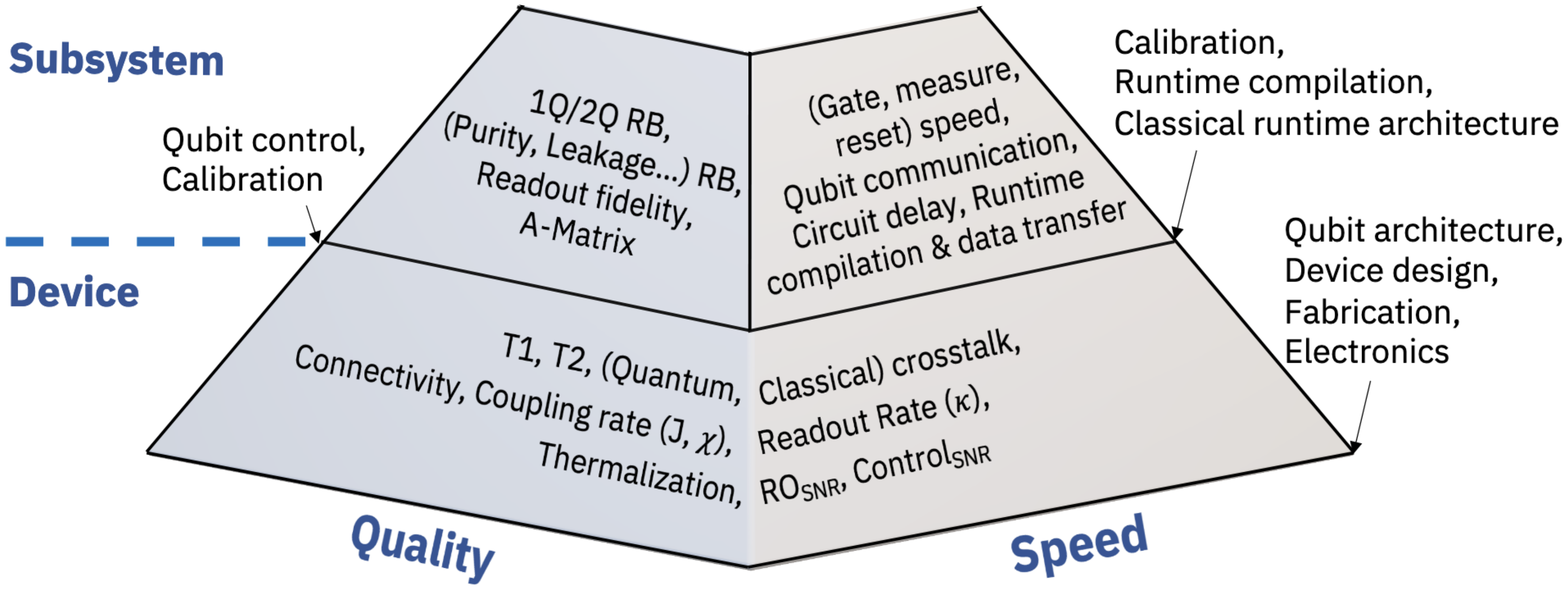
Quantum Benchmarking II

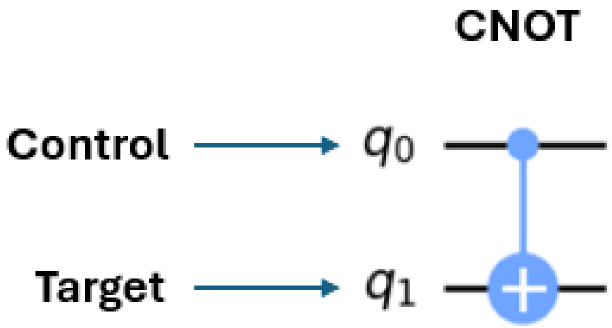
Majo Lozano
Quantum Hardware Engineer
IBM Quantum

What have we covered so far?



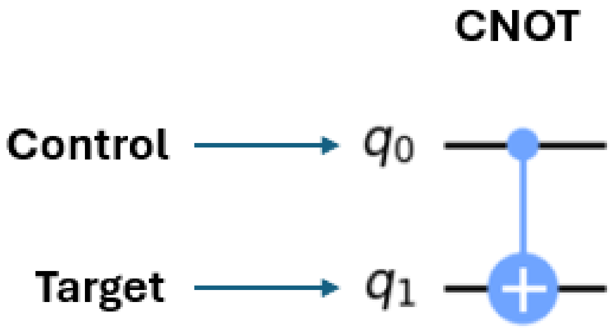
Hellinger Fidelity of a Bell State

- Benchmarks a 2Q entangling gate
- CNOT



Hellinger Fidelity of a Bell State

- Generates a bell state using an odd number of CNOTS
- 4 distinct bell states (maximally states) entangled states)
- Compares with ideal results



$$|\Phi^+\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}}$$

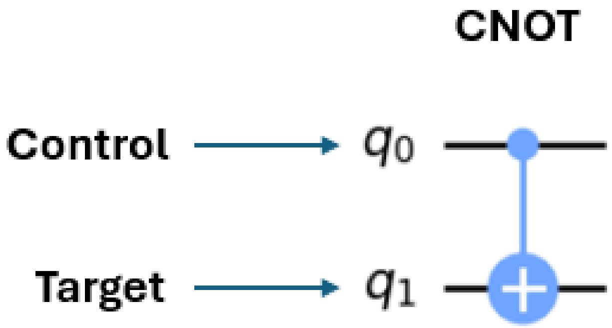
$$|\Phi^-\rangle = \frac{|00\rangle - |11\rangle}{\sqrt{2}}$$

$$|\Psi^+\rangle = \frac{|01\rangle + |10\rangle}{\sqrt{2}}$$

$$|\Psi^-\rangle = \frac{|01\rangle - |10\rangle}{\sqrt{2}}$$

Hellinger Fidelity of a Bell State

- Hellinger Fidelity given as (1-H^2)^2 where H is the Hellinger distance
- Distance between 2 count distributions (ideal case and experimental case)
- Sensitive to qubit initialization, 1Q and 2Q errors, and readout errors



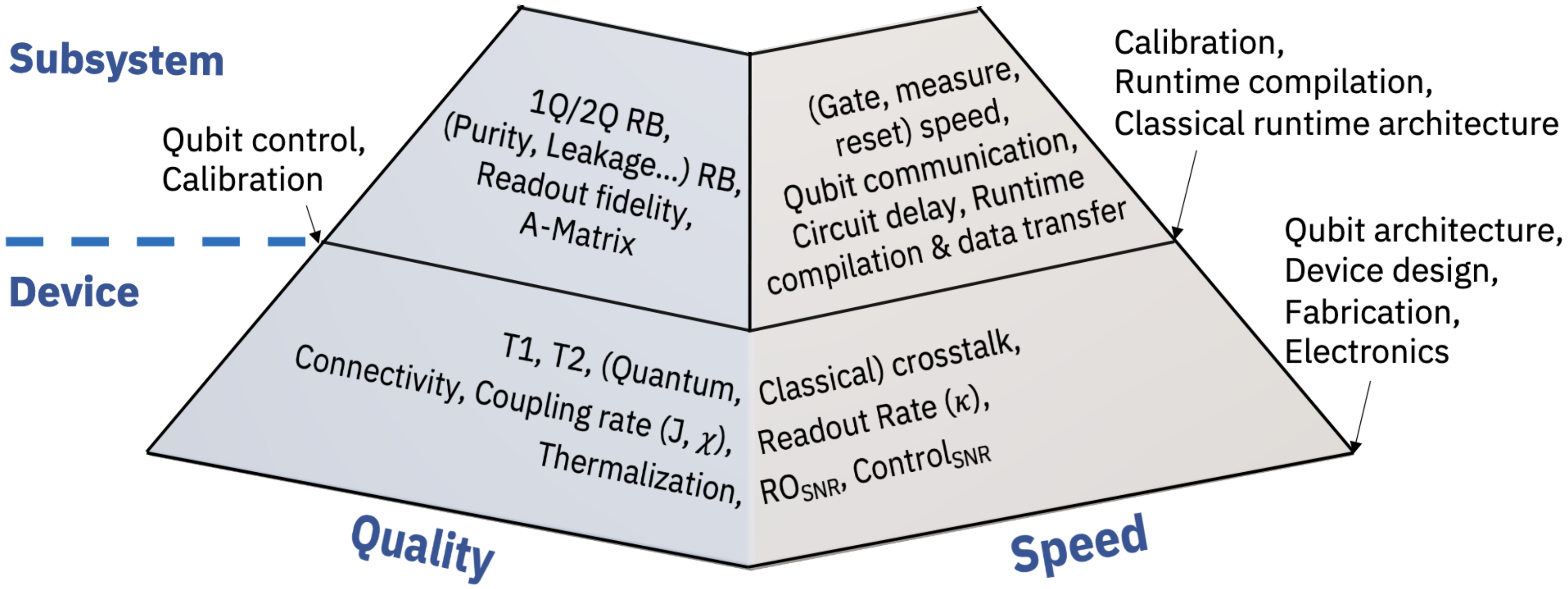
$$|\Phi^+\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}}$$

$$|\Phi^-\rangle = \frac{|00\rangle - |11\rangle}{\sqrt{2}}$$

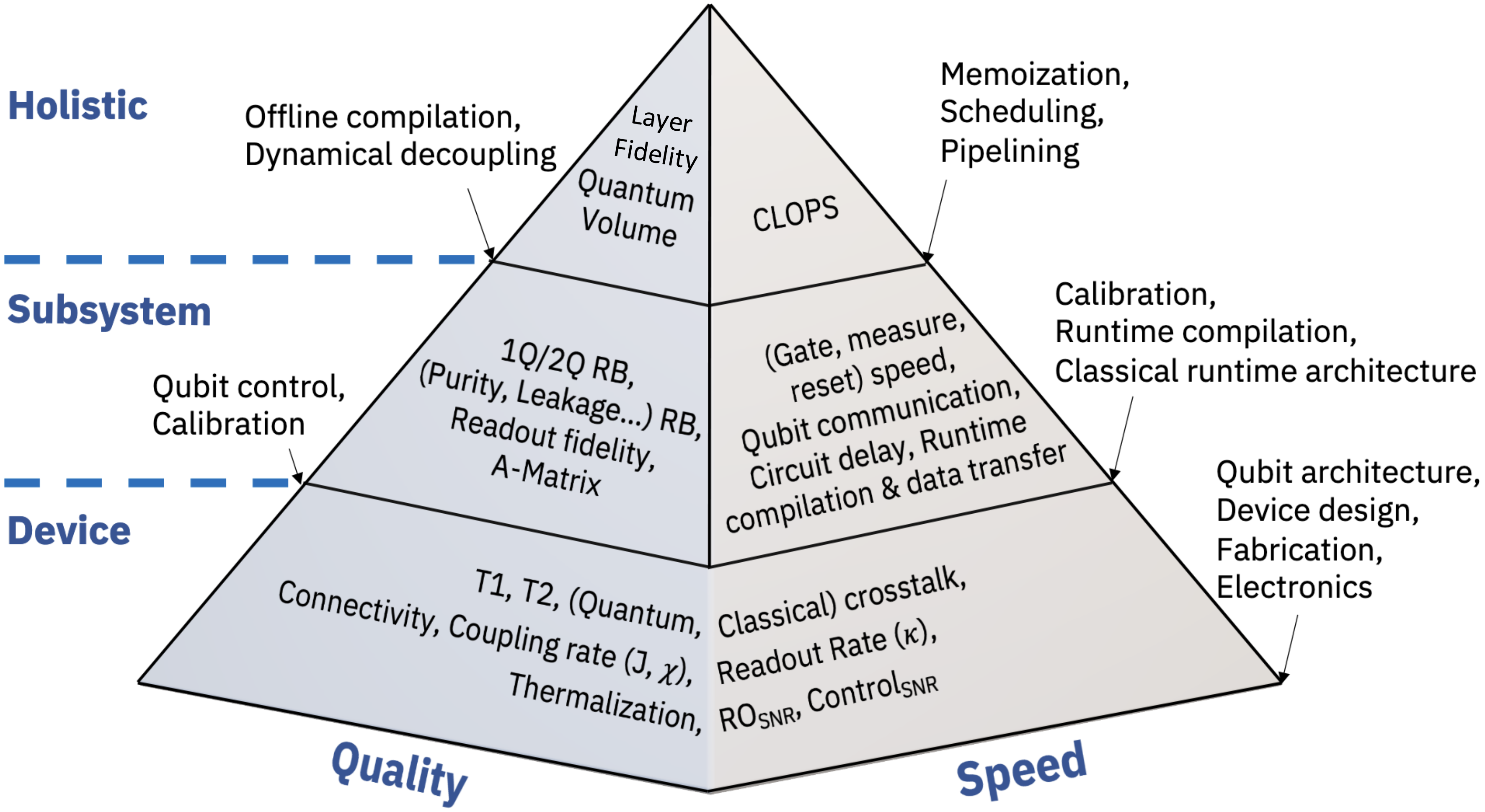
$$|\Psi^+\rangle = \frac{|01\rangle + |10\rangle}{\sqrt{2}}$$

$$|\Psi^-\rangle = \frac{|01\rangle - |10\rangle}{\sqrt{2}}$$

What have we covered so far?

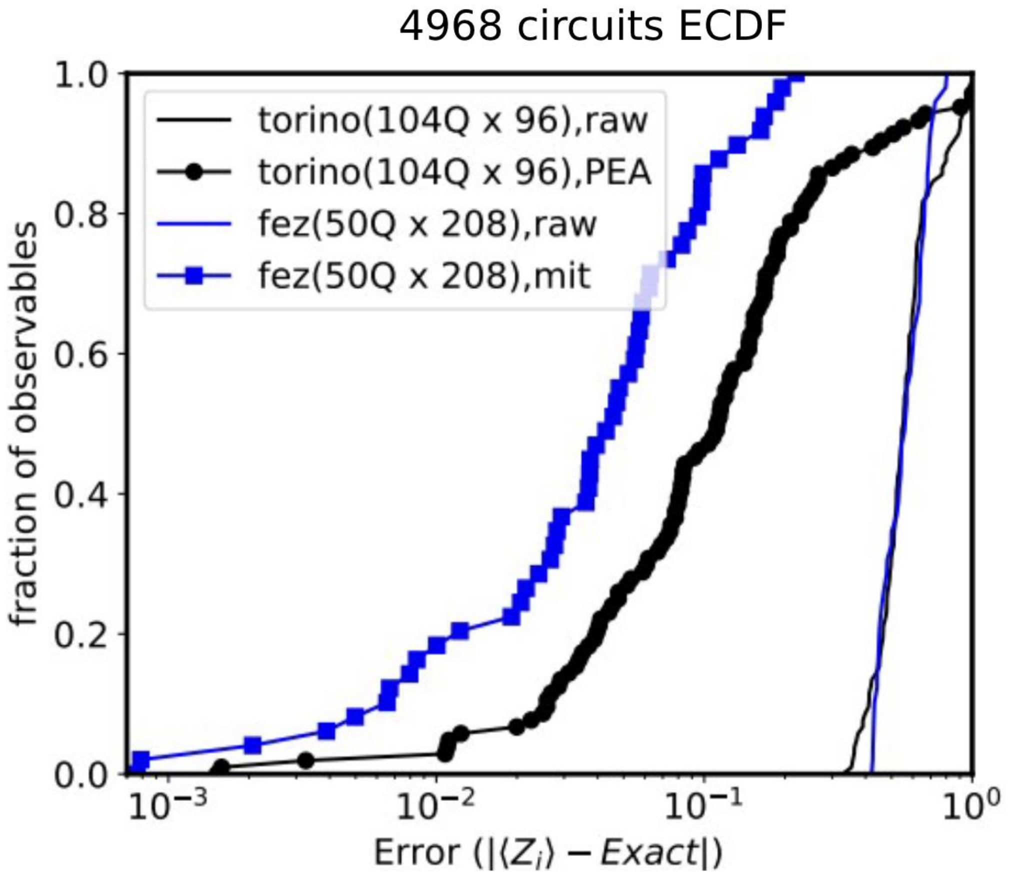


What have we covered so far?



Mirror Circuit (MC)

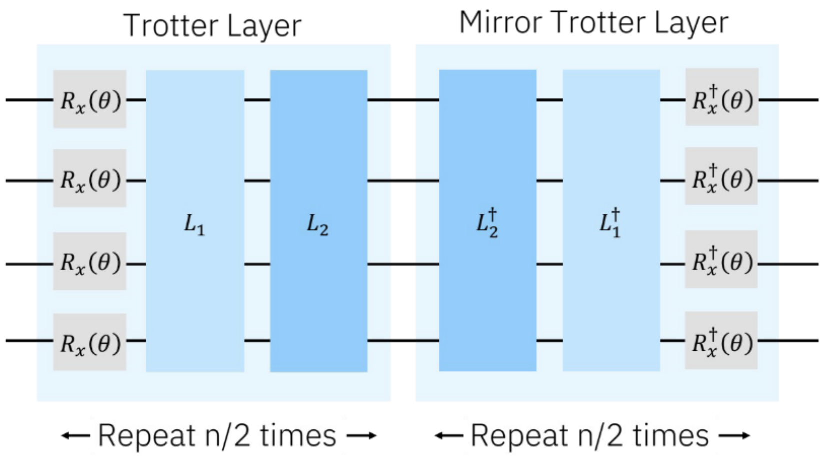
- Primitive quality benchmark
- Indicates the ability to deliver accurate observables for utility-scale circuits (5K gates)
- **Idea**: Apply a utility-scale circuit followed by its inverse. Useful for testing error mitigation (we require estimation tasks with easy-to-compute ideal answers)



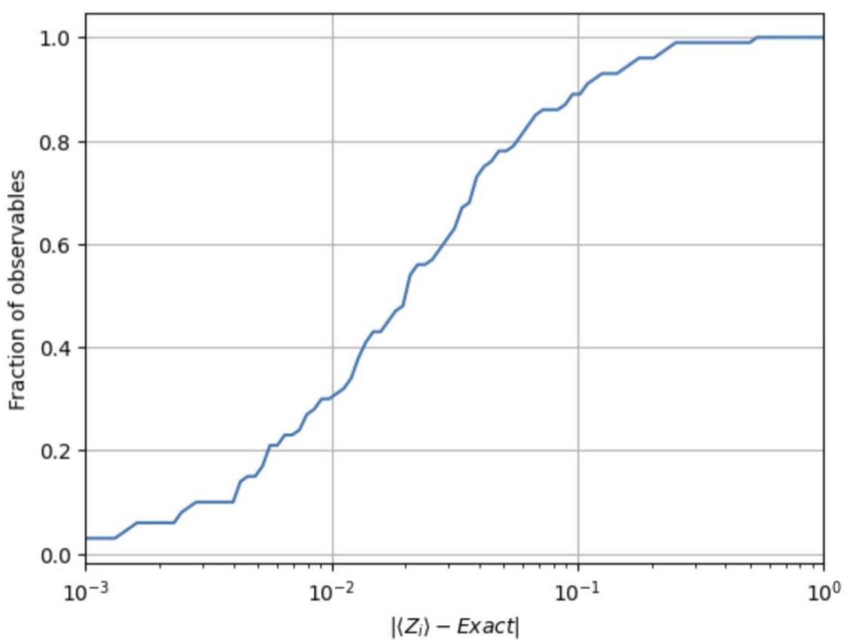
More than 40% of 1Q observables within 10% of ideal value at 5k scale.

Mirror Circuit (MC)

- **Utility-scale circuit**: evolution of a 1D Ising chain (trotterized time) followed by its inverse
- Effective action is equivalent to the **identity**
- Easy to detect whether the returned observables are accurate
- Plot: distribution (observables - exact case). We care about observables within 10% of the ideal solution



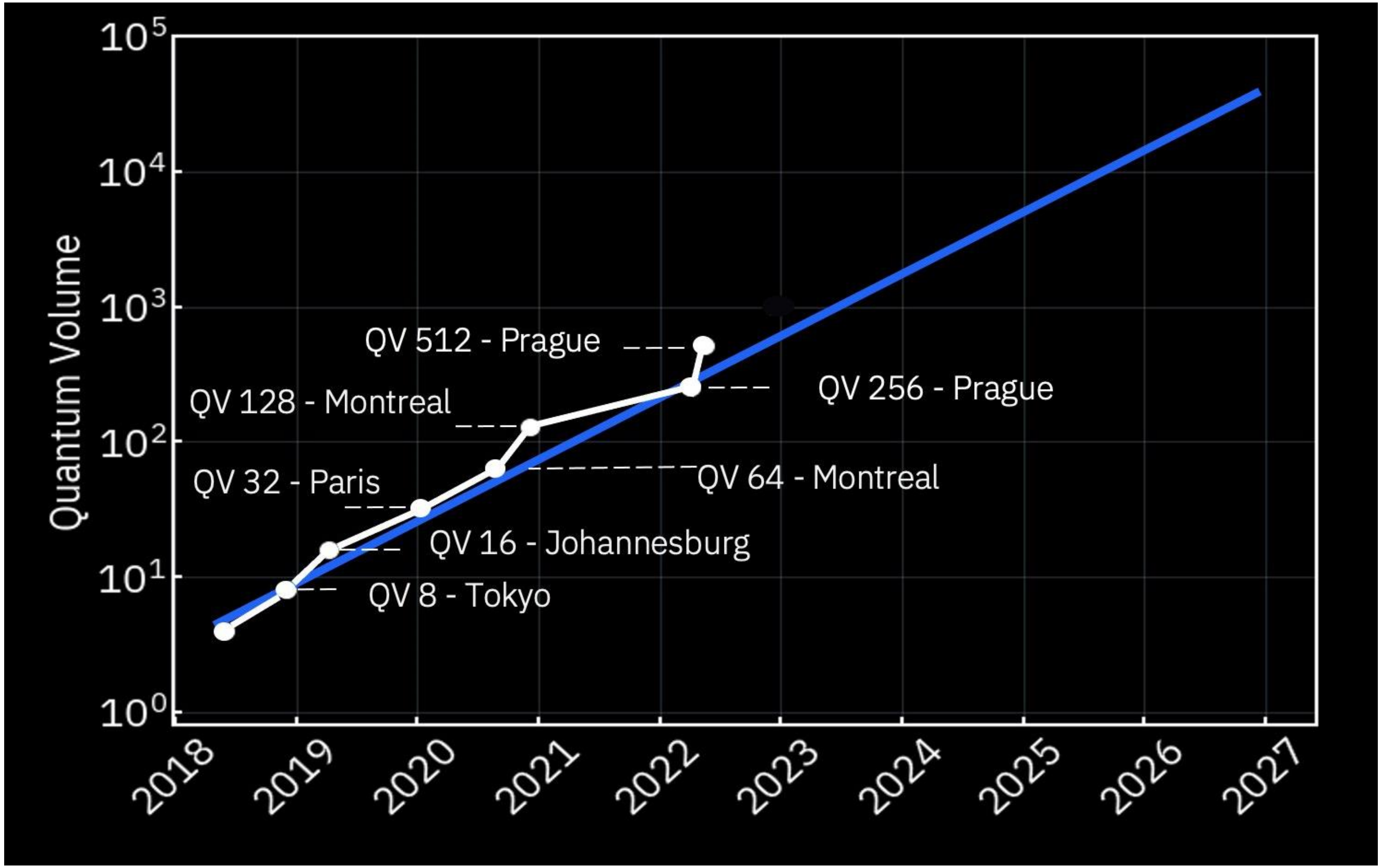
Recent Heron R2 data

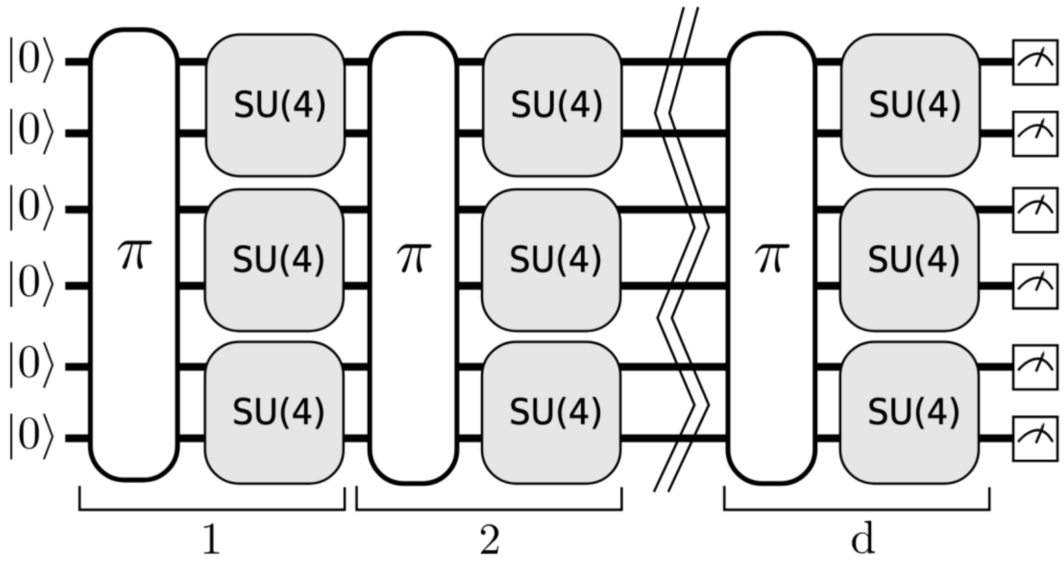


More than 85% of observables within 10% of ideal value at 5k scale

Quantum Volume (QV)

- Holistic metric (quality)
- Indicates how faithfully a “square” quantum circuit of depth d can be implemented
- Sensitive to coherence, gate fidelity, and measurement fidelity

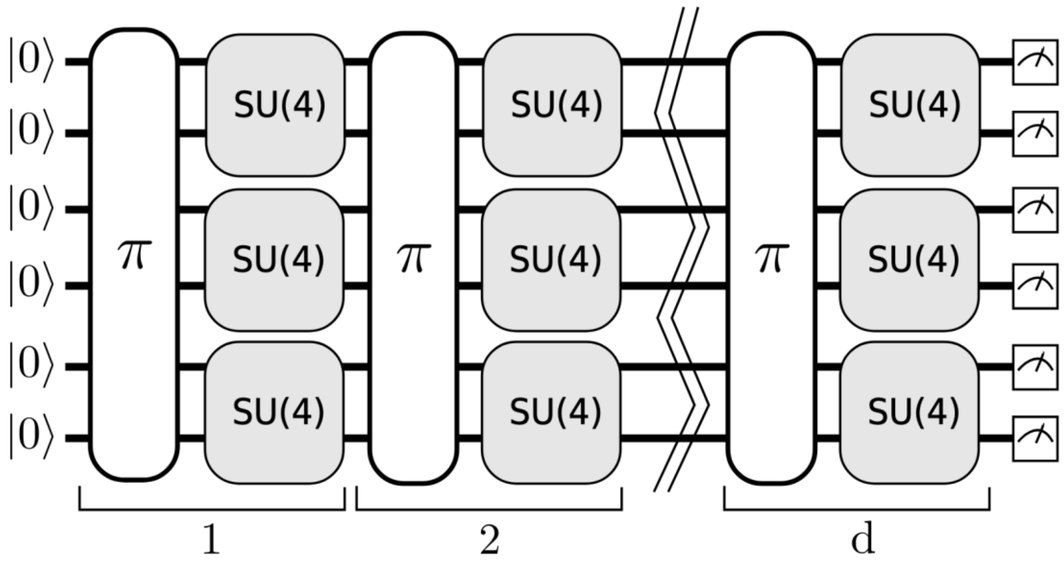




1)

Define a QV circuit consisting of d layers (random permutations of the qubit labels followed by 2Q gates)

Source: https://arxiv.org/pdf/1811.12926



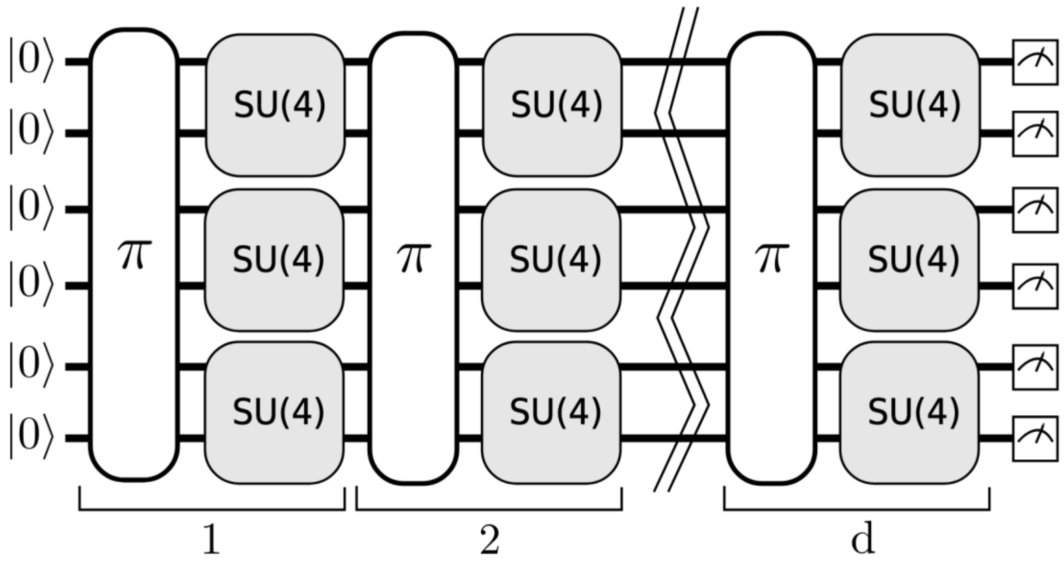
1)

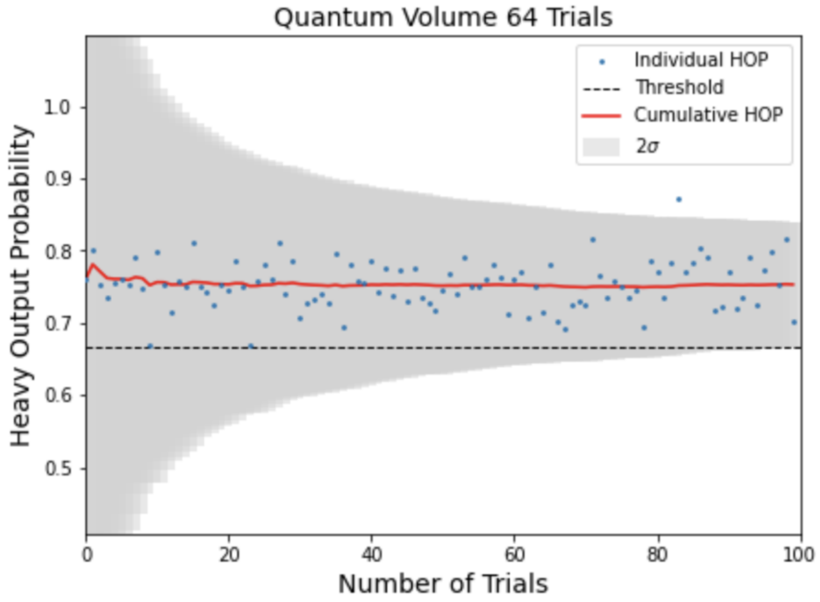
Define a QV circuit consisting of d layers (random permutations of the qubit labels followed by 2Q gates)

2)

Simulate ideal circuit results

Source: https://arxiv.org/pdf/1811.12926





1)

Define a QV circuit consisting of d layers (random permutations of the qubit labels followed by 2Q gates)

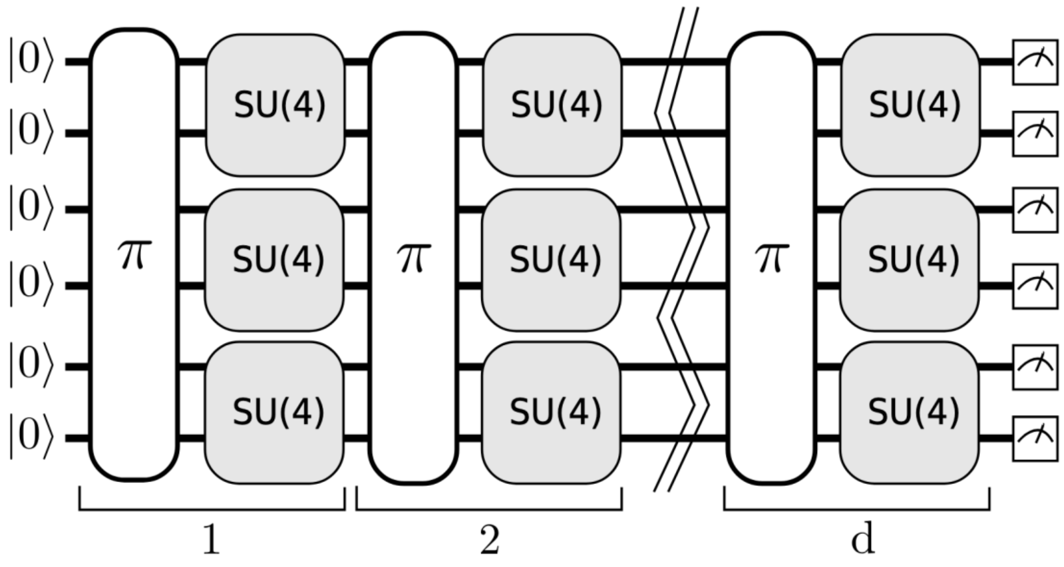
2)

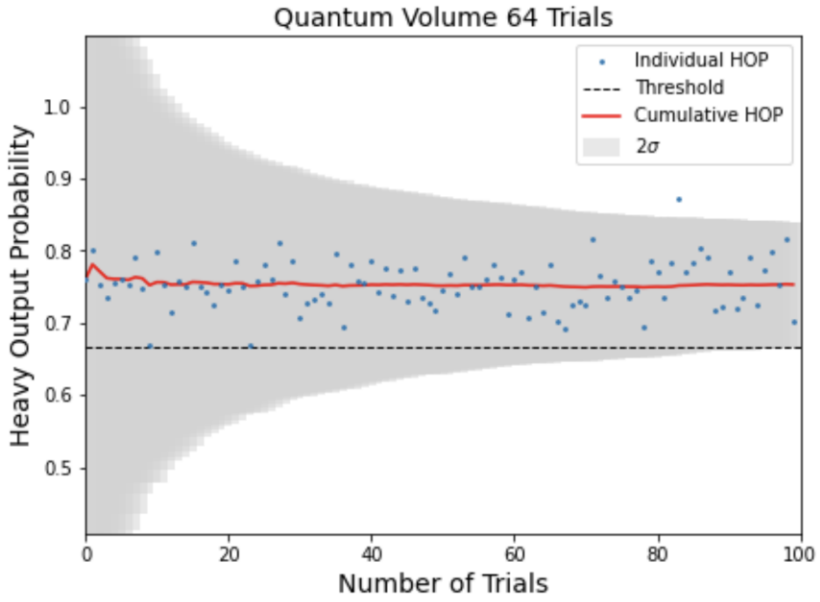
Simulate ideal circuit results

3)

Run the QV circuit and compare the heaty output states with the ideal simulation results

Source: https://arxiv.org/pdf/1811.12926





1)

Define a QV circuit consisting of d layers (random permutations of the qubit labels followed by 2Q gates)

2)

Simulate ideal circuit results

3)

Run the QV circuit and compare the heaty output states with the ideal simulation results

4)

The largest QV circuit of depth d that produces more than 2/3 of heavy outputs determines the QV on a system → 2^d

Source: https://arxiv.org/pdf/1811.12926

Some points to consider regarding QV

(stay tuned for Layer Fidelity, another holistic benchmark that addresses these points)

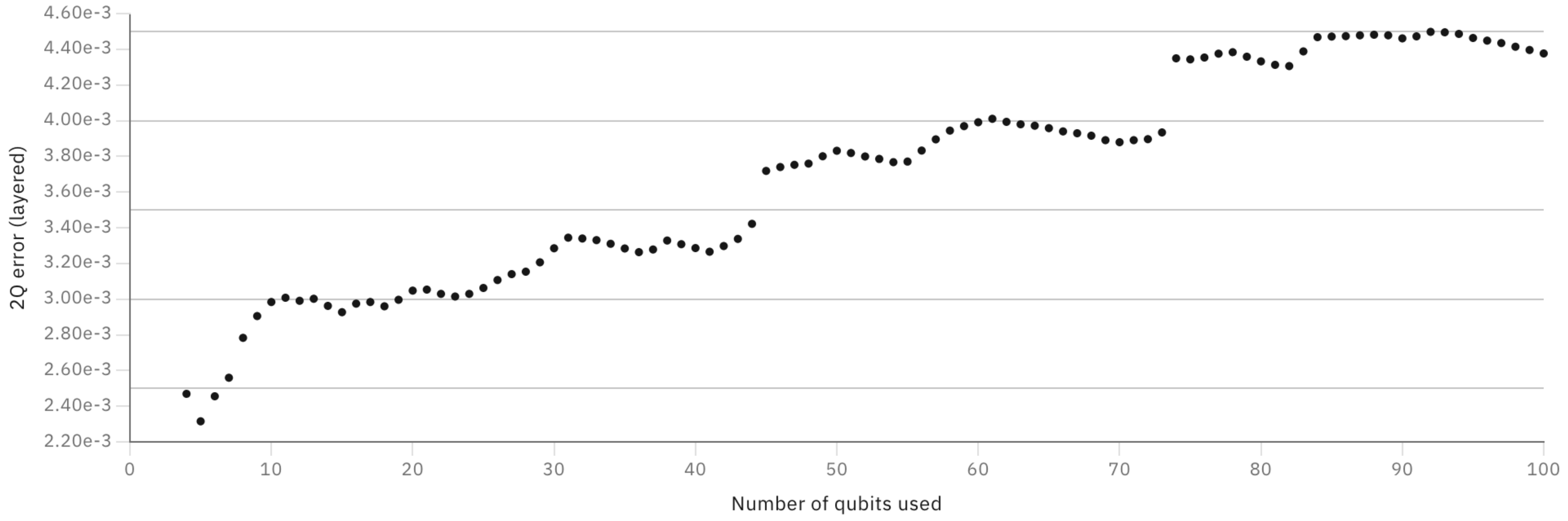
- Is a discrete pass / fail test

- Does not provide individual gate information

- For devices with more qubits than log2(QV), QV is not the best representative number of overall quality

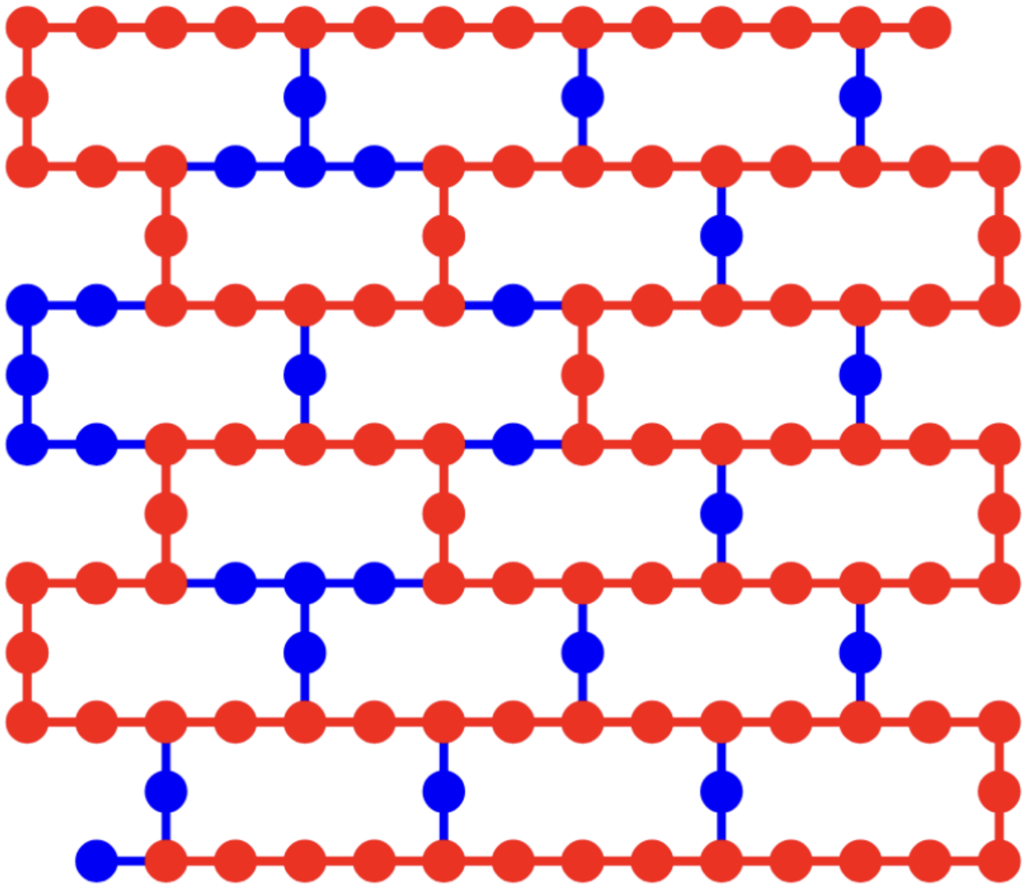
Layer Fidelity (LF)

- Holistic metric (quality)
- Great complement to QV
- Captures fidelity of N fully-connected qubits over M layers
- Insensitive to state preparation and measurement errors



Layer Fidelity (LF)

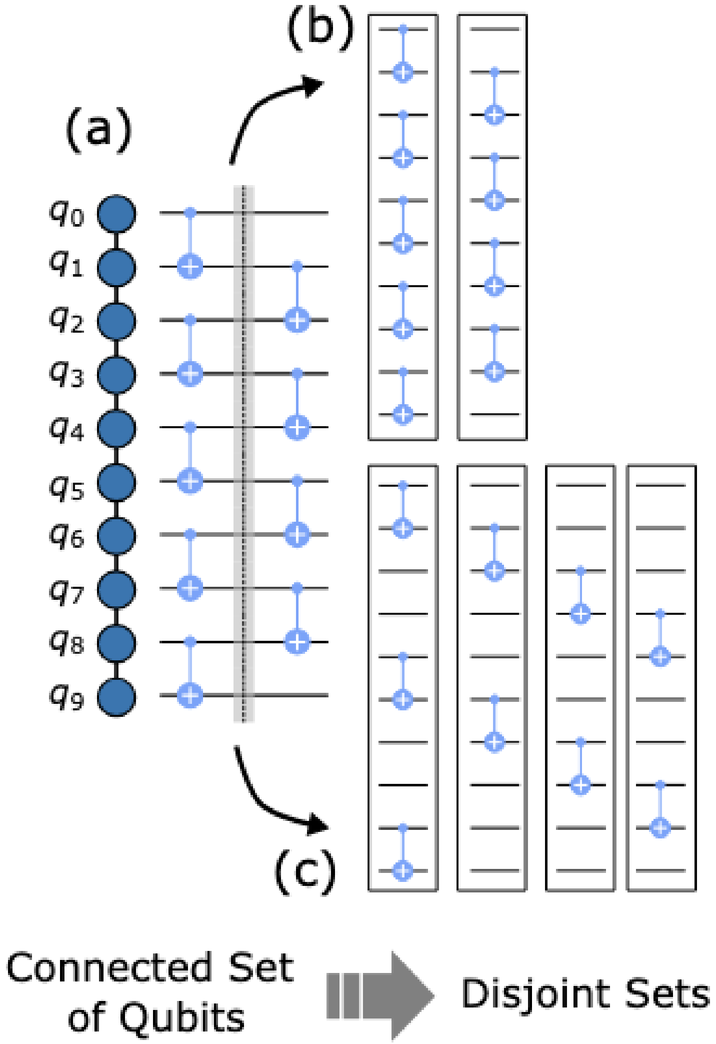
- Size of N is flexible
- N can be closer to the device size → captures device wide performance



Source: https://arxiv.org/pdf/2311.05933

Layer Fidelity (LF)

- Partition a set of qubits into M disjoint layers (M=2 for a linear chain)

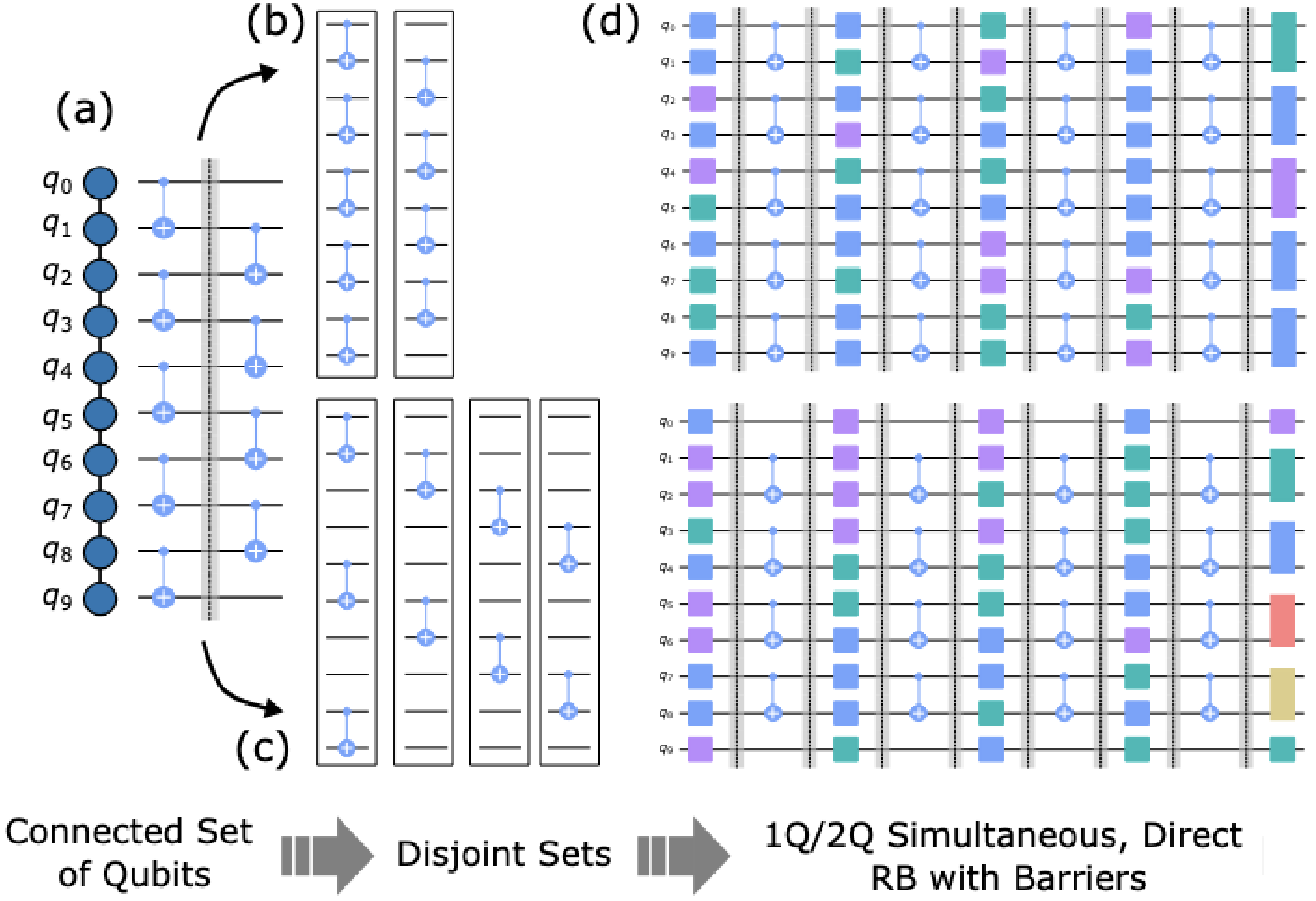


Source: https://arxiv.org/pdf/2311.05933

Layer Fidelity (LF)

On the disjoint layers:

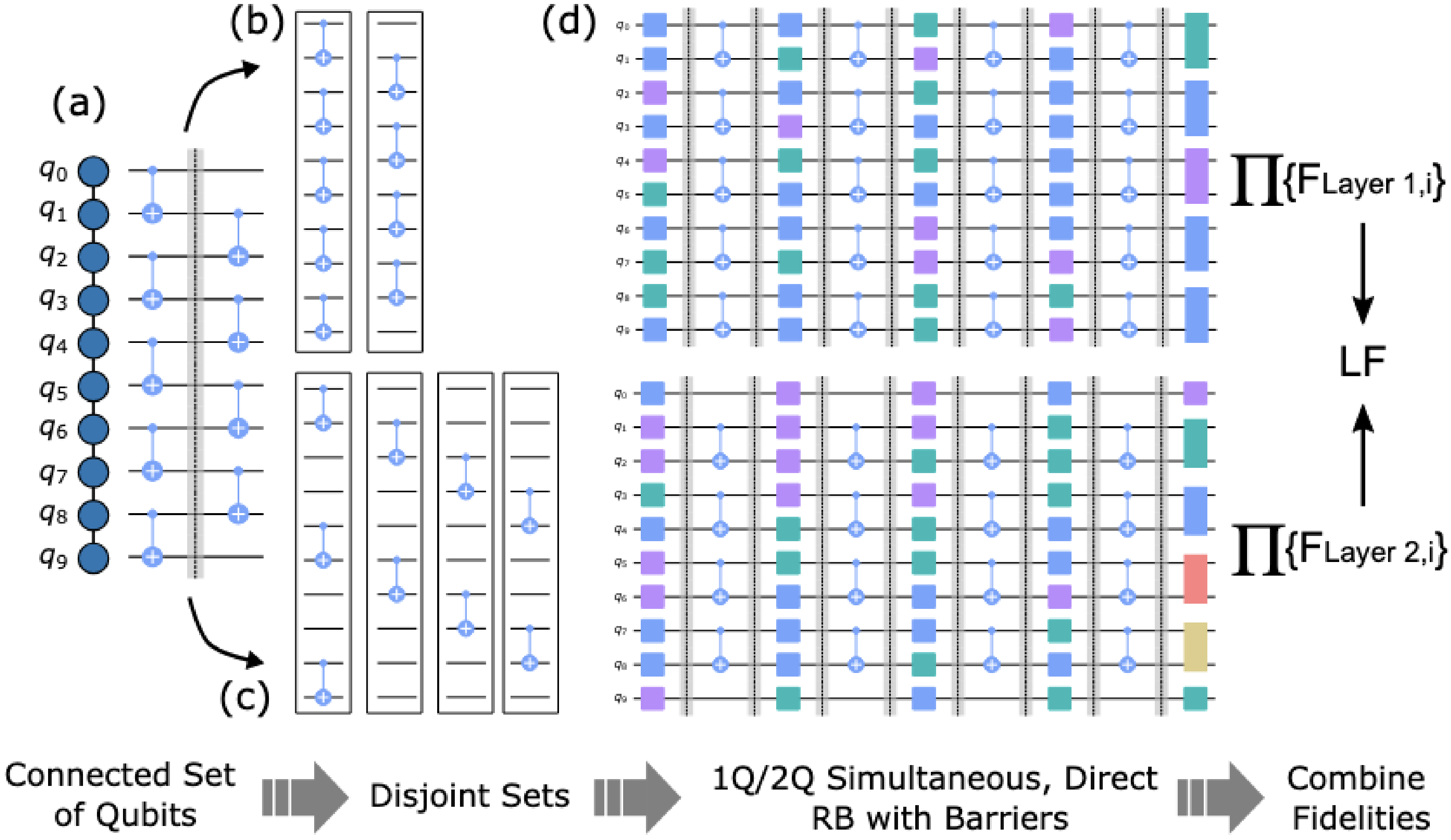
- Run layers of simultaneous direct RB (imposing a timing barrier between 2Q operations, and twirling through random 1Q Clifford layers)
- Captures crosstalk



Source: https://arxiv.org/pdf/2311.05933

Layer Fidelity (LF)

- Get individual process fidelities and multiply them to get the overall layer fidelity



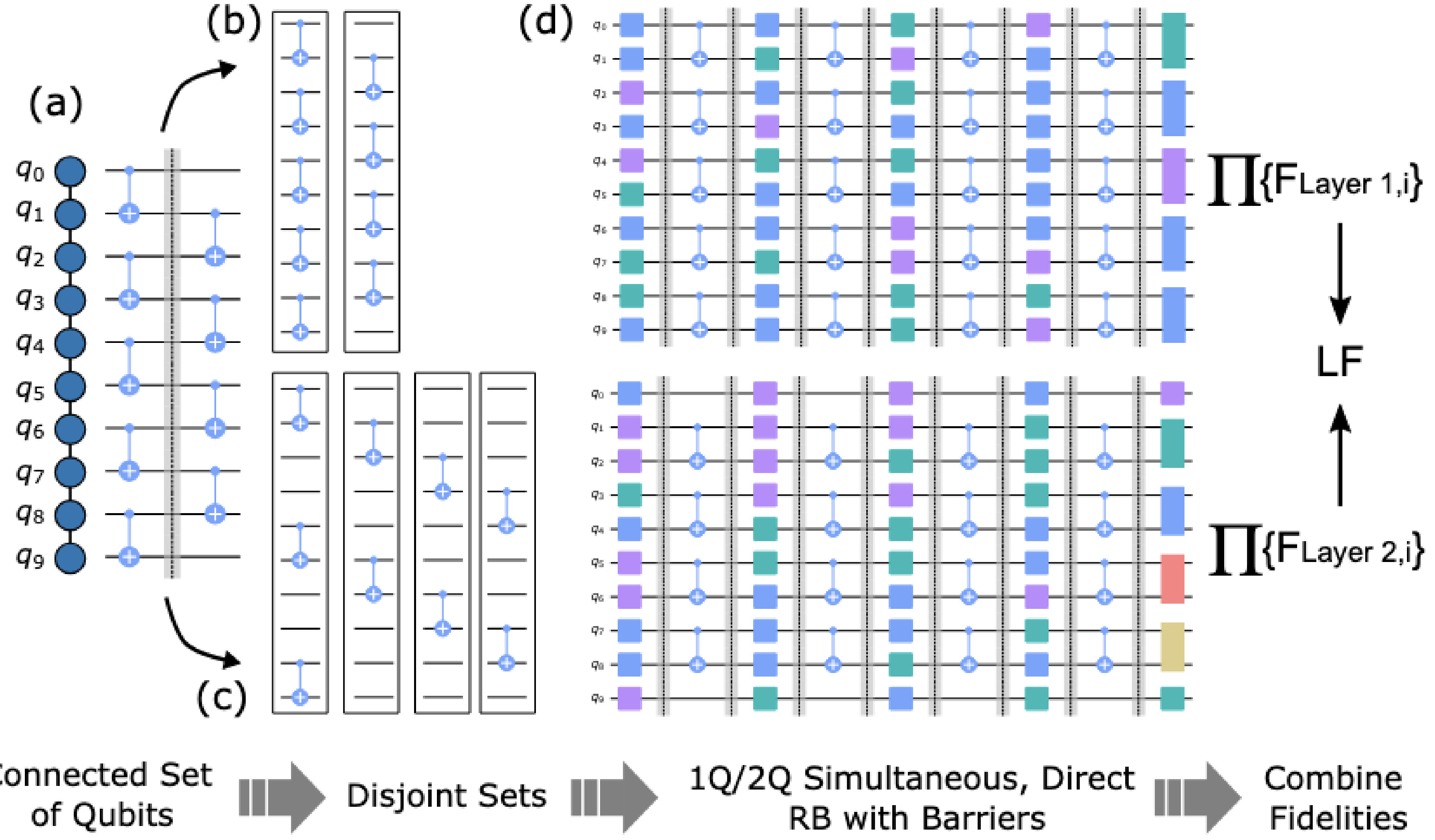
Source: https://arxiv.org/pdf/2311.05933

Layer Fidelity (LF)

$$LF_m = \prod_j F_{j,m},$$

$$LF = \prod_m^M LF_m.$$

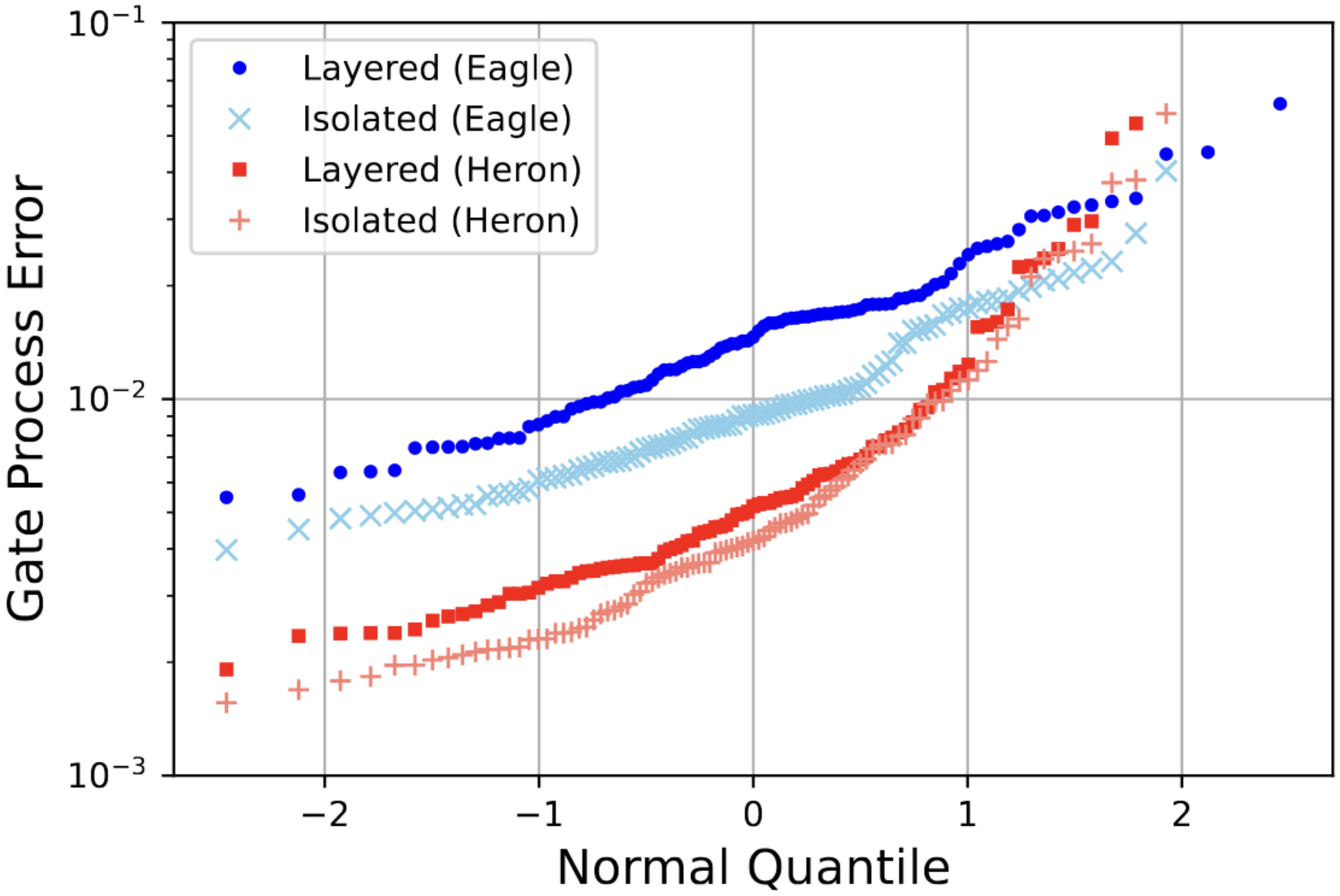
$$\text{EPLG} = 1 - LF^{1/n_{2q}},$$



Source: https://arxiv.org/pdf/2311.05933

Layer Fidelity (LF)

- Error distributions between isolated and layered errors are closer on Herons (virtually eliminates crosstalk)



Source: https://arxiv.org/pdf/2311.05933

Why use Layer Fidelity?

- Is a discrete pass / fail test

- Introduces a size-independent quantity 'Error per Layered Gate' (EPLG)

Why use Layer Fidelity?

- Is a discrete pass / fail test

- Introduces a size-independent quantity 'Error per Layered Gate' (EPLG)

- Does not provide individual gate information

- Provides individual gate errors

Why use Layer Fidelity?

- Is a discrete pass / fail test
- Introduces a size-independent quantity 'Error per Layered Gate' (EPLG)

- Does not provide individual gate information
- Provides individual gate errors

- For devices with more qubits than log2(QV), QV is not the best representative number of overall quality
- Can span across an entire device via layered circuits

Layer Fidelity (LF)

- Typical LF for IBM Heron processors

ibm_**marrakesh**

QPU status	Online
Processor type	Heron r2

Qubits	2Q error (best/layered)	CLOPS
156	1.48e-3/$\boxed{4.38e\text{-}3}$	195K

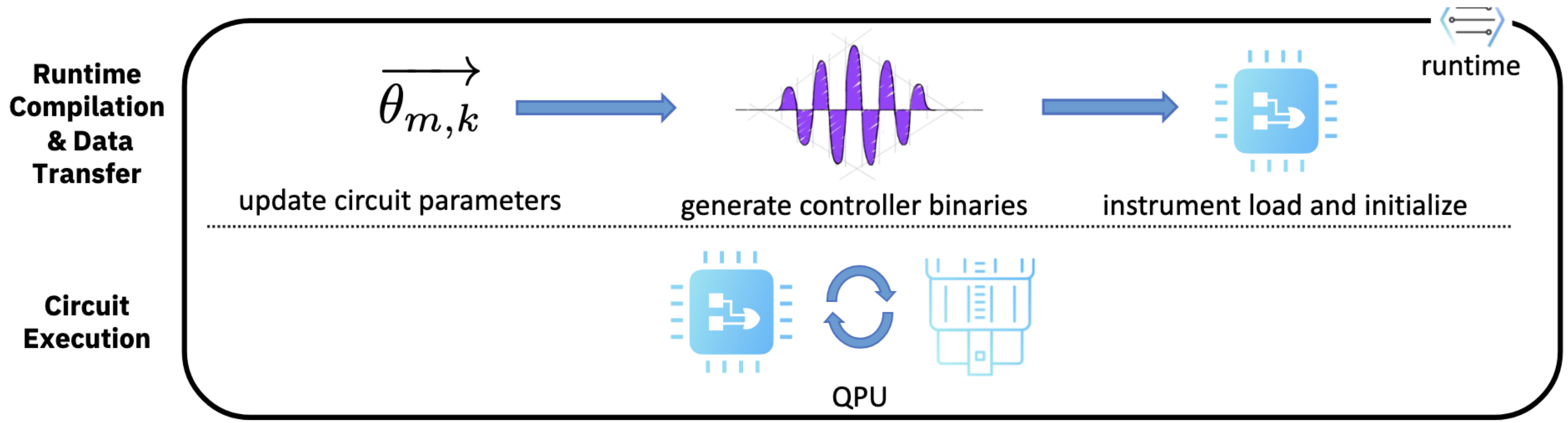
ibm_**fez**

QPU status	Online
Processor type	Heron r2

Qubits	2Q error (best/layered)	CLOPS
156	2.27e-3/$\boxed{5.25e\text{-}3}$	195K

Circuit Layer Operations per Second (CLOPS)

- Holistic metric (speed)
- Measures how many layers can be executed per unit of time



Circuit Layer Operations per Second (CLOPS)

Time break down:

1

Time spent running the circuit: circuit execution

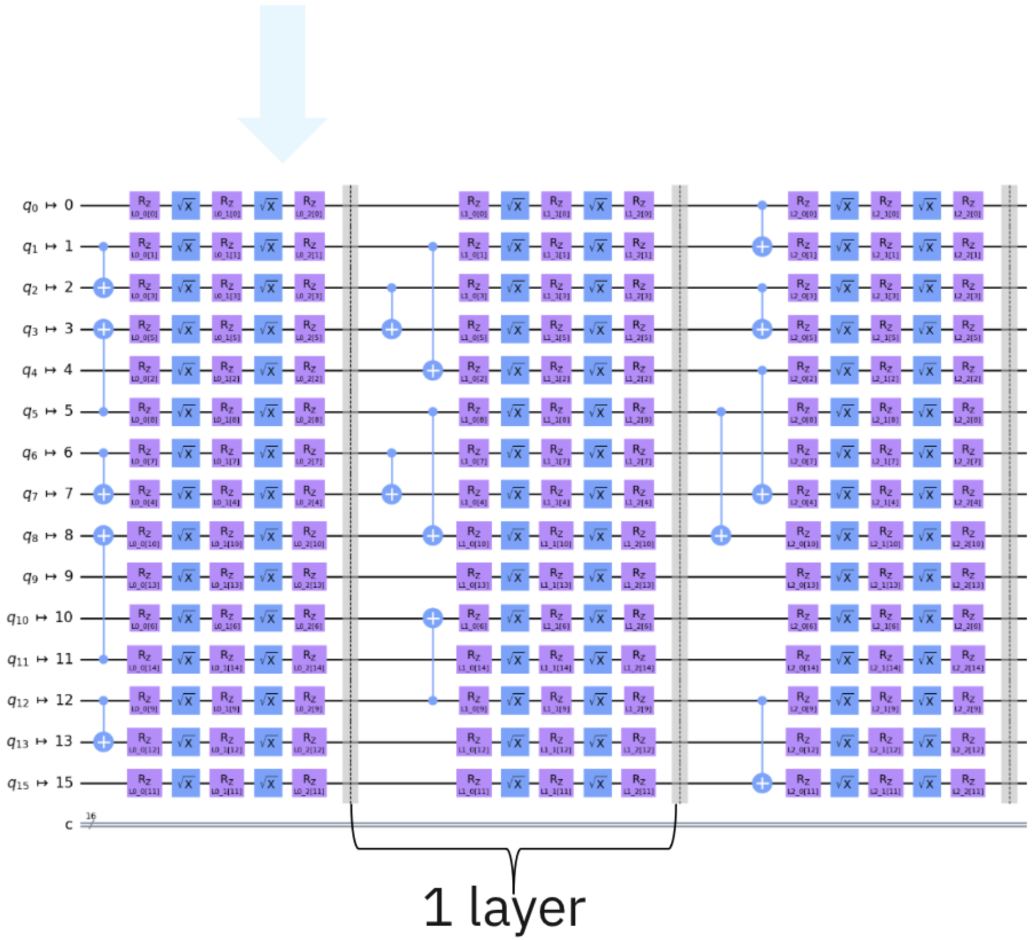
2

Delay time between each shot of each circuit: circuit delay

3

Time spent on preparing the circuits (parameter updates, run-time compilation, waveform generation) as well as data transfers (circuit submission to the backend, instrument initialization, instrument load, return of results to user) : run-time compilation and data transfer

Parameterized single qubit gates

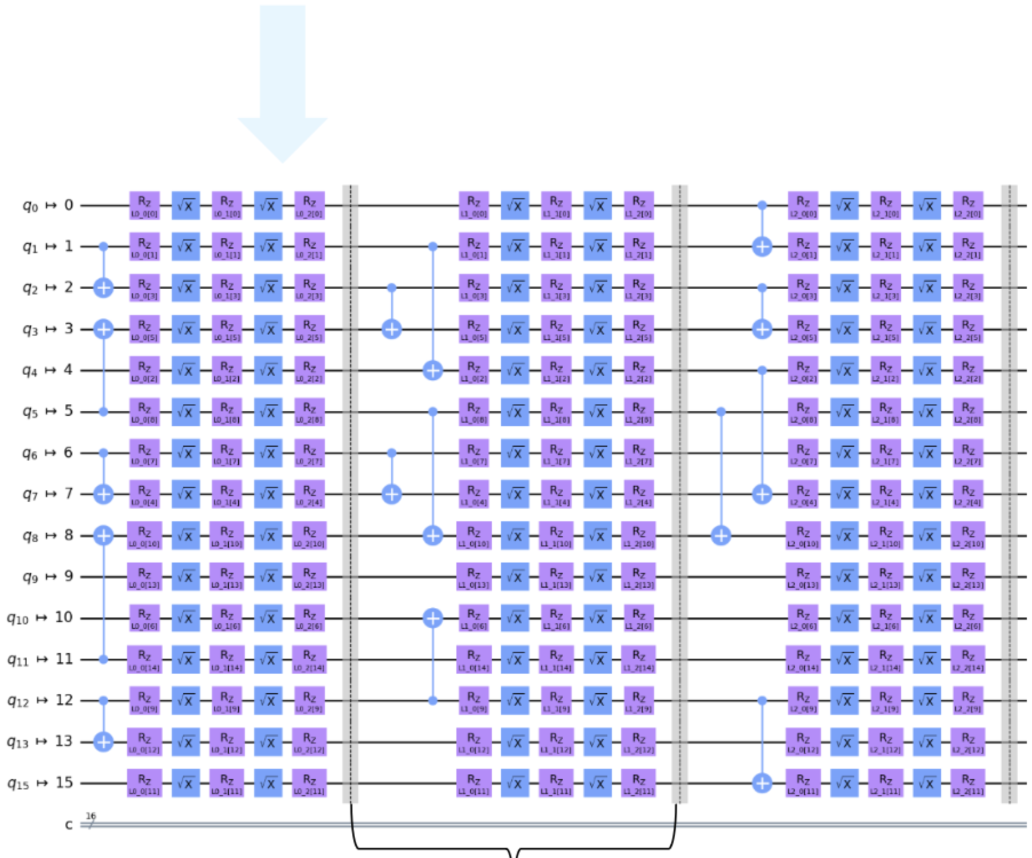


1)

CLOPS measures multiple executions of parametrized layers

The goal is to show that the software and hardware can execute utility scale circuits efficiently

Parameterized single qubit gates



1 layer

1)

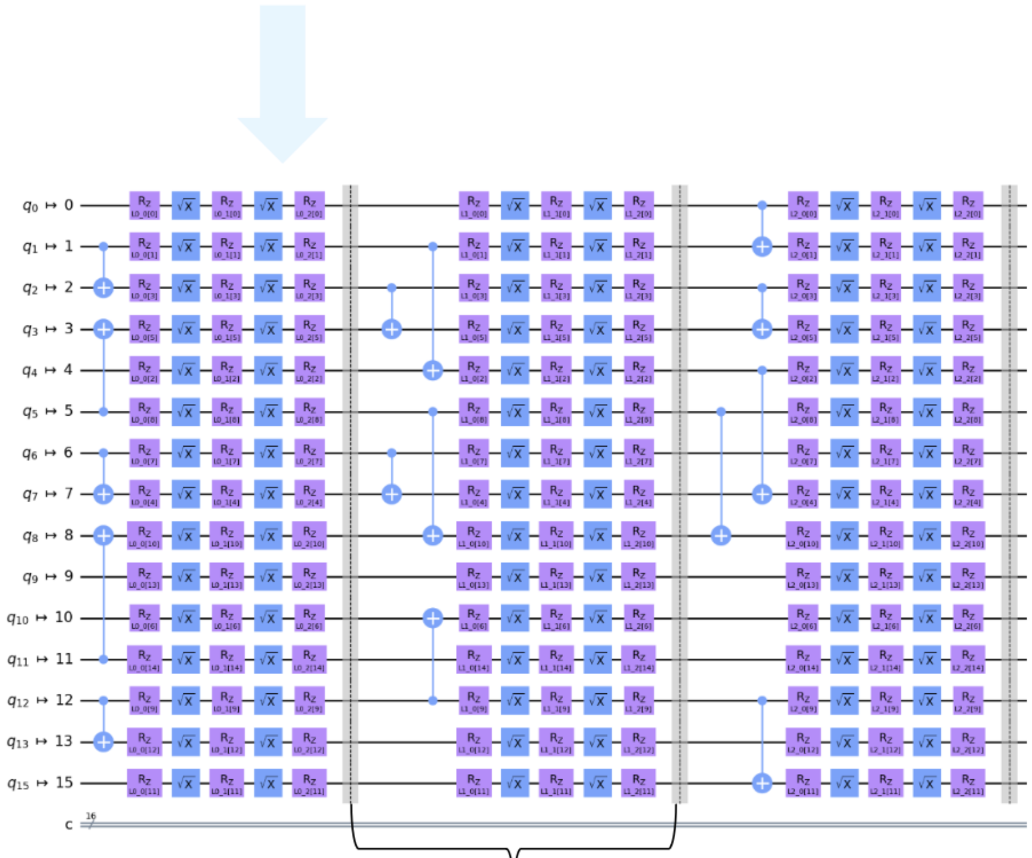
CLOPS measures multiple executions of parametrized layers

The goal is to show that the software and hardware can execute utility scale circuits efficiently

2)

CLOPS can be measured on any circuit size (width/layers). The default is 100x100.

Parameterized single qubit gates



1 layer

1)

CLOPS measures multiple executions of parametrized layers

The goal is to show that the software and hardware can execute utility scale circuits efficiently

2)

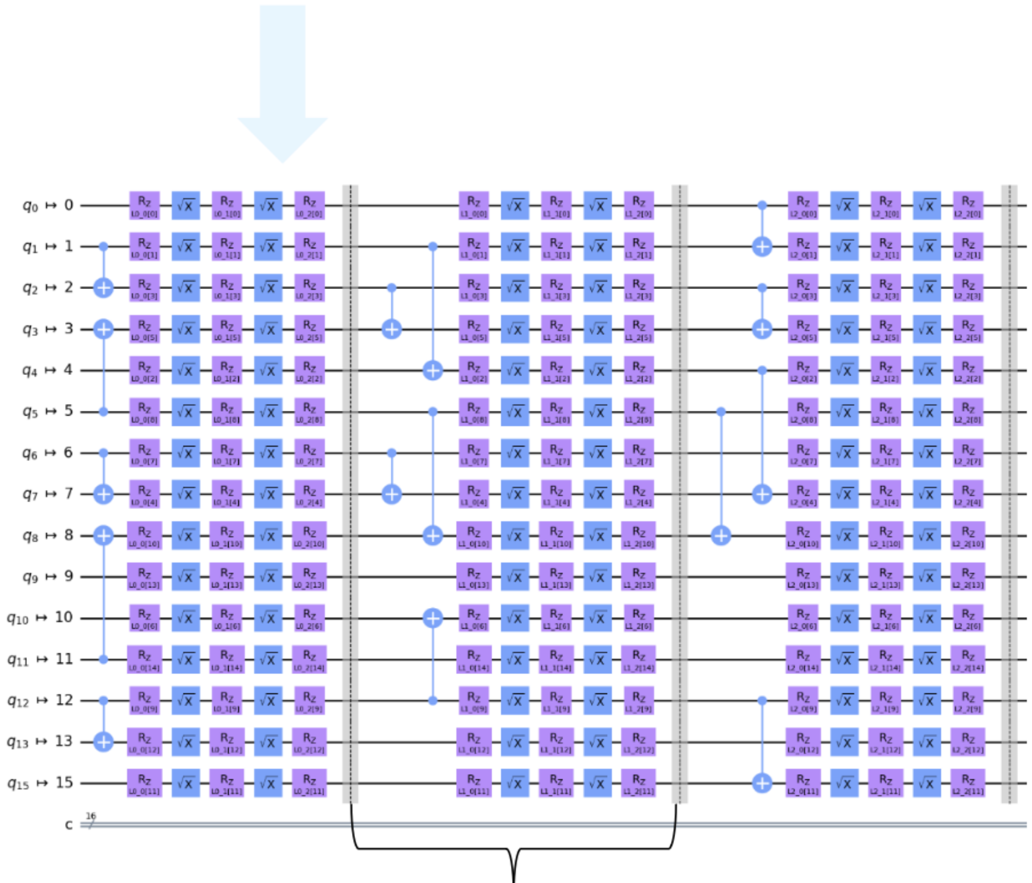
CLOPS can be measured on any circuit size (width/layers). The default is 100x100.

3)

Layers consist of 2Q gates that are executed in parallel, and fully parametrized single qubit rotations on all qubits

Source: https://arxiv.org/pdf/2110.14108

Parameterized single qubit gates



1 layer

1)

CLOPS measures multiple executions of parametrized layers

The goal is to show that the software and hardware can execute utility scale circuits efficiently

2)

CLOPS can be measured on any circuit size (width/layers). The default is 100x100.

3)

Layers consist of 2Q gates that are executed in parallel, and fully parametrized single qubit rotations on all qubits

4)

CLOPS = the number of layers executed per second

Source: https://arxiv.org/pdf/2110.14108

Circuit Layer Operations per Second (CLOPS)

- Typical CLOPS for IBM Heron processors

ibm_**torino**

QPU status	Online
Processor type	Heron r1

Qubits	2Q error (best/layered)	CLOPS
133	1.63e-3/6.09e-3	210K

ibm_**fez**

QPU status	Online
Processor type	Heron r2

Qubits	2Q error (best/layered)	CLOPS
156	2.27e-3/5.25e-3	195K

Notebooks and resources

- GitHub repo: qiskit-device-benchmarking
https://github.com/qiskit-community/qiskit-device-benchmarking
- Layer Fidelity Notebook
https://github.com/qiskit-community/qiskit-device-benchmarking/blob/main/notebooks/layer_fidelity_single_chain.ipynb
- CLOPS Notebook
https://github.com/qiskit-community/qiskit-device-benchmarking/tree/main/qiskit_device_benchmarking/clops
- Mirror Circuit Notebook
https://github.com/qiskit-community/qiskit-device-benchmarking/tree/main/qiskit_device_benchmarking/mirror_test
- Bell State and Identity Tomography
https://github.com/qiskit-community/qiskit-device-benchmarking/blob/main/notebooks/bell_state_tomography.ipynb
- Quantum Volume Notebook
https://github.com/mjlp123/qiskit-device-benchmarking/blob/majo_qv/notebooks/quantum_volume.ipynb

Plot LF and EPLG per chain length

In [14]:
```
lfu.make_lf_eplg_plots(backend=backend,
    exp_data=exp_data,
    chain=qchain,
    machine=machine
)
```