IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 September 2026, accepted 20 September 2026. Date of publication 00 xxxx 0000, date of current version 00 xxxx 0000.

Digital Object Identifier 10.1109/ACCESS.2026.3737721

Quantum Error Correction Survey

MUZAMMIL-E-HAQUE MAHMUD[1], (Member, IEEE), RAJ JAIN[1], (Life Fellow, IEEE), ZEBO YANG[2], (Member, IEEE), CHENLIANG TIAN[1], (Student Member, IEEE), RAMANA KOMPELLA[3], (Member, IEEE), REZA NEJABATI[3], (Senior Member, IEEE), ENEET KAUR[3], (Member, IEEE), AIMAN ERBAD[4], (Senior Member, IEEE), AND MOHAMED M. ABDALLAH[5], (Senior Member, IEEE)

[1]Washington University in St. Louis, St. Louis, MO 63130, USA
[2]Florida Atlantic University, Boca Raton, FL 33431, USA
[3]Cisco Systems, San Jose, CA 95134, USA
[4]Qatar University, Doha, Qatar
[5]Hamad Bin Khalifa University, Doha, Qatar

Corresponding author: Muzammil-E-Haque Mahmud (mmahmud@wustl.edu)

This work was supported in part by Cisco University Research under Grant 6138; and in part by Qatar Research, Development, and Innovation (QRDI) Academic Research under Grant ARG02-0415-240191; and by Qatar National Library.

ABSTRACT This survey examines quantum error correction (QEC) as a cross-layer engineering problem for scalable fault-tolerant quantum computing and protected quantum communication. It covers code and noise models, hardware platforms and physical-integration constraints, real-time decoding and control, system-scale resource limits, and QEC for quantum communication. Across these areas, the survey uses a three-stage analytical lens consisting of physical realizability, real-time executability, and system-scale or end-to-end utility, together with common criteria including logical reliability, timing and throughput, quantum and classical resource overhead, connectivity and data movement, physical-integration and power constraints, and communication rate and fidelity. Overall, the survey shows that scalable QEC depends on coordinated advances across codes, devices, control, decoding, architecture, and communication.

INDEX TERMS Fault-tolerant quantum computing, full-stack quantum error correction, quantum communication, quantum error correction, quantum systems engineering, scalable quantum systems.

I. INTRODUCTION

Quantum computing promises transformative capabilities in simulation, optimization, and secure communication. As quantum processors grow in size and capability, the challenge is increasingly to move from demonstrations of physical-qubit performance toward computations that remain reliable over long runtimes and at useful scales. Small control imperfections and interactions with the environment can accumulate during computation and corrupt the quantum state. Suppressing and managing these errors is therefore a central requirement for scalable quantum computing.

Quantum error correction (QEC) is the framework developed to meet this challenge. At a high level, QEC allows quantum information to be protected and processed reliably even when individual hardware components are noisy. Over the last decade, QEC has moved from a predominantly theoretical topic to an engineering discipline shaped by real devices and real workloads. Early demonstrations established that protected quantum information could be maintained; current efforts increasingly focus on sustaining reliable logical computation over long runtimes and at useful scales.

This shift makes fault tolerance a full-stack problem. Practical QEC must be co-designed with the hardware platforms that implement it, the control and measurement systems that extract error information, and the classical processors that interpret that information under tight latency and power limits. These cross-layer constraints increasingly determine whether a proposed code, decoder, or architecture will remain viable as systems grow.

Motivated by this transition, this survey examines the engineering challenges that shape the path toward scalable, fault-tolerant quantum computing. We take an integrated view across theory, physical platforms, real-time decoding,

The associate editor coordinating the review of this manuscript and approving it for publication was Siddhartha Bhattacharyya.

IEEE Proof

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

system scalability, and quantum communication, with emphasis on how constraints in these areas interact. Our goal is to clarify where current bottlenecks lie and which open problems most strongly shape the path to useful fault-tolerant quantum systems.

Existing QEC surveys and perspectives organize the field around complementary analytical emphases, including foundational and tutorial treatment [1], code families and implementation trade-offs [2], noise-adapted protection [3], learning-based and adaptive QEC [4], and quantum communication [5]. Other studies focus more specifically on real-time decoding and decoder implementation [6], [7], the real-time QEC system stack [8], or algorithm–QEC–hardware co-design for low-overhead fault-tolerant computation [9]. Table 1 compares these selected works using the stated operational criteria, with each classification interpreted only with respect to its named criterion.

The present survey organizes the engineering problem around dependencies among code/noise assumptions, physical implementation, real-time decoding and control, system-scale integration, and protected communication. The four engineering questions below make these dependencies explicit by relating code and device assumptions to embedding and syndrome extraction, decoder workload to service and timing constraints, quantum and classical resources to realizable logical scale, and protected-communication mechanisms to end-to-end rate and fidelity.

More specifically, this cross-layer view is organized around four engineering questions: (i) how do code parameters, physical-noise assumptions, and device/material constraints jointly determine code embedding, syndrome extraction, and viable platform implementations; (ii) when does a decoder remain practically executable once sustained syndrome throughput, backlog, response and tail latency, data movement, accelerator placement, and cryogenic power/energy are included; (iii) how do quantum and ancilla overhead, syndrome-memory pressure, routing/interconnect, packaging, and classical memory and energy jointly limit achievable code distance and logical-system scale for a target logical reliability; and (iv) for protected quantum communication, how do encoding, fusion, decoding, buffering, and encoded swapping constraints propagate across the physical, link, network, and transport layers into end-to-end rate and fidelity.

These questions define a survey-level taxonomy of coupled constraint domains: code/noise, device/material and physical implementation, real-time decoding/control, system-scale integration, and protected communication. These domains are coupled because a design choice or constraint in one domain changes feasible choices, resource requirements, or performance in another.

Across these domains, we use a three-stage analytical lens. First, *physical realizability* asks whether code and noise assumptions remain implementable under device/material properties, connectivity, embedding, syndrome-extraction, packaging, and physical-integration constraints. Second, *real-time executability* asks whether syndrome processing, decoding, data movement, and feedback can satisfy sustained throughput, backlog, response and tail-latency, placement, memory, and power constraints. Third, *system-scale and end-to-end utility* asks what logical scale, reliability, resource efficiency, or protected-communication performance remains achievable after those implementation constraints are included.

Within this lens, the principal recurring evaluation criteria are logical reliability; timing and throughput feasibility; quantum and ancilla resource overhead; connectivity, routing, and data movement; classical memory/compute and thermal/power/physical-integration resources; and system-scale or end-to-end utility, including communication rate and fidelity where applicable. These criteria provide a common engineering basis for the principal code, hardware, decoder, scalability, and communication comparisons in the survey.

The contributions of this paper, to be read with Table 1, are as follows:

- Provide a cross-layer analytical synthesis organized around explicit engineering questions, a coupled-constraint taxonomy, and recurring evaluation dimensions that relate code/noise assumptions to physical implementation, real-time decoding/control, system-scale feasibility, and protected-communication performance.
- Provide a tutorial overview of the fundamental challenges and trade-offs that govern scalable fault tolerance.
- Present a structured survey of decoder families and their real-time architectural implications.
- Analyze scalability across qubit overhead, syndrome cadence, routing/interconnect, and emerging low-overhead directions.
- Provide a unified discussion of QEC for both quantum computation and quantum communication.

The remainder of this paper is organized as follows. Section II outlines foundational concepts and challenges. Section III focuses on hardware and physical implementation. Section IV surveys decoders and real-time constraints. Section V analyzes scalability and resource overhead. Section VI covers noise and decoherence management. Section VII addresses QEC for quantum communication. Section VIII concludes with open problems and a roadmap. To aid navigation, the guiding figures use a common analytical grammar: grey marks the subsection anchor, pink the purpose or outcome, green the pitfall or default state, light blue the decision questions, and gold the key takeaway. These figures condense the subsection's engineering decision logic rather than repeat the surrounding prose. Figure 1 instead uses numbered cycle stages as a workflow and navigation map; its stage colors are visual separators and do not encode quantitative values.

IEEE Proof

TABLE 1. Criterion-level content comparison of selected qec reviews, perspectives, and system-oriented surveys.

Panel I: Code Theory, Noise, Decoding, and Real-Time Control

Study	A. Core QEC / Stabilizer Foundations	B. Standard Quantitative QEC Metrics	C. Approximate / Noise-Adapted QEC	D. Formal Locality–Rate–Distance Bound	E. qLDPC / High-Rate Code Coverage	F. Code-vs.-Hardware / qLDPC Embedding Cost	G. Bosonic / Bias-Tailored QEC	H. Correlated / Structured / Non-Pauli Noise	I. Operator-Level / Calibration Noise Hierarchy	J. Adaptive / ML / Feedback QEC	K. Modern Decoder-Family Breadth	L. Queue / Backlog / Deadline Model	M. Quantitative Decoder Timing / Tails	N. Decoder Hardware Execution	O. Benchmark Comparability / Provenance	P. Syndrome Cadence / Measurement Co-Design	Q. Cryogenic Control / Power / Placement
Wang and Tang [4]	Y	Y	N	N	*P*	*P*	N	Y	N	Y	Y	*P*	*P*	*P*	*P*	*P*	N
Chatterjee *et al.* [1]	Y	Y	N	N	*P*	*P*	*P*	*P*	N	N	*P*	N	*P*	N	N	N	N
Jayashankar and Mandayam [3]	*P*	Y	Y	N	N	N	Y	*P*	Y	N	N	N	N	N	N	N	N
Thakur *et al.* [5]	Y	Y	*P*	N	Y	N	Y	*P*	*P*	*P*	N	N	N	N	N	*P*	N
Battistel *et al.* [6]	Y	Y	N	N	*P*	*P*	N	*P*	*P*	*P*	Y	*P*	Y	Y	*P*	Y	Y
Mahmud and Abdelhadi [2]	Y	Y	*P*	N	Y	N	Y	*P*	N	*P*	Y	N	N	N	N	*P*	N
Chen *et al.* [8]	Y	Y	N	*P*	Y	*P*	N	*P*	*P*	*P*	Y	Y	Y	Y	Y	Y	*P*
This paper	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Panel II: Device/Platform Coupling, Fault Tolerance, System Integration, Communication, and Review Rigor

Study	R. Device-Level Nonidealities–QEC Coupling	S. Materials / Fabrication–QEC Coupling	T. Packaging / 3D / Physical Interconnect	U. Multi-Platform Code-to-Hardware Mapping	V. Fault-Tolerant Logical Operations	W. Eastin–Knill / Transversality–Universality	X. System-Scale Resource Integration	Y. Explicit Cross-Layer QEC Co-Design	Z. Communication QEC	AA. Encoded Repeaters / Network-Layer QEC	AB. General Channel Capacity / Information Theory	AC. Pure-Loss / Repeaterless Capacity Bound	AD. Encoded Communication System Metrics	AE. Reproducible Review Methodology	AF. Explicit Open Problems / Roadmap	AG. Evidence / Implementation Distinction
Wang and Tang [4]	Y	N	N	*P*	N	N	*P*	*P*	N	N	N	N	N	N	Y	*P*
Chatterjee *et al.* [1]	*P*	N	N	*P*	*P*	*P*	*P*	*P*	N	N	N	N	N	N	Y	*P*
Jayashankar and Mandayam [3]	N	N	N	*P*	Y	Y	*P*	N	Y	N	N	N	*P*	N	Y	*P*
Thakur *et al.* [5]	*P*	*P*	N	*P*	*P*	*P*	*P*	*P*	Y	*P*	Y	N	*P*	N	Y	*P*
Battistel *et al.* [6]	Y	N	*P*	*P*	Y	*P*	*P*	*P*	N	N	N	N	N	N	Y	*P*
Mahmud and Abdelhadi [2]	*P*	N	N	*P*	Y	*P*	*P*	*P*	Y	Y	*P*	N	N	Y	Y	Y
Chen *et al.* [8]	Y	N	*P*	Y	Y	*P*	Y	Y	N	N	N	N	N	N	Y	Y
This paper	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	*P*	Y	Y

Classification rule: Y = the complete operational criterion is substantively satisfied; *P* = relevant content is substantively present, but the complete mathematical, quantitative, implementation-level, systems, or cross-layer criterion is not satisfied; N = the specific operational criterion was evaluated and the required substantive treatment was not established. A N applies only to its named operational criterion. For criterion Y (Explicit Cross-Layer QEC Co-Design), complete satisfaction requires an explicit dependency across QEC engineering layers: a design choice or constraint at one layer must be connected to a feasibility condition, resource requirement, or performance consequence at another layer; discussion of multiple layers without such a dependency constitutes partial rather than complete coverage. The entries represent criterion-level content coverage under these operational definitions.

A. REVIEW SCOPE AND METHODOLOGY

This article follows a structured narrative and tutorial engineering-survey methodology spanning QEC and fault-tolerance foundations, noise modeling, hardware and physical implementation, real-time decoding and control, system-scale integration, and protected quantum communication. Candidate literature was identified through multiple scholarly search and publisher platforms, including Google Scholar, IEEE Xplore, and the ACM Digital Library. Screening proceeded from title relevance to abstract relevance and then to full-text technical relevance. Works were retained when full-text assessment established a direct connection to a technical topic, claim, comparison, or engineering question developed in the survey; works whose connection to QEC remained incidental were not used as core support for the corresponding technical analysis.

Foundational and original sources were prioritized for established theorems and canonical constructions, while recent primary studies were prioritized for current hardware, decoder, architecture, and experimental claims where available. Reviews and perspectives were used for synthesis and literature orientation, and recent preprints were retained selectively for rapidly developing topics. Where evidence type affects interpretation, the survey distinguishes theoretical or analytical results, numerical or simulation studies, architecture or synthesis studies, hardware prototypes, and experimental demonstrations.

For quantitative comparisons, source-defined quantities are retained with their workload, noise model, platform, code distance, and implementation context. Author-derived quantities are reported only when compatible source-reported inputs and definitions permit an unambiguous calculation; otherwise values are identified as not reported (NR), not

IEEE Proof

derivable (ND), or not directly comparable (NC), as applicable. Table 10 applies this provenance convention explicitly to decoder benchmarks. For Table 1, each criterion was evaluated against a written operational definition and each classification was interpreted only with respect to that criterion; differences in terminology, section organization, document length, citation count, or number of examples were not used as a basis for classification.

II. FUNDAMENTAL CHALLENGES IN QEC

Quantum error correction (QEC) aims to preserve quantum information in the face of decoherence and imperfect operations while respecting hard hardware limits. We frame the fundamental challenges that govern scalability: locality–rate–distance trade-offs, correlated/non-Pauli noise, real-time decoding complexity, transversality limits, bias-aware and bosonic encodings, cadence constraints, and hardware mapping. Each subsection states the issue concisely, defines the minimum terms, and ends with a brief takeaway. Where several approaches address the same problem, the takeaway compares what each approach improves, what it costs, and which hardware or system constraint determines its practical use.

Core Terms

1) Logical vs. physical qubits: A logical qubit is encoded across many physical qubits so that typical faults can be detected and corrected without collapsing the logical state.

2) Stabilizers and syndromes: Stabilizers are commuting checks whose ± 1 outcomes, called syndromes, reveal the error without revealing the logical state. For CSS codes, H_X and H_Z denote the binary check matrices whose rows specify the X- and Z-type stabilizer generators, respectively; their commutation condition is $H_X H_Z^T = 0$ over $\mathbb{F}_2$.

3) Code parameters: A code is denoted $[[n, k, d]]$: n physical qubits encode k logical qubits with distance d, the minimum fault weight that flips a logical state.

4) Error and timing notation: Throughout Sections II–VII, p_{phys} denotes a physical error probability or rate on the stated operation, location, or noise-model basis; P_L denotes a logical error probability or rate on the stated QEC-cycle, logical-operation, or memory-experiment basis; and τ denotes an execution or processing duration. Subscripts or superscripts identify the timing basis, e.g., τ_{dec}, τ_{round}, and τ_t^{resp}. Source-specific symbols are retained when reproducing reported quantities and are mapped explicitly to these survey-wide conventions.

5) Noise models: Pauli stochastic $X/Y/Z$, amplitude-damping T_1, dephasing T_2, erasure loss with known location, plus correlated and coherent systematic errors; see Section VI.

6) Transversality: Gates that act qubit-wise across a block so single faults do not spread; useful for Cliffords, but not sufficient for universality because of Eastin–Knill.

7) Syndrome extraction and cadence: The per-round measurement schedule and timing budget that must fit device coherence and decoder deadlines; see Sections II-H and IV.

8) Decoder: The classical algorithm/hardware that maps syndromes to corrections under tight latency and power envelopes; see Section IV.

A. WHAT IS QUANTUM ERROR CORRECTION?

This subsection introduces the foundational principles of QEC. It explains how QEC encodes logical information across many physical qubits so that errors can be detected and corrected without collapsing the underlying quantum state.

Qubits, the basic units of quantum information, can exist in superposition, simultaneously representing 0 and 1, and can be entangled so that their states are interdependent. These properties enable quantum algorithms with potential exponential speedups but also make qubits extremely sensitive: brief environmental interactions or small control imperfections can collapse superposition or scramble entanglement, destroying stored information. Because quantum states cannot be copied, as stated by the no-cloning theorem, redundancy must be achieved through entanglement-based encodings rather than simple duplication.

QEC addresses this by encoding a single logical qubit into many physical qubits using a carefully chosen quantum code. For example, surface codes and the Steane code distribute logical information so that most small local errors leave distinct signatures, called syndromes, on auxiliary check qubits called ancillas. Measuring ancillas reveals only the error pattern, not the logical state, thereby preserving coherence. A classical decoder processes the syndrome and applies, or tracks, corrective operations that restore the logical qubit.

Unlike classical error correction, which mainly handles discrete bit flips, quantum errors are richer and effectively continuous. A single fault may act as a bit flip, Pauli X, a phase flip, Pauli Z, or their combination, Pauli Y, and small coherent drifts can accumulate between correction cycles. Directly measuring data qubits would destroy their states, so QEC uses carefully designed circuits and entangled ancillas to perform indirect, syndrome-only measurements. As a result, QEC codes must protect both amplitude and phase information, and they require real-time classical processing throughout repeated correction cycles. Figure 1 maps this recurring workflow and identifies where its principal stages are developed in the survey.

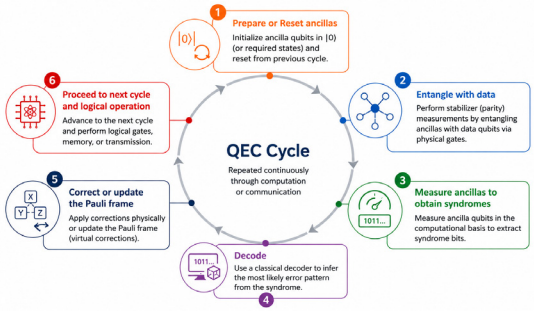


FIGURE 1. QEC Cycle. Ancilla preparation, stabilizer interaction, and syndrome measurement are developed in Sections II-H and III; decoding and Pauli-frame response in Sections II-D and IV; fault-tolerant logical operations in Sections II-E and III-C; and communication-specific QEC in Section VII. The numbered sequence defines the workflow; colors distinguish stages only and carry no quantitative meaning.

In practice, QEC is a full-stack engineering problem. It spans quantum device physics, meaning how different qubit platforms fail; control electronics, meaning how

IEEE Proof

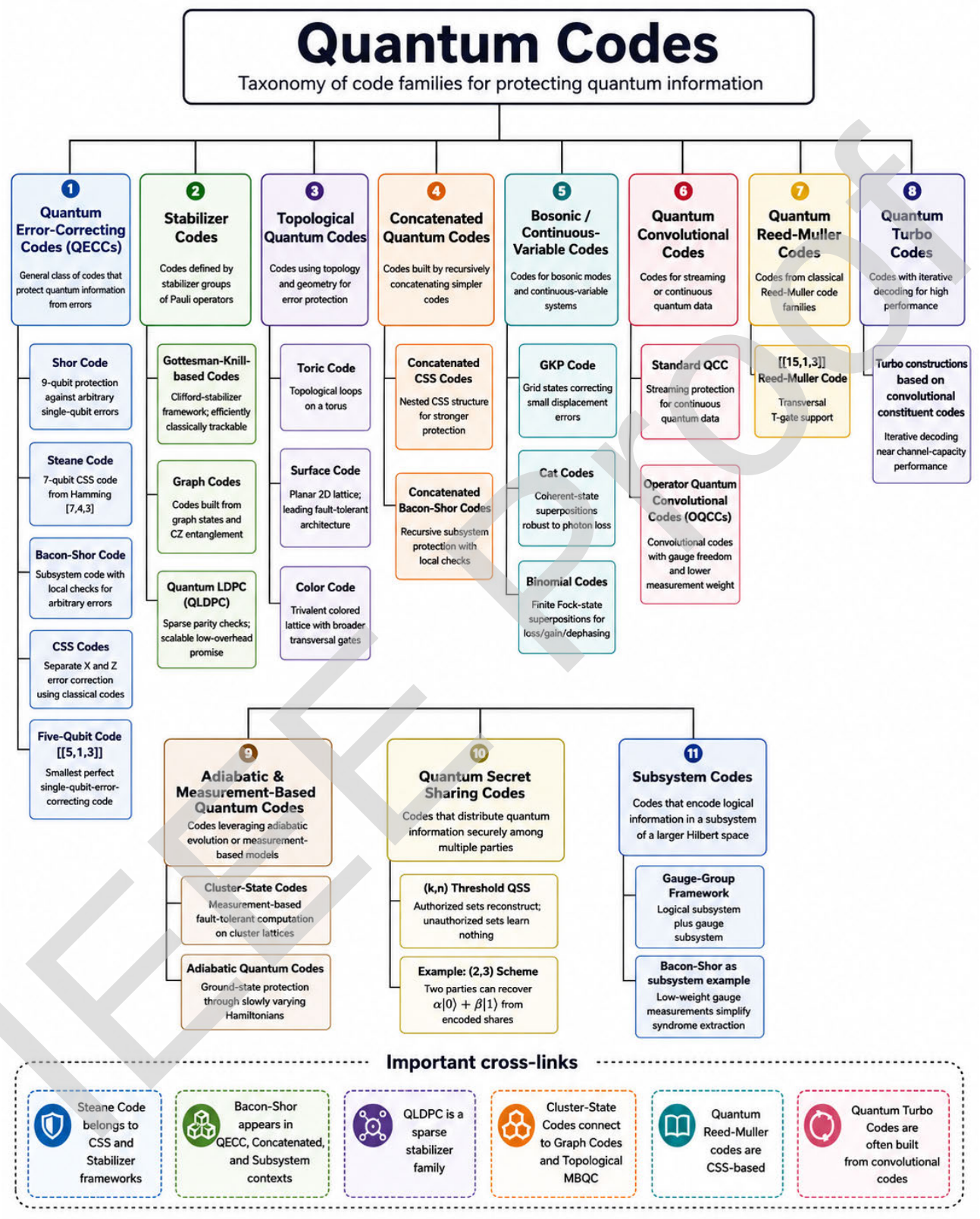


FIGURE 2. Non-exclusive landscape of representative quantum-code families and important cross-links among overlapping constructions [2].

quickly and accurately corrections can be applied; and computer architecture, meaning how data and syndromes move between cryogenic devices and classical decoders. Every code must be co-designed with its hardware platform,

such as superconducting, ion-trap, photonic, or spin-based qubits, to respect constraints on connectivity, timing, and power.

Against this systems context, it is useful to first situate the major quantum-code families encountered throughout the survey. Fig. 2 provides a high-level, non-exclusive landscape of representative constructions and their important cross-links. The categories intentionally overlap: for example, qLDPC codes belong to the stabilizer framework, many topological codes are stabilizer constructions, and subsystem or concatenated structures can intersect with other families. The figure serves as a non-exclusive orientation map of these overlapping code families and their cross-links.

Although these code families differ in mathematical structure, encoding strategy, and implementation assumptions, they are evaluated using a common set of performance and engineering quantities. Table 2 summarizes the principal metrics used throughout this survey to relate code properties to hardware and real-time system constraints.

TABLE 2. Fundamental QEC metrics.

Metric	Meaning
Physical error rate p_{phys}	Physical fault probability or rate on the stated operation, location, or noise-model basis.
Logical error rate P_L	Logical failure probability or rate on the stated QEC-cycle, logical-operation, or memory-experiment basis.
Distance d	Minimum weight of a Pauli operator mapping one codeword to another.
Rate $R = k/n$	Ratio of logical qubits to physical data qubits for an $[[n, k, d]]$ code.
Threshold p_{th}	Maximum p_{phys} below which P_L decreases with increasing code distance under the stated noise model and decoder.
Stabilizer weight	Number of qubits touched by a parity check.
Execution / processing time τ	Duration of the stated operation or processing workload; subscripts or superscripts identify the timing basis.
Decoder latency τ_{dec}	Time between syndrome readout and the corresponding Pauli-frame update.
Locality, or check range/diameter	Maximum geometric span of any stabilizer on the hardware connectivity graph.

Together, the code-family landscape and these common metrics establish the vocabulary used in the remainder of the survey: the former organizes the main families of protection strategies, while the latter provides the quantities needed to compare their reliability, overhead, locality, and real-time implementation requirements.

To make these quantities operational across code families, Table 3 provides a source-conditioned comparison of representative constructions. The table deliberately retains the noise model, workload, decoder, and evidence basis attached to each reported quantity. Thresholds obtained under code-capacity, phenomenological, circuit-level, and experimental conditions are therefore not treated as directly interchangeable. Likewise, n/k denotes code-level data-carrier overhead, whereas N_{hw}/k denotes a broader implementation count only when the cited source reports a sufficiently clear hardware basis.

The comparison reveals three distinct engineering regimes. Surface and color codes retain direct two-dimensional locality but pay $O(d^2)$ code-level overhead per logical qubit. XZZX retains essentially the same geometry while exploiting biased Pauli noise, so its unusually high code-capacity and phenomenological thresholds should not be compared numerically with circuit-level surface-, color-, or BB-code thresholds. Bivariate-bicycle codes reduce code and ancilla overhead substantially in circuit-level simulation but replace strict geometric locality with degree-6, long-range connectivity. Quantum Tanner codes offer constant asymptotic rate and relative distance together with single-shot decoding guarantees, but their expanding connectivity remains a major implementation burden. Finally, concatenated cat–repetition QEC demonstrates a different resource trade: part of the error suppression is performed inside each bosonic mode, making a simple physical-qubit-per-logical-qubit comparison with discrete-qubit codes inappropriate.

B. LOCALITY, RATE, AND DISTANCE TRADE-OFF

This subsection explains how QEC codes balance three interdependent properties: locality, rate, and distance, as summarized in Table 2. Here the rate is $R = k/n$. For the mathematical bound below, locality is used in the stronger Euclidean sense that each commuting check is supported within a region of bounded geometric diameter on a two-dimensional lattice. This is distinct from the qLDPC condition of bounded check weight and bounded Tanner-graph degree: a sparse check graph need not be geometrically local. This distinction anchors the design trade-off shown in Fig. 3.

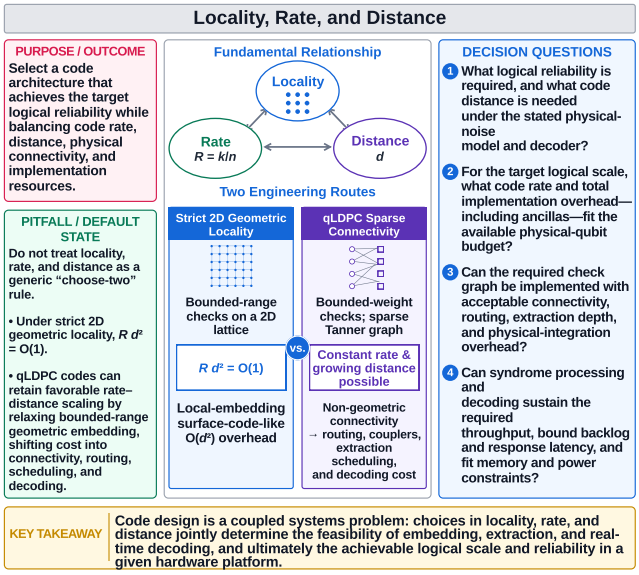


FIGURE 3. Summary of locality, rate, and distance in the form of design questions, pitfalls, and purpose.

For codes whose codespace is specified by geometrically local, pairwise commuting constraints of bounded interaction range on a two-dimensional lattice of finite-dimensional

IEEE Proof

TABLE 3. Source-conditioned quantitative comparison of representative QEC Code families.

Code family	Representative parameters / scaling	Code / implementation overhead	Check weight / locality	Threshold + noise model	Representative logical result + conditions	Connectivity / extraction; decoder implication	Evidence / Ref.
Surface code	Rotated planar family: $[[d^2, 1, d]]$ at the data-code level.	$n/k = d^2$ data qubits. A standard data-plus-measure layout uses $N_{\mathrm{hw}}/k = 2d^2 - 1$. The Willow $d = 7$ experiment used 49 data, 48 measure, and 4 additional leakage-removal qubits: $N_{\mathrm{hw}} = 101$.	Bulk stabilizers have weight 4; geometrically local on a 2D square lattice.	Threshold is circuit- and decoder-dependent. The standard circuit-based comparison used in [10] places the relevant surface-code scale near 0.7%. The Willow result demonstrates below-threshold operation experimentally rather than extracting a universal p_{th}.	Experiment: for $d = 7$, the logical error rate is $P_L = \epsilon_7 = (1.43 \pm 0.03) \times 10^{-3}$ per QEC cycle with a neural decoder, obtained from memory experiments extending to 250 cycles. No single depolarizing-p_{phys} assumption describes the hardware run.	Nearest-neighbor square-grid syndrome extraction. Correlated matching and neural decoding were used offline; a $d = 5$ real-time decoder sustained 1.1 µs QEC cycles for up to 10^6 cycles with 63 µs average response latency.	Experiment [11]
Triangular color code	Representative $6.6.6$ triangular family: $[[n, 1, d]]$ with $n = \Theta(d^2)$.	$n/k = \Theta(d^2)$. N_{hw}/k: NR on a common syndrome-extraction basis in the cited study.	2D local face checks; the triangular $6.6.6$ family contains weight-4 and weight-6 X- and Z-type checks.	0.47% under the standard circuit-level noise model using error-rate-weighted decoding.	NR for an experimental logical-memory error per cycle in the cited work; the study instead provides circuit-level numerical threshold, lattice-surgery decoding, and state-injection analysis.	2D local architecture; weighted decoding graphs and circuit-level lattice-surgery decoding. Higher-weight checks and color-code syndrome processing increase extraction/decoder complexity relative to weight-4 surface-code plaquettes.	Theory / Simulation [12]
XZZX surface code	Locally equivalent to the conventional surface code: $k = O(1), n = O(d^2)$.	Same asymptotic code-level scaling as the surface code. N_{hw}/k: NR as a hardware implementation quantity in the cited work.	Weight-4 local XZZX plaquette checks on a 2D square lattice.	*Code capacity*: $18.7(1)\%$ for depolarizing Pauli noise and approximately 50% for pure X, Y, or Z noise. *Phenomenological*: approaches approximately 10% at large Z-bias with noisy syndrome measurements. These values are not circuit-level thresholds.	NR for a hardware logical-error measurement in the cited study; results are numerical threshold and sub-threshold resource-scaling analyses.	Retains local surface-code connectivity. Bias-aware tensor-network or matching decoding exploits dominant Pauli structure; at infinite Z-bias the decoding problem decomposes into repetition-code-like lines.	Theory / Simulation [13]
Bivariate-bicycle qLDPC	Representative $[[144, 12, 12]]$ BB code.	Code level: $n/k = 144/12 = 12$. With one check ancilla per data qubit, the cited architecture uses 288 total qubits, giving $N_{\mathrm{hw}}/k = 24$.	Weight-6 checks; degree-6 Tanner connectivity. Sparse but not geometrically local; Tanner graph has thickness two.	Approximately 0.7% under the standard circuit-based noise model with the cited fault-tolerant protocol and BP+OSD decoding.	Simulation: for the $[[144, 12, 12]]$ code at $p_{phys} = 10^{-3}$, the reported logical error rate is $P_L = 2 \times 10^{-7}$; simulations use d syndrome cycles for a distance-d code.	Requires six check couplings per qubit, including long-range edges; the Tanner graph decomposes into two planar subgraphs. Syndrome extraction uses seven CNOT layers (depth eight including initialization/measurement), followed by circuit-aware BP+OSD decoding.	Architecture / Simulation [10]
Quantum Tanner qLDPC	Asymptotically good family: $[[n, \Theta(n), \Theta(n)]]$, i.e., constant rate and constant relative distance.	$n/k = \Theta(1)$ asymptotically. N_{hw}/k: NR for a complete fault-tolerant hardware implementation.	Constant-weight parity checks, but the expanding Tanner graph is generally not geometrically local. Explicit layout studies include small instances with check weights in the 6–12 range.	NR for a numerical stochastic circuit-level threshold in the cited single-shot analysis. That result instead proves correction under sufficiently low-weight adversarial data and syndrome errors.	Theory: one noisy syndrome-measurement round suffices for single-shot QEC under the stated adversarial assumptions. A constant number of parallel-decoder iterations per repeated QEC round keeps residual error bounded; $O(\log n)$ iterations yield $\alpha = 0$ in the single-shot residual-error bound.	Long-range expanding connectivity is the principal implementation cost. Sequential and parallel mismatch-decomposition decoders are provably efficient; multilayer superconducting layout studies place Tanner instances near the high-complexity end of the studied qLDPC families at comparable logical efficiency.	Theory / Layout study [14]–[16]
Bosonic cat + repetition	Outer repetition code of distance d over cat-qubit data modes; experiment realizes $d = 5$ with one logical qubit.	Outer-code $n/k = d$ cat data modes. The $d = 5$ device contains 5 storage modes, 4 syndrome-ancilla transmons, 5 stabilization buffer modes, and tunable couplers; a single N_{hw}/k qubit count is NC with two-level-qubit architectures.	Weight-2 outer repetition checks $\bar{X}_i\bar{X}_{i+1}$; nearest-neighbor linear extraction.	NR for a single numerical full-logical circuit threshold. The phase-flip-correcting repetition sector is experimentally demonstrated below threshold, while cat stabilization suppresses rather than outer-code-corrects bit flips.	Experiment: $d = 5$ achieves $P_L = \epsilon_L = 1.65(3)\%$ per QEC cycle at $\lvert\alpha\rvert^2 = 1.5$. Distance-3 sections average $P_L = 1.75(2)\%$ per cycle. These are measured device values, not results at an assumed depolarizing p_{phys}.	Linear storage-mode array with transmon syndrome ancillas and noise-biased CX gates. Repetition decoding corrects phase flips while preservation of the native cat-qubit bias is essential to total logical performance.	Experiment [17]

Comparison convention: n is the number of data carriers entering the code parameters $[[n, k, d]]$; N_{hw} includes syndrome ancillas and other explicitly counted QEC carriers only when the cited source provides a clear implementation count. NR = not reported in the cited source; ND = not derivable from source-reported, workload-matched quantities; NC = not directly comparable under a common resource or noise-model basis. Code-capacity, phenomenological, circuit-level, adversarial, and experimental results are retained under their source-specific assumptions and must not be numerically rank-ordered across models. Simple ratios such as $144/12$ and $288/12$ are derived only from source-reported counts.

particles, Bravyi, Poulin, and Terhal proved the finite-size bound [18]

$$k \le \frac{cn}{d^2} \qquad \Longleftrightarrow \qquad Rd^2 \le c, \tag{1}$$

where c is independent of n, k, and d, and depends only on the locality of the constraints and the local Hilbert-space dimension. Equivalently, the asymptotic scaling is $kd^2 = O(n)$. Thus, under strict two-dimensional geometric locality, a family with increasing distance cannot retain a nonzero asymptotic rate: $d \to \infty$ implies $R = O(d^{-2})$. Conventional planar surface-code patches exhibit $n = \Theta(kd^2)$ scaling and therefore match this trade-off up to constant factors [18].

The geometric-locality assumption is essential. Quantum LDPC locality instead refers to sparse parity checks: bounded check weight and bounded Tanner-graph degree do not require bounded Euclidean interaction range. Indeed, explicit asymptotically good qLDPC families with $k = \Theta(n)$ and $d = \Theta(n)$ exist [19]. These asymptotic qLDPC results do not contradict (1) because their sparse Tanner graphs are not required to admit a bounded-range two-dimensional Euclidean embedding. In a physical processor, the resulting non-geometric connectivity can instead reappear as routing, coupler, stabilizer-scheduling, and decoding overhead.

Recent work explores algebraic product constructions [20], stability and threshold behavior for LDPC codes without Euclidean spatial locality [21], lifting-based improvements in relative code parameters [22], and hardware-oriented bicycle and biplanar layouts [10], [23]. These directions seek to preserve favorable rate and distance while making the required nonlocal connectivity progressively more implementable.

In summary, the rigorous obstruction is not that every quantum code can optimize only two abstract metrics; rather, strict two-dimensional geometric locality imposes $Rd^2 = O(1)$, whereas qLDPC sparsity can support constant rate and

IEEE Proof

growing distance only by relaxing that geometric-embedding assumption.

Can a practically embeddable, efficiently decodable architecture retain the rate–distance advantages of modern qLDPC codes while keeping physical interaction and routing overhead sufficiently local?

C. CORRELATED-NOISE MODELS BEYOND THE THRESHOLD THEOREM

This subsection distinguishes the phenomenological noise models commonly used in QEC analysis from operator-level descriptions of the underlying physical dynamics. In a phenomenological Pauli model, let E_ℓ denote the effective fault assigned to circuit location ℓ. For a homogeneous set of equivalent locations, the independent and identically distributed (IID) approximation assumes

$$\Pr(E_1, \ldots, E_N) = \prod_{\ell=1}^{N} p(E_\ell). \tag{2}$$

Independence is the factorization in (2), while identical distribution means that equivalent locations use the same probability law p; a location-dependent but independent circuit model instead uses $\prod_\ell p_\ell(E_\ell)$. Thus, "Pauli stochastic" and "IID" are not synonymous: correlated Pauli faults are possible when the joint distribution does not factorize. Such phenomenological models are useful for threshold studies and efficient simulation, but a simple factorized model coarse-grains coherent phases, inter-location correlations, and time-dependent device dynamics [24].

At the operator level, a noisy physical operation is represented by a completely positive trace-preserving quantum channel

$$\mathcal{E}(\rho) = \sum_a K_a \rho K_a^\dagger, \qquad \sum_a K_a^\dagger K_a = I, \tag{3}$$

where ρ is the density operator and K_a are Kraus operators [25]. Coherent noise is included as the unitary special case $\mathcal{E}(\rho) = U\rho U^\dagger$, while correlated multi-qubit channels need not factor into independent single-qubit channels. Circuit-level models lie between these descriptions by assigning effective channels to gates, measurements, resets, and idle intervals and, when appropriate, reducing them to Pauli or syndrome-fault probabilities for QEC simulation.

Early threshold analyses include quasi-independent stochastic error models [24], but fault-tolerance theorems also exist for restricted classes of non-Markovian, spatially correlated Hamiltonian noise [26]. Consequently, correlation does not invalidate threshold theory in general; it invalidates a particular IID model when the physical process lies outside that model's assumptions. Spatial correlation, temporal correlation, coherent error, non-stationarity, and non-Markovianity are therefore treated as distinct properties in this survey. Section VI applies this hierarchy to slow flux fluctuations, coherent residual ZZ coupling, and ionizing-radiation-induced phonon/quasiparticle bursts [27]; Fig. 4 distills the resulting hardware-to-QEC design loop.

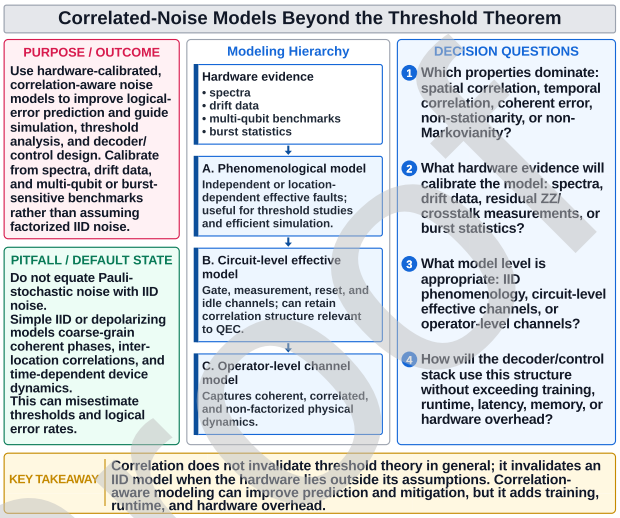


FIGURE 4. Correlation-aware design loop for modeling and decoding.

Section VI develops a hardware-driven taxonomy of coherent, biased, leakage, and correlated error processes, calibrates noise models from spectra, drift data, and multi-qubit benchmarks, and then shows how correlation-aware decoders and control stacks exploit this structure. These models improve logical error predictions and guide mitigation, but they also introduce training, runtime, and hardware overhead for adaptive, correlation-aware decoding [28]. Bridging the gap between such detailed, device-specific noise models and a universal theory of thresholds under correlated noise is an important open direction for QEC theory.

In summary, correlations invalidate the simplest IID threshold assumptions, and while hardware-calibrated, correlation-aware models and decoders can lower logical error rates, they do so at nontrivial overhead.

Uniting noise estimation and adaptive, correlation-aware decoding under a scalable, low-overhead framework remains an open problem in QEC.

D. DECODER COMPLEXITY VS. REAL-TIME DEADLINES

This subsection isolates the fundamental real-time constraint on decoding. At the heart of quantum error correction lies the decoding problem: given a stream of measured syndromes, infer a consistent Pauli frame. Syndrome extraction can continue while this inference is streamed or pipelined, so a Pauli-frame result need not universally be available before the next physical extraction round. Instead, the classical system must provide sufficient sustained service to prevent unbounded decoder backlog and must return results before the logical operations that depend on them.

For generic stabilizer codes, this inference is NP-hard in the worst case, so one cannot rely on globally optimal

IEEE Proof

decoding at scale. The decoder latency τ_{dec}, defined here as the time between syndrome readout and the corresponding Pauli-frame update, therefore remains an important resource, but the one-cycle condition $\tau_{\text{dec}} \leq T_{\text{cycle}}$ is a useful special case rather than a universal streaming-decoder requirement. Section IV-A separates stochastic decoding workload, sustained service capacity, backlog stability, and logical-dependency response deadlines; these must also respect the power and placement constraints of cryogenic hardware in Section III. See Fig. 5.

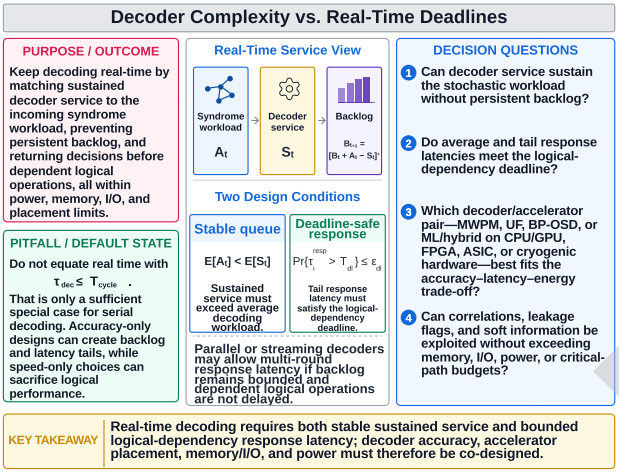


FIGURE 5. Real-time decoding checklist: sustained service versus stochastic workload, backlog and tail-latency limits, buffering and parallelism, and cryo–warm partitioning under the available power and energy budget.

In short, the theoretical challenge is to reconcile intractable worst-case decoding complexity with sustained real-time throughput, bounded response latency, and energy constraints in a deeply layered quantum–classical stack.

A cryogenic-compatible, silicon-level LDPC decoder with sub-µs latency, low power, and provably favorable scaling remains one of the most important open problems in QEC.

E. UNIVERSALITY VS. TRANSVERSALITY: FRAMING THE EASTIN–KNILL PROBLEM

This subsection frames the tension between transversality and universality in fault-tolerant quantum computing. A transversal encoded gate acts independently across the physical subsystems of a code block, so that it does not couple physical subsystems within that block and thereby limits the propagation of physical faults, making transversality attractive for fault tolerance. The Eastin–Knill theorem, however, shows that a quantum code capable of detecting arbitrary errors on individual physical subsystems cannot support a fully universal logical gate set using transversal gates alone [29]. Depending on the code family, useful logical operations, including some Clifford gates, can be transversal, but additional fault-tolerant mechanisms are required to obtain a universal logical gate set. A prominent route is magic-state distillation, which produces higher-fidelity non-Clifford resource states from noisy inputs using Clifford operations and Pauli measurements at additional resource cost [30]; see Fig. 6.

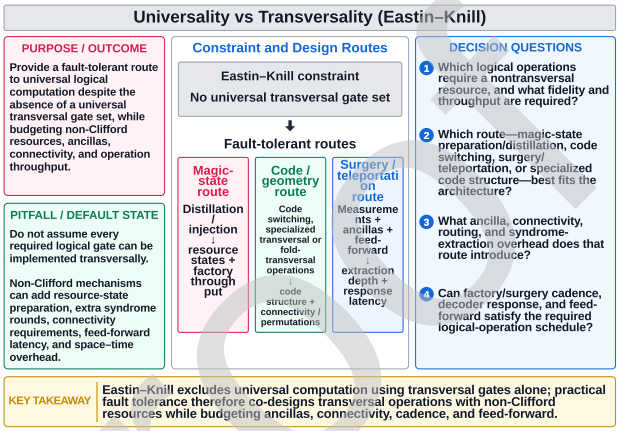


FIGURE 6. Design map for Eastin–Knill: choose the route to non-Cliffords and budget its throughput while avoiding the "everything is transversal" trap.

Code designers also explore algebraic and geometric structures that broaden the space of available quantum codes and logical operations. Torsion in homological CSS constructions can introduce additional logical degrees of freedom [31], while constacyclic constructions over finite chain rings generate new quantum stabilizer-code families through Hermitian and symplectic constructions [32]. These code-construction results broaden the available design space but do not by themselves establish a route around the Eastin–Knill restriction on universal transversal gate sets. More directly, spin codes can be mapped to permutation-invariant multiqubit codes while preserving their transversal gate groups [33], and code symmetries can be exploited to realize additional fold-transversal logical operations in suitable CSS/LDPC codes [34].

Table 4 shows that these approaches address different parts of the Eastin–Knill constraint and incur different implementation costs. Magic-state distillation is broadly applicable but introduces resource-state preparation and repeated distillation overhead. Permutation-invariant codes can enlarge the available transversal gate set at the cost of specialized code structure, while fold-transversal constructions trade code symmetry and qubit-permutation requirements for additional logical Clifford operations. Color-code protocols combine favorable logical-gate structure with weighted decoding, lattice surgery, and state injection, but still require non-Clifford resource preparation for universality [12], [30], [33], [34]. The cited magic-state, permutation-invariant, and fold-transversal results are theoretical constructions, while the cited color-code work includes circuit-level numerical evaluation; none of these four cited works is an experimental demonstration.

In summary, these strategies address different parts of the Eastin–Knill constraint rather than removing it. Some enlarge

IEEE PROOF

TABLE 4. Engineering comparison of representative strategies for navigating the Eastin–Knill constraint.

Strategy	Principal advantage	Implementation / geometry cost	Applicable code / platform	Evidence / Ref.
Magic-state distillation	Produces high-fidelity non-Clifford resource states from noisy inputs using Clifford operations and Pauli measurements, providing a route to universality without requiring a universal transversal gate set.	Requires additional noisy resource states and repeated distillation operations; the resource overhead increases as the target output error is reduced.	Applicable at the protocol level to stabilizer-code architectures that support fault-tolerant Clifford operations and measurements; not tied to a specific hardware platform.	Theory [30]
Permutation-invariant codes with transversal phase gates	Maps suitable spin codes to permutation-invariant multiqubit codes while preserving their transversal gate groups; the resulting constructions include an $((11, 2, 3))$ code with a transversal T gate.	Requires specialized permutation-invariant code structure; the cited work does not establish a hardware-specific implementation or a universal transversal gate set.	Spin-code-derived permutation-invariant multiqubit codes; the cited construction is platform independent.	Theory [33]
Fold-transversal Clifford gates	Uses code dualities to implement additional logical operations in suitable CSS/LDPC codes using transversal gates and qubit permutations; for the $[[30, 8, 3]]$ Bring's code, restricting to four logical qubits yields the full Clifford group on that subspace.	Requires suitable code symmetries and code-preserving qubit permutations; the physical cost of realizing those permutations is architecture dependent and is not experimentally benchmarked in the cited work.	Suitable CSS/LDPC codes; the cited work analyzes the hyperbolic Bring's code as a representative construction.	Theory [34]
Color-code fault-tolerant operations	Error-rate-weighted decoding yields a 0.47% circuit-level threshold for the $6.6.6$ triangular color code, while the same work develops circuit-level lattice-surgery decoding and an improved state-injection protocol.	Requires color-code syndrome extraction and decoding together with lattice-surgery and state-injection operations; non-Clifford universality still requires resource-state preparation rather than a universal transversal gate set.	The $6.6.6$ triangular color code in two-dimensional architectures under the circuit-level noise model considered in the cited work.	Theory / Simulation [12]

TABLE 5. Recent advances in bias-tailored and bosonic codes.

Code/architecture	Mechanism, representative result, and engineering trade-off	Ref.
Domain-wall color code	Domain walls permute excitations across the color-code lattice. In the infinite-bias limit, the X^3Z^3 construction reduces to repetition-code sectors and reaches the 50% code-capacity threshold; at finite bias, its reported code-capacity thresholds match those of XZZX for all single-qubit Pauli channels. These headline results assume perfect stabilizer measurements. A scalable matching-based restriction decoder is provided, but its thresholds remain below the approximate maximum-likelihood results, and fault-tolerant threshold calculations remain open.	[35]
XZZX surface code	A locally equivalent surface-code variant tailored to structured Pauli noise. Its numerically estimated code-capacity threshold matches the zero-rate hashing bound for every single-qubit Pauli channel, with particularly strong performance in biased-noise regimes; under dominant dephasing with unreliable syndrome measurements, the reported fault-tolerant thresholds surpass previously known values. The quantitative advantage therefore depends on the physical noise bias, syndrome-noise model, and decoder rather than on code distance alone.	[13]
LDPC–cat hybrid	Combines sparse phase-flip LDPC checks with strongly noise-biased cat qubits. Under the source assumption of $p_{phys} \equiv \epsilon \approx 0.1\%$ phase-flip error per qubit and operation, 100 logical qubits are implemented in the proposed architecture using 758 cat qubits with $P_L \equiv \epsilon_L \leq 10^{-8}$ per cycle and logical qubit. The QEC layer uses short-range two-dimensional interactions and low-weight stabilizers, while a fault-tolerant universal gate set requires an additional routing layer of cat qubits. The reported performance is an architecture-level numerical result rather than an experimental QEC demonstration.	[36]
Chiral-cat code	Uses engineered higher-order nonlinearities and optical bistability to form code and error manifolds within a single oscillator. The resulting chiral phase-space structure can capture bit-flip events, and numerical analysis reports strong error suppression, including exponential improvement with cat size in the studied regimes while retaining fast-control capability. The approach requires engineered nonlinear processes, parametric control, and suitable operating points, and currently remains a theoretical/numerical proposal.	[37]
Concatenated cat–repetition memory	Experimentally concatenates stabilized bosonic cat qubits with an outer distance-5 repetition code. The phase-flip-correcting repetition code operates below threshold, with minimum measured logical error per cycle of $1.75(2)\%$ for the distance-3 sections and $1.65(3)\%$ for distance 5. Because physical bit flips are not corrected by the outer repetition code, the construction does not possess a proper asymptotic threshold; the present device is limited by intrinsic cat bit/phase-flip errors and ancilla-related error channels.	[17]

the available transversal logical gate set, whereas others provide fault-tolerant nontransversal resources or alter the code and geometry needed to implement logical operations. The practical choice therefore remains tied to code structure, connectivity, syndrome-processing requirements, resource-state overhead, and hardware capabilities, while a universal transversal logical gate set remains excluded by Eastin–Knill.

F. BIAS-TAILORED AND BOSONIC CODES

This subsection examines bias-tailored codes and bosonic encodings, which exploit asymmetric noise to reduce overhead in fault-tolerant quantum computing. Physical noise in leading platforms is often biased: phase-flip Z errors dominate over bit flips in many qubit systems, and photon loss is the primary fault in bosonic modes. Bias-tailored codes allocate more protection to frequent error channels while tolerating rarer ones; bosonic codes encode information directly in harmonic oscillators or photonic modes, using larger Hilbert spaces to suppress specific errors. Together, these approaches can achieve higher thresholds, smaller footprints, and shallower circuits than bias-agnostic designs [13].

Table 5 summarizes representative bias-tailored and bosonic approaches, comparing how each exploits the dominant error structure with the corresponding decoding, control, hardware, or evidence-maturity constraints.

The entries in Table 5 exploit noise structure at different levels of the QEC stack and therefore should not be read as a ranking by a single threshold or overhead

IEEE Proof

figure. Domain-wall and XZZX codes retain qubit-array topological encodings while tailoring their stabilizer structure to biased Pauli noise. LDPC–cat architectures combine sparse coding with a hardware qubit whose native noise is strongly biased, transferring part of the implementation burden to cat stabilization and logical-gate routing. Chiral-cat proposals instead seek additional protection within a single oscillator through engineered nonlinear dynamics, whereas concatenated cat–repetition memories have experimentally demonstrated bosonic error suppression combined with an outer phase-flip code. The relevant engineering choice therefore depends on the physical noise bias, available control and nonlinear elements, syndrome-extraction and decoder assumptions, target logical error rate, and maturity of the required logical operations.

Developing a theory for bridging qubit arrays, bosonic encodings, and LDPC hybrids remains a key open challenge for scaling fault tolerance.

G. NON-PAULI FUNNELS: LEAKAGE, HOOKS, AND NON-CLIFFORD NOISE

In this subsection, we examine error processes that fall outside the standard Pauli error model. Non-Pauli errors—such as leakage, meaning population leaving the computational $|0\rangle/|1\rangle$ subspace, coherent over-rotation, and non-Clifford hook faults that spread across multi-qubit gates—fall outside the standard Pauli error model. They violate stabilizer-code assumptions, distort syndrome statistics, and can limit logical-error scaling even when Pauli error budgets look favorable across superconducting, ion-trap, neutral-atom, and photonic platforms; see Fig. 7 for the design checklist.

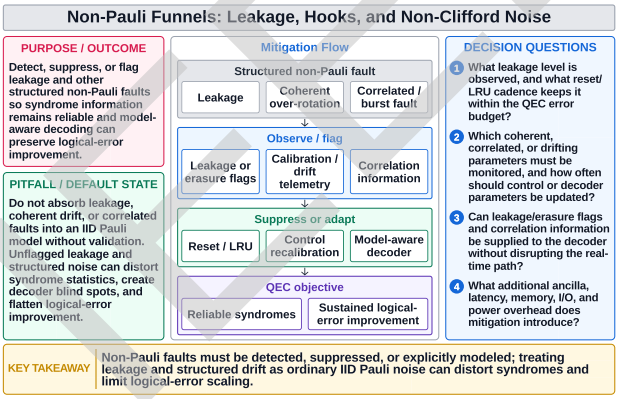


FIGURE 7. Non-Pauli channels corrupt syndromes and stall distance scaling; this checklist focuses design on purge cadence, drift tracking, model-aware decoding, and overhead budgeting under sub-μ s deadlines.

Section VI revisits these mechanisms in detail, using hardware case studies, e.g., leakage removal on transmon arrays, real-time calibration on trapped ions, and context-aware scheduling and compilation, and surveys accelerator-backed decoders and ML-driven controllers that address mixed Pauli-plus-leakage channels and correlated, bursty noise.

Designing a leakage-aware, correlation-robust decoder that natively handles non-Pauli faults while meeting sub-microsecond ($\tau_{\mathrm{dec}} < 1\ \mu s$) deadlines remains the central open challenge.

H. MEASUREMENT-SPEED AND SYNDROME-CADENCE LIMITS

The efficiency of QEC hinges not only on code properties but also on the temporal dynamics of syndrome extraction and processing. In this subsection, we briefly introduce timing constraints for syndrome extraction and how they interact with decoder throughput and hardware coherence.

Two timing quantities dominate: syndrome cadence, meaning the frequency at which stabilizers are read, and measurement latency, meaning the duration of each readout. If cycles exceed coherence times or decoder deadlines, uncorrected faults accumulate, and logical failure rates accelerate. On current superconducting platforms, this implies end-to-end extraction within $\lesssim 1\ \mu s$ [38]. On Sycamore, model-based optimization achieves $\approx$ 500 ns mid-circuit readout of 17 qubits with $\approx$ 1.5% error and suppressed leakage [39].

Cadence and latency, therefore, co-determine whether a code can operate inside the coherence envelope. Approaches such as commutation-aware check partitioning, ancilla reuse, streaming pipelines, and measurement-free primitives have been proposed [40], [41], [42]. Still, their implementation is platform-specific and will be discussed in Section III. For constant-weight quantum LDPC codes, no construction yet achieves both $O(\log n)$ extraction depth and $O(n)$ ancillas under planar connectivity, a central open problem [41], [42]. Table 6 highlights alternate approaches in measurement speed and syndrome cadence.

TABLE 6. Alternate frontiers in measurement speed and syndrome cadence.

Frontier	Key Idea and Impact	References
Measurement-free QEC	Replace mid-circuit readout with coherent copy-and-feedback operations, avoiding the slowest element in many hardware stacks.	[40]
Syndrome-circuit noise spectroscopy, ACES-based	Characterize full-syndrome circuits at code distance 5, enabling targeted pulse-level fixes to shorten effective latency.	

To summarize, measurement speed and cadence must be co-tuned with decoder throughput and coherence limits: if cycles exceed coherence times or the real-time envelope of Section IV, backlog-induced logical failures emerge. Fig. 8 ties the main decision questions and points from this subsection together. While recent proposals have shortened extraction depth or stabilized cadence using techniques such as measurement-free QEC, syndrome-circuit noise

IEEE Proof

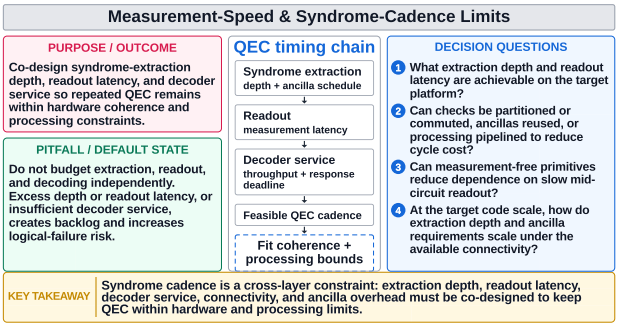


FIGURE 8. **Illustrates how the timing of syndrome extraction and decoding must remain within hardware coherence and processing bounds to prevent backlog and logical failure.**

spectroscopy, and pipelined, correlated-event statistics with feedback-calibrated decoding [40], the following challenge remains.

Designing an efficient, locality-respecting extraction method compatible with large-scale qLDPC codes remains unproven.

I. HARDWARE MAPPING CONSTRAINTS

This subsection highlights how hardware mapping creates a second "threshold problem" for QEC. A code that looks powerful in theory can lose its logical-error advantage when its logical qubits and stabilizer checks are mapped onto a processor with sparse connectivity, dead qubits, or frequency collisions. Long chains of two-qubit exchanges, i.e., SWAP gates that move quantum states across the chip, inflate circuit depth and stress the cadence limits of Section II-H. At the same time, defects and fabrication yield, coherence times, and native connectivity, discussed in Section III, further constrain viable embeddings; see Fig. 9.

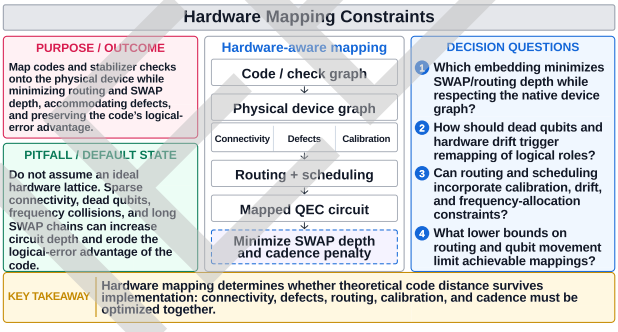


FIGURE 9. **Design frame for hardware-aware code mapping.**

Hardware-aware layouts and compilers mitigate these effects but remain heuristic. Current strategies include cyclic stabilizer embeddings on the device graph [43], semipermeable color-code boundaries that reduce SWAP overhead [44], adaptive qubit mapping that reshuffles logical roles as hardware drifts [45], calibration-aware transpilation that folds error and drift data into circuit scheduling, and global frequency-allocation optimization to avoid crosstalk-inducing collisions [46]. These techniques typically reduce space–time overhead by factors of 2–3, yet still lack rigorous optimality guarantees.

Establishing provable lower bounds on routing and movement overhead under realistic device graphs remains the central theoretical open challenge.

To summarize Section II, the core challenges of QEC arise from coupling code design, realistic noise, and hardware limits: under strict two-dimensional geometric locality, rate and distance obey fundamental scaling constraints, while modern qLDPC constructions can achieve favorable rate–distance scaling by relaxing the corresponding geometric-embedding requirement; correlated and non-Pauli noise invalidate simple IID assumptions; decoder latency and Eastin–Knill's ban on fully transversal universality introduce additional overheads. Bias-tailored and bosonic codes, cadence engineering, and mapping strategies each address part of this space, but a unified framework that combines favorable code scaling, efficient decoding, noise awareness, and practical hardware embedding at scale remains open.

III. HARDWARE AND PHYSICAL IMPLEMENTATION

In this section, we survey hardware and physical implementation challenges for QEC, emphasizing co-design between qubit devices, connectivity, classical control, and packaging. We focus on how platform-dependent figures of merit—coherence distributions, coupling graphs, control latency, thermal load, and wiring overhead—interact with the cadence and decoding constraints of Section IV. The discussion is organized into six subsections on platform constraints and code embedding; leakage, crosstalk, and thermal stability; fault-tolerant gate compatibility; materials and fabrication; cryogenic control; and interconnects and packaging (see Fig. 10 and Table 7).

A. PLATFORM CONSTRAINTS AND CODE EMBEDDING

In this subsection, we explain how native connectivity and interaction graphs constrain code choice, embedding overhead, and practical thresholds. Embedding a code's stabilizer or Tanner graph onto hardware introduces SWAPs, routing congestion, and locality violations that deepen syndrome extraction and raise logical error rates at fixed physical error rates. Thus, connectivity effectively sets a "second threshold," alongside physical-noise limits (Section VI) and cadence/latency limits (Section IV).

On superconducting transmons, planar connectivity makes surface codes a natural fit. On square lattices, embedding is direct; IBM's heavy-hex reduces degree from four to three to lower crosstalk and frequency crowding on 127-qubit Eagle and 433-qubit Ospray devices, but at the cost of extra SWAPs for distance-d patches and a modest depression of effective threshold [47]. On planar and heavy-hex lattices, surface codes embed naturally; crosstalk-aware scheduling keeps

IEEE Proof

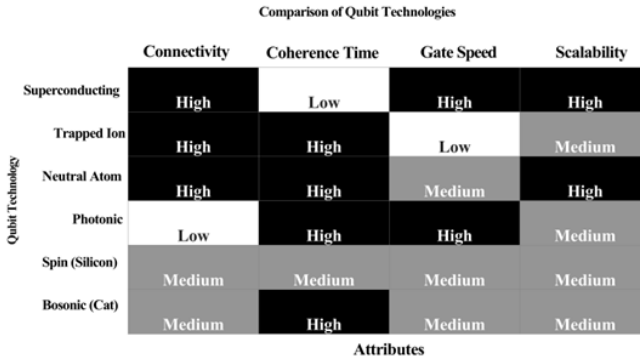


FIGURE 10. Comparison of qubit technologies. Heatmap of platforms versus four QEC-relevant attributes.

cycle times near the $\sim 1\ \mu$ s envelope assumed in Section IV, while subsystem and Floquet codes trade decoding and calibration complexity for better graph matching [48].

Bias-preserving bosonic qubits (e.g., cat states) offload X-error suppression to hardware, reducing the required outer-code distance: a concatenated cat–repetition memory achieved 1.65(3)% logical error per QEC cycle at $d = 5$, compared with approximately 1.75% for its distance-3 sections [17]. Bosonic+LDPC hybrids concatenate modular LDPC checks with cat encodings to reduce overhead toward universal fault tolerance [49].

In trapped ions, all-to-all or reconfigurable connectivity via motional modes and shuttling relaxes embedding constraints. QCCD "racetrack" processors support configurable connectivity with periodic boundary conditions, enabling [[7, 1, 3]] Steane and [[9, 1, 3]] Bacon–Shor codes without SWAP overhead and with improved Steane-style extraction [50], [51]. Recent QCCD traps have demonstrated two logical qubits entangled via transversal CNOT and logical-error suppression by $\approx 10\times$–$500\times$ compared to physical error, with repeated syndrome rounds showing fault-tolerant memory beyond a single cycle [52], [53]. Connectivity is favorable, but gate times in the tens of microseconds still bound the QEC cycle and must be evaluated against the decoding envelope of Section IV.

In photonic platforms, computation is largely measurement-based: QEC is embedded in large cluster states stitched together via probabilistic fusion. Fusion/cluster schemes have realized small logical encodings (e.g., [[4, 2, 2]] in time/frequency modes) [54], [55]. Because there is no long-lived memory, ultrafast feed-forward, delay-line/cavity storage, and low-loss components are essential; decoder latency must track the photonic clock and is therefore directly governed by the real-time constraints of Section IV [56].

Neutral-atom arrays realize 2D-like locality with limited tweezer reconfiguration and native erasure detection, supporting surface/color codes with erasure-aware decoders, though distance scaling is currently routing-limited [57], [58]. Semiconductor spin qubits provide CMOS-compatible footprints but nearest-neighbor exchange and charge noise favor shuttle- or resonator-linked modular layouts and small-distance codes [59], [60]. Bosonic encodings again exploit large oscillator Hilbert spaces but still require outer codes to handle photon loss and residual errors [17].

In summary, each platform's native connectivity and gate set determines its natural code families and practical thresholds. Superconducting transmons favor local surface/surgery layouts, with heavy-hex trading lower crosstalk for extra SWAP overhead; ion traps' near–all-to-all connectivity supports transversal Steane/Bacon–Shor codes with preserved thresholds; and photonic fusion/cluster schemes exploit entanglement at the cost of probabilistic-gate overhead. Neutral atoms, spin qubits, and bosonic encodings are promising but less mature, motivating hybrid strategies such as superconducting processors linked by photonic interconnects or bosonic memories paired with LDPC outer codes [17], [65]. Fig. 11 synthesizes III.A.

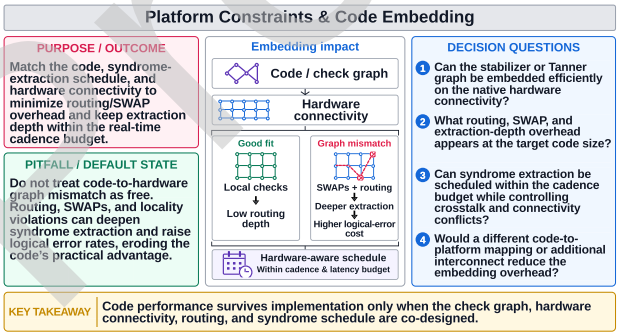


FIGURE 11. Embedding checklist (synthesis of III. A): objective, dominant pitfall, and four questions to align code and schedule with native connectivity and cycle cadence.

B. LEAKAGE, CROSSTALK, AND THERMAL STABILITY

In this subsection, we summarize how leakage, crosstalk, and thermal or slow drift appear in hardware, how they impact QEC thresholds and cadence, and which mitigation hooks tie back to scheduling (Section III-A) and decoding (Section IV).

These metrics feed directly into decoder choice, cadence, and power budgets (Sections II and IV) (see Fig. 12 for a design checklist and Table 8 for definitions).

On superconducting devices, weak anharmonicity causes $|2\rangle$ leakage during fast gates. DRAG pulse shaping, dynamic leakage resets, and LRUs mitigate this; experiments show that inserting LRUs at fixed cadence preserves pseudo-threshold behavior in repeated surface-code cycles [58], [66], [67], [75], [68]. Crosstalk from residual ZZ and microwave bleed-through grows with parallelism and depends on embedding (Section III-A); crosstalk-aware scheduling and compiler constraints reduce correlated faults that would otherwise depress thresholds [69]. Dielectric loss, two-level systems (TLS), quasiparticle bursts, and IR/blackbody loading shorten T_1 and introduce non-IID multi-qubit events [27]; IR filtering, shielding, phonon traps, pumps, and active temperature regulation help stabilize these

IEEE Proof

TABLE 7. QEC hardware platforms and representative evidence.

Platform	Native connectivity	Coherence / storage metric	Representative entangling-operation fidelity / error metric	Measurement / readout metric	Representative QEC / syndrome-cycle or relevant operation time	QEC evidence / maturity	Scalability bottleneck	Ref.
Superconducting	2D square-grid transmon architecture; local CZ couplings support geometrically local surface-code extraction	Willow mean $T_1 = 68\ \mu s$ and $T_{2,\mathrm{CPMG}} = 89\ \mu s$ [11]	Willow median inferred CZ Pauli error 0.29%, measured by cross-entropy benchmarking; this is a reported Pauli-error metric, not a process-fidelity value [61]	Willow median readout errors: 0.53% for $\lvert 0\rangle$ and 1.0% for $\lvert 1\rangle$ [61]	Surface-code QEC cycle: $1.1\ \mu s$ [11]	Experimental below-threshold surface-code memory. For $d = 7$, the 101-qubit code achieved $\epsilon_7 = (1.43 \pm 0.03) \times 10^{-3}$ per cycle and $\Lambda = 2.14 \pm 0.02$. A $d = 5$ code was also operated with real-time decoding for up to 10^6 cycles [11]	Local connectivity; leakage and stray CZ interactions; correlated errors; calibration drift; cryogenic I/O and wiring; real-time decoder throughput as code distance increases [11]	[11], [61]
Trapped ions	Linear ion chain with individually addressed operations and long-range entangling interactions mediated by collective motional modes	Optical-qubit encoding: $T_2 \approx 50\ ms$; ground-state encoding: approximately $5\ ms$. With dynamical decoupling during recooling, no significant dephasing was observed to $60\ ms$, indicating an effective coherence time $> 1\ s$ [51]	Estimated two-qubit error probability $p_{2q} = 0.027$ with standard deviation 0.005, obtained from repeated 16-qubit GHZ-state characterization; not an RB process fidelity [51]	Typical measurement fault probability used for the device model: $p_{meas} = 0.003$ [51]	Two-qubit gate: $322.5\ \mu s$. Mid-circuit sideband cooling requires approximately $15\ ms$; a single complete QEC-round duration is NR [51]	Experimental implementation of multiple rounds of fault-tolerant Steane QEC, including repetition- and seven-qubit-code syndrome extraction [51]	Slow entangling operations relative to solid-state platforms; motional-mode crowding; lengthy mid-circuit recooling; optical-addressing crosstalk and control complexity [51]	[51]
Neutral atoms	Reconfigurable tweezer arrays with atom transport between storage, entangling, readout, and reservoir zones; movement provides programmable connectivity	Coherence in 852-nm traps approximately 1–2 s, depending on the dynamical-decoupling sequence [62]	CZ fidelity approximately 99.5% in repeated benchmarking with $100\ \mu s$ between gates; calibrated CZ performance up to approximately 99.6% is also reported [62]	Non-destructive, loss-resolved readout: $0.46(4)\%$ bit-flip error and $0.24(2)\%$ atom loss [62]	Repeated surface-code QEC round: $4.45\ ms$ [62]	Experimental repeated surface-code QEC on arrays up to 448 atoms. After four QEC rounds, machine-learning decoding with loss information gave $d = 5$ error per round $2.14(13)\times$ lower than $d = 3$; the reported $d = 5$ error per round was $0.62(3)\%$ [62]	Atom loss and replenishment; transport/rearrangement overhead; Rydberg-gate drift and inhomogeneity; imaging and re-initialization latency; optical-control parallelism [62]	[62]
Photonics	Measurement-based graph-state architecture; graph fragments generated by emitters are connected through photon-mediated fusion	Conventional flying-photon T_1/T_2: NC. The atomic emitter qubit used in the cited graph-state experiment had $T_2 \approx 1\ ms$; optical loss rather than stationary-qubit lifetime is the principal photonic storage/transport constraint [55]	Fusion-generated two-emitter Bell state: $F = 0.915 \pm 0.005$. The value depends on photon-arrival-time postselection and ranges from 0.851 ± 0.006 to 0.963 ± 0.008 under the studied criteria [55]	Overall single-photon source-to-detector efficiency was close to 0.5 in the reported experiment [55]	Complete repeated-QEC cycle: NR. The experiment used a $1\ \mu s$ absolute photon-detection window and the principal fusion data used $\lvert\Delta t\rvert \le 250\ ns$ relative-arrival postselection [55]	Experimental deterministic generation and fusion of photonic graph-state fragments, including ring states of up to 6 logical / 8 physical qubits and a tree state of 7 logical / 8 physical qubits. Fusion-based fault-tolerant architectures are established at the protocol level, but a directly comparable repeated full-QEC hardware cycle is NR [55], [63]	End-to-end photon loss; source-to-detector efficiency; photon indistinguishability; fusion success and fidelity; synchronization, buffering, and feed-forward latency [55], [63]	[55], [63]
Spin qubits	Predominantly local exchange coupling between neighbouring gate-defined quantum dots; larger layouts require shuttling or additional coupling/routing mechanisms	A single platform-wide coherence value directly matched to the reported two-qubit benchmark is NR in the cited study	Silicon double-dot CZ gate fidelity $99.65 \pm 0.15\%$, obtained using gate-set tomography; the quoted uncertainty is the reported 2σ interval [64]	Device-specific assignment/readout fidelity is NR in the cited two-qubit-gate benchmark; the experiment reports that algorithmic accuracy was limited partly by readout errors [64]	Repeated QEC-cycle time: NR in the cited study	High-fidelity physical logic benchmark: single- and two-qubit gate fidelities exceed 99.5%, placing the demonstrated two-qubit logic above the commonly cited surface-code gate-fidelity target. The cited experiment does *not* demonstrate repeated QEC [64]	Nearest-neighbour connectivity; charge-noise sensitivity; idling and crosstalk errors; maintaining small-device fidelities across larger dot arrays; scalable initialization and readout [64]	[64]
Bosonic / oscillator	Five storage resonators forming cat-qubit data modes coupled through four transmon syndrome ancillas; nearest-neighbour $\bar{X}_i\bar{X}_{i+1}$ repetition checks	Storage-mode averages: $T_1 > 60\ \mu s$ and $T_2 > 80\ \mu s$ [17]	A conventional CX process fidelity is NR. The cited experiment instead characterizes a noise-biased cat–transmon CX interaction and its preservation of long cat bit-flip times during syndrome extraction [17]	Three-state transmon readout resolves $\lvert g\rangle$, $\lvert e\rangle$, $\lvert f\rangle$ and identifies ancilla erasures. An $\lvert f\rangle \to \lvert e\rangle$ decay has about 50% probability of causing a syndrome-measurement error; using erasure information reduces the effective syndrome-measurement error by more than $2\times$ at $\lvert\alpha\rvert^2 = 1$ [17]	Repeated syndrome/QEC cycle: $2.8\ \mu s$ [17]	Experimental concatenated cat–repetition QEC to $d = 5$. The phase-flip-correcting repetition sector operates below threshold; minimum total logical error was $1.65(3)\%$ per QEC cycle for $d = 5$ [17]	Photon-loss-induced errors; ancilla-induced errors; preserving the native noise bias during entangling operations; stabilization and calibration complexity; resonator, buffer, coupler, and ancilla-mode overhead [17]	[17]

Comparison convention: Entries retain the metric definitions and experimental conditions of the cited primary sources for source-conditioned comparison. A reported error probability is not relabeled as a process fidelity. For photonic and bosonic systems, platform-specific loss, oscillator-lifetime, fusion, and syndrome-readout quantities are used where conventional two-level-qubit metrics are inappropriate. NR = not reported in the cited source on the stated basis; NC = not physically meaningful or not directly comparable on the stated basis. Values from different rows should therefore not be rank-ordered without accounting for differences in device, benchmarking protocol, encoding, workload, and QEC circuit.

TABLE 8. Figures of merit that affect logical performance and cadence.

Metric	QEC relevance and representative sources
Leakage rate	Population outside the computational subspace; mitigation includes leakage-reduction units, leakage-to-erasure conversion, coupler-assisted leakage reduction, and leakage-suppressing pulse design [58], [66]–[68].
Crosstalk spectrum $\chi(\omega)$	Residual ZZ, drive bleed-through, or other simultaneous-operation crosstalk can generate correlated faults; characterization, QEC-aware analysis, and context-aware compilation provide complementary mitigation routes [69]–[71].
Slow-drift and burst statistics	Low-frequency flux drift and quasiparticle or radiation-induced bursts make the effective noise time dependent and can create multi-qubit correlations, motivating explicit monitoring and non-IID modeling [27], [72]–[74].

channels so sub-μ s decoders (Section IV) can assume nearly stationary statistics within cryogenic power limits [72], [76], [77], [78].

In trapped-ion and neutral-atom platforms, leakage is less common, but motional-mode heating and laser spillover can induce correlated over-rotations during parallel gates. Stabilized lasers (reference cavities and feedback), sympathetic cooling, and atom reordering maintain fidelity across rounds, while decoders and ML models that ingest time-varying syndrome statistics help track slow drifts [4], [79]. All-to-all or reconfigurable connectivity eases scheduling pressure (Section III-A) for Steane/Bacon–Shor codes, but mode crowding in long chains caps practical parallelism [69]. In neutral-atom arrays, erasure detection is native, making

IEEE Proof

erasure-aware decoders natural tools for handling atom loss and Rydberg-related erasures [57].

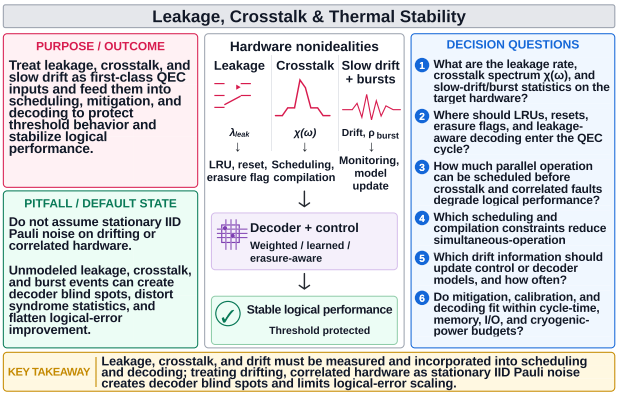


FIGURE 12. Design checklist for non-Pauli channels: measure λ_{leak}, $\chi(\omega)$, Δf_T, and burst statistics; route these signals into LRUs or erasure flags, crosstalk-aware scheduling, and correlation-aware decoders; and operate within parallelism, cycle-time, and cryogenic power limits. Assuming IID Pauli noise on drifting or correlated hardware creates decoder blind spots and can misestimate logical-error scaling.

In fusion/cluster-state photonics, leakage mostly appears as loss/erasure, while crosstalk appears as mode mixing. Both are dominated by component loss and synchronization. Ultrafast feed-forward plus low-loss filtering and delay lines are required to keep decoder latency synchronized with the photonic clock (Section IV) [54], [55], [56]. Erasure-aware surface/LDPC variants are natural choices if loss and timing stay inside this envelope.

In semiconductor spin qubits, leakage through valley or excited states, charge-noise-induced detuning, and local connectivity limits motivate short-distance codes and frequent recalibration, aligned with the cadence and decoder-power budgets of Section IV [60].

In summary, real devices exhibit leakage, correlated crosstalk, and drift that are not symmetric Pauli errors. Treating λ_{leak}, $\chi(\omega)$, and thermal/burst metrics as first-class design inputs—and feeding them into embedding and scheduling (Section III-A) as well as decoder and latency envelopes (Section IV)—preserves thresholds and stabilizes logical performance.

C. FAULT-TOLERANT GATE COMPATIBILITY

In this subsection, we describe how hardware supports logical operations inside a protected code space and how these choices interact with the Eastin–Knill limits of Section II-E and the real-time envelope of Section IV.

We use three metrics: (i) the transversal gate-set supported by the code, (ii) the resource overhead for non-Clifford gates (e.g., magic-state factories), and (iii) the depth and cadence of stabilizer measurements needed to implement gates fault-tolerantly. Because the Eastin–Knill theorem forbids a fully universal transversal gate set on finite-distance stabilizer codes [29], practical architectures combine transversal primitives with lattice surgery, teleportation, and magic-state distillation [30].

Surface-code logical CNOTs are typically implemented by lattice surgery (patch merge/split), while Steane and color codes support native transversal CNOTs. Folded, stacked, or reconfigurable layouts can also make transversal CNOTs available for surface-like codes at the cost of non-planar embeddings [34], [80]. In trapped ions, transversal CNOTs on small Steane/color blocks are natural using global Mølmer–Sørensen interactions and ion shuttling, and logical teleportation or surgery-like operations have been demonstrated [50], [51], [52], [81]. For superconducting processors, deformation-based layouts and virtualized-logical abstractions implement surgery-based CNOTs within the cycle constraints of Section IV.

Non-Clifford gates, such as T, are usually implemented using magic-state injection and distillation or code switching [82]. A recent trapped-ion experiment reached "break-even" fidelity for a logical non-Clifford using small-code magic-state preparation [83]. Bias-preserving bosonic qubits aim to reduce T-gate cost by exploiting strong noise bias [17]. On superconducting hardware, coupler-assisted leakage-reduction mitigates non-Pauli channels that would otherwise inflate factory depth and cadence cost [67]. In code space, 3D color codes offer transversal non-Cliffords in principle, while 2D domain-wall color-code variants tailored to biased noise explore alternative embeddings and trade-offs [35], [82]. Demonstrations of small color-code implementations in ions and superconductors are stepping stones toward these architectures, but still rely on reliable mid-circuit measurements and long-range connectivity.

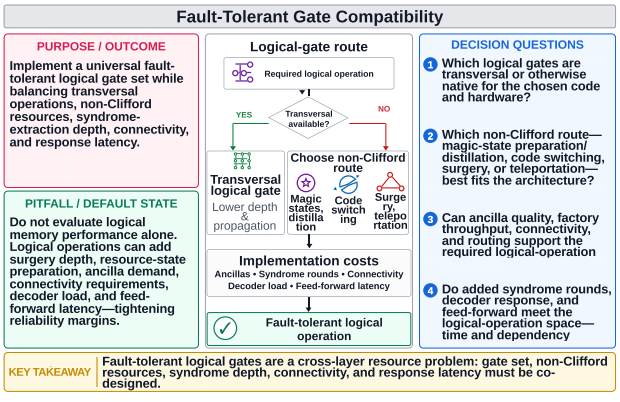


FIGURE 13. Gate-compatibility design map linking available fault-tolerant logical operations to non-Clifford resource overhead, hardware geometry, syndrome cadence, and feedback constraints.

Gate teleportation with logical Bell pairs can overcome limits of native gate sets at the cost of high-fidelity ancilla preparation and fast feed-forward. Such protocols have been demonstrated on ion-trap architectures [52]. Teleportation also enables heterogeneous designs in which different logical blocks run on different hardware, provided that ancilla quality and classical-feedback latency remain within the

IEEE Proof

envelope summarized in Section IV and recent latency studies [84]. Table 9 summarizes core techniques and their QEC implications, and Fig. 13 provides a design checklist.

In summary, hardware must support a high-fidelity stabilizer readout plus at least one practical route to a universal logical gate set. Code choice, embedding, and gate-synthesis protocol are interdependent and must be evaluated jointly against the cadence and decoder-latency/power envelopes of Section IV and the structural constraints of Section II-E [85], [86].

D. MATERIALS SCIENCE AND FABRICATION LIMITS

In this subsection, we examine how material defects and fabrication variability set floors on coherence, uniformity, and yield, and how these limits translate directly into QEC threshold margins and space–time overhead [87].

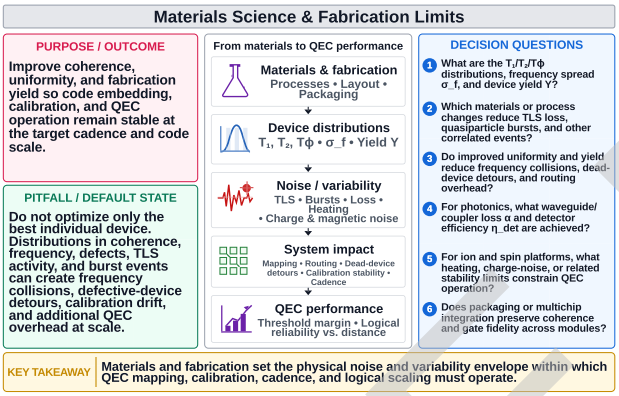


FIGURE 14. Materials and fabrication design map linking coherence, uniformity, yield, loss, and drift to QEC threshold margin, hardware mapping, recalibration, and system-scale overhead.

Relevant metrics include coherence time distributions $T_1/T_2/T_\phi$; frequency spread σ_f from junction-resistance variation; TLS density and quasiparticle-burst rate; device yield Y; and platform-specific loss or heating terms such as ion motional heating rate n, photonic waveguide/coupler loss α, and detector efficiency η_{det}. These quantities determine (i) how far below threshold a platform can realistically operate for a chosen code, (ii) whether mapping and embedding (Section III-A) can avoid frequency collisions and dead-qubit detours, and (iii) whether calibrations remain stable on the timescales required by the cadence and decoder-latency envelope of Section IV.

In superconducting qubits, the main materials drivers are surface/interface TLS, dielectric participation, quasiparticle bursts, IR/blackbody loading, and junction variability. Tantalum electrodes push transmon T_1 into the 0.3–0.5 ms range [88], [89]; "super-sandwich" dielectrics reduce participation and stabilize T_1 [77]; phonon-trap structures suppress and mitigate correlated quasiparticle events [27]; and wafer-scale junction control achieves $<5\%$ resistance spread with near-100% yield on fluxonia, enabling $T_1 \gtrsim 1$ ms and thousand-junction scaling [89]. Flip-chip multichip modules with indium bump bonds have demonstrated 99.1% CZ gates in 56 ns across chips without coherence loss [90]. For QEC, higher T_1 and fewer bursts reduce per-round logical error; tighter σ_f avoids frequency crowding and residual ZZ collisions that would otherwise depress thresholds or force SWAP-heavy rerouting (Section III-A); and stable stacks simplify drift management so that sub-μ s cycles and decoder power budgets (Section IV) remain realistic. Where flip-chip yield is high, multi-chip patches can relieve congestion and SWAP depth without coherence penalties [90].

In ion traps, electrode-surface electric-field noise drives motional heating, while photo-charging and imperfect RF potentials introduce drift and crosstalk. Surface cleaning (ion bombardment/plasma) and cryogenic operation suppress the motional heating rate n, allowing many two-qubit gates to be executed before re-cooling; on-chip superconducting RF resonators stabilize transport fields [52]. Improved UV optics and stabilized lasers further limit slow drift. For QEC, lower n keeps entangling-gate errors below practical thresholds and stabilizes syndrome statistics over many cycles. At the same time, stable transport enables transversal CNOTs on short-distance blocks (Section III-C) without additional depth inflation.

In photonic platforms, waveguide/coupler loss α and detector efficiency η_{det} dominate the effective erasure rate, while integration quality governs synchronization and mode mixing. Low-loss SiN circuits with $\lesssim 0.5\%$ beamsplitter loss and SNSPDs with $\eta_{\text{det}} > 95\%$ and low dark counts have been demonstrated in entangling meshes [91]. Each percentage point of loss reduces fusion fidelity and lowers QEC thresholds for cluster-state and fusion-based schemes [63]. For QEC, lower α and higher η_{det} move the operation toward regimes where erasure-aware surface/LDPC codes have high thresholds, provided feed-forward and decoding latency can keep up with the optical clock (Section IV).

Neutral-atom coherence is limited by Rydberg–surface interactions and optical scattering; improved coatings and vacuum packaging increase stability and support erasure-aware strategies, although parallelism remains bounded by optical crosstalk (Section III-B). Semiconductor spin devices benefit from isotopically enriched silicon, carefully engineered junction stacks, and micromagnets; these have yielded two-qubit fidelities above 99% [64]. For QEC, better uniformity makes small-distance codes with frequent recalibration more attractive, while long-range couplers help reduce routing penalties (see Fig. 14 for the checklist).

In summary, materials and fabrication set the error floor, uniformity, and yield that ultimately determine practical thresholds and space–time overhead. For superconductors, tantalum-class stacks, super-sandwich dielectrics, phonon traps, tight junction control, and flip-chip packaging jointly raise logical margins for surface/surgery layouts and ease embedding [77], [88], [89], [90]. For ions, low motional heating and stable transport enable transversal block codes with predictable cadence [52]. For photonics, low α and high η_{det} are critical for crossing erasure thresholds while

IEEE Proof

keeping feed-forward within decoder budgets [63], [91]. In all cases, materials choices should be treated as inputs to code selection, embedding, and cycle design, not just device-level optimizations [87].

E. CRYOGENIC CONTROL AND REAL-TIME SYSTEMS

In this subsection, we survey the cryogenic control stack that closes the quantum–classical QEC loop and link it directly to the latency and power envelopes of Section IV.

We track three system metrics: (i) end-to-end latency from measurement to corrective actuation or Pauli-frame update, (ii) power dissipation and I/O bandwidth per control/readout channel at cryogenic stages, and (iii) sustained decoding throughput at the target code distance. Latency and jitter appear as idle accumulation and stale Pauli frames; cryogenic power and bandwidth caps bound how many stabilizers and what code distance can be supported before backlogs form (Fig. 15).

Modern processors operate qubits at tens of millikelvins while most electronics remain warm; long cables introduce delay, heat load, and I/O limitations. To relieve this, control and readout are being pushed into the cryostat using cryo-CMOS waveform generation, digitization, and on-die DSP. Prototypes report multi-channel 4.5–5.5 GHz sources and closed-loop feedback at milliwatt-scale power per channel when thermally anchored cold, showing that generation and digitization can move to cryogenic stages without exhausting the thermal budget [60], [92], [93]. This shortens cycles and helps meet per-round logical targets; the same power caps limit how many stabilizers can run in parallel and thus how deep extraction can be at fixed physical error rates.

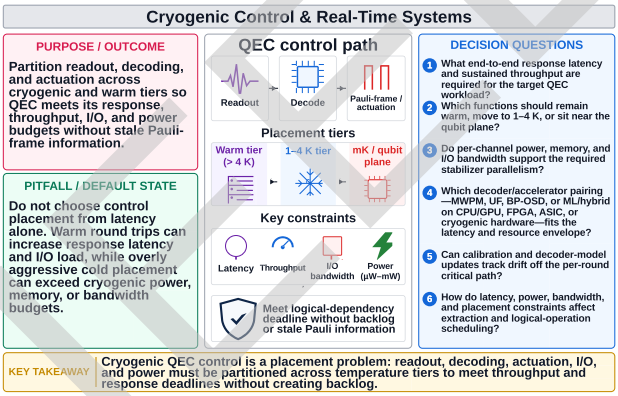


FIGURE 15. Checklist of cryogenic control & real-time budgets.

Routing and triggering at cold stages complete the fast path. Time- and frequency-division multiplexing reduce cable counts; reconfigurable cryogenic switches steer tones and measurements; and single-flux-quantum (SFQ) circuits provide picosecond-class triggers with negligible static power [38], [94], [95], [96]. With appropriate staging, stabilizer layers operate within crosstalk limits (Section III-B), and dynamic power remains within cryogenic budgets, ensuring systems remain within the real-time window of Section IV-A.

A key architectural decision is where to place "intelligence." Keeping readout, decoding, and actuation warm simplifies design but maximizes latency and cryostat I/O. Moving readout cold while decoding remains warm reduces I/O and improves SNR, but still incurs a warm round-trip for decisions. For ≈ 1 μ s cycles on fast platforms, a fully cold loop—cold readout, cold decode, and cold actuation—keeps the Pauli frame current on the hardware timescale. Experiments demonstrating sub-μ s per-round decoding and few-μ s closed-loop feedback show that such loops are feasible when co-designed with schedules and decoder architectures (Section IV) [78], [97], [98], [99], [100], [101]. If decoding remains warm, Pauli-frame updates can lag by multiple rounds, effectively adding idles and mis-timed flips and raising logical error per cycle even when gate fidelities are unchanged.

Decoding, therefore, is a time-consuming load on the control plane. To sustain ~ 1 μ s cycles, streamed, fixed-latency pipelines with bounded jitter are required; reported envelopes for practical distances include $\lesssim 1$ μ s per round, with on-cold memory bandwidth and milliwatt-scale power as dominant scaling constraints [97], [98]. Extra latency or jitter enlarges the unresolved space–time decoding window; backlog converts algorithmic accuracy into delay and can increase operational logical-failure risk or logical-clock time without changing the algebraic code distance d, motivating the throughput–backlog–deadline treatment in Section IV. Non-stationary noise motivates learned or weighted decoders, but parameter updates (edge weights, erasure priors) are best applied at millisecond-to-second timescales based on calibration telemetry, so they track slow drifts and bursts without entering the per-round critical path.

Platform implications follow directly. Superconducting processors benefit from co-located cryogenic sources, digitizers, switches, and fixed-latency decoders that close the loop in hundreds of nanoseconds, enabling surface- and lattice-surgery schedules to remain below threshold at realistic distances [78], [97], [98], [99], [100], [101]. Ion-trap systems emphasize analog stability and deterministic, but slower, control, so warm decoding can suffice if cycles are long and the Pauli-frame lag stays within the error-per-round budget. Photonic and other erasure-dominated stacks push measurement, feed-forward, and often decoding to the detector plane. Hence, decisions keep pace with the optical clock, with on-chip aggregation and inference minimizing path length to the decoder [17], [66], [91].

In short, cryogenic control bridges hardware physics (Section III) and decoding theory (Section II). Bringing generation, digitization, switching, and inference into the cryostat reduces latency, cabling, and thermal load. Still, it imposes strict limits on power and on-cold bandwidth that codes and schedules must obey. With fixed, sub-μ s measurement–decision–actuation pipelines operating within the available tier-specific power and energy budgets, planar

IEEE Proof

TABLE 9. Fault-tolerant gate compatibility.

Technique	Superconducting	Trapped ions	Trade-offs/notes	Ref.
Transversal CNOT in-code	Color-code patches can support transversal CNOTs; surface-code implementations more commonly use surgery, while fold-transversal, stacked, or reconfigurable layouts relax strict planar locality.	Small Steane/color-code blocks can exploit transversal logical gates under favorable connectivity.	Transversal constructions reduce fault propagation but can require nonlocal connectivity, permutations, or reconfigurable layouts.	[12], [34], [44], [80], [102]
Lattice-surgery CNOT	Patch merge/split is a standard surface-code route to logical CNOT.	Surgery-like and teleportation-based logical operations are possible, although transversal gates can be more natural on small ion-code blocks.	Additional syndrome-extraction rounds increase cadence and decoder load.	[52], [102]
Non-Cliffords (T/Toffoli)	Magic-state distillation, code switching, and bias-preserving bosonic approaches provide complementary routes; leakage mitigation can reduce hardware-induced overhead.	Magic-state preparation/distillation and code switching provide fault-tolerant non-Clifford routes.	Ancilla quality, factory throughput, leakage, and feed-forward latency set practical cost.	[17], [67], [82], [83]
Primary QEC consequence	Surgery depth and factory throughput tighten the space–time budget.	Transversal operations can reduce depth, while transport and mode crowding constrain cadence.	Logical-operation feasibility remains coupled to decoder latency, connectivity, and ancilla resources.	[50]–[52], [81], [83]–[85]

surface/surgery layouts align well with current constraints. In contrast, higher-degree LDPC variants must be checked against on-cold memory and I/O limits before deployment [97], [98].

F. INTERCONNECTS, PACKAGING AND INTEGRATION SYSTEMS

In this subsection, we discuss how interconnects and packaging determine whether large-scale QEC processors can be wired, cooled, and synchronized without sacrificing coherence, and how these constraints link to embedding (Section III-A) and decoding (Section IV).

Key constraints are I/O density, per-link thermal load, coherence retention across modules, and link latency or bandwidth for streaming stabilizer rounds. These bounds feed directly into code embedding and routing, into leakage/crosstalk behavior, and into the real-time decoder envelope. Multichip integration is the primary route to overcoming routing congestion and I/O fan-out in superconducting devices. Flip-chip bonding and 3D integration with indium bumps or through-silicon vias for placement control, with coupler chips face-to-face with qubit dies. Two- to four-die modules preserve coherence across chip boundaries, with inter-die couplers achieving >99% gate fidelity and bump resistances in the hundreds of nΩ [60], [92], [94], [103]. This removes the traditional bonding "coherence tax" and enables tileable layouts for lattice surgery [102]. Shorter control paths and reduced SWAP overhead help surface-code schedules stay within the ≈1 μ s cadence window of Sections III and IV. IBM and Google prototypes now use tiled heavy-hex layouts with inter-die couplers that preserve connectivity across chip boundaries while keeping crosstalk and thermal load inside cryogenic budgets [60], [97].

In trapped-ion systems, scaling QCCD traps demands dense vacuum feedthroughs and stable analog distribution. Integrated capacitive filters, multiplexed DACs, and on-chip optics reduce wiring and suppress noise while maintaining precise electrode control [60], [104], [105], [106], [107]. Integrated waveguides deliver beams directly to ions, improving alignment and enabling parallel operations. For QEC, fewer lines and cleaner analog paths reduce motional heating and drift, supporting transversal Steane/Bacon–Shor codes at predictable cadence (Section III-C) without creating decoder backlog, while multiple interaction zones per module increase parallel stabilizer weight before motional-mode crowding becomes the limiter.

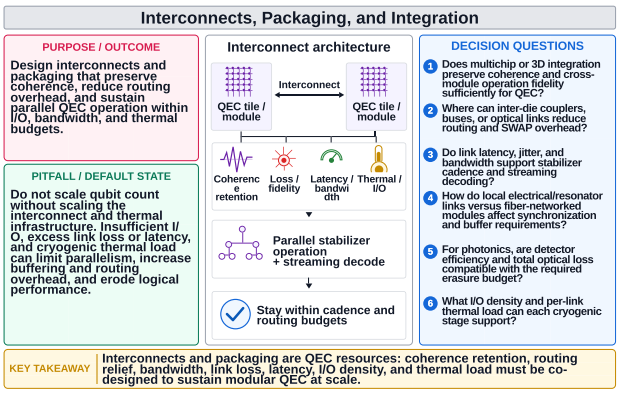


FIGURE 16. Interconnect/packaging design checklist.

For photonic architectures, the dominant constraints are optical loss and synchronization. Integrated platforms now combine sources, interferometers, and SNSPDs on a single chip, with detector efficiencies above 95% and splitter losses below ≈0.5% per element [76], [91]. These parameters set the

IEEE Proof

effective erasure rate: each additional percent of loss reduces fusion fidelity and lowers QEC thresholds for cluster-state and fusion-based codes [63]. Monolithic integration of detectors and sources reduces timing jitter and improves mode stability, both of which are critical for fusion-based quantum computation. Lower loss and feed-forward electronics co-located with detectors keep syndrome cadence aligned with the photonic clock and prevent buffers from growing beyond the decoding envelope of Section IV [66], [91].

Beyond single-module packaging, transducer-mediated links enable distributed fault tolerance. Electro-optic and rare-earth transducers convert microwave or ion excitations into telecom photons for long-distance networking [17], [66]. Two main patterns emerge: (i) local superconducting interconnects—resonator buses or short flex lines—that preserve coherence across multichip tiles but must obey cryogenic power limits (Section III-E); and (ii) optical-fiber interconnects that distribute entanglement across cryostats or racks, introducing a network clock and buffer requirements that must be co-designed with streaming decoders (Section IV). Hybrid schemes link semiconductor spin qubits coherently to superconducting resonators, forming dense "compute tiles" connected by cavity buses; longitudinal couplings enhance readout and mitigate Purcell loss, suggesting modular hybrid processors in which superconductors provide connectivity and spins provide density [38].

In summary (see Fig. 16), interconnects and packaging are QEC resources, not afterthoughts. Multichip superconducting modules and tiled layouts reduce routing overhead and preserve coherence, enabling scalable surface-code designs within timing and power limits. In ion traps, wiring reduction and integrated optics stabilize timing and heating, allowing transversal block codes to run without excess drift. In photonics, detector efficiency and low-loss SiN waveguides are critical for crossing erasure thresholds while keeping feed-forward within the decoder envelope. Across platforms, interconnect density sets the maximum stabilizer weight in parallel, thermal load bounds cadence, and coherence retention determines whether logical error rates stay below threshold. As distributed and hybrid systems emerge, packaging and interconnects will increasingly dictate which codes and embeddings remain viable at scale.

IV. DECODER ARCHITECTURES AND COMPUTATIONAL BOTTLENECKS IN QEC

This section surveys the spectrum of quantum-error-correction (QEC) decoders and the computational bottlenecks they face when syndrome data arrives at MHz rates. We define strict latency–throughput targets and contrast accuracy-optimal but heavy minimum-weight perfect matching (MWPM) decoders with near-linear-time Union-Find (UF), belief propagation (BP/OSD), and ML approaches (autonomous error correction: [108]). We also review FPGA/ASIC accelerators that push latency down to sub-μ s. The core message is that once decoding falls even a few microseconds behind, accumulated physical errors can erase the accuracy gains achieved. Scalable, hardware-aware decoders that balance speed, fidelity, and power/placement constraints (Section III-E) are indispensable for fault-tolerant computing.

A. REAL-TIME DECODING REQUIREMENTS AND BOTTLENECKS

In this subsection, we define the timing and power constraints imposed by MHz-rate QEC cycles and explain why decoders must operate online without creating persistent backlogs. A QEC device typically produces syndrome rounds at an approximately periodic cadence T_{cycle}, whereas the detection-event pattern in each round, and hence the computational work presented to the decoder, is stochastic. We therefore distinguish decoder service capacity, end-to-end response latency, and unfinished decoding work.

Modern superconducting QEC operates at approximately microsecond syndrome cadence, and real-time implementations explicitly distinguish the throughput needed to avoid backlog from the full response latency needed for logical branching [97], [109]. Dedicated accelerators increase the available decoding service through hardware parallelism [84], while AFS and LILLIPUT illustrate the latency–accuracy–resource trade space of hardware decoder design [110], [111]. Online decoding is particularly important when a logical operation depends on a decoded measurement result [97], [109]; related fault-tolerant photonic architectures are discussed in [112].

To formalize backlog without assuming Poisson arrivals, let $A_t \geq 0$ denote the decoding work generated by the syndrome and detection-event pattern of QEC round t, measured in implementation- appropriate work units (e.g., graph/message updates or equivalent accelerator service cycles). Let $S_t \geq 0$ denote the work that the deployed decoder can retire during one interval T_{cycle}, in the same units, and let $B_t \geq 0$ denote unfinished work carried into that interval. A round-synchronous workload recursion is

$$B_{t+1} = [B_t + A_t - S_t]^+, \qquad [x]^+ \equiv \max(x, 0). \tag{4}$$

Thus, rounds remain periodically clocked while their decoding workload is a stochastic batch. Under a stationary and ergodic approximation for the net-input process $A_t - S_t$, with finite first moment, the standard negative-drift condition

$$\mathbb{E}[A_t] < \mathbb{E}[S_t] \tag{5}$$

is the basic queue-stability requirement; if mean incoming work exceeds mean service, backlog cannot remain bounded, while the critical equality case requires additional assumptions [113]. Equation (5) is an average stability condition, not a hard real-time guarantee.

For operations that depend on a decoded result, a separate response deadline is required. We write the engineering target as

$$P_{\text{miss}} = \Pr\{\tau_t^{\text{resp}} > T_{\text{dl}}\} \leq \varepsilon_{\text{dl}}, \tag{6}$$

IEEE Proof

where τ_t^{resp} includes queue waiting, decoding service, and the required communication/control latency, and T_{dl} is the logical-dependency deadline. Consequently, a stable queue can still have unacceptable latency tails. The familiar rule $\tau_{\text{dec}} \leq T_{\text{cycle}}$ is a sufficient special case when one serial decode must finish before the next round; it is not a general stability condition for pipelined or parallel streaming decoders. Parallel-window decoding can sustain the syndrome stream while retaining a multi-round response latency [109], and recent integrated superconducting experiments report sub-μ s mean decoding time per round together with a multi-μ s full feedback response [97].

Correlated or drifting hardware noise enters this model primarily through the workload process A_t, rather than by making the nominal QEC clock Poisson. A minimal statistical description is the time-dependent mean $\mu_A(t) = \mathbb{E}[A_t]$ and temporal covariance $\gamma_A(t, \ell) = \text{Cov}(A_t, A_{t+\ell})$. Changes in $\mu_A(t)$ represent drift in typical load, whereas temporal correlation or bursts can produce consecutive high-work rounds. Experiments show that increased defect rates can increase decoder execution time [97], and recent QEC systems studies report data-dependent decoding latency and bursty, non-stationary decoder demand [114]. Such excursions can increase B_t and P_{miss} even when a long-run stationary approximation satisfies (5), motivating service headroom, buffering, and tail-latency analysis rather than a mean-latency criterion alone. Backlog does not change the algebraic code distance d; operationally, it enlarges the unresolved space–time decoding history and can force stalls, delayed logical decisions, or window truncation, thereby increasing logical-failure risk or logical clock time [109], [114].

Placement is also power-limited. As discussed in Section III-E, cryogenic control and decoding must satisfy tier-specific power constraints. For a stated decoder workload and placement, the per-round decoding energy is

$$E_{round} = \int_0^{T_{budget}} P_{decode}(t)\, dt,$$

subject to

$$\sum P_{tier} \leq P_{cap}(\text{tier}).$$

Because published implementations do not consistently report matched power, timing, and syndrome-volume quantities, Table 10 reports normalized decoder energy only when it is derivable from same-scope source data. Heavier learning or correlation tracking should remain warm unless the deployed cold path satisfies its measured latency and power budget.

We refer to these joint workload, service, response-latency, and power limits as the real-time envelope of Section IV-A and use it throughout the paper. Within this envelope, real-time decoding is a stochastic streaming pipeline: it must provide sufficient service headroom to keep the backlog B_t controlled, provision buffering and parallelism for workload bursts, satisfy the response-deadline target in (6) when logical feed-forward is required, and keep cold-tier inference within the available per-round energy budget. Fig. 17 summarizes these deadline, backlog, and cryo-power constraints for MHz-class online decoding.

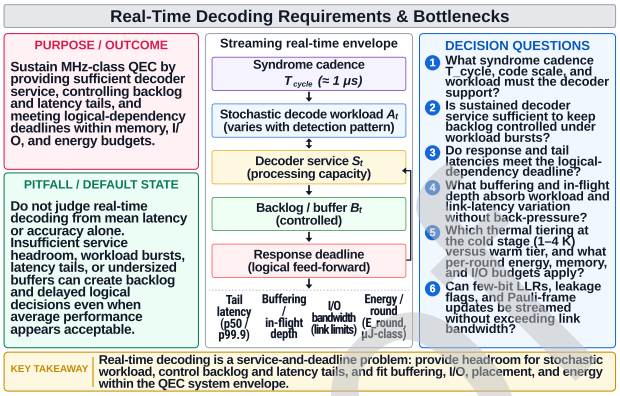


FIGURE 17. Real-time decoding checklist: cycle-time and distance; tail-latency/backlog limits; jitter-absorbing buffers; cryo–warm partitioning with μ J/round caps and few-bit log-likelihood ratios (LLRs) and flags; pitfall (accuracy-first backlog) and objective (deadline-safe, μ J-class, on-time Pauli frame).

B. MINIMUM-WEIGHT PERFECT MATCHING DECODERS

In this subsection, we describe MWPM decoders, which cast decoding as a weighted matching problem and remain the accuracy baseline for surface codes.

Each detection event (defect) is a node in a space–time graph; error likelihoods weight edges, and a minimum-weight perfect matching pairs all defects (or connects them to boundaries), yielding the most likely error configuration. Under both phenomenological noise (abstract X/Z-flip and measurement faults) and circuit-level noise (faults on all gates, idles, and measurements), Blossom-family solvers achieve among the highest known thresholds, which is why early experiments adopted MWPM as the de facto baseline. The drawback is computational: graph updates and memory traffic scale poorly with distance d, and naïve software struggles to sustain MHz-class rounds [84], [115].

Recent work pushes MWPM closer to real-time along two main tracks—algorithmic and hardware—summarized in Fig. 18.

Algorithmic speedups: At low physical error rates, the syndrome graph is sparse. Astrea exploits this sparsity to solve MWPM exactly for $d \approx 7$ with ∼1 ns average and 456 ns worst-case latency on FPGA-class logic; the greedy Astrea-G variant prunes unlikely pairings, remaining near-optimal while holding sub-μ s latency for $d \approx 9$ [75]. Follow-on work extends these ideas to larger graphs [58]. In parallel, Fusion Blossom fuses many partial matchings concurrently, sustaining $\approx 10^6$ rounds/s (≈ 1 μ s/round) for distances in the low-30s around 10^{-3} physical error;

IEEE Proof

streaming variants further reduce queueing delay on many-core hosts [116].

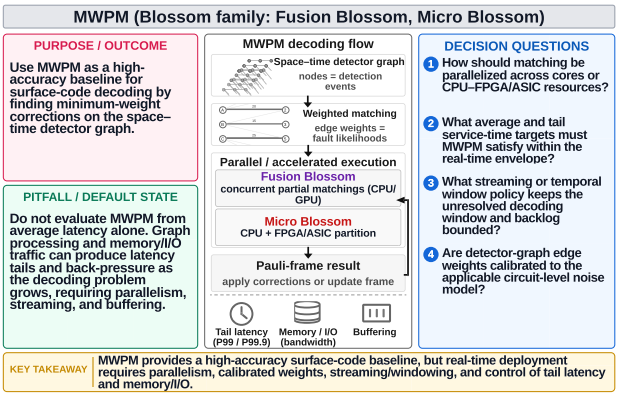


FIGURE 18. MWPM design checklist—decide parallelism, windowing, calibrated circuit-level weights, and service-time policy; pitfall is latency tails/back-pressure; purpose is near-ML matching on the space–time detector graph.

Hardware acceleration: Dedicated logic narrows the latency gap by exploiting fine-grain parallelism. Micro Blossom partitions Blossom between a host CPU and an FPGA/ASIC accelerator, instantiating processing elements for graph nodes and edges. This preserves MWPM optimality while achieving ≈ 0.8 μ s average decoding latency at $d = 13$ on FPGA hardware, at the cost of additional bandwidth, area, and design complexity [84], [97].

In head-to-head comparisons, MWPM remains a top-accuracy reference across many noise models: learned decoders can outperform it on device-calibrated data with crosstalk and leakage, whereas under cleaner simulated noise, MWPM typically matches or exceeds ML performance [97], [117].

The main practical risks are tail latency and memory/I/O bandwidth as d increases—without careful streaming and buffering, rare worst-case rounds can enlarge the unresolved space–time decoding window and raise the probability of buffer or response-deadline violations.

In summary, MWPM offers near-maximum-likelihood accuracy but requires careful algorithmic and hardware acceleration to meet sub-μ s deadlines and keep tail latency within the real-time envelope.

C. UNION-FIND AND RELATED PARTIAL-WEIGHT DECODERS

In this subsection, we describe UF decoders, which greedily grow and merge clusters of detection events to achieve near-linear-time decoding suitable for real-time operation.

UF decoders expand clusters around nearby detection events and merge them using disjoint-set unions; once a cluster touches a boundary or closes a path, the corresponding error chain is applied. This yields almost linear complexity in the number of defects (up to an inverse-Ackermann factor), maps naturally to parallel hardware, and makes UF highly attractive for low-latency decoding [115], [118], [119], [120]. Under phenomenological noise, UF approaches MWPM thresholds; under circuit-level noise, it typically reaches $\approx$60–70% of the MWPM threshold while still exhibiting exponential distance scaling, albeit with a larger logical-error prefactor.

In software, UF is often $\approx 10\times$ faster than MWPM on multicore hosts while maintaining competitive logical error rates at moderate distances. Distributed and parallel UF implementations achieve sub-μ s round latencies for sizable surface codes, with simplifications that keep worst-case time linear and reduce memory pressure so 99.9th-percentile latency remains bounded at MHz cadence [119]. Bias-aware variants such as UIUF (Union-Intersection UF) and IRUF (iterative UF) incorporate noise bias and X/Z correlations; UIUF can provably correct up to about half the code distance in certain biased regimes without losing linear-time behavior [115]. UF-style clustering also extends beyond the surface code to QLDPC on hyperbolic/expander graphs and to erasure settings with known fault locations, where breadth-first traversal variants simplify control and memory access for RTL implementations [120].

UF's operations—local growth, union, and path compression—map cleanly to nearest-neighbor processor lattices and streaming FPGA/ASIC pipelines. Synthesized designs report deterministic, sub-μ s per-round latency with sublinear average time versus distance; even naïve UF retains linear worst-case scaling, keeping area modest and timing predictable [118]. The main limitation is a modest accuracy gap to optimal matching, especially under circuit-level or strongly correlated noise, but this can be narrowed with UIUF/IRUF, lightweight post-processing, or hybrids (e.g., ML scorer $\rightarrow$ UF decision, or UF pre-filter $\rightarrow$ OSD refinement) [115].

In summary (see Fig. 19), UF-style decoders trade a modest accuracy gap for deterministic, low-latency operation and hardware-friendly local processing, making them natural candidates for real-time decoding.

D. BEYOND MATCHING: BP, ML, AND OTHER DECODERS

In this subsection, we describe BP/OSD, machine-learning-based decoders, and related approaches that target QLDPC and hardware-calibrated noise beyond simple matching.

For QLDPC, plain BP is attractive because its message updates parallelize well. While provable linear-time decoders exist for specific good code families [121], for general QLDPC, short cycles and code degeneracy can cause oscillations and inconsistent marginals, so BP alone is usually not sufficient. BP-OSD addresses this by running BP to rank bit reliabilities and then searching a small ordered list of top-k candidates; in reported studies, this achieves

IEEE Proof

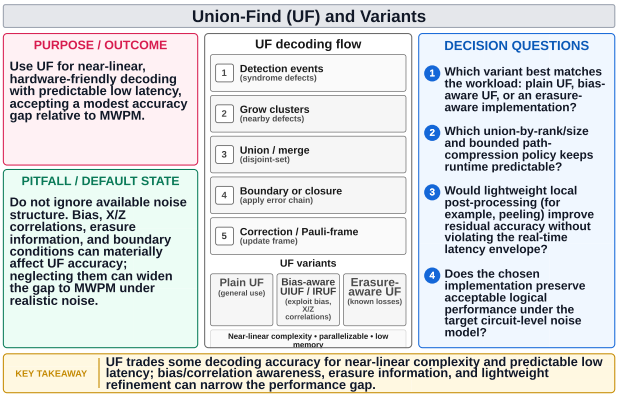


FIGURE 19. UF decoder checklist: choose variant (plain vs bias/erasure-aware UIUF/IRUF), enable union-by-rank with bounded path compression; optional local post-pass for residuals. Pitfall: ignoring bias/erasures widens the MWPM gap. Purpose: deterministic near-linear decoding with sub-μ s latency.

near-optimal performance on small LDPC instances such as bicycle codes under circuit-level noise and has also been exercised on surface codes, with k giving a tunable worst-case bound [11], [122], [123]. Several works combine BP with peeling-style ideas for hypergraph products and related qLDPC codes. At the same time, heuristics such as BP+OTF and ADOSD achieve similar complexity reductions but incur noticeable trade-offs in thresholds or accuracy, especially at a constant rate [11]. Neural-BP variants instead learn damping schedules or edge weights to stabilize updates and cut iteration count; in the literature, they appear as practical refinements, not replacements, of BP-OSD [19], [124] (see Fig. 20).

ML decoders (see Fig. 21) learn hardware-specific correlations and can ingest soft readout values. On device-calibrated data that include crosstalk and leakage, GNN and RNN/Transformer decoders have outperformed an MWPM baseline at small distances (e.g., $d = 3, 5$); on cleaner simulated noise, they are typically comparable to MWPM [58], [97], [117], [125]. The main costs are training and re-tuning, potential distribution mismatch between simulated and live correlated/leaky noise, limited generalization across distances or hardware revisions, and tail latency or on-chip power overheads that complicate strict real-time guarantees [97], [117]. A common pattern is hybridization: use ML as a scorer, prior updater, or correlation tracker, while routing worst-case rounds to a deterministic UF or BP-OSD backstop to keep deadlines and energy budgets bounded.

A further limitation is interpretability and certifiability. The learned decision boundaries of neural decoders are generally difficult to interpret, so rare logical failures can be difficult to diagnose or attribute to specific syndrome features or hardware conditions. Strong average validation performance therefore does not by itself certify worst-case fault-tolerant behavior, particularly under distribution shift or previously unseen correlated, leakage, or burst-error patterns. Confidence calibration and out-of-distribution detection can help expose uncertain inputs and trigger conservative handling, but these mechanisms remain statistical safeguards rather than formal worst-case guarantees. For deployment, an ML decoder should therefore be paired with a deterministic UF, BP–OSD, or MWPM fallback path for low-confidence or out-of-distribution cases, preventing an uncertain learned decision from becoming an uncertified single point of failure.

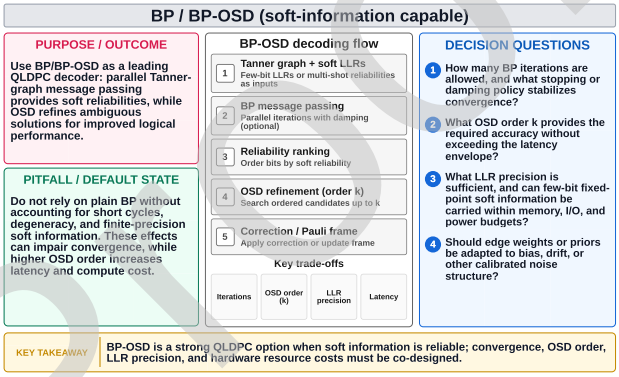


FIGURE 20. BP/BP-OSD checklist: update/OSD-k/LLR quantization; divergence and latency tails; workhorse for (q) LDPC/QLDPC.

Realistic devices further motivate correlation-, leakage-, and soft-information-aware decoding. Lightweight correlation-aware schemes re-weight edges using learned spatial/temporal graphs or temporal filters (e.g., HMMs), with parameter updates kept off the critical path to avoid extra jitter [126]. Leakage-tolerant pipelines treat flagged leakage as biased erasures, schedule periodic leakage-reset operations, and pass leak flags as side-band inputs so implicated checks are down-weighted [58], [97], [117], [125]. Soft information, few-bit LLRs or multi-shot reliabilities, measurably improves BP/BP-OSD thresholds and convergence; carrying LLRs across the cryo-to-warm link and using fixed-point LLR arithmetic in FPGA/ASIC data paths preserves these gains under tight cryogenic power budgets [19].

Other approaches round out the landscape. RG decoders provide multiscale, roughly linear-time heuristics; CA decoders deliver extremely fast local updates at the cost of lower thresholds; MaxSAT/ILP achieves exact maximum likelihood only for small instances and so is mainly a benchmarking tool; and lookup/decision-tree decoders remain practical for very small codes or specialized subroutines [127].

In summary, BP-OSD and soft-information variants are leading options for QLDPC, while ML-based and hybrid decoders capture correlations and leakage effectively when their latency and power costs are kept outside the per-round critical path.

IEEE Proof

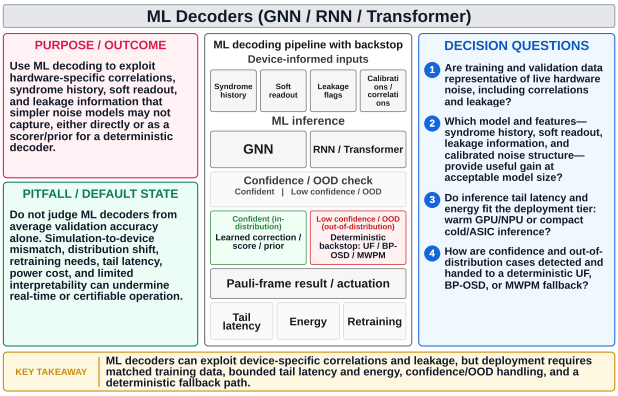


FIGURE 21. ML-decoder checklist: placement/model/features; nondeterministic tails and distribution shift; hybrid with deterministic backstop.

E. HARDWARE REALIZATIONS, PLACEMENT, AND POWER ENVELOPES

In this subsection, we map decoder algorithms to CPU/GPU, FPGA, ASIC, and cryogenic tiers, emphasizing latency and per-round energy constraints. We treat two design constraints from Sections III-E and IV-A: (i) average and tail latency (50th/99.9th percentiles at MHz cadence) and (ii) per-round energy E_{round} that must respect cryogenic power caps [19], [75], [84], [97], [116], [117].

FPGAs provide deterministic streaming and fixed-point datapaths. UF maps cleanly to strictly local grow/union/path-compression pipelines, while BP-OSD can be implemented using systolic message passing with a tunable OSD search. Pipelined MWPM implementations such as Fusion- and Micro-Blossom-style architectures trade additional bandwidth, area, and memory for deterministic acceleration. Compact ML can run cold only when its measured tail latency, memory footprint, and power satisfy the deployed budget; otherwise, ML is better placed in the warm tier with a deterministic decoding backstop [19], [84], [97].

ASIC implementations can further reduce jitter and specialize data movement for a fixed decoder workload. UF and BP-OSD fixed-point pipelines are therefore natural cold-tier candidates, while Micro-Blossom-style node/edge processing can accelerate MWPM at the cost of area and on-chip memory [19], [84], [97]. Because the cited implementations do not consistently report same-scope power, timing, and syndrome-volume measurements, no generic per-round energy class is assigned here; Table 10 reports E_{syn} only where the required source quantities permit an unambiguous derivation.

Below $\lesssim$100 mK, total budgets are $\lesssim$10–100 mW, so only very light logic fits. Cellular-automaton rules and LUT/decision-tree microcode for tiny tiles or flags are appropriate; heavier decoding should be moved up to the cryogenic stage and fed via bounded-tail streaming. Event-driven/neuromorphic fabrics are attractive for implementing local rules, but current toolchains and verification under cryogenic constraints are immature [97].

Regardless of silicon, the control plane must (i) stream syndromes plus few-bit LLRs from readout, (ii) pipeline Pauli-frame updates back to the scheduler, and (iii) provision buffers so worst-case decode never back-pressures measurement; link-latency variability must be hidden inside an in-flight window so that per-round deadlines are met [19], [97].

We refer to these joint workload, service, response-latency, and power limits as the real-time envelope of Section IV-A and use it throughout the paper. Within this envelope, real-time decoding is a stochastic streaming pipeline: it must provide sufficient service headroom to keep the backlog B_t controlled, provision buffering and parallelism for workload bursts, satisfy the response-deadline target in (6) when logical feed-forward is required, and keep cold-tier inference within the available per-round energy budget. Fig. 17 summarizes these deadline, backlog, and cryo-power constraints for MHz-class online decoding.

In summary, UF and BP-OSD are attractive deterministic candidates for cold-tier implementation, MWPM benefits from specialized acceleration, and ML placement should be determined by measured tail-latency, memory, and power constraints rather than by a generic per-round energy class.

F. BENCHMARKS AND OPEN PROBLEMS

In this subsection, we outline a minimal reporting set for decoder benchmarks so that results are comparable and deployment-relevant:

1) Logical error vs distance: report logical error rates as a function of code distance under circuit-level noise that includes spatial/temporal correlations and leakage.

2) Latency statistics and timing basis: report whether a value is response latency, inverse-throughput/service time, or batch latency; state the code distance d and, where defined by the source, the number of syndrome-measurement rounds m represented by the reported timing. Where this basis is explicit, we additionally report the distance-indexed descriptor

$$L_d = \frac{\tau}{md} = \frac{\tau_{\text{round}}}{d}, \qquad [\text{ns}/(\text{round} \cdot d)]. \tag{7}$$

Here, τ denotes the source-defined execution, service, or response duration represented by m syndrome-measurement rounds, and $\tau_{\text{round}} = \tau/m$ is the corresponding per-round descriptor. Because syndrome volume and decoder parallelism need not scale linearly with d, L_d is used only as a descriptive normalization and not as a technology- or algorithm-normalized efficiency metric.

3) Energy and placement: report source-defined power or energy together with implementation scope, evidence type, hardware tier, and workload. Energy per syndrome bit is derived only when the source provides workload-matched

IEEE Proof

quantities,

$$E_{\text{syn}} = \frac{P\tau}{N_{\text{syn}}}, \qquad [\text{pJ/bit}], \tag{8}$$

where P, τ, and N_{syn} must refer to the same implementation scope and operating mode. Otherwise the quantity is marked ND (not derivable from reported data).

4) Soft-information gain: quantify the improvement when decoders receive soft information (e.g., LLRs or analog reliabilities) instead of hard syndromes.

5) Robustness and retuning cost: describe robustness to calibration drift and model mismatch, including the cost of retraining or retuning ML-based decoders.

6) Throughput and data movement: report sustained throughput (rounds/s) at the target distance, along with link bandwidth and buffer depth assumptions on syndrome I/O.

Table 10 re-audits the representative decoder implementations against their primary sources and reports the minimum metadata needed to interpret their timing and power values. Reported values are distinguished from author-derived values; heterogeneous noise models, workload windows, platforms, and evidence types are made explicit. The L_d descriptor is populated only where the source provides an unambiguous round basis. Energy per syndrome bit is marked ND when workload-matched P, τ, and N_{syn} are not all reported.

Open problems that align directly with these benchmarks include: (i) constant-rate QLDPC decoding with guaranteed latency tails under correlated and leaky noise; (ii) leakage-tolerant pipelines (fast flags/resets) with bounded overhead and verified stability; (iii) correlation-aware inference that respects cryogenic energy budgets, favoring lightweight online reweighting over heavy retraining; (iv) ML decoders with certified latency/energy envelopes and out-of-distribution robustness, always paired with deterministic backstops; and (v) end-to-end co-design of readout, few-bit LLR transport, and fixed-point decoding at cryogenic stages under explicit data-motion budgets [19], [97], [117].

V. SCALABILITY AND RESOURCE CONSTRAINTS

Scaling QEC from tens to thousands of logical qubits turns a coding problem into a full-system design constraint. As code distance d grows, ancilla demand and per-round syndrome volume scale as $O(d^2)$, stressing cryogenic I/O and decoder buffers. MHz-class cycle times ($\sim$1 μ s) impose sub-μ s classical deadlines, and sub-100 mK power budgets cap how much decoding and memory can sit near the qubits. We therefore evaluate each technique through four lenses: (i) logical throughput and latency, (ii) qubit/ancilla overhead, (iii) interconnect bandwidth and routing depth, and (iv) classical memory and energy for decoding, always within the real-time envelope defined in Section IV-A.

Taken together, these constraints determine which codes and distances can actually be run as fault-tolerant QEC layers on hardware. In this section, we therefore treat scalability as a joint limit across the code, the quantum device, and the classical decoder, and ask: for a given target logical error rate, how far can we realistically push code distance and logical-qubit count?

Interoperability at this scale also requires stable semantics across the QEC toolchain. The common quantities in Table 2 and the provenance conventions in Table 10 provide the QEC-specific benchmark contract used here: logical-error, threshold, latency, throughput, and energy results remain tied to their code, noise model, workload, timing basis, and implementation context. Across compiler–controller–decoder–hardware boundaries, portability additionally requires explicit exchange of syndrome or detector data (including soft, erasure, or leakage information when used), round/window timing, decoder outputs or Pauli-frame state, and backend connectivity, measurement/reset, timing, and feed-forward capabilities. Existing specifications and standardization efforts cover parts of this path: OpenQASM 3.1 supports real-time classical control, explicit timing, and pulse/calibration descriptions; QIR 2.0 provides an LLVM-based compiler/backend representation with target capability profiles; IEEE P3980 is an active guide project for consistent and interoperable hybrid quantum–classical deployment; and ISO/IEC AWI 63712 is developing quantum-computing hardware benchmarking [131], [132], [133], [134]. None of these efforts by itself defines a complete QEC-specific decoder–controller interchange or benchmark suite, so the QEC definitions and provenance conventions above remain necessary for cross-platform comparison and integration.

A. MOTIVATION AND SCALING CHALLENGES IN QEC

As distance increases, $O(d^2)$ stabilizers and ancillas per logical qubit generate $O(d^2)$ syndrome bits per round, quickly saturating cryogenic I/O and decoder buffers. MHz-class cycles require decoders and control logic that meet strict per-round deadlines; if the classical service rate drops below the syndrome arrival rate, backlogs deepen the effective circuit and increase logical error. Unlike classical coding, fault-tolerant QEC feasibility is therefore set not just by asymptotic code parameters (distance, rate, threshold). It also depends on geometry, thermal management, routing congestion, and timing closure across the quantum–classical loop, which together bound the maximum code distance and number of logical qubits a system can practically support.

B. ANCILLA LOAD AND MEMORY PRESSURE

In this subsection, we discuss how ancilla qubits and syndrome memory limit scalable QEC and how current architectures mitigate these limits. Ancillas enable stabilizer measurement and syndrome extraction but dominate width and wiring: high-frequency measure–reset–reuse cycles raise thermal load, and $O(d^2)$ syndrome bits per round can saturate cryogenic I/O and on-chip buffers [135]. The ancilla schedule is thus directly coupled to the real-time envelope of Section IV-A: whenever the decoder's service rate falls below

IEEE Proof

TABLE 10. QEC decoder benchmarks with workload and provenance.

Decoder	Code / workload / noise	Platform / evidence	Source-defined timing	L_d [ns/(round·d)]	E_{syn} [pJ/bit]	Ref.
LILLIPUT (LUT)	Surface code; $d = 5, m = 2$	FPGA implementation	42 ns response (R)	4.2 (D)*	ND	[111]
Helios (UF)	Surface code; $d = 21$; phenomenological $p_{phys} = 0.1\%$	Xilinx VCU129 FPGA; implementation	11.5 ns/round average (R)	0.55 (D)	ND	[118]
Micro Blossom (MWPM)	Surface code; $d = 13$; circuit-level $p_{phys} = 0.1\%$	Versal FPGA prototype; measured at 62 MHz	0.8 µs average response (R)	NC	ND	[84]
WIT-Greedy	Surface code; $d = 11$; weighted errors	FPGA implementation	< 1 µs at $d = 11$ (R)	ND	ND	[128]
Astrea-G	Surface code; $d = 9$; multi-round decoding window	hardware accelerator implementation	~ 450 ns average (R)	ND	ND	[75]
Fusion Blossom (MWPM)	Rotated surface code; $d = 33$ batch / $d = 21$ stream; circuit-level $p_{phys} = 0.1\%$	64-core CPU; software benchmark	10^6 rounds/s batch; 0.7 ms average stream response (R)	NC	ND	[116]
AFS (UF)	Surface code; measurement-error-tolerant benchmark	hardware microarchitecture/model	42 ns average (R)	ND	ND	[110]
NISQ+ (AQEC)	Surface code; $d = 9$; pure-dephasing model	SFQ; synthesis-based	~ 20 ns maximum (R)	ND	ND	[129]
QULATIS	Surface code; $d \leq 11$; measurement-error/lattice-surgery support	hybrid SFQ/Cryo-CMOS; SPICE/simulation, 2 GHz	source-defined layer/service timing (R)	ND	ND	[130]

R = reported by the cited source; D = derived from source-reported quantities; NR = not reported; ND = not derivable from reported data; NC = not directly comparable. The distance-indexed timing descriptor is $L_d = \tau/(md) = \tau_{round}/d$, where m is the number of syndrome-measurement rounds represented by the source-defined timing. *Fixed-window normalization: arithmetic descriptor only; it should not be interpreted as a streaming response latency. Because syndrome workload and decoder parallelism do not generally scale linearly with d, L_d is not a technology-normalized algorithmic-efficiency metric. Energy per syndrome bit is $E_{\text{syn}} = P\tau/N_{\text{syn}}$ and is reported only when P, τ, and N_{syn} refer to the same implementation scope, operating mode, and workload. ND is used rather than estimating missing quantities. Cross-platform timing represents realized performance under the reported implementation conditions rather than technology-normalized algorithmic performance.

the incoming syndrome rate, queues form, effective round depth grows, and logical errors rise.

Reuse and Time-Multiplexing: Mid-circuit measurement and reuse share a limited ancilla pool over time. Compiler passes such as QOS's QubitReusePass recycle ancillas to shrink circuit width at the cost of additional depth [135]. Trapped-ion QCCD "racetrack" layouts pipeline Yb^+–Ba^+ ancilla chains with deterministic loading and micromotion-hiding resets, preserving coherence but letting transport consume ≈60% of circuit time. These schemes save qubits, yet they lengthen cycles and tighten decoder deadlines because more faults accumulate during idles and transport.

Syndrome Compression and Memory-aware Decoding: Syndrome-only estimation reconstructs detector graphs from histories, compressing data to O(1) parameters per locality and reducing on-chip RAM and thermal load. FPGA decoders such as Collision Clustering sustain sub-μ s latency while keeping backlogs within ≈10 μ s windows, respecting cryogenic power and memory limits [97], [98]. A related syndrome-only calibration procedure shows that fitting the decoder to learned, inhomogeneous error rates improves logical suppression in memory experiments [136].

Layout-, erasure-, and memory-aware extraction: Multi-chip compilers like QuIRC partition lattices into modules joined by entangled bridges, cutting ancilla-patch length by ≈77.8% and circuit depth by ≈51.9% [137]. Passes such as *QubitFreezing*, *GateCutting*, and *WireCutting* reshape the DAG to reduce simultaneous ancilla activation, trading compile time for lower routing congestion and buffer contention [138]. Dual-rail superconducting erasure qubits enable "ancilla-light" extraction: CP-gate checks or single-transmon entangling measurements halve ancilla demand and improve logical suppression by ≈4.2× under strong erasure bias [139], [140]. Decoding remains tractable because erasure registers scale as $O(d^2)$, but memory grows as $O(m\ d^2)$ with hyperedge order m, tightening the ancilla–decoder trade-off.

Stabilizer degeneracy and space–time correlations complicate belief-propagation heuristics, while ML decoders add training cost, dataset shift, and inference energy within the same memory and power budgets [97].

Open questions include cumulative decoherence under aggressive reuse and the feasibility of cryo-compatible SRAM caches (see Section V.C for repetition-cadence optimization and Section III.E for cryogenic memory constraints), all of which ultimately limit the code distances and logical-qubit counts that can be supported at target logical error rates.

C. SYNDROME REPETITION OPTIMIZATION

In this subsection, we discuss the co-design of syndrome repetition depth and decoding resources so that repeated measurements do not waste time or accumulate unnecessary idling errors. Repetition frequency interacts with measurement noise, ancilla reuse, and idling (Section V-B), and must still fit the real-time and power envelope of Section IV.

Adaptive Stopping Rules: Instead of fixed-depth schedules, adaptive schemes terminate once sufficient confidence is achieved. Let t be the repetition depth (number of extra syndrome rounds); for a distance-d code, use $t = \lfloor (d-1)/2 \rfloor$. Real-time analysis of syndrome difference vectors then allows early stopping, requiring on average $O(t)$ rounds while preserving the logical scaling

$$P_L \sim (p_{phys})^{t+1},$$

where P_L is the per-round logical error rate and p_{phys} is the per-operation physical error rate [141]. This reduces idling errors and thermal load but demands variable-depth decoders that can update graphs on the fly within sub-μ s windows.

Mid-Circuit Qubit Reuse: Mid-circuit measurement and reset can reduce the number of simultaneously required physical qubits by allowing measured qubits to be reused later in a circuit. DeCross et al. demonstrated this compilation principle by executing an 80-qubit QAOA circuit on a 20-qubit trapped-ion processor [142]. This is a general

IEEE Proof

circuit-resource result rather than a QEC-specific ancilla-compression benchmark; when applied to syndrome schedules, the corresponding reuse and reset latency must be included explicitly in the cadence budget.

Soft Information and Measurement-Free QEC: Soft-information decoding uses analog measurement reliability rather than only hard syndrome bits and can improve error suppression at fixed code distance [124]. Measurement-free QEC provides a distinct route in platforms with unconditional reset or fresh ancillas by replacing measurement and conditional feed-forward with coherent fault-tolerant primitives [143].

Table 11 summarizes these distinct strategies and the resources they trade. Heavy-hex experiments show that error rates are temporally biased and correlated [140], reinforcing that cadence and repetition should be treated as a first-class co-design axis rather than a fixed "d rounds" default.

Remaining challenges include adaptive pacing that reacts to fluctuating error rates, lightweight hardware–software interfaces to stream diagnostics into controllers, and syndrome-scheduling abstractions that expose timing and ancilla constraints directly to compilers.

D. CONSTANT OVERHEAD CODES AND QUANTUM LDPCS

In this subsection, we examine high-rate quantum LDPC (QLDPC) codes. For an $[[n, k, d]]$ family, constant code rate means $R = k/n = \Omega(1)$, equivalently a constant code-level data-qubit overhead $n/k = O(1)$. Throughout this subsection, "constant overhead" refers only to this code-level quantity; it does not include check ancillas, routing ancillas, long-range couplers, or interconnect layers. Good QLDPC constructions can combine constant rate and growing distance while retaining bounded stabilizer weight and bounded participation degree [144]. This LDPC sparsity is a graph-theoretic property and does not imply bounded-range geometric interactions on a two-dimensional processor.

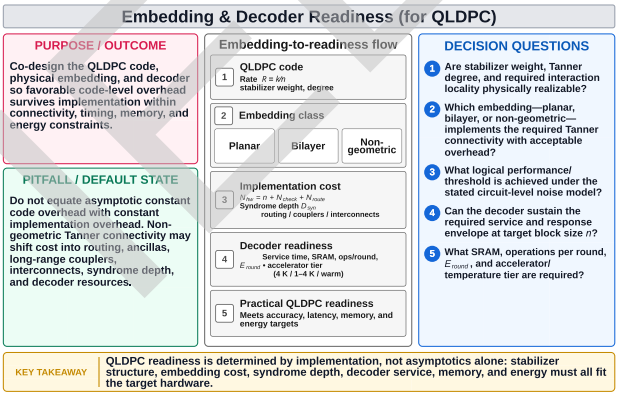


FIGURE 22. Embedding & decoder readiness for QLDPC: stabilizer weight/locality, embedding class, circuit-level threshold, service time at n, and SRAM/ops/Eround by tier.

Code-level versus implementation overhead: Let G_T denote the Tanner graph connecting data and check qubits and let G_H denote the native hardware interaction graph. A Tanner edge that is absent from G_H must be realized by routing, movement, teleportation, or an added nonlocal coupler. We use D_{syn} for the resulting physical two-qubit-layer depth of one syndrome-extraction schedule and

$$N_{\text{hw}} = n + N_{\text{check}} + N_{\text{route}}$$

for architectures in which non-native interactions consume check or routing ancillas. Thus $n/k = 1/R$ measures code-level overhead, whereas N_{hw}/k measures the qubit footprint of a specified implementation; native long-range couplers introduce additional interconnect costs not included in N_{hw}.

There is no universal QLDPC routing-depth law because D_{syn} depends on G_H and on the available nonlocal gate primitive. There is, however, a rigorous geometric obstruction. In the two-dimensional interaction model of Baspin and Krishna, a QLDPC family with $k = \Theta(n)$ and $d = \Theta(n^{\alpha})$ requires $\widetilde{\Omega}(n)$ interactions whose geometric length is $\widetilde{\Omega}(n^{\alpha/2})$, where the tilde suppresses polylogarithmic factors [145]. This is a lower bound on required nonlocal connectivity, not on a particular SWAP schedule: a strictly local device must synthesize those interactions, whereas hardware with native long-range links can trade routing depth for coupler, fabrication, or modular-interconnect complexity.

Code Families: Quantum Tanner codes, derived from left–right Cayley complexes, form asymptotically good QLDPC families with $[[n, \Theta(n), \Theta(n)]]$ parameters [14]. Their constant-weight checks support single-shot decoding [15], but their expanding connectivity is not geometrically local. In particular, the $\alpha = 1$ specialization of the bound above implies $\widetilde{\Omega}(n)$ interactions of geometric length $\widetilde{\Omega}(\sqrt{n})$ under the same two-dimensional embedding assumptions. Recent explicit multilayer superconducting layouts find finite Tanner-code instances with stabilizer weights between 6 and 12 and hardware complexity near the upper end of the studied QLDPC families at comparable logical efficiency [16]. This establishes a substantial connectivity requirement, but not a universal Tanner-code SWAP depth.

Bivariate bicycle (BB) codes have a different and more structured implementation profile. Their Tanner graph has degree six and thickness two, with weight-6 stabilizers [10]. If the required Tanner edges are supplied natively by the proposed two-layer long-range-coupler fabric, syndrome extraction uses seven CNOT interaction layers and n check ancillas [10]. Under the stricter bilayer architecture of Berthusen et al., where two-qubit gates are nearest-neighbor within each layer and nonlocal CNOTs are implemented by teleportation routing, a hand-designed $[[144, 12, 12]]$ embedding requires 12 and 8 routing rounds for its short- and long-range generator sets, respectively, and each routing round has physical depth 11 [146]. That architecture uses $4n = 576$ physical qubits for the $[[144, 12, 12]]$ block, compared with $2n = 288$ data-plus-check qubits when its Tanner connectivity is native [10], [146]. Thus the BB embedding penalty can appear as hundreds of physical circuit layers on strictly local gates, or instead as long-range couplers

IEEE Proof

TABLE 11. Syndrome repetition and resource-reduction strategies.

Strategy	Resource Saved	Main Trade-off	Decoder Implication	Hardware Notes	Evidence / maturity	References
Adaptive stopping	Syndrome rounds and idling	Variable stopping depth and online confidence logic	Decoder must accept a variable number of syndrome rounds	Designed for fault-tolerant syndrome extraction	Theory / Simulation	[141]
Mid-circuit qubit reuse	Simultaneously required physical qubits	Adds measurement/reset and scheduling constraints	If used in QEC schedules, reuse timing must be represented explicitly	General circuit-compilation result; an 80-qubit QAOA circuit was executed on a 20-qubit device using qubit reuse	Experiment (general circuit resource; not QEC-specific)	[142]
Soft-information decoding	Code-distance or measurement-overhead pressure at a target logical error	Requires analog reliability information and soft-decision processing	Decoder consumes soft reliabilities rather than only hard syndromes	Numerically evaluated across surface- and qLDPC-code settings	Simulation	[124]
Measurement-free QEC	Mid-circuit measurement and conditional feed-forward	Requires coherent fault-tolerant primitives plus reset or fresh ancillas	Removes measurement-conditioned decoding from the immediate feedback path	Developed and optimized for neutral-atom-compatible settings	Theory / Protocol design	[143]

and additional wiring layers. Hypergraph-product (HGP) codes support partial-syndrome measurement with small-set-flip decoders that iteratively reduce syndrome weight while maintaining exponential error suppression [147]. Spatially coupled HGP (SC-HGP) codes embed a 2D convolutional structure to suppress short cycles that hinder BP and reach thresholds of 7–9% with rates up to $k/n \approx 0.34$ at block lengths $n \sim 5,000$–$7,000$ [148]. Alternative hardware can shift rather than eliminate this connectivity cost: long-range-connected neutral-atom registers can supply selected nonlocal QLDPC interactions natively [149], whereas modular BB architectures move part of the connectivity burden into inter-module links and compilation [150].

Decoders for QLDPC: Advances in decoders underpin the practical performance of QLDPC codes. Neural belief propagation (NBP) reduces decoding iterations to a few rounds while improving frame error rates. Localized statistics decoding (LSD) dynamically clusters flips with polylogarithmic depth [151]. Graph-neural-network (GNN) decoders implement attention-based message passing in $O(n)$ inference time and can rival or surpass BP+OSD. These approaches mitigate degeneracy and convergence failures of classical BP, but their hardware viability still depends on meeting the latency and energy budgets of Section IV.

Bias-tailored and specialized constructions: Lifted-product codes tuned for biased noise exploit error asymmetries to improve distance and reduce overhead in infinite-bias limits [152], illustrating that code-family choice must match hardware noise characteristics, not only asymptotic scaling. A complementary strategy is to sacrifice some of the ideal asymptotic rate in exchange for strict geometric locality. Hierarchical memories concatenate an outer constant-rate QLDPC code with surface-code patches and admit two-dimensional nearest-neighbor syndrome extraction [153]. For N physical qubits in the resulting hierarchical family, the construction encodes $K = \Omega(N/\log^2 N)$ logical qubits and has a linear-width syndrome circuit with

$$D_{\mathrm{syn}} = O\left(\frac{\sqrt{N}}{r_g}\right),$$

where r_g is the available physical SWAP/interaction range [153]. This is an architecture-specific achievability result, not a universal QLDPC depth law, and illustrates explicitly how geometric locality converts part of the rate advantage into time overhead.

Despite their favorable code-level rate, QLDPC implementation costs cannot be summarized by Tanner-graph non-planarity alone. On strictly local two-dimensional hardware, non-native Tanner edges can increase D_{syn} and N_{hw}; on hardware that supplies those edges natively, the cost instead appears in coupler range, routing tiers, calibration, or modular interconnects. Consequently, $R = \Omega(1)$ by itself establishes neither constant physical syndrome-cycle depth nor a complete constant-overhead hardware architecture.

Open challenges include: (i) sub-µs LDPC decoding under realistic correlated noise; (ii) synthesizing planar embeddings of Tanner/SC-HGP codes into modular layouts; (iii) compiler integration of LDPC scheduling with ancilla routing (Section V); and (iv) noise-tailored codes co-optimized with bias-aware decoders. Because favorable code-level rate and qubit overhead can be offset by architecture-dependent connectivity, routing, syndrome-depth, and decoder-integration costs, the practical comparison with highly local surface-code implementations remains hardware dependent. Bridging theory and hardware therefore requires co-design across codes, compilers, connectivity, and cryogenic decoding hardware.

E. MODULAR ARCHITECTURES AND MULTICHIP SYSTEMS

In this subsection, we describe modular QEC architectures that partition processors into physically distinct units—chiplets, interposers, and 2.5D/3D stacks—because reticle limits, wiring congestion, mK heat removal, and control-line crosstalk cap monolithic scale. We compare three patterns—wired three-tier stacks, 2.5D cavity–memory hybrids, and teleportation-stitched modules—using a common rubric: interconnect density; access or seam latency and error; routing and synchronization cost; decoder coupling; and hardware maturity.

Three-Tier Stacks: In superconducting hardware, a three-stack organization separates the qubit die from a multilayer wiring and signal-processing die through an intermediate superconducting through-silicon-via (TSV) interposer. The interposer provides dense vertical access while physically separating qubits from the multilayer routing stack; experimental two-stack devices have demonstrated qubit control and microwave readout through superconducting TSVs and provide an enabling validation of the full three-stack architecture [154]. The advantage is dense vertical fan-out and functional separation between qubits and routing electronics; the cost moves into TSV and bump fabrication, microwave integrity, alignment, thermalization, and package yield.

IEEE Proof

2.5D Cavity–Memory Hybrids: A different approach reduces the number of actively controlled transmons by attaching resonant, multimode memories to processor qubits. In the Virtualized Logical Qubits architecture, surface-code data are distributed across cavity modes connected to a two-dimensional transmon array; simulations report approximately an order-of-magnitude reduction in transmon count and 6× faster co-located transversal CNOTs than the compared lattice-surgery CNOT, but with an approximately 10× serialization penalty because modes sharing a physical address cannot all be operated in parallel [155]. More recently, an experimental cascaded random-access quantum memory demonstrated arbitrary access to eight superconducting cavity modes with average access infidelity $\lesssim$ 1.5% per mode and support for transversal operations within the memory module [156]. This evidence increases the hardware maturity of the memory primitive, but does not yet constitute a complete fault-tolerant 2.5D QEC processor. The principal trade-off is therefore reduced active-qubit and control-line pressure versus memory-access serialization, cavity lifetime, and many-body or crosstalk errors that grow with memory complexity.

Teleportation-Stitched Modules: A third pattern leaves logical patches in separate modules and connects them through noisy quantum links and teleported inter-module gates. Analytical and numerical results for surface-code patches show that fault-tolerant operation can persist even when the connecting interface is substantially noisier than the module bulk, provided the seam is decoded as part of the protected logical operation [157]. This relaxes monolithic routing constraints but shifts cost into entanglement generation, link errors, synchronization, and asynchronous seam decoding. Toolchains such as SEQC and QuComm make chiplet placement, routing, and inter-node communication explicit compiler objectives [158], [159]. Decoder backplanes must ingest asynchronous seam updates without violating per-round deadlines (Sections IV and V-C).

In summary, the three modular patterns move the scaling bottleneck to different resources. Three-tier stacks trade planar wiring congestion for vertical-interconnect and packaging complexity; 2.5D cavity–memory hybrids trade active-transmon count and control wiring for memory-access serialization and cavity-memory error management; and teleportation stitching trades physical routing for entanglement, link-quality, synchronization, and decoder-seam requirements. The appropriate architecture is therefore determined by which resource is most constrained—wiring density, active control hardware, memory quality, or inter-module link performance—and by whether the resulting access or seam latency remains inside the real-time envelope of Section IV.

F. ROUTING AND INTERCONNECT CONSTRAINTS

In this subsection, we examine routing and interconnect constraints. Routing in QEC systems defines a boundary between abstract code efficiency and physical execution. For the QLDPC codes of Section V-D, the relevant object is the mapping of the Tanner graph G_T to the native hardware graph G_H. When a required Tanner edge is not native, it must be synthesized by SWAPs, teleportation, movement, or another remote-gate primitive. We therefore use the mapped syndrome depth D_{syn}, together with N_{hw}/k, as the primary implementation metrics rather than inferring execution cost from graph non-planarity alone. Additional two-qubit layers increase the number of circuit fault locations and extend the physical syndrome-cycle time; native nonlocal couplers can reduce that time penalty while increasing interconnect complexity.

Local routing and SWAP-based strategies: Generic limited-connectivity compilers such as SABRE optimize SWAP count and circuit depth on a coupling graph $G = (V, E)$. QKNOB provides circuits with known near-optimal transformation costs and benchmarks such algorithms on devices of up to 53 qubits [160]. These results show that practical routers can remain appreciably above the near-optimal SWAP-count or depth cost, but they do not establish a universal $O(n \log n)$ routing law and should not be extrapolated directly to large QLDPC syndrome circuits. For QLDPCs, the code- and architecture-specific D_{syn} results summarized in Section V-D provide the more relevant execution measure.

Physical movement and remote entanglement: In ion-trap and neutral-atom devices, physically moving ions or atoms between gate and storage regions can realize interactions absent from a fixed nearest-neighbor graph. In superconducting architectures, nonlocal QLDPC interactions can instead be synthesized through teleportation routing or supplied by multilayer long-range couplers. These choices have different resource signatures: the former can require routing ancillas and large D_{syn} [146], whereas the latter can preserve shallow interaction schedules at the cost of additional coupler tiers, long physical links, bump transitions, and TSVs [16]. Hence neither Tanner-graph distance nor non-planarity alone determines the physical execution depth.

Long-range fabrics and cryogenic constraints: Long-range connectivity adopts crossbar and bus designs with optical, superconducting, or microwave interposers. Nanophotonic switches and MEMS couplers offer Tb/s-scale optical links across chiplets, while TSVs and indium bump bonding yield vertical interconnects with resistances of $\sim$240 nΩ and high yield at $<$20 μ m pitch. Cryo-CMOS readout chains compress hundreds of analog lines into multiplexed outputs, consuming $<$10 μ W per qubit, and adiabatic logic circuits provide picowatt-scale control without large dynamic heat loads. These techniques increase effective connectivity but must respect cryogenic power and crosstalk limits.

Routing-aware compilation: Routing-aware compilers integrate interconnect constraints directly. Transpilers embed delay, signal fidelity, and wiring congestion into placement and scheduling, using gate fusion, calibration-driven fidelity weights, and integer/SAT-based mapping to trade SWAP count against latency and interconnect overhead. This marks

IEEE Proof

a shift from static nearest-neighbor assumptions toward joint routing–code co-design, where ancilla placement, syndrome extraction, and decoder alignment are all planned with wire-delay constraints in mind.

Taken together, these routing options trade different physical resources. SWAP-based routing preserves a fixed local hardware graph but can increase D_{syn}; teleportation or physical movement shifts cost toward routing ancillas, entanglement generation, transport, and synchronization; and native long-range couplers can preserve shallower interaction schedules while moving cost into coupler tiers, fabrication, calibration, and cryogenic integration. The relevant comparison is therefore mapped syndrome depth and implementation footprint, not graph non-planarity alone.

Remaining challenges include balancing TSV and resonator density against crosstalk and heat dissipation, building unified logical-to-physical compilers that handle ancilla placement and syndrome scheduling under realistic interconnect models, and exploiting new materials and reconfigurable fabrics—photonic buses, microwave interposers, fluxonium chains—so routing becomes a programmable resource rather than a bottleneck that limits achievable code distance and logical-qubit count.

VI. NOISE AND DECOHERENCE MANAGEMENT

This section surveys how realistic, structured noise shapes every layer of QEC. It classifies the dominant error mechanisms, describes how hardware-calibrated and ML-assisted models track drift, explains live decoding and feedback loops, and shows how full-stack co-design can turn noise constraints into architectural design inputs.

A. TYPES AND MODELS OF QUANTUM NOISE

This subsection applies the modeling hierarchy of Section II-C to the main noise families in today's hardware. IID Pauli channels remain a common phenomenological baseline for threshold estimates and Monte Carlo studies, but the Pauli representation and the IID factorization are separate assumptions. Depending on platform and operating point, coherent, biased, leakage, spatially or temporally correlated, and non-stationary mechanisms can require richer channel- or Hamiltonian-level descriptions before they are reduced to an effective QEC noise model.

Table 12 summarizes the noise families used throughout this survey and where they arise in hardware.

1) Noise families and where they arise

Slow calibration drift and coherent control errors can accumulate systematically rather than as independent stochastic flips, so their effect must be characterized and represented explicitly in calibrated noise models [161], [162], [163], [164].

Amplitude-damping and dephasing channels capture T_1 energy relaxation and T_2 phase decay, respectively [25], [161]. When the resulting effective Pauli channel is strongly Z-biased, bias-tailored constructions such as the XZZX surface code can exploit that asymmetry [13].

Leakage—population leaving the $|0\rangle/|1\rangle$ subspace, e.g., transmons populating $|2\rangle$—does not directly trigger syndromes and can spread faults, so experiments now deploy dedicated leakage-reduction units and auxiliary reset qubits to pull leaked qubits back each cycle [58], [67].

As systems scale, several physically distinct mechanisms generate correlated error structure: leakage can create correlations across space and time [165], crosstalk can couple faults across multiple qubits [70], and high-energy radiation can generate phonon-mediated quasiparticle bursts affecting multiple superconducting qubits [73]. Li et al. directly observed muon-induced quasiparticle bursts and associated correlated errors [73]. Separately, Google's long repetition-code experiments observed rare correlated events roughly once every 3×10^9 cycles ($\approx$ hourly), but the physical origin of that residual error floor was not identified [11]. Correlation is also not universally detrimental: particular symmetric multi-qubit correlation patterns can increase surface-code thresholds in specific models [166]. These results show why a single IID depolarizing fit can omit QEC-relevant structure and motivate explicit characterization of correlated error statistics [70], [73], [165], [167].

Operator-level representatives of structured noise

The distinction in Section II-C becomes concrete for the physical mechanisms emphasized here. First, for a flux-sensitive superconducting qubit, a small magnetic-flux fluctuation $\delta\Phi(t)$ produces a frequency fluctuation $\delta\omega_q(t) \simeq (\partial\omega_q/\partial\Phi)\delta\Phi(t)$, corresponding to the stochastic dephasing Hamiltonian $H_\Phi(t) = \hbar\delta\omega_q(t)Z/2$. Within a time window in which the second-order statistics are treated as stationary, define

$$C_\Phi(\Delta t) = \langle \delta\Phi(t)\delta\Phi(t+\Delta t)\rangle\,,$$
$$S_\Phi(\omega) = \int_{-\infty}^{\infty} C_\Phi(\Delta t)e^{-i\omega\Delta t}\,d\Delta t. \tag{9}$$

Measured low-frequency flux noise is commonly fit over the applicable experimental bandwidth by

$$S_\Phi(\omega) \propto |\omega|^{-\alpha}, \qquad \alpha \simeq 1, \tag{10}$$

with device- and frequency-dependent departures from an exact $1/f$ law [74]. The strong low-frequency weight produces slowly varying, temporally correlated detuning across successive operations; this temporal correlation should not by itself be equated with either coherent error or non-Markovianity [169].

Second, residual ZZ crosstalk is naturally an operator-level coherent interaction rather than an independent Pauli-flip process. We write

$$H_{ZZ}(t) = \hbar\,\zeta_{ij}(t)Z_iZ_j,$$
$$U_{ZZ}(t_{int}) = \exp\left[-iZ_iZ_j\int_0^{t_{int}}\zeta_{ij}(t)\,dt\right]. \tag{11}$$

where the coefficient convention is absorbed into ζ_{ij}. A static ζ_{ij} accumulates a coherent, state-dependent two-qubit phase; fluctuations of $\zeta_{ij}(t)$ can additionally introduce stochastic

IEEE Proof

TABLE 12. QEC noise families.

Family	Description	Key assumption	When used
Depolarizing	With some total probability, apply a random Pauli X, Y, or Z error.	Pauli components are assigned by the chosen depolarizing model; IID is an additional assumption when locations are factorized.	Baseline threshold and Monte Carlo studies [25].
Pauli stochastic	Apply I, X, Y, or Z with specified probabilities, possibly with strong bias.	The Pauli representation does not by itself imply independence; the joint distribution may be factorized or correlated.	Biased-noise surface and cat-code studies [13], [17], [36].
Amplitude damping	Excitations relax toward the ground state on a T_1 timescale.	The channel is nonunital and directionally biased in energy.	Relaxation-aware hardware and QEC noise modeling [25], [161].
Phase damping / dephasing	Relative phase coherence decays without the same population relaxation mechanism as amplitude damping.	Low-frequency fluctuations can make the dephasing temporally correlated rather than IID.	Dephasing-dominated and flux-noise regimes [25], [74].
Coherent control error	Systematic over-rotations or unwanted Hamiltonian couplings accumulate phase coherently.	Errors are unitary or contain a coherent component and therefore need not behave as stochastic Pauli faults.	Hardware-calibrated characterization and realistic noise modeling [161]–[164], [168].
Correlated (space/time)	Joint error statistics depend on multiple qubits and/or multiple cycles, including crosstalk, leakage-mediated correlations, and radiation bursts.	The joint process does not factorize over all affected locations or times.	Correlated-error studies on superconducting QEC hardware [70], [73], [165].
Leakage	Population leaves the intended computational subspace.	A leaked state is not represented by an ordinary two-level Pauli fault until an explicit effective reduction is justified.	Leakage detection, removal, and leakage-aware QEC [58], [67], [165].
Measurement / reset	The reported outcome or prepared/reset state is incorrect.	Faults occur in syndrome acquisition or state preparation and can be modeled separately from data-qubit faults.	Repeated syndrome extraction and real-time QEC [11], [97].

time dependence [168]. Surface-code simulations that explicitly retain coherent control and unwanted-interaction errors show why their logical effect need not be inferred correctly from a bare independent-Pauli rate without first validating an effective reduction [161].

Third, radiation-induced quasiparticle bursts are more naturally represented as an event-modulated, time-dependent error or relaxation rate than as a constant IID probability. A minimal representation is

$$\Gamma_i(t) = \Gamma_i^{(0)} + \sum_m a_{im} g_m(t - t_m), \tag{12}$$

where $\Gamma_i^{(0)}$ is the background rate, t_m is the time of burst m, a_{im} is its impact on qubit i, and g_m is its recovery envelope. Equation (12) assumes no particular probability law for the burst times. One event with nonzero a_{im} on many qubits creates spatial correlation, while a recovery envelope extending over several cycles creates temporally clustered errors; changing burst activity also makes the effective channel non-stationary [27], [73].

These examples illustrate the reduction

physical dynamics
↓
effective quantum channel
↓
QEC-level Pauli/syndrome model

Coarse graining, calibration, or Pauli randomization can make the final description efficient, but the validity of that reduction must be checked against the physical regime. Threshold estimates and decoder likelihoods therefore depend not only on an average physical error rate but also on which coherent, spatial, temporal, and time-varying structure is retained in the effective model [11], [161].

2) Simulation, Calibration, and Grey-Box Modeling

Efficient QEC studies often use circuit-level stochastic models for Pauli, erasure, and measurement faults, while amplitude damping, coherent rotations, and unwanted couplings require more general channel or Hamiltonian descriptions before any Pauli reduction is made. Hardware-calibrated parameters can be constrained through benchmarking, realistic-noise simulation, and calibration protocols [161], [162], [163], [164], [170], [171]. Rather than a single generic error rate, these models can assign operation-dependent error budgets and structured-noise parameters that feed QEC simulations and logical-performance predictions.

Where full process tomography is infeasible, grey-box system identification combines explicit physical structure with learned components; Youssry et al. experimentally demonstrated such an approach for quantum-device identification and control [172]. Machine-learning noise spectroscopy provides a related but distinct route for estimating structured and non-Markovian environmental parameters [173]. Temporally correlated low-frequency noise requires models that retain the relevant time structure [74], [169], while adaptive decoding-graph re-weighting can update decoder priors under drifted and correlated noise [174].

B. HARDWARE-CALIBRATED NOISE MODELING AND MACHINE-LEARNING INTEGRATION

This subsection explains how the hardware-calibrated, ML-driven noise models of VI-A are built, kept in sync with

IEEE Proof

drift, and wired into control, so that measured syndromes and calibration data drive both simulators and real-time pulse updates.

Hardware-calibrated noise modeling combines characterization, model-quality testing, and repeated calibration rather than relying on a single generic error probability. Benchmarking and characterization methods assess whether candidate noise models reproduce device behavior, while gate-calibration protocols update parameters as hardware drifts [170], [171], [175]. In surface-code experiments, component-resolved error budgets attribute logical-performance limitations to mechanisms such as gate errors, measurement/reset error, leakage, decoherence, and unwanted interactions [11], [46], [123].

For simulation, the central requirement is to retain the device-specific structure needed by the QEC question under study. Realistic surface-code models, emulator-adjusted noise models, and hardware-aware calibration protocols provide complementary ways to translate measured device behavior into simulation parameters [161], [162], [164].

Table 13 summarizes this calibration stack: model-quality benchmarking constrains the effective description, component-level error budgets identify dominant logical-error contributors, and hardware-calibrated simulation carries those parameters into QEC analysis.

ML is increasingly used to learn noise dynamics from data: neural networks and Gaussian processes predict temporal drift [176]; physics-inspired schemes reconstruct non-Markovian noise from randomized benchmarking and scale via tensor-network structure [177]; and regression and physics-informed models infer process matrices or parameterized dephasing laws from sparse measurements [172], [178].

ML also bridges engineered noise models and hardware: decoders or simulators are trained on simplified noise and then fine-tuned on device data, with transformer-based decoders adapted on crosstalk- and leakage-augmented datasets outperforming purely model-based baselines [117], [179]. This creates an ML-in-the-loop calibration where the processor and classical models co-learn the evolving noise landscape.

Controllers increasingly perform real-time noise adaptation: syndrome patterns indicating, for example, systematic X over-rotation trigger pulse-bias corrections [17], [180], [181]; trapped-ion experiments use slow feedback to tune laser intensities mid-run, while superconducting firmware monitors leakage signatures and triggers targeted resets [173]. On processors such as Quantinuum H2-1 these loops keep over-rotation drift below 3×10^{-3} radians and sustain $\approx$0.95% logical error for distance-5 codes, enabled by FPGA-based control planes that react on microsecond scales [182], [183], [184], [185], [186], [187].

Together, calibrated models, randomized compiling, dynamical decoupling, grey-box identification, and ML-based forecasting turn the VI-A taxonomy into a control stack where learned spectra and correlations directly drive concrete control actions [179].

C. REAL-TIME ERROR CORRECTION, FEEDBACK, AND GREY-BOX SYSTEM IDENTIFICATION

Building on Section IV's real-time envelope, this subsection explains how decoders are embedded directly into continuous quantum–classical control loops, so sustained service, bounded backlog, and response deadlines become system requirements [11], [97], [98]. Early experiments could treat decoding as an offline diagnostic, but as logical lifetimes approach minutes the no-backlog constraints of Section IV force online decoding for mid-circuit operations such as teleportation-based gates, lattice surgery, and magic-state injection [112].

Decoded syndromes now steer logical control flow. IBM's encoded magic-state preparation uses mid-circuit measurements and feedforward to discard faulty attempts and boost success probability of high-fidelity magic states; similar patterns appear in lattice surgery and non-Pauli gate injection, where decoded outcomes directly gate the next logical operation [97], [112].

The accelerators of Sections IV-C–E—Union-Find and MWPM pipelines on FPGA/ASIC and GPU-accelerated decoders—now act as inline controllers, not just offline analyzers. Distributed UF designs reach O(10 ns) per round at distance $d \approx 20$–21 on FPGAs, while custom ASIC and GPU pipelines for MWPM- and BP-class decoders handle high-rate syndrome streams within tight tail-latency and cryogenic-power budgets [118], [167]. Recent collision-clustering ASIC prototypes achieve $\sim$240 ns decode times for $\approx 10^3$-qubit surface-code patches, demonstrating that hardware decoders can keep pace with MHz-scale QEC cycles [84], [98].

Beyond these algorithmic baselines, ML decoders are increasingly treated as data-driven controllers tuned to hardware-specific noise. Message-passing GNN decoders for qLDPC codes and sliding-window schemes achieve strong performance at low complexity [126], [188], while GNN- and transformer-based decoders trained on hardware-like, circuit-level noise (including correlations and bias) can outperform standard matching decoders on surface codes by learning patterns that independent-error models miss [167], [189].

Table 14 illustrates two complementary uses of correlation information: learned surface-code decoding that exploits device-specific syndrome and measurement structure, and joint logical-circuit decoding that models error propagation across interacting logical blocks.

These examples illustrate distinct uses of correlation information in QEC. Learned decoders can exploit device-specific syndrome, readout, leakage, and crosstalk structure, whereas correlated logical-circuit decoders explicitly model error propagation between interacting logical blocks (cf. Section IV-D).

IEEE Proof

TABLE 13. Calibration stack.

Layer	Example Technique	Role in the model
Model-quality benchmarking and calibration	Noise-model quality assessment, characterization benchmarking, and multi-qubit gate calibration [170], [171], [175]	Tests whether an effective model matches device behavior and updates parameters as calibration changes.
Component-level QEC error budgets	Component-resolved gate, measurement/reset, decoherence, leakage, and unwanted-interaction budgets in surface-code experiments [11], [46], [123]	Attributes logical-performance limitations to specific physical mechanisms and constrains simulator parameters.
Hardware-calibrated simulation	Realistic surface-code models, emulator-adjusted noise models, and hardware-aware calibration protocols [161], [162], [164]	Maps measured or calibrated device behavior into structured noise models used for QEC simulation and design.

TABLE 14. Decoders managing noise correlations.

Controller	Role in the loop	Result	Evidence / maturity	Ref.
Recurrent transformer decoder (Google DeepMind–Google Quantum AI)	Surface-code decoder using syndrome histories, soft measurement information, and leakage information; approximate synthetic-noise training is followed by adaptation using a limited set of experimental samples.	Outperforms other state-of-the-art decoder baselines on Sycamore distance-3 and distance-5 data and retains an advantage in realistic simulations through distance 11.	Experiment + Simulation	[117]
Correlated logical-circuit decoding	Jointly decodes interacting logical blocks by representing deterministic error propagation through transversal entangling gates in a decoding hypergraph; the study compares an exact maximum-likelihood decoder with the efficient belief-HUF approximation.	Improves decoding of transversal Clifford and non-Clifford gates and, in simulated deep Clifford circuits, permits substantially fewer syndrome rounds between entangling gates while reducing overall space–time cost.	Theory / Simulation	[190]

Beyond decoding, physical-layer characterization and control can act upstream of QEC by reducing or reshaping the noise presented to syndrome extraction. Circuit-aware dynamical-decoupling methods provide one such mechanism. For example, GraphDD performs real-time, circuit-specific embedding of dynamical-decoupling sequences into executable circuits and is designed to refocus quasistatic dephasing and crosstalk errors during idle periods while minimizing additional single-qubit control operations [191]. This type of error suppression can lower the physical error burden presented to QEC, although it does not in general remove the fixed syndrome-extraction cadence or its associated classical data path.

Grey-box system identification provides a complementary capability by combining explicit physical structure with learned components. Youssry et al. experimentally demonstrated a grey-box approach that reconstructs otherwise inaccessible Hamiltonian and unitary information from quantum-device data and uses the identified model to design control [172]. Related machine-learning noise-spectroscopy methods infer parameters of structured and non-Markovian environments; for example, Scarpetta et al. use learned regressors to estimate system–bath coupling and spectral-density properties in two-level non-Markovian models [173]. These characterization methods can therefore inform calibration, pulse design, and noise-aware control, although they should be distinguished from an experimentally demonstrated closed-loop QEC decoder.

Together with the architectures and power/latency budgets of Section IV, these developments motivate a layered real-time error-management architecture. Accelerator-backed decoders operate on syndrome streams to produce timely logical decisions; learned and correlated decoders exploit device-specific or cross-logical error structure; and physical-layer characterization and control act to suppress, reshape, or identify noise before it propagates into logical failure. QEC therefore increasingly interacts with active characterization and control rather than functioning only as a passive, post-hoc correction layer.

D. IMPLICATIONS OF STRUCTURED NOISE ON QEC CODES AND ARCHITECTURE

This subsection argues that structured, platform-specific noise now drives full-stack co-design: codes, compilers, control hardware, and decoders are chosen to exploit or mitigate measured bias, leakage, and correlations. Table 15 highlights three active thrusts that align decoders and compilers with measured spatiotemporal correlations.

1) NOISE-MATCHED CODES

IBM's heavy-hexagon code tailors degree-3 connectivity to fixed-frequency transmons, reducing ZZ crosstalk and frequency crowding at the expense of some code capacity; combining surface-code X-stabilizers with Bacon–Shor Z-stabilizers and flag-qubit decoding has achieved logical break-even and beyond on 127- and 433-qubit devices [47], [194]. Bias-tailored codes such as the XZZX surface code reach higher thresholds when noise is dominated by a single Pauli (e.g., dephasing) by orienting checks and choosing stabilizer weights to preferentially correct the dominant error [152], [195]. Similar co-design appears in bosonic codes, which exploit predominantly loss/erasure noise, and in proposed Majorana-based codes whose topological qubits exhibit different error symmetries.

2) NOISE-AWARE COMPILERS AND SCHEDULES

Compilers and schedulers are becoming noise-aware [196]: calibration-driven mapping assigns critical logical qubits to physical qubits with the best T_1 and lowest error rates; gate-selection passes choose between, for example, direct versus

IEEE Proof

TABLE 15. Exploratory thrusts in managing correlated and bursty noise.

Research thrust	Key technique and evidence	References
Temporal-error forecasting and data–syndrome decoding	Time-series models can forecast error patterns in repeated single-qubit circuit outcomes, while generalized quantum data–syndrome codes incorporate data and syndrome errors jointly in belief-propagation decoding. These are complementary results rather than a single demonstrated closed-loop QEC controller.	[180] ,[192]
Context-aware suppression of correlated noise	Hardware-aware compilation identifies spatial and temporal contexts in which coherent correlated errors arise and applies dynamical decoupling or error compensation. The reported experiments show an 18.5% increase in layer fidelity for a candidate 10-qubit circuit layer relative to context-unaware suppression.	[71]
Burst detection and architectural containment	Radiation-induced phonon/quasiparticle dynamics can generate spatiotemporally correlated errors; direct measurements identify cosmic-ray-induced correlated events, while modular processor studies show that physical separation can substantially reduce correlations between modules.	[72], [73],[193]

echoed CZ implementations based on detailed error profiles to reduce crosstalk [197]; routing steers SWAP paths around "hot" regions with higher error density even at the cost of extra hops [198]; and dynamical scheduling interleaves gates with dynamical-decoupling pulses on idles so neighbors' operations do not leave them unprotected [71], [199].

3) HARDWARE-LEVEL SUPPRESSION

At the hardware-control layer, structured noise motivates explicit suppression schemes: on Google's 105-qubit Willow processor, a data-qubit leakage-removal (DQLR) protocol using a specialized Leakage-iSWAP to move $|2\rangle$ population from data to ancilla, followed by fast measurement-induced reset each cycle, reduces data-qubit leakage from $\sim$1% to $\lesssim$0.1% and cuts steady-state leakage by an order of magnitude, improving the effective error rates seen by the code [58], [67]. Tunable-coupler architectures similarly isolate qubits when idle to suppress coherent ZZ interactions at the cost of more complex calibration; recent QEC experiments run couplers "off" except during intentional two-qubit gates and use resonator-bus or resonator-induced-phase designs to either engineer long-range connectivity or suppress stray couplings that would otherwise generate correlated errors [200], [201], [202], [203], [204].

4) CROSS-STACK CO-DESIGN AND CORRELATED EVENTS

Code choice feeds back into classical architecture: codes with very high syndrome density or global measurements (as in some LDPC families) can overwhelm available decoder networks, while the connectivity graph constrains which codes can be implemented with reasonable overhead [97], [100], [205], [206]. Programs such as DARPA's US2QC therefore evaluate end-to-end stacks, with some approaches deliberately selecting qubits with slightly higher physical error rates but advantageous noise biases that specific codes can exploit [17]. FPGA-in-the-loop decoders, such as Riverlane's cryo-proximate units, place dedicated decoding hardware physically close to the processor so distance-50 surface codes producing tens of thousands of syndrome bits per microsecond can be processed in real time rather than bottlenecked by generic off-chip interconnects.

Even with an error-suppression factor $\Lambda > 2$ (logical error more than halved per distance step), Google observes a residual floor near 10^{-10} per cycle from unexplained correlated events, which would bound achievable logical error rates for algorithms targeting $\approx 10^{-10}$ per operation [11]. Proposed responses span improved shielding and modular processor layouts to localize chip-scale events [193], muon-veto or cosmic-ray detectors to flag dangerous shots [73], and architectural flags or module-based QEC protocols that detect and discard "bad" cycles or confine catastrophic events to limited regions.

Measured noise spectra and correlations therefore inform choices from hardware layout and cryogenic control to compilers, codes, and decoders, making structured noise a primary design constraint for scalable fault-tolerant architectures [6], [11], [100].

VII. QEC IN QUANTUM COMMUNICATION

A. MOTIVATION AND DISTINCT CHALLENGES

Quantum error correction (QEC) for optical communication is dominated by photon loss. At the physical-channel level, neglecting excess thermal noise, an optical fiber or free-space attenuation link is modeled by the single-mode pure-loss bosonic channel

$$\hat{b} = \sqrt{\eta}\,\hat{a} + \sqrt{1-\eta}\,\hat{e}, \qquad 0 \le \eta \le 1, \tag{13}$$

where $\hat{a}$ and $\hat{b}$ are the transmitted and received modes, respectively, $\hat{e}$ is an environmental vacuum mode, and η is the channel transmissivity [207]. For fiber attenuation α expressed in dB per unit distance, $\eta(L) = 10^{-\alpha L/10}$. When a discrete photonic encoding reliably heralds that a particular carrier is missing, the QEC layer can coarse-grain this physical loss into a known-location erasure; unheralded loss, dark counts, and mode mismatch are not ideal erasures.

The relevant information-theoretic limit depends on the communication task. For reliable transmission of arbitrary quantum information without two-way classical assistance, the energy-unconstrained pure-loss quantum capacity is [208]

$$Q(\mathcal{L}_\eta) = \max\left\{0, \log_2\left(\frac{\eta}{1-\eta}\right)\right\} \quad \text{qubits/mode}. \tag{14}$$

Hence $Q(\mathcal{L}_\eta) = 0$ for $\eta \le 1/2$. By contrast, allowing arbitrary adaptive endpoint operations and unlimited two-way classical communication, but no intermediate quantum repeater, the pure-loss channel has the exact two-way capacities [207]

$$Q_2(\mathcal{L}_\eta) = D_2(\mathcal{L}_\eta) = K(\mathcal{L}_\eta) = -\log_2(1-\eta), \tag{15}$$

IEEE Proof

where Q_2 is the two-way-assisted quantum capacity, D_2 the entanglement-distribution capacity, and K the secret-key capacity. Their equality is specific to this channel and resource model; these quantities are not interchangeable in general. In the high-loss limit, (15) becomes $K(\eta) \simeq \eta/\ln 2 \simeq 1.44\eta$ bits per optical mode, so the optimal repeaterless rate decreases exponentially with fiber distance.

Equations (14) and (15) provide the information-theoretic reference for the remainder of this section. QEC can convert heralded loss into correctable erasures and improve an achievable logical transmission rate, but a protocol using only the two endpoints remains subject to the corresponding direct-channel capacity. Quantum repeaters change the resource model by dividing the end-to-end path into elementary lossy links joined by intermediate quantum operations; they may therefore surpass the direct end-to-end repeaterless curve without violating (15) [209], [210], [211]. We refer to Section VI for the broader noise taxonomy and focus here on loss-dominated, asynchronous, memory-limited optical links. At the broader 6G systems layer, recent work also explores quantum-assisted federated learning with post-quantum task distribution and federated multi-agent learning for EMI-resilient UAV–RIS control [212], [213]. These application- and control-plane approaches are complementary to, but analytically distinct from, the QEC mechanisms considered here; they motivate interfaces among protected quantum links, classical security, and AI-native network control.

At the **physical layer**, bosonic encodings, binomial and cat codes, and Gottesman–Kitaev–Preskill (GKP), place logical qubits in single oscillators so that amplitude damping appears as photon-number-parity flips or approximate Gaussian displacements, correctable via in-line parity or homodyne-based syndrome measurements [214], [215], [216], [217].

At the **link layer**, entanglement is generated and protected via Bell-state measurements (BSMs) and fusion on photonic cluster/graph states; failed fusions and missing photons are treated as classical erasure syndromes so that logical information can still percolate through redundant graphs [209].

The **network layer** composes multi-hop connectivity using encoded swapping and routing—repeater-graph and tree-emission designs, and concatenated ring-graph pipelines—to maintain logical entanglement across lossy segments [210], [211].

The **transport layer** consumes encoded Bell pairs to deliver end-to-end logical states via error-corrected teleportation, and Section VII-E relates these mechanisms to secret-key-rate expressions for quantum key distribution (QKD) [209].

In summary, communication reframes QEC around erasure-tolerant state preparation and measurement-driven protection. Advantages include one-way compatibility and reduced dependence on deep quantum memories; costs include probabilistic fusion that increases buffering and synchronization pressure and imposes strict deadlines on near-detector decoders, directly linked to the real-time constraints of Section IV and to system-level metrics such as logical entanglement rate, path fidelity, fusion success probability, and buffer depth formalized in Section VII-E [218], [219].

B. LOSS-TOLERANT OPTICAL QEC STRATEGIES

At the physical layer, bosonic encodings place a logical qubit in a single harmonic oscillator so that loss leaves a structured footprint that simple measurements can reveal [214], [215], [216], [217].

Binomial and cat codes arrange codewords so that single-photon loss flips photon-number parity, detectable via nondestructive parity checks and correctable with a unitary recovery step, giving first-order protection against loss with modest hardware overhead [214], [215]. GKP codes embed logical states on a lattice in phase space so that pure loss appears as small Gaussian displacements; SUM-gate–based primitives and modular quadrature measurements estimate and correct these shifts, with performance governed by channel transmissivity and the effective displacement variance [216], [217]. These encodings support in-line parity or homodyne-based syndrome extraction as photons are created, buffered, or converted, and they compose cleanly with link-layer fusion, which can consume parity flags or displacement estimates as soft information [209], [216], [217].

Table 16 compares representative loss-tolerant photonic QEC and repeater strategies, focusing on the role of fusion, graph-state resources, and loss-tolerant entanglement distribution.

Beyond single-mode encodings, concatenated GKP streams and small graph or tree encodings act as photonic block codes for erasures, enabling one-way, memory-lean communication that remains resilient near GKP displacement thresholds and interfaces naturally with fusion-based link layers [209], [217].

In summary, physical-layer schemes tie directly to timing and local decoding: parity and GKP syndromes must be processed near detectors within a fusion window to avoid deep buffers, and time-bin synchronization and indistinguishability constrain how many rounds can be stacked before link-layer operations, linking physical design to latency and buffering limits discussed later in this section.

C. CLUSTER-STATE AND FUSION-BASED PROTOCOLS

In this subsection, we talk about network layer. Measurement-based networking prepares photonic cluster (graph) states and uses linear-optics fusion—BSMs on pairs of modes—to stitch incoming photons to locally prepared fragments. Because each fusion is probabilistic and photons can be lost in fiber or free space, failures and missing detections are treated as erasure-style syndromes and tracked classically, so logical information can still propagate through the graph. Logical qubits are embedded in redundant subgraphs—tree structures and topological clusters—whose connectivity is

IEEE Proof

TABLE 16. Loss-tolerant photonic QEC and repeater strategies.

Approach	Key idea and reported evidence	Evidence / maturity	References
Fusion-based fault-tolerant computation	Builds fault-tolerant computation from entangling fusion measurements on small resource states, with probabilistic fusion and photon loss handled by the QEC protocol. A reported ballistic construction tolerates 10.4% loss per fusion, corresponding to 2.7% loss per individual photon in the source model.	Architecture / Simulation	[63]
Experimental graph-state fusion	Deterministically generated photonic graph-state fragments are fused into larger programmable graph states; ring and tree graph states with up to eight qubits were demonstrated experimentally.	Experiment	[55]
Generalized repeater graph states	All-photonic repeater graph states use redundant connectivity to make entanglement swapping robust to probabilistic linear-optical Bell-state measurements; generalized constructions establish exploitable ebits at finite rate with high probability.	Theory / Performance analysis	[210]
Concatenated ring-graph repeaters	A one-way repeater architecture concatenates photonic ring graph codes and linear-optical Bell-state measurements to address both transmission loss and operational errors while controlling stationary-qubit requirements.	Architecture / Numerical analysis	[211]

chosen so that, even with fusion failures and per-photon loss, a high-fidelity path remains across the link layer [209]. The effective syndrome is the pattern of successful/failed fusions and detected/missing photons, which a classical decoder interprets to preserve the logical state while advancing the graph [209].

Designers tune tree depth and branching so that at least one path of successful fusions survives typical loss and BSM failure, embedding trees in lattices above percolation thresholds to suppress logical loss [209], [220]. Repeater-graph states (RGS) encode classical erasure-code structure directly into the photonic graph so that k logical ebits can be recovered from n transmitted photons even at high loss, with success curves governed by the chosen erasure code and post-selection rules [209], [220].

Parity-check and fusion-assisted schemes add global photon-number-parity checks so that BSMs can be heralded even when one mode is lost; adaptive feed-forward enables repeated fusion attempts within a shallow local memory, raising the effective success probability while keeping the architecture link-layer and one-way. Tree/RGS-style fusion networks and parity-assisted BSMs are now common for one-way repeater proposals because they natively tolerate loss and minimize deep memories, while more elaborate topological-cluster constructions are explored when higher effective distances or stronger thresholds are required [209], [211], [220].

Cluster/fusion protocols thus treat failed fusions as data, not disasters: they align naturally with one-way operation and built-in erasure tracking but increase buffer depth and impose strict synchronization and decoder-deadline pressure near detectors, directly tying their practicality to the timing and latency constraints discussed in Section IV [219], [221].

D. ENCODED REPEATERS AND SWAPPING (NETWORK LAYER)

At the network layer, encoded repeaters stitch multi-hop paths that preserve logical entanglement across lossy segments without relying on slow, two-way purification [21], [186], [211]. Entanglement swapping is performed on encoded states, so BSMs and Pauli-frame updates act on the logical code space; redundancy catches typical losses and failed BSMs on the fly, while erasure flags and parity bits guide decoders to maintain a consistent logical state across hops [21], [186], [211].

RGS pre-distribute photonic graphs that implement both the entanglement resource and an erasure-resilient embedding, with local BSMs/fusions extending logical connectivity hop by hop [210]. Deterministic tree-state emission replaces deep delay lines with matter-qubit emitters that sequentially grow small trees providing local redundancy against loss and probabilistic BSMs [21]. Concatenated ring-graph repeaters compose small graph codes into larger cycles, using adaptive fusion and error detection to maintain high logical fidelity at realistic loss with only a modest number of stationary qubits per node [211]. In each case, swapping occurs at the logical level, and the network exports a labeled encoded path while lower layers continue supplying soft information from bosonic and graph encodings [217].

These architectures interact tightly with timing and resources: buffer depth at each node must cover arrival skew between adjacent links, and near-detector decoders must process fusion/BSM syndromes within the local window so that Pauli-frame updates and route decisions do not stall upstream links [186], [219]. One-way encoded repeaters reduce round-trip signaling—classical coordination mainly announces erasures, graph-stitching outcomes, and path readiness—so improved reach and rate come at the cost of more photons per ebit and tighter synchronization across hops [21], [186], [211].

In summary, encoded swapping therefore moves entanglement operations into the logical code space, reducing reliance on deep memories and multi-round purification while tightening resource and timing budgets, as quantified later via logical entanglement rate, path fidelity, and buffer depth.

E. PERFORMANCE METRICS AND SYSTEMS-LEVEL CONSIDERATIONS

For communication-centric QEC, we make a small set of system metrics explicit and reuse them across layers. The logical entanglement rate R_{ent} is the end-to-end rate at which usable, encoded Bell pairs are delivered to the transport layer; path fidelity F_{path} is the fidelity of those delivered

IEEE Proof

logical pairs after all hops and corrections; fusion success probability P_{fuse} is the effective BSM or fusion success on the codes actually used at the link layer (including any heralding [222] or repetition); buffer depth L_{buf} is the number of time bins (or stored modes) a node must hold to hide asynchrony and decoding time; and p_{dark} denotes the dark-count probability per detection window that can corrupt fusion outcomes. These are implementation metrics rather than channel capacities.

To compare an implementation with the information-theoretic limits in Section VII-A, its rate must use the same normalization. If $\dot{M}$ optical modes (channel uses) are launched per second, define

$$r_{\text{ent}} = \frac{R_{\text{ent}}}{\dot{M}}, \; r_{\text{key}} = \frac{R_{\text{key}}}{\dot{M}}, \tag{16}$$

where R_{key} is an achievable QKD secret-key rate in bits/s. The quantities r_{ent} and r_{key}, expressed per optical mode, can be compared with $D_2(\mathcal{L}_\eta)$ and $K(\mathcal{L}_\eta)$ under matching channel and resource assumptions; F_{path}, logical error probability, P_{fuse}, and L_{buf} cannot themselves be compared with a capacity. Because reported repeater studies use different source clocks, segment spacings, detector models, and node resources, raw rates in pairs/s or bits/s should not be interpreted as a fraction of capacity without this normalization. Higher detector efficiency and better mode indistinguishability can raise P_{fuse}, while tighter timing can reduce L_{buf}; together these affect the achievable R_{ent} and QKD goodput [218].

Physical-layer encodings determine how loss becomes a syndrome and how fast it can be extracted. Binomial and cat codes use parity measurements to flag single-photon loss; faster, higher-visibility parity checks reduce buffering before fusion and increase the chance that link-layer operations see "fresh" states [214], [215]. GKP codes convert loss to small Gaussian displacements, where better squeezing, cleaner homodyne readout, and low electronic noise sharpen displacement estimates and reduce mis-corrections that would otherwise depress F_{path} [216], [217]. In all cases, syndrome-extraction time is a first-order lever for L_{buf} at repeater nodes because fusion/BSM scheduling must wait for local decisions before advancing to the next time bin [219].

At the **link layer**, graph design and heralding shape P_{fuse} and the variance of inter-arrival times. Tree and RGS trade more photons (more loss opportunities) for built-in redundancy so at least one fusion path survives, making the effective P_{fuse} a function of both device parameters and graph branching/depth [209]. Parity-assisted fusion raises the chance of a usable BSM by detecting partial events and enabling retries within the same timing envelope, which can increase R_{ent} without deepening memories when decoder decisions return within the fusion window. Synchronization—time-bin alignment and interferometric stability—controls the tails of L_{buf}. Long tails inflate buffering and expose states to more dark counts, lowering F_{path}.

Network-layer architectures convert these local link statistics into multi-hop performance. Encoded swapping accrues soft information (erasure flags, parity bits, displacement estimates) across hops so that classical side information helps decoders keep F_{path} high without expensive two-way purification, provided per-hop decoder deadlines are met [21], [211]. Designs such as deterministic tree emission and concatenated ring-graph repeaters explicitly budget arrival skew and decoding time to keep buffers shallow while maintaining high-fidelity stitching across many hops, thereby stabilizing R_{ent} as distance grows [21], [211]. When physical encodings (e.g., cat/binomial or GKP) are present on the carriers, their soft syndromes can be passed upward so that network-layer decisions (route advancement, selective discard) are better informed, improving both F_{path} and R_{ent} [217].

Transport-layer services translate these ingredients into application outcomes. For direct point-to-point QKD over a pure-loss end-to-end channel of transmissivity η_{tot}, the asymptotic secret-key rate per optical mode cannot exceed $K(\mathcal{L}_{\eta_{\text{tot}}})$ in (15) [207]. A repeater-assisted path can exceed this *direct-link* curve because intermediate quantum nodes divide the path into elementary channels and change the allowed communication resource model, not because QEC removes the underlying photon loss. The achievable end-to-end rate then depends jointly on elementary-link loss, QEC and swapping success, local operational errors, source rate, and repeater resources [209], [210], [211]. A fair comparison should therefore report r_{key} or r_{ent} together with the direct repeaterless benchmark and the required photons, memories, and repeater nodes rather than infer capacity from F_{path} alone. For distributed or blind computation, reliable and scheduled delivery of logical Bell pairs additionally constrains algorithm wall-clock time, making predictable buffers and timely Pauli-frame updates (Section IV) important alongside the information-theoretic rate [209], [219].

In summary, physical encodings decide what the syndrome is and how fast it is learned; link-layer graphs and heralding set the probability and timing of successful fusions; network-layer swapping and routing determine how these statistics compose over distance; and transport converts them into keys and logical-throughput. The benefit of this metric-driven view is transparent trade-offs and consistent comparison across proposals, at the cost of disciplined synchronization, near-detector decoding under tight deadlines, and careful budgeting of photons and stationary qubits to keep L_{buf} bounded while pushing R_{ent} and F_{path} higher.

F. CHALLENGES, TRADE-OFFS, AND RESEARCH ROADMAP

Communication-centric QEC faces system bottlenecks across layers. Photon indistinguishability and time-bin synchronization (including small spectral drift or picosecond-scale jitter) limit how large a cluster/graph or bosonic lattice can be driven before interference visibility falls. This effectively collapses code distance and forces practical designs to

IEEE Proof

cap branching, limit concatenation depth, or insert active stabilization between fusion stages [21]. Asynchronous link arrivals and probabilistic BSMs drive buffer depth or trigger skip/deferral rules, directly impacting logical entanglement rate and path fidelity via buffer-depth and fusion-success metrics [209]. Decoder speed and placement are unforgiving: link-layer fusion produces streaming syndromes that must be processed near detectors so Pauli-frame updates and route advancement land within the same timing window; soft information from bosonic encodings (e.g., GKP displacement estimates) helps only if processed within those deadlines [219]. Finally, hardware stability at scale—source brightness, mode matching across hops, and low-noise parity or homodyne readout—sets ceilings on feasible graph sizes and cat/binomial/GKP operating points [217].

These tensions drive concrete, layer-aware trade-offs. At the physical layer, stronger encodings (larger cats, deeper binomials, tighter GKP grids) expose clearer syndromes but demand better squeezing, parity visibility, or homodyne SNR, and beyond a point additional rounds inflate latency and buffer depth [216], [217]. At the link layer, more redundancy (trees, RGS) boosts the probability that at least one fusion path succeeds, but every added photon is another loss opportunity; parity-assisted fusions recover some efficiency by heralding partial events and enabling retries, provided decoders return decisions within the fusion window [209]. At the network layer, encoded swapping reduces two-way signaling and purification yet demands tighter end-to-end synchronization and more photons per delivered ebit; deterministic tree emission and concatenated ring-graph repeaters explicitly budget skew and decoding time to keep buffers shallow while maintaining high stitching fidelity across many hops [21], [211]. At the transport layer, error-corrected teleportation stabilizes session integrity for QKD and distributed computing but magnifies upstream timing and dark-count issues into lower key rates or longer wall-clock for remote gates [209].

Hybrid strategies balance these pressures. Purified encoding (PE) runs a small number of purification rounds and then encodes, improving end-to-end fidelity when channel noise dominates local gate noise, but it incurs latency and classical coordination, paying off only when the logical-error reduction outweighs the extra round-trip time [186]. Bosonic-to-discrete concatenations (e.g., GKP-to-surface) pass analog syndromes up the stack, allowing higher layers to make informed keep/discard decisions and preserve throughput under variable loss without deep quantum memory, contingent on near-detector decoding [217], [223]. Source engineering—deterministic emitters with GHz repetition and narrow linewidths—reduces mode mismatch and shrinks synchronization budgets, directly lifting fusion success and stabilizing buffers [21].

A practical roadmap follows: (i) co-design synchronization and scheduling with graph depth and bosonic rounds so buffers remain bounded; (ii) invest in near-detector, analog/soft-information decoding that consumes parity/displacement estimates within Section IV's deadlines [219]; (iii) match encodings to hardware strengths—cat/binomial where parity readout is strong, GKP where squeezing and homodyne are mature, tree/RGS where deterministic emitters exist, and concatenated ring-graphs when stationary-qubit budgets are tight but stitching fidelity must be high [21], [209], [211]; and (iv) evaluate designs using the shared metrics—logical entanglement rate, path fidelity, fusion success probability, and buffer depth—while additionally reporting achievable rate per optical mode against the appropriate direct-link capacity benchmark under matched channel and resource assumptions, so improvements at one layer do not hide costs at another [207], [219].

Overall, higher layers increase resource cost (more photons, tighter synchronization, faster decoders) but buy robustness and scalability via one-way operation and erasure tolerance, so feasible systems emerge only from layer-mapped co-design that respects decoder deadlines and synchronization budgets while exploiting opportunities such as analog decoding, deterministic sources, and modest-depth concatenations.

VIII. CONCLUSION

Quantum error correction has crossed a watershed moment. What began as a largely theoretical discipline is now an aggressively multi-layer engineering effort whose progress is paced as much by materials science, cryogenic electronics, and real-time classical processing as by code design itself. Across the field, the community has *established theoretically or demonstrated experimentally, depending on the component*, many of the ingredients needed for fault-tolerant quantum computing: 1) theoretical code constructions and numerical or architecture-level studies of constant-rate qLDPC and hypergraph-product families that can reduce code-level qubit overhead; 2) theoretical, numerical, and experimental results on bias-tailored and bosonic encodings that exploit asymmetric noise, including experimental cat–repetition memories; 3) FPGA-, ASIC-, and cryogenic-electronics-based decoder implementations and prototypes that process syndrome streams within stringent latency budgets under specific workloads; and 4) experimental control primitives that detect, track, or suppress drift, leakage, and related device errors. These results establish substantial pieces of the fault-tolerance stack, but at different levels of evidence maturity and not yet as one unified fault-tolerant system.

Yet this survey also makes clear that no single stack has unified all these advances. Three gaps discussed next now set the agenda for the rest of the decade.

A. CORRELATION-AWARE THRESHOLDS AND DECODERS

Classical proofs and most Many simplified threshold analyses and simulators use IID Pauli noise, while experiments reveal long-range crosstalk, $1/f$ drift, and radiation-induced bursts that lie outside that model. Threshold theorems already cover some restricted correlated and non-Markovian noise classes,

IEEE Proof

but extending decoder-aware theory to broader physically motivated spatio-temporal correlations—and matching it with ASIC-grade inference engines—remains an important step toward trustworthy logical-error budgets.

B. SCALABLE, HARDWARE-NATIVE CODE FAMILIES

Constant-rate LDPC codes deliver linear distance on paper, but their sparse Tanner graphs are challenging for planar processors; conversely, surface-code patches embed cleanly but fall short in terms of rate. The field needs a code family that is simultaneously local, efficient, and strong—and comes with a proven, sub-μ s decoder—before million-qubit machines can be built economically.

C. FULL-STACK INTEGRATION FOR USEFUL WORKLOADS

Individual vendors already publish roadmaps that target tens of thousands of logical qubits and error rates below 10^{-12} by 2030. But those numbers only matter if data movement, magic-state factories, and inter-module links keep pace. Photonic and modular architectures illustrate the difficulty: loss-tolerant codes and asynchronous buffers work in principle, yet their success hinges on detector jitter, dark counts, and sub-100 ns decoder loops that current prototypes barely meet.

D. OUTLOOK

Bridging these gaps will require treating QEC as a systems science rather than a sequence of point optimizations. Four directions look especially promising:

1. Co-designed lattices and electronics; just as high-performance computing pairs cache hierarchies with numerical kernels, next-generation QEC should co-optimize stabilizer layout, schedule, and cryogenic silicon so that decoding complexity tracks physical cycle time rather than code distance.
2. Adaptive, learning-enhanced control. Neural decoders that re-weight edges on the fly already outperform static MWPM in correlated noise; coupling those predictors to real-time calibration loops can push logical-error suppression deeper into the 10^{-6} to 10^{-7} regime without extra qubits.
3. Hybrid bosonic-qubit architectures. Experimental cat–repetition memories demonstrate that bosonic encodings can suppress selected physical error channels while an outer code corrects the remaining dominant errors [17]. LDPC–cat architecture studies and chiral-cat proposals further indicate that bosonic encodings may shift part of the dominant error-suppression burden into the oscillator hardware, potentially reducing the requirements placed on an outer discrete-variable code [36], [37]. However, these latter results are presently theoretical or numerical rather than experimental QEC demonstrations, so their system-level overhead advantages remain to be validated in hardware.
4. Standardized benchmarks and open tooling. A common set of correlation-rich noise traces, decoder latency targets, and energy-per-cycle figures would let researchers compare proposals on equal footing and accelerate the transition from laboratory prototypes to deployable subsystems.

In short, the pieces of the fault-tolerance puzzle are on the table; the next challenge is to assemble them into coherent, resource-balanced systems. If theory can deliver correlation-robust thresholds, industry can tame cryogenic integration, and the community can agree on transparent benchmarks, the next five years should witness the first fault-tolerant logical processors executing dynamic multi-qubit algorithms—a milestone that will redefine both quantum computing and the classical infrastructure that enables it.

ACKNOWLEDGMENT

The statements made here are solely the responsibility of the authors.

REFERENCES

[1] A. Chatterjee, K. Phalak, and S. Ghosh, "Quantum error correction for dummies," in *Proc. IEEE Int. Conf. Quantum Comput. Eng. (QCE)*, Bellevue, WA, USA, Sep. 2023, pp. 70–81, doi: 10.1109/QCE57702.2023.00017.

[2] I. Mahmud and A. Abdelhadi, "Quantum codes: A comprehensive survey of techniques, challenges, and future directions," *IEEE Access*, vol. 13, pp. 214352–214386, 2025, doi: 10.1109/ACCESS.2025.3646024.

[3] A. Jayashankar and P. Mandayam, "Quantum error correction: Noise-adapted techniques and applications," *J. Indian Inst. Sci.*, vol. 103, no. 2, pp. 497–512, Apr. 2023, doi: 10.1007/s41745-022-00332-x.

[4] Z. Wang and H. Tang, "Artificial intelligence for quantum error correction: A comprehensive review," 2024, *arXiv:2412.20380*.

[5] V. S. Thakur, A. Kumar, M. Magarini, K. Dev, and O. A. Dobre, "Quantum communication and information technologies: A survey on foundation, error correction, NISQ, and networks," *TechRxiv*, Sep. 2025, doi: 10.36227/techrxiv.175735780.05700466/v1.

[6] F. Battistel, C. Chamberland, K. Johar, R. W. J. Overwater, F. Sebastiano, L. Skoric, Y. Ueno, and M. Usman, "Real-time decoding for fault-tolerant quantum computing: Progress, challenges and outlook," *Nano Futures*, vol. 7, no. 3, Aug. 2023, Art. no. 032003, doi: 10.1088/2399-1984/aceba6.

[7] J.-N. Li, T.-H. Wei, Z.-H. Mei, Q. Zhou, W.-C. Kong, J. Zhang, Y.-C. Wu, and G.-P. Guo, "Real-time decoding for surface code," *Adv. Quantum Technol.*, vol. 9, no. 5, May 2026, Art. no. e00806, doi: 10.1002/qute.202500806.

[8] Y. Chen, C.-Y. Luan, P. Zheng, X. Zeng, J.-Y. Hou, Z. Fu, Y. Jin, F. Wang, G. Yang, and D. Lv, "Real-time quantum error correction system stack: Architecture, algorithms, and engineering practice," 2026, *arXiv:2605.30765*.

[9] H. Zhou, M. Cain, and M. D. Lukin, "Opportunities in full-stack design of low-overhead fault-tolerant quantum computation," *Nature Comput. Sci.*, vol. 5, no. 12, pp. 1110–1119, Dec. 2025, doi: 10.1038/s43588-025-00895-6.

[10] S. Bravyi, A. W. Cross, J. M. Gambetta, D. Maslov, P. Rall, and T. J. Yoder, "High-threshold and low-overhead fault-tolerant quantum memory," *Nature*, vol. 627, no. 8005, pp. 778–782, Mar. 2024, doi: 10.1038/s41586-024-07107-7.

[11] R. Acharya et al., "Quantum error correction below the surface code threshold," *Nature*, vol. 638, no. 8052, pp. 920–926, Feb. 2025, doi: 10.1038/s41586-024-08449-y.

[12] J. Zhang, Y.-C. Wu, and G.-P. Guo, "Facilitating practical fault-tolerant quantum computing based on color codes," *Phys. Rev. Res.*, vol. 6, no. 3, Jul. 2024, Art. no. 033086, doi: 10.1103/physrevresearch.6.033086.

[13] J. P. Bonilla Ataides, D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, "The XZZX surface code," *Nature Commun.*, vol. 12, no. 1, Apr. 2021, Art. no. 2172, doi: 10.1038/s41467-021-22274-1.

IEEE Proof

[14] A. Leverrier and G. Zémor, "Quantum tanner codes," in *Proc. IEEE 63rd Annu. Symp. Found. Comput. Sci. (FOCS)*, Denver, CO, USA, Oct. 2022, pp. 872–883, doi: 10.1109/FOCS54457.2022.00117.

[15] S. Gu, E. Tang, L. Caha, S. H. Choe, Z. He, and A. Kubica, "Single-shot decoding of good quantum LDPC codes," *Commun. Math. Phys.*, vol. 405, no. 3, p. 85, Mar. 2024, doi: 10.1007/s00220-024-04951-6.

[16] M. Mathews, L. Pahl, D. Pahl, V. L. Addala, C. Tang, W. D. Oliver, and J. A. Grover, "Placing and routing quantum LDPC codes in multilayer superconducting hardware," *npj Quantum Inf.*, vol. 12, no. 1, p. 114, Apr. 2026, doi: 10.1038/s41534-026-01243-w.

[17] H. Putterman et al., "Hardware-efficient quantum error correction via concatenated bosonic qubits," *Nature*, vol. 638, no. 8052, pp. 927–934, Feb. 2025, doi: 10.1038/s41586-025-08642-7.

[18] S. Bravyi, D. Poulin, and B. Terhal, "Tradeoffs for reliable quantum information storage in 2D systems," *Phys. Rev. Lett.*, vol. 104, no. 5, Feb. 2010, Art. no. 050503, doi: 10.1103/physrevlett.104.050503.

[19] P. Panteleev and G. Kalachev, "Asymptotically good quantum and locally testable classical LDPC codes," in *Proc. 54th Annu. ACM SIGACT Symp. Theory Comput.*, Rome, Italy, Jun. 2022, pp. 375–388, doi: 10.1145/3519935.3520017.

[20] L. Golowich and T.-C. Lin, "Quantum LDPC codes with transversal non-Clifford gates via products of algebraic codes," in *Proc. 57th Annu. ACM Symp. Theory Comput.*, Jun. 2025, pp. 689–696, doi: 10.1145/3717823.3718139.

[21] Y. Li, N. O'Dea, and V. Khemani, "Perturbative stability and error-correction thresholds of quantum codes," *PRX Quantum*, vol. 6, no. 1, Feb. 2025, Art. no. 010327, doi: 10.1103/prxquantum.6.010327.

[22] V. Guemard, "Lifts of quantum CSS codes," *IEEE Trans. Inf. Theory*, vol. 71, no. 7, pp. 5418–5442, Jul. 2025, doi: 10.1109/TIT.2025.3552211.

[23] L. Voss, S. J. Xian, T. Haug, and K. Bharti, "Multivariate bicycle codes," *Phys. Rev. A, Gen. Phys.*, vol. 111, no. 6, Jun. 2025, Art. no. L060401, doi: 10.1103/ll5p-z88p.

[24] E. Knill, R. Laflamme, and W. H. Zurek, "Resilient quantum computation: Error models and thresholds," *Proc. Roy. Soc. London. A, Math., Phys. Eng. Sci.*, vol. 454, no. 1969, pp. 365–384, Jan. 1998, doi: 10.1098/rspa.1998.0166.

[25] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, 10th ed. Cambridge, U.K.: Cambridge Univ. Press, 2010, doi: 10.1017/CBO9780511976667

[26] D. Aharonov, A. Kitaev, and J. Preskill, "Fault-tolerant quantum computation with long-range correlated noise," *Phys. Rev. Lett.*, vol. 96, no. 5, Feb. 2006, Art. no. 050504, doi: 10.1103/physrevlett.96.050504.

[27] A. Bargerbos, L. J. Splitthoff, M. Pita-Vidal, J. J. Wesdorp, Y. Liu, P. Krogstrup, L. P. Kouwenhoven, C. K. Andersen, and L. Grünhaupt, "Mitigation of quasiparticle loss in superconducting qubits by phonon scattering," *Phys. Rev. Appl.*, vol. 19, no. 2, Feb. 2023, Art. no. 024014, doi: 10.1103/physrevapplied.19.024014.

[28] A. deMarti iOlius, P. Fuentes, R. Orús, P. M. Crespo, and J. E. Martinez, "Decoding algorithms for surface codes," *Quantum*, vol. 8, p. 1498, Oct. 2024, doi: 10.22331/q-2024-10-10-1498.

[29] B. Eastin and E. Knill, "Restrictions on transversal encoded quantum gate sets," *Phys. Rev. Lett.*, vol. 102, no. 11, Mar. 2009, Art. no. 110502, doi: 10.1103/physrevlett.102.110502.

[30] S. Bravyi and J. Haah, "Magic-state distillation with low overhead," *Phys. Rev. A, Gen. Phys.*, vol. 86, no. 5, Nov. 2012, Art. no. 052329, doi: 10.1103/physreva.86.052329.

[31] S. Novák, "Homological quantum error correction with torsion," 2024, *arXiv:2405.03559*.

[32] Y. Tang, T. Yao, H. Xu, and X. Kai, "New quantum codes from constacyclic codes over finite chain rings," *Quantum Inf. Process.*, vol. 23, no. 9, p. 311, Aug. 2024, doi: 10.1007/s11128-024-04519-2.

[33] E. Kubischta and I. Teixeira, "Permutation-invariant quantum codes with transversal generalized phase gates," *IEEE Trans. Inf. Theory*, vol. 71, no. 1, pp. 485–498, Jan. 2025, doi: 10.1109/TIT.2024.3487964.

[34] N. P. Breuckmann and S. Burton, "Fold-transversal Clifford gates for quantum codes," *Quantum*, vol. 8, p. 1372, Jun. 2024, doi: 10.22331/q-2024-06-13-1372.

[35] K. Tiurev, A. Pesah, P.-J. H. S. Derks, J. Roffe, J. Eisert, M. S. Kesselring, and J.-M. Reiner, "Domain wall color code," *Phys. Rev. Lett.*, vol. 133, no. 11, Sep. 2024, Art. no. 110601, doi: 10.1103/physrevlett.133.110601.

[36] D. Ruiz, J. Guillaud, A. Leverrier, M. Mirrahimi, and C. Vuillot, "LDPC-cat codes for low-overhead quantum computing in 2D," *Nature Commun.*, vol. 16, no. 1, p. 1040, Jan. 2025, doi: 10.1038/s41467-025-56298-8.

[37] A. Labay-Mora, A. Mercurio, V. Savona, G. L. Giorgi, and F. Minganti, "Chiral cat code: Enhanced error correction induced by higher-order nonlinearities," *Phys. Rev. Res.*, vol. 8, no. 3, Sep. 2026, Art. no. 033254, doi: 10.1103/kt69-mwf7.

[38] P. A. Spring, L. Milanovic, Y. Sunada, S. Wang, A. F. van Loo, S. Tamate, and Y. Nakamura, "Fast multiplexed superconducting-qubit readout with intrinsic Purcell filtering using a multiconductor transmission line," *PRX Quantum*, vol. 6, no. 2, Jun. 2025, Art. no. 020345, doi: 10.1103/prxquantum.6.020345.

[39] A. Bengtsson, A. Opremcak, M. Khezri, D. Sank, A. Bourassa, K. J. Satzinger, S. Hong, C. Erickson, B. J. Lester, K. C. Miao, A. N. Korotkov, J. Kelly, Z. Chen, and P. V. Klimov, "Model-based optimization of superconducting qubit readout," *Phys. Rev. Lett.*, vol. 132, no. 10, Mar. 2024, Art. no. 100603, doi: 10.1103/physrevlett.132.100603.

[40] S. Heußen, D. F. Locher, and M. Müller, "Measurement-free fault-tolerant quantum error correction in near-term devices," *PRX Quantum*, vol. 5, no. 1, Feb. 2024, Art. no. 010333, doi: 10.1103/prxquantum.5.010333.

[41] N. Berthusen, S. J. S. Tan, E. Huang, and D. Gottesman, "Adaptive syndrome extraction," *PRX Quantum*, vol. 6, no. 3, Jul. 2025, Art. no. 030307, doi: 10.1103/ps3r-wf84.

[42] A. Chatterjee, A. Ghosh, and S. Ghosh, "Quantum prometheus: Defying overhead with recycled ancillas in quantum error correction," in *Proc. 26th Int. Symp. Quality Electron. Design (ISQED)*, San Francisco, CA, USA, Apr. 2025, pp. 1–7, doi: 10.1109/ISQED65160.2025.11014328.

[43] I. A. Simakov and I. S. Besedin, "Low-overhead quantum error-correction codes with a cyclic topology," *Phys. Rev. A, Gen. Phys.*, vol. 111, no. 1, Jan. 2025, Art. no. 012444, doi: 10.1103/physreva.111.012444.

[44] M. S. Kesselring, J. C. Magdalena de la Fuente, F. Thomsen, J. Eisert, S. D. Bartlett, and B. J. Brown, "Anyon condensation and the color code," *PRX Quantum*, vol. 5, no. 1, Mar. 2024, Art. no. 010342, doi: 10.1103/prxquantum.5.010342.

[45] W. Sun, X. Li, L. Yu, Z. Wang, G. Chen, D. Zheng, and G. Yang, "HAQA: A hardware-guided and fidelity-aware strategy for efficient qubit mapping optimization," *EPJ Quantum Technol.*, vol. 13, no. 1, p. 61, Apr. 2026, doi: 10.1140/epjqt/s40507-026-00505-y.

[46] P. V. Klimov et al., "Optimizing quantum gates towards the scale of logical qubits," *Nature Commun.*, vol. 15, no. 1, p. 2442, Mar. 2024, doi: 10.1038/s41467-024-46623-y.

[47] C. Benito, E. López, B. Peropadre, and A. Bermudez, "Comparative study of quantum error correction strategies for the heavy-hexagonal lattice," *Quantum*, vol. 9, p. 1623, Feb. 2025, doi: 10.22331/q-2025-02-06-1623.

[48] S. Gu, A. Retzker, and A. Kubica, "Fault-tolerant quantum architectures based on erasure qubits," *Phys. Rev. Res.*, vol. 7, no. 1, Mar. 2025, Art. no. 013249, doi: 10.1103/physrevresearch.7.013249.

[49] H. Putterman, K. Noh, R. N. Patel, G. A. Peairs, G. S. MacCabe, M. Lee, S. Aghaeimeibodi, C. T. Hann, I. Jarrige, G. Marcaud, Y. He, H. Moradinejad, J. Clai Owens, T. Scaffidi, P. Arrangoiz-Arriola, J. Iverson, H. Levine, F. G. S. L. Brandão, M. H. Matheny, and O. Painter, "Preserving phase coherence and linearity in cat qubits with exponential bit-flip suppression," *Phys. Rev. X*, vol. 15, no. 1, Mar. 2025, Art. no. 011070, doi: 10.1103/physrevx.15.011070.

[50] M. Ye and N. Delfosse, "Quantum error correction for long chains of trapped ions," *Quantum*, vol. 9, p. 1920, Nov. 2025, doi: 10.22331/q-2025-11-27-1920.

[51] L. Postler, F. Butt, I. Pogorelov, C. D. Marciniak, S. Heußen, R. Blatt, P. Schindler, M. Rispler, M. Müller, and T. Monz, "Demonstration of fault-tolerant steane quantum error correction," *PRX Quantum*, vol. 5, no. 3, Aug. 2024, Art. no. 030326, doi: 10.1103/prxquantum.5.030326.

[52] C. Ryan-Anderson et al., "High-fidelity teleportation of a logical qubit using transversal gates and lattice surgery," *Science*, vol. 385, no. 6715, pp. 1327–1331, Sep. 2024, doi: 10.1126/science.adp6016.

[53] M. Ringbauer, M. Hinsche, T. Feldker, P. K. Faehrmann, J. Bermejo-Vega, C. L. Edmunds, L. Postler, R. Stricker, C. D. Marciniak, M. Meth, I. Pogorelov, R. Blatt, P. Schindler, J. Eisert, T. Monz, and D. Hangleiter, "Verifiable measurement-based quantum random sampling with trapped ions," *Nature Commun.*, vol. 16, no. 1, p. 106, Jan. 2025, doi: 10.1038/s41467-024-55342-3.

IEEE Proof

[54] M. L. Chan, T. J. Bell, L. A. Pettersson, S. X. Chen, P. Yard, A. S. Sørensen, and S. Paesani, "Tailoring fusion-based photonic quantum computing schemes to quantum emitters," *PRX Quantum*, vol. 6, no. 2, Apr. 2025, Art. no. 020304, doi: 10.1103/prxquantum.6.020304.

[55] P. Thomas, L. Ruscio, O. Morin, and G. Rempe, "Fusion of deterministically generated photonic graph states," *Nature*, vol. 629, no. 8012, pp. 567–572, May 2024, doi: 10.1038/s41586-024-07357-5.

[56] Z. Li, K. Xue, J. Li, L. Chen, R. Li, Z. Wang, N. Yu, D. S. L. Wei, Q. Sun, and J. Lu, "Entanglement-assisted quantum networks: Mechanics, enabling technologies, challenges, and research directions," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2133–2189, 4th Quart., 2023, doi: 10.1109/COMST.2023.3294240.

[57] T. M. Graham, L. Phuttitarn, R. Chinnarasu, Y. Song, C. Poole, K. Jooya, J. Scott, A. Scott, P. Eichler, and M. Saffman, "Midcircuit measurements on a single-species neutral alkali atom quantum processor," *Phys. Rev. X*, vol. 13, no. 4, Dec. 2023, Art. no. 041051, doi: 10.1103/physrevx.13.041051.

[58] S. Vittal, P. Das, and M. Qureshi, "ERASER: Towards adaptive leakage suppression for fault-tolerant quantum computing," in *Proc. 56th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Toronto, ON, Canada, Oct. 2023, pp. 509–525, doi: 10.1145/3613424.3614251.

[59] A. Siegel, A. Strikis, and M. Fogarty, "Towards early fault tolerance on a 2× n array of qubits equipped with shuttling," *PRX Quantum*, vol. 5, no. 4, Nov. 2024, Art. no. 040328, doi: 10.1103/prxquantum.5.040328.

[60] J. Yoo, Z. Chen, F. Arute, S. Montazeri, M. Szalay, C. Erickson, E. Jeffrey, R. Fatemi, M. Giustina, M. Ansmann, E. Lucero, J. Kelly, and J. C. Bardin, "Design and characterization of a <4-mW/Qubit 28-nm cryo-CMOS integrated circuit for full control of a superconducting quantum processor unit cell," *IEEE J. Solid-State Circuits*, vol. 58, no. 11, pp. 3044–3059, Nov. 2023, doi: 10.1109/JSSC.2023.3309317.

[61] M. Will, T. A. Cochran, E. Rosenberg, B. Jobst, N. M. Eassa, P. Roushan, M. Knap, A. Gammon-Smith, and F. Pollmann, "Probing non-equilibrium topological order on a quantum processor," *Nature*, vol. 645, no. 8080, pp. 348–353, Sep. 2025, doi: 10.1038/s41586-025-09456-3.

[62] D. Bluvstein et al., "A fault-tolerant neutral-atom architecture for universal quantum computation," *Nature*, vol. 649, no. 8095, pp. 39–46, Jan. 2026, doi: 10.1038/s41586-025-09848-5.

[63] S. Bartolucci, P. Birchall, H. Bombín, H. Cable, C. Dawson, M. Gimeno-Segovia, E. Johnston, K. Kieling, N. Nickerson, M. Pant, F. Pastawski, T. Rudolph, and C. Sparrow, "Fusion-based quantum computation," *Nature Commun.*, vol. 14, no. 1, Feb. 2023, Art. no. 912, doi: 10.1038/s41467-023-36493-1.

[64] X. Xue, M. Russ, N. Samkharadze, B. Undseth, A. Sammak, G. Scappucci, and L. M. K. Vandersypen, "Quantum logic with spin qubits crossing the surface code threshold," *Nature*, vol. 601, no. 7893, pp. 343–347, Jan. 2022, doi: 10.1038/s41586-021-04273-w.

[65] D. Delali Kwesi Wayo, L. Goliatt, and D. Ganji, "Linear optics to scalable photonic quantum computing," 2025, *arXiv:2501.02513*.

[66] M. N. H. Chow, V. Buchemmavari, S. Omanakuttan, B. J. Little, S. Pandey, I. H. Deutsch, and Y.-Y. Jau, "Circuit-based leakage-to-erasure conversion in a neutral-atom quantum processor," *PRX Quantum*, vol. 5, no. 4, Dec. 2024, Art. no. 040343, doi: 10.1103/prxquantum.5.040343.

[67] X. Yang et al., "Coupler-assisted leakage reduction for scalable quantum error correction with superconducting qubits," *Phys. Rev. Lett.*, vol. 133, no. 17, Oct. 2024, Art. no. 170601, doi: 10.1103/physrevlett.133.170601.

[68] E. Hyyppä, A. Vepsäläinen, M. Papič, C. F. Chan, S. Inel, A. Landra, W. Liu, J. Luus, F. Marxer, C. Ockeloen-Korppi, S. Orbell, B. Tarasinski, and J. Heinsoo, "Reducing leakage of single-qubit gates for superconducting quantum processors using analytical control pulse envelopes," *PRX Quantum*, vol. 5, no. 3, Sep. 2024, Art. no. 030353, doi: 10.1103/prxquantum.5.030353.

[69] P. Zhao, K. Linghu, Z. Li, P. Xu, R. Wang, G. Xue, Y. Jin, and H. Yu, "Quantum crosstalk analysis for simultaneous gate operations on superconducting qubits," *PRX Quantum*, vol. 3, no. 2, Apr. 2022, Art. no. 020301, doi: 10.1103/prxquantum.3.020301.

[70] Z. Zhou, A. Ji, and Y. Ding, "Surface code error correction with crosstalk noise," 2025, *arXiv:2503.04642*.

[71] A. Seif, H. Liao, V. Tripathi, K. Krsulich, M. Malekakhlagh, M. Amico, P. Jurcevic, and A. Javadi-Abhari, "Suppressing correlated noise in quantum computers via context-aware compiling," in *Proc. ACM/IEEE 51st Annu. Int. Symp. Comput. Archit. (ISCA)*, Buenos Aires, Argentina, Jun. 2024, pp. 310–324, doi: 10.1109/ISCA59077.2024.00031.

[72] E. Yelton, C. P. Larson, V. Iaia, K. Dodge, G. La Magna, P. G. Baity, I. V. Pechenezhskiy, R. McDermott, N. A. Kurinsky, G. Catelani, and B. L. Plourde, "Modeling phonon-mediated quasiparticle poisoning in superconducting qubit arrays," *Phys. Rev. B, Condens. Matter*, vol. 110, no. 2, Jul. 2024, Art. no. 024519, doi: 10.1103/PhysRevB.110.024519.

[73] X. Li, J. Wang, Y.-Y. Jiang, G.-M. Xue, X. Cai, J. Zhou, M. Gong, Z.-F. Liu, S.-Y. Zheng, D.-K. Ma, M. Chen, W.-J. Sun, S. Yang, F. Yan, Y.-R. Jin, S. P. Zhao, X.-F. Ding, and H.-F. Yu, "Cosmic-ray-induced correlated errors in superconducting qubit array," *Nature Commun.*, vol. 16, no. 1, May 2025, Art. no. 4677, doi: 10.1038/s41467-025-59778-z.

[74] D. A. Rower et al., "Evolution of 1/f flux noise in superconducting qubits with weak magnetic fields," *Phys. Rev. Lett.*, vol. 130, no. 22, May 2023, Art. no. 220602, doi: 10.1103/physrevlett.130.220602.

[75] S. Vittal, P. Das, and M. Qureshi, "Astrea: Accurate quantum error-decoding via practical minimum-weight perfect-matching," in *Proc. 50th Annu. Int. Symp. Comput. Archit.*, Orlando, FL, USA, Jun. 2023, pp. 1–16, doi: 10.1145/3579371.3589037.

[76] S. Joshi and S. Moazeni, "Scaling up superconducting quantum computers with cryogenic RF-photonics," *J. Lightw. Technol.*, vol. 42, no. 1, pp. 166–175, Jan. 1, 2024, doi: 10.1109/JLT.2023.3311806.

[77] E. I. Malevannaya, V. I. Polozov, A. I. Ivanov, A. R. Matanin, N. S. Smirnov, V. V. Echeistov, D. O. Moskalev, D. A. Mikhalin, D. E. Shirokov, Y. V. Panfilov, I. A. Ryzhikov, A. V. Andriyash, and I. A. Rodionov, "An engineering guide to superconducting quantum circuit shielding," 2025, *arXiv:2505.14274*.

[78] P.-A. Mouny, M. Benhouria, V. Yon, P. Dufour, L. Huang, Y. Beilliard, S. Rochette, D. Drouin, and P. Ronagh, "Towards a cryogenic CMOS-memristor neural decoder for quantum error correction," in *Proc. IEEE Int. Conf. Quantum Comput. Eng. (QCE)*, Montreal, QC, Canada, Sep. 2024, pp. 1258–1263, doi: 10.1109/QCE60285.2024.00149.

[79] M. V. Larsen et al., "Integrated photonic source of gottesman–kitaev–preskill qubits," *Nature*, vol. 642, no. 8068, pp. 587–591, Jun. 2025, doi: 10.1038/s41586-025-09044-5.

[80] Z.-H. Chen, M.-C. Chen, C.-Y. Lu, and J.-W. Pan, "Transversal logical Clifford gates on rotated surface codes with reconfigurable neutral atom arrays," 2024, *arXiv:2412.01391*.

[81] K. Mayer et al., "Benchmarking logical three-qubit quantum Fourier transform encoded in the steane code on a trapped-ion quantum computer," 2024, *arXiv:2404.08616*.

[82] S. Heußen and J. Hilder, "Efficient fault-tolerant code switching via one-way transversal CNOT gates," *Quantum*, vol. 9, p. 1846, Sep. 2025, doi: 10.22331/q-2025-09-03-1846.

[83] S. Dasu, S. Burton, K. Mayer, D. Amaro, J. A. Gerber, K. Gilmore, D. Gresh, D. DelVento, A. C. Potter, and D. Hayes, "Breaking even with magic: Demonstration of a high-fidelity logical non-Clifford gate," 2025, *arXiv:2506.14688*.

[84] Y. Wu, N. Liyanage, and L. Zhong, "Micro blossom: Accelerated minimum-weight perfect matching decoding for quantum error correction," in *Proc. 30th ACM Int. Conf. Architectural Support Program. Lang. Operating Syst.*, Mar. 2025, pp. 639–654, doi: 10.1145/3676641.3716005.

[85] D. J. Williamson and T. J. Yoder, "Low-overhead fault-tolerant quantum computation by gauging logical operators," *Nature Phys.*, vol. 22, no. 4, pp. 598–603, Apr. 2026, doi: 10.1038/s41567-026-03220-8.

[86] E. Swaroop, T. Jochym-O'Connor, and T. J. Yoder, "Universal adapters between quantum low-density parity check codes," *PRX Quantum*, vol. 7, no. 1, Feb. 2026, Art. no. 010324, doi: 10.1103/1g44-jp62.

[87] R. Gao, F. Wu, H. Sun, J. Chen, H. Deng, X. Ma, X. Miao, Z. Song, X. Wan, F. Wang, T. Xia, M. Ying, C. Zhang, Y. Shi, H.-H. Zhao, and C. Deng, "The effects of disorder in superconducting materials on qubit coherence," *Nature Commun.*, vol. 16, no. 1, p. 3620, Apr. 2025, doi: 10.1038/s41467-025-58745-y.

[88] H. Mishra, S. Bonam, V. Kumar, and S. G. Singh, "Novel wafer-level ta-ta direct thermocompression bonding for 3D integration of superconducting interconnects for scalable quantum computing system," *IEEE Electron Device Lett.*, vol. 45, no. 11, pp. 2221–2224, Nov. 2024, doi: 10.1109/LED.2024.3453174.

UNDER PEER REVIEW

[89] E. Bhatia, H. Frost, N. Pieniazek, J. Nalaskowski, J. Mucci, W. Collison, B. Martinick, G. Telhu, B. Egan, S. Olson, P.-Y. Hung, I. Wells, C. Johnson, T. Murray, K. Musick, and S. P. Rao, "Enabling 300 mm wafer-scale fabrication of superconducting quantum devices," in *Proc. 35th Annu. SEMI Adv. Semiconductor Manuf. Conf. (ASMC)*, Albany, NY, USA, May 2024, pp. 1–6, doi: 10.1109/ASMC61125.2024.10545509.

[90] M. Field, A. Q. Chen, B. Scharmann, E. A. Sete, F. Oruc, K. Vu, V. Kosenko, J. Y. Mutus, S. Poletto, and A. Bestwick, "Modular superconducting-qubit architecture with a multichip tunable coupler," *Phys. Rev. Appl.*, vol. 21, no. 5, May 2024, Art. no. 054063, doi: 10.1103/physrevapplied.21.054063.

[91] N. Budinger, A. Furusawa, and P. van Loock, "All-optical quantum computing using cubic phase gates," *Phys. Rev. Res.*, vol. 6, no. 2, Jun. 2024, Art. no. 023332, doi: 10.1103/physrevresearch.6.023332.

[92] D. Underwood et al., "Using cryogenic CMOS control electronics to enable a two-qubit cross-resonance gate," *PRX Quantum*, vol. 5, no. 1, Feb. 2024, Art. no. 010326, doi: 10.1103/prxquantum.5.010326.

[93] D. Minn, K. Kang, J. Lee, S. Bae, B. Kim, J. Lee, and J.-Y. Sim, "A 40 nm cryo-CMOS homodyne-demodulation readout SoC for superconducting qubits," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 72, no. 3, pp. 1003–1016, Mar. 2025, doi: 10.1109/TCSI.2024.3518472.

[94] M. Bakr, S. D. Fasciati, S. Cao, G. Campanaro, J. Wills, M. Alghadeer, M. Piscitelli, B. Shteynas, V. Chidambaram, and P. J. Leek, "Multiplexed readout of superconducting qubits using a three-dimensional reentrant-cavity filter," *Phys. Rev. Appl.*, vol. 23, no. 5, May 2025, Art. no. 054089, doi: 10.1103/physrevapplied.23.054089.

[95] A. Di Giovanni, A. S. Aasen, J. Lisenfeld, M. Gärttner, H. Rotzinger, and A. V. Ustinov, "Benchmarking the quality of multiplexed qubit readout beyond assignment fidelity," *Phys. Rev. Appl.*, vol. 24, no. 4, Oct. 2025, Art. no. 044043, doi: 10.1103/dpft-lxtx.

[96] S. Vervoort, L. Nulens, D. A. D. Chaves, H. Dausy, S. Reniers, M. Abouelela, I. P. C. Cools, A. V. Silhanek, M. J. Van Bael, B. Raes, and J. Van de Vondel, "DC-operated Josephson junction arrays as a cryogenic on-chip microwave measurement platform," *Commun. Phys.*, vol. 8, no. 1, p. 292, Jul. 2025, doi: 10.1038/s42005-025-02188-w.

[97] L. Caune et al., "Demonstrating real-time and low-latency quantum error correction with superconducting qubits," *Nature Commun.*, vol. 17, no. 1, p. 7383, Jun. 2026, doi: 10.1038/s41467-026-73331-6.

[98] B. Barber, K. M. Barnes, T. Bialas, O. Buğdayci, E. T. Campbell, N. I. Gillespie, K. Johar, R. Rajan, A. W. Richardson, L. Skoric, C. Topal, M. L. Turner, and A. B. Ziad, "A real-time, scalable, fast and resource-efficient decoder for a quantum computer," *Nature Electron.*, vol. 8, no. 1, pp. 84–91, Jan. 2025, doi: 10.1038/s41928-024-01319-5.

[99] J. C. Brennan, J. Barbosa, C. Li, M. Ahmad, F. Imroze, C. Rose, W. Karar, M. Stanley, H. Heidari, N. M. Ridler, and M. Weides, "Classical interfaces for controlling cryogenic quantum computing technologies," *APL Quantum*, vol. 2, no. 4, Dec. 2025, Art. no. 041501, doi: 10.1063/5.0273490.

[100] D. Camps, E. Rrapaj, K. Klymko, B. Austin, and N. J. Wright, "Evaluation of the classical hardware requirements for large-scale quantum computations," in *Proc. ISC High Perform. Res. Paper (39th Int. Conf.)*, Hamburg, Germany, May 2024, pp. 1–12, doi: 10.23919/ISC.2024.10528937.

[101] M. G. Bautista-Jurney, P. Gonzalez-Guerrero, and A. Butko, "SFQ-driven pulse-phase sequence generator for superconducting qubit control," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)*, Milan, Italy, Jun. 2025, pp. 1163–1169, doi: 10.1109/IPDPSW66978.2025.00182.

[102] D. Litinski, "A game of surface codes: Large-scale quantum computing with lattice surgery," *Quantum*, vol. 3, p. 128, Mar. 2019, doi: 10.22331/q-2019-03-05-128.

[103] T. Mayer, H. Bender, S. J. K. Lang, Z. Luo, J. Weber, C. Moran Guizan, C. Dhieb, D. Zahn, L. Schwarzenbach, W. Hell, M. Andronic, A. Drost, K. Neumeier, W. Lerch, L. Nebrich, A. Hagelauer, I. Eisele, R. N. Pereira, and C. Kutter, "3D-integrated superconducting qubits: CMOS-compatible, wafer-scale processing for flip-chip architectures," 2025, *arXiv:2505.04337*.

[104] T. Oshio, R. Nishimoto, T. Higuchi, K. Hayasaka, K. Koike, S. Morisaka, T. Miyoshi, R. Ohira, and U. Tanaka, "Development of a bipolar 50 v output digital-to-analog converter system for ion-shuttling operations," *J. Appl. Phys.*, vol. 137, no. 14, Apr. 2025, Art. no. 144401, doi: 10.1063/5.0252358.

[105] R. Ohira, S. Morisaka, I. Nakamura, A. Noguchi, and T. Miyoshi, "Multiplexed control at scale for electrode arrays in trapped-ion quantum processors," in *Proc. IEEE Int. Conf. Quantum Comput. Eng. (QCE)*, Albuquerque, NM, USA, Aug. 2025, pp. 1351–1359, doi: 10.1109/QCE65121.2025.00150.

[106] C. Mordini, A. Ricci Vasquez, Y. Motohashi, M. Müller, M. Malinowski, C. Zhang, K. K. Mehta, D. Kienzler, and J. P. Home, "Multizone trapped-ion qubit control in an integrated photonics QCCD device," *Phys. Rev. X*, vol. 15, no. 1, Feb. 2025, Art. no. 011040, doi: 10.1103/physrevx.15.011040.

[107] J. Kwon, W. J. Setzer, M. Gehl, N. Karl, J. Van Der Wall, R. Law, M. G. Blain, D. Stick, and H. J. McGuinness, "Multi-site integrated optical addressing of trapped ions," *Nature Commun.*, vol. 15, no. 1, p. 3709, May 2024, doi: 10.1038/s41467-024-47882-5.

[108] Y. Li, Q. Mei, Q.-X. Jie, W. Cai, Y. Li, Z. Liu, Z.-J. Chen, Z. Xie, X. Cheng, X. Zhao, Z. Luo, M. Zhang, X.-B. Zou, C.-L. Zou, Y. Lin, and J. Du, "Beating the break-even point with autonomous quantum error correction," 2025, *arXiv:2504.16746*.

[109] L. Skoric, D. E. Browne, K. M. Barnes, N. I. Gillespie, and E. T. Campbell, "Parallel window decoding enables scalable fault tolerant quantum computation," *Nature Commun.*, vol. 14, no. 1, p. 7040, Nov. 2023, doi: 10.1038/s41467-023-42482-1.

[110] P. Das, C. A. Pattison, S. Manne, D. M. Carmean, K. M. Svore, M. Qureshi, and N. Delfosse, "AFS: Accurate, fast, and scalable error-decoding for fault-tolerant quantum computers," in *Proc. IEEE Int. Symp. High-Perform. Comput. Archit. (HPCA)*, Seoul, South Korea, Apr. 2022, pp. 259–273, doi: 10.1109/HPCA53966.2022.00027.

[111] P. Das, A. Locharla, and C. Jones, "LILLIPUT: A lightweight low-latency lookup-table decoder for near-term quantum error correction," in *Proc. 27th ACM Int. Conf. Architectural Support Program. Lang. Operating Syst.*, Lausanne, Switzerland, Feb. 2022, pp. 541–553, doi: 10.1145/3503222.3507707.

[112] P. Renault, P. Yard, R. C. Pooser, M. Eaton, and H. A. Zaidi, "End-to-end switchless architecture for fault-tolerant photonic quantum computing," *Quantum*, vol. 9, p. 1796, Jul. 2025, doi: 10.22331/q-2025-07-14-1796.

[113] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Math. Proc. Cambridge Phil. Soc.*, vol. 58, no. 3, pp. 497–520, Jul. 1962, doi: 10.1017/s0305004100036781.

[114] S. Maurya, A. Molavi, A. Albarghouthi, and S. Tannu, "A case for elastic quantum error correction decoders," in *Proc. 21st Eur. Conf. Comput. Syst.*, Edinburgh, U.K., Apr. 2026, pp. 514–531, doi: 10.1145/3767295.3803584.

[115] T.-H. Lin and C.-Y. Lai, "Union-intersection union-find for decoding depolarizing errors in topological codes," *IEEE J. Sel. Areas Inf. Theory*, vol. 6, pp. 163–175, 2025, doi: 10.1109/JSAIT.2025.3581810.

[116] Y. Wu and L. Zhong, "Fusion blossom: Fast MWPM decoders for QEC," in *Proc. IEEE Int. Conf. Quantum Comput. Eng. (QCE)*, Bellevue, WA, USA, Sep. 2023, pp. 928–938, doi: 10.1109/qce57702.2023.00107.

[117] J. Bausch, A. W. Senior, F. J. H. Heras, T. Edlich, A. Davies, M. Newman, C. Jones, K. Satzinger, M. Y. Niu, S. Blackwell, G. Holland, D. Kafri, J. Atalaya, C. Gidney, D. Hassabis, S. Boixo, H. Neven, and P. Kohli, "Learning high-accuracy error decoding for quantum processors," *Nature*, vol. 635, no. 8040, pp. 834–840, Nov. 2024, doi: 10.1038/s41586-024-08148-8.

[118] N. Liyanage, Y. Wu, S. Tagare, and L. Zhong, "FPGA-based distributed union-find decoder for surface codes," *IEEE Trans. Quantum Eng.*, vol. 5, pp. 1–18, 2024, doi: 10.1109/TQE.2024.3467271.

[119] F. Valentino, B. Branchini, D. Conficconi, D. Sciuto, and M. D. Santambrogio, "QUEKUF: An FPGA union find decoder for quantum error correction on the toric code," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 18, no. 3, pp. 1–26, Sep. 2025, doi: 10.1145/3733239.

[120] M. C. Löbl, S. X. Chen, S. Paesani, and A. S. Sørensen, "Breadth-first graph traversal union-find decoder," 2024, *arXiv:2407.15988*.

[121] I. Dinur, M.-H. Hsieh, T.-C. Lin, and T. Vidick, "Good quantum LDPC codes with linear time decoders," in *Proc. 55th Annu. ACM Symp. Theory Comput.*, Orlando, FL, USA, Jun. 2023, pp. 905–918, doi: 10.1145/3564246.3585101.

[122] C.-F. Kung, K.-Y. Kuo, and C.-Y. Lai, "Efficient approximate degenerate ordered statistics decoding for quantum codes via reliable subset reduction," 2024, *arXiv:2412.21118*.

[123] R. Acharya et al., "Suppressing quantum errors by scaling a surface code logical qubit," *Nature*, vol. 614, no. 7949, pp. 676–681, Feb. 2023, doi: 10.1038/s41586-022-05434-1.

IEEE Proof

[124] J. Majaniemi and E. S. Matekole, "Reducing quantum error correction overhead using soft information," *Quantum Sci. Technol.*, vol. 11, no. 2, Mar. 2026, Art. no. 025024, doi: 10.1088/2058-9565/ae4e55.
[125] Y. Zeng, W. Qin, Y.-H. Chen, C. Gneiting, and F. Nori, "Neural-network-based design of approximate gottesman-kitaev-preskill code," *Phys. Rev. Lett.*, vol. 134, no. 6, Feb. 2025, Art. no. 060601, doi: 10.1103/physrevlett.134.060601.
[126] A. Gong, S. Cammerer, and J. M. Renes, "Toward low-latency iterative decoding of QLDPC codes under circuit-level noise," 2024, *arXiv:2403.18901*.
[127] K. R. Ott, B. Hetényi, and M. E. Beverland, "Decision-tree decoders for general quantum LDPC codes," 2025, *arXiv:2502.16408*.
[128] W. Liao, Y. Suzuki, T. Tanimoto, Y. Ueno, and Y. Tokunaga, "WIT-greedy: Hardware system design of weighted ITerative greedy decoder for surface code," in *Proc. 28th Asia South Pacific Design Autom. Conf.*, Tokyo, Japan, Jan. 2023, pp. 209–215, doi: 10.1145/3566097.3567933.
[129] A. Holmes, M. R. Jokar, G. Pasandi, Y. Ding, M. Pedram, and F. T. Chong, "NISQ+: Boosting quantum computing power by approximating quantum error correction," in *Proc. ACM/IEEE 47th Annu. Int. Symp. Comput. Archit. (ISCA)*, May 2020, pp. 556–569, doi: 10.1109/ISCA45697.2020.00053.
[130] Y. Ueno, M. Kondo, M. Tanaka, Y. Suzuki, and Y. Tabuchi, "QULATIS: A quantum error correction methodology toward lattice surgery," in *Proc. IEEE Int. Symp. High-Perform. Comput. Archit. (HPCA)*, Seoul, South Korea, Apr. 2022, pp. 274–287, doi: 10.1109/HPCA53966.2022.00028.
[131] OpenQASM Project. (May 2024). *OpenQASM 3.1 Specification*. ver. 3.1.0. [Online]. Available: https://openqasm.com/versions/3.1/
[132] QIR Alliance. *Quantum Intermediate Representation (QIR) Specification*. ver. 2.0. Accessed: Aug. 29, 2026. [Online]. Available: https://github.com/qir-alliance/qir-spec/blob/main/specification/README.md
[133] IEEE Standards Association, *Guide for General Application of Hybrid Quantum-Classical Computing Technology*, Standard IEEE Project P3980, active PAR, Mar. 2026. [Online]. Available: https://standards.ieee.org/ieee/3980/12497/
[134] ISO/IEC, *Quantum Technologies—Quantum Computing—Hardware Benchmarking*, Standard ISO/IEC AWI 63712, stage 10.99, Jul. 2026. [Online]. Available: https://www.iso.org/standard/94929.html
[135] E. Giortamis, F. Romão, N. Tornow, and P. Bhatotia, "QOS: Quantum operating system," in *Proc. 19th USENIX Symp. Oper. Syst. Des. Implement. (OSDI)*, Boston, MA, USA, Jul. 2025, pp. 429–447. [Online]. Available: https://www.usenix.org/conference/osdi25/presentation/giortamis
[136] E. Takou and K. R. Brown, "Estimating decoding graphs and hypergraphs of memory quantum error-correction experiments," *Phys. Rev. A, Gen. Phys.*, vol. 112, no. 5, Nov. 2025, Art. no. 052414, doi: 10.1103/qcz4-nx4r.
[137] C. Guinn, S. Stein, E. Tureci, G. Avis, C. Liu, S. Krastanov, A. A. Houck, and A. Li, "Co-designed superconducting architecture for lattice surgery of surface codes with quantum interface routing card," 2023, *arXiv:2312.01246*.
[138] F. Zilk, A. Tundo, V. De Maio, and I. Brandic, "Breaking down quantum compilation: Profiling and identifying costly passes," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Kalamata, Greece, Jul. 2025, pp. 1–6, doi: 10.1109/ISVLSI65124.2025.11130327.
[139] R. Finkelstein, R. B.-S. Tsai, X. Sun, P. Scholl, S. Direkci, T. Gefen, J. Choi, A. L. Shaw, and M. Endres, "Universal quantum operations and ancilla-based read-out for tweezer clocks," *Nature*, vol. 634, no. 8033, pp. 321–327, Oct. 2024, doi: 10.1038/s41586-024-08005-8.
[140] S. Gicev, L. C. L. Hollenberg, and M. Usman, "Quantum computer error structure probed by quantum error correction syndrome measurements," *Phys. Rev. Res.*, vol. 6, no. 4, Dec. 2024, Art. no. 043249, doi: 10.1103/physrevresearch.6.043249.
[141] T. Tansuwannont, B. Pato, and K. R. Brown, "Adaptive syndrome measurements for shor-style error correction," *Quantum*, vol. 7, p. 1075, Aug. 2023, doi: 10.22331/q-2023-08-08-1075.
[142] M. DeCross, E. Chertkov, M. Kohagen, and M. Foss-Feig, "Qubit-reuse compilation with mid-circuit measurement and reset," *Phys. Rev. X*, vol. 13, no. 4, Dec. 2023, Art. no. 041057, doi: 10.1103/physrevx.13.041057.
[143] S. Veroni, M. Müller, and G. Giudice, "Optimized measurement-free and fault-tolerant quantum error correction for neutral atoms," *Phys. Rev. Res.*, vol. 6, no. 4, Dec. 2024, Art. no. 043253, doi: 10.1103/physrevresearch.6.043253.
[144] A. Anshu, N. P. Breuckmann, and C. Nirkhe, "NLTS Hamiltonians from good quantum codes," in *Proc. 55th Annu. ACM Symp. Theory Comput.*, Orlando, FL, USA, Jun. 2023, pp. 1090–1096, doi: 10.1145/3564246.3585114.
[145] N. Baspin and A. Krishna, "Quantifying nonlocality: How outperforming local quantum codes is expensive," *Phys. Rev. Lett.*, vol. 129, no. 5, Jul. 2022, Art. no. 050505, doi: 10.1103/physrevlett.129.050505.
[146] N. Berthusen, D. Devulapalli, E. Schoute, A. M. Childs, M. J. Gullans, A. V. Gorshkov, and D. Gottesman, "Toward a 2D local implementation of quantum low-density parity-check codes," *PRX Quantum*, vol. 6, no. 1, Jan. 2025, Art. no. 010306, doi: 10.1103/prxquantum.6.010306.
[147] N. Berthusen and D. Gottesman, "Partial syndrome measurement for hypergraph product codes," *Quantum*, vol. 8, p. 1345, May 2024, doi: 10.22331/q-2024-05-14-1345.
[148] S. Yang and R. Calderbank, "Spatially-coupled QLDPC codes," *Quantum*, vol. 9, p. 1693, Apr. 2025, doi: 10.22331/q-2025-04-07-1693.
[149] L. Pecorari, S. Jandura, G. K. Brennen, and G. Pupillo, "High-rate quantum LDPC codes for long-range-connected neutral atom registers," *Nature Commun.*, vol. 16, no. 1, p. 1111, Jan. 2025, doi: 10.1038/s41467-025-56255-5.
[150] T. J. Yoder, E. Schoute, P. Rall, E. Pritchett, J. M. Gambetta, A. W. Cross, M. Carroll, and M. E. Beverland, "Tour de gross: A modular quantum computer based on bivariate bicycle codes," 2025, *arXiv:2506.03094*.
[151] T. Hillmann, L. Berent, A. O. Quintavalle, J. Eisert, R. Wille, and J. Roffe, "Localized statistics decoding for quantum low-density parity-check codes," *Nature Commun.*, vol. 16, no. 1, Sep. 2025, Art. no. 8214, doi: 10.1038/s41467-025-63214-7.
[152] J. Roffe, L. Z. Cohen, A. O. Quintavalle, D. Chandra, and E. T. Campbell, "Bias-tailored quantum LDPC codes," *Quantum*, vol. 7, p. 1005, May 2023, doi: 10.22331/q-2023-05-15-1005.
[153] C. A. Pattison, A. Krishna, and J. Preskill, "Hierarchical memories: Simulating quantum LDPC codes with local gates," *Quantum*, vol. 9, p. 1728, May 2025, doi: 10.22331/q-2025-05-05-1728.
[154] D. R. W. Yost, M. E. Schwartz, J. Mallek, D. Rosenberg, C. Stull, J. L. Yoder, G. Calusine, M. Cook, R. Das, A. L. Day, E. B. Golden, D. K. Kim, A. Melville, B. M. Niedzielski, W. Woods, A. J. Kerman, and W. D. Oliver, "Solid-state qubits integrated with superconducting through-silicon vias," *npj Quantum Inf.*, vol. 6, no. 1, p. 59, Jul. 2020, doi: 10.1038/s41534-020-00289-8.
[155] C. Duckering, J. M. Baker, D. I. Schuster, and F. T. Chong, "Virtualized logical qubits: A 2.5D architecture for error-corrected quantum computing," in *Proc. 53rd Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2020, pp. 173–185, doi: 10.1109/MICRO50266.2020.00026.
[156] Z. Li, E. Gupta, F. Zhao, R. Banerjee, Y. Lu, T. Roy, A. Oriani, A. Vrajitoarea, S. Chakram, and D. I. Schuster, "A cascaded random access quantum memory," *Nature Phys.*, vol. 22, Sep. 2026, doi: 10.1038/s41567-026-03418-w.
[157] J. Ramette, J. Sinclair, N. P. Breuckmann, and V. Vuletić, "Fault-tolerant connection of error-corrected qubits with noisy links," *npj Quantum Inf.*, vol. 10, no. 1, Jun. 2024, Art. no. 58, doi: 10.1038/s41534-024-00855-4.
[158] M. Jessica Jeng, N. Vuk Maruszewski, M.-T. Lau, C. Selna, M. Gavrincea, K. N. Smith, and N. Hardavellas, "SEQC: Stratify-elaborate quantum compilation towards modular hybrid architectures," 2025, *arXiv:2501.08478*.
[159] A. Wu, Y. Ding, and A. Li, "QuComm: Optimizing collective communication for distributed quantum computing," in *Proc. 56th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Toronto, ON, Canada, Oct. 2023, pp. 479–493, doi: 10.1145/3613424.3614253.
[160] S. Li, X. Zhou, and Y. Feng, "Benchmarking quantum circuit transformation with QKNOB circuits," *IEEE Trans. Quantum Eng.*, vol. 6, pp. 1–15, 2025, doi: 10.1109/TQE.2025.3527399.
[161] M. Katsuda, K. Mitarai, and K. Fujii, "Simulation and performance analysis of quantum error correction with a rotated surface code under a realistic noise model," *Phys. Rev. Res.*, vol. 6, no. 1, Jan. 2024, Art. no. 013024, doi: 10.1103/physrevresearch.6.013024.
[162] J. A. Bravo-Montes, M. Bastante, G. Botella, A. del Barrio, and F. García-Herrero, "A methodology to select and adjust quantum noise models through emulators: Benchmarking against real backends," *EPJ Quantum Technol.*, vol. 11, no. 1, p. 71, Oct. 2024, doi: 10.1140/epjqt/s40507-024-00284-4.

IEEE Proof

[163] V. Tripathi, D. Kowsari, K. Saurav, H. Zhang, E. M. Levenson-Falk, and D. A. Lidar, "Benchmarking quantum gates and circuits," *Chem. Rev.*, vol. 125, no. 12, pp. 5745–5775, Jun. 2025, doi: 10.1021/acs.chemrev.4c00870.

[164] Y. Zhu, J. Cheng, B. Li, K. Liu, Y. Zhou, H. Wang, Y. Ding, and Z. Liang, "Hardware-aware calibration protocol for quantum computers," in *Proc. 52nd Annu. Int. Symp. Comput. Archit.*, Tokyo, Japan, Jun. 2025, pp. 241–256, doi: 10.1145/3695053.3731036.

[165] M. McEwen et al., "Removing leakage-induced correlated errors in superconducting quantum error correction," *Nature Commun.*, vol. 12, no. 1, p. 1761, Mar. 2021, doi: 10.1038/s41467-021-21982-y.

[166] S. Wang, Y. Yan, Z. Xia, and X.-B. Wang, "Effect of multiqubit correlated noise on surface-code thresholds," *Phys. Rev. A, Gen. Phys.*, vol. 112, no. 6, Dec. 2025, Art. no. 062419, doi: 10.1103/hjqn-lyh2.

[167] J. Hines and T. Proctor, "Pauli noise learning for mid-circuit measurements," *Phys. Rev. Lett.*, vol. 134, no. 2, Jan. 2025, Art. no. 020602, doi: 10.1103/physrevlett.134.020602.

[168] Z. Ni, S. Li, L. Zhang, J. Chu, J. Niu, T. Yan, X. Deng, L. Hu, J. Li, Y. Zhong, S. Liu, F. Yan, Y. Xu, and D. Yu, "Scalable method for eliminating residual ZZ interaction between superconducting qubits," *Phys. Rev. Lett.*, vol. 129, no. 4, Jul. 2022, Art. no. 040502, doi: 10.1103/physrevlett.129.040502.

[169] B. Gulácsi and G. Burkard, "Temporally correlated quantum noise in driven quantum systems with applications to quantum gate operations," *Phys. Rev. Res.*, vol. 7, no. 2, Apr. 2025, Art. no. 023073, doi: 10.1103/physrevresearch.7.023073.

[170] S. Raimund Maschek, J. Schwitalla, M. Franz, and W. Mauerer, "Make some noise! Measuring noise model quality in real-world quantum software," in *Proc. IEEE Int. Conf. Quantum Softw. (QSW)*, Helsinki, Finland, Jul. 2025, pp. 1–11, doi: 10.1109/QSW67625.2025.00010.

[171] D. Fan et al., "Calibrating quantum gates up to 52 qubits in a superconducting processor," *npj Quantum Inf.*, vol. 11, no. 1, Feb. 2025, Art. no. 33, doi: 10.1038/s41534-025-00983-5.

[172] A. Youssry, Y. Yang, R. J. Chapman, B. Haylock, F. Lenzini, M. Lobino, and A. Peruzzo, "Experimental graybox quantum system identification and control," *npj Quantum Inf.*, vol. 10, no. 1, p. 9, Jan. 2024, doi: 10.1038/s41534-023-00795-5.

[173] J. M. Scarpetta, J. H. Reina, and M. Hjorth-Jensen, "Machine learning non-Markovian two-level quantum noise spectroscopy," *Phys. Rev. Res.*, vol. 7, no. 4, Dec. 2025, Art. no. 043285, doi: 10.1103/2lzl-vpjd.

[174] H. Wang, P. Liu, Y. Liu, J. Gu, J. Baker, F. T. Chong, and S. Han, "DGR: Tackling drifted and correlated noise in quantum error correction via decoding graph re-weighting," 2023, *arXiv:2311.16214*.

[175] M. L. Dahlhauser and T. S. Humble, "Benchmarking characterization methods for noisy quantum circuits," *Phys. Rev. A, Gen. Phys.*, vol. 109, no. 4, Apr. 2024, Art. no. 042620, doi: 10.1103/physreva.109.042620.

[176] T. B. Adeniyi and S. A. P. Kumar, "Adaptive neural network for quantum error mitigation," *Quantum Mach. Intell.*, vol. 7, no. 1, Jan. 2025, Art. no. 13, doi: 10.1007/s42484-024-00234-4.

[177] X. Zhang, Z. Wu, G. A. L. White, Z. Xiang, S. Hu, Z. Peng, Y. Liu, D. Zheng, X. Fu, A. Huang, D. Poletti, K. Modi, J. Wu, M. Deng, and C. Guo, "Learning and forecasting open quantum dynamics with correlated noise," *Commun. Phys.*, vol. 8, no. 1, p. 29, Jan. 2025, doi: 10.1038/s42005-025-01944-2.

[178] L. Daguerre and M. Sarovar, "Real-time adaptation of quantum noise channel estimates," *Phys. Rev. A, Gen. Phys.*, vol. 111, no. 6, Jun. 2025, Art. no. 062609, doi: 10.1103/physreva.111.062609.

[179] A. Rahman, D. J. Egger, and C. Arenz, "Learning how to dynamically decouple by optimizing rotational gates," *Phys. Rev. Appl.*, vol. 22, no. 5, Nov. 2024, Art. no. 054074, doi: 10.1103/physrevapplied.22.054074.

[180] K.-Y. Kuo and C.-Y. Lai, "Generalized quantum data-syndrome codes and belief propagation decoding for phenomenological noise," *IEEE Trans. Inf. Theory*, vol. 71, no. 3, pp. 1824–1840, Mar. 2025, doi: 10.1109/TIT.2025.3529773.

[181] L. Shirizly, D. Meirom, M. Carroll, and H. Landa, "Feedforward suppression of readout-induced faults in quantum error correction," *Phys. Rev. A, Gen. Phys.*, vol. 112, no. 5, Nov. 2025, Art. no. L050602, doi: 10.1103/ght4-yqmb.

[182] K. Hémery, K. Ghanem, E. Crane, S. L. Campbell, J. M. Dreiling, C. Figgatt, C. Foltz, J. P. Gaebler, J. Johansen, M. Mills, and S. A. Moses, "Measuring the Loschmidt amplitude for finite-energy properties of the Fermi-Hubbard model on an ion-trap quantum computer," *PRX Quantum*, vol. 5, no. 3, Aug. 2024, Art. no. 030323, doi: 10.1103/PRXQuantum.5.030323.

[183] Y. Yu, K. Yan, D. Biswas, V. N. Zhang, B. Harraz, C. Noel, C. Monroe, and A. Kozhanov, "In situ midcircuit qubit measurement and reset in a single-species trapped-ion quantum computing system," *Phys. Rev. Res.*, vol. 7, no. 4, Dec. 2025, Art. no. 043355, doi: 10.1103/qfvd-93lw.

[184] C. Ryan-Anderson, J. G. Bohnet, K. Lee, D. Gresh, A. Hankin, J. P. Gaebler, D. Francois, A. Chernoguzov, D. Lucchetti, N. C. Brown, and T. M. Gatterman, "Realization of real-time fault-tolerant quantum error correcti," *Phys. Rev. X*, vol. 11, no. 4, Dec. 2021, Art. no. 041058, doi: 10.1103/PhysRevX.11.041058.

[185] J. Liu, A. Zang, M. Suchara, T. Zhong, and P. D. Hovland, "Hardware-software co-design for distributed quantum computing," in *Proc. 62nd ACM/IEEE Design Autom. Conf. (DAC)*, San Francisco, CA, USA, Jun. 2025, pp. 1–6, doi: 10.1109/DAC63849.2025.11132538.

[186] P. Pathumsoot, T. Tansuwannont, N. Benchasattabuse, R. Satoh, M. Hajdušek, P. Chaiwongkhot, S. Suwanna, and R. Van Meter, "Boosting end-to-end entanglement fidelity in quantum repeater networks via hybridized strategies," in *Proc. Int. Conf. Quantum Commun., Netw., Comput. (QCNC)*, Kanazawa, Japan, Jul. 2024, pp. 49–56, doi: 10.1109/QCNC62729.2024.00054.

[187] N. D. Stuyck, A. E. Seedhouse, S. Serrano, T. Tanttu, W. Gilbert, J. Y. Huang, F. Hudson, K. M. Itoh, A. Laucht, W. H. Lim, and C. H. Yang, "Silicon spin qubit noise characterization using real-time feedback protocols and wavelet analysis," *Appl. Phys. Lett.*, vol. 124, no. 11, Mar. 2024, Art. no. 114003, doi: 10.1063/5.0179958.

[188] A. deMarti iOlius, I. E. Martinez, J. Roffe, and J. E. Martinez, "An almost-linear time decoding algorithm for quantum LDPC codes under circuit-level noise," *npj Quantum Inf.*, vol. 12, p. 152, Jun. 2026, doi: 10.1038/s41534-026-01292-1.

[189] M. Lange, P. Havström, B. Srivastava, I. Bengtsson, V. Bergentall, K. Hammar, O. Heuts, E. Van Nieuwenburg, and M. Granath, "Data-driven decoding of quantum error correcting codes using graph neural networks," *Phys. Rev. Res.*, vol. 7, no. 2, May 2025, Art. no. 023181, doi: 10.1103/PhysRevResearch.7.023181.

[190] M. Cain, C. Zhao, H. Zhou, N. Meister, J. P. B. Ataides, A. Jaffe, D. Bluvstein, and M. D. Lukin, "Correlated decoding of logical algorithms with transversal gates," *Phys. Rev. Lett.*, vol. 133, no. 24, Dec. 2024, Art. no. 240602, doi: 10.1103/PhysRevLett.133.240602.

[191] P. Coote, R. Dimov, S. Maity, G. S. Hartnett, M. J. Biercuk, and Y. Baum, "Resource-efficient context-aware dynamical decoupling embedding for arbitrary large-scale quantum algorithms," *PRX Quantum*, vol. 6, no. 1, Feb. 2025, Art. no. 010332, doi: 10.1103/PRXQuantum.6.010332.

[192] M. Saghafi and L. Mili, "Predictive time-series analysis of single-qubit quantum circuit outcomes for a superconducting quantum computer: Forecasting error patterns," *IEEE Access*, vol. 13, pp. 40115–40132, 2025, doi: 10.1109/ACCESS.2025.3546259.

[193] X. Wu, Y. J. Joshi, H. Yan, G. Andersson, A. Anferov, C. R. Conner, B. Karimi, A. M. King, and S. Li, "Mitigating cosmic-ray-like correlated events with a modular quantum processor," *Phys. Rev. Appl.*, vol. 24, no. 4, Oct. 2025, Art. no. 044022, doi: 10.1103/4ctq-r6w6.

[194] R. S. Gupta, N. Sundaresan, T. Alexander, C. J. Wood, S. T. Merkel, M. B. Healy, M. Hillenbrand, T. Jochym-O'Connor, J. R. Wootton, T. J. Yoder, A. W. Cross, M. Takita, and B. J. Brown, "Encoding a magic state with beyond break-even fidelity," *Nature*, vol. 625, no. 7994, pp. 259–263, Jan. 2024, doi: 10.1038/s41586-023-06846-3.

[195] L. Pecorari and G. Pupillo, "Quantum low-density parity-check codes for erasure-biased atomic quantum processors," *Phys. Rev. A, Gen. Phys.*, vol. 112, no. 5, Nov. 2025, Art. no. 052417, doi: 10.1103/mgkt-ctv8.

[196] J. Kalloor, M. Weiden, E. Younis, J. Kubiatowicz, B. De Jong, and C. Iancu, "Quantum hardware roofline: Evaluating the impact of gate expressivity on quantum processor design," in *Proc. IEEE Int. Conf. Quantum Comput. Eng. (QCE)*, Montreal, QC, Canada, Sep. 2024, pp. 805–816, doi: 10.1109/QCE60285.2024.00100.

[197] C. G. Yale, R. Rines, V. Omole, B. Thotakura, A. D. Burch, M. N. H. Chow, M. Ivory, and D. Lobser, "Noise-aware circuit compilations for a continuously parameterized two-qubit gateset," *Phys. Rev. Appl.*, vol. 24, no. 2, Aug. 2025, Art. no. 024057, doi: 10.1103/3cmg-5rk7.

IEEE PROOF

[198] F. Wagner, D. J. Egger, and F. Liers, "Optimized noise suppression for quantum circuits," *INFORMS J. Comput.*, vol. 37, no. 1, pp. 22–41, Jan. 2025, doi: 10.1287/ijoc.2024.0551.

[199] B. Wang, P. Wang, Y. Zhang, X. Wang, Z. Zhou, L. Bai, and Y. Wang, "Towards dynamic spatial–temporal graph learning: A decoupled perspective," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2024, vol. 38, no. 8, pp. 9089–9097, doi: 10.1609/aaai.v38i8.28759.

[200] J. Zou, S. Bosco, and D. Loss, "Spatially correlated classical and quantum noise in driven qubits," *npj Quantum Inf.*, vol. 10, p. 46, Apr. 2024, doi: 10.1038/s41534-024-00842-9.

[201] E. Y. Egorova, A. S. Kazmina, I. A. Simakov, I. N. Moskalenko, N. N. Abramov, and D. A. Kalacheva, "Three-mode tunable coupler for superconducting two-qubit gates," *Phys. Rev. Appl.*, vol. 23, no. 6, Jun. 2025, Art. no. 064056, doi: 10.1103/2h4m-mg2p.

[202] T.-M. Li, J. C. Zhang, B. J. Chen, K. Huang, H. T. Liu, Y. X. Xiao, C. L. Deng, G. H. Liang, C. T. Chen, and Y. Liu, "High-precision pulse calibration of tunable couplers for high-fidelity two-qubit gates in superconducting quantum processors," *Phys. Rev. Appl.*, vol. 23, no. 2, Feb. 2025, Art. no. 024059, doi: 10.1103/PhysRevApplied.23.024059.

[203] H. Zhang, C. Ding, D. K. Weiss, Z. Huang, Y. Ma, C. Guinn, S. Sussman, S. P. Chitta, and D. Chen, "Tunable inductive coupler for high-fidelity gates between fluxonium qubits," *PRX Quantum*, vol. 5, no. 2, May 2024, Art. no. 020326, doi: 10.1103/PRXQuantum.5.020326.

[204] X. Deng, W. Zheng, X. Liao, H. Zhou, Y. Ge, J. Zhao, D. Lan, X. Tan, Y. Zhang, S. Li, and Y. Yu, "Long-range (*ZZ*) interaction via resonator-induced phase in superconducting qubits," *Phys. Rev. Lett.*, vol. 134, no. 2, Jan. 2025, Art. no. 020801, doi: 10.1103/PhysRevLett.134.020801.

[205] N. Baspin and D. Williamson, "Wire codes," *Quantum*, vol. 10, p. 2083, Apr. 2026, doi: 10.22331/q-2026-04-24-2083.

[206] G. P. Gehér, O. Crawford, and E. T. Campbell, "Tangling schedules eases hardware connectivity requirements for quantum error correction," *PRX Quantum*, vol. 5, no. 1, Mar. 2024, Art. no. 010348, doi: 10.1103/PRXQuantum.5.010348.

[207] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, "Fundamental limits of repeaterless quantum communications," *Nat. Commun.*, vol. 8, p. 15043, Apr. 2017, doi: 10.1038/ncomms15043.

[208] M. M. Wolf, D. Pérez-García, and G. Giedke, "Quantum capacities of bosonic channels," *Phys. Rev. Lett.*, vol. 98, no. 13, Mar. 2007, Art. no. 130501, doi: 10.1103/PhysRevLett.98.130501.

[209] K. J. Wo, G. Avis, F. Rozpedek, M. F. Mor-Ruiz, G. Pieplow, T. Schröder, L. Jiang, A. S. Sørensen, and J. Borregaard, "Resource-efficient fault-tolerant one-way quantum repeater with code concatenation," *npj Quantum Inf.*, vol. 9, p. 123, Dec. 2023, doi: 10.1038/s41534-023-00792-8.

[210] B. Li, K. Goodenough, F. Rozpędek, and L. Jiang, "Generalized quantum repeater graph states," *Phys. Rev. Lett.*, vol. 134, no. 19, May 2025, Art. no. 190801, doi: 10.1103/PhysRevLett.134.190801.

[211] L. Pettersson and A. S. Sørensen, "Long-distance quantum communication using concatenated ring graph codes," *Phys. Rev. Appl.*, vol. 24, no. 4, Oct. 2025, Art. no. 044090, doi: 10.1103/fxcg-xxry.

[212] M. Bilal Yaseen, F. Wan, S. Asif, and J. Wan, "QFL-SecEdge: A quantum federated learning framework with postquantum cryptographic task distribution for ultrareliable mission-critical 6G-IoT networks," *IEEE Internet Things J.*, vol. 13, no. 13, pp. 29513–29526, Jul. 2026, doi: 10.1109/JIOT.2026.3685072.

[213] M. B. Yaseen, F. Wan, N. M. Murad, R. Song, and F. Siddique, "A novel framework for EMI-resilient UAV-RIS communications using federated multi-agent reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 75, no. 9, pp. 21374–21387, Sep. 2026, doi: 10.1109/TVT.2026.3689153.

[214] S. Omanakuttan, V. Buchemmavari, J. A. Gross, I. H. Deutsch, and M. Marvian, "Fault-tolerant quantum computation using large spin-cat codes," *PRX Quantum*, vol. 5, no. 2, Jun. 2024, Art. no. 020355, doi: 10.1103/PRXQuantum.5.020355.

[215] V. V. Sivak, A. Eickbusch, B. Royer, S. Singh, I. Tsioutsios, S. Ganjam, A. Miano, B. L. Brock, A. Z. Ding, L. Frunzio, S. M. Girvin, R. J. Schoelkopf, and M. H. Devoret, "Real-time quantum error correction beyond break-even," *Nature*, vol. 616, no. 7955, pp. 50–55, Apr. 2023, doi: 10.1038/s41586-023-05782-6.

[216] A. L. Grimsmo and S. Puri, "Quantum error correction with the Gottesman-Kitaev-Preskill code," PRX Quantum, vol. 2, no. 2, Jun. 2021, Art. no. 020101, doi: 10.1103/PRXQuantum.2.020101.

[217] J. Hastrup and U. L. Andersen, "Analysis of loss correction with the Gottesman-Kitaev-Preskill code," *Phys. Rev. A, Gen. Phys.*, vol. 108, no. 5, Nov. 2023, Art. no. 052413, doi: 10.1103/PhysRevA.108.052413.

[218] F. Grünenfelder, A. Boaron, G. V. Resta, and M. Perrenoud, "Fast single-photon detectors and real-time key distillation enable high secret-key-rate quantum key distribution systems," *Nature Photon.*, vol. 17, pp. 422–426, May 2023, doi: 10.1038/s41566-023-01168-2.

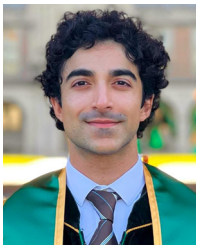
[219] F. Thiele, N. Lamberty, T. Hummel, N. A. Lange, L. M. Procopio, A. Barua, S. Lengeling, V. Quiring, C. Eigner, C. Silberhorn, and T. J. Bartley, "Cryogenic feedforward of a photonic quantum state," *Optica*, vol. 12, no. 5, pp. 720–727, May 2025, doi: 10.1364/optica.551287.

[220] S. K. Borah, A. K. Pradhan, N. Raveendran, M. Pacenti, and B. Vasić, "Fault tolerant decoding of QLDPC-GKP codes with circuit level soft information," in *Proc. IEEE Int. Conf. Quantum Comput. Eng. (QCE)*, Albuquerque, NM, USA, Aug. 2025, pp. 2505–2514, doi: 10.1109/QCE65121.2025.00271.

[221] M. Iuliano, M. C. Slater, A. J. Stolk, and M. J. Weaver, "Qubit teleportation between a memory-compatible photonic time-bin qubit and a solid-state quantum network node," *npj Quantum Inf.*, vol. 10, p. 107, Nov. 2024, doi: 10.1038/s41534-024-00910-0.

[222] T. H. Haug, A. F. Kockum, and R. Van Laer, "Heralding entangled optical photons from a microwave quantum processor," *Phys. Rev. Appl.*, vol. 22, no. 3, Sep. 2024, Art. no. 034068, doi: 10.1103/PhysRevApplied.22.034068.

[223] T. Müller, T. Alexander, M. E. Beverland, M. Bühler, B. R. Johnson, T. Maurer, and D. Vandeth, "Improved belief propagation is sufficient for real-time decoding of quantum memory," 2025, *arXiv:2506.01779*.

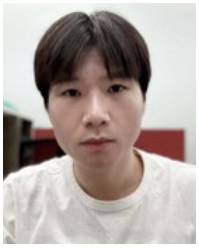


MUZAMMIL-E-HAQUE MAHMUD (Member, IEEE) received the B.S. and M.S. degrees in computer science from Washington University in St. Louis, St. Louis, MO, USA, in 2022. He has professional experience in artificial intelligence, machine learning, and software engineering. His research interests include artificial intelligence, quantum computing, quantum networking, quantum entanglement, quantum memory, quantum error correction, and security.

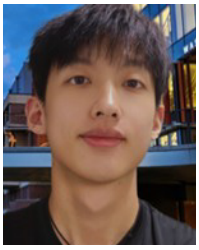


RAJ JAIN (Life Fellow, IEEE) received the B.E. degree in electrical engineering from APS University, India, the M.E. degree from Indian Institute of Science, Bengaluru, India, and the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA, USA, with a focus on computer science. He is currently the Barbara J. and Jerome R. Cox, Jr., Professor with Washington University in St. Louis, St. Louis, MO, USA. He co-founded Nayna Networks Inc., and has held significant roles in academia and industry. He is known as one of the most cited authors in computer science and is the author of *The Art of Computer Systems Performance Analysis*. He is a fellow of ACM and AAAS. He has received the 2018 James B. Eads Award and the 2017 ACM SIGCOMM Lifetime Achievement Award.

IEEE Proof



ZEBO YANG (Member, IEEE) received the M.E. degree in computer engineering from Waseda University, Tokyo, Japan, in 2019, and the Ph.D. degree in computer science from Washington University in St. Louis, St. Louis, MO, USA, in 2025. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA. From 2011 to 2017, he held technical roles at Tencent, Baidu, and DJI. From 2019 to 2025, he was a Research Assistant with Washington University in St. Louis. His research interests include quantum computing, quantum networks, security, and machine learning.



CHENLIANG TIAN (Student Member, IEEE) received the B.S. degree in computing from Macao Polytechnic University, Macao, China, in 2018. He is currently pursuing the Ph.D. degree in computer science with Washington University in St. Louis. Since 2025, he has been a Research Assistant with Washington University in St. Louis. His research interests include quantum computing and networking.



RAMANA KOMPELLA (Member, IEEE) received the B.Tech. degree in computer science and engineering from Indian Institute of Technology Bombay, Mumbai, India, in 1999, the M.S. degree in computer science and engineering from Stanford University, Stanford, CA, USA, in 2001, and the Ph.D. degree in computer science and engineering from the University of California, San Diego, CA, in 2007. He is currently a Distinguished Engineer and the Head of Research with Cisco's Emerging Tech and Incubation Group, where he leads university research collaborations. His research interests include data centers, cloud performance optimization, and router algorithmics. He has been an ACM Member, since 2007. He received the NSF CAREER Award, in 2011, and has received best paper awards at conferences, including ACM SOCC.



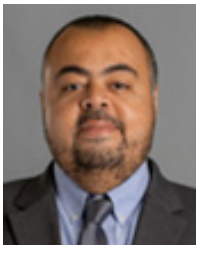
REZA NEJABATI (Senior Member, IEEE) is currently the Head of Quantum Research and Quantum Laboratory, Cisco, where he focuses on advancing quantum networking. Previously, he was the Chair Professor of networks and the Head of the High-Performance Networks Group, University of Bristol, U.K. He is also a Visiting Professor and the Cisco Chair with the Cisco Center for Intent-Based Networking, Curtin University, Australia. He established globally recognized research in autonomous and quantum networks, co-founded Zeetta Networks Ltd., and received the IEEE Charles Kao Award, in 2016, for contributions to 5G, smart cities, quantum communications, and future internet experimentation.



ENEET KAUR (Member, IEEE) received the Ph.D. degree from Louisiana State University, in 2020, with a focus on quantum information and communications. She is currently a Researcher with the Quantum Research Group, Cisco. She then spent a year and a half at the Institute of Quantum Computing, University of Waterloo, delving into quantum repeater designs. Then, she spent a year working in the College of Optical Sciences in Arizona on quantum networks. Her current work focuses on applications of quantum networks to distributed quantum computing.



AIMAN ERBAD (Senior Member, IEEE) received the B.Sc. degree in computer engineering from the University of Washington, Seattle, WA, USA, in 2004, the master's degree in embedded systems and robotics from the University of Essex, Colchester, U.K., in 2005, and the Ph.D. degree in computer science from The University of British Columbia, Vancouver, BC, Canada, in 2012. He is currently a Professor of computer engineering and the Vice President for Research and Graduate Studies with Qatar University, Doha, Qatar. His research interests include cloud and edge computing, edge intelligence, networks intelligence, the Internet of Things, and blockchains. He is also an editor and the program chair for various journals and conferences and a Senior Member of ACM.



MOHAMED M. ABDALLAH (Senior Member, IEEE) received the B.Sc. degree from Cairo University, Giza, Egypt, in 1996, and the M.Sc. and Ph.D. degrees from the University of Maryland at College Park, College Park, MD, USA, in 2001 and 2006, respectively. From 2006 to 2016, he held academic and research positions with Cairo University and Texas A&M University at Qatar, Doha, Qatar. He is currently a Founding Faculty Member with the rank of Professor with the College of Science and Engineering, Hamad Bin Khalifa University, Doha. He has published more than 150 papers in journals and conferences, four book chapters, and co-invented four patents.

IEEE PROOF