## Benchmarking Quantum Processor Performance at Scale

David C. McKay,\* Ian Hincks, Emily J. Pritchett, Malcolm Carroll, Luke C. G. Govia, and Seth T. Merkel *IBM Quantum* (Dated: November 13, 2023)

As quantum processors grow, new performance benchmarks are required to capture the full quality of the devices at scale. While quantum volume is an excellent benchmark, it focuses on the highest quality subset of the device and so is unable to indicate the average performance over a large number of connected qubits. Furthermore, it is a discrete pass/fail and so is not reflective of continuous improvements in hardware nor does it provide quantitative direction to large-scale algorithms. For example, there may be value in error mitigated Hamiltonian simulation at scale with devices unable to pass strict quantum volume tests. Here we discuss a scalable benchmark which measures the fidelity of a connecting set of two-qubit gates over N qubits by measuring gate errors using simultaneous direct randomized benchmarking in disjoint layers. Our layer fidelity can be easily related to algorithmic run time, via  $\gamma$  defined in Ref. [1] that can be used to estimate the number of circuits required for error mitigation. The protocol is efficient and obtains all the pair rates in the layered structure. Compared to regular (isolated) RB this approach is sensitive to crosstalk. As an example we measure a N=80 (100) qubit layer fidelity on a 127 qubit fixed-coupling "Eagle" processor (ibm\_sherbrooke) of 0.26(0.19) and on the 133 qubit tunable-coupling "Heron" processor (ibm\_montecarlo) of 0.61(0.26). This can easily be expressed as a layer size independent quantity, error per layered gate (EPLG), which is here  $1.7 \times 10^{-2} (1.7 \times 10^{-2})$  for ibm\_sherbrooke and  $6.2 \times 10^{-3} (1.2 \times 10^{-2})$  for ibm\_montecarlo.

The development of quantum benchmarks enables improvements to be tracked across devices and technologies so that reasonable inferences on performance can be made. In Ref. [2], some properties of quantum benchmarks were discussed, and that a suite of benchmarks should be designed to address quality, speed and scale, altogether describing performance. There are few suggested speed benchmarks besides CLOPS [3]; however, for quality and scale there are many proposals. Generally, the quality is signified by having high fidelity gates (the underlying operations of the device) over a large set of connected qubits with low crosstalk. The size of the set is a benchmark of scale. Such quality can be measured discretely, by individually benchmarking the gate, or holistically, e.g., by running large representative circuits with well known outputs.

Individual gate quality is typically measured by variants of randomized benchmarking [4, 5] (RB). For RB, one selects a random sequence of Clifford gates, constructs a circuit by appending the inverse of the sequence (also a Clifford), decomposes this circuit into the native gate set of a device, and then runs the circuit on said device. The decay of the measured polarization (of any Pauli-Z operator) versus sequence length averaged over many random sequences is straightforwardly related to the average gate error. Because these sequence lengths can be very deep, small errors can be measured that are not dependent on state-preparation and measurement (unlike tomography). Measured in this way, we obtain fine-grained information about the device since we have error rates on each discrete gate element. However, important features of the noise can be missed depending on

the way these are measured. Specifically, in a connected device, if we measure isolated two-qubit (2Q) gate pairs using RB, we potentially overlook crosstalk terms. This issue was addressed in the simultaneous RB protocol [6], yet there are still ambiguities in the implementation.

Conversely, running test algorithms/structured circuits can give a holistic view of gate quality; however, it is very specific to the type of circuits selected. There have been proposed families of circuits as benchmarks [7–16]; however, it remains an open question how to connect performance on one such benchmark to another. As such, benchmarks based on randomized circuits, such as quantum volume (QV) [17], cross-entropy benchmarking (XEB) [18], mirror RB [19], and inverse-free (binary) RB [20], are often believed to give a better overview of average performance. In particular, QV is a stringent test of the device which is defined as  $QV = 2^N$  when some subset of N qubits can pass the QV test – to measure a heavy output probability greater than 2/3<sup>rd</sup> for circuits with N random, all-to-all connected, SU(4) layers. It is straightforward to compare QV across different qubit technologies, and its performance has been linked to the performance of quantum error correcting codes [21].

However, as with all benchmarks, there are limits to QV. For one, it reports on the performance of the best subset of qubits on a device; it is a "high-flier" benchmark. For devices with more qubits than  $\log_2(\mathrm{QV})$ ,  $\mathrm{QV}$  is not a good representative number of overall quality. For example, in superconducting qubits the largest quantum volume is 512 (9 qubits) [22] and in ion traps 524288 (19 qubits) [23]. Yet, there are devices being constructed at scales far beyond these thresholds and so  $\mathrm{QV}$  is not capturing quality across the full scale of the device. Secondly,  $\mathrm{QV}$  (and also XEB) requires classical computation of the circuits, and so is limited to scales where that is tractable

<sup>\*</sup> dcmckay@us.ibm.com

– generally thought to be about 50 qubits and 50 layers of gates (see recent review Ref. [24] and 53 × 20 simulation on advanced high-performance computing (HPC) [25]). And thirdly, as a discrete benchmark, QV does not indicate continuous changes in gate improvement. Finally, QV measures a specific type of unstructured square circuit; however, many near term algorithms, such as the variational quantum eigensolver (VQE), quantum approximate optimization algorithm (QAOA), and Trotterized dynamics (see, e.g., Ref [26] for a review), are based on the idea of a repetitive layer of gates. Similarly, quantum error correction (QEC) relies on a repetitive structure of the application of parallel gates and measurements to perform code checks and detect errors (see, e.g., Ref. [27] for a QEC inspired benchmark).

Layered circuits lend themselves well to applying the techniques of error migitation [1, 28]. Error mitigation is a post processing technique that makes a tradeoff with speed to improve quality, i.e., by running more instances of the circuit with different noise profiles to purify the final result. Therefore, there are compelling reasons to provide a benchmark that spans across an entire device via layered circuits and which reveals continuous information as a complement to QV. While XEB, mirror RB and binary RB can probe layered circuits, they require high-weight measurements that do not reveal information about individual gates. Furthermore, XEB has similar classical computational limitations as QV and the output fidelity can be optimized over any N-qubit unitary in each layer. This adds flexibility to XEB, but makes device to device and application to application comparisons difficult.

To address these points, we propose an alternative benchmark called layer fidelity (LF), which combines the ideas of simultaneous [6] and direct [29] randomized benchmarking and is summarized graphically in Fig. 1. For a given fully connected set of 2Q gates, we partition them into M layers where the 2Q gates are disjoint. When in disjoint layers, we can construct simultaneous direct randomized benchmarking sequences for these gates with alignment barriers and measure individual 1Q and 2Q fidelities. From these disjoint fidelities we can use the product to estimate the full layer fidelity over N qubits. Given this measurement we have enough information to estimate the layer fidelity of all embedded layers of size  $\langle N \rangle$ . To normalize to a size-independent quantity, we introduce error per layered gate (EPLG),  $EPLG = 1 - LF^{1/n_{2Q}}$  where  $n_{2Q}$  is the number of twoqubit gates (typically N-1 for a linear chain of qubits), which is representative of the process error of a gate in these layered circuits. A similar quantity, the dressed two-qubit pauli error (measured from XEB), was defined in Ref. [30]. Lending support to the LF, Ref. [30] shows a threshold between a quantity similar to LF and the ability to classically simulate random circuits with layered structure.

We will discuss the full algorithm in § I and show data on two IBM devices (127 qubit and 133 qubit) in § II. In contrast to other protocols that Pauli-twirl a repeated layer [1, 31–34], the procedure for calculating LF requires fewer circuits. However, we can still relate LF to a mitigation metric under most conditions,  $\gamma = 1/LF^2$ ;  $\gamma$  links noise to the number of probabalistic error cancellation circuits [1] required for a depth  $\delta$  circuit,  $O(\gamma^{2\delta})$  [35]. We show data comparing LF and  $\gamma$  in § A and further discussion of the bounds is in § E. LF is similarly linked to the quantity measured by mirror RB [19]; we show data comparing LF and mirror RB in § A, and we compare simulations of RB, LF and mirror in § B.

### I. LAYER FIDELITY PROTOCOL

An overview of the protocol is visualized in Fig. 1, and here we outline the steps of the protocol,

- 1. Select a set of N qubits  $\{q_i\}$  with a connected set of Clifford two qubit gates  $U = \{U_{ij}\}$  (e.g., CNOT) such that the set of two qubit gates plus arbitrary single-qubit gates define a universal gate set over  $\{q_i\}$ . Part (a) of Fig. 1.
- 2. Split the full layer into M disjoint layers with  $\{U_{ij}\}_m$  such that  $U = \sum_m^M \{U_{ij}\}_m$  where  $\{U_{ij}\}_m$  have no overlapping qubits. The set of idle qubits are  $\{q_i\}_m$ . Example disjoint layers shown in (b) and (c) of Fig. 1.
- 3. Measure the errors on  $\{U_{ij}\}_m$  and  $\{q_i\}_m$  in the disjoint layers using simultaneous direct randomized benchmarking sequences, (d) of Fig. 1.
- 4. From each measured decay we obtain a process fidelity  $F_i = \frac{1+(d^2-1)\alpha}{d^2}$  where d is the dimension of the decay space (d=2 for 1Q, d=4 for 2Q) and  $\alpha$  is the RB decay rate. The layer fidelity per disjoint layer is

$$LF_m = \prod_j F_{j,m},\tag{1}$$

and the full layer fidelity is

$$LF = \prod_{m}^{M} LF_{m}.$$
 (2)

We define a normalized quantity, the error per layered gate,

$$EPLG = 1 - LF^{1/n_{2q}}, (3)$$

where  $n_{2q}$  is the number of 2Q gates in all the layers, e.g., N-1 for the minimal set of connected gates

There are a few considerations for the protocol:

•  $\{U_{ij}\}$  are typical two-qubit gates such as CNOT, CZ, and iSWAP and variations from those that differ by single qubit gates, e.g., ECR  $(e^{-i\frac{\pi}{4}ZX})$ .

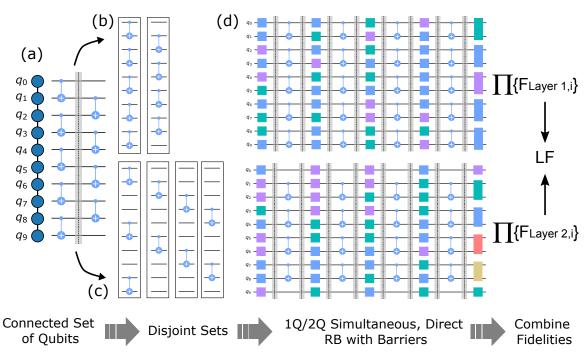


FIG. 1: (a) Here we consider a linear chain of qubits with nearest neighbor coupling for which a connecting set of gates is comprised of a disjoint layer of gates starting on qubit 0 followed by a disjoint layer of gates starting on qubit 1. The disjoint layers (b) can either be the maximally simultaneous sets, but could alternatively be a more sparse set (c) split into more disjoint layers. (d) For the disjoint layer set of (b) this requires two simultaneous direct RB experiments here shown for depth l=4 with the last layer the inverses in each disjoint space. We measure decay curves as a function of l and fit to extract the fidelities, which are then multiplied together as given by Eqn. 2 to obtain the layer fidelity.

- There is no unique decomposition of disjoint layers, but the error of all qubits must be measured, including idle qubits, i.e., qubits without a two-qubit gate in that disjoint layer. We show some data comparing different disjoint layer decompositions in § A.
- A requirement of the protocol is that barriers must be enforced at the layer of two-qubit gates (all gates before the barrier must complete before the circuit can proceed). That is, we apply a set of randomizing single qubit Clifford gates on all qubits, a barrier across all qubits, the layer of disjoint two-qubit gates, then another barrier, and repeat this l times. At the end, we invert each disjoint set, and measure the ground state population of each akin to simultaneous randomized benchmarking [6]. Since the sub-layers are disjoint there is no mixing and we get well-defined decay curves of the ground state population versus l. The use of barriers keeps each layer consistent with how it would appear in the full (i.e. non-disjoint) layer.
- Dynamic decoupling is allowed.
- *LF* makes a Markovianity assumption and is a benchmark insensitive to state preparation and

- measurement (SPAM) errors; however, the contribution of measurement error can be trivially added by taking the product of the measurement assignment fidelities.
- The layer fidelity of a device for N qubits is defined as the maximum layer fidelity measured on the device (practical considerations are discussed in § II).

The goal of the protocol is to measure the fidelity of the full layer defined in the first step of the protocol, for example, (a) of Fig. 1. Formally, the fidelity of that layer is the trace of the Pauli Transfer Matrix (PTM) between the noisy experimental map and the inverse ideal map,

$$F = \text{Tr}(R_{\text{ideal}}^{-1} R_{\text{exp}}) / d^2, \tag{4}$$

where  $R_{ij} = \text{Tr}(P_i\Lambda[P_j])/d$ ,  $P_i$  are the Pauli matrices, and  $\Lambda$  is the process map for the layer. In the limit of no crosstalk, Eqn. 1 is exact, but Eqn. 2 is not because the product of traces is not the trace of the product; however, for small errors this is a good approximation (§ D) and is a lower bound. This is also true for the layer fidelity: after j repetitions of the layer,  $LF^j$  is approximately the true fidelity until  $LF^j$  gets small. With crosstalk, Eqn. 2 and Eqn. 4 are not identical and for specific crosstalk terms (see § C) the layer fidelity will be a lower bound (the crosstalk error terms are double

counted). Because a general treatment of all cases is not possible, we turn to numerics (§ B) with various noise models. We compare layer fidelity to theory (Eqn. 4) and to the fidelity measured from mirror RB [19], which is a protocol to measure the layer fidelity by building a circuit of l/2 layers to which the reverse circuit is appended, and the polarization of the output is measured versus l.

The advantage of mirror circuit RB is that it does capture all crosstalk terms; however with two distinct disadvantages compared to layer. First, with layer fidelity we obtain more information: a detailed set of error rates for each  $\{U_{ij}\}_m$  and  $\{q_i\}_m$ . Second, the signal to noise of layer fidelity is higher since we are measuring the individual error rates versus the error rate of the entire layer (a weight-n measurement). Any protocol that requires the estimation of high-weight observables, which LF avoids, is unscalable because, with enough qubits, the signal will be unmeasurably small even for short protocol depths. Overall, the numerics support the assertion that layer fidelity is capturing the majority of the crosstalk terms for realistic noise models, and layer fidelity and mirror RB agree well in an experimental test (§ A). By fitting all decay terms [36, 37] available to us in our layer fidelity benchmark, we could properly better account for some of these crosstalk terms. However, the added complexity, exposure to measurement errors, and loss of signalto-noise need more careful consideration. As an aside, the advantages shown here for layer fidelity over mirror benchmarking should also hold for binary RB [20] as well.

As mentioned, one advantage of the layer fidelity protocol is that it gives access to the discrete fidelities of the underlying gates. One utility of this is that we can easily measure the layer fidelity on smaller subsets of the measured set  $\{q_i\}$ . We can simply calculate the smaller subset by omitting qubits outside the set in the calculation of LF, i.e., change the indices of Eqn. 1. A gate may extend outside the new subset, so we assume that the gate fidelity is shared equally between those subsets and calculate the fidelity of that qubit in the layer as  $F^{1/2}$ . In most geometries the layer fidelity is optimally measured on a long 1D line of qubits (chain of qubits) as this only requires two disjoint layers, i.e., the gates first starting at  $Q_0$  (even set) and then at  $Q_1$  (odd set) as shown in (b) of Fig. 1. When defined on a line, measuring subsets is particularly straightforward as it is a sliding window of qubits inside the larger 1D chain. Although it is not guaranteed that we find the optimal value of layer fidelity using this subspace method, this can be used as a lower bound. While the line is the densest application of gates possible, based on the definition set forth, the gates can be measured over more disjoint layers, so long as idle qubit errors are accounted for; we show data in § A that splitting over more layers is worse due to the increased duration. In certain geometries, such as a star, more disjoint layers will necessarily be required: this problem is equivalent to constructing an edge-coloring of the coupling graph, and Vizing's theorem guarantees that we

require no more distinct layers than the degree of the graph plus one.

Layer fidelity can be easily related to other metrics which quantify the error models on a layer, such as  $\gamma$  [1], which is defined as

$$\gamma = e^{\sum_{k \in K} 2\lambda_k},\tag{5}$$

where  $\lambda_k > 0$  are the Pauli generator terms in the Lindblad model of the Pauli-twirled noise. While in general all Pauli-twirled error terms exist in Eqn. 5, approximations are made to make the calculation tractable, for example, in Ref. [1] the terms are truncated to all physical connections in the device, and Pauli benchmarking is required to learn them. Even still, this requires many more circuits than are required to measure layer fidelity. The two quantities are easily related for well-behaved noise; for depolarizing noise  $\gamma$  is given by,

$$\gamma_D = \prod_i \left( \frac{16 \times F_i - 1}{15} \right)^{-15/8}$$

$$= \prod_i \alpha_i^{-15/8} \tag{6}$$

which in the limit of  $\alpha$  close to 1 is,

$$\gamma = \frac{1}{LF^2}. (7)$$

We derive this and discuss the bounds in § E, and we show some data comparing  $\gamma$  and LF in § A. Note that this  $\gamma$  is defined over the disjoint layers used to measure layer fidelity, which are at least depth 2. To estimate  $\gamma$  on a  $N_{\gamma}$  qubit  $(N_{\gamma} \text{ even})$ , depth 1 layer  $(\delta = 1)$ , we can use EPLG (Eqn. 3),

$$\gamma_{\delta=1} = (1 - \text{EPLG})^{-N_{\gamma}}, \tag{8}$$

and

$$\bar{\gamma}_{\delta=1} = \gamma_{\delta=1}^{2/N_{\gamma}} = (1 - \text{EPLG})^{-2}.$$
 (9)

The accuracy of this estimate will depend on the the layer structures being similar between the layer for  $\bar{\gamma}$  and the layer used to measure LF/EPLG. If  $\gamma_{\delta=1}$  is defined on the same disjoint layer as used for the layer fidelity measurement, then it can also be calculated directly from the disjoint layer fidelity Eqn. 1.

## II. DATA

As mentioned, the layer fidelity for a subset of N qubits on a device  $(LF_N)$  is defined as the maximum layer fidelity over all K (N-qubit) subsets. Practically it will be impossible to measure all sets on a large device, for example the number of length 100 chains on a 127Q heavy hex device such as ibm\_sherbrooke is 313,980. Therefore, in practice we need heuristic methods to measure

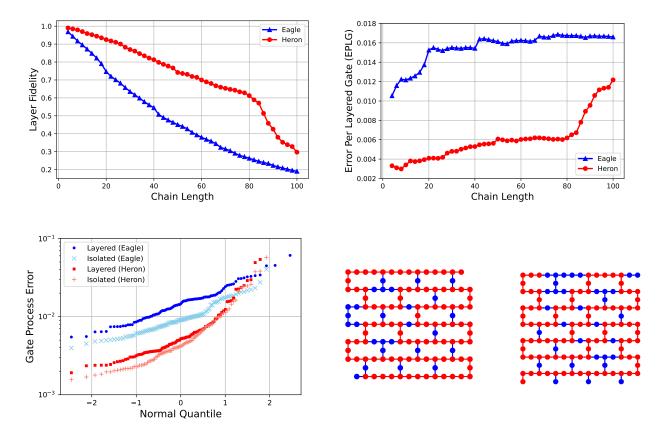


FIG. 2: (Top Left) Layer fidelity for the 127 qubit ibm\_sherbrooke "Eagle" processor (blue triangles) and the 133 qubit ibm\_montecarlo "Heron" processor (red circles) taken using the procedure outlined in the main text for various chain lengths up to 100 qubits. (Top Right) The same data converted to error per layered gate (EPLG). (Bottom Left) Quantile plot of the individual gate errors measured from the best 100 qubit chain from simultaneous direct RB ("layered") versus the backend reported gate errors ("isolated") on the same chain. Errors are reported as process error ( $\epsilon_p$ ) as opposed to average gate error ( $\epsilon_g$ ) where  $\epsilon_p = \frac{d+1}{d}\epsilon_g$ . Both devices have among the lowest gate error measured on a superconducting device, noting the minimum isolated gate error (process error) on ibm\_sherbrooke (Eagle) of  $3.2(4.0) \times 10^{-3}$  and on ibm\_montecarlo (Heron) of  $1.2(1.6) \times 10^{-3}$  (Bottom Right) The 100 qubit chain (red) overlaid on the ibm\_sherbrooke (left) and ibm\_montecarlo (right) device layout schematics.

the optimal layer fidelity. Initial estimates of the layer fidelity can be made with the isolated two-qubit fidelities [38] and from there candidate sets can be measured. One of the bigger considerations here is that the layer fidelity imposes a fixed length on all gates of the disjoint layer equal to the longest gate (see the simulations in the appendix § B), and so the estimates from isolated RB fidelities must take that into consideration. Typically, this is done by omitting edges of the graph with gates that are much longer than the average. We use a heuristic protocol given by the following procedure:

1. Assuming a list of gate errors measured from isolated RB is available, calculate the layer fidelity for each  $N_{max}$  qubit linear string, where  $N_{max}$  is selected to be at least the length of the longest desired chain. In this step long gates may be omitted from the graph as they are known to make the layer fidelity much worse. Find the set with the highest predicted LF (set 1), then find the set with the least overlap with set 1 and the highest predicted LF of that subset (set 2). Repeat this again to find a third set.

- 2. Measure the errors from simultaneous direct RB (described in Fig. 1) in those 3 sets (at least 6 disjoint layers for 1D chains) and calculate the LF (Eqn. 2) from the measured data for each  $N < N_{max}$  by looking at subchains within the sets.
- 3. For each value of N take the largest LF from all the subchains measured. For example, if  $N_{max} = 100$  and N = 50 then there are 150 possible sub-chains.
- 4. Plot LF vs N, convert to error per layered gate (EPLG) as EPLG =  $1 LF^{1/n_{2q}}$  where  $n_{2q}$  is the number of 2Q gates.
- 5. Since this covers a heuristic number of chains, more chains at different lengths can be measured "ad-

hoc" and if the LF of those chains is larger, they will supplant the previously measured values.

We show typical data taken on a 127 qubit "Eagle" processor ibm\_sherbrooke (native two-qubit CX gate using cross-resonance) and 133 qubit "Heron" processor ibm\_montecarlo (native two-qubit CZ gate using tunablecoupler actuation) in Fig. 2. To measure the fidelities we perform the simultaneous direct RB sequences described previously with 300 shots per circuit, 6 randomizations and l = [1, 10, 20, 30, 40, 60, 80, 100, 125, 150, 200,400] (Eagle) and l = [1, 10, 20, 30, 40, 60, 80, 100,125, 150, 200, 400, 750] (Heron). For ibm\_sherbrooke most gate lengths are 533 ns, but as described in the heuristics for picking the gate sets, three edges with gate lengths >700 ns were removed from consideration. For ibm\_montecarlo the gate lengths were about an equal mix of 84 ns and 104 ns and none were removed from consideration due to gate length. All circuits were generated in Qiskit and run through the IBM Quantum cloud interface. In the top plot we show the layer fidelity and the error per layered gate as a function of chain length. The chain of qubits on the ibm\_sherbrooke and ibm\_montecarlo devices for the N = 100 data is shown in red on the bottom right plot.

One of the advantages of this layer fidelity measurement is the access to individual gate errors which can be used for further analysis or fidelity estimates. In particular, we can perform a comparison between isolated RB and the errors from layer fidelity RB which can be a proxy for crosstalk errors. We note here that the isolated RB data is, in fact, a variant of simultaneous RB where there is a distance of at least one idle qubit between all two qubit gate pairs and there are no barriers. As we show in simulations in the appendix § B isolated RB trivially eliminates some crosstalk, such as always on ZZ between pairs of qubits for fixed coupling architectures. The data bears this out as the middle plot of Fig. 2 shows a distinct increase in the error per gate on the "Eagle" processor when run in layers versus isolated RB mode. Conversely, these errors are greatly alleviated in the "Heron" processor since the coupling between neighboring qubits can be turned off when not required for two-qubit gate operation.

### III. CONCLUSIONS

In this manuscript we discussed a benchmark for quantum processors at scale - the layer fidelity. The layer fidelity follows naturally from standard randomized benchmarking procedures, is crosstalk aware, fast to measure over a large number of qubits, has high signal to noise and gives fine-grained information. We demonstrated the key components of the layer fidelity metric with measurements on the 127 qubit Eagle processor ibm\_sherbrooke and 133 qubit Heron processor ibm\_montecarlo. Using simulation and data we showed (§ A and § B) that there is good agreement with mirror randomized benchmarking - a complementary technique for measuring layers - over a number of error models. The layer fidelity links easily with other methods of characterizing layers, such as Pauli learning for  $\gamma$ . We leave a few open questions outside the scope of this manuscript, such as whether more advanced data fitting can improve agreement with the exact layer fidelity, the predictive power of LF with differently structured circuits, the limits of twirling in LF, and how to extend to layered circuits with mid-circuit measurements. Finally, we note that, like all benchmarks layer fidelity should be considered as one piece of information towards full device characterization.

### ACKNOWLEDGMENTS

The authors would like to thank Andrew Wack for helpful discussions, Kevin Krsulich for Qiskit help, Samantha Barron for help with the Pauli learning, and James Wootton and Blake Johnson for manuscript comments. Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-21-1-0002. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

E. v. d. Berg, Z. K. Minev, A. Kandala, and K. Temme, Probabilistic error cancellation with sparse pauli-lindblad models on noisy quantum processors, arXiv preprint arXiv:2201.09866 (2022).

<sup>[2]</sup> M. Amico, H. Zhang, P. Jurcevic, L. Bishop, P. Nation, A. Wack, and D. C. McKay, Defining standard strategies for quantum benchmarks, arXiv preprint arXiv:2303.02108 (2023).

<sup>[3]</sup> A. Wack, H. Paik, A. Javadi-Abhari, P. Jurcevic, I. Faro, J. M. Gambetta, and B. R. Johnson, Quality, speed, and scale: three key attributes to measure the perfor-

mance of near-term quantum computers, arXiv preprint arXiv:2110.14108 (2021).

<sup>[4]</sup> E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, Phys. Rev. A 85, 042311 (2012).

<sup>[5]</sup> J. Helsen, I. Roth, E. Onorati, A. Werner, and J. Eisert, General framework for randomized benchmarking, PRX Quantum 3, 020357 (2022).

<sup>[6]</sup> J. M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, J. A. Smolin, J. M. Chow, C. A. Ryan, C. Rigetti, S. Poletto, T. A. Ohki, M. B. Ketchen, and

- M. Steffen, Characterization of addressability by simultaneous randomized benchmarking, Phys. Rev. Lett. **109**, 240504 (2012).
- [7] T. Lubinski, S. Johri, P. Varosy, J. Coleman, L. Zhao, J. Necaise, C. H. Baldwin, K. Mayer, and T. Proctor, Application-oriented performance benchmarks for quantum computing, arXiv preprint arXiv:2110.03137 (2021).
- [8] K. Mesman, Z. Al-Ars, and M. Möller, Qpack: Quantum approximate optimization algorithms as universal benchmark for quantum computers, arXiv preprint arXiv:2103.17193 (2021).
- [9] J. R. Finžgar, P. Ross, L. Hölscher, J. Klepsch, and A. Luckow, Quark: A framework for quantum computing application benchmarking, in 2022 IEEE International Conference on Quantum Computing and Engineering (QCE) (IEEE, 2022) pp. 226–237.
- [10] T. Tomesh, P. Gokhale, V. Omole, G. S. Ravi, K. N. Smith, J. Viszlai, X.-C. Wu, N. Hardavellas, M. R. Martonosi, and F. T. Chong, Supermarq: A scalable quantum benchmark suite, in 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA) (IEEE, 2022) pp. 587–603.
- [11] V. Zhang and P. D. Nation, Characterizing quantum processors using discrete time crystals, arXiv preprint arXiv:2301.07625 (2023).
- [12] A. Kurlej, S. Alterman, and K. M. Obenland, Benchmarking and analysis of noisy intermediate-scale trapped ion quantum computing architectures, in 2022 IEEE International Conference on Quantum Computing and Engineering (QCE) (IEEE, 2022) pp. 247–258.
- [13] T. Lubinski, C. Coffrin, C. McGeoch, P. Sathe, J. Apanavicius, and D. E. B. Neira, Optimization applications as quantum performance benchmarks, arXiv preprint arXiv:2302.02278 (2023).
- [14] M. Kordzanganeh, M. Buchberger, M. Povolotskii, W. Fischer, A. Kurkin, W. Somogyi, A. Sagingalieva, M. Pflitsch, and A. Melnikov, Benchmarking simulated and physical quantum processing units using quantum and hybrid algorithms, arXiv preprint arXiv:2211.15631 (2022).
- [15] P. S. Mundada, A. Barbosa, S. Maity, T. Stace, T. Merkh, F. Nielson, A. R. Carvalho, M. Hush, M. J. Biercuk, and Y. Baum, Experimental benchmarking of an automated deterministic error suppression workflow for quantum algorithms, arXiv preprint arXiv:2209.06864 (2022).
- [16] A. Li, S. Stein, S. Krishnamoorthy, and J. Ang, Qasmbench: A low-level quantum benchmark suite for nisq evaluation and simulation, ACM Transactions on Quantum Computing 10.1145/3550488 (2022).
- [17] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta, Validating quantum computers using randomized model circuits, Physical Review A 100, 032328 (2019).
- [18] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Characterizing quantum supremacy in nearterm devices, Nature Physics 14, 595 (2018).
- [19] T. Proctor, S. Seritan, K. Rudinger, E. Nielsen, R. Blume-Kohout, and K. Young, Scalable randomized benchmarking of quantum computers using mirror circuits, Phys. Rev. Lett. 129, 150502 (2022).
- [20] J. Hines, D. Hothem, R. Blume-Kohout, B. Whaley, and T. Proctor, Fully scalable randomized benchmarking

- without motion reversal (2023), arXiv:2309.05147 [quant-ph].
- [21] C. H. Baldwin, K. Mayer, N. C. Brown, C. Ryan-Anderson, and D. Hayes, Re-examining the quantum volume test: Ideal distributions, compiler optimizations, confidence intervals, and scalable resource estimations, Quantum 6, 707 (2022).
- [22] J. Gambetta, Qv 512 announcement (2022).
- [23] Quantinuum, Quantinuum h-series quantum computer accelerates through 3 more performance records for quantum volume: 2<sup>1</sup>7, 2<sup>1</sup>8, and 2<sup>1</sup>9 (2023).
- [24] X. Xu, S. Benjamin, J. Sun, X. Yuan, and P. Zhang, A herculean task: Classical simulation of quantum computers (2023).
- [25] Y. Liu, Y. Chen, C. Guo, J. Song, X. Shi, L. Gan, W. Wu, W. Wu, H. Fu, X. Liu, D. Chen, G. Yang, and J. Gao, Validating quantum-supremacy experiments with exact and fast tensor network contraction (2022).
- [26] A. M. Dalzell, S. McArdle, M. Berta, P. Bienias, C.-F. Chen, A. Gilyén, C. T. Hann, M. J. Kastoryano, E. T. Khabiboulline, A. Kubica, G. Salton, S. Wang, and F. G. S. L. Brandão, Quantum algorithms: A survey of applications and end-to-end complexities (2023), arXiv:2310.03011 [quant-ph].
- [27] M. Liepelt, T. Peduzzi, and J. R. Wootton, Enhanced repetition codes for the cross-platform comparison of progress towards fault-tolerance (2023), arXiv:2308.08909 [quant-ph].
- [28] S. Ferracin, A. Hashim, J.-L. Ville, R. Naik, A. Carignan-Dugas, H. Qassim, A. Morvan, D. I. Santiago, I. Siddiqi, and J. J. Wallman, Efficiently improving the performance of noisy quantum computers (2022), arXiv:2201.10672 [quant-ph].
- [29] T. J. Proctor, A. Carignan-Dugas, K. Rudinger, E. Nielsen, R. Blume-Kohout, and K. Young, Direct randomized benchmarking for multiqubit devices, Phys. Rev. Lett. 123, 030503 (2019).
- [30] A. Morvan and et. al, Phase transition in random circuit sampling (2023), arXiv:2304.11119 [quant-ph].
- [31] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, Characterizing large-scale quantum computers via cycle benchmarking, Nat. Comm. 10, 5347 (2019).
- [32] J. Helsen, X. Xue, L. M. K. Vandersypen, and S. Wehner, A new class of efficient randomized benchmarking protocols, npj Quantum Inf 5, 71 (2019).
- [33] S. Kimmel, M. P. da Silva, C. A. Ryan, B. R. Johnson, and T. Ohki, Robust extraction of tomographic information via randomized benchmarking, Phys. Rev. X 4, 011050 (2014).
- [34] A. Carignan-Dugas, D. Dahlen, I. Hincks, E. Ospadov, S. J. Beale, S. Ferracin, J. Skanes-Norman, J. Emerson, and J. J. Wallman, The error reconstruction and compiled calibration of quantum computing cycles (2023).
- [35]  $\delta$  here is the number of repeated full layers which are used to compute  $\gamma$ . If  $\delta$  is the traditionally defined circuit depth,  $\gamma$  is a geometric mean over the disjoint layers and is computed from LF as such.
- [36] D. C. McKay, A. W. Cross, C. J. Wood, and J. M. Gambetta, Correlated randomized benchmarking (2020).
- [37] R. Harper, S. Flammia, and J. Wallman, Efficient learning of quantum noise, Nat. Phys. 16, 1184 (2020).

- [38] P. D. Nation and M. Treinish, Suppressing quantum circuit errors due to system variability, arXiv preprint arXiv:2209.15512 (2022).
- [39] P. Murali, J. M. Baker, A. Javadi-Abhari, F. T. Chong, and M. Martonosi, Noise-adaptive compiler mappings for noisy intermediate-scale quantum computers, in *Proceed*ings of the twenty-fourth international conference on architectural support for programming languages and operating systems (2019) pp. 1015–1029.
- [40] J. Emerson, R. Alicki, and K. Życzkowski, Scalable noise estimation with random unitary operators, Journal of Optics B: Quantum and Semiclassical Optics 7, S347 (2005).
- [41] L. H. Pedersen, N. M. Møller, and K. Mølmer, Fidelity of quantum operations, Physics Letters A 367, 47 (2007).
- [42] K. X. Wei, E. Magesan, I. Lauer, S. Srinivasan, D. F. Bogorin, S. Carnevale, G. A. Keefe, Y. Kim, D. Klaus, W. Landers, N. Sundaresan, C. Wang, E. J. Zhang, M. Steffen, O. E. Dial, D. C. McKay, and A. Kandala, Hamiltonian engineering with multicolor drives for fast entangling gates and quantum crosstalk cancellation, Phys. Rev. Lett. 129, 060501 (2022).
- [43] B. K. Mitchell, R. K. Naik, A. Morvan, A. Hashim, J. M. Kreikebaum, B. Marinelli, W. Lavrijsen, K. Nowrouzi, D. I. Santiago, and I. Siddiqi, Hardware-efficient microwave-activated tunable coupling between superconducting qubits, Phys. Rev. Lett. 127, 200502 (2021).
- [44] D. Greenbaum, Introduction to quantum gate set tomography (2015), arXiv:1509.02921 [quant-ph].
- [45] M. A. Nielsen, A simple formula for the average gate fidelity of a quantum dynamical operation, Physics Letters A 303, 249 (2002).

### Appendix A: Additional Data

Here we provide some additional support for layer fidelity by comparing to the mirror RB protocol [19] which embeds the full layer directly in a mirror circuit. Our specific mirror circuit is comprised of a random 1Q Clifford layer, then the first disjoint layer of two-qubit gates, a second random 1Q Clifford layer, then the second disjoing layer of two-qubit gates. In this way the number of 1Q gates is the same between mirror RB and layer RB when constructing the full layer fidelity. We consider two versions of mirror, one which is a direct mirror (just the forward and reverse circuit) and the second version, more faithful to Ref. [19], includes a random Pauli layer between the forward and reverse circuits. We measure the polarization S as defined in Ref. [19]. Our data comparison is on 20 qubits of the ibm\_peekskill device, which is a 27 qubit fixed-coupling "Falcon" processor and the path is shown in Fig. 3. We perform 10 randomizations and measure 2000 shots. We compare to layer fidelity data from 6 randomization and 300 shots. We can see that the agreement between mirror and layer is quite good and that there is some discrepancy between mirror with and without the Pauli layer. Since our error model in the device is fixed, we investigate the comparison of layer and mirror further with simulations in § B.

Another aspect of the layer protocol is that the choice

of disjoint sets is not unique, as shown in Fig. 1 for 2 versus 4 layers of the chain. Therefore, we take data in the same set of qubits as above splitting the layer fidelity into 2, 4, 6 and 10 disjoint layers. Because of the increased total duration, more disjoint layers leads to lower fidelity. However, this statement will be architecture dependent; there are certain types of crosstalk terms that occur during simultaneous gates that are large enough to offset the longer duration of the circuit, e.g., this was probed on IBM devices in Ref. [39]. In this case, splitting into more disjoint layers is a sensible approach. Another scenario is that the architecture does not allow more than a certain number of simultaneous gates at a time.

In the main text we relate the layer fidelity to a quantity relevant for error mitigation,  $\gamma$ , defined in Eqn. 5. The relation is given in Eqn. 7, and although supported in theory by well behaved noise models (see § E), here we do an experimental comparison on a 16 qubit section of ibm\_peekskill. The layer fidelity data is taken according to the procedure outlined in the main text and the direct  $\gamma$  data is taken according to the procedure in Ref. [1]. For the LF data we take 178 circuits (13 depth points  $\times$  6 randomizations  $\times$  2 disjoint layers) and for the  $\gamma$ data we take 14,000 circuits (14 depth points  $\times$  1000 basis rotations × randomizations). This ratio of circuits demonstrates why layer fidelity is a quick method for estimating gamma. The data is shown in Fig. 4 and the agreement is reasonable; a more comprehensive study including error bars and minimizing time variations of the device properties is left for a future study.

#### Appendix B: Simulations

Here we compare simulations between isolated RB, simultaneous RB, layer fidelity RB, and mirror RB with a variety of error models. The circuits are generated as gates (in the decomposition of X90, X0 [idle], Rz [arbitrary Z rotations and CX gates and then converted to a schedule based on the single-qubit gate being the smallest unit of time; the two-qubit gates are converted into fractional time steps of either 5 or 8 single-qubit gate times. Because the Rz gates are zero time, they are considered their own gate slices, and so a finite time step can consist of the 3 possible gates on each qubit and so for four qubits there are roughly 81 unique four qubit unitaries to construct. We may add coherent error terms to each unitary, e.g., an overrotation or a ZZ crosstalk. We then perform a density matrix simulation, where at each time step the unitary is applied followed by a discrete  $T_1/T_2$  map on each qubit. Breaking the unitary and incoherent evolution into time steps is an implementation of Trotterized simulation of their concurrent evolution. We compare the error extracted from these sequences to the theoretical errors by adding the coherent [40, 41] and

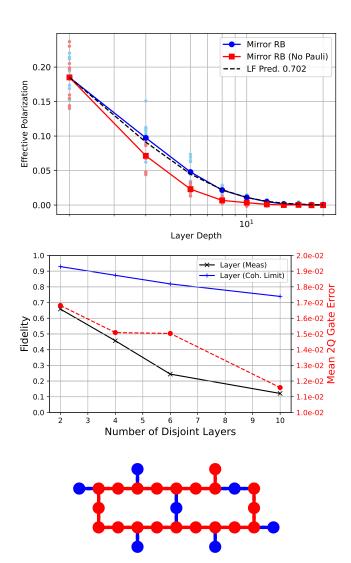


FIG. 3: (Top) Comparing mirror RB data versus layer depth l to the predicted decay from layer fidelity measured on the same set of qubits; for this data LF = 0.702 and so the dashed line is  $0.702^{l}$ . The different mirror RB curves either include (blue) or do not include (red) a random Pauli layer. (Middle) The layer fidelity versus the number of disjoint layers used in the protocol (black). The fidelity decreases as the number of layers increases because the total duration is longer. We estimate this effect by just considering the fidelity decrease due to decoherence (blue). More layers does decrease the mean gate error (dashed, red) due to lower crosstalk, but overall this is not enough to improve because of the increased length. (Bottom) Qubits used on ibm\_peekskill are [23, 24, 25, 22, 19, 16, 14, 11, 8, 5, 3, 2, 1, 4, 7, 10, 12, 15, 18, 17].

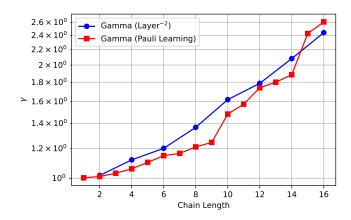


FIG. 4: Comparing  $\gamma$  measured from layer fidelity (blue, circles) and Eqn. 7 to  $\gamma$  measured using Pauli-learning [1] (red, squares). Measured on ibm\_peekskill for the connected set of qubits [19, 22, 25, 24, 23, 21, 18, 15, 12, 13, 14, 11, 8, 5, 3, 2] (even and odd disjoint layers).

incoherent errors.

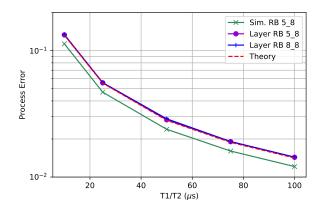
$$\epsilon_{U} = 1 - \frac{\text{Tr} \left[U_{ideal}^{*}U\right]^{2}}{d^{2}}$$

$$\epsilon_{\Lambda} = 1 - \prod_{i} \left(\frac{1}{4} + \frac{1}{2}e^{-t_{g}/T_{2,i}} + \frac{1}{4}e^{-t_{g}/T_{1,i}}\right).$$
(B2)

where  $t_g$  is the gate length and both these errors are process errors.

In the first set of simulations we only consider incoherent errors and perform the simulation on the even layer of the 4Q set as shown in the top of Fig. 5. We consider two different scenarios, one where the two gates in the layer  $(CX_{01} \text{ and } CX_{23})$  are different lengths (5 time units for  $CX_{01}$  and 8 time units for  $CX_{23}$ ) and another scenario where both gates are 8 time units. Trivially, simultaneous RB gives the wrong answer for the error of the layer because there are no enforced barriers between the different two-qubit gates. The different layer fidelities are the same because of the barrier. This illustrates how the layer fidelity enforces the layer to be as long as the longest gate for all qubits. The theory agrees well, once we include the 1.5 single qubit gates per layer (so the layer is considered 8+1.5 units in length). The length unit is 50 ns. We then continue the simulation for both layers (with all three gates having length 8) and compare to mirror RB (bottom of Fig. 5). For comparison here the mirror layer has a set of random 1Q Cliffords before each layer of two-qubit gates so that the total gate counts are the same in layer and mirror. The agreement between the two is near exact.

Next we investigate the more interesting case of coherent crosstalk error. We take  $T1 = T2 = 50\mu s$  (same unit



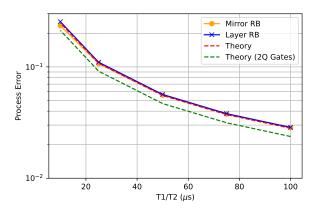


FIG. 5: (Top) Simulation of the even layer with incoherent errors  $(T_1 = T_2)$  and the gate unit length of 50 ns. As described in the main text, when the gate lengths are different simultaneous RB trivially gives the wrong answer. (Bottom) Similar simulation for the two layers with incoherent errors  $(T_1 = T_2)$  comparing mirror to layer and the agreement is exact. There are two theory curves in the bottom plot; in the red curve the single qubit gates and the two qubit gate layers are included in calculating the total incoherent error (there are on average 1.5 single qubit gate layers per two-qubit gate layer). The green theory curve is the error if we just consider the two-qubit layer. For agreement with theory, the single qubit gates in the layer must be considered. There are 10 random sequences in each simulation.

time length as before, 50 ns) and vary the ZZ interaction rate,  $e^{-i2\pi\xi_{ZZ}|11\rangle\langle11|}$ , between qubits 0 and 3  $(ZZ_{03})$  and qubits 1 and 2  $(ZZ_{12})$ . This error is out of the disjoint subspace. We consider two versions of this ZZ crosstalk; one version where the ZZ is "always-on", and another where it only occurs during simultaneous two-qubit gate operation. All the gate lengths are the same (8 units).

First, we look at just layer 1 (top Fig. 6) with alwayson ZZ and compare isolated RB, simultaneous RB and layer RB. Trivially the isolated RB is not affected by the

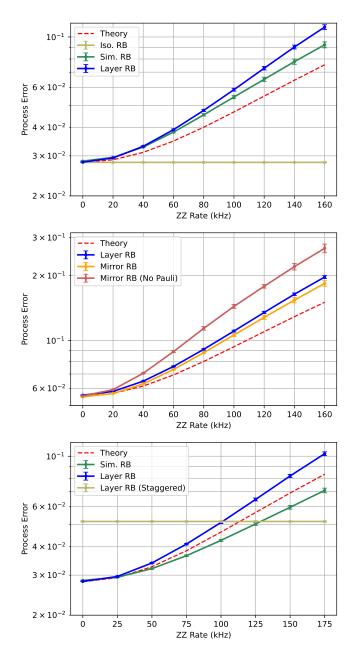


FIG. 6: (Top) Simulation of the even layer vs ZZ rate with the same length gate (8 units) comparing isolated RB, simultaneous RB and layer RB. (Middle) Simulation of the full layer vs ZZ rate comparing layer RB, mirror RB and mirror RB without a Pauli layer between mirrors. (Bottom) Simulation of the even layer vs ZZ, where the ZZ is only applied when there are simultaneous 2Q gates. If we stagger the gates then the crosstalk term disappears, but the overall baseline error is higher. There are 30 random sequences in each simulation.

ZZ interaction, demonstrating that it's a poor method for assessing crosstalk. Simultaneous RB and Layer RB

are reasonably similar, with the caveat from Fig. 5 that if the gate lengths are different simultaneous RB will not reflect the layer properly.

Next we consider the full layer with always-on ZZ and compare layer RB to mirror RB. We look at two flavors of mirror RB, the first is to exactly mirror the circuit ("no pauli") and the second is to more faithfully execute the mirror circuit with a random Pauli layer between the original circuit and its mirror. For this crosstalk the "no Pauli" mirror reports a much higher error whereas the layer and mirror (with Pauli) are in very close agreement.

Finally, we look at layer 1, with ZZ that is only activated by simultaneous 2Q gates, noting that such a crosstalk could be activated by the physics described in Ref. [42, 43]. Here there is a greater divergence between the layer fidelity and simultaneous RB results because simultaneous RB does not enforce strictly running the 2Q gates at the same time. Furthermore, we see that if we run a layer where the 2Q gates are staggered, this crosstalk term trivially vanishes, although the layer has higher baseline error since it's longer. This elucidates why sometimes it can be beneficially to run nonsimultaneous gates for crosstalk as seen in Ref. [39]. This particular version of crosstalk is not necessarily representative, but serves as an example for a family of similar crosstalk terms that activate with simultaneous 2Q gates. We note that there is some ambiguity in calculating the theory curves for these plots, we use Eqn. B1, but since the errors are in the single qubit layer as well we approximate the errors by assuming the single layers and two qubit layers add.

Ultimately the comparison we desire is between layer fidelity and mirror fidelity. In the plots we see the two are fairly close, but the space of possible unitary errors is very large. Therefore, we take a scattershot look at a variety of different error terms and compare between the two methods, as summarized in Fig. 7. The general trend appears is that layer and mirror measure very similar errors, layer fidelity tends to measure slightly higher error than mirror (consistent with the discussion in the next section), whereas without the Pauli layer mirror always measures a larger error.

# Appendix C: Crosstalk and Layer Fidelity

Here we consider the layer fidelity protocol with a single Pauli weight-2 coherent crosstalk term. There are two subspaces k and j that both have  $n_k$  and  $n_j$  qubits and there is a crosstalk term of the form,

$$U = e^{-i\alpha P_x} \approx \mathcal{I} - i\alpha P_x - \frac{\alpha^2}{2} \mathcal{I}$$
 (C1)

where  $P_x$  is a weight-2 Pauli spanning k and j and  $\alpha$  is small so we take the small  $\alpha$  expansion. The true fidelity

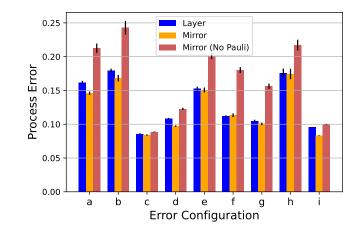


FIG. 7: Comparing layer fidelity, mirror RB and mirror RB without a Pauli layer for measuring the process error with several different coherent error scenarios. As in the other simulations  $T_1 = T_2 = 50 \mu s$  and the unit time is 50 ns. (a) Always on 150 kHz ZZ rate between  $0_{-1}$  and  $2_{-3}$ . (b) Always on 150 kHz ZZ rate between 0.3 and 1.2 (same as Fig. 6). (c) Simultaneous only 150 kHz ZZ rate between 0\_1 and 2\_3. (d) Simultaneous only 150 kHz ZZ rate between 0.3 and 1.2. (e) Always on 100 kHz ZZ rate between 0.1, 1.2 and 2\_3 (all the connected qubits). (f) Z error applied after every time slice of 0.02,  $e^{-i0.02Z/2}$ . (g) 10% over rotation on all two-qubit gates. (h) 10% over rotation on all two-qubit gates and a 10% under rotation on all 1Q gates. (i) Drive crosstalk of 10% (IY and ZY) from qubit 1 to qubit 2 when applying the  $CX_{01}$  gate, from qubit 2 to qubit 1 when applying the CX<sub>23</sub> gate, and

from qubit 1 to qubit 0 when applying the  $CX_{12}$  gate.

of the layer (idles and this crosstalk term) is

$$F_U = \frac{Tr(U)^2}{4^{n_k + n_j}}$$

$$\approx 1 - \alpha^2$$
(C2)

$$\approx 1 - \alpha^2$$
 (C3)

Now, what if we do simultaneous RB and are able to twirl k and j (mythically here without additional problems), from the simultaneous paper we know that the decay parameter is  $\text{Tr}[\Pi_k R_U]/\text{Tr}[\Pi_k]$  where  $\Pi_k$  are the Pauli's just in k (e.g. if  $n_k = 2, n_i = 2$ , this would be the 15  $XIII, YIII, \dots, IXII, \dots, XXII$ ). Calculating the PTM terms (remember we only need the on-diagonal

terms) and leaving off the  $1/2^{n_k+n_j}$  in front of the trace,

$$(R_U)_{i,i} = \text{Tr}[P_i U^{\dagger} P_i U] \tag{C4}$$

$$(R_U)_{i,i} = \operatorname{Tr}[P_i(\mathcal{I}(1 - \frac{\alpha^2}{2}) + i\alpha P_x)P_i$$

$$(\mathcal{I}(1 - \frac{\alpha^2}{2}) - i\alpha P_x)] \tag{C5}$$

$$= \operatorname{Tr}[\mathcal{I}(1 - \frac{\alpha^2}{2})^2 + i\alpha(1 - \frac{\alpha^2}{2})(P_i P_x P_i - P_x) + \alpha^2 P_i P_x P_i P_x)]$$
 (C6)

$$\approx \operatorname{Tr}[\mathcal{I}(1-\alpha^2) + \alpha^2 P_i P_r P_i P_r)] \tag{C7}$$

because the middle terms have trace zero. So if  $[P_i, P_x] =$ 0 then the above is 1, and if they don't commute then the above is  $1-2\alpha^2$ . In the space for the decay parameter of k then, there are  $4^{n_k-1/2}$  elements with  $1-2\alpha^2$  (because there are 2 Pauli's in k that don't commute with  $P_X$ which is weight 2 but only has 1 weight in k) and the rest are 1. The fidelity in space k is then

$$F_{U,k} = 1 - 2\frac{4^{n_k - 1/2}}{4^{n_k}}\alpha^2$$
 (C8)  
=  $1 - \alpha^2$  (C9)

$$= 1 - \alpha^2 \tag{C9}$$

So the estimate is independent of  $n_k$  and therefore  $F_{U,k} =$  $F_{U,j}$ . And then since we multiply the two fidelities together to estimate  $F_U$ ,

$$\tilde{F}_U = (1 - \alpha^2)^2 \tag{C10}$$

$$\approx 1 - 2\alpha^2$$
 (C11)

which is a lower fidelity than the true fidelity Eqn. C3.

This analysis also holds by the same arguments for a Pauli stochastic error channel of the form

$$\mathcal{E}(\rho) = (1 - \alpha^2)\rho + \alpha^2 P_x \rho P_x, \tag{C12}$$

with  $P_x$  defined as before. For this error channel we again find that  $F_{\mathcal{E}} = 1 - \alpha^2$ , and  $F_{\mathcal{E},k} = F_{\mathcal{E},j} = 1 - \alpha^2$ , such that the layer fidelity is a lower bound for the true process fidelity  $F_{\mathcal{E}}$ .

### Appendix D: Combining Process Fidelities

Here we summarize some properties of the process fidelity which have been shown in other sources for reference. The process fidelity is defined as the trace of the superoperators (see, e.g. Ref [44] for a summary), in Pauli form,

$$R_{ij} = \frac{\text{Tr}\left[P_i\Lambda[P_j]\right]}{d} \tag{D1}$$

$$F_p = \frac{\text{Tr}\left[R_{\text{ideal}}^{-1}R\right]}{d^2} \tag{D2}$$

which is related to the average gate fidelity [45]

$$F_g = \frac{dF_p + 1}{d + 1} \tag{D3}$$

where  $\epsilon_g = 1 - F_g$  is the average gate error and is often quoted from the gate error from randomized benchmarking. If there are two disjoint subspaces that have process fidelities  $F_{p,0}$  and  $F_{p,1}$ , then the fidelity of the combined system is  $F_{p,0}F_{p,1}$ , which is the property we have used to build up each disjoint layer fidelity. It is, however, not true that process fidelities multiply across layers (for simplicity assuming the ideal is the identity),

$$F_{p,q} = \frac{\operatorname{Tr}\left[R_p R_q\right]}{d^2} \tag{D4}$$

$$\neq \frac{\operatorname{Tr}\left[R_{p}\right]}{d^{2}} \times \frac{\operatorname{Tr}\left[R_{q}\right]}{d^{2}}$$
 (D5)

However this is approximately true for small errors of diagonal maps, which can be shown by a simple expansion. Practically this means that the fidelity of the layer repeated to multiple depths is fairly well approximated until the fidelity drops below a percent.

## Appendix E: Relating $\gamma$ to LF

In this section we relate the LF to  $\gamma$  as defined in Ref. [1]. This is a useful metric for error mitigation since it indicates the number of circuit randomizations required to perform probabilistic error mitigation.  $\gamma$  is defined for Pauli diagonal noise model, and although we define LF for a depolarizing model, we will do a general comparison here for a Pauli diagonal noise model. As a reminder the definition of  $\gamma$  in terms of the PTM elements defined in the above section is,

$$\gamma = e^{2\sum_k \lambda_k} \tag{E1}$$

$$R_i = e^{-2\sum_{\langle k \rangle_i} \lambda_k} \tag{E2}$$

$$F_p = \frac{1 + \sum_{i=1}^{4^n - 1} R_i}{4^n}$$
 (E3)

where  $\langle k \rangle_i$  is the sum over k where  $[P_i, P_k] \neq 0$  and  $\lambda_k$ are generators of a Lindblad equation that are small for small errors. In that limit, it's straightforward to expand the exponentials,

$$R_i \approx 1 - 2 \sum_{\langle k \rangle_i} \lambda_k$$
 (E4)

$$F_p \approx \frac{1 + \sum_{i}^{4^n - 1} \left(1 - 2 \sum_{\langle k \rangle_i} \lambda_k\right)}{4^n}$$
 (E5)

$$= 1 - \sum_{k} \lambda_k \tag{E6}$$

$$\approx e^{-\sum_k \lambda_k}$$
 (E7)

$$= \gamma^{-1/2} \tag{E8}$$

using the fact that  $\sum_{i} \sum_{\langle k \rangle_i} = \sum_{k} \sum_{\langle i \rangle_k}$ and  $\sum_{\langle i \rangle_k} = 2^n$ .

Next we explore the correspondence of  $\gamma$  to more commonly used gate metrics such as the diamond norm or the average gate fidelity with more rigor and provide bounds. For a Pauli channel, both the average gate fidelity and the diamond norm have simple deviations from the spectral properties of the process matrix  $\Lambda$ , where  $\Lambda = e^{\mathcal{L}}$  [4]. In particular, for a Pauli Channel we can write the average gate error and diamond norm as

$$\varepsilon(\Lambda) = \frac{1 - \frac{\text{Tr}(\Lambda)}{d^2}}{1 + \frac{1}{d}},\tag{E9}$$

$$\|\Lambda\|_{\diamond} = 2\left(1 + \frac{1}{d}\right)\varepsilon(\Lambda) = 2\left(1 - \frac{\operatorname{Tr}(\Lambda)}{d^2}\right).$$
 (E10)

For a Pauli channel, we can derive either metric from the arithmetic average of the eigenvalues of  $\Lambda$ .

Note that the form in Eqn. 5 is derived from a Lindbladian generator,  $\mathcal{L}(\rho) = \sum_k \lambda_k (P_k \rho P_k - \rho)$ . This allows us to express  $\gamma$  in terms of the spectrum of  $\mathcal{L}$ . That is

$$\operatorname{Tr}(\mathcal{L}) = -\sum \lambda_k d^2 \implies \gamma = e^{\frac{-2\operatorname{Tr}(\mathcal{L})}{d^2}}$$
 (E12)

Through the exponential map we arrive at

$$\gamma = \det(\Lambda)^{\frac{-2}{d^2}} \tag{E13}$$

That is gamma is related to the geometric mean of the eigenvalues of  $\Lambda^2$ .

### 1. depolarizing channels

For a depolarizing channel, the spectrum of  $\Lambda$  is a single 1 and  $(d^2-1)$ ,  $(1-\alpha)$ 's. In this case

$$\frac{\operatorname{Tr}(\Lambda)}{d^2} = \frac{1}{d^2} + \left(1 - \frac{1}{d^2}\right)(1 - \alpha) \tag{E14}$$

$$\det(\Lambda)^{\frac{1}{d^2}} = (1 - \alpha)^{1 - \frac{1}{d^2}} \tag{E15}$$

In the limit of large d these both converge to  $(1 - \alpha)$ . In terms of this depolarizing parameter  $\alpha$  we have,

$$\varepsilon(\Lambda) \approx \frac{\alpha}{1 + \frac{1}{d}},$$
 (E16)

$$\|\Lambda\|_{\diamond} \approx 2\alpha,$$
 (E17)

$$\gamma \approx (1 - \alpha)^{-2}.\tag{E18}$$

Alternatively, we can express the gate fidelity and diamond norm in terms of gamma as

$$\varepsilon(\Lambda) \approx \frac{1 - 1/\sqrt{\gamma}}{1 + \frac{1}{d}},$$
 (E19)

$$\|\Lambda\|_{\diamond} \approx 2(1 - 1/\sqrt{\gamma}).$$
 (E20)

### 2. The small error limit

Let's assume  $\Lambda$  is very close to the identity, i.e., the spectrum contains terms  $1-\epsilon_j$ . Let's define  $\bar{\epsilon} \equiv \frac{1}{d^2} \sum_j \epsilon_j$ .

$$\frac{\text{Tr}(\Lambda)}{d^2} = 1 - \bar{\epsilon} \tag{E21}$$

$$\det(\Lambda)^{\frac{1}{d^2}} = \prod_j (1 - \epsilon_j)^{\frac{1}{d^2}} = 1 - \bar{\epsilon} + \mathcal{O}(\epsilon^2)$$
 (E22)

Once again we are in the limit where the arithmetic and geometric means are the same, which again yields

$$\varepsilon(\Lambda) \approx \frac{1 - 1/\sqrt{\gamma}}{1 + \frac{1}{d}},$$
 (E23)

$$\|\Lambda\|_{\diamond} \approx 2(1 - 1/\sqrt{\gamma}).$$
 (E24)

#### 3. Bounds

The process fidelity of a superoperator is the arithmetic mean of its eigenvalues. On the other hand, Eqn. E13 established that  $\gamma^{-\frac{1}{2}}$  is equal to the geometric mean of the eigenvalues. To make  $F_p$  and  $\gamma$  more easily comparible this section chooses to work in terms of  $\gamma^{-\frac{1}{2}}$ 

We start with Theorem 1 which provides upper and lower bounds for  $\gamma^{-\frac{1}{2}}$  in terms of  $F_p$ . Although the lower bound appears complicated, it is extremely close to  $\sqrt{2F_p-1}$  on all of  $[\frac{1}{2},1]$ , which can therefore be used as a proxy for most practical purposes. Especially note that both the upper and lower bounds are independent of the dimension  $d=2^n$ . Following the theorem, we provide natural families of channels that saturate the upper bound, and nearly saturate the lower bound. For high fidelity layers, say above  $F_p=0.9$ , it will be seen that  $F_p\approx \gamma^{-1/2}$ .

**Theorem 1.** Suppose  $\Lambda$  is a CPTP Pauli channel with a process fidelity  $F_p = \operatorname{Tr} \Lambda/d^2$ , and  $\gamma = \det(\Lambda)^{-2/d^2}$ . Then it holds that

$$F_p - 1 + 2\lambda_0(1 - F_p) + (2F_p - 1)^{\lambda_0} \le \gamma^{-\frac{1}{2}} \le F_p \text{ (E25)}$$

where

$$\lambda_0 = \frac{\log(2 - 2F_p) - \log(-\log(2F_p - 1))}{\log(2F_p - 1)}.$$
 (E26)

*Proof.* The upper bound on  $\gamma^{-\frac{1}{2}}$  follows directly from a standard application of Jensen's inequality; the geometric mean of positive numbers cannot exceed their arithmetic mean.

To show the lower bound, first observe that all of the Pauli fidelities  $f_a$  of  $\Lambda$  lie in the interval  $[2F_p-1,1]$ . This was shown in Ref. [31], but we repeat the brief argument

here for completeness. We can express the Pauli fidelity  $f_a$  in terms of the Kraus probabilities as

$$f_a = \sum_b (-1)^{\langle a,b \rangle} p_b = \sum_{b:\langle a,b \rangle = 0} p_b - \sum_{b:\langle a,b \rangle \neq 0} p_b$$
$$= 2 \sum_{b:\langle a,b \rangle = 0} p_b - 1, \tag{E27}$$

where we have used the CPTP condition  $\sum_b p_b = 1$  to write the sum of those  $p_b$  where b does not commute with a as one minus the sum of those that do. Now clearly  $\sum_{b:\langle a,b\rangle=0} p_b \geq p_I$ , and moreover  $p_I = 4^{-n} \sum_a (-1)^{\langle I,a\rangle} f_a = F_p$ , hence

$$f_a \ge 2F_p - 1. \tag{E28}$$

We can now apply Lemma 1 (below) with  $c = F_p - 1$  and d = 1 to get the stated inequality.

The geometric mean and arithmetic mean agree exactly when their arguments are equal, which means that  $F_p = \gamma^{-\frac{1}{2}}$  exactly when all of the Pauli fidelities are equal, which, for TP channels, only happens with the identity channel. However, globally depolarizing channels are the next best thing as all values but one are equal. For a globally depolarizing channel with non-trivial Pauli fidelities  $\alpha$ , we have

$$F_p = \frac{1 + (4^n - 1)\alpha}{4^n}$$
 and  $\gamma^{-\frac{1}{2}} = \alpha^{(4^n - 1)/4^n}$ , (E29)

which are very close to equal even for moderate n; see the upper curve in Fig. 8.

The next family of channels we consider are tensor products of two-qubit depolarizing channels, each with strength  $\alpha$ . Assuming n is even and we have the tensor product of n/2 such channels, we get

$$F_p = \sum_{k=0}^{n/2} \alpha^k 15^k \binom{n/2}{k} / 4^n \quad \text{and} \quad \gamma^{-\frac{1}{2}} = \alpha^{15n/32},$$
(E30)

which, as with global depolarizing channels (see Fig. 8), are very close to equal even for moderate n.

The previous two families of channels have had Pauli fidelites that are tightly concentrated. To saturate the lower bound of Theorem 1, we will need to instead choose a channel that maximizes the variance of the Pauli fidelities. As seen in the proof of Lemma 1, this is done by making the Pauli fidelities strongly bimodal, concentrating roughly half of them at 1, and the other half at another value less than 1. This can be done by choosing a  $\Lambda$  to have a Kraus map with only one non-trivial Pauli,

$$\Lambda(\rho) = p\rho + (1 - p)P\rho P. \tag{E31}$$

In this case,  $f_a = 1$  whenever a commutes with P, but 2p - 1 otherwise. In this case we get

$$F_p = p$$
 and  $\gamma^{-\frac{1}{2}} = \sqrt{2p - 1}$ , (E32)

where we note the independence of n. A physically relevant example of such a noise model is one where a single subsystem has a high error that dominates every other error, for example, by taking  $P = XIII \cdots I$ , where the first qubit is faulty. The geometric mean (viz.  $\gamma$ ) is good at capturing this outlier, but the arithmetic mean (viz.  $F_p$ ) is not.

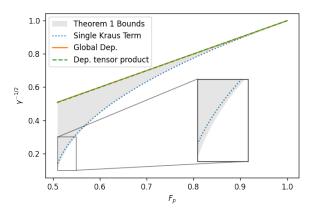


FIG. 8: For any fixed process fidelity on the x-axis, the grey region above it represents the range of values of  $\gamma^{-\frac{1}{2}}$  that are physically consistent by some Pauli channel with that process fidelity. The three curves depict three families of Pauli channels that saturate the bounds of the grey region, two on the top and one on the bottom. The global depolarizing curve (orange) is shown for n=10, but would look identical for any other n not too close to 1. Likewise, the 5-tensor product of 2-qubit depolarizing (dashed green) curve also corresponds to n=10, but would look similar for more qubits. The lower bound is (just about; see inset) saturated by channels for which there is only one non-trivial term in the Kraus representation, representative of noise models where there is a single outlying subsystem, such as  $\Lambda(\rho) = p\rho + (1-p)XIIII\rho XIIII$  (dotted blue).

**Lemma 1.** Suppose that 0 < c < d are real numbers and fix a positive integer N. Define  $a : [0, \infty)^N \to \mathbb{R}$  by  $a(x) = \sum_i x_i/N$  and  $g : [0, \infty)^N \to \mathbb{R}$  by  $g(x) = \prod_i x_i^{1/N}$ . Then restricting to the hyperrectangular region  $D = [c, d]^N$ , we have

$$\max_{x \in D} (a(x) - g(x)) \le f(\lambda_0)$$
 (E33)

where  $f(\lambda) = \lambda c + (1 - \lambda)d + c^{\lambda}d^{1-\lambda}$  and

$$\lambda_0 = \frac{\log(\log(d/c)) - \log((d-c)/d)}{\log(d/c)}.$$
 (E34)

*Proof.* Since a is linear and g is concave, a-g must be convex. Therefore, the maximum of a-g is acheived on

the extreme points of the convex set D, which are given by  $E = \{a, b\}^N$ . That is, we have  $\max_{x \in D} (a - g)(x) = \max_{e \in E} (a - g)(e)$ .

Now, for any  $e \in E$ , there exists some  $0 \le m \le N$  such

that  $(a-g)(e) = (mc + (N-m)d)/N + (c^m d^{N-m})^{1/N}$ . Therefore,  $\max_{e \in E} (a-g)(e) \le \max_{0 \le \lambda \le 1} f(\lambda)$ . Standard calculus show that f is concave on [0,1] and acheives a maximum value at  $\lambda_0$ , which proves the inequality of this lemma.  $\square$