

# Quantum simulation with just-in-time compilation

Stavros Efthymiou<sup>1</sup>, Marco Lazzarin<sup>1</sup>, Andrea Pasquale<sup>1,2</sup>, and Stefano Carrazza<sup>2,3,1</sup>

<sup>1</sup>Quantum Research Centre, Technology Innovation Institute, Abu Dhabi, UAE.

<sup>2</sup>TIF Lab, Dipartimento di Fisica, Università degli Studi di Milano and INFN Sezione di Milano, Milan, Italy.

<sup>3</sup>CERN, Theoretical Physics Department, CH-1211 Geneva 23, Switzerland.

Quantum technologies are moving towards the development of novel hardware devices based on quantum bits (qubits). In parallel to the development of quantum devices, efficient simulation tools are needed in order to design and benchmark quantum algorithms and applications before deployment on quantum hardware. In this context, we present a first attempt to perform circuit-based quantum simulation using the just-in-time (JIT) compilation technique on multiple hardware architectures and configurations based on single-node central processing units (CPUs) and graphics processing units (GPUs). One of the major challenges in scientific code development is to balance the level of complexity between algorithms and programming techniques without losing performance or degrading code readability. In this context, we have developed qibojit: a new module for the Qibo quantum computing framework, which uses a just-in-time compilation approach through Python. We perform systematic performance benchmarks between our JIT approach and a subset of relevant publicly available libraries for quantum computing. We show that our novel approach simplifies the complex aspects of the implementation without deteriorating performance.

## 1 Introduction

The growing interest in quantum technologies for computational tasks which could exceed classical devices performance has received a boost thanks to the availability of noisy intermediate-scale quantum (NISQ) devices [1] and recent promising results [2, 3]. We observe important steps towards the development of stable and efficient quantum processing units (QPUs), following the gate-based model of quantum computation [4–7] or quantum annealing [8, 9].

Despite the effort in QPU technology development, aspects involving theory and modeling do still require classical simulation of quantum computing to develop new algorithms and applications. High performance quantum simulation serves as a testing and profiling tool for the development of quantum algorithms,

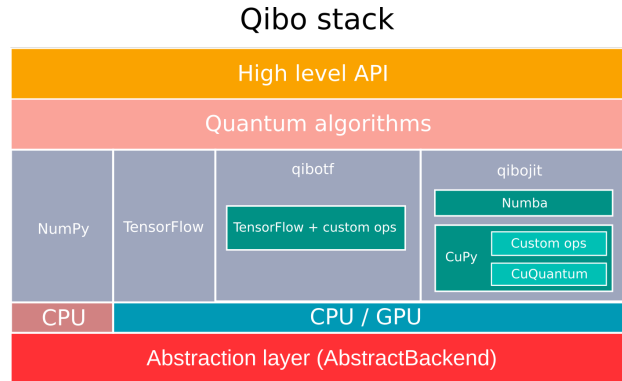


Figure 1: Schematic view of the Qibo structure design.

while from an experimental point of view it provides a reference for benchmarks and error simulation.

In quantum computing the state  $\psi$  of a system of  $n$  qubits is represented by a vector of  $2^n$  complex probability amplitudes in the computational basis. In Schrödinger’s approach of quantum simulation [10, 11], each gate is applied to the state via the following matrix multiplication

$$\psi'(\sigma_1, \dots, \sigma_n) = \sum_{\tau'} G(\boldsymbol{\tau}, \boldsymbol{\tau}') \psi(\sigma_1, \dots, \tau'_1, \dots, \sigma_n) \quad (1)$$

where the gate targeting  $n_{\text{tar}}$  qubits is represented by the  $2^{n_{\text{tar}}} \times 2^{n_{\text{tar}}}$  complex matrix  $G(\boldsymbol{\tau}, \boldsymbol{\tau}') = G(\tau_1, \dots, \tau_{n_{\text{tar}}}, \tau'_1, \dots, \tau'_{n_{\text{tar}}})$  and  $\sigma_i, \tau_i \in \{0, 1\}$ . The numerical solution to Eq. 1 requires the manipulation of state vectors of size  $2^n$ , which scales exponentially with the number of qubits, and the subsequent linear algebra operations related to the application of unitary gates. Thus, quantum simulation tools on classical hardware need to take into account both challenges and provide efficient solutions.

In this context, we have developed the Qibo [12–14] framework, an open-source, full stack API written in Python, which supports circuit-based quantum simulation, adiabatic evolution simulation and quantum hardware control [15]. The Qibo structure since release 0.1.7 is shown in Fig. 1. The high-level API and pre-coded quantum algorithms are implemented following a backend agnostic approach. Each backend provides specialized methods to achieve maximum performance on multiple devices, including hardware accelerators, such as multi-threading CPU, GPU and

multi-GPU configurations.

The main disadvantage associated with the development and maintenance of scientific software with parallel computing and hardware acceleration support is the need to maintain a large code-base of algorithms defined in compiled languages (such as Fortran, C++ and CUDA). This requires a non-negligible level of programming experience for the developer. Furthermore, testing and deployment of these codes requires custom workflows which should build pre-compiled binaries for a target subset of platforms and architectures.

To address these issues, we published the `qibojit` backend [16] which supports efficient circuit-based quantum simulation through just-in-time (JIT) [17, 18] compilation, with Python as input programming language interface. The Python JIT approach provides to the code developer the possibility to maintain a modern project layout, with automatic documentation, testing workflows and standard deployment procedure with minor changes to the algorithmic part of the code. The code readability and homogeneity are preserved, and it makes the installation on different platforms easier. In this paper we first present the layout adopted by `qibojit` and then perform a systematic benchmark to quantify the impact on performance for quantum computing tasks.

Finally, it is important to highlight that similar quantum simulators are implemented by other research collaborations and companies. Some examples included in the benchmark section of this work are Qiskit [19] from IBM, Cirq and qsim [20, 21] from Google, ProjectQ [22, 23] by ETH Zürich, HybridQ [24] by NASA, Qulacs [25] and QCGPU [26].

The paper is organized as follows. In Sec. 2 we present the technical details of the `qibojit` implementation as a module for the Qibo framework, highlighting the code design and structure. The Sec. 3 presents performance benchmarks of all Qibo backends as well as other quantum simulators. Finally, in Sec. 4 we present our conclusion and outlook.

## 2 Methodology

Qibo provides multiple backends for implementing the matrix multiplication of Eq. 1 which are based on different technologies including pre-compiled binaries and just-in-time compilation as shown in Fig. 1. All backends inherit from the `AbstractBackend` class and define its properties and methods using primitives provided by Python libraries, such as NumPy [27], and custom operations coded in Python or low-level languages such as C++ and CUDA. The abstract methods include general algebraic and linear algebra operations, such as element wise vector operations, tensor products, eigenvalue and eigenvector methods, as well as specialized operations for applying gates to state vectors and density matrices, following Eq. 1.

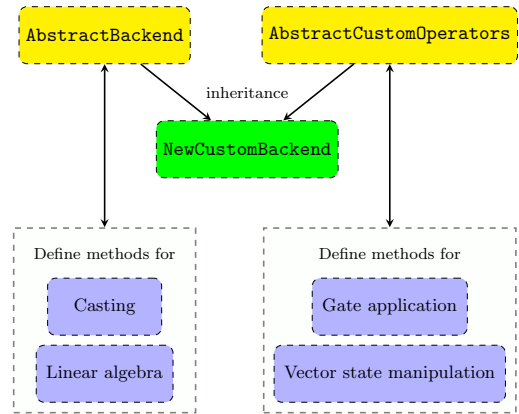


Figure 2: Flowchart describing how to implement a new custom backend.

This abstraction layer allows us to disentangle the core Qibo code and applications from a specific backend or Python library. Furthermore, it provides a layout for users that would like to define a new backend that is compatible with Qibo. This can be done just by inheriting from the `AbstractBackend` and the `AbstractCustomOperator` and defining their abstract methods as shown in Fig. 2.

The basic backends included in the Qibo package, distributed from PyPI [28] and conda-forge [29], are `numpy` and `tensorflow`. These backends implement basic algebraic operations using primitives of the underlying library, NumPy [27] and TensorFlow [30]. The specialized operations for applying gates are based on the `einsum` method which is exposed as a primitive in both libraries. This provides satisfactory performance when simulating circuits up to 20 qubits. The `numpy` backend is available by default when installing Qibo and it is designed to support a high number of architectures, including `arm64`, and thus be deployed in multiple contexts, including laboratory devices. The optional `tensorflow` backend provides moderate performance and the possibility to perform automatic differentiation which is useful for quantum machine learning applications.

To efficiently simulate circuits with a larger number of qubits, we extend these basic backends with custom operators. In particular, `tensorflow` is extended by `qibotf` [31] and `numpy` is extended by `qibojit` [32]. These are not included in the basic Qibo library but can be installed as separate Python packages. These backends keep using their parent libraries (NumPy and TensorFlow) for basic algebraic and linear algebra operations, however the gate application methods are replaced by custom operators. Unlike the `einsum` approach, which duplicates the state vector while applying a gate, custom operators perform in-place updates. This reduces both memory requirements and execution time since custom operators modify directly the initial state vector based on the gates applied.

Furthermore, the custom operators exploit the sparsity of matrices associated with some common operations, such as Pauli gates and controlled gates, to reduce the number of operations required to apply each gate. In particular, if the application of a specific gate modifies only a few components of the initial state, using custom operators we update directly these particular elements, avoiding the matrix multiplication of Eq. 1.

The basic custom operator defines the application of an arbitrary single-qubit gate to a state vector. An example of this operator for the `qibojit` backend is shown below.

---

```

from numba import njit, prange

@njit(parallel=True, cache=True)
def apply_gate_kernel(state, gate, target):
    """Operator that applies an arbitrary one-qubit gate.

    Args:
        state (np.ndarray): State vector of size (2 **
        nqubits,).
        gate (np.ndarray): Gate matrix of size (2, 2).
        target (int): Index of the target qubit.
    """
    k = 1 << target
    # for one target qubit: loop over half states
    nstates = len(state) // 2
    for g in prange(nstates):
        # generate index with fast binary operations
        i1 = ((g >> m) << (m + 1)) + (g & (k - 1))
        i2 = i1 + k
        state[i1], state[i2] = (gate[0, 0] * state[i1] + \
                                gate[0, 1] * state[i2],
                                gate[1, 0] * state[i1] + \
                                gate[1, 1] * state[i2])
    return state

```

---

Additional operators that follow a similar approach are used to apply gates with more target qubits, as well as controlled gates. All these operators take advantage of multi-threading CPUs and GPUs by parallelizing the loop over state elements, the cost of which scales exponentially with the number of qubits. Furthermore, we provide specialized operators for applying Pauli X, Y and Z gates and the SWAP gate, which use more simplified kernels inside the loop. In order to simulate real measurements, we provide a custom operator for collapsing and re-normalizing states and a method for sampling shot frequencies based on Metropolis algorithm [33]. Both `qibotf` and `qibojit` define the same custom operators but use different technologies to interface them with Python. These are analyzed in what follows.

In `qibotf` we use TensorFlow custom operators written in C++ and CUDA. These need to be compiled before the execution, a step that typically improves performance but could complicate installation and make it very device specific. Nevertheless, this is the first custom backend released for Qibo and includes multi-threading CPU, GPU and multi-GPU support.

The latest backend added to Qibo is `qibojit`, which implements custom operators based on a just-in-time compilation approach. `qibojit` also supports

multi-threading CPU, GPU and multi-GPU configurations.

For CPU we write operators in Python using Numba’s [17] `njit` decorator with a set of signatures for each function that specify both return and argument types. This decorator compiles the Python code using LLVM. Moreover, the loop over the state elements is parallelized using Numba’s `numba.prange` method. To further speed up the circuit execution the appropriate indices for each update are generated on-the-fly using fast binary operations.

For GPU, we choose Cupy [18] as the main driver, which enables us to follow the CPU approach based on on-the-fly compilation. We also tried different GPU backends, including Numba and Jax [34] for Python or in C++ Eigen3 [35], ViennaCL [36] and NVIDIA thrust [37]. The main problems with these options concern the lack of linear algebra operations and the difficulties in writing custom operators. We decided not to choose Numba because we observe a significant overhead when simulating circuits with a small number of qubits. The implementation of the custom operators in the Cupy backend was performed using the `RawKernel` method, which allows us to define custom CUDA kernels written in C++, which are compiled using `nvcc` [38] at their first invocation and cached for each device. This method also takes care of exposing these compiled kernels to Python. Another positive aspect of Cupy is the compatibility with AMD ROCm, which enables us to run Qibo on setups with ROCm-compatible GPUs.

We also provide a different GPU simulator, within the `qibojit` backend, which is based on `cuQuantum` [39], a quantum simulation library from NVIDIA. Its addition to Qibo was facilitated since the main driver is Cupy which is already employed by `qibojit`. This backend replaces the custom kernels with primitives from the `cuQuantum` library. The main advantage is the fact that we no longer need to write C++ or CUDA code to achieve good performances with large number of qubits. However, using an external library instead of custom operators comes at the cost of having less control over the code and there can also be some missing features that need to be included manually. In our case, if a particular operator is not defined in the `cuQuantum` library, the compatibility with Cupy allows us to fall back to the custom operators of the Cupy backend to maintain good performances without complicating the code.

For a full list of primitives and models for quantum computing simulation available in Qibo 0.1.7 please refer to [12] and Sec. 3 in [14].

### 3 Benchmarks

In this section we compare performance of the different Qibo backends and other open-source libraries on various tasks, including simulation of quantum

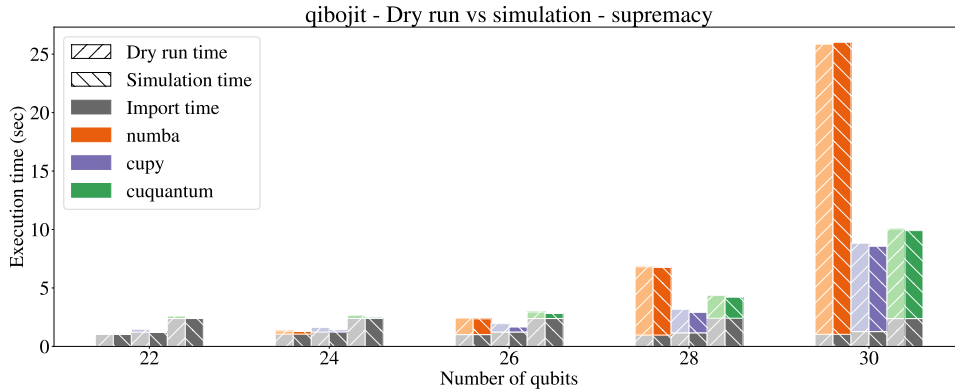


Figure 3: Comparison between import, dry run and simulation times for the three platforms of the qibojit backend.

Name	Version	Distribution
qibo	0.1.7	pip
qibojit	0.0.4	pip
qibotf	0.0.6	pip
tensorflow	2.8.0	pip
numba	0.55.1	pip
cuda-toolkit	11.6.0	conda
cupy	10.1.0	conda
cuquantum	0.1.0.30	conda-forge
cuquantum-python	0.1.0.0	conda-forge

Table 1: Versions of Qibo and its dependencies used in the benchmarks.

circuits as well as adiabatic time evolution. The benchmarks were performed with an AMD EPYC 7742 CPU with 128 threads and 2TB of RAM and a NVIDIA RTX A6000 GPU with 48GB of memory, unless otherwise noted. Qibo was installed in a Python 3.9 conda environment with the dependencies shown in Table 1. The source code used to generate the results in this section is publicly available in the following repository [40].

### 3.1 Circuit Simulation

The quantum circuits used in our benchmarks are shown in Table 2. All circuits are defined using the OpenQASM [41, 42] language and ported to each simulation library. Some libraries allow importing circuits directly from OpenQASM, while for other libraries we coded the parsing manually. The qft, variational and bv circuits are defined directly in OpenQASM using the corresponding gates. The supremacy circuit is created using Cirq [43] and the qv circuit using Qiskit [44] and are both ported to OpenQASM in order to be converted to the different libraries.

When benchmarking libraries which involve just-in-time compilation it is important to distinguish the first execution because it will involve a compilation or loading of cached binaries and therefore will be

slower than subsequent executions in the same run time. In what follows, we call this first run as *dry run* and any subsequent run as *simulation*. Fig. 3 shows the difference between these two runs for simulating the supremacy circuit using the different platforms (numba, cupy and cuquantum) of Qibo’s qibojit backend. We observe that the difference between the first (dry run) and second (simulation) run is negligibly small on CPU (numba) but slightly higher on GPU (cupy, cuquantum). Note that qibojit implements a caching algorithm for custom operators which are generated during installation, thus in this case negligible performance differences between dry and simulation run-times are expected. Furthermore, a constant of about one second is required to import the library, which can be relevant (comparable or larger than execution time) for simulation of small circuits. This is unlikely to impede practical usage as it is only a small constant overhead that is independent of the total simulation load.

In Fig. 4 we show how the dry run (left) and simulation (right) time to execute the qft circuit scales with the number of qubits for different Qibo backends. These plots show the total time a user would experience when simulating the circuit, which includes the library import, allocation of the circuit and gate objects and finally execution on the specified hardware. Up to 20 qubits this is dominated by import time and lightweight CPU backends such as numpy are the optimal choice. For larger circuits, the custom qibojit and qibotf backends which take advantage of multi-threading CPU and GPU architectures provide a much more favorable scaling. Moreover, qibojit provides better performance than qibotf despite its code simplicity. We also note that qibojit and qibotf perform in-place updates, in contrast to numpy and tensorflow which duplicate the state vector, a feature that reduces both time and memory requirements significantly.

In order to quantify the advantages associated with the JIT approach in Table 3 we show both memory footprints and execution times separated in dry run and simulation when executing a 26 qubits qft circuit.

Name	Notation	Source	Depth	Gates	Depth*	Gates*
Quantum Fourier Transform [45]	qft	Qibo	60	480	58	450
Variational [46]	variational	Qibo	4	90	2	30
Supremacy [2, 47]	supremacy	Cirq	4	98	2	22
Quantum Volume [48]	qv	Qiskit	7	165	1	15
Bernstein-Vazirani [49]	bv	Qibo	32	89	29	29

Table 2: Description of circuits used in the benchmarks. The circuit depths and the number of gates shown are referred to 30 qubits circuits. In the last two columns we show the circuit depths and the number of gates after applying gate fusion.

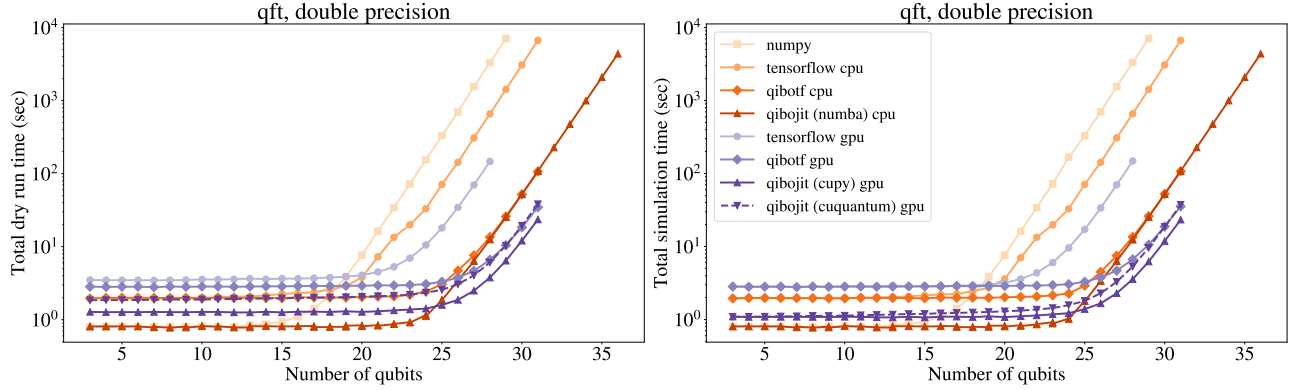


Figure 4: Total dry run (left) and simulation (right) time scaling with the number of qubits for simulating the qft circuit using different Qibo backends.

We observe that the considerable reduction of the execution times does not imply an increase in memory usage. In fact, the backends that implement custom operators and in-place updates, `qibojit` and `qibotf`, require less memory in order to perform the simulation compared to the `tensorflow` and the `numpy` backend due to the multiple copies of the state vector employed by these two.

In Fig. 5 we show the total dry run and simulation times scaling with the number of qubits in the circuit, but now focusing on the `qibojit` backend and using it on different hardware configurations. As mentioned earlier, CPU is preferable for smaller circuits due to faster import times, which are dominating execution time at this region. We observe that `qibojit` can reach high qubit values thanks to its state vector in-place memory updates and it can operate on multiple systems, including commercial solutions such as ATOS QLM [50] hardware. For circuits with more than 25 qubits the exponential scaling starts to appear, and high-end GPUs provide an advantage. Lower end GPUs, such as the NVIDIA GTX 1650 do not seem to provide any advantage over a powerful CPU and their limited memory (4 GB) prohibits the simulation of circuits with more than 27 qubits. The newest NVIDIA RTX A6000 is the fastest of our devices. Moreover, in order to test `qibojit` performance on AMD ROCm GPUs we include numbers for the AMD Radeon VII with 16GB which confirms competitive results.

To take full advantage of GPU acceleration for large

circuits, Qibo provides the possibility to simulate circuits on multiple GPU devices. This is useful because the maximum number of qubits that can be simulated in a GPU is limited by its internal memory. In Qibo’s multi-gpu scheme, the full state vector is stored in the host memory, which is typically larger than the GPU memory and only slices of it are transferred to the GPUs for calculation. For technical details of this implementation we refer to Sec. 2.5 Ref. [12]. The multi-gpu scheme can be used with multiple physical GPU devices, if available, but also with a single GPU that is re-used for multiple state slices during the calculation. The multi-gpu feature is supported by the `qibojit` and `qibotf` backends only.

In Fig. 6 we plot the times for simulating different circuits of 32 qubits using multiple GPUs. The NVIDIA DGX workstation [51] with four Tesla V100 (32GB) GPUs, Intel Xeon E5-2698 CPU and 256GB of RAM was used for this benchmark. For each circuit we distribute the execution to one, two or four physical GPUs. The state slices are processed in parallel when different physical devices are used, while they are processed sequentially if the same physical device is used. Therefore, increasing the number of physical devices leads to better performance for both backends. We also observe that, even though for `qibotf` there is no significant variation between dry run and subsequent simulations, for `qibojit` dry run appears slower, particularly when multiple physical devices are used. This happens because parts of the calculation are not executed in parallel in different devices

backend	$\Delta m$ (MB)	m(MB)	dry run(s)	simulation(s)
qibojit (cupy)	695.55	1093.55	1.92	0.60
qibojit (cuquantum)	406.86	1804.09	1.035	0.86
qibotf (GPU)	654.49	3260.28	0.76	0.76
tensorflow (GPU)	1469.14	4072.92	31.89	30.58
qibojit (numba)	903.23	1146.02	2.67	2.57
qibotf (CPU)	735.21	1276.16	2.52	2.34
tensorflow (CPU)	4845.74	5385.39	147.01	146.77
numpy (CPU)	3005.98	3248.69	697.12	698.65

Table 3: Memory usage, dry run and simulation times for different backends when simulating the qft circuit with 26 qubits. m denotes the maximum memory usage during the execution, while  $\Delta m$  represents the difference between m and the memory required for importing the dependencies.

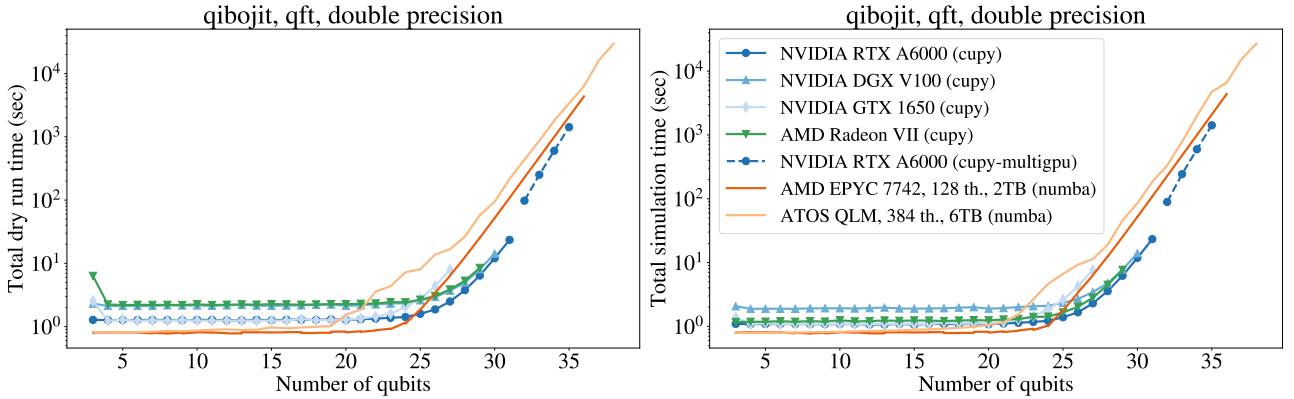


Figure 5: Total dry run (left) and simulation (right) time scaling with the number of qubits for simulating the qft circuit on different devices.

Name	Version	Precision	Hardware
Qibo	0.1.7	single/double	CPU/GPU
Qiskit	0.34.2 [52]	single/double	CPU/GPU
Qulacs	0.3.0	double	CPU/GPU
ProjectQ	0.7.1	double	CPU
qsimcirq	0.12.0	single	CPU/GPU
QCGPU	0.1.1	single	GPU
HybridQ	0.8.1	single/double	CPU/GPU

Table 4: Simulation libraries used in the benchmarks.

during the just-in-time compilation step.

Finally, we performed comparisons with other open-source Python libraries for quantum simulation. The libraries used are shown in Table 4 with their corresponding versions. We focused on libraries that are compatible with multiple general-purpose, high-performance hardware configurations, including multi-threading CPU and GPU. Some libraries allow switching between single (complex64) and double (complex128) precision, while others support only a specific precision, therefore we provide different comparisons for each case.

In Fig. 7 we plot the comparison with different simulation libraries. In each case we use all libraries that

support that precision, and we benchmark each circuit from Table 2 for 20 and 30 qubits on all supported hardware configurations (multi-threading CPU and GPU). Optimizations such as gate fusion were disabled on all libraries for this benchmark and the qibojit backend (numba/cupy) was used for Qibo. We do not include results for qsimcirq in this section, as it is not possible to disable gate fusion for this library. We find that Qibo is slightly slower, though still competitive when compared to other libraries, for circuits of 20 qubits. This is primarily due to the import time and the time required to load the kernels from disk, associated with just-in-time compilation. These times are comparable to execution time for small circuits. The situation is reversed for 30-qubit circuits where kernel loading times are less relevant and Qibo is considerably fast, particularly on GPU.

### 3.2 Gate Fusion

Gate fusion [53–55] is a commonly used approach to speed up simulation of quantum circuits. Multiple gates are fused together by multiplying their underlying matrices and applying them to the state vector as a single gate. This is preferable to naively applying the gates one-by-one, particularly when simulat-

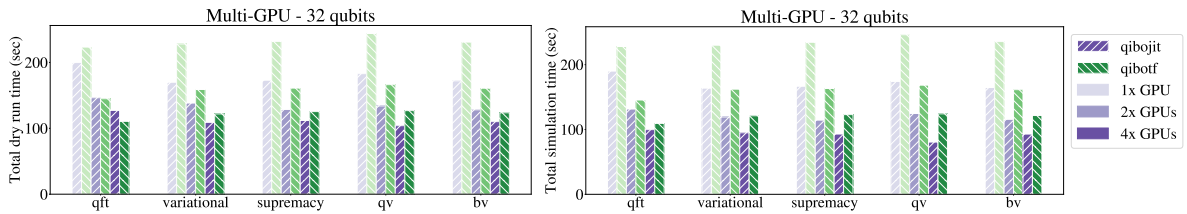


Figure 6: Total dry run (left) and simulation (right) time for simulating 32-qubit circuits using multiple GPUs.

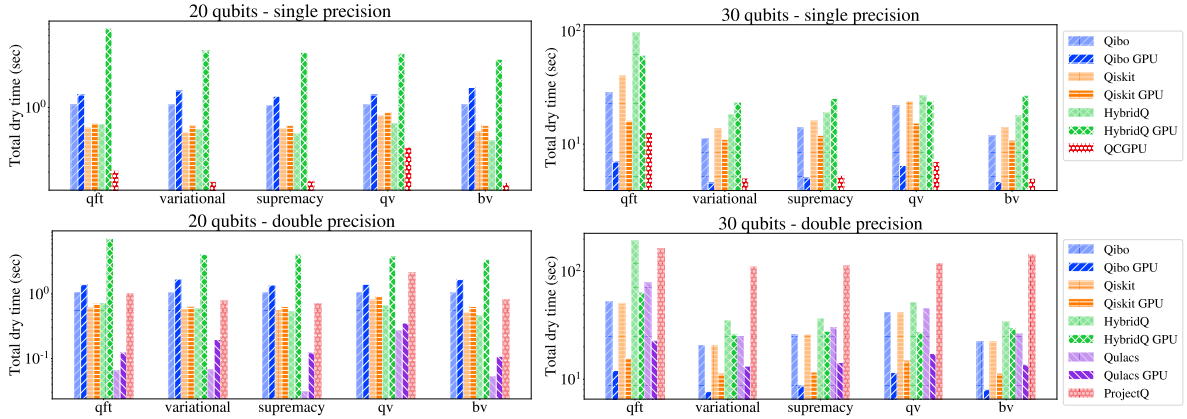


Figure 7: Total dry run time for simulating different circuits of 20 qubits (left) and 30 qubits (right), using libraries that support single (top) and double (bottom) precision.

ing circuits with many qubits, because multiplying small gate matrices is computationally cheaper than multiplying a gate matrix with the exponentially large state vector. Qibo provides a simple algorithm for fusing gates up to two target qubits. This works by iterating over the circuit gates and greedily combining one-qubit and two-qubit gates that act on the same target qubits.

In Fig. 8 we compare all `qibojit` platforms for simulating different 30-qubit circuits with and without fusion. The depth and the number of gates before and after the fusion are shown in Table 2. Fusion provides significant speed-up, particularly when using CPU. However, it is important to note that this speed-up depends on the circuit that is simulated. For example, gate fusion does not help much in the qft circuit. In Fig. 9 we compare different libraries on simulating different circuits with gate fusion enabled. The maximum number of target qubits for a fused gate was set to two for all libraries, in order to be consistent with the fusion algorithm used in Qibo. Other libraries support fusion with higher maximum number of target qubits. The most significant advantage appears when switching from no fusion to two-qubit fusion. Further increasing the maximum number of qubits in fused gates up to about five, may provide additional advantage for some circuits.

### 3.3 Adiabatic Evolution

Qibo provides functionality for simulation of unitary time evolution under arbitrary Hamiltonians. A spe-

cial case is adiabatic evolution, a typical method for finding ground states of Hamiltonians [56, 57], which is provided as a special Qibo model. The trivial algorithm for unitary time evolution calculates the exponential of the Hamiltonian matrix  $e^{-iH(t)\delta t}$  at each time  $t$  where  $\delta t$  is a pre-defined time step. This approach is not feasible for large systems as the Hamiltonian matrix for  $n$  qubits has size  $2^n \times 2^n$ . For such cases, an alternative approach based on the Trotter decomposition [58] is provided in Qibo and can be used out-of-the-box, with the decomposition being handled automatically by the library.

Here we benchmark the adiabatic evolution with the transverse-field Ising model (TFIM) as the target Hamiltonian, starting from the easy to prepare Hamiltonian that is sum of Pauli X operators. We use a system of 10 qubits and plot the scaling of execution time with the time step  $\delta t$  used in evolution, using both the matrix exponentiation (Fig. 10) and Trotter decomposition (Fig. 11 top) methods. With the Trotter decomposition we can also simulate a 20-qubit system (Fig. 11 bottom) which is intractable when using the full Hamiltonian matrix exponentiation. As expected, the full exponentiation is computationally heavier and requires more time. Moreover, this approach does not make use of custom operators but is based on numpy (CPU) and cupy (GPU) primitives for the `qibojit` backend and tensorflow (CPU and GPU) primitives for the tensorflow and `qibotf` backends. In contrast, when the Trotter decomposition is used, the time evolution is decomposed into a circuit of unitary gates and all functionalities presented in

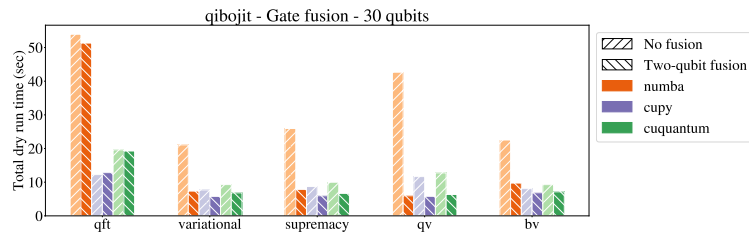


Figure 8: Dry run time for simulating different circuits of 30 qubits using qibojit with and without gate fusion.

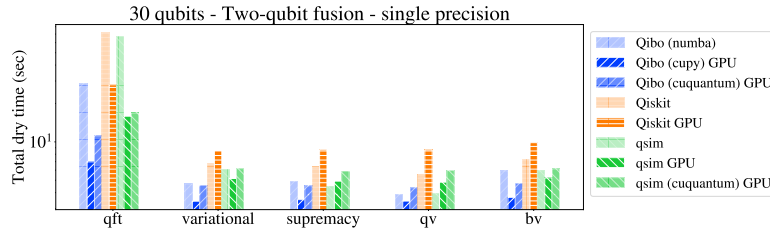


Figure 9: Dry run time for simulating different circuits of 30 qubits using different simulation libraries with gate fusion enabled.

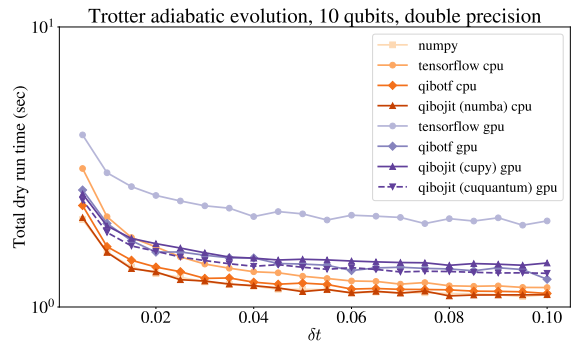
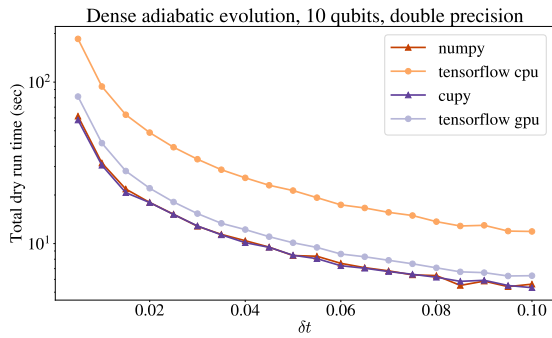


Figure 10: Total dry run time for simulating adiabatic evolution of 10 qubits using the full Hamiltonian matrix.

Sec. 3.1 can be used. CPU is faster than GPU when simulating 10 qubits, however the situation is reversed for 20 qubits, similar to what we observed in circuit simulation.

## 4 Conclusion

In this work we present the implementation of `qibojit`, a just-in-time compiled quantum simulator module for Qibo, with support on multi-threading CPU and hardware accelerators (GPU and multi-GPU). We show that the modular backend agnostic layout provided by Qibo simplifies the inclusion of new modules with minor costs in terms of development time and maintainability.

Following the benchmark results presented in Sec. 3 we can confirm that `qibojit` performance is acceptable and the impact of dry run is negligible in most cases. The possibility to share the state vector representation with external libraries such as `cuQuantum`, enhances furthermore the capabilities of this module by allowing to obtain immediate performance benefits

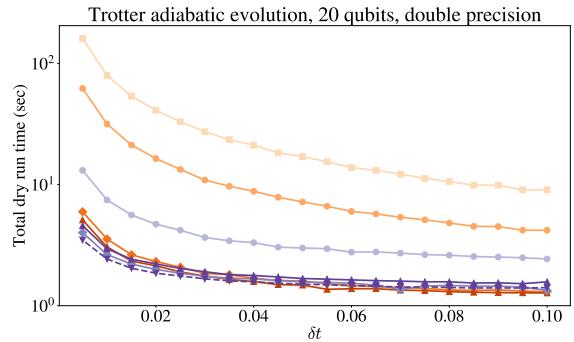


Figure 11: Total dry run time for simulating adiabatic evolution of 10 qubits (top) and 20 qubits (bottom) using the Trotter decomposition.



from external specialized implementations.

In the short term, for quantum simulation we plan to explore the implementation of alternative techniques to state vector simulation, while for QPU support we are testing the framework on multiple real quantum hardware configurations.

## Acknowledgments

We thank the NVIDIA Corporation team for supporting this project. The authors would like to thank Christian Hundt for technical discussions about GPU technology. We thank the Qibo team members for testing the code and providing feedback concerning the results presented in this manuscript. S.C. thanks Sofia Vallecorsa and CERN's QTI for granting access to the ATOS QLM hardware.

## References

- [1] J. Preskill, *Quantum* **2**, 79 (2018).
- [2] F. Arute *et al.*, *Nature* **574**, 505 (2019).
- [3] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, *et al.*, *Science* **370**, 1460 (2020).
- [4] Google Research, *Google AI Quantum* (2017).
- [5] IBM Research, *IBM Quantum Experience* (2016).
- [6] Rigetti, *Rigetti Computing* (2017).
- [7] Intel Corporation, *Intel Quantum Computing* (2017).
- [8] D-Wave Systems, *The Quantum Computing Company* (2011).
- [9] D-Wave Systems, *D-Wave Neal*.
- [10] I. L. Markov, A. Fatima, S. V. Isakov, and S. Boixo, arXiv preprint arXiv:1807.10749 (2018).
- [11] J. Chen and *et al.*, Classical simulation of intermediate-size quantum circuits (2018), arXiv:1805.01450 [quant-ph].
- [12] S. Efthymiou, S. Ramos-Calderer, C. Bravo-Prieto, A. Pérez-Salinas, D. García-Martín, A. Garcia-Saez, J. I. Latorre, and S. Carrazza, *Quantum Science and Technology* **7**, 015018 (2021).
- [13] The Qibo team, *qiboteam/qibo: Qibo*.
- [14] S. Carrazza, S. Efthymiou, M. Lazzarin, and A. Pasquale, in *20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research: AI Decoded - Towards Sustainable, Diverse, Performant and Effective Scientific Computing* (2022) arXiv:2202.07017 [quant-ph].
- [15] The Qibo team (2022).
- [16] S. Carrazza, S. Efthymiou, M. Lazzarin, and A. Pasquale, *qiboteam/qibojit: qibojit*.
- [17] S. K. Lam, A. Pitrou, and S. Seibert, in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* (2015) pp. 1–6.
- [18] R. Okuta, Y. Unno, D. Nishino, S. Hido, and C. Loomis, in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)* (2017).
- [19] G. Aleksandrowicz *et al.*, *Qiskit: An open-source framework for quantum computing* (2019).
- [20] Cirq Developers, *Cirq* (2021), See full list of authors on Github: <https://github.com/quantumlib/Cirq/graphs/contributors>.
- [21] Quantum AI team and collaborators, *qsim* (2020).
- [22] D. S. Steiger, T. Häner, and M. Troyer, *Quantum* **2**, 49 (2018).
- [23] T. Häner, D. S. Steiger, K. Svore, and M. Troyer, *Quantum Science and Technology* **3**, 020501 (2018).
- [24] S. Mandrà, J. Marshall, E. G. Rieffel, and R. Biswas, in *2021 IEEE/ACM Second International Workshop on Quantum Computing Software (QCS)* (2021) pp. 99–109.
- [25] Y. Suzuki, Y. Kawase, Y. Masumura, Y. Hiraga, M. Nakadai, J. Chen, K. M. Nakanishi, K. Mitarai, R. Imai, S. Tamiya, T. Yamamoto, T. Yan, T. Kawakubo, Y. O. Nakagawa, Y. Ibe, Y. Zhang, H. Yamashita, H. Yoshimura, A. Hayashi, and K. Fujii, *Quantum* **5**, 559 (2021).
- [26] A. Kelly, arXiv preprint arXiv:1805.00988 (2018).
- [27] T. Oliphant, *Guide to NumPy* (2006).
- [28] <https://pypi.org/project/qibo>.
- [29] <https://anaconda.org/conda-forge/qibo>.
- [30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems* (2015), software available from tensorflow.org.
- [31] <https://github.com/qiboteam/qibotf>.
- [32] <https://github.com/qiboteam/qibojit>.
- [33] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
- [34] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, *JAX: composable transformations of Python+NumPy programs* (2018).

- [35] G. Guennebaud, B. Jacob, *et al.*, Eigen v3, <http://eigen.tuxfamily.org> (2010).
- [36] K. Rupp, P. Tillet, F. Rudolf, J. Weinbub, A. Morhammer, T. Grasser, A. Jüngel, and S. Selberherr, *SIAM Journal on Scientific Computing* **38**, S412 (2016), <https://doi.org/10.1137/15M1026419>.
- [37] NVIDIA, Thrust (2020).
- [38] NVIDIA, nvcc (2022).
- [39] NVIDIA, cuQuantum SDK (2021).
- [40] S. Eftymiou, M. Lazzarin, S. Carrazza, and A. Pasquale, [qiboteam/qibojit-benchmarks: benchmarks v0.0.1](https://github.com/qiboteam/qibojit-benchmarks) (2022).
- [41] A. W. Cross, L. S. Bishop, J. A. Smolin, and J. M. Gambetta, arXiv e-prints , arXiv:1707.03429 (2017), arXiv:1707.03429 [quant-ph] .
- [42] A. W. Cross, A. Javadi-Abhari, T. Alexander, N. de Beaudrap, L. S. Bishop, S. Heidel, C. A. Ryan, J. Smolin, J. M. Gambetta, and B. R. Johnson, arXiv preprint arXiv:2104.14722 (2021).
- [43] [https://github.com/quantumlib/Cirq/blob/master/cirq-core/cirq/experiments/random\\_quantum\\_circuit\\_generation.py](https://github.com/quantumlib/Cirq/blob/master/cirq-core/cirq/experiments/random_quantum_circuit_generation.py).
- [44] <https://qiskit.org/documentation/stubs/qiskit.circuit.library.QuantumVolume.html>.
- [45] D. Coppersmith, An approximate Fourier transform useful in quantum factoring (2002), arXiv:quant-ph/0201067 [quant-ph] .
- [46] Circuit consisting of alternating layers of parametrized RY rotations and entangling CZ gates.
- [47] I. L. Markov, A. Fatima, S. V. Isakov, and S. Boixo, arXiv preprint arXiv:1807.10749 10.48550/arXiv.1807.10749 (2018).
- [48] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta, *Physical Review A* **100**, 10.1103/physreva.100.032328 (2019).
- [49] E. Bernstein and U. Vazirani, *SIAM Journal on Computing* **26**, 1411 (1997), <https://doi.org/10.1137/S0097539796300921> .
- [50] ATOS, Quantum Learning Machine.
- [51] NVIDIA team, NVIDIA DGX Station.
- [52] Qiskit-aer 0.10.3, qiskit-aer-gpu 0.10.2.
- [53] M. Smelyanskiy, N. P. D. Sawaya, and A. Aspuru-Guzik, qHiPSTER: The quantum high performance software testing environment (2016), arXiv:arXiv:1601.07195 [quant-ph] .
- [54] M. B. *et al.*, TensorFlow Quantum: A software framework for quantum machine learning (2020), arXiv:arXiv:2003.02989 [quant-ph] .
- [55] S. V. I. *et al.*, Simulations of quantum circuits with approximate noise using qsim and Cirq (2021), arXiv:arXiv:2111.02396 [quant-ph] .
- [56] T. Kadowaki and H. Nishimori, *Physical Review E* **58**, pp. 5355–5363 (1998).
- [57] E. Crosson and A. W. Harrow, 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS) 10.1109/focs.2016.81 (2016).
- [58] S. Paeckel and *et al.*, *Annals of Physics* **411**, pp. 167998 (2019).