Progressive Optimization of Width-Indexed CNOT–RZ Circuits in the MatrixMul LV16 Benchmark

@jackylee0424, @gnuchev, @edi3on, @runeape-sats

Abstract. Block encoding supplies a data-input primitive for qubitization and quantum singular-value transformation. Arbitrary rotation synthesis dominates the modeled fault-tolerant cost. This benchmark encodes that cost with a 64× volume weight. The public specification defines its target operationally, through a per-width reference oracle and a probe contract. It does not fix the block-encoding tuple in closed form. Claims here are scoped to that operational contract. We report a 7.30× decrease in public benchmark score across a width-indexed target family. The progression used declared-width reduction, a 17-qubit minimum-width rebuild, shared-parity scheduling, and a small closed parity-trace search. The final circuit was validated against the benchmark's public probe suite at width 17. It has score 38,782.1608. The benchmark also exhibits numerical path dependence in its matrix-product-state (MPS) validator. The 17-qubit rebuild era accounts for about 99.77% of the total baseline-to-final score reduction. All later refinements equal about 1.41% of the rebuild-era score. The final trace step improved the previous best by about 0.010%. These proportions place width and rebuild engineering at the center of the result.

1 Introduction

Representing a 2^{16}-dimensional operator as a circuit changes the data interface for large linear algebra. A block encoding embeds a scaled matrix in a projected block of a unitary. Qubitization and quantum singular-value transformation then consume that representation [1, 2]. Circuit synthesis therefore links structured matrix descriptions to quantum matrix arithmetic. FABLE gives a related route from matrix data to approximate block-encoding circuits [4].

A dense classical comparison sets the scale without implying a runtime. For $N = 65{,}536$, dense real GEMM requires $2N^3 = 5.6295 \times 10^{14}$ floating-point operations. This is about 562.95 TFLOP of total work. Three dense complex128 $N \times N$ operands occupy about 192 GiB before workspace. The benchmark operator is sparse. Structure-exploiting classical baselines would therefore do far less work than dense GEMM.

Dense execution at this scale is a systems-engineering problem. The 192 GiB of operands exceeds the memory of a single conventional accelerator. It may require a large-memory server or a multi-accelerator node. Beyond a single node the distributed HPC stack applies, with interconnects, scheduling, communication, and data movement. The structured sparse target admits much cheaper classical operations than dense materialization. The calculation is therefore an illustrative scale comparison, not a runtime baseline. The quantum representation concentrates the target description into 17 logical qubits at minimum width. The benchmark score is only a circuit-cost proxy under a fault-tolerant weighting. It is not a demonstrated end-to-end runtime advantage, and we make no quantum-speedup claim.

The benchmark score is

$$S = q\sqrt{VD}, \tag{1}$$

where q is declared width. Weighted volume V charges 64 for each R_Z and one for each local Clifford or CNOT. Weighted depth D charges 16 time units for each R_Z and one for each local Clifford or CNOT. A CNOT on non-adjacent wires charges an additional 6 per unit of excess distance in V, and routed CNOTs likewise expand D. Among the frontier rows below, only the 42-qubit baseline incurs this routing term. This model reflects the cost of approximating arbitrary rotations in a fault-tolerant gate set [3]. The public score fell from 283,024.0672 to 38,782.1608, or 7.30×.

1.1 Target specification

The intended target is a block encoding of a 2^{16}-dimensional sparse operator. The general acceptance form uses a width-specific logical isometry $W_q : \mathbb{C}^{2^{16}} \to (\mathbb{C}^2)^{\otimes q}$,

$$\left\| A_q - \alpha_q W_q^\dagger U_q W_q \right\| \le \epsilon_q, \tag{2}$$

where U_q is the candidate unitary. The isometry W_q carries the code-space and workspace convention. The 42-qubit register comprises 32 two-rail matrix-ladder qubits and 10 block-encoding workspace qubits. A plain zero-ancilla projector $(\langle 0^a| \otimes I)$ with $a = q - 16$ therefore does not describe that layout.

The public materials define the target operationally. They fix a per-width reference instruction stream and centered-angle generator, together with the probe contract below. They do not define A_q, α_q, ϵ_q, or W_q in closed form. Width enters the angle computation. Different declared widths therefore define different reference oracles, not implementations of one unchanged unitary. The public metadata describes a sparse, Laplacian-style ladder structure [8].

The paper's claims are scoped accordingly. This work performs same-width oracle-circuit optimization under the operational probe contract. Eq. (2) is the intended reading of the benchmark. Its closed-form specification is an open item, restated in the Limitations.

Operational acceptance uses 9,024 deterministic numerical test states. These are probes, not measurement shots. Every qubit receives one state from $\{0, 1, +, -, i, -i\}$. Both infidelity and norm difference must not exceed 10^{-8}. The test is a finite, structured regression suite visible during optimization. It is not an operator-equivalence certificate. At width 17, the full state dimension is $2^{17} = 131{,}072$, which exceeds the probe count.

Conventions are fixed here for the remainder of the paper. Qubit labels follow OpenQASM indices in ascending MPS-site order, with q_0 first. The first CNOT operand is control and the second is target. OpenQASM uses $R_Z(\theta) = \mathrm{diag}(e^{-i\theta/2}, e^{i\theta/2})$. The verifier's normalized fidelity ignores global phase. Gates are drawn from $\{H, X, Y, Z, R_Z, \mathrm{CX}\}$, with CNOT synonymous with CX.

2 Related Work

Block encodings connect circuit representations to broader quantum algorithms. Low and Chuang introduced qubitization for Hamiltonian simulation [1]. Gilyén, Su, Low, and Wiebe developed quantum singular-value transformation and matrix arithmetic [2]. Camps and Van Beeumen proposed FABLE for approximate block-encoding circuit generation [4]. Ross and Selinger analyzed ancilla-free Clifford+T approximation of z rotations [3]. That synthesis cost motivates the benchmark's rotation weight.

Parity networks provide the closest circuit-synthesis context. Amy, Azimzadeh, and Mosca study CNOT-phase circuits and their CNOT complexity [6]. Patel, Markov, and Hayes synthesize linear reversible maps over GF(2) [7]. Our local search also evolves binary linear maps. It differs by attaching required phase events to replay boundaries and demanding closed traces. No standard-toolchain baseline has yet been run under the same gate set and objective. That comparison remains open.

The trace representation draws on the Dialog abstraction of Khattar and coauthors [5]. Dialog stores ordered invertible transitions and supports replay with implicit access. We inherit only that structural abstraction. We do not inherit its extended-Euclidean construction or its storage claims.

3 Progressive Synthesis of the Frontier

The frontier advanced by changing its main optimization variable over time. Width came first. A minimum-width rebuild then supplied nearly all remaining score reduction. Parity scheduling and numerical stabilization refined that base. A Clifford-tableau gauge and a linear Dialog trace produced the current frontier.

3.1 Declared-width reduction

The width-reduction era instantiated separate members of the target family. @jackylee0424 established the 42-qubit generated baseline. @runeape-sats then constructed direct targets at widths 41, 40, 36, 35, 34, and 32. Each generator used its declared width inside the centered-angle domains. These rows therefore compare benchmark scores across distinct targets. Each width has its own direct generator.

The 32-qubit construction used one preparation layer and four deterministic rounds. Each round applied per-wire phases, adjacent parity gadgets, and fixed logical mixers. Its circuit contained 598 gates. Its score was 119,818.9568.

3.2 Minimum-width reconstruction

The 17-qubit rebuild by @edi3on created the largest frontier step. It declared the allowed minimum width and rebuilt all four rounds for that target. Each round applied centered single-wire phases on 17 wires. It also applied one parity gadget on each adjacent pair. Fixed logical mixers remained in place.

The rebuild used 61 Hadamards, 154 rotations, and 128 nearest-neighbor CNOTs. It reduced weighted volume to 10,045 and weighted depth to 533. Its score was 39,335.7556. This stage, rather than the later local search, drove the headline reduction.

3.3 Parity networks and numerical stabilization

@gnuchev jointly synthesized the last two parity gadgets. Shared-control networks exposed both adjacent ZZ parities on a shorter critical path. Successive fourteen-, ten-, and eight-CNOT tails traded algebraic cost against MPS behavior. Identity CNOT pairs acted as verifier-specific gauge resets. They are numerical-path findings, not general synthesis rules.

The eight-CNOT tail reached score 38,860.8250. It kept 154 rotations while reducing weighted depth to 520. Its public-probe margin was narrow, showing that exact binary parity reasoning did not predict bounded-bond MPS behavior.

3.4 Tableau gauge and linear trace

@jackylee0424 next tracked CNOT and Hadamard action in a signed Clifford tableau. The two parity phases shared one weighted-depth layer. An explicit identity gauge retained numerical behavior. This step reached 38,786.0207.

The current refinement removes the tableaus from the diagonal subproblem. The internal name is the *linear Dialog trace*. Claims below use the fuller name *Dialog-inspired closed parity-trace search*. It exposes the same phase boundary with a smaller state description. It also makes a trailing $H; H$ identity pair auditable and removable.

Contributions. @runeape-sats developed the declared-width reduction sequence. @edi3on rebuilt the target at the 17-qubit minimum. @gnuchev developed shared parity networks and MPS-oriented numerical stabilization.

@jackylee0424 established the baseline, then developed the Clifford-tableau gauge and closed parity trace.

4 The Trace Method in Depth

Dialog-inspired closed parity-trace search turns a local gate-word problem into a replay problem over GF(2). The selected region covers lanes q_{14}, q_{15}, q_{16}. It must expose two adjacent parities and then restore the original basis. The following mixer is emitted only after closure.

4.1 State model and search

Let $x = [x_{14}, x_{15}, x_{16}]^{\mathsf{T}}$. A replay boundary stores $B_k \in \mathrm{GF}(2)^{3\times 3}$, with $B_0 = I$. Row i gives the parity currently held on lane i. Each nearest-neighbor CNOT is an invertible row operation. Replay updates the state as $B_{k+1} = E_k B_k$.

We name four transitions

$$a = \mathrm{CX}(14,15), \qquad A = \mathrm{CX}(15,14),$$
$$b = \mathrm{CX}(15,16), \qquad B = \mathrm{CX}(16,15).$$

The search enumerated words of length four through eight over $\{a, A, b, B\}$. B does not appear in the selected trace but belongs to the enumerated space. It rejected words that failed closure. It queried every remaining boundary for parity masks 110 and 011. Phase events were inserted only when both masks were available. The search closed 3,608 words and rendered 2,092 distinct score-eligible candidates.

4.2 Selected trace

The selected trace is

$$a\ a\ b\ A\ [L\ R]\ b\ A, \tag{3}$$

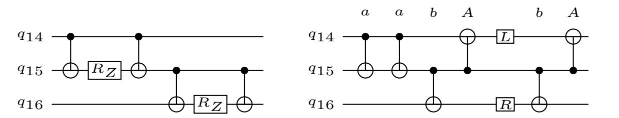
where L and R are the two required rotations. Table 1 shows the replay. Both rotations occur at boundary four. Figure 1 contrasts the tail before and after the rewrite.

k	Event	lane 14	lane 15	lane 16
0	start	x_{14}	x_{15}	x_{16}
1–2	$a\,a$	x_{14}	x_{15}	x_{16}
3	b	x_{14}	x_{15}	$x_{15} \oplus x_{16}$
4	A	$x_{14} \oplus x_{15}$	x_{15}	$x_{15} \oplus x_{16}$
4	L, R	rotate	—	rotate
5	b	$x_{14} \oplus x_{15}$	x_{15}	x_{16}
6	A	x_{14}	x_{15}	x_{16}

Table 1: Replay-boundary evolution of the three parity rows.

4.3 Correctness invariants

Three invariants establish the ideal parity skeleton. First, the two requested parities appear simultaneously at boundary four. An R_Z on each exposed lane therefore applies the corresponding parity phase. Second, replay closes with $B_6 = I$. The product of the transition matrices is the identity. Third, the $H; R_Z; H$ mixer is emitted after closure. It never enters the parity-row model.



(a) serial gadgets, two layers (b) closed trace, one layer

Figure 1: Three-wire tail before and after the rewrite. Panel (a) runs two serial CX; R_Z; CX parity gadgets, on (q_{14}, q_{15}) and then (q_{15}, q_{16}), so the two rotations sit in separate layers. Panel (b) is the closed trace of Eq. (3); letters match Table 1, and both rotations L and R occur at boundary four in one layer.

Frontier era	Lead	q	V	D	Score
Generated baseline	@jackylee0424	42	28,470	1,595	283,024.0672
Declared-width reduction	@runeape-sats	32	17,860	785	119,818.9568
Minimum-width rebuild	@edi3on	17	10,045	533	39,335.7556
Shared parity and stabilization	@gnuchev	17	10,049	520	38,860.8250
Clifford-tableau gauge	@jackylee0424	17	10,049	518	38,786.0207
Closed parity trace	@jackylee0424	17	10,047	518	38,782.1608

Table 2: Progression of the public benchmark frontier. The Lead column credits the GitHub username whose accepted circuit ended each era.

Implicit access avoids constructing the full unitary. If a lane holds $p(x) = v \cdot x$, then an R_Z on that lane adds the phase indexed by v. Restoring $B = I$ returns each computational basis bit to its original lane. The argument proves the unbiased local rewrite within the stated circuit model.

Two retained features concern the validator instead of ideal synthesis. The leading $a\,a$ pair is exactly the identity but changes the MPS update path. A -2.262829×10^{-6} bias is also applied to one rotation angle. Shifting one R_Z by δ changes its unitary by at most $2|\sin(\delta/4)| \simeq 1.13 \times 10^{-6}$ in operator norm.

The error budget must be stated in the verifier's own metrics. Operator-norm distance is not the quantity the probe contract bounds. Worst-case pure-state infidelity from the bias alone is $\sin^2(\delta/2) \simeq 1.28 \times 10^{-12}$. That lies below the 10^{-8} infidelity threshold in exact arithmetic. The exact state-norm change from an R_Z perturbation is zero, because R_Z is unitary and diagonal. The bias contributes nothing to the norm tolerance in exact arithmetic. Remaining observed error is attributable to structural differences, bounded-bond MPS truncation, and floating-point execution paths. The biased circuit is approximate rather than exactly equivalent, but the angle perturbation alone satisfies the infidelity threshold in exact arithmetic. Both retained features remain verifier-specific numerical stabilizers.

5 Results

The score reduction is dominated by declared width and the 17-qubit rebuild.

Cross-width rows compare different members of the width-indexed target family. Same-width rows refine the width-17 target. Table 2 gives an era-level frontier. Every score row is recomputable as $q\sqrt{VD}$.

The authors contributed complementary levers. @runeape-sats reduced declared width across the family. @edi3on supplied the minimum-width construction. @gnuchev shortened the parity critical path and studied MPS stability. @jackylee0424 supplied the tableau and trace refinements.

The contribution proportions quantify those roles. The 17-qubit rebuild era accounts for about 99.77% of the baseline-to-final reduction. All post-rebuild refinements combined equal about 1.41% of the rebuild-era score. The final trace step improved the previous best by about 0.010%. Its value is a current frontier technique, not the driver of the 7.30× headline.

The final circuit contains 345 gates at ordinary depth 83. Its gate counts are 61 Hadamards, 154 rotations, and 130 CNOTs. Weighted volume is 10,047 and weighted depth is 518. These values give score 38,782.1608 through Eq. (1). The circuit was validated against the benchmark's public probe suite at declared width 17.

Rotations consume 9,856 of the final weighted volume. The remaining terms are 130 from CNOTs and 61 from Hadamards. Rotations therefore fix about 98% of observed V. This decomposition gives only a conditional ceiling on Clifford-only gains. Under the benchmark weights, at fixed width, and near the observed depth, deleting Clifford volume changes the score only through $\sqrt{V}$. It is not a joint lower bound on score. Clifford changes can also alter scheduling depth.

6 Discussion and Future Work

The next search should expand trace space without losing exact replay checks. Meet-in-the-middle search can index partial words by boundary matrix. Boundary annotations can record available parities and scheduling time. Larger diagonal CNOT-phase regions can share parity work across more than two rotations. Commuting transitions and identity subtraces can be quotiented algebraically. Numerical gauges should remain explicit search variables.

Larger score changes require algorithm-level rotation elimination. Rotation folding, phase-polynomial simplification, or a different block-encoding construction could reduce the fixed R_Z count. These changes must preserve the declared-width target's centered angles or budget their approximation. Standard synthesis tools should also be evaluated with the same gates, connectivity, weights, and acceptance suite.

The contest format contributes a useful methodology. A public cost function turns circuit choices into comparable hypotheses. A visible lineage preserves which optimization lever changed. Repeated validation exposes numerical behavior that symbolic parity checks cannot predict. The same visibility also creates the limitations below.

6.1 Limitations

The formal target tuple is not fixed in closed form. The public materials do not define A_q, α_q, ϵ_q, or the logical isometry W_q of Eq. (2). The claims above are therefore scoped to the operational probe contract of Section 1.1. A closed-form specification of the block-encoding relation remains an open item.

Finite public probes were reused during optimization. This creates overfitting risk. The 9,024 product states are structured and visible. They do not include hidden entangled holdouts. No exact statevector, symbolic, or unitary-invariant independent check has yet been completed.

Acceptance also depends on the numerical execution path. The error budget has three separate parts. The deliberate angle bias contributes the metric-aligned perturbation of Section 4.3. Bounded-bond MPS truncation contributes representation error. Floating-point order contributes platform and compiler variation. A future study should measure and allocate these terms separately.

No standard-toolchain baseline has been run under the same gate set and objective. The reported frontier therefore compares community constructions, not tool classes. The conditional Clifford ceiling also relies on the observed rotation count, fixed width, benchmark weights, and near-observed depth.

6.2 Reproducibility and data availability

Every score row is reproducible from Table 2 as $q\sqrt{VD}$. Circuits, verifier, target metadata, and the probe suite are public in the contest repository [8]. The repository state containing the accepted final circuit is pinned by the acceptance commit `74109105baba3f9337b4289c5b8deaa2520411b1` of 2026-08-12. Candidate and verifier artifact hashes and a machine-readable validation bundle remain open items.

6.3 Conclusion

Progressive synthesis reduced the public benchmark score 7.30× across its width-indexed target family. Declared-width engineering and the 17-qubit rebuild produced the result's scale. The rebuild era accounts for about 99.77% of the total reduction. Later refinements combined equal about 1.41% of the rebuild-era score. The final Dialog-inspired closed parity-trace search step contributed about 0.010% over the previous best. This proportionality directs future work toward rotation removal and broader structural changes.

Acknowledgments

The contributions were AI-assisted using OpenAI GPT-5/5.5 and the Hermes Agent harness. This draft was auto-assembled and is not peer reviewed.

References

[1] G. H. Low and I. L. Chuang. "Hamiltonian Simulation by Qubitization." arXiv:1610.06546, 2016.

[2] A. Gilyén, Y. Su, G. H. Low, and N. Wiebe. "Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics." arXiv:1806.01838, 2018.

[3] N. J. Ross and P. Selinger. "Optimal ancilla-free Clifford+T approximation of z-rotations." arXiv:1403.2975, 2014.

[4] D. Camps and R. Van Beeumen. "FABLE: Fast Approximate Quantum Circuits for Block-Encodings." arXiv:2205.00081, 2022.

[5] T. Khattar, N. Shutty, C. Gidney, A. Zalcman, N. Yosri, D. Maslov, R. Babbush, and S. P. Jordan. "Verifiable Quantum Advantage via Optimized DQI Circuits." arXiv:2510.10967, 2025, Sec. II.B.3.

[6] M. Amy, P. Azimzadeh, and M. Mosca. "On the CNOT-complexity of CNOT-PHASE circuits." arXiv:1712.01859, 2017.

[7] K. N. Patel, I. L. Markov, and J. P. Hayes. "Efficient Synthesis of Linear Reversible Circuits." arXiv:quant-ph/0302002, 2003.

[8] MatrixMul. "Matrix Multiplication via Block-Encoding Quantum Circuit." github.com/matrixmul/lv16-block-encoding, acceptance commit `74109105baba3f9337b4289c5b8deaa2520411b1`, 2026.