Comments on Benchmarking of Quantum Computers

Scott Pakin
Computer, Computational, and Statistical Sciences Division
Los Alamos National Laboratory
pakin@lanl.gov

February 1, 2019

Abstract

Injecting some consistency into the terminology and methodology used to evaluate quantum-computing hardware is a worthwhile endeavor. This document represents my personal feedback on "An IEEE Framework for Metrics and Benchmarks of Quantum Computers" and other thoughts I have on the subject.

1 Document feedback

In "An IEEE Framework for Metrics and Benchmarks of Quantum Computers", the Computational Models part makes sense. The Use Cases start to get a bit fuzzy. And I find the Technology Layers rather confusing.

Use cases When I think of use case, I don't think of specific devices as the primary discriminator but rather as something providing native support for solving a class of problems. For example, I might think of "optimization", "sampling from a Boltzmann distribution", "simulating quantum systems", and "general-purpose computing" as use cases rather than Table 1's "NISQ", "QA", "quantum simulators", and "FTQEC" devices.

Technology layers Is this supposed to represent pieces of hardware or software whose performance one might want to measure? Some of the terminology is confusing. For example, defining a "physical register" as something that "stores qubits" strikes me as rather vague. Does this mean a chip? A package? A complete system? All of those "store qubits" in some sense. It may help to start specifying technology layers with just those things that contribute to execution speed or solution quality: qubit implementation, qubit count, decoherence time and other noise properties, operation set (e.g., gate set for a gate-model system or Hamiltonian properties for an annealing-based system), qubit interconnect topology, operation reconfiguration latency, coefficient precision/accuracy, etc. I see Table 3 entries like "manufacturing" being extraneous to metrics/benchmarks except as it affects the preceding properties.

2 Other comments

I hope this metrics and benchmarking effort is being coordinated with IEEE's Quantum Computing Nomenclature Working Group to ensure consistent usage of terminology.

As far as metrics go, some thought should be given to empirical performance vs. asymptotic performance. Gate-model systems are the darling of most people in the quantum-computing arena because their performance advantages can be proven from a complexity-theory standpoint while no such proofs yet exist for annealing-based systems. However, annealing-based systems tend to be larger and arguably more resilient to noise. Hence, benchmarking/metrics need to distinguish between comparing ideal implementations of various quantum computers with real-world, point-in-time implementations.

One metric to consider is some notion of “goodput”,[1] essentially “valid results per unit time”. Because NISQ devices so commonly give wrong results, it may be worth determining empirically how many times on average a given problem needs to be run on a given system to achieve a correct answer with high likelihood.

It would be worthwhile comparing programs run on quantum computers to programs run on classical computers. This could produce interesting trend lines related to when the former truly exhibits value over the latter. The challenge lies in making a fair comparison. Naturally, the best classical algorithm/implementation should be used for comparisons. However, is it fair for a single quantum processor be compared against a single classical core/socket? Should price be kept constant (e.g., comparing a $40,000,000 quantum computer to $40,000,000 worth of classical hardware)? Should power be kept constant (e.g., comparing a quantum computer that draws 25 kW to 25 kW worth of classical hardware)? Energy? Or is time to (correct) solution all that anyone cares about?

Be sure always to measure end-to-end performance. There are numerous classical overheads involved in running a program on a quantum computer, and these should not be ignored when reporting empirical measurements. Getting a problem onto a quantum processor and getting the results back to a classical host are far from being zero-time operations and should therefore contribute to the total time being reported. In particular, there can be substantial classical setup costs needed for software to map a logical problem onto a physical device. I would recommend specifying benchmarks at a relatively high level: a rich gate set for gate-model quantum computers and a generic, Ising-model Hamiltonian for annealing-based quantum computers—assuming all-to-all connectivity in both cases. Then include in any empirical measurements the classical overheads needed to express these logical problems in terms of the hardware’s physical capabilities. This should lead to an equitable comparison of densely connected vs. sparsely connected systems and systems with rich gate sets vs. limited gate sets.

[1] `https://en.wikipedia.org/wiki/Goodput`