

Simulation of Quantum Many-Body Dynamics with Tensor Processing Units: Floquet Prethermalization

Alan Morningstar^{1,2,*} Markus Hauru² Jackson Beall² Martin Ganahl² Adam G.M. Lewis²
Vedika Khemani³ and Guifre Vidal²

¹*Department of Physics, Princeton University, Princeton, New Jersey 08544, USA*

²*Sandbox@Alphabet, Mountain View, California 94043, USA*

³*Department of Physics, Stanford University, Stanford, California 94305, USA*



(Received 20 November 2021; accepted 20 April 2022; published 11 May 2022)

Tensor processing units (TPUs) are specialized hardware accelerators developed by Google to support large-scale machine-learning tasks but they can also be leveraged to accelerate and scale other linear-algebra-intensive computations. In this paper, we demonstrate the usage of TPUs for massively parallel classical simulations of quantum many-body dynamics on long time scales. We apply our methods to study the phenomenon of *Floquet prethermalization*, i.e., exponentially slow heating in quantum spin chains subject to high-frequency periodic driving. We simulate the dynamics of $L = 34$ qubits for over 10^5 Floquet periods, corresponding to circuits with 4×10^6 nearest-neighbor two-qubit gates. The circuits simulated have no additional symmetries and represent a pure-state evolution in the full 2^L -dimensional Hilbert space. This is achieved by distributing the computation over 128 TPU cores. On that size TPU cluster, we find speed-ups in wall-clock run time of 230 times and 15 times when compared to reference CPU and single-graphics-processing-unit (GPU) simulations, respectively, for shorter-time 30-qubit simulations that can be handled by all three platforms. We study the computational cost of the simulations, as a function of both the number of qubits and the number of TPU cores used, up to our maximum capacity of $L = 40$ qubits, which requires a “full pod” of 2048 TPU cores with tens of terabytes of memory in total. For these simulations, an eight-TPU-core machine is comparable to a single A100 GPU and thus the full TPU pod is comparable to a machine with hundreds of top-of-the-line GPUs. However, the TPU pod is more energy and cost efficient and readily accessible (via Google Cloud), unlike such large many-GPU configurations. We also study the accumulation of numerical error as a function of circuit depth in very deep circuits. Our work demonstrates that TPUs can offer significant advantages for state-of-the-art simulations of quantum many-body dynamics.

DOI: [10.1103/PRXQuantum.3.020331](https://doi.org/10.1103/PRXQuantum.3.020331)

I. INTRODUCTION

The study of out-of-equilibrium quantum dynamics is an exciting research frontier. Recent years have seen many important developments on foundational questions, ranging from the onset of chaos and thermalization in isolated quantum systems [1,2], to mechanisms for understanding the breakdown of thermalization, such as many-body localization [3–5] or quantum scarring [6,7] and to the discovery of novel nonequilibrium phases of matter [8].

Research in this area involves the study of the *out-of-equilibrium dynamics* of *highly excited* many-body quantum systems, which lies outside the purview of most traditional frameworks and techniques of many-body quantum theory, thereby necessitating the development of a new conceptual and technical toolkit.

Simulation of the dynamics of a general quantum many-body system on a classical computer is prohibitively costly due to the exponential scaling of the dimension of the quantum state space with system size. This reality, along with the fact that quantum-mechanical systems are of both fundamental interest and practical importance, was the original motivation to begin the longstanding investigation and development of computers that operate quantum mechanically [9]. While significant progress has been made in developing quantum simulators across a variety of platforms [10–20], much research in quantum dynamics still relies heavily on classical numerical simulations.

*alanmorningstar@princeton.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

In parallel to the development of quantum hardware, the rise in prominence of computational science and engineering—and in particular, data-intensive machine learning—has led to advances in classical computing, namely, the integration of specialized *hardware accelerators* into supercomputing systems, to meet these growing needs. The most common example of a hardware accelerator is the graphics processing unit (GPU), which is a widely available type of general-purpose processor capable of performing many arithmetic operations in parallel. Another is Google’s development of the tensor processing unit (TPU) and the JAX and TENSORFLOW software libraries, as a means of accelerating and scaling the computations needed to train and evaluate large neural networks with accessible tools [21–25]. Unlike GPUs, TPUs are not general-purpose processors; they are specialized application-specific integrated circuits (ASICs) for deep-learning workloads that are heavy on matrix multiplication and vector operations.

Given these advances in classical-computing capabilities, it is natural to explore the application of TPUs to other challenging computational tasks [26–32]. Since TPUs are specifically designed for large-scale linear algebra, the classical simulation of quantum systems is a particularly suitable application for them. More specifically, the strength of using TPUs for simulation of the dynamics of quantum systems comes from two contributions: (1) TPU cores have hardwired matrix multiplication units that allow them to contract arrays (with some constraints on their shapes) faster and with higher energy efficiency than on contemporary CPUs and GPUs [22]; and (2) arrays can be distributed over thousands of directly linked TPU cores, and operated on easily within the single-program-multiple-data (SPMD) programming paradigm [33]. Data transfer between the cores is carried out via fast intercore interconnects (ICIs), and this makes multi-TPU configurations extraordinarily performant at scale. These ingredients naturally allow for the accelerated application of local operators to wave functions that are distributed over many TPU cores and other useful operations. Similar efforts have been made to leverage GPUs to accelerate the classical simulation of quantum circuits [34–46], also resulting in large gains over CPU simulations. TPUs exist because they are more performant and more energy efficient (and thus cost effective) than GPUs for many large-scale machine-learning tasks. It is therefore promising to begin to explore their application to quantum simulation.

In this work, which is the first of a set of related papers [47–54], we demonstrate the application of TPUs to accelerate and scale the classical simulation of quantum many-body dynamics. Specifically, we study dynamics in Floquet, i.e., time-periodic, quantum circuits as a technically illustrative and physically motivated application. Periodic driving has become a central workhorse of

quantum simulation, both as a tool for the “Floquet engineering” of interesting band structures and Hamiltonians [55] and as a means of realizing novel intrinsically nonequilibrium phases of matter such as time crystals [8,56–62]. Indeed, the observation of such novel phases of driven systems has recently been identified [63] and experimentally demonstrated [64] as an exceptionally well-suited near-term application of digital quantum simulators for many-body physics. Thus, pushing the boundaries of the classical simulation of Floquet systems is not only motivated by a fundamental theoretical interest in quantum dynamics but also by a need to benchmark quantum simulators and to raise the bar for evaluating quantum advantage.

We focus on the phenomenon of *Floquet prethermalization* for our demonstration. The periodic driving of an interacting many-body system generically leads to thermalization to a featureless infinite-temperature state [65,66]; the time dependence of the Hamiltonian breaks the usual conservation of energy, and infinite temperature is the appropriate maximum-entropy equilibrium in the absence of any additional conservation laws. The heating to infinite temperature can be averted in the presence of disorder induced many-body localization [67,68]. Alternatively, at sufficiently high driving frequencies ω , the heating can be slowed down to a rate that is exponentially small in ω , corresponding to a heating time $t_h \sim e^{\omega/\omega_0}$ [69–77]. This high-frequency prethermal regime with exponentially slow heating is challenging to observe in finite-size numerics [75,78]: one needs driving frequencies that are large compared to the local energy scales of approximately J in the problem but small compared to the extensive many-body bandwidth of approximately JL [in one-dimensional (1D) systems of length L], which does not present a large dynamic range in systems of size amenable to numerical simulation. Furthermore, exponentially long time evolutions are needed and the full 2^L -dimensional wave function must be used (rather than a truncated tensor-network representation) because of the extensive amount of entanglement generated by the dynamics. For these reasons, the Floquet prethermalization problem is a well-motivated choice for demonstrating the technical advantages of using TPUs to accelerate and scale up classical simulations of the quantum dynamics of Floquet circuits.

The rest of this work is structured as follows. In Sec. II, we provide an overview of the conceptual principles of quantum simulation algorithms for TPUs. Then, in Sec. III, we describe the model that we use and provide a brief review of Floquet prethermalization. Following this, in Sec. IV we present the physics-centric results of our numerical simulations. Then, in Sec. V, we move on to a study of the computational cost of, and accumulation of numerical errors in, the simulations done in this work. Finally, we summarize our work and discuss

potential future applications of TPU simulations to quantum dynamics research in Sec. VI.

II. PRINCIPLES OF TPU ALGORITHMS

In later sections, we study the long-time dynamics of periodically driven quantum spin chains using TPU simulations, so here we first provide a high-level overview of TPUs and the principles of using TPUs for these simulations.

The computing platform we work with is a cluster of 2^{N_g} connected TPU v3 cores, with $N_g \in [3, 11]$ (number of cores $\in [8, 2048]$). The connectivity of the cores is toroidal and each core has 16 GB of high-bandwidth memory. A full 2048-core v3 “pod” therefore has 32 TB of memory. Since the simulation of a larger number of qubits L requires exponentially more memory (in L), we adjust N_g based on the total memory needed and/or the total number of TPUs we want working simultaneously. In general, use of the highest fraction of memory possible (use of the smallest cluster possible) results in the highest efficiency. On TPUs, matrix multiplication occurs natively in half precision but single and double precision can be emulated and all of our simulations in this work are done in single precision.

For concreteness, we consider the example of $N_g = 3$ (eight TPU cores). Thus the cores can be labeled by 3-bit binary numbers $c_1c_2c_3 \in \{000, 001, \dots, 111\}$. The central object relevant to our simulations is the full wave function of L qubits, which is distributed over all of the cores in the following way. The wave function is the array of amplitudes $\psi_{b_1 \dots b_L} = \langle b_1 \dots b_L | \psi \rangle$, where $b_1 \dots b_L$ is a bit string that labels the computational basis states. The subset of these amplitudes that we store on a particular core is then simply the subset for which the first three qubits, termed the *global qubits*, match the 3-bit label ($c_1c_2c_3 = b_1b_2b_3$) of the core, i.e., each core stores its own branch of the wave function where the global qubits are in a definite computational state. For example, we store the branch of the wave function where the three global qubits are in state $|101\rangle$ on the TPU core labeled by 101 (the sixth core), which corresponds to the array $\psi_{101b_4 \dots b_L}$ existing locally on that core. In this way, each core only stores an array of size 2^{L-3} (or, more generally, 2^{L-N_g}) locally. For this reason, we refer to the $L - N_g$ final qubits as *local qubits*. Since 2^{29} single-precision complex floats take up 4.3 GB of memory, the 16 GB of memory per core is enough to store three separate wave functions with $L - N_g = 29$ local qubits. The maximum total number of qubits we can do simulations with is therefore $L = 40$, because the largest TPU v3 cluster corresponds to $N_g = 11$ global qubits (2048 cores). See Fig. 1 for a depiction of the distributed wave function.

We use the JAX software library [23] to write SPMD-style code to execute our simulations on the TPU clusters. Recall that SPMD (single program multiple data) means that the same set of instructions is run on each TPU core

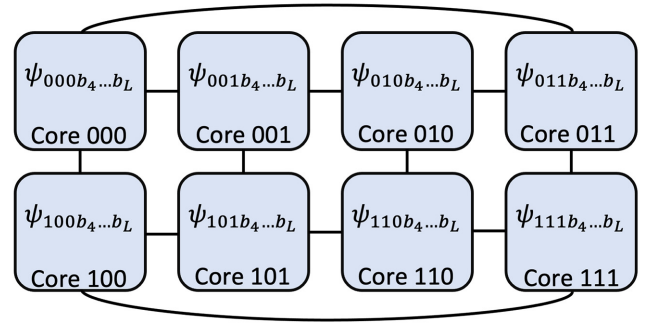


FIG. 1. The distribution of the many-body wave function over TPU cores. There are eight TPU cores depicted ($N_g = 3$), labeled by the binary strings 000, 001, $\dots$, 111. On each core, a branch of the full wave function is stored, where the state of the global qubits is fixed according to the TPU core. Lines between TPU cores indicate the toroidal pattern of intercore connectivity.

but each core hosts a different set of data to input into those instructions. Notably, the “single program” can include intercore communication as long as it is symmetric under permutation of the cores. This is the type of parallelism that is available in JAX (see the documentation for available functionality, tutorials, etc. [79]) and it can be used to do all of the necessary operations, e.g., prepare initial states, apply local operators, compute inner products, etc. TPUs have a native matrix multiplication unit and data storage pattern that works optimally for arrays for which the shape satisfies certain criteria: the basic rules of thumb are that the final two dimensions of an array must have sizes that are multiples of 8 and 128, respectively, and that contractions must occur on indices for which the size is a multiple of 128. Note that $8 = 2^3$ (three qubits) and $128 = 2^7$ (seven qubits), so these are practical constraints about how many qubits must be grouped together in a wave-function array and during operations such as applying local operators to wave functions. But here we ignore that complication for the sake of clarity. Below, we illustrate the principles of TPU algorithms for quantum simulation via two simple examples.

Let us first consider an inner product between two distributed wave functions ψ and ϕ . This can be written as (recall that $N_g = 3$ here)

$$\langle \psi | \phi \rangle = \sum_{c_1c_2c_3} \sum_{b_4 \dots b_L} \psi_{c_1c_2c_3b_4 \dots b_L}^* \phi_{c_1c_2c_3b_4 \dots b_L}. \quad (1)$$

When evaluating this, the inner sum over local qubits occurs first locally on each core, i.e., an inner product (which is a “single program”) of the branches of the wave functions on each TPU core (the “multiple data”) is evaluated resulting in a single number on each core that can be indexed by the label of that core $c_1c_2c_3$ (another “multiple-data” form). The outer sum over the state of the global

qubits in Eq. (1) is then evaluated simply by collecting all of these numbers from all of the individual cores and summing them. Each core does this identically (another “single program” that runs the same on each core). This can be done using the JAX PSUM operation [79], which is one of the few operations that encodes SPMD-type intercore communication.

As a second and final example, we consider applying a local two-qubit gate. Again, in practice, to get optimal performance from the TPUs we must fuse gates into operators acting on certain groups of seven or more qubits but we ignore that detail here for conceptual clarity. In the case that the gate acts only on local qubits, the action of applying the gate (single program) can be performed independently on each branch of the wave function (multiple data) that is local to each core. For example, applying a gate to qubits 7 and 8 results in the updated local (to each core labeled by $c_1 c_2 c_3$) branches being

$$(U\psi)_{c_1 c_2 c_3 b_4 \dots b_L} = \sum_{\alpha\beta} U_{b_7 b_8 \alpha\beta} \psi_{c_1 c_2 c_3 b_4 \dots \alpha\beta \dots b_L}. \quad (2)$$

Note that the same sum over α and β can occur independently on each core with the data from the gate and the locally stored branch of the wave function. The result is an updated wave function that has the same distribution pattern as the original one. In the case where the gate must be applied to some of the global qubits, we use the JAX PSWAPAXES operation [79] (which allows intercore communication) to effectively swap the global qubits that label the TPU cores with the local qubits that do not. This reduces this case to the previous one of applying a gate to local qubits.

Much of the effort in writing these algorithms goes into complying with the aforementioned constraints implied by requiring optimal performance from the TPU matrix multiplication unit. More discussion of this can be found in subsequent works [47–49] but what is provided here is all that is needed for the simulations in the following sections.

III. MODEL AND REVIEW

In this section, we provide both a concrete model that we work with for the rest of the paper and a review of the relevant physics of Floquet prethermalization.

A. Floquet-circuit model

Our model system consists of L spin-1/2 (qubit) degrees of freedom arranged in a 1D array with open boundaries. The stroboscopic time evolution of the state of the system is generated by repeatedly applying a Floquet unitary U_F , which is the unitary time-evolution operator over one period, i.e.,

$$|\psi(nT)\rangle = U_F^n |\psi(0)\rangle, \quad (3)$$

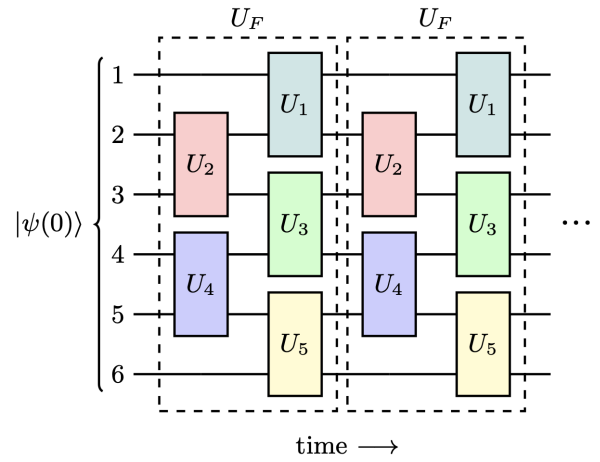


FIG. 2. The time-periodic quantum circuit generating the dynamics of our model system, illustrated on $L = 6$ qubits. The qubits are numbered top to bottom and time runs left to right. The system is prepared in the state $|\psi(0)\rangle$ described later in the main text [Eq. (9)]. Each gate U_k takes time $T/2$ to be executed, during which its two qubits are evolving under an associated two-qubit Hamiltonian $2H_k$ [Eq. (5)]. All gates in the same layer are applied simultaneously, so the two-layer repeating unit (the Floquet unitary U_F) lasts for time T .

where T is the physical time duration of U_F (the Floquet period) and $n \in \mathbb{N}$. The frequency is $\omega = 2\pi/T$.

We choose the Floquet unitary to be a two-layer quantum circuit of close-packed two-qubit unitary gates U_k acting on qubits k and $k+1$ (see Fig. 2). The gates alternate between acting on even and odd bonds in each layer. For even L , this takes the form

$$U_F = \prod_{k=1}^{L/2} U_{2k-1} \prod_{k=2}^{L/2} U_{2k-2}. \quad (4)$$

Each gate U_k enacts Hamiltonian time evolution on its two qubits under the Hamiltonian $2H_k$ for time $T/2$,

$$U_k = \exp(-iH_k T), \quad (5)$$

where we work in units in which $\hbar = 1$.

Since our circuit is time periodic (Floquet), with period T , and the gates are formulated in terms of two-qubit Hamiltonians H_k , we can also view the system as evolving under the following time-dependent local Hamiltonian $H(t)$, with $H(t) = H(t+T)$:

$$H(t) = \begin{cases} 2H_e, & n \leq \frac{t}{T} < n + \frac{1}{2}, \\ 2H_o, & n + \frac{1}{2} \leq \frac{t}{T} < n + 1, \end{cases} \quad (6)$$

$$H_e \equiv \sum_{k=2}^{L/2} H_{2k-2} \quad H_o \equiv \sum_{k=1}^{L/2} H_{2k-1}.$$

During the first half of each period, $H(t) = 2H_e$, and during the second half, $H(t) = 2H_o$, where H_e and H_o contain the even- and odd-bond terms of the time-averaged Hamiltonian operator:

$$\tilde{H} = \sum_{k=1}^{L-1} H_k. \quad (7)$$

The $L - 1$ independent two-qubit Hamiltonians H_k are each chosen by first sampling a random Hermitian 4×4 matrix from the Gaussian unitary ensemble (GUE) and then shifting and scaling that matrix to be traceless with Frobenius norm $J = \sqrt{2}$, which sets the microscopic energy scale J and the units of energy and time that we use. Thus our model is actually an ensemble of random Floquet circuits that has a single tuning parameter T (or ω). Random Floquet circuits have been used as models of both quantum chaos [80] and many-body localization (MBL) [81] and we emphasize that the randomness in our model does not induce any form of MBL. In order to avoid unnecessary ensemble averaging, we only use a single sample from the ensemble in this paper, because a single sample is already a good representative of a generic interacting driven system. This means that we sample a long list of two-qubit Hamiltonians once and then use the first $L - 1$ terms from that list for an L -qubit system. This is a desirable property when assessing convergence in L , while avoiding the need to average over samples [82].

We choose to use this random Floquet-circuit model because circuits are computationally convenient to simulate compared to continuous-time Hamiltonian evolution that cannot be written as a circuit, where an artificial discretization scheme is needed. Circuits are also the natural mode of operation of digital quantum simulators. Note that this circuit model can also be viewed as a Trotterization [83–86] of evolution under the time-averaged (and time-independent) Hamiltonian $\tilde{H}$. However, here we study the full dynamics of U_F even outside of the regime of very small T (large ω) where the Trotterization approximates evolution due to $\tilde{H}$. Our choice of random H_k is so as to study a minimal model with spatial locality and periodic driving but no other physical features (such as symmetries, topology, etc.), in order to capture general universal features of high-frequency Floquet prethermalization.

B. Floquet prethermalization

As mentioned earlier, for any (finite) drive frequency ω , if L is large enough, a generic driven many-body system thermalizes to a featureless “infinite-temperature” thermal state [65,66]. This means that the reduced density matrix of any small subsystem will be proportional to the identity. However, if $\omega \gg J$, i.e., the drive frequency is much larger than the local energy scale of the system, then the transitions that allow the system to exchange energy with

the drive will involve the rearrangement of many local degrees of freedom; an estimate for how many is $m \sim \omega/J$ because each local transition is on an energy scale J and the drive couples to energy ω . In that case, the effective rates for these processes will be exponentially suppressed in m because there are approximately m off-resonant virtual states involved and thus the thermalization rate (time) will be exponentially small (large) in ω/J [69–72].

More precisely, the Floquet unitary of our model is given by $U_F = \exp(-iH_o T) \exp(-iH_e T)$ but one can always define a (nonunique) Floquet Hamiltonian H_F such that $U_F \equiv \exp(-iH_F T)$, i.e., the stroboscopic dynamics are identical to evolution generated by a time-independent Hamiltonian H_F . Since, in the large- L limit, the system always thermalizes to an infinite-temperature equilibrium, the simultaneous eigenstates of U_F and H_F are all infinite-temperature states and thus H_F is a highly nonlocal unphysical Hamiltonian. However, if $\omega \gg J$, then there is a *local* Hamiltonian $\tilde{H}_F$ such that $\|U_F - \exp(-i\tilde{H}_F T)\|$ is exponentially small in ω/J and the dynamics of the system are well approximated by the time-independent Hamiltonian $\tilde{H}_F$ out to times that are exponentially long in ω/J . This is one of the main results of several mathematical works on this topic [70–73].

In order to approximate $\tilde{H}_F$, one can employ a high-frequency Baker-Campbell Hausdorff (BCH) expansion to obtain

$$\tilde{H}_F = H_o + H_e - \frac{iT}{2}[H_o, H_e] + O(T^2 H^3), \quad (8)$$

where H_o and H_e are the odd- and even- k terms in Eq. (6). This is an asymptotic series that eventually diverges but it appears to be converging up to a finite order $m_{\text{opt}} \sim \omega/J$, which defines an effective Hamiltonian, $\tilde{H}_F$, obtained from truncating the series. The zeroth-order term in $\tilde{H}_F$ is the time-averaged Hamiltonian $\tilde{H} = H_o + H_e$, which provides a useful approximate notion of energy in our system, which we use throughout the rest of this work. The exponentially long thermalization time is detected by monitoring the evolution of $\langle \tilde{H} \rangle$ over time, which is $\langle \tilde{H}_F \rangle$ up to relative corrections of $O(JT)$ [absolute corrections of $O(JTH)$]. Note that once the frequency becomes comparable to the (extensive) many-body bandwidth, a high-frequency expansion can be convergent, in which case one obtains a sensible local $\tilde{H}_F$ governing the dynamics for all time; however, this is not expected to happen for finite frequencies for systems in the large- L limit.

This discussion illustrates why it is numerically challenging to observe prethermalization: simulations need to reach long enough times to study the exponentially slow processes involved and large enough system sizes to avoid finite-size effects. Thus this problem provides a suitable testing ground for the applications of TPUs to quantum many-body dynamics.

IV. SIMULATIONS AND RESULTS

We now present numerical evidence for the predicted exponential-in-frequency thermalization times characteristic of Floquet prethermalization. For all of the simulations in this paper, we initialize the system in the ground state of H_0 . This is a product state of local two-qubit ground states (gs):

$$|\psi(0)\rangle = \bigotimes_{k \text{ odd}} |\text{gs}(H_k)\rangle. \quad (9)$$

Since the H_k are uncorrelated and traceless in our model, the energy of this initial state with respect to the full time-averaged Hamiltonian is close to the sum of local ground-state energies of the odd terms. We choose this initial state because it is simple and is at a finite energy density corresponding to starting with a finite temperature initial state with respect to $\bar{H}$.

As accessible indicators of the slow approach to thermal equilibrium and the existence of a quasiconserved $\bar{H}_F$, we track both the mean and standard deviation of energy with respect to $\bar{H}$:

$$E(t) = \langle \psi(t) | \bar{H} | \psi(t) \rangle \quad (10)$$

$$\sigma_E(t) = \sqrt{\langle \psi(t) | \bar{H}^2 | \psi(t) \rangle - E(t)^2}. \quad (11)$$

In Fig. 3, we show E and σ_E as a function of time t out to $t = 10^5$ for a system of $L = 34$ qubits and for several evenly spaced values of $\omega \in [5, 8]$. By eye, we can see that as ω increases uniformly, the thermalization time scale increases uniformly on a logarithmic scale, implying that the thermalization rate scales like $\Gamma(\omega) = \exp(a\omega + b)$ within the observable window of frequencies.

In order to make the frequency dependence quantitative, we extract thermalization rates from the curves in the top panel of Fig. 3: for each curve (frequency), we determine a decay rate via $\Gamma = 1/(t_2 - t_1)$, where t_k is defined as the time that satisfies $E(t_k) = E(0) \exp(-k)$. Later time intervals would work too, e.g., $\Gamma = 1/(t_3 - t_2)$, but we avoid the earliest such interval ($t \in [t_0, t_1]$) because it contains the rapid early time dynamics of $E(t)$ before it settles into a slowly decaying “prethermal plateau.” In Fig. 4, we plot these thermalization rates and show that, indeed, they satisfy an exponential scaling $\Gamma(\omega) = \exp(a\omega + b)$ with $a = -2.48(7)$ and $b = 10.9(5)$.

Our simulations are in agreement with the theoretical expectation that Γ is exponentially suppressed in ω at large enough frequencies and with previous experimental [76,77] and numerical [75,78,87] work on other systems. We note here that prior numerical studies of this effect have been performed with system sizes of $L \leq 24$ and times $t \leq 10^4$ (in microscopic units), so the use of TPUs allows

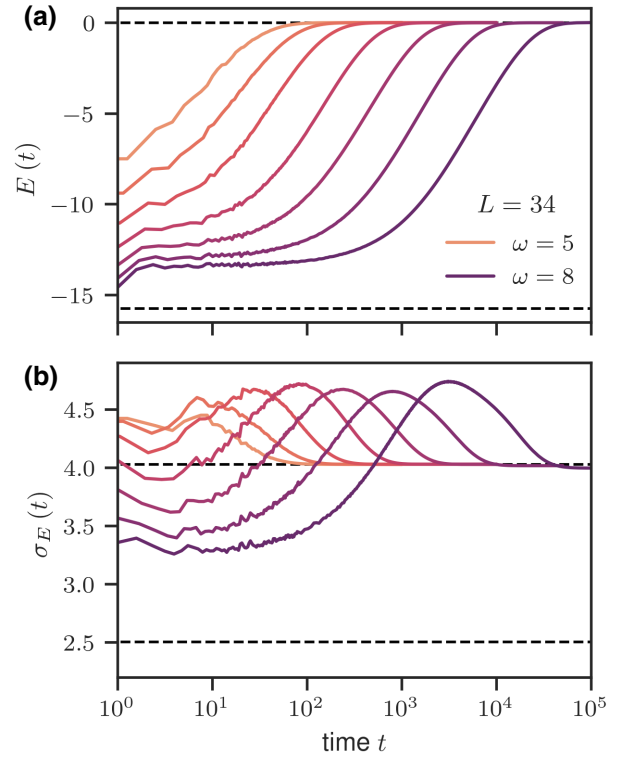


FIG. 3. The mean (a) and standard deviation (b) of the energy, with respect to $\bar{H}$, versus time. Light to dark (left to right) curves correspond to $\omega \in [5, 8]$ in steps of $1/2$. All data are for $L = 34$ qubits. The dashed black lines indicate both initial values and infinite-temperature equilibrium values (where the density matrix is $\rho \propto \mathbb{I}$) for comparison. Data are collected after every Floquet period up to $t = 10^2$ and then after every tenth period from then on. Note that the conversion from time t to number of gates g is $g = (L - 1)t/T$, where $T = 2\pi/\omega$, so the rightmost curve corresponds to a total of 4.2×10^6 two-qubit gates.

us to significantly extend the system sizes and times studied by about ten qubits and a factor of 10 in time. This is a rough estimate and not a direct comparison to previous work due to differences in the specific model being simulated and the computational resources consumed (our Floquet-circuit model is indeed easier to simulate than noncircuit models, which is one of the reasons for using it). A more direct comparison of computational resources required for simulation of our chosen model on CPUs versus GPUs versus TPUs is provided in the following section in order to put our work into context.

In Fig. 5 we demonstrate sufficient convergence in the number of qubits L to extract large- L decay rates. We fix the frequency of the drive to $\omega = 8$, the largest frequency that we consider, and plot $E(t)/E(0)$ over time for system sizes $L \in [22, 34]$. The convergence of these curves at the larger values of L ensures that the thermalization rates shown in Fig. 4, which are for systems with $L = 34$, are converged to the $L \rightarrow \infty$ result. In Sec. V, we demonstrate

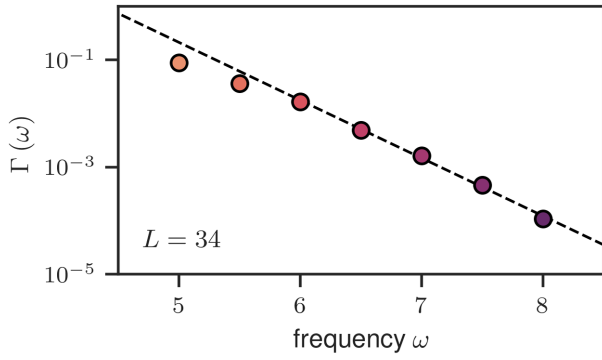


FIG. 4. The thermalization rate versus the frequency. The marker colors match the corresponding curves in Fig. 3 and are a redundant encoding of ω , indicated on the horizontal axis. The dashed line is a fit excluding the two leftmost points and results in $\Gamma = \exp(a\omega + b)$ with $a = -2.48(7)$ and $b = 10.9(5)$. The thermalization rates are extracted as described in the main text.

simulations of up to 38 qubits for much shorter times, in order to study the computational cost of these simulations.

Some clarifying comments on our presented data are in order. First, the initial energy $E(0)$ is indeed independent of ω , even though it may not appear that way in Fig. 3 due to the logarithmic horizontal scale. This is because all simulations start in the same state [Eq. (9)] and we are using only the zeroth-order effective Hamiltonian $\tilde{H}_F^{(0)} = \bar{H}$ to compute the energy. The fact that $E(t)$ does vary significantly with ω but slowly with t at early times is a reflection of the fact that $\bar{H}$ is not a very good approximation to the effective Hamiltonian $\tilde{H}_F$, even in the regime in which the heating time is exponential in ω . Another indicator that $\bar{H}$ is quite different from the optimal effective Hamiltonian is that σ_E takes values above the infinite-temperature equilibrium value (bottom panel of Fig. 3). In general, a

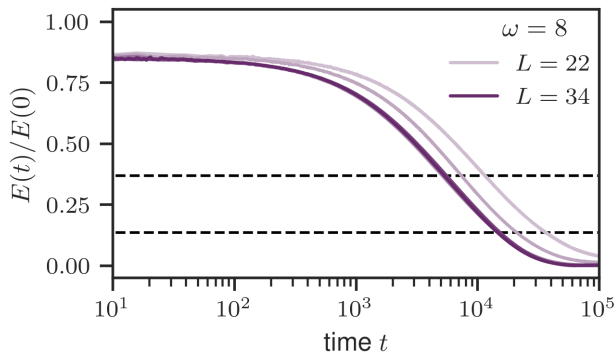


FIG. 5. The normalized energy versus time. Light to dark (right to left) curves correspond to $L \in [22, 34]$ in steps of 2. All data are for $\omega = 8$, the largest frequency we consider in Fig. 3. The dashed horizontal lines are at the thresholds e^{-1} and e^{-2} , which are used to extract the thermalization rate Γ , as described in the main text.

thermal state of a Hamiltonian H has a larger energy variance at infinite temperature than at any finite temperature. The fact that we see larger values of σ_E is fully consistent with the system being in a quasistatic (with slowly evolving temperature) thermal state of an ideal prethermal Hamiltonian $\tilde{H}_F$ that differs from $\bar{H}$ by corrections that are suppressed in ω , as expected. Finally, the rightmost curve in the bottom panel of Fig. 3, corresponding to $\omega = 8$, converges to a value that is visibly slightly different from the infinite-temperature value. This is because at $\omega = 8$ the system is just starting to enter the finite-size regime, for $L = 34$, where the system has difficulty exchanging energy with such a high-frequency drive. It is interesting to note that this effect is present for σ_E but not for E and thus higher moments of the energy distribution seem to be more sensitive to any residual nonthermalizing dynamics.

V. COMPUTATIONAL COST AND NUMERICAL ACCURACY

The workhorse of our simulation method is the application of the two-layer Floquet unitary circuit to the wave function, $|\psi(nT)\rangle = U_F|\psi([n-1]T)\rangle$, using TPUs to accelerate and scale the computation. Application of the Hamiltonian is also crucial, for computing E and σ_E , but relies on the same technology. We therefore provide a summary of the wall-clock time it takes to evolve an L -qubit wave function by one cycle of the circuit in Fig. 6. The timing is done by simulating 100 successive periods and dividing the total execution time by 100. Since we are not simulating out to long times, as in Sec. IV, we provide results for simulations of up to $L = 38$ qubits done on clusters with up to 512 TPU cores and extrapolate to our maximum capacity of $L = 40$ on 2048 TPU cores. We also perform reference CPU and GPU simulations on up to 30 qubits. Of course, there is no single correct metric for comparing CPU to GPU to TPU simulations but we believe that our reference CPU and GPU simulations are more performant than what most researchers in the quantum dynamics and thermalization community use [78,90,91] and our purpose here is to demonstrate the application of TPUs in that setting.

In Fig. 6 we see that, for the same number of qubits, simulations on an eight-TPU-core machine are about 25 times faster than the reference simulations using the QSIMCIRQ package on a standard eight-CPU-core workstation (Intel Xeon @ 2.3 GHz with 64 GB of RAM) and faster than or comparable to the QSIMCIRQ simulations performed on one A100 GPU. For example, for $L = 30$, our reference CPU (GPU) calculation takes 6.3 s (0.41 s) per cycle, whereas the eight-TPU-core machine takes 0.25 s. A significant further speed-up is shown by distributing the computation over 128 TPU cores, which results in a wall-clock time that is 230 times and 15 times shorter than the reference CPU and GPU calculations, respectively.

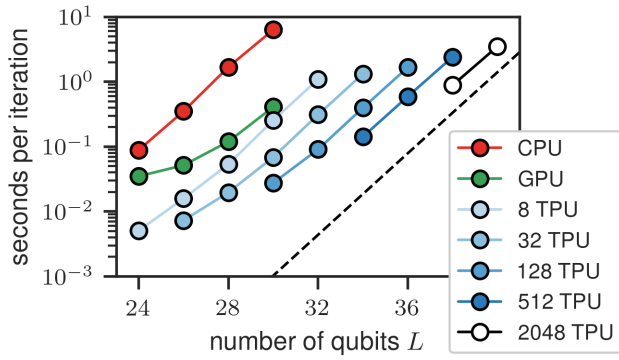


FIG. 6. The wall-clock run time for applying the Floquet unitary circuit (two layers of gates) to an L -qubit wave function. The red data at the top left are reference values obtained by simulating the circuit on a standard eight-CPU-core workstation using QSIMCIRQ [88,89]. The green data are QSIMCIRQ simulations run on one NVIDIA A100 GPU. The blue data, from left to right, correspond to eight, 32, 128, and 512-TPU-core clusters. The slope of the dashed line indicates the expected asymptotic (large L) scaling of the computational cost, $\propto L^{2.2}$. The unfilled markers at the top right are extrapolated from the top two points in each series and represent our maximum capacity simulations on 2048 TPU cores. All simulations are performed with single-precision arithmetic. All TPU calculations in this paper are performed on the third generation of TPUs: “TPU v3.”

Beyond this speed-up, the easy distribution of the computation over many TPUs (see Sec. II) allows for a higher total memory (up to 32 TB) and thus larger system sizes (up to $L = 40$). As shown in Fig. 6, the cost of distributing the calculation over many TPUs is quite tolerable and therefore our method is highly scalable. For example, the $L = 32$ simulation is done on eight, 32, and 128-TPU-core clusters and the speed-up factors in going from eight to 32, and 32 to 128 TPU cores are 3.5 and 3.4, compared to the factor of 4 we would expect if inter-TPU communication was negligible. Another way to see this feature is by looking at the top point in each series of TPU data. By scaling up the number of TPU cores by a factor of 2^6 , we are able to scale the computation from 32 to 38 qubits (the Hilbert space grows by a factor of 2^6 here) while only increasing the wall-clock time by a factor of 2.2. In the ideal case where inter-TPU communication is negligible, this factor would be approximately $38/32 \approx 1.2$, because the total number of gates increases proportionally with the number of qubits.

Since, for these simulations, an eight-TPU-core machine is comparable to an A100 GPU, we expect that a full TPU pod of 2048 cores is comparable to a configuration of hundreds of GPUs (currently, the largest multi-GPU configuration available on Google Cloud has 16 A100 GPUs). Thus large TPU clusters seem to have strengths of being more practically available and more energy and cost efficient than comparable GPU-based machines. We

also note that the latest TPU v4 pod has twice the matrix-multiplication throughput per chip, and 4 times as many chips (connected via state-of-the-art interconnects), as the v3 used in this work. Thus TPUs are a state-of-the-art platform for the kind of quantum simulations we study in this work.

Beyond the cost of TPU simulations, we must also verify their accuracy, especially for the kind of long-time simulations presented in Sec. IV, because TPUs are designed for fast low-precision matrix multiplication, among other things [21]. While double-precision (64-bit) arithmetic can be emulated at a cost (about 11 times slower than single precision), the simulations presented in this work are all performed in single precision (32-bit). Because of this, the numerical errors are larger than in double-precision simulations, which are the standard in quantum dynamics research. While the required precision is task dependent and there are quantum simulations that require higher precision than single or even double, single-precision arithmetic is sufficient for many calculations and comes with the advantage of being less costly. In order to study the accumulation of error and validate our earlier results, we perform an “echo” procedure. We first time evolve the usual initial state $|\psi(0)\rangle$ [Eq. (9)] out to time $t = t_f$ and then reverse the time evolution back to time $t = 0$, using the time-reversed Floquet unitary $U_F^\dagger$, in order to obtain $|\psi'(0)\rangle$. Since the forward and backward evolutions are carried out with a finite numerical precision, $|\psi(0)\rangle$ and $|\psi'(0)\rangle$ are not exactly the same. We quantify the difference via the “overlap error,” $|1 - \langle\psi(0)|\psi'(0)\rangle|$. We check on smaller-scale CPU simulations that this is a good way to estimate $|1 - \langle\psi(2t_f)|\psi'(2t_f)\rangle|$, where here the reference state $|\psi(2t_f)\rangle$ is given by double-precision simulation and $|\psi'(2t_f)\rangle$ by single precision. Since the overlap (fidelity) is a good measure of the proximity of two wave functions and the wave function contains all of the information about a quantum system, this is a robust way of tracking the accumulation of numerical error over time. In Fig. 7, we show the overlap error resulting from this echo procedure as a function of the total simulated time, $2t_f$, at $L = 34$ and $\omega = 8$. From Fig. 7, we see that, indeed, our simulations are accurate up to an overlap error of about 10^{-7} at early times (as expected with 32-bit precision) and at the latest time, $t = 10^5$, the overlap error reaches 10^{-3} . We also find that the *relative* error in $E(0)$ and $\sigma(0)$, when comparing $|\psi'(0)\rangle$ to $|\psi(0)\rangle$, is comparable to the overlap error for these simulations. This serves to validate our earlier data and confirm the applicability of single-precision TPU simulations for studying Floquet prethermalization and other similar problems. We note that some problems do require more than single-precision arithmetic and in that case TPUs may offer an advantage of scalability even though emulating double precision is costly, or TPUs may not be appropriate in their current form for those problems.

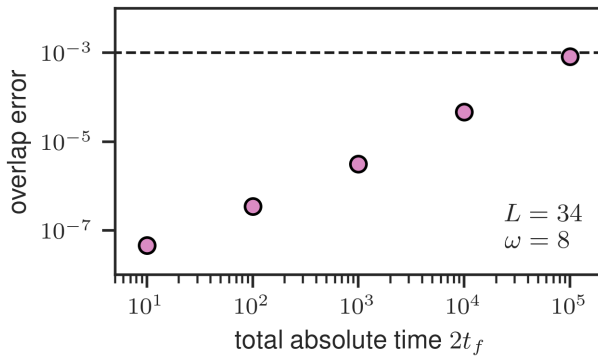


FIG. 7. The accumulation of error over time. The error on the wave function is quantified via the overlap of initial and final states in an “echo” procedure, where the wave function is evolved out to time t_f and then the evolution is reversed using the time-reversed Floquet unitary $U_F^\dagger$ in order to get back to time $t = 0$. The quantity we show is $|1 - \langle \psi(0) | \psi'(0) \rangle|$, where $|\psi'(0)\rangle$ is the state obtained from evolving $|\psi(0)\rangle$ out to time t_f and back. All simulations are run at $\omega = 8$ and $L = 34$. The dashed black line is at the 0.1% error level, for reference.

VI. SUMMARY AND DISCUSSION

The main goal of this work is to demonstrate a new approach to the classical simulation of quantum many-body dynamics that leverages Google’s TPUs and associated software, to accelerate and scale the computation.

We first give an overview of the principles of TPU algorithms for quantum simulation and then briefly review the phenomenon of Floquet prethermalization and exponentially slow heating in periodically driven quantum spin chains, and propose a minimal Floquet-circuit model that exemplifies this phenomenon. We apply our methods to demonstrate this exponential scaling over longer times and on larger systems than have previously been accessible: In this study, we compute nontrivial dynamics, accurately, on long time scales (over 10^5 Floquet cycles) for systems of 34 qubits using a 128-TPU-core cluster and show that by distributing the computation over 2048 TPU cores, systems with up to 40 qubits are accessible. We also demonstrate a 230-times (15-times) speed-up in wall-clock run time when comparing a distributed TPU simulation to a reference simulation performed on a CPU (GPU). This is intended to provide familiar points of reference and not as proof that TPUs are the most performant hardware. For that, an extensive comparison to multi-GPU configurations would be necessary. However, as mentioned earlier, we expect that a full TPU pod is comparable to a machine with hundreds of top-of-the-line GPUs and such configurations are not readily available. We emphasize that in this problem, entanglement quickly becomes nearly maximal and so keeping track of all 2^L amplitudes of the wave function is necessary.

The problem of studying Floquet prethermalization is a good example of the challenging computational tasks that arise in quantum many-body dynamics research and it provides a suitable testing ground for demonstrating the application of TPUs in this field. TPU simulations widen the finite-size and finite-time windows that numerical simulations can access, thereby allowing for higher-quality data and more thorough studies of dynamical phenomena and phase transitions in quantum many-body systems. In the future, TPU simulations of the full wave function, as we do here, could be useful for many topical studies of quantum dynamics, wherein our understanding of many fundamental issues is limited by finite system size and time studies [92–100].

A natural extension of this work is to employ TPU methods for performing state-of-the-art classical simulations for the purposes of benchmarking simulations done on quantum hardware and in assessing quantum advantage. For this, entanglement may not be near maximal as it is for the Floquet prethermalization problem, so it would be promising to also develop TPU-accelerated versions of tensor-network methods. The development of tensor-network methods may also benefit from having larger scale exact simulations, like the ones performed in this work, against which to compare.

ACKNOWLEDGMENTS

This research was supported with Cloud TPUs from Google’s TPU Research Cloud (TRC). Sandbox is a team within the Alphabet family of companies, which includes Google, Verily, Waymo, X, and others. The GPU simulations presented in this paper were performed on computational resources managed and supported by Princeton Research Computing. G.V. is a Canadian Institute for Advanced Research (CIFAR) fellow in the Quantum Information Science Program and a Distinguished Visiting Research Chair at Perimeter Institute. Research at Perimeter Institute is supported by the Government of Canada through the Department of Innovation, Science, and Economic Development and by the Province of Ontario through the Ministry of Research, Innovation, and Science. V.K. acknowledges support from the Sloan Foundation through a Sloan Research Fellowship, the Packard Foundation through a Packard Fellowship, and by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Early Career Award No. DE-SC0021111.

-
- [1] A. Polkovnikov, K. Sengupta, A. Silva, and M. Vengalattore, Colloquium: Nonequilibrium dynamics of closed interacting quantum systems, *Rev. Mod. Phys.* **83**, 863 (2011).

- [2] L. D'Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics, *Adv. Phys.* **65**, 239 (2016).
- [3] R. Nandkishore and D. A. Huse, Many-body localization and thermalization in quantum statistical mechanics, *Annu. Rev. Condens. Matter Phys.* **6**, 15 (2015).
- [4] D. A. Abanin, E. Altman, I. Bloch, and M. Serbyn, Colloquium: Many-body localization, thermalization, and entanglement, *Rev. Mod. Phys.* **91**, 021001 (2019).
- [5] F. Alet and N. Laflorencie, Many-body localization: An introduction and selected topics, *C. R. Phys.* **19**, 498 (2018).
- [6] S. Moudgalya, B. A. Bernevig, and N. Regnault, Quantum many-body scars and Hilbert space fragmentation: A review of exact results, *ArXiv:2109.00548* (2021).
- [7] M. Serbyn, D. A. Abanin, and Z. Papić, Quantum many-body scars and weak breaking of ergodicity, *Nat. Phys.* **17**, 675 (2021).
- [8] F. Harper, R. Roy, M. S. Rudner, and S. L. Sondhi, Topology and broken symmetry in Floquet systems, *Annu. Rev. Condens. Matter Phys.* **11**, 345 (2020).
- [9] R. P. Feynman, Simulating physics with computers, *Int. J. Theor. Phys.* **21**, 467 (1982).
- [10] I. M. Georgescu, S. Ashhab, and F. Nori, Quantum simulation, *Rev. Mod. Phys.* **86**, 153 (2014).
- [11] T. D. Ladd, F. Jelezko, R. Laflamme, Y. Nakamura, C. Monroe, and J. L. O'Brien, Quantum computers, *Nature* **464**, 45 (2010).
- [12] I. Bloch, J. Dalibard, and S. Nascimbène, Quantum simulations with ultracold quantum gases, *Nat. Phys.* **8**, 267 (2012).
- [13] R. Blatt and C. F. Roos, Quantum simulations with trapped ions, *Nat. Phys.* **8**, 277 (2012).
- [14] C. Monroe, W. C. Campbell, L.-M. Duan, Z.-X. Gong, A. V. Gorshkov, P. W. Hess, R. Islam, K. Kim, N. M. Linke, and G. Pagano, *et al.*, Programmable quantum simulations of spin systems with trapped ions, *Rev. Mod. Phys.* **93**, 025001 (2021).
- [15] A. Aspuru-Guzik and P. Walther, Photonic quantum simulators, *Nat. Phys.* **8**, 285 (2012).
- [16] A. A. Houck, H. E. Türeci, and J. Koch, On-chip quantum simulation with superconducting circuits, *Nat. Phys.* **8**, 292 (2012).
- [17] J. Preskill, Quantum computing in the NISQ era and beyond, *Quantum* **2**, 79 (2018).
- [18] E. Altman, K. R. Brown, G. Carleo, L. D. Carr, E. Demler, C. Chin, B. DeMarco, S. E. Economou, M. A. Eriksson, and K.-M. C. Fu, *et al.*, Quantum Simulators: Architectures and Opportunities, *PRX Quantum* **2**, 017003 (2021).
- [19] Y. Alexeev, D. Bacon, K. R. Brown, R. Calderbank, L. D. Carr, F. T. Chong, B. DeMarco, D. Englund, E. Farhi, and B. Fefferman, *et al.*, Quantum Computer Systems for Scientific Discovery, *PRX Quantum* **2**, 017001 (2021).
- [20] D. Awschalom, K. K. Berggren, H. Bernien, S. Bhawe, L. D. Carr, P. Davids, S. E. Economou, D. Englund, A. Faraon, and M. Fejer, *et al.*, Development of Quantum Interconnects (QuICs) for Next-Generation Information Technologies, *PRX Quantum* **2**, 017002 (2021).
- [21] N. Jouppi, D. Yoon, G. Kurian, S. Li, N. Patil, J. Laudon, C. Young, and D. Patterson, A domain-specific supercomputer for training deep neural networks, *Commun. ACM* **63**, 67 (2020).
- [22] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, and A. Borchers, *et al.*, in *Proceedings of the 44th Annual International Symposium on Computer Architecture, ISCA '17* (Association for Computing Machinery, New York, 2017), p. 1.
- [23] J. Bradbury, R. Frostig, P. Hawkins, M. James 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, <http://github.com/google/jax>.
- [24] R. Frostig, M. Johnson, and C. Leary, Compiling machine learning programs via high-level tracing, *Systems for Machine Learning*, 23, 2018.
- [25] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, *et al.*, in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (USENIX Association and Savannah, GA, USA, 2016), p. 265.
- [26] F. Belletti, D. King, K. Yang, R. Nelet, Y. Shafi, Y.-F. Chen, and J. Anderson, Tensor processing units for financial Monte Carlo, *ArXiv:1906.02818* (2020).
- [27] Q. Wang, M. Ihme, Y.-F. Chen, and J. Anderson, A TensorFlow simulation framework for scientific computing of fluid flows on tensor processing units, *ArXiv:2108.11076* (2021).
- [28] Z. Pan and P. Mishra, Hardware acceleration of explainable machine learning using tensor processing units, *ArXiv:2103.11927* (2021).
- [29] T. Lu, T. Marin, Y. Zhuo, Y.-F. Chen, and C. Ma, in *2021 IEEE 18th International Symposium on Biomedical Computing (2021)*, p. 783.
- [30] C. Ma, T. Marin, T. J. Lu, Y. F. Chen, and Y. Zhuo, Nonuniform fast Fourier transform on TPUs, 2021.
- [31] T. Lu, Y.-F. Chen, B. Hechtman, T. Wang, and J. Anderson, Large-scale discrete Fourier transform on TPUs, *ArXiv:2002.03260* (2020).
- [32] F. Huot, Y.-F. Chen, R. Clapp, C. Boneti, and J. Anderson, High-resolution imaging on TPUs, *ArXiv:1912.08063* (2019).
- [33] Single program, multiple data (SPMD): a parallel programming paradigm whereby a single set of instructions (program) is evaluated simultaneously on multiple distinct inputs (data) hosted on a set of parallel processes.
- [34] A. Li, O. Subasi, X. Yang, and S. Krishnamoorthy, in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* (2020), p. 1.
- [35] T. Vincent, L. J. O'Riordan, M. Andrenkov, J. Brown, N. Killoran, H. Qi, and I. Dhand, JET: Fast quantum circuit simulations with parallel task-based tensor-network contraction, *ArXiv:2107.09793* (2021).
- [36] S. Efthymiou, S. Ramos-Calderer, C. Bravo-Prieto, A. Pérez-Salinas, D. García-Martín, A. García-Saez, J. I. Latorre, and S. Carrazza, QIBO: A framework for quantum

- simulation with hardware acceleration, *Quantum Sci. Technol.* **7**, 015018 (2021).
- [37] Y. Suzuki, Y. Kawase, Y. Masumura, Y. Hiraga, M. Nakadai, J. Chen, K. Nakanishi, K. Mitarai, R. Imai, and S. Tamiya, *et al.*, QULACS: A fast and versatile quantum circuit simulator for research purpose, *Quantum* **5**, 559 (2021).
- [38] T. Jones, A. Brown, I. Bush, and S. C. Benjamin, Quest and high performance simulation of quantum computers, *Sci. Rep.* **9**, 10736 (2019).
- [39] X.-Z. Luo, J.-G. Liu, P. Zhang, and L. Wang, Yao.jl: Extensible, efficient framework for quantum algorithm design, *Quantum* **4**, 341 (2020).
- [40] S. V. Isakov, D. Kafri, O. Martin, C. Vollgraf Heidweiller, W. Mruczkiewicz, M. P. Harrigan, N. C. Rubin, R. Thomson, M. Broughton, and K. Kissell, *et al.*, Simulations of quantum circuits with approximate noise using QSIM and CIRQ, *ArXiv:2111.02396* (2021).
- [41] F. Pan and P. Zhang, Simulating the Sycamore quantum supremacy circuits, *ArXiv:2103.03074* (2021).
- [42] S. Mandrà, J. Marshall, E. G. Rieffel, and R. Biswas, in *2021 IEEE/ACM Second International Workshop on Quantum Computing Software (QCS)* (IEEE, and St. Louis, MO, USA, 2021), p. 99.
- [43] A. Kelly, Simulating quantum computers using opencl, *ArXiv:1805.00988* (2018).
- [44] C. Huang, F. Zhang, M. Newman, J. Cai, X. Gao, Z. Tian, J. Wu, H. Xu, H. Yu, and B. Yuan, *et al.*, Classical simulation of quantum supremacy circuits, *ArXiv:2005.06787* (2020).
- [45] J. Gray, QUIMB: A PYTHON library for quantum information and many-body calculations, *J. Open Source Softw.* **3**, 819 (2018).
- [46] B. Villalonga, D. Lyakh, S. Boixo, H. Neven, T. S. Humble, R. Biswas, E. G. Rieffel, A. Ho, and S. Mandrà, Establishing the quantum supremacy frontier with a 281 pflop/s simulation, *Quantum Sci. Technol.* **5**, 034003 (2020).
- [47] A. G. M. Lewis, J. Beall, M. Ganahl, M. Hauru, S. B. Mallick, and G. Vidal, Large scale distributed linear algebra with tensor processing units, *ArXiv:2112.09017* (2021).
- [48] M. Ganahl, *et al.*, Tensor Processing Units for Simulating Quantum Circuits, *Sandbox@Alphabet* (to be published).
- [49] M. Hauru, A. Morningstar, J. Beall, M. Ganahl, A. Lewis, and G. Vidal, Simulation of quantum physics with tensor processing units: Brute-force computation of ground states and time evolution, *ArXiv:2111.10466* (2021).
- [50] R. Shillito, *et al.*, Classical simulation of superconducting quantum hardware using Tensor Processing Units, *Sandbox@Alphabet* (to be published).
- [51] R. Pederson, *et al.*, Tensor processing units for quantum chemistry, *Sandbox@Alphabet* (to be published).
- [52] R. Song, *et al.*, Simulation of spin light emitting diodes using tensor processing units, *Sandbox@Alphabet* (to be published).
- [53] M. Ganahl, *et al.*, Density matrix renormalization group using tensor processing units, *Sandbox@Alphabet* (to be published).
- [54] E. Gustafson, B. Holzman, J. Kowalkowski, H. Lamm, A. C. Y. Li, G. Perdue, S. Boixo, S. Isakov, O. Martin, and R. Thomson, *et al.*, Large scale multi-node simulations of $\mathbb{Z}_2$ gauge theory quantum circuits using google cloud platform, *ArXiv:2110.07482* (2021).
- [55] T. Oka and S. Kitamura, Floquet engineering of quantum materials, *Annu. Rev. Condens. Matter Phys.* **10**, 387 (2019).
- [56] M. Bukov, L. D'Alessio, and A. Polkovnikov, Universal high-frequency behavior of periodically driven systems: From dynamical stabilization to Floquet engineering, *Adv. Phys.* **64**, 139 (2015).
- [57] T. Oka and H. Aoki, Photovoltaic Hall effect in graphene, *Phys. Rev. B* **79**, 081406(R) (2009).
- [58] M. S. Rudner, N. H. Lindner, E. Berg, and M. Levin, Anomalous Edge States and the Bulk-Edge Correspondence for Periodically Driven Two-Dimensional Systems, *Phys. Rev. X* **3**, 031005 (2013).
- [59] P. Titum, E. Berg, M. S. Rudner, G. Refael, and N. H. Lindner, Anomalous Floquet-Anderson Insulator as a Nonadiabatic Quantized Charge Pump, *Phys. Rev. X* **6**, 021013 (2016).
- [60] V. Khemani, A. Lazarides, R. Moessner, and S. Sondhi, Phase Structure of Driven Quantum Systems, *Phys. Rev. Lett.* **116**, 250401 (2016).
- [61] D. V. Else, B. Bauer, and C. Nayak, Floquet Time Crystals, *Phys. Rev. Lett.* **117**, 090402 (2016).
- [62] H. C. Po, L. Fidkowski, T. Morimoto, A. C. Potter, and A. Vishwanath, Chiral Floquet Phases of Many-Body Localized Bosons, *Phys. Rev. X* **6**, 041070 (2016).
- [63] M. Ippoliti, K. Kechedzhi, R. Moessner, S. L. Sondhi, and V. Khemani, Many-Body Physics in the NISQ Era: Quantum Programming a Discrete Time Crystal, *PRX Quantum* **2**, 030346 (2021).
- [64] X. Mi, M. Ippoliti, C. Quintana, A. Greene, Z. Chen, J. Gross, F. Arute, K. Arya, J. Atalaya, and R. Babbush, *et al.*, Observation of time-crystalline eigenstate order on a quantum processor, *ArXiv:2107.13571* (2021).
- [65] L. D'Alessio and M. Rigol, Long-time Behavior of Isolated Periodically Driven Interacting Lattice Systems, *Phys. Rev. X* **4**, 041048 (2014).
- [66] A. Lazarides, A. Das, and R. Moessner, Equilibrium states of generic quantum systems subject to periodic driving, *Phys. Rev. E* **90**, 012110 (2014).
- [67] P. Ponte, Z. Papić, F. Huveneers, and D. A. Abanin, Many-Body Localization in Periodically Driven Systems, *Phys. Rev. Lett.* **114**, 140401 (2015).
- [68] A. Lazarides, A. Das, and R. Moessner, Fate of Many-Body Localization Under Periodic Driving, *Phys. Rev. Lett.* **115**, 030402 (2015).
- [69] D. A. Abanin, W. De Roeck, and F. Huveneers, Exponentially Slow Heating in Periodically Driven Many-Body Systems, *Phys. Rev. Lett.* **115**, 256803 (2015).
- [70] T. Mori, T. Kuwahara, and K. Saito, Rigorous Bound on Energy Absorption and Generic Relaxation in Periodically Driven Quantum Systems, *Phys. Rev. Lett.* **116**, 120401 (2016).
- [71] T. Kuwahara, T. Mori, and K. Saito, Floquet-Magnus theory and generic transient dynamics in periodically driven

- many-body quantum systems, *Ann. Phys. (NY)* **367**, 96 (2016).
- [72] D. A. Abanin, W. De Roeck, W. W. Ho, and F. Huveneers, Effective Hamiltonians, prethermalization, and slow energy absorption in periodically driven many-body systems, *Phys. Rev. B* **95**, 014112 (2017).
- [73] D. Abanin, W. De Roeck, W. W. Ho, and F. Huveneers, A rigorous theory of many-body prethermalization for periodically driven and closed quantum systems, *Commun. Math. Phys.* **354**, 809 (2017).
- [74] D. V. Else, B. Bauer, and C. Nayak, Prethermal Phases of Matter Protected by Time-Translation Symmetry, *Phys. Rev. X* **7**, 011026 (2017).
- [75] D. J. Luitz, R. Moessner, S. L. Sondhi, and V. Khemani, Prethermalization Without Temperature, *Phys. Rev. X* **10**, 021046 (2020).
- [76] A. Rubio-Abadal, M. Ippoliti, S. Hollerith, D. Wei, J. Rui, S. L. Sondhi, V. Khemani, C. Gross, and I. Bloch, Floquet Prethermalization in a Bose-Hubbard System, *Phys. Rev. X* **10**, 021044 (2020).
- [77] P. Peng, C. Yin, X. Huang, C. Ramanathan, and P. Capellaro, Floquet prethermalization in dipolar spin chains, *Nat. Phys.* **17**, 444 (2021).
- [78] F. Machado, G. D. Kahanamoku-Meyer, D. V. Else, C. Nayak, and N. Y. Yao, Exponentially slow heating in short and long-range interacting Floquet systems, *Phys. Rev. Res.* **1**, 033202 (2019).
- [79] JAX documentation, <https://jax.readthedocs.io/en/latest/> (accessed: 2022-01-15).
- [80] A. Chan, A. De Luca, and J. T. Chalker, Solution of a Minimal Model for Many-Body Quantum Chaos, *Phys. Rev. X* **8**, 041019 (2018).
- [81] C. Sünderhauf, D. Pérez-García, D. A. Huse, N. Schuch, and J. I. Cirac, Localization with random time-periodic quantum circuits, *Phys. Rev. B* **98**, 134204 (2018).
- [82] To be sure that we do not choose an atypical sample with, for example, very weak interactions or very strong single-qubit potentials anywhere, we split each H_k into its component noninteracting and interacting pieces and check that the norms of these pieces are $\lesssim 1$ and approximately 1, respectively.
- [83] S. Lloyd, Universal quantum simulators, *Science* **273**, 1073 (1996).
- [84] M. Suzuki, General theory of fractal path integrals with applications to many-body theories and statistical physics, *J. Math. Phys.* **32**, 400 (1991).
- [85] G. Vidal, Efficient Simulation of One-Dimensional Quantum Many-Body Systems, *Phys. Rev. Lett.* **93**, 040502 (2004).
- [86] M. Heyl, P. Hauke, and P. Zoller, Quantum localization bounds Trotter errors in digital quantum simulation, *Sci. Adv.* **5**, eaau8342 (2019).
- [87] C. Fleckenstein and M. Bukov, Thermalization and prethermalization in periodically kicked quantum spin chains, *Phys. Rev. B* **103**, 144307 (2021).
- [88] Quantum AI team and collaborators, QSIM, 2020.
- [89] Cirq Developers, CIRQ, 2021.
- [90] P. Weinberg and M. Bukov, QuSpin: A PYTHON package for dynamics and exact diagonalisation of quantum many body systems part I: Spin chains, *SciPost Phys.* **2**, 003 (2017).
- [91] T. L. M. Lezama and J. H. Bardarson, Apparent slow dynamics in the ergodic phase of a driven many-body localized system without extensive conserved quantities, *Phys. Rev. B* **99**, 161106(R) (2019).
- [92] J. Šuntajs, J. Bonča, T. Prosen, and L. Vidmar, Quantum chaos challenges many-body localization, *Phys. Rev. E* **102**, 062144 (2020).
- [93] M. Kiefer-Emmanouilidis, R. Unanyan, M. Fleischhauer, and J. Sirker, Slow delocalization of particles in many-body localized phases, *Phys. Rev. B* **103**, 024203 (2021).
- [94] D. Sels and A. Polkovnikov, Dynamical obstruction to localization in a disordered spin chain, *ArXiv:2009.04501* (2020).
- [95] D. A. Abanin, J. H. Bardarson, G. De Tomasi, S. Gopalakrishnan, V. Khemani, S. A. Parameswaran, F. Pollmann, A. C. Potter, M. Serbyn, and R. Vasseur, Distinguishing localization from chaos: Challenges in finite-size systems, *Ann. Phys. (NY)* **427**, 168415 (2021).
- [96] P. Sierant, D. Delande, and J. Zakrzewski, Thouless Time Analysis of Anderson and Many-Body Localization Transitions, *Phys. Rev. Lett.* **124**, 186601 (2020).
- [97] R. K. Panda, A. Scardicchio, M. Schulz, S. R. Taylor, and M. Žnidarič, Can we study the many-body localisation transition?, *EPL (Europhys. Lett.)* **128**, 67003 (2020).
- [98] D. J. Luitz and Y. Bar Lev, Absence of slow particle transport in the many-body localized phase, *Phys. Rev. B* **102**, 100202 (2020).
- [99] A. Morningstar, L. Colmenarez, V. Khemani, D. J. Luitz, and D. A. Huse, Avalanches and many-body resonances in many-body localized systems, *ArXiv:2107.05642* (2021).
- [100] P. Sierant and J. Zakrzewski, Can we observe the many-body localization?, *ArXiv:2109.13608* (2021).