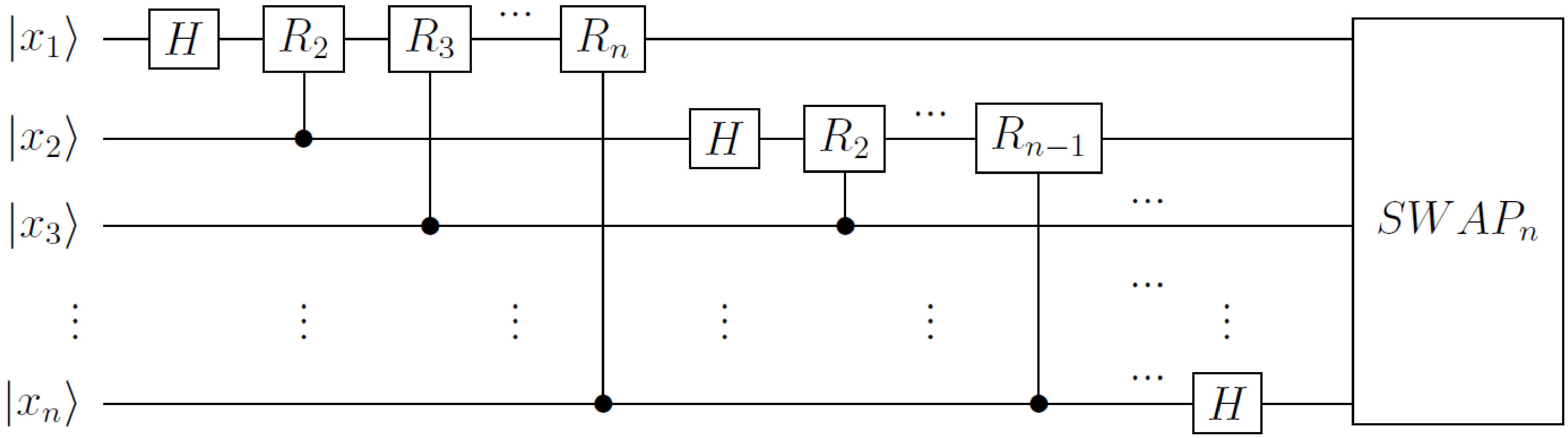
- We want each tensor to have as less open edges as possible, and meanwhile reduce the maximum #qubits.



Credit: https://pennylane.ai/qml/demos/tutorial_quantum_circuit_cutting/

Impact of Topology

- We want each tensor to have as less open edges as possible, and meanwhile reduce the maximum #qubits.



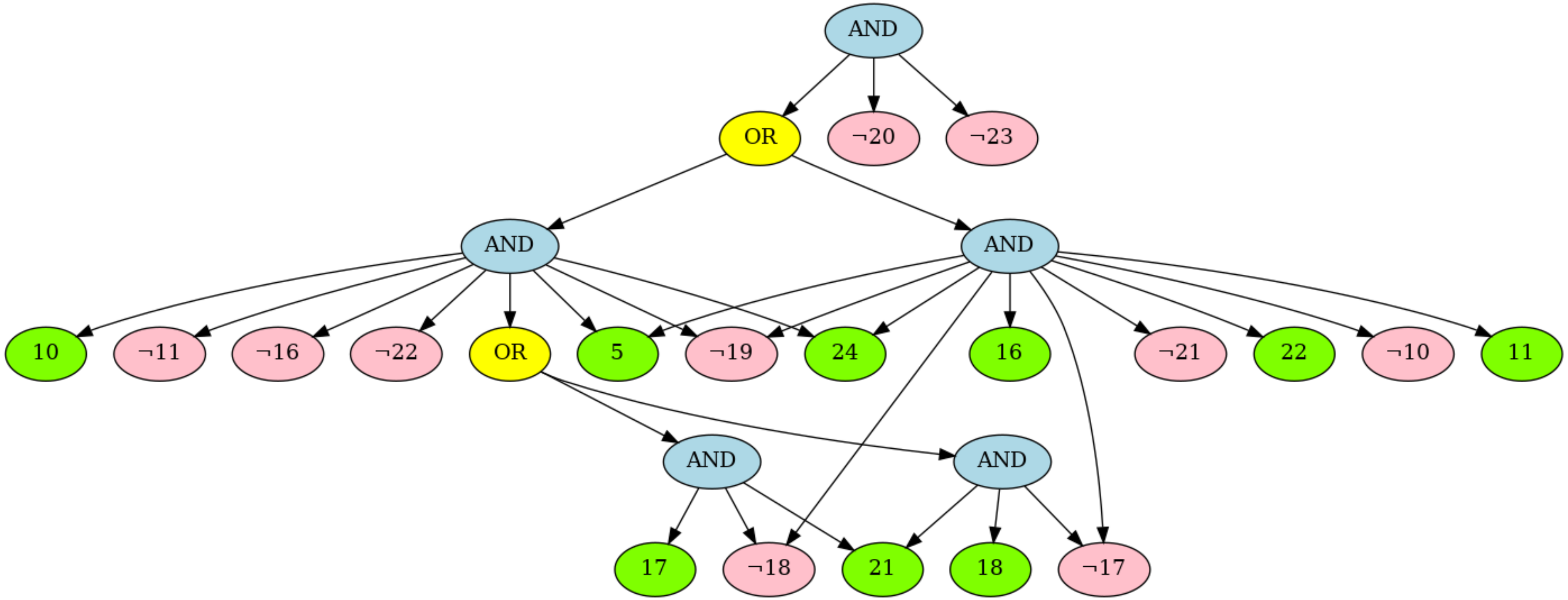
Credit: https://medium.com/colibritd-quantum/getting-to-know-quantum-fourier-transform-ae60b23e58f4

Impact of Determinism

- Clifford gates' Conditional Probability Distributions(CPD) is deterministic.
  - If an n-qubit unitary matrix is Clifford, the tensor size is $4^n \times 4^n$, and there are only 4^n non-zero weights.
  - Clifford gates stabilize Pauli strings. In other words, Clifford gate will only do a permutation of all Pauli strings.
- For a non-Clifford gate, like T gate, the Conditional Probability Distributions is not deterministic.

	I	X	Y	Z
I	1	0	0	0
X	0	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	0
Y	0	$\frac{-1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	0
Z	0	0	0	1

Knowledge Compilation

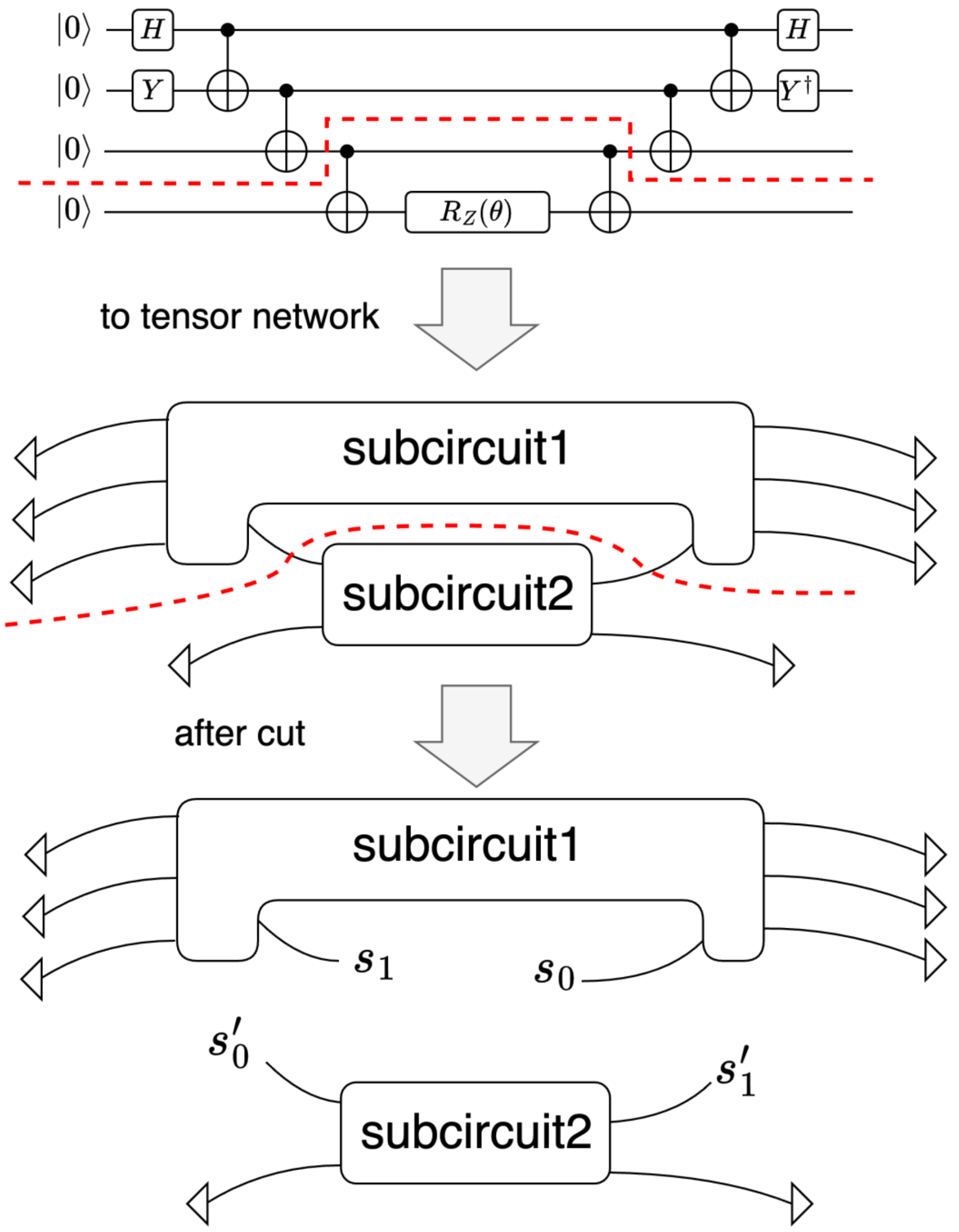


Knowledge Compilation

- We can know which entries have zero weight.
- Subcircuit 1 has 4 non-zero weights.
- Subcircuit 2 has 8 non-zero weights.

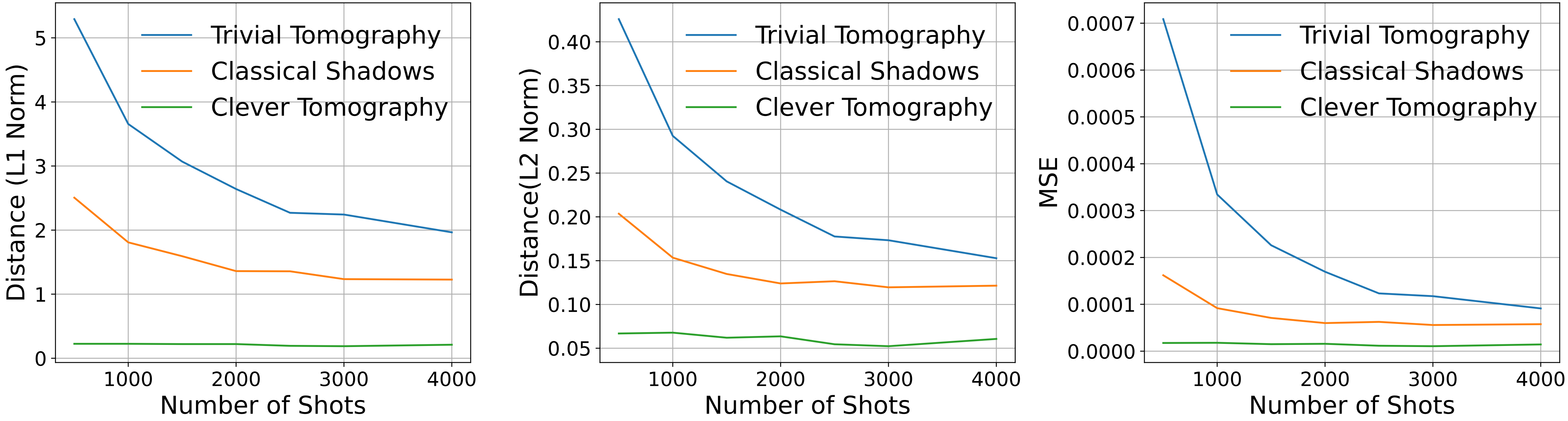
s0	s1	w
I	I	?
I	X	?
I	Y	?
I	Z	?
…	…	…

s'0	s'1	w
I	I	?
I	X	?
I	Y	?
I	Z	?
…	…	…

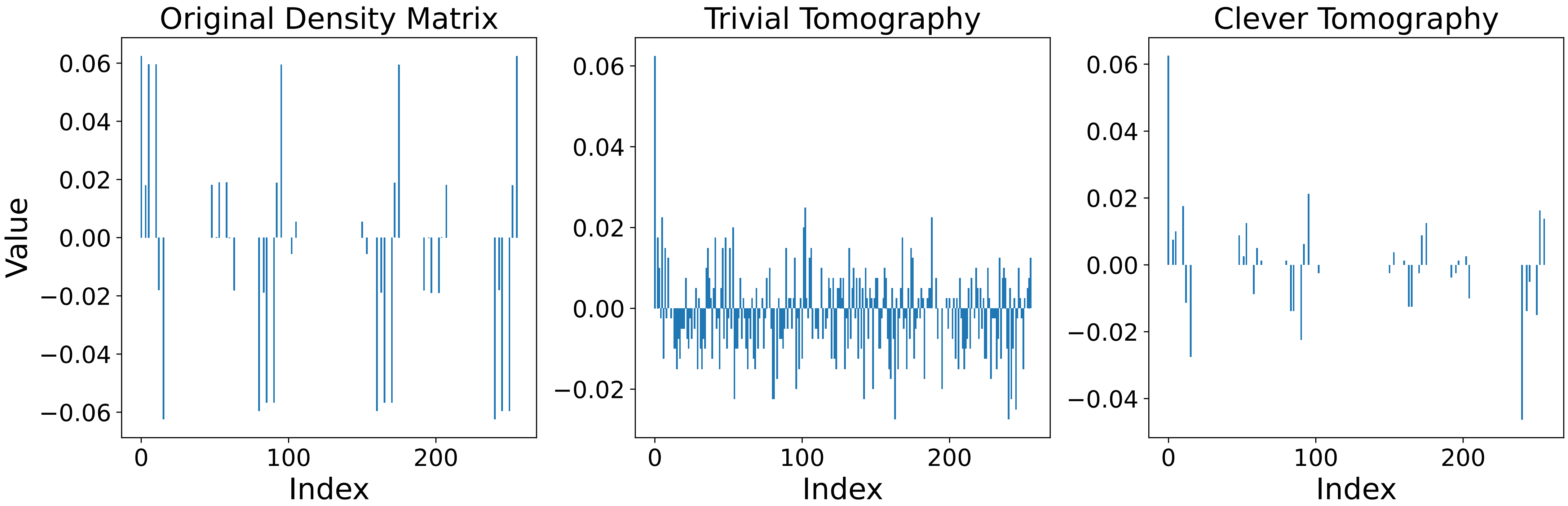


Knowledge Compilation

- Save subcircuit executions!
  - Subcircuit 1: 16 settings→4 settings.
  - Subcircuit 2: 16 settings→8 settings.

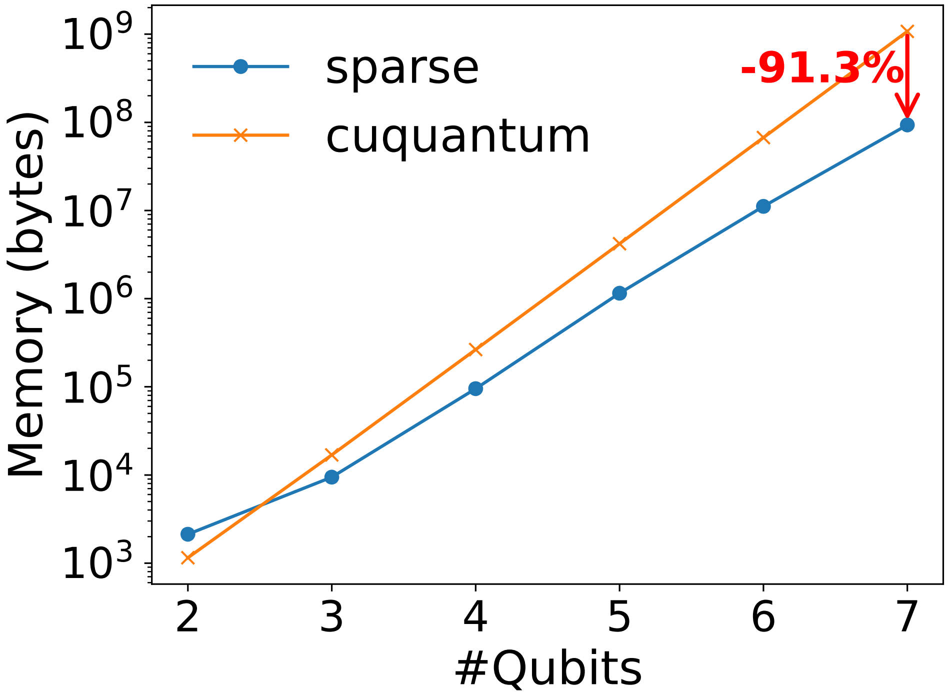


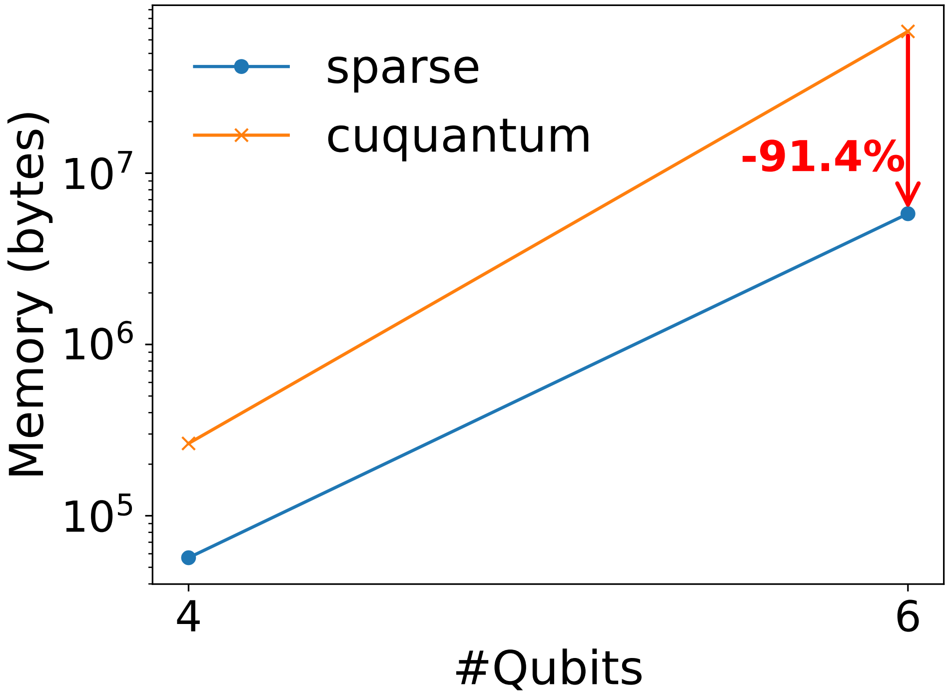
Error Mitigation



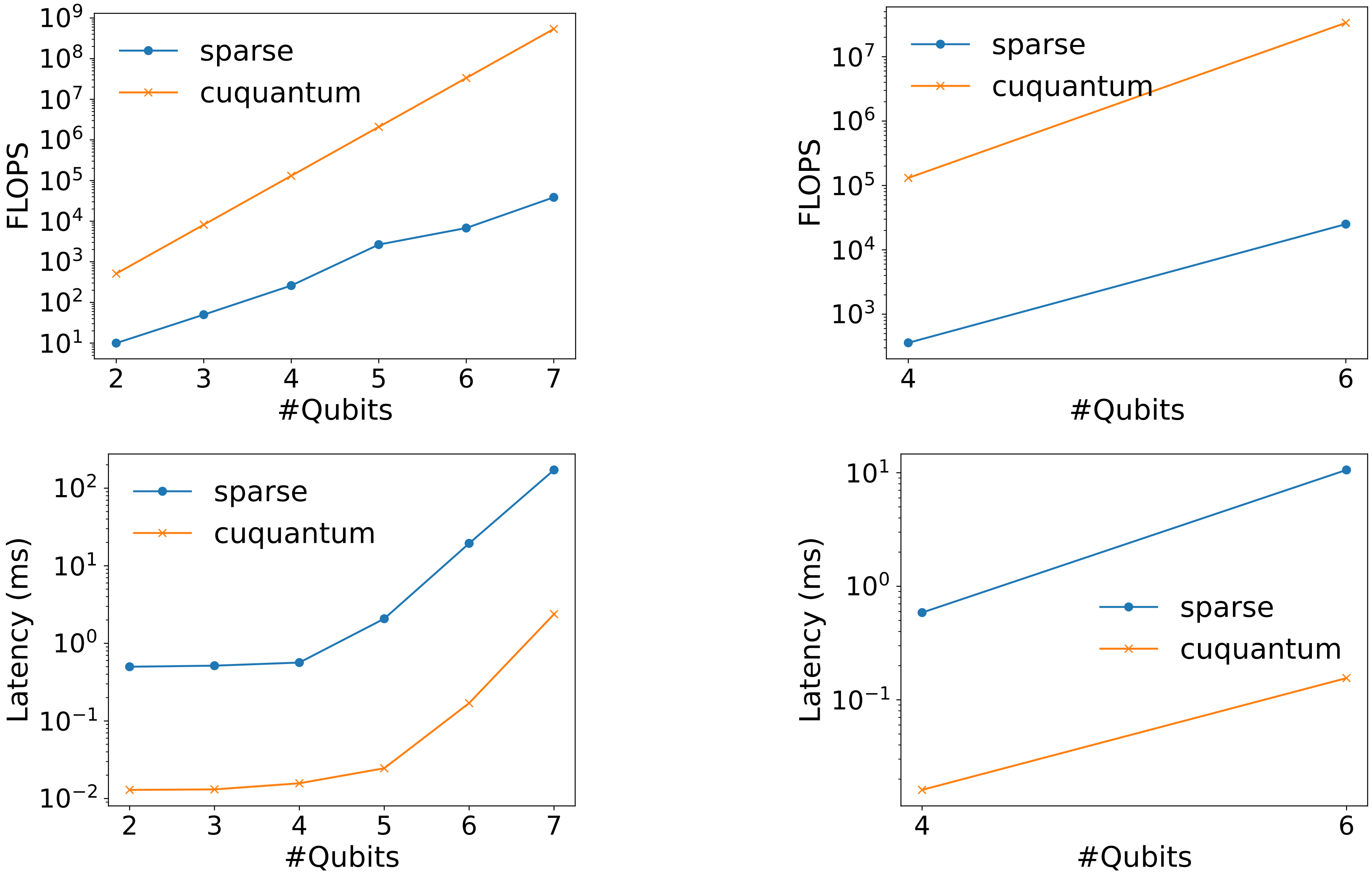
Impact of Sparsity

- We know the tensors are actually very sparse.
- We propose using sparse tensor contraction:





Impact of Sparsity



Impact of Sparsity

- For bigger-sized problems, it's even more sparse.

task name	number of qubits	max #edges	max sparsity	cuQuantum memory footprint	pgmQC memory footprint	memory footprint reduction
VQE	12	12	10.3%	128 MB	26.5 MB	79.3%
VQE	14	14	0.56%	2 GB	22.9 MB	98.88%
QFT	20	20	0.0009%	8 TB	190 MB	99.998%
GHZ	10	10	0.1%	8 MB	17 KB	99.8%
GHZ	20	20	0.00003%	8 TB	5 MB	99.999%
W State	20	20	0.006%	8 TB	1 GB	99.99%
Erdos	20	20	0.00005%	8 TB	37.7 MB	99.999%
Supremacy	16	16	1.0 %	34 GB	714 MB	97.9%
Sycamore	16	16	0.001 %	34 GB	1 MB	99.997%

TABLE I: Memory footprint comparison between cuQuantum and pgmQC for different quantum tasks.

Conclusion

- From the intuition that Clifford gates stabilize Pauli strings:
  - Tensors in quantum simulation are sparse.
  - In circuit cutting, sparsity can save the effort to create the tensor (subcircuit executions) and contract the tensor (classical postprocessing).
  - Proposed clever tomography to reduce the effort of subcircuit executions and mitigate errors.
  - Proposed using sparse tensor contraction to save memory footprint during classical postprocessing.

Thank you!