— CHAPTER 07 · FAULT TOLERANCE

Quantum error correction — surface codes to qLDPC.

The most technically dense chapter — and the one where pre-2024 textbooks are most stale.

WHAT YOU'LL LEARN

- Why no-cloning forces syndrome measurement
- Read a below-threshold plot (Willow d=3,5,7)
- Simulate a distance-5 surface code in Stim

— WHY WE NEED QEC AT ALL

0.1% per-gate sounds tiny. Run it *a thousand times* — most circuits fail.

0.1%

PER-GATE ERROR

Best 2026 two-qubit fidelity, ~99.9%. One in a thousand operations is wrong.

1,000

GATES IN A USEFUL CIRCUIT

VQE on a small molecule, a Shor demo, even a moderate QAOA depth — all easily a thousand gates.

≈63%

PROBABILITY AT LEAST ONE ERROR

$1 - (0.999)^{1000}$. The noiseless answer is buried in the noise tail.

- Decoherence (T1, T2)
- Gate infidelity
- Measurement error
- Three independent failure modes

— THE APPARENT PARADOX

Classical EC copies bits. Quantum can't — so we measure *the parity, not the state*.

CLASSICAL (EASY)

Triple-modular redundancy

Send the bit three times: 0 → 000. Flip one in transit and majority-vote recovers it. Copying is free, measurement is non-destructive. We've done this since the 1950s.

QUANTUM (HARD)

No-cloning + collapse

$|\psi\rangle \to |\psi\rangle|\psi\rangle|\psi\rangle$ is forbidden (no-cloning theorem). And measuring $|\psi\rangle$ destroys the superposition. Naive redundancy is impossible on both counts.

THE WORKAROUND — STABILISER MEASUREMENT

Read parity via an ancilla, leave the logical state alone

Encode one logical qubit across many physical qubits. Use ancilla qubits to measure joint operators — like $Z \otimes Z \otimes Z \otimes Z$ — that *commute* with the logical information. The ancilla tells you whether an error happened and where, without revealing the encoded state. That bit-string is the *syndrome*.

Stabiliser formalism · Ancilla-based readout · Syndrome → decoder → correction

— THE DOMINANT CODE, 2010 – 2023

2D grid of physical qubits. Bigger grid, *exponentially* fewer logical errors.

DISTANCE 3

9 data + 8 ancilla

Corrects 1 error. Smallest non-trivial surface code. ~17 physical qubits per logical qubit.

DISTANCE 5

25 data + 24 ancilla

Corrects 2 errors. ~49 physical per logical. Willow's middle step in the 2024 scan.

DISTANCE 7

49 data + 48 ancilla

Corrects 3 errors. ~97 physical per logical. Willow's largest demonstrated code in Dec 2024.

THE THRESHOLD THEOREM

Below ~1% physical error per gate — bigger d wins

If physical error rate p is below the code's threshold p_{th}, logical error rate scales like $(p / p_{th})^{(d+1)/2}$. Surface-code threshold ≈ 1% — comfortably above 2026 hardware. Above threshold, more qubits make things worse.

- ~d^2 physical per logical
- Nearest-neighbour only
- Threshold ≈ 1%
- Corrects $\lfloor (d-1)/2 \rfloor$ errors

— VENDOR SPOTLIGHT · NATURE · DECEMBER 2024

Google Willow — adding qubits actually *helped*, for the first time on real silicon.

D = 3

~3.0 × 10^{-2}

Logical error per cycle. Smallest code, highest error — as expected.

D = 5

~1.4 × 10^{-2}

Roughly half. Adding the next ring of qubits paid off.

D = 7

~6.5 × 10^{-3}

Half again. Below-threshold confirmed.

THE Λ (LAMBDA) PARAMETER

Λ ≈ 2.14 per +2 distance

Each step from d to d+2 cut the logical error rate by roughly 2.14×. The threshold theorem demands a constant factor > 1; Willow delivered it. 105 transmon qubits. *Nature*, Dec 2024 — the QEC result, not the random-circuit-sampling number, is the milestone.

105 transmons · d = 3, 5, 7 · Λ ≈ 2.14 · Below threshold ✓

— VENDOR SPOTLIGHT · IBM QUANTUM

The Gross code — 12 logical qubits in 144 physical. A *ten-fold* overhead cut versus surface code.

SURFACE CODE (FOR THE SAME 12 LOGICAL, D=12)

~1,452 physical qubits

Each logical needs ~121 physical at distance 12. Times 12 logicals — and that's before routing, ancillas, and magic-state factories.

GROSS CODE [[144, 12, 12]]

144 physical qubits

Bivariate Bicycle qLDPC code. Same distance, same logical count — order-of-magnitude fewer physical qubits. The catch: stabiliser checks span long-range connections, not just nearest neighbours.

WHY THIS CHANGES THE IBM ROADMAP

Loon c-couplers → Kookaburra → Starling

Surface codes ran 2010–2023. qLDPC overtakes when you can wire long-range couplers between modules — exactly the engineering bet behind Loon (2025) and the Kookaburra (2026) qLDPC memory tile. Starling 2028-2029 is the first machine designed around this trade.

- [[144, 12, 12]]
- Bivariate Bicycle
- ~10× overhead reduction
- Needs long-range couplers

— VENDOR SPOTLIGHT · SEPTEMBER 2024

Logical qubits running production circuits — *not plots.*

12

LOGICAL QUBITS

Encoded on the Quantinuum H2 trapped-ion processor. Real, not simulated.

22×

CIRCUIT ERROR REDUCTION

Logical circuit error rate vs the equivalent circuit on bare physical qubits. The encoding paid for itself.

1st

END-TO-END CHEMISTRY ON LOGICAL

Hybrid quantum-classical chemistry simulation executed entirely on encoded logical qubits.

WHY IT MATTERS

First time encoding was a net win, not a tax

For most of QEC history, running an algorithm on logical qubits was *worse* than running it on physical — the overhead exceeded the gain. The Sep 2024 Microsoft + Quantinuum result flipped that for a real workload. Pre-2024 textbooks called this "the QEC break-even"; 2026 calls it shipped.

Quantinuum H2 ions · Microsoft codes · Sep 2024

— VENDOR SPOTLIGHT · NATURE 2024

QuEra — 96 logical qubits with *transversal* gates and the first logical magic-state distillation.

96

LOGICAL QUBITS

Encoded on neutral-atom hardware. Largest logical-qubit count publicly demonstrated.

1st

LOGICAL MAGIC-STATE DISTILLATION

The piece you need to make T-gates fault-tolerant. Now demonstrated on real hardware.

Nature

2024

Peer-reviewed primary publication, not a vendor blog claim.

THE TRICK — TRANSVERSAL GATES

Apply the same physical op to every qubit in the block

When you do, errors can't spread within a code block, because each physical qubit only ever talks to its counterpart in another block. Neutral-atom arrays can be reconfigured between rounds, so the transversal trick is natural there in a way it isn't on a fixed superconducting grid.

- Neutral atoms (Rydberg)
- Reconfigurable array
- Transversal gates
- Logical magic-state distillation

— THE HIDDEN COST OF FTQC

Clifford gates are free. *T-gates* are why FTQC chips will be huge.

● CLIFFORD GATES (CHEAP)

H, S, CNOT, Pauli

Native to the surface code. Transversal or lattice-surgery. The bad news — a Clifford-only circuit is classically simulable (Gottesman-Knill theorem). You can't compute anything quantum-useful with just these.

● T-GATE (EXPENSIVE)

π/8 phase rotation

Needed for universality. Cannot be done transversally in the surface code. Must be injected from a pre-prepared "magic state" — and those states arrive noisy and have to be *distilled*.

● THE OVERHEAD

10× – 100× the logical computation's qubit budget

A Shor's-algorithm chip running on logical qubits typically spends the majority of its physical qubits in magic-state factories, not on the algorithm itself. This is why FTQC resource estimates land at millions of physical qubits for thousands of logical — most of them are making T-states.

- Gottesman-Knill: Clifford = classical
- T-gate needed for universality
- Distillation = most of the chip

— FRAMEWORK ANCHOR · STIM 1.X (CRAIG GIDNEY)

One snippet — a distance-5 surface code, *10 rounds*, decoded by minimum-weight matching.

● PYTHON · STIM + PYMATCHING

Generate, sample, decode

```
import stim, pymatching, numpy as np

circuit = stim.Circuit.generated(
    "surface_code:rotated_memory_z",
    distance=5, rounds=10,
    after_clifford_depolarization=0.001,
    after_reset_flip_probability=0.001,
    before_measure_flip_probability=0.001,
    before_round_data_depolarization=0.001,
)
dem = circuit.detector_error_model(decompose_errors=True)
matcher = pymatching.Matching.from_detector_error_model(dem)

sampler = circuit.compile_detector_sampler()
detections, observables = sampler.sample(shots=100_000, separate_observables=True)
predictions = matcher.decode_batch(detections)
logical_err = np.mean(np.any(predictions != observables, axis=1))
print(f"logical error per shot @ d=5: {logical_err:.2e}")
```

- Clifford-only → millions of qubits OK
- p = 0.001 per Clifford
- PyMatching = MWPM decoder
- Runs in seconds on a laptop

— TRY IT YOURSELF · ~15 MINUTES · FREE, LOCAL

Reproduce *exponential suppression* on your laptop — no QPU required.

SETUP

pip install stim pymatching numpy

Stim is Clifford-only, so even d=9 with 50 rounds runs in seconds on a CPU. No GPU, no cloud account.

RUN

Sweep d ∈ {3, 5, 7}, p = 0.001, shots = 100k

Use the slide-10 snippet, change `distance=`. Record logical error rate at each distance.

WHAT TO LOOK FOR

Each step from d to d+2 should drop logical error by > 2×

At p = 0.001 you are comfortably below the surface-code threshold, so the $(p/p_{th})^{(d+1)/2}$ scaling should hold. If the ratio collapses to ~1, you're above threshold — drop p to 0.0005 and try again. You have just reproduced, on a laptop, the qualitative result Google published in *Nature*.

- Free, local, ~5 min runtime
- Plot logical err vs d
- Next: FTQC roadmaps