Editors' Suggestion | Featured in Physics

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

(Received 11 March 2021; revised 20 July 2021; accepted 3 August 2021; published 28 September 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 2 048-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.

DOI: 10.1103/PhysRevLett.127.140503

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 Fig. 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 Håstad [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 μs, and a reaction time of 10 μs): 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].

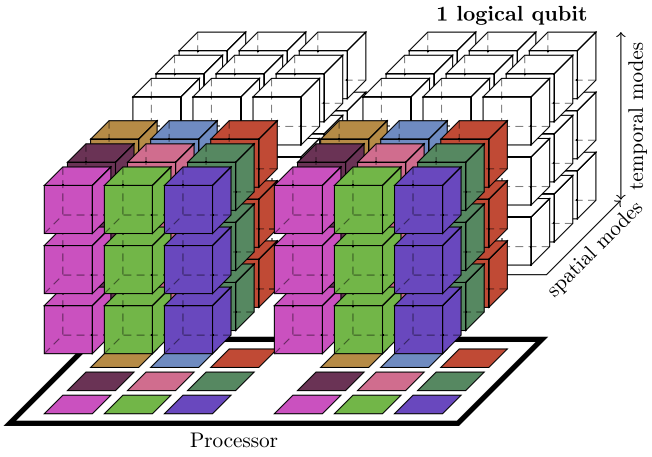

FIG. 1. Quantum computer architecture using a processor made with a 2D grid of qubits and a memory operating as a qubit register where the address of each qubit is specified by a temporal and spatial index. Only (dressed) logical qubits are represented; additional ancillary qubits are used for measuring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

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, 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 2 048-bit RSA integer in 177 days using a multimode memory with a storage time of about 2 hours and a processor including 13 436 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 integers 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 $h = g^{(pq-p-q+1)/2} \times g^{(p+q-2)/2} \equiv g^{(p+q-2)/2} \bmod N$ 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 $g^r \equiv 1 \bmod N$) 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 $(e_1, e_2) \to g^{e_1} h^{-e_2}$ 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 Supplemental Material [21], Sec. A).

Number of gates.—The modular exponentiation needed in Ekerå and Håstad's algorithm, i.e., the operation $|e\rangle|1\rangle \mapsto |e\rangle|g^e \bmod N\rangle$, with the input e and the output $g^e \bmod N$ 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 the Supplemental Material [21], Sec. B for details). Each modular addition is obtained with a standard adder circuit at the cost of a specific representation—the coset representation (see the Supplemental Material [21], Sec. 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 the Supplemental Material [21], Sec. B). For the controlled addition, we can use a semiclassical 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 the Supplemental Material [21], Sec. 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 the Supplemental Material [21], Sec. B). This cost can however be reduced using windowed arithmetic circuits [44]. The basic idea consists of grouping the

bits of e by blocks (each including 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 the Supplemental Material [21], Sec. D, the cost of exponentiation is dominated in this case by $2[n_e(n+m)n/(w_e w_m)]$ 1-qubit gates, $[2^{w_e+w_m}n + 12(n+m)][n_e(n+m)/(w_e w_m)]$ CNOTs, and $4[n_e(n+m)^2/(w_e w_m)]$ 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 [45].

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 [46], contrary to standard ways of operating the surface code [47].

The two codes are based on a shared geometrical structure: a large tetrahedron constructed from elementary tetrahedrons (see the Supplemental Material [21], Sec. 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 [48]. 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 the Supplemental Material [21], Sec. 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 [[49], Eq. (4)]

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

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 the Supplemental Material [21], Sec. E4, the run-time and the resource as a function of the code threshold.

Architecture.—For simplicity, the tetrahedral structure of the error correction (see the Supplemental Material [21], Sec. E) can be included into a large cube in which physical qubits are now represented by elementary cubes (see Fig. 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 [46,50]. 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 $n_{\text{qubits}} = 2 \times 2 \times [(3d^2 + 2d - 3)/2]$, corresponding to two logical qubit slices (see the Supplemental Material [21], Sec. 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 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 $t_c = 1\ \mu$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 $t_{\mathrm{exp}} \times n_{\mathrm{qubits}}$. $t_{\mathrm{exp}} = t/p_s$ 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 the Supplemental Material [21], Sec. F. Our estimation suggests that the factorization of a 2 048-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 the Supplemental Material [21], Sec. 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 the Supplemental Material [21], Sec. 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.

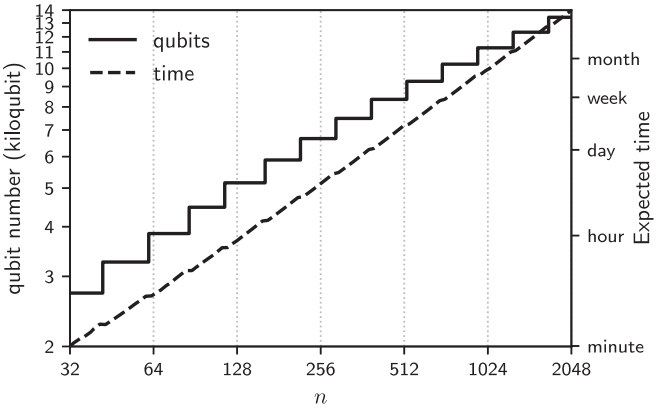

FIG. 2. Number of qubits in the processor and run-time to factor n-bit RSA integers with a computer architecture using a multimode memory.

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. [51], 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 $C = g^2\bar{N}/(\kappa\Gamma) = \alpha L \times \mathcal{F} = 1$ [52]. Once absorbed, the microwave field can be re-emitted by time reversing the inhomogeneous dephasing using a spin echo technique [53]. 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 $C = 1$ [51]. In the regime $\kappa \gg g\sqrt{\bar{N}} \gg \Gamma$, the memory bandwidth is given by 4Γ [51], meaning that any input with a spectrum, say, 10 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 [51]. Assuming that the storage efficiency is unchanged if the storage time is 100 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 the 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 (3 002 spatial modes and 5 temporal modes) and a processor 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 (6650) 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 the Supplemental Material [21], Sec. 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 [54–56].

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.

*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, Annu. Rev. Condens. Matter Phys. **11**, 369 (2020).

[2] L. Lamata, A. Parra-Rodriguez, M. Sanz, and E. Solano, Digital-analog quantum simulations with superconducting circuits, Adv. Phys. **3**, 1457981 (2018).

[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).

[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).

[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 (London) **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 *Proceedings of the 33rd International Symposium on Computer Architecture (ISCA'06)* (IEEE, New York, 2006), pp. 378–390, https://dx.doi.org/10.1109/ISCA.2006.32.

[8] Z.-L. Xiang, S. Ashhab, J.-Q. You, and F. Nori, Hybrid quantum circuits: Superconducting circuits interacting with other quantum systems, Rev. Mod. Phys. **85**, 623 (2013).

[9] G. Kurizki, P. Bertet, Y. Kubo, K. Mølmer, D. Petrosyan, P. Rabl, and J. Schmiedmayer, Quantum technologies with hybrid systems, Proc. Natl. Acad. Sci. U.S.A. **112**, 3866 (2015).

[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, C.R. Phys. **17**, 693 (2016).

[11] H. Bombín and M. A. Martin-Delgado, Exact topological quantum order in $D = 3$ and beyond: Branyons and brane-net condensates, Phys. Rev. B **75**, 075103 (2007).

[12] P. W. Shor, Algorithms for quantum computation: discrete logarithms and factoring, in *Proceedings of the 35th Annual Symposium on Foundations of Computer Science* (IEEE, New York, 1994), pp. 124–134, https://dx.doi.org/10.1109/SFCS.1994.365700.

[13] M. Ekerå and J. Håstad, Quantum algorithms for computing short discrete logarithms and factoring RSA integers, in *Post-Quantum Cryptography*, Lecture Notes in Computer Science, edited by T. Lange and T. Takagi, Vol. 10346 (Springer International Publishing, New York, 2017), pp. 347–363, https://dx.doi.org/10.1007/978-3-319-59879-6_20.

[14] R. L. Rivest, A. Shamir, and L. Adleman, A method for obtaining digital signatures and public-key cryptosystems, Commun. ACM **21**, 120 (1978).

[15] W. Diffie and M. E. Hellman, New directions in cryptography, IEEE Trans. Inf. Theory **22**, 644 (1976).

[16] Information Technology Laboratory, *Digital Signature Standard (DSS)* (National Institute of Standards and Technology (NIST), 2013), https://dx.doi.org/10.6028/NIST.FIPS.186-4.

[17] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, SIAM J. Comput. **26**, 1484 (1997).

[18] M. Ekerå, Modifying shor's algorithm to compute short discrete logarithms, https://eprint.iacr.org/2016/1128.

[19] M. Ekerå, On post-processing in the quantum algorithm for computing short discrete logarithms, https://eprint.iacr.org/2017/1122.

[20] M. Ekerå, Quantum algorithms for computing general discrete logarithms and orders with tradeoffs, https://eprint.iacr.org/2018/797.

[21] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.140503 for details about the logical circuit, error correction and implementation, which includes Refs. [22–43].

[22] R. B. Griffiths and C.-S. Niu, Semiclassical Fourier Transform for Quantum Computation, Phys. Rev. Lett. **76**, 3228 (1996).

[23] C. Gidney, Approximate encoded permutations and piecewise quantum adders, arXiv:1905.08488.

[24] S. A. Cuccaro, T. G. Draper, S. A. Kutin, and D. P. Moulton, A new quantum ripple-carry addition circuit, arXiv:quant-ph/0410184.

[25] C. Gidney, Halving the cost of quantum addition, Quantum **2**, 74 (2018).

[26] V. Vedral, A. Barenco, and A. Ekert, Quantum networks for elementary arithmetic operations, Phys. Rev. A **54**, 147 (1996).

[27] C. Zalka, Shor's algorithm with fewer (pure) qubits, arXiv:quant-ph/0601097.

[28] 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, Phys. Rev. X **8**, 041015 (2018).

[29] 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).

[30] T. G. Draper, Addition on a quantum computer, arXiv:quant-ph/0008033.

[31] T. G. Draper, S. A. Kutin, E. M. Rains, and K. M. Svore, A logarithmic-depth quantum carry-lookahead adder, Quantum Inf. Comput. **6**, 351 (2006).

[32] D. Poulin, Stabilizer Formalism for Operator Quantum Error Correction, Phys. Rev. Lett. **95**, 230504 (2005).

[33] P. Zanardi, D. A. Lidar, and S. Lloyd, Quantum Tensor Product Structures are Observable Induced, Phys. Rev. Lett. **92**, 060402 (2004).

[34] 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).

[35] 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, California Institute of Technology, 2018, https://dx.doi.org/10.7907/059V-MG69.

[36] A. Kubica and N. Delfosse, Efficient color code decoders in $d \geq 2$ dimensions from toric code decoders, arXiv:1905.07393.

[37] A. Kubica, M. E. Beverland, F. Brandão, J. Preskill, and K. M. Svore, Three-Dimensional Color Code Thresholds via Statistical-Mechanical Mapping, Phys. Rev. Lett. **120**, 180501 (2018).

[38] F. Boudot, P. Gaudry, A. Guillevic, N. Heninger, E. Thomé, and P. Zimmermann, Factorization of RSA-250, 2020, https://lists.gforge.inria.fr/pipermail/cado-nfs-discuss/2020-February/001166.html.

[39] S. Bravyi and R. König, Classification of Topologically Protected Gates for Local Stabilizer Codes, Phys. Rev. Lett. **110**, 170503 (2013).

[40] C. Simon, H. de Riedmatten, M. Afzelius, N. Sangouard, H. Zbinden, and N. Gisin, Quantum Repeaters with Photon Pair Sources and Multimode Memories, Phys. Rev. Lett. **98**, 190503 (2007).

[41] 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, New York, 2021), http://meetings.aps.org/Meeting/MAR21/Session/B31.3.

[42] 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, Phys. Rev. Lett. **107**, 220501 (2011).

[43] 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, Phys. Rev. Lett. **125**, 210505 (2020).

[44] C. Gidney, Windowed quantum arithmetic, arXiv:1905.07682.

[45] Code is available at https://github.com/ElieGouzien/factoring_with_memory.

[46] H. Bombín, Gauge color codes: optimal transversal gates and gauge fixing in topological stabilizer codes, New J. Phys. **17**, 083002 (2015).

[47] E. T. Campbell, B. M. Terhal, and C. Vuillot, Roads towards fault-tolerant universal quantum computation, Nature (London) **549**, 172 (2017).

[48] H. Bombín, Single-Shot Fault-Tolerant Quantum Error Correction, Phys. Rev. X **5**, 031043 (2015).

[49] B. J. Brown, N. H. Nickerson, and D. E. Browne, Fault-tolerant error correction with the gauge color code, Nat. Commun. **7**, 12302 (2016).

[50] S. J. Devitt, W. J. Munro, and K. Nemoto, Quantum error correction for beginners, Rep. Prog. Phys. **76**, 076001 (2013).

[51] M. Afzelius, N. Sangouard, G. Johansson, M. U. Staudt, and C. M. Wilson, Proposal for a coherent quantum memory for propagating microwave photons, New J. Phys. **15**, 065008 (2013).

[52] M. Afzelius and C. Simon, Impedance-matched cavity quantum memory, Phys. Rev. A **82**, 022310 (2010).

[53] T. Chanelière, G. Hétet, and N. Sangouard, Quantum optical memory protocols in atomic ensembles, in *Advances In Atomic, Molecular, and Optical Physics* (Elsevier, New York, 2018), Vol. 67, ch. 2, pp. 77–150, https://dx.doi.org/10.1016/bs.aamop.2018.02.002.

[54] 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, Phys. Rev. B **84**, 060501(R) (2011).

[55] 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, J. Phys. B **45**, 124019 (2012).

[56] 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, Phys. Rev. Lett. **110**, 157001 (2013).