



Factoring 2048-bit RSA Integers in 177 Days with 13 436 Qubits and a Multimode Memory

Élie Gouzien * and Nicolas Sangouard †

Université Paris–Saclay, CEA, CNRS, Institut de Physique Théorique, 91191 Gif-sur-Yvette, France

(Dated: September 29, 2021)

We analyze the performance of a quantum computer architecture combining a small processor and a storage unit. By focusing on integer factorization, we show a reduction by several orders of magnitude of the number of processing qubits compared with a standard architecture using a planar grid of qubits with nearest-neighbor connectivity. This is achieved by taking advantage of a temporally and spatially multiplexed memory to store the qubit states between processing steps. Concretely, for a characteristic physical gate error rate of 10^{-3} , a processor cycle time of 1 microsecond, factoring a 2048-bit RSA integer is shown to be possible in 177 days with 3D gauge color codes assuming a threshold of 0.75 % with a processor made with 13 436 physical qubits and a memory that can store 28 million spatial modes and 45 temporal modes with 2 hours’ storage time. By inserting additional error-correction steps, storage times of 1 second are shown to be sufficient at the cost of increasing the run-time by about 23 %. Shorter run-times (and storage times) are achievable by increasing the number of qubits in the processing unit. We suggest realizing such an architecture using a microwave interface between a processor made with superconducting qubits and a multiplexed memory using the principle of photon echo in solids doped with rare-earth ions.

Introduction — Superconducting qubits form building blocks of one of the most advanced platforms for realizing quantum computers [1, 2]. The standard architecture consists of laying superconducting qubits in a 2D grid and computing using only neighboring interactions. Recent estimations showed however that fault-tolerant realizations of various quantum algorithms with this architecture would require millions of physical qubits [3–5]. These performance analyses naturally raise the question of an architecture better exploiting the potential of superconducting qubits.

In developing a quantum architecture we have much to learn from classical architectures. Realizations using trapped ions for example combine processing with storage units [6]. The authors of Ref. [7] realized that key quantum algorithms are mostly sequential meaning that we may only need a small computing block for all the qubits in the storage unit in this architecture. Ongoing experimental efforts aim at exploiting this idea to reduce the number of superconducting qubits in the standard approach to quantum computing by adding a quantum memory implemented with spins or atoms [8–10]. A detailed analysis of the performance of this hybrid architecture is however missing.

We here report on such an analysis by considering a quantum memory that can store multiple spatial transverse and temporal modes. The memory can be thought of as a qubit register in which the address of each qubit is identified by a temporal and a spatial index. When a given qubit needs to be processed, its state is released and mapped into the processor by means of a microwave field in a temporal and spatial mode corresponding to the

qubit address. When the processing is done, the qubit state is mapped back to the memory and stored until another processing operation is needed.

More precisely, we use 3D error-correction codes [11] in which the address of each (dressed) logical qubit is encoded into a 3D structure of physical addresses, two dimensions being encoded in space and one in time (see [Figure 1](#)). Error-correction and logical gates are applied by sequentially releasing physical qubits corresponding to different “horizontal” slices (with different temporal indexes) and by processing each slice (with the same temporal indexes) simultaneously.

We assess the performance of this architecture through a version of Shor’s algorithm [12] proposed by Ekerå and Hästad [13]. The algorithm is a threat for widely used cryptosystems based either on the factorization [14] or the discrete logarithm problem [15, 16]. It can also be considered as a certification tool to check the proper functioning of an actual quantum computer as its outcome can be verified efficiently. Last but not least, the cost of its implementation has been evaluated using plausible physical assumptions for a large scale processor with a standard 2D grid of superconducting qubits (a characteristic physical gate error rate of 10^{-3} , a surface code cycle time of 1 μ s, and a reaction time of 10 μ s): it was estimated that it should be possible to factor a 2048-bit integer, typically used in the Rivest–Shamir–Adleman (RSA) cryptosystem, in 8 hours with 20 million qubits [3].

By taking this estimation as a reference, we estimate the cost of implementing the same version of Shor’s algorithm in terms of physical processing qubit number,

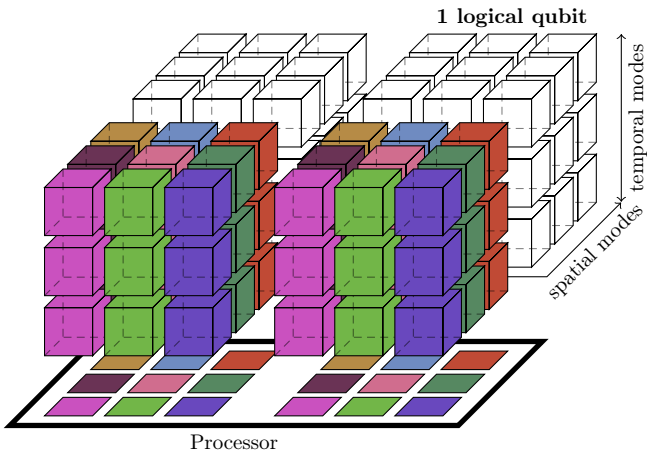


Figure 1. Quantum computer architecture using a processor made with a 2D grid of qubits and a memory operating as a qubit register where the address of each qubit is specified by a temporal and spatial index. Only (dressed) logical qubits are represented; additional ancillary qubits are used for measuring the operators for error correction.

multimode capacity, memory storage time, and run-time. Our evaluation is given in the case where the processor is made with two (dressed) logical qubit slices. Under the assumptions used in Ref. [3] for the gate error rate and the cycle time, we show that it should be possible to factor a 2048-bit RSA integer in 177 days using a multimode memory with a storage time of about 2 hours and a processor including 13436 physical qubits—a reduction by more than 3 orders of magnitude of the number of physical qubits, as compared to the standard architecture without memory [3], at the cost of a ≈ 500 times longer run-time. By inserting additional error-correction steps, we show that the storage time can be significantly reduced at the cost of a slight increase of run-time. We also explain how shorter run-times and storage times are achievable at the cost of increasing the number of qubits in the processing unit. We propose a realization of such an architecture using a microwave interface between a processor made with superconducting qubits and a multiplexed memory using the principle of photon echo in solids doped with rare-earth ions embedded in cavities.

Principles of (a variant of) Shor’s algorithm — Consider the factorization of $N = p \times q$, the product of two prime numbers of similar sizes, p and q . We note n the number of bits involved in the binary representation of N , that is $2^{n-1} \leq N < 2^n$. While no efficient classical factorization algorithm is known, Shor’s algorithm and its variants factor N with a polynomial complexity into n [12, 13, 17–20].

The version of Shor’s factorization algorithm proposed by Ekerå and Håstad [13] starts by randomly selecting an integer g in the multiplicative group of inte-

gers modulo N , \mathbb{Z}_N^* , and defining $h = g^{(N-1)/2}$. As the order of \mathbb{Z}_N^* is $\phi(N) = (p-1)(q-1)$, we have $h = g^{(pq-p-q+1)/2} g^{(p+q-2)/2} \equiv g^{(p+q-2)/2} \pmod{N}$ where the last equivalence is the result of the Chinese remainder theorem. Under the assumption that the order r of g (the smallest non-negative integer such that $g^r \equiv 1 \pmod{N}$) satisfies $r > (p+q-2)/2$, computing the discrete logarithm of h modulo N , as detailed later, yields $l = (p+q-2)/2$. For large N , the assumption is verified with a high probability [13]. Using $N = pq$ and $l = (p+q-2)/2$, where N and l are both known, p and q are recovered by choosing one solution of the equation $N = p(2l+2-p)$, and then exploiting $q = 2l+2-p$.

The discrete logarithm is computed in three steps. First, the exponentiation $(e_1, e_2) \rightarrow g^{e_1} h^{-e_2}$ is applied once on two quantum registers prepared in a superposition of every possible value of e_1 and e_2 , respectively. Two quantum Fourier transforms are then applied independently to the two registers before being measured. Finally, a classical postprocessing extracts the discrete logarithm l of h modulo N from the measurement results. Because the measurements are performed directly after the Fourier transform, the cost of exponentiation largely dominates the cost of Ekerå and Håstad’s algorithm (see Appendix A).

Number of gates — The modular exponentiation needed in Ekerå and Håstad’s algorithm, *i.e.*, the operation $|e\rangle |1\rangle \mapsto |e\rangle |g^e \pmod{N}\rangle$, with the input e and the output $g^e \pmod{N}$ encoded on n_e and n bits, respectively, can be decomposed into n_e multiplications, each being decomposed into $2n$ controlled additions of integers of typical size n and one controlled swap between two registers of size n , giving a total number of $2n_e n$ (n_e) controlled additions (swaps between registers, respectively) (see Appendix B for details). Each modular addition is obtained with a standard adder circuit at the cost of a specific representation—the coset representation (see Appendix C)—adding m additional qubits to the register. A controlled swap operation between two qubits can be performed using two controlled NOTs (CNOTs) and one Toffoli gate. Hence, the total cost for controlled swaps operating on two registers using $n+m$ qubits is of $2(n+m)$ CNOTs and $n+m$ Toffoli gates (see Appendix B). For the controlled addition, we can use a semi-classical adder whose mean cost for integers of size $n+m$ is of $5.5(n+m) - 9$ CNOTs and $2(n+m) - 1$ Toffoli gates (see Appendix B). Given the number of gates in controlled addition and swap operations, the number of additions and swaps in the multiplication, and the number of multiplications in the modular exponentiation, the cost of factorization can easily be estimated (see Appendix B). This cost can however be reduced using windowed arithmetic circuits [21]. The basic idea consists of grouping the bits of e by blocks (each includ-

ing w_e bits) for controlling each multiplication, hence reducing the number of these multiplications. Similarly, for each multiplication input bits are grouped (in blocks including w_m bits) to reduce the number of additions composing it. As detailed in [Appendix D](#), the cost of exponentiation is dominated in this case by $2^{\frac{n_e(n+m)n}{w_e w_m}}$ 1-qubit gates, $[2^{w_e+w_m}n + 12(n+m)]^{\frac{n_e(n+m)}{w_e w_m}}$ CNOTs, and $4^{\frac{n_e(n+m)^2}{w_e w_m}}$ Toffoli gates. We emphasize that this is a first order estimation. In the code used to compute the required resources and find optimal parameters, the complete formulae have been used [\[22\]](#).

Error correction — The error correction is achieved using 3D gauge color codes, a family of subsystem codes [\[11\]](#). A first code admits a transversal implementation of CNOT and Hadamard gates while a second code accepts a transversal implementation of the non-Clifford T gate. Switching between the two codes gives a universal set of gates without the need for state distillation [\[23\]](#), contrary to standard ways of operating the surface code [\[24\]](#).

The two codes are based on a shared geometrical structure: a large tetrahedron constructed from elementary tetrahedrons (see [Appendix E](#) for details). A physical qubit is attributed to each elementary tetrahedron. As in any subsystem codes, the stabilized subspace is split into a tensorial product of the (bare) logical and gauge qubits (the dressed logical qubit includes the bare logical qubit and gauge qubits). A set of operators—generators of gauge operators—are measured, each being the product of (up to six) X (or Z) operators associated to qubits corresponding to tetrahedrons sharing the same edge. From these measurements, the values of stabilizers of the two codes are deduced. In the code used for implementing H and CNOT gates, the stabilizers are defined from the vertices, *i.e.*, the product of X (or Z) operators associated to qubits corresponding to tetrahedrons sharing the same vertex. In the code used for implementing T gates, the stabilizers are defined from the vertices for X operators and from the edges for Z operators. The value of an operator represented by a vertex is classically recovered by multiplying the measurement results of combinations of specific edges ending at the given vertex. Several combinations are possible giving redundancies that can be exploited to achieve fault-tolerant error correction with only one run of measurements [\[25\]](#). The structure of codes in which the stabilized subsystem is the tensor product of the gauge and (bare) logical subsystems guarantees that measurements of gauge operators do not reveal the value of the (bare) logical qubit (see [Appendix E](#)).

To account for the additional resource needed to implement these codes, we use an estimation of the residual

error probability on one logical qubit given in [\[26, eq. \(4\)\]](#)

$$p_{\text{logical}} = A \exp \left[\alpha \log \left(\frac{p}{p_{\text{th}}} \right) d^\beta \right] \quad (1)$$

where $A \approx 0.033$, $\alpha \approx 0.516$, $\beta \approx 0.822$, p is the error probability per physical qubit, d the code distance which is related to the number of physical qubits per logical qubits (see below) and p_{th} the fault-tolerance threshold. While the circuit-level threshold is unknown, we choose $p_{\text{th}} = 0.75\%$ as a working hypothesis and give in [Appendix E 4](#) the run-time and the resource as a function of the code threshold.

Architecture — For simplicity, the tetrahedral structure of the error correction (see [Appendix E](#)) can be included into a large cube in which physical qubits are now represented by elementary cubes (see [Figure 1](#)). The large cubes are stored into the memory and loaded by slices into the processor when they need to be processed. We size the processor such that one slice of two large cubes can be loaded simultaneously, which is convenient to perform 2-qubit gates efficiently. Each gate is immediately followed by an error-correction round on the processed qubits. This is done by reloading again each slice sequentially in the processor and by measuring the gauge generators (before recovering classically the code stabilizers), each of them using up to six 2-qubit gates, one auxiliary qubit and one measurement of this auxiliary [\[23, 27\]](#). Note that the codes of interest are 3D local and the auxiliary qubits only need to keep coherence for the time of loading and measuring two successive slices for successfully performing a stabilizer measurement. Once the syndromes are obtained and the errors are detected, the correction of these errors is delayed and merged with the next operation applied on the qubit to be corrected. Further note that all-to-all connectivity between the logical qubits is achieved if each physical address in the memory can be mapped to three physical qubits in the processor: two for the 2-qubit gates (depending on whether the physical qubit is the logical control or target qubits) and one for the error correction. For achieving a code distance d the number of physical qubits in the processor is $n_{\text{qubits}} = 2 \times 2 \times \frac{3d^2+2d-3}{2}$, corresponding to two logical qubit slices (see [Appendix E](#)) and including the ancillary qubits (essentially one per physical qubit) needed for stabilizer measurements. For a code distance d , we approximate the time it takes to perform one (1-qubit or 2-qubit) logical gate by $2(d-2)t_c$ where t_c is the cycle time of the 2D processor (time to load one qubit slice; to measure the stabilizers, which is longer than the gate operation; and to reload the slice into the memory) and the factor 2 comes from the fact that the gate is immediately followed by an error-correction round.

Cost evaluation — To evaluate the resources required for integer factorization, we consider the total number

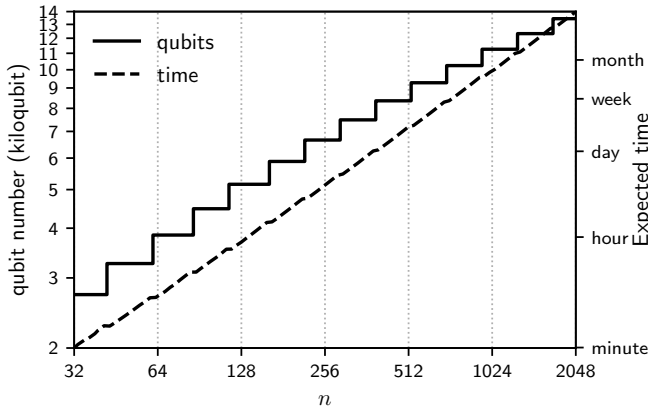


Figure 2. Number of qubits in the processor and run-time to factor n -bit RSA integers with a computer architecture using a multimode memory.

of gates involved in the logical circuit. The total run-time for one attempt is obtained by multiplying the gate number by the time it takes to perform one gate, while the success probability is deduced from the logical error probability [Eq. (1)]. Following Ref. [3], we consider a cycle time of $t_c = 1 \mu\text{s}$ and a mean error per physical qubit and per gate of $p = 10^{-3}$. Note that the mean error per gate now includes errors during reading of and writing into the memory.

The cost evaluation is finally obtained by optimizing the two window parameters w_m and w_e , the coset representation padding m and the code distance d in order to minimize the volume $t_{\text{exp}} \times n_{\text{qubits}}$. $t_{\text{exp}} = \frac{t}{p_s}$ is the average time to obtain the result (several attempts might be necessary), with t the computation time per attempt and p_s the success probability.

Results — The required resources to factor a n -bit RSA integer are presented in Figure 2 and discussed in Appendix F. Our estimation suggests that the factorization of a 2048-bit integer corresponding to the most common RSA key size would be possible in about 177 days with a processor having only 13 436 qubits. Concerning the memory, we made the hypothesis of an error per cycle of $p = 10^{-3}$, including the reading and writing error. As previously discussed, we need a memory for which each mode can be mapped to three different qubits of the processor. We estimated the maximum time between storage and readout of the same qubit to be less than 2 hours. A memory with a storage time of at least 2 hours is however not necessary as error-correction steps can be implemented periodically at the cost of increasing the run-time. Error correction of all the qubits stored in the memory is estimated to take 186 ms with a processor having 13 436 bits, meaning that the storage time simply needs to be longer than 186 ms. Applying a correction

every second for example would increase the run-time by about 23%. Note also that both the run-time and storage time can be reduced by increasing the size of the processor (see Appendix F). We also estimated that 28 million spatial modes and 45 temporal modes need to be stored. Note that the number of stored modes does not enter in the volume and is thus not optimized (see Appendix G). Note also that qubit addresses in the memory can be identified by temporal indexes only at the cost of longer run-time when photon-echo type protocols are used, *cf.*, below for a concrete example.

Implementation — Our proposal provides a viable solution to get rid of the individual control of millions of qubits but the challenge now relies on the realization of an efficient multimode quantum memory. As shown in Ref. [28], such a memory could be implemented using a solid-state spin ensemble (\bar{N} spins with an inhomogeneous spectral broadening Γ), resonantly coupled (with single spin coupling rate g) to a frequency tunable single-mode microwave resonator (of length L and with damping rate κ to an external transmission line). The resonator serves to enhance microwave absorption and re-emission by the spins. In particular, unit efficiency absorption of a microwave field can be realized if the finesse \mathcal{F} of the resonator matches the single-path absorption αL of spins $\mathcal{F} = (\alpha L)^{-1}$, *i.e.*, if the cooperativity $C = \frac{g^2 \bar{N}}{\kappa \Gamma} = \alpha L \times \mathcal{F} = 1$ [29]. Once absorbed, the microwave field can be re-emitted by time reversing the inhomogeneous dephasing using a spin echo technique [30]. Detuning the resonator off and on resonance at the right time, the spin coherence is recovered, leading to a noise-free, unit re-emission probability of the stored photon if $C = 1$ [28]. In the regime $\kappa \gg g\sqrt{\bar{N}} \gg \Gamma$, the memory bandwidth is given by 4Γ [28], meaning that any input with a spectrum, say, ten times thinner *i.e.*, $4\Gamma/10$ can be stored with close to unit efficiency. Furthermore, the time duration during which an optical coherence can be preserved is limited by the inverse of the homogeneous linewidth γ_h [28]. Assuming that the storage efficiency is unchanged if the storage time is hundred times shorter than γ_h^{-1} , this means that the number of temporal modes that can be stored with almost unit efficiencies is roughly given by $\Gamma/(250\gamma_h)$. Interestingly, a well-identified temporal mode can be released while keeping all the other modes in the memory by appropriately detuning the resonator off and on resonance with the spins at the cost of introducing a dead time between two readouts of half the duration of the stored train of pulses on average.

To give an idea of what could be realized in a near future, we estimate that it should be possible to factor 35 in about 1 min using the exact algorithm presented here (with windowed arithmetic and 3D color codes) and a setup combining a memory for storing 38 logical qubits (3002 spatial modes and 5 temporal modes) and a pro-

cessor with 316 physical qubits (we estimate that more than 60 000 qubits would be needed with a standard 2D grid and surface code). If instead of using a spatially and temporally multiplexed memory, the qubits are stored in the same spatial mode and are identified by (6 650) temporal addresses only, we evaluate the same factorization to be possible in about 1 day using a memory bandwidth $4\Gamma = 2\pi \times 48$ MHz and taking into account the corresponding dead time between two memory readouts. In this case, error correction of all the qubits stored in the memory is estimated to take 132 ms meaning the storage time needs to be longer than 132 ms. For a memory bandwidth $4\Gamma = 2\pi \times 120$ MHz, the same factorization would take 9 hours, and error correction is estimated to take 53 ms. As discussed in [Appendix H](#), these requirements can realistically be met with a realization of the memory protocol described before combining a solid doped with rare-earth and a superconducting microwave resonator [[31–33](#)].

Conclusion — We have shown that the use of a quantum memory for quantum computing is appealing as unprocessed qubits can be loaded into the memory which significantly reduces the size of the processor compared with standard architectures where all qubits are kept in the processor. All-to-all connectivity between logical qubits is reached if each address in the memory can be mapped to only 3 qubits in the processor. The use of a memory allows one to exploit a 3D code on a 2D processor. If we allow each memory mode to be mapped to any qubit in the processor, all-to-all connectivity between physical qubits can be obtained, hence offering many opportunities for error correction and for implementing algorithms with gates operating between non-neighboring qubits.

We acknowledge M. Afzelius, J.-D. Bancal, P. Bertet, E. Flurin, P. Sekatski, X. Valcarce and J. Zivy for stimulating discussions and/or for critically reviewing the manuscript. We acknowledge funding by the Institut de Physique Théorique (IPhT), Commissariat à l'Énergie Atomique et aux Energies Alternatives (CEA) and the Region Île-de-France in the framework of DIM SIRTEQ.

Appendix A: Semi-classical Fourier transform

We here discuss the semi-classical Fourier transform presented in [[34](#)] and show that its cost is negligible. The standard way to perform the Fourier transform on n_e qubits is shown in [Figure 3a](#): it requires a sequence of one Hadamard and controlled phase gates for each qubit. In Shor's algorithm as well as in Ekerå and Håstad's version of Shor's algorithm, the qubits are measured right

after the Fourier transform, hence explaining the measurements of each qubit at the end of gate sequences in [Figure 3a](#). The simple rearrangement presented in [Figure 3b](#) shows that the measurement can be performed right after the Hadamard provided that the following phase gates are classically controlled by the result of this measurement, see [Figure 3c](#). In this case, the successive classically controlled phase gates operating on the same qubit can be merged together, leading to a circuit with one phase gate, one Hadamard gate and one measurement per qubit. When this semi-classical Fourier transform operates on a register made with n_e qubits (the number of bits of the exponent), its cost is linear in n_e and is thus negligible compared to the cubic complexity of the exponentiation.

Appendix B: Decomposition of the exponentiation into elementary gates

In this appendix, we aim to give a clear view of how to decompose the modular exponentiation into elementary gates. The presented method is intended to be simple to understand, but not optimal. A more efficient one is presented in [Appendix D](#).

1. Decomposition of a modular exponentiation into additions

The modular exponentiation needed in Ekerå and Håstad's algorithm, *i.e.*, the operation $|e\rangle|1\rangle \mapsto |e\rangle|g^e \bmod N\rangle$, with the input e and the output $g^e \bmod N$ encoded on n_e and n bits respectively, can be implemented from controlled modular additions as we show now. For simplicity, we omit the modulo in this paragraph.

Let $e_{n_e-1} \dots e_i \dots e_0$ be the binary form of e . The exponentiation can first be seen as a sequence of multiplications

$$g^e = \prod_{i=1}^{n_e-1} g^{2^i e_i} = \prod_{i=1}^{n_e-1} \left[g^{2^i} \right]^{e_i} \quad (\text{B1})$$

where each multiplication is controlled by the bit value e_i . [Figure 4](#) shows an implementation of such a multiplication in which a quantum register encoding the integer x ends up into an encoding of $x \times g^{2^i e_i}$. It uses two controlled product-additions, *i.e.*, the operation letting $|y\rangle|z\rangle$ unchanged if $|e_i\rangle = |0\rangle$ and mapping $|y\rangle|z\rangle$ into $|y\rangle|z + y \times \gamma\rangle$ ($(y, z, \gamma) \rightarrow (x, 0, g^{2^i})$) for

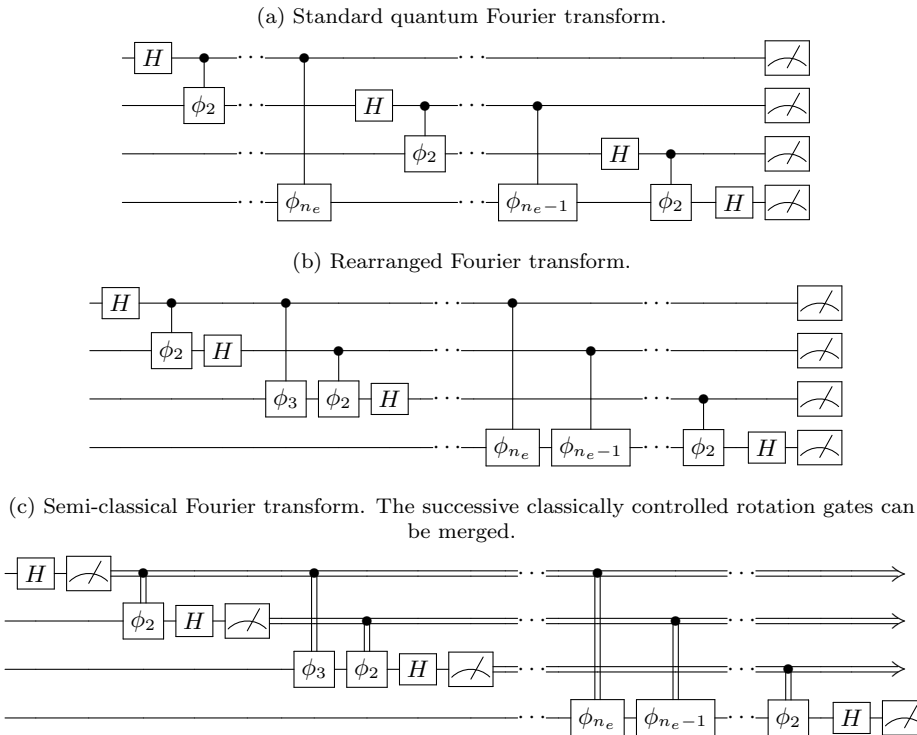


Figure 3. Different versions of the Fourier transform followed by measurements. They are used to convince the reader that the number of gates in the Fourier transform is negligible with respect to the cost of the exponentiation. These three versions are based on the phase gates ϕ_k defined as a 2×2 matrix with diagonal elements $(1, e^{\frac{2\pi i}{2^k}})$ and zeros off diagonal. Note that the control and target qubits can be reversed in the representation of each controlled phase gate without changing the result.

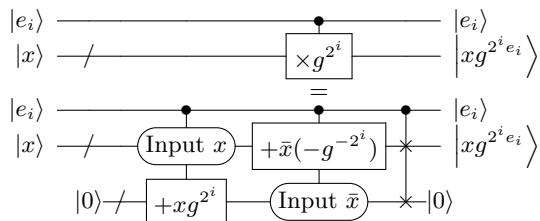


Figure 4. Principle of a modular multiplication circuit transforming a quantum register encoding the integer x into a state encoding $x \times g^{2^i e_i}$. A first product-addition operation transforms auxiliary qubits in $|0\rangle$ into $|x \times g^{2^i}\rangle$ if $|e_i\rangle = |1\rangle$. Then, a product-addition applies $+\bar{x}(-g^{-2^i})$ with $\bar{x} = x \times g^{2^i}$ into the register encoding x if $|e_i\rangle = |1\rangle$. A final swapping is applied if $|e_i\rangle = |1\rangle$ to put the quantum register into $|x \times g^{2^i e_i}\rangle$ and resets the auxiliary qubits to $|0\rangle$. Note that all the operations are performed modulo N .

the first product-addition appearing in [Figure 4](#) and $(y, z, \gamma) \rightarrow (\bar{x}, x, -g^{-2^i})$ for the second one, where the negative power stands for multiplicative inverse modulo N when $|e_i\rangle = |1\rangle$. In case $|e_i\rangle = |1\rangle$, the mapping is performed by considering the binary representation

$y_{n-1} \dots y_0$ of y and by rewriting the product as

$$y \times \gamma = \sum_{j=0}^{n-1} \gamma 2^j y_j = \sum_{j=0}^{n-1} [\gamma 2^j] y_j. \quad (\text{B2})$$

As y_j is either 0 or 1, the controlled product-addition can be implemented by a sequence of additions, each of them controlled both by the values of bits $|y_j\rangle$ and $|e_i\rangle$. [Figure 5](#) shows explicitly the decomposition of the first product-addition appearing into each multiplication of the exponentiation.

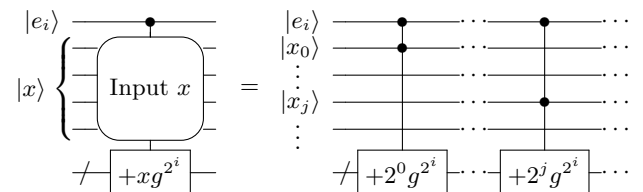


Figure 5. Decomposition of the first product-addition appearing in each element of the decomposition of the exponentiation into multiplications, see [Figure 4](#).

We deduce that the modular exponentiation requires n_e multiplications, each being decomposed into $2n$ controlled additions and 1 controlled swap between two registers, giving to a total number of $2n_e n$ (n_e) controlled

additions (swaps between registers respectively). Each addition needs to be modular, which can be obtained with a specific representation and a standard adder circuit.

2. Coset representation

The basic idea of the coset representation for adding $2^j\gamma$ to a quantum register encoding the integer z is to extend the register for z with m additional qubits and to encode it into the state $\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |z + kN\rangle$. Except at the bounds, this state is invariant under the addition of N . This implies:

$$\begin{aligned} & \frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |(z + 2^j\gamma + kN) \bmod 2^{n+m}\rangle \\ & \approx \frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |(z + 2^j\gamma \bmod N) + kN\rangle \end{aligned}$$

i.e. the modular addition of $2^j\gamma$ in the register of z can be performed with a standard adder, at the cost of a small error which is exponentially suppressed when increasing m [35]. Note that the resource needed to initialize the register is negligible with respect to the resource taken to implement the adder, see Appendix C. Taking into account the increase in register size, this means that $2n_e(n+m)$ controlled additions and n_e controlled register swaps are needed for realizing Ekerå and Håstad's algorithm.

3. Controlled operations

A controlled swap operation between two qubits (Fredkin gate) can be performed using two CNOTs and one Toffoli gates, see Figure 6b. Hence the total cost for controlled swaps operating on two registers prepared in the coset representation of integers (encoded each with $n+m$ qubits) is of $2(n+m)$ CNOTs and $n+m$ Toffoli gates.

For the controlled addition, note first that since we use the coset representation of integers, a circuit for controlled addition modulo a power of two is sufficient to implement a controlled modular addition. Such an addition can be implemented with the semi-classical adder presented in Figure 6a, which is inspired by Refs. [4, 36, 37]. It shows the basic circuit taking a classical value $2^j\gamma$ and a register encoding z' and returning $2^j\gamma$ and $z' + 2^j\gamma$ if the

two controlled qubits $|e_i\rangle$ and $|x_j\rangle$ are both in state $|1\rangle$. When such an addition is applied on a quantum register encoding z' using $n+m$ qubits, the block in the dashed box of Figure 6a is repeated $n+m-2$ times, giving a mean cost of $5.5(n+m)-9$ CNOTs and $2(n+m)-1$ Toffoli gates.

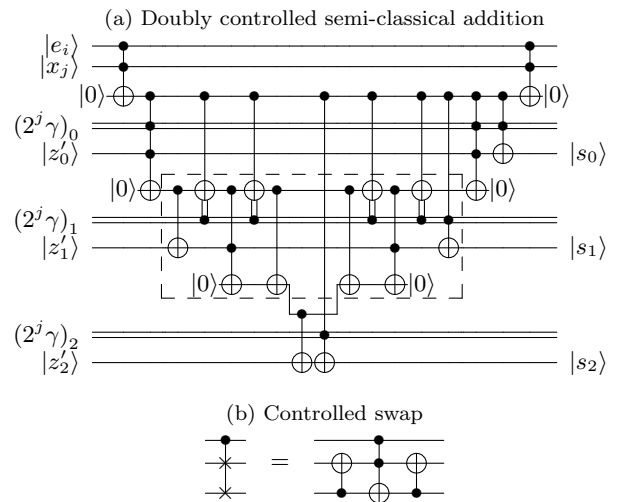


Figure 6. Controlled operations. (a): semi-classical adder taking each bit of the classical value $2^j\gamma = \sum_{k=0}^{2^m-1} 2^k (2^j\gamma)_k$ and the three qubits register encoding z' as inputs and returning $(2^j\gamma)_k$ and $|s_k\rangle = |(z' + 2^j\gamma \cdot e_i \cdot x_j)_k\rangle$. The block in the dashed box uses in average 5.5 CNOTs and 2 Toffoli gates. (b): Fredkin gate implemented with a Toffoli and two CNOT gates. The controlled swap between registers (as required in Figure 4) is obtained by applying it to each pair of qubits.

4. Number of gates

Given the number of gates in the controlled addition and swap operations, the number of additions and swaps in the multiplication and the number of multiplications in the modular exponentiation, we estimate that factorization takes at leading order $11n_e(n+m)^2$ CNOTs and $4n_e(n+m)^2$ Toffoli gates.

Appendix C: Coset representation

Modular addition is typically implemented with variants of the addition: an addition, a comparison, a controlled correction and the clean-up of ancillary qubits [38]. As exposed in main text, coset representation of integers, introduced by Zalka [39] and formalized by Gidney [35], can be used to approximate the modular addition with a single standard adder circuit.

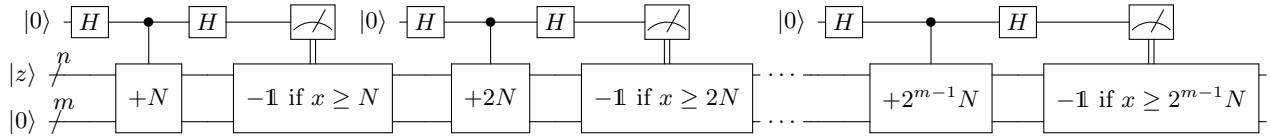


Figure 7. Preparation proposed in [35, Fig. 1] of a quantum register with $n + m$ qubits in the state $\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |z + kN\rangle$ as requested in the initialization of the coset representation. The first controlled operation adds the integer N to the register made with $n + m$ qubits provided that the ancillary qubit is in state $|1\rangle$. The first classically controlled operation aims to change the phase of the input state encoded in $n + m$ qubits if and only if the result of the measurement is 1 and the number encoded in the $n + m$ qubits is larger or equal than N . In case one of the two conditions is not met, the input state is unchanged.

The basic idea of the coset representation for adding $2^j\gamma$ modulo N to a quantum register encoding the integer z is to extend the register for z with m additional qubits and to encode it into the state $\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |z + kN\rangle$. Except at the bounds, this state is invariant under the addition of N . This implies

$$\begin{aligned} & \frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |(z + 2^j\gamma + kN) \bmod 2^{n+m}\rangle \\ & \approx \frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |(z + 2^j\gamma \bmod N) + kN\rangle \quad (\text{C1}) \end{aligned}$$

i.e. the modular addition of $2^j\gamma$ in the register of z can be performed with a standard adder (modulo 2^{n+m}), at the cost of a small error which is exponentially suppressed when increasing m [35]. Note also that the precision is improved if instead of adding $2^j\gamma$, one adds $2^j\gamma \bmod N$ (which does not change the result of the sum since we consider the sum modulo N). This is possible each time the quantity to add is known classically.

The goal of the first subsection is to show that the resource needed to extend the register encoding $|z\rangle$ into the state $\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |z + kN\rangle$, as requested in this representation, is negligible with respect to the resource taken to implement the modular exponentiation. In the second subsection, we show that the coset representation is compatible with the modular multiplication circuit presented in the main text.

In the two next subsections, the coset representation is considered for additions modulo N ; n is the number of bits encoding N , and m the number of qubits added to the register for the coset representation.

1. Initialization

Starting from a register with n qubits in state $|z\rangle$, the initialization of the coset representation consists in

preparing the state $\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |z + kN\rangle$ in an extended register of size $n + m$. This is done by performing successive additions, each controlled by an ancillary qubit prepared in the state $\frac{|0\rangle+|1\rangle}{\sqrt{2}}$ (m ancillary qubits in total) which is then uncomputed, see Figure 7. The controlled addition is performed using the circuit presented in Figure 6a with only one control qubit. The uncomputation of the ancillary qubit is based on a measurement and depending on the result, a conditioned correction is realized, see Figure 7. Let us detail the uncomputation of the first ancillary qubit presented in Figure 7. When the result of the measurement is 0, the register made with $n + m$ qubits is projected into $\frac{1}{\sqrt{2}}(|z\rangle + |z + N\rangle)$. When the result is 1, the register state is $\frac{1}{\sqrt{2}}(|z\rangle - |z + N\rangle)$ and the operation -1 needs to be applied to the component $|z + N\rangle$, *i.e.*, when the state of the register encodes an integer larger than N . In order to implement the conditioned operations for decomputing the m ancillary qubits, we need to compare the value x encoded in the quantum register of size $n + m$ and an integer y known classically satisfying $0 < y \leq 2^{m-1}N < 2^{n+m}$ (see the last uncomputation in Figure 7) *i.e.* that can be written with $n + m$ bits. This comparison is implemented using the circuit presented in Figure 8a. First, the value $2^{n+m} - y = y'$ is computed classically. Then the last carry of the sum of x and y' is computed with a circuit derived from the addition. If the value of this carry is 1, we conclude that $x \geq y$, otherwise $x < y$. A Z gate is thus applied on the qubit encoding the last carry, before uncomputing the carries. The register ends up in state $\pm|x\rangle$ depending on the relative value between x and y , as desired.

Each controlled addition and correction costs $O(n + m)$ gates. This operation is repeated m times, giving a total cost of the coset representation initialization of the order $O(m(n + m))$.

In the modular exponentiation algorithm, the two registers at the bottom of Figure 4 need to be prepared initially in $x = 1$ and 0 respectively. Initializing them in the coset representation $\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |1 + kN\rangle$ and

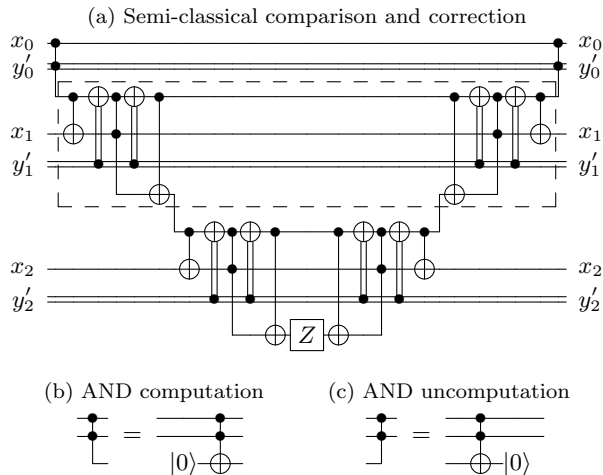


Figure 8. (a): Circuit inspired from [4, Fig. 17] which compares the integer x encoded in $n + m$ qubits and the integer $y < 2^{n+m}$ known classically, and returns $-|x\rangle$ if and only if $x \geq y$. This is done in three steps: i) compute the carries of $y' + x$ with $y' = 2^{n+m} - y$, ii) apply a Z operation on the last carry and iii) uncompute the carries.

(b) and (c): circuits defining the notations used to compute and uncompute an AND operation, as introduced in [37, 40] where the authors give efficient implementations in terms of T (or $\frac{\pi}{4}$) gates. When only one quantum control appear, it uses a CNOT instead of a Toffoli gate, and it can be removed by directly using the control bit instead of the ancillary.

$\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |kN\rangle$ takes $O(m(n+m))$ gates which is negligible compared to the cubic cost of the full exponentiation. Note however, that the cost of this initialization is taken into account in our script for the evaluation of the whole algorithm cost.

2. Compatibility with the multiplication

When computing the multiplications from sequences of two product-additions (see Figure 4 of main text), the input register encoding x and the ancillary register are used both as control and target of the product-additions. We here check that having the control register encoded in the coset representation is not a problem for performing the multiplication.

Let us consider the first product-addition used to implement the multiplication shown in the bottom part of Figure 4. In the coset representation, the input x and ancillary registers are in the state $\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |x+kN\rangle \sum_{k'=0}^{2^m-1} |0+k'N\rangle$ meaning that after the product-addition, their state ends up in $\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} \sum_{k'=0}^{2^m-1} |x+kN\rangle |(x+kN)g^{2^i} + k'N \bmod 2^{n+m}\rangle$.

As $kNg^{2^i} + k'N$ is a multiple of N , the obtained state is very close to $\frac{1}{\sqrt{2^m}} \sum_{k=0}^{2^m-1} |x+kN\rangle \sum_{k'=0}^{2^m-1} |xg^{2^i} + k'N\rangle$ thanks to the coset representation itself, cf. Eq. (C1). The latter corresponds to the desired state.

Appendix D: Windowed arithmetic

In order to reduce the number of multiplications and additions in the exponentiation algorithm, we use windowed arithmetic circuits [21]. They consist in grouping the bits of e for controlling each multiplication, hence reducing the number of multiplications. Similarly, for each multiplication, the input bits are grouped to reduce the number of additions composing each multiplication.

The next subsection shows the details of the decomposition of the exponentiation into elementary additions of the form $+T_k \bmod N$ where the quantity T_k depends on the value of an integer k . These specific additions are implemented in three steps, that are presented in separated subsequent subsections.

1. Windowed exponentiation and multiplication

Let us start by specifying the notations. We label the binary form of e as

$$e_{n_e-1} \dots e_{i+w_e} \overbrace{e_{i+w_e-1} \dots e_i}^{e_{i:i+w_e}} \dots e_2 e_1 e_0 \quad (D1)$$

i.e. e_j is the j th bit of e . Let also $e_{i:i+w_e}$ be defined as

$$e_{i:i+w_e} = \sum_{j=i}^{i+w_e-1} 2^{j-i} e_j \quad (D2)$$

i.e. $e_{i:i+w_e}$ is the number whose bit decomposition is given by the bits of e starting at index i and taking w_e bits. The strategy for computing the exponentiation using windowed arithmetic consists in decomposing exponent e in terms of numbers $e_{i:i+w_e}$

$$e = \sum_{\substack{0 \leq i < n_e \\ i \equiv 0 \pmod{w_e}}} 2^i e_{i:i+w_e}, \quad (D3)$$

such that

$$g^e = \prod_{\substack{0 \leq i < n_e \\ i \equiv 0 \pmod{w_e}}} g^{2^i e_{i:i+w_e}}. \quad (D4)$$

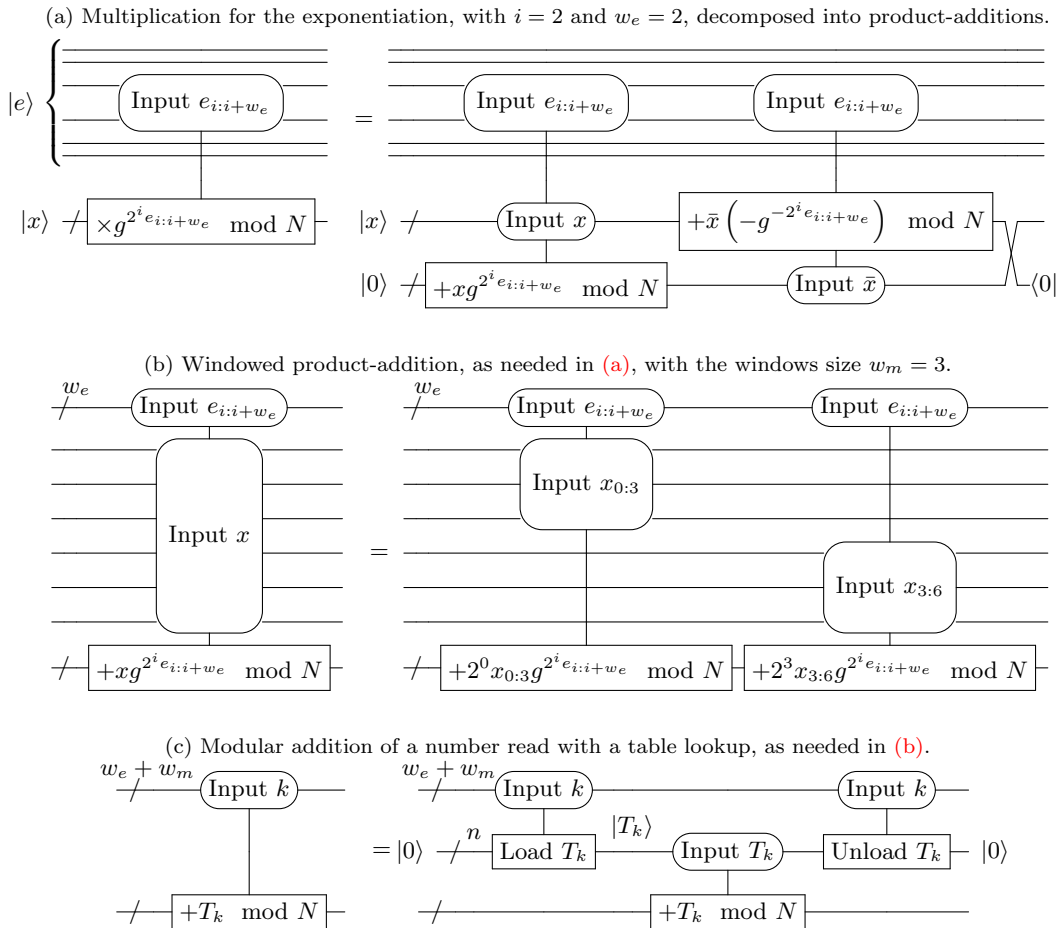


Figure 9. Windowed arithmetic subcircuits for the modular exponentiation. When not specified, the register size is $n+m$ qubits (register encoded into the coset representation of integers).

The comparison with the decomposition of g^e presented in Eq. (B1) clearly shows that windowed exponentiation divides the number of multiplications by w_e .

As for the standard algorithm, the multiplications of the product (D4) are implemented successively and each multiplication is decomposed into a sequence of two product-additions, as shown in Figure 9a. The difference is that the added number now depends on the number $e_{i:i+w_e}$.

The product-addition is also performed in a windowed way [21]. Figure 9b shows in particular how the first product-addition needed for each multiplication is performed using windows for input x of size $w_m = 3$.

Figure 9c finally shows the implementation of an addition $+T_k \bmod N$ of a quantity T_k that depends on the value k . It requires three steps. First, the number T_k is loaded into an ancillary register. Second, this number is unconditionally added to the desired register and finally the ancillary register is cleaned up. Note that the value of T_k (given by $T_{k_1, k_2} = 2^i k_1 g^{2^i k_2}$, with $k_1 = e_{i:i+w_e}$ and

$k_2 = x_{i:i+w_m}$, k being the concatenation of k_1 and k_2) to be added is known classically. Its addition being realized modulo N , its value can be computed modulo N before being loaded. n bits are thus sufficient to encode T_k .

Loading a value T_k into a quantum register is done using a quantum table lookup circuit which we discuss right after. The subsequent subsection is dedicated to the task aiming to unload the value T_k and reset the register in state $|0\rangle$. The last subsection is dedicated to the requested addition.

2. Table lookup

The quantum table lookup proposed in [40], produces the following operation on basis states: $|k\rangle |x\rangle \mapsto |k\rangle |x \oplus T_k\rangle$ with \oplus the bitwise XOR operator. For state preparation, as required in the first step of the operation presented in Figure 9c, the target register starts in the state $|0\rangle$ such that control and target registers end

in $|k\rangle|T_k\rangle$. The circuit presented in **Figure 10** shows the principle of this operation with registers for k and T_k composed respectively of 3 and 5 qubits.

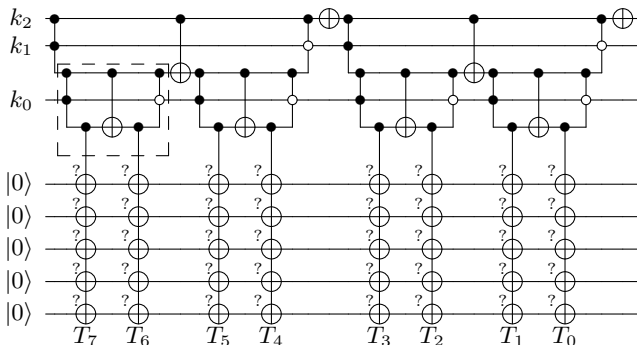


Figure 10. Example of a quantum table lookup. For a basis state $|k\rangle$ specifying the address of the number T_k from a classical table, the quantum table lookup maps basis states $|k\rangle|0\rangle$ into $|k\rangle|T_k\rangle$. Here k and the output are composed respectively of 3 and 5 qubits. The notations for the AND computation and uncomputation is presented in **Figure 8**. Black and white circles are controls on the $|1\rangle$ and $|0\rangle$ states respectively. The question mark on the controlled NOT means that a controlled NOT is applied on qubit i only when the i th bit of T_k takes the value 1.

Concretely, the numbers T_k specify the set of controlled NOT to be used (the question mark on the controlled NOT means that a controlled NOT is applied on qubit i only when the i th bit of T_k takes the value 1). The circuit operating on the bits k_i of k prepares the last ancillary qubit (line 5 from the top) in the state $|1\rangle$ at the time (specified by k) where the gates corresponding to T_k are applied, and $|0\rangle$ otherwise. The building block of the circuit is boxed in **Figure 10**. It uses 1 CNOT, 1 AND computation and uncomputation. Given that k is encoded into the number of bits $w_e + w_n$ and can thus take $2^{w_e + w_n}$ different values, the number of blocks in the upper part of **Figure 10** is given by $\sum_{j=1}^{w_e + w_n - 1} 2^j = 2^{w_e + w_n} - 2$. This means that $2^{w_e + w_n} - 2$ CNOT gates, $2^{w_e + w_n} - 2$ AND computations and uncomputations are needed to implement these blocks. Moreover, the number of controlled multi-NOT gates to load the value T_k is given by $2^{w_e + w_n}$, each gate being decomposed into $n/2$ CNOT in average since T_k takes n bits. When including the (two) NOT gates operating on the highest bit of k , we conclude that the table lookup uses 2 NOT gates, $2^{w_e + w_n} - 2 + 2^{w_e + w_n - 1}n$ CNOT gates, $2^{w_e + w_n} - 2$ AND computations and uncomputations (corresponding to $2 \times (2^{w_e + w_n} - 2)$ Toffoli gates).

3. Table unlookup

The purpose of the table unlookup operation (last step in **Figure 9c**) is to map the state $\sum_k \alpha_k |k\rangle|T_k\rangle$ into $\sum_k \alpha_k |k\rangle$, where α_k are some complex coefficients. A natural way to do this mapping is to apply again the lookup operation described in the previous subsection. Since the lookup operates on the computational basis following $|k\rangle|x\rangle \mapsto |k\rangle|x \oplus T_k\rangle$ where \oplus stands for the bitwise XOR operator, by linearity it maps $\sum_k \alpha_k |k\rangle|T_k\rangle \mapsto \sum_k \alpha_k |k\rangle|0\rangle$, the latter corresponding to the desired state when simply discarding the qubits previously encoding the numbers T_k .

However, a more efficient measurement-based technique is possible, as shown in Ref. [41, Appendix C] and improved in Ref. [21]. The principle consists in starting by measuring the register encoding T_k in the X basis before applying a phase shift conditioned on the result of measurements. For a more detailed explanation, let us start to expand the qubits encoding the numbers T_k in bits indexed by j ($(T_k)_j$ being the j th bit of T_k). The state before the uncomputation can be written as

$$\sum_k \alpha_k |k\rangle \otimes \left| (T_k)_j \right\rangle_j. \quad (\text{D5})$$

Let us now focus on a specific qubit indexed by j^* . We label $\mathcal{K}_0 = \{k \mid (T_k)_{j^*} = 0\}$ and $\mathcal{K}_1 = \{k \mid (T_k)_{j^*} = 1\}$. The state before the uncomputation can be rewritten as

$$\left[\sum_{k \in \mathcal{K}_0} \alpha_k |k\rangle \otimes \left| (T_k)_j \right\rangle_j \right] |0\rangle_{j^*} + \left[\sum_{k \in \mathcal{K}_1} \alpha_k |k\rangle \otimes \left| (T_k)_j \right\rangle_j \right] |1\rangle_{j^*}. \quad (\text{D6})$$

By applying a Hadamard gate on the j^* th qubit, we obtain

$$\frac{1}{\sqrt{2}} \left[\sum_{k \in \mathcal{K}_0} \alpha_k |k\rangle \otimes \left| (T_k)_j \right\rangle_j + \sum_{k \in \mathcal{K}_1} \alpha_k |k\rangle \otimes \left| (T_k)_j \right\rangle_j \right] |0\rangle_{j^*} + \frac{1}{\sqrt{2}} \left[\sum_{k \in \mathcal{K}_0} \alpha_k |k\rangle \otimes \left| (T_k)_j \right\rangle_j - \sum_{k \in \mathcal{K}_1} \alpha_k |k\rangle \otimes \left| (T_k)_j \right\rangle_j \right] |1\rangle_{j^*}. \quad (\text{D7})$$

Hence, if the measurement of the j^* th qubit yields 0, the qubit is properly uncomputed. If the result is 1, a phase

shift needs to be applied on states corresponding to the indexes $k \in \mathcal{K}_1$.

This uncomputation is successively applied to all the qubits encoding the numbers T_k . Let t_j be the measurement result of the j th qubit. The state after all the measurements is given by

$$\sum_k \alpha_k \sigma_k |k\rangle, \quad (\text{D8})$$

with $\sigma_k = \prod_j (-1)^{t_j (T_k)_j}$. We now label

$$\mathcal{K} = \{k \mid \sigma_k = -1\}. \quad (\text{D9})$$

In order to recover the desired state, we need to correct selectively the phase of terms $|k\rangle$ for which $k \in \mathcal{K}$.

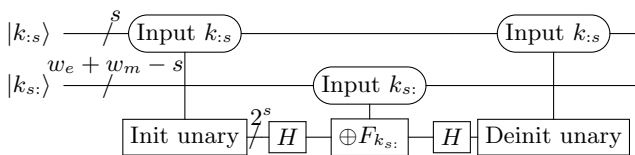


Figure 11. Representation of the four steps proposed in Ref. [21] to selectively change the phase of components $|k\rangle$ in the state given in (D8) when the index k belongs to \mathcal{K} (D9). The central operation is a table lookup with the values $F_{k_s} = \sum_{j=0}^{2^s-1} 2^j \delta(j + 2^s k_s)$ where $\delta(\cdot)$ is the indicator function of \mathcal{K} .

The selective phase correction is done in four steps [21], as shown in Figure 11. First, the control register which uses $w_e + w_m$ qubits in state $|k\rangle$ is split in two groups. The first group is made with s qubits in state $|k:s\rangle$. The second group takes the remaining $w_e + w_m - s$ qubits in state $|k_s:\rangle$, such that $|k\rangle = |k_s:\rangle \otimes |k:s\rangle$ and $k = k_s + 2^s k_{s:}$. The second step consists in writing the integer k_s in an ancillary register in the unary representation: a register with 2^s qubits representing a number k_s with the state of the qubit number k_s being $|1\rangle$ and all the other qubits in the state $|0\rangle$. The qubits in state $|k_s:\rangle$ and the ancillary qubits are then used as control and target qubits for a lookup circuit where the controlled multi-NOT gates are replaced by controlled multi-Z gates. Finally, the ancillary register is uncomputed. The circuit used to initialize the ancillary register is shown in Figure 12a. The one used to put it back in its initial state is given in Figure 12b.

Starting from x encoded in s qubits, the conversion to the unary representation takes 1 NOT gate, $2^s - 1$ CNOT gates and $2^s - 1$ AND computation. The conversion back to the binary representation takes 1 NOT gate, $2^s - 1$ CNOT gates and $2^s - 1$ AND uncomputation. Given that k is encoded in $w_e + w_m$ bits, and that a choice $s = \lfloor \frac{w_e + w_m}{2} \rfloor$ is judicious to minimize the number of gates, the change of phase of components $|k\rangle$ takes

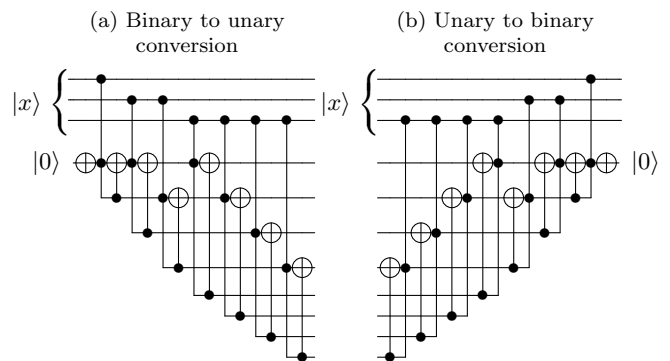


Figure 12. (a): representation of the circuit proposed in Ref. [21] for preparing a copy in an ancillary register of an integer x in a unary representation starting from an encoding of x in a control register in the binary representation. The first not operation prepares the first qubit in the ancillary register in state $|1\rangle$. The first AND computation writes the result of an AND operation between the first bit of x and the bit 1 encoded in the first qubit of the ancillary register into the second qubit of the ancillary register. In case the state of the latter is $|1\rangle$, the state of the first qubit of the ancillary register is changed to $|0\rangle$. The combination of AND and CNOT operations is successively repeated until the desired qubit of the ancillary register is in state $|1\rangle$. (b): representation of the circuit proposed in Ref. [21] to erase the value in the ancillary register while keeping the integer x into the control register. The circuits (a) and (b) corresponds to the first and third operations needed for the selective phase correction operation presented in Figure 11.

$2^{\lfloor \frac{w_e + w_m}{2} \rfloor + 1} + 4$ 1-qubit gates, $2^{w_e + w_m - 1} + 2^{\lfloor \frac{w_e + w_m}{2} \rfloor + 1} + 2^{\lceil \frac{w_e + w_m}{2} \rceil} - 4$ CNOTs and $2^{\lfloor \frac{w_e + w_m}{2} \rfloor} + 2^{\lceil \frac{w_e + w_m}{2} \rceil} - 3$ ANDs (1 NOT gate, $2^{\lfloor \frac{w_e + w_m}{2} \rfloor} - 1$ CNOT gates and $2^{\lfloor \frac{w_e + w_m}{2} \rfloor} - 1$ AND computation for the unary conversion, $2 \times 2^{\lfloor \frac{w_e + w_m}{2} \rfloor}$ Hadamard gates around the table lookup, 2 NOT gates, $2^{\lceil \frac{w_e + w_m}{2} \rceil} - 2 + 2^{w_e + w_m - 1}$ CNOT gates, $2^{\lceil \frac{w_e + w_m}{2} \rceil} - 2$ AND computations and uncomputations for the lookup circuit and 1 NOT gate, $2^{\lfloor \frac{w_e + w_m}{2} \rfloor} - 1$ CNOT gates and $2^{\lfloor \frac{w_e + w_m}{2} \rfloor} - 1$ AND uncomputation for the binary conversion). Including the additional n Hadamard gates and n measurements on T_k , we conclude that the table unlookup takes $2^{\lfloor \frac{w_e + w_m}{2} \rfloor + 1} + n + 4$ 1-qubit gates, $2^{w_e + w_m - 1} + 2^{\lfloor \frac{w_e + w_m}{2} \rfloor + 1} + 2^{\lceil \frac{w_e + w_m}{2} \rceil} - 4$ CNOTs and $2^{\lfloor \frac{w_e + w_m}{2} \rfloor} + 2^{\lceil \frac{w_e + w_m}{2} \rceil} - 3$ ANDs.

4. Standard adder

As we use the coset representation of integers with windowed arithmetic operations, a circuit for unconditional addition modulo a power of two is sufficient to implement a modular addition. The adder we use, which is

described in [37] and optimized from [36] for use with T gates, is presented in Figure 13. It is thrifty in gate number and ancillary qubits, at the cost of being deeper than other circuits [42, 43], which is not a disadvantage for our architecture.

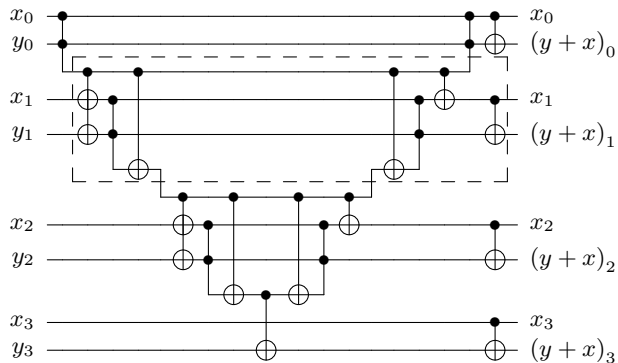


Figure 13. Adder modulo 2^4 from [37], using the same notations as in Figure 8. The building block (boxed) is repeated two times, for the qubits numbers 1 and 2, while the first and last use a simplified subcircuit.

As presented in Figure 9c, the adder needs to add a number T_k taking n qubits into a register with $n + m$ qubits. To achieve this, either the first register for T_k is extended with qubits in the $|0\rangle$ state, either we use carry propagation blocs for the last qubits. Such blocs are identical to the ones of semi-classical adder with classical input 0; see [4, Fig. 17] for an example of such a circuit. For gate counting, the first solution is taken into account.

The cost of the addition circuit (Figure 13) is $6(n + m) - 9$ CNOT gates and $n + m - 1$ AND computations and uncomputations.

5. Cost estimation

In summary, the parameters of the logical circuit for computing the modular exponentiation are

n number of bits of the exponentiated number g

n_e number of bits of the exponent e

w_e window size for the exponentiation

w_m window size for the multiplication

m number of qubits added by the coset representation

The aim of this subsection is to give an estimate of the number of gates needed to implement this circuit. In

order to keep the evaluation independent of the error correction choice, we express the cost in terms of the number of 1-qubit, 2-qubit gates and AND computation and uncomputation [44].

The modular exponentiation consists in n_e/w_e multiplications, each multiplication using 2 product addition and a swap and each product addition is implemented with $(n+m)/w_m$ lookups, additions and unlookups. Note that the swap operation is realized by simply relabeling the register, hence is for free. According to the counts obtained from previous subsections, the cost of the exponentiation is dominated — in the limit $n \rightarrow \infty$, $n_e = O(n)$, w_e and w_m constant — by: $2^{\frac{n_e(n+m)n}{w_e w_m}}$ 1-qubit gates, $(2^{w_e+w_m}n + 12(n+m)) \frac{n_e(n+m)}{w_e w_m}$ CNOTs, and $2^{\frac{n_e(n+m)^2}{w_e w_m}}$ AND computations and uncomputations (translatable into $4^{\frac{n_e(n+m)^2}{w_e w_m}}$ Toffoli gates). Note that when considering the universal gate set T, S, H, X, Y, Z , CNOT, controlled-Z and their conjugate, according to Fig. 4 of Ref. [40] the AND computation and uncomputation costs in average 8 1-qubit gates and 3.5 2-qubit gates. The total cost of the exponentiation is hence given at the leading order by $2^{\frac{n_e(n+m)n}{w_e w_m}} (9n + 8m)$ 1-qubit gates and $(2^{w_e+w_m}n + 19(n+m)) \frac{n_e(n+m)}{w_e w_m}$ 2-qubit gates. In the code used to compute the required resources and find the optimal parameters, the complete formula have been used [22].

Appendix E: Error correction

This appendix is dedicated to 3D gauge color codes. The first subsection is dedicated to the principle of subsystem codes. The second subsection describes the geometrical structure of 3D gauge color codes. The last subsection provides a detailed description of the cut of the code structure that is used to process and correct the logical qubits.

1. Subsystem codes

Subsystem stabilizer codes [45] are defined by three subgroups of the Pauli group: the stabilizer, gauge and logical (also designated as *bare logical* in [23]) operator groups, such that the stabilizer group is the center of the gauge group up to phases, $i\mathbb{1}$ is included in the gauge group, the operators from the gauge and logical groups commutes, and the normalizer of the stabilizer group is the product of gauge and logical groups. We invite the reader to look at Refs. [23, 45] for an explicit construction

of those groups from canonical generators of the Pauli group. The stabilizer group plays the standard role of stabilizers, *i.e.* divides the total Hilbert space \mathcal{H} into a direct sum of orthogonal subspaces $C \oplus C^\perp$ where C — the stabilized subspace — corresponds to the eigenspace $+1$ of all stabilizers. The gauge and logical groups decompose the stabilized subspace C into a tensor product of the logical qubits space A and the gauge qubits space B [46], that is, the Hilbert space is decomposed as

$$\mathcal{H} = \underbrace{(A \otimes B)}_C \oplus C^\perp.$$

The gauge group acts trivially on the logical qubits and is the Pauli group of the gauge qubits while the logical group acts trivially on the gauge qubits and is the Pauli group of the logical qubits (up to phases). This ensures that gauge operator measurements don't modify the logical qubits.

A gauge fixing operation consists in switching from a code to another one such that the new stabilizer group includes the original one while being included into the original gauge group, while keeping unmodified the logical group. The decomposition associated to the original code

$$\mathcal{H} = \underbrace{(A \otimes B)}_C \oplus C^\perp$$

then becomes of the form

$$\mathcal{H} = (A \otimes B') \oplus \underbrace{(A \otimes B'')}_{C^\perp} \oplus C^\perp$$

where B' is the new gauge qubit space. As a consequence, a valid code-word for this new code is also valid for the initial one. The passage of the latter to the new code is done by measuring the generators of the gauge group, the results of these measurements giving the correction to apply on $B' \oplus B''$ to remove the components on B'' .

For 3D gauge color codes, code switching allows a transversal error-corrected implementation of a universal set of gates [23].

2. Code geometrical structure

The geometrical structure of the 3D gauge color codes is described in detail in Section 3.1 of [23]. It takes a large tetrahedron, itself decomposed into elementary tetrahedrons, see Figure 14 for an example. Four extra points ($v_i, i \in \{1, 2, 3, 4\}$) are then added outside the large tetrahedron, one point in front of each facet of the large tetrahedron. Elementary tetrahedrons are finally added between those extra points and the vertices at the surface

of the large tetrahedron, see Fig. 4b of [47] for an illustration. The vertices of elementary tetrahedrons are colored with 4 different colors such that adjacent vertices get a different color. Each elementary tetrahedron represents a physical qubit.

The measured operators are the gauge generators for the code used to implement the H and CNOT gates — the (1,1) code (see [23]). These generators are described by the edges: each operator is the product of X or Z operators of the elementary tetrahedrons adjacent to a given edge (each operator implies up to 6 physical qubits).

The stabilizer generators of the (1,1) and (1,2) codes (the (1,2) code refers to the code used to implement the T gate [23]) which are described by the vertices and edges, are deduced from the values of measured operators. More precisely, the operator corresponding to a vertex can be written as the product of the operators corresponding to edges starting at the given vertex and ending on vertices of a common color. Three choices of color are possible, allowing one to recover in three different ways an operator corresponding to a vertex. This redundancy can be used for achieving fault-tolerant error correction in only one measurement of the (gauge) operators related to the edges [23].

Let n_{code} be the index of the code which is the number of vertices of the same color on one edge of the large tetrahedron (denoted as n in [23]). The code distance is given by $d = 2n_{\text{code}} + 1$, and the number of physical qubits is $1 + 4n_{\text{code}} + 6n_{\text{code}}^2 + 4n_{\text{code}}^3 = \frac{d^3+d}{2}$ [23, 47].

3. Slicing of the code structure

To process the information, the code structure is decomposed into slices, each slice being map successively into the 2D processor. While several cuts in slices are possible, we choose slices orthogonal to two faces (see Figure 14). The processor need to be sized to fit in the larger slice, that join the edge not included into any of the two faces to the middle of the opposing edge — the magenta slice in Figure 14.

With the lattice described in Ref. [23], the central slice corresponds to the elementary tetrahedrons for which all vertices coordinates satisfy $x + z = n_{\text{code}} - 2$ or $x + z = n_{\text{code}} - 1$ (the elementary tetrahedrons between the two plans defined by the previous equations). Note that the number of slices is given by $d - 2$.

The number of elementary tetrahedrons included in this slice is counted by considering three tetrahedron sets,

see Figure 14 and Figure 15. Two sets correspond to the elementary tetrahedrons having a facet at the interplay between two slices (Figure 15a (Figure 15b) is associated to the elementary tetrahedrons with one facet at the interplay between the magenta slice and the green (cyan) slice). The last set is associated to the elementary tetrahedrons having no facet at the interplay between two slices (Figure 15c). One can check that the two first sets

include $\sum_{k=1}^{2n_{\text{code}}-2} k = 2n_{\text{code}}^2 - 3n_{\text{code}} + 1$ elementary tetra-

hedrons while the last set has $(2n_{\text{code}} - 1) + \sum_{k=0}^{n_{\text{code}}-2} 2(2k +$

$1) = 2n_{\text{code}}^2 - 2n_{\text{code}} + 1$ elementary tetrahedrons. They are $16n_{\text{code}} - 2$ additional elementary tetrahedrons resulting from the 4 added points in the construction of the code. In total, the maximum number of elementary tetrahedron for one slice of the code structure is $6n_{\text{code}}^2 + 8n_{\text{code}} + 1$. Since we consider a processor that can process up to two slices (associated to two different logical qubits) and accounting for the ancillary subsystems needed to measure the gauge generators by a simple factor of two, we obtain the number of physical qubits in the processor specified in the main text. For more details, see the ancillary file `tetrahedron_3_bis.scad` [22], where each tetrahedron color corresponds to a given slice, the larger being the magenta one.

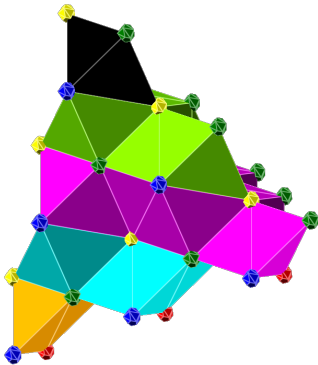


Figure 14. Code geometrical structure for $n_{\text{code}} = 3$ (without the extra points $(v_i, i \in \{1, 2, 3, 4\})$). Each slice has been represented with a specific color. The larger slice is with the magenta elementary tetrahedrons. The figure shows that the maximum number of slices involved in an operator corresponding to an edge is 2.

4. Threshold of 3D gauge color codes

The value of the threshold for 3D gauge color codes has been evaluated in a few references that we now discuss.

In order to clarify on the context, let us first remind that there are three main definitions of error-

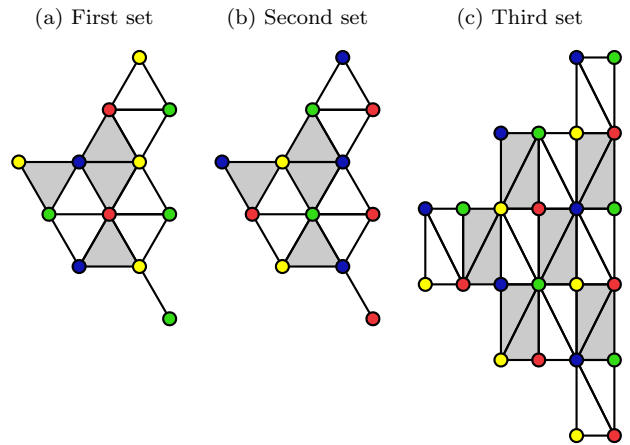


Figure 15. Decomposition of the central slice for $n_{\text{code}} = 3$ (magenta slice in the tetrahedron presented in Figure 14). Each subfigure corresponds to a set of elementary tetrahedrons of the central slice, seen from different point of views. On (a) and (b), each triangle corresponds to an elementary tetrahedron. On (c) each small rectangle correspond to an elementary tetrahedron.

rate threshold used for stabilizer codes in the literature: code-capacity, phenomenological and circuit-level. Code-capacity thresholds assume perfect measurements of stabilizers. Phenomenological thresholds model faulty-measurements as bit-flip errors on stabilizer measurement outcomes. Circuit-level thresholds model errors occurring at any stage of stabilizer measurement circuits.

In Ref. [26] a clustering decoding scheme is presented and by including a phenomenological noise to the measurement outputs, the authors estimate a code-capacity threshold of 0.46 % and phenomenological threshold to about 0.31 %, suggesting an upper bound on the circuit-level threshold. Note however that the underlying lattice considered in Ref. [26] (cubic lattice) is different from the one we have considered (body centered cubic lattice (bcc)).

A more recent decoding algorithm is presented in [48, 49] using the bcc lattice, but the authors give an estimate of the code capacity threshold of 0.77 % only. By the way, a slightly better code capacity threshold of 0.80 % has been estimated in Ref. [47] under the same assumptions.

Finally, statistical arguments have been used in Ref. [50] to estimate code-capacity threshold of 3D gauge color codes with ideal decoding to around 1.9 %. This suggests that an appropriate decoder could significantly improve the value of the code-capacity threshold and hence of the phenomenological and circuit-level thresholds.

Since we believe that the determination of the circuit-level threshold goes beyond the scope of this work, the

run-time and resource needed to factor a 2048-bit RSA integer are given in the main text under the assumption of a threshold of 0.75%. Since this choice is somehow arbitrary, we give the evolution of run-time and resource as a function of the threshold in [Figure 16](#). More precisely, they are given as a function of the ratio p/p_{th} between the physical error probability per cycle p and the fault-tolerant threshold p_{th} which is the only relevant quantity at first order. For $p_{\text{th}} = 0.75\%$ and an error probability per cycle and per physical qubit of 10^{-3} , this ratio p/p_{th} is given by ≈ 0.13 .

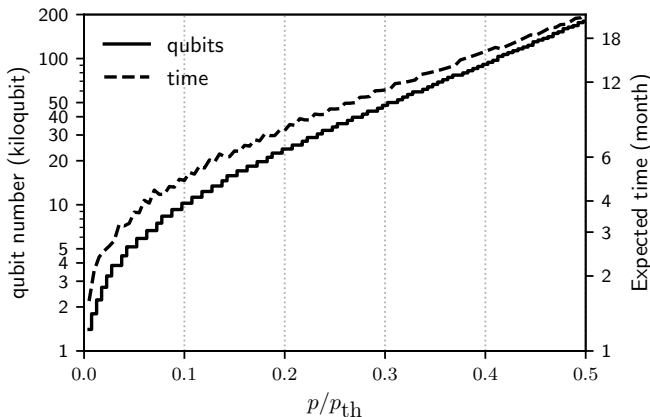


Figure 16. Number of qubits in the processor and run-time to factor of 2048-bit RSA integers in function of the ratio between the physical qubit error and the fault-tolerant code threshold.

We emphasize that the value of the threshold for 3D gauge color codes does not change the take home message of the whole paper, namely that the use of a quantum memory in quantum computing strongly reduces the number of qubits in the processing unit. Even when considering for example a circuit-level threshold of 0.2% and a error probability per operation of 10^{-3} , the use of a quantum memory reduces the number of qubits by two orders of magnitude in the processor compared to an architecture without memory for factoring 2048-bit RSA integers (the same conclusion holds when considering the standard approach using surface code, see [Appendix F 3](#)).

Appendix F: Results and possible improvements

We presented in the main text the resources needed to factor 2048-bit RSA integers corresponding to the most common RSA key size. In the first subsection of this appendix, we discuss the factorization of RSA integers of various sizes. The second subsection is dedicated to a discussion on ways to reduce the run-time to factor RSA integers and in particular, on the trade-off between

the number of physical qubits in the processor and the run-time.

1. Optimal parameters to factor n -bit RSA integers

The resources and parameters needed to factor RSA integers encoded in n bits are specified in [Table I](#). In particular, we consider the factorization of RSA integers with $n = 6$ bits, the number of bits needed to factor 35. We also consider $n = 829$ which corresponds to the largest RSA integer factorized so far [\[51\]](#).

2. Trade-off between qubits and run-time

We have estimated that an average run-time of 177 days is needed to factor a 2048-bit RSA number. There are several ways to reduce this number, most of them coming at the cost of using more qubits in the processor. The items below present several ways separately.

- Due to the tetrahedral geometry of the code structure, only one third of the processor qubits are used during the error-correction steps in average. A factor 3 in time could thus be saved by making use of them.
- The logical circuit can be parallelized in several ways, giving a speed-up roughly proportional to the increase in qubit numbers in the processor. More precisely:
 - Some operations in the adder can be parallelized (see [Figure 13](#)). The controlled NOT operations aligned vertically can be applied at the same time.
 - The run-time is dominated by the time spent to implement the CNOT gates of the quantum lookup circuit (see [Figure 10](#)) and they are easily parallelizable. A full parallelization, would reduce the factorization of 2048-bit RSA integers to about 27 days, at the cost of using about 12 million qubits in the processor.
 - Oblivious carry runways allows parallelization of the adders [\[35\]](#).
 - Other type of adders could exploit further parallelizations, for instance lookahead adders [\[43\]](#).

n	n_e	m	w_e	w_m	d	n_{qubits}	t_{exp}	logical qubits	total modes	spatial modes	temporal modes	all memory correction
6	6	4	3	2	7	316	1 min	38	6 650	3 002	5	95 μs
8	9	8	3	2	13	1 060	2 s	58	64 090	15 370	11	319 μs
16	21	11	3	2	17	1 796	10 s	99	244 035	44 451	15	742 μs
128	189	19	3	3	29	5 156	50 min	571	6 971 339	736 019	27	8 ms
256	381	21	3	3	33	6 660	7 hours	1 089	19 585 665	1 813 185	31	17 ms
512	765	24	3	3	37	8 356	2 days	2 122	53 782 090	4 432 858	35	37 ms
829	1 242	26	3	3	41	10 244	11 days	3 396	117 097 476	8 697 156	39	66 ms
2 048	3 029	30	3	3	47	13 436	177 days	8 284	430 229 540	27 825 956	45	186 ms

Table I. For different integer sizes n and corresponding exponent size n_e ($\sim 1.5n$), the table presents the optimal set of parameters, processor size and computation run-time, and the memory requirements.

- During a product-addition operation, the different additions can be parallelized by computing separately partial sums.
- During the exponentiation, the different multiplications can be parallelized by computing separately partial products.
- The qubit number can be reduced using another slicing of the code structure, at the cost of a longer computation time. For example, if one chooses to cut the tetrahedron by slices parallel to a facet of this tetrahedron, we estimate that a 2 048-bit RSA integer could be factorized with 6 628 qubits in the processor and 354 days.

3. Decoupling the gain from 3D gauge color code and multimode memory

Two new design elements have been proposed in this manuscript, the use of 3D Gauge color codes and an architecture using a multi-mode memory. We here separate them out and get insight into the improvements from each.

The main motivation to use 3D gauge color codes is to get rid of the magical state factory needed for implementing non-Clifford gates in surface code. However, the transversality of T gate on 3D gauge color codes is strongly linked with the dimensionality, and 2D color codes can't directly achieve it [23, 52]. There is no direct way to make use of a 3D color code on a 2D grid.

The main advantage brought by the memory is to unload qubits from the processing unit to the memory. Using a memory in the standard approach for example (2D grid and surface code), we estimate that a RSA-2 048 integer can be factorized with a 2D surface code in about 68 days using a memory that can store up to 5 million modes and a processor with 184 thousand qubits, 180 thousand being dedicated to the magical state factory and 4 thousand to the logical qubits on the processor. The additional reduction in the processor size in our

approach comes from the fact that there is no need for magic state distillation in the use 3D gauge color codes. The number of qubits in the processor is kept small because the qubits are released from the memory and process slice by slice.

Appendix G: Memory requirements to factor RSA-2 048 integers

We would like to first emphasize that the main objective of our project was to evaluate accurately the performance of an architecture in which unprocessed qubits are stored in a quantum memory. The standard approach suffers from the need of millions of individually controlled qubits and several research entities are dedicating large teams of engineers to tackle this challenge. We have shown through Shor's algorithm that the use of a quantum memory reduces significantly the number of qubits in the processor though a significant change in the way the information is processed and protected against errors. Our results hence provide a solution to an engineering problem and turns it into a physics problem: the implementation of a faithful and multimode memory. Before discussing the requirements on the memory in detail, let us clearly define the notion of multimode memory [53].

From an algorithm point of view, "spatial modes" are stored modes that can be accessed in constant time, while "temporal modes" can only be sequentially recoverable (first stored, first release). In the proposed implementation based on spin-echo, temporal modes correspond to different time slots, photons arriving in different time bins being remitted sequentially after spin refocusing. Spatial modes correspond to either different spatial (transverse) modes of a cavity or to different cavities (with possibility to combine both). As discussed in the main text, it is possible to use temporal multiplexing only, at the cost of increasing the run-time.

We now estimate that the factorization of RSA-2 048

integers with the proposed architecture would take a memory with the following characteristics:

- A large multimode capacity to store 28 million spatial modes, each spatial mode being used to store 45 temporal modes. We stress that the number of modes in the memory has not been optimized (only the number of qubits in the processor and the run-time are optimized). Note also that different choices of processing and error-correction protocols may lead to compromises between the numbers of processing qubits and multimode capacity, if needed. For example, we estimate that RSA-2048 integers can be factorized in 68 days with a 2D surface code using a memory that can store up to 5 million modes and a processor with 184 thousand qubits in the processor.
- Storage time greater than 186 ms. More precisely, we estimate that the maximum storage time between two readouts of the same qubits is less than 2 hours. A memory with a storage time of at least two hours is however not necessary as error-correction steps can be implemented periodically at the cost of increasing the run-time. Error correction of all the qubits stored in the memory is estimated to take 186 ms with a processor having 13 436 qubits, meaning that the storage time simply needs to be longer than 186 ms. Applying a correction every second for example would increase the run-time by about 23%.
- Error probability for a transfer to memory, storage, and retrieval less than 0.1%. Note that this requirement for a complete cycle of write/read from memory is likely very conservative. Indeed, the threshold value of error correction is mainly determined by the errors happening during the stabilizer measurements. We thus conjecture that the error correction could handle higher error rate for those specific operations. The effect of this strongly dissymmetric noise between the memory/processor operations is still under investigation, and we choose to stick to the conservative hypothesis for this article.
- The information stored in a given memory mode can be mapped to 3 qubits of the processor: two for the 2-qubit gates (depending on whether the physical qubit is the logical control or target qubits) and one for the error correction and 1-qubit gates. No need for an all to all connectivity.

Appendix H: Realization combining a rare-earth doped solid and a superconducting resonator

For implementing a multimode memory with a spin-echo technique, materials doped with rare-earth, such as Erbium Er^{3+} provide an appealing example since these ions have doubly-degenerate Zeeman states which split when an external magnetic field is applied. Several manuscripts have reported on the successful coupling between the crystal $\text{Er}^{3+}:\text{Y}_2\text{SiO}_5$ and a superconducting microwave resonator [31–33]. Ref. [33] in particular reported on the strong coupling with a collective coupling rate $g\sqrt{N} = 2\pi \times 34$ MHz and an inhomogeneous linewidth $\Gamma = 2\pi \times 12$ MHz. This results in a very high absorption coefficient $\alpha = 4.0 \text{ m}^{-1}$. If we assume a $L = \lambda/2$ cavity, unit absorption and re-emission efficiencies are obtained if the quality factor is $Q = F = 2\pi/(\alpha\lambda) \approx 26$ for a 5 GHz cavity. In this low-Q regime, $\kappa \gg \Gamma$ and a coherence time of a few hundreds of microseconds would translate into a multimode capacity of a few tens of modes. By working with crystals having lower doping concentrations, the coherence time can be significantly increased [54], while still reaching the impedance matching point with low-Q resonators. In this case, a few thousand modes might realistically be stored very efficiently. Rare-earth doped materials is not the only option and other candidates such as negatively charged nitrogen vacancy color centers in diamond [55] or bismuth donors in silicon [56] may be even more promising.

* elie.gouzien@cea.fr

† <https://quantum.paris>

- [1] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, Superconducting Qubits: Current State of Play, *Annual Review of Condensed Matter Physics* **11**, 369 (2020), 1905.13641.
- [2] L. Lamata, A. Parra-Rodriguez, M. Sanz, and E. Solano, Digital-analog quantum simulations with superconducting circuits, *Advances in Physics: X* **3**, 1457981 (2018), 1711.09810.
- [3] C. Gidney and M. Ekerå, How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits, *Quantum* **5**, 433 (2021), 1905.09749.
- [4] Y. R. Sanders, D. W. Berry, P. C. S. Costa, L. W. Tessler, N. Wiebe, C. Gidney, H. Neven, and R. Babbush, Compilation of Fault-Tolerant Quantum Heuristics for Combinatorial Optimization, *PRX Quantum* **1**, 020312 (2020), 2007.07391.
- [5] J. Lee, D. W. Berry, C. Gidney, W. J. Huggins, J. R. McClean, N. Wiebe, and R. Babbush, Even More Efficient Quantum Computations of Chemistry Through Tensor Hypercontraction, *PRX Quantum* **2**, 030305 (2021), 2011.03494.

- [6] D. Kielpinski, C. Monroe, and D. J. Wineland, Architecture for a large-scale ion-trap quantum computer, *Nature* **417**, 709 (2002).
- [7] D. D. Thaker, T. S. Metodi, A. W. Cross, I. L. Chuang, and F. T. Chong, Quantum Memory Hierarchies: Efficient Designs to Match Available Parallelism in Quantum Computing, in *33rd International Symposium on Computer Architecture (ISCA'06)* (IEEE, 2006) pp. 378–390, [quant-ph/0604070](https://arxiv.org/abs/quant-ph/0604070).
- [8] Z.-L. Xiang, S. Ashhab, J.-Q. You, and F. Nori, Hybrid quantum circuits: Superconducting circuits interacting with other quantum systems, *Reviews of Modern Physics* **85**, 623 (2013), [1204.2137](https://arxiv.org/abs/1204.2137).
- [9] G. Kurizki, P. Bertet, Y. Kubo, K. Mølmer, D. Petrosyan, P. Rabl, and J. Schmiedmayer, Quantum technologies with hybrid systems, *Proceedings of the National Academy of Sciences* **112**, 3866 (2015), [1504.00158](https://arxiv.org/abs/1504.00158).
- [10] C. Grezes, Y. Kubo, B. Julsgaard, T. Umeda, J. Isoya, H. Sumiya, H. Abe, S. Onoda, T. Ohshima, K. Nakamura, I. Diniz, A. Auffeves, V. Jacques, J.-F. Roch, D. Vion, D. Esteve, K. Mølmer, and P. Bertet, Towards a spin-ensemble quantum memory for superconducting qubits, *Comptes Rendus Physique* **17**, 693 (2016), [1510.06565](https://arxiv.org/abs/1510.06565).
- [11] H. Bombín and M. A. Martin-Delgado, Exact topological quantum order in $D = 3$ and beyond: Branyons and brane-net condensates, *Physical Review B* **75**, 075103 (2007), [cond-mat/0607736](https://arxiv.org/abs/cond-mat/0607736).
- [12] P. W. Shor, Algorithms for quantum computation: discrete logarithms and factoring, in *Proceedings 35th Annual Symposium on Foundations of Computer Science* (IEEE Comput. Soc. Press, 1994) pp. 124–134.
- [13] M. Ekerå and J. Hästad, Quantum Algorithms for Computing Short Discrete Logarithms and Factoring RSA Integers, in *Post-Quantum Cryptography*, Lecture Notes in Computer Science, Vol. 10346, edited by T. Lange and T. Takagi (Springer International Publishing, 2017) pp. 347–363, [1702.00249](https://arxiv.org/abs/1702.00249).
- [14] R. L. Rivest, A. Shamir, and L. Adleman, A method for obtaining digital signatures and public-key cryptosystems, *Communications of the ACM* **21**, 120 (1978).
- [15] W. Diffie and M. E. Hellman, New directions in cryptography, *IEEE Transactions on Information Theory* **22**, 644 (1976).
- [16] Information Technology Laboratory, *Digital Signature Standard (DSS)* (2013).
- [17] P. W. Shor, Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer, *SIAM Journal on Computing* **26**, 1484 (1997), [quant-ph/9508027](https://arxiv.org/abs/quant-ph/9508027).
- [18] M. Ekerå, Modifying Shor’s algorithm to compute short discrete logarithms (2016), <https://eprint.iacr.org/2016/1128>.
- [19] M. Ekerå, On post-processing in the quantum algorithm for computing short discrete logarithms (2017), <https://eprint.iacr.org/2017/1122>.
- [20] M. Ekerå, Quantum algorithms for computing general discrete logarithms and orders with tradeoffs (2018), <https://eprint.iacr.org/2018/797>.
- [21] C. Gidney, Windowed quantum arithmetic, [1905.07682](https://arxiv.org/abs/1905.07682) (2019).
- [22] Code is available at https://github.com/ElieGouzien/factoring_with_memory.
- [23] H. Bombín, Gauge color codes: optimal transversal gates and gauge fixing in topological stabilizer codes, *New Journal of Physics* **17**, 083002 (2015), [1311.0879](https://arxiv.org/abs/1311.0879).
- [24] E. T. Campbell, B. M. Terhal, and C. Vuillot, Roads towards fault-tolerant universal quantum computation, *Nature* **549**, 172 (2017), [1612.07330](https://arxiv.org/abs/1612.07330).
- [25] H. Bombín, Single-Shot Fault-Tolerant Quantum Error Correction, *Physical Review X* **5**, 031043 (2015), [1404.5504](https://arxiv.org/abs/1404.5504).
- [26] B. J. Brown, N. H. Nickerson, and D. E. Browne, Fault-tolerant error correction with the gauge color code, *Nature Communications* **7**, 12302 (2016), [1503.08217](https://arxiv.org/abs/1503.08217).
- [27] S. J. Devitt, W. J. Munro, and K. Nemoto, Quantum error correction for beginners, *Reports on Progress in Physics* **76**, 076001 (2013), [0905.2794](https://arxiv.org/abs/0905.2794).
- [28] M. Afzelius, N. Sangouard, G. Johansson, M. U. Staudt, and C. M. Wilson, Proposal for a coherent quantum memory for propagating microwave photons, *New Journal of Physics* **15**, 065008 (2013), [1301.1858](https://arxiv.org/abs/1301.1858).
- [29] M. Afzelius and C. Simon, Impedance-matched cavity quantum memory, *Physical Review A* **82**, 022310 (2010), [1004.2469](https://arxiv.org/abs/1004.2469).
- [30] T. Chanelière, G. Hétet, and N. Sangouard, Quantum Optical Memory Protocols in Atomic Ensembles, in *Advances In Atomic, Molecular, and Optical Physics*, Vol. 67 (Elsevier, 2018) Chap. 2, pp. 77–150, [1801.10023](https://arxiv.org/abs/1801.10023).
- [31] P. A. Bushev, A. K. Feofanov, H. Rotzinger, I. Protopopov, J. H. Cole, C. M. Wilson, G. Fischer, A. V. Lukashenko, and A. V. Ustinov, Ultralow-power spectroscopy of a rare-earth spin ensemble using a superconducting resonator, *Physical Review B* **84**, 060501 (2011), [1102.3841](https://arxiv.org/abs/1102.3841).
- [32] M. U. Staudt, I.-C. Hoi, P. Krantz, M. Sandberg, M. Simoen, P. A. Bushev, N. Sangouard, M. Afzelius, V. S. Shumeiko, G. Johansson, P. Delsing, and C. M. Wilson, Coupling of an erbium spin ensemble to a superconducting resonator, *Journal of Physics B: Atomic, Molecular and Optical Physics* **45**, 124019 (2012), [1201.1718](https://arxiv.org/abs/1201.1718).
- [33] S. Probst, H. Rotzinger, S. Wünsch, P. Jung, M. Jerger, M. Siegel, A. V. Ustinov, and P. A. Bushev, Anisotropic Rare-Earth Spin Ensemble Strongly Coupled to a Superconducting Resonator, *Physical Review Letters* **110**, 157001 (2013), [1212.2856](https://arxiv.org/abs/1212.2856).
- [34] R. B. Griffiths and C.-S. Niu, Semiclassical Fourier Transform for Quantum Computation, *Physical Review Letters* **76**, 3228 (1996), [quant-ph/9511007](https://arxiv.org/abs/quant-ph/9511007).
- [35] C. Gidney, Approximate encoded permutations and piecewise quantum adders, [1905.08488](https://arxiv.org/abs/1905.08488) (2019).
- [36] S. A. Cuccaro, T. G. Draper, S. A. Kutin, and D. P. Moulton, A new quantum ripple-carry addition circuit, [quant-ph/0410184](https://arxiv.org/abs/quant-ph/0410184) (2004).
- [37] C. Gidney, Halving the cost of quantum addition, *Quantum* **2**, 74 (2018), [1709.06648](https://arxiv.org/abs/1709.06648).
- [38] V. Vedral, A. Barenco, and A. Ekert, Quantum networks for elementary arithmetic operations, *Physical Review A* **54**, 147 (1996), [quant-ph/9511018](https://arxiv.org/abs/quant-ph/9511018).
- [39] C. Zalka, Shor’s algorithm with fewer (pure) qubits, [quant-ph/0601097](https://arxiv.org/abs/quant-ph/0601097) (2006).
- [40] R. Babbush, C. Gidney, D. W. Berry, N. Wiebe, J. McClean, A. Paler, A. Fowler, and H. Neven, Encoding Electronic Spectra in Quantum Circuits with Linear T Complexity, *Physical Review X* **8**, 041015 (2018), [1805.03662](https://arxiv.org/abs/1805.03662).
- [41] D. W. Berry, C. Gidney, M. Motta, J. R. McClean, and R. Babbush, Qubitization of Arbitrary Basis Quantum

- Chemistry Leveraging Sparsity and Low Rank Factorization, *Quantum* **3**, 208 (2019), 1902.02134.
- [42] T. G. Draper, Addition on a Quantum Computer, [quant-ph/0008033](https://arxiv.org/abs/quant-ph/0008033) (2000).
- [43] T. G. Draper, S. A. Kutin, E. M. Rains, and K. M. Svore, A logarithmic-depth quantum carry-lookahead adder, *Quantum Information and Computation* **6**, 351 (2006), [quant-ph/0406142](https://arxiv.org/abs/quant-ph/0406142).
- [44] Due to the measurement-based uncomputation, it is often more efficient to implement the non-Clifford operations through AND computation. In case of direct implementation of Toffoli gates, the circuit cost could be slightly reduced.
- [45] D. Poulin, Stabilizer Formalism for Operator Quantum Error Correction, *Physical Review Letters* **95**, 230504 (2005), [quant-ph/0508131](https://arxiv.org/abs/quant-ph/0508131).
- [46] P. Zanardi, D. A. Lidar, and S. Lloyd, Quantum Tensor Product Structures are Observable Induced, *Physical Review Letters* **92**, 060402 (2004), [quant-ph/0308043](https://arxiv.org/abs/quant-ph/0308043).
- [47] M. E. Beverland, A. Kubica, and K. M. Svore, Cost of Universality: A Comparative Study of the Overhead of State Distillation and Code Switching with Color Codes, *PRX Quantum* **2**, 020341 (2021), 2101.02211.
- [48] A. M. Kubica, *The ABCs of the color code: A study of topological quantum codes as toy models for fault-tolerant quantum computation and quantum phases of matter*, Ph.D. thesis (2018).
- [49] A. Kubica and N. Delfosse, Efficient color code decoders in $d \geq 2$ dimensions from toric code decoders, 1905.07393 (2019).
- [50] A. Kubica, M. E. Beverland, F. Brandão, J. Preskill, and K. M. Svore, Three-Dimensional Color Code Thresholds via Statistical-Mechanical Mapping, *Physical Review Letters* **120**, 180501 (2018), 1708.07131.
- [51] F. Boudot, P. Gaudry, A. Guillevic, N. Heninger, E. Thomé, and P. Zimmermann, *Factorization of RSA-250* (2020).
- [52] S. Bravyi and R. König, Classification of Topologically Protected Gates for Local Stabilizer Codes, *Physical Review Letters* **110**, 170503 (2013), 1206.1609.
- [53] C. Simon, H. de Riedmatten, M. Afzelius, N. Sangouard, H. Zbinden, and N. Gisin, Quantum Repeaters with Photon Pair Sources and Multimode Memories, *Physical Review Letters* **98**, 190503 (2007), [quant-ph/0701239](https://arxiv.org/abs/quant-ph/0701239).
- [54] M. Le Dantec, M. Rancic, E. Flurin, D. Vion, D. Esteve, P. Bertet, P. Goldner, T. Chanelière, B. Sylvain, S. Lin, and R. B. Liu, Twenty millisecond electron-spin coherence in an erbium doped crystal, in *Bulletin of the American Physical Society* (American Physical Society, 2021).
- [55] Y. Kubo, C. Grezes, A. Dewes, T. Umeda, J. Isoya, H. Sumiya, N. Morishita, H. Abe, S. Onoda, T. Ohshima, V. Jacques, A. Dréau, J.-F. Roch, I. Diniz, A. Auffeves, D. Vion, D. Esteve, and P. Bertet, Hybrid Quantum Circuit with a Superconducting Qubit Coupled to a Spin Ensemble, *Physical Review Letters* **107**, 220501 (2011), 1110.2978.
- [56] V. Ranjan, J. O'Sullivan, E. Albertinale, B. Albanese, T. Chanelière, T. Schenkel, D. Vion, D. Esteve, E. Flurin, J. J. L. Morton, and P. Bertet, Multimode Storage of Quantum Microwave Fields in Electron Spins over 100 ms, *Physical Review Letters* **125**, 210505 (2020), 2005.09275.