



Opinions Libres

le blog d'Olivier Ezratty

Inside Google Willow

On December 9th 2024, Google made the news with **announcing** its 105-qubit superconducting Willow chip, with some **more details here**. It was linked to the release of a new **Nature paper** on this chip and extravagant claims about this quantum computer's ability to solve some specific problem in a couple minutes, that would take 10^{25} years to be solved on a classical supercomputer.

This drove a lot of comments around the world and the web, from positioning this announcement as a major breakthrough to being a huge marketing ploy. Comments are available with slides, posts and videos from **John Preskill, Scott Aaronson, Brian Lenahan, Bob Sutor, Alan Ho** from Qolab, **Maciej Malinowski, Michel Kurek**, and **Sabine Hossenfelder**, and more are yet to come.

However, forming an educated opinion on this announcement is a tedious and very technical task. I consolidate here some insights on these announcements and position them on the road towards fault-tolerant quantum computers. FTQC is about delivering on the many promises related to solving problems in healthcare, energy, climate, optimizations and whatever market. Google's Willow is a step forward in the creation of logical qubits, but it is just one step in a long journey that will take at least a decade.

Since I published first this post on December 12th, I updated it with the following:

- Explained the difference between “below threshold” and “breakeven”.
- Explained what is “exponential” in the error correction implemented in Willow.
- Provided more details on the cross-entropy benchmark and suggested a more appropriate benchmark, a large GHZ state.
- Estimated how many Willow chip equivalent would be required to break a RSA-2048 key (a lot!).
- Added some references to the Nature paper feedback from the referees.
- Added a reference to the largest GHZ test done as of December 2024.
- Corrected in February 2025 a significant mistake in my post, noticing that Willow is only implementing logical memory correction but not any logical single or two-qubit gate correction. Thanks to **Christophe Vuillot** from Inria who found this error in December.

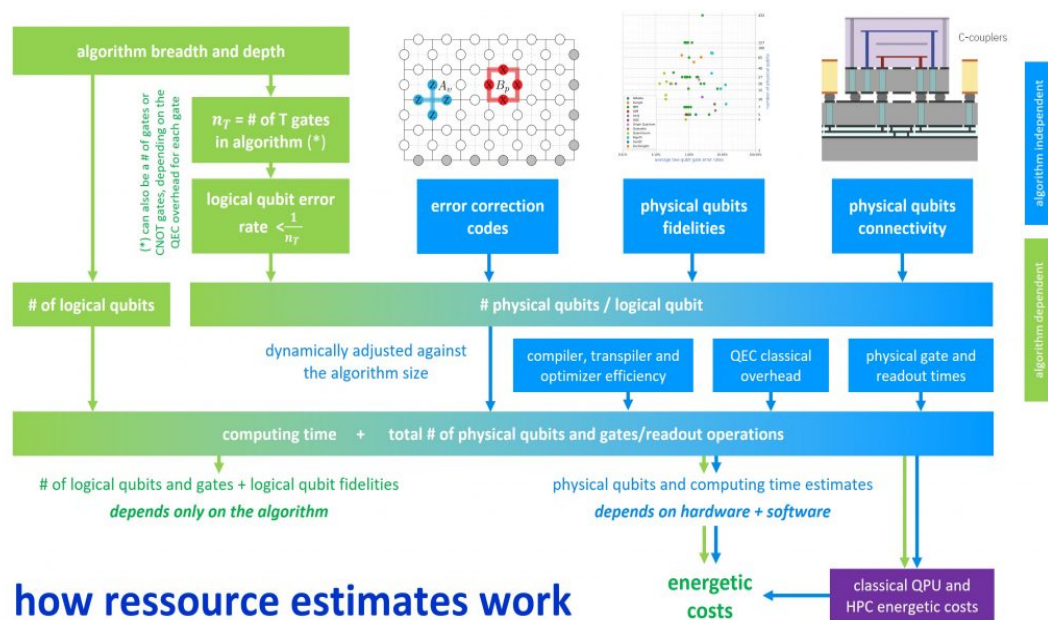
The FTQC quest

All qubit types are known to be noisy and don't enable the execution of large quantum algorithm circuits. Current quantum computers with between a dozen and a couple hundred qubits are limited in scope. They

belong to the so-called NISQ-era computing era (noisy intermediate scale quantum), based on using variational shallow quantum algorithms that are yet to deliver a quantum advantage compared to their equivalent best-in-class classical solutions. Several approaches are competing here. It starts with gate-based quantum computers with over 100 physical qubits like IBM's Heron r2 with 156 qubits and using the quantum error mitigation technique, which runs classical post-processing to extend the capacity of their noisy physical qubits. Trapped-ions quantum computers like Quantinuum's H2 system have better qubit gate fidelities but are currently limited to 56 physical qubits. Another option is to use analog quantum computers like the ones from D-Wave, Pasqal and QuEra, which are keen to solve relevant combinatorial and decision problems but have not yet proven a clear superiority against the best classical solutions, particularly in the heuristics and approximate solutions realm. However, these systems can also solve various physics problems (spin models and Fermi-Hubbard models in condensed matter physics) of great interest for fundamental research in various fields (magnetism, high-temperature superconductors, exotic quantum phases like spin liquids).

The accepted wisdom is that quantum computers able to deliver on their promise of solving intractable problems on classical systems will require quantum error correction techniques and the creation of so-called logical qubits, and fault-tolerant quantum approaches (FTQC, fault-tolerant quantum computers). This is, however, costly. The redundancy factor is huge. Depending on the algorithm size and qubit types, a "logical qubit" will be based on between 100 and 10K physical qubits. And the most well-known useful algorithms, like quantum chemistry simulations, need a couple of thousands of these logical qubits.

One key thing to understand is that there is no such thing as an "absolute" logical qubit. Quantum error correction maps a set of logical qubits on a much larger number of physical qubits, but this number depends on several factors: the size of the algorithm, the physical qubits fidelities (for single and two qubit gates but also qubit readout), the connectivity between the physical qubits, and the combination of quantum error correction codes that are used. This is described in the figure below, that is also in the 2024 edition of **my book**.



What is not well known is that a practical fault-tolerant quantum computer will rely on several error correction codes. There are different codes, for example, to correct single qubit gates and two-qubit gates, and there are different codes to correct simple gates (aka "Clifford gates", like the X, Y, Z and H gates) and non-Clifford gates like the T gate (which is a key enabler of exponential quantum speedups). On top of that, with surface codes, logical qubits are not isolated on a specific part of the chip. Several logical qubits are distributed and intertwined in a given a chip, or, may even be distributed over several chips.

The challenges to build these FTQC quantum computers are manifold:

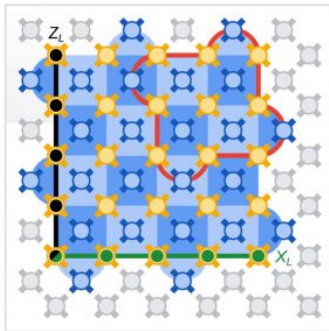
- Create systems with a very large number of high-quality physical qubits.
- Sustain gate fidelities at scale.
- Implement efficient error correcting codes covering a universal qubit gate set. That is technical but very important.
- Implement error correction in real-time.
- At some point, given the qubit chip size and quality limitations, interconnect several chips together, using entanglement sharing.
- Depending on the qubit modality, find ways to scale control electronics, cabling, cryogenics, laser power and the like.
- Improve existing algorithms to reduce their resource requirements. Another key goal is to develop quantum algorithms solving combinatorial optimization and decision problems with an exponential speedup.
- Develop new techniques assembling supercomputers and quantum computers to solve complicated problems like with quantum chemistry.

Put together, these are enormous scientific and engineering challenges. It will take at least a decade to succeed with finding solutions to all these challenges. And Google's recent performance is just one step along this long and winding road.

The writing was on the wall

Now let's put into context what Google announced on December 9th, 2024.

What was announced was already hidden in Google's roadmap and scientific papers dating from Summer 2022, so, 2,5 years ago. In **Suppressing quantum errors by scaling a surface code logical qubit** by Rajeev Acharya et al, arXiv then Nature, July 2022-February 2023 (44 pages), Google explained that it planned to create a logical qubit with better fidelities than their physical qubits with using a distance-7 surface code and about 98 qubits. In 2022, it had only 72 qubits at hand with the second generation of Sycamore chips. In 2022, the plan was to release that new chip in 2023. It took a bit longer to create a workable version of the chip. They probably had several iterations before creating a version where the distance-7 surface code could work successfully. Also, they embedded the already mentioned gap engineering feature which was invented more recently.



- Data qubit (d^2)
- Measure qubit ($d^2 - 1$)
- Unused
- Subset distance-3

Sycamore 72-qubit processor distance-5 logical qubit

Suppressing quantum errors by scaling a surface code logical qubit
by Rajeev Acharya et al, Google AI, July 2022-February 2023.

nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

nature > news > article

NEWS | 22 February 2023

Google's quantum computer hits key milestone by reducing errors

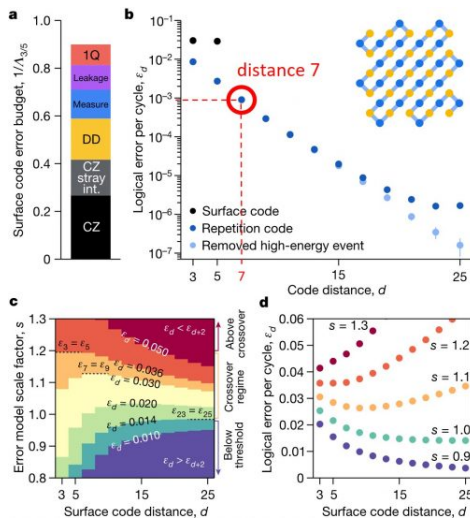
Researchers demonstrate for the first time that using more qubits can lower error rate of quantum calculations.

Davide Castelvecchi

logical qubit was not yet better than
underlying physical qubits

it was expected to reach this with >100 qubits
in 2023, and published in August 2024

That 2022 paper was about creating a “better logical qubit”, but it wasn’t yet as good as the underlying physical qubits. Google’s communication was very ambiguous for that respect and drove a lot of confusion. Their logical qubit was then based on a distance-5 surface code and 49 physical qubits running on this 72-qubit version of Sycamore.



Towards large-scale quantum error correction

To understand how our surface code results project forwards to future devices, we simulate the logical error performance of surface codes ranging from distance-3 to 25, while also scaling the physical error rates shown in Fig. 1c. For efficiency, the simulation considers only Pauli errors. Figure 4c,d illustrates the contours of this parameter space, which has three distinct regions. When the physical error rate is high (for example, the initial runs of our surface code in Fig. 3c), logical error probability increases with increasing system size ($\epsilon_{d+2} > \epsilon_d$). On the other hand, low physical error rates show the desired exponential suppression of logical error ($\epsilon_{d+2} < \epsilon_d$). This threshold behaviour can be subtle³⁸, and there exists a crossover regime in which, owing to finite-size effects, increasing system size initially suppresses the logical error per cycle before later increasing it. We believe our experiment lies in this regime.

Although our device is close to threshold, reaching algorithmically relevant logical error rates with manageable resources will require an error-suppression factor $\Lambda_{d(d+2)} \gg 1$. On the basis of the error budget and simulations in Fig. 4, we estimate that component performance must improve by at least 20% to move below threshold, and substantially improve beyond that to achieve practical scaling. However, these projections rely on simplified models and must be validated experimentally, testing larger code sizes with longer durations to eventually realize the desired logical performance. This work demonstrates the first step in that process, suppressing logical errors by scaling a quantum error-correcting code—the foundation of a fault-tolerant

Then, the August 2024 paper unfolded this 105 qubit chip later named Willow, and another 72 chip: **Quantum error correction below the surface code threshold** by Rajeev Acharya, Frank Arute, Michel Devoret, Edward Farhi, Craig Gidney, William D. Oliver, Pedram Roushan et al, Google, arXiv, August 2024 (27 pages).



It dealt with two achievements related to correcting error on a quantum memory:

- Correcting logical errors on the 72-qubit chip, and with real-time error correction.
- Correcting logical errors with the 105-qubits chip, but not in real time.

Both chips were below threshold results, meaning, logical memory fidelities being better than physical memory fidelities, in line with their prediction laid out in their 2022 paper.

By the way, the Google 2024 paper is a huge Russian dolls construction, referring to many other papers from the company, describing various relatively recent techniques used in their logical qubits experiment (examples below). This shows that what Google achieved is the result of a lot of work from many teams. The Nature paper has over 240 coauthors, showing the scale of their effort, including some third parties like William D. Oliver from MIT Lincoln Labs.

tional fidelities compared to our previously reported processors [17, 26]. The qubits have a mean operating T_1 of 68 μ s and T_2 ,CPMG of 89 μ s, which we attribute to improved fabrication techniques, participation ratio engineering, and circuit parameter optimization. Increasing coherence contributes to the fidelity of all of our operations which are displayed in Fig. 1b.

We also make several improvements to decoding, employing two types of offline high-accuracy decoders. One is a neural network decoder [27], and the other is a harmonized ensemble [28] of correlated minimum-weight perfect matching decoders [29] augmented with matching synthesis [30]. These run on different classical hardware, offering two potential paths towards real-time decoding with higher accuracy. To adapt to device noise, we fine-tune the neural network with processor data [27] and apply a reinforcement learning optimization to the matching graph weights [31].

We operate a distance-7 surface code memory comprising 49 data qubits, 48 measure qubits, and 4 additional leakage removal qubits, following the methods in Ref. [17]. Summarizing, we initiate surface code operation by preparing the data qubits in a product state in either the X_L or Z_L basis of the ZXXZ surface code [32]. We then repeat a variable number of cycles of error cor-

rection, during which measure qubits extract parity information from the data qubits to be sent to the decoder. Following each syndrome extraction, we run data qubit leakage removal (DQLR) [33] to ensure that leakage to higher states is short-lived. We measure the state of the logical qubit by measuring the individual data qubits and then check whether the decoder's corrected logical measurement outcome agrees with the initial logical state.

From surface code data, we can characterize the physical error rate of the processor using the bulk error detection probability [34]. This is the proportion of weight-4 stabilizer measurement comparisons that detect an error. The surface code detection probabilities are $p_{\text{det}} = (7.7\%, 8.5\%, 8.7\%)$ for $d = (3, 5, 7)$. We attribute the increase in detection probability with code size to finite size effects [24] and parasitic couplings between qubits. We expect both effects to saturate at larger processor sizes [35].

We characterize our surface code logical performance by fitting the logical error per cycle ε_d up to 250 cycles, averaged over the X and Z bases. We average the performance of 9 different distance-3 subgrids and 4 different distance-5 subgrids to compare to the distance-7 code. Finally, we compute the error suppression factor Λ using linear regression of $\ln(\varepsilon_d)$ versus d . With our neural

Then, the August 2024 arXiv became a **Nature paper** after a typical peer-reviewing process. The **reviewers feedback** is 34-pages long! It gives interesting insights on the λ (lambda) factor improvements achieved by Google that I describe a little bit in this post. The ping-pong discussion between the paper authors and referees also shows some dissent, referee #2 writing: “*The authors have minimally addressed my concerns and*

questions, and when they did respond, their replies often did not lead to any changes in the manuscript. Therefore, I still do not recommend the publication“. But the three other referees where OK with the paper and the author’s feedback to their comments.

What changes happened? Seemingly, not much in the paper, with minor edits based on the referees feedback. But certainly, good communication and marketing. Since error correction is a complicated matter, Google decided to switch the discussion to its absurd claims on its randomized benchmark comparing Willow with a supercomputer and its “septillion years” computing time. We will explain later why it doesn’t make much sense and is even contradictory with the logical qubit achievement. Indeed, why would we need so much overhead with quantum error correction if we can already get so much speedup with the available chip? Why is the shift from “useless” but still “amazingly fast” quantum computing to “useful computing” so difficult? If you are not a specialist, you are already lost in the understanding quantum maze. And some non-specialized commentators easily fall into the trap.

nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

[nature](#) > [news](#) > article

NEWS | 09 December 2024

‘A truly remarkable breakthrough’: Google’s new quantum chip achieves accuracy milestone

Error-correction feat shows quantum computers will get more accurate as they grow larger.

By [Davide Castelvecchi](#)

Willow news and left-overs

Let’s dig into what’s new in this Willow chip and what is the road ahead to really deliver on the FTQC promise.

Physical qubit fidelities. In August 2024, Google had not released characterization data for their 105-qubit chip, which was used for half of the experiments in their paper. It was **published** in December 2024 with the table below. The figures of merit seem to correspond to physical error rates despite the title “Quantum Error Correction (chip 1)”. It means that these are the physical characteristics of the chip used to test logical qubits.

Willow System Metrics	
Number of qubits	105
Average connectivity	3.47 (4-way typical)
Quantum Error Correction (Chip 1)	
Single-qubit gate error ¹ (mean, simultaneous)	0.035% ± 0.029%
Two-qubit gate error ¹ (mean, simultaneous)	0.33% ± 0.18% (CZ)
Measurement error (mean, simultaneous)	0.77% ± 0.21% (repetitive, measure qubits)
Reset options	Multi-level reset (1> state and above) Leakage removal (2> state only)
T ₁ time (mean)	68 μs ± 13 μs ²
Error correction cycles per second	909,000 (surface code cycle = 1.1 μs)
Application performance	$\Lambda_{3,5,7} = 2.14 \pm 0.02$

I just made a comparison with the latest IBM Heron r2 Marrakech 156-qubit system and Quantinuum H2 56-qubit system. For most figures of merit, Heron r2 is better than Google Willow, but Quantinuum is even better with gate fidelities, coherence times (T1, T2) and connectivity. One thing not shown here that disadvantages Quantinuum and trapped ions in general are gate times, which are several orders of magnitude slower than with superconducting qubits. But you could make other comparisons like the ratios between coherence times (T1/T2) and average gate times, which tend to favor trapped ions, at the expense of way longer computing times, that are not entirely offset by a better qubit connectivity (many-to-many at small scale). Willow chip's physical qubits coherence time is multiplied by 5 thanks to the error correction implemented in their experiment.

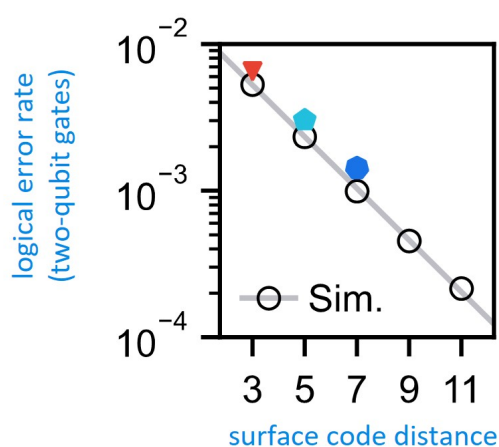
Chips	# qubits	1 qubit fidelity	2 qubit fidelity	readout fidelity	T1 (μs)	T2 (μs)	Connectivity
Heron r2 Marrakech	156	99.974%	99.78%	98.50%	191	107	1 to 2 and 1 to 3
Google Willow	105	99.965%	99.67%	99.23%	68	N/A	1 to 4
Quantinuum	56	99.971%	99.84%	99.85%	>1mn	4s	many to many

Quantum Memory: at this stage, Willow is made of quantum memory qubits that are not yet suitable for universal, fault-tolerant computation. First, no logical gates were implemented (X, CNOT, ...). On top of this, there is no support for T-gates or Toffoli gates error correction via magic state distillation, which is very resource-intensive. These gates are essential for achieving exponential acceleration, as they play a critical role in quantum Fourier transforms and modular exponentiations which rely on arbitrary rotation gates constructed with T-gates. These are particularly necessary for quantum phase estimation (QPE)-type chemical simulation algorithms and linear system-solving algorithms like HHL or those used in solving partial differential equations in engineering and computational fluid mechanics. This is a major upcoming milestone for Google and other platforms.

Surface code. Google used a XXXZ surface code that was created by Pablo Bonilla Ataides back in 2020. Their surface code has 4 leakage removal qubits which they call “qubit leakage removal” (DQLR). They don't indicate how this resource scales with the size of the surface code. See **The XXXZ Surface Code** by J. Pablo Bonilla Ataides, David K. Tuckett, Stephen D. Bartlett, Steven T. Flammia, and Benjamin J. Brown, arXiv, September 2020-April 2021 (6 pages).

Exponential what? Google says that their error reduction is “exponential”. It is matching theory with a λ factor of 2.14, which is the error reduction ratio when the code distance d increases by 2, from 5 to 7, and then, doubling the number of physical qubits, from 49 to 97, according to $(2d^2 - 1)$, d being the code distance. See **Improved accuracy for decoding surface codes with matching synthesis** by Cody Jones, Google, arXiv, August 2024 (16 pages).

That’s what is exponential here, meaning that as you decrease exponentially the logical qubit error rate, the required physical qubits number increases more or less polynomially. It shows up in the following chart, extracted from Google’s paper. In the X axis, you have the surface code distance, given the related number of qubits scales polynomially with this number. And in the Y axis, you have the logical memory error rate. Google plans that, as the surface code distance increases, logical error rates will follow the theoretical curve (small black circles). Google even expects to grow this λ factor thanks to improvements in qubit fidelities, which may enable better scaling capabilities with larger dimension surface codes. But nobody knows yet what will happen with larger numbers of physical qubits. Only experiments can tell.



Real-Time syndrome decoding: Google performed real-time error correction on a logical memory qubit using a distance-5 surface code on a 72-qubit chip, which is impressive. The distance-7 surface code on the Willow 105-qubit chip does not yet work in real time. The decoding time scales quadratically with the surface code distance. Google has not yet estimated the classical computational resources required to correct larger codes. In their Willow experiment, they used a 64-core AMD Ryzen Threadripper PRO 5995WX with 62 GB of RAM and a real-time version of Linux (mentioned page 14 of the **referee paper**).

They used machine learning and reinforcement learning for syndrome decoding, combining five different decoders including one using the matching synthesis technique and the Libra decoder, from Google, but with a less efficient correction yielding 0.171% errors. Real-time decoding of the syndrome measurements per cycle increases quadratically with code distance. They could not implement a real-time decoder with a distance-7 surface code as a result. Decoding was done in software and future designs may adopt hardware decoding, which would be much faster, more scalable and not sacrifice latency to accuracy (topic mentioned page 11 of the **referee paper**).

Some recent work related to real-time decoding was done by Riverlane and Rigetti this year, using FPGAs but working only with small distance surface codes. In the future, Riverlane plans to shift from FPGAs to more efficient and scalable ASICs electronics to support larger distance surface codes. See **Demonstrating real-time and low-latency quantum error correction with superconducting qubits** by Laura Caune, Earl T. Campbell, Alexander D. Hill et al, Rigetti, Riverlane, arXiv, October 2024 (11 pages).

Also, Google DeepMind created AlphaQubit in 2024, which relies on a Transformer-based reinforcement

learning technique to implement surface code syndrome decoding. But its scalability for large code distances remains to be checked. They evaluated its cost up to a distance 11 code, which is way below what is required for utility-grade FTQC. Beyond that, the Transformer requires increasing amounts of data which may be difficult to capture. See **Learning high-accuracy error decoding for quantum processors** by Johannes Bausch, Craig Gidney, Demis Hassabis, Sergio Boixo, Hartmut Neven, Pushmeet Kohli et al, Google, DeepMind, Nature, November 2024 (28 pages).

Correlated errors reduction: the 72-qubit and 105-qubit 2024 Google chips exhibit two significant innovations: mitigation of leakage errors using four dedicated qubits and substantial reduction of errors caused by cosmic rays through the technique Josephson junction gap engineering. For that respect, Google ran “repetition codes” up to distance-29 with as many qubits to correct only one type of (flip) error. They found that the logical performance was limited by these rare, correlated error events occurring approximately once every hour, or 3×10^9 cycles. It creates a floor of 10^{-10} for logical qubit fidelities. This is better than the 10^{-6} contribution of correlated errors coming from cosmic rays. This technique stems from work led by the French researcher Michel Devoret, who has been working for Google since 2023, with results published in **Resisting high-energy impact events through gap engineering in superconducting qubit arrays** by Matt McEwen, Michel Devoret, Alex Opremcak et al, Google, UCSB, arXiv, February 2024 (14 pages). Google’s efforts to improve circuit quality are progressing, aided by their in-house cleanroom.

Below threshold or breakeven? Both expressions are used in the Google paper, the first one being in the title. What’s the difference?

“under threshold” or “below threshold” corresponds to the condition where the qubit memory physical error rates are below a critical value that make it possible to create logical qubits with logical errors being below physical errors. With the surface code, the number of required physical qubits is infinite at this level. You need better qubit fidelities so that the overhead becomes reasonable, usually with an error rate of 10% of the threshold error rate. Then, as you grow the size of the surface code, expressed in a distance d , you reduce the error rate of the logical qubits.

“breakeven” corresponds to the situation when logical qubit’s performance are better than with physical qubits used without error correction. It can be expressed as having a logical qubit coherence time that is longer than the physical qubit coherence time. It also accounts to accounting for the whole overhead of implementing quantum error correction.

In Google’s case, the created logical qubits in Willow are both “under threshold” (relatively easy) and “breakeven” (much harder).

FTQC needs: To create a logical memory qubit with an error rate of 10^{-6} , Google estimates the need for 1,457 physical qubits, using a distance-27 surface code. For a 10^{-10} error rate, that number would grow to about 3,720 physical qubits and a code distance of 39. But this doesn’t consider the cost of correcting T gates. This relates to the “exponential” efficiency of error correction. But still, these are large numbers. Google’s plans are to create chips with up to 10K physical qubits. It means that only a few logical qubits will fit in these chips, whether we are in the MegaQOPs or multi-TeraQOPs range (QOP stands for quantum operations per seconds).

In September 2024, Google researchers proposed an efficient way to implement magic state “cultivation” (instead of distillation) for the creation of logical T gates. It could potentially reach logical fidelities of 2×10^{-9} , and even 4×10^{-11} but at the condition of halving the depolarizing qubit noise to 5×10^{-4} which remains challenging. See **Magic state cultivation: growing T states as cheap as CNOT gates** by Craig Gidney, Noah

Shutty, and Cody Jones, Google, arXiv, September 2024 (33 pages).

Building a 10K superconducting chip is a huge challenge. The academic fab that Google created in Santa Barbara will probably not be enough for many reasons I can't detail here. It is very well described in Qolab's recent blueprint. The startup was created by Google alumnis including John Martinis and Alan Ho. See **How to Build a Quantum Supercomputer: Scaling Challenges and Opportunities** by Masoud Mohseni, John M. Martinis et al, Qolab, HPE, Applied Materials, Synopsys, and many others, arXiv, November 2024 (64 pages).

Interconnect. This will necessitate interconnect techniques, at a minimum based on microwave guides, which suffer from losses. This challenge affects all players working with transmon-type superconducting qubits. Bosonic qubits, such as Alice&Bob's approach, may require fewer physical qubits to achieve comparable logical fidelities but have not yet been demonstrated in this regime. FTQC algorithms for practical use cases require fidelities between 10^{-8} and 10^{-20} , demanding even more robust logical qubits. Way above 4,000 physical transmon qubits may be needed to create logical qubits suitable for workloads that provide significant value to businesses, particularly in quantum chemistry applications.

Other logical qubits. Google is not the first to deliver logical qubits under break-even. It already happened with trapped ions back in 2021 and 2024 and with cold atoms in December 2023 with the Harvard/MIT/QuEra famous experiment led by Mikhail Lukin. See **Fault-tolerant control of an error-corrected qubit** by Laird Egan, Christopher Monroe et al, Maryland University, Duke University, January 2021 (9 pages), **Entangling four logical qubits beyond break-even in a nonlocal code** by Yifan Hong et al, University of Colorado Boulder and Quantinuum, arXiv, June 2024 (6 pages) and **Logical quantum processor based on reconfigurable atom arrays** by Dolev Bluvstein, Mikhail D. Lukin et al, Nature, December 2023 (42 pages). There is still a huge plus coming from Google: it didn't use any post-selection or post-rejection trick that QuEra and Quantinuum are using. They can repeat the correction cycle a great number of times. So, on paper, it scales much better. This is way better than Microsoft/Quantinuum **logical qubit** that can correct errors only once.

Challenges ahead for Google

Willow is just an interim step in a long roadmap towards FTQC.

Based on the above text, here is a new and more detailed consolidation of the challenges left to build useful FTQC quantum computers:

- Implement error correction for actual gates, like an **X gate** and **CZ/CNOT two-qubit gate**.
- Implement error correction for **T gates** with a minimum overhead cost.
- Implement **syndrome decoding in real-time** for logical qubits with a distance > 5 .
- Build **larger chips** with the same fidelities as with Willow, up to 10K qubits, and with low variability.
- Develop **reliable interconnect** solutions between these chips, using whether microwave connectivity (fixed, flexible) or optical photon transduction which is very complicated. We are in the very early stage here with very low TRLs (technology readiness levels).
- Fix **correlated error** beyond the 10^{-10} level, and yet to discover new sources of such correlated errors.
- Scale **cabling and cryoelectronics** to accommodate for large numbers of qubits in a cryostat.
- **Energetics, cost and price** of these FTQC quantum computers remains an open question.

This is a lot of work, just at the quantum and enabling technologies hardware levels! And not just for Google, but for the whole quantum industry.

The cross-entropy randomized benchmark

Now, let's talk about these benchmarks and claims comparing Willow and supercomputers.

The story is old. Google played with this in 2019 with its first “supremacy” claims using Sycamore and 53 qubits. Back then. Their Cross-Entropy Benchmarking (XEB) combined a set of random quantum gates with a homogeneous distribution. This last part scanned all the possible values (2^{53}) of qubits superpositions. In the supremacy regime, the so-called quantum computation had a 0.14% chance to produce the right results. It was executed 3 million times to generate an average measurement mitigating this low fidelity. The XEB benchmark was running 20 quantum gates cycles. Google stated then that they could run their “algorithm” in 2.5 days instead of 10,000 years on the IBM Summit supercomputer.

On October 21, 2019, IBM researchers **published a paper** in which they questioned Google's performance, stating that they could run their algorithm in 2.5 days instead of 10,000 years on the IBM Summit supercomputer. This would require adding 64 PB of SSD to the supercomputer, which they had not actually tested. That's about 7 racks full of SSDs at 2019 capacity. Later, several emulations of Google's noisy qubits could replicate their 2019 supremacy experiment from 2020 to 2022, using tensor networks.

In 2022, Google upgraded its experiment with using 70 of their latest Sycamore 72-qubit processor. The related paper did show that their 2019 supremacy was short-lived. It is now possible to emulate it classically in just 6.18 seconds, instead of the 2.5 minutes of the quantum experiment. This paper did show that the 2019 experiment could now be implemented classically in just 6 seconds on the DoE Frontier supercomputer, producing one million circuit samples. So, it was faster than the 3 minutes of the Google experiment. With 70 qubits, the classical timescale moved to 47 years. See **Phase transition in Random Circuit Sampling** by A. Morvan et al, Google AI, April-December 2023 (45 pages).

Exp.	1 amp.	1 million noisy samples		
	FLOPs	FLOPs	XEB fid.	Time
SYC-53 [9]	$6.44 \cdot 10^{17}$	$2.60 \cdot 10^{17}$	$2.24 \cdot 10^{-3}$	6.18 s
ZCZ-56 [10]	$6.24 \cdot 10^{19}$	$6.40 \cdot 10^{19}$	$6.62 \cdot 10^{-4}$	25.3 min
ZCZ-60 [11]	$1.32 \cdot 10^{21}$	$1.41 \cdot 10^{23}$	$3.66 \cdot 10^{-4}$	38.7 days
This work	$4.74 \cdot 10^{23}$	$6.27 \cdot 10^{25}$	$1.68 \cdot 10^{-3}$	47.2 yr

In Summer 2024, a China team led by Jian-Wei Pan even beat Google at its own game with emulating the 53-qubit experiment perfectly an order of magnitude faster and will less energy. In **Achieving Energetic Superiority Through System-Level Quantum Circuit Simulation** by Rong Fu, Jian-Wei Pan, et al, arXiv, June 2024 (11 pages) and **Leapfrogging Sycamore: Harnessing 1432 GPUs for 7 times Faster Quantum Random Circuit Sampling** by Xian-He Zhao, Jian-Wei Pan, Ming-Cheng Chen et al, arXiv, June 2024 (8 pages).

Now, when you upgrade the supremacy experiment from 70 to 105 qubits, it becomes even more difficult to emulate it classically with tensor networks. The couple minutes in Willow computing time become 10^{25} years of classical computing, way beyond the age of the universe, with existing available RAM. With infinite RAM, it would be reduced to about 300 million years, which is still large. The number of gate layers that was 20 in 2019 becomes 40 in Willow, thanks to better physical qubit fidelities. But Google didn't publish any scientific preprint documenting this new “supremacy”. We know that the probability of success is 0.1%, a little lower than the 0.14% of the 2019 experiment. By the way, nobody can check the results of this random computation

since, as **Scott Aaronson writes**, it would also require 10^{25} years of classical computing.

On top of that, Google didn't use the same 105-qubit chip. It used some "chip 2" with different figures of merit. The two-qubit physical error rates are even better with 0.14% instead of 0.33% in "chip 1" but with using a different gate, an iSWAP gate. iSWAP gates have higher fidelities than CZ gates when using tunable couplers, which is the case in Google's Willow chip.

Random Circuit Sampling (Chip 2)	
Single-qubit gate error ¹ (mean, simultaneous)	$0.036\% \pm 0.013\%$
Two-qubit gate error ¹ (mean, simultaneous)	$0.14\% \pm 0.052\%$ (iswap-like)
Measurement error (mean, simultaneous)	$0.67\% \pm 0.51\%$ (terminal, all qubits)
Reset options	Multi-level reset (1> state and above) Leakage removal (2> state only)
T_1 time (mean)	$98 \mu\text{s} \pm 32 \mu\text{s}^2$
Circuit repetitions per second	63,000
Application performance	XEB fidelity depth 40 = 0.1%
Estimated time on Willow vs classical supercomputer	5 minutes vs. 10^{25} years

¹Operation errors measured with randomized benchmarking techniques and reported as "average error"

²Chip 1 and 2 exhibit different T_1 due to a tradeoff between optimizing qubit geometry for electromagnetic shielding and maximizing coherence

Now, I have to explain you why this randomized benchmark is not "actual" computing from a computer science standpoint.

The cross-entropy randomized benchmarking used by Google is a sort of physical experiment that is hard to simulate classically. From a computational standpoint, this benchmark has no useful input data since the quantum gates are random and it doesn't yield any useful output data corresponding to some problem to solve. It just outputs some useless data distribution. It is a giant and costly random numbers generator. This is even crazy given that the challenge to emulate the Willow chip is about simulating classically all its defaults since it's full of noise! By the way, Scott Aaronson himself proposed a QRNG application for these supremacy experiments, in **Certified Randomness from Quantum Supremacy** by Scott Aaronson, and Shih-Han Hung, arXiv, March 2023 (84 pages). We should then consider the various Device Independent QRNG solutions which provide similar services and at a much lower cost.

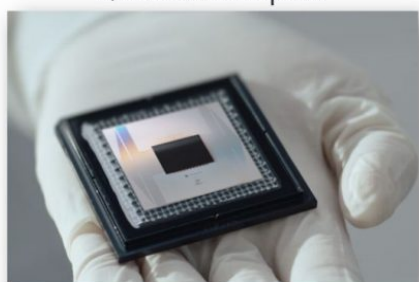
Back in 2019 and 2020, after its supremacy experiment, Google started to publish case studies with its 53-qubit

Sycamore processor. They were using between 15 and 23 qubits. They couldn't use the whole 53 qubits chip because they were too noisy. And 15 to 23 qubits could be easily emulated on your own laptop. See for example **Hartree-Fock on a superconducting qubit quantum computer** by Google AI Quantum and Collaborators, April 2020 (27 pages) and **Quantum Approximate Optimization of Non-Planar Graph Problems on a Planar Superconducting Processor** by Google AI Quantum and Collaborators, April 2020 (17 pages). We can suspect that Google researchers will publish in 2025 some NISQ experiments with Willow that may use a little more qubits, but not necessarily in a quantum advantage regime. It will be interesting to see whether Google and its academic partners will use quantum error mitigation techniques like IBM, which extend the size of the quantum circuits usable in NISQ mode (without error correction). IBM's plan is to exceed 10K gates in these circuits. Their current recent record is with 5K gates over 133 to 156 physical qubits (Heron r1 and r2).

You can play with other analogies to describe the absurdity of the XEB benchmark used by Google: simulating a drop of water at the atomic level with its $\sim 10^{21}$ water molecules (there are obviously simpler ways to do that, using computational fluid dynamics, interface modeling, etc.), a laser traversing a bottle of milk (an idea from Alain Aspect) or a toaster, per the nice and fancy idea by **Todd Austin** (below). These are all physical systems too complicated to simulate classically, but they don't provide much actual computing power. But I have to admit that these comparisons may be exaggerated since it doesn't deal with qubits and qubit gates.

Achieving Classical Supremacy?

Quantum Computer



Year invented: 2024
 Program: Random quantum circuit
 Runtime: 5 min
 Speedup over supercomputer: 1.05×10^{30}

Maillard Computer



Year invented: 1893
 Program: Maillard reaction
 Runtime: 3 min
 Speedup over supercomputer : 3.5×10^{35}

Another way to look at the situation is to evaluate the resources needed to break a 2048-bit RSA key with such a system using Peter Shor's famous integer factoring algorithm, that was invented in 1994. A minimum of 4,000 logical qubits are needed with an error rate below 10^{-12} . This results in requiring logical qubits with a minimum of 4,000 physical qubits each. So you'd roughly need 16M physical qubits vs Willow's 105 qubits. Simple maths tells you that it is the equivalent of 160,000 Willow chips, and it is about the same for running interesting chemical simulations. Of course, you can count on some future optimizations that may reduce this footprint. But, as mentioned above in this paper, scaling quantum computing is not just a matter of brute force. It will uncover new challenges, particularly at the chip interconnect level.

A really interesting benchmark, but not a "consumer" one, would be to rate the fidelity of the largest GHZ state possible. This is a quantum state superposing two values from the computational basis: $|0000 \dots 0000\rangle$ and $|1111 \dots 1111\rangle$, with N zeros and ones, N being equal or inferior to the number of qubits in the chip. Such a benchmark should have a related fidelity, and without using the post-selection trick used by QuEra and Quantinuum. I would suspect that the largest GHZ state possible with an efficiency over 50% would be quite small on Willow. A GHZ benchmark is a litmus test of the capability of a processor to handle entanglement at

scale. For N qubits, it requires at least N gate cycles, depending on the qubit connectivity (a set of Hadamard gates on all qubits and then a series of CNOT gates for entangling qubits).

So far, the best GHZ results have been obtained with trapped-ions quantum computers thanks to their high qubit fidelities and many-to-many connectivity. Quantinuum **reached** a GHZ of 50 qubits in December 2024 on their H2 processor which hosts 56 ion qubits, although with using an error detection technique. It was announced during the Q2B Santa Clara conference, the same week as Google's Nature paper.

Many worlds? Please, one is enough!

The last and most absurd claim from Google was this sentence found on their blog: that their experiment “*lends credence to the notion that quantum computation occurs in many parallel universes, in line with the idea that we live in a multiverse, a prediction first made by David Deutsch*”. This belongs to the Many Worlds Interpretation of quantum physics which is still debated (look at some references about it in my book). Sorry guys! There was only one Willow chip used in the experiment. You don't need “many worlds” to explain how superconducting qubits operate.

Willow's performance on this benchmark is astonishing: It performed a computation in under five minutes that would take one of today's [fastest supercomputers](#) 10^{25} or 10 septillion years. If you want to write it out, it's 10,000,000,000,000,000,000,000,000 years. This mind-boggling number exceeds known timescales in physics and vastly exceeds the age of the universe. *It lends credence to the notion that quantum computation occurs in many parallel universes, in line with the idea that we live in a multiverse, a prediction first made by David Deutsch.*

These latest results for Willow, as shown in the plot below, are our best so far, but we'll continue to make progress.

Challenges ahead

Let's end this by answering some simpler questions:

- Is Google making progress? Yes, definitively, but they are not the only ones implementing logical qubits.
- Is this Google announcement showing some accelerated or exponential progress? I don't think so. They followed their own roadmap with being a little late. Moving from 53 to 105 qubits in 5 years is far from being some exponential progress, particularly given millions of physical qubits are required to do something useful.
- Are the Internet (with RSA) and the Bitcoin threatened? Not more than before. It will take at least a long decade before a Google or other quantum computer can assemble over 4,000 logical qubits with high logical fidelities (better than 10^{-10} error rates) that are necessary to break current asymmetric encryption.

Again, creating fault-tolerant quantum computers is a very long journey, where patience is a virtue.

Cet article a été publié le 12 décembre 2024 et édité en PDF le 18 février 2025.
(cc) Olivier Ezratty – “Opinions Libres” – <https://www.oezratty.net>