

A High Performance Compiler for Very Large Scale Surface Code Computations

George Watkins^{1,2}, Hoang Minh Nguyen², Keelan Watkins³, Steven Pearce², Hoi-Kwan Lau^{3,4}, and Alexandru Paler¹

¹Department of Computer Science, Aalto University, 00076 Espoo, Finland

²School of Computing Science, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6

³Department of Physics, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6

⁴Quantum Algorithms Institute, Surrey, B.C., Canada V3T 5X3

We present the first high performance compiler for very large scale quantum error correction: it translates an arbitrary quantum circuit to surface code operations based on lattice surgery. Our compiler offers an end to end error correction workflow implemented by a pluggable architecture centered around an intermediate representation of lattice surgery instructions. Moreover, the compiler supports customizable circuit layouts, can be used for quantum benchmarking and includes a quantum resource estimator. The compiler can process millions of gates using a streaming pipeline at a speed geared towards real-time operation of a physical device. We compiled within seconds 80 million logical surface code instructions, corresponding to a high precision Clifford+T implementation of the 128-qubit Quantum Fourier Transform (QFT). Our code is open-sourced at <https://github.com/latticesurgery-com>.

1 Introduction

Applying surface quantum error correcting codes (QECCs) efficiently to large computations is challenging in terms of classical computing resources necessary for the compilation process. Compilers tailored for QECC are only starting

George Watkins: invio.george@gmail.com

Hoang Minh Nguyen: hoangminh98@gmail.com

Keelan Watkins: keelan.w@outlook.com

Steven Pearce: stevenp@sfu.ca

Hoi-Kwan Lau: hklau.physics@gmail.com

Alexandru Paler: alexandrupaler@gmail.com

to appear, often with significant limitations with respect to the scale of the circuits that can be handled or the compilation time.

Large scale QECC compilation is a necessity, because practical algorithms, like Shor's and Grover's assume high-quality qubits with a very low error rate [1], but we are unlikely to obtain hardware (physical) qubits with such fidelity in the near future [2]. QECCs solve this issue by using a large number of error prone physical qubits to encode higher fidelity *logical* qubits. For example, a quantum factoring algorithm needs roughly 1000 qubits to factor a 1000-bit number and millions of gates [3, 4]. Consequently, practical algorithms require very large scale quantum computers, while only some carefully crafted examples of problems where quantum hardware has an advantage with small devices exist [5].

Surface codes are a family of QECCs that require low qubit connectivity and a reasonably high hardware error rate (such as between 0.1% and 1%) to create good logical (computational) qubits [6, 7, 8] and only require degree four nearest neighbour connectivity. These properties make them a promising option for error correcting devices with a couple hundred logical qubits. Physical devices with compatible layouts have already been made or proposed, albeit on a small scale [5, 9, 10, 11, 12]. Examples of larger scale quantum circuits protected by surface QECCs were compiled manually in [13, 14]. The complexity of optimising surface code circuits has been shown to be related to NP-hardness [15, 16].

We present and demonstrate the extremely high scalability of our efficient QECC compiler. This is a step forward for quantum software: we create a streaming pipeline and a compilation

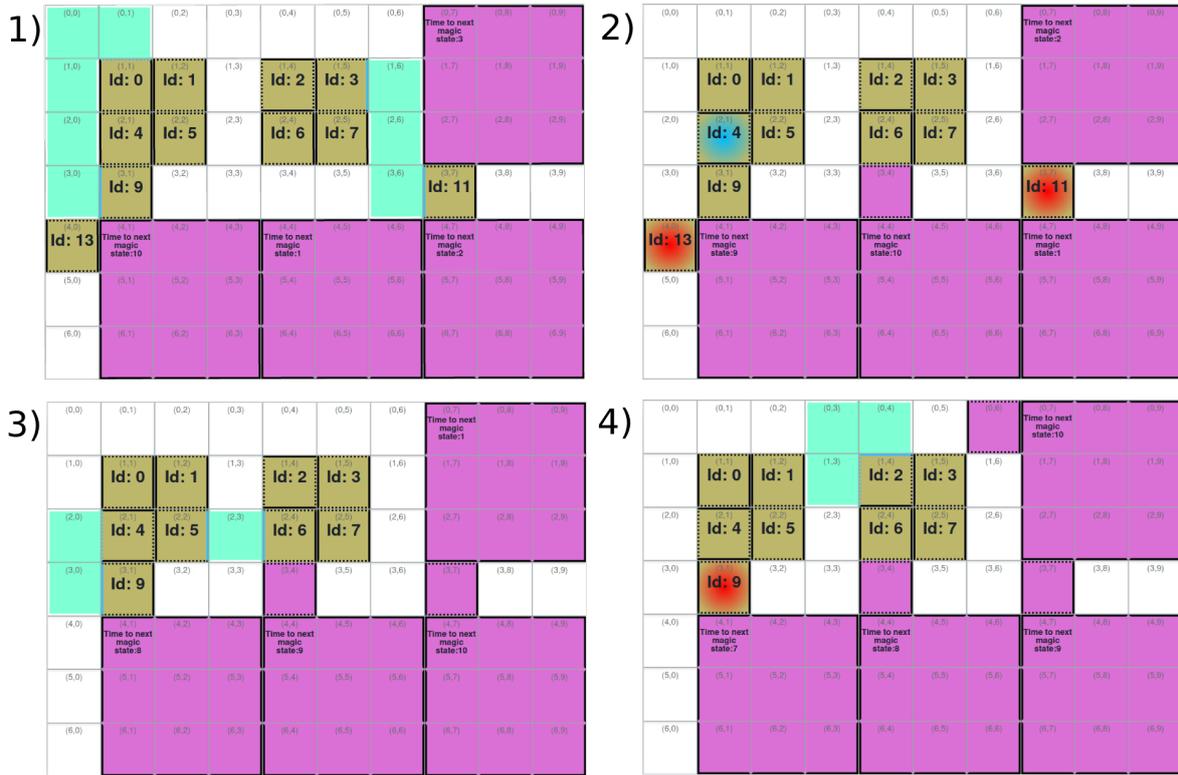


Figure 1: Example output of from the compiler: sequence of discrete time steps, called *slices*, of a surface code computation. Each slice has an associated time instant when it is taking place. The slices are obtained after mapping a circuit to a layout, where patches end up being used for holding error-corrected logical qubits (brown), distillation procedures (pink), routing merge and split operations of patches (blue), or not used (white). 1) There is a merge and split operation (blue) taking place between logical qubits 0 and 3. 2) Qubits 11 and 13 are measured; 3) logical operation between qubits 4 and 9, and logical operation between 5 and 6, and the bottom right distillation is outputting a distilled state; 4) qubit 9 is measured, merge and split between 1 and 2, and the right most distillation region is outputting a distilled state.

environment for the compilation and optimisation of very large scale QECCs. Our high performance pipeline makes it possible to process extremely large circuits (would not fit in memory). We can compile directly, in a streaming process, by reading and writing to mass storage. Streaming enables the real-time operation of our compiler, meaning that this tool may be integrated in the classical control software necessary to operate quantum computers [17].

This paper is organised as follows: In Sec. 2 we introduce the concepts necessary for presenting the compilation methods and workflow. Sec. 3 describes the two-stage compilation pipeline that consists of gate level processing (Sec. 3.2) and logical operation routing (Sec. 3.3). The latter includes also a fast method to perform state vector simulation that takes into account the entangling and disentangling action of the lattice surgery operations. Finally, Sec. 4 illustrates the perfor-

mance of our compiler. We compile within seconds a high-precision 128-qubit Quantum Fourier Transform (QFT) [18]. To the best of our knowledge, this is the largest-scale compilation of this kind.

2 Background

This section introduces the necessary background details for describing the compilation process. The application of error correction to quantum circuits resembles the process well known to classical computer scientists of program compilation: the *compiler* reads code in a programming language (higher level quantum gates) and outputs machine instructions (lattice surgery quantum gates).

We opted for flexibility and developed a compiler with a well-defined intermediate representation to separate circuit pre-processing from sur-

face code instruction layout. Surface code instructions for large scale computations at present is interesting for at least two purposes, one is being able to produce reliable resource estimates, and the other is to start preparing for when we will have such devices, so that hardware engineers can start designing devices with instruction sets for error correction in mind.

We assume the reader is familiar with the basic concepts of quantum computing and quantum information [18, 19]. We assume the conventional meaning for common quantum gates (Phase gates S and T, Hadamard gate H, CNOT, Toffoli) and the Pauli matrices (I, X, Y, Z). By the phase rotation gate $R_Z(\theta)$ we mean:

$$R_Z(\theta) = \begin{bmatrix} 1 & 0 \\ 0 & e^{i\frac{\theta}{2}} \end{bmatrix}$$

We will frequently use Pauli product rotations, for which we assume the following: given an axis P (which may be a Pauli matrix or a tensor product of Pauli matrices) we denote by $P(\theta) = \exp(-i\theta P) = \cos(\theta)I - i\sin(\theta)P$. Note that under this convention, $R_P(\theta) = P(\frac{\theta}{2})$ for $P = X, Z$. Also, when the Pauli matrices appear with sub-indices, e.g. $Z_1Z_2Z_3Z_4$ in Fig. 2, we mean the tensor product $X \otimes Z$ of the Pauli matrices applied to qubits indexed 1 and 2. Similarly, we use gate $R_X(\theta) = HR_Z(\theta)H$.

2.1 Surface Codes

A major challenge with the current generation of quantum computers is the occurrence of errors while performing computations. Errors may occur because of control system faults or stray interaction with the environment. A proposed solution for avoiding errors are Quantum Error Correcting Codes (QECC). These codes add some degree of fault tolerance to computations by using many *physical qubits* to form fewer but more reliable abstract *logical qubits*[22]. Surface codes are a family of QECCs that aim at improving computational fidelity by entangling physical qubits in a *physical lattice* [23, 18]. This kind of codes, with topological properties, was first theorized with exotic particles known as “anyons” [24]. Surface codes are appealing because they are well understood, and feature a high error threshold. In near future, quantum computing hardware with thousands of qubits

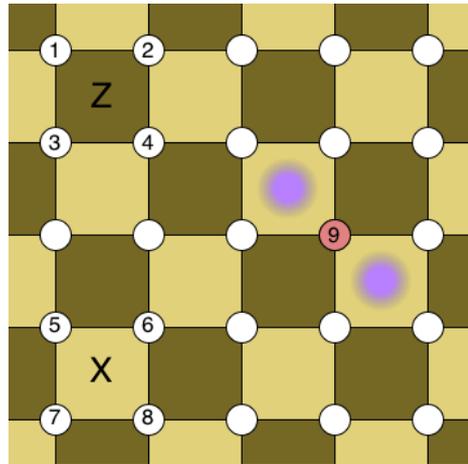


Figure 2: A graphical depiction of surface code layout. The white circles are *data* qubits protected from errors by measuring stabilizers around them. The squares in the lighter and darker shades of yellow represent stabilizer measurements. For example, the squares marked with Z and X represent the $Z_1Z_2Z_3Z_4$ and $X_5X_6X_7X_8$ stabilizer measurements respectively. If an error occurs in a data qubit, such as a phase flip occurring on 9, the X stabilizers around it will pick it up by changing outcome (*syndromes*, highlighted in purple). There are advanced methods to decode sets of errors (e.g [6, 20, 21]). Errors can either be corrected on the spot or tracked classically by inverting later readouts. This cycle of detecting, decoding and correcting is referred to as the *surface code cycle*.

might be realized [25, 26, 12] which would be able to operate a surface code cycle on a lattice of qubits.

The key step of surface code error detection is stabilizer measurement, as shown by the shaded squares in Fig. 2. These measurements act as parity checks on bit flips or phase flips of a square lattice of *data qubits*. The surface code and its cycle (the sequence of quantum gates applied for enforcing the code constraints) only tell us how to protect a lattice from error. The surface code distance indicates how much error is tolerated [27].

2.2 Logical Qubits and Logical Operations

Logical (computational) qubits are encoded by “cutting out” portions of a device’s physical lattice into *patches*, which are cluster states error corrected by the surface code cycle. This encoding of logical qubits is known as the *planar code* [27, 28]. Patches have boundaries outside of which they don’t interact, except when performing certain logical operations (Sec. 2.2). Fig. 3 outlines how patches relate to the surface code.

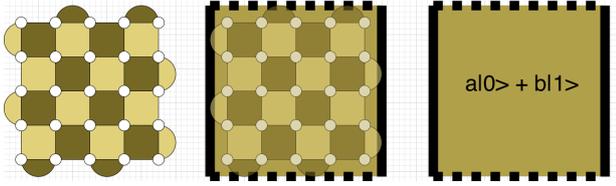


Figure 3: Abstracting physical qubits to patches. We omit the details of stabilizers and data qubits that make up patches, and instead represent distance-independent features. It is always possible to compute back these details about stabilizers from the output format and code distance. The picture to the left shows how this abstract representation relates to the physical implementation, and to the right there is a fully abstract patch, which has its own logical state. The different stabilizers on the boundaries yield two different kinds of boundaries, which are often referred to as *rough* and *smooth*.

The patch-based approach has been shown to be a resource-efficient choice for quantum error correction [29, 30, 31].

We will be looking at square patches with two kinds of boundaries that encode a single qubit. Patch size is proportional to code distance and the performance of the decoding algorithm (e.g. [6, 21]). For our intents, it suffices to know that the size of the patches will depend on the physical error rate of the device, length of the computation and desired success rate of the logical computation. In Sec. 4 we estimate the resources [32] necessary to execute the compiled output.

Having obtained logical qubits, we require a method to perform operations between them. Table 1 offers an overview of all the surface code operations supported by our compiler at the logical level. Some logical operations are performed directly on patches: Pauli X and Z [33], and Hadamard gates [29], can be implemented in this way and are called *transversal* operations. It is also possible to directly initialize a patch in the $|0\rangle$ or $|+\rangle$ states and to measure in the X or Z basis [31]. For the remaining operations needed to complete a universal gate set we use lattice surgery [29]. This protocol achieves entangling multibody measurements by merging and splitting patches.

We use these measurements along with prepared ancillae states (and corresponding patches) to implement CNOT as shown in [29], and the S and the T gates (Fig. 12 in Appendix). T gates utilize a *magic state*, $|m\rangle = \frac{1}{\sqrt{2}}(|0\rangle + e^{\frac{i\pi}{4}}|1\rangle)$,

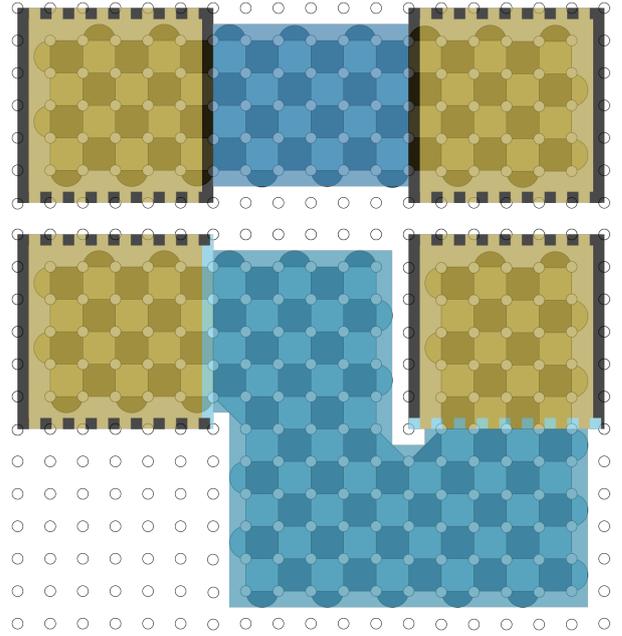


Figure 4: Lattice surgery of patches. Patches are *merged* by activating the stabilizer measurements with the data qubits between them (blue regions). This operation causes the two patches to become one, hence losing a degree of freedom and projecting the logical state into a subspace. After stopping the stabilizer measurements and measuring the mediating data qubits, the patches are *split*. Overall, this operation is equivalent to a logical multi body measurement [29]. The observable depends on the boundaries: rough for X and smooth for Z. This figure shows measurements of the observables $Z \otimes Z$ (top) and $Z \otimes X$ (bottom)

which in the surface code cannot be initialized directly with a high fidelity. These states have to be prepared by *distillation*. There are several protocols for magic state distillation [34, 35], but for our compilation purposes it suffices to acknowledge the fact that these distillations occupy some amount of space on the device’s lattice and that they have a certain duration in time: distillation regions are described by their *bounding box* which includes a time axis for how long it will take to produce the next magic state.

2.3 Related Work

Compilers for surface codes have been previously presented in the literature, and most of the times, the compilation problem has been decoupled from the challenges of optimising the resulting circuits. In general, automatic optimisation is performed by implementing heuristic algorithms for the efficient layout of the logi-

cal operations, and this includes parallelizing as many as possible operations, using fewer patches for the routing etc. Surface code computations can be implemented through braiding (e.g. Surf-braid [36]) or lattice surgery like the tool presented herein (e.g. OpenSurgery [37] or the compilers from [38, 39, 40]).

Our compiler is distinct from the others in the following ways. Our compiler’s source and target are similar to OpenSurgery, but improves on the compilation time performance, offers new optimizations, adds the ability to customize layouts, and handles parallel magic state distillation.

The compilers from [39, 40] focus specifically on routing long range surface code interactions. While we do tackle such problem, as it is necessary for our overall compilation goal, the focus of this project is broader in scope and we organise our compilation into a very modular, highly efficient pipeline which can handle both short and long range interactions.

Our compiler supports very large scale layouts through a layout specification and the compiler can automatically map large-scale circuits to the layouts. In contrast, the compiler from [38] is a small scale procedure that explores the trade offs of different layouts for mapping algorithms onto surface code architectures.

Our compiler is modular and can include manual optimisation techniques, for example, by replacing existing gate decompositions, or reconfiguring the bounding boxes of the distillation sub-circuits. For completeness, manually obtained surface code layouts with techniques such as the AutoCCZ for optimizing ripple carry adders where presented for example by [14]. Finally, one last approach to quantum compilers worth mentioning are variational compilers [41], which share with our project the challenges of circuit pre-processing.

Compared to existing surface code compilers, our tool extends the state of the art by including at least the following novelties:

- support for an intermediate language for compiling high-level circuits from different languages (e.g. Q#, Cirq, Qiskit) and descriptions (e.g. Clifford+T, multibody measurements);
- highly configurable layouts for qubits, routing space and multiple parallel distillation procedures;

Operation	Method
Patch initialization in the $ 0\rangle$ and $ +\rangle$ states	Direct Initialization of data qubits [29]
Single patch measurements	Direct measurement of data qubits [29]
Pauli X and Z	Transversal in surface codes [6]
Hadamard gates	Transversal in planar code patches [29]
Entangling multi body measurements	Lattice surgery merges and splits, mediated by ancilla patches for routing [31]
Boundary Rotation	Patch deformation[31]
S gates	Lattice surgery with twist defects [42]
Preparation of Magic states $ m\rangle = 0\rangle + e^{\frac{i\pi}{4}} 1\rangle$	Distillation in dedicated regions [35]

Table 1: The list of logical surface code operations supported by the compiler. The operations are formalized into *logical lattice instructions* (LLI), which serve as a central intermediate representation to our compiler. LLI decouples the pre-processing to surface code instructions from laying them out on an abstract lattice.

- very high-speed, configurable routing heuristics which can be easily replaced with more sophisticated approaches including based on machine learning;
- pipelined, modular design that is compatible with distributed computing platforms such that compilations and optimisations can be performed on multi-core/parallel computers.

3 Methods

We address the problem of taking a circuit specified in a machine readable format, and converting the circuit to the surface code operations outlined in Table 1. For small circuits it is easy enough to perform such conversion by hand, but automation is necessary for large scale circuits.

Our *compiler* is a computer program that reads text in a *source* formal language and outputs machine code in another language, called the *target*. In our case the source is a quantum circuit in a

subset of OpenQASM 2.0 [43] (Sec. 5.1), while the target is a JSON logical operation instructions (Sec. 2.2 and Fig. 1).

We implemented our compiler and the source code is open sourced at <https://github.com/latticesurgery-com>. In order to improve the readability of this paper, we keep the engineering and implementation details to a minimum, and point the interested reader to the open sourced code. The latter is written with modern C++ features which increase the comprehension of the code’s functionality.

The compiler is continuously tested and verified for functional correctness with modern continuous integration, while practical performance plays a significant role. Our compiler offers a wide range of configuration options, ranging from optimization heuristics, intermediate representations of the computations, as well as flexible layouts.

3.1 The Compilation Pipeline

The compiler operates a two-stage pipeline (Fig. 5): 1) a pre-processing stage, and 2) a layout and routing stage. The two stages communicate through an intermediate representation we refer to as *logical lattice instructions* (LLI from Table 1). The LLI contains all the information about the logical operations happening on the lattice, but none about the physical locations of the patches, or about routing and distillation regions. The physical qubit lattice will be operated according to LLI instructions (Table 1).

The first stage, the *gate level processing stage*, operates mostly at the logical circuit level. We resort to a universal gate set based on surface code operations. We gradually process the input circuit’s gates to align with our surface code instructions. Once the circuit is in a suitable format (only Clifford+T gates or certain Pauli rotations), the circuit maps 1-to-1 with surface code operations and is written down as LLI.

The second stage is the *slicer*. Herein, the LLI are combined with a layout specification in space and time (Fig. 7). The LLI language is circuit layout agnostic, meaning that the mapping of the logical qubits to the physical lattice may have a great impact on the efficiency of the compiled circuit. The result of these steps is a “sequence” of *slices* of the physical lattice. The slices depict the state of the computation at each point in time,

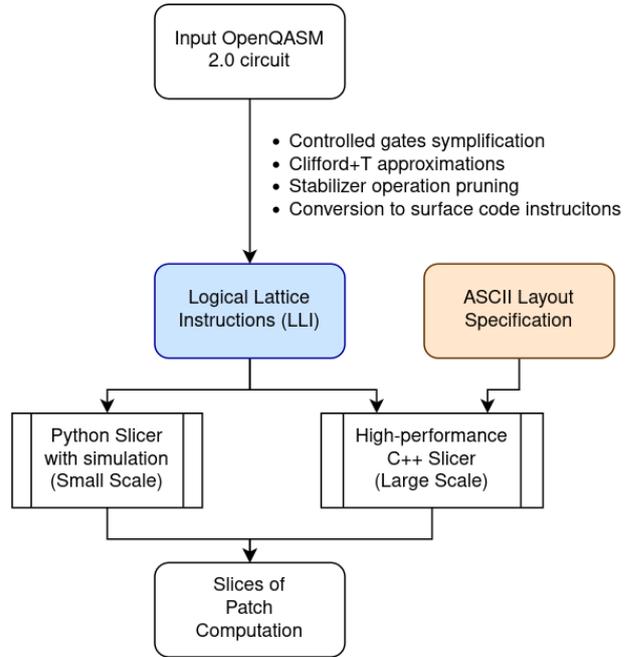


Figure 5: The pipeline as implemented in the compiler.

as shown in (Fig. 1). We offer two such slicers: one written in Python, geared towards the verification of small scale circuits (Sec. 3.3.1) and a high performance one written in C++ for large scale circuits (Sec. 3.3.2).

3.2 Gate Level Processing

The first stage takes a logical circuit specified in our own minimal dialect of OpenQASM 2.0. We offer two ways to pre-process the circuit: 1) with Pauli rotations and Pauli product measurements, and 2) directly with higher level quantum gates such as Toffoli gates. In both cases, we first parse the circuit into a list of gates, using either Qiskit [44], PyZX [45], a custom parser or a combination of the three depending on the circuit.

The gate list expression of the input circuit might use gates which are not supported by the error-correction procedure. In this step we reduce the gate set so that it easily translates to LLI. Our custom parser is able to break down very small angle rotations, such as $Z(\frac{\pi}{2^{i28}})$ by symbolic processing of the argument. These rotations are needed to compile, for example, a 128-qubit quantum Fourier transformation (QFT) circuit. After parsing, the list of gates is passed through the pipeline to the next stage.

First, controlled gates are broken down to CNOTs and single qubit rotations using the identity in Fig. 13. The circuit now only has single

qubit Clifford gates, CNOTs and single qubit rotations. At the last stage of pre-processing in the gate model single qubit rotations smaller than $\frac{\pi}{4}$ are approximated to single qubit Clifford+T gates.

It is possible to convert controlled-rotation gates to Clifford operations plus some small angle $Z(\theta)$ rotations (Figure 13 in Appendix). The latter are not Clifford+T and are difficult to perform in a fault-tolerant way [46]. We achieve arbitrary $Z(\theta)$ rotations by approximating them with Clifford+T gates, for which we leverage the Gridsynth package [47] which outputs approximations constituted of sequences of H, X, Z, S and T gates. The T gates are performed by consuming magic states, which are prepared in dedicated *distillation regions*[33, 35].

We utilize two methods to convert the Gridsynth approximation to LLI. The first is to directly apply the gates with the methods of Table 1: H, X and Z transversally, S with a twist and T as $Z(\frac{\pi}{8})$ rotation as shown in Fig. 12. The second approach, we refer to as *Pauli rotation compression*, is shown in Fig. 6, and consists of interpreting the gate sequences returned by the Gridsynth approximations as a sequence of Pauli rotations of varying angles.

The direct application of gates is simpler and results in the same Clifford corrective terms for every rotation. With Pauli rotation compression the Clifford corrective terms change for every angle, thus more complex classical control would be required by a downstream stage. In the Appendix we present an algorithm for Clifford gate optimization.

3.3 Slices and Routing

To overcome the logistical challenges of structuring a computation on surface code device, we arrange the computation in space and time. Space structure is given by partitioning the physical lattice into square *cells*. A cell may hold or may not hold a patch, be part of a distillation region, or may be used for routing, but patches, distillation regions and routing areas are always placed in accordance to cell boundaries (Figs. 1, 3 and 7).

Time structure is given by thinking of the computation in terms of *slices*. Surface code computations can be viewed as 3D structures in space-time [48, 37, 14], and a slice is a plane through the structure at a fixed time value (Fig. 1). In a

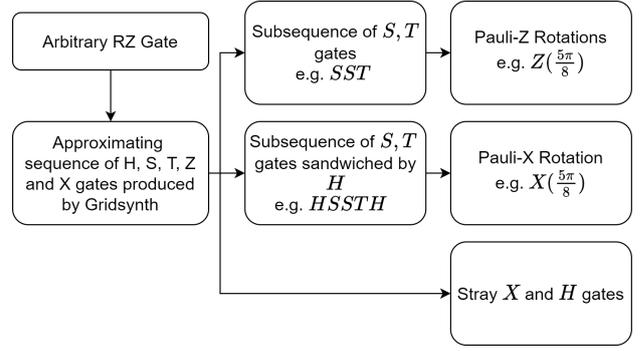


Figure 6: Pauli rotation compression of Clifford+T approximations of small angle rotations E.g. $R_Z(\frac{\pi}{2^{128}})$: gate sequences obtained from Gridsynth are interpreted in the Pauli Frame by breaking it into subsequences. For instance, the sequence $HSHTSHX$ would be split as HSH, TS, H, X and would become the sequence $X\frac{\pi}{4}Z\frac{3\pi}{8}HX$

nutshell, a slice is a temporally discretized partition of the computation (*clock timesteps* in Litinski [31] or *moments* in Google Cirq [49] terminology, for example). Each slice represents a snapshot of the the LLIs that are happening simultaneously on the lattice – slice duration is given by the duration of the slowest LLI.

Routing is the problem of deciding how the cells of a slice are allocated to patches or reserved for other purposes. Finding optimal layouts has a great impact on the depth of the computation. Different layouts can for example be used to trade off space for time [38]. Layouts may need to change depending on the task during an algorithm. For example, the oracle in Grover’s search algorithm may be very different from the implementation of the diffusion operator [50]. We defined our own *configurable layout specification* (Fig. 7). The compiler reads a text file containing the layout specification and produces slices with patches arranged accordingly.

3.3.1 On-the-fly, Functionally Verified Slicer

Our first slicer supports the real-time, on-the-fly functional verification for correctness. This slicer can be used as a preliminary verification of smaller scale lattice surgery circuits. The slicer and the simulation operate on an array of patches of variable length and assumes that all magic states have been prepared ahead of time. The verified slicer is very powerful when it comes to understanding the details of small computation and we used it in the development of the com-



Figure 7: ASCII specification for patch layouts. Q indicates a patch holding a logical qubit, r marks cells that are reserved for routing (the cyan “snakes” of Fig. 1). Numbers 0 to 9 are used to identify distillation regions. The boundaries of the distillation regions are computed by connected components search for same numbers, so it is possible to have more than 10 distillation regions. Magic states produced by these regions are queued in the r cells neighbouring a distillation block. Finally, A marks cells reserved to allocate new ancillae patches in the states, such as the $|+\rangle$ states that mediate CNOTs or the places for the Y eigenstates used by $\frac{\pi}{4}$ Pauli rotations. The layout file format is described in the Appendix.

piler.

The simulator, called the *Lazily Tensorized State-vector Simulation* (LTSvS), has the major feature of being able to simulate patch states at the LLI instruction level, such as simulating multi body measurements and Pauli operator gates. LTSvS tensors at the matrix level only when strictly required, otherwise just tracks the fact that the global state is given by a tensor product of sub vectors. LTSvS offers a great performance advantage over naïve state vector simulation: our simulator doesn’t expand the full state vector of all logical qubits on the lattice. In particular, qubits that are known to not be entangled, because they were just initialized or measured, are automatically tracked in separate sub-state vectors. Qubits may be entangled within a sub-state vector. An example of unentangled qubits is the array of magic states waiting to be used or ancillae patches.

Methodologically, the LTSvS simulator is very similar to the matrix-product-states (MPS) simulation techniques [51], which are efficient on circuits with low counts of entangling gates. Compared to the MPS simulators, e.g. from Qiskit [44], ours is fine tuned for computations

with many ancillae and measurements, can handle classical control and can be executed in parallel with the compilation process.

3.3.2 High Performance Slicer

The main goal of our compiler is to handle very large scale circuits with thousands of logical qubits and millions of LLIs. At this scale every CPU clock cycle and every byte are precious. Our high performance compiler is written in C++ because it comes with zero cost abstraction [52].

The first step of the slicer is to read a layout file (Fig. 7) in order to create an abstract layout representation that describes the device layout. The layout is used to initialize a slice template which will be reused for the routing. The template will be recomputed once the layout dictates this. Our implementation of slice processing keeps memory usage to a minimum because $O(1)$ slices are ever kept in memory by the slicer itself. Moreover, this representation is stored in a high performance data structure based on bit-streams and hash-maps. The representation will be used for computing routes using a variant of Dijkstra’s algorithm.

The slicer streams LLI from text or standard input, updating the slice with each instruction, evolving the slices over time. Since the slicer can also stream read from standard input and write to standard output, its possible to implement external programs (e.g. Python scripts or other command line tools) that visit slices by reading from standard input. Given the capability of evolving the lattice state, the slice processing functionality is implemented by defining a C++ functor to visit all slices.

The streamed evolution of slices includes managing distillation, queuing magic states [48, 53], initializing ancillae and LLI operations. A user may collect statistics on slices, such as magic state queue and routing space usage in seconds, without having to store in terabytes of memory that slow down the processing. At the same time this functor approach has the advantage of hiding the implementation details from the client so that they can focus on the processing functionality.

To place routing regions, we used our own implementation of Dijkstra’s algorithm, which is implemented *in place*, such that our tool can

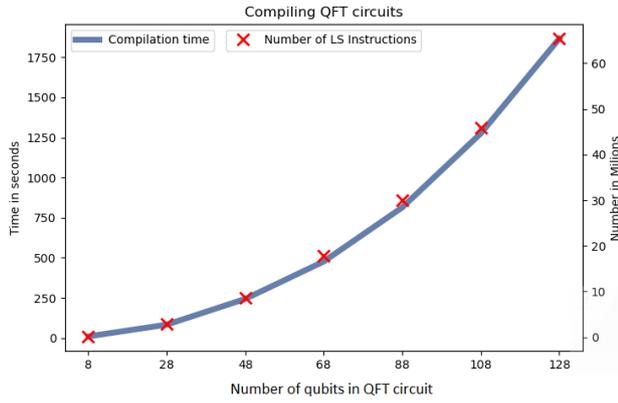


Figure 8: Time taken to compile a QFT with the C++ slicer on a laptop (Intel i5350U, 8GB RAM) and number of LLI instructions for different QFT sizes. The Clifford+T implementation of the QFT requires thousands of gates for each controlled rotation, to retain the rotation accuracy we set (10^{-41}).

search the lattice without constructing a graph of it. Our implementation has close to zero cost overheads with respect to memory and CPU instructions needed to translate back and forth between the lattice layout and the graph needed for performing Dijkstra’s algorithm. To further speed up routing, we employ a *cached routing* technique where previously computed routes are saved and reused later.

4 Results

We present results for compiling very large quantum circuits, and focus on scalability and resource estimation.

4.1 128-Qubit QFT

To validate the performance of our compiler and high performance slicer we took a circuit that has wide spread use and presents technical fault-tolerant execution challenges. The Quantum Fourier Transform (QFT) is a crucial component providing quantum speedup to algorithms such as Shor’s algorithm and quantum phase estimation [18]. The fault-tolerant implementation of the QFT is challenging because of the presence of small angle controlled rotations. For the QFT to retain the desired level of precision, these have to be approximated by a long sequence of Clifford+T, which results in a very long computation. We set Gridsynth’s precision to 10^{-41} for the Clifford+T approximations for small angle

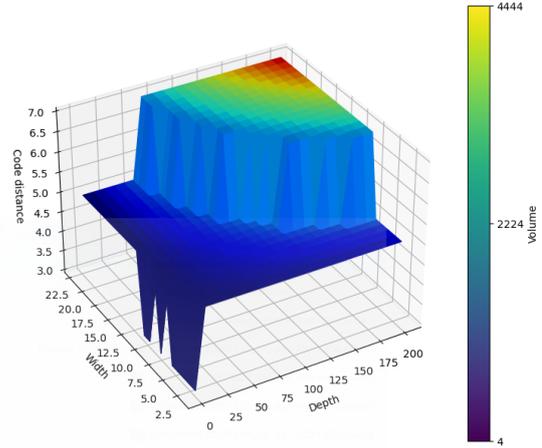


Figure 9: A snapshot of the resource requirement landscape for random H, T and CNOT circuits. The horizontal axes show circuit width (number of qubits) and depth (number of gates). The vertical axis shows the code distance required to execute the desired circuit with a success rate of 99%. The colour scale represents the space-time volume of the computation, which relates closely with code distance.

rotations, which results in thousands of gates. Such number was chosen as it is 3 orders of magnitude less than the smallest angle rotation in our circuit $\frac{\pi}{2^{128}} \approx 10^{-38}$, after expanding out the controlled rotations (Sec. 3.2).

The number of controlled rotations increases quadratically with the number of qubits the QFT is applied to. Thus, at 128 qubits and after small angle rotation approximation, the QFT circuit has more than 80 Million LLI without gate to Pauli compression. The number of LLI includes Clifford corrective terms that are meant to be applied depending on measurement outcomes. Thanks to concurrent magic state distillation, there are no idle slices waiting for magic states to be produced. We used the high performance slicer to compile the 128-qubit QFT: for example, laying out the slices for the roughly 80 million LLI of the 128-qubit QFT takes less than 15 minutes on an ordinary laptop. The generation of LLI of a QFT on 128 qubits takes negligible time (under 10s on a laptop). Fig. 8 illustrates the performance of the C++ slicer for the QFT128 circuit.

4.2 Resource Estimation

A challenging problem in the community of fault-tolerant QC is determining the amount of physical resource that is necessary to carry out a logical computation with a certain degree of precision. Such resource is often quantified by physical qubits over time – often called a *space time volume*. The depth of the circuit and the required magic state fidelity affect the code distance, which in turn affect the number of physical qubits required. Moreover, the degree of parallelization achieved at the routing stage will affect computation depth.

Our compiler includes a prototypical resource estimator for surface code computations. We use the Qentiana [54] software to estimate such values, and computed some code distances for randomized circuits of H, T and CNOT gates (Fig. 9).

5 Conclusion

We introduced and described a compiler for lattice surgery quantum circuits and showcased some of the results achieved with it. We motivated the design choices behind our two stage pipeline. The first stage included how input circuits are parsed, pre-processed, reduced to Clifford+T and viewed as Pauli rotations. The second stage focuses on laying out circuits on physical devices, which presents substantial performance challenges.

We demonstrated the compiler’s performance by compiling a 128-qubit QFT. We believe this is a notable achievement: despite its widespread appearance in algorithms, to the best of our knowledge, no surface code compiler is able to handle such large scale circuits. We also showcased the compiler’s ability to estimate resource requirements, in particular patch code distance, which is promising in the perspective of quantum benchmarking.

Our project is laying the foundation for a full-stack quantum circuit compilation framework. Future work will leverage hybrid classical and quantum instruction sets such as LLVM/QIR to program high performance classical control while integrating the Quantum Processing Unit (QPU) instructions.

Acknowledgements

We thank Tyler LeBond for contributing important code supporting the improved layout functionality, and Niki Loppi of the NVIDIA AI Technology Center Finland for his help with the implementation. We thank Varun Seshadri for his feedback and helpful discussions.

George Watkins and Alexandru Paler were with funding from the Defense Advanced Research Projects Agency [under the Quantum Benchmarking (QB) program under award no. HR00112230007 and HR001121S0026 contracts]. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Hoang Minh Nguyen, Keelan Watkins and George Watkins have been supported by the Unitary.fund. Hoi-Kwan Lau acknowledges support from the Canada Research Chair. Alexandru Paler acknowledges a Google Gift 2023, and funding received from the Finnish-American Research and Innovation Accelerator, one of eight global pilots funded by the Finnish Ministry of Education and Culture.

References

- [1] Martin Suchara, John Kubiawicz, Arvin Faruque, Frederic T. Chong, Ching-Yi Lai, and Gerardo Paz. “Qure: The quantum resource estimator toolbox”. In 2013 IEEE 31st International Conference on Computer Design (ICCD). Pages 419–426. (2013).
- [2] John Preskill. “Quantum Computing in the NISQ era and beyond”. *Quantum* **2**, 79 (2018).
- [3] S. Parker and M. B. Plenio. “Efficient factorization with a single pure qubit and $\log N$ mixed qubits”. *Phys. Rev. Lett.* **85**, 3049–3052 (2000).
- [4] Craig Gidney and Martin Ekerå. “How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits”. *Quantum* **5**, 433 (2021).
- [5] Frank Arute, Kunal Arya, Ryan Babush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell,

- et al. “Quantum supremacy using a programmable superconducting processor”. *Nature* **574**, 505–510 (2019).
- [6] Austin G Fowler, Matteo Mariantoni, John M Martinis, and Andrew N Cleland. “Surface codes: Towards practical large-scale quantum computation”. *Physical Review A* **86**, 032324 (2012).
- [7] David S Wang, Austin G Fowler, and Lloyd CL Hollenberg. “Surface code quantum computing with error rates over 1%”. *Physical Review A* **83**, 020302 (2011).
- [8] Yu Tomita and Krysta M Svore. “Low-distance surface codes under realistic quantum noise”. *Physical Review A* **90**, 062320 (2014).
- [9] Sebastian Krinner, Nathan Lacroix, Ants Remm, Agustin Di Paolo, Elie Genois, Catherine Leroux, Christoph Hellings, Stefania Lazar, Francois Swiadek, Johannes Herrmann, Graham J. Norris, Christian Kraglund Andersen, Markus Müller, Alexandre Blais, Christopher Eichler, and Andreas Wallraff. “Realizing repeated quantum error correction in a distance-three surface code” (2021).
- [10] Christian Kraglund Andersen, Ants Remm, Stefania Lazar, Sebastian Krinner, Nathan Lacroix, Graham J. Norris, Mihai Gabureac, Christopher Eichler, and Andreas Wallraff. “Repeated quantum error detection in a surface code”. *Nature Physics* **16**, 875–880 (2020).
- [11] Rajeev Acharya, Igor Aleiner, Richard Allen, Trond I Andersen, Markus Ansmann, Frank Arute, Kunal Arya, Abraham Asfaw, Juan Atalaya, Ryan Babbush, et al. “Suppressing quantum errors by scaling a surface code logical qubit”. *Nature* **614**, 676–681 (2023).
- [12] J. Eli Bourassa, Rafael N. Alexander, Michael Vasmer, Ashlesha Patil, Ilan Tzitrin, Takaya Matsuura, Daiqin Su, Ben Q. Baragiola, Saikat Guha, Guillaume Dauphinais, Krishna K. Sabapathy, Nicolas C. Menicucci, and Ish Dhand. “Blueprint for a Scalable Photonic Fault-Tolerant Quantum Computer”. *Quantum* **5**, 392 (2021).
- [13] Michael Hanks, Marta P. Estarellas, William J. Munro, and Kae Nemoto. “Effective compression of quantum braided circuits aided by zx-calculus”. *Phys. Rev. X* **10**, 041030 (2020).
- [14] Craig Gidney and Austin G. Fowler. “Flexible layout of surface code computations using autoccz states” (2019). [arXiv:1905.08916](https://arxiv.org/abs/1905.08916).
- [15] Daniel Herr, Franco Nori, and Simon J Devitt. “Optimization of lattice surgery is np-hard”. *npj Quantum Information* **3**, 35 (2017).
- [16] Kunihiro Wasa, Shin Nishio, Koki Suetsugu, Michael Hanks, Ashley Stephens, Yu Yokoi, and Kae Nemoto. “Hardness of braided quantum circuit optimization in the surface code”. *IEEE Transactions on Quantum Engineering* **4**, 1–7 (2023).
- [17] Alexandru Paler. “Aggregated control of quantum computations: When stacked architectures are too good to be practical soon”. *Computer* **53**, 74–78 (2020).
- [18] Michael A. Nielsen and Isaac L. Chuang. “Quantum computation and quantum information”. Cambridge University Press. (2000).
- [19] N David Mermin. “Quantum computer science: an introduction”. Cambridge University Press. (2007).
- [20] Savvas Varsamopoulos, Koen Bertels, and Carmen G. Almudever. “Decoding surface code with a distributed neural network-based decoder”. *Quantum Machine Intelligence* **2**, 1–12 (2020).
- [21] Mark Shui Hu and David Elkouss. “Quasi-linear time decoding algorithm for topological codes with high error threshold”. *Master’s thesis, TU Delft* (2020).
- [22] Barbara M. Terhal. “Quantum error correction for quantum memories”. *Rev. Mod. Phys.* **87**, 307–346 (2015).
- [23] Joschka Roffe. “Quantum error correction: an introductory guide”. *Contemporary Physics* **60**, 226–245 (2019).
- [24] A.Yu. Kitaev. “Fault-tolerant quantum computation by anyons”. *Annals of Physics* **303**, 2–30 (2003).

- [25] Google Research. “Google quantum ai journey” (2022).
- [26] IBM Research. “Ibm quantum roadmap” (2022).
- [27] S. B. Bravyi and A. Yu. Kitaev. “Quantum codes on a lattice with boundary” (1998). [arXiv:quant-ph/9811052](#).
- [28] Eric Dennis, Alexei Kitaev, Andrew Landahl, and John Preskill. “Topological quantum memory”. [Journal of Mathematical Physics](#) **43**, 4452–4505 (2002).
- [29] Dominic Horsman, Austin G Fowler, Simon Devitt, and Rodney Van Meter. “Surface code quantum computing by lattice surgery”. [New Journal of Physics](#) **14**, 123011 (2012).
- [30] Hendrik Poulsen Nautrup, Nicolai Friis, and Hans J Briegel. “Fault-tolerant interface between quantum memories and quantum processors”. [Nature communications](#) **8**, 1–8 (2017).
- [31] Daniel Litinski. “A game of surface codes: Large-scale quantum computing with lattice surgery”. [Quantum](#) **3**, 128 (2019).
- [32] Daniel Herr, Alexandru Paler, Simon J. Devitt, and Franco Nori. “Time versus hardware: Reducing qubit counts with a (surface code) data bus” (2019). [arXiv:1902.08117](#).
- [33] Sergey Bravyi and Alexei Kitaev. “Universal quantum computation with ideal clifford gates and noisy ancillas”. [Phys. Rev. A](#) **71**, 022316 (2005).
- [34] Sergey Bravyi and Jeongwan Haah. “Magic-state distillation with low overhead”. [Phys. Rev. A](#) **86**, 052329 (2012).
- [35] Daniel Litinski. “Magic state distillation: Not as costly as you think”. [Quantum](#) **3**, 205 (2019).
- [36] Alexandru Paler. “Surfbraid: A concept tool for preparing and resource estimating quantum circuits protected by the surface code” (2019). [arXiv:1902.02417](#).
- [37] Alexandru Paler and Austin G. Fowler. “Opensurgery for topological assemblies”. In 2020 IEEE Globecom Workshops (GC Wkshps. [Pages 1–4](#). (2020).
- [38] Lingling Lao, Bas van Wee, Imran Ashraf, J van Someren, Nader Khammassi, Koen Bertels, and Carmen G Almudever. “Mapping of lattice surgery-based quantum circuits on surface code architectures”. [Quantum Science and Technology](#) **4**, 015005 (2018).
- [39] Fei Hua, Yanhao Chen, Yuwei Jin, Chi Zhang, Ari Hayes, Youtao Zhang, and Eddy Z. Zhang. “Autobraid: A framework for enabling efficient surface code communication in quantum computing”. In MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture. [Page 925–936](#). MICRO ’21 New York, NY, USA (2021). Association for Computing Machinery.
- [40] Michael Beverland, Vadym Kliuchnikov, and Eddie Schoute. “Surface code compilation via edge-disjoint paths”. [PRX Quantum](#) **3**, 020342 (2022).
- [41] Xiaosi Xu, Simon C. Benjamin, and Xiao Yuan. “Variational circuit compiler for quantum error correction”. [Phys. Rev. Appl.](#) **15**, 034068 (2021).
- [42] Daniel Litinski and Felix von Oppen. “Lattice surgery with a twist: Simplifying clifford gates of surface codes”. [Quantum](#) **2**, 62 (2018).
- [43] Andrew W. Cross, Lev S. Bishop, John A. Smolin, and Jay M. Gambetta. “Open quantum assembly language” (2017). [arXiv:1707.03429](#).
- [44] H Abraham et al. “Qiskit: An open-source framework for quantum computing” (2021).
- [45] Aleks Kissinger and John van de Wetering. “PyZX: Large Scale Automated Diagrammatic Reasoning”. In Bob Coecke and Matthew Leifer, editors, Proceedings 16th International Conference on Quantum Physics and Logic, Chapman University, Orange, CA, USA., 10-14 June 2019. [Volume 318 of Electronic Proceedings in Theoretical Computer Science](#), pages 229–241. Open Publishing Association (2020).
- [46] Tomas Jochym-O’Connor, Aleksander Kubica, and Theodore J. Yoder. “Disjointness of stabilizer codes and limitations on

fault-tolerant logical gates”. *Phys. Rev. X* **8**, 021047 (2018).

- [47] Neil J. Ross and Peter Selinger. “Optimal ancilla-free clifford+t approximation of z-rotations”. *Quantum Info. Comput.* **16**, 901–953 (2016).
- [48] Alexandru Paler, Simon J Devitt, and Austin G Fowler. “Synthesis of arbitrary quantum circuits to topological assembly”. *Scientific reports* **6**, 1–16 (2016).
- [49] Cirq Developers. “Cirq”. *Zenodo* (2021).
- [50] Samuel Jaques, Michael Naehrig, Martin Roetteler, and Fernando Virdia. “Implementing grover oracles for quantum key search on aes and lowmc”. In *Advances in Cryptology – EUROCRYPT 2020: 39th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, May 10–14, 2020, Proceedings, Part II*. Page 280–310. Berlin, Heidelberg (2020). Springer-Verlag.
- [51] Guifré Vidal. “Efficient classical simulation of slightly entangled quantum computations”. *Phys. Rev. Lett.* **91**, 147902 (2003).
- [52] Bjarne Stroustrup. “Keynote speech: Abstraction and the c++ machine model”. In *Proceedings of the First International Conference on Embedded Software and Systems*. Page 1–13. ICES’04Berlin, Heidelberg (2004). Springer-Verlag.
- [53] Alexandru Paler, Austin G Fowler, and Robert Wille. “Synthesis of arbitrary quantum circuits to topological assembly: Systematic, online and compact”. *Scientific reports* **7**, 1–16 (2017).
- [54] Daniel Herr, Franco Nori, and Simon J Devitt. “Lattice surgery translation for quantum computation”. *New Journal of Physics* **19**, 013034 (2017).
- [55] JQ-Authors. “Command-line JSON processor”. <https://github.com/stedolan/jq> (2022).
- [56] Taewan Kim and Byung-Soo Choi. “Efficient decomposition methods for controlled-r n using a single ancillary qubit”. *Scientific Reports* **8** (2018).
- [57] Scott Aaronson and Daniel Gottesman. “Improved simulation of stabilizer circuits”. *Phys. Rev. A* **70**, 052328 (2004).

Appendix

5.1 OpenQASMMin

The natively supported subset of OpenQASM 2.0 instructions is presented in Table 2. Our compiler can parse a small subset of OpenQASM 2.0 instead of LLI. We call this type of assembly OpenQASMMin. In general, OpenQASMMin should be valid OpenQASM, with the restrictions below:

- Program must begin with `OPENQASM 2.0;`
- Only one register is allowed (whether the names match will not be checked)
- At most one gate per line
- Single qubit gates must be in the form `g q[n]`; where `g` is one of `h,x,z,s,t` and `n` is a non-negative integer
- `rz(expr)` and `crz(expr)` where `expr` has form `pi/m` or `n*pi/m` for `n, m` integers. No whitespace.
- CNOTs must be in the form `cx q[n],q[m]`; where `n` and `m` are non-negative, and target is first;
- No classical control is supported;
- No measurement operators are supported;
- Only inline comments, e.g. `cx q[0],q[7]; //cnot on 0 and 7.`

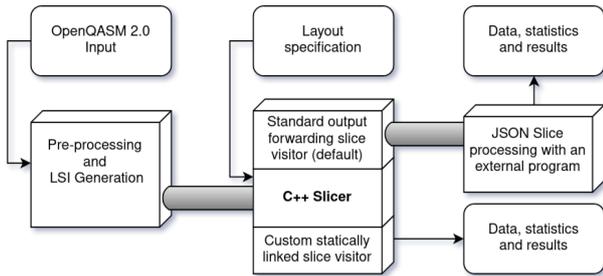
Table 2: OpenQASMMin: Supported instructions

Classical registers, barriers and include directives are ignored.	
Unsupported instructions and gates raise an exception.	
<code>h, x, z, s, sdg, t, tdg</code>	Single qubit gates with the usual meaning. The <code>-dg</code> suffix stands for dagger.
<code>cnot</code> or <code>cx, cz</code>	Controlled X and Z gates.
<code>rx(theta), rz(theta)</code>	R_Z and R_X gates. Argument <code>theta</code> has to have form $[N*\pi/D]$ for integers <code>N</code> and <code>D</code> , with the <code>D</code> being a non negative power of two. Both <code>N</code> and <code>D</code> can be of arbitrary size.
<code>crx(theta), crz(theta)</code>	Controlled R_Z and R_X gates. The argument has the same form as for the R_Z and R_X gates.
<code>qreg</code>	Only one quantum register is supported.

5.2 Compiler Pipeline and Operating Systems

It is possible to accept a wider range of circuits with additional processing by using Qiskit and PyZX. For example, we have successfully processed Grover circuits with multi qubit gates and Toffoli based adders by decomposing additional gates such as `mcx`, `ccx` and `rcx` which are not in the natively supported set. This kind of gate decomposition is done on a case by case basis.

Figure 10 is a diagram of how the C++ slicer can take advantage of the operating systems ability to broker messages by using POSIX pipes and an example shell command to run such a pipeline.



E.g.: `./qft_to_stdout.py 128 | lsqecc_slicer -l 12by12.txt | jq .[] [] [] | grep null | wc -l`

Figure 10: `qft_to_stdout.py` is our tool to stream LLI for a high precision QFT, the 128 argument indicates the number of qubits. `lsqecc_slicer -f json` is our C++ slicer, which lays out the LLI instructions and outputs a stream of JSON slices to `stout`, the `-l 12by12.txt` flag tells it which layout to use. Finally `jq` [55] is a streaming JSON processor that can extract information from the sequence of slices, here combined with some POSIX utilities to count the total number of routing cells.

5.3 Circuit Simulation and Gate Decomposition

For verification purposes, we require a state-vector snapshot of the logical state of the lattice computation at every time-step. To meet both goals, we use a form of state-vector tracking that stores patch states without tensoring non-entangled states together. This approach greatly extends what is possible to verify and debug compared to a naive state-vector simulator. Since this simulation is for debugging and verification purposes it is also able to detect when simulated states are numerically close enough to some common state such as

$|0\rangle, |+\rangle, |m\rangle$ and display them as such. Figure 11 is a graphical depiction of a circuit with intermediate states in a *Lazily Tensored State-vector Simulation (LTSvS)*.

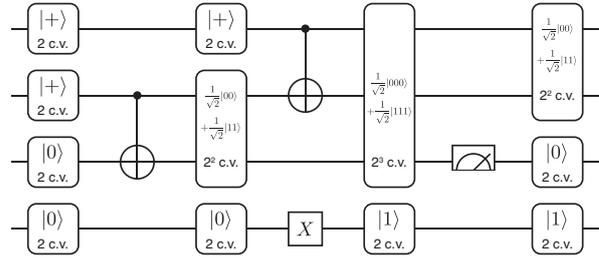


Figure 11: Simulating a deep lattice surgery computation is challenging, because there are constantly patches being entangled and measured out, but at a given time few are actually entangled. Each state is represented by certain number of complex variables (c.v.).

Conversion of gates and Pauli rotations to multi-body measurements and other LLI are presented in Fig. 12.

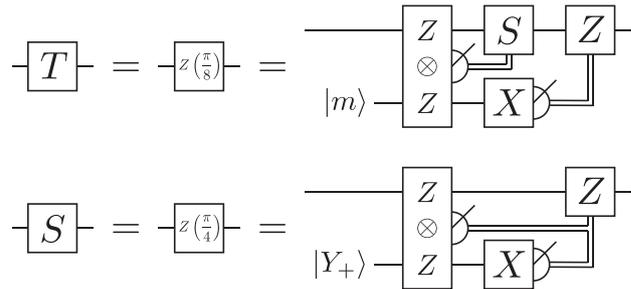


Figure 12: This figure only shows rotations by $\frac{\pi}{8}$ and $\frac{\pi}{4}$, but other angles with the same denominator are also possible by adjusting the classical logic controlling the corrective terms that follow. The $\frac{\pi}{8}$ Pauli rotations consume a distilled magic state, while the $\frac{\pi}{4}$ rotations consume a positive Y eigenstate which is prepared by applying a twist based measurement[42].

In Figure 13, controlled rotations are common in circuit such as the QFT. The first step towards converting them to fault-tolerant instructions is breaking them down into single qubit rotations and CNOTs [56]. Single qubit rotations of angles greater than $\frac{\pi}{8}$ have to be approximated, while CNOTs we implement with lattice surgery [29].

Fig. 14 illustrates the efficiency of the rotation compression technique we described in Fig. 6 from Sec. 3.2. How much compression we get exactly depends on the types of gate sequences appearing in the approximations (e.g. both `HSSSTH` and a lone `T` compress to a single Pauli rotation).

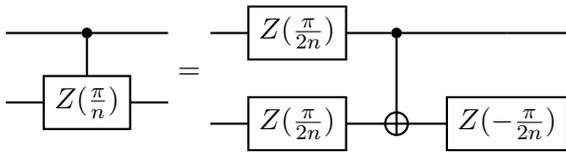


Figure 13: Decomposing controlled rotations into CNOTs and single qubit rotations.

Testing on rotations from $R_Z(\frac{\pi}{2^8})$ to $R_Z(\frac{\pi}{2^{128}})$ we observed a factor of ≈ 2.5 fewer LLI per small angle rotation – this is an illustration of our optimisation heuristics.

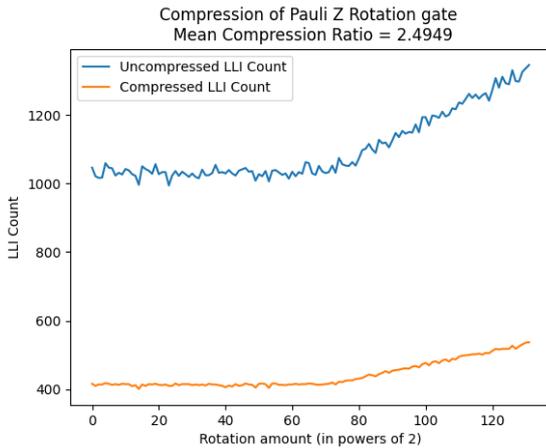


Figure 14: Benchmarking the decomposition techniques for arbitrary rotations to surface code instructions with and without Pauli rotation compression. The benchmark circuit is a single $R_Z(\frac{\pi}{2^n})$ rotation. It is possible to see how grouping gates that rotate in the same basis (as shown in Fig. 6) drastically reduces the number of required surface code instructions.

5.4 Clifford Elimination

According to the Gottesman-Knill theorem, it is possible to efficiently simulate circuits which only contain a particular set of gates, known as *Clifford* gates [18, 57]. It is natural to ask whether it is possible to leverage classical computing to reduce the load on the QPU when processing such circuits. Litinski [31] outlined an algorithm to remove the Clifford part of the circuit at compile time, when all we care about are measurement outcomes (i.e. we are not compiling the circuit to be a state preparation routine). We call this algorithm the Litinski Transform (LT) and provide an implementation in the following manner.

The first step of LT is to convert each gate of the input circuit into a sequence of rotations

about $\frac{\pi}{2}$, $\frac{\pi}{4}$ or $\frac{\pi}{8}$, or multibody measurements with Pauli product observables. Of these blocks, only the $\frac{\pi}{8}$ rotations are not Clifford, so we apply Litinski’s commutation rules to bring them all to the front of the circuit. Next the $\frac{\pi}{2}$ and $\frac{\pi}{4}$ rotations are commuted past the previously end-of-circuit measurements. Since this case we only care about measurement outcomes, the Clifford rotations that are now after measurement can be discarded.

5.5 Layout File Format

The layout determines how computation elements are placed on the lattice. Herein we describe the technical details, limitations, examples and future enhancements of the layout format introduced in Sec.3.3. The online source code repository includes more layout examples.

Structure. The purpose of our layouts is to define the structure of a lattice intended for lattice surgery operations, abstracting the details of the physical implementation. The layout is stored as an ASCII plain text file, in order to make it easy for humans to view and edit and to enable good portability (i.e. special tools needed to edit). The layout is specified by a grid of ASCII characters where each cell has a specific meaning based on the character it contains. For example, the text `QrQ` represents two logical qubits separated by an inactive routing space:



The layout is always assumed to be rectangular. If a layout file is not rectangular in content, then the compiler will assume that the bounding box of its contents is available and pad with empty routing space.

Qubits Q: Represents a patch holding a logical qubit encoded using the surface code planar code. The boundaries are assumed to be rough north-south, and smooth east-west, so it’s the users responsibility that connectivity between each qubit is possible.

Routing r: In their default or *quiescent* state, these cells are inactive. This means they do not actively participate in quantum computations. However, when required, the contents of these routing cells can be “activated” to facilitate long-range merges and splits between distant Q cells,

as shown in Fig. 4.

Ancilla A: These cells are reserved to allocate new ancillae patches. Ancillae patches are auxiliary qubits used in quantum computations to assist in the construction of measurement-based gates. These patches can be in states like the $|+\rangle$ state, which mediates CNOT operations, or they can be places for the Y eigenstates used by $\frac{\pi}{4}$ Pauli rotations.

Distillation regions numbers 0–9: Distillation regions are specialized areas designated for the production of magic states, as defined in section 2.2. These regions are represented by areas with the same number in the ASCII layout. The extents of these regions are identified by running a connected components search on areas with matching numbers. This means that contiguous cells with the same number are considered part of the same distillation region.

The magic states produced by these regions are output to a neighboring `r` cell. Distillation time is assumed to be the same for every distillation region. The compiler makes no assumptions about the internal operations or processes occurring within distillation regions. It is the user’s responsibility to ensure that the size and configuration of a region is correct.

Example 1. The layout below supports two logical qubits, and routing. No two-qubit gates can be applied immediately, because there are no `A` patches.

```
QrQ
rrr
```

Example 2. Seven qubits and four distillation regions can operate on the layout below. The `A` ancilla can be used for performing CNOTs between the logical qubits, for example.

```
rrrrrrr444
rQrQr444
rQrQAr444
rrrrrrrrr
r11222333
r11222333
r11222333
```

Planned future extensions of the layout file include: a) incorporation of multi-cell patches support; b) qubit indexing enhancements; c) initialization of square patches with alternate boundary configurations; d) introduction of a in browser editor equipped with syntax highlighting.