Standardizing Metrics and Benchmarks for Quantum Computers

Joshua Combes and **Marcus P. da Silva**
Rigetti Quantum Computing
(8th of February 2019)

Executive summary:
Many aspects of building quantum computers can be categorized as a research effort while some aspects are closer to engineering. The theory and practice of benchmarking quantum computers, compilers, and virtual machines is, at this stage, a research endeavor. Standardising too early could harm progress in the development and exploration of these devices. We expect that in 5 years it would be reasonable to endeavor to standardize benchmarks for near term devices and applications. Until then, the community is best served by encouraging innovation and exploration of many different benchmarks.

Introduction:
After a quarter century of theoretical and experimental investigations there is genuine progress towards building intermediate scale quantum computers. As a result, global investments in quantum technologies are skyrocketing. It is likely that, in the next 18 months, devices of order 100 qubits will be publicly available and in the next 3-5 years one can realistically expect devices with 200 to 500 qubits.

Pressings question for all interested parties are:

- **Q1** For what problems will quantum computers outperform other computing methods in some way?
- **Q2** When will we see these problems solved on real quantum hardware, and exhibiting said advantageous performance?
- **Q3** How does the answer to **Q1** and **Q2** depend on *realistic* errors and numbers of qubits?

It is reasonable to use benchmarks as a valuable tool for tracking progress towards these questions, but here we argue that a *premature* focus on standardizing a set of benchmarks is likely to *slow* progress towards the important scientific goal of demonstrating quantum advantage.

The crux of the issue is that it is easy to propose methods to benchmark the performance of quantum devices, but it is difficult to ensure that these represent operationally meaningful metrics, especially because applications that demonstrate quantum advantage in near term devices have yet to be proposed.

On the longer time scale required to reach fault-tolerance, algorithms like Shor's algorithm will demonstrate quantum advantage. However, applications for near and medium term devices have only recently gained attention of the scientific community. While there are certainly many promising proposals, e.g., variants of variational quantum eigensolver (VQE) and the quantum

approximate optimization algorithm (QAOA), the problem of demonstrating commercially relevant quantum advantage over classical computers remains largely open. We will refer to the yet to be developed algorithms and applications as *potential applications*. Deciding on benchmarks before the potential applications and/or problem instances have identified is unlikely to yield useful guideposts on the road to quantum advantage.

The more pernicious problem is that fixing a set of benchmarks before the potential applications have been identified runs the risk of encouraging premature optimization aligned to the wrong metrics, which in turn can hold back progress towards quantum advantage. This problem is not specific to one platform or quantum computer implementation: different platforms may demonstrate quantum advantage with different problems that are best matched to their architectural and physical idiosyncrasies. Attempting to create a metric that is agnostic of these details may single out one platform and prematurely "prune" the diversity that may be crucial for the success of this field.

Ultimately, there is no argument at this stage against all quantum hardware technologies succeeding, and no clear argument for which near-term quantum algorithms may yield advantage over classical algorithms. The focus of the quantum computing community should be to explore these possibilities and the bring clarity to these questions before committing to any particular hardware, architectural or algorithmic path.

Bad benchmarks:
Here we use an example to illustrate a general point. One might have low level access to near term quantum hardware and try to use a classical optimization (or a machine learning) algorithm to design a better pulse sequence to implement, e.g., a Toffoli gate. The optimization requires a benchmark metric, and one natural choice is the average error in the corresponding classical truth table (the correspondence between bitstring inputs and outputs), see e.g. [Linke et al, PNAS 114 13 3305 (2017)]. If the optimization algorithm is sophisticated enough, this will result in a sequence of operations corresponding to the Margolus gate, which is much shorter than the textbook (optimal) recipe for a Toffoli gate, but which has the same truth table. If this optimized gate is to be used in a quantum algorithm that expects the Toffoli gate, but that acts on coherent superpositions of computational states, the observed behavior will be very different from what one would naively infer from the truth table.

This concretely illustrates how bad benchmarks may lead to a pathologically suboptimal solution. Goodhart's law will also undoubtedly apply if we fixate on a single metric, even if it is more reasonable than the one above. Imperfections in quantum computers are multifaceted, as are the useful features in these devices. For example, while qubit connectivity may be highly beneficial at small problem sizes, some architectures may trade that off for lower effort in scaling to 100s or 1000s of qubits. Different architectural choices may then lead to fundamentally different modalities of errors, such as crosstalk or loss of qubits in transport. Research in quantum benchmarking is only now awakening to the importance of *holistics metrics* (that can capture global behavior with few assumptions about the nature of the errors) instead of *diagnostic metrics* (that can help an experimentalist debug a quantum computer). It is our belief that IEEE should be encouraging efforts that will explore the development of such holistic

metrics, and efforts connected them to near term quantum applications, before any useful discussion of standardization can happen.

Benchmarking quantum computers is a research problem:
This might seem surprising as some parts of the field of quantum characterization, validation, and verification (QCVV) are decades old. However access to more than 8 qubits, until recently, was a very rare commodity. The early research led to a great set of methods to benchmark one or two qubits. The idea was benchmarking at that level would allow one to "bootstrap" up to reasoning about performance of larger devices. This thinking was too reductionist. Some techniques miss many important errors (e.g., crosstalk or non-Markovianity). Other techniques (e.g., process tomography) cannot scale to larger systems. We are only now starting to see the first round of protocols (e.g., DRB and quantum volume) that can capture holistic circuit level performance. However, even these metrics are implicitly associated with the assumption that quantum devices must be essentially noiseless to be useful. Moreover there are glaring holes in the literature regarding temporal variation. How should we measure and report of stability of quantum computers? If a gate is reported as having an 99% fidelity over a 48 hour period, how far and how often will it deviate from that number? How does that impact the performance of variational algorithms? None of these have been addressed in the literature.

Benchmarking near term algorithms and applications:
A useful starting point in this effort is to focus on the few known near term algorithms, such as QAOA, VQE, and QML algorithms such as quantum kitchen sinks (QKS). Because these algorithms have features that allow them to run on noisy devices, it is unclear when there is a benefit to moving from one particular device size (with some gateset, connectivity, and error rate) to a larger device with slightly higher error rates or different connectivity. Even the simple question of when one realization of QAOA or VQE is outperforming another has not been widely addressed. Several credible possibilities for performance metrics arise: time to optimal solution, time to approximate solution, convergence rate of the solution quality, etc. Pulling this thread, more open questions arise: how reproducible and stable should the experiments be on a time scale of seconds/minutes/hours, so that a variational algorithm like QAOA or VQE can meaningfully optimize the variational parameter? How does that depend on the problem size?

Recommendations:
It should be apparent that there are more questions than answers in this quest. It is easy to come up with benchmarks but it is difficult to make sure they are good. The experts in benchmarking quantum systems, the QCVV community, have been actively working on these problems. The research is not at a stage where benchmarks should codified and standardized; we should instead aim to formulate the relevant question for the performance of near term algorithms, and explore ways to measure the resources that impact this performance. This effort is non-trivial, and could use IEEE's leadership and experience in organizing workshops focused on addressing these questions, and bringing clarity to the most promising ways to answer them.