Metrics and Benchmarks for Quantum Processors: State of Play

Robin Blume-Kohout and **Kevin Young**
Quantum Performance Laboratory
Sandia National Laboratories; Albuquerque, NM 87185 and Livermore, CA 94550

November 1, 2018

A compelling narrative has taken hold as quantum computing explodes into the commercial sector: *Quantum computing in 2018 is like classical computing in 1965*. In 1965 Gordon Moore wrote his famous paper about integrated circuits [1], saying :

> *At present, [minimum cost] is reached when 50 components are used per circuit. But… the complexity for minimum component costs has increased at a rate of roughly a factor of two per year… by 1975, the number of components per integrated circuit for minimum cost will be 65,000.*

This narrative is both appealing (we want to believe that quantum computing will follow the incredibly successful path of classical computing!) and plausible (2018 saw IBM, Intel, and Google announce 50-*qubit* integrated chips). But it is also deeply misleading.

Here is an alternative: *Quantum computing in 2018 is like classical computing in 1938.* In 1938, John Atanasoff and Clifford Berry built the very first electronic digital computer. It had no program, and was not Turing-complete. Vacuum tubes – the standard "bit" for 20 years – were still 5 years in the future. ENIAC and the achievement of "computational supremacy" (over hand calculation) wouldn't arrive for 8 years, despite the accelerative effect of WWII. Integrated circuits and the information age were more than 20 years away.

Neither of these analogies is perfect. Quantum computing *technology* is more like 1938, while the level of funding and excitement suggest 1965 (or later!). But the point of the cautionary analogy to 1938 is simple: **Quantum computing in 2018 is a *research* field. It is far too early to establish metrics or benchmarks for performance. The best role for neutral organizations like IEEE is to encourage and shape research *into* metrics and benchmarks, so as to be ready when they become necessary**.
This white paper presents the evidence and reasoning for this claim. We explain what it means to say that quantum computing is a "research field", and why metrics and benchmarks for quantum processors also constitute a research field. We discuss the potential for harmful consequences of prematurely establishing standards or frameworks. We conclude by suggesting specific actions that IEEE or similar organizations can take to accelerate the development of good metrics and benchmarks for quantum computing.

The maturity of quantum computing technology

The IEEE Framework on Metrics and Benchmarks [2] seeks

> *to outline a framework by which the continuing progress in quantum engineering can be monitored by the broader quantum computing community… to guide the decisions of policy makers and technology stakeholders as well to monitor the overall growth of the quantum research community.*

It's appealing to believe that progress can be monitored, and decisions guided, by a short list of numbers. In some areas of technology and engineering, they can. But not in

quantum computing right now. At this time, there is no substitute for deep, extensive technical knowledge of the field. A CEO, investor, or program manager who wishes to monitor progress in quantum computing, but lacks this knowledge personally, has only one option: hire an expert agent.

This stems from the immaturity of the field. Today's most advanced quantum processors are like infants. Metrics and benchmarks that are useful for adult humans (e.g., IQ or SAT scores) are blatantly inapplicable to an infant. An infant's whole purpose is to grow into an adult. Monitoring its progress requires skills and knowledge totally different from what's needed to evaluate an adult. Children and immature technologies both progress counterintuitively and sometimes even appear to regress (losing baby teeth or entering adolescence).

Today, there are as yet no *useful* quantum computers. The noisy 5-50 qubit quantum processors that exist in 2018 may or may not deserve to be called "quantum computers". What we do have is several distinct technologies and roadmaps leading toward genuinely useful quantum computers. The leading qubit technologies are superconducting circuits, trapped ions, and semiconductor nanostructures. Within each category, several very different approaches are being explored. Each faces *unique* engineering obstacles. And for each architecture, the truly critical question is "Will it surmount its specific engineering obstacles, possibly thanks to unpredicted lateral innovation?" Improvement at easily measured and/or cross-platform metrics is less significant, and a poor measure of progress.

These are the hallmarks of a technology in its *research* phase, rather than its *production* phase. Until recently, the key research question was "Can quantum computing work at all?" Establishing a consensus answer ("yes") is tremendously exciting. But the immediate next step is grappling with the equally hard and critical question "What approach will work?" We do not yet have answers to fundamental, existential questions like "Will it use microwaves or lasers?", "Will it run at millikelvin temperatures or 77K?", or "What architecture can reduce crosstalk to tolerable levels?" These are not variations on "What approach works best?" The community is asking "What approach works *at all*?" Until research reveals an answer, "progress" will resemble exploration of a tree, not a road.

To appreciate just how immature quantum computing technology is, take a moment to consider the state of computing in 1965 (when Moore's paper appeared). Electronic computers had existed for 27 years, and achieved "computational supremacy" – outperforming rooms of dedicated calculating savants – as early as 1944 with Colossus and ENIAC. Vacuum tubes had been established as the "bit" of choice for 15 years, followed by discrete transistors for another 10. Mass-produced commercial computers had been in continuous operation solving important problems since 1951.

In contrast, the quantum analogue to vacuum tubes hasn't yet been identified. Quantum processors are still impractically bulky (individual qubits are deceptively tiny, but each one requires a dedicated *control* apparatus occupying liters of space). The field is still working toward "quantum supremacy" – i.e., outperforming existing technology at a highly contrived and useless problem. Solving any *useful* problem remains further in the future. A quantum computer has never solved a real-world computational problem that couldn't be solved more easily another way. Quantum computation is genuinely, excitingly *promising* – but like regular computation in 1938 (or an infant!), that *promise* is all it offers now.

Commercial excitement notwithstanding, quantum computing is still mainly *research*. The past two years have been genuinely revolutionary, because the challenges have shifted

from physics to engineering. Five years ago, quantum processors were physics experiments. Today, they are engineering experiments. But this should not obscure the fact that this remains cutting-edge research – whose goal is to determine what can and cannot be done – *not* predictable production engineering. Cutting-edge 10-20 qubit processors – e.g., IBM's QX, Rigetti's QCS, or those recently commissioned by the US Department of Energy – are known as *testbeds*. But this term should not be misinterpreted. Testbed-class quantum processors are not reliable harnesses on which new quantum software can be evaluated. They are risky, bleeding-edge engineering experiments, meant to test fundamental aspects of quantum engineering, especially control and integration.

Unique challenges to benchmarking quantum processors

The main contention of this paper is this: **Benchmarking quantum processors remains, necessarily, a research endeavor**. This stems in part from the fact that quantum computing itself is research, but not entirely. Few sensible metrics for quantum performance are known – even in the scientific literature – and even fewer benchmarking techniques have been proposed. Almost all of these are recent, speculative, and the subject of fierce and productive debate. None of them are clearly suitable even for today's 10-20 qubit processors (much less the larger ones expected soon). Most aspects of performance have not been addressed at all (the few widely used benchmarking techniques are focused on error performance). There is no consensus yet within the community of experts on how to quantify or assess performance.

It's tempting and easy to take inspiration from benchmarks for classical computers (which *are* mature and well-developed), and propose some kind of synthetic benchmark for quantum computers. But standard synthetic benchmarks are based on real-world applications. Even LINPACK, now seen as relatively artificial, measures performance on the real-world practical task of performing dense linear algebra. Application-based benchmarks for quantum computers face two critical problems:

1. No applications giving practical advantage have been found for current near-future quantum computers. Many candidates and/or heuristics have been discussed, but it remains unclear which might yield genuinely useful advantage in the future.
2. Realistic examples of the candidate applications that *might* be relevant are still too big or too complex to run on any existing quantum processors.

These problems are not insurmountable. Several research groups, including ours, are tackling them. The future application landscape is growing and becoming more clear, quantum processors are growing in capability, and we're working on ways to cram "representative" quantum circuits into the limitations of extant processors. But this remains, clearly and blatantly, *research*. In 5 years, we expect that research will have flowered and paid off, and mature techniques will appear and be candidates for *de facto* or consensus standards. But application-based benchmarks aren't even possible today.

Instead, most of the effort to date has gone into *component-level* benchmarking of individual qubits and quantum logic gates. This is now a mature field of research, and several techniques are well-known – quantum tomography, direct fidelity estimation, robust phase estimation, randomized benchmarking, gate-set tomography. But this, too, remains a research area. The most well-tested and broadly used technique is randomized benchmarking (RB), yet in the past year nearly a dozen papers have either (1) revealed new, unexpected aspects of RB's behavior or interpretation, or (2) proposed significantly

new and arguably improved ways to do RB. Vigorous debate about all of the known protocols is ongoing in the community and scientific journals.

This doesn't mean that those protocols shouldn't be used. But it does preclude standardization – either *de facto*, or imposed. Scientific debate, reproduction, and extended peer review are essential steps in this process. They lead to the emergence of consensus, at which point organizations like IEEE are needed to shepherd, curate, and codify best practices. But at this time, there *are* no accepted best practices – and this is for good reasons. It is still not clear – even to experts – what properties of components (qubits, gates, etc) are necessary or desirable for quantum processors. RB emphasizes certain properties, while gate-set tomography emphasizes others. Until further research (ongoing right now!) establishes how component-level properties determine overall performance, it's impossible to choose wisely between techniques.

Identifying best practices will also require a significant corpus of real-world usage. In the past two years, the number of viable quantum processors has grown tremendously. Various benchmarking techniques are being tried out on them, and that corpus is now growing. Within 5 years, sufficient data and experience will have been accumulated to guide wise choices among them. But today, that corpus of experience is insufficient.

Bad benchmarks are worse than no benchmarks

Metrics and benchmarks *could* be established now. But they would almost certainly be bad ones. Bad benchmarks would harm the field, stifle innovation at a time when innovation is needed more than anything else, and create perverse incentives to direct R&D efforts toward unproductive goals. In a classic article from 1987 in *IEEE Spectrum* [3], Dongarra *et al* write:

> *Although benchmarks are essential in performance evaluation, simple-minded application of them can produce misleading results. In fact,* ***bad benchmarking can be worse than no benchmarking at all.***

They go on to emphasize that a useful benchmark must model the device's real-world workload. As we explained in the previous section, this is currently impossible – current quantum processors can't run any real-world applications, and nobody knows what the first useful real-world applications will be, or what aspects of future processors they'll stress.

Several benchmarks have been explored in the research literature. They include tiny implementations of Shor's and Grover's algorithms, randomized benchmarking, gate-set tomography, the circuits defined by IBM's "quantum volume" metric [4], and an array of other algorithms (see, e.g. [5]). This research activity is highly desirable! It simultaneously tests and stresses experimental quantum hardware and the candidate benchmarks themselves. It contributes to an intellectual ferment from which continually better ideas emerge. But each of these benchmarks is sensitive to particular limitations, and insensitive to others. If one was "established" and promoted, even as a *de facto* standard, researchers and engineers would be incentivized to optimize their hardware to perform well on that particular test – which, almost certainly, would *not* measure the properties that will be important for future applications (since those aren't known yet).

This is not a hypothetical scenario. At various times, computers have been optimized for clock speed (GHz), Whetstones, BogoMIPS, and LINPACK. The digital camera industry focused obsessively on megapixel count for nearly a decade. Cars emphasized horsepower (to the detriment of fuel economy). Elementary education today is heavily

influenced by Common Core standards, and strong incentives to "teach to the test" at the expense of creativity.

All of these *de facto* metrics were recognized by experts to be flawed and actively detrimental to useful progress, but history shows that it's surprisingly difficult to resist the lure of "benchmarketing". We are concerned that the same phenomenon may already be appearing in quantum computing, as many experimentalists optimize their qubit gates for randomized benchmarking (RB). Like Whetstones or megapixels, RB is a real and useful benchmark – but no single number is likely to accurately capture quantum performance, and so optimizing exclusively for *any* single number is probably unwise.

Roles for IEEE

So far, we have argued that (1) quantum computing itself is at a far lower Technology Readiness Level than is sometimes suggested, and therefore neither requires nor is suitable for the adoption of standards for metrics and benchmarks; and (2) benchmarking and characterizing quantum processors is itself a subject of active research, containing too many essential yet as-yet-unanswered questions to support consensus or standards.

What, then, are valuable roles for IEEE – and specifically for the IEEE Framework on Metrics and Benchmarks?

A framework is a rigid thing – a scaffolding upon which solid structures can be built. *Quantum metrics and benchmarks are, at this time, too fluid for a framework*. Frameworks presume that at least *some* foundational questions are clearly established and beyond debate. This is not yet the case here. Most of the assumptions that would have to go into a useful framework are, themselves, subjects of current debate and uncertainty!

A better metaphor is "scaffolding" – temporary, ad hoc frameworks. Across the United States, various research groups are erecting scaffolds for benchmarking, which they describe in talks and publish in journals. They are creative, diverse, competitive. Their proponents argue at conferences and in journals. This is the marketplace of ideas, where research is performed and honed. It describes the state of play in the field of metrics and benchmarks today – experts generating ideas, debating them, and identifying their flaws.

IEEE can play many valuable roles in this process. In the next 3-6 years, a shaky consensus will begin to emerge. IEEE can nudge researchers (in academia, government, and industry) toward creating the pieces of this consensus, and toward fitting those pieces together and agreeing on them. IEEE can hasten that consensus by supporting research, advocating for specific kinds of research that will generate better benchmarks and/or validate proposed ones, and curating research results via conferences, special meetings, and journals. And once that consensus begins to emerge – first, a consensus on a framework, and later a consensus on specific metrics and benchmarks – IEEE will play a critical role in assembling and documenting it, and facilitating negotiations between stakeholders.

Acknowledgements and Disclaimers: Any subjective views or opinions that might be expressed herein do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia (NTESS), LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Bibliography

[1] G. Moore, “Cramming more components onto integrated circuits.” *Electronics* 38.8 (1965)

[2] IEEE Framework on Metrics and Benchmarking draft document version 0.2. Retrieved October 16, 2018 from https://docs.google.com/document/d/1iNETuK_XOA-HpwkCf4_H6xjVir7tWNVh1Y-jIEMfhWM

[3] J. Dongarra *et al*, "Computer benchmarking: Paths and pitfalls." *IEEE Spectrum* 24.7 (1987): 38-43.

[4] L. Bishop *et al*, “Quantum Volume.” 2017. Downloaded from https://pdfs.semanticscholar.org/650c/3fa2a231cd77cf3d882e1659ee14175c01d5.pdf

[5] N. Linke, "Experimental comparison of two quantum computing architectures." *PNAS* 114.13 (2017): 3305-3310.