Fidelity decay and error accumulation in quantum volume circuits

Nadir Samos Sáenz de Buruaga[1], Rafał Bistroń[2,3], Marcin Rudziński[2,3],
Rodrigo Miguel Chinita Pereira[1], Karol Życzkowski[2,4], and Pedro Ribeiro[1]
[1] *CeFEMA, LaPMET, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*
[2] *Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, ul. Łojasiewicza 11, 30-348 Kraków, Poland*
[3] *Doctoral School of Exact and Natural Sciences, Jagiellonian University, ul. Łojasiewicza 11, 30-348 Kraków, Poland and*
[4] *Center for Theoretical Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warszawa, Poland*
(Dated: April 17, 2024)

We provide a comprehensive analysis of fidelity decay and error accumulation in faulty quantum circuit models. We devise an analytical bound to the average fidelity between the desired and faulty output states, accounting for errors that may arise during the implementation of two-qubit gates and multi-qubit permutations. We demonstrate that fidelity decays exponentially with both the number of qubits and circuit depth, and determine the decay rates as a function of the parameterized probabilities of the two types of errors. These decay constants are intricately linked to the connectivity and dimensionality of the processor's architecture. Furthermore, we establish a robust linear relationship between fidelity and the heavy output frequency used in Quantum Volume tests to benchmark quantum processors, under the considered errors protocol. These findings pave the way for predicting fidelity trends in the presence of specific errors and offer insights into the best strategies for increasing Quantum Volume.

Introduction.— Developing a full-fledged quantum computer faces a substantial technological challenge: precisely controlling large numbers of quantum degrees of freedom (qubits) while ensuring their adequate isolation to uphold collective quantum coherence and large-scale entanglement. Although present-day quantum computers are still noisy and prone to errors, the progress of the last decade has been substantial. However, regardless of the improvements to come, quantum processors will inevitably operate amidst a finite degree of errors and noise.

Irrespectively of the way to tackle these difficulties, for example, to enhance error-correcting codes or to directly utilize the capabilities of the so-called noisy intermediate-scale quantum era (NISQ) [1], there is a pressing demand for larger and more reliable quantum computers (QCs). Despite the lack of consensus on whether quantum supremacy [2, 3] has been achieved with current technologies [4], scaling is essential alongside with benchmarking tools to assess computer performance and characterize error accumulation, which is vital for ensuring the trustworthiness of QCs.

Although quantum tomography is the ultimate characterization tool [5], the number of experiments needed to store a state of L qubits grows exponentially with the number of qubits, making it unfeasible to assess large computers.

The practical inapplicability of quantum tomography requires the development of alternative benchmarking tools that should remain agnostic to specific technological implementation of qubits, such as superconducting circuits [6], trapped ions [7], Rydberg atoms [8], quantum dots [9], photonics [10], etc., while also being capable of assessing the scaling capabilities of the processor. Both requirements are met with randomized benchmarking protocols [11–16], which originally characterized the decay rate of quantum fidelity [17], $\mathcal{F} = \langle\Psi|\tilde{\rho}|\Psi\rangle$, quantifying the discrepancy between the outcome state of an ideal quantum computer $|\Psi\rangle$ and the actual implemented density matrix $\tilde{\rho}$, where $\tilde{A}$ denotes hereafter the faulty (real) version of the physical quantity A.

Another standard benchmarking protocol is the Quantum Volume (QV) [18, 19], which, roughly speaking, counts the maximum number of qubits that can be properly entangled in a circuit. More precisely, given a QC made of L qubits and T layers. The QV corresponds to the dimension of the Hilbert space associated with the largest square circuit ($2^{\min(L,T)}$) that can exceed a specified threshold ϵ, according to a passing criterion. Originally, it was related to gate fidelities [20],[18], however, the currently adopted standard, proposed in [19], relays on the heavy output generation problem [21].

To ensure agnosis with respect to the details of the qubit nature and to capture the influence of many variables such as L, T, architecture, imperfections in the gates, connectivity, and derived crosstalk errors, the QV test was designed to be implemented in a random circuit depicted in Fig. 1 (a) where each layer $\tau = 1, \ldots, T$ implements the unitary operation $U_\tau = V_\tau \Pi_\tau$ specified in Fig. 1 (b). For this layer configuration, we refer to as *original*. Π_τ is an operator applying a random permutation $P \in S_L$ that interconnects the qubits accordingly, followed by a unitary operation V_τ made of random two-qubit gates sampled from the circular unitary ensemble (CUE).

There is a growing literature on numerical simulations

and real experiments with NISQ computers that calculate the QV [22–32]. However, to our knowledge, there has not been a rigorous analytical study of the dependence of the aforementioned criteria on different noise sources or of its scalability.

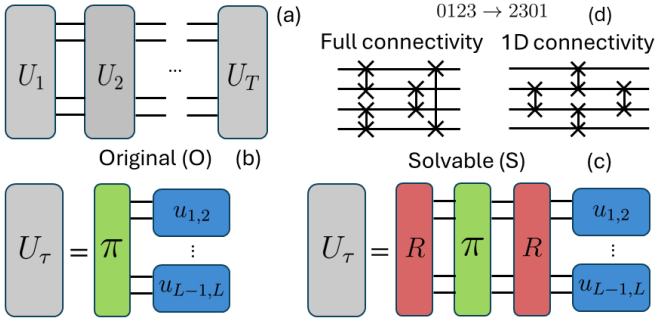


FIG. 1: Schematic of the randomized circuit considered **(a)**. In the *original* circuit **(b)**, each layer of unitaries is composed of a random permutation and faulty 2-qubit gates. The *solvable* **(c)** is obtained from the previous by acting with a unitary gate R before and after the permutation. In **(d)** we depict an example of permutation decomposition by SWAP gates in two different architectures: full connectivity, where interactions between all qubits are allowed, and 1D connectivity, where only the nearest neighbors can interact.

The purpose of this work is to establish bounds for error accumulation in the original quantum volume circuit of width L and depth T, depicted in Fig. 1 (a,b), as quantified by both fidelity $\mathcal{F}$ and heavy output frequency h_U scores. We analyze the effects of implementing *faulty 2-qubit gates* poisoned by generic unstructured noise drawn from the Gaussian Unitary Ensemble (GUE). Furthermore, we explore *faulty permutations* by assuming $\tilde{\Pi}$ is implemented by a combination of malfunctioning *swap* gates $\tilde{S}$ that interchange nearest-neighbour qubits within a specific architecture.

Our main result is an analytical expression for fidelity decay as a function of the number of qubits and layers and noise, that takes into account parametrization for faulty 2-qubit gates and swaps as well as effective processor connectivity. We offer evidence that the rather intuitive conjecture in Ref. [19] regarding the increase in QV with connectivity holds for these generic circuits. These results allow one to identify the noise source whose reduction most effectively enhances the quantum volume.

Average Quantum Fidelity.— The main object we compute is the average fidelity between the faulty and ideal states $\overline{\mathcal{F}} = \overline{|\langle\Psi|\tilde{\Psi}\rangle|^2}$. We obtain it analytically by evaluating the overlap of an averaged superoperator acting in a four-copy Hilbert space $\mathcal{H}(2^{4L})$. We also double-check our results numerically. To present our finding we first introduce the ket vectorization, and the compact notation that emphasizes the dependence between the four copies $\|+^{\mathbf{14;23}}\rangle := \sum_{\boldsymbol{nm}} \|\boldsymbol{mnnm}\rangle$, where $\boldsymbol{m} = m_1, \ldots m_L$, $m = 0, 1$. Equipped with this, as we detail in the Appendix A,

$$\overline{\mathcal{F}} = \frac{1}{2^L} \langle +^{\mathbf{14;23}} \| \, (\boldsymbol{\mathcal{U}})^T \, \| +^{\mathbf{1234}} \rangle , \tag{1}$$

with

$$\boldsymbol{\mathcal{U}} = \overline{\left[\tilde{U} \otimes \tilde{U}^* \otimes U \otimes U^*\right]},$$

and where we have assumed that each layer τ is sampled independently, and the average is taken over all computational basis states $\|+^{\mathbf{1234}}\rangle := \sum_{\boldsymbol{n}} \|\boldsymbol{nnnn}\rangle$, which also correspond to the *entanglement fidelity* of a channel [33]. We use bold calligraphic typeface symbols hereafter to represent averaged superoperators acting on the four-copy Hilbert space.

Error Modeling.— As mentioned, each layer of the original circuit has the form $U_\tau = V_\tau \Pi_\tau$ consisting of the alternation of random permutations $\Pi \in S_L$, followed by a unitary operation $V_\tau = \bigotimes_{r=1}^{L/2} u_{2r-1,2r}$ with $u_{r,r'} \in \mathrm{CUE}(4)$. Therefore, we distinguish between errors that arise in the permutation operator Π caused by faulty swaps and errors in the computation itself V_τ.

To simulate the errors within the permutation operator $\tilde{\Pi}$, we first decompose it as a combination of *swap* operators S interchanging nearest qubits within some architecture – see Fig. 1 (d). The error scheme assumes that each swap S is not performed exactly but with some too-short or too-long impulse $S \to S^\beta$, where the power β is drawn from the Gaussian distribution with mean equal 1 and variance σ^2 independently for all swaps. Such modelling of error is natural, for example, if one uses $\sqrt{S}$ as a universal 2-qubit gate because in such a case the swap gate S is elementary and hence fast, so the natural source of error is the imperfection in $\sqrt{S}$ itself.

In the Appendix C, we illustrate how the faulty permutation $\tilde{\Pi}$ can be conveniently reinterpreted, after averaging the swap powers β, as a composition of the same swaps S, with each swap having the probability $p = \left[1 - \exp(-\pi^2\sigma^2/2)\right]/2$ of being omitted independently.

Concerning the errors in the random gates, each two-qubit gate is modified by a generic noise $\tilde{V} = \bigotimes_{r=1}^{L/2} \tilde{u}_{2r-1,2r}$ with $\tilde{u}_{2r-1,2r} = e^{i\alpha h_{2r-1,2r}} u_{2r-1,2r}$, where $h_{2r,2r-1} \in \mathrm{GUE}(4)$ and $\alpha \geq 0$.

Hence, we are interested in computing the average fidelity Eq. (1) in terms of the two noise parameters $\overline{\mathcal{F}(\alpha, p)}$. Notice that this is a challenging task, since due to the random permutations, each layer cannot be treated independently.

Solvable Model.— We now present a solvable model that exhibits a qualitative fidelity decay with L and T akin to the original circuit. The solvable circuit, presented in Fig. 1 (c) is derived from the original, presented in Fig. 1 (b) by introducing *faultless* random unitaries R_τ and $R_{\tau+1} \in \mathrm{CUE}(2^L)$ before and after each

permutation Π_τ. The ability to average each R-layer independently enables us to explicitly compute $\mathcal{F}$ using Eq. (1). Despite notable differences between the circuits, particularly in terms of their expressibility or entanglement capability [34], we present extensive numerical evidence that both circuits exhibit the same responses to the considered noise models. A simplified rationale for this agreement is as follows: although implementing CUE(2^L) matrices with two-qubit layers is exponentially costly [35], their cumulative action decorrelates a typical state for the solvable model in a similar manner as random two-qubit gates. This remarkable agreement is evident in Figs. 2 (a) and (b), where the fidelity is plotted as a function of number of layers T or noise level α, respectively, assuming no permutation errors. In all subsequent figures, points represent numerical simulations of the original model, while solid curves correspond to Eq. (4) (below) obtained for the solvable model.

We now outline the calculation leading to an analytical closed form of the fidelity for the solvable model. Detailed derivations are provided in the Appendix A. Each layer average, denoted as $\boldsymbol{\mathcal{U}}$, can be decomposed into the contribution of permutation $\boldsymbol{\mathcal{P}}$, embedded by CUE averages $\boldsymbol{\mathcal{R}}$, and the contribution of faulty operations $\boldsymbol{\mathcal{V}}$. Their combined action can be expressed in terms of an effective spin-1/2 orthonormal basis, denoted as $\|\Uparrow\rangle, \|\Downarrow\rangle$, which remains invariant under $\boldsymbol{\mathcal{R}}$ [36, 37]:

$$\boldsymbol{\mathcal{R}}\,\boldsymbol{\mathcal{P}}\,\boldsymbol{\mathcal{R}} = \|\Uparrow\rangle\langle\Uparrow\| + \frac{\delta-1}{4^L-1}\,\|\Downarrow\rangle\langle\Downarrow\|\,, \tag{2}$$

with $\delta = \overline{(\mathrm{tr}\tilde{\Pi}(p)\Pi^T)^2}$. After a straightforward calculation, we obtain $\delta(p) = \overline{4^{m(P,p)}}$, where $m(P,p)$ is the number of cycles in the permutation $Q(p)P^{-1}$, with $Q(p)$ corresponding to faulty implementation of permutation P. The contribution of faulty operations $\boldsymbol{\mathcal{V}}$ are obtained by conveniently assembling each 2-qubit gate superoperator

$$\begin{aligned}\boldsymbol{u}_{2r-1,2r} = {} & \|\Uparrow_{2r-1,2r}\rangle\langle\Uparrow_{2r-1,2r}\| \\ & + \frac{4f(\alpha)+1}{5}\,\|\Downarrow_{2r-1,2r}\rangle\langle\Downarrow_{2r-1,2r}\|\,,\end{aligned} \tag{3}$$

where in this case the orthonormal basis spans the subspace of a four-copy $\mathcal{H}(2^4)$ of two qubits. The effect of noise is manifested through the function $f(\alpha)$ that is closely related to the spectral form factor of GUE(4) [38] (see App. B). In the most relevant limit $\alpha \ll 1$, $f(\alpha) \approx e^{-5\alpha^2}$.

Combining the contribution of both sources of noise, we obtain the main result of this work consisting of an upper bound for the average fidelity – for derivation details see App.B,

$$\overline{\mathcal{F}} = \left(1-\frac{1}{2^L}\right)\left(\frac{(\delta-1)(\Delta-1)}{(4^L-1)^2}\right)^T + \frac{1}{2^L}. \tag{4}$$

The noise parameter $\delta = \delta(p)$ is given above Eq. (2), while $\Delta = \Delta(\alpha) \approx 2^L\,(3f(\alpha)+1)^{L/2}$. The last term corresponds to the fidelity between two random L-qubit pure states [39].

Faulty operations.— Let us consider that all permutations Π_τ are noiseless and that the only source of error comes from the two-qubit gates. This corresponds to fix $\delta(0) = 4^L$ in Eq. (4). For small α and large L we can write $\overline{\mathcal{F}} \sim \exp(-15\alpha^2 LT/8)$. The validity of this approximation can be seen in Fig. 2(b). Furthermore, we show in Fig. 2 (a) that both circuits, original and solvable, behave correspondingly even for shallow circuits.

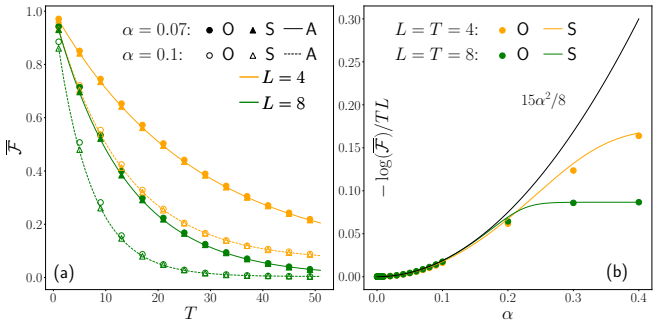


FIG. 2: Fidelity evolution for the system sizes $L = 4$ and 8 with for the *solvable* (S) and *original* (O) circuit together with the analytical prediction (A), both for $p = 0$. **(a)** - Decay as a function of the number of layers T. The solid lines correspond to the analytical result for the (S) model with $\alpha = 0.07$ whereas the dashed to $\alpha = 0.1$. **(b)** - Exponent in fidelity as a function of the parameter α for square circuits. Asymptotically, $\overline{\mathcal{F}} \sim e^{-15\alpha^2 LT/8}$.

Faulty permutations.— To understand the permutation error factor $\delta(p)$ we have to first grasp the decomposition of a random permutation operator Π for a given architecture. The subject of decomposing permutations into swaps within certain graph architectures – token swapping – is already quite extensively studied –see for example [40–43]. However, to the best of our knowledge, they were not considered from the perspective of faulty behavior of the entire quantum circuit.

We acknowledge that in real-live implementation of quantum volume circuits those permutations are modified and optimized [27] However, we noticed that for the discussed scenarios the optimal modifications of permutations in the circuit one by one change the results only by a parameter-independent constant. Thus for clarity, we restrict our attention only to the implementation of permutations.

Full connectivity. Let us first consider the idealistic situation in which each pair of qubits is connected and can be swapped in one move. For such a case the *optimal* way of decomposing permutation can be established giving the minimal number of swaps organized in a minimal number of layers. Each permutation P can be decomposed into a set of cycles $\{C_i\}$ of k_i elements and each cycle C_i can be decomposed into $k_i - 1$ swaps combined in just two layers [40], as shown in the Appendix C. Thus, to

implement permutations of L qubits on average one performers only $L - H_L \approx L - \log L - \gamma$, where $H_L \approx \log L + \gamma$ are Harmonic numbers describing the average number of cycles in permutation [40] and γ is Euler's constant.

The structure of the above-discussed implementation of Π in only two layers gives one more crucial property. In the derived probabilistic model (see App. C) with each swap omitted with probability p, one omission in $\tilde{\Pi}(p)$ corresponds directly to the disappearance of one cycle in $Q(p)P^{-1}$. In other words, each omitted swap in $\tilde{\Pi}(p)$ introduces a new error instead of possibly mutating or suppressing previous ones.

Thanks to the properties described above the error factor $\delta(p)$ can be directly calculated

$$\delta_{full} \approx 4^L e^{-\frac{3}{4}p(L-\log L-\gamma)}, \tag{5}$$

neglecting small corrections in higher powers of p. This corresponds, in the absence of other errors, to fidelity decay in the pattern:

$$\overline{\mathcal{F}}_{full} \approx e^{-\frac{3}{4}pTL\left(1-\frac{\log L+\gamma}{L}\right)} \approx e^{-\frac{3}{4}pTL} \tag{6}$$

in the limit of the large number of qubits L and layers T.

1D Architecture. Full connectivity serves only as a theoretical limiting case, which is why we focus on different models, first 1 dimensional line. The generalization of this model into 2 or higher dimensional lattices will be considered later and discussed in detail in the Appendix D.

For 1D, and other lattice-like architectures new difficulty emerges. Since in typical permutation Π the qubits usually have to be "moved" by the number of nodes proportional to the length scale of architecture, the number of layers and additional swaps in which Π is decomposed have to grow with the size of the circuit as well, resulting in the stacking of multiple swaps one onto the other. Thus the omission of some swap in $\tilde{\Pi}(p)$ can modify previous errors instead of introducing a new one. For clarity, we refer to the regime of "spare errors" when the frequency of such cases is negligible. Then we may state:

$$\delta \gtrsim 4^L e^{-\frac{3}{4}p\overline{w(L)}}, \tag{7}$$

Where $\overline{w(L)}$ is an average number of swaps in permutations of L qubits. With the increase of error probability p this approximation deviates from the exact value of the error factor and starts serving as a lower bound – see App. D.

Based on the above discussion our focus is on *"optimal"* implementations of Π in the sense of minimal necessary number of swaps. Simultaneously, we also want to minimize the number of layers in Π implementation, since they translate to longer implementation time, hence larger chances of memory errors.

For 1D architecture the decomposition of permutation P which gives the minimal number of swaps is obtained by the odd-even sort (known also as brick-sort or parity-sort algorithm) [44]. The maximal number of swaps is $L(L-1)/2$ and the distribution of a minimal number of swaps is given by the Mahonian distribution [45] which very quickly converges to Gaussian distribution. Moreover, the maximal number of layers in Π implementation is equal L which is also an asymptotic limit for the minimal necessary number of layers.

Summarizing, in the regime of sparse errors for 1D architecture the error factor is given by:

$$\delta_{1D} \gtrsim 4^L e^{-\frac{3}{4}p\frac{L(L-1)}{4}}, \tag{8}$$

Which corresponds, in the absence of other errors, to fidelity decay in the pattern:

$$\overline{\mathcal{F}}_{1D} \gtrsim e^{-\frac{3}{16}pTL^2\left(1-\frac{1}{L}\right)} \approx e^{-\frac{3}{16}pTL^2} \tag{9}$$

in the limit of the large number of cubits L and layers T.

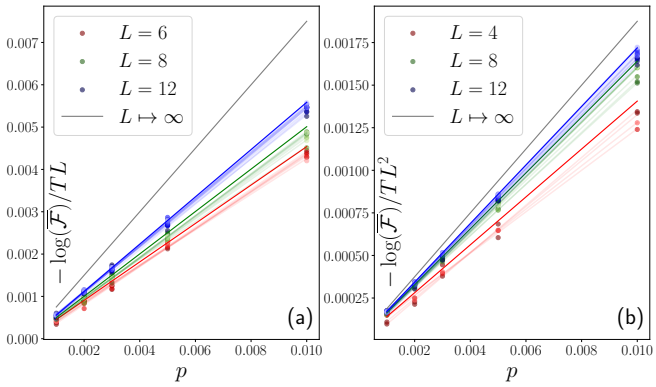


FIG. 3: Fidelity decay as a function of swap ommission probability p for quantum volume circuits with perfect two-qubit gates $\alpha = 0$ with L qubits and from 1 to L layers (from thinner to stronger shade). **(a)** – fully connected architecture, **(b)** – linear architecture. To compare different numbers of qubits L and layers T the index in fidelity decay is divided by general trends. Dots with shaded lines correspond to numerical results with fitted behavior, continuous lines to theoretical predictions for a given number of qubits (6), (9) and grey line the asymptotic behavior for many qubits.

Other Architectures. The above discussion of 1D architecture can be generalized for 2D and higher dimensional cubic lattices. Although the exact optimal implementations are not known, the complexity in the number of swaps and layers required can be directly derived. In the Appendix D, we present lower bounds for the average necessary number of swaps and layers to implement a permutation. For example, in 2D square lattice of L qubits, the typical permutation "moves" any qubit by a distance proportional to the length of a square $\sqrt{L}$. So if we allow only nearest neighbor swaps, the average minimal number of swaps cannot be smaller than $\overline{w(L)} \propto L^{3/2}$.

Furthermore, we also constructed a complex algorithm to decompose permutations on square or higher

dimensional architectures with the desired complexity, presented in App. D, which implementation is available at the github repository. The study of this algorithm lets us establish a lower bound for the error factor $\delta_{2D} \gtrsim 4^L e^{-9pL^{3/2}/8}$ for square circuit and $\delta_{dD} \gtrsim 4^L e^{-\frac{3}{4}(d-1/2)pL^{1+\frac{1}{d}}}$ for d-dimensional circuit. In the absence of other errors, this leads to the bounds for the fidelity decay,

$$\overline{\mathcal{F}}_{2D} > e^{-\frac{9}{8}pTL^{3/2}}, \quad \overline{\mathcal{F}}_{dD} > e^{-\frac{3}{4}(d-1/2)pTL^{1+1/d}} \tag{10}$$

in the limit of the large number of qubits L and layers T.

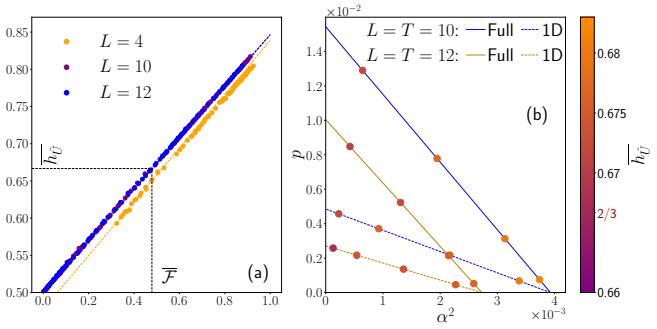


FIG. 4: **(a)** Linear relation collapse between the heavy output frequency and fidelity for the *original* circuit (dots) and solvable circuit (full lines) for various $T \geq L$, 1D and fully connected architectures, and different values of α and p. The dotted lines indicate a postulated linear relation. The threshold value $h^*_{\tilde{U}} = 2/3$ corresponds to a fidelity of $\mathcal{F}^* \approx 0.479$ (black dashed lines). **(b)** Threshold contour curves for $QV = L = T$ for 1D (dashed line) and fully connected (full line) architectures obtained by solving $\overline{\mathcal{F}}(\alpha, p) = \mathcal{F}^*$ with Eq. (4). The colored points represent the corresponding numerical average $\overline{h_{\tilde{U}}}$ of the original circuit Fig. 1 (b).

Heavy output frequency vs Fidelity.— In this section, we present the arguments for the generalization of the discussed result from fidelity into a more operational framework. To do so, we refer to Quantum Volume, which is a widely used benchmarking tool for assessing quantum processors [18, 19].

The standard measure in the Quantum Volume test, heavy-output frequency $h_{\tilde{U}}$, compares the outputs of faulty and ideal states represented in the computational basis. Heavy output frequency $h_{\tilde{U}}$ is calculated as the average sum of probabilities measured in the experiment, $\tilde{p}(m) = |\langle m|\tilde{\rho}|m\rangle|$, that exceed the median of the ideal distribution $p(m) = |\langle m|\Psi\rangle|^2$. QCs achieving an average $\overline{h_{\tilde{U}}}$ greater than a certain threshold, $\epsilon = 2/3$, with statistical accuracy deemed reliable [21].

The numerical experiments performed (Appendix E) strongly indicate that for the original circuit and the noise models discussed $\overline{\mathcal{F}}$ and $\overline{h_{\tilde{U}}}$ have a linear relation, similar as stated in [27]. These coincidences suggest that the obtained dependence may be a more general phenomenon, accurate e.g. for isotropic noise. In Fig. 4(a) we illustrate how this relationship together with its asymptotic. A detailed discussion of our findings is presented in Appendix E.

Plugging the limiting value of fidelity corresponding to $h^*_{\tilde{U}} = 2/3$ into the analytical expression Eq. (4) allows us to obtain the threshold line in the space of parameters α and p below which a circuit of L qubits passes the quantum volume test with QV equal to 2^L. Fig. 4(b) shows this threshold line for circuits $L = 10, 12$ and for the two architectures, together with discrepancies obtained from the numerical study of the heavy output frequency on the original circuit. For a quantum processor characterized by (α, p) tuple, our results can be used not only to compute its QV but also to infer what is the best strategy to increase it, e.g. by decreasing α or p.

Concluding remarks.— We have quantified error accumulation in random quantum circuits with faulty two-qubit gates and permutations, with error rates parametrized by α and p respectively, and assuming a processor architecture where qubit connectivity is characterized by an effective dimension d. The average fidelity (4) for a large number of qubits and layers converges to

$$\overline{\mathcal{F}} \approx \exp\left(-\frac{15}{8}\alpha^2 LT\right)\exp\left(-\frac{3}{4}\left(d-\frac{1}{2}\right)pL^{1+\frac{1}{d}}T\right). \tag{11}$$

This expression is of practical interest for identifying the primary error source that must be mitigated to enhance computational figures of merit such as Quantum Volume. We establish this result by providing numerical evidence for the linear dependence between average fidelity and heavy output frequency in random benchmarking circuits. The obtained formula formalizes the viewpoint presented for example in [19], that denser connectivity in quantum computer architecture leads to better performance.

Our results pave the way to a general framework for understanding how fidelity decay and error accumulation are affected by errors in two-qubit gates and permutations, across varying levels of connectivity and dimensional configurations of quantum architectures. This framework also serves as a starting point for considering other kinds of errors, such as nonunitary, memory, crosstalk and implementation-specific errors.

Acknowledgements. It is a pleasure to thank Ryszard Kukulski, and Javier Molina-Villaplana for fruitful discussions. We also acknowledge Tomaž Prosen and Sergiy Denisov for early discussion that triggered this work. This work realized within the DQUANT QuantERA II Programme was supported by the National Science Centre, Poland, under the contract number 2021/03/Y/ST2/00193, and by FCT-Portugal Grant Agreement No. 101017733 [46] It received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101017733. PR, NB, and RP acknowledge further support from FCT-Portugal through Grant No. UID/CTM/04540/2020.

Appendix A: Average fidelity computation

In this section, we derive the expression Eq. (1) of the main document. The faultless circuit yields a state $|\Psi\rangle$

$$|\Psi\rangle = U_T U_{T-1}, \ldots, U_1 |\psi_0\rangle, \tag{A1}$$

whereas the more realistic circuit produces

$$|\tilde{\Psi}\rangle = \tilde{U}_T \tilde{U}_{T-1}, \ldots, \tilde{U}_1 |\psi_0\rangle, \tag{A2}$$

where $\tilde{U}_\tau$ corresponds to the faulty realization of the clean layer unitary U_τ. In particular, we always can write:

$$\tilde{U}_\tau = V_\tau U_\tau, \tag{A3}$$

where V_τ is the unitary error. As discussed in the previous section, we are interested in computing the average of the fidelity

$$\mathcal{F} = \left|\langle\Psi|\tilde{\Psi}\rangle\right|^2. \tag{A4}$$

To do so, it is useful to use the vectorized notation.

$$|\psi^T\rangle = \langle\psi|^T \tag{A5}$$

$$\|\psi\phi\rangle = |\psi\rangle \otimes |\phi^T\rangle. \tag{A6}$$

Equipped with it we can write the fidelity in a 4-copy Hilbert $\mathcal{H}(2^{4L})$ space:

$$\begin{aligned}
\mathcal{F} &= \left|\langle\Psi|\tilde{\Psi}\rangle\right|^2 \\
&= \langle\psi_0| U_1^\dagger ... U_T^\dagger \tilde{U}_T .... \tilde{U}_1 |\psi_0\rangle \langle\psi_0| \tilde{U}_1^\dagger ... \tilde{U}_T^\dagger U_T .... U_1 |\psi_0\rangle \\
&= \sum_{\boldsymbol{nm}} \langle\psi_0| U_1^\dagger ... U_T^\dagger |\boldsymbol{m}\rangle \langle\boldsymbol{m}| \tilde{U}_T .... \tilde{U}_1 |\psi_0\rangle \langle\psi_0| \tilde{U}_1^\dagger ... \tilde{U}_T^\dagger |\boldsymbol{n}\rangle \langle\boldsymbol{n}| U_T .... U_1 |\psi_0\rangle \\
&= \sum_{\boldsymbol{nm}} \langle\boldsymbol{mn}\| \left(\tilde{U}_T .... \tilde{U}_1 \otimes \tilde{U}_T^* ... \tilde{U}_1^*\right) |\psi_0\rangle \otimes \left|\psi_0^T\right\rangle \langle\boldsymbol{nm}\| U_T .... U_1 \otimes U_T^* ... U_1^* \|\psi_0\psi_0\psi_0\psi_0\rangle \\
&= \sum_{\boldsymbol{nm}} \langle\boldsymbol{mnnm}\| \prod_\tau^{\leftarrow} \left[\tilde{U}_\tau \otimes \tilde{U}_\tau^* \otimes U_\tau \otimes U_\tau^*\right] \|\psi_0\psi_0\psi_0\psi_0\rangle,
\end{aligned} \tag{A7}$$

where

$$\sum_{\boldsymbol{n}} |\boldsymbol{n}\rangle = \left(\sum_{m=0,1} |m\rangle\right)^{\otimes L}.$$

The average fidelity is then

$$\overline{\mathcal{F}} = \sum_{\boldsymbol{nm}} \langle\boldsymbol{mnnm}\| \left(\overline{\left[\tilde{U} \otimes \tilde{U}^* \otimes U \otimes U^*\right]}\right)^T \|\psi_0\psi_0\psi_0\psi_0\rangle = \sum_{\boldsymbol{nm}} \langle\boldsymbol{mnnm}\| \left(\boldsymbol{\mathcal{U}}\right)^T \|\psi_0\psi_0\psi_0\psi_0\rangle, \tag{A8}$$

where we have assumed that each layer is sampled independently, so we can remove the subindex τ.

It is convenient to introduce these two different notations

$$\sum_{\boldsymbol{nm}} \|\boldsymbol{mnnm}\rangle = \left\|_+^{\mathbf{14;23}}\right\rangle = \left(\right), \quad \text{and} \quad \sum_{\boldsymbol{nm}} \|\boldsymbol{mmnn}\rangle = \left\|_+^{\mathbf{12;34}}\right\rangle = \begin{pmatrix} \\ \end{pmatrix}, \tag{A9}$$

and to average the above quantity also with respect to all the all the computational basis states $|\psi_0\rangle = |\boldsymbol{m}\rangle$

$$\sum_{\boldsymbol{n}} \|\boldsymbol{nnnn}\rangle = \left\|_+^{\mathbf{1234}}\right\rangle = \begin{pmatrix} \\ \\ \end{pmatrix}. \tag{A10}$$

Hence, the final expression of the average fidelity is

$$\overline{\mathcal{F}} = \frac{1}{2^L} \left\langle_+^{\mathbf{14;23}}\right\| \left(\boldsymbol{\mathcal{U}}\right)^T \left\|_+^{\mathbf{1234}}\right\rangle, \tag{A11}$$

Appendix B: Average fidelity solvable model circuit

In this section, we will obtain the expression Eq. (4) of the main document corresponding to the average fidelity of the solvable model introduced in the main text, where each layer of the solvable model consists of a faulty permutation $\tilde{\Pi}$ embedded between two faultless large unitaries $R_1, R_2 \in \mathrm{CUE}(2^L)$ belonging to the circular unitary ensemble, i.e., sampled uniformly according to the Haar measure, and a gate made of nonoverlapping random faulty 2-qubit unitaries $\tilde{V} = \bigotimes_{r=1}^{L/2} \tilde{u}_{2r-1,2r}$. Therefore, we have the average of the superoperator $\boldsymbol{\mathcal{U}}$ in Eq. (A11) can be further decompose:

$$\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{V}}\, \boldsymbol{\mathcal{R}}\, \boldsymbol{\mathcal{P}}\, \boldsymbol{\mathcal{R}}\,, \tag{B1}$$

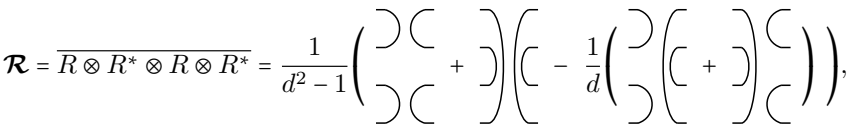
where

$$\boldsymbol{\mathcal{V}} = \overline{\tilde{V} \otimes \tilde{V}^* \otimes V \otimes V^*} \tag{B2}$$

$$\boldsymbol{\mathcal{P}} = \overline{\tilde{\Pi} \otimes \tilde{\Pi} \otimes \Pi \otimes \Pi} \tag{B3}$$

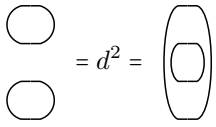
$$\boldsymbol{\mathcal{R}} = \overline{R \otimes R^* \otimes R \otimes R^*} \tag{B4}$$

On what follows, we will describe in detail each of the above quantities. First, the averages in the circular unitary ensemble have been widely studied and computed by means of Weingarten calculus [36, 47]. In particular, following the tensor network perspective introduced in [48], we have that given $R \in \mathrm{CUE}(d)$,

$\boldsymbol{\mathcal{R}} = \overline{R \otimes R^* \otimes R \otimes R^*} =$  (B5)

where we have used the diagrammatic notation introduced above in Eq. (A9).

Observe that the states $\|+^{\mathbf{14;23}}\rangle$ and $\|+^{\mathbf{12;34}}\rangle$ are not orthogonal:

$\langle +^{\mathbf{12;34}} \| +^{\mathbf{12;34}} \rangle =$  $= d^2 =$ $= \langle +^{\mathbf{14;23}} \| +^{\mathbf{14;23}} \rangle$

$\langle +^{\mathbf{12;34}} \| +^{\mathbf{14;23}} \rangle =$  $= d =$  $= \langle +^{\mathbf{14;23}} \| +^{\mathbf{12;34}} \rangle\,.$ (B6)

Hence, we find an orthogonal basis via the Gram-Schmidt procedure:

$$\|\Uparrow\rangle = \frac{1}{d} \|+^{\mathbf{12;34}}\rangle \qquad \|\Downarrow\rangle = \frac{1}{\sqrt{d^2-1}} \left(\|+^{\mathbf{14;23}}\rangle - \frac{1}{d} \|+^{\mathbf{12;34}}\rangle \right). \tag{B7}$$

On this basis, the expression Eq. (B5) takes the simpler form

$$\boldsymbol{\mathcal{R}} = \|\Uparrow\rangle \langle\Uparrow\| + \|\Downarrow\rangle \langle\Downarrow\|\,. \tag{B8}$$

It is clear from the above expression that $\boldsymbol{\mathcal{R}}$ is a projector since $\boldsymbol{\mathcal{R}}^k = \boldsymbol{\mathcal{R}}$, $k \in \mathbb{N}$. This property will be convenient in what follows.

1. Computing $\boldsymbol{\mathcal{V}}$

On this subsection we aim to compute the corresponding average contribution of the faulty 2-qubit gates

$$\begin{aligned} \boldsymbol{\mathcal{V}} &= \overline{\bigotimes_{r=1}^{L/2} \tilde{u}_{2r-1,2r} \otimes \bigotimes_{r=1}^{L/2} \tilde{u}^*_{2r-1,2r} \otimes \bigotimes_{r=1}^{L/2} u_{2r-1,2r} \otimes \bigotimes_{r=1}^{L/2} u^*_{2r-1,2r}} = \bigotimes_{r=1}^{L/2} \overline{\tilde{u}_{2r-1,2r} \otimes\ \tilde{u}^*_{2r-1,2r} \otimes\ u_{2r-1,2r} \otimes\ u^*_{2r-1,2r}} \\ &= \bigotimes_{r=1}^{L/2} \boldsymbol{u}_{2r-1,2r}. \end{aligned} \tag{B9}$$

As discussed in the main text, each faulty gate is modeled as a random unitary $u_{r,r'} \in \mathrm{CUE}(4)$ poissoned with an unstructured and generic noise given by a random unitary whose generator belongs to the Gaussian Unitary Ensemble $\tilde{u}_{2r-1,2r} = e^{i\alpha h_{2r-1,2r}} u_{2r-1,2r}\ h_{2r,2r-1} \in \mathrm{GUE}(4)$ where $\alpha \geq 0$. Therefore, the local average of the 4-copied unitaries can be further decomposed.

$$\begin{aligned} \boldsymbol{u}_{2r,2r-1} &= \overline{\tilde{u}_{2r-1,2r} \otimes\ \tilde{u}^*_{2r-1,2r} \otimes\ u_{2r-1,2r} \otimes\ u^*_{2r-1,2r}} \\ &= \left(\overline{e^{i\alpha h_{2r-1,2r}} \otimes\ e^{-i\alpha h_{2r-1,2r}} \otimes\ \mathbb{1} \otimes\ \mathbb{1}}\right)\left(\overline{u_{2r-1,2r} \otimes\ u^*_{2r-1,2r} \otimes\ u_{2r-1,2r} \otimes\ u^*_{2r-1,2r}}\right). \end{aligned} \tag{B10}$$

We identify the rightmost average as the one outlined in the previous section $\boldsymbol{\mathcal{R}}$ particularized for $d = 4$. Therefore, it remains to compute the average $\overline{e^{i\alpha h_{2r-1,2r}} \otimes\ e^{-i\alpha h_{2r-1,2r}}}$ with respect to the GUE measure. To keep the discussion more general, let us compute this quantity for a general $H \in \mathrm{GUE}(d)$.

$$\overline{e^{i\alpha H} \otimes e^{-i\alpha H^*}} = \int dH e^{-\mathrm{tr}\frac{H^2}{2}} \left(e^{-i\alpha H} \otimes e^{i\alpha H^*}\right), \quad dH = \prod_{i=1}^{d} dH_{ii} \prod_{i>j} \frac{1}{\sqrt{2}} \Re H_i j \Im H_i j. \tag{B11}$$

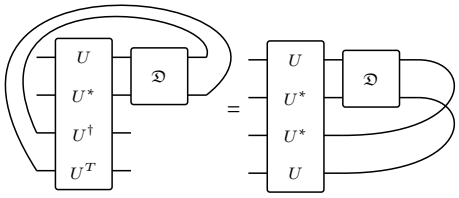
In terms of the unitary that diagonalizes $e^{i\alpha H} = U D_\lambda U^\dagger$, with

$$D_\lambda = \mathrm{diag}(e^{i\alpha\lambda_1} \cdots e^{i\alpha\lambda_d}) \tag{B12}$$

we can decompose Eq. (B11) in terms of the average with the Haar measure discussed above of the unitaries, and the average with respect to the GUE measure of the eigenvalues:

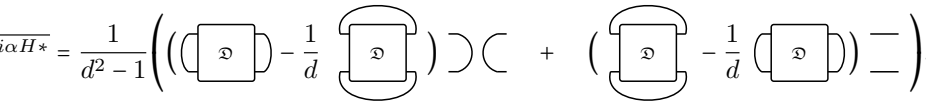
$$\overline{e^{i\alpha H} \otimes e^{-i\alpha H^*}} = \overline{(U \otimes U^*)\overline{(D_\lambda \otimes D^*_\lambda)}(U^\dagger \otimes U^T)} = \overline{(U \otimes U^*)\mathfrak{D}(U^\dagger \otimes U^T)}. \tag{B13}$$

In terms of the diagrammatic notation, we can write:

$\overline{e^{i\alpha H} \otimes e^{-i\alpha H^*}} =$ 

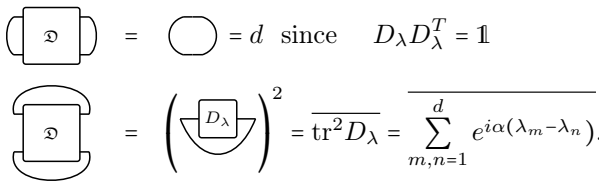
(B14)

After interchanging rows three and four, we identify the first box as the average computed above $\boldsymbol{\mathcal{R}}$ Eq. (B5). Hence, it can be shown that

$\overline{e^{i\alpha H} \otimes e^{-i\alpha H*}} = \frac{1}{d^2-1}\Big($ 

$\Big).$ (B15)

The above expression can be further simplified taking into account



(B16)

We recognize in the second row the definition of the *spectral form factor* (SFF) [49, 50]: the Fourier transform of two point function, which is the simplest correlation function that displays universality. In this context, the strength of the noise α plays the role of time. For the GUE, there is a translational invariance between eigenvalues, and we can eliminate the specific dependence on n, m:

$$f_d(\alpha) \coloneqq f_d(\lambda_n, \lambda_m, \alpha) = \int \prod_{i=1}^{d} d\lambda_i P_{\mathrm{GUE}}(\lambda_1, \ldots, \lambda_d) e^{i\alpha(\lambda_m - \lambda_n)} = \int C \int \prod_{i=1}^{d} (d\lambda_i e^{-\frac{\lambda_i^2}{2}}) \prod_{\lambda_i > \lambda_j} (\lambda_i - \lambda_j)^2 e^{i\alpha(\lambda_m - \lambda_n)},\ n \neq m, \tag{B17}$$

where P_{GUE} is GUE probability density function. We postpone the computation of this expression for the next section. Now, it is enough to write the SFF as

$$\overline{\mathrm{tr}^2 D_\lambda} = d(1 + (d-1) f_d(\alpha))$$

Thus, the average Eq. (B11) is

$\overline{e^{i\alpha H} \otimes e^{-i\alpha H^*}} = \frac{1 - f_d(\alpha)}{d+1}$  $+ \frac{d f_d(\alpha) + 1}{d+1}$  . (B18)

Finally, to find $\boldsymbol{u}_{2r-1,2r}$ Eq.(B10), we multiply Eq. (B18) (previosly adding two identities/ horizontal lines to the diagrams) and Eq. (B5):

$\overline{(e^{i\alpha H} \otimes e^{-i\alpha H^*} \otimes \mathbb{1} \otimes \mathbb{1})(R \otimes R^* \otimes R \otimes R^*)} = \frac{1}{d^2-1}\Big\{(1 + \frac{f_d(\alpha)-1}{d(d+1)})$  $+ \frac{d f_d(\alpha)+1}{d+1}\Big($ $\Big)\Big($ $- \frac{1}{d}\Big)\Big($ $- \frac{1}{d}\Big)\Big)\Big\}$

$= \frac{1}{d^2-1}\Big\{(1 + \frac{f_d(\alpha)-1}{d(d+1)}) \left\|+^{\mathbf{12;34}}\right\rangle\left\langle+^{\mathbf{12;34}}\right\| + \frac{d f_d(\alpha)+1}{d+1}\Big(\left\|+^{\mathbf{14;23}}\right\rangle\left\langle+^{\mathbf{14;23}}\right\| - \frac{1}{d}\left\|+^{\mathbf{12;34}}\right\rangle\left\langle+^{\mathbf{14;23}}\right\| - \frac{1}{d}\left\|+^{\mathbf{14;23}}\right\rangle\left\langle+^{\mathbf{12;34}}\right\|\Big)\Big\},$

where we have used the state notation Eq. (A9) and ommited the labeling of the gates, since the above result is valid for all dimensions. In terms of the spin basis 1/2 introduced above, Eq. (B7) we find

$$\overline{(e^{i\alpha H} \otimes e^{-i\alpha H^*} \otimes \mathbb{1} \otimes \mathbb{1})(R \otimes R^* \otimes R \otimes R^*)} = \|\Uparrow\rangle\langle\Uparrow\| + \frac{d f_d(\alpha)+1}{d+1}\|\Downarrow\rangle\langle\Downarrow\|. \tag{B19}$$

Observe that the average is also diagonal in this basis, but the $\Downarrow$ sector depends on $f_4(\alpha)$, and consequently the projector property that the $\boldsymbol{\mathcal{R}}$ possesses Eq. (B8) is not valid anymore.

To conclude, we shall reintroduce the labelling of the unitaries and restrict the case for $d = 4$, yielding:

$$\boldsymbol{\mathcal{V}} = \bigotimes_{r=1}^{L/2} \|\Uparrow_{2r-1,2r}\rangle\langle\Uparrow_{2r-1,2r}\| + \frac{4f_4(\alpha)+1}{5}\|\Downarrow_{2r-1,2r}\rangle\langle\Downarrow_{2r-1,2r}\| \tag{B20}$$

In the remainder of this subsection, we compute explicitly $f_d(\alpha)$.

a. *Computing $f_d(\alpha)$*

In this subsection, we finally address the computation of the function $f_d(\alpha)$, intimately connected with the spectral form factor. Whereas the large L limit is well known; see, for instance, [38, 51], the exact computation is less standard. See [52]quantity for arbitrary d and α, although in practice, we are interested in $d = 4$ and $\alpha \ll 1$. This is, in a time-evolution Hamiltonian perspective, we are interested in the non-universal regime, where the nature of the matrix ensemble takes a great relevance. For a complete introduction to the random matrix, see, for example, [51, 53]. Here we start by taking into account the definition of the n-point function:

$$\rho^{(n)}(\lambda_1, \ldots, \lambda_n) = \int d\lambda_{n+1} \ldots d\lambda_d P_{\mathrm{GUE}}(\lambda_1, \ldots, \lambda_d), \tag{B21}$$

we see that

$$f_d(\alpha) = \int \rho^{(2)}(\lambda_m, \lambda_n) e^{i\alpha(\lambda_m - \lambda_n)},\ n \neq m. \tag{B22}$$

The term $\prod_{\lambda_i > \lambda_j}(\lambda_i - \lambda_j)^2$ that appears with the Jacobian is called Vandermonde determinant $\Delta_d(\boldsymbol{\lambda})$:

$$\Delta_d^2(\boldsymbol{\lambda}) = (\lambda) \prod_{\lambda_i > \lambda_j} (\lambda_i - \lambda_j)^2 = \det\left[\{\lambda_j^{i-1}\}_{i,j=1}^d\right]^2 = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_d \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1^{d-1} & \lambda_2^{d-1} & \cdots & \lambda_d^{d-1} \end{vmatrix}^2 \tag{B23}$$

Now we can harness the properties of the determinant. That is,

$$\Delta_d(\boldsymbol{\lambda}) = \det\left[\{\lambda_j^{i-1}\}_{i,j=1}^d\right] = \det\left[\{P_{i-1}(\lambda_j)\}_{i,j=1}^d\right], \tag{B24}$$

where $P_k(x)$ $k = 0, \ldots, d-1$ are a family of monic orthogonal polynomials. In our case, the Hermite family is suitable. The reason is that they are orthogonal to the weight $e^{-\frac{x^2}{2}}$ that appears in our probability density function $P_{\rm GUE}$:

$$\int_{-\infty}^{\infty} dx e^{-\frac{x^2}{2}} H_m(x) H_n(x) = \sqrt{2\pi} n! \delta_{mn}. \tag{B25}$$

In order to play with orthonormal objects, we shall define the Hermite wavefunctions:

$$\Psi_m(x) = \frac{1}{(2\pi)^{1/4}} \frac{1}{\sqrt{m!}} e^{-\frac{x^2}{4}} H_m(x). \tag{B26}$$

Therefore, we write:

$$f_d(\alpha) = C \int \prod_{i=1}^{d} d\lambda_i \det\left[\{e^{-\frac{\lambda_j^2}{4}} H_{i-1}(\lambda_j)\}_{i,j=1}^{d}\right] \det\left[\{e^{-\frac{\lambda_j^2}{4}} H_{i-1}(\lambda_j)\}_{i,j=1}^{d}\right] e^{i\alpha(\lambda_m - \lambda_n)} \tag{B27}$$

$$= (2\pi)^{d/2} \prod_{k=1}^{d-1} k! C \int \prod_{i=1}^{d} d\lambda_i \det\big[\{\Psi_{i-1}(\lambda_j)\}_{i,j=1}^{d}\big] \det\big[\{\Psi_{i-1}(\lambda_j)\}_{i,j=1}^{d}\big] e^{i\alpha(\lambda_m - \lambda_n)}, \ n \neq m. \tag{B28}$$

Taking into account that $\det[A]\det[B] = \det[AB]$ and $\det[A] = \det\big[A^T\big]$:

$$f_d(\alpha) = (2\pi)^{d/2} \prod_{k=1}^{d-1} k! C \int \prod_{i=1}^{d} d\lambda_i \det\left[\{\sum_{k=0}^{d-1} \Psi_k(\lambda_i)\Psi_k(\lambda_j)\}_{i,j=1}^{d}\right] e^{i\alpha(\lambda_m - \lambda_n)} \tag{B29}$$

$$= C_d \int \prod_{i=1}^{d} d\lambda_i \det\big[\{K_d(\lambda_i, \lambda_j)\}_{i,j=1}^{d}\big] e^{i\alpha(\lambda_m - \lambda_n)}, \ n \neq m, \tag{B30}$$

where we have collected the constants, and

$$K_d(x,y) = \sum_{k=0}^{d-1} \Psi_k(x)\Psi_k(y) \tag{B31}$$

is the reproducing Hermite kernel. The reason is that it satisfies the property:

$$K_d(x,y) = \int_{-\infty}^{\infty} du K_d(x,u) K_d(u,y). \tag{B32}$$

This property allows the $d \times d$ determinant to be sequentially simplified by reducing its dimension. It can be shown that

$$C_d \int d\lambda_n \det\big[\{K_d(\lambda_i, \lambda_j)\}_{i,j=1}^{n}\big] = C_d (K_d(x,x) - n + 1) \det\big[\{K_d(\lambda_i, \lambda_j)\}_{i,j=1}^{n-1}\big], \quad K_d(x,x) = d. \tag{B33}$$

Hence, iterating the above expression d times, we obtain the normalization constant of the probability density function $C_d = 1/d!$, and we rewrite the 2-point function Eq. (B21)

$$\rho^{(2)}(\lambda_m, \lambda_n) = \frac{(d-n)!}{d!} \begin{vmatrix} K_d(\lambda_m, \lambda_m) & K_d(\lambda_m, \lambda_n) \\ K_d(\lambda_n, \lambda_m) & K_d(\lambda_n, \lambda_n) \end{vmatrix} \tag{B34}$$

Using the above expression in Eq. (B22) yields

$$f_d(\alpha) = \frac{(d-2)!}{d!} \iint d\lambda_m d\lambda_n \begin{vmatrix} K_d(\lambda_m, \lambda_m) & K_d(\lambda_m, \lambda_n) \\ K_d(\lambda_n, \lambda_m) & K_d(\lambda_n, \lambda_n) \end{vmatrix} e^{i\alpha(\lambda_m - \lambda_n)}$$

$$= \frac{1}{d(d-1)} \left(\left(\int dx K_d(x,x) e^{-i\alpha x} \right)^2 - \left(\int dx dy K_d(x,y) e^{-i\alpha(x-y)} \right)^2 \right). \tag{B35}$$

We have to compute the following integral:

$$\int_{-\infty}^{\infty} dx e^{-\frac{x^2}{2}} H_k(x) H_{k'}(x) e^{\pm i\alpha x}. \tag{B36}$$

We make use of the following three results:

1. Given a function $f(x)$ that vanishes at infinity, the integral

$$\int_{-\infty}^{\infty} dx e^{-\frac{x^2}{2}} H_k f(x) = \int_{-\infty}^{\infty} dx e^{-\frac{x^2}{2}} f^{(k)}(x), \quad f^{(k)}(x) = \frac{d^k}{dx^k} f(x), \tag{B37}$$

as can be seen by integrating k times by parts.

2. The generalized Leibniz rule:

$$(f \times g)^{(k)} = \sum_{l=0}^{k} \frac{k!}{l!(k-l)!} f^{(k-l)} g^{(l)}. \tag{B38}$$

3. The following recursion relation that the Hermite polynomials hold:

$$H_n^{(m)}(x) = \frac{n!}{(n-m)!} H_{n-m}(x). \tag{B39}$$

To solve the integral Eq. (B36), we first use the identity Eq. (B37) with $f(x) = H_{k'}(x) e^{\pm i\alpha x}$, following by the Leibniz Eq.(B38) rule, and finally Eq. (B39):

$$\int_{-\infty}^{\infty} dx e^{-\frac{x^2}{2}} H_k(x) H_{k'}(x) e^{\pm i\alpha x} = \sum_{l=0}^{k} \frac{k!k'!(\pm i\alpha)^{k'-k+2l}}{l!(k-l)!(k'-k+l)!} \sqrt{2\pi} e^{-\frac{\alpha^2}{2}} \tag{B40}$$

We use the above result to obtain $f_d(\alpha)$. It is interesting to distinguish between the contribution of the connected and disconnected parts of the (Fourier transform of) the two-point function. The disconnected part is obtained by particularizing $k = k'$:

$$\Big(\int_{-\infty}^{\infty} dx K_d(x,x) e^{\pm i\alpha x}\Big)^2 = \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{d-1} \frac{1}{k!} \int_{-\infty}^{\infty} dx e^{-x^2/2} H_k^2(x) e^{-i\alpha x} = e^{-\alpha^2} \Big(\sum_{k=0}^{d-1} k! \sum_{l=0}^{k} \frac{(\pm i\alpha)^{2l}}{(l!)^2 (k-l)!}\Big)^2 \tag{B41}$$

The connected part is:

$$\begin{aligned}\int_{-\infty}^{\infty} dx dy K_d(x,y) K_d(y,x) e^{\pm i\alpha(x-y)} &= \frac{1}{2\pi} \sum_{k=0}^{d-1} \sum_{k'=0}^{d-1} \frac{1}{k!k'!} \Big(\int_{-\infty}^{\infty} dx e^{-x^2/2} e^{i\alpha x} H_k(x) H_{k'}(x)\Big) \\ &\quad \times \Big(\int_{-\infty}^{\infty} dx e^{-y^2/2} e^{-i\alpha x} H_k(y) H_{k'}(y)\Big) \\ &= e^{-\alpha^2} \Big(\sum_{k=0}^{d-1} \sum_{k'=0}^{d-1} k!k'!(-1)^{k'-k+1} \alpha^{k'-k} \sum_{l=l'=0}^{k} \frac{\alpha^{2(l+l')}}{l!l'!(k-l)!(k-l')!(k'-k+l)!(k'-k+l')!}\Big).\end{aligned} \tag{B42}$$

So we arrive to the final result:

$$f_d(\alpha) = \frac{e^{-\alpha^2}}{d(d-1)} \Big\{ \Big(\sum_{k=0}^{d-1} k! \sum_{l=0}^{k} \frac{(\pm i\alpha)^{2l}}{(l!)^2 (k-l)!}\Big)^2 - \Big(\sum_{k=k'=0}^{d-1} k!k'!(-1)^{k'-k+1} \alpha^{k'-k} \Big| \sum_{l=0}^{k} \frac{(i\alpha)^{2l}}{l!(k-l)!(k'-k+l)!} \Big|^2 \Big) \Big\} \tag{B43}$$

In particular, we find that

$$f_4(\alpha) = \frac{1}{36} e^{-\alpha^2} \left(-\alpha^{10} + \frac{25\alpha^8}{2} - 64\alpha^6 + 138\alpha^4 - 144\alpha^2 + 36 \right). \tag{B44}$$

Remarkably, it can be seen that $f_d(\alpha) = 1 - (d+1)\alpha^2 + o(\alpha^3)$, so we shall propose a more manageable expression:

$$f_d(\alpha) \sim e^{-(d+1)\alpha^2}. \tag{B45}$$

The validity of this approximation can be examined in Fig. 5

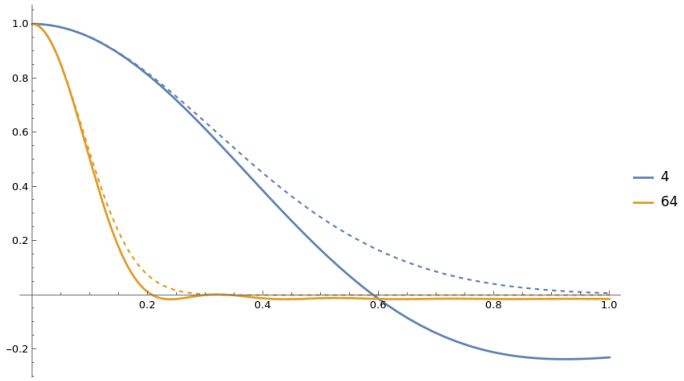
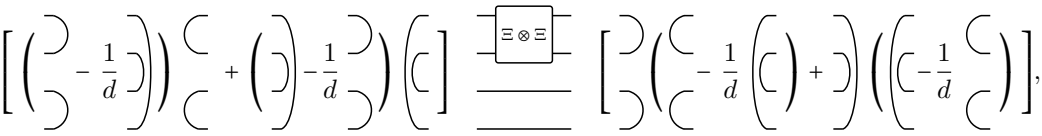


FIG. 5: $f_d(\alpha)$ for $d = 4$ and $d = 64$. The thick lines correspond to the analytic result Eq. (B43), and the dashed lines to Eq. (B45).

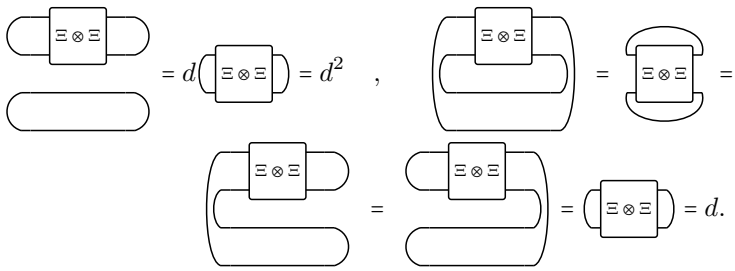
2. Computing $\mathcal{R}\,\mathcal{P}\,\mathcal{R}$

Now we shall focus on the average of the faulty permutation contribution $\mathcal{P}$. The large unitaries that *encapsulate* the permutation contribution allow us to assume that the effect of each permutation is decorrelated, making a great simplification. As a first step, we harness the left/right invariance of the Haar measure:

$$\begin{aligned}\mathcal{R}\,\mathcal{P}\,\mathcal{R} &= \left(\overline{R\otimes R^*\otimes R\otimes R^*}\right)\left(\overline{\tilde{\Pi}\otimes\tilde{\Pi}\otimes\Pi\otimes\Pi}\right)\left(\overline{R\otimes R^*\otimes R\otimes R^*}\right)\\ &= \left(\overline{R\otimes R^*\otimes R\otimes R^*}\right)\left(\overline{\tilde{\Pi}\Pi\Pi^T\otimes\tilde{\Pi}\Pi\Pi^T\otimes\mathbb{1}\otimes\mathbb{1}}\right)\left(\overline{R\otimes R^*\otimes R\otimes R^*}\right).\end{aligned} \tag{B46}$$

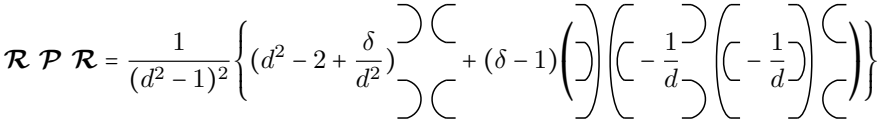
$\mathcal{R}\,\mathcal{P}\,\mathcal{R} =$  (B47)

where $\Xi = \tilde{\Pi}\Pi\Pi^T$. Using Eq. (B16) we find that



(B48)

Therefore, we find that

$\mathcal{R}\,\mathcal{P}\,\mathcal{R} =$  (B49)

Finally, we write it in terms of the orthogonal basis Eq.(B7):

$$\mathcal{R}\,\mathcal{P}\,\mathcal{R} = \|\Uparrow\rangle\langle\Uparrow\| + \frac{\delta-1}{d^2-1}\|\Downarrow\rangle\langle\Downarrow\|\,. \tag{B50}$$

Recall that $\delta = \overline{\mathrm{tr}^2\tilde{\Pi}\Pi\Pi^T}$. In the next section, we give details regarding the computation of this average and the nature itself of the faulty permutation $\tilde{\Pi}$.

3. Putting all together

We have all the ingredients to compute the average fidelity Eq. (A11), but they need to be written on a common basis: while $\boldsymbol{\mathcal{RPR}}$ is written in terms of the total spins $\Uparrow, \Downarrow$, $\boldsymbol{\mathcal{V}}$ is written in terms of local spins $\Uparrow_{2r-1,2r}, \Downarrow_{2r-1,2r}$ with $r = 1, \cdots, L/2$ as can be seen in Eq. (B20). To do so, we introduce the notation $\|m, i\rangle$ where m counts the number of spins $\Downarrow_{2r-1,2r}$ and i labels the degeneracy. We shall write all the quantities, namely $\|\Uparrow\rangle$, $\|\Downarrow\rangle$, $\boldsymbol{\mathcal{U}}$ in terms of this basis.

The easiest starting point is $\|\Uparrow\rangle$, since

$$\|\Uparrow\rangle = \otimes_{r=1}^{L/2} \|\Uparrow_{2r-1,2r}\rangle = \|0,0\rangle\,. \tag{B51}$$

The state $\|\Downarrow\rangle$ is a complicated combination consequence of being a superposition state (see Eq. (B7)) Equivalently (see Eq. (B7))

$$\begin{aligned}\|\Downarrow\rangle = \frac{1}{\sqrt{4^L-1}}\left(\bigotimes_{r=1}^{L/2}\left\|+^{\mathbf{14;23}}_{2r-1,2r}\right\rangle - \frac{1}{2^L}\bigotimes_{r=1}^{L/2}\left\|+^{\mathbf{12;34}}_{2r-1,2r}\right\rangle\right) &= \frac{1}{\sqrt{4^L-1}}\left(\bigotimes_{r=1}^{L/2}\left(\|\Uparrow_{2r-1,2r}\rangle + 15^{1/2}\,\|\Downarrow_{2r-1,2r}\rangle\right) - \|0,0\rangle\right)\\ &= \frac{1}{\sqrt{4^L-1}}\sum_{m=1}^{L/2} 15^{m/2}\sum_{i=1}^{\binom{L/2}{m}}\|m,i\rangle\,,\end{aligned} \tag{B52}$$

where in the second equality we have inverted Eq. (B7) and particularized for $d = 4$. It is important to note that the summation index starts from 1 and not from 0 in the final result.

Now we express $\boldsymbol{\mathcal{V}}$ Eq. (B20) and $\boldsymbol{\mathcal{R}}\ \boldsymbol{\mathcal{P}}\ \boldsymbol{\mathcal{R}}$ Eq. (B50) in the new basis $\{\|m,i\rangle\}$:

$$\begin{aligned}\boldsymbol{\mathcal{V}} &= \bigotimes_{r=1}^{L/2}\|\Uparrow_{2r-1,2r}\rangle\langle\Uparrow_{2r-1,2r}\| + \frac{4f_4(\alpha)+1}{5}\|\Downarrow_{2r-1,2r}\rangle\langle\Downarrow_{2r-1,2r}\| = \sum_{m=0}^{L/2}\left(\frac{4f_4(\alpha)+1}{5}\right)^m\sum_{i=1}^{\binom{L/2}{m}}\|m,i\rangle\langle m,i\|\\ \boldsymbol{\mathcal{R}}\ \boldsymbol{\mathcal{P}}\ \boldsymbol{\mathcal{R}} &= \|\Uparrow\rangle\langle\Uparrow\| + \frac{\delta-1}{d^2-1}\|\Downarrow\rangle\langle\Downarrow\| = \|0,0\rangle\langle 0,0\| + \left(\frac{\delta-1}{4^L-1}\right)\left(\frac{1}{4^L-1}\right)\sum_{m,n=1}^{L/2}15^{\frac{m+n}{2}}\sum_{i,j=1}^{\binom{L/2}{m}\binom{L/2}{n}}\|m,i\rangle\langle n,j\|\,.\end{aligned} \tag{B53}$$

Observe that the summation index in $\boldsymbol{\mathcal{V}}$ runs from 0 in this case. Hence, we are finally able to write the average of one layer:

$$\begin{aligned}\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{V}}\ \boldsymbol{\mathcal{R}}\ \boldsymbol{\mathcal{P}}\ \boldsymbol{\mathcal{R}} &= \left(\sum_{m=0}^{L/2}\left(\frac{4f_4(\alpha)+1}{5}\right)^m\sum_{i=1}^{\binom{L/2}{m}}\|m,i\rangle\langle m,i\|\right)\left(\|0,0\rangle\langle 0,0\| + \left(\frac{\delta-1}{4^L-1}\right)\left(\frac{1}{4^L-1}\right)\sum_{n,p=1}^{L/2}15^{\frac{n+p}{2}}\sum_{j,k=1}^{\binom{L/2}{n}\binom{L/2}{p}}\|n,j\rangle\langle p,k\|\right)\\ &= \|0,0\rangle\langle 0,0\| + \left(\frac{\delta-1}{4^L-1}\right)\left(\frac{1}{4^L-1}\right)\sum_{m,n=1}^{L/2}15^{\frac{m+n}{2}}\left(\frac{4f_4(\alpha)+1}{5}\right)^m\sum_{i,j=1}^{\binom{L/2}{m}\binom{L/2}{n}}\|m,i\rangle\langle n,j\|\,,\end{aligned} \tag{B54}$$

where we have used that $\langle m,i\,\|n,j\rangle = \delta_{m,n}\delta_{i,j}$. Indeed, we are able to exponentitate the above quantity thanks to the fact that the crossed terms vanish:

$$\begin{aligned}(\boldsymbol{\mathcal{U}})^T &= \|0,0\rangle\langle 0,0\| + \left[\left(\frac{\delta-1}{4^L-1}\right)\left(\frac{1}{4^L-1}\right)\sum_{m=1}^{L/2}15^m\left(\frac{4f_4(\alpha)+1}{5}\right)^m\binom{L/2}{m}\right]^{T-1}\\ &\times\left[\left(\frac{\delta-1}{4^L-1}\right)\left(\frac{1}{4^L-1}\right)\sum_{m,n=1}^{L/2}15^{\frac{m+n}{2}}\left(\frac{4f_4(\alpha)+1}{5}\right)^m\sum_{i,j=1}^{\binom{L/2}{m}\binom{L/2}{n}}\|m,i\rangle\langle n,j\|\right]\end{aligned} \tag{B55}$$

To finally find the average fidelity, we just need compute the overlaps $\left\langle+^{\mathbf{14;23}}\right\|\boldsymbol{\mathcal{U}}^T\left\|+^{\mathbf{1234}}\right\rangle$. First,

$$\left\langle+^{\mathbf{14;23}}\,\|0,0\rangle = \frac{1}{2^L}\left\langle+^{\mathbf{14;23}}\,\right\|+^{\mathbf{12;34}}\right\rangle = 1,$$

where we have used Eq. (B51) and Eq. (B6). It is straightforward to check that

$$\begin{aligned}\left\langle+^{\mathbf{14;23}}\right\|\left(\sum_{m=1}^{L/2}15^{\frac{m}{2}}\left(\frac{4f_4(\alpha)+1}{5}\right)^m\sum_{i=1}^{\binom{L/2}{m}}\|m,i\rangle\right) &= \sum_{m,n=1}^{L/2}15^{\frac{m+n}{2}}\left(\frac{4f_4(\alpha)+1}{5}\right)^m\sum_{i,j=1}^{\binom{L/2}{m}\binom{L/2}{n}}\langle n,j\,\|m,i\rangle\\ &= \sum_{m=1}^{L/2}15^m\left(\frac{4f_4(\alpha)+1}{5}\right)^m\binom{L/2}{m}\end{aligned} \tag{B56}$$

Also, from Eq. (B6) and Eq. (B7) it is clear that

$$\langle 0,0 \left\|_{+}^{\mathbf{1234}}\right\rangle = \frac{1}{2^L}\left\langle {}_{+}^{\mathbf{12;34}} \left\|_{+}^{\mathbf{1234}}\right\rangle = 1, \tag{B57}$$

and that $\langle \Downarrow_{2r-1,2r} \left\|_{+_{2r-1,2r}}^{\mathbf{1234}}\right\rangle = 3/\sqrt{15}$. Then we can fin the last overlap:

$$\left(\sum_{m=1}^{L/2} 15^{\frac{m}{2}} \left(\frac{4f_4(\alpha)+1}{5}\right)^m \sum_{j=1}^{\binom{L/2}{n}} \langle n,j\|\right)\left\|_{+}^{\mathbf{1234}}\right\rangle = \sum_{m=1}^{L/2}\left(\frac{4f_4(\alpha)+1}{5}\right)^m 3^m \binom{L/2}{m} \tag{B58}$$

Hence, we finally obtain:

$$\overline{\mathcal{F}} = \frac{1}{2^L}\Big(1+\Big[\Big(\frac{\delta-1}{4^L-1}\Big)\frac{\Big(15^{\frac{4f_4(\alpha)+1}{5}}+1\Big)^{L/2}-1}{4^L-1}\Big]^T(2^L-1)\Big) \tag{B59}$$

$$= \frac{1}{2^L}+\Big[\Big(\frac{\delta-1}{4^L-1}\Big)\Big(\frac{(12f_4(\alpha)+4)^{L/2}-1}{4^L-1}\Big)\Big]^T\Big(1-\frac{1}{2^L}\Big). \tag{B60}$$

Defining $\Delta(\alpha) = 2^L(3f_4(\alpha)+1)^{L/2}$, we get to the final result Eq. (4) of the main text.

Appendix C: Permutations implementation

In this section, we discuss the structure of permutations and their implementation. Our results let us then establish the analytical formula for the error factor δ of permutation for the fully connected architecture as a starting point to for discussing outer architectures.

1. Imperfect SWAP gates

We address the imperfections in SWAP gate operations, characterizing each swap S in a permutation Π as imprecise due to variations in impulse duration. Such an model of error is natural, if one uses $\sqrt{S}$ as the universal 2-qubit gate (instead of, for example, CNOT), in which case the S gate is a simple composition of two fundamental gates $S = \sqrt{S}\sqrt{S}$ and each of them could be performed imperfectly.

Specifically, a swap operation S is modified to $S \to S^\beta = P_S - e^{i\pi\beta}P_A$, where β follows a Gaussian distribution with a mean of 1 and variance σ, independently for each swap. P_S and P_A are projectors on symmetric and antisymmetric subspaces respectively. For any density matrix ρ the average over imperfect swapping reads

$$\int_{-\infty}^{\infty} d\beta\ S^\beta\rho(S^\beta)^\dagger\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\beta-1)^2}{2\sigma^2}} = \int_{-\infty}^{\infty} d\beta\ \left[(P_S\rho P_S)+(P_A\rho P_S)e^{i\pi\beta}+(P_S\rho P_A)e^{-i\pi\beta}+(P_A\rho P_A)\right]\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\beta-1)^2}{2\sigma^2}} =$$
$$= (P_S\rho P_S)-(P_A\rho P_S)e^{-\frac{1}{2}\pi^2\sigma^2}-(P_S\rho P_A)e^{-\frac{1}{2}\pi^2\sigma^2}+(P_A\rho P_A) = p\rho+(1-p)\ S\rho S$$

Which turns out to simplify into a probabilistic mixture with swapping probability $1-p$ and no swapping probability p, where

$$p = \frac{1}{2}(1-e^{-\frac{1}{2}\pi^2\sigma^2}) \tag{C1}$$

Thus from now on, we will use such an integrated scenario keeping in mind the original source of errors. Since the model of imperfect swapping directly corresponds to the probabilistic occurrence of swaps with $p \in [0,1/2]$ and larger values of p are futile from the perspective of quantum computer performance, we restrict ourselves to p to such a range.

This model allows us to perform various estimations of the error factor $\delta(p) = \delta$ as in (B49). In the simplest one of those, when we consider the implementation of a generic permutation of L qubit system by expanding in the following series (averaging over all permutations of L elements)

$$\delta(p) = \frac{1}{L!}\sum_{P\in\Sigma_L}\langle \mathrm{Tr}[\tilde{\Pi}(p)\Pi^\top]^2\rangle_p = \frac{1}{L!}\sum_{P\in\Sigma_L}(1-p)^{w(P)}\mathrm{Tr}[\Pi\Pi^\top]^2+\cdots > 4^L(1-p)^{w(P)} \tag{C2}$$

where P is a permutation of L qubits, Π is the corresponding operator and $\tilde{\Pi}(p)$ is a faulty realization of Π. In the second step, we extracted the cases with all swaps executed properly. Note that the only parameters are the probability p of not performing a swap and the number of swaps w necessary to implement a permutation.

The analytical expressions for the number of swaps w necessary to implement each permutation are not known for most architectures. In such cases, we may further bound Eq. (C2) further in the following way. Let m_w be a number of permutations which demand w swaps to implement, then: +

$$\delta(p) \geq \frac{4^L}{L!} \sum_{P \epsilon \Sigma_L} (1-p)^w = \frac{4^L}{L!} \sum_w m_w (1-p)^w \geq \frac{4^L}{L!} \sum_w m_w (aw+b) \tag{C3}$$

Where $f(w) = aw+b$ is a tangent line to a function $(1-p)^w$ in a point with w equal its average over all permutations $w = \overline{w}$. More precisely, $a = \ln(1-p)(1-p)^{\overline{w}_L}$, $b = (1-\ln(1-p)\overline{w}_L)(1-p)^{\overline{w}_L}$. Because $f(w) = aw+b$ is tangent to a convex function, the last inequality in (C3) is obvious. The remaining calculations are quite simple:

$$\delta(p) \geq \frac{4^L}{L!} \sum_w m_w (aw+b) = \frac{4^L}{L!} \left(a \sum_w m_w w + b \sum_w m_w \right) = \frac{4^L}{L!} (a\ L!\ \overline{w}_m + b\ L!) = a\overline{w} + b = 4^L (1-p)^{\overline{w}_L} = 4^L e^{-q\overline{w}_L} \tag{C4}$$

with $q = -\ln(1-p) \approx p \approx \frac{\pi^2 \sigma^2}{4}$ for small p. Therefore to know a lower-bound one just has to know the average number of swaps.

2. Fully connected architecture

In this section we provide a detailed computation of Eq. (6) of the main text. We consider a fully connected quantum architecture that allows arbitrary qubit interactions. In this model, any permutation can be decomposed into at most L swap operations. The decomposition process involves organizing the permutation into cycles, where swaps within different cycles can be executed simultaneously. Each cycle can be transformed into 2-cycles using one layer of swaps, with the transformation process described as follows: starting with any two adjacent elements, subsequent swaps are performed between the next nearest elements until the cycle is completed. Since a 2-cycle can be reduced to the identity with a single swap, every permutation can be implemented with exactly two layers. See the figure 6, decomposition of cycle C, and for more more detailed discussion see [40].

Moreover, because each cycle of m elements involves only $m-1$ swaps, the number of necessary swaps equals L minus the number of cycles. Fortunately, the average number of cycles is given by Harmonic numbers $H_L = \sum_{k=1}^{L} \frac{1}{k} \approx \ln L + \gamma$, where $\gamma \approx 0.577$ is the Euler number. Therefore the average number of necessary swaps grows as

$$\overline{w}_L = L - H_L \approx L - (\ln L + \gamma) \tag{C5}$$

with the dimension L.

More precisely, the distribution of the permutations of L elements with m cycles is given by the Stirling numbers $\left[{L \atop m}\right]$. The number of swaps is given by $k = L - m$ since for each cycle one swap can be omitted, as argued before.

To derive the formula for the error factor $\delta(p)$ (Eq. (5) of the main text), we aim to translate the number of omitted swaps in $Q(p)$ into the number of cycles in $Q(p)P^{-1}$ for each permutation P and its faulty realization $Q(p)$. The next step is to average the formula for the error factor over all realizations of Q and finally over all permutations P.

Let us start by considering any k element cycle C (constructed by $k-1$ swaps) from permutation P, name its unperfect realization $D(p)$ and the number of omitted swaps as $l(p)$. We also want to emphasize that in order to maintain the connection between the number of cycles in qubit permutation $\tilde{P}$ and its trace, we treat each node unmoved by $\tilde{P}$ as a 1-cycle. According to the optimal implementation of the cycle presented in figure 6 each node in the cycle C except two is connected with two "neighbouring" nodes via swaps. Thus, one can enumerate nodes in cycle C, starting from 0, starting from the node which is moved only by the second layer. Then tag the node connected to it by second-layer swap as 1, next tag the node connected to node 1 by first-layer swap as 2, next tag the node connected to node 2 in the second layer as 3 and so on, see fig 6.

Notice that in the composition C^{-1} with its faulty realization D the errors in the first layer results in corresponding swaps, each connecting nodes with numbers $2m+1$ and $2m+2$ for some integer m, "sandwiched" by the second layer. Moreover, the errors in the second layer can be rewritten as additional swaps, connecting nodes with numbers $2m+2$ and $2m+3$ for some integer m, cancelling appropriate swaps in the perfect second layer (see fig 6 first equality). In the next step, we combine two perfect second layers with the errors from the first layers sandwiched between them. This results in the shift of each extra first-layer swap from the pair nodes $2m+1$, $2m+2$ into the pair of nodes $2m$,

$2m+3$ (see fig 6 second equality, blue lines). Finally, we multiply those transformed extra first-layer swaps by the swaps corresponding the the second-layer errors. Notice that each of the modified first-layer error lowers the number of cycles in the product $C^\top D$ by one since each of them combines two 1-cycles into one 2-cycle. Moreover, each second-layer error connects two cycles into one larger cycle, because, by the previous discussion, it cannot cancel the first-layer swap.

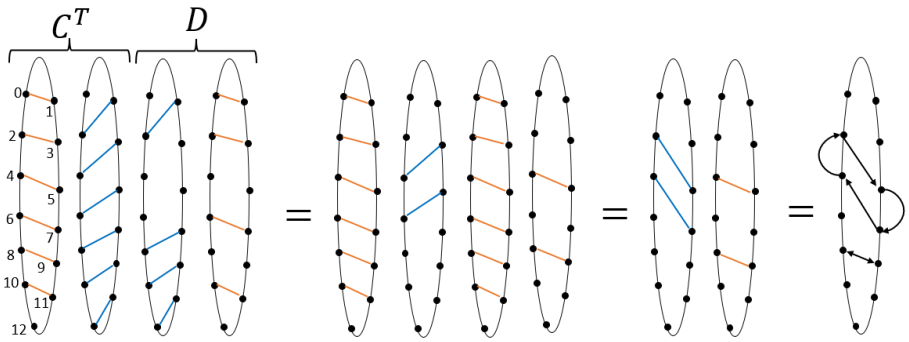


FIG. 6: Graphic representation of decomposed 13-cycle C combined with its decomposed faulty realization D. Each line corresponds to a swap and each arrow to the shift in the final product. In the first step, the errors from the first (blue) layer are combined to the extra swaps and the errors in the second (red) layer in D are decomposed by additional outer swaps. Next, the "perfect outer layer" moves the inner swaps, as described in the text and the outer errors complete them to cycles.

Thus, we established a linear dependence between the number of omitted swaps $l(p)$ and the number of cycles in $D(p)$: $m(C,p)$.

$$m(C,p) = k - l(p) = k - \sum_l \binom{k-1}{l} p^l (1-p)^l = k - (k-1)p = k(1-p) + p$$

To consider the entire permutation P one must add the $m(C,p)$ for all cycles C in P.

$$m(P,p) = \sum_{cycles} k(1-p) + p = L(1-p) + cp \tag{C6}$$

where c is a number of cycles in P. Therefore, the average error factor factorizes in the following way:

$$\delta_{full}(p) = \langle\langle \mathrm{Tr}[\Pi^\top \tilde{\Pi}(p)]^2 \rangle_p \rangle_P = \langle\langle 4^{m(P,p)} \rangle_p \rangle_P = \langle\langle \prod_{C_i} 4^{m(C_i,p)} \rangle_p \rangle_P$$

Where the inner average is taken over all false realizations of $Q(p)$ of P and the outer average is taken over all permutations P on L elements, and in the last step we used the decomposition of permutation P into its cycles C_i. Let us start with the calculation of the middle average. Consider one chosen cycle C of k elements, assuming that the positions of omitted swaps are uncorrelated:

$$\langle 4^{m(C,p)} \rangle_p = \sum_{l=0}^{k-1} \binom{k-1}{l} p^l (1-p)^{k-1-l} 4^{k-l} = 4(4-3p)^{k-1} \tag{C7}$$

Because swaps in different cycles are uncorrelated we may combine the above calculation into an average of a trace of the entire permutation:

$$\delta_{full}(p) = \langle\langle \mathrm{Tr}[\Pi^\top \tilde{\Pi}(p)]^2 \rangle_p \rangle_P = \langle \prod_{C_i} 4(4-3p)^{k_i - 1} \rangle_P = \langle 4^c (4-3p)^{L-c} \rangle_P = \cdots \tag{C8}$$

where c is the number of cycles in the permutation. Because the distribution of number of cycles is given by Stirling numbers, we follow the calculations:

$$\delta_{full}(p) = \sum_{c=1}^{L} \frac{1}{L!} \begin{bmatrix} L \\ c \end{bmatrix} 4^c (4-3p)^{L-c} = 4^{f(p)L} \sum_{c=1}^{L} \frac{1}{L!} \begin{bmatrix} L \\ c \end{bmatrix} 4^{g(p)c}$$

with

$$f(p) = \log_4(4-3p) \approx 1 - \frac{3p}{4\log(4)} - \frac{9p^2}{32\log(4)} + O(p^3) \quad \text{and} \quad g(p) = 1 - f(p) = \frac{3p}{4\log(4)} + \frac{9p^2}{32\log(4)} + O(p^3).$$

So, one clearly sees that for small p both $f(p)$ and $g(p)$ can be assumed to be linear with very small corrections for reasonably small p. For example, for the extremal value of $p = 0.5$, the corrections from all nonlinear terms sum to only $0.068\cdots$, whereas the $f(p)$ takes a value $f(0.5) \approx 0.661$.

Since from the study general bounds (C4), we know that (at least for small p) the fidelity should decay exponentially as $\delta_{full}(p) \approx 4^L e^{-\alpha(L)Lp}$, using this assumption we might derive the value of $\alpha(L)$ using equation (C8). Notice that by the above mentioned *anzatz* we can express parameter $\alpha(L)$ as:

$$\begin{aligned}
\alpha(L) &\approx -\frac{\partial \log(\delta(p))}{\partial p}\Big|_{p=0} \frac{1}{L} = -\frac{\frac{\partial \delta(p)}{\partial p}\big|_{p=0}}{\delta(0)} \frac{1}{L} = \\
&= -\frac{1}{4^L L}\left(4^{f(0)L}\log(4)Lf'(p)|_{p=0} \sum_{c=1}^{L} \frac{1}{L!}\begin{bmatrix} L \\ c \end{bmatrix} 4^{g(0)c} + 4^{f(0)} \sum_{c=1}^{L} \frac{1}{L!}\begin{bmatrix} L \\ c \end{bmatrix} 4^{g(0)c}\log(4)g'(p)|_{p=0}c\right) = \\
&= -\frac{\log(4)}{4^L L}\left(4^L L f'(0) \sum_{c=1}^{L} \frac{1}{L!}\begin{bmatrix} L \\ c \end{bmatrix} + 4^L \sum_{c=1}^{n} \frac{1}{L!}\begin{bmatrix} L \\ c \end{bmatrix} g'(0)c\right) = -\log(4)\left(f'(0) + g'(0)\frac{1}{n}\sum_{c=1}^{L} \frac{1}{L!}\begin{bmatrix} L \\ c \end{bmatrix} c\right) = \\
&= -\log(4)\left(f'(0) + g'(0)\frac{H_L}{L}\right) \approx -\log(4)\left(f'(0) + g'(0)\frac{\log L + \gamma}{L}\right) = \\
&= \frac{3}{4}\left(1 - \frac{\log L + \gamma}{L}\right)
\end{aligned}$$

Thus for large n (and reasonably small p), up to the corrections of order $O(\log(L)/L)$, the decay constant is equal $\alpha = 3/4$, so the average error factor $\delta(p)$ behaves as

$$\delta_{full}(p) \approx 4^L e^{-\frac{3}{4}pL\left(1-\frac{\log L+\gamma}{L}\right)} \approx 4^L e^{-\frac{3}{4}pL} \tag{C9}$$

The quadratic and higher corrections in p for this formula originate from the higher orders of $f(p)$ and $g(p)$ expansion and a mixed term which decays as $O(\log(L)/L)$, thus as we have shown can be neglected to a good approximation.

Appendix D: Other architectures

From now on we focus our attention on simple models of physical architectures: 1D architecture of all qubits connected in line, 2D architecture with qubits placed in a square lattice and 3D and higher dimensional architectures with qubits placed in a cubic lattice.

For each of those scenarios, we discuss the optimal, or almost optimal way to decompose permutations given the connectivity. Next given the error model discussed above, we calculate the formula for error factor δ from the formula of fidelity (B60), which for small errors is close to exact, and for larger errors gives a lower bound.

In the quantum volume test, one has the freedom to modify and optimize the circuit as long as its overall action on quantum states is the same. Thus at the end of each subsection, we present also an estimated bound for error factor $\delta(p)$ given that the permutations are not explicitly implemented but each of them is separately "optimized" during implementation.

1. Linear architecture

In this section, we discuss linear architecture with only nearest-neighbour interactions. Thus, the permutation must be implemented as a composition of swap operators $S_{i,i+1}$ between i and $i+1$ qubits. One way of decomposition of a given permutation into such swaps is the brick sort algorithm (also known as left-right sort or parallel bubble sort) [44]. This approach guarantees the minimal number of swaps involved and gives an upper bound for a number of layers (equal to a number of qubits). Hence, if necessary, we will focus on this decomposition and apply it to the generic permutation.

The distribution of the number of swaps w_s necessary to implement a typical permutation σ of n elements is known as Mahonian distribution [45], and very quickly approximates the Gaussian distribution, with average $\overline{w}_n$ and variance $\mathrm{Var}(w_s)$ equal:

$$\overline{w}_L = \frac{L(L-1)}{4}, \qquad \mathrm{Var}(w_L) = \frac{2L^3+3L^2-5L}{72} \tag{D1}$$

see [45]. The number of layers in each permutation could be bounded by the longest path in this permutation. This can be further lower bounded if we consider only right-moving paths. Then number of permutations of n elements, with such path of length k number is given by $T(n,k) = \max_i(\sigma_i - i) = T(n,k) = k!((k+1)^{n-k} - k^{n-k})$. The average length of the longest right-moving path, using Ramanujan P-function, is given by [54]

$$P(L) = \sum_{k=0}^{L-1} \frac{kT(L,k)}{k!} \approx (L-1) - \sqrt{\frac{(L-1)\pi}{2}} + O(1). \tag{D2}$$

Thus the number of layers one needs to use for the decomposition of standard permutation tends to the maximal number of layers - L.

From a computational point of view, the simplest efficient way to obtain a decomposition of a given permutation σ using a minimal number of swaps and a small number of layers is to utilize a brick sort algorithm known also as odd-even sort or parallel bubble sort [44]. One just needs to sort the permutation save the sequence of swaps and then apply it layer by layer in the inverse order on qudits. Since a brick sort may not be optimal, it gives an upper bound on a number. Moreover, since it is a parallelization of bubble sort, it can also be upper bounded by n layers.

Later we will use one more distribution $T(L,k)$ [`https://oeis.org/A324225`] of the number of signed paths of length k in all permutations of n elements:

$$T(L,k) = (L-|k|)(L-1)!\ ,$$

By convention, we will describe right-moving paths with positive k and left-moving paths with negative k. Using this distribution, one can, for example, calculate the average (unsigned) length of the path averaged over all permutations of L elements

$$\overline{n}_L = \frac{1}{L!L}\sum_k |k|T(L,k) = \frac{L^2-1}{3L}\ . \tag{D3}$$

For general architecture, the direct connection between the number of omitted swaps in $Q(p)$ and the cycles of $Q(p)P^{-1}$ is unfortunately no longer valid. This is so because, contrary to the implementation of fully connected architecture, the omitted swaps can be stacked on top of each other, so that new errors do not create new cycles in $Q(p)P^{-1}$ but modify the existing one.

Nevertheless, below we try to calculate the average Fidelity decay in scenarios where the errors are so sparse, later called *sparse error regime*, that the abovementioned phenomena occur with negligible frequency. One sees, that when there is on average more than one error in each layer: $p \geq \overline{l}/\overline{w} \approx 1/L$, the discussion below is almost certainly not valid. However, when there are on average one or fewer errors in the implementation of standard permutation: $p \leq 1/\overline{w} \approx 1/L^2$, there is a large chance that we did not exploit our assumptions too drastically.

We also emphasize that since further from sparse error regimes new omitted swaps have a smaller chance of reducing the error factor $\delta(p)$, thus the approximation we are deriving is in fact a lower bound for the error factor $\delta(p)$.

Thus, for sparse error regime (suitable small p), we assume that number of cycles in $\tilde{P}$, $m(P,p)$, is given by

$$m(P,p) \approx L - l(p)$$

where L is the number of qubits, and $l(p)$ the number of omitted swaps. Hence, one clearly sees that for $l \approx L$, so average one error in a layer, $p \approx 1/L$, this approximation cannot be true. The average overall realizations for one permutation gives

$$\langle 4^{m(P,p)}\rangle_p \approx \sum_{l=0}^{w}\binom{w}{l}p^l(1-p)^{w-l}4^{L-l} = 4^L\left(1-\frac{3p}{4}\right)^w = 4^L 4^{f(p)w} \tag{D4}$$

Where

$$f(p) = \log_4(1 - 3p/4) \approx -\frac{3p}{4\log(4)} - \frac{9p^2}{32\log(4)} + O(p^3),$$

so exactly as before taking our assumptions of small p into account we may safely assume that $f(p)$ is linear.

Since from the study of general bounds (C4), we know that (at least for small p) the error factor should behave as $\delta(p)_{1D} \approx 4^L e^{-\alpha(L)L^2 p}$. using this assumption we might derive the value of $\alpha(L)$ from the equation (D4). This results in

$$\alpha(L) \approx -\frac{\partial \log(\delta(p))}{\partial p}\Big|_{p=0} \frac{1}{L^2} = -\frac{\frac{\partial \delta(p)}{\partial p}\big|_{p=0}}{\delta(0)} \frac{1}{L^2} = -\frac{1}{4^L L^2}\left(4^L \sum_{w=0}^{L(L-1)/2} 4^{f(0)w} \rho(w) \log(4) f'(p)|_{p=0} w\right) =$$
$$= -\frac{\log(4) f'(0)}{L^2}\left(\sum_{w=0}^{L(L-1)/2} \rho(w) w\right) = -\frac{\log(4) f'(0)}{L^2} \frac{L(L-1)}{4} = \frac{3}{16}\left(1 - \frac{1}{L}\right)$$

Where $\rho(w)$ are weights coming from the Mahonian distribution, and so is the expectation value of $\overline{w}$. Thus, for large L and appropriately small $p \ll \frac{1}{L}$, up to the corrections of order $O(1/L)$, the decay constant is equal $\alpha = 3/16$, so the average fidelity behaves as

$$\delta(p)_{1D} \gtrsim 4^L e^{-\frac{3}{16} p L^2 (1 - \frac{1}{L})} \approx 4^L e^{-\frac{3}{16} p L^2} \tag{D5}$$

The above derivation can be generalized in a straight-forward way into

$$\delta(p)_{1D} \gtrsim 4^L e^{-\frac{3}{4} p \overline{w}_L} \tag{D6}$$

where $\overline{w}_L$ is the average number of swaps in permutations of L elements for a given architecture.

a. Upper bound of error factor δ for optimized permutation in 1D

As mentioned at the beginning of the section, in real-life quantum volume tests the permutations do not need to be exactly implemented. Only the performance of the entire circuit needs to agree. This does not mean, however, that we cannot leverage the information about the average behaviour of permutations and about 1D architecture.

In particular, if permutation brings together a pair of distant qubits on which a 2-qubit gate is applied, then no matter how good one's transpiler is, those qubits need to be moved to each other. Therefore to calculate *general minimal* number of swaps to implement "optimized" permutation we first need to calculate the average distance $\overline{dist}$ between two qubits that are moved to each other.

Consider a permutation represented by a permutation matrix P. Then a pair of elements that are moved to each other corresponds to a pair of rows in P and the distance between those elements is the distance between columns in which there are 1 in those two rows. Because we consider the average over all permutations, without loss of generality we can the first two rows. Moreover, for each distance - each occupied pair of columns - the number of permutations is exactly the same. Hence the formula for the distance is given by:

$$\overline{dist} = \frac{\sum_{i \neq j=1}^{L} |i - j|}{\sum_{i \neq j=1}^{L} 1} = \frac{\sum_{ij}^{L} |i - j|}{L(L-1)} = \frac{\frac{1}{3} L\left(L^2 - 1\right)}{L(L-1)} = \frac{1}{3}(L+1)$$

Finally, for each pair of cubits, the sum of distances travelled by two of those must be a least the original distance between them. Thus we can give a strict and always true bound for the average number of swaps for 1D architecture

$$\overline{w}_L \geq \lfloor L/2 \rfloor \ (\overline{dist} - 1) \approx \frac{L(L-2)}{6} \tag{D7}$$

which, using (D5) gives the following upper bound for error factor for small p:

$$\delta(p) < 4^L e^{-\frac{1}{8} p L (L-2)} \approx 4^L e^{-\frac{1}{8} p L^2} \tag{D8}$$

2. Square cube and hypercube architectures

The most natural generalization of linear architecture is square architecture, where the L qubits are arranged in a square of size $\sqrt{L} \times \sqrt{L}$, and the swaps are allowed between nearest neighbours in both axes. In this subsection, we will discuss such a case and along the way, we generalize presented results for higher dimensional cubes.

For such an arrangement, one can easily derive a lower bound of an average number of necessary swaps $\overline{w}_L$ and layers $\bar{t}_L$ to implement permutations of L elements on a square.

Firstly let's consider the average number of layers $\bar{t}_L$. The average length of the longest right-moving path in one dimension was (asymptotically) given by $(L-1)+\sqrt{(L-1)\pi/2}+O(1)$, and in two dimensions elements can move also in "top-bottom" direction, effectively skipping $\sqrt{L}$ elements at once. Thus the average longest right-bottom-path, and therefore the average number of layers to implement a permutation, is bounded from below by

$$\bar{t}_L \geq \frac{L-1}{\sqrt{L}} + \sqrt{\frac{(L-1)\pi}{2L}} + O(L^{-1/2}) = O(\sqrt{L}). \tag{D9}$$

In the case of higher-dimensional cubes of dimension d the above expression generalizes to

$$\bar{t}_L \geq \frac{L-1}{L^{1-\frac{1}{d}}} + \frac{\sqrt{(L-1)\pi}}{\sqrt{2}L^{1-\frac{1}{d}}} + O(L^{-\frac{d-1}{d}}) = O(L^{\frac{1}{d}}) \tag{D10}$$

by the analogous arguments.

To study a lower bound on the average number of swaps $\overline{w}_L$ necessary to implement permutations of L elements let us notice that each swap can decrease the sum length of all paths in a permutation ($\sum_{i=1}^{L} n_{i,P}^{2D}$) by at most 2. Thus the number of swaps in permutation π: w_π cannot be smaller than half of the sum of the length of all paths. Averaging this relation over all permutations of n elements we obtain:

$$\overline{w}_L \geq \frac{1}{L!} \sum_{P \in \Sigma_L} \frac{1}{2} \sum_{i=1}^{L} n_{i,P}^{2D} \tag{D11}$$

In the next step, we once again bound the length of the path in $2D$ architecture, by its length on a line which, using (D3), gives us:

$$\overline{w}_L \geq \frac{1}{L!} \sum_{P \in \Sigma_L} \frac{1}{2} \sum_{i=1}^{L} \frac{1}{\sqrt{L}} n_{i,P} = \frac{1}{2}\sqrt{L}\,\overline{n}_L = \frac{L^2-1}{6\sqrt{L}} = O(L^{\frac{3}{2}}) \ . \tag{D12}$$

The generalization into higher-dimensional cubes of dimension d gives us

$$\overline{w}_L \geq \frac{1}{2} L^{\frac{d-1}{d}}\,\overline{n}_L = \frac{L^2-1}{6L^{1-\frac{1}{d}}} = O(L^{1+\frac{1}{d}}) \ . \tag{D13}$$

a. *Hypercube sorting*

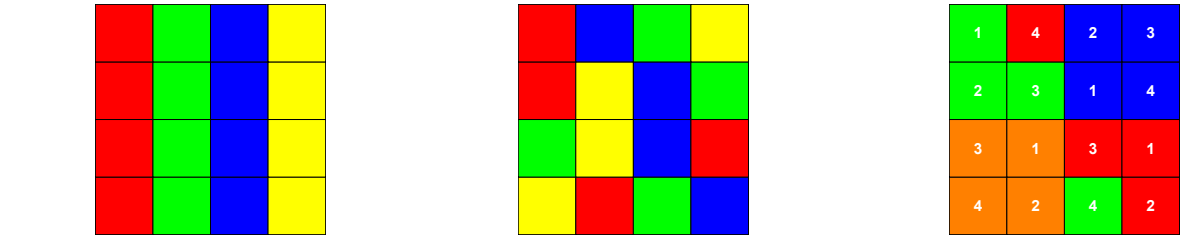
Below we describe an algorithm which gives an efficient upper bound for the average necessary number of layers and swaps simultaneously to implement a permutation using hypercube architectures. The implementation of the algorithm in the Python language is provided in the github repository. The main idea of this algorithm is aligned with [40] theorem 4.3, but our derivation is self-sustained and fully structural thanks to the properties of discussed architectures. Moreover, due to explicit construction, we can argue simultaneously about both the necessary number of swaps and layers to implement a permutation.

As we already argued the problem of implementing a permutation is equivalent to the problem of sorting the inverse of that permutation, thus we focus on the second one for convenience.

Let us start with a square. If all elements from each column were in the correct columns, one would just perform brick sort in each column, see 7, thus the maximal number of layers would be equal $\sqrt{L}$, the maximal number of swaps $\sqrt{L} \times \frac{\sqrt{L}(\sqrt{L}-1)}{2} \approx L^{3/2}/2$. If some elements are in the wrong column, but in each row, there are elements from all columns, one must first sort the rows, again by brick sort. Thus placing all elements in the correct columns, and simplifying the problem to the previous one, see 7b.

In general, however, elements which should be placed in one column are randomly scattered through the entire square, see 7c. We claim that the general case can be reduced to the one described in the above paragraph. One may mark each element in a square by natural numbers from 1 to $\sqrt{L}$ in such a way, that in each column there are all marks (without repetitions) and elements which should be in one column are marked with all marks (without repetitions). The proof that such enumeration is always possible, and an explicit algorithm for such enumeration, is placed at the end of the section for clarity. Then after sorting by marks in each column, using brick sort, the elements with marker i end up in row number i, so by the properties of enumeration we reduced the problem to the above-described.

Overall to sort a 2D square of L elements we thus did 3 times parallel brick sort - one in columns one in rows and again one in columns - giving maximally $3\sqrt{L}$ layers and no more than $3L^{3/2}/2$ swaps.



(a) (b) (c)

FIG. 7: In the picture (a), elements on a square lattice with correct columns, the original column for each element is denoted by its colour. In the middle picture (b), columns are incorrect, however, there is only one colour in each row, so one could sort it by brick sort. In the picture (c), a most complex case with appropriate marking is presented.

Now we can iteratively generalize this method to hypercubes of L elements. Similarly as above we mark the elements in d dimensional hypercube by natural numbers form 1 to $L^{\frac{1}{d}}$, such that in each 1 dimensional column there are all markers and the elements which should be in one column are marked by all the marks. Then we perform sorting over marks in each 1 dimensional column. After this sorting, by the properties of marking, in each $L^{\frac{1}{d}}$ of $d-1$ dimensional stratums there are elements with the same marks, thus in each stratum there are elements which should be placed in all columns. The next step is to perform recursive sorting in those $d-1$ dimensional stratums, after which all elements are in the correct columns, so we finish sorting by applying brick sort in each column separately.

It is a simple proof by induction that such a way of sorting gives the following bound on the maximal, hence also the average number of layers and swaps:

$$\overline{t}_L \leq L^{\frac{1}{d}}(2d-1) \quad \text{and} \quad \overline{w}_L < L^{\frac{d+1}{d}}\left(d-\frac{1}{2}\right) \tag{D14}$$

The only missing part in the above-described algorithm is the proof that appropriate enumerations can always be done. For the general case of hypercubes, we prove the following statement

Theorem 1 *Let p be a permutation of $m \times j$ elements organized in the rectangle of size $m \times j$. Then there always exists a way to mark all elements of p by the numbers from 1 to j such that in each column are all markers form 1 to j and the set of elements originated from each column has all the markers form 1 to j.*

a. Proof: If $m = 1$ all marking numbers are 1 so the theorem is trivially satisfied, so in the following, we assume $m > 1$. Each permutation p can be decomposed into a finite number of transpositions, thus we prove the theorem by induction over consecutive transpositions.

As a first inductive step let us notice that if p is an identity, there is a straightforward way of marking: the marker of each element is just its position in the column. Next, we assume that there was some correct marking for the permutation p and that the permutation p' differs from p by one extra transposition of elements.

If those elements had the same marks in p, or they originated from the same column, marking for p is also a valid marking for p'. Moreover, if those elements belong to the same column, but originated from different rows, the valid marking for p' differs from the marking for p by the same transposition.

Now, we come to the last, most complicated, case, where the transposition changing p into p' mixes the elements which originated from different columns, are in different columns, and have different markings.

To construct the new marking for p' we first copy all the marks, except those with values the same as for swapped elements. Then we start to rewrite the marks from one of two columns with exchanged elements. First, we exchange

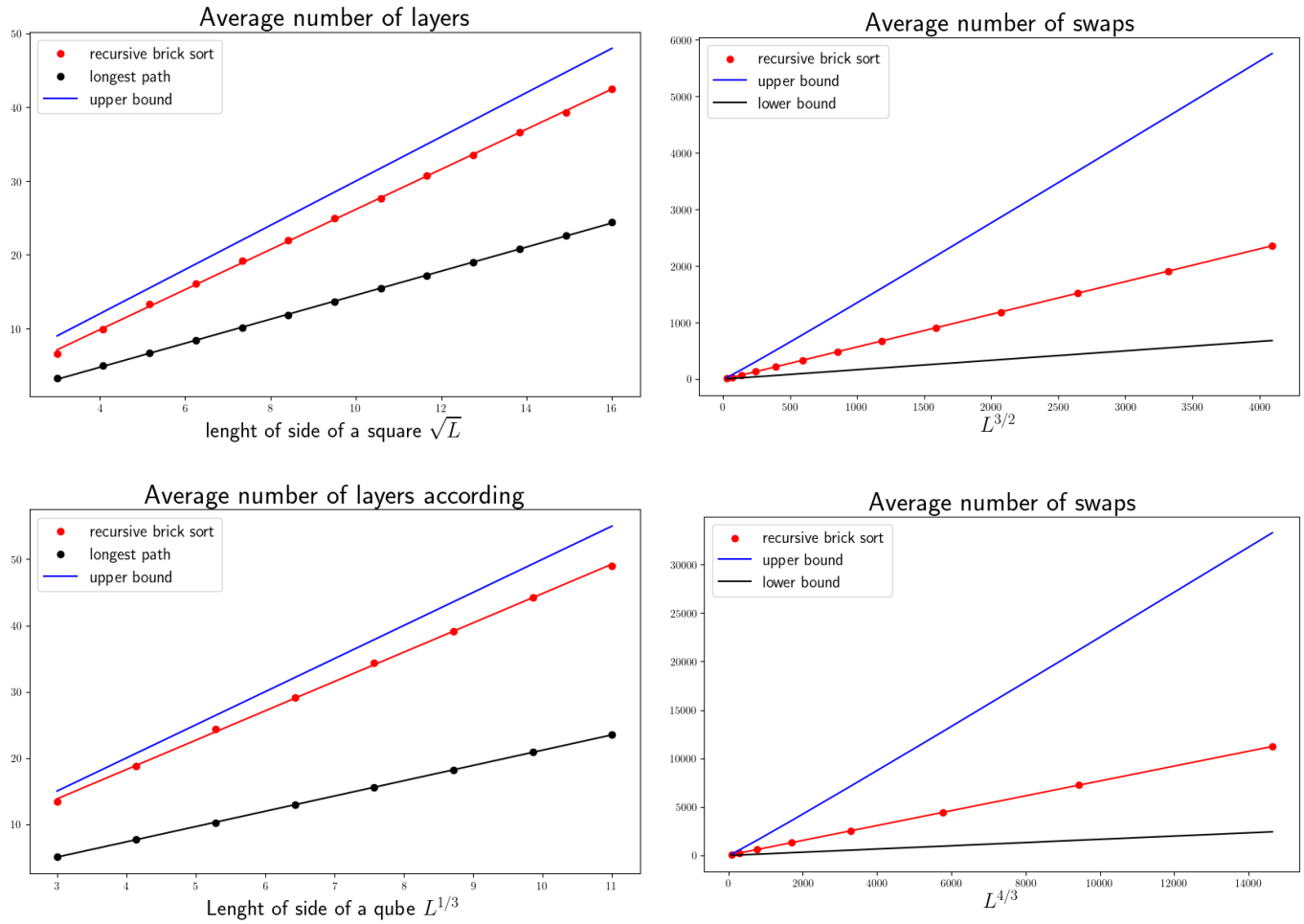


FIG. 8: Plot of the average number of layers (left) and swaps (right) to implement a permutation by the recursive version of the brick sort discussed above for the square of n elements and cube of n elements. The upper bound follows directly from the discussion within the algorithm (D14) and the lower bound from the discussion of the average length of paths (D10)(D13). Each point was obtained as an average over 1000 trials.

those marks which are the same as those of swapped elements in one of the columns. This move resolved the conflict with marks for the first swapped element without affecting the property that in each column there are all marks. However, this exchange of marks created a new error for elements which originated from the same column as the one with just exchanged marks.

So next we look for the element with the same mark from this set, identify its column in this column swap the two marks of interest. This swap resolved the above-mentioned error but potentially created a new one, thus we further proceeded in the discussed manner. Because the number of columns is limited, this procedure finishes under a finite number of steps with all the conflicts resolved. If some marks for p weren't transcribed into marks for p' in this process, the final step is to copy them without any changes. Therefore we constructed the proper marking for the permutation p'. □

The above theorem guarantees the existence of correct marking not only for permutations on a square architecture (for $m = j = \sqrt{L}$) but also on a hypercube ($m = L^{\frac{1}{d}}$, $j = L^{\frac{1-d}{d}}$) since all dimensions except the first (column) one can be flattened without affecting the properties of marking.

Hence, according to to (D6) we obtain the following bounds for error rate $\delta(p)$:

$$\delta(p)_{2\mathrm{D}} \approx 4^L e^{-\frac{9}{8}pL^{\frac{3}{2}}} \quad \text{and} \quad \delta(p)_{d\mathrm{D}} \approx 4^L e^{-\frac{3}{4}pL^{\frac{d+1}{d}}(d-\frac{1}{2})} \tag{D15}$$

b. Upper bound of error factor δ for optimized permutation in 2D and higher dimensions

For square architecture, similarly, as for 1D architecture, we may lower bound the minimal necessary number of swaps by calculating the distance between a pair of qubits that are brought together. All the arguments regarding the average over permutations from 1D case hold still, but now we consider displacement in two dimensions so the average distance between the pair of qubits is given by:

$$\overline{dist}_{2D} = \frac{\sum_{(i_1,i_2)\neq(j_1,j_2)=1}^{\sqrt{L}} |i_1 - j_1| + |i_2 - j_2|}{\sum_{(i_1,i_2)\neq(j_1,j_2)=1}^{\sqrt{L}}} = \frac{\frac{2}{3}L^{3/2}(L-1)}{L(L-1)} = \frac{2}{3}L^{1/2} \ . \tag{D16}$$

This calculation can be easily generalized to d dimensional case:

$$\overline{dist}_{dD} = \frac{\frac{d}{3}L^{2-\frac{1}{d}}\left(L^{\frac{2}{d}}-1\right)}{L(L-1)} \approx \frac{d}{3}L^{1/d} \ . \tag{D17}$$

Hence same as (D7) we can derive a lower bound on the necessary number of swaps to implement an "optimized" permutation. Which gives the following upper bounds for the error factor:

$$\delta(p)_{2\mathrm{D}} \lesssim 4^L e^{-\frac{1}{4}pL^{\frac{3}{2}}} \quad \text{and} \quad \delta(p)_{d\mathrm{D}} \lesssim 4^L e^{-\frac{d}{8}pL^{\frac{d+1}{d}}} \tag{D18}$$

Appendix E: Connection between the Heavy-output frequency and the Fidelity

In this section, we first present the definition and design of a heavy-output frequency test. Then we provide additional evidence for the connection already suggested between heavy output frequency and fidelity [27].

For a given random quantum circuit U and an input state $|\psi_0\rangle$, the output state is denoted as $|\Psi\rangle = U|\psi_0\rangle$. The basis states measured with a probability greater than the median p_{med} of all probabilities are named *heavy outputs*, constituting the *heavy output subspace* represented by H_U.

$$H_U = \{|\boldsymbol{m}\rangle \ \text{s.t.} \quad p_{\boldsymbol{m}} > p_{\mathrm{med}}\}, \tag{E1}$$

where $p_{\boldsymbol{m}} = |\langle\Psi|\boldsymbol{m}\rangle|^2$ denotes the probability of measuring a basis state $|\boldsymbol{m}\rangle$. The heavy-output probability h_U is defined as

$$h_U = \sum_{|\boldsymbol{m}\rangle\in H_U} p_{\boldsymbol{m}}. \tag{E2}$$

This concept is useful for benchmarking quantum computers. Under specific assumptions, it has been demonstrated that no classical algorithm can identify heavy outputs with a probability greater than 2/3 [21]. Consequently, a quantum device's ability to exceed this probability threshold of 2/3 may signify a quantum advantage in sampling, making it a passing criterion in the Quantum Volume test [19]. In the QV test, the heavy output subspace is identified by the classical simulation of the quantum circuit U. A real quantum device executes a corresponding faulty circuit $\tilde{U}$ that generates an outcome state $|\tilde{\Psi}\rangle = \tilde{U}|\psi_0\rangle$. The probabilities of the basis states of the heavy output subspace, determined by classical simulation of a quantum circuit, are measured leading to the faulty heavy output frequency $h_{\tilde{U}}$, which reads

$$h_{\tilde{U}} = \sum_{\boldsymbol{m}\in H_U} \tilde{p}_{\boldsymbol{m}}, \tag{E3}$$

For an ideal, error-free circuit, the asymptotic average heavy output frequency approaches $h_U \to (1+\log(2))/2 \approx 0.85$, compared to 0.5 for a completely depolarized device [21].

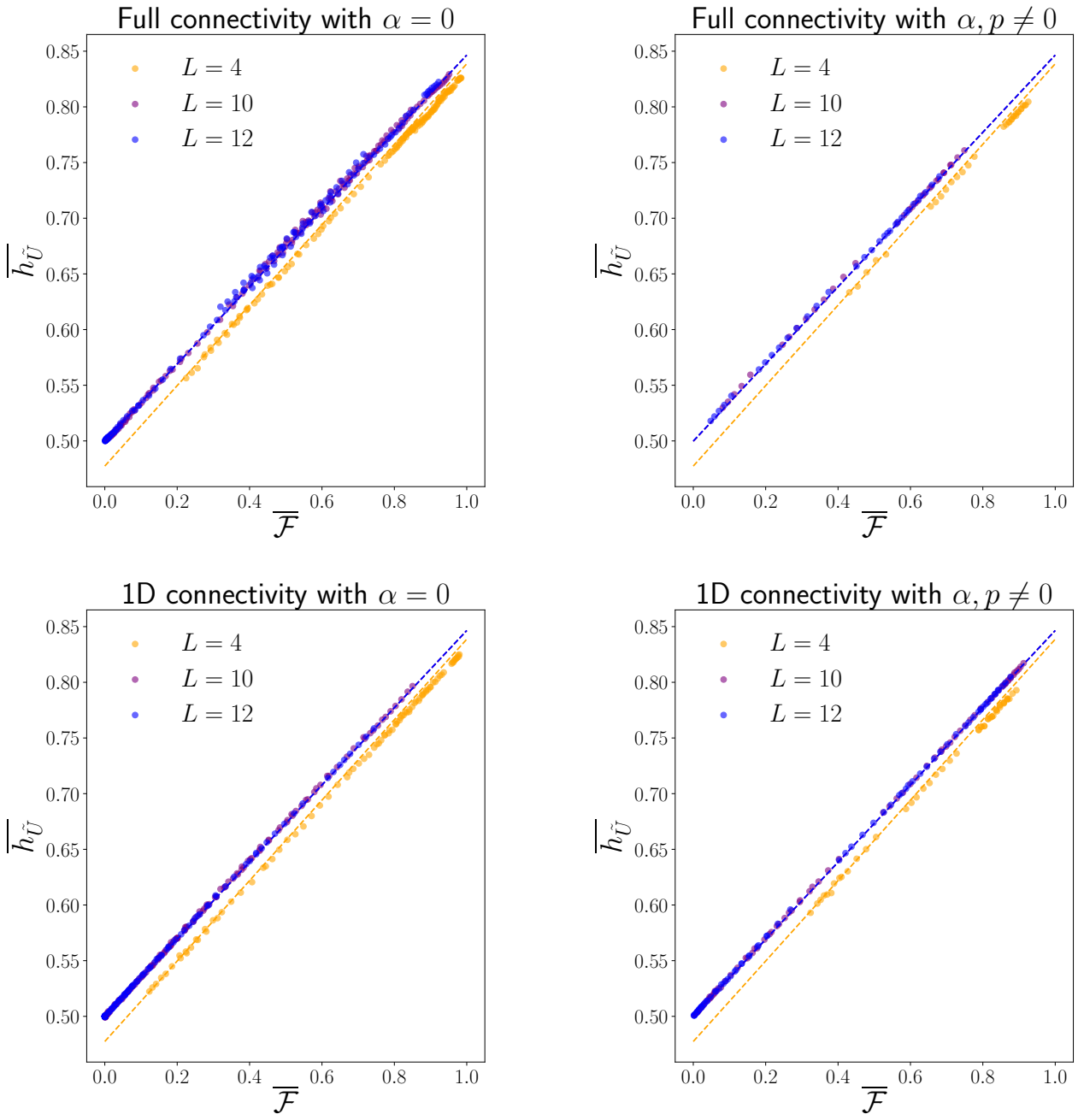


FIG. 9: Plot of the average faulty heavy output frequency as a function of the average fidelity for the two architectures considered. The data points were obtained for several layers varying the parameters α and p, and only consider data with $12 \leq T \leq 20$. Each point was obtained averaging over 5000 iterations for $L = 4$, 2000 for $L = 10$, and 1000 for $L = 12$. For $\alpha, p \neq 0$ and both for Full and 1D connectivities, the values of the parameters (α, p) were (0.003, 0.006), (0.008, 0.0048), (0.04, 0.008), (0.08, 0.008) for all system sizes. Moreover, in simulations for 1D architecture the parameters (0.001, 0.002), (0.002, 0.0012), (0.003, 0.0006), (0.0045, 0.00045) for $L = 10, 12$ were also used. In the case of $\alpha = 0$ and both architectures, the simulations were realized using the parameter $p = n \times 10^{-j}$ with $n = \{1, 2, ..., 9\}$ and $j = \{1, 2, 3\}$ and $p \leq 0.2$.

The likelihood of measuring a specific outcome can be straightforwardly determined by executing a quantum circuit multiple times. However, deriving an analytical expression for the average heavy output frequency in the presence of noise remains challenging. The expression for the average heavy output frequency can be obtained in the spirit of Eq. (A4) is

$$\overline{h_{\tilde{U}}} = \sum_{\boldsymbol{m}} \langle \boldsymbol{mm} \| \overline{\mathcal{P}_H(U_T \dots U_1) \otimes \mathcal{P}_H(U_T \dots U_1) \prod_{\tau}^{\leftarrow} \left[\tilde{U}_\tau \otimes \tilde{U}_\tau^* \right]} \| \psi_0 \psi_0 \rangle \,, \tag{E4}$$

where $|\psi_0\rangle$ is the chosen input state, the $\mathcal{P}_H$ denotes the projection on the appropriate heavy output subspace and in general is dependent both on the gates inside the circuit and the chosen input state. Due to this fact, the averages cannot be factorized even for the uncorrelated errors. Moreover, the unitarity in the circuit is lost as well making the analytical calculations unattainable in general scenarios.

However, it has been achieved for certain simple noise types, such as depolarizing noise, where the dependency was found to be linear in relation to the average fidelity [27]. Conversely, calculating the average fidelity Eq. (A4) between a state produced by a perfect circuit and one affected by noise is analytically more tractable but experimentally challenging. This discrepancy raises the question of how closely these quantities are related.

It turns out that for the discussed types of errors and architectures of the quantum volume circuit, the connection between the fidelity and heavy output frequency can also be stated as a simple function. The relation is given by the linear rescaling such that the lower and upper bounds for both quantities coincide:

$$F = 1 - \frac{2^L - 1}{2^L} \frac{h_U - h_{\tilde{U}}}{h_U - \frac{1}{2}} \,, \tag{E5}$$

where h_U is the average value of heavy output frequency obtained for the ideal scenario with no errors. The numerical evidence supporting this claim is presented in Figure 9.

We note that the same relation was obtained in the case of *global* depolarizing channel [27], which suggests that this simple behavior is general at least for isotropic noise. The fact that standard approximations of heavy output frequency, which stream from the behaviour of fidelity, repeatedly provided the expected results [19][27] additionally support this claim.

[1] J. Preskill, Quantum Computing in the NISQ era and beyond, Quantum **2**, 79 (2018).
[2] 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 near-term devices, Nature Physics **14**, 595 (2018).
[3] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).
[4] F. Pan, K. Chen, and P. Zhang, Solving the Sampling Problem of the Sycamore Quantum Circuits, Physical Review Letters **129**, 090502 (2022).
[5] I. L. Chuang and M. A. Nielsen, Prescription for experimental determination of the dynamics of a quantum black box, Journal of Modern Optics **44**, 2455 (1997).
[6] D. D. Stancil and G. T. Bird, *Principles of Superconducting Quantum Computers* (John Wiley & Sons, Inc, 2022).
[7] I. Pogorelov, T. Feldker, Ch. D. Marciniak, L. Postler, G. Jacob, O. Krieglsteiner, V. Podlesnic, M. Meth, V. Negnevitsky, *et al.*, Compact Ion-Trap Quantum Computing Demonstrator, PRX Quantum **2**, 020343 (2021).
[8] C. S. Adams, J. D. Pritchard, and J. P. Shaffer, Rydberg atom quantum technologies, Journal of Physics B: Atomic, Molecular and Optical Physics **53**, 012002 (2019).
[9] C. Kloeffel and D. Loss, Prospects for Spin-Based Quantum Computing in Quantum Dots, Annual Review of Condensed Matter Physics **4**, 51 (2013).
[10] S. Takeda and A. Furusawa, Toward large-scale fault-tolerant universal photonic quantum computing, APL Photonics **4**, 060902 (2019).
[11] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, Physical Review A **77**, 012307 (2008).
[12] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and Robust Randomized Benchmarking of Quantum Processes, Physical Review Letters **106**, 180504 (2011).
[13] E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, Physical Review A **85**, 042311 (2012).
[14] A. Carignan-Dugas, J. J. Wallman, and J. Emerson, Bounding the average gate fidelity of composite channels using the unitarity, New Journal of Physics **21**, 053016 (2019).
[15] A. W. Cross, E. Magesan, L. S. Bishop, J. A. Smolin, and J. M. Gambetta, Scalable randomised benchmarking of non-Clifford gates, npj Quantum Information **2**, 1 (2016).
[16] J. Helsen, I. Roth, E. Onorati, A. Werner, and J. Eisert, General Framework for Randomized Benchmarking, PRX Quantum **3**, 020357 (2022).

[17] 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).

[18] N. Moll, P. Barkoutsos, L. S. Bishop, J. M. Chow, A. Cross, D. J. Egger, S. Filipp, A. Fuhrer, J. M. Gambetta, *et al.*, Quantum optimization using variational algorithms on near-term quantum devices, Quantum Science and Technology **3**, 030503 (2018).

[19] 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).

[20] L. S. Bishop, S. Bravyi, A. Cross, J. M. Gambetta, and J. Smolin, Quantum Volume, Storage Consortium (2017).

[21] S. Aaronson and L. Chen, Complexity-Theoretic Foundations of Quantum Supremacy Experiments (2016), arxiv:1612.05903 [quant-ph].

[22] IBM Achieves a New Quantum Volume Level of 128 (2020).

[23] P. Jurcevic, A. Javadi-Abhari, L. S. Bishop, I. Lauer, D. F. Bogorin, M. Brink, L. Capelluto, O. Günlük, T. Itoko, *et al.*, Demonstration of quantum volume 64 on a superconducting quantum computing system, Quantum Science and Technology **6**, 025020 (2021).

[24] J. M. Pino, J. M. Dreiling, C. Figgatt, J. P. Gaebler, S. A. Moses, M. S. Allman, C. H. Baldwin, M. Foss-Feig, D. Hayes, *et al.*, Demonstration of the trapped-ion quantum CCD computer architecture, Nature **592**, 209 (2021).

[25] E. Pelofske, A. Bärtschi, and S. Eidenbenz, Quantum Volume in Practice: What Users Can Expect From NISQ Devices, IEEE Transactions on Quantum Engineering **3**, 1 (2022).

[26] I. P. Galanis, I. K. Savvas, and G. Garani, Experimental Approach of the Quantum Volume on Different Quantum Computing Devices, in *Intelligent Distributed Computing XIV*, edited by D. Camacho, D. Rosaci, G. M. L. Sarné, and M. Versaci (Springer International Publishing, Cham, 2022).

[27] 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).

[28] R. LaRose, A. Mari, V. Russo, D. Strano, and W. J. Zeng, Error mitigation increases the effective quantum volume of quantum computers (2022), arxiv:2203.05489 [quant-ph].

[29] Y. Zhang, D. Niu, A. Shabani, and H. Shapourian, Quantum Volume for Photonic Quantum Processors, Physical Review Letters **130**, 110602 (2023).

[30] Achieving Quantum Volume 128 on the Honeywell Quantum Computer.

[31] Honeywell Sets New Record For Quantum Computing Performance.

[32] Quantinuum H-Series quantum computer accelerates through 3 more performance records for quantum volume: 217, 218, and 219.

[33] S. T. Flammia and Y.-K. Liu, Direct Fidelity Estimation from Few Pauli Measurements, Physical Review Letters **106**, 230501 (2011).

[34] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms, Advanced Quantum Technologies **2**, 1900070 (2019).

[35] E. Knill, Approximation by Quantum Circuits (1995), arxiv:quant-ph/9508006.

[36] B. Collins, S. Matsumoto, and J. Novak, The Weingarten Calculus, Notices of the American Mathematical Society **69**, 1 (2022).

[37] M. P. Fisher, V. Khemani, A. Nahum, and S. Vijay, Random Quantum Circuits, Annual Review of Condensed Matter Physics **14**, 335 (2023).

[38] J. Liu, Spectral form factors and late time quantum chaos, Physical Review D **98**, 086026 (2018).

[39] K. Życzkowski and H.-J. Sommers, Average fidelity between random quantum states, Physical Review A **71**, 032313 (2005).

[40] N. Alon, F. R. K. Chung, and R. L. Graham, Routing permutations on graphs via matchings, in *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing - STOC '93* (ACM Press, San Diego, 1993).

[41] F. Wagner, A. Bärmann, F. Liers, and M. Weissenbäck, Improving Quantum Computation by Optimized Qubit Routing, Journal of Optimization Theory and Applications **197**, 1161 (2023).

[42] É. Bonnet, T. Miltzow, and P. Rzążewski, Complexity of Token Swapping and Its Variants, Algorithmica **80**, 2656 (2018).

[43] A. M. Childs, E. Schoute, and C. M. Unsal, Circuit Transformations for Quantum Architectures, in *Leibniz International Proceedings in Informatics (LIPIcs)*, Vol. 135 (Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019).

[44] S. Lakshmivarahan, S. K. Dhall, and L. L. Miller, Parallel Sorting Algorithms, in *Advances in Computers*, Vol. 23, edited by M. C. Yovits (Elsevier, 1984).

[45] E. R. Canfield, S. Janson, and D. Zeilberger, The Mahonian probability distribution on words is asymptotically normal, Advances in Applied Mathematics Special Issue in Honor of Dennis Stanton, **46**, 109 (2011).

[46] `https://doi.org/10.54499/QuantERA/0003/2021`.

[47] J. A. Miszczak and Z. Puchała, Symbolic integration with respect to the Haar measure on the unitary groups, Bulletin of the Polish Academy of Sciences: Technical Sciences; 2017; 65; No 1; 21-27 (2017).

[48] K. Poland, K. Beer, and T. J. Osborne, No Free Lunch for Quantum Machine Learning (2020), arxiv:2003.14103 [quant-ph].

[49] L. Leviandier, M. Lombardi, R. Jost, and J. P. Pique, Fourier Transform: A Tool to Measure Statistical Level Properties in Very Complex Spectra, Physical Review Letters **56**, 2449 (1986).

[50] E. Brézin and S. Hikami, Spectral form factor in a random matrix theory, Physical Review E **55**, 4067 (1997).

[51] M. L. Mehta, *Random Matrices*, 3rd ed., Vol. Volume 142.

[52] A. del Campo, J. Molina-Vilaplana, and J. Sonner, Scrambling the spectral form factor: Unitarity constraints and exact results, Physical Review D **95**, 126008 (2017).

[53] F. Haake, *Quantum Signatures of Chaos*, Springer Series in Synergetics, Vol. 54 (Springer, Berlin, Heidelberg, 2010).

[54] M. Bouvel, L. Cioni, and L. Ferrari, Preimages under the Bubblesort Operator, The Electronic Journal of Combinatorics , P4.32 (2022).