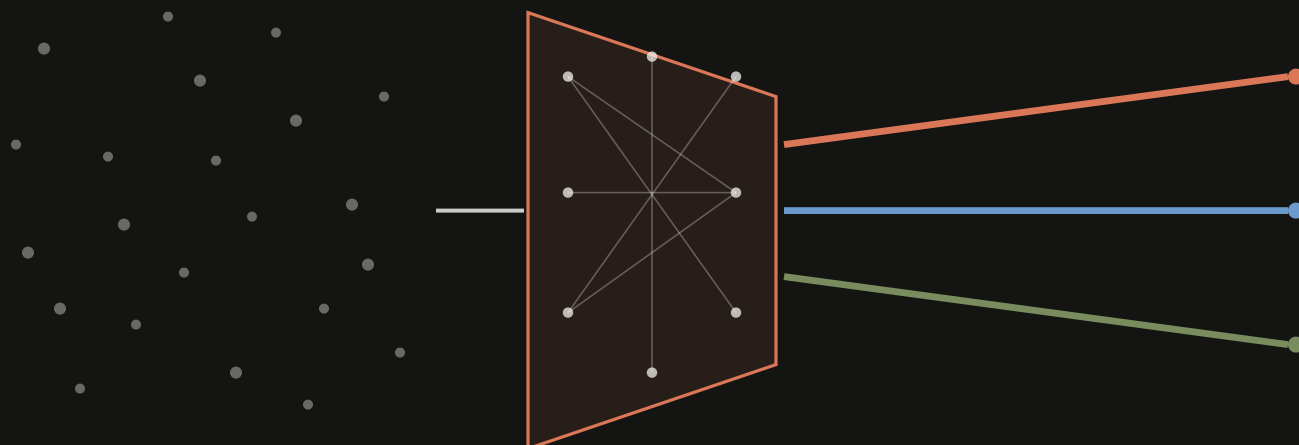


INFLECTION.

THE WEEKLY MAGAZINE OF INNOVATION

ISSUE 02 · FRIDAY, SEPTEMBER 25, 2026 · DEEP DIVE — FAULT-TOLERANT QUANTUM



QUANTUM COMPUTING · ERROR CORRECTION

The qubit race was the *wrong* race

This week two Australian startups showed that a thousand error-corrected qubits could be built from a hundred and fifty thousand ordinary ones — on plain silicon, made in a chip foundry. The number everyone is quoting is the wrong headline. The real news is where the difficulty went.

01

Why the bottleneck just moved from physics to classical latency

02

AI predicts a fusion plasma tearing itself apart — 200ms early

03

A telescope that saved so much fuel it doubled its own life

CONTENTS

03	The Feature — Opener	<i>A thousand clean qubits from silicon</i>
04	The Feature — Breakthroughs	<i>What Pinnacle actually did</i>
05	The Feature — So What	<i>Lab to market, and Field Notes</i>
06	Against the Grain	<i>The bottleneck is now classical</i>
07	Signals	<i>Fusion, aging, and a frugal telescope</i>
08	By the Numbers	<i>Eight figures that frame the week</i>
09	The Long View	<i>Sources & further reading</i>

DISPATCH

Difficulty is conserved. It only changes address.

There is a failure mode specific to technically literate readers. We are good at tracking a number. We watch it fall, we extrapolate the line, and we mistake the number for the problem. The number is almost never the problem.

In 2019 the accepted cost of breaking RSA-2048 on a quantum computer was twenty million physical qubits. By May 2025 Craig Gidney had it under one million. In February 2026 a Sydney startup put it under a hundred thousand. Each fall came faster than the last, and each headline read the same way: *the quantum threat is closer than you thought*.

Read the papers instead of the headlines and a different story appears. Every one of those reductions was purchased. Gidney bought his order of magnitude with approximate arithmetic and denser idle storage — and paid in runtime, eight hours becoming a week. Iceberg Quantum bought its order of magnitude with quantum LDPC codes —

and paid in connectivity and, above all, in the difficulty of decoding them fast enough to matter.

That is the lens for this issue. Difficulty in engineered systems behaves a little like energy: you rarely destroy it, you relocate it. The useful question about any breakthrough is not *how much smaller did the number get* but *which discipline now owns the hard part*. Because that is where the value migrates, where the scarce talent goes, and where — if you are allocating capital — the next moat gets dug.

This week's answer is startling. After a decade in which fault-tolerant quantum computing was a physics problem, the load-bearing constraint has quietly become a classical one: a latency-bound decoding computation that has to finish in ten microseconds, and that nobody has demonstrated at scale. The qubits got cheap. The silicon that watches them did not.

A thousand clean qubits, out of ordinary silicon

On 15 September, Iceberg Quantum and Diraq reported that Iceberg's Pinnacle error-correction architecture runs on Diraq's spin qubits without asking more of the hardware than a conventional surface code would. The result: 1,000 logical qubits from 150,000 physical ones — about ten times fewer than Diraq's own previous estimate.

BY THE INFLECTION DESK • SYDNEY, LEUVEN, SANTA CLARA

Quantum computers do not fail gracefully. A qubit holds its state the way a coin holds an edge-balance: perfectly, until something in the room breathes. The field's answer since the late 1990s has been redundancy — spend many fragile physical qubits to synthesise one reliable *logical* qubit, and accept the exchange rate. For twenty years that rate was set by surface codes, and it was brutal: one logical qubit per patch, the patch growing as the square of the protection you want. Every serious roadmap inherited that arithmetic, and every one ended in the same place — millions of qubits, sometime in the 2030s.

The exchange rate nobody could move

Surface codes are locked at an encoding rate of one logical qubit per d^2 physical ones — a ratio that decays toward zero as you demand more protection. At distance 24, the rotated surface code spends roughly 1,152 physical qubits to buy a single logical one. Gidney's 2025 "yoked" variant compressed idle storage to about 430, but the active patches doing real work still cost around 1,352 apiece.

Quantum low-density parity-check codes — qLDPC — break that lock. They encode k logical qubits into n physical ones at a rate approaching a non-zero constant as distance grows. The mathematics has been understood for years. The obstruction was never density. It was that nobody knew how to *compute* on a qLDPC code without surrendering the very density advantage that made it attractive.

The obstruction, and the assumption

Worse, qLDPC codes demand connections between qubits that are not neighbours. Surface codes are famously mod-

est: every interaction is nearest-neighbour, which maps cleanly onto a flat chip. qLDPC codes want long-range links. On superconducting hardware that is a wiring nightmare — the reason consensus held that qLDPC was a beautiful idea for a machine nobody could build.

Which is the assumption this week's result attacks, from a direction the field spent a decade dismissing. Not superconducting circuits, not trapped ions or neutral atoms, but silicon spin qubits: single electrons parked in transistor-like structures, fabricated on the same 300-millimetre CMOS lines that make the chip in your laptop. Diraq's qubits do not need wires between distant sites. They can be *shuttled* — physically moved across the array. Connectivity stops being a wiring problem and becomes a choreography problem, which is a far better kind of problem to have.

What Pinnacle actually did

Keep the non-locality inside the box

Pinnacle is built from modular processing blocks, and the insight that makes it implementable is almost embarrassingly simple: non-local connectivity is only ever required *within* a block, never across the whole device. Shuttling distances stay short. The Iceberg and Diraq teams then optimised the codes, circuits and shuttling schedules inside each block until the errors accumulated in transit contributed no more to the budget than the physical gates themselves. Decoherence from moving qubits around stopped being a new error source and became a line item already accounted for.

Agreement to within five percent

The test of any architecture paper is what survives contact with a real device. When the teams re-derived physical qubit counts against Diraq's actual hardware constraints, the numbers matched the Pinnacle paper's idealised estimates to within 5 percent. Numerical simulations that included shuttling-dependent noise confirmed the logical performance held. The architecture did not degrade on the way down to hardware — the usual fate of such proposals.

A hundred and one to one

The density numbers are where the surprise lives. Pinnacle's distance-24 generalised-bicycle block encodes 16 logical qubits in 1,020 data and check qubits; add the measurement gadgets and inter-block bridges and a full processing block runs to about 1,620 physical qubits. That is roughly **101 physical qubits per logical qubit** — a thirteen-fold density improvement over the hot surface-code patches it replaces.

Even if this paper's details were all wrong, the overarching point — that non-nearest-neighbour connections are a resource that can substantially reduce qubit overhead — would remain correct.

CRAIG GIDNEY, GOOGLE QUANTUM AI

The magic engine

Universal quantum computation needs a steady supply of high-fidelity "magic states," and producing them usually dominates the floorplan — Gidney's 2025 design gave six dedicated factory modules to the job. Pinnacle folds production into the code blocks themselves. Each magic engine splits its logical qubits into left and right sectors that alternate roles every cycle: one distils while the other injects, then they swap. The pipeline delivers one magic state at roughly 10^{-11} infidelity per logical cycle per processing unit, with about a 6 percent rejection rate at a physical error rate of 10^{-3} .

The number that got the headlines

Put it together and the February preprint reports RSA-2048 factored with roughly 97,000 physical qubits over about a year, or 151,000 over a week, or 471,000 over a single day — all assuming a physical error rate of 10^{-3} , a one-microsecond code cycle and a ten-microsecond classical reaction time. Against 20 million in 2019, it reads like collapse.

The quieter benchmark in the same paper deserves more attention than the cryptography. A 16×16 Fermi-Hubbard simulation — genuine condensed-matter physics, relevant to high-temperature superconductivity — needs an estimated 62,000 physical qubits at 10^{-3} , and 22,000 at 10^{-4} , against 940,000 and 200,000 for surface codes. Runtime: one and a half to three and a half minutes per shot. Minutes, not months.

From lab bench to fab line

The commercial logic here is not about quantum at all. It is about manufacturing. Diraq's qubits are made on imec's 300-millimetre CMOS-compatible platform — the same process technology, the same tooling, the same yield-learning machinery that the semiconductor industry has spent sixty years and a trillion dollars perfecting. In July the two demonstrated coherent operation of an eight-qubit linear array from that line, with two-qubit gate fidelities consistently above 99 percent, the threshold error correction needs.

Eight. That is the honest number, and it sits beside 150,000 like a rebuke. But the argument for silicon was never that it is ahead today. It is that the scaling curve is somebody else's problem, already solved. A CMOS-native quantum chip can in principle hold millions of qubits in a package small enough for a normal datacentre rack, built in fabs that already exist. Diraq's public target is over a million qubits by 2031.

The third party in the announcement is the one to watch. The mapping work was done on NVIDIA's CUDA-Q Logical, an orchestration layer released the day before, which compiles logical circuits down to physical operations on a specific device and generates full-stack resource estimates without rebuilding the toolchain each time. Fermilab used the same layer to compress a fault-tolerant architecture study from five months to three weeks.

That is what changed this week, more than any qubit count. Architecture and hardware can now be co-designed in simulation, against real device constraints, before the device exists. Iceberg — eight people and a \$6 million seed round from LocalGlobe, Blackbird and DCVC — is already mapping Pinnacle onto PsiQuantum's photonics, IonQ's trapped ions and Diraq's spins. The architecture is becoming a portable layer, sold across substrates. Someone is trying to be ARM for quantum computers, and this week they showed the instruction set runs on a second chip.

FIELD NOTES • HOW A LOGICAL QUBIT WORKS

- 1 Spread the information.** One logical qubit is stored across many physical ones, so that no single electron holds anything worth losing. Damage to a few members never destroys the message.
- 2 Ask indirect questions.** Measuring a qubit destroys it, so the machine instead measures *parity checks* — relationships between qubits. The answers, called the syndrome, reveal where an error struck without ever revealing the data.
- 3 Decode, then correct.** A classical computer reads the syndrome stream and infers which error occurred. It must finish before the next round arrives — roughly ten microseconds. This step is ordinary silicon doing ordinary computation, and it is now the hard part.

GLOSSARY

qLDPC code

An error-correcting code whose checks each touch only a few qubits, but not necessarily neighbouring ones. Far denser than surface codes; far harder to decode.

Syndrome decoding

The classical computation that turns a stream of parity-check results into a guess about which errors occurred.

Magic state

A carefully prepared resource state consumed to perform the gates error correction cannot do directly. Traditionally the largest cost centre on the chip.

Shuttling

Physically moving an electron across a silicon array to bring distant qubits together — spin qubits' substitute for long wires.

Quantum computing just became a *classical* hardware business

Here is the read that looks wrong. This week's announcement is being filed under *quantum threat accelerates*. The more useful filing is the opposite: Pinnacle has moved the binding constraint on fault-tolerant quantum computing out of *quantum physics entirely*, and into a latency-bound classical computation. If that is right, the company best positioned to capture quantum computing's value may not build qubits at all.

The argument runs through the decoder. Every architecture in this lineage — Gidney's and Pinnacle's alike — assumes a classical decoder that turns syndrome data into a correction within about ten microseconds. For surface codes that is a solved problem: decoding reduces to minimum-weight matching on a planar graph, which commodity FPGAs do in microseconds. For qLDPC codes it is a categorically different animal. The Tanner graph is non-planar, dense with short cycles, and full of trapping sets that make belief propagation converge to the wrong answer. The standard fix, BP with ordered statistics decoding, needs a matrix inversion with worst-case cubic complexity — on a 1,020-qubit block, inside ten microseconds.

Pinnacle's own simulations do not use a real-time decoder. They use most-likely-error decoding via mixed-integer programming: optimal, exponentially expensive, and never going to run in a control loop. The authors say plainly that building a fast enough decoder is outside the paper's scope. They are not hiding it. The coverage simply buried it.

So follow the scarcity. If logical qubits are cheap in qubits and expensive in real-time classical inference, then the moat is in decoder ASICs, low-latency interconnect, and the compilation layer that binds codes to devices. Which is precisely the layer NVIDIA shipped on 14 September. CUDA-Q Logical is not a quantum product. It is a bid to own the classical control plane of every quantum computer, whoever

builds the qubits — the same play, structurally, that CUDA ran on scientific computing. The qubit vendors may end up in the position of disk manufacturers: essential, capital-intensive, and not where the margin settles.

Why the consensus disagrees — and has a case

Three honest objections. **First, the paper is a preprint.** Not peer-reviewed, and parts of the generalised-bicycle code family are presented as conjecture rather than closed result. The concrete instances are simulated, not proven. Treat 98,000 as a target, not a floor.

Second, the difficulty may simply be unpayable. Gidney's objection is precise: the same reaction time is assumed for a much harder decoding problem. Aaronson's is architectural — qLDPC needs "wildly nonlocal measurements of the error syndromes." A constraint relocated to classical silicon is only good news if classical silicon can actually meet it, and at cubic scaling in ten microseconds, nobody has shown that it can. Rearranged difficulty is not reduced difficulty.

Third, the hardware is eight qubits. Diraq's best foundry-made array holds eight. Pinnacle's estimates also assume month-long sustained fault-tolerant operation, exceeding anything demonstrated by orders of magnitude. And the "surface code killer" still lights its magic engines using fifteen ancillary distance-9 rotated surface codes. No standards body has moved: NIST and the UK's NCSC still point at 2030–2035, and CNSA 2.0's milestones are unchanged.

All three are fair. None of them touches the structural point. Even if every number in the preprint is wrong, the direction is not: connectivity buys qubit overhead, and the bill arrives in classical compute.

200_{ms}

WARNING, BEFORE THE EVENT

ENERGY · FUSION

Princeton's Plasma Physics Laboratory reported that PACMAN — the Plasma Control and Automation Machine Learning Application Network — ran live on the DIII-D tokamak across five experiments. In one, it predicted a tearing mode, the instability that rips a fusion plasma apart, roughly 200 milliseconds before it appeared, and re-shaped the plasma to prevent it forming. Conventional controllers can only detect a tearing mode once it has already started. PACMAN closes its loop every 20 milliseconds. The significance is less about speed than about category: plasma control moves from reaction to anticipation, which is the difference between a physics experiment and a power station.

SOURCE · PRINCETON PLASMA PHYSICS LABORATORY, SEPTEMBER 2026

1

PROTEIN, MANY SYMPTOMS

BIOTECH · AGEING

Researchers reported that declining levels of Menin in the hypothalamus drive a startlingly broad set of ageing phenotypes in mice. Animals with lower Menin developed brain inflammation, memory deficits, impaired balance, reduced bone mass, thinner skin and shortened lifespan. Restoring Menin reversed several of these effects, and supplementing D-serine — an amino acid whose production falls when Menin does, and which regulates signalling between neurons — recovered cognitive performance. The claim worth testing is the architectural one: that a single hypothalamic regulator sits upstream of systems usually studied in isolation.

SOURCE · PLOS BIOLOGY / SCIENCEDAILY, 18 SEPTEMBER 2026

9%

OF THE FUEL BUDGET SPENT

SPACE · INSTRUMENTS

NASA's Nancy Grace Roman Space Telescope was allotted 441 pounds of hydrazine for its first mid-course correction. It used about 40 — under a tenth — and executed the burn with better than 99 percent accuracy. Combined with a launch mass of 17,760 pounds against a budgeted maximum of 21,605, the observatory now carries enough propellant for at least 22 years of science, against a 10-year design life. Roman reaches its L2 halo orbit around early December, with full science operations expected mid-2027. Conservative margins, compounded, bought a second mission for free.

SOURCE · NASA SCIENCE / ROMAN MISSION BLOG, 14 SEPTEMBER 2026

20M → 100k 150,000

The estimated physical qubits needed to break RSA-2048, 2019 to 2026.

GIDNEY & EKERÅ · PINNACLE PREPRINT

Physical qubits for 1,000 logical ones on Diraq's silicon — ten times fewer than its own prior estimate.

ICEBERG QUANTUM · DIRAQ

101

Physical qubits per logical qubit inside a Pinnacle processing block. Surface-code hot patches cost about 1,352.

PINNACLE ARCHITECTURE, ARXIV:2602.11457

10 μ s

The classical decoder reaction time every one of these estimates rests on. Nobody has met it for qLDPC codes.

ASSUMPTION, SECTION III.D

8

Qubits in the largest coherent array Diraq has made on imec's 300mm CMOS line. Two-qubit fidelity above 99%.

DIRAQ · IMEC, JULY 2026

7×

Fermilab's speedup in fault-tolerant architecture development — five months compressed to three weeks.

NVIDIA CUDA-Q LOGICAL

62,000

Qubits for a 16×16 Fermi-Hubbard simulation at 10^{-3} error — minutes per shot, against 940,000 for surface codes.

PINNACLE BENCHMARK

22 yrs

Roman Space Telescope's fuel-limited lifetime after one frugal burn, against a 10-year design life.

NASA SCIENCE

The map is not the constraint

Three stories this week, and one shape underneath them. A fusion controller that anticipates instead of reacting. A telescope that doubled its life by spending less than it was allowed. An error-correction architecture that made qubits cheap by making decoding expensive. None of them is a story about a new physical capability. All three are stories about where a system's real limit turned out to sit.

This is the recurring illusion in technology forecasting. We build roadmaps around the variable we can measure — qubit count, parameter count, transistor density, watts per kilogram — and then mistake the tracked variable for the binding one. The measured variable is usually the one that was hard *last* decade. It stays on the dashboard long after the difficulty has moved somewhere the dashboard does not look.

For fault-tolerant quantum computing, the difficulty has moved. It now lives in a classical decoder that must solve a hard inference problem in ten microseconds, over and over,

for a month without stopping. That is not a physics problem. It is a real-time-systems problem, and it will be solved — or not — by people who have never touched a dilution refrigerator.

Which suggests a discipline, for anyone allocating capital or attention against a technology with a steep-looking curve. Do not ask how fast the headline number is falling. Ask what was traded to make it fall, and who owns the new bill. The falling number tells you a field is alive. The relocated difficulty tells you where it is going.

Eight qubits on a foundry line, and a hundred and fifty thousand needed. That gap is real, and it may not close on anyone's announced schedule. But the interesting thing about this week was never the gap. It was the discovery that crossing it is now partly an ordinary engineering problem — one that ordinary silicon, made by ordinary companies, is allowed to attempt.

SOURCES & FURTHER READING

Iceberg Quantum and Diraq Demonstrate Pinnacle qLDPC Architecture on Spin Qubits — The Quantum Insider, 15 Sep 2026. <https://thequantuminsider.com/2026/09/15/iceberg-quantum-diraq-pinnacle-spin-qubit-hardware/>

The Pinnacle Architecture: Reducing the cost of breaking RSA-2048 to 100,000 physical qubits using quantum LDPC codes — Webster et al., arXiv:2602.11457, 12 Feb 2026. <https://arxiv.org/abs/2602.11457>

Pinnacle Architecture: 100,000 Qubits to Break RSA-2048, but at What Cost? — Marin Ivezić, PostQuantum.com, Mar 2026. <https://postquantum.com/post-quantum/pinnacle-architecture-break-rsa-2048-critical/>

How to factor 2048 bit RSA integers with less than a million noisy qubits — Craig Gidney, arXiv:2505.15917, May 2025. <https://arxiv.org/abs/2505.15917>

NVIDIA Expands Open Source CUDA-Q Platform for Fault-Tolerant Quantum Computing — NVIDIA Newsroom, 14 Sep 2026. <https://nvidianews.nvidia.com/news/nvidia-expands-open-source-cuda-q-platform-for-fault-tolerant-quantum-computing>

Imec and Diraq Demonstrate Eight-Qubit Silicon Spin Array on CMOS-Compatible Platform — The Quantum Insider, 13 Jul 2026. <https://thequantuminsider.com/2026/07/13/imec-diraq-eight-qubit-silicon-spin-qubit-array/>

Industry-compatible silicon spin-qubit unit cells exceeding 99% fidelity — Nature, 2025. <https://www.nature.com/articles/s41586-025-09531-9>

PACMAN AI framework for controlling fusion systems safely makes key decisions in milliseconds — Princeton Plasma Physics Laboratory, Sep 2026. <https://www.pppl.gov/news/2026/pacman-ai-framework-controlling-fusion-systems-safely-makes-key-decisions-milliseconds>

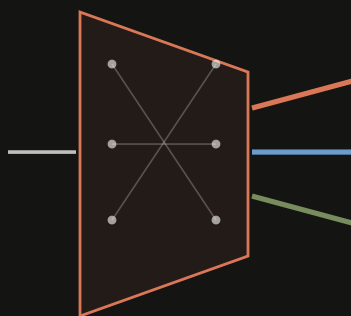
Scientists restore a brain protein and reverse signs of aging in mice — ScienceDaily, 18 Sep 2026. <https://www.sciencedaily.com/releases/2026/09/260918000437.htm>

Fuel Savings Double Potential Lifetime for NASA's Roman Mission — NASA Science, 14 Sep 2026. <https://science.nasa.gov/blogs/roman/2026/09/14/fuel-savings-double-potential-lifetime-for-nasas-roman-mission/>

Iceberg Quantum Unveils Pinnacle Architecture and Raises \$6 Million Seed Round — The Quantum Insider, 13 Feb 2026. <https://thequantuminsider.com/2026/02/13/trifecta-iceberg-quantum-unveils-pinnacle-architecture-claims-sub-100000-qubits-could-break-rsa-2048-and-raises-6-million-seed-round/>

INFLECTION.

THE WEEKLY MAGAZINE OF INNOVATION



NEXT ISSUE · FRIDAY

THE LENS

When a number you track starts falling fast, what did someone trade to make it fall — and who is holding that bill now?

RESEARCHED, WRITTEN & DESIGNED WITH CLAUDE.
TYPESET IN POPPINS & LORA ON THE ANTHROPIC PALETTE.
ISSUE 02 · FRIDAY, SEPTEMBER 25, 2026