ARTICLE OPEN

Check for updates

Resource-efficient simulation of noisy quantum circuits and application to network-enabled QRAM optimization

Luís Bugalho[1,2,3,4]✉, Emmanuel Zambrini Cruzeiro[5], Kevin C. Chen[6,7], Wenhan Dai[7,8], Dirk Englund[6,7] and Yasser Omar[1,2,3]

Giovannetti, Lloyd, and Maccone (2008) proposed a quantum random access memory (QRAM) architecture to retrieve arbitrary superpositions of N (quantum) memory cells via quantum switches and $O(\log(N))$ address qubits. Toward physical QRAM implementations, Chen et al. (2021) recently showed that QRAM maps natively onto optically connected quantum networks with $O(\log(N))$ overhead and built-in error detection. However, modeling QRAM on large networks has been stymied by exponentially rising classical compute requirements. Here, we address this bottleneck by: (1) introducing a resource-efficient method for simulating large-scale noisy entanglement, allowing us to evaluate hundreds and even thousands of qubits under various noise channels; and (2) analyzing Chen et al.'s network-based QRAM as an application at the scale of quantum data centers or near-term quantum internet; and (3) introducing a modified network-based QRAM architecture to improve quantum fidelity and access rate. We conclude that network-based QRAM could be built with existing or near-term technologies leveraging photonic integrated circuits and atomic or atom-like quantum memories.

npj Quantum Information (2023)9:105 ; https://doi.org/10.1038/s41534-023-00773-x

INTRODUCTION

A quantum random access memory (QRAM) is an essential computational primitive for many quantum algorithms. The ability to perform a QRAM query in $\log(N)$ time steps, where $N = 2^n$ is the number of memory cells, implies polynomial speed-ups for applications such as quantum machine learning[1], matrix inversion[2], quantum imaging[3], and quantum searching[4]. Despite its clear importance to quantum information processing, a QRAM has yet to be realized experimentally. Hence, finding a suitable architecture that can be realized in the near future remains an active research subject in the theoretical and experimental domains.

In this article, we present a method to simulate large-scale entanglement accounting for various sources of noise. We are able to efficiently simulate circuits with thousands of qubits under dephasing, amplitude damping, and CNOT errors. Based on our simulation model, we present a QRAM architecture for photonic network-based QRAM based on ref. [5]. The feasibility assessment is based on realistic parameters extracted from recent experiments, which we will refer to throughout the article.

A classical RAM[6] consists of a binary tree leading to a final layer of memory cells, each corresponding to a unique address. The address is represented as a series of bits, with each bit corresponding to a layer of the binary tree. Each bit of an address describes how the bus signal propagates in the layer: to the right or to the left child node. Hence, the nodes of the binary tree act as switches for the address. When provided with a n-bit address, the RAM returns a bit string f_k associated with the memory cell labeled k. This is called the fan-out scheme[7].

A QRAM is the quantum analog of the RAM, similarly consisting of addresses, *quantum* switches, and memory cells in the form of qubits. In particular, with a quantum address state, over the set of address qubits a, given by $|\psi'_{\text{in}}\rangle = \sum_{j=1}^{n} a_j |j\rangle_a$, one can retrieve data from a superposition of memory cells. A QRAM query is defined via the following transformation,

$$|\psi_{\text{in}}\rangle = |\psi'_{\text{in}}\rangle|\emptyset\rangle_b \longrightarrow |\psi_{\text{out}}\rangle = \sum_{j=1}^{N} a_j |j\rangle_a |D_j\rangle_b \quad (1)$$

where $|\emptyset\rangle$ represents an ancillary state over the bus qubit b, which transforms into the retrieved data state after querying. In this article, we will restrict our investigations to classical data, i.e., $|D_j\rangle$ are separable bits. A direct conversion of classical fan-out protocol to the quantum realm is inefficient since it requires maintaining quantum coherence over an exponential number of connections[7].

Three main schemes have been investigated to date: the fan-out scheme that was already described, the bucket brigade model, and the teleportation-based scheme. Important figures of merit for the QRAM are the fidelity of the above transformation and the query time. For a detailed study and comparison of the first two schemes, please refer to ref. [8].

In the bucket brigade (BB) model[7,9], the number of qubits of the device scales as $O(2^n)$, as does the number of gates. Moreover, the original protocol[7] includes an additional third state in each node, called the "wait" state, in order to prevent the exponential scaling of the amount of decoherence with respect to the memory size. However, Hann et al.[8] have shown that the origin of the noise resilience of the BB model is the amount of entanglement among the memory's components and not the presence of the "wait" state, as one can devise a BB model without the "wait" state that still achieves a polynomial scaling of the decoherence with respect to the number of memory addresses n.

More recently, Chen et al. presented a photonic network-based QRAM scheme[5] that makes use of quantum teleportation of addresses from a quantum computer to the QRAM binary tree.

[1]Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal. [2]Physics of Information and Quantum Technologies Group, Centro de Física e Engenharia de Materiais Avançados (CeFEMA), Lisboa, Portugal. [3]PQI – Portuguese Quantum Institute, Lisboa, Portugal. [4]Sorbonne Université, CNRS, LIP6, 4 Place Jussieu, Paris F-75005, France. [5]Instituto de Telecomunicações, Lisbon 1049-001, Portugal. [6]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [7]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]Department of Computer Science, University of Massachusetts, Amherst, Massachusetts 01003, USA. ✉email: luis.bugalho@tecnico.ulisboa.pt

Such a scheme greatly increases the protocol's efficiency by teleporting the registers to the layers (initially prepared in GHZ states) in parallel as opposed to in series, thereby circumventing the event of a single qubit loss collapsing the entire tree state. Additionally, the proposed QRAM maps onto quantum networks, leading to potential applications in distributed quantum computing and sensing.

However, Chen et al. left as an open challenge the simulation of the scheme on large-scale networks since the computational complexity scales exponentially with the number of qubits. In this work, we bypass this problem resorting to more efficient ways of modeling the noise in stabilizer states. Moreover, this method generalizes to other quantum networking tasks with similar constructions, such as protocols for distributed quantum computation.

This comes in line with the fact that distributing entanglement is central in quantum information processing schemes ranging from quantum computing to sensing to communications[10–12]. Simulation of distributed entanglement in a network setting, be it a long-distance network such as a possible future quantum internet[13] or small-distance quantum local area network (QLAN)[14], is important to assess the limitations imposed by near-term quantum technologies. The architecture of the QRAM considered in this paper, building on the photonic network-based QRAM proposed in ref. [5], involves a series of exponentially growing GHZ states, with the largest having as many qubits as there are memory cells. Each GHZ state spans across a physical layer in the QRAM architecture, and the number of nodes per layer grows exponentially with the number of memory cells $2^n \equiv N$ to be addressed, as shown in Fig. 1.

Computer simulations of noisy quantum processes in such a system quickly become computationally intensive[15–17] due to the density matrices growing exponentially in size with the number of qubits. Even though the entire QRAM protocol definition, i.e., the retrieval of data given an input address (see Eq. (1)), requires more than just Clifford operations, creating the routing state over the QRAM architecture only uses Clifford gates. These operations are the ones used to create the GHZ states and to teleport the address state onto the QRAM access layers. Moreover, the operations required to access the QRAM after the routing state is distributed over the routing nodes only grow with the logarithm of the number of qubits of the QRAM, in comparison to the linear amount of operations required to create the routing state. This is the reason why noise in the system mostly comes from the GHZ states before access. In particular, this set of operations to create the routing state can be classically simulated efficiently[16]. This approach enables an explicit and efficient description of all the intermediary states, up to local unitary corrections. In this article, we develop efficient methods to simulate large-scale noisy entanglement by characterizing the impact of noise at all intermediate steps and apply these tools to simulate a noisy QRAM.

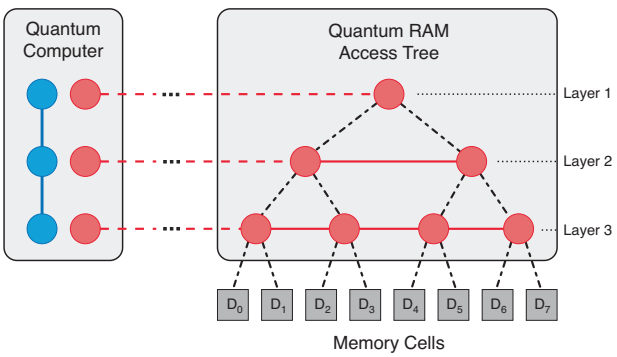

Fig. 1 Overview of a teleportation-based QRAM architecture. A quantum RAM in the form of a binary tree comprises GHZ states for each physical layer. The left-most node of each layer i is entangled with an ancillary qubit in a remote quantum computer, which hosts the query address qubits (blue). Bell state measurement in the quantum computer then teleports the address state onto the access tree. The elementary operations to constructing GHZ states in a photonic integrated circuit (PIC) QRAM are identical to the ones over[5].

There are several architectures for a QRAM. Here, we focus on the optically mediated quantum network-based QRAM architecture introduced in ref. [5], as it offers several key benefits: implementation in quantum networks compatible with envisioned quantum internet architecture and quantum data centers, faster query times and the possibility of executing in a non-local manner by means of teleportation. Hence, this scheme works under any network-like architecture, be it locally (e.g., on a chip) or across large distances (e.g., over a quantum internet). Without loss of generality, we characterize each node of the architecture as one of a spin-photon network that could be implemented in photonic integrated circuits (PIC).

The architecture of the QRAM is similar to previous models, such as BB and the fan-out models. The main difference concerns the execution of the protocol and the resources available at each node. In this architecture, one considers two agents: the quantum computer, which prepares the addresses, and the QRAM or quantum access tree (see Fig. 1). The quantum computer must provide an address state with $n = \log_2 N$ qubits, where N is the total number of memories (for simplicity assume $n \in \mathbb{N}$). The QRAM has a binary tree architecture, with n physical layers, where the kth layer ($k \in \{1, \ldots, n-1\}$) has 2^{k-1} quantum nodes. As we describe next, in each physical layer, all the nodes share a GHZ state, which is used to teleport the address state onto the QRAM itself, allowing for an ancilla qubit to access the memories in the correct superposition.

As for the type of physical implementation chosen, and without loss of generality, we focus on a QRAM implementation involving solid-state spin qubits integrated into PICs, an approach that is promising in terms of scalability. In particular, we consider diamond nanophotonic cavities coupled with silicon-vacancy centers[18,19] as each QRAM tree node. Each emitter contains an electronic spin that directly interacts with the photonic address register qubits and an accompanying nuclear spin acting as a long-lived memory. By entangling the electronic spin with the photon via cavity reflection, consecutive reflection of a photon off two neighboring nodes and subsequent heralding achieves spin-spin entanglement. This remote entangling strategy is repeatedly used to generate a GHZ state across each layer. Such operations are probabilistic (see Fig. 2): the photon has a non-zero probability of being lost to the environment before reflecting off two cavities and arriving at the detector. On the other hand, it is possible to perform close to deterministic two-qubit gates between the electronic and nuclear spin qubits, albeit with a larger error[20,21]. For this reason, we term this architecture teleportation-based deterministic QRAM or TD-QRAM.

In these types of systems, the main contributors to errors are (1) spin phase errors (at rate $1/T_2$), (2) spin-flip errors (at rate $1/T_1$), and (3) errors in hyperfine gates between electron and nuclear spins

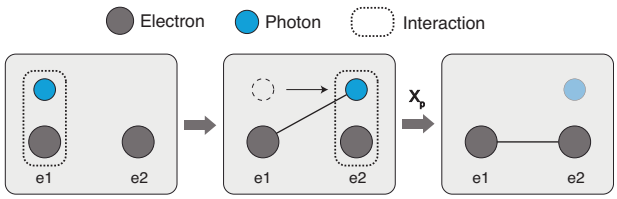

Fig. 2 Probabilistic CNOT. Execution of a CNOT gate between two electrons, e1 and e2, mediated by a photon.

(see Supplementary Table I). We leave out photon-electron interactions, as one could conceive trading off the efficiency η for arbitrarily high fidelity in the cavity-reflection-based scheme proposed in ref. [22] in the high-cooperativity and over-coupling regime.

Hence, we explore different values for T_1, T_2 of both electronic and nuclear spin qubits and p_e and p_n for the probabilities of error in electronic and nuclear spin CNOTs. For the remainder of this article, we set $T_1^n = 100\,T_1^e \equiv 100\,T_1$ and $T_2^n = 100\,T_2^e \equiv 100\,T_2$. Nuclear spins have a higher coherence time as they are much less coupled to the noisy spin-bath compared to electronic spins. Reported values of characteristic times go, experimentally, up to $T_1^e \sim 1\,\mathrm{s}, T_2^e \sim 10\,\mathrm{ms}$[23], and there are theoretical predictions of being able to reach $p_e, p_n = 10^{-2} \sim 10^{-4}$[24,25]. Moreover, we detail other important physical parameters of this type of system used for the simulations in Supplementary Table I.

RESULTS

Simulating the effects of decoherence for a TD-QRAM

To simulate the QRAM initialization protocol, we use NetSquid[26] under the stabilizer formalism and extract all the parameters of the noise channels before implementing them in simulations, for instance: timing parameters for every qubit used throughout the simulation, all the noisy CNOTs with corresponding error probabilities, and to which qubits and at which step it is applied. From here, we compute the fidelity of the final QRAM state by substituting all these values into the expressions presented in the "Methods".

We start by presenting the simulation of a 2^{12}-qubit QRAM in Fig. 3. Here, we detail individually the fidelities of the GHZ state distributed at each physical layer of the QRAM. The fidelity of the full state of the QRAM is given by:

$$F(\,\mathrm{QRAM}\,) = \prod_{i=1}^{n-1} \mathcal{F}\left(\mathrm{Layer}_i, |\mathrm{GHZ}\rangle_{2^{i-1}+1}\right), \quad \text{where } |\mathrm{GHZ}\rangle_q = \tfrac{1}{\sqrt{2}}\left(|0\rangle^{\otimes q} + |1\rangle^{\otimes q}\right) \tag{2}$$

i.e., the fidelity of the entire tree (or the QRAM) is defined as the product of the fidelities of each physical layer (see Supplementary Methods for more details). We distinguish access fidelity from tree fidelity, where the former refers to the fidelity of the state retrieved after accessing the memory cells ($|\psi_{\mathrm{out}}\rangle$ in Eq. (1)), and

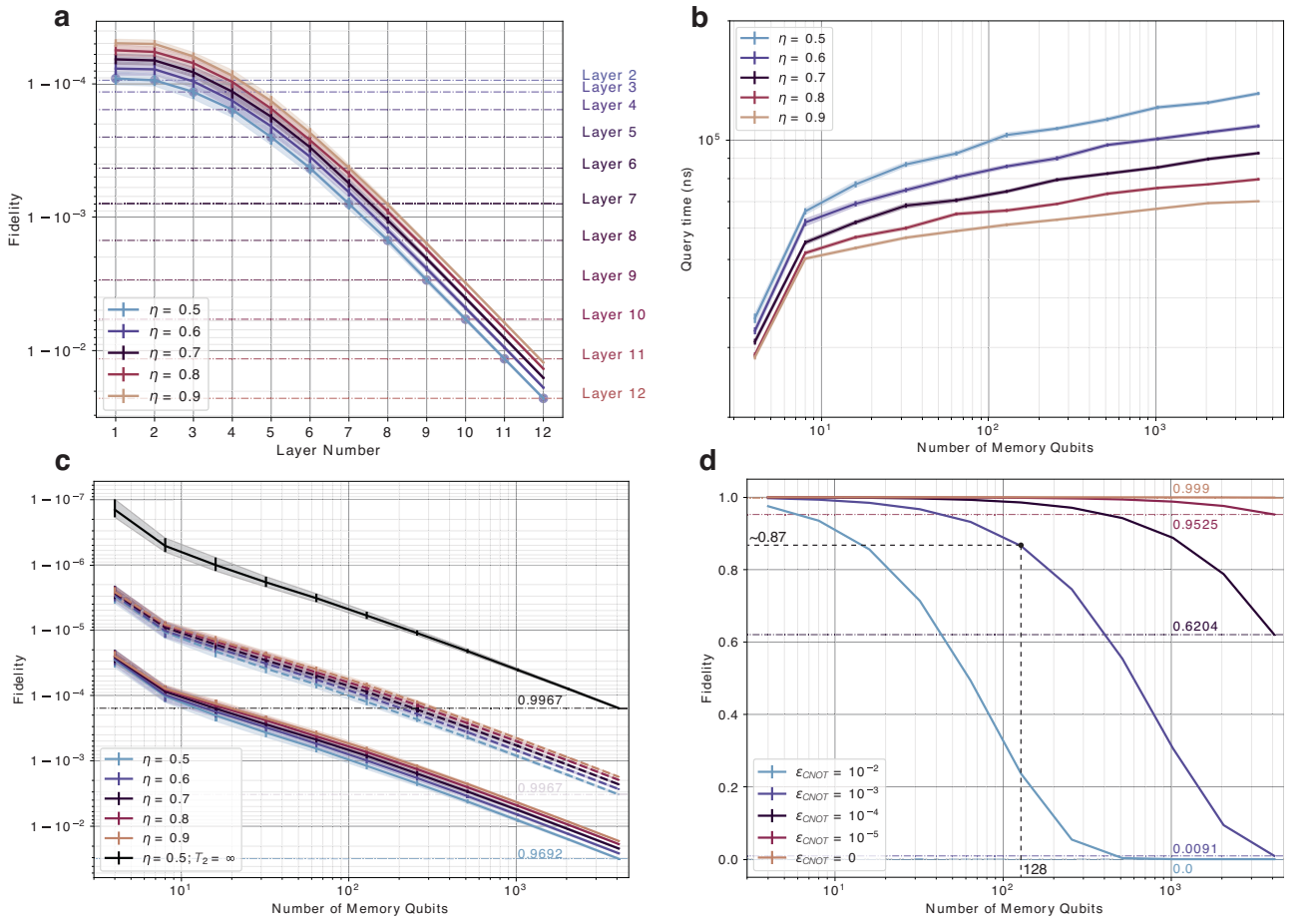

Fig. 3 TD-QRAM Simulations. a TD-QRAM access protocol for 12 layers, with the efficiency of generating a Bell pair swept from $\eta = 50\%$ to $\eta = 90\%$. The noise analysis considers only dephasing and damping errors. The final fidelity is calculated according to Eq. (2), with $T_1 = 20$ ms, $T_2 = 10$ ms, and $\epsilon_{\mathrm{CNOT}} = 0$ for each layer. **b** Query times with varying sizes from 2 layers to 12 layers, and sweeping the efficiency of generating a Bell pair from $\eta = 50\%$ to $\eta = 90\%$. There is an expected logarithmic scaling of the query time with the number of qubits. **c** TD-QRAM noise analysis with dephasing errors, $T_2 = 10$ ms (filled lines) and $T_2 = 100$ ms (traced lines), with fixed amplitude-damping error $T_1 = 2$ s. We consider different QRAM sizes from 2 layers to 12 layers as well as various efficiencies of generating a Bell pair from $\eta = 50\%$ to $\eta = 90\%$. **d** TD-QRAM noise analysis with noisy CNOTs, $p_e = p_n \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, for a QRAM with the number of layers ranging from 2 to 12. The dephasing time is fixed at $T_2 = 100$ ms, and the amplitude-damping time is fixed at $T_1 = 2$ s. The efficiency of generating a Bell pair is fixed at $\eta = 90\%$. The final fidelity mainly depends on the number of noisy CNOTs performed throughout the protocol and has little dependence on the efficiency. All the error bars over the data correspond to the error of the average value over 100 simulations of the protocol.

the latter refers to the multipartite state fidelity of the binary tree constituting the QRAM. Only the access fidelity depends on the address and bus qubits.

One observes an exponential decrease of the fidelity with the number of the layer (notice the logarithmic scaling on the *y*-axis corresponding to the fidelity). This agrees with the GHZ state size increasing exponentially with the number of layers, i.e., scaling 2^k. When one qubit in this multipartite state suffers an error, the entire state is affected.

One critical figure of merit that we extract from the NetSquid simulations is the query time. As demonstrated in ref. [5], the query efficiency scales logarithmically with the number of qubits. Extracting from multiple queries of the QRAM, we obtain the query times (apart from a logarithmic factor derived from making the bus qubit traverse the binary tree) in Fig. 3.

Dephasing and damping errors for TD-QRAM

Considering only the effects of dephasing and amplitude-damping errors in the spin qubits, we take $T_2 = \infty$ ms for amplitude-damping errors only, and then $T_2 = 10$ ms and $T_2 = 100$ ms with a fixed $T_1 = 2$ s[23], see Supplementary Table I. We also set the CNOT error rate to 0. We present the simulation results for the TD-QRAM scheme under memory dephasing for increasing QRAM size, as shown in Fig. 3.

Looking closely at Fig. 3c, one can observe that the effect of amplitude-damping shows an identical behavior to the one of dephasing and amplitude-damping combined, i.e., with the same type of scaling. However, it is residual compared to the effect of dephasing. This is easily explainable by the time-scales of the coherence times of the corresponding noises (T_1 and T_2) in the memory differ by orders of magnitude, with the first, T_1, being usually much longer than the latter, T_2, i.e., $T_1 \sim 1$ s[23]. For this reason, its impact can be neglected relative to other sources of error.

Dephasing, damping and noisy CNOTs for TD-QRAM

The only type of error missing in the analysis is the error derived from the use of noisy CNOTs. Illustrated in Fig. 3, the dephasing and damping errors minimally contribute to infidelity. We now analyze the case for noisy CNOTs on top of fixed $T_1 = 2$ s and $T_2 = 100$ ms (note we now switch to linear scale in the y-axis for the fidelity due to the set of values present for the different simulations). For simplicity, we consider equal CNOT error probability, ϵ_{CNOT}, for both *electronic* and *nuclear* CNOTs, and vary ϵ_{CNOT} from 10^{-5} to 10^{-2}, as shown in Fig. 3:

These simulations show that the CNOT gates dominate the overall error in the QRAM state fidelity in the TD-QRAM. For instance, to access a 128-qubit QRAM, one needs fidelities of the CNOT gates to be somewhere near 99.9% to obtain an access fidelity exceeding 90%. In this architecture, while the query times do not increase linearly with the size of the memory, the errors *do*. Expectedly, applying an error to a single qubit of a GHZ state contributes in the same order for the entire state.

The price to pay for performing CNOTs with such large error rates deterministically could be circumvented by near-perfect yet probabilistic CNOTs[22,27] via cavity-based electron spin-photon interactions, as opposed to deterministic yet error-prone nuclear-electron spin coupling. In light of this, we explore a *hybrid* teleportation-based QRAM architecture in the following section.

Teleportation-based stochastic QRAM

In the TD-QRAM protocol, the entanglement generation and swap (Fig. 8) operation are still probabilistic, given the finite chance of photon loss. Hence, these *probabilistic* CNOTs are done in parallel throughout each physical layer to improve efficiency. After an EPR pair is created between two electron spins, however, transferring entanglement onto the nuclear spins is a deterministic procedure. Thereby, the query time grows sub-linearly. As noted before addressing the TD-QRAM scheme, this *deterministic* CNOT based on nuclear-electron spin interaction mainly dominates the infidelity of the GHZ state, motivating us to contemplate an alternative solution.

Since the decoherence errors from T_1 and T_2 contribute much less to the infidelity relative to electron-nuclear spin CNOT, replacing some of the noisy deterministic CNOTs with probabilistic CNOTs helps improve fidelity despite reducing efficiency. As we will show, this leads to higher QRAM tree state fidelities, albeit with longer query times. We call this architecture 'teleportation-based *stochastic* QRAM', or TS-QRAM.

Relying solely on probabilistic CNOTs in every step of the protocol would be *very* inefficient since the probability of generating a GHZ state diminishes exponentially with the number of nodes. In other words, if one entanglement attempt fails during the construction of a GHZ state, the entire state collapses. Since each linking process is heralded, there are ways to circumvent this by choosing a specified order to perform the CNOTs, similar to entanglement swapping in a repeater chain[28,29]. Here, the probabilistic swapping operations are equivalent to the probabilistic CNOTs, and measuring the middle node is analogous to joining smaller GHZ states to form a larger GHZ state. Abstractly, they describe the same problem, which allows us to use the solutions provided by ref. [29]. Next, we present an in-depth analysis of the trade-off between fidelity and query rate as a function of error rates and physical implements.

Increasing T_1 and T_2

To decrease the number of employed deterministic CNOTs, and taking into account that these always happen when the electronic spins interact with the nuclear spins, it is natural to consider dropping the nuclear spins altogether. This is motivated by the fact that we can perform CNOTs, albeit probabilistically, between the electron spins. The downside is that electron spins suffer from having shorter coherence times than their nuclear counterparts. Still, it is advantageous to consider such schemes to avoid the use of noisier deterministic CNOTs.

To minimize the consequently increased decoherence, one could conceive schemes for increasing the T_1 and T_2 times for the electrons since these are the ones now causing the fidelity bottleneck, together with the required time to query the memory.

Presently, the SiV's electronic spin's T_1 time is shown to be longer than 1 s[23], thereby posing no concern over depolarization. On the other hand, its T_2 coherence time is limited to tens of milliseconds[23], even under dynamical decoupling. The main dephasing mechanism is attributed to the surrounding nuclear spin bath, which is weakly coupled to the electronic spin of interest via hyperfine interaction[30]. A potential avenue to improving the electronic spin's T_2 is therefore to "purify" its environment by materials engineering[31]. By producing SiV in a carbon-13 free matrix, for example, the coherence time may be further extended.

Nevertheless, our numerical analyses of the hybrid scheme show fidelities still exceeding 60% for a reasonable CNOT error rate of 10^{-3} and 1024 memory cells, using a T_2 of 100 ms. For such a result, a probability of success of about 70% for the CNOT is required.

The teleportation-based stochastic QRAM protocol

In the TD-QRAM protocol, there are two steps occurring in parallel across each layer in the QRAM: one for generating EPR pairs across every other node and another for linking all the states into a larger GHZ state via sharing EPR pairs in-between nodes holding the previously shared EPR pairs (see Fig. 8). This could be made in parallel because the linking operations are deterministic.

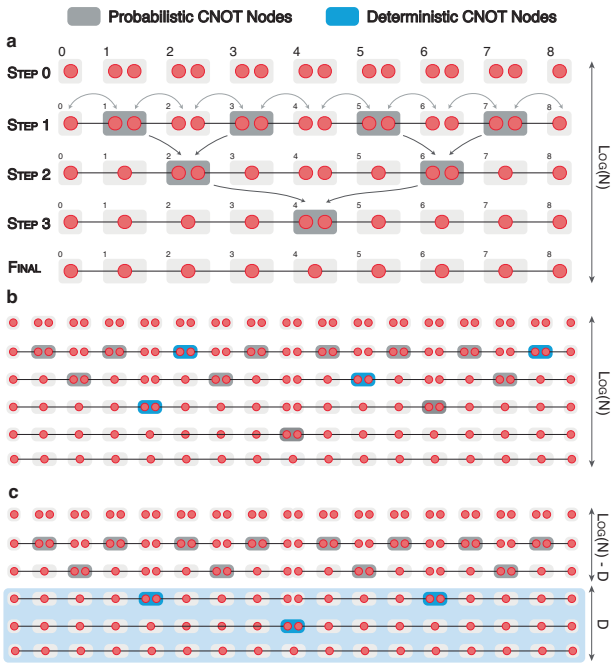

Fig. 4 Binary-tree-like approach of linking nodes and possible placement of deterministic CNOTs. a The arrows represent heralding signals for the subsequent step, and the dark nodes represent the selected nodes for attempting entanglement at each time step. **b** Randomly distributed deterministic nodes across the $\log N$ distribution layers. **c** Intuitively distributed deterministic nodes with D deterministic distribution layers.

In the TS-QRAM protocol, however, we must now consider an order for the linking step that depends on the node's position, similar to the quantum repeater chain problem[26,29,32]. If a linking process fails, the subset of qubits that would have become entangled must be reset. The optimal strategy is then performing the linking process in a binary-tree-like approach[29]. This binary-tree order for the linking processes means that now a heralding signal for a successful link must be exchanged within the tree. Each parenting node will have two children, the right-child and the left-child. Each node only attempts entanglement if it receives heralding signals from both children nodes that have been successfully entangled themselves. Figure 4 illustrates this procedure and defines the order.

Moreover, as mentioned before, the advantage of the TS-QRAM protocol is that probabilistic CNOTs are used to minimize state infidelity. One might consider the optimal placements for the deterministic CNOTs to maximize the GHZ state fidelity across each physical layer. We further introduce having an additional *distribution* layer. This is the layer of the order binary-tree at which a linking step is attempted, as shown in Fig. 4. These abstract layers are only needed to describe the order of the linking steps and help illustrate the optimal placements for the deterministic CNOTs.

For this reason, we present two possible options to solve the placement of deterministic nodes problem: the first is randomly choosing a set of nodes to be deterministic, regardless of their distribution layer. The second option is choosing the nodes that attempt to link entanglement at the higher steps since if those attempts are unsuccessful, they take the biggest toll on the protocol requiring re-attempting every preceded step. We illustrate these two possible options in Fig. 4.

As we will verify later, we need a much smaller number of deterministic nodes if we place them in higher-level distribution layers. We first present simulation results for both cases.

Simulating the effects of decoherence for a TS-QRAM

Using the aforementioned results, we compare the TD-QRAM protocol with the TS-QRAM, which includes both probabilistic EPR pair generation and deterministic linking. For comparison purposes, we start by assuming all probabilistic CNOTs (*Prob* (node being deterministic) $\equiv P_d = 0\%$) have unit efficiency in the latter scheme. If the deterministic and the probabilistic CNOTs are of the same order in gate time, then the binary-tree approach is bound to be more time-consuming, considering its greater number of entanglement attempts. However, the probabilistic CNOT based on cavity reflection is typically several orders of magnitude faster than the deterministic CNOT (10^1 ns vs. 10^5 ns). We therefore present both the QRAM's query time and its fidelity for both schemes in Fig. 5, assuming perfect deterministic CNOTs. We also consider the case where the CNOT efficiency is less than unity for comparison.

Before moving on to the noise simulations, we delve into the query times. It is not obvious that now the query times scale logarithmically (or even poly-logarithmically) since the efficiency of the distributed CNOTs can increase the query times depending on the order of the linking steps. In fact, Fig. 5 already shows a non-logarithmic behavior when considering a completely probabilistic protocol. If one were to choose sequential linking steps, the query times would increase exponentially with the efficiency. By choosing the scheme demonstrated in Fig. 4, we are able to reduce this to polynomial scaling[26]. However, depending on the noise parameters, this increase in time, compared to the initial two-step scheme, might not be wanted, as we will verify next. In Fig. 5, we present the query times for different efficiencies of the distributed CNOT, under the two possible hybrid schemes, with different numbers of deterministic CNOTs placed strategically (see Fig. 4).

We start by verifying that, for a random placement of the deterministic nodes, there is no clear dependence on the number of deterministic CNOTs. The reason is that, when choosing random placements for the deterministic nodes, the best order for the linking steps immediately changes and is no longer a binary tree. There already exist algorithms[29] that use linear programming to solve an identical problem of finding the best order to attempt entanglement swapping along a chain, which is virtually identical to our problem. However, the polynomial scaling of these algorithms in terms of the number of nodes of the chain makes it unsuitable for exponentially growing chains. For the intuitive placement of the deterministic nodes, this is not the case, as choosing only the top layers of the linking tree does not change the best order to do the linking. We also consider varying the efficiency of the distribution of the Bell pairs, as shown in Supplementary Fig. 1.

Dephasing and damping errors for a TS-QRAM

We start by considering the case where there are no deterministic CNOTs and vary the dephasing and damping parameters, T_2 and T_1, respectively. Note that, as expected, the query times have increased by orders of magnitude (see in greater detail in Supplementary Fig. 2), hence the extent of decoherence in the memories. Moreover, to overcome the necessity of performing noisy deterministic CNOTs, the qubits used are now the electronic spin qubits, whose dephasing and damping times are much smaller than their nuclear counterparts, thereby limiting the fidelity of the QRAM tree state. For this reason, we analyzed a wide range of possible values for T_1^e and T_2^e: {20 ms, 200 ms, 2 s, 20 s} and {10 ms, 100 ms, 1 s, 10 s}, respectively. In this scenario of having only probabilistic distributed CNOTs, we analyze for multiple CNOT efficiencies η and T_2 values, fixing $T_1 = 2$ s, as its contribution to the error is negligible compared to the T_2. In Fig. 6, we observe infidelity values scaling exponentially with the number of qubits for a completely probabilistic execution of the hybrid

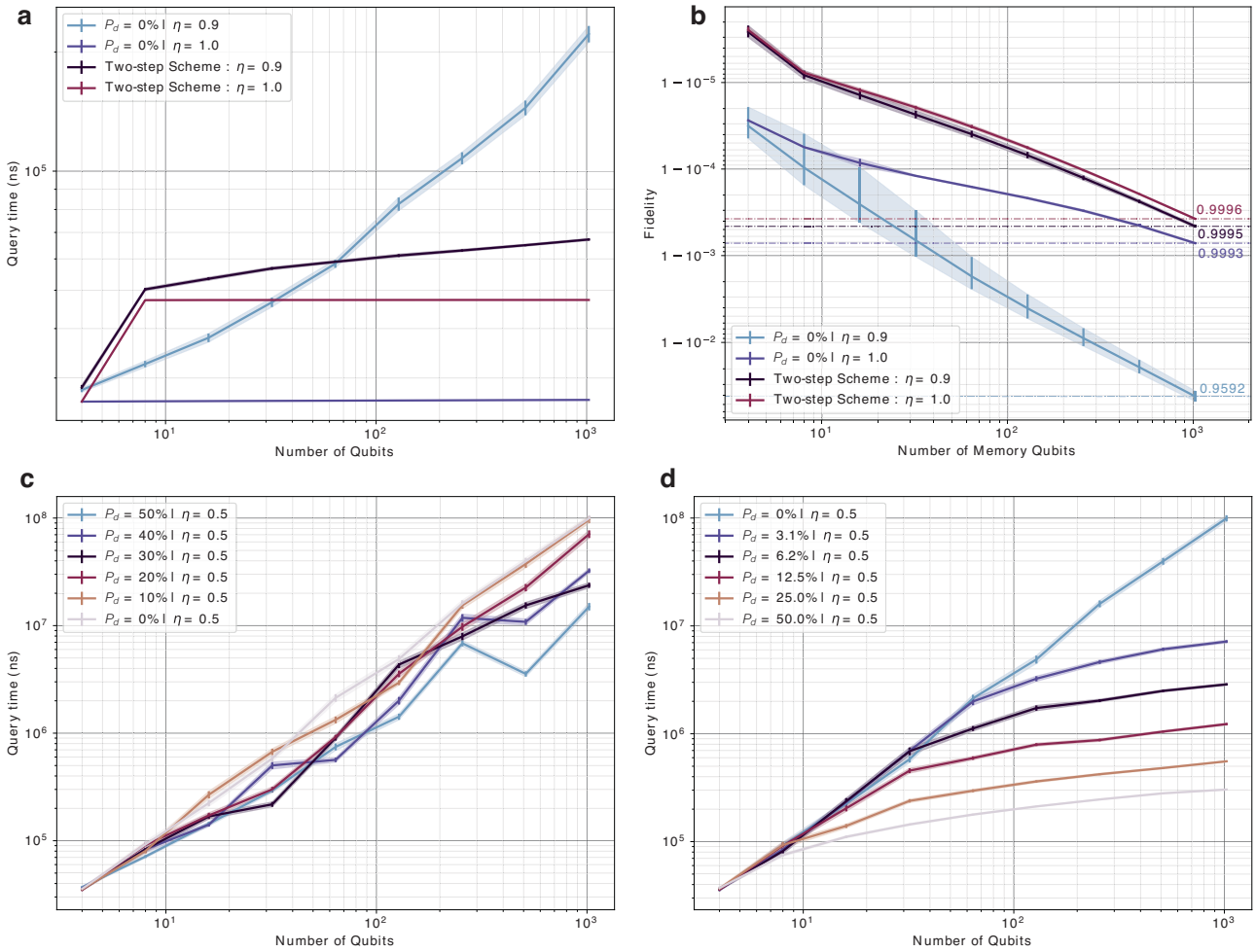

Fig. 5 TS-QRAM simulations for protocol comparison. a Query times for accessing a QRAM and **b** Fidelity of access of a QRAM for a completely probabilistic hybrid scheme ($P_d = 0\%$) and comparison under identical efficiencies of the distribution of Bell pairs for the TD-QRAM (two-step) scheme. $T_1 = 2$ s, $T_2 = 100$ ms, and $\epsilon_{CNOT} = 0$. **c** Query time scaling for randomly distributed deterministic nodes under the regular binary tree ordering and **d** and placed at higher-level steps for the linking tree, comparing between the two deterministic CNOT placement strategies for distributing the GHZ states in the TS-QRAM scheme (see Fig. 4). Notice that for the non-random placement strategy, the ratio of deterministic nodes P_d is approximately given by $P_d \approx 2^{-(\log_2(N)-D)-1}$. In both cases, the efficiency of the distribution of a Bell pair was set at $\eta = 0.5$. All the error bars over the data correspond to the error of the average value over 100 simulations of the protocol.

protocol. Only for memory coherence times on the second timescale, i.e., $T_2 = 1$ s, does the fidelity reach around 80% under a CNOT efficiency of $\eta = 0.5$. For other combinations of parameters, we refer to Supplementary Figs. 3–5.

Dephasing, damping and noisy CNOTs for TS-QRAM

Here, we explore adding some noisy deterministic CNOTs to counteract the effect of the decoherence for longer periods of time. As seen previously, the better location for these deterministic CNOTs are the nodes that perform the linking step at higher levels of the linking tree. In our simulations, we evaluate different values of the first deterministic layer $\log_2(N) - D \in \{2, 3, 4, 5, 6\}$. The results are presented in Fig. 7.

Depending on the CNOT error, the TS-QRAM scheme can surpass the fidelities of access of the TD-QRAM scheme under high enough T_2 times in the order of seconds. For other possible sets of parameters, we refer again to Supplementary Figs. 3–5.

DISCUSSION

In this article, we introduce a method to simulate large quantum networks in an open system model. Specifically, this approach enables us to model networks comprising hundreds of stationary qubits by modeling decoherence processes as noisy channels with spin-dephasing errors, spin-flip errors, and noisy CNOT gates. When applied to the challenging but important problem of network-based QRAM, we find that the qubit depth of memory calls in the recently proposed TD-QRAM architecture becomes limited by CNOT errors. To overcome this bottleneck, we propose a modified network-enabled QRAM in which the noisy deterministic gates of ref. [20] are replaced by *heralded probabilistic* CNOT gates, which can sharply reduce gate errors. This scheme, TS-QRAM, trades increased query time for improved memory access fidelity and/or memory depth. The TS-QRAM protocol makes use of already demonstrated elements (see Supplementary Table I), suggesting the viability of near-term demonstrations in platforms of solid-state color centers as well as potentially other atomic memory modalities.

An outstanding problem relates to the compounding loss of photonic qubits with increasing memory depth. Since teleportation-based QRAM[5] has shown that distributed quantum computers naturally map onto quantum networks, error correction schemes proposed for the former may be applied to address the issue of photon loss for the latter. Approaches include (1) photonic forward error correction using, for example, 2D photonic cluster

states[33–38] and (2) error-corrected cluster states[39–41]. We leave the exploration of error correction schemes in the context of QRAM for future studies.

METHODS

Discrete-time-event-based simulations with NetSquid

Given the complexity of a quantum network and its formulations, a tool such as NetSquid[28] is essential to simulating a QRAM. NetSquid is capable of defining intricate discrete-time-event-based protocols, with a number of steps and operations that are executed conditioned on the signaling and heralding of prior processes. Furthermore, NetSquid can simulate quantum circuits, providing methods for (1) stabilizer circuits, with simpler and faster execution, of complexity $\mathcal{O}(m^2)$, where m is the number of qubits; (2) graph states formalism, with possibly even faster execution, in $\mathcal{O}(d^2)$, where $\log m < d < m$ and m is again the number of qubits; (3) density matrix formalism, which is slower in execution, in $\mathcal{O}(2^{3m})$; (4) sparse density matrix formalism that relies on sparse matrix codes to speed up the execution.

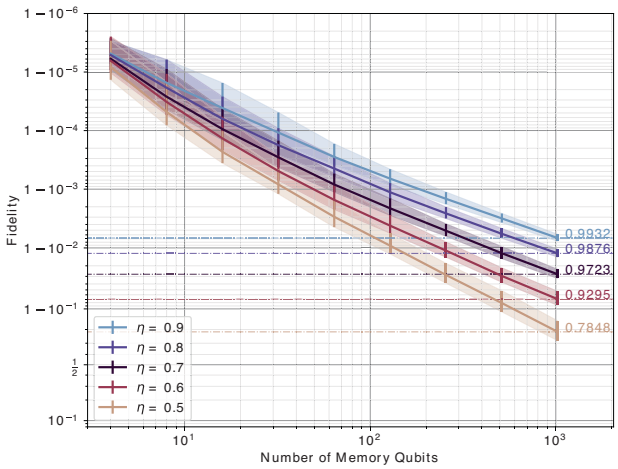

Fig. 6 Fidelity scaling for a dephasing time $T_2 = 1$ s and amplitude-damping time $T_1 = 2$ s. The simulations are for the completely probabilistic execution of the linking step ($P_d = 0\%$), meaning there are no deterministic CNOTs being executed to create the GHZ states within each layer of the QRAM. We present different simulations for several possible values for the efficiency of each distributed CNOT (i.e., the probability of success of each of the distributed CNOT), namely $\eta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. All the error bars over the data correspond to the error of the average value over 100 simulations of the protocol.

Instead of using the density matrix formalism from NetSquid, we begin by retrieving all the noise information in each step of the protocol from a *noiseless* discrete-time event simulation, resorting to the stabilizer formalism in polynomial time. The noise information is constituted by the time qubits spent decohering, together with the information about the channels of decoherence that would have been applied in real noisy simulations, both for waiting times and gate errors. We then incorporate the extracted information to estimate the effects of decoherence at each step *a posteriori*. With this information at hand, we have access to the time-evolved state of the QRAM tree at *all* steps of the protocol, which allows us to reconstruct the noise that would have been applied in the system in a noisy simulation. To reconstruct the density matrix, we find the analytical expressions for the density matrices of smaller parts of the system and how they evolve after the required operations under a set of noise channels. We express these as fundamental building blocks in terms of noise parameters, namely the probability of error and time of decoherence. This is what allows us to postpone the noise calculations to the end of the simulation without losing the effects of the natural stochastic behavior of the protocol. In a way, this can be understood as pre-compiling the error effect on the intermediary states of the protocol to shortcome the exponential complexity of calculating the density matrix at each time step using a quantum simulator.

In the rest of the "Methods" section, we formalize these elementary building blocks for the operations required to create these GHZ states across each physical layer in the QRAM and explain how different types of noise affect each of the intermediate steps, allowing for a reconstruction of the density matrix. We analyzed dephasing, damping and depolarizing channels, and we believe other noise models could be added in a similar manner. The result is an explicit description of the final state of the QRAM access tree state prior to the execution of the teleportation protocol. The error in the state of the access tree encompasses the majority of all the errors of accessing the QRAM, as the number of steps and operations made *after* creating the GHZ state grows with the total number of memory addresses *logarithmically*, whereas the process of generating the GHZ state requires a number of operations linear with the number of memories.

Elementary building blocks for TD-QRAM

The protocol for generating GHZ states across each layer consists of two steps (see Fig. 8):

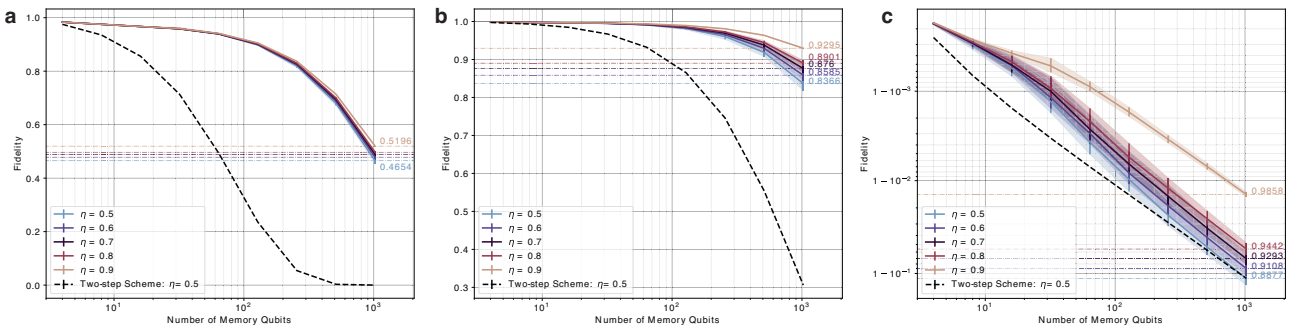

Fig. 7 Fidelity scaling for a dephasing time $T_2 = 1$ s, amplitude-damping time of $T_1 = 2$ s and a varying CNOT error. a $\epsilon_{CNOT} = 10^{-2}$. **b** $\epsilon_{CNOT} = 10^{-3}$. **c** $\epsilon_{CNOT} = 10^{-4}$. The simulations are in the hybrid regime, with 6.2% of deterministic nodes. We present different simulations for several possible values for the efficiency of each distributed CNOT (i.e., the probability of success of each of the distributed CNOT), namely $\eta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. In dashed black lines, we also plot the values for the two-step scheme for the case with $\eta = 0.5$ and with the same T_1, T_2 and ϵ_{CNOT} as the hybrid scheme. All the error bars over the data correspond to the error of the average value over 100 simulations of the protocol.

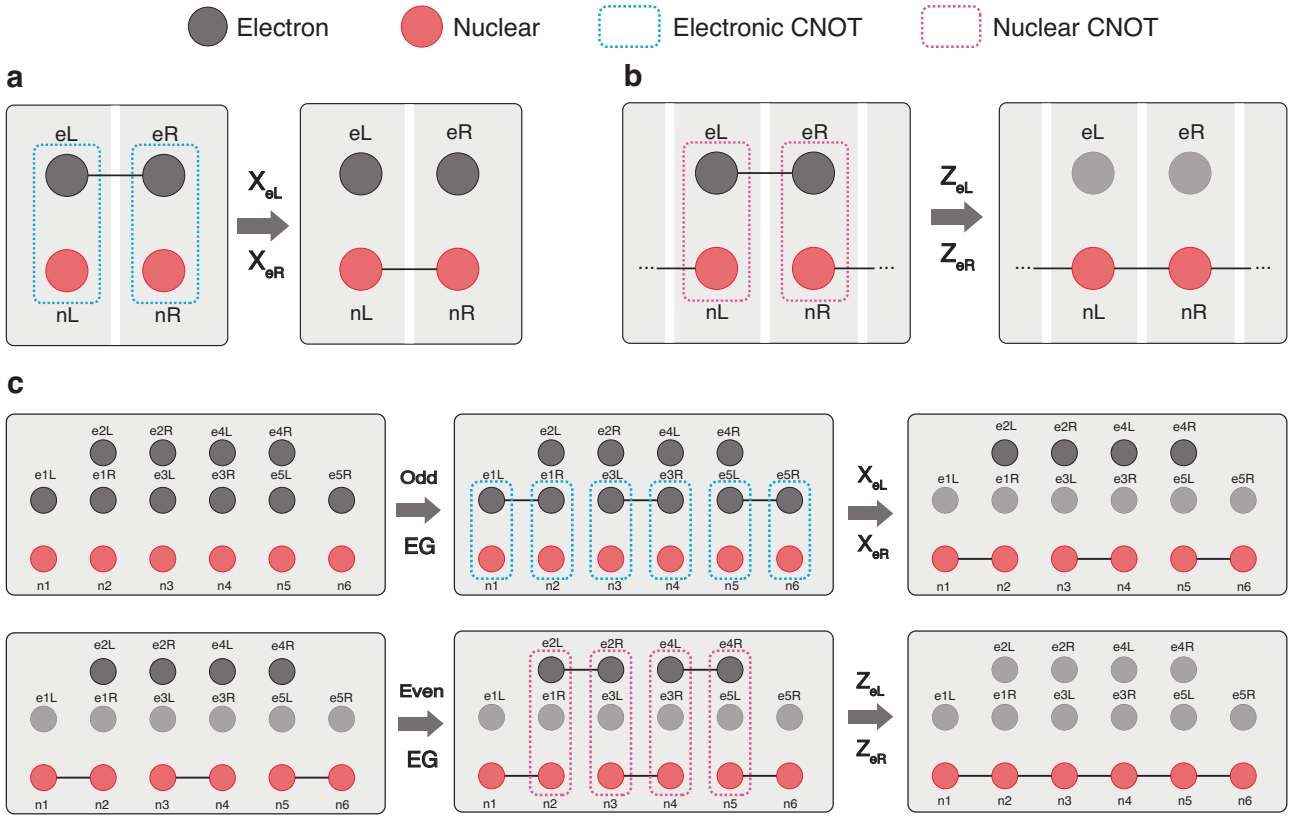

Fig. 8 Building blocks for creating the GHZ state for a chain of 6-qubits excluding the final step of entangling with the quantum computer. a Creating an EPR pair between two electrons, then transferring the entanglement to the nuclear spin qubits. **b** Linking two GHZ states through an entangled pair and a set of operations and measurements. **c** EPR pair creation along the odd-indexed links and transfer followed by linking of pairs by pre-sharing an EPR along the even-indexed links. Note the index of a link is with respect to the left node numbering.

1. Generating entanglement between the odd-indexed links. This entails first distributing photon-mediated heralded entanglement between the electrons, with a certain efficiency η, followed by *electronic* CNOTs being applied with the electron qubit acting as the control and the nuclear qubits as the target. Finally, a measurement of the electron spin in the X basis, with posterior corrections sent to the nuclear qubits.
2. The second step links the entangled pairs, creating a larger GHZ state distributed across each layer. This starts off by generating heralded entanglement, with the same efficiency η, between the even-indexed links, followed by applying *nuclear* CNOTs, where now the control is the nuclear qubit and the target is the electronic qubit. We then make consequent measurements on the Z basis on both electronic qubits, followed by appropriate Pauli corrections.

After the aforementioned steps, each physical layer hosts a GHZ state shared among all the nodes. Subsequently, each physical layer extends its (2^{k-1})-GHZ state into a $(2^{k-1}+1)$-GHZ state by sharing an additional entangled pair between an outermost node in the QRAM layer and the quantum computer, which holds the address state. After performing a Bell state measurement and corresponding corrections, the address state is teleported to the QRAM. Lastly, the memories can be accessed in superposition to complete the QRAM protocol.

EPR pair creation and transferring for TD-QRAM

The first step to creating a GHZ state across each layer is to share entanglement between neighboring nodes. EPR pairs are created by performing a distributed CNOT gate between these nodes' electronic spin qubits, mediated by a photon. Despite the process being probabilistic with an efficiency dependent on the experimental implementation, it is a *heralded* entanglement. Hence, the presence (absence) of photon detection informs the success (failure) of the entangling attempt. After this CNOT is applied between the electronic spins, an EPR pair is created and transferred to the nuclear spins in each node via a deterministic *electronic* CNOT.

Given the different operations and various types of qubits involved, we introduce noise sources in the system to estimate the protocol's fidelity. We consider amplitude damping, dephasing, and CNOT gate errors for both electronic and nuclear spin qubits.

In this step, illustrated in Fig. 8, the following takes place:

1. eL and eR decohere for a duration of time t_{eL} and t_{eR}, respectively;
2. An *electronic* CNOT is applied between eL and nL, with an error probability of p_e,
3. An *electronic* CNOT is applied between eR and nR with an error probability of p_e,
4. nL and nR decohere for a duration of time t_{nL} and t_{nR}, respectively.

Hence, these parameters, plus the parameters associated with the physical systems, namely the T_1 and T_2 times, govern the final form of the entangled pairs. Using the notation $\epsilon(\sigma) = 1 - e^{-\sigma}$ and $\overline{\epsilon}(\sigma) = 1 - \epsilon(\sigma) = e^{-\sigma}$ for parameters that are functions of other physical parameters, namely the elapsed times and coherence times. We will use ϵ for parameters that go to zero in the absence of noise, as is the case for $\epsilon(\cdot)$ and p_n. We also further assume

$\epsilon \ll 1$. We then apply the following sequence of noise channels (check Supplementary Methods for details on the parameters):

1. Apply a Dephasing channel with probability $\epsilon(t_{eL}/T_2^e) \equiv \epsilon_{eL}^{(2)}$ and $\epsilon(t_{eR}/T_2^e) \equiv \epsilon_{eR}^{(2)}$ to electronic spin qubits eL and eR, respectively;
2. Apply an Amplitude-damping channel with probability $\epsilon(t_{eL}/T_1^e) \equiv \epsilon_{eL}^{(1)}$ and $\epsilon(t_{eR}/T_1^e) \equiv \epsilon_{eR}^{(1)}$ to electronic spin qubits eL and eR, respectively;
3. Apply Depolarizing channels with probability p_e to all qubits after applying CNOTs (modeling a noisy CNOT);
4. Apply a Dephasing channel with probability $\epsilon(t_{nL}/T_2^n) \equiv \epsilon_{nL}^{(2)}$ and $\epsilon(t_{nR}/T_2^n) \equiv \epsilon_{nR}^{(2)}$ to nuclear spin qubits nL and nR, respectively;
5. Apply an Amplitude-damping channel with probability $\epsilon(t_{nL}/T_1^n) \equiv \epsilon_{nL}^{(1)}$ and $\epsilon(t_{nR}/T_1^n) \equiv \epsilon_{nR}^{(1)}$ to nuclear spin qubits nL and nR, respectively;

The final state for each entangled pair becomes:

$$\frac{1}{2}\begin{pmatrix} 1-\mu & 0 & 0 & \nu \\ 0 & \mu & 0 & 0 \\ 0 & 0 & \mu & 0 \\ \nu & 0 & 0 & 1-\mu \end{pmatrix} \tag{3}$$

where

$$\mu = \frac{1 - f\left(\epsilon_{eL}^{(1)}, \epsilon_{eR}^{(1)}\right)(1-p_e)^2 g\left(\epsilon_{nL}^{(1)}, \epsilon_{nR}^{(1)}\right) - \epsilon_{nL}^{(1)}\epsilon_{nR}^{(1)}}{2}, \tag{4}$$

$$\nu = \bar{\epsilon}_{eL}^{(2)} \cdot \bar{\epsilon}_{eR}^{(2)} \cdot \bar{\epsilon}_{nL}^{(2)} \cdot \bar{\epsilon}_{nR}^{(2)} \cdot \sqrt{\bar{\epsilon}_{eL}^{(1)} \cdot \bar{\epsilon}_{eR}^{(1)} \cdot \bar{\epsilon}_{nL}^{(1)} \cdot \bar{\epsilon}_{nR}^{(1)}} \cdot (1-p_e)^4 \tag{5}$$

and

$$\begin{aligned} f(\epsilon_1, \epsilon_2) &= 1 - \epsilon_1 - \epsilon_2 + 2\epsilon_1\epsilon_2, \\ g(\epsilon_1, \epsilon_2) &= (1-\epsilon_1)(1-\epsilon_2) \end{aligned} \tag{6}$$

For intuition regarding the ϵ function, consider the following two limits: (1) $\sigma \to 0$ in the noiseless regime where the memory coherence time goes to infinity (no decoherence) and (2) $\sigma \to \infty$ where there only exists noise and all the information is scrambled. In these limits, we retrieve: $\lim_{\sigma\to 0} \epsilon(\sigma) = 0, \lim_{\sigma\to\infty} \epsilon(\sigma) = 1, \lim_{\sigma\to 0} \bar{\epsilon}(\sigma) = 1$ and $\lim_{\sigma\to\infty} \bar{\epsilon}(\sigma) = 0$.

Linking of Bell pairs for TD-QRAM

The following step is crucial to extending entanglement from bipartite to GHZ states across the entire physical layer of the QRAM. It relies on using an entangled pair to combine two GHZ states of smaller sizes into a larger GHZ state, whose number of qubits equals the sum of each of the elementary GHZ states (i.e., n_1-GHZ linked with a n_2-GHZ becomes a (n_1+n_2)-GHZ state).

In this step, we account for decoherence before applying CNOTs, therefore entering the previous expressions for the form of each pair. The decoherence to be analyzed in this step stems from:

1. A nuclear CNOT gate on eL and nL with probability of error p_n;
2. A nuclear CNOT gate on eR and nR with probability of error p_n;
3. Nuclear qubits nL and nR decohere after a CNOT for t'.

Additionally, for each block, we analyze the impact of decoherence by applying the following noise channels:

1. Apply depolarizing channels with probability p_n to all qubits (eL, eR, nL, nR) after applying CNOTs (modeling a noisy CNOT);
2. Apply a dephasing channel with probability $\epsilon(t'_{nL}/T_2^n) \equiv \epsilon_{nR}^{\prime(2)}$ and $\epsilon(t'_{nR}/T_2^n) \equiv \epsilon_{nL}^{\prime(2)}$ to nuclear spin qubits nL and nR, respectively;
3. Apply an amplitude-damping channel with probability $\epsilon(t'_{nL}/T_1^n) \equiv \epsilon_{nR}^{\prime(1)}$ and $\epsilon(t'_{nR}/T_1^n) \equiv \epsilon_{nL}^{\prime(1)}$ to nuclear spin qubits nL and nR, respectively;

Note that all the following calculations are now lower bounds for the fidelity, as the calculation of the full analytical expressions grows exponentially with the number of qubits. Because of this, we keep only the terms up to $\mathcal{O}(\epsilon)$. In Supplementary Methods, we detail and test the validity of our approximations.

The final GHZ state in each layer is described by a matrix with the following form:

$$\frac{1}{2}\begin{pmatrix} \rho_{00} & 0 & \dots & 0 & \rho_{01} \\ 0 & \epsilon & \dots & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ 0 & 0 & \dots & \epsilon & 0 \\ \rho_{10} & 0 & \dots & 0 & \rho_{11} \end{pmatrix} \tag{7}$$

where all ϵ terms are of at least order $\mathcal{O}(\epsilon)$ and do not contribute to infidelity, as they are orthogonal to the GHZ state.

The diagonal elements that we consider are only the first and the last, as the remaining ones have at least $\mathcal{O}(\epsilon)$ and, when expanding to a larger GHZ state, contribute in $\mathcal{O}(\epsilon^2)$ or higher orders, hence negligibly affecting the fidelity.

Let us first consider the form of the state after executing the linking protocol in a noiseless manner, with previously noisy states, as the ones that result from the entangling step given by Eq. (3). Starting with the simple case of a 4-qubit GHZ state built from three states of the form of Eq. (3), with parameters $(\mu_j, \nu_j), j = 1, 2, 3$ respectively, the final matrix is:

$$\frac{1}{2}\begin{pmatrix} \bar{\mu}_1\bar{\mu}_2\bar{\mu}_3 & 0 & \dots & 0 & \nu_1\nu_2\nu_3 \\ 0 & \bar{\mu}_1\mu_2\bar{\mu}_3 & \dots & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ 0 & 0 & \dots & \mu_1\mu_2\bar{\mu}_3 & 0 \\ \nu_1\nu_2\nu_3 & 0 & \dots & 0 & \bar{\mu}_1\bar{\mu}_2\bar{\mu}_3 \end{pmatrix} \tag{8}$$

where we, again, denote a bar over a variable as 1 minus itself, $\bar{\mu}_i \equiv 1 - \mu_i$. Note that each of the μ_i comes from one of the pairs used to create the GHZ state, as these pairs are solely described by two numbers (μ_i, ν_i) (see Eq. (3)). There exists a rule for each entry in the diagonal, which we detail in Supplementary Methods, and the same rule holds for any number of qubits of the final state. The GHZ diagonal entries then become:

$$\rho_{00} = \rho_{11} = \bar{\mu}_1\bar{\mu}_2\bar{\mu}_3 \tag{9}$$

Now, adding the effect of the noisy CNOTs on the state, we calculate the diagonal terms that are shown to be identical, given by:

$$\begin{aligned} \rho'_{00} &= \rho'_{11} = \left(1 - \tfrac{p_n}{2}\right)^2(1-\mu_1)(1-\mu_2)(1-\mu_3) \\ &\quad - p_n\left(1 - \tfrac{p_n}{2}\right)\left(1 - \mu_1 - \tfrac{p_n}{2}\right)\left(1 - \mu_2 - \tfrac{p_n}{2}\right)(1-2\mu_3) \\ &\quad + \mathcal{O}(\epsilon^3) \\ &= \left[\left(1 - \tfrac{p_n}{2}\right)^2 - p_n\left(1 - \tfrac{p_n}{2}\right)\right](1-\mu_1)(1-\mu_2)(1-\mu_3) \\ &\quad + \mathcal{O}(\epsilon^2) \\ &\equiv h(p_n)(1-\mu_1)(1-\mu_2)(1-\mu_3) + \mathcal{O}(\epsilon^2) \end{aligned} \tag{10}$$

where we recall that every term with $p_n, \mu_i \ll 1$ converges to zero in the noiseless limit. For the other diagonal entries, we multiply them by $h(p_n)$.

Finally, incorporating memory decoherence after CNOTs, we perform another approximation. For the diagonal terms, only the

damping channel plays a role. The first and last entries of the diagonal become:

$$\begin{aligned}\rho''_{00} &= h(p_n)\left(\bar{\mu}_1\bar{\mu}_2\bar{\mu}_3 + \epsilon'^{(1)}_{nL}\mu_1\mu_2\bar{\mu}_3 + \epsilon'^{(1)}_{nR}\bar{\mu}_1\mu_2\mu_3\right) + \mathcal{O}(\epsilon^4)\\ \rho''_{11} &= \rho_{11}\left(1-\epsilon'^{(1)}_{nL}\right)\left(1-\epsilon'^{(1)}_{nR}\right)\end{aligned} \tag{11}$$

In this approximation, the extra terms that appear for the first entry are already of order $\mathcal{O}(\epsilon^3)$ and could be neglected.

Lastly, we compute the off-diagonal terms by multiplying every contribution from each noise channel applied in the correct manner. The expression is given by:

$$\rho''_{01} = \rho''_{10} = \nu_1\nu_2\nu_3 \cdot \bar{\epsilon}'^{(2)}_{nL}\bar{\epsilon}'^{(2)}_{nR}\sqrt{\bar{\epsilon}'^{(2)}_{nL}\bar{\epsilon}'^{(2)}_{nR}} \cdot (1-p_n)^2 f(p_n,p_n) \tag{12}$$

When extending the linking protocol to a larger number of qubits, the expressions maintain their form. We only need to add all the terms in a similar manner as in the case of the 4-GHZ state. The complete analysis is detailed in Supplementary Methods.

Thus far, we show how three types of noise (one for the two-qubit operations and two for individual memories) influence the final state of the GHZ states generated across each physical layer of the QRAM access tree. Note that we always present the full expressions accounting for all the noise channels. In fact, if we include a specific noise channel or a subset of what we have considered, we may simply set the parameters corresponding to other noises to zero. For example, it is straightforward to verify that setting p_e and p_n to zero and T_1 to infinity recovers the case for only having dephasing, thus affecting only the off-diagonal terms. The same is valid for all the other noises.

Modified protocol and building blocks for TS-QRAM

We now consider an alternative architecture that enables different subsets of each layer to create GHZ states independently. As illustrated in Fig. 9, this architecture assumes two electron spins and one nuclear spin (instead of each node of the QRAM having an electronic spin and a nuclear spin assumed for the TD-QRAM). As we will show, this architecture still retains similar building blocks as the aforementioned TD-QRAM protocol.

EPR creation for TS-QRAM

As in the non-hybrid version of the protocol, the first step to creating an EPR pair between two physically separated electronic spins is sending a photon that interacts with them sequentially. A subsequent measurement heralds the successful production of a spin-spin EPR pair. Notably, there are no deterministic CNOTs applied to transfer the qubit states onto the nuclear spins, as we only work with the electron spins at this stage.

The final state shared between the electronic spins is the one of Eq. (3) in the limit of the absence of *electronic* CNOT error ($p_e \to 0$) and altering the memory decoherence noise from nuclear to electronic ($\epsilon_{nL} \to \epsilon'_{eL}$ and $\epsilon_{nR} \to \epsilon'_{eR}$):

$$\frac{1}{2}\begin{pmatrix} 1-\mu & 0 & 0 & \nu \\ 0 & \mu & 0 & 0 \\ 0 & 0 & \mu & 0 \\ \nu & 0 & 0 & 1-\mu \end{pmatrix} \tag{13}$$

where

$$\mu = \frac{1 - f\left(\epsilon^{(1)}_{eL}, \epsilon^{(1)}_{eR}\right) g\left(\epsilon'^{(1)}_{eL}, \epsilon'^{(1)}_{eR}\right) + \epsilon'^{(1)}_{eL}\epsilon'^{(1)}_{eR}}{2}, \tag{14}$$

$$\nu = \bar{\epsilon}^{(2)}_{eL} \cdot \bar{\epsilon}^{(2)}_{eR} \cdot \bar{\epsilon}'^{(2)}_{eL} \cdot \bar{\epsilon}'^{(2)}_{eR} \cdot \sqrt{\bar{\epsilon}^{(1)}_{eL} \cdot \bar{\epsilon}^{(1)}_{eR} \cdot \bar{\epsilon}'^{(1)}_{eL} \cdot \bar{\epsilon}'^{(1)}_{eR}} \tag{15}$$

and again,

$$\begin{aligned} f(\epsilon_1,\epsilon_2) &= 1-\epsilon_1-\epsilon_2+2\epsilon_1\epsilon_2,\\ g(\epsilon_1,\epsilon_2) &= (1-\epsilon_1)(1-\epsilon_2)\end{aligned} \tag{16}$$

where we use the same abbreviation $\bar{\epsilon}(\sigma) = 1-\epsilon(\sigma) = e^{-\sigma}$.

Linking pairs in the probabilistic scenario for a TS-QRAM

TS-QRAM differs from TD-QRAM in that the operation of linking pairs has a non-unity probability of succeeding—let us call this probability p_{CNOT}. Moreover, it is executed in a similar way as that of creating an EPR pair:

1. Interact photon γ with the left electronic spin qubit eL, executing a local CNOT,
2. Send the single photon γ to the right cavity,
3. Interact the photon γ with the right electronic spin qubit eR, executing a local CNOT,

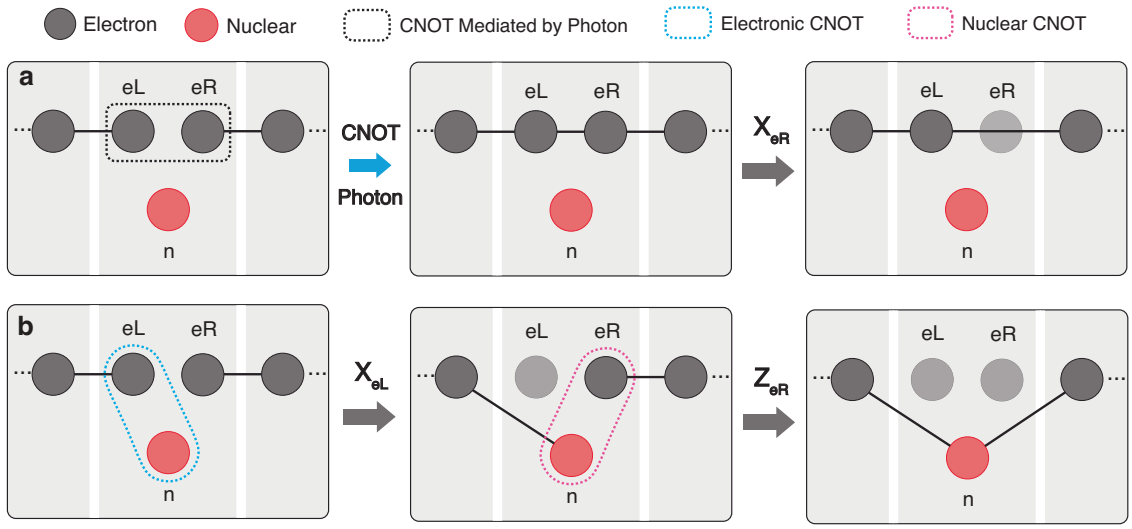

Fig. 9 Possible protocols for linking smaller GHZ states into a larger GHZ state both probabilistically and deterministically. a Probabilistic CNOT Protocol mediated by a photon. **b** Deterministic CNOT Protocol, consisting of a nuclear CNOT between the nuclear spin ancilla and the left electronic spin qubit.

4. Measure the photon γ,
5. Measure the right (or left) electronic spin in X.

Importantly, both cavities *belong to the same node* in this step. This results in a controlled gate applied between the right and left electronic spin qubits. Unlike before, it is still necessary to measure one of the nodes' electronic spin qubits, as the state has twice the number of qubits as the final state (we chose to measure the right electron, but one could choose to keep the right and measure the left instead; the choice is arbitrary and translates to the same practical outcome). This measurement should be on the X basis in order to not destroy the entanglement shared among all the qubits and rendering the state useless. Moreover, a correction must be made depending on the outcome of the measurement of the electronic spin qubit and the photonic qubit.

Afterward, the GHZ states shared between the left and right nodes are linked into a larger GHZ state via an intermediary node. Inside this intermediary node, its left electronic spin merges into the larger GHZ state. We again take into account the previous calculations for detailing the density matrix of the final state. The decoherence sources are now only provenient from the memories of where each qubit is being held (which we chose to be the left cavity of the node). As we used near-perfect probabilistic CNOTs mediated by a photon, only its memory affects the state fidelity. Thus, for the remainder of the protocol, we:

1. Apply a dephasing channel with probability $\epsilon(t''_{\mathrm{eL}}/T_2^{\mathrm{e}}) \equiv \epsilon_{\mathrm{eL}}^{\prime\prime(2)}$ to electronic spin qubit eL;
2. Apply an amplitude-damping channel with probability $\epsilon(t''_{\mathrm{eL}}/T_1^{\mathrm{e}}) \equiv \epsilon_{\mathrm{eL}}^{\prime\prime(1)}$ to electronic spin qubit eL.

Importantly, the following calculations are again lower-bound approximations for fidelity. Performing the calculations for a simple link of two entangled pairs described by Eq. (13), with parameters $(\mu_j, \nu_j), j = 1, 2$, the final density matrix of the 3-GHZ state, prior to any memory decoherence, is:

$$\frac{1}{2}\begin{pmatrix} \overline{\mu}_1\overline{\mu}_2 & 0 & \dots & 0 & \nu_1\nu_2 \\ 0 & \overline{\mu}_1\mu_2 & \dots & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ 0 & 0 & \dots & \overline{\mu}_1\mu_2 & 0 \\ \nu_1\nu_2 & 0 & \dots & 0 & \overline{\mu}_1\overline{\mu}_2 \end{pmatrix} \tag{17}$$

Adding the memory decoherence accounting for both dephasing and amplitude-damping leads to a matrix similar in form to one shown in Eq. (7), except with entries changing to:

$$\begin{aligned} \rho'_{00} &= \overline{\mu}_1\overline{\mu}_2 + \epsilon_{\mathrm{eL}}^{\prime\prime(1)}\mu_1\mu_2 + \mathcal{O}(\epsilon^4) \\ \rho'_{11} &= \overline{\mu}_1\overline{\mu}_2\left(1 - \epsilon_{\mathrm{eL}}^{\prime\prime(1)}\right) \\ \rho'_{10} &= \rho'_{01} = \nu_1\nu_2\overline{\epsilon}_{\mathrm{eL}}^{\prime\prime(2)}\sqrt{1 - \epsilon_{\mathrm{eL}}^{\prime\prime(1)}} \end{aligned} \tag{18}$$

Linking pairs in the deterministic scenario for TS-QRAM

We next describe a scheme for deterministic CNOT gates. Since the composition of each node is now different, the operations one needs to execute to deterministically link smaller GHZ states into larger GHZ states changes as well:

1. Apply a deterministic *electronic* CNOT controlled by the left electronic spin qubit eL and targeted at the nuclear spin qubit n,
2. Measure the left electronic spin qubit eL in X,
3. Apply a deterministic nuclear CNOT controlled by the nuclear spin qubit n and targeted at the right electronic spin qubit eR,
4. Measure the right electronic spin qubit eR in Z.

Measurement-conditioned corrections result in a GHZ state consisting of the nuclear spin and the remaining electronic spin qubits. Notice that by not involving the photon-mediated CNOT, this has been done in a deterministic fashion. In this case, we must consider additional errors, namely those that arise from using deterministic *electronic* and *nuclear* CNOTs. The sequence of noise channels becomes:

1. Apply depolarizing channels with probability p_{e} to the electronic spin qubit eL and to the nuclear spin qubit n,
2. Apply depolarizing channels with probability p_{n} to the nuclear spin qubit n and to the electronic spin qubit eR,
3. Apply a dephasing channel with probability $\epsilon(t'_n/T_2^{\mathrm{n}}) \equiv \epsilon_n^{\prime(2)}$ to nuclear spin qubit n,
4. Apply an amplitude-damping channel with probability $\epsilon(t'_n/T_1^{\mathrm{n}}) \equiv \epsilon_n^{\prime(1)}$ to nuclear spin qubit n.

From here, we calculate the final state's density matrix. Performing the calculations for the same simple link of two entangled pairs described by Eq. (13), with parameters $(\mu_j, \nu_j), j = 1, 2$, the final density matrix of the 3-GHZ state, prior to any memory decoherence and without any CNOT errors is the same as Eq. (17). Adding the effect of the CNOTs leads to:

$$\begin{aligned} \rho'_{00} &= \rho'_{11} = (1-p)^2\overline{\mu}_1\overline{\mu}_2 + \tfrac{p}{2}\left(1 - \tfrac{p}{2}\right) \\ \rho'_{10} &= \rho'_{01} = \nu_1\nu_2(1-p)^3\left(1 - \tfrac{p}{2}\right) \end{aligned} \tag{19}$$

where we set $p_{\mathrm{e}} = p_{\mathrm{n}} \equiv p$. In fact, all diagonal entries can be decomposed into terms of the form $(1-p)^2\,\mathrm{diag}(\rho) + \mathbb{1}p/2(1-p/2)$. Using this fact, we incorporate the posterior amplitude-damping noise channels:

$$\begin{aligned} \rho''_{00} &= \tilde{h}(p)\cdot\left(\overline{\mu}_1\overline{\mu}_2 + \epsilon_{\mathrm{eL}}^{\prime(1)}\mu_1\mu_2\right) + \mathcal{O}(\epsilon^3) \\ \rho''_{11} &= \tilde{h}(p)\cdot\overline{\mu}_1\overline{\mu}_2\left(1 - \epsilon_{\mathrm{eL}}^{\prime(1)}\right) + \mathcal{O}(\epsilon^3) \\ \rho''_{10} &= \rho''_{01} = \nu_1\nu_2(1-p)^3\left(1 - \tfrac{p}{2}\right)\sqrt{1 - \epsilon_{\mathrm{eL}}^{\prime(1)}} \end{aligned} \tag{20}$$

where $\tilde{h}(p) = (1-p)^2 + p/2(1-p/2)$. In the Supplementary Methods, we present the derivation for a chain of an arbitrary number of channels, as well as proof of the validity of our approximations.

Validity of the results

In this section, we discuss our methods and provide empirical proof for our noise analysis robustness. In our work, we have considered an analysis of a discrete protocol, where each gate takes a fixed amount of time, and qubits stay in memory waiting for instructions from the protocol. Given the discreteness of the problem, solving the master equations in each of the specific time periods where the noise actually happens is equivalent to applying the corresponding noise channels.

To support our claims and demonstrate the equivalence of using the full density matrix and our methods of postponing the noise analysis to the end of the protocol simulation, we present one additional figure. This figure shows the comparison of the fidelity of the GHZ states distributed at each of the layers of the QRAM for a small QRAM (5 layers, ~36 qubits) obtained from our simulation methods and those derived using the density matrix formalism. We present the comparison for simulations for the TD-QRAM, as for TS-QRAM, we use the same methods, just under different assumptions over which noise channels are applied. Let F_{DM} be the fidelity calculated using the full density matrix and F_{S} be the one from our simulation methods. The figure showcases how the difference of the fidelities calculated using the density matrix and our methods evolve for different layers (with increasing amounts of

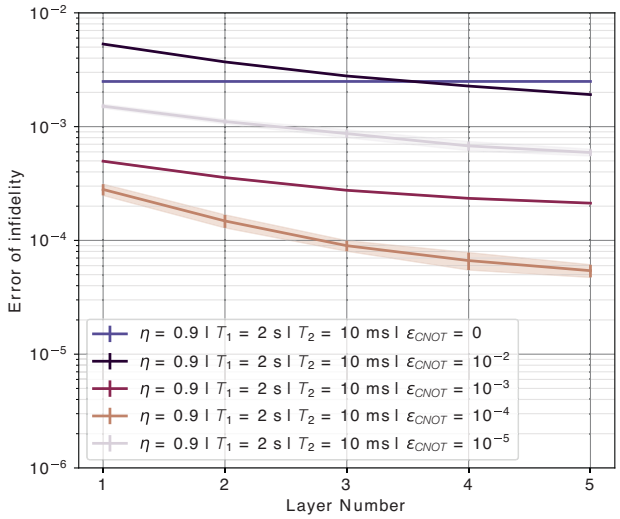

Fig. 10 Comparison between the infidelity calculated using a full density matrix simulation and our methods. Simulations made under a TD-QRAM protocol execution up to 5 layers for different combinations of error parameters. All the error bars over the data correspond to the error of the average value over 100 simulations of the protocol.

qubits), weighted by the overall error (or infidelity of the state), i.e., $|F_{DM} - F_S|/(1 - F_{DM})$.

One can observe in Fig. 10 that the error is always less than 1%, showing a trend of either maintaining or decreasing as the size of the system increases. This provides empirical proof of the power of our methods across the different scenarios.

DATA AVAILABILITY

The data supporting the results in this work are available from L.B. at https://github.com/luisbugalho/HeraldedQRAM.

CODE AVAILABILITY

The code supporting the results in this work is available from L.B. at https://github.com/luisbugalho/HeraldedQRAM.

Received: 10 February 2023; Accepted: 28 September 2023;
Published online: 20 October 2023

REFERENCES

1. Biamonte, J. et al. Quantum machine learning. *Nature* **549**, 195–202 (2017).
2. Harrow, A. W., Hassidim, A. & Lloyd, S. Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.* **103**, 150502 (2009).
3. Kiani, B. T., Villanyi, A. & Lloyd, S. Quantum medical imaging algorithms. Preprint at https://arxiv.org/abs/2004.02036 (2020).
4. Grover, L. K. A fast quantum mechanical algorithm for database search. In *Proc. Twenty-Eighth Annual ACM Symposium on Theory of Computing, STOC '96*, 212–219 (Association for Computing Machinery, 1996).
5. Chen, K. C., Dai, W., Errando-Herranz, C., Lloyd, S. & Englund, D. Scalable and high-fidelity quantum random access memory in spin-photon networks. *PRX Quantum* **2**, 030319 (2021).
6. Jaeger, R. & Blalock, T. *Microelectronic Circuit Design* 4th edn (McGraw-Hill Education, 1997).
7. Giovannetti, V., Lloyd, S. & MacCone, L. Architectures for a quantum random access memory. *Phys. Rev. A* **78**, 052310 (2008).
8. Hann, C. T., Lee, G., Girvin, S. & Jiang, L. Resilience of quantum random access memory to generic noise. *PRX Quantum* **2**, 020311 (2021).
9. Giovannetti, V., Lloyd, S. & MacCone, L. Quantum random access memory. *Phys. Rev. Lett.* **100**, 160501 (2008).
10. Tóth, G. Multipartite entanglement and high-precision metrology. *Phys. Rev. A* **85**, 022322 (2012).
11. Sidhu, J. S. & Kok, P. A geometric perspective on quantum parameter estimation. *AVS Quantum Sci.* **2**, 014701 (2019).
12. Murta, G., Grasselli, F., Kampermann, H. & Bruß, D. Quantum conference key agreement: a review. *Adv. Quantum Technol.* **3**, 2000025 (2020).
13. Wehner, S., Elkouss, D. & Hanson, R. Quantum internet: a vision for the road ahead. *Science* **362**, eaam9288 (2018).
14. Alshowkan, M. et al. Reconfigurable quantum local area network over deployed fiber. *PRX Quantum* **2**, 040304 (2021).
15. Van den Nest, M. Simulating quantum computers with probabilistic methods. *Quantum Inf. Comput.* **11**, 784–812 (2011).
16. Jozsa, R. & van den Nest, M. Classical simulation complexity of extended Clifford circuits. *Quantum Inf. Comput.* **14**, 633–648 (2014).
17. Takahashi, Y., Takeuchi, Y. & Tani, S. Classically simulating quantum circuits with local depolarizing noise. *Theor. Comput. Sci.* **893**, 117–132 (2021).
18. Bhaskar, M. K. et al. Experimental demonstration of memory-enhanced quantum communication. *Nature* **580**, 60–64 (2020).
19. Wan, N. H. et al. Large-scale integration of artificial atoms in hybrid photonic circuits. *Nature* **583**, 226–231 (2020).
20. Nguyen, C. T. et al. An integrated nanophotonic quantum register based on silicon-vacancy spins in diamond. *Phys. Rev. B* **100**, 165428 (2019).
21. Bradley, C. E. et al. Robust quantum-network memory based on spin qubits in isotopically engineered diamond. *NPJ Quantum Inf.* **8**, 122 (2022).
22. Chen, K. C., Bersin, E. & Englund, D. A polarization encoded photon-to-spin interface. *NPJ Quantum Inf.* **7**, 1–6 (2021).
23. Sukachev, D. D. et al. Silicon-vacancy spin qubit in diamond: a quantum memory exceeding 10 ms with single-shot state readout. *Phys. Rev. Lett.* **119**, 223602 (2017).
24. Duan, L.-M. & Kimble, H. J. A scheme for preparation of multi-atom entanglement by detecting the cavity decay and analysis of its implementation. In *Proc. SPIE Quantum Communications and Quantum Imaging*, Vol. 5161, 40–47 (SPIE, 2004).
25. Calderon-Vargas, F. A. et al. Fast high-fidelity entangling gates for spin qubits in Si double quantum dots. *Phys. Rev. B* **100**, 035304 (2019).
26. Coopmans, T., Brand, S. & Elkouss, D. Improved analytical bounds on delivery times of long-distance entanglement. *Phys. Rev. A* **105**, 012608 (2022).
27. Duan, L.-M. & Kimble, H. J. Scalable photonic quantum computation through cavity-assisted interactions. *Phys. Rev. Lett.* **92**, 127902 (2004).
28. Coopmans, T. et al. NetSquid, a NETwork Simulator for QUantum Information using Discrete events. *Commun. Phys.* **4**, 164 (2021).
29. Dai, W., Peng, T. & Win, M. Z. Optimal remote entanglement distribution. *IEEE J. Sel. Areas Commun.* **38**, 540–556 (2020).
30. Childress, L. et al. Coherent dynamics of coupled electron and nuclear spin qubits in diamond. *Science* **314**, 281–285 (2006).
31. Findler, C., Lang, J., Osterkamp, C., Nesládek, M. & Jelezko, F. Indirect overgrowth as a synthesis route for superior diamond nano sensors. *Sci. Rep.* **10**, 22404 (2020).
32. Brand, S., Coopmans, T. & Elkouss, D. Efficient computation of the waiting time and fidelity in quantum repeater chains. *IEEE J. Sel. Areas Commun.* **38**, 619–639 (2020).
33. Pichler, H., Choi, S., Zoller, P. & Lukin, M. D. Universal photonic quantum computation via time-delayed feedback. *Proc. Natl Acad. Sci. USA* **114**, 11362–11367 (2017).
34. Larsen, M. V., Guo, X., Breum, C. R., Neergaard-Nielsen, J. S. & Andersen, U. L. Deterministic generation of a two-dimensional cluster state. *Science* **366**, 369–372 (2019).
35. Russo, A., Barnes, E. & Economou, S. E. Generation of arbitrary all-photonic graph states from quantum emitters. *New J. Phys.* **21**, 055002 (2019).
36. Pant, M., Towsley, D., Englund, D. & Guha, S. Percolation thresholds for photonic quantum computing. *Nat. Commun.* **10**, 1070 (2019).
37. Uppu, R. et al. Scalable integrated single-photon source. *Sci. Adv.* **6**, eabc8268 (2020).
38. Michaels, C. P. et al. Multidimensional cluster states using a single spin-photon interface coupled strongly to an intrinsic nuclear register. *Quantum* **5**, 565 (2021).
39. Nickerson, N. H., Fitzsimons, J. F. & Benjamin, S. C. Freely scalable quantum technologies using cells of 5-to-50 qubits with very lossy and noisy photonic links. *Phys. Rev. X* **4**, 041041 (2014).
40. Nemoto, K. et al. Photonic architecture for scalable quantum information processing in diamond. *Phys. Rev. X* **4**, 031022 (2014).
41. Choi, H., Pant, M., Guha, S. & Englund, D. Percolation-based architecture for cluster state creation using photon-mediated entanglement between atomic memories. *NPJ Quantum Inf.* **5**, 104 (2019).

ACKNOWLEDGEMENTS

L.B., E.Z.C., and Y.O. thank the support from Fundação para a Ciência e a Tecnologia (FCT, Portugal), namely through projects UIDB/04540/2020 and UIDB/50008/2020. L.B. acknowledges the support of FCT through scholarship BD/05268/2021 and of the

PEPR integrated project EPiQ ANR-22-PETQ-0007 part of Plan France 2030. E.Z.C. also acknowledges funding by FCT through project 2021.03707.CEECIND/CP1653/CT0002. W.D. is supported by the National Science Foundation to the Computing Research Association for the CIFellows 2020 Program. K.C.C. and D.E. acknowledge funding support by the National Science Foundation (NSF) Engineering Research Center for Quantum Networks (CQN), awarded under cooperative agreement number 1941583, and the ARO MURI on 'Theory and Engineering of Large-Scale Distributed Entanglement' (W911NF2110325). D.E. further acknowledges support from the NSF C-Accel program, grant number 2040695.

AUTHOR CONTRIBUTIONS

L.B., E.Z.C., W.D. and K.C. performed the simulation of the protocol and consequent analysis. All authors discussed the analysis of the data and contributed to writing or proofreading the manuscript. Y.O. and E.D. supervised the project.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41534-023-00773-x.

Correspondence and requests for materials should be addressed to Luís Bugalho.