The limits of quantum circuit simulation with low precision arithmetic

S. I. Betelu

UNT (adjunct)
DataVortex (chief scientist)

Oct 2020

How to simulate ideal quantum computers and why

The normalized wave function in a circuit of Q qubits is written as

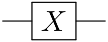
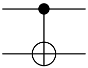
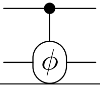
$$|\psi\rangle = \sum_{k=0}^{N-1} c_k |k\rangle, \qquad \sum_{k=0}^{N-1} |c_k|^2 = 1$$

and the output is the result of a series of matrix multiplications,

$$|\psi_t\rangle = U_t \cdot U_{t-1} \cdot \ldots \cdot U_2 \cdot U_1 \cdot |\psi_0\rangle$$

- $N = 2^Q$ is the number of terms. Cannot simulate $Q > 50$ (quantum supremacy)
- $|k\rangle$ are the computational basis states (orthogonal unit vectors).
- U_i are *quantum gates*, $N \times N$ *unitary* matrices $U_i^* U_i = I$
- Current quantum computers (IBM Q, Rigetti, Google) very primitive, only way to design and test new quantum algorithms is with simulation

Typical elementary gates (universal)

Gate	Circuit	Matrix
NOT		$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
Hadamard		$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$
Controlled NOT		$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
Controlled phase		$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{i\phi} \end{pmatrix}$

Most useful gates can be represented as tensor products of more elementary gates $U = U_1 \times U_2 \times ... \times U_s$ and linear operations

Practical implementation of gates

How gates operating on qubits p, q are implemented: "$\leftarrow$" represents assignment and "$\leftrightarrow$" swapping. Parentheses contain the binary index k of c_k and the dots indicate unaffected bits. H is Hadamard's gate and CP are the controlled phase gates.

$$|\psi\rangle = c_{00}|00\rangle + c_{01}|01\rangle + c_{10}|10\rangle + c_{11}|11\rangle$$

Gate	**Operation**
$H(q)$	$c(..,0_q,..) \leftarrow \frac{1}{\sqrt{2}}\left(c(..,0_q,..) + c(..,1_q,..)\right)$ $c(..,1_q,...) \leftarrow \frac{1}{\sqrt{2}}\left(c(..,0_q,..) - c(..,1_q,..)\right)$
CNOT (p,q)	$c(..,1_p,..,0_q,..) \leftrightarrow c(..,1_p,..,1_q,..)$
CP(p,q)	$c(..,1_p,..,1_q,..) \leftarrow e^{i\pi/2^m} c(..,1_p,..,1_q,..)$
SWAP(p,q)	$c(..,1_p,..,0_q,..) \leftrightarrow c(..,0_p,..,1_q,..)$

K. De Raedt, K. Michielsen, H. De Raedt, B. Trieu, G. Arnold, M. Richter, Th. Lippert, H. Watanabe, N.Ito, "Massively parallel quantum computer simulator", Computer Physics Communications, 176, 2, (2007)

Example circuit: Quantum Fourier Transform (QFT)

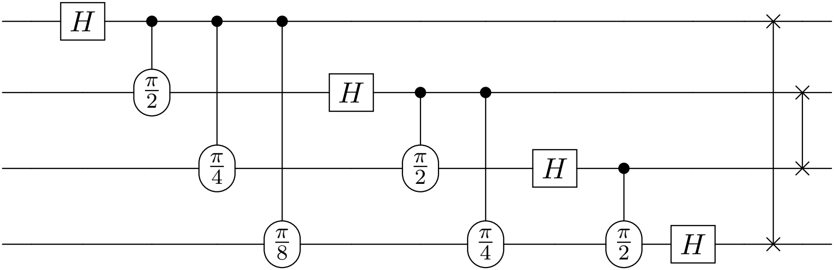
Input:

$$|\psi_0\rangle = \sum_{k=0}^{N-1} c_k |k\rangle$$

The $N = 2^Q$ output coefficients are the usual FT

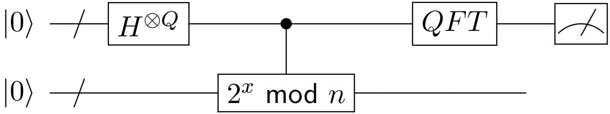
$$|\psi_t\rangle = \sum_{k=0}^{N-1} f_k |k\rangle, \qquad f_k = \frac{1}{\sqrt{N}} \sum_{l=0}^{N-1} c_l e^{2\pi i k l / N}$$

Time complexity $T = O(Q^2)$ versus classical FFT $O(N \log_2 N) = O(Q 2^Q)$



A famous algorithm: Shor's algorithm

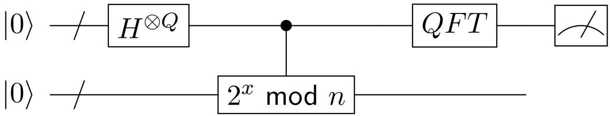
- Problem: factorize $n = p \cdot q$
- Find Q such that $n^2 \leq 2^Q < 2n^2$
- Take into account that the period of the function $f(x) = a^x \bmod n$ divides Euler's totient function $\phi(n) = (p-1)(q-1)$
- Take the FT and measure the position of a peak, then do some math (classical continued fraction expansion) to find the factors.



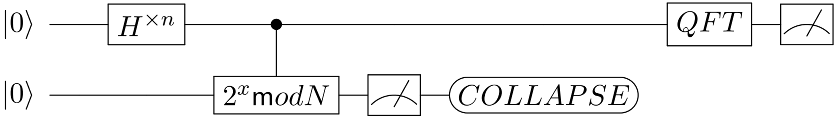
PW Shor, "Polynomial-Time algorithms for prime Factorization and Discrete Logarithms on a Quantum Computer", SIAM J. Comp 1997

Quantum simulation benchmark (QuanSimBench)

- Simplify Shor's algorithm
- Factorize increasing integers until memory exhausted
- Only simulates AQFT: same as QFT but with fewer phases
- Generate data of measured $f(x) = a^x \bmod n$ classically.
- Open source **https://github.com/datavortex/QuanSimBench**



Deferred measurement does not change QFT peaks



Thus result is equivalent to load data after first measurement

Other methods for ideal quantum circuit simulations

G: number of gates, D: depth of circuit, M: memory usage, T: time

- Schrodinger's formulation: full vector states (this work)
  $|\psi\rangle = \sum_{k=0}^{N-1} c_k |k\rangle \qquad |\psi(t)\rangle = U_G \cdot U_{G-1} ... U_2 \cdot U_1 \cdot |\psi_0\rangle$

  $$T = O(G2^Q)$$
  $$M = O(2^Q)$$

  feasible for random states and large depths.
- Feynman path integration (very slow)

  $$T = O(2^G)$$
  $$M = O(G + Q)$$

- Tensor contraction family: time-space tradeoff, good for low entropy states, problematic for large depths and random states

  $$T = O(Q2^{Q-k}(2D)^{k+1})$$
  $$M = O(2^{Q-k} log(D))$$

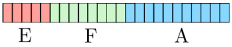
Saving memory with log-polar low precision format

Encode quantum state

$$|\psi\rangle = \sum_{k=0}^{N-1} c_k |k\rangle$$

$$c_k \approx T(c_k) = \exp\left(-\left(e_k + \frac{f_k}{2^F}\right) + 2\pi i \frac{a_k}{2^A}\right), \tag{1}$$

The complex amplitudes are encoded with E bits for the integer part of the exponent, F bits for the fraction and A bits for the argument.



bits per coefficient: $B = E + F + A$

- Rounding error is uniformly distributed
- Simplifies mathematical analysis
- Some phase gates are exact $\pi/2^k, k < A$.
- More accurate than pairs of floats for given number of bits.
- Drawback: slower, not native CPU conversions
- Use lookup tables and interpolation to speed up

Log-polar versus pair of floating point numbers

Polar format more regular, simpler error statistics and allows to compute phase gates $P(\pi/2^k)$ without error.

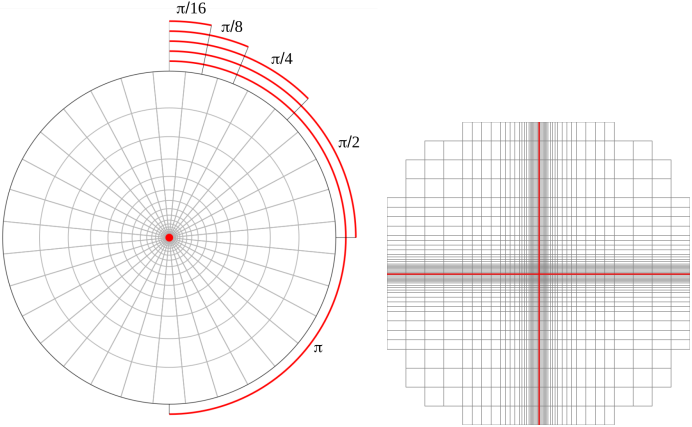


Figure: Very low precision format with $E = 2, F = 2, A = 5$ (9 bits) versus floats with 3 bits of exponent and 2 of mantissa (10 bits). Red are underflows.

Distribution of rounding errors is uniform

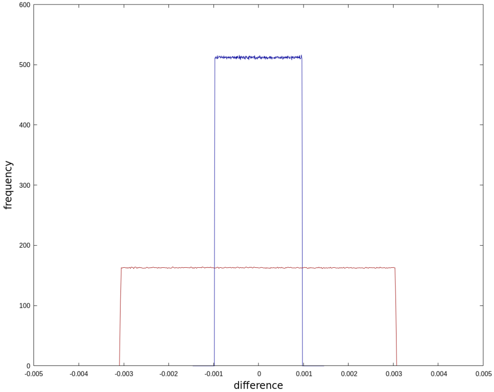


Figure: Empirical histograms of the rounding errors for the logarithm of the modulus (high rectangle) and for the argument of Eq. (1) for $Q = 20$, $E = 5, F = 9$ and $A = 10$. They are uniformly distributed when the real and imaginary parts of the coefficients c_k are random because the rounded binary digits after the least significant digit are random. This is not true for floating point formats.

Main result: cumulative error after G error-prone gates

Define the cumulative error,

$$\sigma^2 = \left\| |\psi_G\rangle - |\psi_{G,exact}\rangle \right\|^2$$

and assuming the initial condition has maximum entropy,

$$\boxed{\sigma^2(E,F,A,G) \approx \left(\phi + (1-\phi)\frac{2^{-2F} + 4\pi^2 2^{-2A}}{12}\right) G,}$$

where the normalization error due to underflows is

$$\phi = 1 - (N\mu^2 + 1)e^{-N\mu^2}$$

and the smallest representable modulus is

$$\mu = \min |c_k| = \exp(-2^E + 2^{-F})$$

For non-random states, an upper bound for the error (loose bound)

$$\sigma^2 \leq \frac{2^{-2F} + 4\pi^2 2^{-2A}}{4} G^2$$

Optimal triplets E, F, A with respect of the expected value of the conversion error for random states, computed by brute force minimization of the conversion error with the constraint $E + F + A = B$.

	$Q = 20$	$Q = 30$	$Q = 40$	$Q = 50$
B	E, F, A	E, F, A	E, F, A	E, F, A
12	4, 3, 5	4, 3, 5	4, 3, 5	5, 2, 5
14	4, 4, 6	4, 4, 6	4, 4, 6	5, 3, 6
16	4, 5, 7	4, 5, 7	4, 5, 7	5, 4, 7
18	4, 6, 8	4, 6, 8	5, 5, 8	5, 5, 8
20	4, 7, 9	4, 7, 9	5, 6, 9	5, 6, 9
22	4, 8, 10	4, 8, 10	5, 7, 10	5, 7, 10
24	4, 9, 11	4, 9, 11	5, 8, 11	5, 8, 11
26	4, 10, 12	4, 10, 12	5, 9, 12	5, 9, 12
28	4, 11, 13	4, 11, 13	5, 10, 13	5, 10, 13
30	4, 12, 14	4, 12, 14	5, 11, 14	5, 11, 14
32	4, 13, 15	4, 13, 15	5, 12, 15	5, 12, 15
34	4, 14, 16	4, 14, 16	5, 13, 16	5, 13, 16
36	4, 15, 17	4, 15, 17	5, 14, 17	5, 14, 17

Not all gates generate the same errors: effective gates

$$G = \sum_{g=1}^{n} \beta_g, \tag{2}$$

n is the total number of gates, β_g is the fraction of coefficients affected by gate g,

Gate type	β_g
$X, Z^{1/k}$ $(k < A)$, CNOT, SWAP, TOFF	0
$Z^{1/k}$ $(k \geq A)$	$1/2$
H, $X^{1/k}$, $Y^{1/k}$ $(k > 2), U_3(\theta, \lambda, \phi)$	1
Last row with k controls	$1/2^k$

Table: Fraction of coefficients affected by rounding error for typical gates.

Sketch of derivation

Compute the expected value of

$$\varepsilon_c^2 = \| T|\psi\rangle - |\psi\rangle \|^2 = \sum_{k=0}^{N-1} |T(c_k) - c_k|^2 =$$

$$\sum_{|c_k|<\mu} |c_k|^2 + \sum_{|c_k|\geq\mu} |c_k|^2 \left| e^{\epsilon_k + i\gamma_k} - 1 \right|^2 \approx \phi + (1-\phi)\frac{2^{-2F} + 4\pi^2 2^{-2A}}{12}$$

using uniform distribution of $-2^{-F}/2 \leq \epsilon_k \leq 2^{-F}/2$ and $-\pi 2^{-A} \leq \gamma_k \leq \pi 2^{-A}$ and that $p = |c_k|^2$ are distributed according to Porter-Thomas distribution with PDF $f(p) \approx N e^{-pN}$

For the cumulative error we use unitariness and the recurrence

$$|\varepsilon_{t+1}\rangle = U_t |\varepsilon_t\rangle + |\tau_t\rangle.$$

How many gates can we compute

- The *fidelity* is defined as

$$\Phi = |\langle \psi_G | \psi_{G,exact} \rangle|^2$$

- related to σ^2 as

$$\Phi \geq \left(1 - \sigma^2/2\right)^2$$

- A barely tolerable result has $\sigma^2 = 1/4$ represents a fidelity of $\Phi \geq 0.765$ (this would be the probability of success of an algorithm if the final state had only one coefficient $c_k \neq 0$).
- Number of error-prone gates we can compute high entropy states

$$G_{random} < \frac{12\sigma^2}{2^{-2F} + 4\pi^2 2^{-2A}}.$$

- Only counts error-prone gates, many gates are error free.

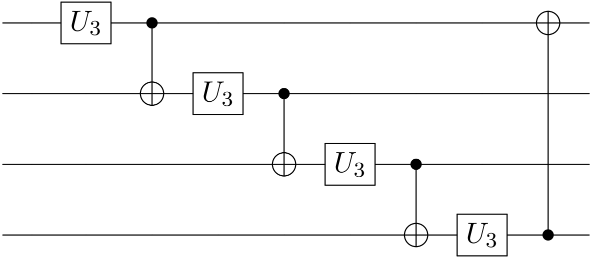
How many gates can be computed with low precision

B	ε_c^2	G_{random}
8	1.35e-01	2
12	8.42e-03	30
16	5.26e-04	475
20	3.29e-05	7600
24	2.06e-06	121599
28	1.28e-07	1.94e+06
32	8.03e-09	3.11e+07
36	5.02e-10	4.98e+08
40	3.14e-11	7.96e+09

Table: Typical values of one-conversion errors ε_c^2 and maximum number of error prone gates for $\sigma = 1/2$ and $Q = 50$ for random states using the optimal triplets.

Random circuit test: a circuit hard to simulate

- Generates entangled maximum entropy state after $C \approx 7$ cycles
- Test the ability of a simulation (or quantum computer) to "hold" a maximally entangled state
- Each cycle rotates all qubits in the Bloch sphere with the rotation gate $U_3(\pm\pi/2, \pm\pi/4, \pm\pi/4)$ and random signs.



$$U_3(\theta, \lambda, \phi) = \begin{pmatrix} \cos\frac{\theta}{2} & -e^{i\lambda}\sin\frac{\theta}{2} \\ e^{i\phi}\sin\frac{\theta}{2} & e^{i(\lambda+\phi)}\cos\frac{\theta}{2} \end{pmatrix}$$

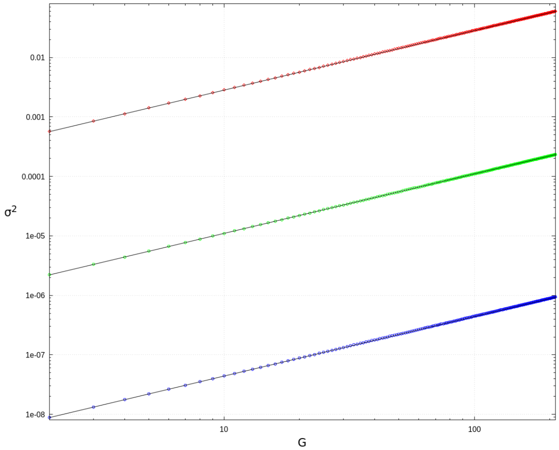


Figure: Growth of the numerical cumulative error (points) for a uniformly distributed, random initial condition, as a function of the number of error prone gates G, compared with the model (lines), with $Q = 30$ for triplets E, F, G: $4, 5, 7$ (top line), $4, 9, 11$ (middle) and $4, 13, 15$ (bottom). The error is computed by comparing the output with low precision $|\psi_G\rangle$ with a computation with double precision as a proxy for the exact solution $|\psi_{ex}\rangle$.

Resulting coefficients distribution for random circuits

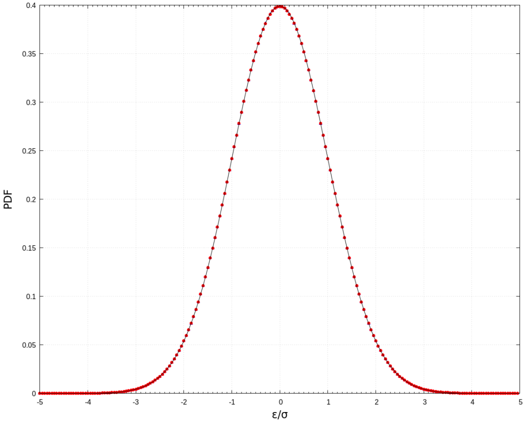


Figure: Starting with a uniform random initial condition we run 7 cycles twice, first with double precision and then with low precision. These are the histograms of the normalized errors of the real part of the coefficients, $\mathrm{Re}(c_{k,double} - c_{k,lowprec})$ for $E = 4, F = 9, A = 11$ (points). The distribution is approximately normal with standard deviation σ.

Back and forth test

```
// CREATE A RANDOM STATE
for i=1,C
    for q=1,Q
        k= Q*i+q
        U3(q, t(k), l(k), p(k) )
        CNOT(q, (q+1)%Q )
    end
end
NORMALIZE
// RUN IN REVERSE ORDER TO RESTORE IC
for i=C,1
    for q=Q,1
        k= Q*i+q
        CNOT(q, (q+1)%Q )
        U3(q, -t(k), -p(k), -l(k) )
    end
end
NORMALIZE
```

Back and forth test

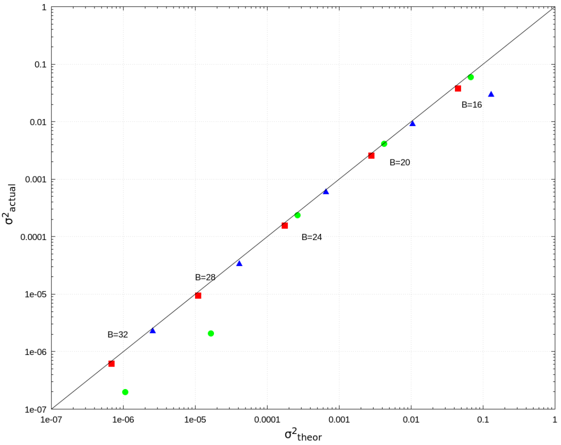


Figure: Random algorithm test, 4 cycles forth and then 4 cycles for the inverse. Comparison of the actual error (y-axis) and the theoretical error (x-axis). Red squares: 20 qubits, brown circles: 30 qubits, blue triangles: 40 qubits. The bits per coefficient are indicated on the labels, with optimal triplets E, F, A .

Reducing errors on amplitudes: normalization

When the normalization deteriorates,

$$\sum_{k=0}^{N-1} |c_k|^2 \neq 1$$

must renormalize each time the total probability departs from unity with random factors $-2^{-F-1} < \delta_k < 2^{-F-1}$

$$c'_k = \frac{c_k}{\| |\psi\rangle \|} e^{\delta_k},$$

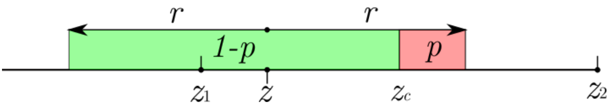
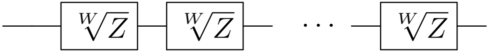


Figure: Let $z = \ln \frac{|c_k|}{\| |\psi\rangle \|}$ and $z_1 < z_2$ be two consecutive discrete logarithms with separation $z_2 - z_1 = 2^{-F} = 2r$ and $z_1 < z < z_2$. We want to round z to the closest of z_1 or z_2. After we add a uniformly distributed random number δ to z, with $-r \leq \delta < r$, the numbers to the right of $z_c = (z_1 + z_2)/2$ are rounded to z_2 with probability $p = (z - z_1)/(2r)$ and the numbers to the left of z_c are rounded to z_1 with probability $1 - p$, thus $\mathbb{E}(round(z + \delta)) = (1 - p)z_1 + pz_2 = z$.

Reducing rounding errors on phases

Systematic errors accelerates the growth of total error. Below is a potentially failing circuit after $W > 2^A$ applications of the gate. The problem is solved by multiplying the amplitudes with carefully chosen random factors



$$c'_k = c_k \exp(\delta_k + i\gamma_k), \tag{3}$$

with $-2^{-F-1} < \delta_k < 2^{-F-1}$ and $-\pi 2^{-A} < \gamma_k < \pi 2^{-A}$
In other algorithms normalization may be necessary as well.

Final remarks

Other tests performed

- Quantum Fourier Transform
- Grover's algorithm
- Simplified Shor's algorithm (quansimbench)

Open problems

- Is it optimal? (preliminary work says no, but it is close)
- How to speedup?
- Translate to tensor contraction formulations
- Solve partial differential equations of high dimensionality

ACKNOWLEDGMENTS

Many thanks to the University of North Texas HPC Research IT Services for providing access to TALON 3 computer, to Datavortex Technologies that supported this work and provided the computer system Hypatia, to the Texas Advanced Computing Center (TACC), University of Texas at Austin, for providing access to the Stampede 2 system.

- S. Betelu, "Quansimbench: a benchmark for HPC quantum circuit simulations", https://github.com/datavortex/QuanSimBench
- S. Betelu, "C and MPI simulation of quantum circuits with low precision arithmetic", https://github.com/datavortex/lowprecisionqubits