Evaluating the performance of quantum process units (QPUs) at large width and depth

Community Call - Unitary Foundation

March 12, 2025 I Alejandro Montanez-Barrera | FZJ - JSC

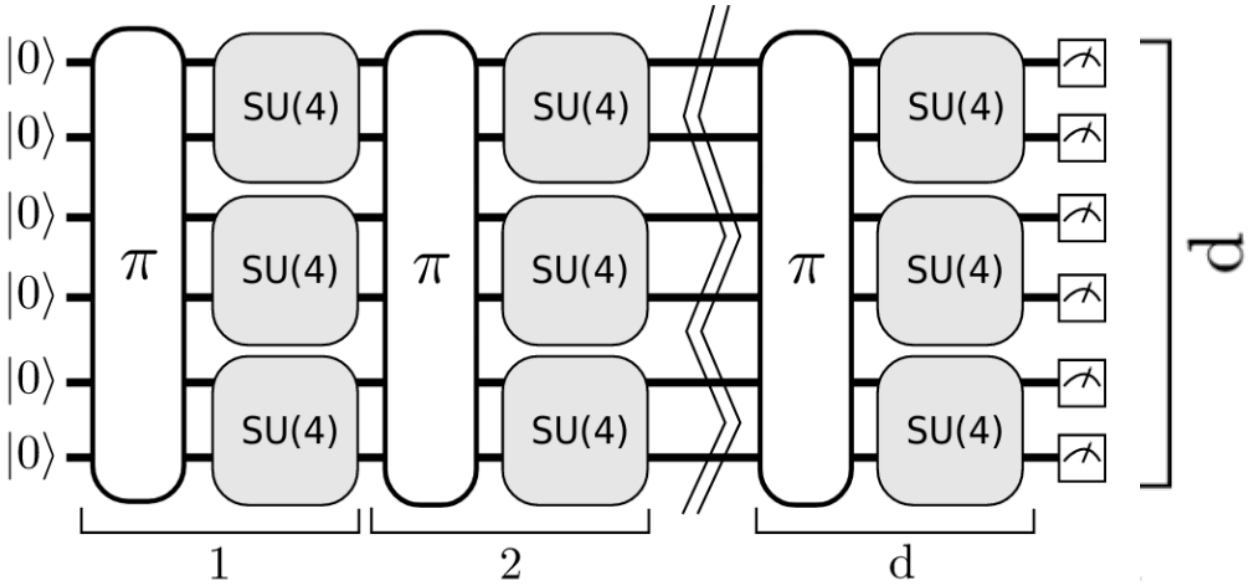
https://arxiv.org/abs/2502.06471

Outline

- Quantum Benchmarks
- The LR-QAOA Benchmark
- The problem behind the LR-QAOA Benchmark.
- The random limit threshold
- 1D-Chain Benchmark
- Native layout (NL) benchmark
- Fully connected (FC) Benchmark
- Conclusions

Benchmarks in Quantum computing

(a) QUANTUM VOLUME

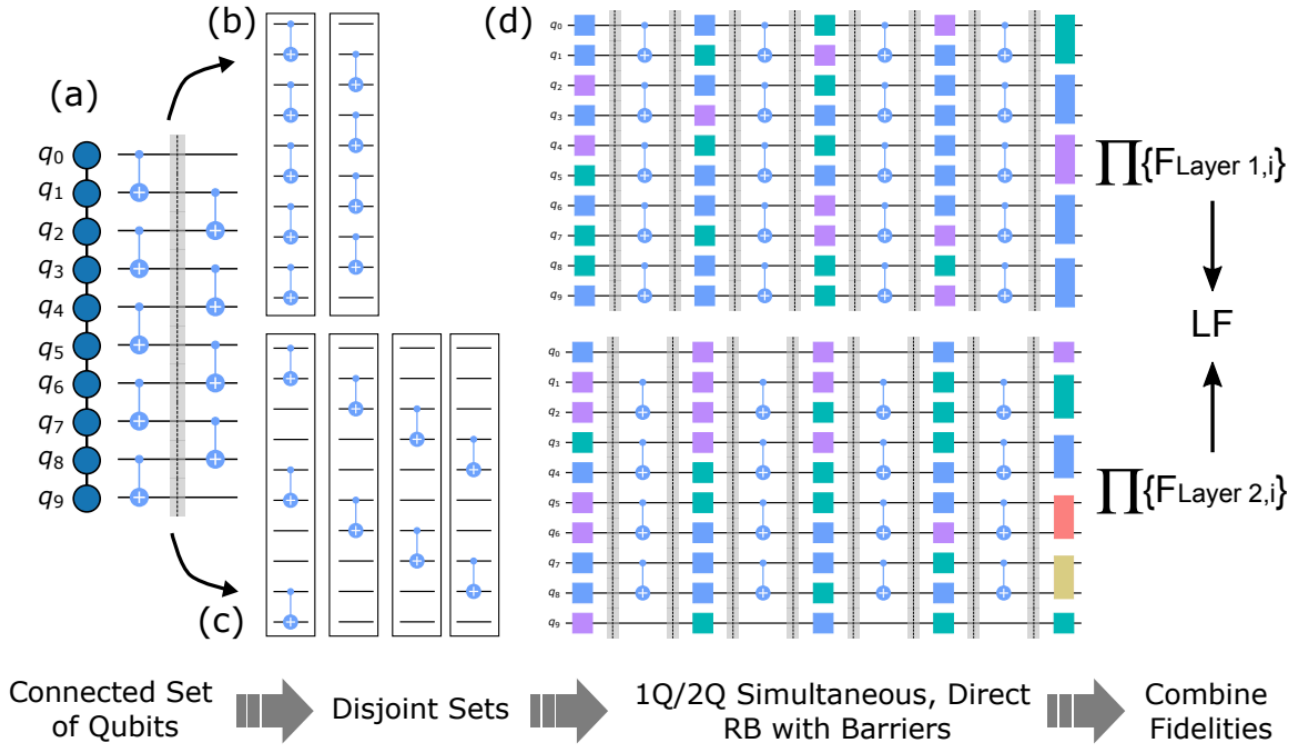


https://doi.org/10.1103/PhysRevA.100.032328

- Applied to a squared circuit
- Gives a holistic sense of the QPU performance
- It has been widely applied by quantum computing companies: IBM, IQM, Quantinuum, and Alpine Quantum Technologies.
- It is not a scalable performance benchmark and we lose information about shorter-depth

Benchmarks in Quantum computing

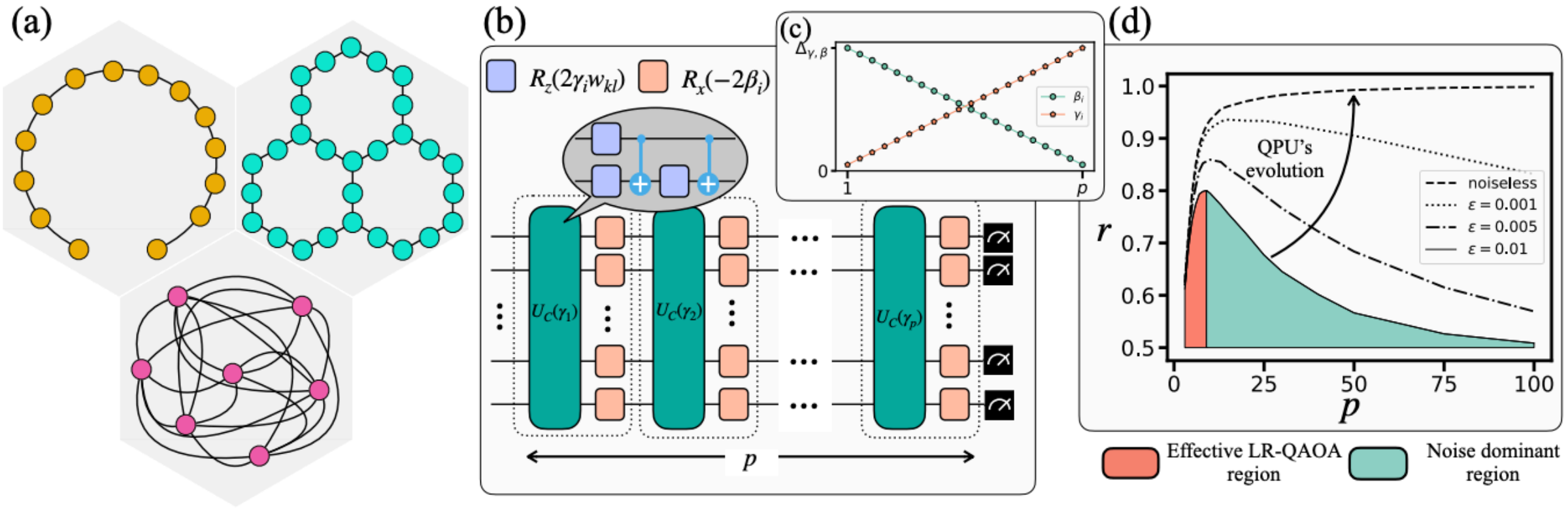
ERROR PER LAYERED GATE (EPLG)



- Applied to a rectangular circuit
- Gives a holistic sense of the QPU performance
- It is only given by IBM
- It is scalable performance benchmark.
- Lose information about connected layers

https://arxiv.org/pdf/2311.05933

The Linear Ramp Quantum Approximate Optimization Algorithm (LR-QAOA)

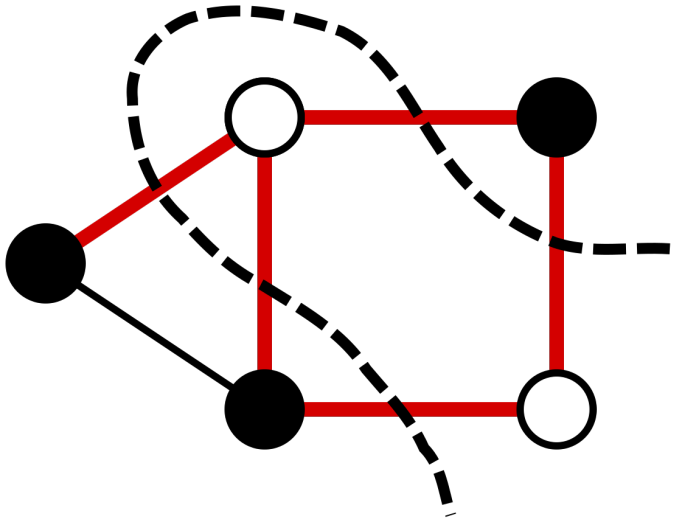


(a) Graph topologies
(b) QAOA algorithm
(c) Linear ramp protocol for QAOA
(d) Expected performance with noise

We applied this benchmark methodology to ~~19~~21 different QPUs from ~~5~~6 different vendors, IQM, IBM, Rigetti, IonQ, Quantinuum, and OriginQ using 5 to 156 qubits and up to p=10,000.

The problem behind LR-QAOA

The weighted maxcut (WMC) problem involves determining the partition of the vertices in an undirected graph so that the total weight of the edges between the two sets is maximized.



https://en.wikipedia.org/wiki/Maximum_cut

$$G = (V, E)$$

$$H_C = \sum_{(i,j)\in E} w_{ij}\sigma_i\sigma_j, \qquad \text{Hamiltonian}$$

$$\sigma_i \in \{-1, 1\} \qquad \text{Variables}$$

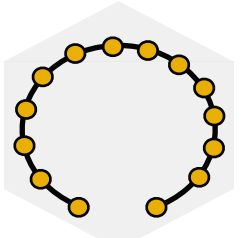
$$r = \frac{\sum_{k=1}^{n} H_C(s_k)/n}{C(s^*)} \qquad \text{Approximation ratio}$$

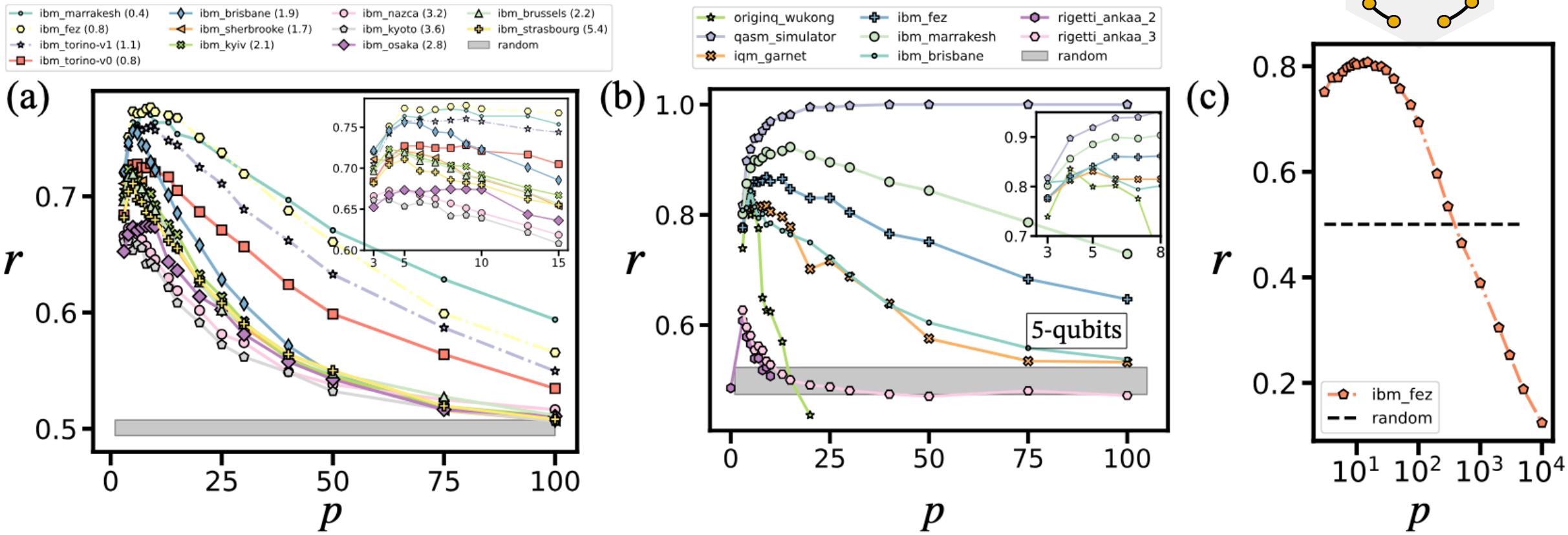
s_k sample solution

s^* optimal solution

n Samples

LR-QAOA on a 1D-Chain graph



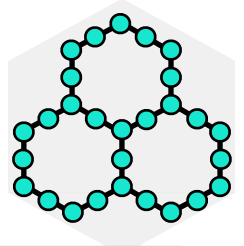


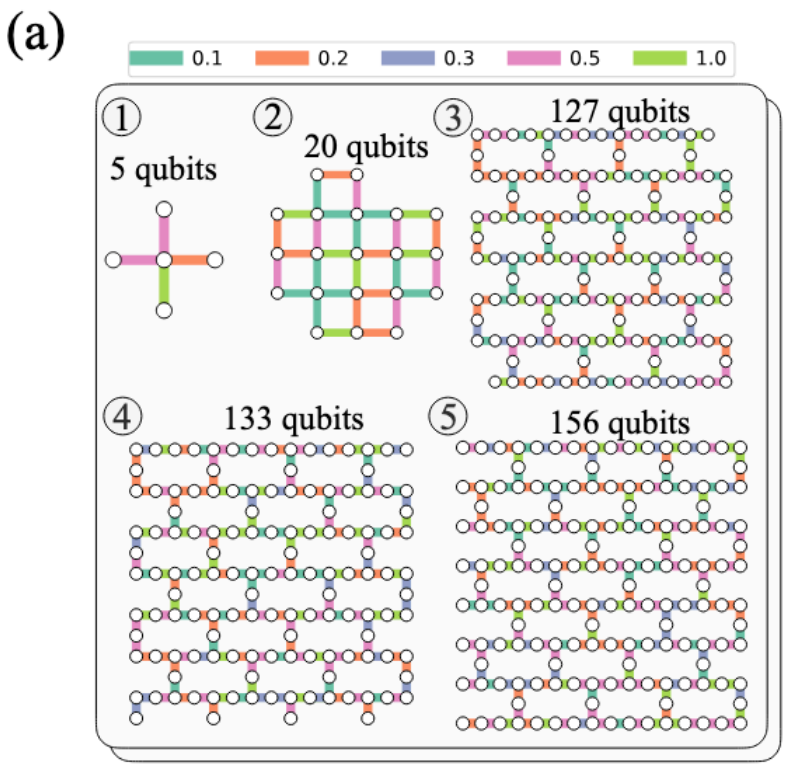
(a)Performance on a 100 1D-Chain IBM devices

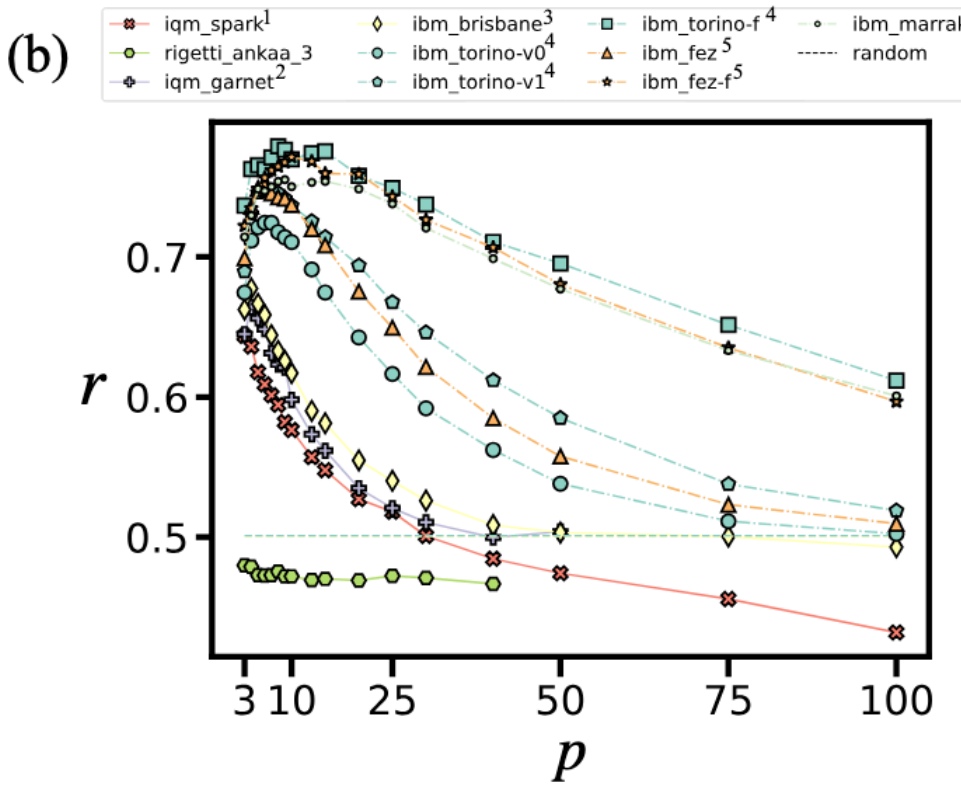
(b) Performance on the best 5 qubits from different superconductive-based QPUs.

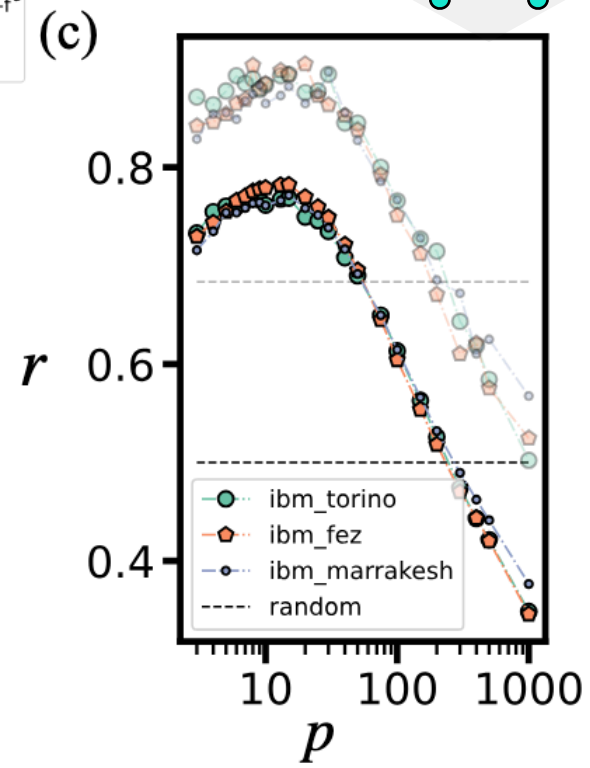
(c) Performance of ibm_fez over a 100 qubit line with up to p=10.000 layers

LR-QAOA on a Native layout graph



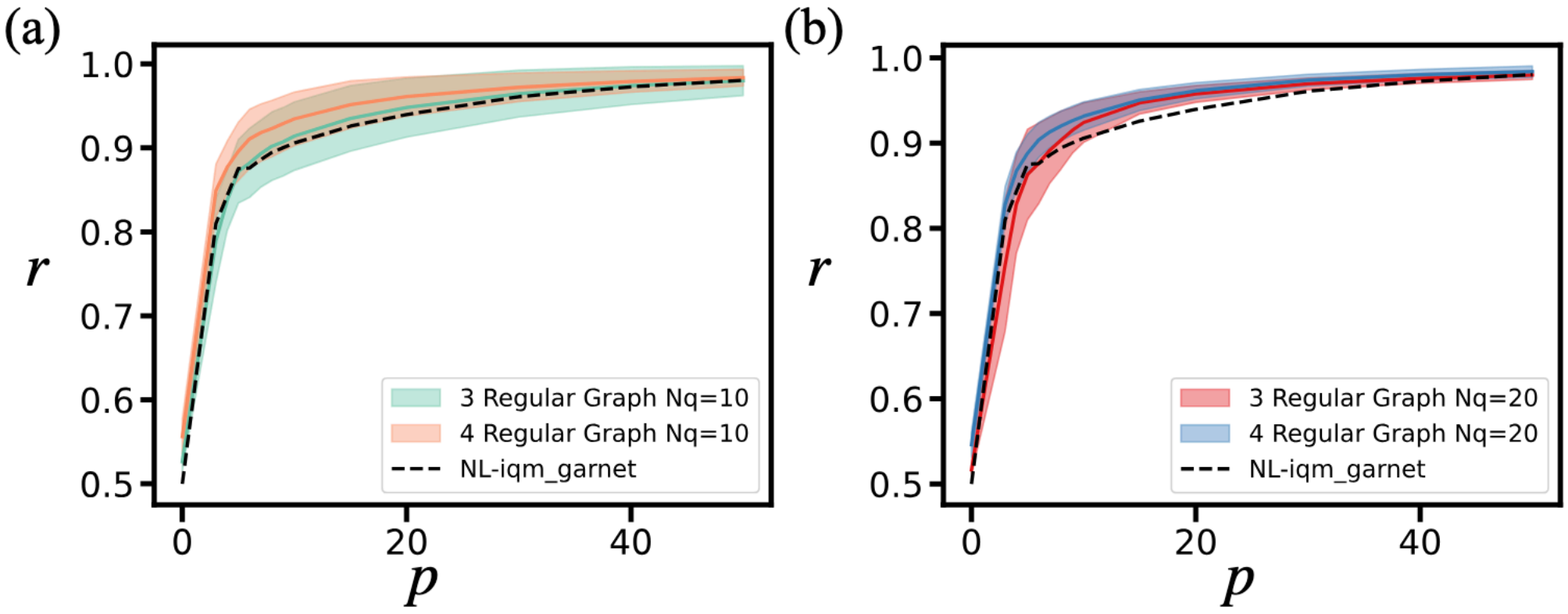






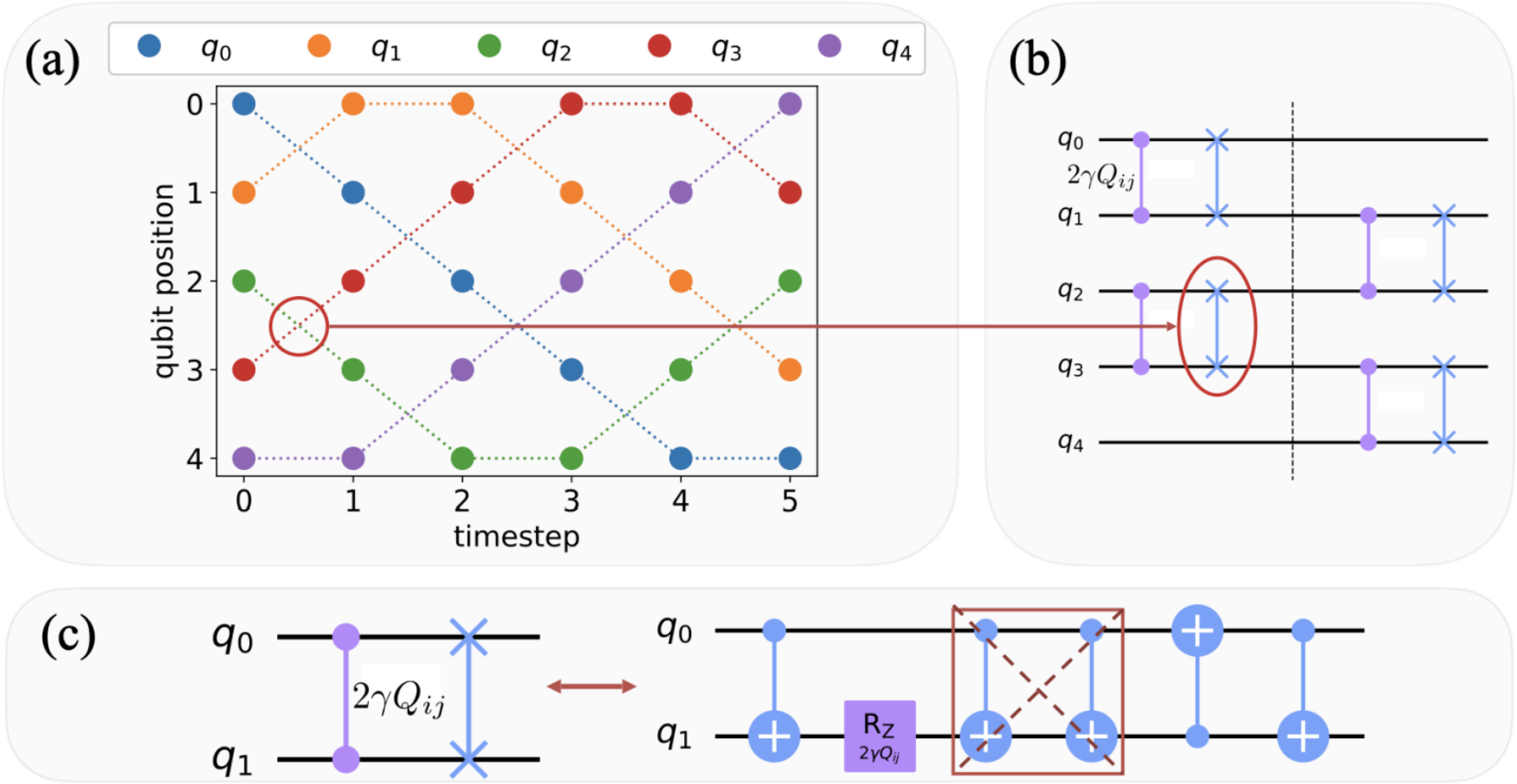
(a) Different QPUs topologies
(b) Performance on different devices
(c) Performance of IBM Heron devices using fractional gates up to p=1000 layers.

Does the connectivity in the native layout graph matter?



Regular graph experiments do not show an appreciable dependence of the approximation ratio performance on the number of qubits or connectivity.

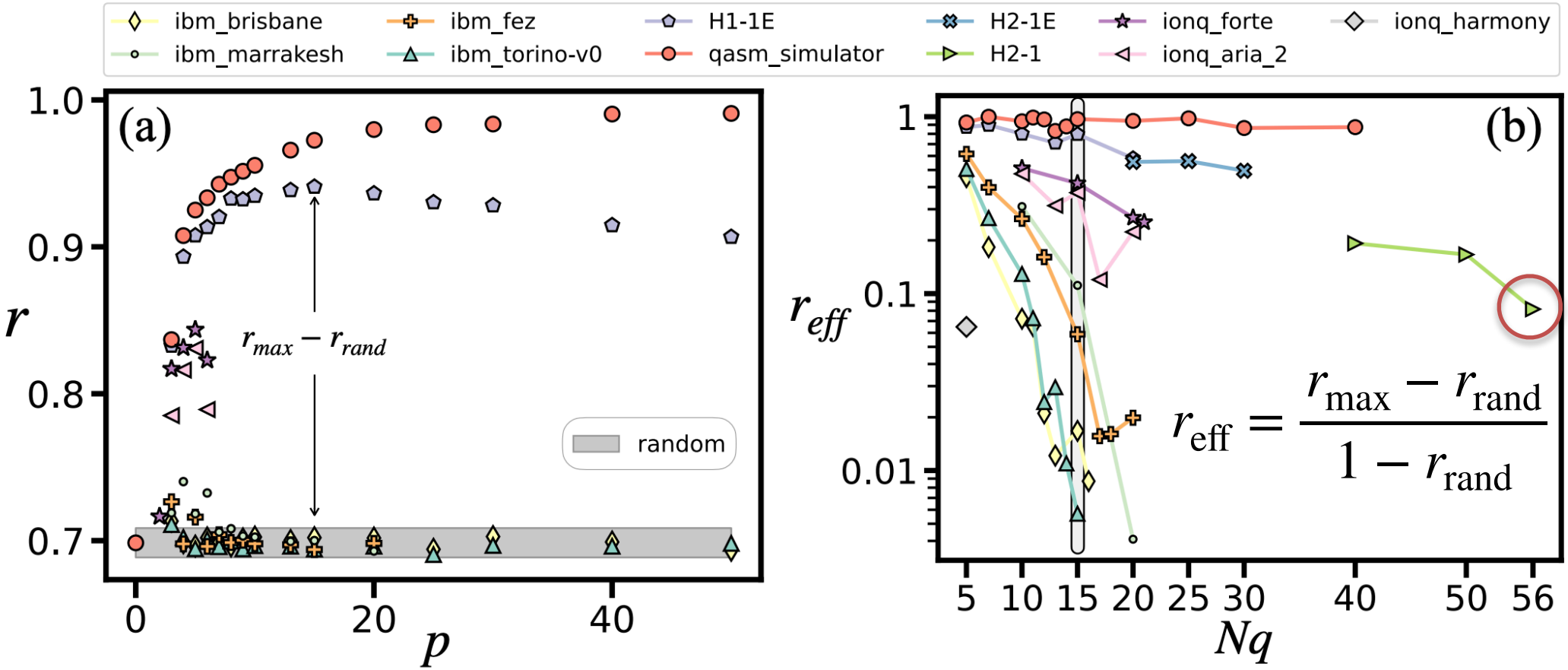
From a fixed layout to a fully connected

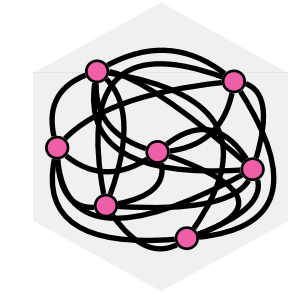


Using a swap strategy we can convert a 1D-Chain graph into a fully connected graph.

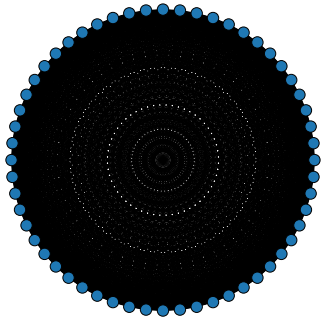
We need 3 times more 2-qubit gates to implement this protocol.

LR-QAOA on Fully Connected (FC) problems





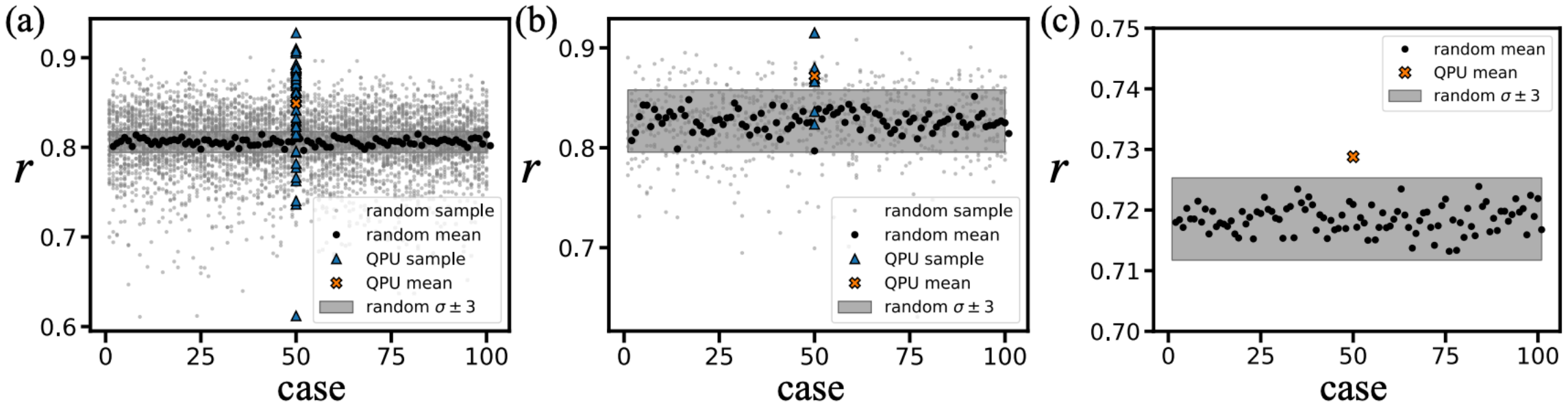
1540 edges



(a) FC for a 15-qubit Weighted MaxCut problem
(b) Effective approximation ratio

Distinguishing successful results

To certify if the result of a QPU is still meaningful, we compare the approximation ratio for the LR-QAOA WMC problem given by the samples of the QPU to those coming from a random sampler.



(a) H2-1 50-qubit, 50 samples, and p=4, (b) 56-qubit, 7 samples, and p=3, (c) ibm_fez 20-qubit, 1000 samples, and p=3.

Conclusions

- We holistically benchmarked 21 QPUs from 6 vendors using LR-QAOA, evaluating their performance on different graph topologies and testing scalability in qubit count and circuit depth.
- IBM QPUs show significant improvements from Eagle to Heron generations, while IonQ and Quantinuum maintain performance through generations and offer better gate fidelity but suffer from slow execution times.
- Our results highlight key bottlenecks in quantum hardware, emphasizing the need for advancements in circuit depth, execution speed, and gate fidelity to support large-scale quantum algorithms.