A Fault-Tolerant Million Qubit-Scale Distributed Quantum Computer

Junpyo Kim
Seoul National University
Seoul, Republic of Korea
junpyo.kim@snu.ac.kr

Dongmoon Min
Seoul National University
Seoul, Republic of Korea
dongmoon.min@snu.ac.kr

Jungmin Cho
Seoul National University
Seoul, Republic of Korea
jungmin.cho@snu.ac.kr

Hyeonseong Jeong
Seoul National University
Seoul, Republic of Korea
hyeonseong.jeong@snu.ac.kr

Ilkwon Byun
Seoul National University
Seoul, Republic of Korea
ik.byun@snu.ac.kr

Junhyuk Choi
Seoul National University
Seoul, Republic of Korea
junhyuk.choi@snu.ac.kr

Juwon Hong
Seoul National University
Seoul, Republic of Korea
jw.hong@snu.ac.kr

Jangwoo Kim[*]
Seoul National University
Seoul, Republic of Korea
jangwoo@snu.ac.kr

Abstract

A million qubit-scale quantum computer is essential to realize the quantum supremacy. Modern large-scale quantum computers integrate multiple quantum computers located in dilution refrigerators (DR) to overcome each DR's unscaling cooling budget. However, a large-scale multi-DR quantum computer introduces its unique challenges (i.e., slow and erroneous inter-DR entanglement, increased qubit scale), and they make the baseline error handling mechanism ineffective by increasing the number of gate operations and the inter-DR communication latency to decode and correct errors. Without resolving these challenges, it is impossible to realize a fault-tolerant large-scale multi-DR quantum computer.

In this paper, we propose a million qubit-scale distributed quantum computer which uses a novel error handling mechanism enabling fault-tolerant multi-DR quantum computing. First, we apply a low-overhead multi-DR error syndrome measurement (ESM) sequence to reduce both the number of gate operations and the error rate. Second, we apply a scalable multi-DR error decoding unit (EDU) architecture to maximize both the decoding speed and accuracy. Our multi-DR error handling SW-HW co-design improves the ESM latency, ESM errors, EDU latency, and EDU accuracy by 3.7 times, 2.4 times, 685 times, and $6.1 \cdot 10^{10}$ times, respectively.

With our scheme applied to assumed voltage-scaled CMOS and mature ERSFQ technologies, we successfully build a fault-tolerant million qubit-scale quantum computer.

CCS Concepts: • **Computer systems organization → Quantum computing**; • **Hardware → Quantum error correction and fault tolerance**.

Keywords: Fault-tolerant quantum computing, Distributed quantum computing, Quantum error correction, Cryogenic computing, Single flux quantum (SFQ)

ACM Reference Format:
Junpyo Kim, Dongmoon Min, Jungmin Cho, Hyeonseong Jeong, Ilkwon Byun, Junhyuk Choi, Juwon Hong, and Jangwoo Kim. 2024. A Fault-Tolerant Million Qubit-Scale Distributed Quantum Computer. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24), April 27-May 1, 2024, La Jolla, CA, USA*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3620665.3640388

1 Introduction

Quantum computing is a promising computing paradigm to innovate various areas, and it requires thousands of logical qubits to run practical quantum applications (e.g., identify electron structures of FeMo-co [59]). However, due to the error-prone nature of qubits and gates, modern large-scale quantum computers apply error correction methods which can construct a single fault-tolerant logical qubit using thousands of noisy physical qubits. Therefore, it is essential to develop a large-scale fault-tolerant quantum computer (FTQC) consisting of more than one million physical qubits with a correspondingly scalable error correction method applied.

[*] Corresponding author.

ASPLOS '24, April 27-May 1, 2024, La Jolla, CA, USA

ACM ISBN 979-8-4007-0385-0/24/04
https://doi.org/10.1145/3620665.3640388

From this perspective, the first key requirement is to build a million qubit-scale quantum computer. Recent studies proposed scalable quantum control methods which exploit various temperatures, device technologies, and microarchitecture optimizations [7, 73]. However, even with the advanced device technologies (i.e., voltage-scaled CMOS [14] or ERSFQ [51]) used, their scaling is still limited to below 100K qubits due to the tight cooling budget of a dilution refrigerator (DR).

To build a scalable quantum computer, academia and industry now aim to build a distributed quantum computer by integrating multiple DRs (multi-DR QC). A multi-DR QC uses a reliable long-range entanglement technology to integrate multiple DRs so that it can run a single large quantum application using the qubits distributed over multiple distant DRs [65]. A multi-DR QC can continuously scale its cooling capacity by integrating more DRs with long-range interactions. The leading QC industries (e.g., Google, IBM) also agree on this scaling direction, and thus include multi-DR QCs in their roadmaps [2, 6].

However, a large-scale multi-DR QC incurs its own critical challenges in maintaining the efficiency of the baseline error handling mechanism. Therefore, a multi-DR QC suffers from more severe logical errors than single-DR systems. First, slow and error-prone inter-DR two-qubit operations (i.e., EPR pair generation) significantly increase the error rate of the error syndrome measurement (ESM). In addition, as there exist a limited number of interconnections between DRs, ESM conducts many SWAP operations on them which increases the physical error rate. Second, slow inter-DR communications and the large qubit scale significantly reduce the decoding speed and accuracy of the error decoding unit (EDU). Without resolving the error handling challenges, even the state-of-the-art control architecture fails (i.e., logical error of 1.0) to support a million qubit-scale multi-DR QC.

In this paper, we realize the distributed quantum computers that can successfully run practical million qubit-scale quantum applications. With our novel error handling mechanisms derived from hardware-software co-design, we resolve all the challenges regarding the (1) ESM errors, (2) error decoding latency, and (3) decoding accuracy of multi-DR QCs.

First, we design a multi-DR aware ESM sequence to minimize the latency and physical error of the ESM. Specifically, we propose a quantum-classical hybrid ESM that minimizes the number of gates for the ESM including EPR pair generations. Next, as the limited number of inter-DR interconnects introduces many SWAP operations, we overlap the SWAPs and ESM operation in a pipelined manner. Finally, we replace the expensive SWAP (i.e., three CZ + six H gates [76]) with our EPR-pair specialized SWAP, which consists of only one iSWAP and one $S^\dagger$ gates. With these three solutions, we significantly reduce the inter-DR ESM latency and physical error per ESM by 3.7 times and 2.4 times, respectively.

Second, we propose a scalable EDU architecture that decodes the errors from distributed million qubits faster than the ESM execution. To achieve such a high decoding speed, we first propose a parallel nine-patch-granular decoding following our observation that an error chain can reach up to adjacent eight logical qubit patches at best. In addition, to reduce and hide the inter-DR communication latency, we propose a 4K & 300K hybrid EDU. It offloads the inter-DR patches' decoding to 300K EDUs, and decodes their errors in parallel with intra-DR patches' decoding at 4K. With the two solutions, we reduce the decoding latency by 509.0 times and 684.9 times, assuming 4K CMOS and ERSFQ, respectively.

Lastly, to improve the decoding accuracy of the fast but error-prone spike-based EDU, we propose an ensemble decoding. In our analysis, the spike-based EDU is error-prone as it usually falls to the local optimum due to its heuristic and greedy approach (e.g., consider only one token-allocating order). To mitigate this problem, the ensemble decoding runs eight EDUs with different token orders and qubit-type priorities and then chooses the best. With this idea, we improve the accuracy of the spike-based EDU by $6.06 \cdot 10^{10}$ times.

Our evaluation shows that the distributed quantum computer equipped with our error handling mechanisms successfully ensures a low logical error to run the practical quantum chemistry application (i.e., FeMo-co) over two million qubits. Note that a multi-DR QC fails to achieve this goal even without just one of our ESM, decoding latency, and decoding accuracy optimizations. Therefore, we emphasize the importance of resolving all these challenges to realize a million qubit-scale distributed quantum computer.

In summary, our work makes the following contributions:

- **Multi-DR aware ESM design**: To the best of our knowledge, this is the first work to design an ESM that addresses the unique challenges of the multi-DR QCs.
- **Scalable distributed EDU**: It is also the first to propose a fast and accurate EDU architecture that successfully decodes errors from a million distributed qubits using parallel, 4K & 300K hybrid, and ensemble EDU.
- **Fault-tolerant million qubit-scale QC**: A million qubit-scale quantum computer equipped with our error handling mechanisms guarantees a fault-tolerant execution of practical quantum algorithms. We believe that our work marks a milestone for realizing a next-generation large-scale distributed quantum computer.

2 Background and Motivation

2.1 Large-scale QC using multiple DRs

To solve classically-intractable and practical problems, we need a fault-tolerant quantum computer (FTQC) that consists of more than a million qubits. For this purpose, architects should prepare a scalable quantum control system supporting more than a million qubits in advance. It is timely to consider such a control system because leading industries aim to build a million-qubit quantum computer within a few years (e.g., 2029 of Google [2, 75] and 2033 of Microsoft [72]).

Following this motivation, previous studies [7, 73] have proposed scalable quantum control systems assuming all superconducting qubits operate in a single DR. The quantum control system mainly consists of two key components: quantum control processor (QCP), a classical digital processor to support FTQC operations, and quantum-classical interface (QCI), an interfacing electronics to control and measure qubits. The previous works have explored the scalable QCPs and QCIs while considering (1) cooling power budget and (2) logical error constraints. However, even with the advanced device technologies (e.g., 7nm voltage-optimized 4K CMOS, ERSFQ) and state-of-the-art DR (i.e., Bluefors KIDE [5], 4.5W of 4K power budget), they cannot support even 100,000 qubits due to the tight cooling power budget.

In this context, the quantum computer using multiple DRs (multi-DR QC) has emerged as a promising solution to scale the control system further. In the multi-DR QC, qubits are distributed over multiple DRs but abstracted as a single large quantum computer with inter-DR two-qubit operations. As we can put more DRs in this approach, the cooling power budget does not limit the scalability anymore. With its great potential, IBM and Google clarify that they will develop their next-generation quantum computers using multiple DRs [2, 6].

However, we still need an extremely low logical error rate which becomes the major concern in the multi-DR scenario. Therefore, we focus on minimizing the logical error of the large-scale fault-tolerant distributed quantum computers.

2.2 Fault-tolerant quantum computing overview

We target fault-tolerant quantum computing (FTQC) based on the surface code with lattice surgery, one of the prominent quantum error correction (QEC) protocols [26, 37, 62, 63].

2.2.1 Quantum error correction with surface code. For FTQC, we build and maintain a logical qubit through QEC [27, 38, 108]. QEC is an iterative process of error syndrome measurement (ESM), error decoding, and error correction. Figure 1(a) shows the overall process of QEC with an example of a surface-code logical qubit (or patch) with code distance three (denoted by $d = 3$).

Error syndrome measurement. First, we apply a quantum circuit called ESM d times (or rounds) to the two types of physical qubits; data qubits (solid circle; D1~4) containing state information and ancilla qubits (open circle; X or Z) used for extracting error information. Figure 1(b) shows two ESMs that entangle ancilla qubits with their adjacent data qubits and then measure them. As a result, we can discretize continuous data-qubit errors (e.g., gate error, decoherence error) into X or Z errors. More importantly, the obtained ancilla qubit measurements (or error syndromes) provide hints for these errors. Specifically, every Z (or X) error flips the adjacent X (or Z)-ancilla measurements, and thus a chain of errors flips the syndromes in its two endpoints.

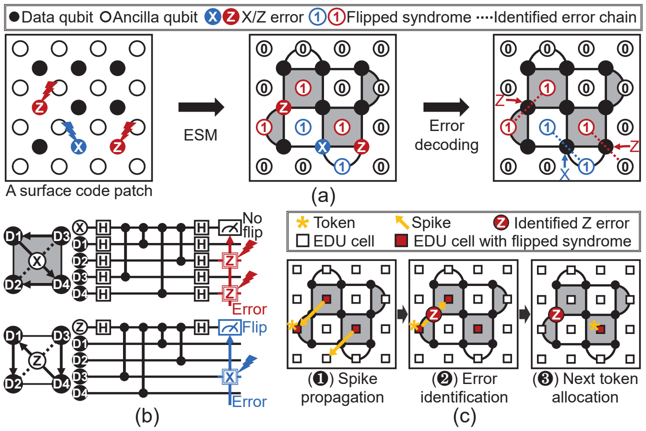


Figure 1. (a) Overall process of quantum error correction, (b) error syndrome measurement (ESM), and (c) error decoding

Error decoding. Next, by using the d-round error syndromes, we decode the types and locations of data-qubit errors. For this purpose, an error decoding algorithm aims to pair the nearest flipped syndromes and identifies the most likely error chains connecting them. Figure 1(c) shows an example based on the state-of-the-art spike-based error decoding unit (EDU) [7, 103, 104]. The EDU consists of an array of EDU cells where each cell tracks the error syndromes for its dedicated ancilla qubit. First, the EDU gives a token to the highest priority EDU cell among the cells with the flipped syndrome and lets the others send spikes toward the token (❶). Next, when the token-allocated cell receives a spike, it reflects and propagates the spike to the original sender. While forwarding the spike back, the EDU naturally identifies the error chain connecting the token and the sender (❷). Lastly, if the spike arrives at the origin, the EDU removes the matched syndromes and allocates a token to the next-priority cell (❸). The EDU iterates the above process until no flipped error syndrome remains.

Error correction. Finally, we correct the identified data-qubit errors virtually. Specifically, we do not physically apply error-correcting X (or Z) gates but track the errors using Pauli frame [85]. As a result, we can proceed to the next ESM before the decoding of the previous ESM finishes.

2.2.2 Logical quantum operation with lattice surgery. With lattice surgery, we can run an arbitrary quantum circuit with $\text{PPR}_{\frac{\pi}{8}}$ operations while suppressing the errors [62]. A $\text{PPR}_{\frac{\pi}{8}}$, a universal gate, mainly consists of three ESMs: Initializing ESM, Merging ESM, and Splitting ESM (Figure 2).

Initializing ESM (ESM_{INIT}). We should prepare two auxiliary logical qubits (i.e., A1 in the magic state and A2 in the zero state) for every $\text{PPR}_{\frac{\pi}{8}}$. For this purpose, we run ESM for all the logical qubits to initialize A1 and A2.

Merging ESM ($\text{ESM}_{\text{MERGE}}$). Following the Pauli product operator (i.e., P_1P_2 in Figure 2(a)), we merge target logical

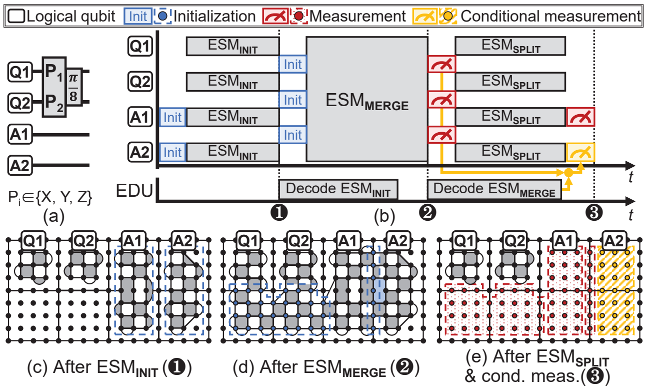


Figure 2. (a) Quantum circuit, (b) timeline, and (c-e) qubit lattice of PPR$_{\frac{\pi}{8}}$

qubits by running ESMs for all surface-code patches connecting the logical qubits (or intermediate patches). Note that we should synchronize the ESM operations of all merged qubits with the worst-case ESM latency to apply next operations.

Splitting ESM (ESM$_{\text{SPLIT}}$). A PPR$_{\frac{\pi}{8}}$ finishes with the conditional measurement on A2, which requires the error decoding results of ESM$_{\text{MERGE}}$. Therefore, we need at least one additional ESM after measuring the intermediate patches.

2.2.3 Factors affecting logical error rate.

Even though FTQC can dramatically suppress the errors on logical qubits and operations (i.e., logical error), we still can suffer from the logical errors due to four factors. These factors are related to the ESM sequence or error decoding.

ESM latency and gate-induced error. If too many physical errors occur during an ESM, QEC can fail due to error chains longer than $\frac{d}{2}$. From this perspective, a longer ESM latency incurs more logical errors by increasing the decoherence error. Similarly, the logical error rate increases if an ESM includes more quantum gates or error-prone operations.

Error decoding latency. When an error decoding is slower than the ESM execution, it incurs a critical system failure due to the backlog problem [36, 100]. For example, if the decoding for ESM$_{\text{MERGE}}$ finishes later than the ESM$_{\text{SPLIT}}$ in Figure 2, we should put more ESMs before the conditional measurement. As the additional ESMs delay the decoding of the next ESM$_{\text{MERGE}}$ again, the number of additional ESMs exponentially increases for every PPR$_{\frac{\pi}{8}}$. As a result, the system fails to run a quantum algorithm as the logical error rate reaches 1.0. To prevent this failure, the worst-case (or maximum) decoding latency should be lower than the ESM latency [18, 19, 36, 105].

Error decoding accuracy. An error decoding algorithm's accuracy also significantly affects the logical error rate. In general, there are trade-offs between accuracy and speed; the more accurate the decoding, the slower the decoding. Note that accurate decoding can reduce the logical error rate only when the decoding is faster than the ESM execution.

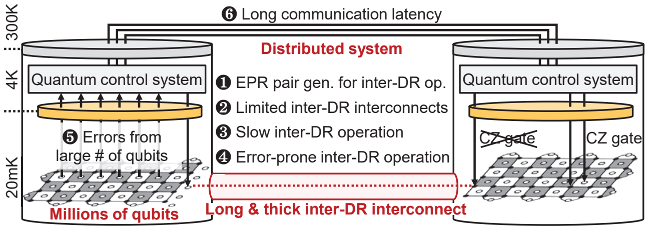

300K
4K
20mK
❻ Long communication latency
Distributed system
Quantum control system
Quantum control system
❶ EPR pair gen. for inter-DR op.
❷ Limited inter-DR interconnects
❸ Slow inter-DR operation
❹ Error-prone inter-DR operation
❺ Errors from large # of qubits
~~CZ gate~~
CZ gate
Millions of qubits
Long & thick inter-DR interconnect

Figure 3. Characteristics of the million qubit-scale multi-DR system (❶-❻)

2.3 Challenges in designing large-scale multi-DR QC

A quantum computer with multiple DRs (multi-DR QC) is promising for system scaling. However, in a large-scale multi-DR QC, ensuring a low logical error is extremely challenging due to characteristics of the large-scale multi-DR QC. We introduce its six unique characteristics and how they aggravate the four factors affecting the logical error rate. Figure 3 shows the six characteristics of the large-scale multi-DR QC.

2.3.1 Challenge of high physical error during ESM.

To entangle two qubits in the separated DRs, we need a long and thick inter-DR interconnect to couple them. The long and thick interconnect causes four unique characteristics that increase the physical-qubit errors during an ESM.

EPR pair generation for inter-DR operation (❶). We should use a new two-qubit operation called EPR pair (i.e., $\frac{(|00\rangle+|11\rangle)}{\sqrt{2}}$) generation as we cannot apply a CZ gate to distant qubits [67]. As a result, for the ESM of inter-DR patches, we should implement a CZ gate between distant qubits with the multiple gates including the EPR pair generation (e.g., distributed CZ [61]). However, it requires many more gates and increases the ESM latency and gate-induced error rate.

Limited inter-DR interconnects (❷). An inter-DR interconnect is inevitably thick to maintain 20mK for the long-range connection. Therefore, we have limited connections between DRs, much fewer than the number of qubits to be entangled [6]. To execute the inter-DR ESM in this scenario, we should insert SWAP gates to move an EPR pair to the target qubits. However, the additional SWAP gates incur longer ESM latency and higher gate-induced error rate.

Slow & error-prone inter-DR operation (❸, ❹). The inter-DR entanglement through long-range coupling (i.e., EPR pair gen.) is much slower and more error-prone (e.g., 391ns with 4% error [65]) than CZ gate (e.g., 33ns with 0.19% error [67]). As a result, the inter-DR ESM suffers from higher ESM latency and gate-induced error than the single-DR case.

2.3.2 Challenge of low decoding speed and accuracy.

We should successfully perform FTQC on million qubits distributed over the multiple DRs. However, it poses challenges to achieving fast and accurate error decoding.

Errors from large number of qubits (❺). Due to the large number of qubits, a massive number of error syndromes

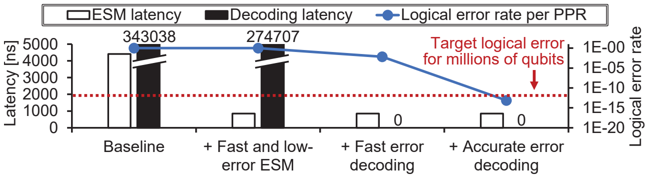


Figure 4. Necessity of all the low-error ESM sequence, fast error decoding, and accurate error decoding

are generated every ESM round. Therefore, we need an accurate EDU capable of decoding the large amount of errors within the ESM latency. However, existing error decoding algorithms proportionally slow down with the increasing qubit scale. Even worse, any design choice for accelerating an EDU inevitably reduces the decoding accuracy due to general trade-offs.

Long communication latency (❻). From the scalability and feasibility perspectives, we consider a distributed quantum control system running at 4K rather than the centralized 300K system [6, 7, 46, 69, 73]. In this scenario, EDUs in the adjacent DRs should communicate for every clock cycle to find boundary-crossing error chains. However, this communication takes a long time (e.g., 55ns according to our setup in Section 3.2.1) and significantly increases the decoding latency.

2.4 Research goal

The above six characteristics eventually incur system failure by increasing the logical error rate. Even the state-of-the-art control system (Baseline in Figure 4) fails to meet our target logical error rate for running a million qubit-scale application (Red line in Figure 4). Therefore, we aim to realize a fault-tolerant quantum computer by resolving all these challenges regarding higher physical error, higher decoding latency, and lower decoding accuracy.

We clarify our research goal with a thorough analysis in Figure 4 (detailed setup in Section 3). To achieve the target logical error rate, we need all the fast and low-error ESM sequence, fast error decoding, and accurate error decoding. In Figure 4, we assume ordinary single-DR ESM sequence (+ fast and low-error ESM), zero EDU latency (+ fast error decoding), and optimized spike-based EDU that all of our EDU solutions applied (+ accurate error decoding). Note that without even one of these three, we cannot ensure a low logical error enough to run a practical quantum program.

In this paper, we achieve our goal as follows. First, we design a multi-DR aware ESM sequence to reduce the ESM latency and gate overhead (Section 4). Second, we propose a scalable EDU architecture that accurately decodes a million qubits' errors within the ESM latency (Section 5). Finally, we show our solutions successfully achieve a low logical error, sufficient to realize the true quantum supremacy (Section 6).

Table 1. Evaluation setup

Target scalability specification							
		Target qubit scale		Target logical error△		# of PPR op.	Code distance
2030 (3D Jellium)		215,040		7.14E-10		1.40E+7	31
2033 (FeMo-co)		2,138,112		7.50E-13		1.34E+10	
Multi-DR system specification							
Cooling capacity			# of inter-DR connection		Inter-DR communication latency		# of qubits per DR
4K	100mK	20mK					
4.5W	300μW	90μW	11		55ns		92,160
Control hardware (QCP & QCI) specification							
QCI power*		Baseline EDU power*				EDU Frequency	
4K CMOS		4K CMOS		4K ERSFQ		300K & 4K CMOS	4K ERSFQ
4.15W		42.6mW		88.0mW		1.5GHz	21GHz
Physical-qubit error specification							
		Gate operation				Decoherence time	
Operation type		1Q	2Q	EPR gen.	RO	T1	T2
Error	2030	3.02E-5	2.09E-4	5.47E-3	1.49E-3	3,218μs	2,104μs
	2033	1.40E-5	7.65E-5	2.07E-3	8.44E-4	8,100μs	4,464μs
Latency (ns)		25	50	391	517	-	

△Logical error per PPR operation *Per-DR runtime power consumption

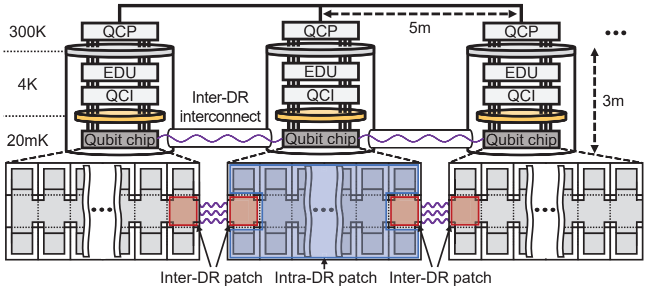


Figure 5. Overview of our multi-DR QC setup

3 Evaluation Methodology

We describe our target qubit scale and logical error, multi-DR QC setup, and simulation infrastructure. Table 1 summarizes our setup. Note that our solutions in Sections 4 and 5 effectively improve the logical error rate regardless of the setup.

3.1 Target qubit scale and logical error

To realize a true sense of innovation, we target a 99% success rate [31, 59, 62] for practical applications requiring more than 100,000 and 1,000,000 qubits. For the target applications, we consider 3D Jellium [52] and FeMo-co [59], which require 215,040 qubits and 2,138,112 qubits at $d = 31$. We set the target logical error of the 215,040 and 2,138,112 qubit systems to $7.14 \cdot 10^{-10}$ and $7.50 \cdot 10^{-13}$, respectively, by following their required number of logical operations (i.e., 1% failure rate / number of PPR$_{\frac{\pi}{8}}$s [62]).

3.2 Multi-DR quantum computer setup

Figure 5 illustrates our target multi-DR quantum computer with the multi-DR system, control hardware, and qubit plane setups.

3.2.1 Multi-DR system setup.

Dilution refrigerator. We follow the specifications (i.e., size, cooling capacity) of Bluefors KIDE [5], which will be used in the next-generation IBMQ System Two [40].

Inter-DR communication latency. We set the inter-DR communication latency to 55ns based on the height of KIDE (3m), inter-DR distance in previous experiments (5m) [65], and signal-transfer speed of coaxial cable (5ns/m) [87].

Number of inter-DR interconnects. We set the number of inter-DR connections (i.e., long-range qubit coupling between adjacent DRs) to 11 by assuming that three qubits at the $d = 31$ patch boundary share one interconnect. We provide a sensitivity analysis for the inter-DR connections in Figure 20.

3.2.2 Control hardware (QCI & QCP) setup.

We operate the QCI and EDU at 4K, and other QCP units at 300K, following the guidelines of the previous works [7, 73].

4K QCI. We assume the 7nm voltage-optimized 4K CMOS QCI and adopt the frequency and power consumption values reported in the previous work [73].

4K EDU. We evaluate both 4K CMOS and ERSFQ-based EDUs to show our solutions in Section 5 are effective regardless of the device technologies. To derive the frequency and power of a 4K CMOS-based EDU, we first run XQsim and then project the obtained result to 7nm technology following ITRS roadmap [96]. For an ERSFQ-based EDU, we also use XQsim with MITLL SFQ5ee library [88].

3.2.3 Qubit plane setup.

Surface-code patch setup. We set the surface-code distance (d) to 31 by following the reference of our target applications [59]. For the patch topology (i.e., block type), we adopt the compact block with three patch rows, where we use top and bottom rows to map logical qubits and use the middle row as intermediate patches to merge logical-qubit patches [62].

Number of physical qubits per DR. As the 4K power consumption limits the single-DR scalability in our setup, we calculate the number of physical qubits (i.e., 92,160) and patches (i.e., 45) inside each DR by dividing the 4K cooling capacity with the 4K power per qubit. Note that the 4K power is also the scalability bottleneck of the previous works [3, 7, 46, 73, 79].

Physical-qubit error. To model the realistic physical-qubit errors (i.e., gate errors, decoherence time) of the target qubit-scale quantum computers, we first estimate the years to realize the target-qubit scales and then obtain the expected errors with the projection. Figure 6 shows the growing number of qubits and decreasing errors with the data of representative quantum computers in recent years. For the projection, we use the quantum computers of academia and industry with the largest number of qubits, highest decoherence time, and lowest errors, for each year. [1, 4, 8, 11–13, 15–17, 20, 22, 32, 33, 39, 41–43, 45, 49, 53, 54, 56, 60, 64–68, 78, 83, 84, 86, 90, 91, 94, 97, 102, 106, 112, 114]. As the target qubit-scale quantum computers are expected to be realized in 2030 and 2033 (Figure 6(a)), we derive the gate errors and decoherence times in 2030 and 2033 (Figure 6(b), (c)). Note that the physical error per ESM derived from these values ($1.4 \cdot 10^{-3}$ and $5.5 \cdot 10^{-4}$ for the intra-DR patches in 2030 and 2033) are similar to the widely-used values in FTQC research [18, 19, 62, 82, 104, 105].

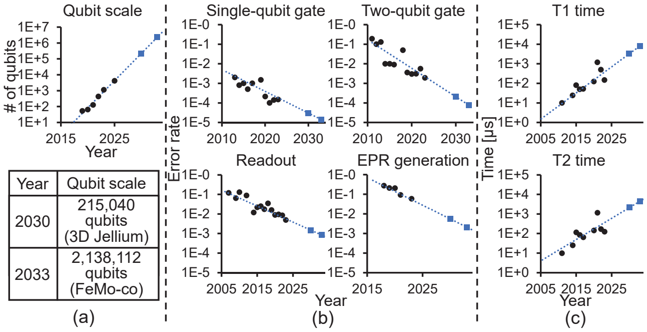


Figure 6. Trend of (a) the number of qubits, (b) gate error rate, and (c) decoherence time of major quantum computers over time

3.2.4 Baseline ESM & EDU setup.

Baseline ESM. Figure 8(a) shows our baseline ESM sequence for the multi-DR system. First, we use the distributed CZ operation [109] for the inter-DR CZ gates in the original ESM sequence. We need one EPR pair and several single-qubit and CZ gates for each distributed CZ operation. Second, considering the limited inter-DR interconnections, we insert SWAP gates to move a generated EPR pair from the interconnection point (i.e., purple circle) to other qubits. For the gate latency, we follow the specification of previous works on the intra-DR gates [73] and EPR pair generation [65].

Baseline EDU. As our EDU baseline, we adopt the state-of-the-art spike-based EDU, the fastest and the only hardware decoder supporting lattice surgery [7, 104]. We start with the recently proposed fast EDU, which can directly allocate a token to the highest-priority flipped syndrome [7]. The 4K CMOS and ERSFQ-based baselines run at 1.5GHz and 21GHz to decode the errors of intra-DR patches. On the other hand, when decoding the inter-DR patches, the EDU switches its clock frequency to 18.2MHz (i.e., the reciprocal of the inter-DR communication latency; 55ns) because an EDU should wait for the spikes from adjacent DRs for every clock cycle.

3.3 Simulation infrastructure

Figure 7 shows our simulation infrastructure to derive ESM & decoding latencies and physical & logical error rates.

3.3.1 ESM latency and physical error rate per ESM.

We obtain the ESM latency and physical error rate per ESM from QIsim, the validated QCI simulation framework [73]. For the target ESM sequence, QIsim runs timing simulation and derives the ESM latency considering the gate latency, QCI microarchitecture, and true dependency [89]. At the same time, QIsim outputs the physical error rate per ESM based on the gate, measurement, and decoherence error rates.

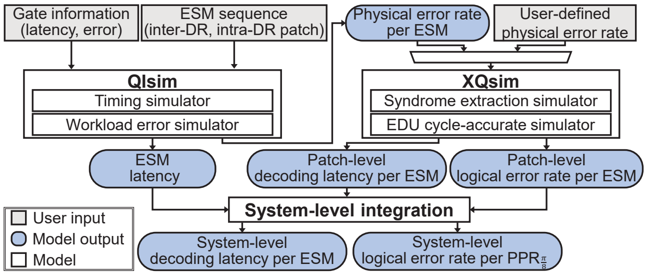

Gate information (latency, error)
ESM sequence (inter-DR, intra-DR patch)
Physical error rate per ESM
User-defined physical error rate
QIsim
Timing simulator
Workload error simulator
XQsim
Syndrome extraction simulator
EDU cycle-accurate simulator
ESM latency
Patch-level decoding latency per ESM
Patch-level logical error rate per ESM
System-level integration
□ User input
○ Model output
□ Model
System-level decoding latency per ESM
System-level logical error rate per PPR$_{\frac{\pi}{8}}$

Figure 7. Simulation infrastructure

We get the results for both inter-DR and intra-DR patches by running QIsim with their corresponding ESM circuits.

3.3.2 Patch-level decoding latency and logical error.

We obtain the patch-level per-ESM decoding latency and logical error rate using XQsim, the validated QCP simulation framework [7]. By taking the physical error rate per ESM from QIsim output, XQsim first generates the d-round error syndromes with Stim [30]. Next, XQsim derives the decoding latency and logical error rate at the patch level by running an EDU timing simulation with the obtained syndromes.

To derive the extremely low logical error at a low physical error range, we adopt the widely-used error projection [24, 25, 27]. We first obtain the logical error at the relatively high physical error range (i.e., 0.008~0.012), and then project the results to the low physical error range. For each data point, we run Monte-Carlo simulation of d-round ESM by 10^5 times. We use the well-known fitting function (i.e., $c_1 * (p/p_{th})^{c_2*(d+1)/2}$) of the previous works [27, 36], where p_{th} is the threshold and p is the physical error rate per ESM.

3.3.3 System-level decoding latency and logical error.

We obtain the system-level maximum decoding latency per ESM and system-level logical error per PPR$_{\frac{\pi}{8}}$ as follows.

First, we define the maximum decoding latency as the worst-case latency occurring only once during the workload execution. Following the definition, we set the probability of observing the maximum latency to the reciprocal of the number of PPR operations. Next, we obtain the system-level latency distribution from the patch-level latency distributions through random sampling, and then derive the maximum decoding latency with the target probability by using the widely-used fitting function (i.e., Gamma distribution function [58, 95]).

Meanwhile, we obtain the system-level logical error rate by summing up the error rates of patches participating in the three ESMs for a PPR$_{\frac{\pi}{8}}$ [62]. Considering the backlog problem, we set the logical error to 1.0 when the system-level decoding latency for ESM$_{\text{MERGE}}$ is longer than the ESM latency.

We provide the validation results of our system-level modeling methodology in Section 7.2.

4 Multi-DR aware ESM Sequence

To achieve the target logical errors, we should minimize the physical error rate per ESM. Unfortunately, the baseline's physical error rate is high due to its long ESM latency and severe gate-induced errors. Therefore, in this section, we reduce the physical error with three solutions. Figure 8 provides an overview of our solutions with their change of qubit mapping, ESM sequence, and its timeline. Figure 9 shows the (a) ESM latency and (b) physical error of inter-DR qubits and (c) other qubits while accumulatively applying our solutions.

4.1 Solution-#1. Quantum-classical hybrid ESM

(1) Limitation: Slow and error-prone distributed CZ. As shown in Figure 8(a), the baseline uses the distributed CZ gate as we should emulate the inter-DR CZ gate with EPR pairs. However, the distributed CZ gate needs too many gates, including slow and error-prone operations (i.e., one EPR pair generation and two measurements). As a result, the inter-DR qubits suffer from the huge ESM latency (4,415ns) and physical error rate (0.0301) (Figure 9(a), (b)). Note that the slow inter-DR ESM increases the overall ESM latency, because all the merged qubits should wait until the end of the entire ESM.

(2) Solution. To avoid using the distributed CZ gates, we propose a quantum-classical hybrid ESM sequence (Figure 8(b)). The key idea of our solution is to offload the gate overhead of the quantum computation domain to the classical computation domain. Specifically, the proposed sequence first generates an EPR pair, runs ordinary ESM sequence, and measures two ancilla qubits at different DRs in the quantum computation domain. As two measurements can be executed in parallel, there is no additional ESM latency overhead. Then, at the classical computation domain, we obtain the error syndrome by XORing the measurement results. We confirm with Qiskit simulation that our quantum-classical hybrid ESM sequence has the same functionality as the baseline [81].

(3) Result. Our hybrid solution needs only minimum number of gates for the main ESM operations (i.e., white boxes in Figure 8(b) timeline). In addition, it consumes only one EPR pair for each inter-DR error syndrome, while the baseline requires two. As a result, our scheme reduces the ESM latency and physical error per ESM of inter-DR qubits and other qubits by 1.78 times, 2.08 times, and 1.20 times, respectively (Figure 9).

4.2 Solution-#2. Pipelined EPR pair preparation

(1) Limitation: Slow EPR pair preparation. Even though the hybrid ESM resolves the challenge of the distributed CZ gates, it still suffers from the slow EPR pair preparation (i.e., blue boxes in Figure 8(b)). Due to the limited number of inter-DR connections, three physical qubits should share one connection using SWAP gates. The SWAP gates incur long EPR pair preparation latency and huge physical error.

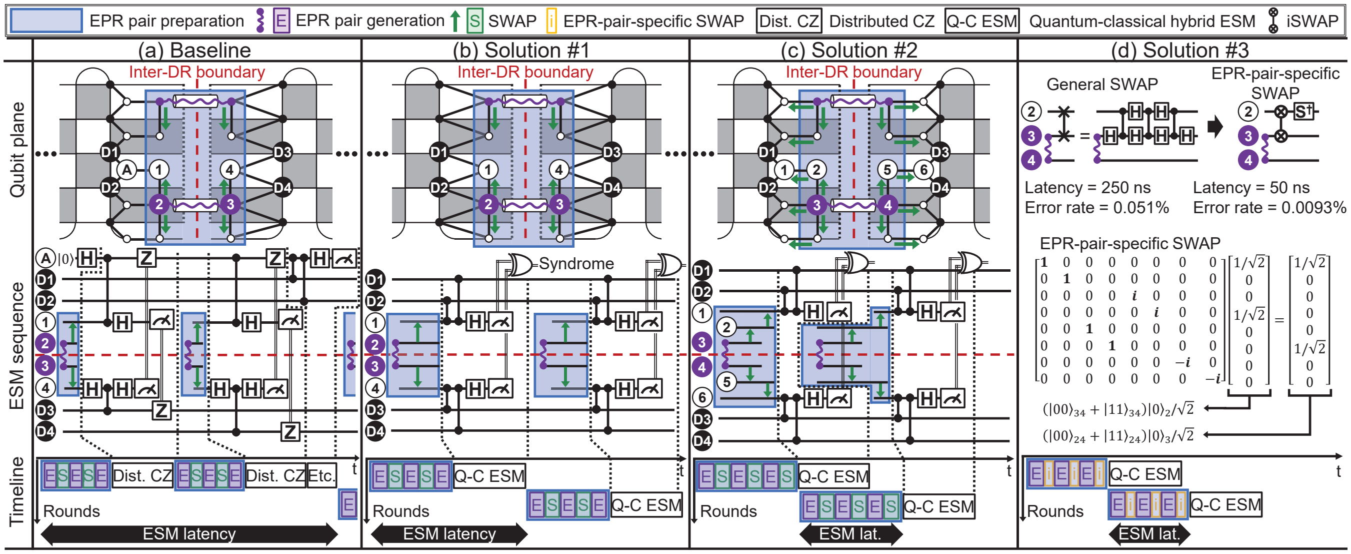


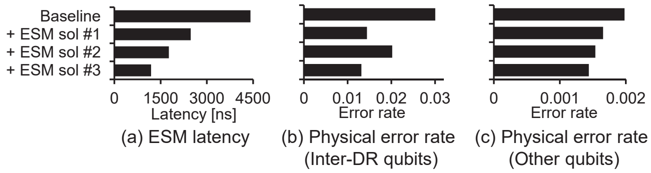
Figure 8. Qubit mapping, ESM sequence, and execution timeline of baseline and our multi-DR aware solutions; (a) Baseline that uses a distributed CZ gate to emulate an inter-CZ gate, (b) Solution #1 that introduces a quantum-classical hybrid ESM sequence to remove the distributed CZ gates, (c) Solution #2 that hides the EPR preparation latency by pipelining the EPR pair preparation and main ESM operation, and (d) Solution #3 that converts a general SWAP into a newly-proposed EPR-pair-specific SWAP operation for the physical error reduction

(2) Solution. To hide the preparation latency, we propose a pipelined EPR pair preparation (Figure 8(c)). The key idea is to run the EPR pair preparation and main ESM operation in parallel by separating them with qubit mapping modification. Specifically, we first shift the qubit mapping of each DR by one qubit to get additional qubits, which will be used as EPR pair buffers. As a result, the executions of the two operations become independent. Then, in a pipelined manner, we prepare the EPR pairs using the additional buffer qubits while running the main ESM operations.

(3) Result. Our pipelining solution enables ESMs to overlap the main ESM operations and EPR pair preparations by using few additional buffer qubits. As a result, we reduce the ESM latency by 1.40 times (i.e., +ESM sol #1 vs. +ESM sol #2 in Figure 9(a)) with 0.067% of negligible qubit overhead. Note that we only change the qubit mapping on the same physical qubit topology (i.e., 2D array [1, 23]) for every solution.

4.3 Solution-#3. EPR-pair-specific SWAP operation

(1) Limitation: Expensive SWAP operation. Even with the shorter ESM latency, the pipelining solution adversely increases the physical error of inter-DR qubits (Figure 9(b)). Specifically, the changed qubit mapping requires additional SWAP gates to move the EPR pairs further and thus incurs huge gate-induced errors (Figure 8(c)). In addition, we still cannot completely hide the EPR pair preparation latency due to the slow SWAP gate. Therefore, we need a fast and low-error SWAP gate to minimize the physical error rate.



(a) ESM latency (b) Physical error rate (Inter-DR qubits) (c) Physical error rate (Other qubits)

Figure 9. (a) ESM latency (ESM_{MERGE}) and (b) physical error per ESM of inter-DR qubits and (c) other qubits

(2) Solution. To reduce the huge SWAP overhead, we propose an EPR-pair-specific SWAP operation (Figure 8(d)). The key insight of our solution is to devise a specialized SWAP for EPR pairs. Specifically, we find that "iSWAP + $S^{\dagger}$" gates work as SWAP if one target qubit is a member of an EPR pair and the other is in $|0\rangle$ state (Matrix representation in Figure 8(d)). First, the iSWAP gate is faster and less error-prone than the CZ gate thanks to its simpler operation [97]. In addition, QCIs can realize the $S^{\dagger}$ (i.e., $-\frac{\pi}{2}$ z-axis rotation) without delay and error (i.e., virtual Z gate [70, 73]). Therefore, our solution greatly reduces the latency and error of each SWAP gate by 80% and 74.07%. We conservatively set the latency and error of the iSWAP gate to those of the CZ gate (50ns, 0.021%).

(3) Result. With the low SWAP overhead, our EPR-pair-specific SWAP reduces the ESM latency and physical error of inter-DR and other qubits by 1.49 times, 1.54 times, and 1.07 times, respectively (+ESM sol #3 in Figure 9).

Overall, with our three ESM solutions, we achieve 3.71 times shorter ESM latency and 2.29 times and 1.38 times

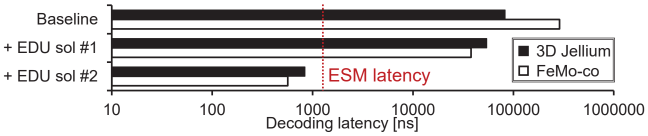


Figure 10. Maximum decoding latency of 4K CMOS-based EDU, decreasing with our solutions for low decoding latency

lower physical error rate for inter-DR and other qubits, respectively. Note that our multi-DR targeted ESM latency (1,189ns) is comparable to the single-DR ESM latency (851ns), which emphasizes the effectiveness of our solutions.

5 Multi-DR Aware & Scalable EDU Design

Even with the proposed low-error ESM sequence, we still cannot run our target workloads (i.e., 3D Jellium, FeMo-co) due to the slow and inaccurate error decoding (Figure 4). Therefore, we propose the multi-DR aware and scalable EDU architecture, which can accurately decode errors from millions of qubits within the ESM latency. For our EDU analyses, we use our multi-DR aware ESM sequence proposed in Section 4.

5.1 Solutions for low error decoding latency

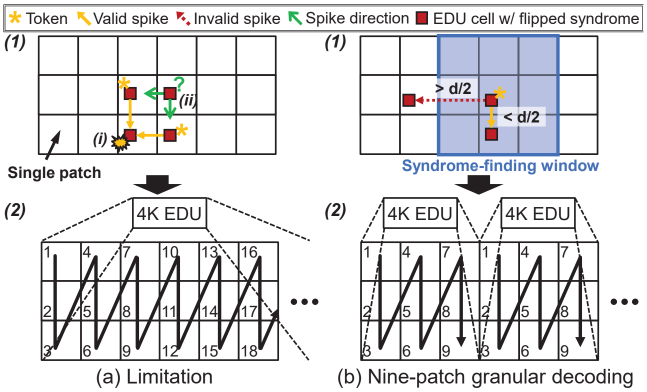
Figure 10 shows the maximum decoding latency of 4K CMOS EDU where baseline's latency is 242.6 times longer than the ESM latency. As a result, a quantum computer fails to run any practical algorithm due to the severe backlog problem (Section 2.2.3). Therefore, we propose two EDU solutions to reduce the decoding latency. Note that we evaluate the ERSFQ EDU as well as the 4K CMOS EDU in Section 6.

5.1.1 Solution-#1. Parallel nine-patch-granular EDU.

(1) Limitation: Sequential decoding. Due to the large number of qubits to be decoded at ESM_{merge}, we should run EDUs in parallel to finish the decoding within the ESM latency. However, no previous work decodes multiple errors in parallel due to the collision problem.

Figure 11(a) illustrates the collision problem when decoding multiple errors in parallel. First, when we naively allocate an EDU for every patch and run all the EDUs in parallel, one flipped syndrome can match with the two same-distance syndromes in different patches (Figure 11(a)-(1)-(i)). In addition, it is challenging to set the spike-propagating direction of a flipped syndrome if multiple tokens exist in the different directions (Figure 11(a)-(1)-(ii)). To avoid the collisions, the previous works [7, 104] sequentially decode multiple patches one by one, and thus suffer from the huge decoding latency (Figure 11(a)-(2)).

(2) Solution. To enable the parallel decoding without collision, we propose a nine-patch-granular decoding (Figure 11(b)). The key observation of our solution is that the length of a correctable error chain is always shorter than $\frac{d}{2}$ according to the definition of the surface code. That is, for a given flipped syndrome (or token), we only have to search for the syndromes in the adjacent patches, up to nine patches including the token-allocated patch (syndrome-finding window in Figure 11(b)-(1)).



(a) Limitation (b) Nine-patch granular decoding

Figure 11. (a) Sequential decoding due to the collision problem and (b) our parallel nine-patch-granular decoding

Following our observation, we allocate independent EDUs in nine-patch granularity and decode errors in parallel (Figure 11(b)-(2)). In addition, to prevent collisions, every EDU synchronously decodes the nine patches in the same order. In Figure 11(b)-(2), each patch's number indicates the decoding order. As a result, we decode every nine-patch group in parallel without collision and reduce the decoding time complexity from $O(n)$ to $O(1)$, where n is the number of logical qubits.

(3) Result. With the parallel decoding, our solution significantly reduces the maximum decoding latency of 4K CMOS EDU by 34.4% (54,374ns) and 86.9% (37,852ns) for 3D Jellium and FeMo-co, respectively (Baseline vs. +EDU sol #1 in Figure 10). As FeMo-co requires 9.94 times more logical qubits, its latency reduction is much higher than the 3D Jellium case.

5.1.2 Solution-#2. 4K & 300K hybrid EDU.

(1) Limitation: Inter-DR communication latency. Even with the proposed parallel decoding, we still suffer from the huge error decoding latency mainly due to the long inter-DR communication latency. Specifically, to identify an error chain crossing an inter-DR boundary, the EDU should wait for the spike from adjacent patches inside a different DR for every clock cycle. Therefore, to decode patches bordering the inter-DR boundary (e.g., #7, #8, and #9 of DR1 and #1, #2, and #3 of DR2 in Figure 12(a)), the EDU should run at 18.2MHz (i.e., the reciprocal of 55ns inter-DR communication latency). We emphasize that the error decoding latency for these patches occupies a significant portion (99.39%) of the total maximum decoding latency.

(2) Solution. To resolve the inter-DR communication challenge, we propose a 4K & 300K hybrid EDU architecture (Figure 12(b)) with three key ideas. First, we change the qubit

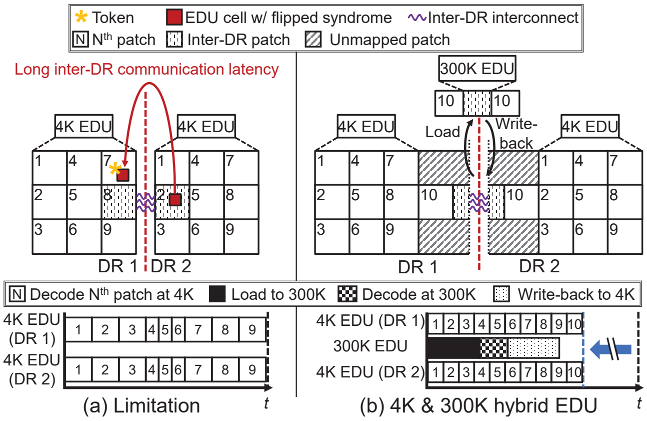


Figure 12. (a) Slow error decoding due to the long inter-DR communication latency and (b) our fast 4K & 300K hybrid EDU

mapping to minimize the number of boundary-crossing error chains. As a result, a single inter-DR patch becomes the only patch requiring the spike information from the two DRs. Second, we offload the inter-DR patches' decoding to the 300K EDUs to decode their errors with one-time inter-DR communication (i.e., load to 300K and write-back to 4K). Finally, we hide the slow inter-DR communication and 300K decoding by running the 300K and 4K EDUs in parallel. Note that now the 4K EDUs can operate with their maximum clock frequency (i.e., 21GHz for ERSFQ and 1.5GHz for 4K CMOS).

Figure 12(b) also shows the detailed execution timeline. While the 4K EDUs decode #1~9 patches, the 300K EDU gets the error syndromes of the inter-DR and #10 patches, decodes them, and moves the results back to 4K. Next, if the 300K decoding results are available after decoding #1~9 patches, the 4K EDUs finally decode the #10 patches.

(3) Result. Our 4K & 300K hybrid solution drastically reduces the maximum decoding latency of the 4K CMOS-based EDU by 98.5% (839.6ns) and 98.5% (566.7ns) for 3D Jellium and FeMo-co, respectively. With our novel solutions, we achieve the decoding latency lower than the ESM latency (for both 4K CMOS and ERSFQ technologies in Figure 16). Therefore, we successfully overcome the backlog problem even in the million qubit-scale multi-DR quantum computers.

5.2 Solution for high error decoding accuracy

We improve the decoding accuracy of the EDU as the final step to achieve our target logical errors (Figure 4).

5.2.1 Solution-#3. Ensemble EDU.

(1) Limitation: Huge logical error of spike-based EDUs. With thorough analyses, we identify the two critical reasons that incur the low decoding accuracy in the spike-based EDU. In short, the existing spike-based EDUs consider only a single setup and usually fall to a local optimum. As a result, they incur much more logical errors than other global-optimum exploring algorithms (e.g., MWPM [21]). First, its decoding result is sensitive to a token-allocating order. For example, the EDU can produce correct or wrong results according to the token order, even with the same syndromes; the forward token order (FWD) is correct in Figure 13(a) but the backward order (BWD) is correct in Figure 13(b). Second, no spike-based EDU considers various qubit-type priorities when constructing an error chain. A token cell uses the qubit-type priority to handle cases when multiple spikes from different qubits arrive at the token cell at the same time. Specifically, if the token receives two spikes from an ancilla and data qubit simultaneously, we determine a qubit to reflect the spike based on the qubit-type priority. In many cases, it is more accurate to give priority to ancilla-qubit errors because ancilla qubits are subject to more error-prone operations (e.g., measurement, EPR pair generation). However, preferring data-qubit errors sometimes results in more accurate decoding. Therefore, to properly decode various error patterns, the EDU should consider both qubit-type priorities (i.e., data qubit-type, ancilla qubit-type priority).

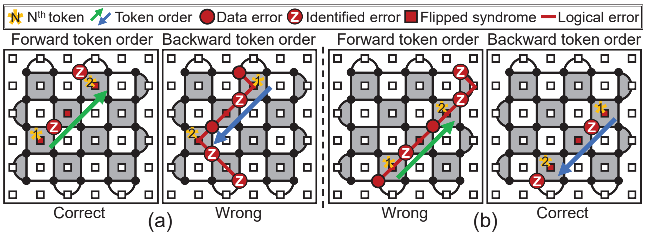


Figure 13. Sensitive decoding result for different token orders; forward order is correct in (a) but backward order is correct in (b).

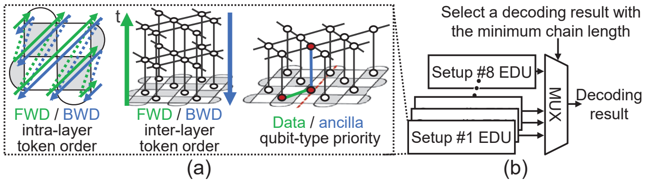
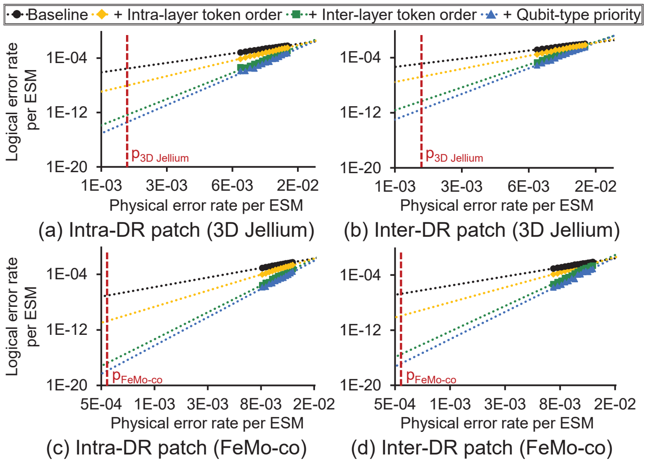


Figure 14. Ensemble EDU; (a) various token-allocating order and qubit-type priority setups and (b) the microarchitecture

(2) Solution. To improve the decoding accuracy, we propose an ensemble EDU architecture. The key idea of this solution is to consider various token-allocating orders and qubit-type priorities simultaneously with multiple EDUs, and then choose the best result based on the global feature (i.e., total length of error chains). As a result, the ensemble EDU produces more accurate results close to the global optimum.

Figure 14 shows the microarchitecture of our ensemble EDU with its various token-order and qubit-priority considerations. First, the ensemble EDU consists of eight EDUs with different token-order and priority setups. For the token-order setups, we consider two orders inside an ESM layer



(a) Intra-DR patch (3D Jellium) (b) Inter-DR patch (3D Jellium)

(c) Intra-DR patch (FeMo-co) (d) Inter-DR patch (FeMo-co)

Figure 15. Patch-level logical error rate of the ensemble EDU; (a) Intra-DR and (b) inter-DR patch results for 3D Jellium; (c) Intra-DR and (d) inter-DR patch results for FeMo-co

(i.e., intra-layer order), and two more orders with the different layer directions (i.e., inter-layer order). In addition, we also consider the qubit-type priority with two different setups. Specifically, if a token receives two spikes from an ancilla qubit and a data qubit simultaneously, we construct an error chain according to the priority. As a result, our ensemble EDU runs eight EDUs with eight different setups (i.e., $2 \times 2 \times 2$) and chooses the best result with the minimum total length of the error chains.

(3) Result. Figure 15 shows our ensemble EDU's patch-level logical error rates while accumulatively considering the intra-layer token order, inter-layer token order, and qubit-type priority. The x-axis indicates average physical error rate per ESM, and each data point is derived by our simulation.

First, by considering the different intra-layer token orders, our ensemble EDU reduces the logical error of intra-DR and inter-DR patches by 4,686 times and 1,617 times, respectively (Baseline vs. +Intra-layer order, FeMo-co). Second, thanks to the inter-layer token order consideration, it further reduces the logical error of intra-DR and inter-DR patches by $1.2 \cdot 10^6$ times and $3.8 \cdot 10^5$ times, respectively (+Intra-layer order vs. +Inter-layer order, FeMo-co). Finally, with the different qubit-priority setups, we additionally reduce the logical error of intra-DR and inter-DR patches by 12.6 times and 22.3 times, respectively (+Inter-layer order vs. +Qubit priority, FeMo-co).

As a result, the ensemble EDU achieves $7.1 \cdot 10^{10}$ times and $1.4 \cdot 10^{10}$ times of huge logical error reduction in intra-DR and inter-DR patches (vs. Baseline, FeMo-co). This significant error reduction originates from the (1) higher probability of finding a global optimum with various setups and (2) global feature consideration enabled by the ensemble (i.e., choosing a setup with the minimum total error-chain length), which have been ignored in the existing spike-based EDUs.

6 Final Evaluation

6.1 ESM latency, decoding latency, and logical error

To clarify the effectiveness of our solutions, we evaluate the ESM latency, error decoding latency, and logical error rate while accumulatively applying each solution. Figure 16(a) and (b) show the results with 4K CMOS and ERSFQ-based EDUs for 3D Jellium and FeMo-co, respectively. Note that the 3D Jellium and FeMo-co are our target applications with different qubit scales (i.e., 215,040 qubits and 2,138,112 qubits), target logical error rates, and physical-qubit error setups.

First, the three ESM solutions greatly reduce the ESM latency and physical error rate. In sum, they reduce the ESM latency by 3.71 times for both 3D Jellium and FeMo-co. In addition, our ESM solutions suppress the physical error of the inter-DR qubits by 2.29 times and 2.37 times for 3D Jellium and FeMo-co, respectively. However, due to the decoding latency longer than the ESM latency, both targets fail with logical error rate 1.0 (i.e., syndrome backlog problem).

Second, our EDU solutions #1 and #2 significantly reduce the decoding latency. Specifically, they reduce the decoding latency of 4K CMOS and ERSFQ-based EDUs by 98.7 times and 123.3 times for 3D Jellium, and 509.0 times and 684.9 times for FeMo-co, respectively. As a result, our target systems achieve the decoding latency lower than the ESM latency and thus overcome the backlog problem.

Finally, our EDU solution #3 drastically reduces the logical error rate. With various token-order and qubit-type priority considerations, it reduces the logical error rate per $PPR_{\frac{\pi}{8}}$ by $3.4 \cdot 10^7$ times and $6.1 \cdot 10^{10}$ times for 3D Jellium and FeMo-co, respectively. As a result, we achieve the logical error lower than the target logical error rate in both systems.

In summary, with three ESM and three EDU solutions, we successfully achieve our goal, realizing a fault-tolerant million qubit-scale quantum computer with multiple DRs.

6.2 Necessity of all of our ESM and EDU solutions

To clarify that all of our solutions are necessary, we evaluate the logical error rates when some of our solutions are absent. Figure 17(a) and (b) show the results for 3D Jellium and FeMo-co with the ERSFQ-based EDU setup. First, without the physical error rate improvement (i.e., no ESM sol #1,2,3), we fail to achieve our goal for both 3D Jellium and FeMo-co, due to the 25,323 times and 122 times higher logical error rates, respectively. Next, without improving decoding latency (i.e., no EDU sol #1,2), we cannot achieve our logical-error goal for both setup with 1.0 of logical error rate. Lastly, without the improvement of the decoding accuracy (i.e., no EDU sol #3), the logical error rates become $3.3 \cdot 10^7$ times and $8.9 \cdot 10^9$ times higher than the logical error targets of 3D Jellium and FeMo-co, respectively. Therefore, all of our solutions are necessary to realize a fault-tolerant million qubit-scale quantum computer.

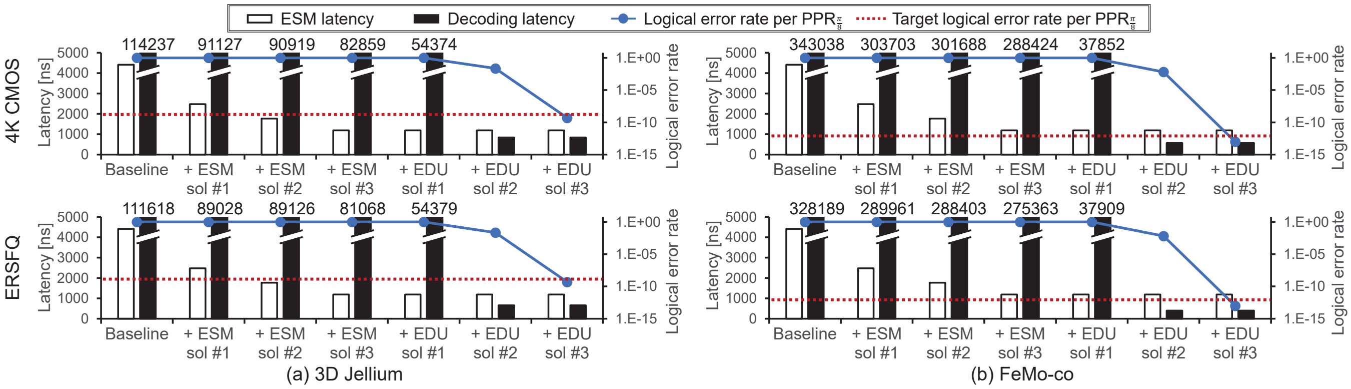


Figure 16. ESM latency, system-level decoding latency per ESM, and system-level logical error per $\text{PPR}_{\frac{\pi}{8}}$ for (a) 3D Jellium and (b) FeMo-co

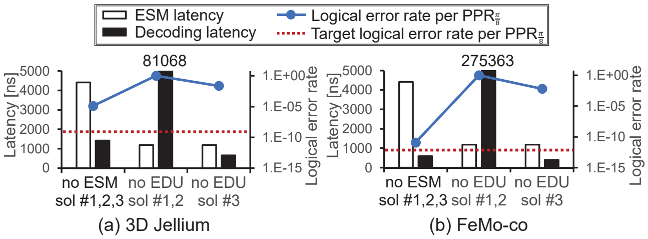


Figure 17. System-level results without the ESM solutions #1,2,3, EDU solutions #1,2, or EDU solution #3; (a) 3D Jellium and (b) FeMo-co with ERSFQ-based EDUs

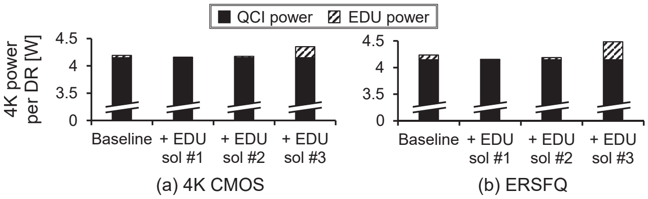


Figure 18. Total 4K runtime power of the system with (a) 4K CMOS and (b) ERSFQ-based EDUs for each EDU solution

6.3 Runtime power consumption at 4K domain

Figure 18 shows the runtime power at 4K (i.e., sum of QCI and EDU power) with our EDU solutions targeting the million-qubit application (FeMo-co). We evaluate the 4K power because it determines the number of qubits inside each DR.

The 4K EDU power changes to 11.5mW, 25.4mW, and 203.3 mW for the 4K CMOS and 6.4mW, 39.4mW, and 336.6mW for the ERSFQ while applying our EDU solutions #1, #2, and #3, respectively. Our final EDU designs consume 3.82 times more power than the baseline due to the 685 times faster decoding and eight times overhead for the ensemble. However, the portion of the EDU power is negligible (7.51%) over the total power. Therefore, the increased EDU power does not affect the number of qubits inside each DR.

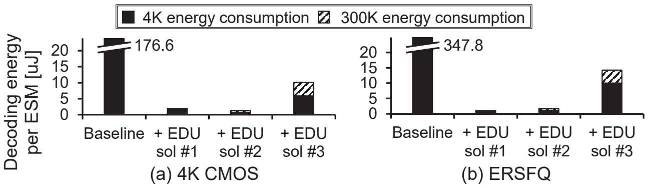


Figure 19. Decoding energy per ESM of (a) 4K CMOS and (b) ERSFQ-based EDUs for each solution

6.4 EDU energy consumption per ESM

Figure 19 shows the decoding energy per ESM with our EDU solutions (FeMo-co). We break down the energy consumption at 300K and 4K for the thorough analysis.

First, the EDU solution #1 reduces the energy by 318 times because it only activates adjacent patches (i.e., syndrome-finding window in Figure 11(b)-(1)) for each token rather than entire patches like the baseline. Next, the EDU solution #2 slightly increases the energy consumption by 1.56 times due to the 300K EDUs and the one more patch to be decoded in the inter-DR 4K EDU (i.e., patch #10). Finally, the EDU solution #3 increases the energy by 8.37 times as the ensemble EDU runs the eight EDUs in parallel. However, its decoding energy is still 24 times lower than the baseline.

7 Discussion

7.1 Generality of our solutions

Although we focus on the two setups for clear description, our solutions are generally applicable over various setups.

Different qubit technologies. Our solutions are also applicable to other qubit technologies. For example, as ESM solutions #1,2,3 reduce the overhead of slow and erroneous EPR pair generation in ESM sequences, they are also beneficial to the qubit technologies that utilize EPR-pair generation (e.g., Transmon, trapped-ion qubits [74, 77], and spin qubit [44]). In addition, a fault-tolerant quantum computer typically requires a large number of qubits regardless of the

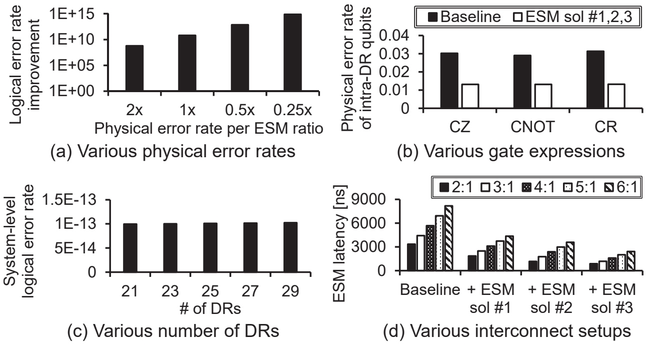


Figure 20. Effectiveness of our solutions for different (a) physical error rates per ESM, (b) two-qubit gate expressions, (c) number of DRs, and (d) interconnect setups

qubit technologies [27, 47]. In this context, EDU solutions #1 and #3 are also effective for other qubit technologies as our solutions significantly increase the speed and accuracy of large-scale EDUs, independent of the specific qubit technologies. EDU solution #2 is also applicable to different qubit technologies if they use many DRs, e.g., spin qubit [34].

Different decoding algorithms. Our EDU solutions #1 and #2 are effective for other decoding algorithms (e.g., MWPM, Union Find) because they also suffer from the collision and inter-DR communication problems when decoding in parallel. In addition, regardless of the decoding algorithm, ESM solutions #1,2,3 are applicable when using EPR-pair generation, because the decoder design is orthogonal to the ESM sequence.

Different physical-qubit errors. Our ESM & EDU solutions significantly reduce the ESM latency, decoding latency, and logical error regardless of the detailed physical-qubit error setup. Figure 20(a) shows the logical error improvement of our solutions with various physical error rates. The graph clearly shows that our solutions are highly effective even with twice higher physical error rate than that of FeMo-co (i.e., reducing logical error by $5.7 \cdot 10^8$ times). Furthermore, as the improvement becomes larger at the lower physical errors, we expect our solutions will be more effective in the future (thanks to the continuously reducing error rates).

Different two-qubit gate expressions. Our ESM solutions are generally applicable to various two-qubit gate expressions. Figure 20(b) shows the physical error rate of intra-DR qubits for various ESM sequences encoded with CZ, CNOT, and CR gates, respectively. For the ESM with CNOT gates, we use ESM sequences in [27]. For the ESM with CR gates, we first adopt the CNOT-based ESM sequence [27] and then convert its CNOT gates into a set of single-qubit gates and CR gates following the converting rule of the previous work [55]. The graph clearly shows that our ESM solutions #1,2,3 greatly reduce the physical error rate for every gate type.

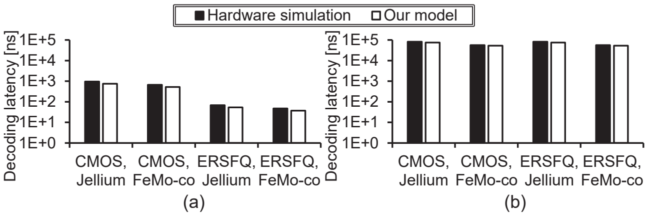


Figure 21. Validation result of our system-level latency modeling against hardware simulation; (a) Nine merged patches inside a DR, and (b) Nine merged patches between two DRs

Different number of DRs and interconnects. Our ESM solutions are effective regardless of the number of DRs and interconnects per DR. Figure 20(c) shows the logical error reduction for various number of DRs. The number of DRs determines the number of qubits inside a single DR and the ratio between inter-DR and intra-DR patches. Note that the different number of DRs represents the different setups in refrigerator power budget, system power budget, and QCI & QCP power consumption. Regardless of the setup, our solutions successfully reduce the logical errors. Figure 20(d) shows the ESM latency reduction for various degree of interconnect sharing. The figure also shows that our ESM solutions greatly reduce the ESM latency for every degree of interconnect sharing.

7.2 Accuracy of our system-level modeling

As we estimate the system-level results (i.e., multi-patch results) by projecting the single-patch results, we validate our system-level modeling methodology by comparing its result with the outcome of multiple-patch-scale hardware simulation. As we follow the widely-used methodology for the system-level logical error prediction [62], we validate the methodology for the system-level decoding latency only.

For the multi-patch hardware simulation, we run ESM operation and error decoding for nine merged patches (18,432 qubits), which consist of only intra-DR patches or seven intra-DR and two inter-DR patches. The two setups represent the patches inside a DR and across DRs, respectively. We validate our methodology at the nine-patch scale because the overall maximum decoding latency after adopting our solution (i.e., parallel nine-patch-granular decoding) becomes the same as the maximum decoding latency of nine patches. We first obtain the latency distribution by running enough shots ($4.0 \cdot 10^5$ shots for each setup, a total of $3.2 \cdot 10^6$ shots), fit the distribution using the widely adopted fitting function (i.e., Gamma distribution function [58, 95]), and derive the maximum decoding latency corresponding to the target probability.

Figure 21(a) and (b) show the validation results for the nine intra-DR patches and the patches including two inter-DR patches, respectively. Our model accurately predicts the overall maximum decoding latency with up to 21.0% and 6.7%

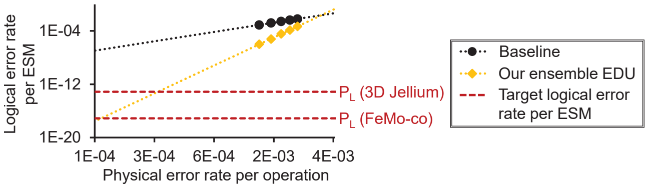


Figure 22. Patch-level logical error rate of the ensemble EDU in terms of physical error rate per operation

of low prediction error, respectively. Note that the prediction error of the system-level latency is similar to the error of inter-DR patches (6.7%) as the decoding latency for inter-DR patches dominates the total system-level decoding latency (as shown in Figure 21).

7.3 Analysis with physical error rate per operation

Even though many previous works show their EDU accuracy with physical error per ESM [18, 19, 36, 82, 98, 104], there are also several previous works that analyze their result with physical error per operation [27, 105, 108]. Therefore, to enable complete comparison with the previous works, we also analyze our decoder's accuracy based on physical error rate per operation. Figure 22 shows our EDU's patch-level logical error rates where the x-axis indicates the physical error of qubits per operation (p_{op}). We derive each data point from our simulation while setting the error of every gate to p_{op}. Our system requires p_{op} of $2.6 \cdot 10^{-4}$ (3D Jellium) and $1.1 \cdot 10^{-4}$ (FeMo-co) to achieve the logical-error targets. Note that our solutions increase the required p_{op} to achieve target logical error rates by 152 times and 864 times for 3D Jellium and FeMo-co, respectively, compared to the baseline.

7.4 Novelty against prior hierarchical & ensemble EDUs

With the importance of error decoding, several previous works proposed hierarchical decoders and a neural ensemble decoder as their solution. We will discuss the differences between our EDU solutions and existing decoder designs in the rest of this section.

Hierarchical decoding. Our EDU solution #2 is orthogonal to the previous hierarchical approaches. While the previous works target to reduce the bandwidth [82, 93], latency [10, 93], and logical errors [29, 71] inside single-DR systems, our hybrid solution aims to reduce communication overhead between DRs in the multi-DR system. Furthermore, to achieve their goal, prior hierarchical EDUs decode the errors of the same qubits twice in a hierarchical manner. On the other hand, the key idea of our solution is to decode the errors of the different qubits in parallel at 4K and 300K.

Neural ensemble decoding. Our EDU solution #3 takes a totally different approach and resolves the limited scalability and effectiveness of the previous neural ensemble decoding [92]. Specifically, the neural ensemble decoder runs like a Tournament predictor [50], rather than the ensembled way: (1) it first predicts the best decoder candidate using the neural-network (NN) based classifier and (2) runs only the selected decoder to decode errors. However, this idea cannot support even a single d=13 surface-code patch due to the exponentially increasing classifier training overhead [92]. In addition, its logical error reduction is limited, as it cannot utilize other decoder's results. On the other hand, our ensemble decoder runs oppositely: (1) first runs all decoder candidates in parallel and (2) selects the best decoding result. Our key idea is to use the global feature (i.e., the total length of error chains) in the post-selection so that we benefit from other decoders without relying on the NN. As a result, our ensemble decoder successfully achieves 10^{10} of logical error reduction at d=31.

7.5 Limitation and future direction of this work

Even though we successfully build a million qubit-scale distributed quantum computer, there are several limitations and future research directions of this work.

First, our CMOS and SFQ-based implementation requires an improvement of the device fabrication technology. Specifically, our SFQ-based EDU needs 200M JJs, and we believe that it will be realized soon considering the state-of-the-art fabrication technology (with 150M JJs/cm^2) and its rapidly evolving trend (twice more JJs per 1.5 years) [101]. In addition, our 4K CMOS-based EDU needs aggressive voltage scaling enabled by cryogenic environments. We also believe that it will be realized soon considering the recent interests and demonstrations of the voltage-scaled CMOS chips at cryogenic temperatures (e.g., TSMC [14], ARM [80]).

Second, it will be challenging to synchronize 300K and 4K EDUs when we actually deploy our hybrid solution. Therefore, it could be a nice future research direction to investigate the cost-effective 4K-300K synchronization techniques.

Lastly, even though our decoder is the only decoder design satisfying both latency and logical error constraints of million qubit systems, its decoding accuracy is still lower than that of the minimum weight perfect matching (MWPM) algorithm. For example, while the MWPM algorithm theoretically achieves an effective distance of 16 at the d=31 logical-qubit patch [36], our ensembled decoder achieves the effective distance of 11. In addition, the accuracy gap becomes larger at lower physical error due to our lower effective distance. Nevertheless, the state-of-the-art MWPM decoders also suffer from an order of magnitude slower error decoding than ours [35, 110], or cannot support larger code distance due to limited scalability (i.e., d=11 of Astrea [105]). Therefore, a future research direction is to find better combination of ensemble-decoder setup achieving lower logical error. For the MWPM decoders, it will also be a nice direction to accelerate their decoding speed while achieving high scalability and maintaining low error.

8 Related work

Scalable quantum computer system. [46] proposed a scalable SFQ-based QCI supporting single and two-qubit operations. [3, 9, 14, 48, 111] fabricated 4K CMOS-based QCIs. [28] introduced the overview of the fault-tolerant QCP. [99] highlighted and resolved the scalability problem of quantum computers due to the 300K-4K wire bandwidth. [7, 73] developed a modeling tool and proposed the fully-implemented 59,000-qubit-scale QCP and 64,000-qubit-scale QCI designs, respectively.

EDU design. [36, 103] proposed the spike-based EDU designs supporting batch and online decoding. [104] developed the spike-based EDUs supporting both online decoding and lattice surgery. [18, 105] proposed the look-up table-based MWPM decoders and [19] proposed the EDU design running Union-Find algorithm. [82] developed the hierarchical EDU design with MWPM and light-weighted EDU.

Multi-DR system. [6] showed their future quantum computer systems using multiple DRs and EPR pairs. [57, 113] developed the mechanisms to entangle the inter-DR qubits in EPR states. By using the entanglement, [107, 109] proposed the compiler for NISQ-targeted distributed quantum computers.

However, no previous work proposed a million qubit multi-DR FTQC systems with multi-DR aware ESM and EDU solutions. To the best of our knowledge, this work is the first to propose the multi-DR aware ESM and EDU designs toward the distributed fault-tolerant quantum computers.

9 Conclusion

In this work, we proposed a million qubit-scale distributed quantum computer with our multi-DR aware error handling mechanisms. First, we developed a multi-DR aware ESM sequence to reduce the number of gates and latency. Second, we proposed a multi-DR aware and scalable EDU to maximize both decoding speed and accuracy. Our evaluation showed that our ESM and EDU designs significantly reduced the logical error and successfully supported a practical quantum application over two million qubits.

Acknowledgments

This work was supported in part by National Research Foundation of Korea (NRF) funded by the Korean Government under Grants NRF-2019R1A5A1027055, NRF-2021R1A2C3014131, NRF-2021M3F3A2A02037893, and NRF-2019-Global Ph.D. Fellowship Program, by the Creative Pioneering Researchers Program through Seoul National University, and by the Automation and Systems Research Institute (ASRI) and Inter-university Semiconductor Research Center (ISRC) at Seoul National University.

References

[1] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C. Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando G. S. L. Brandao, David A. Buell, Brian Burkett, Yu Chen, Zijun Chen, Ben Chiaro, Roberto Collins, William Courtney, Andrew Dunsworth, Edward Farhi, Brooks Foxen, Austin Fowler, Craig Gidney, Marissa Giustina, Rob Graff, Keith Guerin, Steve Habegger, Matthew P. Harrigan, Michael J. Hartmann, Alan Ho, Markus Hoffmann, Trent Huang, Travis S. Humble, Sergei V. Isakov, Evan Jeffrey, Zhang Jiang, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Paul V. Klimov, Sergey Knysh, Alexander Korotkov, Fedor Kostritsa, David Landhuis, Mike Lindmark, Erik Lucero, Dmitry Lyakh, Salvatore Mandrà, Jarrod R. McClean, Matthew McEwen, Anthony Megrant, Xiao Mi, Kristel Michielsen, Masoud Mohseni, Josh Mutus, Ofer Naaman, Matthew Neeley, Charles Neill, Murphy Yuezhen Niu, Eric Ostby, Andre Petukhov, John C. Platt, Chris Quintana, Eleanor G. Rieffel, Pedram Roushan, Nicholas C. Rubin, Daniel Sank, Kevin J. Satzinger, Vadim Smelyanskiy, Kevin J. Sung, Matthew D. Trevithick, Amit Vainsencher, Benjamin Villalonga, Theodore White, Z. Jamie Yao, Ping Yeh, Adam Zalcman, Hartmut Neven, and John M. Martinis. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, Oct 2019.

[2] Joseph Bardin. Beyond-classical computing using superconducting quantum processors. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 65, pages 422–424. IEEE, 2022.

[3] Joseph C. Bardin, Evan Jeffrey, Erik Lucero, Trent Huang, Sayan Das, Daniel Thomas Sank, Ofer Naaman, Anthony Edward Megrant, Rami Barends, Ted White, Marissa Giustina, Kevin J. Satzinger, Kunal Arya, Pedram Roushan, Benjamin Chiaro, Julian Kelly, Zijun Chen, Brian Burkett, Yu Chen, Andrew Dunsworth, Austin Fowler, Brooks Foxen, Craig Gidney, Rob Graff, Paul Klimov, Josh Mutus, Matthew J. McEwen, Matthew Neeley, Charles J. Neill, Chris Quintana, Amit Vainsencher, Hartmut Neven, and John Martinis. Design and characterization of a 28-nm bulk-cmos cryogenic quantum controller dissipating less than 2 mw at 3 k. *IEEE Journal of Solid-State Circuits*, 54(11):3043–3060, 2019.

[4] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and John M. Martinis. Superconducting quantum circuits at the surface code threshold for fault tolerance. *Nature*, 508(7497):500–503, Apr 2014.

[5] Bluefors. We made 1 000 qubits for quantum computing possible. https://bluefors.com/blog/we-made-1-000-qubits-for-quantum-computing-possible/, 2023. [Online Accessed, 09-August-2023].

[6] Sergey Bravyi, Oliver Dial, Jay M Gambetta, Darío Gil, and Zaira Nazario. The future of quantum computing with superconducting qubits. *Journal of Applied Physics*, 132(16), 2022.

[7] Ilkwon Byun, Junpyo Kim, Dongmoon Min, Ikki Nagaoka, Kosuke Fukumitsu, Iori Ishikawa, Teruo Tanimoto, Masamitsu Tanaka, Koji Inoue, and Jangwoo Kim. Xqsim: modeling cross-technology control processors for 10+ k qubit quantum computers. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pages 366–382, 2022.

[8] P. Campagne-Ibarcq, E. Zalys-Geller, A. Narla, S. Shankar, P. Reinhold, L. Burkhart, C. Axline, W. Pfaff, L. Frunzio, R. J. Schoelkopf, and M. H. Devoret. Deterministic remote entanglement of superconducting circuits through microwave two-photon transitions. *Phys. Rev. Lett.*, 120:200501, May 2018.

[9] Sudipto Chakraborty, David J. Frank, Kevin Tien, Pat Rosno, Mark Yeck, Joseph A. Glick, Raphael Robertazzi, Ray Richetta, John F. Bulzacchelli, Devin Underwood, Daniel Ramirez, Dereje Yilma, Andrew Davies, Rajiv V. Joshi, Shawn D. Chambers, Scott Lekuch, Ken Inoue, Dorothy Wisnieff, Christian W. Baks, Donald S. Bethune, John

Timmerwilke, Thomas Fox, Peilin Song, Blake R. Johnson, Brian P. Gaucher, and Daniel J. Friedman. A cryo-cmos low-power semi-autonomous transmon qubit state controller in 14-nm finfet technology. *IEEE Journal of Solid-State Circuits*, 57(11):3258–3273, 2022.

[10] Christopher Chamberland, Luis Goncalves, Prasahnt Sivarajah, Eric Peterson, and Sebastian Grimberg. Techniques for combining fast local decoders with global decoders under circuit-level noise. *Quantum Science and Technology*, 8(4):045011, 2023.

[11] Liangyu Chen, Hang-Xi Li, Yong Lu, Christopher W. Warren, Christian J. Križan, Sandoko Kosen, Marcus Rommel, Shahnawaz Ahmed, Amr Osman, Janka Biznárová, Anita Fadavi Roudsari, Benjamin Lienhard, Marco Caputo, Kestutis Grigoras, Leif Grönberg, Joonas Govenius, Anton Frisk Kockum, Per Delsing, Jonas Bylander, and Giovanna Tancredi. Transmon qubit readout fidelity at the threshold for quantum error correction without a quantum-limited amplifier. *npj Quantum Information*, 9(1):26, Mar 2023.

[12] Yu Chen, C. Neill, P. Roushan, N. Leung, M. Fang, R. Barends, J. Kelly, B. Campbell, Z. Chen, B. Chiaro, A. Dunsworth, E. Jeffrey, A. Megrant, J. Y. Mutus, P. J. J. O'Malley, C. M. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, Michael R. Geller, A. N. Cleland, and John M. Martinis. Qubit architecture with high coherence and fast tunable coupling. *Phys. Rev. Lett.*, 113:220502, Nov 2014.

[13] Zijun Chen, Julian Kelly, Chris Quintana, R. Barends, B. Campbell, Yu Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Lucero, E. Jeffrey, A. Megrant, J. Mutus, M. Neeley, C. Neill, P. J. J. O'Malley, P. Roushan, D. Sank, A. Vainsencher, J. Wenner, T. C. White, A. N. Korotkov, and John M. Martinis. Measuring and suppressing quantum state leakage in a superconducting qubit. *Phys. Rev. Lett.*, 116:020501, Jan 2016.

[14] H.-L. Chiang, R. A. Hadi, J.-F. Wang, H.-C. Han, J.-J. Wu, H.-H. Hsieh, J.-J. Horng, W.-S. Chou, B.-S. Lien, C.-H. Chang, Y.-C. Chen, Y.-H. Wang, T.-C. Chen, J.-C. Liu, Y.-C. Liu, M.-H. Chiang, K.-H. Kao, B. Pulicherla, J. Cai, C.-S. Chang, K.-W. Su, K.-L. Cheng, T.-J. Yeh, Y.-C. Peng, C. Enz, M.-C. F. Chang, M.-F. Chang, H.-S. P. Wong, and I. P. Radu. How fault-tolerant quantum computing benefits from cryo-cmos technology. In *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pages 1–2, 2023.

[15] Jerry M. Chow, A. D. Córcoles, Jay M. Gambetta, Chad Rigetti, B. R. Johnson, John A. Smolin, J. R. Rozen, George A. Keefe, Mary B. Rothwell, Mark B. Ketchen, and M. Steffen. Simple all-microwave entangling gate for fixed-frequency superconducting qubits. *Phys. Rev. Lett.*, 107:080502, Aug 2011.

[16] Jerry M Chow, Jay M Gambetta, Andrew W Cross, Seth T Merkel, Chad Rigetti, and M Steffen. Microwave-activated conditional-phase gate for superconducting qubits. *New Journal of Physics*, 15(11):115012, nov 2013.

[17] Jerry M. Chow, Jay M. Gambetta, Easwar Magesan, David W. Abraham, Andrew W. Cross, B. R. Johnson, Nicholas A. Masluk, Colm A. Ryan, John A. Smolin, Srikanth J. Srinivasan, and M. Steffen. Implementing a strand of a scalable fault-tolerant quantum computing fabric. *Nature Communications*, 5(1):4015, Jun 2014.

[18] Poulami Das, Aditya Locharla, and Cody Jones. Lilliput: A lightweight low-latency lookup-table decoder for near-term quantum error correction. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '22, page 541–553, New York, NY, USA, 2022. Association for Computing Machinery.

[19] Poulami Das, Christopher A. Pattison, Srilatha Manne, Douglas M. Carmean, Krysta M. Svore, Moinuddin Qureshi, and Nicolas Delfosse. Afs: Accurate, fast, and scalable error-decoding for fault-tolerant quantum computers. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 259–273, 2022.

[20] R. Dassonneville, T. Ramos, V. Milchakov, L. Planat, É. Dumur, F. Foroughi, J. Puertas, S. Leger, K. Bharadwaj, J. Delaforce, C. Naud, W. Hasch-Guichard, J. J. García-Ripoll, N. Roch, and O. Buisson. Fast high-fidelity quantum nondemolition qubit readout via a nonperturbative cross-kerr coupling. *Phys. Rev. X*, 10:011045, Feb 2020.

[21] Eric Dennis, Alexei Kitaev, Andrew Landahl, and John Preskill. Topological quantum memory. *Journal of Mathematical Physics*, 43(9):4452–4505, 08 2002.

[22] A. Dewes, F. R. Ong, V. Schmitt, R. Lauro, N. Boulant, P. Bertet, D. Vion, and D. Esteve. Characterization of a two-transmon processor with individual single-shot qubit readout. *Phys. Rev. Lett.*, 108:057002, Feb 2012.

[23] Yongshan Ding, Pranav Gokhale, Sophia Fuhui Lin, Richard Rines, Thomas Propson, and Frederic T. Chong. Systematic crosstalk mitigation for superconducting qubits via frequency-aware compilation. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 201–214, 2020.

[24] Austin G Fowler. Optimal complexity correction of correlated errors in the surface code. *arXiv preprint arXiv:1310.0863*, 2013.

[25] Austin G. Fowler, Simon J. Devitt, and Cody Jones. Surface code implementation of block code state distillation. *Scientific Reports*, 3(1):1939, Jun 2013.

[26] Austin G. Fowler and Craig Gidney. Low overhead quantum computation using lattice surgery, 2019.

[27] Austin G. Fowler, Matteo Mariantoni, John M. Martinis, and Andrew N. Cleland. Surface codes: Towards practical large-scale quantum computation. *Phys. Rev. A*, 86:032324, Sep 2012.

[28] X. Fu, M. A. Rol, C. C. Bultink, J. van Someren, N. Khammassi, I. Ashraf, R. F. L. Vermeulen, J. C. de Sterke, W. J. Vlothuizen, R. N. Schouten, C. G. Almudever, L. DiCarlo, and K. Bertels. A microarchitecture for a superconducting quantum processor. *IEEE Micro*, 38(3):40–47, 2018.

[29] Spiro Gicev, Lloyd CL Hollenberg, and Muhammad Usman. A scalable and fast artificial neural network syndrome decoder for surface codes. *Quantum*, 7:1058, 2023.

[30] Craig Gidney. Stim: a fast stabilizer circuit simulator. *Quantum*, 5:497, July 2021.

[31] Craig Gidney and Martin Ekerå. How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits. *Quantum*, 5:433, April 2021.

[32] Simon Gustavsson, Olger Zwier, Jonas Bylander, Fei Yan, Fumiki Yoshihara, Yasunobu Nakamura, Terry P. Orlando, and William D. Oliver. Improving quantum gate fidelities by using a qubit to measure microwave pulse distortions. *Phys. Rev. Lett.*, 110:040502, Jan 2013.

[33] Johannes Heinsoo, Christian Kraglund Andersen, Ants Remm, Sebastian Krinner, Theodore Walter, Yves Salathé, Simone Gasparinetti, Jean-Claude Besse, Anton Potočnik, Andreas Wallraff, and Christopher Eichler. Rapid high-fidelity multiplexed readout of superconducting qubits. *Phys. Rev. Appl.*, 10:034040, Sep 2018.

[34] Daniel B. Higginbottom, Alexander T. K. Kurkjian, Camille Chartrand, Moein Kazemi, Nicholas A. Brunelle, Evan R. MacQuarrie, James R. Klein, Nicholas R. Lee-Hone, Jakub Stacho, Myles Ruether, Camille Bowness, Laurent Bergeron, Adam DeAbreu, Stephen R. Harrigan, Joshua Kanaganayagam, Danica W. Marsden, Timothy S. Richards, Leea A. Stott, Sjoerd Roorda, Kevin J. Morse, Michael L. W. Thewalt, and Stephanie Simmons. Optical observation of single spins in silicon. *Nature*, 607(7918):266–270, Jul 2022.

[35] Oscar Higgott and Craig Gidney. Sparse blossom: correcting a million errors per core second with minimum-weight matching. *arXiv preprint arXiv:2303.15933*, 2023.

[36] Adam Holmes, Mohammad Reza Jokar, Ghasem Pasandi, Yongshan Ding, Massoud Pedram, and Frederic T. Chong. Nisq+: Boosting quantum computing power by approximating quantum error correction. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pages 556–569, 2020.

[37] Dominic Horsman, Austin G Fowler, Simon Devitt, and Rodney Van Meter. Surface code quantum computing by lattice surgery. *New Journal of Physics*, 14(12):123011, dec 2012.

[38] Fei Hua, Yanhao Chen, Yuwei Jin, Chi Zhang, Ari Hayes, Youtao Zhang, and Eddy Z. Zhang. Autobraid: A framework for enabling efficient surface code communication in quantum computing. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '21, page 925–936, New York, NY, USA, 2021. Association for Computing Machinery.

[39] IBM. Ibm's roadmap for scaling quantum technology. https://research.ibm.com/blog/ibm-quantum-roadmap, 2020. [Online Accessed, 09-August-2023].

[40] IBM. Ibm scientists cool down the world's largest quantum-ready cryogenic concept system. https://research.ibm.com/blog/goldeneye-cryogenic-concept-system, 2022. [Online Accessed, 09-August-2023].

[41] IBM. Ibmq. https://quantum-computing.ibm.com, 2023. [Online Accessed, 05-July-2023].

[42] Evan Jeffrey, Daniel Sank, J. Y. Mutus, T. C. White, J. Kelly, R. Barends, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. Megrant, P. J. J. O'Malley, C. Neill, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and John M. Martinis. Fast accurate state measurement with superconducting qubits. *Phys. Rev. Lett.*, 112:190504, May 2014.

[43] X. Y. Jin, A. Kamal, A. P. Sears, T. Gudmundsen, D. Hover, J. Miloshi, R. Slattery, F. Yan, J. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver. Thermal and residual excited-state population in a 3d transmon qubit. *Phys. Rev. Lett.*, 114:240501, Jun 2015.

[44] Hamza Jnane, Brennan Undseth, Zhenyu Cai, Simon C Benjamin, and Bálint Koczor. Multicore quantum computing. *Physical Review Applied*, 18(4):044064, 2022.

[45] J. E. Johnson, C. Macklin, D. H. Slichter, R. Vijay, E. B. Weingarten, John Clarke, and I. Siddiqi. Heralded state preparation in a superconducting qubit. *Phys. Rev. Lett.*, 109:050506, Aug 2012.

[46] Mohammad Reza Jokar, Richard Rines, Ghasem Pasandi, Haolin Cong, Adam Holmes, Yunong Shi, Massoud Pedram, and Frederic T. Chong. Digiq: A scalable digital controller for quantum computers using sfq logic. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 400–414, 2022.

[47] N Cody Jones, Rodney Van Meter, Austin G Fowler, Peter L McMahon, Jungsang Kim, Thaddeus D Ladd, and Yoshihisa Yamamoto. Layered architecture for quantum computing. *Physical Review X*, 2(3):031007, 2012.

[48] Kiseo Kang, Donggyu Minn, Seongun Bae, Jaeho Lee, Seokhyeong Kang, Moonjoo Lee, Ho-Jin Song, and Jae-Yoon Sim. A 40-nm cryo-cmos quantum controller ic for superconducting qubit. *IEEE Journal of Solid-State Circuits*, 57(11):3274–3287, 2022.

[49] J. Kelly, R. Barends, A. G. Fowler, A. Megrant, E. Jeffrey, T. C. White, D. Sank, J. Y. Mutus, B. Campbell, Yu Chen, Z. Chen, B. Chiaro, A. Dunsworth, I.-C. Hoi, C. Neill, P. J. J. O'Malley, C. Quintana, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and John M. Martinis. State preservation by repetitive error detection in a superconducting quantum circuit. *Nature*, 519(7541):66–69, Mar 2015.

[50] Richard E Kessler. The alpha 21264 microprocessor. *IEEE micro*, 19(2):24–36, 1999.

[51] D. E. Kirichenko, S. Sarwana, and A. F. Kirichenko. Zero static power dissipation biasing of rsfq circuits. *IEEE Transactions on Applied Superconductivity*, 21(3):776–779, 2011.

[52] Ian D. Kivlichan, Craig Gidney, Dominic W. Berry, Nathan Wiebe, Jarrod McClean, Wei Sun, Zhang Jiang, Nicholas Rubin, Austin Fowler, Alán Aspuru-Guzik, Hartmut Neven, and Ryan Babbush. Improved Fault-Tolerant Quantum Simulation of Condensed-Phase Correlated Electrons via Trotterization. *Quantum*, 4:296, July 2020.

[53] Morten Kjaergaard, Mollie E. Schwartz, Ami Greene, Gabriel O. Samach, Andreas Bengtsson, Michael O'Keeffe, Christopher M. McNally, Jochen Braumüller, David K. Kim, Philip Krantz, Milad Marvian, Alexander Melville, Bethany M. Niedzielski, Youngkyu Sung, Roni Winik, Jonilyn Yoder, Danna Rosenberg, Kevin Obenland, Seth Lloyd, Terry P. Orlando, Iman Marvian, Simon Gustavsson, and William D. Oliver. Programming a quantum computer with quantum instructions, 2020.

[54] Philip Krantz, Andreas Bengtsson, Michaël Simoen, Simon Gustavsson, Vitaly Shumeiko, W. D. Oliver, C. M. Wilson, Per Delsing, and Jonas Bylander. Single-shot read-out of a superconducting qubit using a josephson parametric oscillator. *Nature Communications*, 7(1):11417, May 2016.

[55] Philip Krantz, Morten Kjaergaard, Fei Yan, Terry P Orlando, Simon Gustavsson, and William D Oliver. A quantum engineer's guide to superconducting qubits. *Applied physics reviews*, 6(2), 2019.

[56] H. Krauter, D. Salart, C. A. Muschik, J. M. Petersen, Heng Shen, T. Fernholz, and E. S. Polzik. Deterministic quantum teleportation between distant atomic objects. *Nature Physics*, 9(7):400–404, Jul 2013.

[57] P. Kurpiers, P. Magnard, T. Walter, B. Royer, M. Pechal, J. Heinsoo, Y. Salathé, A. Akin, S. Storz, J.-C. Besse, S. Gasparinetti, A. Blais, and A. Wallraff. Deterministic quantum state transfer and remote entanglement using microwave photons. *Nature*, 558(7709):264–267, Jun 2018.

[58] Pierre Simon Laplace. *Théorie analytique des probabilités*. Courcier, 1814.

[59] Joonho Lee, Dominic W. Berry, Craig Gidney, William J. Huggins, Jarrod R. McClean, Nathan Wiebe, and Ryan Babbush. Even more efficient quantum computations of chemistry through tensor hypercontraction. *PRX Quantum*, 2:030305, Jul 2021.

[60] N. Leung, Y. Lu, S. Chakram, R. K. Naik, N. Earnest, R. Ma, K. Jacobs, A. N. Cleland, and D. I. Schuster. Deterministic bidirectional communication and remote entanglement generation between superconducting qubits. *npj Quantum Information*, 5(1):18, Feb 2019.

[61] Ying Li and Simon C. Benjamin. Hierarchical surface code for network quantum computing with modules of arbitrary size. *Phys. Rev. A*, 94:042303, Oct 2016.

[62] Daniel Litinski. A Game of Surface Codes: Large-Scale Quantum Computing with Lattice Surgery. *Quantum*, 3:128, March 2019.

[63] Daniel Litinski and Felix von Oppen. Lattice Surgery with a Twist: Simplifying Clifford Gates of Surface Codes. *Quantum*, 2:62, May 2018.

[64] A. Lupaşcu, S. Saito, T. Picot, P. C. de Groot, C. J. P. M. Harmans, and J. E. Mooij. Quantum non-demolition measurement of a superconducting two-level system. *Nature Physics*, 3(2):119–123, Feb 2007.

[65] P. Magnard, S. Storz, P. Kurpiers, J. Schär, F. Marxer, J. Lütolf, T. Walter, J.-C. Besse, M. Gabureac, K. Reuer, A. Akin, B. Royer, A. Blais, and A. Wallraff. Microwave quantum link between superconducting circuits housed in spatially separated cryogenic systems. *Phys. Rev. Lett.*, 125:260502, Dec 2020.

[66] François Mallet, Florian R. Ong, Agustin Palacios-Laloy, François Nguyen, Patrice Bertet, Denis Vion, and Daniel Esteve. Single-shot qubit readout in circuit quantum electrodynamics. *Nature Physics*, 5(11):791–795, Nov 2009.

[67] Fabian Marxer, Antti Vepsäläinen, Shan W. Jolin, Jani Tuorila, Alessandro Landra, Caspar Ockeloen-Korppi, Wei Liu, Olli Ahonen, Adrian Auer, Lucien Belzane, Ville Bergholm, Chun Fai Chan, Kok Wai Chan, Tuukka Hiltunen, Juho Hotari, Eric Hyyppä, Joni Ikonen, David Janzso, Miikka Koistinen, Janne Kotilahti, Tianyi Li, Jyrgen Luus, Miha Papic, Matti Partanen, Jukka Räbinä, Jari Rosti, Mykhailo Savytskyi, Marko Seppälä, Vasilii Sevriuk, Eelis Takala, Brian Tarasinski, Manish J. Thapa, Francesca Tosto, Natalia Vorobeva, Liuqi Yu, Kuan Yen Tan, Juha Hassel, Mikko Möttönen, and Johannes Heinsoo. Long-distance transmon coupler with cz-gate fidelity above 99.8%. *PRX Quantum*, 4:010314, Feb 2023.

[68] Elisha Siddiqui Matekole, Yao-Lung Leo Fang, and Meifeng Lin. Methods and results for quantum optimal pulse control on superconducting qubit systems. In *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 600–606, 2022.

[69] R McDermott, M G Vavilov, B L T Plourde, F K Wilhelm, P J Liebermann, O A Mukhanov, and T A Ohki. Quantum–classical interface based on single flux quantum digital logic. *Quantum Science and Technology*, 3(2):024004, jan 2018.

[70] David C. McKay, Christopher J. Wood, Sarah Sheldon, Jerry M. Chow, and Jay M. Gambetta. Efficient z gates for quantum computing. *Phys. Rev. A*, 96:022330, Aug 2017.

[71] Kai Meinerz, Chae-Yeun Park, and Simon Trebst. Scalable neural decoder for topological surface codes. *Physical Review Letters*, 128(8):080505, 2022.

[72] Microsoft. Microsoft achieves first milestone towards a quantum supercomputer. https://cloudblogs.microsoft.com/quantum/2023/06/21/microsoft-achieves-first-milestone-towards-a-quantum-supercomputer, 2023. [Online Accessed, 09-August-2023].

[73] Dongmoon Min, Junpyo Kim, Junhyuk Choi, Ilkwon Byun, Masamitsu Tanaka, Koji Inoue, and Jangwoo Kim. Qisim: Architecting 10+ k qubit qc interfaces toward quantum supremacy. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–16, 2023.

[74] C. Monroe, R. Raussendorf, A. Ruthven, K. R. Brown, P. Maunz, L.-M. Duan, and J. Kim. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Phys. Rev. A*, 89:022317, Feb 2014.

[75] Hartmut Neven. Quantum summer symposium 2020 opening keynote. https://www.youtube.com/watch?v=TJ6vBNEQReU, 2020. [Online Accessed, 09-August-2023].

[76] Michael A Nielsen and Isaac L Chuang. Quantum computation and quantum information. *Phys. Today*, 54(2):60, 2001.

[77] Ramil Nigmatullin, Christopher J Ballance, Niel de Beaudrap, and Simon C Benjamin. Minimally complex ion traps as modules for quantum communication and computing. *New Journal of Physics*, 18(10):103028, oct 2016.

[78] Hanhee Paik, D. I. Schuster, Lev S. Bishop, G. Kirchmair, G. Catelani, A. P. Sears, B. R. Johnson, M. J. Reagor, L. Frunzio, L. I. Glazman, S. M. Girvin, M. H. Devoret, and R. J. Schoelkopf. Observation of high coherence in josephson junction qubits measured in a three-dimensional circuit qed architecture. *Phys. Rev. Lett.*, 107:240501, Dec 2011.

[79] Jongseok Park, Sushil Subramanian, Lester Lampert, Todor Mladenov, Ilya Klotchkov, Dileep J. Kurian, Esdras Juarez-Hernandez, Brando Perez Esparza, Sirisha Rani Kale, Asma Beevi K. T., Shavindra P. Premaratne, Thomas F. Watson, Satoshi Suzuki, Mustafijur Rahman, Jaykant B. Timbadiya, Saksham Soni, and Stefano Pellerano. A fully integrated cryo-cmos soc for state manipulation, readout, and high-speed gate pulsing of spin qubits. *IEEE Journal of Solid-State Circuits*, 56(11):3289–3306, 2021.

[80] Divya Prasad, Manoj Vangala, Mudit Bhargava, Arnout Beckers, Alexander Grill, Davide Tierno, Krishnendra Nathella, Thanusree Achuthan, David Pietromonaco, James Myers, Matthew Walker, Bertrand Parvais, and Brian Cline. Cryo-computing for infrastructure applications: A technology-to-microarchitecture co-optimization study. In *2022 International Electron Devices Meeting (IEDM)*, pages 23.5.1–23.5.4, 2022.

[81] Qiskit contributors. Qiskit: An open-source framework for quantum computing, 2023.

[82] Gokul Subramanian Ravi, Jonathan M. Baker, Arash Fayyazi, Sophia Fuhui Lin, Ali Javadi-Abhari, Massoud Pedram, and Frederic T. Chong. Better than worst-case decoding for quantum error correction. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS 2023, page 88–102, New York, NY, USA, 2023. Association for Computing Machinery.

[83] Matthew Reagor, Christopher B. Osborn, Nikolas Tezak, Alexa Staley, Guenevere Prawiroatmodjo, Michael Scheer, Nasser Alidoust, Eyob A. Sete, Nicolas Didier, Marcus P. da Silva, Ezer Acala, Joel Angeles, Andrew Bestwick, Maxwell Block, Benjamin Bloom, Adam Bradley, Catvu Bui, Shane Caldwell, Lauren Capelluto, Rick Chilcott, Jeff Cordova, Genya Crossman, Michael Curtis, Saniya Deshpande, Tristan El Bouayadi, Daniel Girshovich, Sabrina Hong, Alex Hudson, Peter Karalekas, Kat Kuang, Michael Lenihan, Riccardo Manenti, Thomas Manning, Jayss Marshall, Yuvraj Mohan, William O'Brien, Johannes Otterbach, Alexander Papageorge, Jean-Philip Paquette, Michael Pelstring, Anthony Polloreno, Vijay Rawat, Colm A. Ryan, Russ Renzas, Nick Rubin, Damon Russel, Michael Rust, Diego Scarabelli, Michael Selvanayagam, Rodney Sinclair, Robert Smith, Mark Suska, Ting-Wai To, Mehrnoosh Vahidpour, Nagesh Vodrahalli, Tyler Whyland, Kamal Yadav, William Zeng, and Chad T. Rigetti. Demonstration of universal parametric entangling gates on a multi-qubit lattice. *Science Advances*, 4(2):eaao3603, 2018.

[84] M. D. Reed, L. DiCarlo, B. R. Johnson, L. Sun, D. I. Schuster, L. Frunzio, and R. J. Schoelkopf. High-fidelity readout in circuit quantum electrodynamics using the jaynes-cummings nonlinearity. *Phys. Rev. Lett.*, 105:173601, Oct 2010.

[85] L. Riesebos, X. Fu, S. Varsamopoulos, C. G. Almudever, and K. Bertels. Pauli frames for quantum computer architectures. In *Proceedings of the 54th Annual Design Automation Conference 2017*, DAC '17, New York, NY, USA, 2017. Association for Computing Machinery.

[86] M. A. Rol, C. C. Bultink, T. E. O'Brien, S. R. de Jong, L. S. Theis, X. Fu, F. Luthi, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, D. Deurloo, R. N. Schouten, F. K. Wilhelm, and L. DiCarlo. Restless tuneup of high-fidelity qubit gates. *Phys. Rev. Appl.*, 7:041001, Apr 2017.

[87] Yves Salathé, Philipp Kurpiers, Thomas Karg, Christian Lang, Christian Kraglund Andersen, Abdulkadir Akin, Sebastian Krinner, Christopher Eichler, and Andreas Wallraff. Low-latency digital signal processing for feedback and feedforward in quantum computing and communication. *Phys. Rev. Appl.*, 9:034011, Mar 2018.

[88] Lieze Schindler, Johannes A Delport, and Coenrad J Fourie. The coldflux rsfq cell library for mit-ll sfq5ee fabrication process. *IEEE Transactions on Applied Superconductivity*, 32(2):1–7, 2021.

[89] Zheng Shan, Yu Zhu, and Bo Zhao. A high-performance compilation strategy for multiplexing quantum control architecture. *Scientific Reports*, 12(1):7132, May 2022.

[90] Sarah Sheldon, Lev S. Bishop, Easwar Magesan, Stefan Filipp, Jerry M. Chow, and Jay M. Gambetta. Characterizing errors on qubit operations via iterative randomized benchmarking. *Phys. Rev. A*, 93:012301, Jan 2016.

[91] Sarah Sheldon, Easwar Magesan, Jerry M. Chow, and Jay M. Gambetta. Procedure for systematically tuning up cross-talk in the cross-resonance gate. *Phys. Rev. A*, 93:060302, Jun 2016.

[92] Milap Sheth, Sara Zafar Jafarzadeh, and Vlad Gheorghiu. Neural ensemble decoding for topological quantum error-correcting codes. *Physical Review A*, 101(3):032338, 2020.

[93] Samuel C Smith, Benjamin J Brown, and Stephen D Bartlett. Local predecoder to reduce the bandwidth and latency of quantum error correction. *Physical Review Applied*, 19(3):034050, 2023.

[94] Aaron Somoroff, Quentin Ficheux, Raymond A. Mencia, Haonan Xiong, Roman Kuzmin, and Vladimir E. Manucharyan. Millisecond coherence in a superconducting qubit. *Phys. Rev. Lett.*, 130:267001, Jun 2023.

[95] E. W. Stacy. A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192, 1962.

[96] Aaron Stillmaker and Bevan Baas. Scaling equations for the accurate prediction of cmos device performance from 180nm to 7nm. *Integration*, 58:74–81, 2017.

[97] Youngkyu Sung, Leon Ding, Jochen Braumüller, Antti Vepsäläinen, Bharath Kannan, Morten Kjaergaard, Ami Greene, Gabriel O. Samach, Chris McNally, David Kim, Alexander Melville, Bethany M. Niedzielski, Mollie E. Schwartz, Jonilyn L. Yoder, Terry P. Orlando, Simon

Gustavsson, and William D. Oliver. Realization of high-fidelity cz and *zz*-free iswap gates with a tunable coupler. *Phys. Rev. X*, 11:021058, Jun 2021.

[98] Yasunari Suzuki, Takanori Sugiyama, Tomochika Arai, Wang Liao, Koji Inoue, and Teruo Tanimoto. Q3de: A fault-tolerant quantum computer architecture for multi-bit burst errors by cosmic rays. In *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 1110–1125. IEEE, 2022.

[99] Swamit S. Tannu, Zachary A. Myers, Prashant J. Nair, Douglas M. Carmean, and Moinuddin K. Qureshi. Taming the instruction bandwidth of quantum computers via hardware-managed error correction. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-50 '17, page 679–691, New York, NY, USA, 2017. Association for Computing Machinery.

[100] Barbara M. Terhal. Quantum error correction for quantum memories. *Rev. Mod. Phys.*, 87:307–346, Apr 2015.

[101] Sergey K Tolpygo, Justin L Mallek, Vladimir Bolkhovsky, Ravi Rastogi, Evan B Golden, Terence J Weir, Leonard M Johnson, and Mark A Gouker. Progress toward superconductor electronics fabrication process with planarized nbn and nbn/nb layers. *IEEE Transactions on Applied Superconductivity*, 33(5):1–12, 2023.

[102] S. Touzard, A. Kou, N. E. Frattini, V. V. Sivak, S. Puri, A. Grimm, L. Frunzio, S. Shankar, and M. H. Devoret. Gated conditional displacement readout of superconducting qubits. *Phys. Rev. Lett.*, 122:080502, Feb 2019.

[103] Yosuke Ueno, Masaaki Kondo, Masamitsu Tanaka, Yasunari Suzuki, and Yutaka Tabuchi. Qecool: On-line quantum error correction with a superconducting decoder for surface code. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 451–456, 2021.

[104] Yosuke Ueno, Masaaki Kondo, Masamitsu Tanaka, Yasunari Suzuki, and Yutaka Tabuchi. Qulatis: A quantum error correction methodology toward lattice surgery. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 274–287, 2022.

[105] Suhas Vittal, Poulami Das, and Moinuddin Qureshi. Astrea: Accurate quantum error-decoding via practical minimum-weight perfect-matching. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, ISCA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[106] Chenlu Wang, Xuegang Li, Huikai Xu, Zhiyuan Li, Junhua Wang, Zhen Yang, Zhenyu Mi, Xuehui Liang, Tang Su, Chuhong Yang, Guangyue Wang, Wenyan Wang, Yongchao Li, Mo Chen, Chengyao Li, Kehuan Linghu, Jiaxiu Han, Yingshan Zhang, Yulong Feng, Yu Song, Teng Ma, Jingning Zhang, Ruixia Wang, Peng Zhao, Weiyang Liu, Guangming Xue, Yirong Jin, and Haifeng Yu. Towards practical quantum computers: transmon qubit with a lifetime approaching 0.5 milliseconds. *npj Quantum Information*, 8(1):3, Jan 2022.

[107] Anbang Wu, Yufei Ding, and Ang Li. Collcomm: Enabling efficient collective quantum communication based on epr buffering. *arXiv preprint arXiv:2208.06724*, 2022.

[108] Anbang Wu, Gushu Li, Hezi Zhang, Gian Giacomo Guerreschi, Yufei Ding, and Yuan Xie. A synthesis framework for stitching surface code with superconducting quantum devices. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22, page 337–350, New York, NY, USA, 2022. Association for Computing Machinery.

[109] Anbang Wu, Hezi Zhang, Gushu Li, Alireza Shabani, Yuan Xie, and Yufei Ding. Autocomm: A framework for enabling efficient communication in distributed quantum programs. In *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 1027–1041, 2022.

[110] Yue Wu and Lin Zhong. Fusion blossom: Fast mwpm decoders for qec. *arXiv preprint arXiv:2305.08307*, 2023.

[111] Xiao Xue, Bishnu Patra, Jeroen P. G. van Dijk, Nodar Samkharadze, Sushil Subramanian, Andrea Corna, Brian Paquelet Wuetz, Charles Jeon, Farhana Sheikh, Esdras Juarez-Hernandez, Brando Perez Esparza, Huzaifa Rampurawala, Brent Carlton, Surej Ravikumar, Carlos Nieva, Sungwon Kim, Hyung-Jin Lee, Amir Sammak, Giordano Scappucci, Menno Veldhorst, Fabio Sebastiano, Masoud Babaie, Stefano Pellerano, Edoardo Charbon, and Lieven M. K. Vandersypen. Cmos-based cryogenic control of silicon quantum circuits. *Nature*, 593(7858):205–210, May 2021.

[112] Fei Yan, Simon Gustavsson, Archana Kamal, Jeffrey Birenbaum, Adam P. Sears, David Hover, Ted J. Gudmundsen, Danna Rosenberg, Gabriel Samach, S. Weber, Jonilyn L. Yoder, Terry P. Orlando, John Clarke, Andrew J. Kerman, and William D. Oliver. The flux qubit revisited to enhance coherence and reproducibility. *Nature Communications*, 7(1):12964, Nov 2016.

[113] Y. P. Zhong, H.-S. Chang, K. J. Satzinger, M.-H. Chou, A. Bienfait, C. R. Conner, É Dumur, J. Grebel, G. A. Peairs, R. G. Povey, D. I. Schuster, and A. N. Cleland. Violating bell's inequality with remotely connected superconducting qubits. *Nature Physics*, 15(8):741–744, Aug 2019.

[114] Youpeng Zhong, Hung-Shen Chang, Audrey Bienfait, Étienne Dumur, Ming-Han Chou, Christopher R. Conner, Joel Grebel, Rhys G. Povey, Haoxiong Yan, David I. Schuster, and Andrew N. Cleland. Deterministic multi-qubit entanglement in a quantum network. *Nature*, 590(7847):571–575, Feb 2021.