Performance modelling facing disruptive technologies on the horizon

Gerhard Wellein

Erlangen National High Performance Computing Center (NHR@FAU)

Department of Computer Science
Friedrich-Alexander-Universität Erlangen-Nürnberg

Georg Hager
Christie L. Alappat
Florian Lange

Agenda

1. Analytical, resource-based, first-principles performance models – where we are

2. Disruptive technologies on the horizon: Quantum…..

Disruptive, innovative, revolutionary,….

https://en.wikipedia.org/wiki/Disruptive_innovation:

In business theory, **disruptive innovation** is innovation that creates a new market and value network or enters at the bottom of an existing market and eventually displaces established market-leading firms, products, and alliances.[1] The concept was developed by the American academic Clayton Christensen and his collaborators beginning in 1995,[2][*full citation needed*]

CRAY Vector

Attack of the Killer-Micros

CRAY MPPs

COTS-Clusters

GPUs, GPGPUs, AI-GPGPUs

Analytical, resource-based, first-principles performance models – where we are

Analytical, Resource-Based, First-Principles Performance Model?

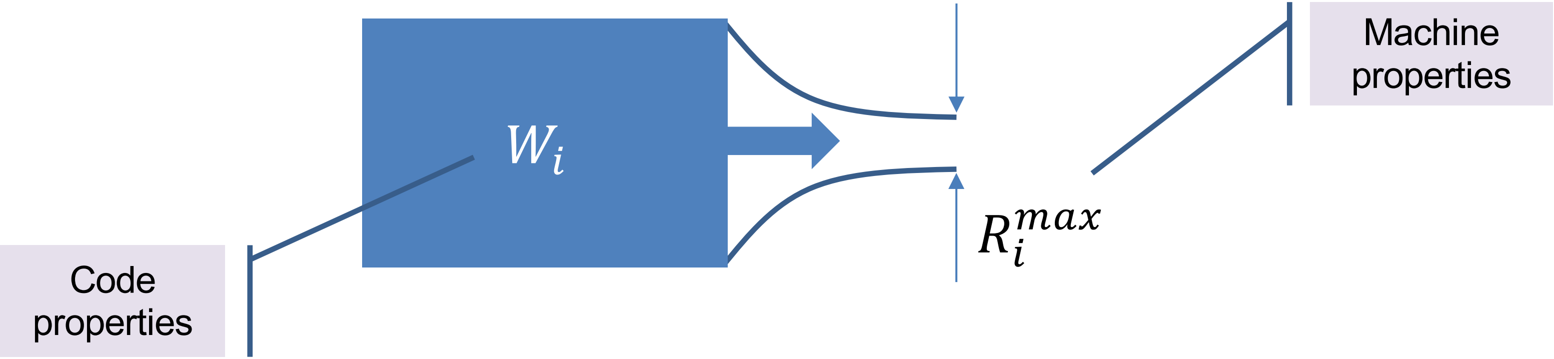
a.k.a. white-box models

A mathematical representation of hardware-software interaction based on simplified machine and application models, which predicts the performance or runtime of a program using hardware resource limits and code requirements

Resource bottlenecks

What is the maximum performance when limited by a bottleneck?

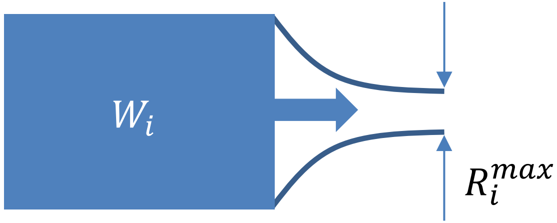
- Resource bottleneck $\boldsymbol{i}$ delivers resources at maximum rate $\boldsymbol{R_i^{max}}$
- $\boldsymbol{W_i}$ = needed amount of resources (Instructions, FLOPs, Data Volume,…)



Resource bottlenecks

Minimum runtime due to bottleneck i:

$$T_i = \frac{W_i}{R_i^{max}} + \lambda_i$$



- Multiple bottlenecks?
  → multiple minimum runtimes: $T_{\min} = f(T_1, \ldots T_n)$

🤔

- Overall performance:

$$P_{\max} = \frac{W}{T_{\min}}$$

Simple two-bottleneck models for single loops

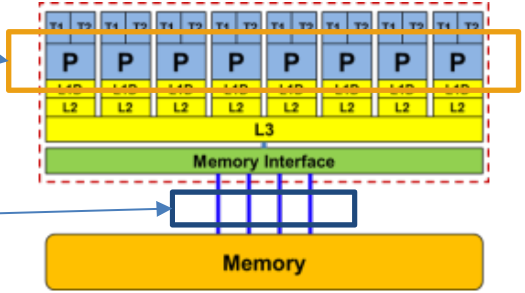
```
#pragma omp parallel for
for(i=0; i<10^7; ++i)
  a[i] = a[i] + s * c[i];
```

$$W_{flops} = 2\times10^{7}\ \text{flops}$$

$$W_{BW} = 3\times8\times10^{7}\ \text{bytes}$$

$$R_{flops}^{max} = 192\,\frac{\text{Gflops}}{\text{s}}$$

$$R_{BW}^{max} = 40\,\frac{\text{Gbyte}}{\text{s}}$$



8-core CPU
(3 GHz Intel Sandy Bridge)

$$T_{flops} = \frac{2\times10^{7}\ \text{flops}}{192\,\frac{\text{Gflops}}{\text{s}}} = 104\ \mu s$$

$$T_{BW} = \frac{2.4\times10^{8}\ \text{bytes}}{40\,\frac{\text{Gbyte}}{\text{s}}} = 6.0\ \text{ms}$$

Bottleneck models for single loops

How do we reconcile the multiple bottlenecks?

I.e., what is the functional form of $f(T_1, \dots T_n)$?

→ pessimistic (no overlap): $f(T_1, \dots T_n) = \sum_i \; T_i$

→ optimistic (full overlap): $f(T_1, \dots T_n) = \max(T_1, \dots T_n)$

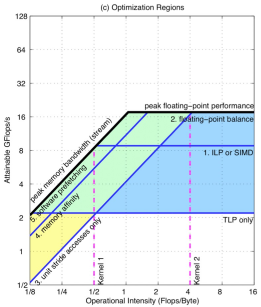
Roofline model
(Williams et al., 2008)
$P = \min(P_{peak}, I * b_S)$

Our example (two bottlenecks): $T_{\min} = \max\left(T_{flops}, T_{BW}\right) = 6\text{ ms}$

Maximum performance ("light speed"): $P_{upper} = \frac{2\times10^{7}}{6.0\times10^{-3}} \frac{\text{flops}}{\text{s}} = 3.3\ \text{Gflop/s}$

Analytic modelling – where we are: Examples

Roofline Model

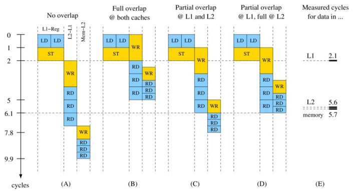


S. Williams, A. Waterman, D. Patterson (2009)
DOI:10.1145/1498765.1498785

Energy:
J. W. Choi, D. Bedard, R. Fowler, R. Vuduc
(2013) DOI: 10.1109/IPDPS.2013.77.

Cache-Aware:
A. Ilic, F. Pratas, L. Sousa (2014)
DOI: 10.1109/L-CA.2013.6.

Execution Cache Memory Model


No overlap
Full overlap @ both caches
Partial overlap @ L1 and L2
Partial overlap @ L1, full @ L2
Measured cycles for data in ...
L1−Reg
L2−L1
Mem−L2
LD
ST
WR
RD
0
1
2
5
6.1
7.8
9.9
cycles
(A)
(B)
(C)
(D)
(E)
L1 2.1
L2 5.6
memory 5.7

Hager, Treibig, Habich, Wellein (2016)
DOI: 10.1002/cpe.3180.

Power/Energy:
Hofmann, Hager, Fey (2018).
https://doi.org/10.1007/978-3-319-92040-5_2

$$\frac{W}{f(T_1, \dots, T_n)}$$

Communication models

LogP and variants

Proven/useful for

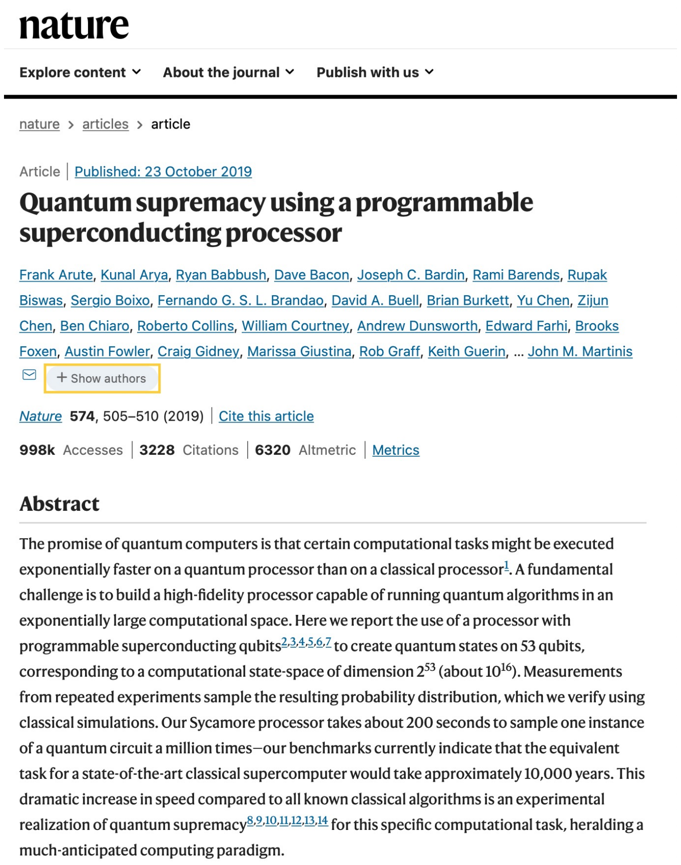
- CPU-type
- GPU-type
- Vector-type

Data + Flops/Instructions – Throughputs / Latencies

Disruptive technologies on the horizon: Quantum…..

Florian Lange

The Quantum hype



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 23 October 2019

Quantum supremacy using a programmable superconducting processor

Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C. Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando G. S. L. Brandao, David A. Buell, Brian Burkett, Yu Chen, Zijun Chen, Ben Chiaro, Roberto Collins, William Courtney, Andrew Dunsworth, Edward Farhi, Brooks Foxen, Austin Fowler, Craig Gidney, Marissa Giustina, Rob Graff, Keith Guerin, ... John M. Martinis

+ Show authors

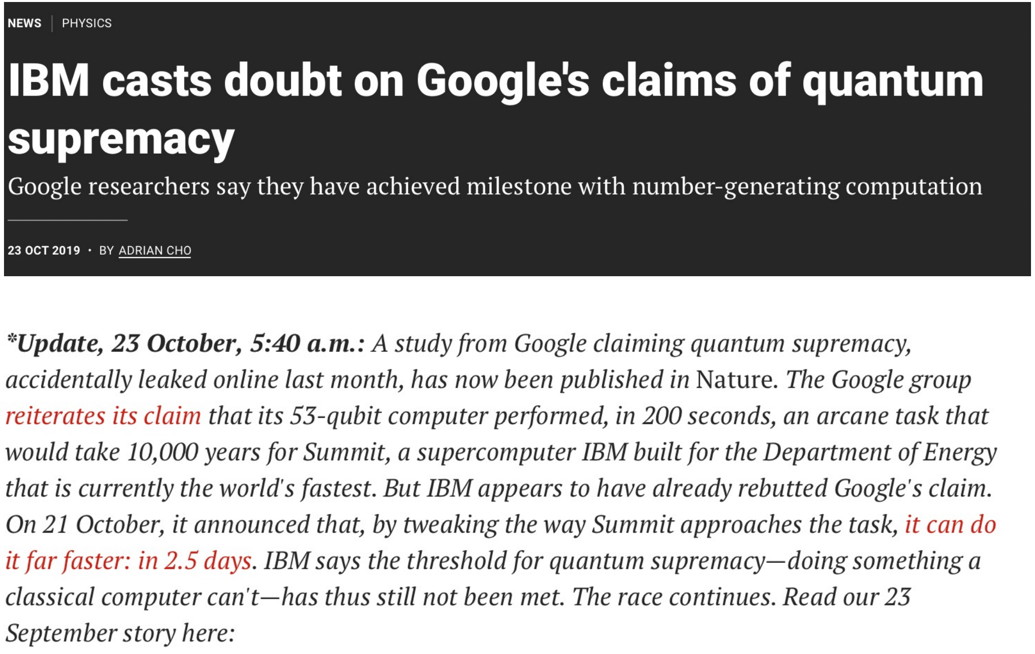
Nature **574**, 505–510 (2019) | Cite this article

998k Accesses | **3228** Citations | **6320** Altmetric | Metrics

Abstract

The promise of quantum computers is that certain computational tasks might be executed exponentially faster on a quantum processor than on a classical processor[1]. A fundamental challenge is to build a high-fidelity processor capable of running quantum algorithms in an exponentially large computational space. Here we report the use of a processor with programmable superconducting qubits[2,3,4,5,6,7] to create quantum states on 53 qubits, corresponding to a computational state-space of dimension 2^{53} (about 10^{16}). Measurements from repeated experiments sample the resulting probability distribution, which we verify using classical simulations. Our Sycamore processor takes about 200 seconds to sample one instance of a quantum circuit a million times—our benchmarks currently indicate that the equivalent task for a state-of-the-art classical supercomputer would take approximately 10,000 years. This dramatic increase in speed compared to all known classical algorithms is an experimental realization of quantum supremacy[8,9,10,11,12,13,14] for this specific computational task, heralding a much-anticipated computing paradigm.

ww.science.org



NEWS | PHYSICS

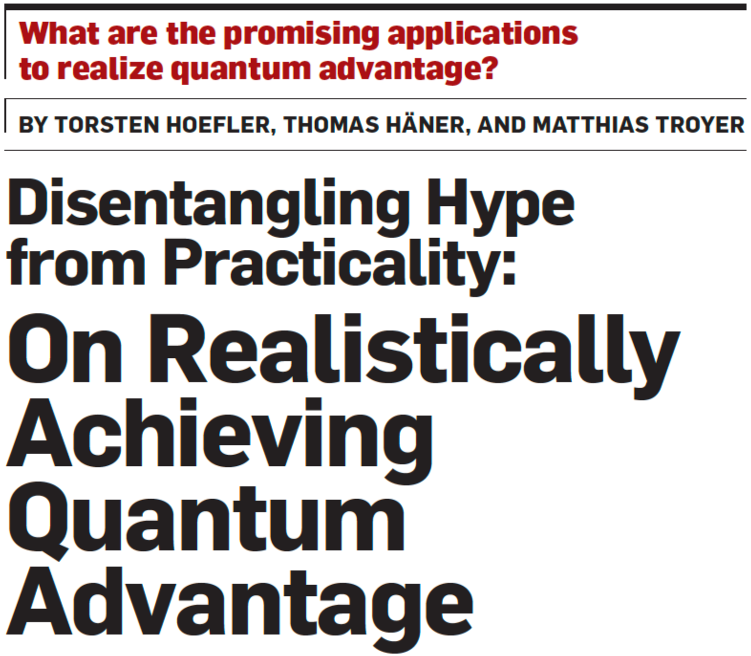
IBM casts doubt on Google's claims of quantum supremacy

Google researchers say they have achieved milestone with number-generating computation

23 OCT 2019 • BY ADRIAN CHO

***Update, 23 October, 5:40 a.m.:** A study from Google claiming quantum supremacy, accidentally leaked online last month, has now been published in* Nature. *The Google group reiterates its claim that its 53-qubit computer performed, in 200 seconds, an arcane task that would take 10,000 years for Summit, a supercomputer IBM built for the Department of Energy that is currently the world's fastest. But IBM appears to have already rebutted Google's claim. On 21 October, it announced that, by tweaking the way Summit approaches the task, it can do it far faster: in 2.5 days. IBM says the threshold for quantum supremacy—doing something a classical computer can't—has thus still not been met. The race continues. Read our 23 September story here:*

FP arithmetics and data movement ←→ Quantum computer



What are the promising applications to realize quantum advantage?

BY TORSTEN HOEFLER, THOMAS HÄNER, AND MATTHIAS TROYER

Disentangling Hype from Practicality:

On Realistically Achieving Quantum Advantage

Favorable assumptions on quantum technology:

- Cost of FPxy computations excessively high → no FP calculations
- Data transfer costs 10000x higher → very restricted data transfer

Quantum advantage ←→ Complexity classes

Communications of the ACM, May 2023, Vol. 66 No. 5, Pages 82-87. DOI:10.1145/3571725

Performance Modelling of QC

The easy / nice answer:

- Quantum Computer is an accelerator to an HPC system (which becomes “serial“ part)
- Assume some quantum quantum acceleration factor (x_{QA}) and use Amdahl‘s law

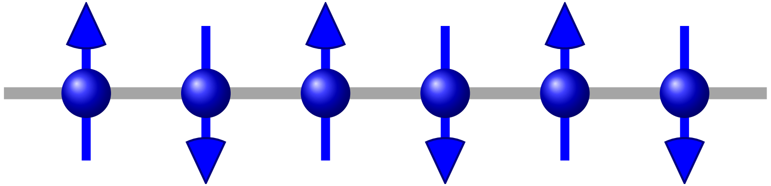
$$S_P(QC) = \frac{T_{HPC}}{T_{HPC+QC}} = \frac{1}{s + \frac{1-s}{x_{QA}}}$$

Quantum Simulations

- Use quantum computers to simulate quantum systems

- Spin (1/2) systems ←→ Qbits

$|\dots 010101 \dots\rangle$ ←



antiferromagnetic state

- General state: $\mathbf{2^N}$ degrees of freedom / base states

Exponential Complexity

$$|\psi\rangle = \sum_{n\in\{0,1\}^N} \psi_n |n_1 n_2 \dots n_N\rangle, \qquad \psi_n \in \mathbb{C}$$

Quantum Simulations: Time Evolution – XXZ model

- Hamiltonian of XXZ model:

$$H = \sum_{j=1}^{N-1} \left[\frac{1}{2}\left(S_j^+ S_{j+1}^- + S_j^- S_{j+1}^+\right) + \Delta S_j^z S_{j+1}^z\right],$$

$$\Delta \in \mathbb{R}$$

$$S_j^+ |\ldots 0 \ldots\rangle = |\ldots 1 \ldots\rangle$$
$$S_j^- |\ldots 1 \ldots\rangle = |\ldots 0 \ldots\rangle$$
$$S_j^z \left|\ldots n_j \ldots\right\rangle = \frac{1}{2}(-1)^{n_j} \left|\ldots n_j \ldots\right\rangle$$

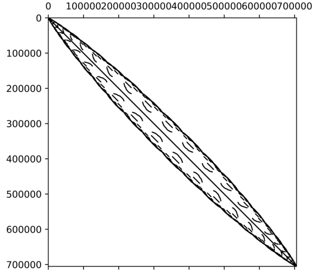
- time-evolution operator solves time dependent Schrödinger equation:

$$|\psi(t)\rangle = U(t)|\psi(0)\rangle\,, \qquad U(t) = e^{-itH} \approx \sum_{k=0}^{m} a_k H^k\,,\ a_k \in \mathbb{C}$$

Quantum Simulations: Time Evolution – Classical Approach

- Hamiltonian is mapped to (large) sparse matrix

$$H = \sum_{j=1}^{N-1} \left[\frac{1}{2} \left(S_j^+ S_{j+1}^- + S_j^- S_{j+1}^+ \right) + \Delta S_j^z S_{j+1}^z \right]$$



- Compute time evolution by calculating sparse matrix polynomials

$$y = \sum_{k=0}^{m} a_k H^k x$$

$|\psi(t)\rangle$

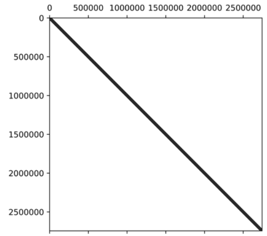
$|\psi(0)\rangle$

Sparse Matrix-Vector Multiplication (SpMV)

- Performance Modelling & Cache Blocking

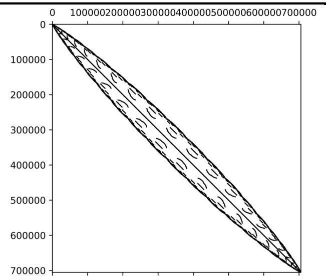
e.g., Alapatt et al., DOI: 10.1109/TPDS.2022.3223512

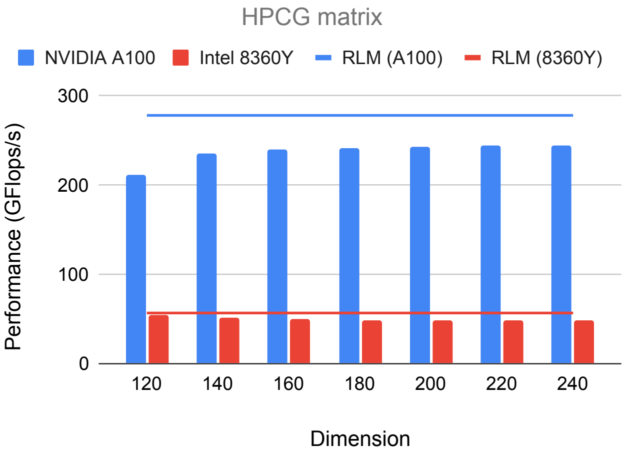
Quantum simulations: time evolution – classical approach

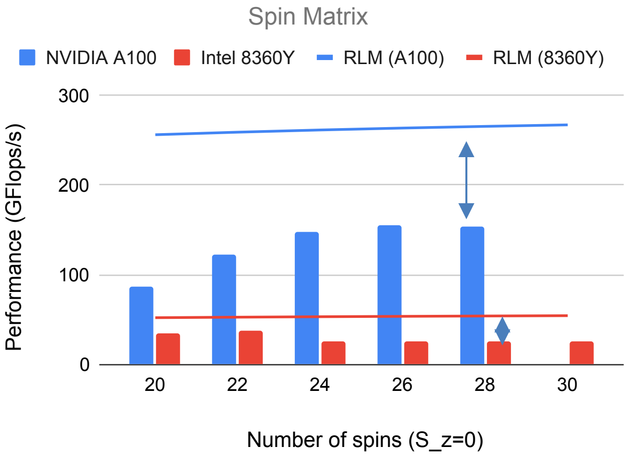


Upper performance bound (RLM) for SpMV

$$P = b_S/B_C \ with \ B_C = \left(6 + \frac{14}{N_{nzr}}\right)\frac{B}{F}$$







Strong irregular vector access!

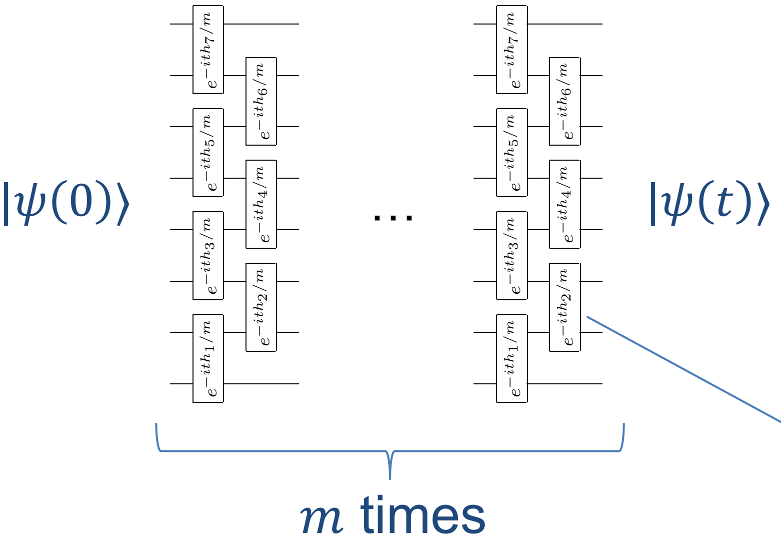
Exponential Complexity

Simple quantum algorithm for time evolution

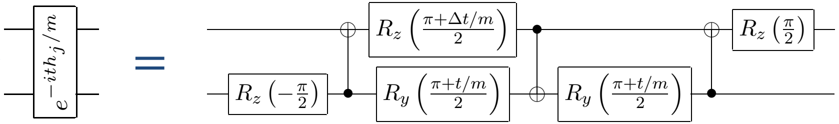
- Prepare initial state $|\psi(0)\rangle$ on the quantum computer (may not be trivial)
- Decompose time-evolution operator into 2-qubit gates via Suzuki-Trotter

$$e^{-itH} = \left[\left(\prod_{j=2,4,\dots} e^{-ith_j/m}\right)\left(\prod_{j=1,3,\dots} e^{-ith_j/m}\right)\right]^m + O(t^2/m)$$

$$h_j = \frac{1}{2}\left(S_j^+ S_{j+1}^- + S_j^- S_{j+1}^+\right) + \Delta S_j^z S_{j+1}^z$$



Each $e^{-ith_j/m}$ is split into a product of elementary gates available on quantum hardware, e.g.:

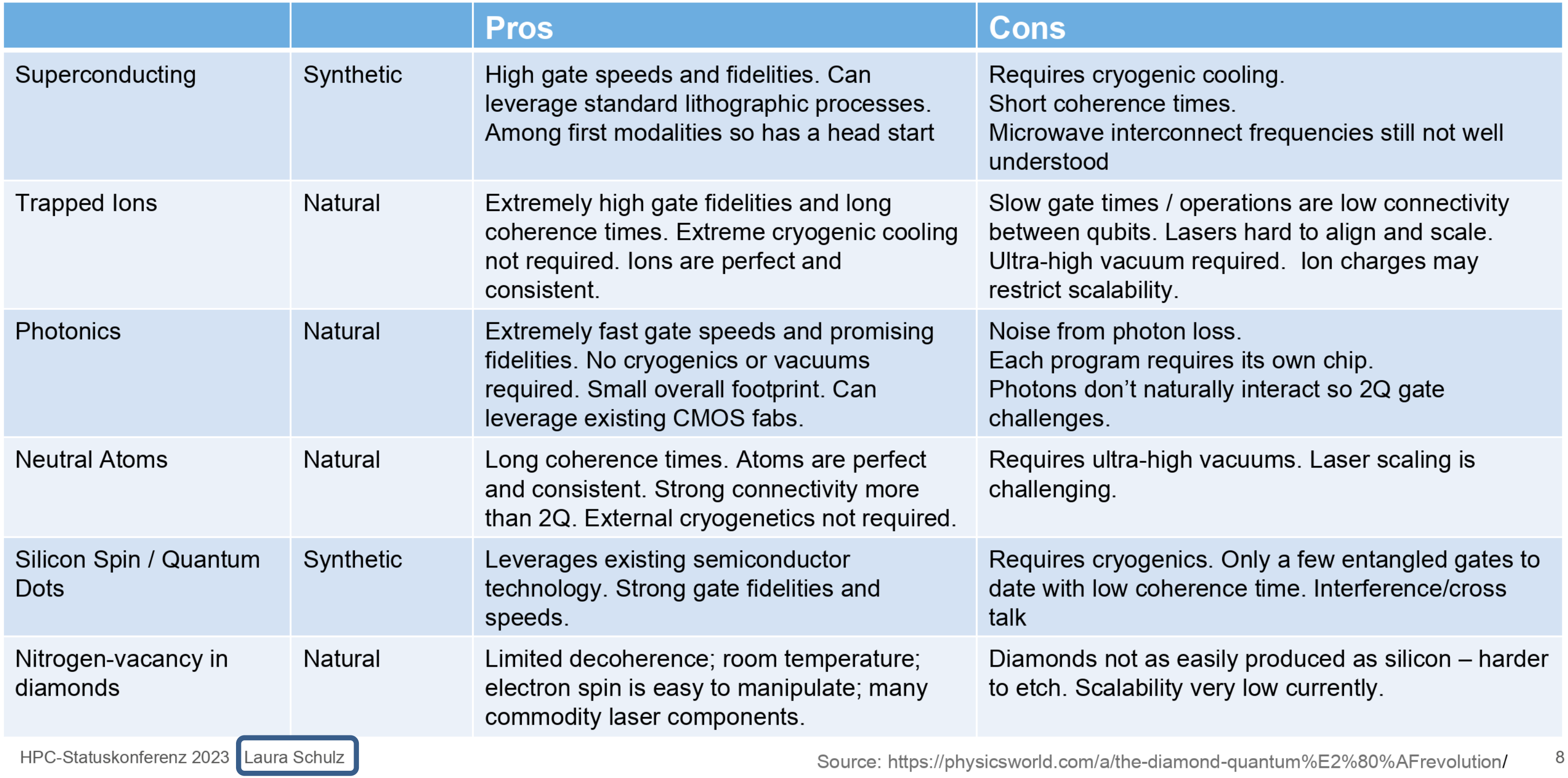


$e^{-ith_j/m}$
$=$
$R_z\left(-\frac{\pi}{2}\right)$
$R_z\left(\frac{\pi+\Delta t/m}{2}\right)$
$R_y\left(\frac{\pi+t/m}{2}\right)$
$R_y\left(\frac{\pi+t/m}{2}\right)$
$R_z\left(\frac{\pi}{2}\right)$

Quantum speedup

- For general lattice Hamiltonians: Time evolution up to an error ϵ can be simulated using $O(Nt \operatorname{polylog}(Nt/\epsilon))$ gates
  → exponential speedup compared to known classical methods
  J. Haah et al., SIAM J. Comput. SPECIAL SECTION FOCS (2018)
- Gate errors and noise limit the reachable time scales on current hardware
- Error-mitigation techniques may help enable a quantum advantage on near-term quantum computers
  - Heisenberg chain
    M. Urbanek et al., PRL 127, 270502 (2021)
  - transverse-field Ising model $(N = 127)$
    Y. Kim et al., Nature 618, 500–505 (2023)

Which Quantum Computer to use?



		Pros	Cons
Superconducting	Synthetic	High gate speeds and fidelities. Can leverage standard lithographic processes. Among first modalities so has a head start	Requires cryogenic cooling. Short coherence times. Microwave interconnect frequencies still not well understood
Trapped Ions	Natural	Extremely high gate fidelities and long coherence times. Extreme cryogenic cooling not required. Ions are perfect and consistent.	Slow gate times / operations are low connectivity between qubits. Lasers hard to align and scale. Ultra-high vacuum required. Ion charges may restrict scalability.
Photonics	Natural	Extremely fast gate speeds and promising fidelities. No cryogenics or vacuums required. Small overall footprint. Can leverage existing CMOS fabs.	Noise from photon loss. Each program requires its own chip. Photons don’t naturally interact so 2Q gate challenges.
Neutral Atoms	Natural	Long coherence times. Atoms are perfect and consistent. Strong connectivity more than 2Q. External cryogenetics not required.	Requires ultra-high vacuums. Laser scaling is challenging.
Silicon Spin / Quantum Dots	Synthetic	Leverages existing semiconductor technology. Strong gate fidelities and speeds.	Requires cryogenics. Only a few entangled gates to date with low coherence time. Interference/cross talk
Nitrogen-vacancy in diamonds	Natural	Limited decoherence; room temperature; electron spin is easy to manipulate; many commodity laser components.	Diamonds not as easily produced as silicon – harder to etch. Scalability very low currently.

HPC-Statuskonferenz 2023 Laura Schulz

Source: https://physicsworld.com/a/the-diamond-quantum%E2%80%AFrevolution/

8

Summary / Conlusions

- Many open issues in “performance“ modelling of quantum computers
  - Classical concepts of data, computation, latency, throughput need to be reconsidered
  - Techonolgy not yet fixed → time scales, reliability,…
  - Noisy qubits → statistical modelling

- Similar questions arise with other disruptive technologies

- Performance Modelling for QC – right time to start with?!