Parallel-Driving for Fast Quantum Computing Under Speed Limits

Evan McKinney
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
evm33@pitt.edu

Chao Zhou
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
chz78@pitt.edu

Mingkang Xia
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
mix20@pitt.edu

Michael Hatridge
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
hatridge@pitt.edu

Alex K. Jones
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
akjones@pitt.edu

ABSTRACT

Increasing quantum circuit fidelity requires an efficient instruction set to minimize errors from decoherence. The choice of a two-qubit (2Q) hardware basis gate depends on a quantum modulator's native Hamiltonian interactions and applied control drives. In this paper, we propose a collaborative design approach to select the best ratio of drive parameters that determine the *best basis gate* for a particular modulator. This requires considering the *theoretical computing power of the gate* along with the *practical speed limit* of that gate, given the modulator drive parameters. The practical speed limit arises from the couplers' tolerance for strong driving when one or more pumps is applied, for which some combinations can result in higher overall speed limits than others. Moreover, as this 2Q basis gate is typically applied multiple times in succession, interleaved by 1Q gates applied directly to the qubits, the speed of the 1Q gates can become a limiting factor for the quantum circuit, particularly as the pulse length of the 2Q basis gate is optimized. We propose *parallel-drive* to drive the modulator and qubits simultaneously, allowing a richer capability of the 2Q basis gate and in some cases for this 1Q drive time to be absorbed entirely into the 2Q operation. This allows increasingly short duration 2Q gates to be more practical while mitigating a significant source of overhead in some quantum systems. On average, this approach can decrease circuit duration by 17.8% and decrease infidelity for random 2Q gates by 10.5% compared to the currently best reported basic 2Q gate, $\sqrt{\texttt{iSWAP}}$.

CCS CONCEPTS

• **Hardware** → **Quantum computation**; *Emerging tools and methodologies.*

KEYWORDS

Transpilation, Basis Gate, Weyl Chamber

ISCA '23, June 17–21, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0095-8/23/06…$15.00
https://doi.org/10.1145/3579371.3589075

ACM Reference Format:
Evan McKinney, Chao Zhou, Mingkang Xia, Michael Hatridge, and Alex K. Jones. 2023. Parallel-Driving for Fast Quantum Computing Under Speed Limits. In *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23), June 17–21, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3579371.3589075

1 INTRODUCTION

Quantum Computers (QCs) leverage quantum superposition and entanglement which, unlike classical computers, allows the QC core computing element, or *qubit*, to conceptually interact with all other qubits, simultaneously. This provides the promise of solving problems that that are intractable for classical computers. Currently realized QCs are part of the Noisy Intermediate-Scale Quantum (NISQ) era. NISQ machines with more than a hundred qubits can be readily created [1]; however, the qubit interactions remain limited to small neighborhoods and these quantum operations—or quantum *gates*—have limited fidelity. While these "noisy" quantum operations continue to improve, even the best gates typically do not exceed 99.9% fidelity [2–5].

Quantum interactions are realized through qubit-qubit coupling. Coupling is possible when there is a physical connection between the qubits and is governed by a *modulator*. These modulators range from as simple as capacitive couplings to more elaborate nonlinear circuits [1, 6, 7]. The major source of error in superconducting QC hardware, which is at the heart of machines by IBM and Google, comes from qubit decoherence. Thus, continued improvements in quantum gate capabilities and speeds are required to increase feasible circuit depth.

A critical component to building better quantum circuits is to identify the *best basis gate* that can be realized by the modulator. The reason for selecting a single basis gate is that calibrating gates is an expensive process. Otherwise, one could just calibrate every single possible gate, or at least each gate that is required for a particular quantum workload. Also, gates must to be calibrated independently between each pair of qubits as each pair will require different parameters and frequencies to be addressed uniquely. Moreover, gate calibrations are finicky processes that drift over time, requiring periodic re-calibration [8]. Thus, a single gate calibration or determining multiple gate parameters as a function of the calibrated gate is necessary to make this process tractable [9].

However, the metric for determining the best basis gate may not be clear. A standard metric for determining the quality of a gate is to

calculate its Haar score, which is its aggregate coverage of all two-qubit gates, as represented in a 3D space by the Weyl Chamber [10]. While this is a good representation of the computational power of the gate, many quantum algorithms tend to be reduced to the CNOT family of gates to complete their computational work [11]. The remainder of the circuit typically requires the non-entangling SWAP gate, primarily to move data on the machine's interconnection topology. Thus, a basis gate that best optimizes these two operations of CNOT and SWAP is also a useful metric.

In this work, we consider parametrically driven interactions, in which far off-resonant drives activate an effective two-body interaction between a pair of qubits [12–14]. To create the set of all possible basis gates we explore the coupler in the form of its Hamiltonian expression and the drive parameters that can be used to tune different 2Q basis gates. By adjusting the ratio of gain and conversion terms of the Hamiltonian that naturally implements the iSWAP family of gates, it is possible to also *directly implement* the CNOT and B families of gates, as well as other more exotic gates (see Section 2 below). However, the pulse times to implement these gates is a function of the drive capacity of the modulator.

To further complicate basis gate selection, these parametrically-driven gates depend on an actuator/modulator that has an inherent *speed limit* due to factors such as fridge heating or disrupting parametric coupling [15]. The parametric drive terms, *e.g.*, the gain and conversion terms, both contribute towards the speed limit, but combine in a non-linear way. Thus, finding the fastest basis gate can become an optimization function of both the *theoretical computing power* of the gate and the *pulse time* of that gate due to the physical speed limit of that particular ratio of drive parameters.

Moreover, when approaching gate decomposition of the selected basis gate, current approaches only consider the frequency and speed of two-qubit (2Q) gates. For more accurate analysis, it is important to consider the inherent speed limits of the device. For instance, previous work [16] has shown that using a continuum of a basis gate can achieve increasingly efficient decomposition. However, if a smaller duration basis gate is used, it must be balanced against the increased relative cost of one-qubit (1Q) gates required for decomposition using traditional techniques. Figure 1 shows that using increasingly small fractional basis gates for decomposition remains practical up to a limit of $\sqrt{\texttt{iSWAP}}$ shown for examples of 1Q gates consuming either 10% (red dotted line) or 25% (blue dotted line) the duration of a 2Q gate.

This paper introduces *a collaboratively designed quantum architecture* that includes *a new methodology* for analyzing gate costs to select the appropriate 2Q basis gate while proposing a novel *parallel-drive* technique that drives both the modulator and the qubits directly. The parallel-drive approach allows the selected 2Q basis gate to increase in Haar volume, and in some cases, can eliminate the need for interspersed 1Q gates improving decomposition and overall circuit durations.

Parallel-drive is illustrated in Figure 2 where the technique turns the straight Cartan trajectories of 2Q gates into curves in the Weyl Chamber, eliminating some of the vertices represented 1Q gates. Using parallel-drive it is possible to calibrate two basis gates, one with and without parallel-drive, to realize the benefit of both a smaller duration 2Q basis gate as well as eliminating many 1Q gates. In particular, this paper makes the following contributions:

- We characterize simultaneous application of two basic parametric interactions to implement 2Q gates and articulate the various 2Q gate families that can be realized.
- We observe that partially pulsed gates, *e.g.*, $\sqrt{\texttt{iSWAP}}$, can be more efficient than the full pulse gate, *e.g.*, iSWAP. However, when using smaller fractions of a gate, the overhead of the 1Q gates becomes more significant.
- We present a parallel-drive methodology to improve the agility of a basis gate by concurrently driving the modulator and the participating qubits. We show that parallel-drive can improve the computing capability of a basis gate and provide the potential to remove interleaved 1Q gates in repeated application of 2Q gates.

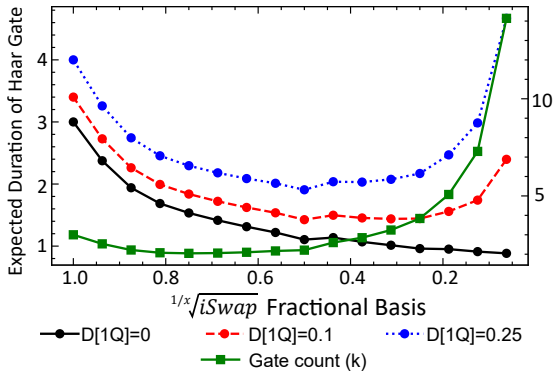
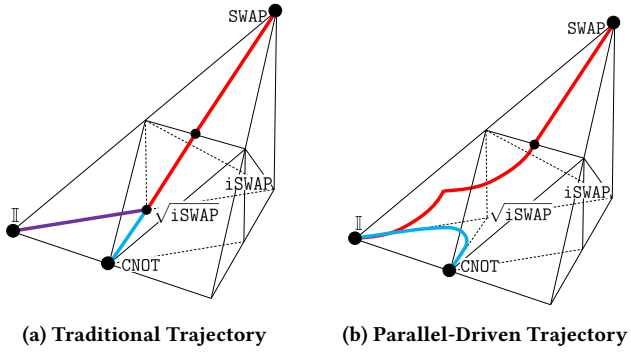


Figure 1: The optimal Haar gate in the iSwap-family changes as function of the 1Q gate times. This is because the number of gate applications increases for smaller cost 2Q basis gates, which trades off with interleaved 1Q layers. $\sqrt{\texttt{iSWAP}}$ minimizes duration costs for non-negligible 1Q gates.



(a) Traditional Trajectory **(b) Parallel-Driven Trajectory**

Figure 2: Cartan trajectories [17] for CNOT (blue) and SWAP (red) using $\sqrt{\texttt{iSWAP}}$ basis. Trajectories represent the total accumulated unitary transformation over time, beginning at Identity $\mathbb{I}$ and ending at the target gate U_T. Black dots represent interleaved 1Q gates where orientation can be changed.

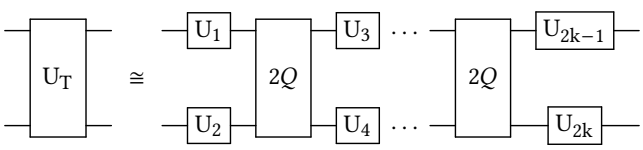


Figure 3: Generic decomposition 2Q unitary ← 2Q basis gate.

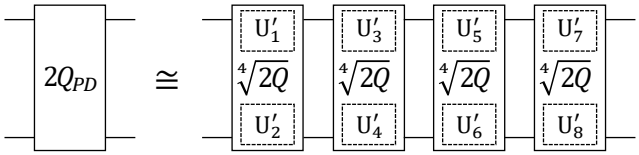


Figure 4: Parallel-drive decomposition for a 2Q unitary

- We present a detailed study of using speed limits and parallel-drive to reduce circuit delay for important quantum computing workloads including an improved decomposition equivalency for iSWAP and CNOT built with $\sqrt[4]{\texttt{iSWAP}}$.

In the next section we explore the basis gate design space from a modulator by exploring parameters of the Hamiltonian.

2 HAMILTONIAN DESIGN SPACE

Fundamentally, quantum gates are unitary matrix operations, or unitaries, that act on quantum states. In general, 1Q and 2Q gates form the building blocks of quantum circuits [11]. A native quantum gate set, analogous to a classical computer's instruction set, defines which unitary operations are available to use on a machine. The available gates depends on the engineered Hamiltonian of the system, which is related to the unitary, described by Schrödinger's equation, $U(t) = e^{-i\hat{H}t/\hbar}$. In superconducting QCs, parametric driving on a qubit-coupling mechanism provides control over the Hamiltonian to activate the desired unitary and corresponding gate.

Using Cartan's KAK decomposition [18, 19], an arbitrary 2Q gate can be built from repeated applications of a universal 2Q basis gate with interleaved 1Q gates (Figure 3). Simple techniques for gate decomposition use this interleaving template and via an exact analytical solution [20, 21], or an approximate numerical optimizer [22–24], find a solution to the 1Q gates for a variable number of repetitions. We refer to a *basis template* as a quantum circuit that interleaves the basis gate K times. To perform decomposition, the template is instantiated with the sufficient size K.

Crucially, the proper selection of basis gate determines the overall complexity of the transpiled quantum algorithm, as different basis gates may require comparatively larger or smaller K in decomposition. Moreover, each basis gate has a latency depending on the system's physical interactions. **For this reason, characterizing the set of candidate basis gates requires reasoning about both their decomposition efficiency, $K[U_B]$, as well as their hardware latency, $D[U_B]$.**

As discussed in Section 1, the set of possible 2Q gates is represented geometrically by the Weyl Chamber [25–27], where locally-equivalent 2Q gates, differing only by 1Q gates, are mapped to the same coordinate.

This comes from the assumption that any *locally equivalent* gate for a particular 2Q gate has the same entangling power and decomposition efficiency. For example, CZ and CX/CNOT can be considered the same equivalent gates in this context. Also, the unitary conjugates are reflected over the x-axis, which is like executing the gate backwards, so essentially it is only necessary to plot gates on the left side of the chamber. In this work, references to a 2Q basis gate may refer generally to the set of locally equivalent gates with matched computational power; however, references to 2Q target gates include the additional 1Q costs of local basis translation required for algorithm correctness. To reason about decomposition, we plot a gate's coverage volume, which are regions that span all gates buildable by a template. The use of monodromy polytopes [10] analytically creates the coverage sets, so we can reason about spanning volumes of K gate applications, decide if a gate is contained in a template, and output its weighted volume.

Revisiting Figure 2, conceptually, the pair of 1Q gates orients the trajectory in a particular direction, then the 2Q basis gate traverses the Weyl Chamber to a new point. Using traditional decomposition (Figure 2a) to implement a CNOT or SWAP functionality using $\sqrt{\texttt{iSWAP}}$ follows the template in Figure 3. The first leg in the Weyl Chamber is identical, shown as purple. As the point of interest was not yet reached, the direction is re-oriented (1Q gates) before drawing the next line. CNOT is reached in two steps, but the process repeats for SWAP until the point is reached, which happens on the third step. This process is akin to a car driving on established straight roads, including intersections at fixed points where the car can select new roads to follow.

Parallel-drive replaces a 2Q gate with several partially pulsed 2Q gates, which have unique simultaneous drives to the qubits, shown in Figure 4 for four time steps. This allows the trajectories to become curves, akin to steering while driving, which allows U_3 and U_4 to be eliminated when implementing CNOT and SWAP from a parallel-drive iSWAP family of gates (Figure 2b).

2.1 Flexible Realization of Gates with Parametric Couplings

Parametric or tunable couplers in superconducting qubit systems can be driven with external flux and/or microwave fields to create a wide variety of two-qubit gates [14, 17, 28–30]. Two families of interactions used in parametric amplification are photon exchange and two-mode squeezing/gain, which can naturally realize iSWAP gates among two-level or anharmonic qubits. These couplers provide a versatile and flexible way to implement various gate operations, making them a valuable tool in superconducting qubit systems [31].

These couplers can be driven to produce a wide variety of 2Q gates, especially those in the Weyl Chambers' floor, which are just combinations of simultaneous gain and conversion driving[1]. We can write such a combination Eq. 1,

$$\hat{H} = g_c(e^{i\phi_c}a^\dagger b + e^{-i\phi_c}ab^\dagger) + g_g(e^{i\phi_g}ab + e^{-i\phi_g}a^\dagger b^\dagger), \quad (1)$$

[1]Note, careful attention must be paid to the nonlinearity and encoding of the qubit states being used. For instance, the same parametric interaction among qubits realized as transmons produces different gates than high-Q cavities, for which the latter produces Fock states.

where g_c, g_g and ϕ_c, ϕ_g represent the pump-controlled amplitude and phase, respectively, such that g_c, ϕ_c result from difference (conversion) driving and g_g, ϕ_g are from sum (gain) driving.

Each choice of control parameters yields a continuum of gates. To illustrate the flexibility of jointly driving multiple interactions simultaneously, the case with both couplings are at non-zero strength and both pump phases are set to zero arrives at the following unitary:

$$U(t) = \begin{bmatrix} \cos\theta_g & 0 & 0 & -i\sin\theta_g \\ 0 & \cos\theta_c & -i\sin\theta_c & 0 \\ 0 & -i\sin\theta_c & \cos\theta_c & 0 \\ -i\sin\theta_g & 0 & 0 & \cos\theta_g \end{bmatrix} \quad (2)$$

such that $\theta_c = g_c t, \theta_g = g_g t$, where t is the driving time.

By varying the interaction strengths g_c and g_g at a fixed $t = 1$. The iSWAP gate in this language is given by setting θ_c or θ_g to $\frac{\pi}{2}$, yielding Eq. 3.

$$\hat{H} = \frac{\pi}{2}(a^\dagger b + ab^\dagger) \text{ or } \hat{H} = \frac{\pi}{2}(ab + a^\dagger b^\dagger), \quad (3)$$

while the CNOT gate can be realized by setting $\theta_c = \theta_g = \frac{\pi}{2}$, yielding Eq. 4.

$$\hat{H} = \frac{\pi}{4}(a^\dagger b + ab^\dagger) + \frac{\pi}{4}(ab + a^\dagger b^\dagger). \quad (4)$$

There is a continuous set of possible unitary operators that can be naturally realized by this Hamiltonian. By visualizing this in the Weyl Chamber (Figure 5a) these two points of interest appear at both ends of the yellow band, with the iSWAP at the tip and CNOT along the baseline at the $\frac{\pi}{2}$ point. In fact, the theoretical power of this Hamiltonian covers the entire base plane of the Weyl Chamber with different points reachable in different ratios of θ_g and θ_c and total angle $\theta_g + \theta_c$.

A vital question, then, is which combination of drives yields the best gate? There are several important factors for selecting this gate such as the decompositional efficiency of the gate and the pulse time of the gate. There is evidence that fractional pulse duration gates can be more efficient (*e.g.*, $\sqrt{\texttt{iSWAP}}$ vs iSWAP), further reducing pulse time. In the next section, we detail the physical processes which enable parallel-drive using fractional basis gates.

2.2 Realization of Parallel Parametric Gates

Transmon Hamiltonians have been explored to optimize pulses for creation of specific gates or algorithms [32–35]. In our work, we modify the "Conversion-Gain" Hamiltonian discussed in the previous section by appending single-qubit X-gates with drive amplitudes $\epsilon_1(t), \epsilon_2(t)$, each described by $D[2Q]/D[1Q]$ discrete time steps (Eq. 5). Essentially this creates parallel 1Q gates to occupy the duration of the 2Q gate, each with a distinct amplitude (Figure 4).

$$\hat{H} = g_c(e^{i\phi_c} a^\dagger b + e^{-i\phi_c} ab^\dagger) + g_g(e^{i\phi_g} ab + e^{-i\phi_g} a^\dagger b^\dagger) + \epsilon_1(t)(a + a^\dagger) + \epsilon_2(t)(b + b^\dagger) \quad (5)$$

By allowing this extension there are two important outcomes that reduce the overall circuit latency: (1) the basis gate coverage region can be enriched and (2) 1Q gates and their sequential delay may be able to be *absorbed* into the 2Q gate operation, to improve overall circuit time.

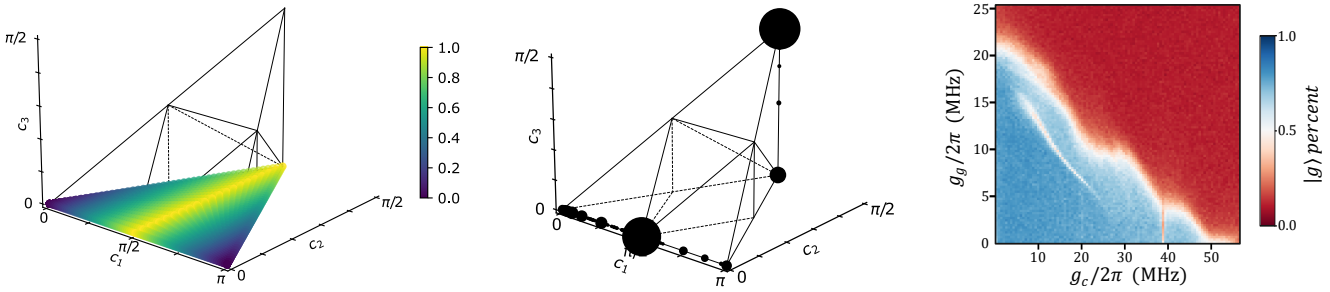
We propose a calibration strategy that sets the basis gate to the same pulse duration of a 1Q gate. Specifically, for the Superconducting Nonlinear Asymmetric Inductive eLement (SNAIL) modulator [36] used in our prototyping experiments, we recommend using an $\sqrt[4]{\texttt{iSWAP}}$ gate, which satisfies $\mathrm{D}[1Q] = 0.25 = \mathrm{D}[2Q]$. Parallel-drive does introduce a frequency Kerr-shift on the qubits, which requires adjusting the parallel-driven gate for a qubit frequency shift [14, 37]. Unlike optimal-control protocols that require complex calibration due to the continuously changing control requirements, parallel-drive activates/deactivates qubit drive pulses simultaneously during these discrete time steps. Thus, calibration requires tuning two gates, regular $\sqrt[4]{\texttt{iSWAP}}$ and $\sqrt[4]{\texttt{iSWAP}}$ under this constant frequency shift under parallel-drive.

This frequency shift is proportional to the fourth-order coupling term. Existing calibration schemes, such as interleaved-randomized and cross-entropy benchmarking, can fine-tune frequency drives for non-Clifford gates without affecting overhead. Modulators like the SNAIL utilize third-order coupling while minimizing the fourth-order coupling term to avoid frequency crowding such that parallel-drive should not sacrifice qubit fidelity. Unfortunately, parallel-drive may create additional crosstalk in the IBM cross-resonance gate due to its dependence on fourth-order coupling. However, the new IBM initiative to build machines with parametric couplers may allow for high-fidelity parallel-drive. In the next section we discuss methods to evaluate this decomposition efficiency.

2.3 Gate Score Methodology

In order to optimize the choice of control parameters, it is first necessary to reason about the unitaries' decomposition efficiency. We compare two methodologies to quantify the decomposition efficiency of a gate: uniform gate distribution and algorithm-sampled distributions. While decomposition determines the number of iterations required of a basis gate to realize a target unitary U_T, recall from Figure 5a, different realizable gates from the Hamiltonian require different pulse times. To represent both aspects of a gate we define $K_{U_B}[U_T]$ as the number of basis gates (U_B) to build the target (U_T) and $D_{U_B}[U_T]$ as the normalized duration to build a target using the basis. The expectation $\mathbb{E}$ operator signifies the cost averaged over the random Haar distribution.

The Haar measure [38], is used to construct a uniform distribution of 2Q gates. Conceptually, it is a density function inside the Weyl Chamber which weights the perfect entangler interior region more heavily than the exterior $\mathbb{I}$ (identity) and SWAP vertices. It is used to build a Haar score, a common metric to quantify the decomposition power of a basis set. The Haar score is the expected number of gates (K) to generate Haar random 2Q gates. In other words, it is a volume-weighted average over the basis template's spanning regions to achieve full Weyl Chamber coverage. This is demonstrated in Figure 6, by plotting the K-template spanning region for some popular 2Q gates. The iSWAP gate (Figure 6a) can reach the bottom plane in $k = 2$ and the entire volume in $k = 3$. The $\sqrt{\texttt{iSWAP}}$ gate (Figure 6b) actually has better coverage at $k = 2$ with a shorter pulse time. The popular CNOT (Figure 6c) has similar coverage behavior as iSWAP, which is reasonable as both are Clifford gates. The B gate (Figure 6e) minimizes $\mathbb{E}$[Haar] because it spans the entirety of the region in $k = 2$ (green), whereas $\sqrt{\texttt{CNOT}}$



(a) Set of gates natively produced by conversion and gain parametric driving. The color bar indicates $\theta_g + \theta_c$, normalized by $\pi/2$.

(b) Frequency of gates from set of 16-qubit benchmarks transpiled onto an 4×4 square lattice topology.

(c) Demonstration of limitation of gain and conversion coefficients (g_g and g_c) when both processes are turned on.

Figure 5: Analysis of basis gate choice including, gate range and timing, gate usage by application, and impact of drive ratio.

(Figure 6d) does not completely span the chamber until six steps ($k = 6$, yellow). $\sqrt{\text{B}}$ sits between $\sqrt{\texttt{iSWAP}}$ and $\sqrt{\texttt{CNOT}}$ requiring $k = 4$ (orange).

Unfortunately, the $\mathbb{E}[\text{Haar}]$ score fails to capture that in practice, gates are not uniformly distributed. Algorithms are written primarily using CPhase gates, implemented by *controlled* unitaries, analytically decomposed into 2Q CNOT gates. The reason why CPhase gates are ubiquitous in algorithm design may be explained by the Quantum Singular Value Transform (QSVT), a key subroutine of

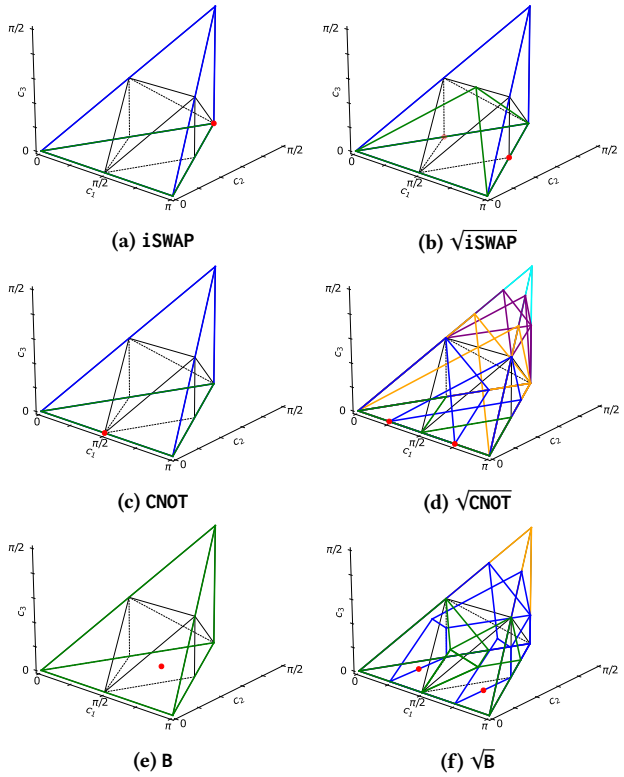


Figure 6: Gate coverage sets. red: $k = 1$, green: $k = 2$, blue: $k = 3$, orange: $k = 4$, purple: $k = 5$, cyan: $k = 6$

Grover's Search, Phase Estimation, and Hamiltonian Simulation circuits [39]. The QSVT subroutine encodes an operator $\boldsymbol{A}$ as a block inside a larger unitary matrix, $\boldsymbol{U}$. When $\boldsymbol{A}$ is a unitary matrix, then $\boldsymbol{U}$ becomes a controlled $\boldsymbol{A}$ operator, which naturally decomposes into CPhase 2Q gates [11]. Even algorithms which use different subroutines, such as Quantum Approximate Optimization Algorithm (QAOA), still rely on the CNOT for their own reasons. In QAOA, the Hamiltonian cost function maps to states that are diagonal in the computational basis, such that the canonical expansion is into ZZ gates [40]. Simply put, creating new quantum algorithms is such a difficult task, most known algorithms are variations of the same subroutine, which happens to use controlled-gate operators [41].

Moreover, qubit connectivity topologies necessitate data movement via SWAP gates. Due to the limited connection topologies of NISQ superconducting QCs of square lattice and heavy hex, SWAP gates are required to move data into qubits in the same neighborhood. It has been shown that these gates can dominate transpiled gate counts [42, 43].

The frequency of SWAP gates naturally depends on the coupling topology. For simplicity, we consider a 4x4 square lattice topology as the target coupling map. For a representative set of quantum benchmarks including QFT, QAOA, Adder, Multiplier, GHZ, Hidden Linear Function, and VQE, but excluding the special case of Quantum Volume, the workloads were mapped to this topology using the Qiskit v0.20.2 transpiler with -O3 (optimization level 3), inducing the necessary SWAPs. The results are displayed in a "shot-chart" that increases the size of the plotted gates relative to its frequency in the workloads, as shown in Figure 5b. From this experiment, the most frequent targeted gates are SWAP followed by CNOT, with iSWAP as a more distant third. Interestingly, there is a significant usage of CNOT *family* gates, which show up along the Weyl Chamber baseline.

Thus, an alternative gate scoring function introduces $V(U_B)$, which weights the decomposition cost of target, U_T, using the basis gate duration, $D_{U_B}[U_T]$, by the frequency of the target gate, for instance as shown in Figure 5b. The best basis gate would minimize this weighted cost as shown in Eq. 6.

$$V(U_B) = \sum_{U_T} f(U_T) D_{U_B}[U_T] \tag{6}$$

Table 1: Decomposition Gate Counts (k). Each value is determined by the spanning regions from Figure 6 Best value reported in blue, worst value reported in red.

	iSWAP	$\sqrt{\text{iSWAP}}$	CNOT	$\sqrt{\text{CNOT}}$	B	$\sqrt{\text{B}}$
K[CNOT]	2	2	1	2	2	2
K[SWAP]	3	3	3	6	2	4
$\mathbb{E}$[K[Haar]]	3	2.21	3	3.54	2	2.50
K[$W(.47)$]	2.53	2.53	2.06	4.12	2	3.06

As gates must typically be calibrated prior to knowing the circuits they will be programmed to implement, a simplified distribution might only consider and weight the dominating CNOT and SWAP gates, which, by extension, will generally be true for any CPhase algorithm deployed to the device. We fit the value λ as ratio of CNOT to the total of CNOT and SWAP gates using our benchmark workloads as illustrated in Figure 5b.

$$W(U_B, \lambda) = \lambda * D_{U_B}[\text{CNOT}] + (1-\lambda) * D_{U_B}[\text{SWAP}] \tag{7}$$

This ratio, $\lambda = 731/(731+828) \approx 0.47$. Therefore, the weighted function $W(U_B, 0.47)$ serves to optimize basis gate selection over quantum workload circuits. We go on to refer to this weighted distribution of gates as $W(\lambda = .47)$.

Table 1 compares the decomposition cost of the six common gates from Figure 6 in terms of number of gates to realize target gates of SWAP and CNOT, as well as Haar and our empirical W distributions. The best performing gate for Haar is the B because it can span the Weyl Chamber in $k = 2$, however, $\sqrt{\text{iSWAP}}$ and $\sqrt{\text{B}}$ perform well with $k = 2.21$ and $k = 2.5$, respectively. The W cost function requires a gate that is good at both CNOT and SWAP. While the K function is useful for reasoning about theoretical computational capabilities, the D function better compares the implementation of these gates as it considers pulse times and their impact on the decomposition cost, which we explore in the next section.

2.4 Speed Limit Scaled Duration Costs

Although each of the discussed candidate basis gates, as well as many other gates, are natively produced by conversion/gain Hamiltonians, different combinations of drives require different duration pulse sequences. It follows from Eq. 1 and Eq. 2, to realize a specific gate with fixed θ_c and θ_g, the interaction strengths $g_{c,g}$ are inversely proportional to time t. Thus, a unitary is realizable with the shortest duration when the interaction strengths are as a strong as possible.

However, in a real physical system, the effective g_c and g_g coefficients cannot be infinitely large due to physical limitations. Examples include over-driving the modulator, which can result in drive lines causing heating, instability in non-linear objects, "bright-state"-ing, bifurcation, chaos, population leakage [44], among others. Understanding the various physical mechanisms in determining these speed limits with the goal to improve them is an ongoing research effort both for parametrically driven qubit gates and the related field of parametrically driven amplifiers [37, 45, 46].

The maximum magnitude is specific to the system being used. In general, it can be described with a Speed Limit Function (SLF) which describes the valid operating range for variable drive strengths. The SLF represents the boundary of the regions where the parameters *obey the speed limit* and coupling operates correctly versus where the *speed limit was exceeded* and the unitary gate fails.

2.4.1 Characterizing Gate Speed Limits. To illustrate a concrete example of how the speed limit appears in a parametric coupling system and inform our co-design study, we swept the g_g, g_c drive strengths for a SNAIL modulator [36] coupled with transmon qubits. The gain-, and conversion-only experiments were first performed individually between a qubit and the SNAIL coupler mode to find the maximum g_c when $g_g = 0$ and vice versa. This calibrates the relations between the drive amplitudes and the g coefficients. Then, the pumps were detuned from the on-resonance frequencies (so that the drive affects the SNAIL but we perform no two-qubit gate) and applied simultaneously to the SNAIL coupler at different amplitude combinations. The result of this study is shown in Figure 5c.

To monitor the speed limit, which manifests as a *break point* of the SNAIL coupler, a second qubit that also couples to the SNAIL mode is used. This second test qubit is prepared in the ground state and measured immediately after the gain and conversion pumps were turned off. Excitation of this second monitoring qubit signals exceeding the speed limit, in which the SNAIL coupler transitions to a (at present poorly understood) chaotic behavior and creates photons in both itself and coupled modes, illustrated by the red region in the Figure 5c. The blue region indicates the monitoring qubit remained in the ground state and represents our proxy for all the feasible g_g and g_c combinations that can be used to construct 2Q gates. Finally, the SLF of interest is illustrated as the boundary between the blue coupling region and the red non-coupling region, shown as the white line. A few characteristics of interest from Figure 5c: first, g_c can be driven much harder than g_g and second, the SLF is non-linear.

To capture this experimental information for determining the best basis gate, unitaries described by the values g_c, g_g and time t, a gate can be visualized as a line from the origin with the same g_c to g_g ratio that intersects with the SLF to define g_c^{max} and g_g^{max}. The ratio of change in drive strength is accompanied by inverse scaling of t to find t^{min}. This process is described in Algorithm 1. We normalized the speed limit to eliminate any dependencies on hardware-specific gate durations by uniformly scaling it based on the fastest iSWAP intersection, either on the x-axis or y-axis. We adjust the scaling such that the first x- or y-intercept to reach $\pi/2$ sets the fastest iSWAP to $t^{min} = 1$. Now, rather than reporting $D[U_B]$ in units of time, we use units proportional to one (1) iSWAP duration, colloquially referred to as *a single pulse*. In the next section we update our decomposition analysis using the gate speed limits.

2.4.2 Circuit Decomposition Costs. Integrating the SLF into duration efficiency combines the theoretical and practical aspects of gate counts to predict circuit latency. Speed-limited duration of the same popular basis gates reported previously are contained in Table 2. Compared to the theoretical gate counts where CNOT and B both outperformed $\sqrt{\text{iSWAP}}$, the speed analysis explains why, in practice, $\sqrt{\text{iSWAP}}$ becomes the more optimized basis gate, as $\sqrt{\text{iSWAP}}$ has lowest consistent pulse cost for Haar score (1.11) and the best weighted W score (1.27).

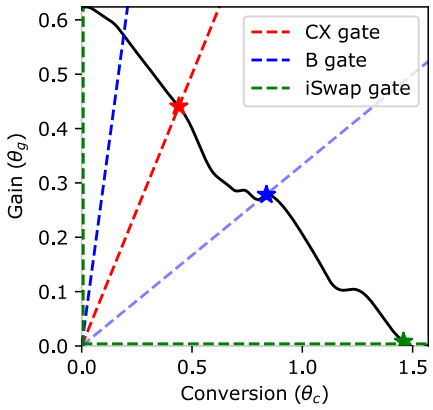


Figure 7: Hardware speed limit characterization. For each gate, drive parameters are optimized by finding the intersection of the speed limit with the gate's conversion and gain drive rates.

Table 2: Decomposition duration. D_{Basis} is the normalized pulse duration for each candidate basis gate based on the SLF. Each decomposition score is computed using Table 1 and Algorithm 1

Basis	iSWAP	$\sqrt{\text{iSWAP}}$	CNOT	$\sqrt{\text{CNOT}}$	B	$\sqrt{\text{B}}$
	SNAIL Characterized Speed Limit					
D_{Basis}	**1.00**	**0.50**	**1.78**	**0.89**	**1.40**	**0.70**
D[CNOT]	2.00	1.00	1.78	1.78	2.81	1.41
D[SWAP]	3.00	1.50	5.35	5.35	2.81	2.81
$\mathbb{E}$[Haar]	3.00	1.11	5.35	3.17	2.81	1.76
D[W(.47)]	2.53	1.27	3.67	3.67	2.81	2.15

To find the best basis gate for implementing these target unitaries, the speed limit functions are plotted in Figure 7. As the gate family is defined by the ratio between g_c and g_g terms, we plot the ratios from the origin for the CNOT (CX) (blue dotted lines), B (red dotted line) and iSWAP (green dotted lines, which are along the x-axis [conversion] and y-axis [gain]), gate families.

Algorithm 1 Scale gate scores using speed limit function

Input: $SLF, U_B(\theta_c, \theta_g), K_{U_B}[U_T], D[1Q]$
Find the largest g_c and g_g which produces the input U
$\beta \leftarrow \theta_g/\theta_c$
Find intersection of $g_g = \beta g_c$ with $SLF(g_c)$ by solving
$\begin{cases} g_g^{max} = \beta g_c^{max} \\ g_g^{max} = SLF(g_c^{max}) \end{cases}$
Scaling time using updated strengths
$t^{min} \leftarrow \theta_c/g_c^{max}$
Scale decomposition cost by duration
$D_{U_B}[U_T] \leftarrow K_{U_B}[U_T] * t^{min} + (K_{U_B}[U_T] + 1) * D[1Q]$
return $D_{U_B}[U_T]$

Interestingly, for our characterized system, because CNOT is a comparatively slower gate in terms of pulse length, it is actually faster to realize CNOT using 2 $\sqrt{\text{iSWAP}}$ gates rather than a directly calibrated CNOT in the modulator. Due to the strong preference for conversion drives in the SNAIL, iSWAP is the obvious choice for a basis gate. However, recalling Table 2, basis gates from the same gate family can yield significantly different results depending on the pulse length, *e.g.*, $\sqrt{\text{iSWAP}}$. The preference for conversion drives may not always hold true in some iSWAP modulators, *including* the SNAIL device, where the behavior depends on the characteristics of the different resonant frequencies of the device. Therefore, moving forward we utilize a more generic characterization, where $g_c + g_g \leq L$, resulting in a Linear SLF of $g_g = L - g_c$.

Furthermore, we must keep in mind from decomposition rules (Figure 3), templates include interleaved 1Q gates. Our results indicate that for negligible 1Q gate duration, the optimal basis gate is much closer to Identity $\mathbb{I}$, than for non-negligible 1Q gates, which tend to be much closer to $\sqrt{\text{iSWAP}}$, CNOT, and B. In the next section we discuss the impact of basis gate selection and fractional pulse lengths as impacted by 1Q gates.

2.5 Interleaving 1Q Gates

The trend to treat 1Q gates as negligible is due to the relative simplicity to engineer single qubit interactions and as such, for them to be less likely to be a significant source of error. Prior work confirms that $\sqrt{\text{iSWAP}}$ is the more efficient basis gate when only considering 2Q gate costs [21, 43]. However, when decoherence over time is the primary source of error, we find there is an important tradeoff between faster basis gates and increased K-template lengths (Figure 1), thus, the accumulated 1Q gate count impacts total duration more for fractional basis gates.

In practice, 1Q gates can be quite fast, *e.g.* around 10% [47] the duration of the basis gate ($D[1Q] = 0.1$). In other systems with very fast 2Q gates, the 1Q gates are as much as twice as slow as the full pulse 2Q gate [2, 4], depending on the modulator. We treat all 1Q gates as having the same duration, which can be made possible using virtual Z-gates [48]. Now the overall duration of a U_B decomposition can be expressed as in Eq. 8, which sums both the 2Q and 1Q durations for K repetitions. To show the impact on decomposition, Table 3 shows the decomposition efficiency for the linear speed limit when 1Q gates are 25% of the speed of a full pulse 2Q gate. Similar calculations for other speed limits follow the same trends.

$$D_{U_B}[U_T] = K_{U_B}[U_T] t^{min} + (K_{U_B}[U_T] + 1) D[1Q] \tag{8}$$

The total circuit delay can be calculated as in described in Eq. 9, where the pulse delay from Eq. 8 is summed for all gates on the critical path of the full circuit.

$$D_{U_B}[\text{Circuit}] = \sum_{U_T \text{ on Critical Path}} D_{U_B}[U_T] \tag{9}$$

Based on the results of this analysis, it is clear that for a linear speed limit, $\sqrt{\text{iSWAP}}$ is the most duration optimized basis gate (Figure 1). Furthermore, this methodology offers useful insights for experimentalists when constructing their own basis gates, given their own Hamiltonian design-space and 1Q gate speeds. This is especially pertinent, demonstrated by our hardware speed limit,

Table 3: Decomposition duration with $D[1Q] = 0.25$. Each value is computed using Eq. 8 with linear SLF.

	iSWAP	$\sqrt{\text{iSWAP}}$	CNOT	$\sqrt{\text{CNOT}}$	B	$\sqrt{\text{B}}$
D[CNOT]	2.75	1.75	1.50	1.75	2.75	1.75
D[SWAP]	4.00	2.50	4.00	4.75	2.75	3.25
$\mathbb{E}$[D[Haar]	4.00	1.91	4.00	2.91	2.75	2.13
D[W(.47)]	3.41	2.15	2.83	3.34	2.75	2.55

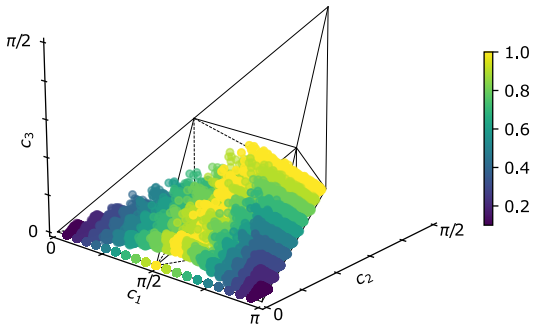


Figure 8: Gates natively produced by conversion and gain parametric driving with parallel 1Q gate drives ($K = 1$). Color bar indicates the sum of the g_c and g_g, normalized to $\pi/2$.

when there is a strong preference to using one kind of interaction. Moreover, as fraction of the 2Q pulse becomes shorter (approaches Identity $\mathbb{I}$), the 1Q gate duration becomes more significant such that the theoretical limit of small pulses of iSWAP becomes capped, with the limit remaining closer to $\sqrt{\text{iSWAP}}$. In the next section, we introduce parallel-driven gates as a means of improving the basis gate coverage volumes, and consequently reducing duration costs by effectively executing the 2Q and 1Q gates simultaneously.

3 PARALLEL 1Q DRIVE FOR BASIS OPTIMIZATION

In the previous section, we explored the use of variable conversion/gain drives to achieve different basis gates for varying pulse lengths, with optimized driving power ratios determined by the speed limit. Another method to achieve a selection of basis gates is by simultaneously driving the qubits involved in the 2Q gate, which enables part of the "steering" work of the interleaved 1Q layers to be carried out in parallel during the 2Q gate operation. This is possible because the drive to the modulator governing the 2Q interactions is distinct from the drive to the qubits implementing the 1Q gates. With the "parallel-drive" approach, only one 2Q basis gate needs to be calibrated in two conditions: with and without 1Q qubit driving.

3.1 Computing Parallel-Drive Coverage Sets

To demonstrate the additional computing capabilities of basis gates using parallel-driven 1Q gates, we first show the increased set of primitive basis gates ($K = 1$) found by sweeping the free variables of the Hamiltonian (Eq. 5), plotted in Figure 8 (compared to Figure 5a). The important outcome is that the parallel-driven basis gates extend off the bottom plane into the volume of the Weyl Chamber, which guarantees that our basis templates with parallel-drive, K', will be able to build some targets with fewer iterations than without, K_0, *e.g.*, $K' \leq K_0$. This translates to an advantage in Haar Score.

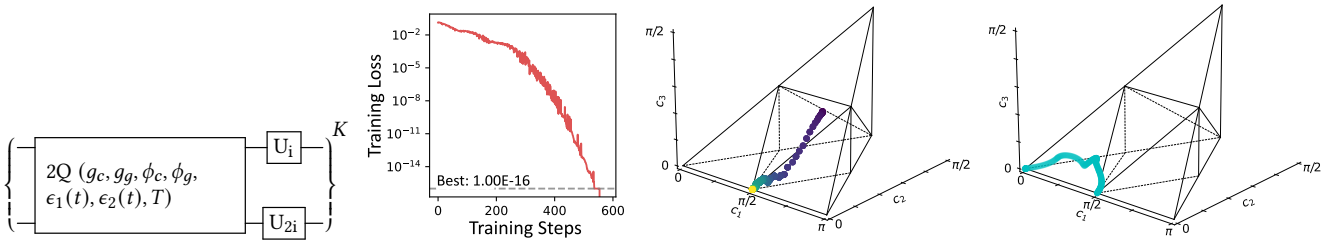
As the analytical volume coverage calculation of monodromy polytopes cannot support 2Q gates with parallel-drive, we developed a numerical procedure to estimate the expanded regions. First the coverage region is seeded with randomized parameters. Second, based on the loosely articulated region, target points outside the region are identified to help clarify the outer boundaries. Using an optimizer to these target points a point at the outside edge of the region can be found. The approach is described in Algorithm 2, which constructs a polytope using the `lrs` [49] backend via the convex hulls defined by a set of Weyl Chamber coordinates. In order to preserve convexity, we partition the coordinate list into left and right sides of the Weyl Chamber ($c_1 = \pi/2$), with convex hulls created separately. Finally, we specifically target exterior points $\mathbb{I}$, CNOT, iSWAP and SWAP as these gates are unlikely to be reached via Haar uniform randomization. To optimize the template to reach each exterior point, we adapt the strategy from previous work [43, 50], and use the Nelder-Mead optimization method [51] with a Makhlin invariant functional approach [33, 52]. The free variables are phase ($\phi_{c,g}$) and 1Q drive amplitude ($\epsilon_1(t), \epsilon_2(t)$) bounded by $(0, 2\pi)$, for each K iterations of the template, as shown in Figure 9a.

We consider four discrete 1Q drive time steps when building the extended volumes. This corresponds $D[1Q] = 0.25$, for a full pulse iSWAP, hence $D[2Q] = 1$. Previous work has explored driving 1Q gates with many more time steps [32]; however, in our experimentation, four time steps provides sufficiently similar coverage sets as compared to 250 time steps, but in a more reasonable computing time. For example, Figure 9b plots the norm training loss of an iSWAP basis converging to CNOT in less than 600 iterations, and Figure 9c plots the updated coordinate after each iteration. The final (yellow) coordinate successfully converges to the target CNOT at $(\pi/2, 0, 0)$. Note that arbitrarily small error is possible with increased training iterations and 1Q time steps. In the next section, we explore the impact of parallel-drive on Weyl Chamber coverage.

3.2 Impact of Parallel-Drive on Decomposition

The extended volumes for each of the six comparative basis gates are reported in Figure 10. The first major difference from the traditional coverage sets (Figure 6) is that the $K = 1$ in red has increased from being only local to the basis gate into a non-zero volume. Second, each K spanning region is a superset of its original coverage volumes. Third, no gate reaches 100% coverage in less template repetitions than before, which highlights the inherent difficulty of optimizing the SWAP gate.

Using the same procedure as before, the coverage volumes can be used to find the K and D costs shown in Tables 4 and 5, respectively. Note, internal 1Q gates are unnecessary if the target gate is identical to multiple fractional copies of the basis gate, *e.g.*, iSWAP formed from two $\sqrt{\text{iSWAP}}$s. While, this seems trivial for traditional circuits where a straight line in the Weyl Chamber continues twice as far, when adding parallel-drive, this property becomes more important as this line becomes a volume in the Weyl Chamber, providing more opportunities to eliminate interleaved 1Q gates.

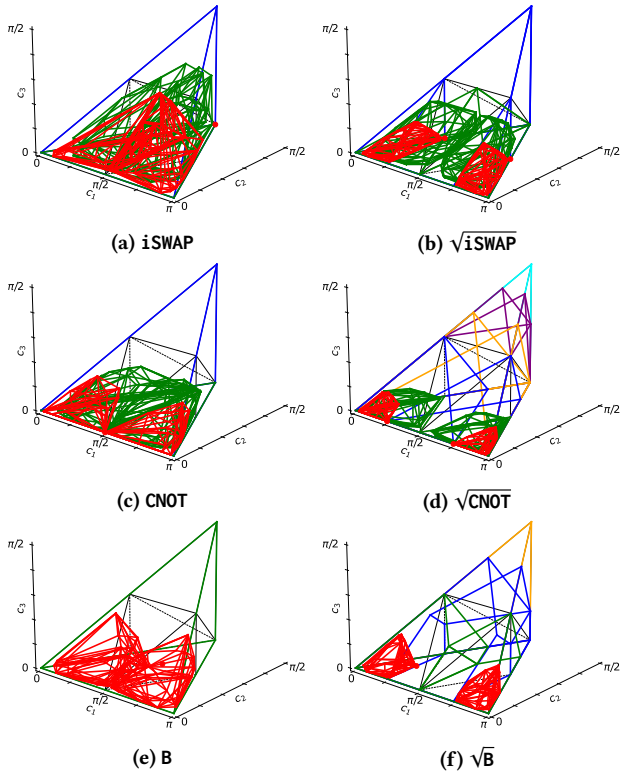


(a) Decomposition template **(b) Iterative gate coordinate** **(c) Convergence plot** **(d) Optimized cartan trajectory**

Figure 9: (a) The decomposition template given to the Nelder-Mead optimizer. To bound the coverage regions we attempt to converge to exterior Weyl Chamber points. (b)–(c) $K = 1$ iSWAP is verified to contain CNOT by optimizing over $\epsilon_1(t)$ and $\epsilon_2(t)$. (d) The resulting parallel-drive unitary evolution.

Algorithm 2 Method for calculating approximate improved volumes from parallel-drive

```
Basis Template ← g_c, g_g, T
k ← 0
while Coverage Volume not 100% do
    k ← k + 1
    Coordinate List ← []
    Randomly Generate Coverage Points
    for N iterations do
        Template ← Random(φ_c, φ_g, ε_1(t), ε_2(t))
        U ← Evaluate(Template)
        (x,y,z) ← Convert U to Weyl coordinate
        Coordinate List ← (x,y,z)
    end for
    Train for Exterior Coordinates
    for target in (I, CNOT, SWAP, iSWAP) do
        Save every coordinate along training path
        Coordinate List ← Template.optimize(target)
    end for
    Coordinate List partitioned into left and right
    Convex Hulls ← Coordinate List
    Basis.Polytope[k]← Convex Hulls
end while
return Haar Volume(Basis.Polytopes)
```



(a) iSWAP **(b) $\sqrt{\text{iSWAP}}$** **(c) CNOT** **(d) $\sqrt{\text{CNOT}}$** **(e) B** **(f) $\sqrt{\text{B}}$**

Figure 10: Parallel 1Q drive extended gate coverage sets. N=3000 random samples. red: $k = 1$, green: $k = 2$, blue: $k = 3$, orange: $k = 4$, purple: $k = 5$, cyan: $k = 6$

We use this property to build joint coverage sets between iSWAP and $\sqrt{\text{iSWAP}}$ which create decomposition rules using either gate in the decomposition template. Interestingly, predominately $\sqrt{\text{iSWAP}}$ sees a significant advantage from this procedure, as $K = 1$ iSWAP partially covers the perfect entangling region, which is heavily favored by the Haar distribution. For instance, the $K = 1$ iSWAP volume contains the point $(\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{4})$, which renders the $K = 2$ $\sqrt{\text{iSWAP}}$ coverage of the same point unnecessary, and eliminates the duration from the 1Q gate in the decomposition. Both $\sqrt{\text{CNOT}}$ and $\sqrt{\text{B}}$ obey the same rule, but with smaller overlapping volumes. This makes the advantage present, but less significant. After applying parallel-drive to improve the computing power of each basis gate, we continue to find that $\sqrt{\text{iSWAP}}$ is the best candidate for a basis gate. Next, we will build explicit decomposition rules into this basis, implement them into a transpilation scheme, and report improved simulated fidelities on quantum algorithm benchmarks.

4 PARALLEL-DRIVE FOR ISWAP-FAMILY

Our work has shown the advantage in calibrating a basis gate with the smallest fraction of total pulse time that does not compromise fidelity. This can improve Haar score by reducing unnecessary computational work done by longer duration gates. Additionally, our approach does not incur significant calibration overhead, as

Table 4: Extended basis decomposition gate count cost (K). Value determined by extended spanning regions (Figure 10)

	iSWAP	$\sqrt{\text{iSWAP}}$	CNOT	$\sqrt{\text{CNOT}}$	B	$\sqrt{B}$
K[CNOT]	1	2	1	2	1	2
K[SWAP]	2	3	3	6	2	4
$\mathbb{E}$[K[Haar]]	1.35	2.17	2.33	3.52	1.75	2.50
K[W(.47)]	1.53	2.53	2.06	3.65	1.53	3.06

Table 5: Extended basis decomposition duration cost (D) using parallel-drive, (D[1Q]=.25, Linear SLF). Fractional basis scores are calculated using the joint spanning regions between themselves and the full basis gate, selecting the lowest cost template.

	iSWAP	$\sqrt{\text{iSWAP}}$	CNOT	$\sqrt{\text{CNOT}}$	B	$\sqrt{B}$
D[CNOT]	1.5	1.5	1.5	1.5	1.5	1.5
D[SWAP]	2.75	2.25	4	4	2.75	2.75
$\mathbb{E}$[D[Haar]]	1.94	1.71	3.16	2.88	2.44	2.06
D[W(.47)]	2.16	1.90	2.83	2.83	2.16	2.16

only a minimal basis set must be calibrated. We take advantage of the fact that basis gates naturally combine to create larger fractions of themselves, *i.e.*, a $\sqrt{\text{iSWAP}}$ and iSWAP can be constructed by two and four $\sqrt[4]{\text{iSWAP}}$s, respectively, to reason about joint coverage sets. Short basis gates are useful for building gates near Identity $\mathbb{I}$, such as the small controlled-phase rotations that appear in QFT; capable of combination to take long strides, *e.g.*, to SWAP; and can take advantage of parallel-drive to boost computational power.

4.1 Parallel-Drive for Improving CNOT and SWAP

The methodology of creating coverage sets for parallel-driven gates allows us to easily improve decomposition templates via inspection. The CNOT-family and SWAP decomposition rules are given in Figure 11 and Figure 12[2]. The iSWAP as a function of $\epsilon_1(t), \epsilon_2(t)$ is constructed using the approach from Figure 4 using iSWAP in place of the generic 2Q gate. Recall, both the CNOT and SWAP decompositions are demonstrated graphically in Figure 2, where the parallel-drive is responsible for the curve in the trajectory. In a full transpilation scheme, the optimizer would be required to fit the exterior 1Q gate parameters, but for the purpose of simulating duration-dependent fidelity, the actual solution is unnecessary.

The Weyl Chamber does not represent distances and pulse costs uniformly and may mistakenly convey that a shorter, direct trajectory from $\mathbb{I}$ to U_T reduces the required 2Q basis duration, *e.g.* building CNOT with parallel-driven $\sqrt{\text{iSWAP}}$. However, there is a persistent requirement that 1 total iSWAP pulse durations appear in the decomposition to be able to reach CNOT, likewise 1.5 total iSWAP pulse duration is required to reach SWAP. These inherent costs are more rigorously detailed using quantum resource theories [53], and explain that 2Q decomposition can only be further optimized by removing the 1Q gate delays, but never a shorter 2Q time *i.e.*, the fundamental invariant related to "computing power."

[2]For all other gates, the gate coverage set is used as a lookup table for the required template size.

This inherent relationship between iSWAP and CNOT is depicted in Figure 13, such that a fractional duration iSWAP always contains the same fractional duration CNOT. For instance, K=2 with $\sqrt{\text{iSWAP}}$ or K=1 with parallel-drive iSWAP both reach CNOT. Of course, $\sqrt{\text{iSWAP}}$ is still the more powerful basis despite this relation to CNOT for containing additional Weyl Chamber volume. Both iSWAP and CNOT are special perfect entanglers, and interestingly a non-entangling SWAP gate can be used to convert back and forth between gates [54].

4.2 Simulated Fidelity Improvements

We utilize a circuit fidelity model that captures the primary source of error as decoherence in time following the methodology of prior work [42, 43]. The fidelity of a final qubit state, $\mathcal{F}_Q$, over a single path or wire in the circuit exponentially decays as a function of the ratio of circuit duration time and the qubit's T_1 (relaxation rate). Then the total circuit fidelity, $\mathcal{F}_T$ is given by the composite final qubit state, and thus is exponential with number of qubits. For this reason, even small improvements in circuit duration cascade into improved path and total circuit fidelities.

$$\mathcal{F}_Q = e^{-D[\text{Circuit}]/T_1} \tag{10}$$

$$\mathcal{F}_T = \prod_{i=1}^{N} \mathcal{F}_{Q_i} \tag{11}$$

Our transpilation scheme uses the SLF normalized durations $D[U_B]$, which are converted back into units of time by multiplying by the iSWAP duration. To quantify these improvements we choose $D[\text{iSWAP}] = 100$ ns and $D[1Q] = 25$ ns with qubit lifetime $T_1 = 100$ µs, which is consistent with transmon qubits using a SNAIL modulator [14]. Using these values, the improvements from the reduced duration decompositions over CNOT, SWAP, and Haar random targets are given in Table 6. The baseline uses previously derived analytical $\sqrt{\text{iSWAP}}$ decomposition rules [21]. Note that although exterior gates are used in the CNOT decomposition to make it perfectly equivalent, quantum algorithms often have 1Q gates before and after CNOT gates. Therefore, the circuit's 1Q gates and the decomposition substitution's 1Q gates would naturally combine for an even lower cost than represented here.

In our transpilation flow, we start by consolidating runs of all unitary blocks into 2Q gates and inducing SWAPs on a 4×4 square-lattice topology[3]. We then decompose each gate into the $\sqrt{\text{iSWAP}}$ basis with the pulse duration calculated by the provided SLF. Decomposition uses predefined substitutions for gates locally equivalent to CNOT-family and SWAP gates (see Figure 2b). If a rule is not known, we load the iSWAP and $\sqrt{\text{iSWAP}}$ extended coverage sets to construct a minimum size K template. Finally, we consolidate consecutive 1Q gates and report the remaining durations on each path. Our decomposition improvements using parallel-drive led to an average relative reduction in duration of 17.8% for the set of quantum algorithm workloads, as determined by selecting the best outcome from 10 transpiler runs, reported in Table 7. The Quantum Volume results were further averaged through additional runs due to the random nature of the algorithm.

[3]A CNOT followed by a SWAP on the same qubit pair is equivalent to an iSWAP which appears with non-negligible frequency, see Figure 5b at iSWAP.

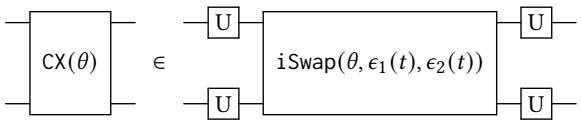


Figure 11: Decomposition template for CNOT into $\sqrt{\texttt{iSWAP}}$. A solution is $\epsilon_1 = 3, \epsilon_2 = 0$ for all time steps.

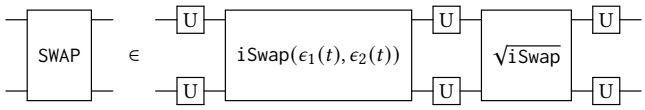


Figure 12: Decomposition template for SWAP into $\sqrt{\texttt{iSWAP}}$. A suitable solution for the parallel-drives is $\epsilon_1 = \pi, \epsilon_2 = \pi$ for all time steps. The interior set of 1Q gates is expected to be unnecessary if derived more precisely.

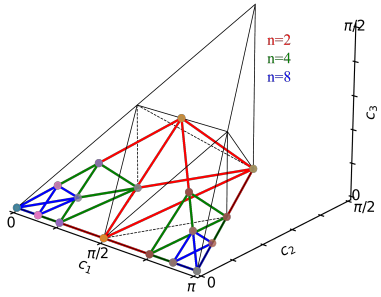


Figure 13: Illustrating the $K = 2$ coverage of $\sqrt[n]{\texttt{iSWAP}}$ for $n \in \{2, 4, 8\}$ which can realize $\sqrt[m]{\texttt{CNOT}}$ for $m \in$ {1,2,4}.

Table 6: Gate infidelities, $1 - \mathcal{F}_Q$ (D[1Q]=.25, Linear SLF)

U_T	Baseline	Optimized	% Improved
CNOT	0.0035	0.0030	14.3
SWAP	0.0050	0.0045	9.98
$\mathbb{E}$[Haar]	0.0038	0.0034	10.5
W(.47)	0.0043	0.0038	11.62

Table 7: Transpilation results (D[1Q]=.25, Linear SLF). Baseline and Optimized columns report total circuit duration in D[2Q]=1 normalized units. Duration, $\mathcal{F}_Q$, and $\mathcal{F}_T$ columns are reported as the relative % improvement between the baseline and optimized durations.

Benchmark	Baseline	Optimized	Duration	$\mathcal{F}_Q$	$\mathcal{F}_T$
QV	133.0	118.4	11.22	1.50	27.0
VQE_L	25.75	21.5	16.50	0.43	7.04
GHZ	31.75	27.00	14.96	0.48	7.90
HLF	102.3	88.00	13.94	1.43	25.6
QFT	149.5	120.3	19.53	2.96	59.5
Adder	175.0	144.3	17.57	3.12	63.6
QAOA	197.8	147.8	25.25	5.12	122
VQE_F	333.3	286.8	13.95	4.76	110
Multiplier	1065.25	770.76	27.64	34.2	11000

The average relative reduction in duration is directly related to our improvement method, while the relative path and total fidelity improvements depend on the baseline duration, mock gate durations, and qubit lifetime. Shallower circuits inherently have higher fidelities, and thus, their improvement potential is limited compared to deeper circuits with lower fidelities. For instance, the Quantum Volume improves from 0.875 to 0.889 in terms of path fidelities (a 1.5% improvement), which, due to its exponential relationship in the number of qubits, results in an increase from 0.119 to 0.151 (a 21% improvement) in total fidelity. In contrast, the shortest VQE_L algorithm path fidelities baseline of 0.975 only improve to 0.979 (0.4% improvement), leading to a total fidelity increase from 0.662 to 0.709 (6.6% improvement). Finally, our W(.47) metric predicts an average 11.6% reduction in duration. Our experimentally demonstrated 17.8% actually outperforms this case due to additional improvements to CNOT where the decomposition template's exterior 1Q gates could be merged or eliminated.

5 CONCLUSION

In this paper, we formally characterized the optimal basis gate for a parametric coupler under hardware speed limitations. The results indicate that, despite the $\sqrt{\texttt{iSWAP}}$ being close to optimal prior to our analysis, it can still be improved by utilizing parallel 1Q gates. This small improvement leads to a notable enhancement in fidelity as the number of qubits increases. Our co-design evaluated uniform Haar gates and circuit-based gate sets, finding that for realistic cost functions, such as our experimentally-determined SNAIL-coupler data, the $\sqrt{\texttt{iSWAP}}$ gate performed the best in nearly all scenarios.

Initially, gate count scores favored the B gate, but after considering the cost of direct generation through multiple simultaneous parametric drives, the $\sqrt{\texttt{iSWAP}}$ gate was the most efficient. The introduction of parallel-drive and related transpilation optimizations reduced the gate duration for most basis gates and improved the pulse time for the $\sqrt{\texttt{iSWAP}}$ gate. The iSWAP family was uniquely enhanced through joint parallel-drive extended coverage sets, yielding significant improvements in fidelity due to faster circuit execution.

In future work, we aim to expand our parallel-drive transpilation flow to further enhance compilation strategies for quantum algorithms and test them on various quantum systems with differing speed limit characterizations and dynamics. Moreover, detailed studies of improvement of parallel-drive volume versus calibration complexity for different quantum machine targets, including studying calibration complexity, while expanding the flexibility to handle continuously variable drive parameters, similar to optimal-control theory methods, are important next steps.

ACKNOWLEDGMENTS

This work is supported by the University of Pittsburgh via a SEEDER grant, by the Charles E. Kaufman Foundation via a new initiative grant, by the DOE via the C2QA collaboration, and NSF Award CNS-1822085. CZ, MX, and MJH are partially supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers Co-Design Center for Quantum Advantage under contract DE-SC0012704.

REFERENCES

[1] Jerry Chow, Oliver Dial, and Jay Gambetta. "IBM Quantum breaks the 100-qubit processor barrier". In: *[Available Online] https://research.ibm.com/blog/127-qubit-quantum-processor-eagle* (2021).

[2] Frank Arute et al. "Quantum supremacy using a programmable superconducting processor". In: *Nature* 574.7779 (2019), pp. 505–510.

[3] Shaowei Li et al. "Realisation of high-fidelity nonadiabatic CZ gates with superconducting qubits". In: *npj Quantum Information* 5.1 (2019), pp. 1–7.

[4] Peng Zhao et al. "High-Contrast Z Z Interaction Using Superconducting Qubits with Opposite-Sign Anharmonicity". In: *Physical Review Letters* 125.20 (2020), p. 200503.

[5] Ilya N Moskalenko et al. "High fidelity two-qubit gates on fluxoniums using a tunable coupler". In: *npj Quantum Information* 8.1 (2022), pp. 1–10.

[6] Jerry M Chow et al. "Simple all-microwave entangling gate for fixed-frequency superconducting qubits". In: *Physical review letters* 107.8 (2011), p. 080502.

[7] Abhinav Kandala et al. "Error mitigation extends the computational reach of a noisy quantum processor". In: *Nature* 567.7749 (2019), pp. 491–495.

[8] Caroline Tornow et al. "Minimum quantum run-time characterization and calibration via restless measurements with dynamic repetition rates". In: *Physical Review Applied* 17.6 (2022), p. 064061.

[9] David Rodriguez Perez et al. "Error-divisible two-qubit gates". In: *arXiv preprint arXiv:2110.11537* (2021).

[10] Eric C Peterson, Gavin E Crooks, and Robert S Smith. "Fixed-depth two-qubit circuits and the monodromy polytope". In: *Quantum* 4 (2020), p. 247.

[11] Michael A Nielsen and Isaac Chuang. *Quantum computation and quantum information*. 10th ed. Cambridge University Press, 2011.

[12] Anirudh Narla et al. "Robust concurrent remote entanglement between two superconducting qubits". In: *Physical Review X* 6.3 (2016), p. 031036.

[13] N. Leung et al. "Deterministic bidirectional communication and remote entanglement generation between superconducting qubits". In: *npj Quantum Inf* 5.1 (Feb. 2019), pp. 1–5. ISSN: 2056-6387. DOI: 10.1038/s41534-019-0128-0. URL: https://www.nature.com/articles/s41534-019-0128-0 (visited on 07/28/2021).

[14] Chao Zhou et al. "A modular quantum computer based on a quantum state router". In: *arXiv:2109.06848* (2021).

[15] Chao Zhou et al. "Understanding the speed limits of parametrically pumped quantum gates". In: *Bulletin of the American Physical Society* (2022).

[16] Eric C Peterson, Lev S Bishop, and Ali Javadi-Abhari. "Optimal synthesis into fixed xx interactions". In: *Quantum* 6 (2022), p. 696.

[17] S. Lin et al. "Let Each Quantum Bit Choose Its Basis Gates". In: *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2022, pp. 1042–1058. DOI: 10.1109/MICRO56248.2022.00075. URL: https://doi.ieeecomputersociety.org/10.1109/MICRO56248.2022.00075.

[18] Navin Khaneja and Steffen Glaser. "Cartan Decomposition of SU (2^ n), Constructive Controllability of Spin systems and Universal Quantum Computing". In: *arXiv preprint quant-ph/0010100* (2000).

[19] Robert R Tucci. "An introduction to Cartan's KAK decomposition for QC programmers". In: *arXiv preprint quant-ph/0507171* (2005).

[20] Timjan Kalajdzievski and Nicolás Quesada. "Exact and approximate continuous-variable gate decompositions". In: *Quantum* 5 (2021), p. 394.

[21] Cupjin Huang et al. "Quantum instruction set design for performance". In: *arXiv preprint arXiv:2105.06074* (2021).

[22] Péter Rakyta and Zoltán Zimborás. "Approaching the theoretical limit in quantum gate decomposition". In: *Quantum* 6 (2022), p. 710.

[23] Ethan Smith et al. "Leap: Scaling numerical optimization based synthesis using an incremental approach". In: *ACM Transactions on Quantum Computing* (2021).

[24] Liam Madden and Andrea Simonetto. "Best approximate quantum compiling problems". In: *ACM Transactions on Quantum Computing* 3.2 (2022), pp. 1–29.

[25] Yuriy Makhlin. "Nonlocal properties of two-qubit gates and mixed states, and the optimization of quantum computations". In: *Quantum Information Processing* 1.4 (2002), pp. 243–252.

[26] S Balakrishnan and R Sankaranarayanan. "Characterizing the geometrical edges of nonlocal two-qubit gates". In: *Physical Review A* 79.5 (2009), p. 052339.

[27] Jun Zhang et al. "Geometric theory of nonlocal two-qubit operations". In: *Physical Review A* 67.4 (2003), p. 042313.

[28] Youngkyu Sung et al. "Realization of High-Fidelity CZ and ZZ-Free iSWAP Gates with a Tunable Coupler". In: *Phys. Rev. X* 11 (2 June 2021), p. 021058. DOI: 10.1103/PhysRevX.11.021058. URL: https://link.aps.org/doi/10.1103/PhysRevX.11.021058.

[29] Tanay Roy et al. "Realization of two-qubit quantum algorithms on a programmable superconducting processor". In: *arXiv preprint arXiv:2211.06523* (2022).

[30] D. K. Weiss et al. "Fast high-fidelity gates for galvanically-coupled fluxonium qubits using strong flux modulation". In: *arXiv preprint arXiv:2207.03971* (2022).

[31] Ananda Roy and Michel Devoret. "Introduction to parametric amplification of quantum signals with Josephson circuits". In: *Comptes Rendus Physique* 17.7 (2016). Quantum microwaves / Micro-ondes quantiques, pp. 740–755. ISSN: 1631-0705. DOI: https://doi.org/10.1016/j.crhy.2016.07.012. URL: https://www.sciencedirect.com/science/article/pii/S1631070516300640.

[32] Paul Kairys and Travis S Humble. "Efficient Quantum Gate Discovery with Optimal Control". In: *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE. 2021, pp. 413–418.

[33] Michael H. Goerz et al. "Krotov: A Python implementation of Krotov's method for quantum optimal control". In: *SciPost Phys.* 7 (2019), p. 80. DOI: 10.21468/SciPostPhys.7.6.080.

[34] Xin Wang, Edwin Barnes, and S Das Sarma. "Improving the gate fidelity of capacitively coupled spin qubits". In: *npj Quantum Information* 1.1 (2015), pp. 1–7.

[35] Nathan Earnest, Caroline Tornow, and Daniel J Egger. "Pulse-efficient circuit transpilation for quantum applications on cross-resonance-based hardware". In: *Physical Review Research* 3.4 (2021), p. 043088.

[36] NE Frattini et al. "3-wave mixing Josephson dipole element". In: *Applied Physics Letters* 110.22 (2017), p. 222603.

[37] Gangqiang Liu et al. "Josephson parametric converter saturation and higher order effects". In: *Applied Physics Letters* 111.20 (2017), p. 202603.

[38] Paul Watts, Maurice O'Connor, and Jiri Vala. "Metric structure of the space of two-qubit gates, perfect entanglers and quantum control". In: *Entropy* 15.6 (2013), pp. 1963–1984.

[39] John M Martyn et al. "Grand unification of quantum algorithms". In: *PRX Quantum* 2.4 (2021), p. 040203.

[40] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. "A quantum approximate optimization algorithm". In: *arXiv preprint arXiv:1411.4028* (2014).

[41] Peter W Shor. "Why haven't more quantum algorithms been found?" In: *Journal of the ACM (JACM)* 50.1 (2003), pp. 87–90.

[42] Pranav Gokhale et al. "Faster and more reliable quantum swaps via native gates". In: *arXiv preprint arXiv:2109.13199* (2021).

[43] Evan McKinney et al. "Co-Designed Architectures for Modular Superconducting Quantum Computers". In: *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. 2023, pp. 759–772. DOI: 10.1109/HPCA56546.2023.10071036.

[44] LS Theis, F Motzoi, and FK Wilhelm. "Simultaneous gates in frequency-crowded multilevel systems using fast, robust, analytic control shapes". In: *Physical Review A* 93.1 (2016), p. 012324.

[45] Luca Planat et al. "Understanding the saturation power of Josephson parametric amplifiers made from SQUID arrays". In: *Physical Review Applied* 11.3 (2019), p. 034014.

[46] Sahel Ashhab et al. "Speed limits for two-qubit gates with weakly anharmonic qubits". In: *Physical Review A* 105.4 (2022), p. 042614.

[47] Maika Takita et al. "Demonstration of weight-four parity measurements in the surface code architecture". In: *Physical review letters* 117.21 (2016), p. 210505.

[48] David C McKay et al. "Efficient Z gates for quantum computing". In: *Physical Review A* 96.2 (2017), p. 022330.

[49] David Avis. "Living with lrs". In: *Japanese Conference on Discrete and Computational Geometry*. Springer. 1998, pp. 47–56.

[50] Lingling Lao et al. "Designing Calibration and Expressivity-Efficient Instruction Sets for Quantum Computing". In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)* (June 2021). DOI: 10.1109/isca52012.2021.00071. URL: http://dx.doi.org/10.1109/ISCA52012.2021.00071.

[51] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[52] Paul Watts et al. "Optimizing for an arbitrary perfect entangler. I. Functionals". In: *Physical Review A* 91.6 (2015), p. 062306.

[53] Eric Chitambar and Gilad Gour. "Quantum resource theories". In: *Reviews of modern physics* 91.2 (2019), p. 025001.

[54] Gavin E Crooks. "Gates, states, and circuits". In: (2020).

Received 22 November 2022; revised 21 February 2023; revised 3 April 2023; accepted 10 April 2023