Quantum Fan-out: Circuit Optimizations and Technology Modeling

Pranav Gokhale*, Samantha Koretsky†, Shilin Huang‡, Swarnadeep Majumder‡,
Andrew Drucker†, Kenneth R. Brown‡, Frederic T. Chong†*
*Super.tech
†University of Chicago
‡Duke University

2021 IEEE International Conference on Quantum Computing and Engineering (QCE) | 978-1-6654-1691-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/QCE52317.2021.00045

***Abstract*—Instruction scheduling is a key compiler optimization in quantum computing, just as it is for classical computing. Current schedulers optimize for data parallelism by allowing simultaneous execution of instructions, as long as their qubits do not overlap. However, on many quantum hardware platforms, instructions on overlapping qubits can be executed simultaneously through *global interactions*. For example, while fan-out in traditional quantum circuits can only be implemented sequentially when viewed at the logical level, global interactions at the physical level allow fan-out to be achieved in one step. We leverage this simultaneous fan-out primitive to optimize circuit synthesis for NISQ (Noisy Intermediate-Scale Quantum) workloads. In addition, we introduce novel quantum memory architectures based on fan-out.**

Our work also addresses hardware implementation of the fan-out primitive. We perform realistic simulations for trapped ion quantum computers. We also demonstrate experimental proof-of-concept of fan-out with superconducting qubits. We perform depth (runtime) and fidelity estimation for NISQ application circuits and quantum memory architectures under realistic noise models. Our simulations indicate promising results with an asymptotic advantage in runtime, as well as 7–24% reduction in error.

***Keywords*-quantum computing; trapped ions; NISQ; global interactions**

I. Introduction

Instruction scheduling is a powerful compiler technique in both classical and quantum computing. In the classical realm, scheduling techniques such as pipelining, Single Instruction Multiple Data (SIMD), and Out-of-order execution have led to continued gains in processing power. These scheduling techniques are designed to preserve a program's logical correctness by respecting constraints known as *hazards*.

Just as in the classical setting, quantum computing is also amenable to instruction scheduling. In fact, due to the short lifetimes of qubits in the NISQ (Noisy Intermediate-Scale Quantum) era [67], scheduling to reduce latency is critical for successful execution [13, §II. E.]. The potential of quantum instruction scheduling was recently exemplified by Google's Quantum Supremacy result [68], which experimentally demonstrated a task soluble in seconds on a 53 qubit computer that is argued to likely require days [66] on a supercomputer. At the core of the Supremacy result is a *coupler activation* [68] schedule that maximizes simultaneous resource utilization.

A number of papers [34], [35], [43], [58] in the architecture community have studied quantum scheduling, inspired by techniques from the classical setting. One principle underlying these papers is **exclusive activation**: a qubit can be involved in at most one operation per timestep [35]. In architectural terms, this is a *structural hazard* [40]. Under exclusive activation, schedulers optimize for data parallelism by simultaneously executing instructions on disjoint qubits. However, there are natural limits to such schedulers, since instructions on overlapping qubits must be serialized.

Our work begins with a simple but consequential observation: the structural hazard of exclusive activation is not actually enforced by most quantum hardware. In fact, it is often *more* natural for a quantum processor to simultaneously execute multiple operations on shared qubits through *global interactions*. The building block of our work is the fan-out operation depicted in Figure 1. This operation can be understood purely classically. The four CNOT (Controlled-NOT) gates at the left each comprise a control (•) and a target (⊕), and the target is flipped iff the control qubit is 1. This operation performs fan-out for classical input states: when the targets are initialized to 0, the control bit gets copied to the targets.

While exclusive activation would serialize the four CNOT instructions as depicted on the left, underlying quantum hardware can naturally perform these interactions simultaneously, as depicted on the right. This form of Single Instruction Multiple Data (SIMD) parallelism arises only after discarding structural hazards that don't manifest in

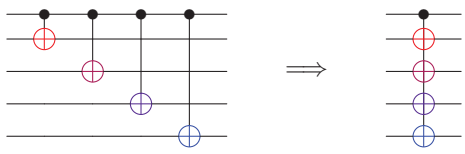


Figure 1: Device level fan-out allows a NOT to the bottom four targets iff the top control is on. While exclusive activation induces serialization (left), quantum hardware can implement fan-out simultaneously (right) in a single step.

978-1-6654-1691-7/21/$31.00 ©2021 IEEE
DOI 10.1109/QCE52317.2021.00045

hardware. As we demonstrate later, the fan-out building block generalizes to efficiently-scheduled circuit synthesis for the ubiquitous Controlled-U operation. Henceforth in this paper, fan-out will refer to simultaneous operation on the right of Figure 1.

We begin in Section II with a survey of prior work. The three subsequent sections capture our core contributions:

- Section III: We generalize the simultaneous fan-out primitive into a **circuit synthesis procedure to schedule Controlled-U** operations with an asymptotic depth advantage.
- Section IV: We leverage this circuit synthesis procedure to **optimize NISQ circuits** (which rely on Controlled-U). We also introduce **novel quantum memory architectures**.
- Section V: We perform **technology modeling** of simultaneous fan-out on trapped ion qubits.

Section VI presents results for several benchmarks. Section VII proposes an implementation of fan-out on superconducting qubits and demonstrates experimental proof-of-concept. Sections VIII concludes. To aid understanding, we link to interactive in-browser demos in Quirk [24] for important circuits.

II. Prior Work

Our work builds on top of prior work from the (1) computer architecture, (2) computer science theory, and (3) physics communities. At a high level, the priorities of the work in each community can be characterized as follows:

1) architects have devised intelligent schedulers/circuit synthesis tools, but they assume a false structural hazard by overlooking global interactions
2) theorists have devised intelligent circuit constructions assuming global gates, but they don't consider NISQ workloads or device-level operation
3) physicists have studied global interactions, but usually in an ad hoc fashion separated from computation and NISQ workloads

Our work unites insights from all three disciplines to devise a circuit synthesis tool that leverages global interactions to accelerate NISQ workloads.

A. Computer Architecture

Amongst architects, a number of papers [34], [35], [39], [43], [58], [83] have studied instruction scheduling in quantum computers. These papers all assume some structural hazard against simultaneous execution of overlapping qubits. [34], [35] provides the most formal description of this hazard, terming it as the principle of exclusive activation which forbids a qubit from being involved in more than one operation per timestep. Moreover, hardware-dependent considerations such as crosstalk [51], [61] further narrow the scope of when operations can be parallelized. For example, on superconducting hardware, `CNOT(a,b); CNOT(c,d);` may be forbidden simultaneously, even though the `CNOT` gates are disjoint.

In other architectural work such as [58] and [39], the authors provide examples for obtaining data parallelism on disjoint instructions. However, in both papers, the examples ultimately incur serialization upon encountering gates on overlapping qubits. As we will demonstrate in Section III, this serialization is unnecessary.

Finally, [43] describes exclusive activation as a data dependency, since the no-cloning theorem [89] prevents copying a qubit to participate in multiple instructions simultaneously. This is indeed a valid perspective. Regardless, we will demonstrate that the underlying problem is in fact addressable with the fan-out primitive.

B. Computer Science Theory

Quantum fan-out has also been studied from a complexity theory lens. [41] proved that the $\mathrm{QNC}^0_\mathrm{f}$ circuit class with unbounded fan-out is powerful for fault-tolerant applications such as Shor's factoring algorithm [76]. Other applications of fan-out to arithmetic operations such as addition, OR, and modulus are considered in [30], [80], [81]. Finally, [82] shows that under widely-held complexity theory assumptions, fan-out in quantum circuits can increase the hardness of classical simulability. Our work revisits these theory results with NISQ workloads and underlying device physics in mind.

C. Physics

The engineering of global interactions on N qubits has been well studied in device physics communities. A common benchmark for global interactions is the preparation of the *GHZ state* [31], a task which is essentially equivalent to fan-out. Experimentally, global interactions have been used to prepare the GHZ state on a variety of leading qubit technologies including Trapped Ion [52], Neutral Atom [63], and NMR [19]. Implementation on NV center qubits has been proposed as well [29]. Notably, superconducting qubits, which are the current leader in hardware scale, were not previously known to support simultaneous overlapping interactions. However, in Section VII, we experimentally demonstrate simultaneous fan-out on superconducting qubits.

Global interactions have already been noted by physicists for their application to Hamiltonian simulation, an important quantum algorithm. As early as 2005, [94] noted that global interactions enable constant depth parity measurement, an important building block for Hamiltonian simulation. Later work [50], [56] further optimized the procedure. Recently this year, three papers [32], [69], [92] have applied global interactions to building blocks of longer-term fault tolerant quantum computers. These papers demonstrate that the Generalized Toffoli operation can be performed in constant time with global interactions, whereas otherwise linear or log depth is required [27].

Very recent papers have adopted an interdisciplinary approach, combining insights from physics and architecture. For example, [56]—which inspired our work—describes fan-out as SIMD parallelism. Also, a recent trapped ion hardware paper [33] describes global interactions as a form of Multiple Instruction Multiple Data (MIMD) parallelism. Our work continues this architectural perspective, while also focusing on NISQ circuit optimizations and further refining the underlying technology models based on recent experimental developments.

III. Controlled-U Synthesis

The basic building block of our work is the simultaneous fan-out operation depicted on the right side of Figure 1. Two important considerations arise in evaluating the applicability of fan-out. The first is whether the simultaneous implementation via global interactions truly achieves a linear speedup over serialized CNOTs. As described in Sections V and VII, experimental results from hardware assert this is indeed the case. The second consideration is how fidelity is affected by simultaneous fan-out versus serialized CNOTs. Our results in Sections V and VII indicate a modest improvement in fidelity from simultaneous fan-out.

We focus on a circuit synthesis procedure that uses fan-out to optimize the Controlled-U operation, described below. This operation is ubiquitous in NISQ algorithms, and each application in Section IV is an instance of Controlled-U. As we will describe in this section, our circuit synthesis procedure yields a Controlled-U implementation that is scheduled to align CNOT gates into a single fan-out step. This yields asymptotic improvements in circuit depth.

The controlled-U operation is depicted at the left of Figure 2. As in other controlled operations like CNOT, the U operation should be applied if and only if the top control qubit is $|1\rangle$. Here we consider the case when U is an operation on multiple qubits. Therefore, U itself has a decomposition into gates, shown under the blue overlay. Our results are applicable for any decomposition basis, but we focus on the decomposition into the universal set of single-qubit + CNOT gates, since quantum algorithms are typically expressed in this form. In the example, U has a width of four qubits and a depth of two layers. The first layer contains four disjoint single qubit gates, and the second layer contains two disjoint CNOTs.

Under exclusive activation, implementation of Controlled-U is bottlenecked by the dependence of each controlled gate on the single control qubit. Thus, the parallel two-layer implementation of U collapses into a serial implementation of Controlled-U as depicted at the right of Figure 2. The amount of serialization is proportional to the width of U, so that the effective depth of a Controlled-U operation under serialization is $O(\text{Depth} \times \text{Width})$. In many workloads, the width greatly exceeds depth, so this serialization is very harmful.

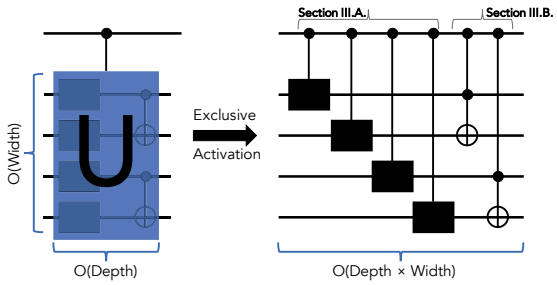


Figure 2: Left: general form of controlled-U. Right: under exclusive activation, adding the control induces serialization and multiplies the effective Depth by the Width.

It is not immediately obvious how fan-out can help speed up Controlled-U. Whereas fan-out is a SIMD operation, Controlled-U is a MIMD operation, since the gates in U are arbitrary. However, we can resolve this difficulty be decomposing gates into a form amenable to 'alignment' of CNOTs into a single fan-out step. This circuit synthesis procedure has two underlying cases. The first, Shared-Control Single Qubit Gates, supports the simultaneous execution of multiple Controlled-U_i gates with a shared control qubit. This procedure applies to the first layer of U in Figure 2. The second, Shared-Control Toffoli's, supports the simultaneous execution of multiple Controlled-CNOTs with a shared control qubit. These double-controlled NOTs are referred to as Toffoli's. The Shared-Control Toffoli's case applies to the second layer of U in Figure 2.

In practice, arbitrary U's will also contain mixed layers that contain both single-qubit gates and CNOTs. This general case can be handled by unifying the synthesis procedures for Shared-Control Single Qubit Gates and Shared-Control Toffoli's. It is not presented here for brevity, but was implemented in our code.

Table I compares the time (depth) and space (ancilla qubits) costs of implementing Controlled-U. Our work, which uses fan-out, is optimal with $O(D)$ depth (and very small constants) and 0 ancilla qubits. The status quo approach of serialization incurs $O(ND)$ depth which is harmful because $N >> D$ in many applications. Past work in [41] and [54] has proposed alternative approaches for parallelizing circuits using global interactions. In the best case, where a "basis-change" is cheap and efficiently computable, [41] matches our $O(D)$ depth. However, it is extremely expensive in space, requiring $O(N^2)$ ancilla qubits. Finally, [54] provides a numerical optimization technique for compiling Controlled-U down to the minimal possible depth. In this sense, it could achieve the $O(D)$ lower bound. However, the numerical optimization for compilation has exponential cost—simply defining the optimization problem involves specifying a $2^N \times 2^N$ sized matrix. Moreover, the optimization itself is expensive, and convergence to $O(D)$

depth is not guaranteed.

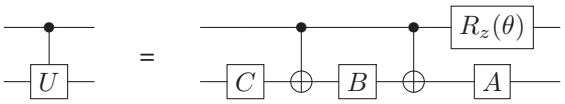
	Depth	Ancilla Qubits
Our Work (with fan-out)	**O(D)**	**0**
Serialization	$O(ND)$	0
[41] (if cheap basis-change)	$O(D)$	$O(N^2)$
[54] ($\Omega(2^N)$ compile time)	$O(D)$?	0

Table I: Cost of implementing a controlled-U operation in time (depth) and space (ancilla qubits). U has a depth of D and width of N qubits.

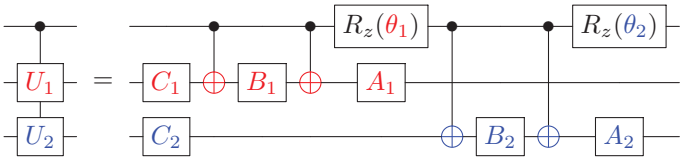
While our procedure achieves the best possible asymptotic space and time costs, it is not as general as [41] and [54]. Our procedure only addresses the special case of Controlled-U parallelization, whereas [41] and [54] apply to the parallelization of any commuting gates or the depth reduction of any unitary, respectively. Nonetheless, our specialization is justified because the Controlled-U template is ubiquitous in NISQ workloads.

A. Shared-Control Single Qubit Gates

Here, we consider how to simultaneously execute controlled single-qubit gates with a shared control, as in the first layer of U in Figure 2. This is a form of MIMD parallelism with overlapping data, but we only have access to the fan-out SIMD primitive. However, we can make progress by invoking the following well-known identity [62] for decomposing controlled single-qubit gates. It shows that for any single-qubit gate U, the Controlled-U operation can be implemented by using CNOT as the only two-qubit gate. Specifically there exist (trivially computable) single-qubit gates A, B, C, and an angle θ, such that



Let us consider applying this identity to a small example: attempting to parallelize Controlled-U_1 and Controlled-U_2 targeting two different qubits. The result is shown below, with colors used for disambiguation.



It appears that applying the circuit identity led to minimal improvements—only C_2 can slide left to execute simultaneously with controlled-U_1. The rest of the blue gates are unable to parallelize, because they are blocked by an apparent dependency on the $R_z(\theta_1)$ gate. However, we can see that the apparent dependence of the blue CNOTs on the $R_z(\theta_1)$ is actually a false dependence, because R_z gates commute with controls. By commuting the $R_z(\theta_1)$ gate to the end of the circuit, we attain the final result in Figure 3.

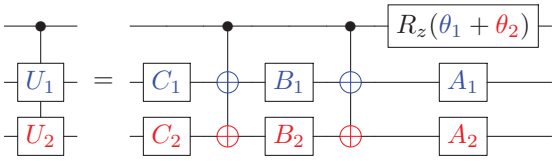
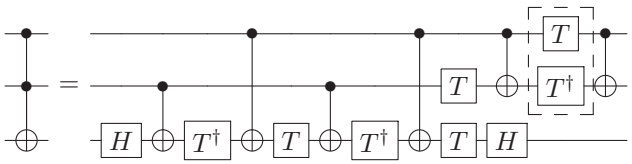


Figure 3: Simultaneous execution of shared-control single qubit gates, using the fan-out primitive. This decomposition has constant (5 layer) depth, independent of width.

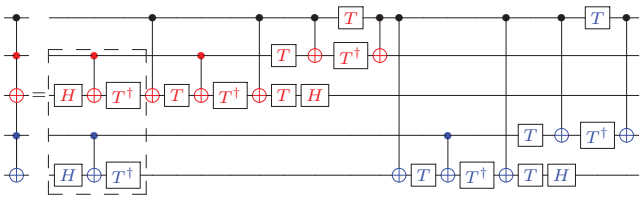
We have now demonstrated simultaneous execution of shared-control U_1 and U_2 on overlapping data (top+middle and top+bottom qubits respectively), using the fan-out primitive. This pattern extends *ad infinitum* to more qubits—the total depth will always consist of five layers: two fan-out layers and three single-qubit gate layers. For certain gates, the cost could be reduced even further. For instance, for $U = Z$, it is known that the Controlled-Z operation can be implemented with just a single CNOT [62].

B. Shared-Control Toffoli's

The second piece needed for optimized Controlled-U synthesis is simultaneous execution of shared-control Toffoli's. Here, we seek to simultaneously execute multiple Toffoli (Controlled-CNOT) gates, where the CNOTs are disjoint but the additional control is shared across the CNOTs, as in the second layer of U in Figure 2. Since Toffoli is a three-qubit operation, it must first be decomposed into single-qubit gates and CNOTs. The standard [62] decomposition is shown next. T and $T^\dagger$ are shorthand for $R_z(\frac{\pi}{8})$ and $R_z(\frac{-\pi}{8})$ respectively.



The boxed group with T and $T^\dagger$ is one example of data parallelism. This level of data parallelism is referred to as a coarse-grained schedule in past architectural work [39]. Next, let us consider applying the Toffoli decomposition to a small example: attempting to simultaneously execute two shared-control Toffoli's, where the CNOTs are disjoint. This exact example is also considered in Figure 4 of [39]. The result is shown below, with colors used again for disambiguation.



As indicated by the boxed layers, only three gates from the blue Toffoli were able to parallelize with the execution of the red Toffoli. This level of parallelization, which results in 21 layers of depth, is referred to as fine-grained scheduling in [39]. While it is slightly better than coarse-grained scheduling, it still linearly serializes the depth. However, we can again leverage commutativity relationships to proceed further and exploit our fan-out primitive.

Notice that the dependency between the right-most red CNOT and the subsequent blue CNOT is in fact a false dependency. These two gates commute since their targets are different. After transposing the two gates, we encounter a T gate that commutes with the control of the blue CNOT, since T is a R_z-type gate. Repeating such commutative transpositions, we can push the blue CNOT to the left to align into a single fan-out. The rest of the blue circuit can be handled similarly, resulting in the final form presented in Figure 4. Since $T = R_z(\frac{\pi}{8})$, the $T \times T$ gate at the top right is just a single $R_z(\frac{\pi}{4})$ gate.

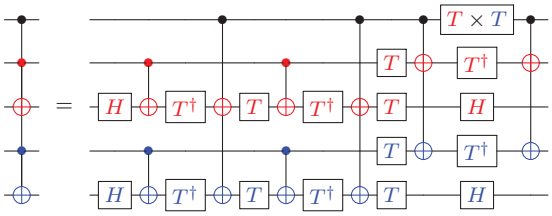


Figure 4: Simultaneous execution of shared-control Toffoli's using the fan-out primitive. This decomposition has constant (12 layer) depth, independent of width. Quirk Link.

The design in Figure 4 extends naturally to more qubits. Regardless of the number of qubits, the depth of the circuit is always 12 layers. Since the depth of a single Toffoli operation is also 12 layers, this means that our shared-control Toffoli's synthesis is optimal. For the circuits we will encounter in the following sections, the number of Toffoli's spans the entire circuit. Therefore the depth cost of the other approaches is $O(N)$, versus our $12 = O(1)$ constant depth.

The combination of simultaneous shared-control single qubit gates and Toffoli's enables a depth-optimized execution schedule for any Controlled-U. Moreover, the multiplicative constants for our circuit synthesis are small. Shared-control single qubit gates incur a depth of just 5 layers, which matches worst case depth. Shared-control Toffoli's incur no depth expansion relative to a single Toffoli and are thus optimal. The resulting Controlled-U circuit synthesis procedure was implemented in our code. In the following section, we apply the Controlled-U procedure to optimize several NISQ-important quantum circuits, which are all fundamentally Controlled-U operations. While our approach is already asymptotically optimal with low constants, in some cases we can reduce the depth constants even further. This is exemplified by the SWAP Test, which we discuss next.

IV. Applications

We now examine how Controlled-U circuit synthesis can be leveraged to optimize NISQ circuits. We also apply fan-out to develop novel quantum memory architectures. Table II summarizes the spacetime advantages of our work (using simultaneous fan-out) for the applications surveyed in this Section.

Application	Spacetime costs
SWAP Test between two $k = \frac{N-1}{2}$ qubit registers	(0 ancilla for all)
Our work	$14 = O(1)$ depth
Serialized	$\sim 14k = O(N)$ depth
Coarse-grained sched. [41]	$\sim 12k = O(N)$ depth
Fine-grained sched. [41]	$\sim 9k = O(N)$ depth
Hadamard Test; N-qubit circuit; U has depth D	
Our work	$O(D)$ depth, 0 ancilla
Other approaches (Table I)	$O(ND)$ depth, $O(N^2)$ ancilla, or $\Omega(2^N)$ compile time
Explicit Memory with n index qubits and bitwidth W	
Our work	$O(n)$ depth, 0 ancilla
Bucket-Brigade QRAM [5]	$O(W2^n)$ depth, 0 ancilla
Parallel QRAM [17]	$O(Wn)$ depth, $O(2^n)$ ancilla
Implicit Memory with n index qubits and bitwidth W	($\sim 1 \cdot n$ ancilla for both)
Our work	$O(2^n)$ depth
QROM [6]	$O(W2^n)$ depth

Table II: Summary of space (ancilla qubits) and time (depth) costs for different applications. Our work leverages the ubiquity of simultaneous fan-out to attain asymptotic advantages.

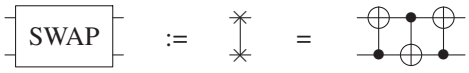
A. SWAP Test

One of the most important [74] procedures in quantum computing, especially NISQ machine learning algorithms, is the calculation of inner products between quantum states. This inner product reports the *overlap* or similarity between states. For two qubit registers $|A\rangle$ and $|B\rangle$, this overlap is

denoted by $|\langle A|B\rangle|^2$. For equal states $|\langle A|B\rangle|^2 = 1$, and for orthogonal states $|\langle A|B\rangle|^2 = 0$.

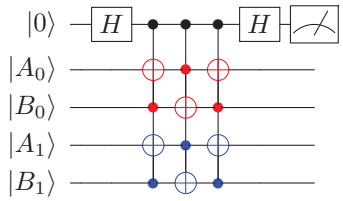
The calculation of this overlap is a procedure known as the SWAP Test. The SWAP Test features heavily in NISQ applications such as quantum kernel classification, which was introduced in [73] and realized experimentally on IBM's quantum hardware in [38]. These quantum kernel methods are noise resilient and amenable to noise mitigation [38]. Further work [23] has introduced kernels that have strong complexity theory foundations for hardness of classical simulability. All of these kernel methods require the evaluation of inner product overlaps. The SWAP Test is also integral to cost function evaluation in NISQ-friendly deep quantum neural networks [7]. In the near-term (and in fact current-term), experimental sequences in quantum sensing [93] are essentially overlap measurements.

The SWAP Test has a very simple form. It is essentially just the case of Controlled-U with U = SWAP. First, we examine the decomposition of a SWAP between two qubits:



This decomposition is equivalent to the triple XOR sequence for in-place SWAPs of classical bits. For a SWAP Test, we need to perform this U = SWAP sequence not just between two individual qubits, but between two registers of qubits. Moreover, the SWAP is controlled on an ancilla qubit. The SWAP Test also requires a Hadamard-sandwich around the controls, and a measurement of the ancilla. After executing such a circuit, the overlap between the two registers is related by a simple function to the probability of measuring $|0\rangle$ on the ancilla. Repeated executions can therefore estimate the overlap to a desired precision.

Let us concretely consider the example of a SWAP Test on two two-qubit registers, $|A = A_1A_0\rangle$ and $|B = B_1B_0\rangle$. To disambiguate the gates, we have used colors and interleaved the bit ordering of the $|A\rangle$ and $|B\rangle$ registers below:



Under standard serialization of the shared-control gates, the depth is 63 at best from fine-grained scheduling. However, our Controlled-U synthesis procedure, specifically the shared-control Toffoli's decomposition, is directly applicable here. The resulting SWAP Test depth is $3 \times 12 = 36$ (ignoring the two Hadamard gates). Moreover, our procedure always yields a constant depth of 36 layers regardless of the circuit width N, whereas serialized approaches scale as $O(N)$.

While this asymptotic advantage is already appealing, we can attain even further cost reductions to our constants via a circuit identity. It can be shown that the outer two controls on the ancilla qubit can be removed [22], [62]. After this optimization, the final circuit has a depth of just 14 layers, regardless of the size of the SWAP Test. To illustrate for larger N, this Quirk Link shows an interactive SWAP Test circuit for computing the overlap of two four-qubit registers, with an ancilla qubit on the top.

1) Interference Circuit: Recent work has explored alternatives to the traditional SWAP Test, with the aim of reducing spacetime costs. The most promising one is the interference circuit [72], [74], which halves the qubit width requirement. Whereas the traditional SWAP Test requires $2k+1$ qubits to compute the overlap of two k-qubit registers, the interference circuit only requires $k+1$ qubits. In order to use the interference circuit, we must know the sequences of gates U_A and U_B that can create $|A\rangle$ and $|B\rangle$, respectively. In practice, this is indeed the case for useful applications. The interference circuit has the following simple form shown in Figure 5. As in the traditional SWAP Test, the overlap is a simple function of the probability of measuring $|0\rangle$ on the ancilla.

The open-control (open circle) on U_B activates on $|0\rangle$ and can therefore be replaced with an ordinary control surrounded by NOT ($\oplus$) gates. Therefore our Controlled-U is directly applicable to the interference circuit, and it allows overlap calculation with no asymptotic depth overhead relative to U_A and U_B. This is again a linear $O(N)$ speedup via fan-out.

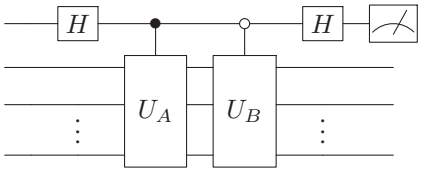


Figure 5: The interference circuit computes the overlap between k-qubit states, $|A\rangle$ and $|B\rangle$, with just $k+1$ qubits.

B. Hadamard Test

The SWAP Test is a specific case of a more general procedure called the Hadamard Test. The Hadamard Test has a very simple and familiar form shown in Figure 6. This is essentially just the Controlled-U operation we focused on in Section III. Moreover, the SWAP Test is just the case where U = SWAP. Selecting other U makes the Hadamard Test give rise to a wide variety of applications. We list our benchmarked applications in Table III. There are numerous additional applications of the Hadamard Test, such as training Quantum Boltzmann Machines [88], gradient evaluation [18], [36], [59], [71], and Jones polynomial approximation [2].

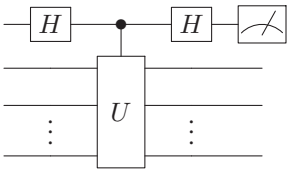


Figure 6: Circuit for the Hadamard Test.

Application	Description
Variational Quantum Linear Solver [10], [42], [91]	Algorithm for solver large linear systems using NISQ hardware
Matrix elements of group representation [14], [46]	Group theory problem; U is essentially the Quantum Fourier Transform
Entanglement spectroscopy [45]	Computation of entanglement spectrum of arbitrary quantum states
Controlled Density Matrix Exponentiation (DME) [47]	Several appliations, e.g. for private quantum software [55]

Table III: Applications of the Hadamard Test. Each corresponds to a different choice of Controlled-U.

C. Quantum Memory Architectures

Next, we investigate the use of fan-out to improve the implementation of quantum memory, which speeds up or enables many quantum algorithms [84], [87]. The high-level function of a quantum memory is similar to that of a classical memory: n index bits enumerate over 2^n memory cells. Following the notation of [4], we denote the n index bits as the $|b\rangle$ register and the 2^n memory cells as the $|m\rangle$ register. As in the classical case, we expect that setting the index register to $|i\rangle$ should allow us to retrieve the ith memory cell, $|m_i\rangle$. However, for a quantum memory, we also require the retrieval to work over superpositions of index qubits. For example, setting $|b\rangle$ to $\frac{1}{\sqrt{2}}[|000\rangle + |111\rangle]$ should retrieve the superposition, $\frac{1}{\sqrt{2}}[|m_0\rangle + |m_7\rangle]$.

In this section, we apply the fan-out primitive to both explicit and implicit quantum memories, which we define below. We demonstrate considerable improvements—exponential and linear respectively—over prior work, as summarized in Table II. These improvements are important because the cost of quantum memory is often the principal bottleneck for realizing practical speedups. While it remains unclear if quantum memory architectures will be feasible [1], [5], [9], [67] even for future fault-tolerant devices, our proposed improvements at least justify a re-assessment of the feasibility.

1) Explicit Quantum Memory: In an explicit quantum memory, the 2^n memory cells are each explicitly stored in qubit registers. In this sense, an explicit quantum memory is akin to a 2^n–to–1 multiplexer or data selector from classical electronics. As discussed, the quantum variant should extend to the case where select lines are in superposition. Moreover, each of the 2^n memory cells is stored in a qubit register, so each memory cell can itself contain a quantum (superposition) state.

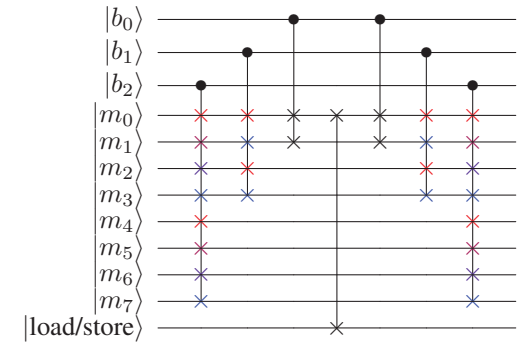


Figure 7: Architecture for an explicit quantum memory with $n = 3$ index qubits and $2^n = 8$ memory cells of bitwidth $W = 1$. Quirk demo.

The dominant architecture for this explicit quantum memory is termed Quantum Random Access Memory. The bucket brigade design of QRAM was introduced in [25], [26] and cast to the quantum circuit model in [5]. This bucket brigade QRAM requires $\sim 2 \cdot 2^n$ qubits and $O(W2^n)$ depth. Later work [17] was able to parallelize execution to achieve $O(Wn)$ depth, but requires an additional $\sim 6 \cdot 2^n$ ancilla qubits. We now present a novel architecture for explicit quantum memory that requires only $O(n)$ depth, with 0 ancilla qubits.

Figure 7 shows our architecture for $n = 3$ index qubits. There are $2^3 = 8$ explicit memory cells, each of single-qubit bitwidth $W = 1$. At a high level, the circuit performs a "migration" of the target memory cell into $|m_0\rangle$. Consider for example $|\vec{b} = 101\rangle$, which should access $|m_5\rangle$. The control on the MSB performs a SWAP between $|m_{7654}\rangle$ and $|m_{3210}\rangle$, moving $|m_5\rangle$ into $|m_1\rangle$. The control on the middle index does not activate, but the control on the LSB is activated and SWAPs $|m_1\rangle$ into the $|m_0\rangle$ destination. Finally, this qubit is swapped into the |load/store⟩ register. The right half of the circuit reverses the earlier migrations, restoring the other memory cells to their original locations.

The efficiency of this architecture is enabled by the simultaneous execution of controlled SWAPs, which in turn is enabled by the fan-out primitive. As a result, the circuit depth is only $O(n)$. Moreover, while our example shows the $W = 1$ bitwidth case, it is apparent that with simultaneous fan-out, W is irrelevant to depth. By contrast, serialization would impose an additional linearity in W.

During the preparation of this paper, another proposal was published for $O(n)$-depth and ancilla-free explicit quantum

memory [64], which matches our asymptotic costs.

2) Implicit Quantum Memory: Next we consider implicit quantum memory. In this model, the 2^n memory cells represent classical (non-superposition) data that is known in advance. In such a case, there is no need to waste qubits to represent the classically-known memory cells. Instead, the memory can be stored implicitly through the classical control, a memory architecture that has been referred to as Quantum Read Only Memory [6].

Figure 8 shows an example implicit memory storing the first four prime numbers: $\{00 \to 2, 01 \to 3, 10 \to 5, 11 \to 7\}$. The resulting circuit has a simple form, enumerating all 2^n indices and associating each index with a corresponding pattern of $\oplus$ gates. Without fan-out, implicit memory has $O(W2^n)$ depth via the unary iteration optimization in [6]. However, simultaneous fan-out obviates the scaling in W. This is appealing, because for datasets such as images, the bitwidth (W) of each record exceeds the number of records.

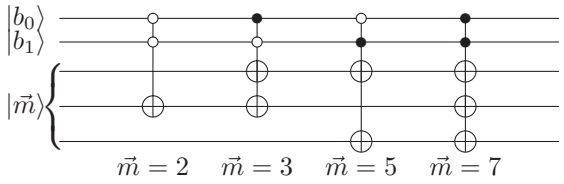


Figure 8: Implicit memory storing the first four prime numbers. The $W = 3$ bitwidth memory is implicitly defined through classical control, based on the pattern of $\oplus$'s. For anticipated applications, W can be large.

V. Technology Modeling: Trapped Ion

In this section, we model the implementation of fan-out on trapped ion quantum computers. Trapped ions feature long qubit coherence times [86] and gate fidelities exceeding 99.99% and 99.9% for single- and two- qubit gates on current hardware [12], [21]. Furthermore, all N qubits can be simultaneously entangled via a global interaction known as the Global Mølmer–Sørensen (GMS) gate [60], [78]. Recent work [8], [56] has explicitly demonstrated how GMS is essentially equivalent to simultaneous fan-out. Moreover, in the past year, experimental work has merged demonstrating pulse shaping for global interactions [20], [33], [52] to support the use of GMS both for fan-out and for parallel two-qubit gates on disjoint qubits. Our focus here is on studying differences in speed and fidelity between simultaneous fan-out versus $N - 1$ serialized CNOTs. For brevity and to maintain a focus on architectural themes, we omit many physical implementation details here.

Regarding the potential speedup, [8], [33], [79] assert that simultaneous fan-out via GMS is indeed linearly faster than serialized CNOTs. To evaluate the fidelity impact, we performed numerical simulations of fan-out via GMS for $N = 2$ to 8 qubits. We constructed a realistic error model that accounts for two sources of noise: overrotation and laser dephasing. Overrotation occurs due to the fact that the angle θ of the Mølmer-Sørensen rotation is sensitive to motional frequency drifts, and it has higher-order dependence on the motional states [16], [85], [90]. An overrotation error can be modeled by replacing θ by $(1+\epsilon)\theta$, where ϵ denotes the overrotation rate. Laser dephasing arises from fluctuations of the optical path length [49], [85], [90].

For current trapped ion hardware, we conservatively estimate typical overrotation rates of 5%. We modeled GMS interaction times of 100 μs [8], contrasted against 80 ms laser coherence time [85]. To evaluate the sensitivity of our results to these parameters, we also modeled under three future scenarios: 5x lower overrotation rate, 5x longer laser coherence, and both improvements. Our simulations were performed using master-equation simulation in QuTiP [44]. We performed stochastic simulation over 100k runs per scenario. The fidelity results are shown in Figure 9.

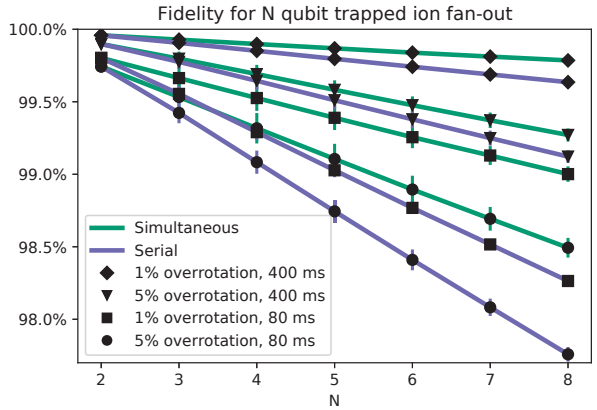


Figure 9: Simulation results for fan-out on trapped ion hardware. Sensitivity analysis performed under four {overrotation rate, laser coherence time} scenarios. For each scenario, we simulated fidelity for simultaneous versus serial. Results averaged across 100k stochastic runs per scenario, executed with 50k CPU-core hours on a large computing cluster.

Conceptually, overrotation errors affect simultaneous and serial equally. Meanwhile, laser dephasing affects serial more adversely, because the laser dephasing effect on the control qubit accumulates over the additional time required for $N - 1$ consecutive CNOTs. Although simultaneous always outperforms serial on our simulations, the exact fidelity advantage is dependent on the parameter settings. For current technology (•), simultaneous has an almost 1% higher fidelity for $N = 8$. For the scenario with 5x longer laser coherence (▼), simultaneous has almost no fidelity advantage over serial. For the scenario with 5x lower overrotation (■), simultaneous again has a nearly 1%

fidelity advantage over serial. Also, across all scenarios, the advantage of simultaneous fan-out increases for larger N, which is encouraging. The simulation results roughly agree with experimental work. For example [52] observed 93.4% fidelity inclusive of State Preparation and Measurement (SPAM) errors for a 4-qubit fan-out executed on hardware over a year ago, and Figure 9 suggests SPAM-exclusive fidelity of 99.3% on current hardware. As cloud access to trapped ion hardware emerges over the coming year, it will be possible to experimentally validate these simulations.

VI. Results

A. Methodology

We evaluated the exact depth reduction for eight applications: SWAP Tests (both traditional and interference circuit), Hadamard Tests (all four applications in Table III), and memory architectures (both explicit and implicit). We compiled each benchmark, across a wide range of circuit widths, using both our fan-out based approach (Simultaneous) and the standard serialized approach with no fan-out (Serial). The results are plotted in Figure 10.

In Figure 11, we also evaluated the fidelity advantage of simultaneous fan-out for the five most NISQ-friendly benchmarks. For each benchmark type, we found the largest circuit instance with fan-out of at most 8 qubits, matching the largest fan-out we simulated in Figure 9. Then, we estimated fidelity with a coarse metric: for each circuit, we assigned each gate a fidelity based on the current hardware "5% overrotation, 80 ms laser coherence" simulation in Figure 9. Multiplying together these gate fidelities gives an approximation for the total circuit fidelity (i.e. 1 - infidelity). We also performed this multiplication under the "1% overrotation, 80 ms laser coherence" future scenario with 5x lower overrotation. While these estimates are less accurate than full density matrix simulation, as we performed in Figure 9, they are informative from an Amdahl's Law perspective. In particular, single- and two- qubit gates are equally penalized in the Simultaneous and Serial circuits, so the Simultaneous circuits can only perform better when there are large fan-out gates.

B. Discussion

As mentioned in Section V, simultaneous fan-out does genuinely give a linear speedup over serialization. Therefore, the depth reductions in Figure 10 translate directly to faster time-to-solution. For four of the eight benchmarked applications, the underlying U has constant depth, so our Simultaneous circuit also has constant depth. For the other four benchmarks, the underlying U has $\Omega(N)$ depth, so both Simultaneous and Serial have increasing depth with N. However, Simultaneous' scaling is still lower than Serial's by a linear factor.

Among our benchmarks, Variational Quantum Linear Solver and Controlled Density Matrix Exponentiation have particularly high fidelity advantage via Simultaneous Fan-out. Our results also demonstrate the sensitivity to the underlying trapped ion hardware's error parameters. For example, VQLS has a 13.9% Serial→Simultaneous infidelity reduction on current hardware versus a 20.9% reduction on future hardware with 5x lower overrotation.

On current hardware, fidelity is the primary system bottleneck. As such, the fidelity improvement of simultaneous fan-out justifies its use in NISQ machines. 7–24% reductions in infidelity on 8-qubit circuits are equivalent to months of hardware progress. As a practical message to hardware providers, we emphasize that exposing global interactions to software will lead to substantial improvements in both fidelity and speed for NISQ applications.

VII. Technology Modeling: Superconducting

Global interaction can be realized for many technologies, but superconducting qubits—which are currently the frontrunner in commercial activity—are a notable exception. To the best of our knowledge, there are no prior implementations of fan-out on superconducting devices. In this section, we demonstrate that superconducting quantum computers can in fact perform simultaneous fan-out.

We first examine the implementation of a CNOT with superconducting qubits. The CNOT is not a natural physical interaction between qubits. Instead, it is performed through a sequence of more primitive physical interactions between qubits. On Google and Rigetti superconducting quantum hardware, CNOT can be realized by a sequence of iSWAP interactions, which are similar to ordinary SWAPs. However, this seems incompatible with simultaneous fan-out, which conceptually requires concurrent reads on the control qubit. By contrast, iSWAP performs both reads and writes on the control qubit since its state is swapped with the target.

An alternative two-qubit interaction called Cross-Resonance [65], [70] is better suited. The Cross-Resonance interaction is used to perform CNOT gates on IBM's devices. It has less restrictive hardware requirements than iSWAP, so Cross-Resonance could be performed on Google and Rigetti hardware as well. Critically, the Cross-Resonance interaction only reads the control qubit, so it does not suffer the immediate barrier to fan-out that iSWAP does.

Although the control qubit *state* is unaffected during Cross-Resonance, the interaction requires (somewhat counterintuitively) driving the control qubit with a microwave pulse. However, by setting this microwave pulse to the frequency of the target qubit, the target qubit will rotate conditioned on the state of the control qubit. This physical interaction easily converts to CNOT through a single-qubit postprocessing gate.

Figure 12 illustrates how we can extend this Cross-Resonance interaction to engineer fan-out. In this example, qubit 3 is the control and qubits 2 and 5 are the two targets. To perform the CNOT from 3 to 2 (5), we would drive qubit

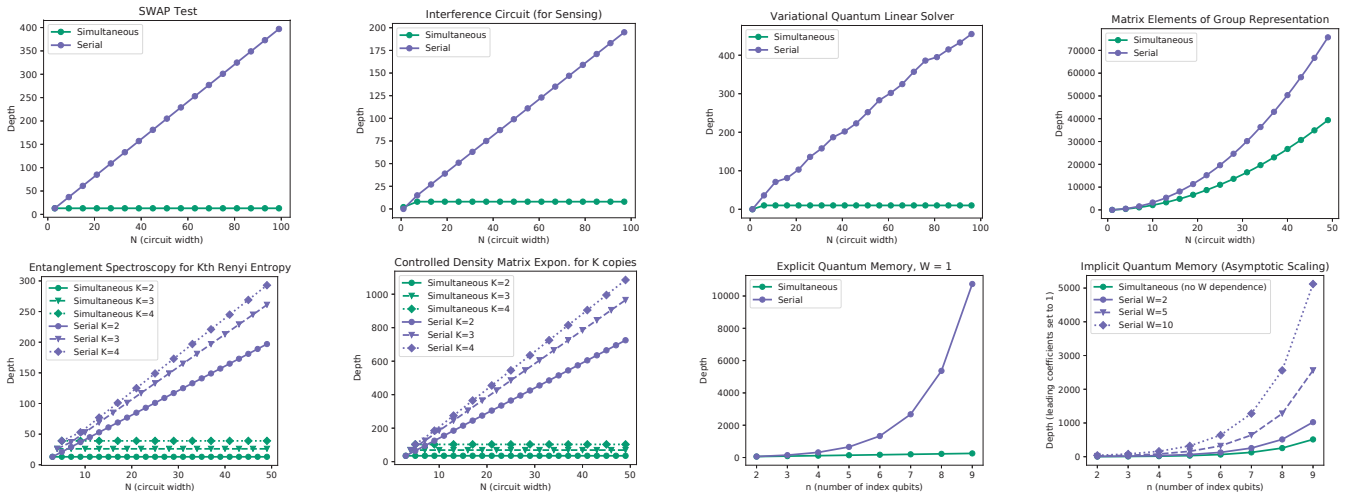


Figure 10: Depth (lower is better) for SWAP Test, Hadamard Test, and memory architecture benchmarks. We compare circuits compiled with our Controlled-U circuit synthesis procedure (which uses simultaneous fan-out) versus circuits that serialize the CNOTs.

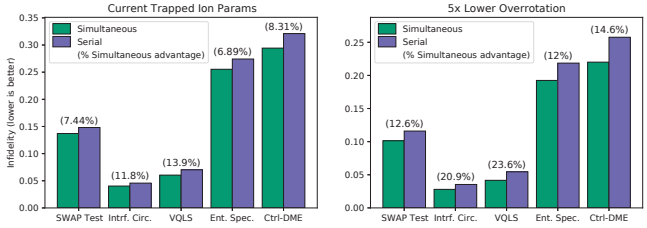


Figure 11: Infidelity estimates for five benchmarks.

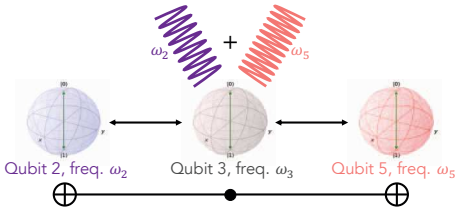


Figure 12: Schematic of fan-out using Cross-Resonance on superconducting qubits. The control qubit (3) is driven with the sum of waves at the targets' frequencies, ω_2 and ω_5.

3 with microwave at frequency ω_2 (ω_5). However, if we instead drive qubit 3 with the *summation* of two sine waves at frequencies ω_2 and ω_5, then we effectively perform both CNOTs simultaneously. The resulting pulse sequence has a linear speedup over serialization, as desired.

We experimentally realized this specific example of fan-out from qubit 3 to qubits 2 and 5 using IBM's Paris quantum computer. We performed our experiment using OpenPulse [3], [28], [57], an interface that enables low-level access of quantum computers through Arbitary Waveform Generators (AWGs). This level of access is required since we need to drive qubit 3 with an unconventional sum-of-waves pulse. We also use a technique called sideband modulation, which is needed since the qubit 3 drive is configured to oscillate at ω_3 by default. Moreover, in practice, high fidelity Cross-Resonance interactions require an echo sequence [15] and active cancellation pulses on the target qubits [53], [75]. Additionally, we had to calibrate a phase offset for the sideband to compensate for accumulated phase on the coaxial cable transitioning from room temperature electronics to the fridge [3], [48].

Figure 13 shows our experimental results. We generated the GHZ state, $\frac{|000\rangle+|111\rangle}{\sqrt{2}}$, by performing NOT on qubit 3 and then fanning out its state to qubits 2 and 5. Ideally, this would result in $|000\rangle$ and $|111\rangle$ each with 50% probability. With simultaneous fan-out, we achieved 31% and 29% respectively. Serialization achieved 42% and 36% respectively.

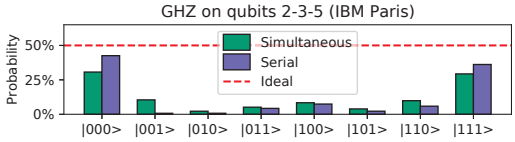


Figure 13: OpenPulse results from 8000×2 repetitions on IBM Q Paris. The ideal output is 50% $|000\rangle$ and 50% $|111\rangle$.

While the GHZ state produced with serial fan-out is better than the one produced with simultaneous fan-out, we emphasize that the simultaneous version ran almost twice as fast. This speedup is encouraging, because superconducting qubits have short coherence lifetimes, so faster operations lead to significant fidelity improvements [13, §II. E.]. Moreover, when we consider larger width circuits, faster fan-out

on a subset of qubits can improve the quality of the other qubits which decohere for less time. Finally, anticipated increases to the sampling rate of AWGs should improve the fidelity of the simultaneous fan-out operation. Most importantly, our experiment affirms that simultaneous fan-out is possible at all on superconducting qubits.

A. Scalability

An immediate barrier to scaling our procedure to more target qubits is that each control-target pair must be connected in hardware. On superconducting qubit platforms, connectivity is typically sparse. For example, on IBM Q Paris's device topology, the maximum degree is 3, and most qubits are connected to just one or two neighbors. Scaling the connectivity will be a challenge. However, we note that fan-out does not require all-to-all connectivity. Instead, we require a star topology, where a single (control) qubit is connected to every other qubit. Such star topologies have been realized experimentally with 10 qubits connected to a single bus [77]. Moreover, star topology is also useful for Hamiltonian simulation circuits [37], so there are numerous other quantum subroutines that would also benefit.

A second consideration is that summing waves for each target qubit's frequency (as in Figure 12) will not scale since the maximum amplitude of AWGs is power-constrained. We propose two possible solutions to this. On frequency tunable devices (where ω_q for each qubit can be controlled), we can simply tune all target qubits to a common frequency during fan-out. Then, the control qubit can be driven at this single common frequency, bypassing the summation of multiple waves. The other solution pertains to fixed-frequency devices. Here, we propose that rectangular-topology qubits could be fabricated with frequencies according to a checkerboard pattern. In such an arrangement, just two colors (frequencies) are needed to ensure no frequency collisions between neighboring qubits. During fan-out, the control qubit can be driven at the sum of just two frequencies, averting the scalability issue.

While these proposed solutions are sound in theory, practical realization will be challenging due to experimental nuances. For example, current qubit fabrication technologies are imprecise and stochastic [11], so fabricating qubit frequencies in a checkerboard pattern will be difficult. Thus, more experimental progress will be needed to scale fan-out on superconducting hardware. These hardware-software codesign considerations are valuable in closing the gap from NISQ hardware to practical applications. We propose further work to evaluate simultaneous fan-out with superconducting qubits.

VIII. Conclusion

At a high level, this work validates the importance of hardware-software codesign. Our core result is driven from the hardware $\rightarrow$ software observation that the exclusive activation structural hazard is not necessary in quantum computing. By exploiting simultaneous fan-out, we are able to synthesize optimized circuit schedules for Controlled-U, which is important in NISQ workloads. In the software $\rightarrow$ hardware direction, our results suggest a number of priorities for future hardware development—in particular, the importance of exposing global interactions. Moreover, our demonstration of simultaneous fan-out in superconducting qubits they could be brought to parity with trapped ions.

In current systems, our results affirm a linear speedup from fan-out. In the NISQ era, algorithms will require millions of iterations [27], so quantum execution speedups translate to direct reductions in time-to-solution. This opportunity is particularly pronounced on trapped ions, which operate at relatively slow kHz speeds. In addition to the circuit execution speedup, our simulations show 7–24% infidelity reductions from simultaneous fan-out. This is validated by our trapped ion simulation with a realistic noise model. Our experimental results from superconducting qubits are also promising, though our emphasis is on the mere fact that simultaneous fan-out is possible at all on superconducting qubits.

Acknowledgment

We are grateful to Ali Javadi-Abhari, Dave Schuster, and Dripto Debroy for helpful suggestions. This work is funded in part by EPiQC, an NSF Expedition in Computing, under grants CCF-1730082/1730449; in part by STAQ under grant NSF Phy-1818914; in part by DOE grants DE-SC0020289 and DE-SC0020331; and in part by NSF OMA-2016136 and the Q-NEXT DOE NQI Center. In addition, this work is funded in part by

This material is based upon work supported by the National Science Foundation under Grant No. 2110860 and by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under Award Number DE-SC0021526.

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. We also acknowledge the University of Chicago's Research Computing Center for their support of this work.

Disclosure: Fred Chong is Chief Scientist at Super.tech and an advisor to Quantum Circuits, Inc.

References

[1] S. Aaronson, "Read the fine print," *Nature Physics*, vol. 11, no. 4, pp. 291–293, 2015.

[2] D. Aharonov, V. Jones, and Z. Landau, "A polynomial quantum algorithm for approximating the jones polynomial," *Algorithmica*, vol. 55, no. 3, pp. 395–421, 2009.

[3] T. Alexander, N. Kanazawa, D. J. Egger, L. Capelluto, C. J. Wood, A. Javadi-Abhari, and D. McKay, "Qiskit pulse: Programming quantum computers through the cloud with pulses," *arXiv preprint arXiv:2004.06755*, 2020.

[4] I. Arad and Z. Landau, "Quantum computation and the evaluation of tensor networks," *SIAM Journal on Computing*, vol. 39, no. 7, pp. 3089–3121, 2010.

[5] S. Arunachalam, V. Gheorghiu, T. Jochym-O'Connor, M. Mosca, and P. V. Srinivasan, "On the robustness of bucket brigade quantum ram," *New Journal of Physics*, vol. 17, no. 12, p. 123010, 2015.

[6] R. Babbush, C. Gidney, D. W. Berry, N. Wiebe, J. McClean, A. Paler, A. Fowler, and H. Neven, "Encoding electronic spectra in quantum circuits with linear t complexity," *Physical Review X*, vol. 8, no. 4, p. 041015, 2018.

[7] K. Beer, D. Bondarenko, T. Farrelly, T. J. Osborne, R. Salzmann, and R. Wolf, "Efficient learning for deep quantum neural networks," *arXiv preprint arXiv:1902.10445*, 2019.

[8] A. Bermudez, X. Xu, R. Nigmatullin, J. O'Gorman, V. Negnevitsky, P. Schindler, T. Monz, U. G. Poschinger, C. Hempel, J. Home, F. Schmidt-Kaler, M. J. Biercuk, R. Blatt, S. Benjamin, and M. Muller, "Assessing the progress of trapped-ion processors towards fault-tolerant quantum computation," *Physical Review X*, vol. 7, p. 041061, 2017.

[9] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, 2017.

[10] C. Bravo-Prieto, R. LaRose, M. Cerezo, Y. Subasi, L. Cincio, and P. J. Coles, "Variational quantum linear solver: A hybrid algorithm for linear systems," *arXiv preprint arXiv:1909.05820*, 2019.

[11] M. Brink, J. M. Chow, J. Hertzberg, E. Magesan, and S. Rosenblatt, "Device challenges for near term superconducting quantum processors: frequency collisions," in *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2018, pp. 6–1.

[12] K. R. Brown, A. C. Wilson, Y. Colombe, C. Ospelkaus, A. M. Meier, E. Knill, D. Leibfried, and D. J. Wineland, "Single-qubit-gate error below 10- 4 in a trapped ion," *Physical Review A*, vol. 84, no. 3, p. 030303, 2011.

[13] J. Cheng, H. Deng, and X. Qian, "Accqoc: Accelerating quantum optimal control based pulse generation," *arXiv preprint arXiv:2003.00376*, 2020.

[14] P. J. Coles *et al.*, "Quantum algorithm implementations for beginners," *ArXiv*, vol. abs/1804.03719, 2018.

[15] A. D. Córcoles, J. M. Gambetta, J. M. Chow, J. A. Smolin, M. Ware, J. Strand, B. L. Plourde, and M. Steffen, "Process verification of two-qubit quantum gates by randomized benchmarking," *Physical Review A*, vol. 87, no. 3, p. 030301, 2013.

[16] D. Debroy, M. Li, M. Newman, and K. R. Brown, "Stabilizer slicing: Coherent error cancellations in ldpc codes," *arXiv preprint arXiv:1810.01040*, 2018.

[17] O. Di Matteo, V. Gheorghiu, and M. Mosca, "Fault-tolerant resource estimation of quantum random-access memories," *IEEE Transactions on Quantum Engineering*, vol. 1, pp. 1–13, 2020.

[18] M. Q. Documentation, "Estimategradient operation," Available at https://docs.microsoft.com/en-us/qsharp/api/qsharp/microsoft.quantum.machinelearning.estimategradient.

[19] S. Dogra, K. Dorai, and Arvind, "Experimental construction of generic three-qubit states and their reconstruction from two-party reduced states on an nmr quantum information processor," *Physical Review A*, vol. 91, no. 2, p. 022312, 2015.

[20] C. Figgatt, A. Ostrander, N. M. Linke, K. A. Landsman, D. Zhu, D. Maslov, and C. Monroe, "Parallel entangling operations on a universal ion-trap quantum computer," *Nature*, vol. 572, no. 7769, pp. 368–372, 2019.

[21] J. P. Gaebler, T. R. Tan, Y. Lin, Y. Wan, R. Bowler, A. C. Keith, S. Glancy, K. Coakley, E. Knill, D. Leibfried, and D. J. Wineland, "High-fidelity universal gate set for be 9+ ion qubits," *Physical review letters*, vol. 117, no. 6, p. 060505, 2016.

[22] J. C. Garcia-Escartin and P. Chamorro-Posada, "Swap test and hong-ou-mandel effect are equivalent," *Physical Review A*, vol. 87, no. 5, p. 052330, 2013.

[23] R. Ghobadi, J. S. Oberoi, and E. Zahedinejhad, "The power of one qubit in machine learning," *arXiv preprint arXiv:1905.01390*, 2019.

[24] C. Gidney, "Quirk quantum circuit simulator," *A drag-and-drop quantum circuit simulator. URL: https://algassert. com/quirk*, 2016.

[25] V. Giovannetti, S. Lloyd, and L. Maccone, "Architectures for a quantum random access memory," *Physical Review A*, vol. 78, no. 5, p. 052310, 2008.

[26] ——, "Quantum random access memory," *Physical review letters*, vol. 100, no. 16, p. 160501, 2008.

[27] P. Gokhale, J. M. Baker, C. Duckering, N. C. Brown, K. R. Brown, and F. T. Chong, "Asymptotic improvements to quantum circuits via qutrits," in *Proceedings of the 46th International Symposium on Computer Architecture*, 2019, pp. 554–566.

[28] P. Gokhale, A. Javadi-Abhari, N. Earnest, Y. Shi, and F. T. Chong, "Optimized quantum compilation for near-term algorithms with openpulse," *arXiv preprint arXiv:2004.11205*, 2020.

[29] G. Goldstein, P. Cappellaro, J. Maze, J. Hodges, L. Jiang, A. S. Sørensen, and M. Lukin, "Environment-assisted precision measurement," *Physical review letters*, vol. 106, no. 14, p. 140502, 2011.

[30] F. Green, S. Homer, C. Moore, and C. Pollett, "Counting, fanout, and the complexity of quantum acc," *arXiv preprint quant-ph/0106017*, 2001.

[31] D. M. Greenberger, M. A. Horne, and A. Zeilinger, "Going beyond bell's theorem," in *Bell's theorem, quantum theory and conceptions of the universe*. Springer, 1989, pp. 69–72.

[32] K. Groenland, F. Witteveen, K. Schoutens, and R. Gerritsma, "Sequences of molmer-sorensen gates can implement controlled rotations using quantum signal processing techniques," *arXiv preprint arXiv:2001.05231*, 2020.

[33] N. Grzesiak, R. Blümel, K. Beck, K. Wright, V. Chaplin, J. M. Amini, N. C. Pisenti, S. Debnath, J.-S. Chen, and Y. Nam, "Efficient arbitrary simultaneously entangling gates on a trapped-ion quantum computer," *arXiv preprint arXiv:1905.09294*, 2019.

[34] G. G. Guerreschi, "Scheduler of quantum circuits based on dynamical pattern improvement and its application to hardware design," *arXiv preprint arXiv:1912.00035*, 2019.

[35] G. G. Guerreschi and J. Park, "Two-step approach to scheduling quantum circuits," *Quantum Science and Technology*, vol. 3, no. 4, p. 045003, 2018.

[36] G. G. Guerreschi and M. Smelyanskiy, "Practical optimization for hybrid quantum-classical algorithms," *arXiv preprint arXiv:1701.01450*, 2017.

[37] K. Gui, T. Tomesh, P. Gokhale, Y. Shi, F. T. Chong, M. Martonosi, and M. Suchara, "Term grouping and travelling salesperson for digital quantum simulation," *arXiv preprint arXiv:2001.05983*, 2020.

[38] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.

[39] J. Heckey, S. Patil, A. JavadiAbhari, A. Holmes, D. Kudrow, K. R. Brown, D. Franklin, F. T. Chong, and M. Martonosi, "Compiler management of communication and parallelism for quantum computation," in *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2015, pp. 445–456.

[40] J. L. Hennessy and D. A. Patterson, *Computer architecture: a quantitative approach*. Elsevier, 2011.

[41] P. Høyer and R. Špalek, "Quantum fan-out is powerful," *Theory of computing*, vol. 1, no. 1, pp. 81–103, 2005.

[42] H.-Y. Huang, K. Bharti, and P. Rebentrost, "Near-term quantum algorithms for linear systems of equations," *arXiv preprint arXiv:1909.07344*, 2019.

[43] A. JavadiAbhari, S. Patil, D. Kudrow, J. Heckey, A. Lvov, F. T. Chong, and M. Martonosi, "Scaffcc: Scalable compilation and analysis of quantum programs," *Parallel Computing*, vol. 45, pp. 2–17, 2015.

[44] J. R. Johansson, P. D. Nation, and F. Nori, "Qutip 2: A python framework for the dynamics of open quantum systems," *Computer Physics Communications*, vol. 184, no. 4, pp. 1234–1240, 2013.

[45] S. Johri, D. S. Steiger, and M. Troyer, "Entanglement spectroscopy on a quantum computer," *Physical Review B*, vol. 96, no. 19, p. 195136, 2017.

[46] S. P. Jordan, "Fast quantum algorithms for approximating some irreducible representations of groups," *arXiv preprint arXiv:0811.0562*, 2008.

[47] M. Kjaergaard *et al.*, "A quantum instruction set implemented on a superconducting quantum processor," *arXiv: Quantum Physics*, 2020.

[48] S. Krinner, S. Storz, P. Kurpiers, P. Magnard, J. Heinsoo, R. Keller, J. Luetolf, C. Eichler, and A. Wallraff, "Engineering cryogenic setups for 100-qubit scale superconducting circuit systems," *EPJ Quantum Technology*, vol. 6, no. 1, p. 2, 2019.

[49] P. J. Lee, K.-A. Brickman, L. Deslauriers, P. C. Haljan, L.-M. Duan, and C. Monroe, "Phase control of trapped ion quantum gates," *Journal of Optics B: Quantum and Semiclassical Optics*, vol. 7, no. 10, p. S371, 2005.

[50] D. Leibfried and D. J. Wineland, "Efficient eigenvalue determination for arbitrary pauli products based on generalized spin-spin interactions," *Journal of Modern Optics*, vol. 65, no. 5-6, pp. 774–779, 2018.

[51] G. Li, Y. Ding, and Y. Xie, "Towards efficient superconducting quantum processor architecture design," *arXiv preprint arXiv:1911.12879*, 2019.

[52] Y. Lu, S. Zhang, K. Zhang, W. Chen, Y. Shen, J. Zhang, J.-N. Zhang, and K. Kim, "Global entangling gates on arbitrary ion qubits," *Nature*, vol. 572, no. 7769, pp. 363–367, 2019.

[53] E. Magesan and J. M. Gambetta, "Effective hamiltonian models of the cross-resonance gate," *arXiv preprint arXiv:1804.04073*, 2018.

[54] E. A. Martinez, T. Monz, D. Nigg, P. Schindler, and R. Blatt, "Compiling quantum algorithms for architectures with multi-qubit gates," *New Journal of Physics*, vol. 18, no. 6, p. 063029, 2016.

[55] I. Marvian and S. Lloyd, "Universal quantum emulator," *arXiv preprint arXiv:1606.02734*, 2016.

[56] D. Maslov and Y. Nam, "Use of global interactions in efficient quantum circuit constructions," *New Journal of Physics*, vol. 20, no. 3, p. 033018, 2018.

[57] D. C. McKay, T. Alexander, L. Bello, M. J. Biercuk, L. Bishop, J. Chen, J. M. Chow, A. D. Córcoles, D. Egger, S. Filipp, J. Gomez, M. R. Hush, A. Javadi-Abhari, D. Moreda, P. Nation, B. Paulovicks, E. Winston, C. J. Wood, J. Wootton, and J. M. Gambetta, "Qiskit backend specifications for openqasm and openpulse experiments," *ArXiv*, vol. abs/1809.03452, 2018.

[58] T. S. Metodi, D. D. Thaker, A. W. Cross, F. T. Chong, and I. L. Chuang, "Scheduling physical operations in a quantum information processor," in *Quantum Information and Computation IV*, vol. 6244. International Society for Optics and Photonics, 2006, p. 62440T.

[59] K. Mitarai and K. Fujii, "Methodology for replacing indirect measurements with direct measurements," *Physical Review Research*, vol. 1, no. 1, p. 013006, 2019.

[60] K. Mølmer and A. Sørensen, "Multiparticle entanglement of hot trapped ions," *Physical Review Letters*, vol. 82, no. 9, p. 1835, 1999.

[61] P. Murali, D. C. McKay, M. Martonosi, and A. Javadi-Abhari, "Software mitigation of crosstalk on noisy intermediate-scale quantum computers," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020, pp. 1001–1016.

[62] M. A. Nielsen and I. Chuang, "Quantum computation and quantum information," 2002.

[63] A. Omran, H. Levine, A. Keesling, G. Semeghini, T. T. Wang, S. Ebadi, H. Bernien, A. Zibrov, H. Pichler, S. Choi, J. Cui, M. Rossignolo, P. Rembold, S. Montangero, T. Calarco, M. Endres, M. Greiner, V. Vuletić, and M. D. Lukin, "Generation and manipulation of schrödinger cat states in rydberg atom arrays," *Science*, vol. 365, pp. 570 – 574, 2019.

[64] A. Paler, O. Oumarou, and R. Basmadjian, "Constant depth bucket brigade quantum ram circuits without introducing ancillae," *arXiv preprint arXiv:2002.09340*, 2020.

[65] G. Paraoanu, "Microwave-induced coupling of superconducting qubits," *Physical Review B*, vol. 74, no. 14, p. 140504, 2006.

[66] E. Pednault, J. A. Gunnels, G. Nannicini, L. Horesh, and R. Wisnieff, "Leveraging secondary storage to simulate deep 54-qubit sycamore circuits," *arXiv preprint arXiv:1910.09534*, 2019.

[67] J. Preskill, "Quantum computing in the nisq era and beyond," *Quantum*, vol. 2, p. 79, 2018.

[68] G. A. Quantum and collaborators, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, pp. 505–510, 2019.

[69] S. Rasmussen, K. Groenland, R. Gerritsma, K. Schoutens, and N. Zinner, "Single-step implementation of high-fidelity n-bit toffoli gates," *Physical Review A*, vol. 101, no. 2, p. 022308, 2020.

[70] C. Rigetti and M. Devoret, "Fully microwave-tunable universal gates in superconducting qubits with linear couplings and fixed transition frequencies," *Physical Review B*, vol. 81, no. 13, p. 134507, 2010.

[71] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, "Evaluating analytic gradients on quantum hardware," *Physical Review A*, vol. 99, no. 3, p. 032331, 2019.

[72] M. Schuld, M. Fingerhuth, and F. Petruccione, "Implementing a distance-based classifier with a quantum interference circuit," *arXiv preprint arXiv:1703.10793*, 2017.

[73] M. Schuld and N. Killoran, "Quantum machine learning in feature hilbert spaces," *Physical review letters*, vol. 122, no. 4, p. 040504, 2019.

[74] M. Schuld and F. Petruccione, *Supervised learning with quantum computers*. Springer, 2018, vol. 17.

[75] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, "Procedure for systematically tuning up cross-talk in the cross-resonance gate," *Physical Review A*, vol. 93, no. 6, p. 060302, 2016.

[76] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM review*, vol. 41, no. 2, pp. 303–332, 1999.

[77] C. Song, K. Xu, W. Liu, C. Yang, S. Zheng, H. Deng, Q. Xie, K. Huang, Q. Guo, L. Zhang, P. Zhang, D. Xu, D. Zheng, X. Zhu, H. Wang, Y. Chen, C.-Y. Lu, S. Han, and J. Pan, "10-qubit entanglement and parallel logic operations with a superconducting circuit." *Physical review letters*, vol. 119 18, p. 180511, 2017.

[78] A. Sørensen and K. Mølmer, "Quantum computation with ions in thermal motion," *Physical review letters*, vol. 82, no. 9, p. 1971, 1999.

[79] R. J. Spiteri, M. Schmidt, J. Ghosh, E. Zahedinejad, and B. C. Sanders, "Quantum control for high-fidelity multi-qubit gates," *New Journal of Physics*, vol. 20, no. 11, p. 113009, 2018.

[80] Y. Takahashi and S. Tani, "Collapse of the hierarchy of constant-depth exact quantum circuits," *computational complexity*, vol. 25, no. 4, pp. 849–881, 2016.

[81] Y. Takahashi, S. Tani, and N. Kunihiro, "Quantum addition circuits and unbounded fan-out," *arXiv preprint arXiv:0910.2530*, 2009.

[82] Y. Takahashi, T. Yamazaki, and K. Tanaka, "Hardness of classically simulating quantum circuits with unbounded toffoli and fan-out gates," *Quantum Information & Computation*, vol. 14, no. 13-14, pp. 1149–1164, 2014.

[83] D. Venturelli, M. Do, E. Rieffel, and J. Frank, "Compiling quantum circuits to realistic hardware architectures using temporal planners," *Quantum Science and Technology*, vol. 3, no. 2, p. 025004, 2018.

[84] G. Verdon, M. Broughton, and J. Biamonte, "A quantum algorithm to train neural networks using low-depth circuits," *arXiv preprint arXiv:1712.05304*, 2017.

[85] Y. Wang, S. Crain, C. Fang, B. Zhang, S. Huang, Q. Liang, P. H. Leung, K. R. Brown, and J. Kim, "High-fidelity two-qubit gates using a mems-based beam steering system for individual qubit addressing," *arXiv preprint arXiv:2003.12430*, 2020.

[86] Y. Wang, M. Um, J. Zhang, S. An, M. Lyu, J.-N. Zhang, L.-M. Duan, D. Yum, and K. Kim, "Single-qubit quantum memory exceeding ten-minute coherence time," *Nature Photonics*, vol. 11, no. 10, pp. 646–650, 2017.

[87] N. Wiebe, A. Kapoor, and K. M. Svore, "Quantum deep learning," *arXiv preprint arXiv:1412.3489*, 2014.

[88] N. Wiebe and L. Wossnig, "Generative training of quantum boltzmann machines with hidden units," *arXiv preprint arXiv:1905.09902*, 2019.

[89] W. K. Wootters and W. H. Zurek, "A single quantum cannot be cloned," *Nature*, vol. 299, no. 5886, pp. 802–803, 1982.

[90] Y. Wu, S.-T. Wang, and L.-M. Duan, "Noise analysis for high-fidelity quantum entangling gates in an anharmonic linear paul trap," *Physical Review A*, vol. 97, no. 6, p. 062325, 2018.

[91] X. Xu, J. Sun, S. Endo, Y. Li, S. C. Benjamin, and X. Yuan, "Variational algorithms for linear algebra," *arXiv preprint arXiv:1909.03898*, 2019.

[92] D. Yu, Y. Gao, W. Zhang, J. Liu, and J. Qian, "Scalability and high-efficiency of an $(n+1)$-qubit toffoli gate sphere via blockaded rydberg atoms," *arXiv preprint arXiv:2001.04599*, 2020.

[93] S. Zaiser, T. Rendler, I. Jakobi, T. Wolf, S.-Y. Lee, S. Wagner, V. Bergholm, T. Schulte-Herbrüggen, P. Neumann, and J. Wrachtrup, "Enhancing quantum sensing sensitivity by a quantum memory," *Nature communications*, vol. 7, p. 12279, 2016.

[94] B. Zeng, D. Zhou, and L. You, "Measuring the parity of an n-qubit state," *Physical review letters*, vol. 95, no. 11, p. 110502, 2005.