Noise-resistant quantum state compression readout

Chen Ding\(^1\), Xiao-Yue Xu\(^1\), Yun-Fei Niu\(^1\), Shuo Zhang\(^1\), Wan-Su Bao\(^1\,*\), and He-Liang Huang\(^1,2,3,4\)*

\(^1\)Henan Key Laboratory of Quantum Information and Cryptography, Zhengzhou 450000, China; \(^2\)Hefei National Research Center for Physical Sciences at the Microscale and School of Physical Sciences, University of Science and Technology of China, Hefei 230026, China; \(^3\)Shanghai Research Center for Quantum Science and CAS Center for Excellence in Quantum Information and Quantum Physics, University of Science and Technology of China, Shanghai 201315, China; \(^4\)Hefei National Laboratory, University of Science and Technology of China, Hefei 230088, China

Received July 22, 2022; accepted September 23, 2022; published online February 2, 2023

Qubit measurement is generally the most error-prone operation that degrades the performance of near-term quantum devices, and the exponential decay of readout fidelity severely impedes the development of large-scale quantum information processing. Given these disadvantages, we present a quantum state readout method, named compression readout, that naturally avoids large multi-qubit measurement errors by compressing the quantum state into a single qubit for measurement. Our method generally outperforms direct measurements in terms of accuracy, and the advantage grows with the system size. Moreover, because only one-qubit measurements are performed, our method requires solely a fine readout calibration on one qubit and is free of correlated measurement error, which drastically diminishes the demand for device calibration. These advantages suggest that our method can immediately boost the readout performance of near-term quantum devices and will greatly benefit the development of large-scale quantum computing.

quantum compression readout, qubit measurement, error mitigation, quantum computing, noisy intermediate-scale quantum

PACS number(s): 03.67.Ac, 03.67.Lx, 03.67.Pp

1 Introduction

The physical implementations of quantum computing have developed tremendously [1-14]. In particular, the milestone of quantum computational advantage (also known as quantum supremacy) has been reached using superconducting and optical systems [2-4, 15], which marks a tendency of accelerating development in quantum computing. Presently, with the capacity to implement near-term quantum computing applications, such as quantum machine learning [16-22], cloud quantum computing [23-27], and quantum simulation [28-34], we have entered the noisy intermediate-scale quantum era.

Several hurdles must be solved before quantum computing becomes widespread and reaches its full potential. Maintaining high-fidelity quantum operations while increasing the number of qubits is the main task of current quantum computing. To run large applications, the sources of noise, in-
including gate error, readout error, decoherence, and cross-talk, must be carefully suppressed. Taking a state-of-the-art (SOTA) quantum processor Zuchongzhi [4] as an example, the final fidelity of a random quantum circuit with 56 qubits and 20 cycles is only 0.0672%, although the quantum operation has reached a high fidelity (the single-qubit gate error is 0.14%, the two-qubit gate error is 0.59%, and the readout error is 4.52%). Moreover, we find that the readout error, which is the main error of this processor, is an order of magnitude higher than the gate error. Because the fidelity exponentially decays as the number of qubits increases, and readout is inevitable to obtain results in quantum computing, reducing readout error is essential for large-scale quantum computing implementation.

Generally, noisy state readout is modeled as probabilistic transitions \( p_{\text{exp}} = T p_{\text{ideal}} \) among the basis states, where \( p_{\text{ideal}} \) is the ideal readout populations, \( p_{\text{exp}} \) is the experimentally measured results, \( T \) is the transition matrix, and \( T_{ij} \) represents the transition probability from state \( |i⟩ \) to \( |j⟩ \). The transition matrix error mitigation (TMEM) methods [35-39] infer \( T \) by calibration and apply its inverse \( T^{-1} \) on \( p_{\text{exp}} \). For multi-qubit readout, a large part of the noise can be suppressed by mitigating the readout noise of each qubit. However, as the processor’s system size continuously grows and the qubit-qubit couplings become more complex, the correlated readout noise among qubits becomes nonnegligible. To completely suppress the noises, a general method needs an exponential calibration effort to determine the binary matrix \( n \times n \) transition matrix \( n \) is the system size), which limits the scalability of error mitigation. Some efficient TMEM methods are hence proposed to speedup the calibration [37-39]. Nevertheless, these methods assume the locality of correlated readout noises, and thus an amount of accuracy is sacrificed. Furthermore, as the processor’s system size increases, the statistical error and system stability degradation error will grow during the calibration procedure.

In this work, we propose a quantum state readout method to avoid the large readout noise in multi-qubit measurements. The method compresses the quantum state into one qubit and recovers the state amplitude populations from the one-qubit measurement results. Thus, the task of measuring a multi-qubit state is reduced to a single-qubit readout, and the issue of readout fidelity exponentially decaying with the number of qubits is alleviated by lowering the decay rate. Because only the one qubit to be measured needs to be fine-calibrated, large-scale readout calibration is no longer necessary. Meanwhile, correlated readout noises are also naturally avoided. We rigorously prove the noise resistance of compression readout compared to that of direct readout. Numerical experiments show its practicality in its implementation on near-term quantum processors.

2 Results

2.1 Algorithm

Given copies of the \( n \)-qubit quantum state \( |ψ⟩ = \sum_{i=0}^{2^n-1} \alpha_i |i⟩ \), the compression readout algorithm first encodes the amplitude information \( \alpha_i \) into an ancilla qubit through a specially controlled rotation scheme so that the probability of the ancilla qubit being \( |0⟩ \) is a Fourier series with \( \alpha_i \) as coefficients. Then, we measure this ancilla qubit to estimate the \( |0⟩ \) probability. Finally, we recover all \( |\alpha_i|^2 \) according to a trapezoidal rule for discrete integration of the Fourier series. The specific steps are as follows (see Figure 1).

Step 1 Initialization. Preset the number of measurement shots \( N \) on each grid point. Set the integration grid values as \( x_k = \frac{2\pi}{2m+1}, k = 1, \ldots, m \), in which the number of grids is \( m = 2^n - 1 \).

Step 2 Encoding. For \( k \in [1, \ldots, m] \), do the following:
(i) Introduce an ancilla qubit \( |0⟩ \) and perform the controlled rotation operation on \( |ψ⟩(0) \) as:

\[
|ψ⟩(0) \rightarrow \sum_{i=0}^{2^n-1} \alpha_i \left( \cos i x_k |i⟩|0⟩ + \sin i x_k |i⟩|1⟩ \right).
\]

The corresponding circuit is shown in the lower left corner of Figure 1.

(ii) Measure the ancilla qubit in the computational basis for \( N \) shots and yield \( \tilde{A}(x_k) \) as the estimate of \( A(x_k) \), the \( |0⟩ \) probability in the qubit.

Step 3 Decoding. Output the estimators of the populations \( |\alpha_i|^2 \) as:

\[
p_0 \leftarrow \frac{1}{2m+1} \left( 1 - 2m + 4 \sum_{k=1}^{m} \tilde{A}(x_k) \right),
\]

\[
p_{\neq 0} \leftarrow \frac{4}{2m+1} \left( 1 + 2 \sum_{k=1}^{m} \tilde{A}(x_k) \cos 2ix_k \right).
\]

Theorem 1 states the correctness of our algorithm (see the Supporting Information for rigorous proof).

Theorem 1 (Correctness) Given a \( n \)-qubit quantum state \( |ψ⟩ = \sum_{i=0}^{2^n-1} \alpha_i |i⟩ \) with

\[
N \geq \frac{48m^2 + 4m(2m + 1)\epsilon}{(2m + 1)^2\epsilon^2} \log \left( \frac{m}{\eta} \right),
\]

the compression readout algorithm yields estimates of all populations \( |\alpha_i|^2 \) with accuracy \( 1 - \epsilon \) and a success probability of at least \( 1 - \eta \). Mathematically, \( \forall i = 0, \ldots, 2^n - 1 \),

\[
P_i \left( |p_i - |\alpha_i|^2 | \geq \epsilon \right) \leq \eta,
\]

in which \( p_i \) are the estimates.
The correctness of the algorithm can be roughly understood in this way. After the controlled rotation in Step 2, the probability of yielding $|0\rangle$ in the ancilla qubit is

$$A = \sum_{j=0}^{2^n-1} |\alpha_j|^2 \cos^2 jx,$$

$$(5)$$

$${=} \frac{1}{2} (1 + |\alpha_0|^2) + \frac{1}{2} \sum_{j=1}^{2^n} |\alpha_j|^2 \cos 2jx,$$

$$(6)$$

$${=} A(x),$$

$$(7)$$

which is a Fourier series with variable $x$. We can recover its coefficients $|\alpha_j|^2$ by the inverse Fourier transform

$$|\alpha_j|^2 = \begin{cases} \frac{4}{\pi} \int_0^\pi A(x)dx - 1, & j = 0, \\ \frac{4}{\pi} \int_0^\pi A(x) \cos 2jxdx, & j \neq 0. \end{cases}$$

We use a trapezoidal rule to calculate this integration, which is exactly eqs. (2) and (3).

### 2.2 Computational cost

We analyze the complexity of classical data processing and the required number of measurement shots, showing that compression readout has at most a polynomial overhead compared to direct readout.

The classical postprocessing (the inverse Fourier transform $F^{-1}$ in eqs. (2) and (3)) can be reformulated as:

$$p' = \begin{pmatrix} 1 & \ldots & 1 \\ \cos 2x_1 & \ldots & \cos 2x_m \\ \vdots & \ddots & \vdots \\ \cos 2mx_1 & \ldots & \cos 2mx_m \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{pmatrix},$$

in which

$$p' = \begin{pmatrix} 1 & \frac{1}{4}((2m+1)p_0+2m-1) \\ \frac{1}{4}((2m+1)p_1-4) \\ \vdots \\ \frac{1}{4}((2m+1)p_m-4) \end{pmatrix}.$$
the relation between the number of shots and the accuracy and compare it to that of direct readout, we provide Theorem 2 as follows (see proof in the Supporting Information).

**Theorem 2** (The required number of measurement shots) Given a $n$-qubit quantum state, for the statistical error of estimator $p_i, i = 0, \ldots, 2^n - 1$ less than $\epsilon$, the required number of measurement shots $N$ on each grid point for compression readout is

$$N = O\left(\frac{1}{m\epsilon^2}\right),$$

where $m$ is the number of grids. Thus, the total number of measurement shots is $mN = O\left(\frac{1}{\epsilon}\right)$.

In comparison, to achieve a statistical error of estimator $p_i, i = 0, \ldots, 2^n - 1$ less than $\epsilon$, direct readout needs $O\left(\frac{1}{\epsilon}\right)$ measurement shots [43], which is on the same order of magnitude as compression readout. We note that the above scaling is yielded in the noiseless case. Considering noises, direct readout may require much more computational effort to match compression readout, as shown in the numerical experiments.

### 2.3 Noise resistance

We compare the performances of compression readout and direct readout in the presence of noise. We adopt a simple and common bit-flip model [36, 44] for the measurement error: during each shot in the measurement, the readout of each qubit has a probability $\xi$ of flipping its value ($|0\rangle$ goes to $|1\rangle$, $|1\rangle$ goes to $|0\rangle$), and the infidelity of two-qubit gates is $\gamma$. Given an arbitrary $n$-qubit state $|\psi\rangle = \sum \alpha_i |i\rangle$, we consider the error of the readout methods as the total variation distance between the read distribution $p(\psi)$ and the original populations $a = (|\alpha_i|^2)_{i=0,\ldots,2^n-1}$

$$E(\psi) := \|p(\psi) - a\|_{TV} = \frac{1}{2} \sum_{j=0}^{2^n-1} \left| p_j(\psi) - |\alpha_j|^2 \right|.$$ 

**Theorem 3** (Noise resistance) Assume that the measurement error (symmetric bit-flip model) is $\xi$, and the infidelities of two-qubit gates $\gamma$ are near zero. For any $n$-qubit pure state $|\psi\rangle$, the error of compression readout is

$$E_{\text{compression}}(\psi) = O(\xi + \sqrt{n}\gamma).$$

In comparison, given an arbitrary basis state $|i\rangle$, direct readout has an error of

$$E_{\text{direct}}(i) = \|O^{\text{sys}} \cdot i\|_{TV} = \frac{1}{2} \left( 1 - (1 - \xi)^n + \sum_{k=1}^{n} \left( \frac{n}{\xi} \right) (1 - \xi)^{n-k} \xi^k \right) = O(n\epsilon),$$

in which $\epsilon_i$ is the natural basis vector in $\mathbb{R}^{2^n}$, $Q = \begin{pmatrix} 1 - \xi & \xi \\ \xi & 1 - \xi \end{pmatrix}$ is the population transition matrix in each qubit. If the error rates $\xi, \gamma$ are near zero and the system size $n$ is large, $E_{\text{direct}}$ is clearly larger than $E_{\text{compression}}$ for any basis state $|i\rangle$. Therefore, we find compression readout is advantageous for improving the readout accuracy of the basis states of large systems. Moreover, in the Supporting Information, we evaluate $E_{\text{direct}}$ for any pure state $|\psi\rangle$ and show that the advantage of compression readout generally exists.

### 2.4 Numerical results

We simulate the compression readout and direct readout to compare their performance with the near-term quantum device error rates. To simulate the gate errors, we apply the “worst-case” noise channel—the depolarizing channels [45] following each gate in our simulation. Additionally, we apply individual symmetric bit-flip error models for readout in each qubit, as introduced in the assumptions of Theorem 3. By simulating the asymmetric bit-flip measurement error model and the circuit compilation on the nearest-neighbor architectures, the extended numerical results are provided in the Supporting Information.

Figure 2 shows the errors $E_{\text{direct}}$ and $E_{\text{compression}}$ of the two readout methods given various quantum state inputs with increasing system size $n$. We take the measurement error rate $\xi = 0.0452$ and depolarizing probability $\gamma = 0.0063$, converted from error values in the published data of the SOTA quantum processor Zuchongzhi 2.0 [4], as a rough imitation of its system noise. We first simulate the two algorithms in the case of infinite shots, with $n$ ranging from 2 to 1000. The results are shown in Figure 2(a) and (b). Then, we run the algorithms with $10^6$ total measurement shots. The results are shown in Figure 2(c) and (d). We find that the errors of these two methods increase with the system size because of the accumulation of readout and gate errors, while the compression readout is clearly more robust against system expansion. Another phenomenon is that the errors of reading random quantum states are lower than those of reading the states $|1\rangle^{\otimes n}$, which is mainly due to the readout noise causing interchanges of probability among all the measurement results. As mentioned in our theoretical analysis, a random state is more balanced than $|1\rangle^{\otimes n}$. See the Supporting Information for the results with error rates of other SOTA quantum processors.

Furthermore, Figure 3 shows the errors of the two readout methods with various numbers of total measurement shots.
Figure 2 (Color online) Errors of compression readout and direct readout with increasing system size $n$. (a) and (b) are the results for the readout of the states $|1\rangle^{\otimes n}$ and $|0\rangle^{\otimes n}$ in the case of infinite shots. (c) and (d) are the results for the readout of the state $|1\rangle^{\otimes n}$ and random quantum states (sampled from the Haar distribution), with $10^6$ total measurement shots. In (c) and (d), the lines show the average errors of the two methods over ten experiments, the bands show the standard errors of the means, and the circles show the errors in the case of infinite shots for comparison.

Figure 3 (Color online) Errors of compression readout and direct readout with an increasing number of total measurement shots. (a) Readout of 3-qubit state $|1\rangle^{\otimes 3}$; (b) readout of 3-qubit random states; (c) readout of reading 6-qubit state $|1\rangle^{\otimes 6}$; (d) readout of 6-qubit random states, where we also show the magnified plot of a local region for clear comparison. In all subfigures, the lines show the average errors of the two methods over ten experiments, the bands show the standard errors of the means, and the dashed lines show the errors in the cases of infinite shots.
We find that both readout methods’ errors converge to their minima (which are exactly the errors in the case of infinite shots) as the number of total measurement shots increases. Although direct readout may need fewer total measurement shots to converge to its minimum, it needs many more total measurement shots to match the high accuracy of compression readout. As suggested in Figure 3(d), to achieve the same accuracy level of 0.09, direct readout needs approximately 9-fold more \((3.5 \cdot 10^6\) compared with \(3.7 \cdot 10^5\)) total measurement shots than compression readout. The required number of shots for compression readout to outcompete direct readout seems to increase with system size \(n\). However, with \(n\) increasing, maintaining a high readout fidelity itself would require an increasing number of measurement shots. We show in the Supporting Information that the readout error increases with the system size when compression readout slightly outcompetes direct readout, suggesting that the number of shots at the outcompeting points is far fewer than the number required for high-fidelity readout. Thus, compared to direct readout, applying compression readout and obtaining the high-fidelity readout advantage usually requires no extra effort (see detailed analysis in the Supporting Information).

Figure 4 shows the advantage span of compression readout with near-term device error rates, where we marked five SOTA quantum processors: Zuchongzhi 2.0 [4], Zuchongzhi 2.1 [46], Sycamore in 2019 [2], Sycamore in 2021 [47], and System Model H1-2 [48]. We find that the compression readout algorithm outperforms direct readout when the readout error is relatively higher than the gate error. In fact, for most superconducting quantum processors, the readout error is an order of magnitude higher than the gate error. Thus, the performance of current noisy quantum devices could be boosted immediately by our method, as suggested in the maps by the red points in the advantage area (the yellow-blue area under the lines). For the ion-trap quantum processor system model H1-2, its readout error rate (0.0039) and two-qubit gate error rate (0.002453) are relatively close. In this case, the advantage of compression readout emerges as system size increases. In all cases, the advantage and the advantage area of compression readout expand rapidly when system size \(n\) increases. The results in Figures 2-4 indicate that compression readout is well suited for large-scale quantum state readout.

**Figure 4** (Color online) Advantage area map of compressing readout, which shows the ratios of the compression readout errors to the direct readout errors, with varying error rates. Each lattice shows the ratio of average error \((E_{\text{direct}}/E_{\text{compression}})\) over 10 experiments with \(10^6\) total measurement shots. (a) Readout of 3-qubit state \(|1\rangle^3\); (b) readout of 3-qubit random states; (c) readout of 6-qubit state \(|1\rangle^6\); (d) readout of 6-qubit random states. In the maps, the red contour lines show the error rates in which the ratio equals 1. The area under the line is where the compression readout shows its advantage. The red points mark the locations of the five SOTA quantum processors.
3 Discussion and conclusion

Our proposed quantum state compression readout method resists the large noise caused by multi-qubit measurement by encoding the information-to-read into one qubit. This method has an obvious noise resistance effect on current noisy SOTA quantum devices, such as Sycamore and Zuchongzhi, and can naturally overcome the problems of mitigating correlated error and large-scale calibration. Compared with direct readout, the accuracy advantage expands rapidly as the system size increases, making compression readout a promising alternative option for high-fidelity, large-scale quantum state readout. Compression readout outputs a full list of all amplitude populations $|\alpha\rangle^2$, $|\alpha\rangle^0$, $|\alpha\rangle^1$, $|\alpha\rangle^{-1}$ and generally requires fewer measurement shots to achieve the same high performance of direct readout (see demonstration in sect. 2 and Figure 3), which is preferable for applications like quantum state tomography [49], solving linear and nonlinear equations [50, 51], and fitting distributions in quantum machine learning [52].

The readout error during the single-qubit measurement can be easily mitigated using the TMEM method. Additionally, the error of two-qubit gates in the compression circuit can be further mitigated by schemes like zero-noise extrapolation [53-55] and probabilistic error cancellation [56-59]. Therefore, the compression readout method can be improved even further, making it a powerful tool for boosting the performance of SOTA quantum processors and enhancing the quantum advantage.

He-Liang Huang was supported by the Youth Talent Lifting Project (Grant No. 2020-JCJQ-QT-030), National Natural Science Foundation of China (Grants Nos. 11905294, and 12274464), China Postdoctoral Science Foundation, and Open Research Fund from State Key Laboratory of High Performance Computing of China (Grant No. 201901-01). We gratefully acknowledge Hefei Advanced Computing Center for hardware support with the numerical experiments.

Supporting Information

The supporting information is available online at http://phys.scichina.com and https://link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.
