Ten Years of Benchmarking Quantum Computers: Lessons Learned, Insights Gained, and Challenges Ahead

Presentation prepared for the

TQCI International Seminar on Benchmarks for Quantum Computers

24-25 June 2025 - Paris

by

Tom Lubinski

QED-C[*] Standards and Benchmark Committee, Quantum Computing Group Lead

Senior Technical Advisor, Quantum Circuits Inc.

[*] QED-C = Quantum Economic Development Consortium

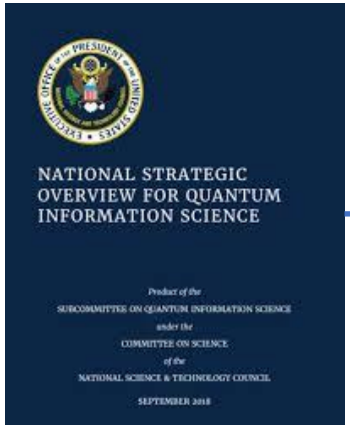
About Me

- From 2019, chaired QED-C Standards and Benchmark Committee (3 yrs)
  - Directed QED-C Computing Group’s App-Oriented Benchmarks project for 3 yrs
  - Organized 4 years of summer intern programs and member involvement
  - Produced 5 papers on concrete app-oriented benchmarking methods

- From 2016, with Quantum Circuits Inc
  - Senior Technical Advisor for 3, Chief Software Architect for 5

- 3 decades prior as CEO and Technical founder of SL Corporation
  - (not the Korean auto parts company)
  - Graphical Modeling System Software for Process Monitoring and Visualization

About Quantum Economic Development Consortium (QED-C)



EXECUTIVE OFFICE OF THE PRESIDENT OF THE UNITED STATES

NATIONAL STRATEGIC OVERVIEW FOR QUANTUM INFORMATION SCIENCE

Product of the
SUBCOMMITTEE ON QUANTUM INFORMATION SCIENCE
under the
COMMITTEE ON SCIENCE
of the
NATIONAL SCIENCE & TECHNOLOGY COUNCIL

SEPTEMBER 2018

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

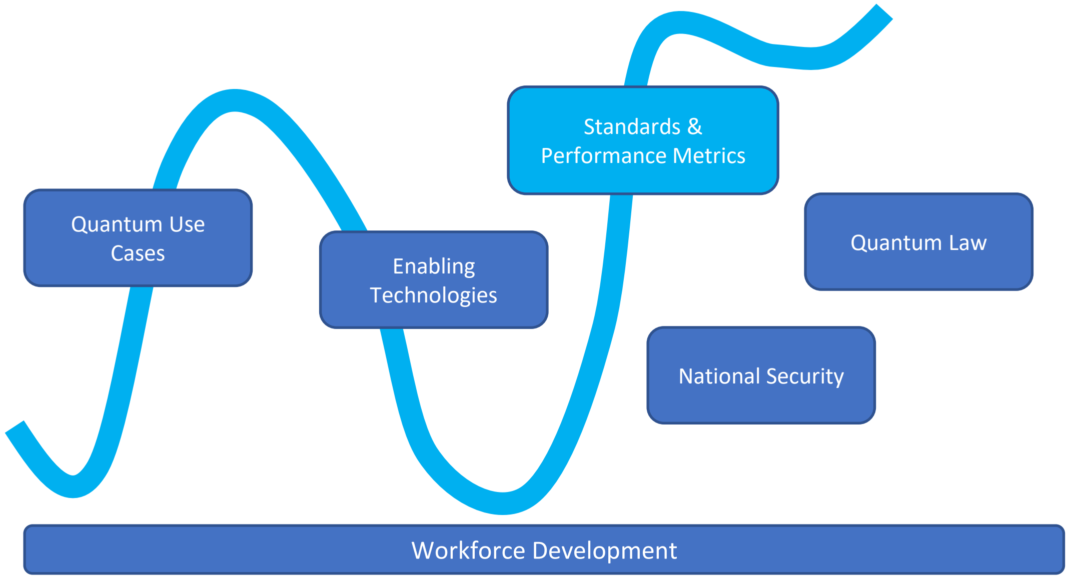
SRI International®

QED·C

- Sep 2018

- **Encourage and facilitate** ***global*** quantum research and development and grow the emerging quantum industry in computing, communications and sensing.

- **Technical Advisory Committees**



Quantum Use Cases

Enabling Technologies

Standards & Performance Metrics

Quantum Law

National Security

Workforce Development

*Standards TAC: Identify ways to **encourage the development of standards and performance metrics** in Quantum Information Science to accelerate commercialization of quantum-based products and services.*

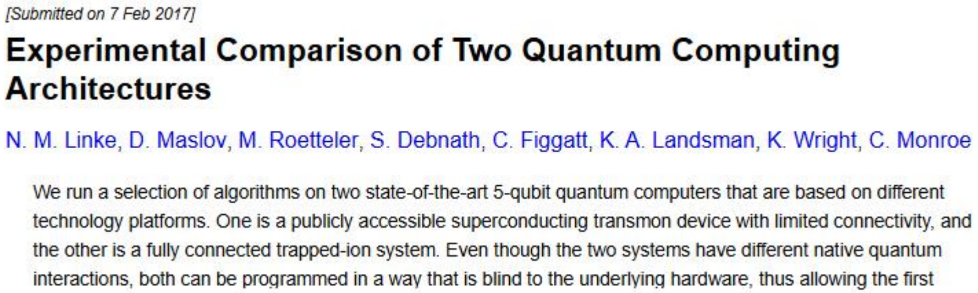
It is likely you have a severe case of BIO!

(Benchmark Information Overload)

So, let’s get to the point …

Ten Years? It’s more than that!

(RB 2008 + 5-Qubit papers 2017) / 3 = ~10 years

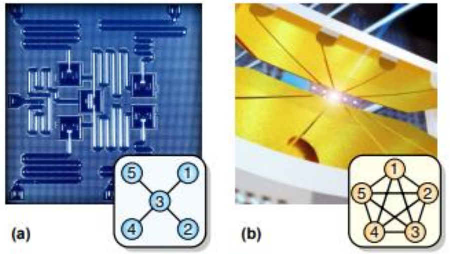


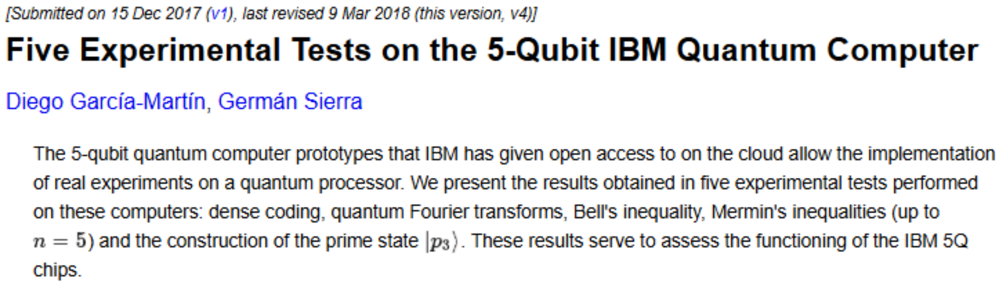
[Submitted on 7 Feb 2017]

Experimental Comparison of Two Quantum Computing Architectures

N. M. Linke, D. Maslov, M. Roetteler, S. Debnath, C. Figgatt, K. A. Landsman, K. Wright, C. Monroe

We run a selection of algorithms on two state-of-the-art 5-qubit quantum computers that are based on different technology platforms. One is a publicly accessible superconducting transmon device with limited connectivity, and the other is a fully connected trapped-ion system. Even though the two systems have different native quantum interactions, both can be programmed in a way that is blind to the underlying hardware, thus allowing the first





[Submitted on 15 Dec 2017 (v1), last revised 9 Mar 2018 (this version, v4)]

Five Experimental Tests on the 5-Qubit IBM Quantum Computer

Diego García-Martín, Germán Sierra

The 5-qubit quantum computer prototypes that IBM has given open access to on the cloud allow the implementation of real experiments on a quantum processor. We present the results obtained in five experimental tests performed on these computers: dense coding, quantum Fourier transforms, Bell's inequality, Mermin's inequalities (up to $n = 5$) and the construction of the prime state $|p_3\rangle$. These results serve to assess the functioning of the IBM 5Q chips.

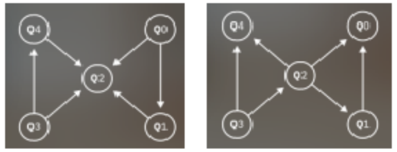
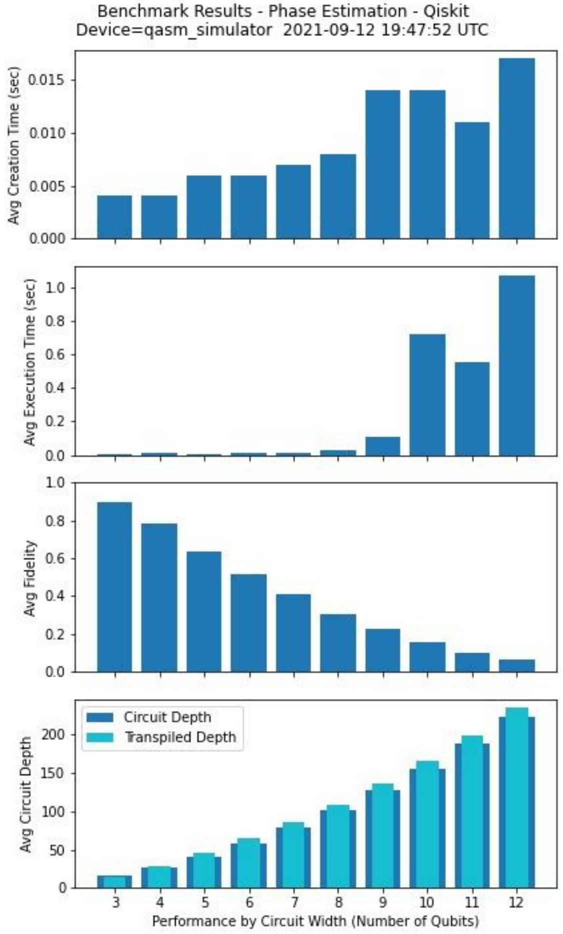


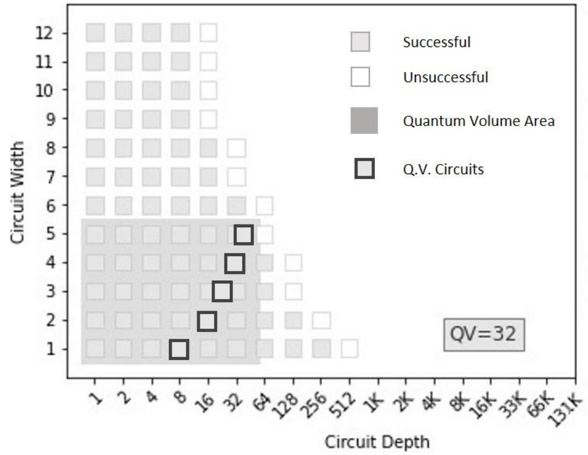
Figure 1. Diagram of the available cNOTs among qubits on the IBM 5Q computers: ibmqx2 (left) and ibmqx4 (right). The qubits are represented by circles, while the cNOTs are arrows pointing from control qubit to target qubit.

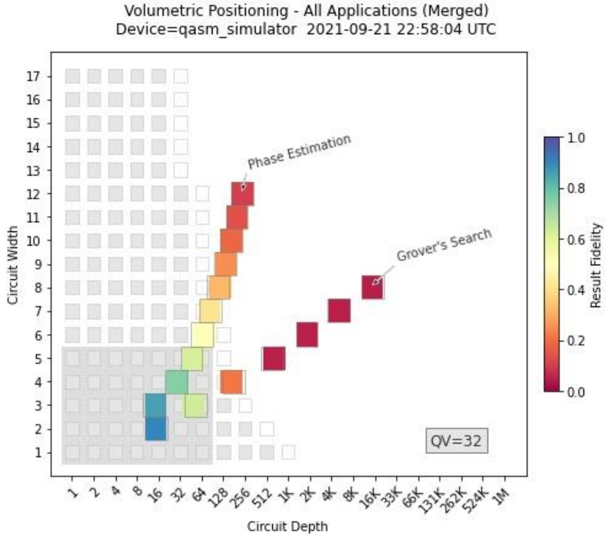
QED-C Application-Oriented Performance Benchmarks for Quantum Computing (2020-21)

Present Run Time and Quality Metrics in Bar Charts and Volumetric Charts by App

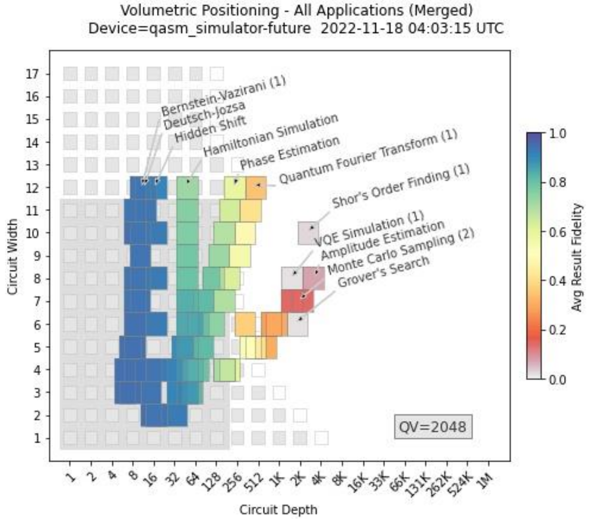

Benchmark Results - Phase Estimation - Qiskit
Device=qasm_simulator 2021-09-12 19:47:52 UTC

Exec Time, Fidelity, Gate Counts over Number of Qubits

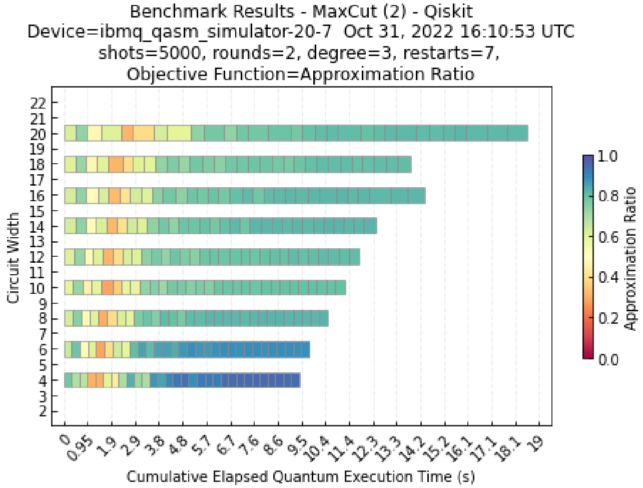



Volumetric Positioning - All Applications (Merged)
Device=qasm_simulator 2021-09-21 22:58:04 UTC

Uses of Volumetric Benchmarking
Sandia QPL - Blume-Kohout, Young, Proctor


Volumetric Positioning - All Applications (Merged)
Device=qasm_simulator-future 2022-11-18 04:03:15 UTC

Fidelity (Single Circuit)


Benchmark Results - MaxCut (2) - Qiskit
Device=ibmq_qasm_simulator-20-7 Oct 31, 2022 16:10:53 UTC
shots=5000, rounds=2, degree=3, restarts=7,
Objective Function=Approximation Ratio

Fidelity + Run Time (Iterative)

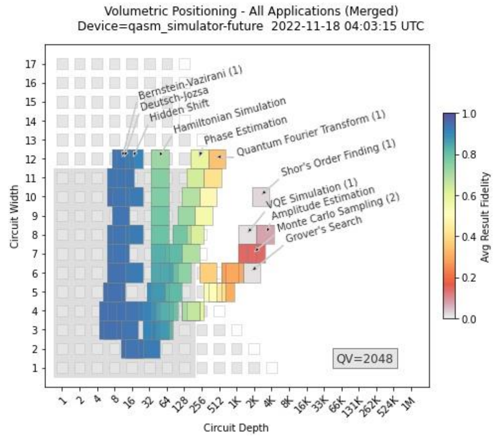
Volumetric Positioning of Application profiles

Timeline of QED-C Benchmark Efforts

Application-Oriented Performance Benchmarks

https://ieeexplore.ieee.org/document/10061574

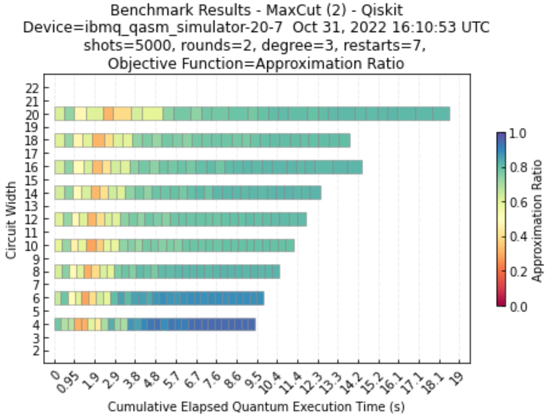
https://arxiv.org/abs/2110.03137



Optimization Applications as Quantum Benchmarks

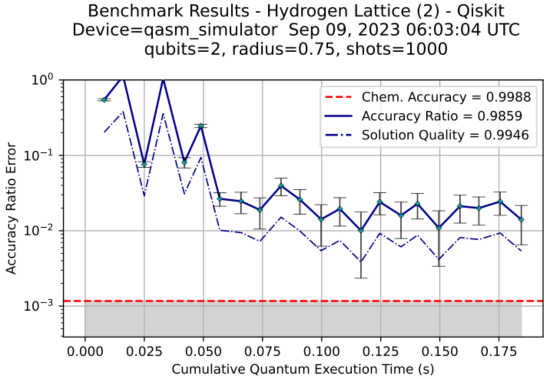
https://dl.acm.org/doi/10.1145/3678184

https://arxiv.org/abs/2302.02278



Quantum Algorithm Exploration with A.O. Benchmarks

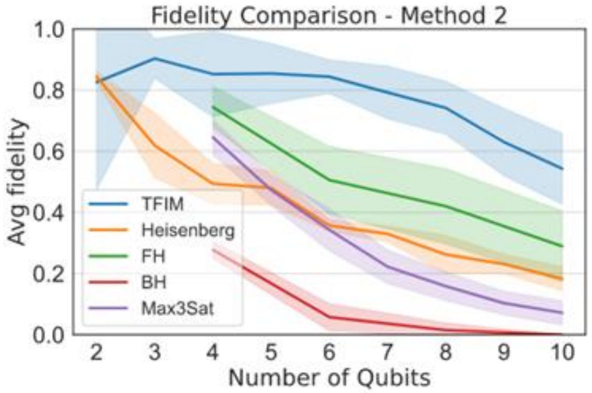
https://arxiv.org/abs/2402.08985



Benchmarking Applications of Hamiltonians using HamLib – Fidelity and Observables

https://arxiv.org/abs/2409.06919

https://arxiv.org/abs/2504.09813



Conception + Planning

2019 - 2020

2021

Algorithm Execution Fidelity and Run Time

2022

Throughput and Application Metrics

2023

Increase Coverage Exploration Features

2024 - 2025

Simulation and Applications

A small fraction of recent benchmarking papers

(sorry if I missed yours, not enough room)

[A Practical Introduction to Benchmarking and Characterization of Quantum Computers]

[A Review and Collection of Metrics and Benchmarks for Quantum Computers: definitions, methodologies and software]

[A Practical Introduction to Benchmarking and Characterization of Quantum Computers]

[How NOT to Fool the Masses When Giving Performance Results for Quantum Computers]

[Benchmarking the performance of quantum computing software]

[Evaluating the performance of quantum processing units at large width and depth]

[Benchmarking quantum computers]

[Benchmarking Quantum Computers: Towards a Standard Performance Evaluation Approach]

[Systematic benchmarking of quantum computers: status and recommendations]

[When Clifford benchmarks are sufficient; estimating application performance with scalable proxy circuits]

[QuAS: Quantum Application Score for benchmarking the utility of quantum computers]

[BenchQC -- Scalable and modular benchmarking of industrial quantum computing applications]

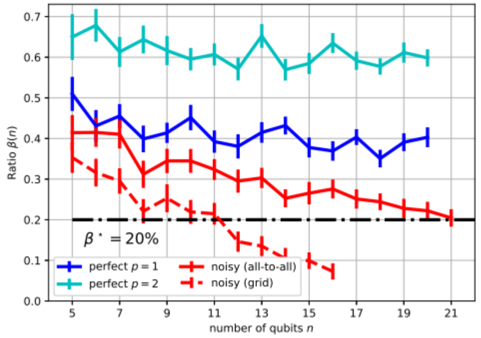
[Demonstrating Scalable Benchmarking of Quantum Computers]

[Defining Standard Strategies for Quantum Benchmarks]

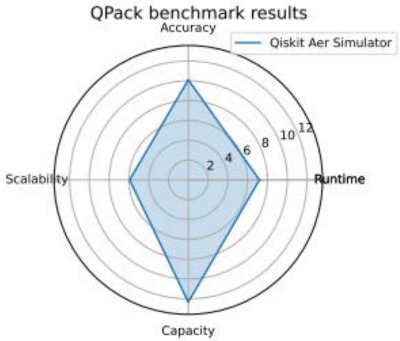
Nearly every topic covered:
methodology,
types of benchmarks,
algorithms,
good vs bad benchmarks,
gaming of results,
detailed metrics definitions,
best practices,
…

Many Emerging Application-Focused Benchmark Toolkits

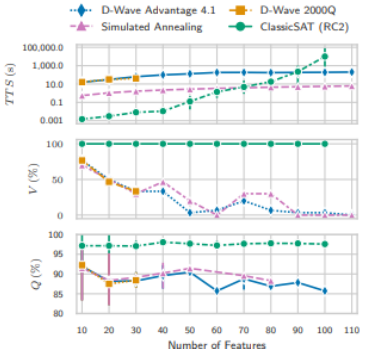
Q-score (Atos)



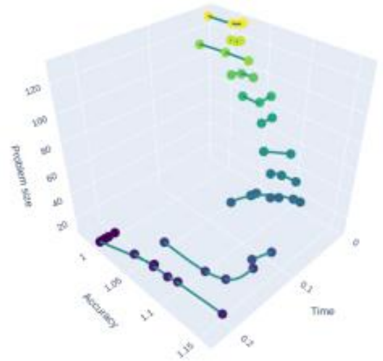
QPack-Scores (Delft)



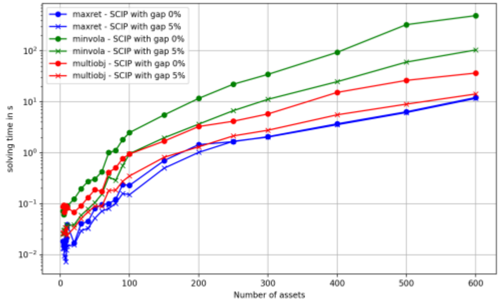
QUARK (BMW)



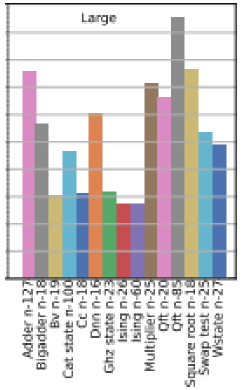
QuAS



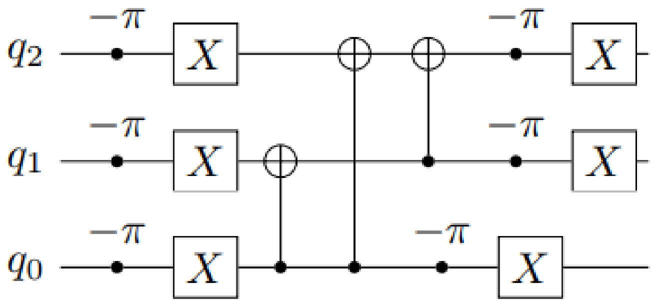
Open QBench



QASMBench (PNNL)



MQT Bench (U. Munich)



And … a fantastic set of presentations in this TQCI conference

What could I possibly add to this amazing volume of scientifically rich benchmark intelligence?

Some perspective on …

Lessons learned

Insights gained

Challenges ahead

(Note: the opinions expressed are the presenter’s alone and have not been reviewed or approved by QED-C, QED-C members, or Quantum Circuits Inc.)

It’s really important to stay focused on the why

- Here, why does not mean more of …

  get visibility, compare systems, measure progress, etc.

- The why may change over time.

- *Consider these public statements about quantum computing*:

Recently, a quantum computer ***solved a problem so complex, it would’ve taken the world’s fastest supercomputer 47 years. The quantum machine? It took seconds.*** Let that sink in. (1)

An annealing computer performs a ***magnetic materials simulation in minutes that would take a million years and more than the world’s annual electricity consumption*** to solve using a classical supercomputer. (2)

In benchmark tests, Willow ***solved a standard computation in <5 mins that would take a leading supercomputer over 10^25 years*** - beyond the age of the universe(!) (3)

(1) Rahul Mishra, Medium 2025
(2) D-Wave website, News Details page
(3) Sundar Pichai, Chief Executive Officer of Alphabet and Google, said in a post on X.

And these comments from experts in the field …

- **Scott Aaronson** -- ***it's only for a few practical problems that we know how to <use quantum computation> in a way that vastly outperforms the best known classical algorithms.***

- **Matthias Troyer** -- the ***number of applications where quantum computers could provide a meaningful advantage was more limited*** than some might have you believe.

- **BCG Forecast for Quantum Computing -- *Quantum computing today provides no tangible advantage over classical computing in either commercial or scientific applications***

https://scottaaronson.blog/?p=8329

https://spectrum.ieee.org/quantum-computing-skeptics#:~:text=The%20main%20promise%20of%20quantum,exactly%20how%20much%20faster%20varies

https://www.bcg.com/publications/2024/long-term-forecast-for-quantum-computing-still-looks-bright

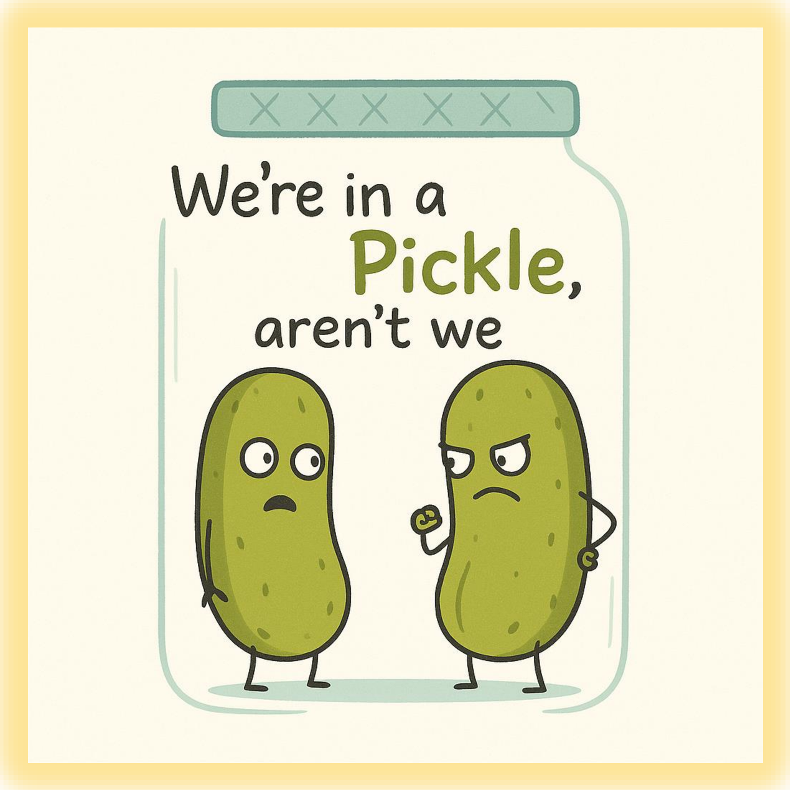
Uh-oh

On s’est mis dans de beaux draps …

… we have landed on a sticky wicket

… we have gotten ourselves into a bit of a pickle

→ Huge expectations, big timeline and technical challenges, and a new (self-inflicted) problem to solve!



How do we get out of this “Pickle” ?

Re-establish trust and credibility in user community

Cast these incredible results as special cases that are “beacons” to the future potential

Re-direct attention … focus on what we have, i.e.

Highlight excitement about the steady and accelerating growth in what we can do now

Be forthright and transparent about the facts

We need reliable, repeatable, ***accessible***, and ***digestible*** benchmarking

We have reliable benchmarking

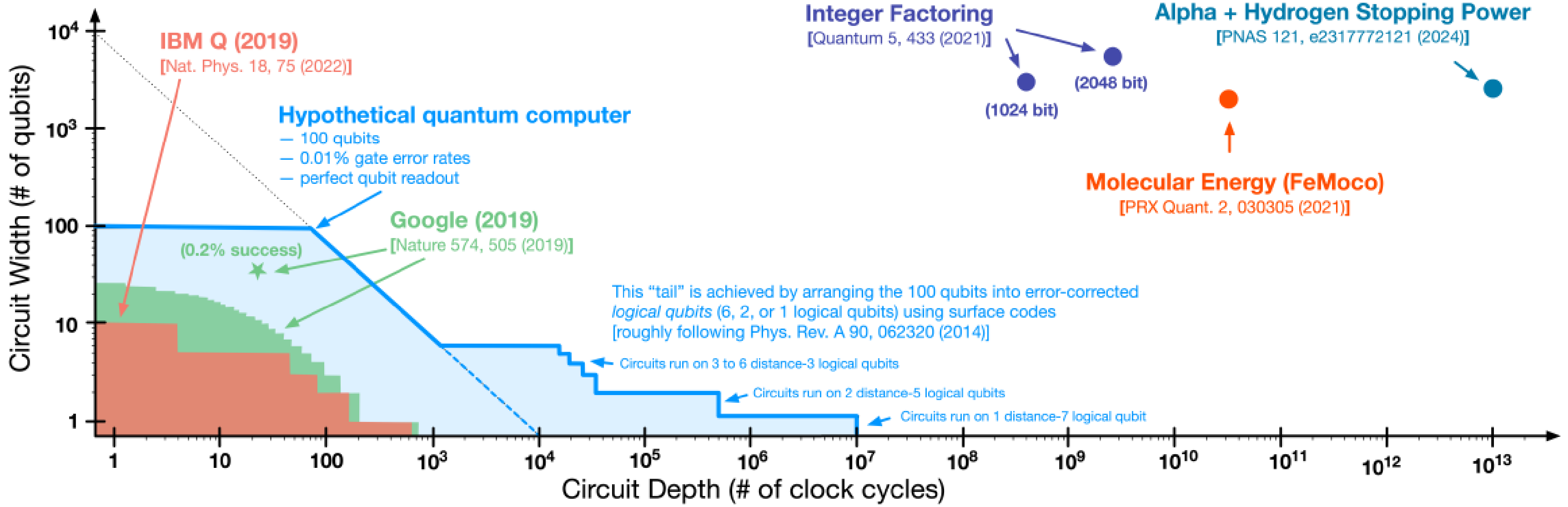
We have all the information we need

- We have the data, we need to work on the presentation

Benchmarking Quantum Computers – Proctor, *et al.*

<Component/System>

Sandia Quantum Performance Laboratory

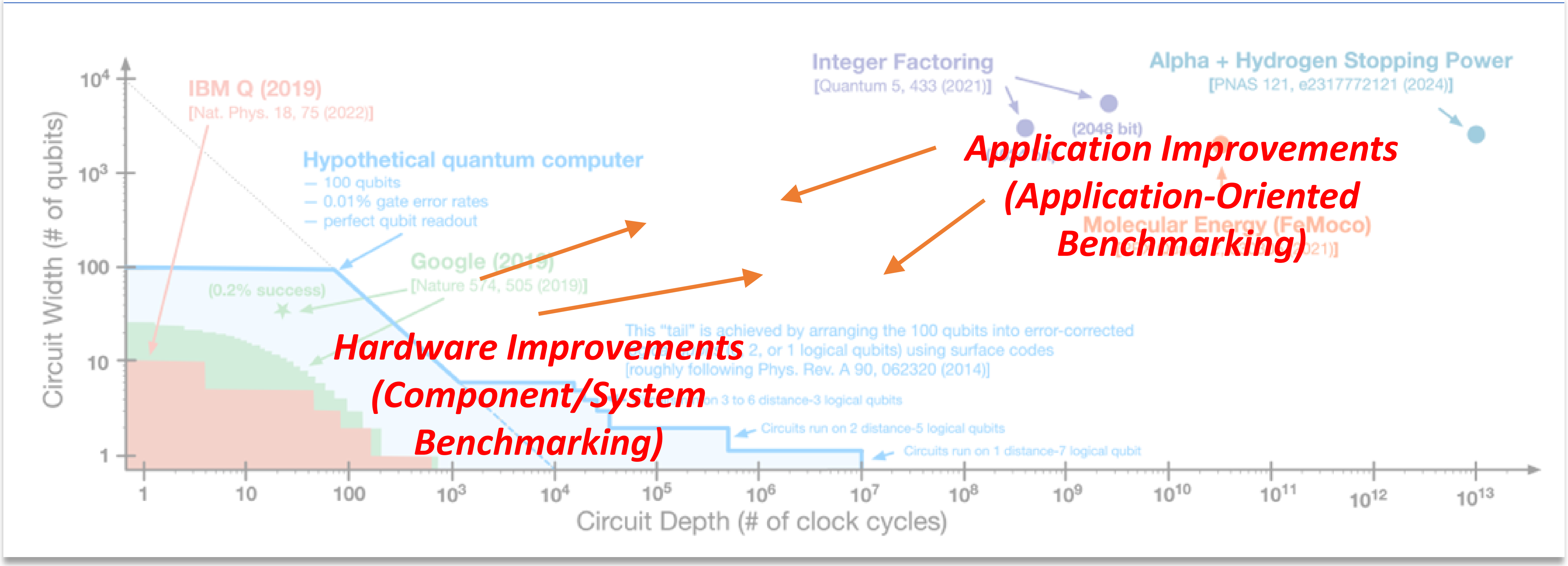


Here it looks like we are far far away

Figure 4. Assessing quantum computer performance via capability. This figure illustrates one way to compare experimentally benchmarked performance against resource estimates for challenge problems, using a multidimensional *capability* metric. Challenge problems and benchmark tasks are represented by the width (some measure of the number of qubits) and depth (some measure of the number of clock cycles) of a quantum circuit that performs the task. Regions indicate the circuits performable by two real-world quantum computers—Google's Sycamore (green) as extrapolated from results in Arute *et al.*[2], and an ensemble of IBM Q devices (pink) benchmarked by our group[37]—and one hypothetical quantum computer (blue) (we use a success threshold of $1/e$). Points indicate constant-factor resource estimates for three candidate challenge problems analyzed in the literature[7–9]. For these problems, width is the number of logical qubits, not accounting for logical qubits used in distillation or routing, and depth is the total number of non-Clifford operations (i.e., Toffoli and/or T gates). These metrics are somewhat crude, but indicate the rough scale of resources required for these challenge problems. We emphasize the wide gulf between that "utility" scale and current state of the art capabilities—logarithmic axes were required to compress both scales into one figure. Plots like this one could enable stakeholders to track and extrapolate the growth of quantum computer capabilities over time, toward eventual achievement of quantum utility.

Benchmarking Quantum Computers - https://arxiv.org/pdf/2407.08828

We can overlay additional information onto this plot



Application Improvements (Application-Oriented Benchmarking)

Hardware Improvements (Component/System Benchmarking)

Here it looks like we might be making progress!

Highlight what we want users to take away

We can’t be just talking amongst ourselves, we need to communicate coherently and uniformly to the outside world.

Slow the uncontrolled proliferation of new projects and github repositories. Collaborate more.

Perhaps we need a “task force” to determine what needs to be done to steer around the iceberg.

How can we communicate about or progress effectively to the necessary audience?

Captain! There’s an iceberg ahead!

Benchmarking quantum computers is “quantum complicated”

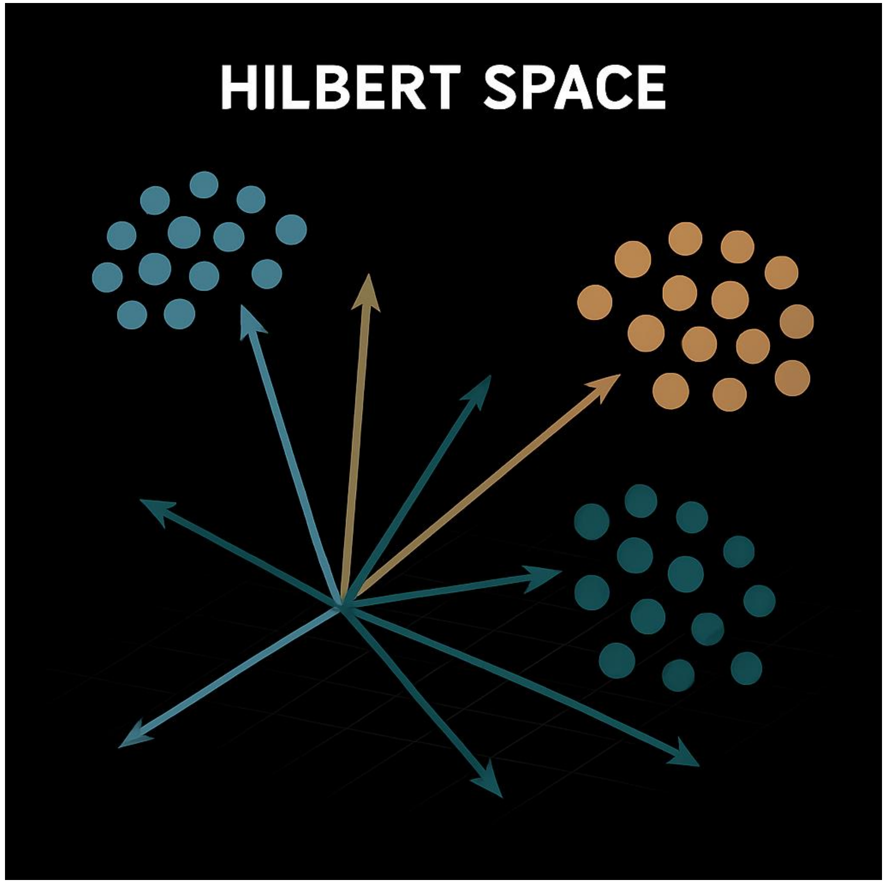
- Classical benchmarking measures speed and resources primarily
- Quantum involves speed, resources, quality, noise properties, hybrid cost, and more

Benchmarking Quantum Computers is like a gigantic Hilbert Space

Many measurement axes

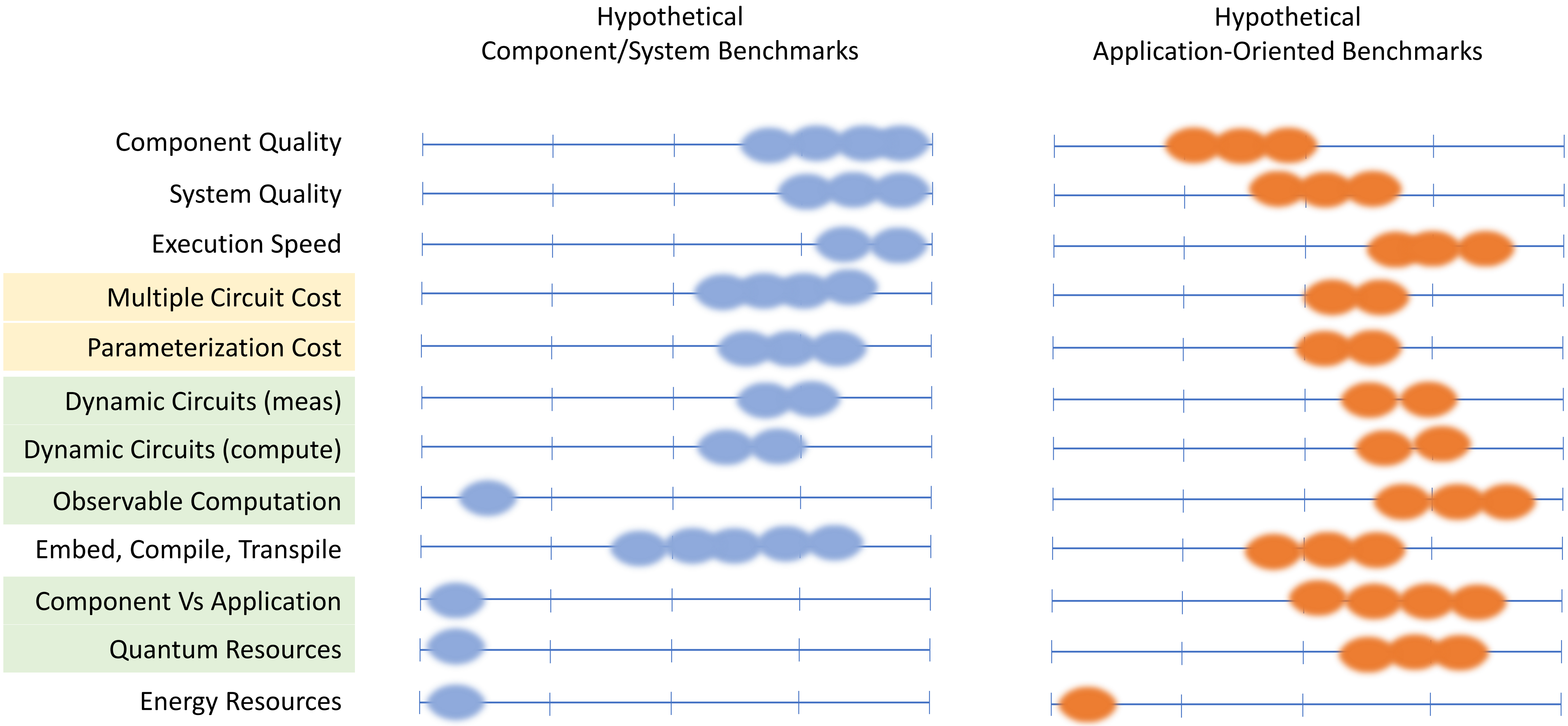
Clusters of benchmark features target specific areas of performance

Overlap in the clusters indicates multiple ways available to benchmark an area



*** Need Clarity about Component-, System-, Application-Level Benchmarks

Many Dimensions to Benchmark Performance by Feature Spectrum



Hypothetical
Component/System Benchmarks

Hypothetical
Application-Oriented Benchmarks

Component Quality

System Quality

Execution Speed

Multiple Circuit Cost

Parameterization Cost

Dynamic Circuits (meas)

Dynamic Circuits (compute)

Observable Computation

Embed, Compile, Transpile

Component Vs Application

Quantum Resources

Energy Resources

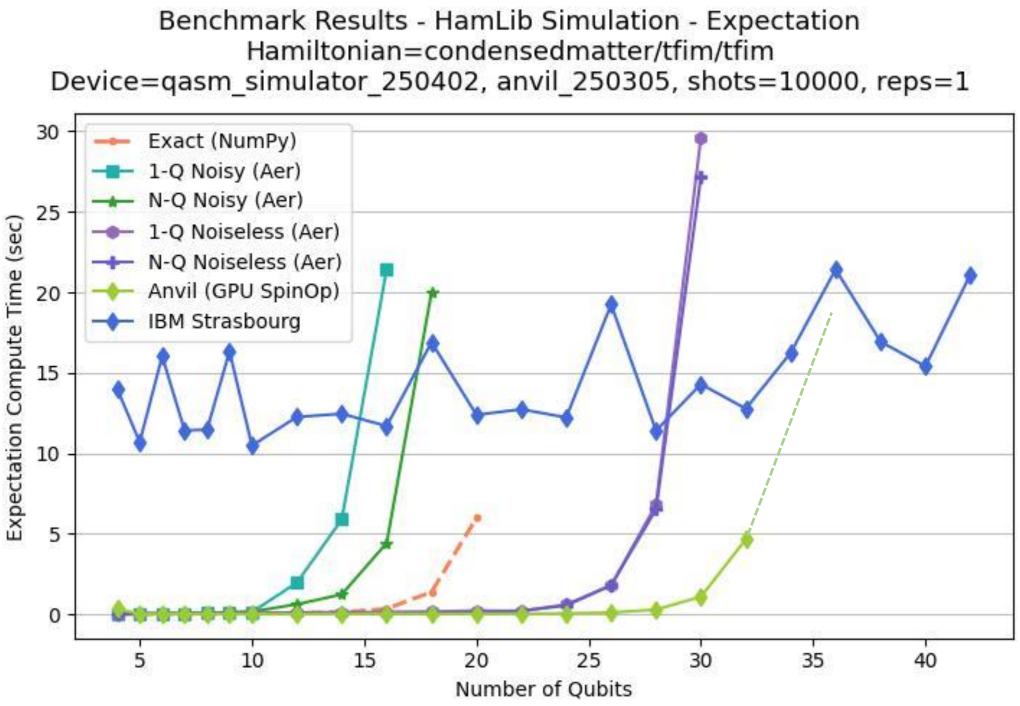
Significant Confusion Results

So many different benchmarks are surfaced and one says they are better than the other, how can users know? Let’s send a coherent message.

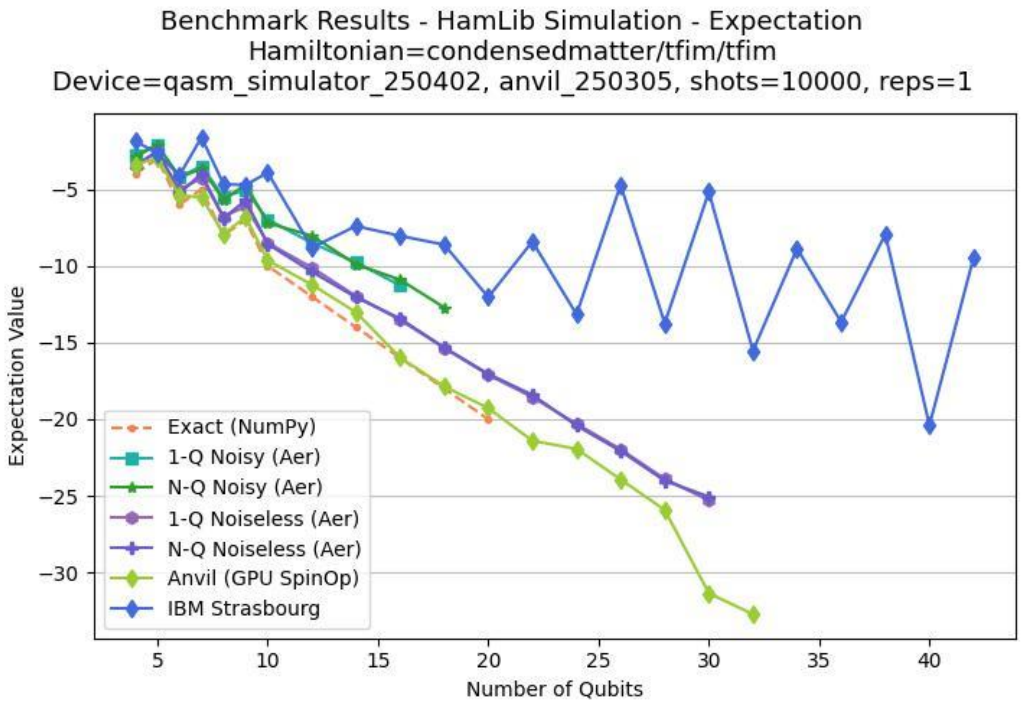
Standards TAC contributed to the confusion by leaving users to assume app-oriented benchmarks assess hardware performance accurately.

Scientific papers can provide rigorous analysis, but the upshot is difficult to digest without boiling it down to simple “sound bits”

Benchmarking Observable Computation Key to Many Algorithms



Benchmark Results - HamLib Simulation - Expectation
Hamiltonian=condensedmatter/tfim/tfim
Device=qasm_simulator_250402, anvil_250305, shots=10000, reps=1



Benchmark Results - HamLib Simulation - Expectation
Hamiltonian=condensedmatter/tfim/tfim
Device=qasm_simulator_250402, anvil_250305, shots=10000, reps=1

Execution Time

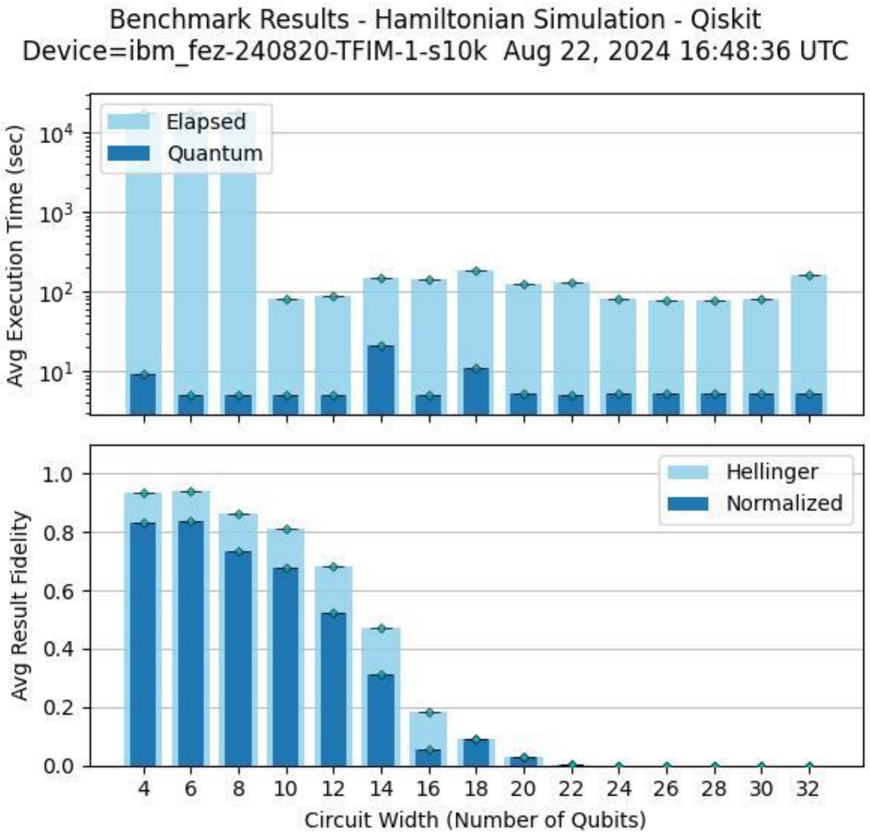
Quality of Result

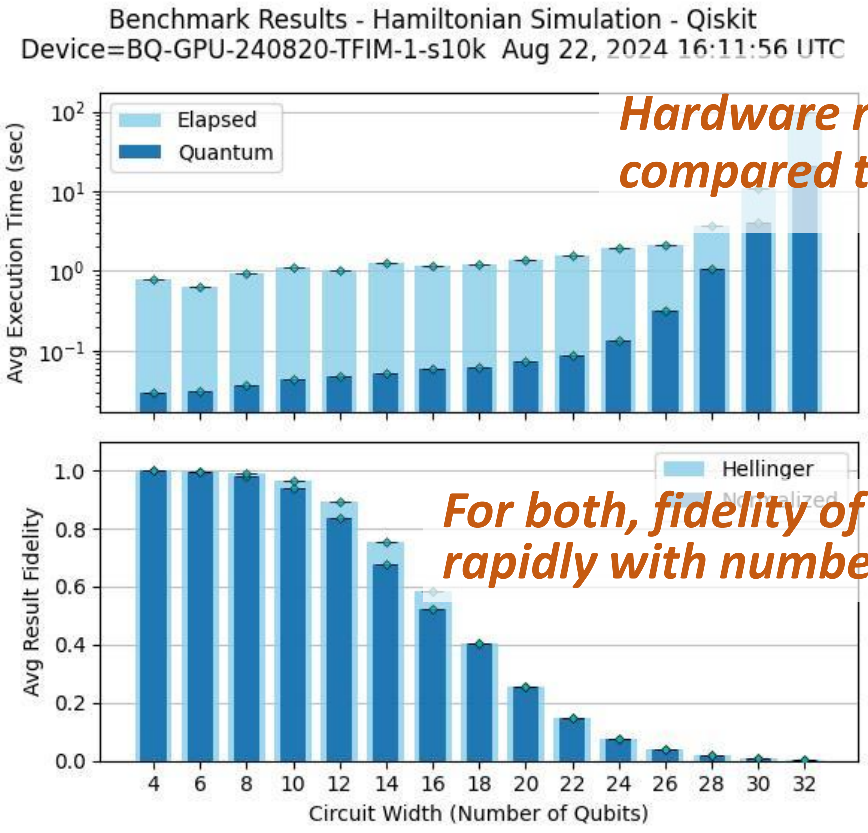
Exploring methods for computing observables – can parallelization help?

https://arxiv.org/abs/2504.09813

The Impact of Fault-Tolerant Quantum Computation

Compare Execution Time and Fidelity - IBM Fez and BlueQubit GPU Sim





Hardware run time is flat compared to GPU simulation

For both, fidelity of execution degrades rapidly with number of qubits

→ What does this suggest about Fault-Tolerant Quantum Computing?

https://arxiv.org/abs/2409.06919

Fault-Tolerant Quantum Computing may not be a panacea

- How well does an ideal GPU simulation mimic FTQC?
  - If algorithms don’t perform well in ideal simulation, how can FTQC do better?

- FTQC uses many more gates, larger exec times
  - E.g. > 50 X more gates for QEC

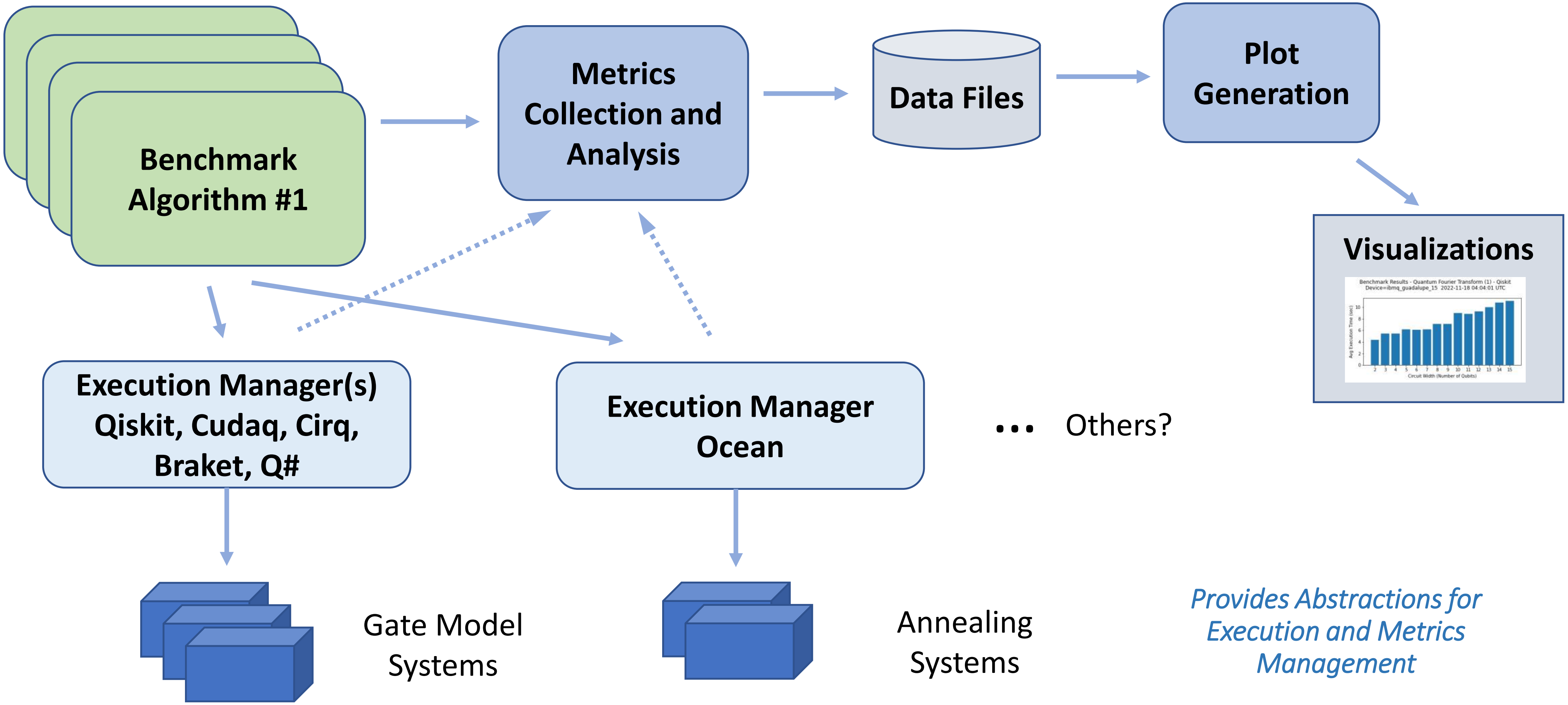
- Error Mitigation techniques will not be important with FTQC

→ ***We MUST improve algorithms in parallel while we improve the hardware***

Collaborate and Consolidate and Focus

- Naïve to think we will all link hands and create utopia
- Perhaps there are consolidation points that can be identified.
- Several of the projects we heard about these 2 days could complement and enhance one another.

QED-C Benchmarking Framework – Architecture



Benchmark Algorithm #1

Metrics Collection and Analysis

Data Files

Plot Generation

Visualizations

Benchmark Results - Quantum Fourier Transform (1) - Qiskit
Device=ibmq_guadalupe_15 2022-11-18 04:04:01 UTC
Avg Execution Time (sec)
0
2
4
6
8
10
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Circuit Width (Number of Qubits)

Execution Manager(s) Qiskit, Cudaq, Cirq, Braket, Q#

Execution Manager Ocean

... Others?

Gate Model Systems

Annealing Systems

Provides Abstractions for Execution and Metrics Management

QED-C Benchmarks are Highly Promiscuous

- Support for Qiskit, CUDAQ, D-Wave, and other API layers
- Integration with HamLib as source of Hamiltonians
- Recent work with NVIDIA to integrate CUDAQ and MPI for multi-GPU
- Current collaboration with Unitary Fund
  - Execution Management via Metriq-gym
  - Metrics Reporting via Metriq-info (TBD)
- Current collaboration with Sandia QPL
  - Integrate QED-C benchmark circuit library with Sandia pyGSTi libraries to enable scalable benchmarks using advanced mirror circuit and subcircuit volumetric benchmarking methods.
- … open to collaboration with other projects

What’s important isn’t always “easy”

From the architects of the IBM System/360 — Gene Amdahl, Fred Brooks, and Gerrit Blaauw in 1964 …

> The real value of an information system is properly measured by answers-per-month, not bits-per-microsecond.

- We can easily benchmark low-level performance—speed, fidelity, and noise.
  (difficult to do it well, of course)
- Measuring application-level usefulness—real answers—is much harder.
- The challenge: link low-level progress to real, potential high-level value.
- Focus on the progress, something we can actually demonstrate.

Amdahl, Gene & Blaauw, Gerrit & Brooks, Jr, Frederick. (2000). Architecture of the IBM System/360. IBM Journal of Research and Development. 44. 21-36. 10.1147/rd.82.0087.

The End

Questions and Comments to Follow