Main results

For a diagonal level-three gate U on n qubits with $\{\mathrm{CNOT}, T\}$ T-count $\delta(U)$, cubic moment tensor of radical dimension d, and quadratic moment matrix of rank r:

- $\delta(U) \geq 2(n-d)$, and $\geq 2n+1$ when $d=0$ (Theorem 19).
- The unitary stabilizer nullity is $\nu = n-d$, so $\delta(U) \geq 2\nu+1$ (Proposition 24).
- $\delta(U) \geq 2(n-d^*-r)$ for any level-three gate, d^* the radical of the full third-power moment tensor (Theorem 21).
- Residue weight $\leq 7 \Rightarrow \delta(U)$ is exact and n-independent; δ is non-additive (Theorem 25, 26).
- $\delta\big(\bigoplus_{i=1}^{m} CCZ_i\big) = 6m+1$ for all m (Theorem 27).

1 Preliminaries

We work over the field $\mathbb{F}_2 = \{0,1\}$ with addition equal to XOR ($1+1=0$) and multiplication equal to AND. Vectors live in $\mathbb{F}_2^t$ and are added coordinatewise. A subspace $W \subseteq \mathbb{F}_2^t$ has $|W| = 2^{\dim W}$.

1.1 Bilinear forms and isotropic subspaces

Definition 1 (dot product, orthogonal complement)**.** For $u, v \in \mathbb{F}_2^t$ the dot product is $\langle u, v\rangle = \sum_{i=1}^{t} u_i v_i \bmod 2$. It is symmetric and bilinear, and it is non-degenerate: if $\langle u, v\rangle = 0$ for all v then $u=0$. The orthogonal complement of W is $W^\perp = \{v : \langle v, w\rangle = 0\ \forall w \in W\}$, and non-degeneracy gives $\dim W + \dim W^\perp = t$.

A feature special to characteristic 2 is that a vector can be orthogonal to itself: $\langle u, u\rangle = \sum_i u_i^2 = \sum_i u_i = \mathrm{wt}(u) \bmod 2$, the parity of its Hamming weight.

Definition 2 (isotropic subspace)**.** W is *totally isotropic* if $\langle u, v\rangle = 0$ for all $u, v \in W$; equivalently $W \subseteq W^\perp$.

Lemma 3. *If $W \subseteq \mathbb{F}_2^t$ is totally isotropic then* $\dim W \leq t/2$.

Proof. $W \subseteq W^\perp$ gives $\dim W \leq \dim W^\perp = t - \dim W$, so $2\dim W \leq t$. □

Example 4. In $\mathbb{F}_2^4$ let $W = \mathrm{span}\{1100, 0011\}$. Then $\langle 1100, 1100\rangle = 2 \equiv 0$, $\langle 1100, 0011\rangle = 0$, $\langle 0011, 0011\rangle \equiv 0$, so W is totally isotropic and $\dim W = 2 = 4/2$, the largest possible.

We will also use the following standard fact about the rank of a form on a subspace.

Lemma 5. *Let $W = \mathrm{span}\{c_1, \ldots, c_n\} \subseteq \mathbb{F}_2^t$ and let G be the $n \times n$ Gram matrix $G_{ab} = \langle c_a, c_b\rangle$. Then the rank of the dot product restricted to W equals* $\mathrm{rank}_{\mathbb{F}_2} G$*, and the radical $W \cap W^\perp$ of the restricted form has dimension* $\dim W - \mathrm{rank}\, G$.

Proof. Pick a basis of W from among the c_a; in that basis the form has a symmetric matrix B whose rank is the rank of the form on W. Writing the full generating set in terms of the basis gives $G = P^\top B P$ with P of full column rank, so $\mathrm{rank}\, G = \mathrm{rank}\, B$. The radical of a symmetric form on a space of dimension $\dim W$ has dimension $\dim W -$ (rank of the form). □

1.2 Boolean functions, parities, and cubic forms

Definition 6 (parity)**.** Identify $y \in \mathbb{F}_2^n$ with the set $\{i : y_i = 1\} \subseteq [n]$. The parity indexed by y is the function $x \mapsto y \cdot x = \bigoplus_{i \in y} x_i$.

Definition 7 (algebraic normal form, degree, cubic form)**.** Every $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is uniquely a XOR of monomials, $f(x) = \bigoplus_{A \subseteq [n]} \hat{f}_A \prod_{i \in A} x_i$ with $\hat{f}_A \in \mathbb{F}_2$; its degree is $\max\{|A| : \hat{f}_A = 1\}$. A *cubic form* is a homogeneous degree-3 function $\bigoplus_{a<b<c} T_{abc}\, x_a x_b x_c$, encoded by the symmetric tensor $T_{abc} \in \mathbb{F}_2$ (with $T_{abc} = 0$ whenever two indices coincide).

1.3 Phase polynomials and the T-count

We use Clifford+T circuits. The Clifford group is generated by H, $S = \mathrm{diag}(1, i)$, and CNOT; the non-Clifford generator is $T = \mathrm{diag}(1, \omega)$ with $\omega = e^{i\pi/4}$. A Hadamard-free $\{\mathrm{CNOT}, T\}$ circuit is diagonal and applies a phase that is a sum of T-rotations on parities of the input:

$$U|x\rangle = \omega^{P(x)}|x\rangle, \qquad P(x) = \sum_{y \in \mathbb{F}_2^n} c_y\,(y \cdot x) \bmod 8, \quad c_y \in \mathbb{Z}_8. \tag{1}$$

A T adds 1 to one coefficient c_y, an S adds 2, a Z adds 4, and two gates on the same parity merge by adding their coefficients. The T-count of the representation is the number of odd c_y, because only odd coefficients require a T and merging cannot increase the count of odd coefficients. Write $\delta(U)$ for the minimum T-count over all representations of U.

Example 8 (the three basic gates)*.* Over $\mathbb{Z}$, hence valid modulo 8,

$$T:\ x_i = \chi_{\{i\}}, \qquad \mathrm{CS}:\ 2x_i x_j = \chi_{\{i\}} + \chi_{\{j\}} - \chi_{\{i,j\}}, \qquad \mathrm{CCZ}:\ 4x_i x_j x_k = \sum_{\varnothing \neq S \subseteq \{i,j,k\}} (-1)^{|S|-1} \chi_{\oplus S},$$

where $\chi_S(x) = S \cdot x$. So CCZ has seven parities (the nonempty subsets of $\{i, j, k\}$), each with an odd coefficient.

1.4 Linear codes and Reed–Muller codes

Definition 9 (linear code, weight, distance, dual)**.** A binary linear code of length L is a subspace $C \subseteq \mathbb{F}_2^L$. The Hamming weight $\mathrm{wt}(v)$ is the number of 1s; the minimum distance is $d = \min_{0 \neq c \in C} \mathrm{wt}(c)$. The dual is $C^\perp = \{v : \langle v, c\rangle = 0\ \forall c \in C\}$, of dimension $L - \dim C$. The syndrome of a vector is its inner products against a basis of $C^\perp$; two vectors have the same syndrome iff they differ by a codeword, that is, lie in the same coset $v + C$. A coset leader is a minimum-weight vector in a coset, and *minimum-distance decoding* asks for the nearest codeword to a given vector. It is NP-hard for general codes.

Definition 10 (puncturing and shortening)**.** Puncturing C at a coordinate deletes that coordinate from every codeword. Shortening keeps only the codewords that are 0 at the coordinate, then deletes it. The two are exchanged by duality: the dual of C punctured at i is $C^\perp$ shortened at i. We write $\mathrm{RM}(r, m)^*$ for $\mathrm{RM}(r, m)$ punctured at the all-zeros coordinate.

Definition 11 (affine flat)**.** A k-flat in $\mathbb{F}_2^m$ is a coset $a + V$ of a k-dimensional subspace; it has 2^k points, and its indicator is a Boolean function of degree exactly $m - k$.

Definition 12 (Reed–Muller code)**.** $\mathrm{RM}(r, m)$ is the evaluation code of all Boolean polynomials of degree at most r in m variables: list the value of each such polynomial at all 2^m points. It has dimension $\sum_{i \le r} \binom{m}{i}$ and minimum distance 2^{m-r}, with minimum-weight codewords the indicators of $(m - r)$-flats, and its dual is $\mathrm{RM}(m - r - 1, m)$.

1.5 The Amy–Mosca correspondence and the moment reformulation

Reduce the coefficients in (1) modulo 2: the *residue* $\mathrm{Res}_2(U) = (c_y \bmod 2)_y \in \mathbb{F}_2^{2^n-1}$ (indexed by nonzero parities) records which parities carry an odd phase. Amy and Mosca proved that two representations implement the same gate iff their residues differ by a codeword of $\mathrm{RM}(n-4,n)^*$, so

$$\delta(U) = \mathrm{dist}\big(\mathrm{Res}_2(U), \mathrm{RM}(n-4,n)^*\big) = \min_{c\in\mathrm{RM}(n-4,n)^*} \mathrm{wt}\big(\mathrm{Res}_2(U)\oplus c\big). \tag{2}$$

It is worth seeing where the code comes from, because the degree $n-4$ is the source of every later constant. A set S of parities (all coefficients 1) costs zero T, that is, composes to a Clifford, exactly when $\sum_{y\in S}(y\cdot x) \equiv 0 \pmod 8$ for all x. For one parity, the identity "XOR equals integer sum with carries" reads, over $\mathbb{Z}$,

$$y\cdot x = \Big(\textstyle\sum_{i\in y} x_i\Big) - 2e_2(x_y) + 4e_3(x_y) - \cdots,$$

where $e_k(x_y)$ is the k-th elementary symmetric polynomial in the bits indexed by y. Modulo $8 = 2^3$ this truncates after e_3. Summing over $y \in S$ and reducing mod 8, the condition "$\equiv 0$ for all x" becomes, one carry order at a time, the requirement that $\mathbf{1}_S$ is orthogonal to the evaluation of every monomial $\prod_{i\in A} y_i$ with $1 \le |A| \le 3$. Those evaluations generate $\mathrm{RM}(3,n)$ shortened at $y = 0$ (shortening removes the constant monomial), so the free relations are its dual, $\mathrm{RM}(n-1-3,n)^* = \mathrm{RM}(n-4,n)^*$. The "$-3$" counts the three carry orders of mod 8, and the "-1" is the shortening.

This motivates working with the syndrome coordinates directly.

Definition 13 (moments)**.** For a set $r' \subseteq \mathbb{F}_2^n \setminus \{0\}$ of parities and $A \subseteq [n]$, the moment is

$$M_A(r') = \big|\{y \in r' : A \subseteq y\}\big| \bmod 2.$$

The moment M_A is exactly the syndrome coordinate of r' against the monomial $\prod_{i\in A} y_i$, so the derivation above gives the following operational form of (2).

Fact 14 (moment form of Amy–Mosca)**.** *The coset of a residue, and hence the gate it represents, is determined by the moments M_A with $1 \le |A| \le 3$, and*

$$\delta(U) = \min\big\{\, |r'| : M_A(r') = s_A(U) \text{ for all } 1 \le |A| \le 3 \,\big\},$$

where $s_A(U)$ is the gate's syndrome. The order-0 moment and all moments with $|A| \ge 4$ are unconstrained.

Example 15 (CCZ in moments). The seven parities of CCZ are the nonempty subsets of $\{1,2,3\}$. Then $M_{\{1\}} = |\{1,12,13,123\}| = 4 \equiv 0$, $M_{\{1,2\}} = |\{12,123\}| = 2 \equiv 0$, and $M_{\{1,2,3\}} = 1$. So all order-1 and order-2 moments vanish and one order-3 moment is 1.

Definition 16 (pure-cubic gate, cubic tensor, radical, quadratic matrix)**.** A gate is *pure-cubic* if $s_A = 0$ for all $1 \le |A| \le 2$. Its *cubic moment tensor* is $T_{abc} = s_{\{a,b,c\}}$. The *radical* of T is

$$R = \Big\{ w \in \mathbb{F}_2^n : \textstyle\sum_a w_a T_{aqr} = 0 \ \forall q, r \Big\}, \qquad d = \dim R,$$

and T is *non-degenerate* when $d = 0$. For a general gate the *quadratic moment matrix* is $Q \in \mathbb{F}_2^{n\times n}$ with $Q_{ab} = s_{\{a,b\}}$ for $a \ne b$ and $Q_{aa} = s_{\{a\}}$, of rank $r = \mathrm{rank}_{\mathbb{F}_2} Q$. Computing d and r is Gaussian elimination, so $O(n^3)$ time.

Finally we record the one coding fact behind the rigidity radius.

Lemma 17 (uniform minimum distance)**.** *For every $n \geq 4$, $\mathrm{RM}(n-4,n)^*$ has minimum distance 15.*

Proof. $\mathrm{RM}(n-4,n)$ has minimum distance $2^{n-(n-4)} = 2^4 = 16$, attained by the indicators of 4-flats. Puncturing at the all-zeros coordinate deletes one symbol: a minimum-weight flat through the origin loses one 1, leaving 15, and no codeword drops lower. So the minimum distance is 15, independent of n. □

2 A computable lower bound

Fix a representation of a gate, with odd parities $y_1, \ldots, y_t$. Let $Y \in \mathbb{F}_2^{t\times n}$ be the matrix whose rows are the y_j, and let $c_a \in \mathbb{F}_2^t$ be its a-th column, the indicator of which parities contain qubit a. The key identities are

$$\langle c_a, c_b\rangle = \sum_j (y_j)_a (y_j)_b = M_{\{a,b\}}, \qquad \langle c_a, c_a\rangle = \sum_j (y_j)_a = M_{\{a\}}, \tag{3}$$

and, for the triple products, $\sum_j (c_a)_j (c_b)_j (c_q)_j = M_{\{a,b,q\}} = T_{abq}$.

2.1 The pure-cubic case

Lemma 18 (isotropic columns)**.** *For a pure-cubic gate, every representation has $t \geq 2\,\mathrm{rank}\,Y$.*

Proof. By (3), a pure-cubic gate has $\langle c_a, c_b\rangle = M_{\{a,b\}} = 0$ for $a \neq b$ and $\langle c_a, c_a\rangle = M_{\{a\}} = 0$. So $W = \mathrm{span}\{c_a\}$ is totally isotropic, and Lemma 3 gives $\dim W \leq t/2$. Since column rank equals row rank, $\mathrm{rank}\,Y = \dim W \leq t/2$. □

Theorem 19 (cubic lower bound)**.** *Let U be a pure-cubic gate on n qubits whose cubic tensor has radical dimension d. Then $\delta(U) \geq 2(n-d)$, and $\delta(U) \geq 2n+1$ if $d = 0$.*

Proof. Let $V = \mathrm{span}\{y_j\}$ be the row space and $\rho = \dim V = \mathrm{rank}\,Y$.

Rank versus radical. If $w \in V^\perp$ then w is orthogonal to every row, so $\sum_a w_a c_a = 0$ in $\mathbb{F}_2^t$. Hence for all q, r,

$$\sum_a w_a T_{qra} = \sum_j (c_q)_j (c_r)_j \Big(\sum_a w_a c_a\Big)_j = 0,$$

so $w \in R$. Thus $V^\perp \subseteq R$, giving $\dim V^\perp \leq d$ and $\rho \geq n-d$. With Lemma 18, $t \geq 2\rho \geq 2(n-d)$.

The extra $+1$ when $d = 0$. Suppose $d = 0$ and, for contradiction, $t = 2n$. Then $\dim W = \rho = n = t/2$, so by Lemma 3 the inclusion $W \subseteq W^\perp$ is equality: $W = W^\perp$ is maximal isotropic. The all-ones vector $\mathbf{1} \in \mathbb{F}_2^t$ satisfies $\langle \mathbf{1}, c_a\rangle = M_{\{a\}} = 0$, so $\mathbf{1} \in W^\perp = W$ and $\mathbf{1} = \sum_{a\in S} c_a$ for some nonempty S. For every pair p, q, since $\mathbf{1}$ is all ones,

$$0 = \langle c_p, c_q\rangle = \sum_j (c_p)_j (c_q)_j \cdot 1 = \sum_{a\in S} T_{pqa},$$

so $\mathbf{1}_S$ contracts the tensor to zero, that is, $\mathbf{1}_S \in R = \{0\}$, forcing $S = \varnothing$, a contradiction. Hence $t \neq 2n$, so $t \geq 2n+1$. □

Example 20. For a single CCZ, $n = 3$ and the tensor has the one entry $T_{123} = 1$, so $d = 0$ and the bound is $2 \cdot 3 + 1 = 7 = \delta(\mathrm{CCZ})$. The bound is tight.

2.2 The general case

The purity hypothesis was used only to make the columns isotropic. For an arbitrary gate the column Gram matrix is the quadratic moment matrix Q, which has some rank r, and a form of rank r still leaves an isotropic radical of codimension r.

Theorem 21 (general lower bound)**.** *For any diagonal third-level gate on n qubits, with quadratic moment matrix of rank r and $d^* = \dim\{w : \sum_a w_a\, s_{\{a,q,r\}} = 0\ \forall q, r\}$ the radical of the full third-power moment tensor (coincident indices give the order-≤ 2 moments),*

$$\delta(U) \geq 2(n - d^* - r).$$

For a pure-cubic gate the order-≤ 2 moments vanish, so $d^* = d$ and this recovers Theorem 19.

Proof. By (3) the Gram matrix of the columns is exactly Q. Put $W = \operatorname{span}\{c_a\}$ and $\rho = \dim W = \operatorname{rank} Y$. By Lemma 5 the radical $W \cap W^{\perp}$ of the form restricted to W has dimension $\rho - r$. This radical is totally isotropic in $\mathbb{F}_2^t$, so by Lemma 3, $\rho - r \leq t/2$, that is, $t \geq 2(\rho - r)$. For the rank step, if $w \in V^{\perp}$ then $\sum_a w_a c_a = 0$, so $\sum_a w_a\, s_{\{a,q,r\}} = \sum_j c_q c_r (\sum_a w_a c_a) = 0$ for all q, r; the sum runs over all a, so coincident indices contribute order-≤ 2 moments via $c_q^2 = c_q$, and thus $w \in R^*$, giving $\rho \geq n - d^*$. Therefore $t \geq 2(\rho - r) \geq 2(n - d^* - r)$. (Using the cubic radical d here would be an error off the pure-cubic locus.) ☐

Proposition 22 (the floor at level k)**.** *For a diagonal gate in the k-th level of the Clifford hierarchy (phase mod 2^k) whose order-$\leq (k-1)$ moments vanish, $\delta_k(U) \geq 2(n - d_k)$, where d_k is the radical dimension of the order-k moment tensor and δ_k counts the finest gate. The proof is the argument of Theorem 19: order-2 moments zero gives isotropic columns and $t \geq 2 \operatorname{rank} Y$, and $V^{\perp}$ lies in the radical of the order-k tensor, giving* $\operatorname{rank} Y \geq n - d_k$*; validity of the order-$\leq k$ moments as complete invariants is Amy–Mosca for $k = 3$. The bound is not tight for $k \geq 4$. The nullity identity persists, however: $\nu = n - d_k$ (verified at $k = 4$), so the floor is 2ν and strictly exceeds the nullity bound $\delta \geq \nu$ at every level; only the exactness against δ is special to $k = 3$.*

Remark 23 (comparison with stabilizer nullity). Stabilizer nullity ν is the standard efficiently computable lower bound on T-count. For the gates above it gives $\nu = n$ (for example $\nu(\mathrm{CCZ}) = 3$), so Theorem 19 is a clean factor-two improvement, and it is computed in the same $O(n^3)$ time. Unlike nullity it is exactly tight on many structured gates (the next two sections). No efficiently computable bound can be tight on *all* cubic gates, since that would solve minimum-distance decoding in $\mathrm{RM}(n-4, n)^*$; the point is that the bound here is exact on the structured targets that occur in practice.

Proposition 24 (stabilizer nullity equals $n - d$, and a uniform $+1$)**.** *For a pure-cubic gate the unitary stabilizer nullity is $\nu = n - d$, and $\delta \geq 2\nu + 1 = 2(n - d) + 1$.*

Proof. The nullity is $\nu = n - \log_2 |A|$, where A is the group of shifts a with $\Delta_a \phi(y) = \phi(y \oplus a) - \phi(y)$ of the form $c + 4\ell(y)$, ℓ affine. For $\phi = 4f$ with f cubic, $\Delta_a \phi = 4\Delta_a f$ and $\Delta_a f$ is affine iff its quadratic part, the bilinear form $(q, r) \mapsto T_{aqr}$, vanishes, that is iff $a \in R$; so $A = R$ and $\nu = n - d$. Aligning R to the last d coordinates by a CNOT relabelling (which preserves δ), the homogeneous cubic has no monomial meeting a radical coordinate, so $U = U' \otimes I_d$ with U' non-degenerate on $n - d$ qubits; the identity factor carries no parities, so $\delta = \delta(U') \geq 2(n - d) + 1$ by the non-degenerate case of Theorem 19. ☐

The comparison is concrete in Table 1. On each non-degenerate target nullity gives $\nu = n$, this bound gives $2n+1$, and the true T-count (the proven $6m+1$ for Toffoli layers, or exact Reed–Muller coset decoding otherwise) equals $2n+1$: the bound is exactly $2\nu + 1$ and exactly tight, while nullity is loose by nearly a factor of two.

gate	n	nullity ν	this bound	true δ	source of δ
Toffoli (CCZ)	3	3	7	7	exact decode
2 parallel Toffolis	6	6	13	13	exact decode
3 parallel Toffolis	9	9	19	19	Theorem 27
4 parallel Toffolis	12	12	25	25	Theorem 27
3-local line f_5	5	5	11	11	exact decode
3-local line f_6	6	6	13	13	exact decode
elem. symmetric e_3 ($n = 5$)	5	5	11	11	exact decode

Table 1: Stabilizer nullity versus this bound versus the true T-count on non-degenerate cubic targets ($d = 0$). The bound equals $2\nu + 1$ and is tight throughout; nullity is off by nearly a factor of two. The nullity column is computed directly (`verify_nullity.py`, self-tested against $\nu(\text{CCZ}) = 3$); decoded δ values self-test against $\delta(\text{CCZ}) = 7$, $\delta(\text{CS}) = 3$ (`compare_bounds.py`).

3 Rigidity radius and non-additivity

Theorem 25 (rigidity radius)**.** *If a third-level gate has residue of Hamming weight $w \le 7$, then $\delta = w$ exactly, for every n, with no search.*

Proof. For any nonzero codeword $c \in \text{RM}(n-4, n)^*$, the triangle inequality and Lemma 17 give

$$\text{wt}(\text{Res}_2 \oplus c) \ge \text{wt}(c) - w \ge 15 - w \ge 8 > w.$$

So the residue itself is the strict minimum-weight coset leader, and by (2), $\delta = w$. The threshold $w \le 7 = \lfloor (15-1)/2 \rfloor$ is the unique-decoding radius. $\square$

Thus every small gate is pinned for all n: $\delta(T) = 1$, $\delta(\text{CS}) = 3$, $\delta(\text{CCZ}) = 7$, and $\delta(\text{CS}\otimes\text{CS}) = 6$. Just past the radius, the weight-15 codewords begin to overlap the residue and cancel mass, and T-count stops being additive.

Theorem 26 (non-additivity)**.** *T-count is not additive over variable-disjoint third-level gates:*

$$\delta(\text{CCZ} \oplus T) = 7 = \delta(\text{CCZ}) \quad (n \ge 4), \qquad \delta(\text{CCZ} \oplus \text{CCZ}) = 13 < 14 = 2\,\delta(\text{CCZ}) \quad (n = 6).$$

Proof. CCZ$\oplus T$. Restricting any representation to the three CCZ qubits (setting the other inputs to 0) gives a representation of CCZ with no more odd parities, so $\delta \ge \delta(\text{CCZ}) = 7$. For the matching upper bound, the residue has weight 8: the seven CCZ parities (nonempty subsets of $\{1, 2, 3\}$) and the singleton $\{4\}$. Take the weight-15 codeword $c_y = \prod_{i \ge 5}(1 \oplus y_i)$, the indicator of the 4-flat span$\{e_1, e_2, e_3, e_4\}$ punctured at the origin; it is 1 at all 15 nonzero parities supported on $\{1, 2, 3, 4\}$, hence at all eight support points, leaving $\text{wt}(\text{Res}_2 \oplus c) = 15 - 8 = 7$. (The single-coordinate flat $c_y = 1 \oplus y_4$ fails: it vanishes at $\{4\}$, covers only the seven CCZ positions, and leaves weight at least 9.)

CCZ $\oplus$ CCZ. On its six qubits this is pure-cubic with $d = 0$, so Theorem 19 gives the lower bound $2 \cdot 6 + 1 = 13$, and the hub-tree construction of Theorem 27 (the case $m = 2$) gives a representation of weight 13. This is strictly below $2\,\delta(\text{CCZ}) = 14$. $\square$

4 Exact cost of Toffoli layers

A Toffoli gate is a CCZ conjugated by Hadamards on the target, and a layer of m disjoint Toffolis is, up to Clifford gates, a layer of m disjoint CCZ gates on $n = 3m$ qubits. Write $\delta_m = \delta\big(\bigoplus_{i=1}^{m} \text{CCZ}_i\big)$.

Theorem 27 (exact Toffoli-layer cost)**.** *For every $m \geq 1$, $\delta_m = 6m+1$. Equivalently, the per-gate T-rate of a parallel Toffoli/*CCZ *layer is exactly* 6*, and T-count* $= 6 \cdot ($*Toffoli count*$) + 1$.

Proof. Lower bound. The layer is pure-cubic, with $T_{pqr} = 1$ exactly when $\{p, q, r\}$ is a block. Its radical is trivial: if $w \in R$, taking q, r to be two qubits of a block forces w to vanish on the third, and ranging over the three pairs of each block gives $w = 0$. So $d = 0$, and Theorem 19 gives $\delta_m \geq 6m + 1$.

Upper bound (hub tree). Root any tree on the m blocks in which every node has at most three children; such a tree exists for all m. Label each edge by a distinct qubit, called a *port*, of the parent block, which is possible since a block has three qubits and at most three children. For a block B let $E(B)$ be the set of ports lying on the path from the root to B, so E accumulates one ancestor port per level and $E(\text{root}) = \varnothing$, and let $\Pi(B)$ be the set of B's own ports that are used by its children. Emit, for each block B, the parities

$$\{\, s \cup E(B) : s \in \mathrm{CCZ}(B),\ s \notin \{\{p\} : p \in \Pi(B)\}\,\},$$

that is, the seven CCZ parities of B, each enlarged by the accumulated ancestor ports $E(B)$, omitting the singleton at every port that B hands to a child.

These parities are distinct across the tree. A parity y emitted by B has $y \cap B = s$, and $E(B)$ is disjoint from B because ancestor ports lie in ancestor blocks; if $|s| \geq 2$ then B is the unique block met by y in two or more qubits, and if $|s| = 1$ then B is the block whose non-port qubit lies in y. Block B emits $7 - |\Pi(B)|$ parities, and summing over the tree with $\sum_B |\Pi(B)| = \#\text{edges} = m - 1$ gives total weight $7m - (m-1) = 6m + 1$.

It remains to check the moments. A parity emitted by B has support in $B \cup E(B)$ and meets any other block in at most one qubit, the port on the connecting edge. Fix A with $|A| \leq 3$.

- If A is the triple of a block B_0, only B_0's full-triple parity covers all three, so $M_A = 1$.
- If A is two qubits of one block, they are covered only by that block's two relevant CCZ subsets, so $M_A = 2 \equiv 0$.
- If $A = \{q\}$ with $q \in B_0$, then B_0 contributes $4 - [q \in \Pi(B_0)]$ parities through q. Moreover, if q is a port, then every parity in the entire subtree S_q hanging from q contains q (because $q \in E$ throughout S_q), and that subtree contributes $\sum_{B \in S_q}(7 - |\Pi(B)|) = 6|S_q| + 1$. So $M_q = 4 \equiv 0$ when q is not a port, and $M_q = 3 + (6|S_q| + 1) = 4 + 6|S_q| \equiv 0$ when it is.
- If A spans two or three blocks, a qubit shared across blocks is covered only when it is a port, and then the same subtree count $6|S| + 1$ cancels the contribution, giving $M_A = 0$.

So $M_A = 1$ exactly on block triples, which is the syndrome of $\bigoplus \mathrm{CCZ}_i$ by Fact 14. The emitted set has weight $6m + 1$, meeting the lower bound, so it is optimal and $\delta_m = 6m + 1$. □

The cancellation works at every depth because a subtree of t blocks emits $6t + 1$ parities, which is always odd, and these all carry the inherited port, balancing modulo 2 the three parities the parent keeps after dropping that port's singleton. Accumulating all ancestor ports, rather than only the immediate one, is what makes this hold below depth one.

Corollary 28 (magic compression)**.** *Implementing m disjoint Toffoli/*CCZ *gates as one parallel layer costs $6m + 1$ distilled magic states, against $7m$ if they are paid for one at a time. The saving is exactly $m - 1$ states, asymptotically one per gate, and it is the most possible because $6m + 1$ is the floor.*

Proof. A single CCZ costs 7 (Theorem 25), so m of them paid separately cost $7m$; the layer costs $\delta_m = 6m + 1$ (Theorem 27); and $7m - (6m + 1) = m - 1$. □

References

[1] Matthew Amy and Michele Mosca. T-count optimization and Reed–Muller codes. *IEEE Transactions on Information Theory*, 65(8):4771–4784, 2019.

[2] Matthew Amy, Dmitri Maslov, Michele Mosca, and Martin Roetteler. A meet-in-the-middle algorithm for fast synthesis of depth-optimal quantum circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 32(6):818–830, 2013.

[3] Jiaqing Jiang and Xin Wang. Lower bound for the T count via unitary stabilizer nullity. *Physical Review Applied*, 19:034052, 2023.

[4] Michael Beverland, Earl Campbell, Mark Howard, and Vadym Kliuchnikov. Lower bounds on the non-Clifford resources for quantum computations. *Quantum Science and Technology*, 5:035009, 2020.

[5] Shawn X. Cui, Daniel Gottesman, and Anirudh Krishna. Diagonal gates in the Clifford hierarchy. *Physical Review A*, 95:012329, 2017.