Physics-Aware, Full-Stack Optimization

Fred Chong
Seymour Goodman Professor, University of Chicago
Chief Scientist for Quantum Software, Infleqtion

Lead PI, the EPiQC Project, an NSF Expedition in Computing
NSF 1730449/1730082/1729369/1832377/1730088
NSF Phy-1818914/OMA-2016136
DOE DE-SC0020289/0020331/QNEXT
ARO W911NF-23-1-0077

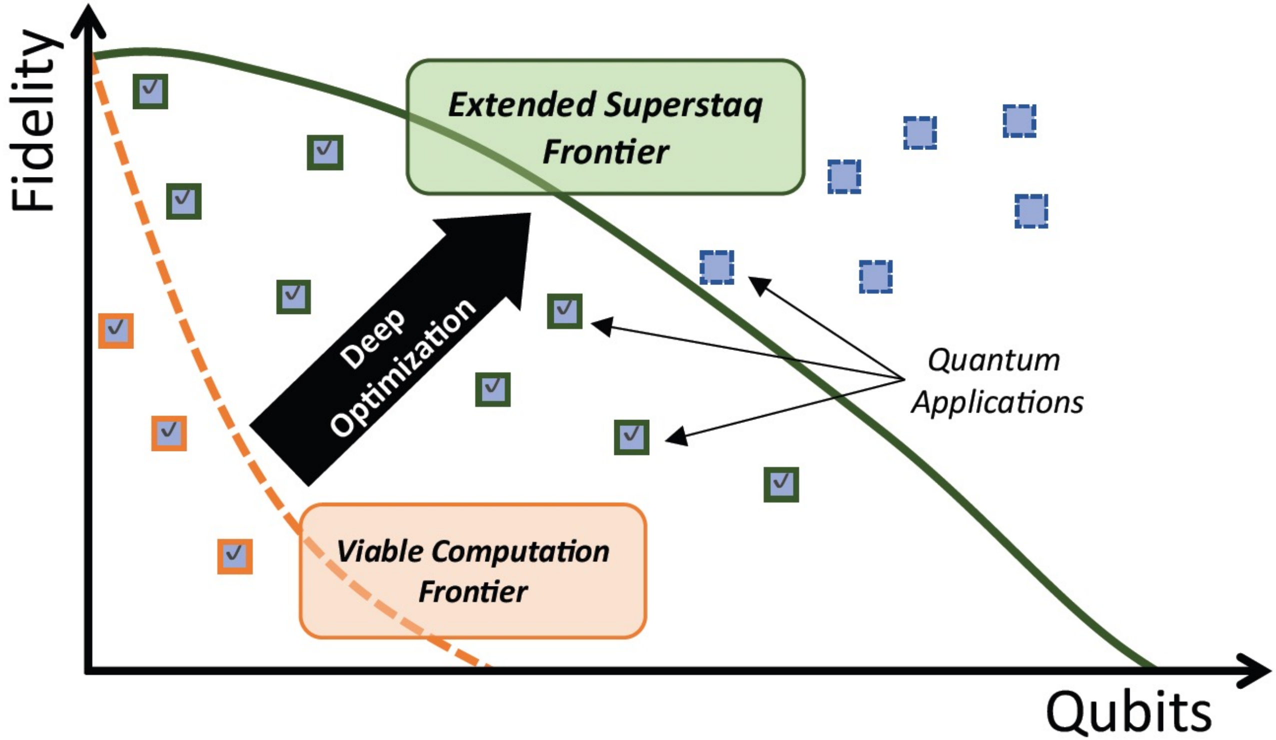
Disclosure: FC is an advisor to QCI

Last 5 Years: Please break abstraction layers!

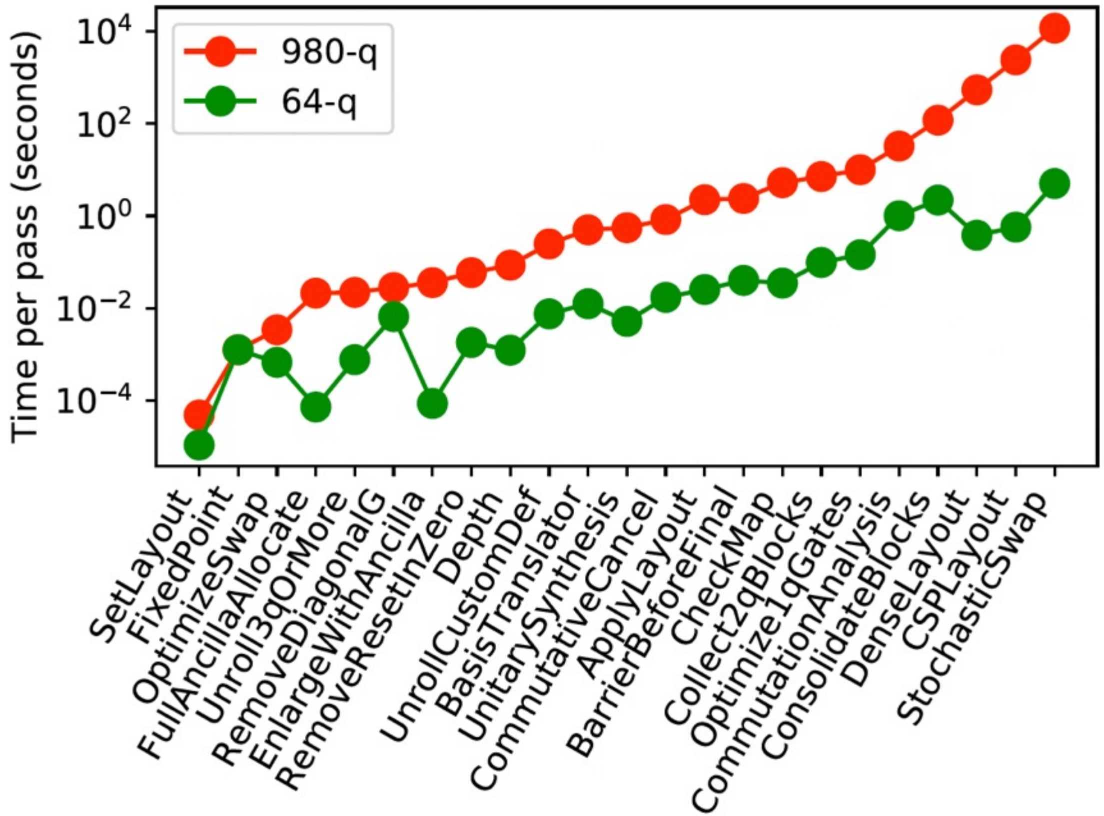
X. Fu et al. (2017)
arXiv:1708.07677

- Stack: rigid layers + interfaces
- Benefits
  - Taming complexity
- **Problems**
  - **Lost opportunities for optimization**
  - **QC stack + layers change**
- We should compile to hardware primitives. Physics first.

Deep Optimization Extends the Computation Frontier

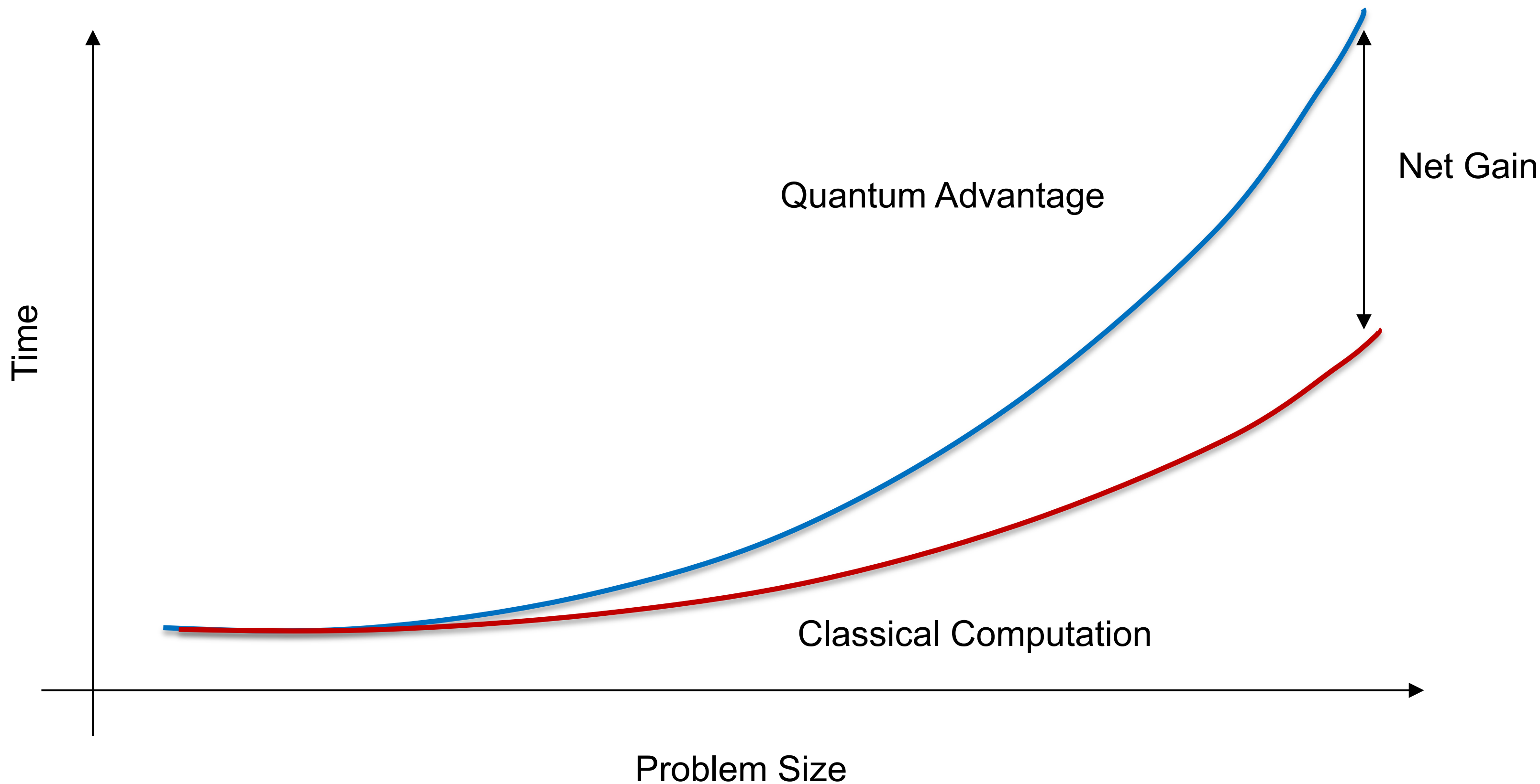


Scalability vs Deep Optimization



Gokul Subramanian Ravi, Kaitlin N. Smith, Pranav Gokhale, Frederic T. Chong: Quantum Computing in the Cloud: Analyzing job and machine characteristics. IISWC 2021: 39-50

Hybrid Quantum-Classical Computing: Contest of Exponentials



The Secret Menu of Quantum Hardware


TACO BELL
CHIPOTLE
MEXICAN GRILL
Enchirito
Buffalo Chicken
Quesarito

MENU

Item		Price
1.	NOT	0.1% error
2.	RZ(θ)	0.01% error
3.	CNOT	Market price

No substitutions permitted.

SECRET HARDWARE PRIMITIVE MENU

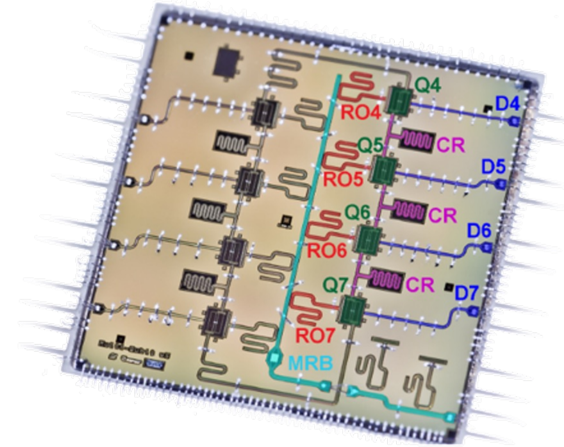
4.	CR(θ)	0.4% error
5.	Parametric $e^{-i\,\theta\,ZZ}$	~1.0% * θ/π
6.	|1> ←→ |2> qutrit transition	0.3% error

Additional options available—ask your friendly physicist.

1. Swap Gates and ECA

- *Optimized SWAP networks with equivalent circuit averaging for QAOA*, Akel Hashim , Rich Rines, Victory Omole, Ravi K. Naik, John Mark Kreikebaum, David I. Santiago, Frederic T. Chong, Irfan Siddiqi, and Pranav Gokhale. Physical Review Research 4, 033028

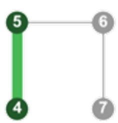
Advanced Quantum Testbed (AQT)



Single-qubit gates: $X_{\pi/2}$ (Rabi, 30ns)

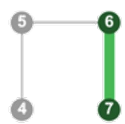
Z_φ (virtual)

Entangling gates:



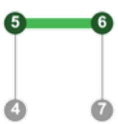
$CZ_{4,5}$: 200ns

$CS^\dagger_{4,5}$: 150ns



$CZ_{6,7}$: 200ns

$CS^\dagger_{6,7}$: 150ns



$CZ_{5,6}$: 200ns

$CS_{5,6}$: 150ns

Parallel gate performance:

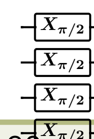
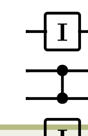
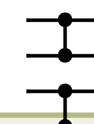
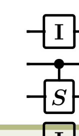
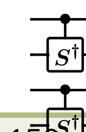
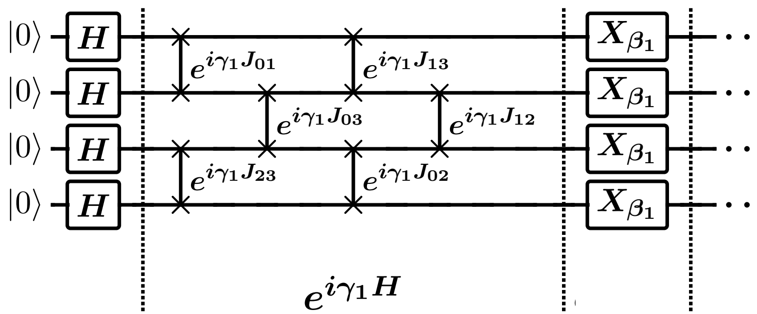
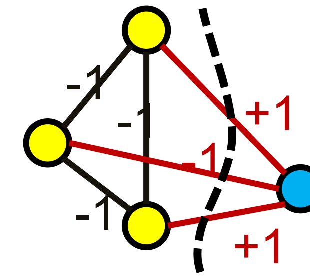
					
gate time	30ns	200ns	200ns	150ns	150ns
ϵ_T (10^{-2})	0.19(8)	1.09(9)	2.3(1)	0.68(9)	1.07(9)

Photo: A. Hashim, et al., *Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor*, arXiv:2010.00215 (2020).

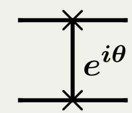
Application: QAOA for Weighted Max-Cut

Problem: Max-cut on a fully connected graph, edge weights ∈ {-1, 1}





⇒ *Efficiently mapped to linear topology*



$$= ZZ\text{-SWAP}(\theta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & e^{i\theta} & 0 \\ 0 & e^{i\theta} & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Equivalent Circuit Applications

Optimized Scheduling:

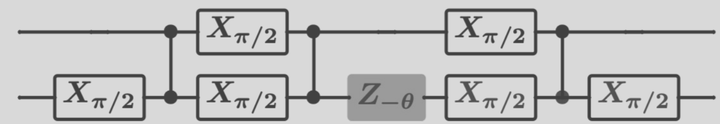
- *Iteratively select decompositions to maximize prior gate cancellation*

Equivalent Circuit Average (ECA):

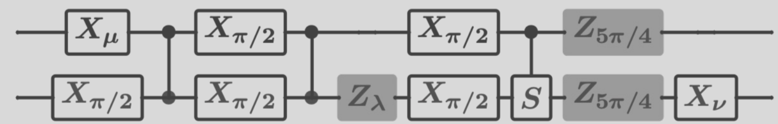
- *Randomly select decompositions to generate M logically equivalent circuits, to mitigate coherent error*

ECA + Optimized scheduling:

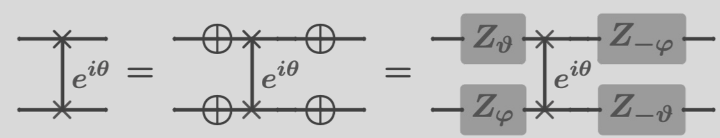
- *Only randomize over decompositions minimizing critical path depth*



32 unique CZ-CZ-CZ decompositions



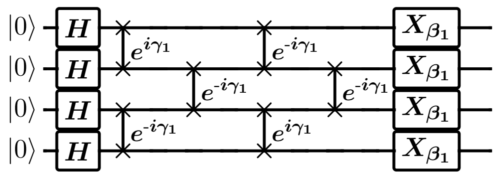
64 unique CZ-CZ-CS decompositions



2 commutation rules

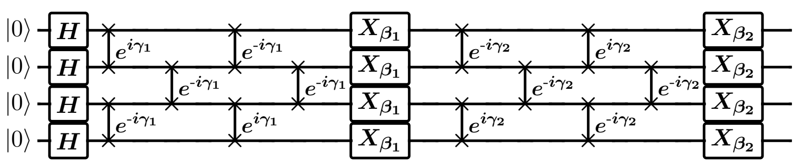
ECA + Optimized Scheduling on the AQT

QAOA, p = 1:

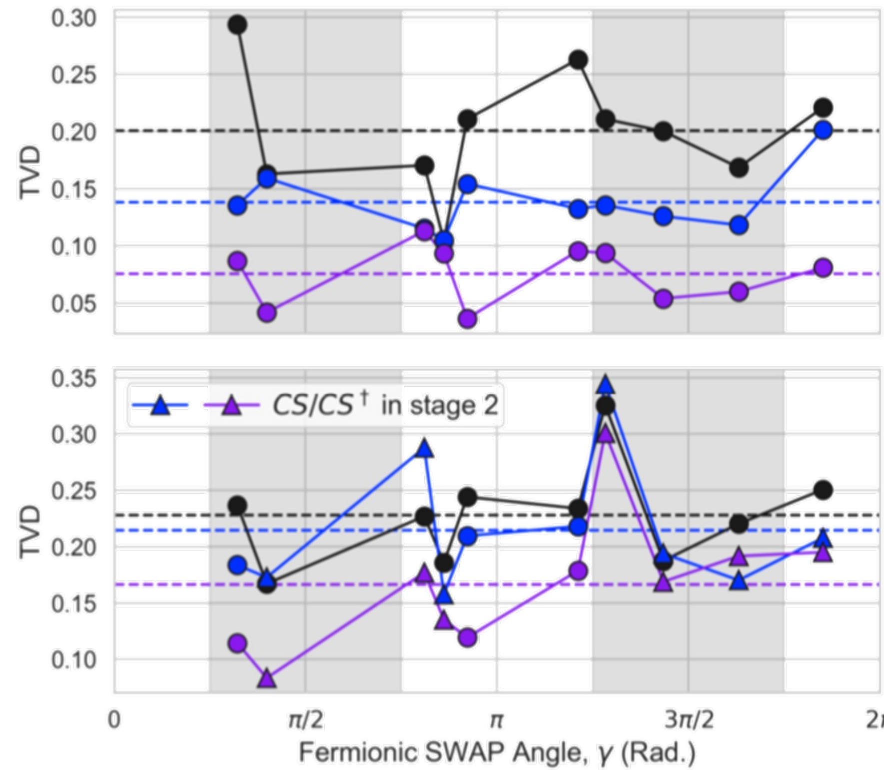


- Gate/schedule opt.: 30% error reduction
- ECA-OPT (M=20): **60% error reduction (TVD)**

QAOA, p = 2:



- Gate/schedule opt.: 7% error reduction
- ECA-OPT (M=20): **30% error reduction**



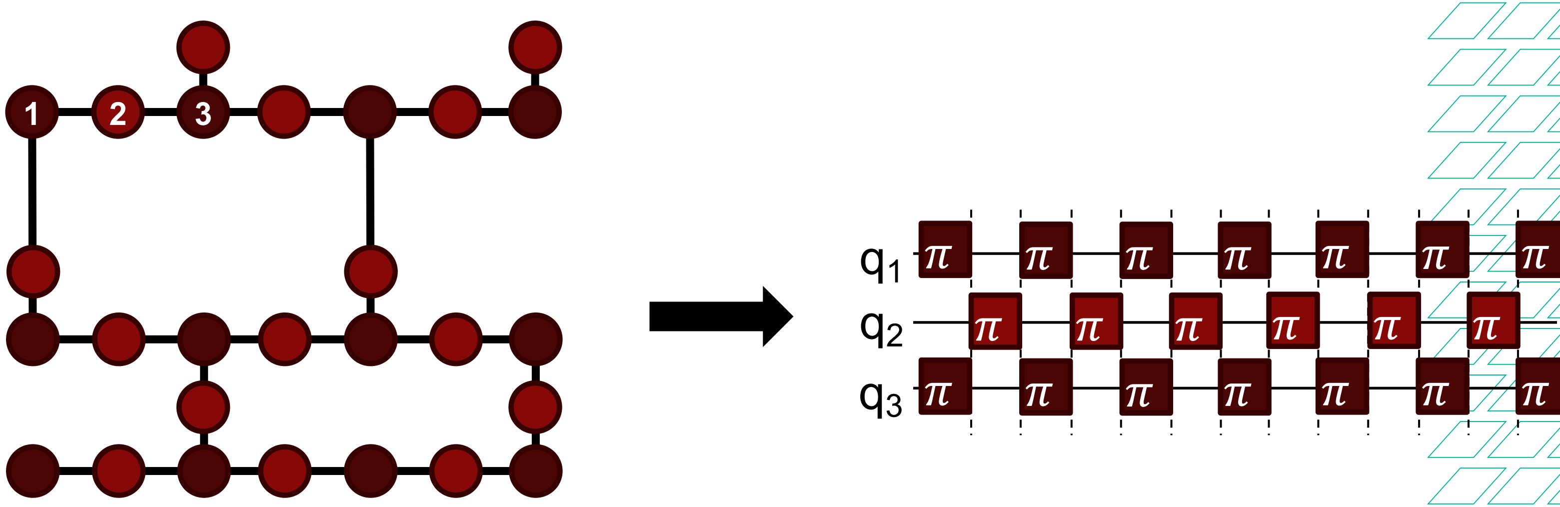
2. SuperstaQ

- *SuperstaQ: Deep Optimization of Quantum Program*, Colin Campbell, Frederic T. Chong, Denny Dahl, Paige Frederick, Palash Goiporia, Pranav Gokhale, Benjamin Hall, Salahedeen Issa, Eric Jones, Stephanie Lee, Andrew Litteken, **Victory Omole**, David Owusu-Antwi, Michael A. Perlin, Rich Rines, Kaitlin N. Smith, Noah Goss, Akel Hashim, Ravi Naik, Ed Younis, Daniel Lobser, Christopher G. Yale, Benchen Huang, Ji Liu. In the IEEE International Conference on Quantum Computing and Engineering, 2023.

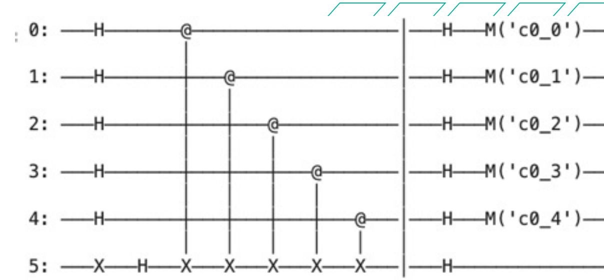
Infleqtion AQT < AIDE | QC >

Context-aware DD Approach

- The device graph informs the schedule of dynamical decoupling sequences
- DD pulses are limited to fixed points at constant intervals, determined by graph coloring, to mitigate unwanted qubit-qubit interaction (i.e. parasitic ZZ coupling)



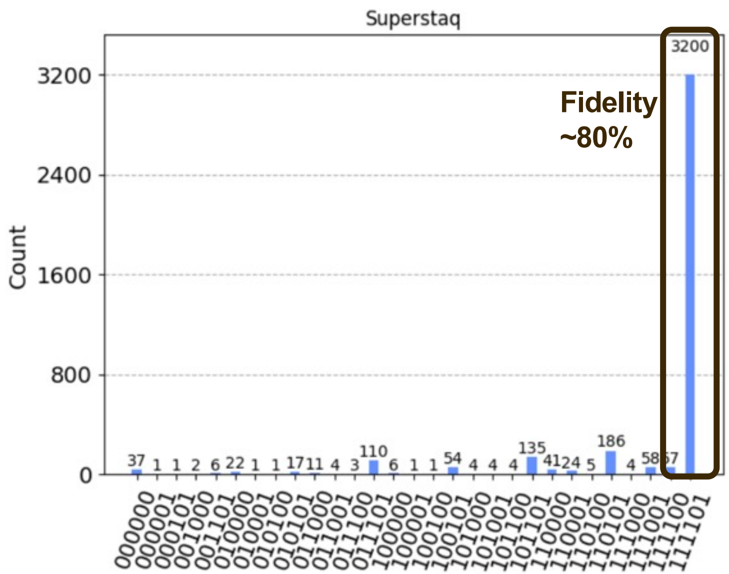
Superstaq Case Study: 6q BV

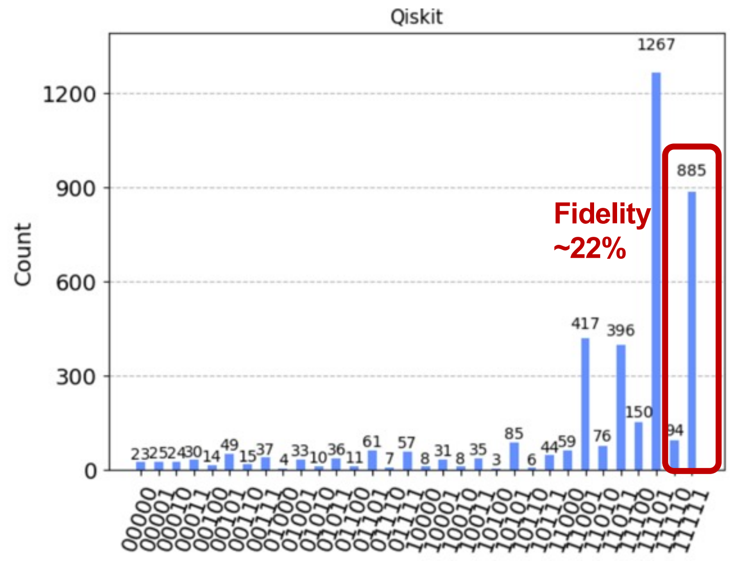


Logical Circuit

Superstaq improvements:

- 2.62x fidelity
- 24.63x relative strength

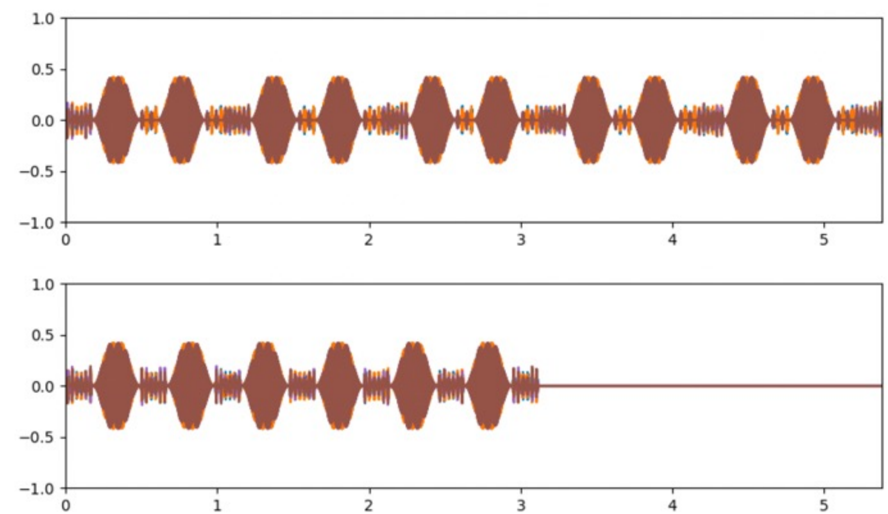




Note: distribution bitstring differences due to logical to physical qubit mapping

3. SuperstaQ: Bring Your Own Gates

Unitaries for Native Gates



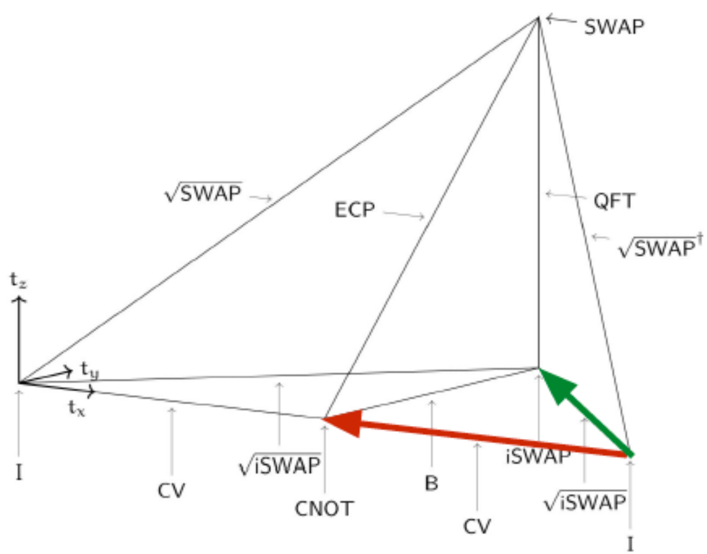
Qutrit Swap (73.5% Fidelity)

	|00⟩	|01⟩	|02⟩	|10⟩	|11⟩	|12⟩	|20⟩	|21⟩	|22⟩
|00⟩	0.63	0.06	0.05	0.02		0.04	0.08	0.09	0.03
|01⟩	0.01	0.01		0.7	0.01	0.09	0.04	0.1	0.04
|02⟩	0.05	0.04	0.02	-0.1	0.04	0.03	0.84	0.05	0.03
|10⟩	0.07	0.69	0.02	0.03		0.02	0.03	0.1	0.03
|11⟩	0.01	0.01	0.03		0.81	0.03	0.06	0.03	0.03
|12⟩	0.08	0.03	0.14	0.07	-0.03	0.03	0.04	0.56	0.07
|20⟩	0.04	0.03	0.65	0.01	0.04	0.02	0.05	0.12	0.04
|21⟩	0.02	0.02	0.01		0.03	0.83	0.03	0.03	0.02
|22⟩	0.02	0.01	0.01		0.01	-0.01	0.03	0.03	0.9

4. Adapting Gates for Qubit Variation

- *Let Each Quantum Bit Choose Its Basis Gates,* S. Lin, S. Sussman, C. Duckering, P. Mundada, J. Baker, R. Kumar, A. Houck, and F. Chong. International Symposium on Microarchitecture, 2022.

2Q basis gates on superconducting qubits



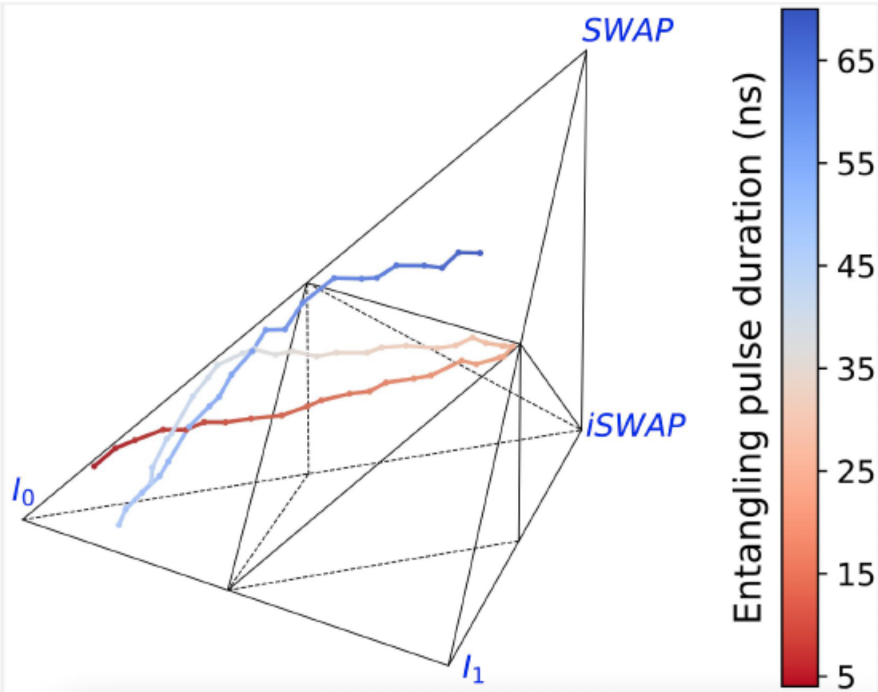
Weyl Chamber using Cartan Coordinates

It's standard practice to engineer XX- or XY-type gates on superconducting transmons

XX: from Identity to the equivalence class of CNOT/CZ. cross-resonance (CR), controlled phase (CPhase), CNOT/CZ

XY: from Identity to the equivalence class of iSWAP. iSWAP, bSWAP

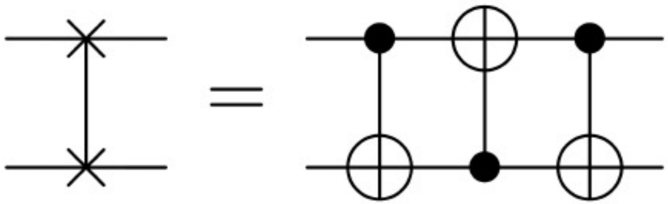
Deviations from standard trajectories



Systematic deviations arise, even when the goal is to engineer a standard gate family.

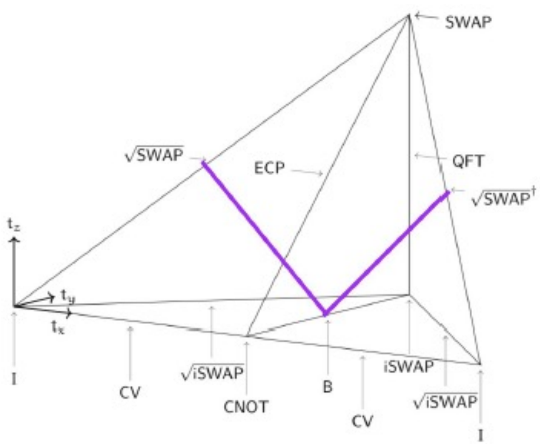
The figure shows an example of a nonstandard trajectory in the Weyl chamber, experimentally measured from a Houck lab device. The fastest perfect entangler on the device is at ~13ns.

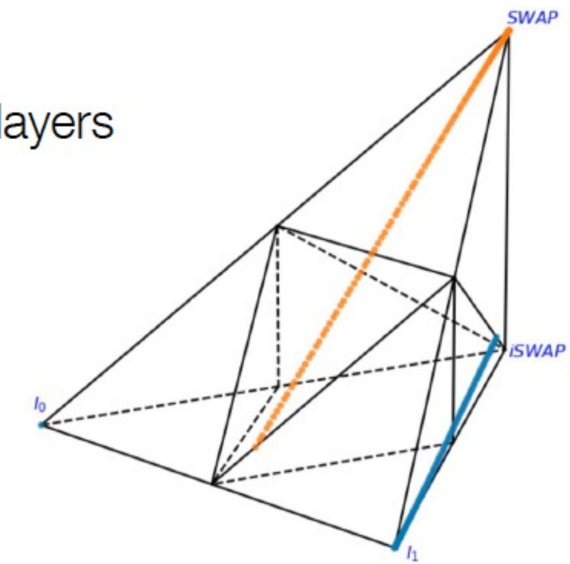
Synthesis of SWAP



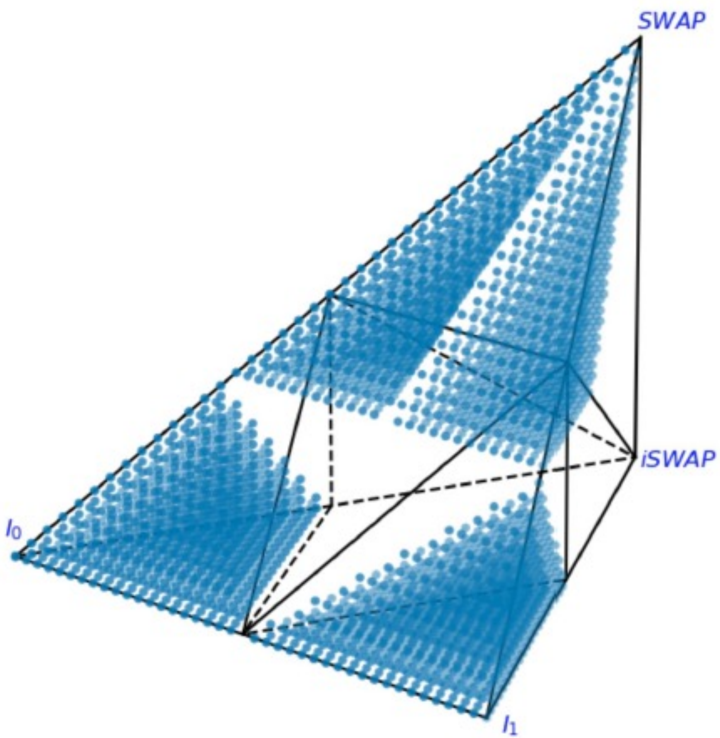
SWAP decomposed into 3 CNOT

SWAP synthesis in 2 layers

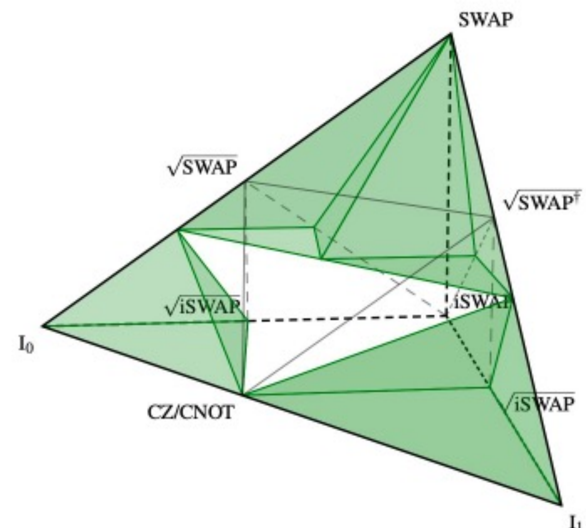




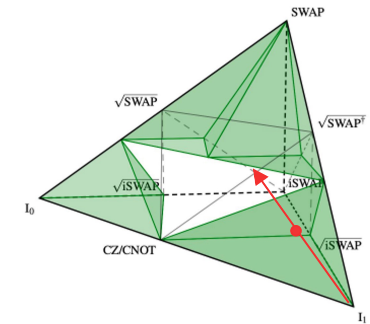
Synthesis of SWAP



Gates that are NOT able to synthesize SWAP in 3 layers



Gate Adaptation Summary



- 8X infidelity reduction for gates
- 2-30X infidelity reduction on circuits
- Need to optimize calibration (gate-set tomography)
- Need larger experimental demonstration
- *Break the abstraction of uniform basis gates*

5. Ququart Gates and Compilation

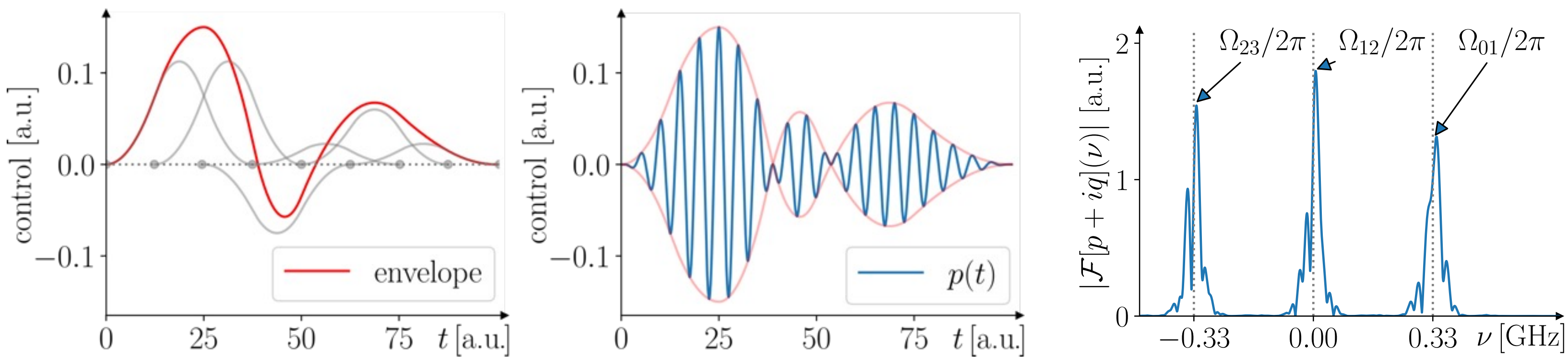
- *Time-Efficient Qudit Gates through Incremental Pulse Re-seeding*, Lennart Maximilian Seifert, Jason Chadwick, Andrew Litteken, Frederic T. Chong and Jonathan M. Baker. IEEE International Conference on Quantum Computing and Engineering, 2022.
- *Qompress: Efficient Compilation for Ququarts Exploiting Partial and Mixed Radix Operations for Communication Reduction*, A. Litteken, L. Seifert, J. Chadwick, N. Nottingham, F. Chong, and J. Baker. International Symposium on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2023.
- *Dancing the Quantum Waltz: Compiling Three-Qubit Gates on Four-Level Architectures*, Andrew Litteken, Lennart Maximilian Seifert, Jason Chadwick, Natalia Nottingham, Tanay, Ziqian Li, David Schuster, Frederic T. Chong, and Jonathan M Baker. International Symposium on Computer Architecture (ISCA), 2023.

Motivation

- 70X device reduction on qutrits for specialized ancilla circuits
- Ququarts would give more general compression/decompression and fast qubit operations inside a ququart
- Theory predicts quadratic scaling of gate duration with radix
- Maybe pulse implementation would be better…

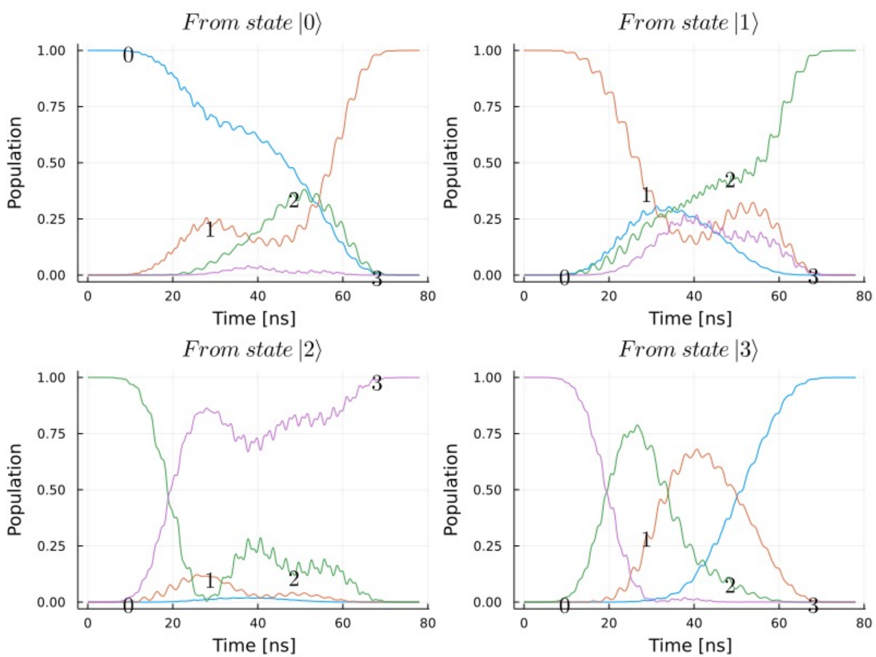
Quantum optimal control for qudits

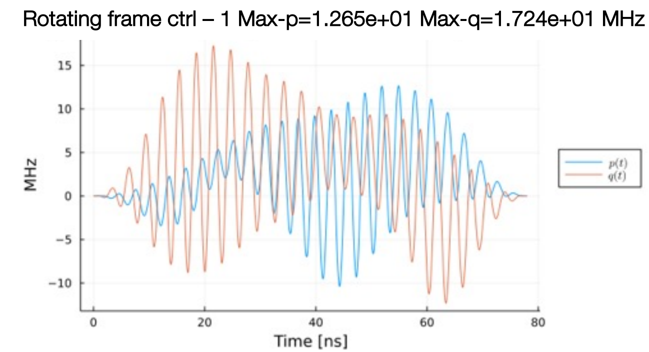
- QOC tools Juqbox and Quandary (LLNL)
  - Helpful parameterization using carrier waves
  - Choose carrier wave frequencies Ω_j resonant with qudit state transitions

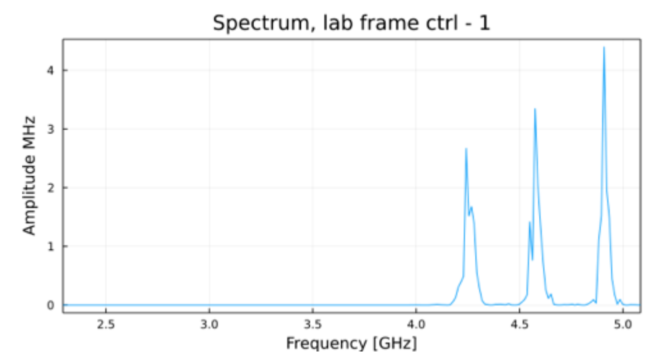


Quantum optimal control for qudits

- Example: QOC for X_4, $F = 99.9\%$

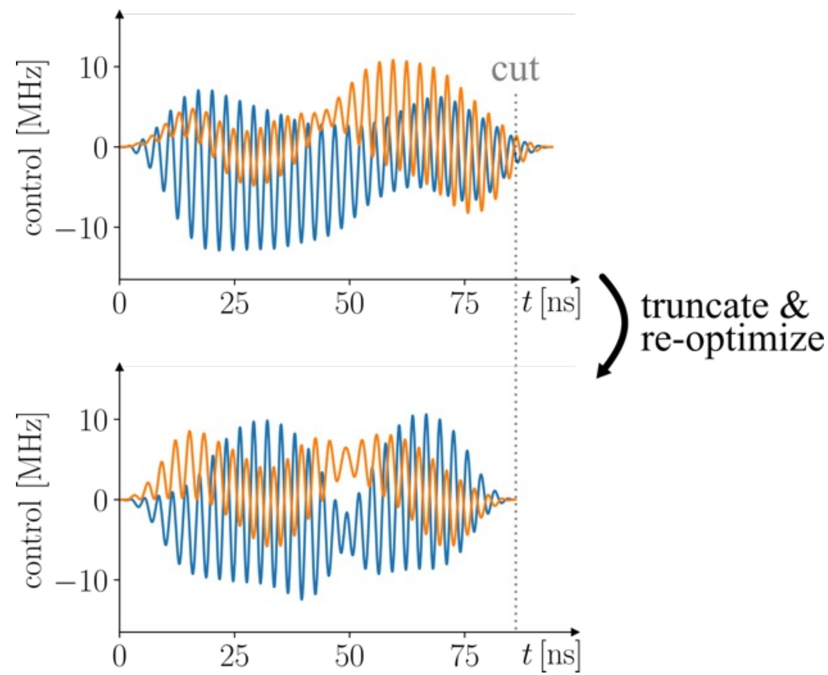






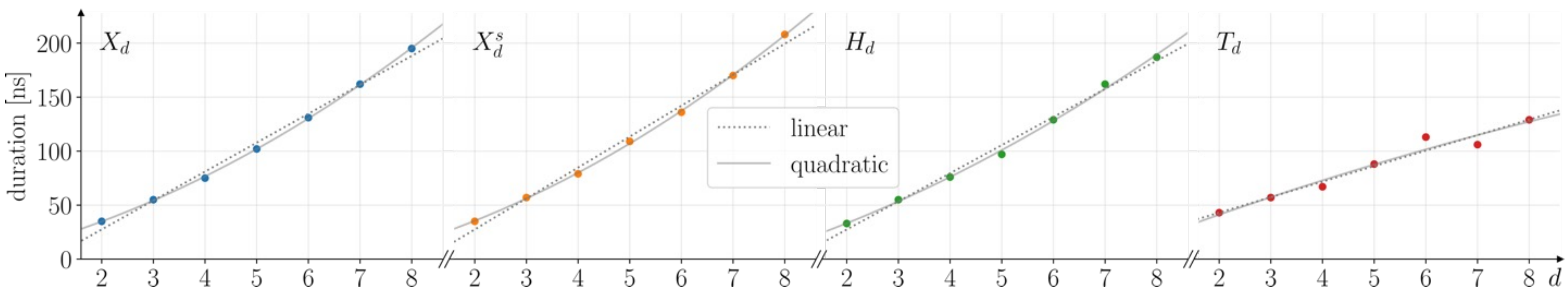
Pulse duration minimization using IPR

- Need to specify pulse duration T for optimization
- Iterative scheme: ***Incremental Pulse Re-seeding*** *(IPR)*
  - Key idea: Reuse optimized pulses from successful/failed iterations as initial guesses to start close to fidelity target



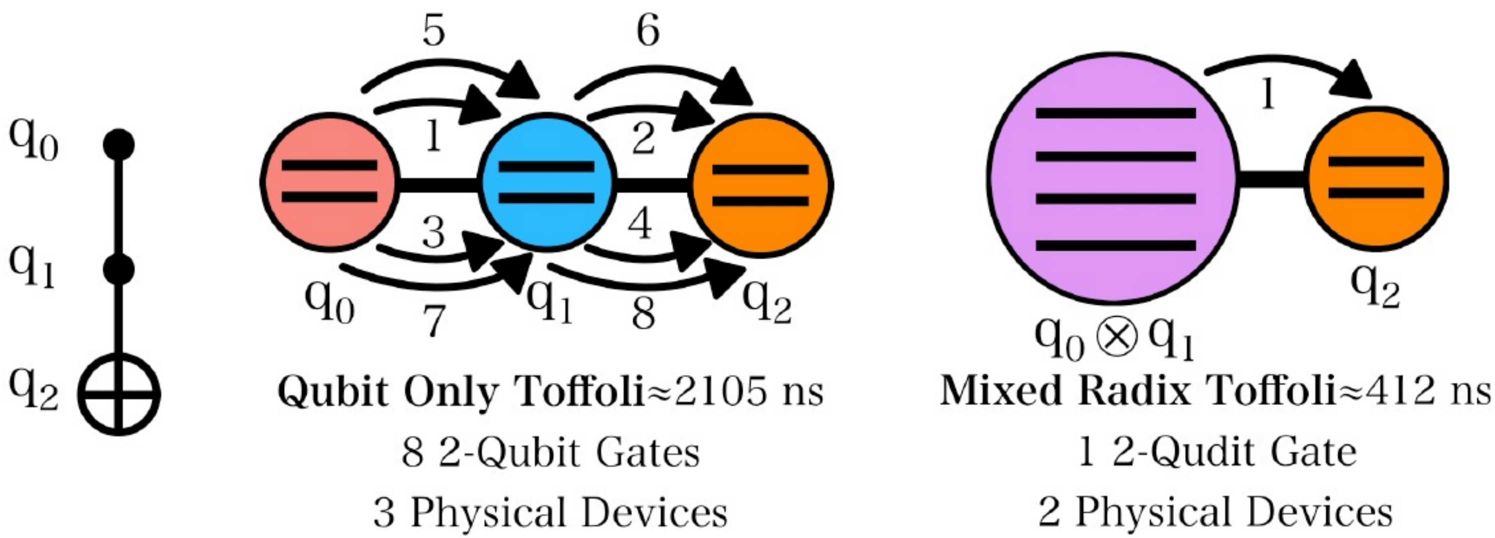
Pulse duration minimization using IPR

- Time-optimized pulse duration scaling: Single-qudit gates
  - Observe near-linear scaling up to $d = 8$ qudits



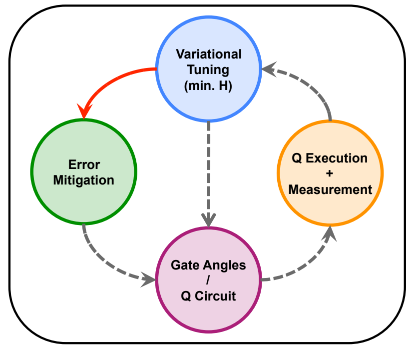
Qudit Conclusion

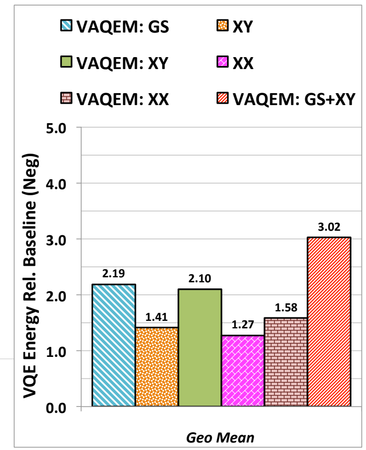
- Near-linear scaling in practical qudit regime $d \leq 8$
- *Enables compiler optimizations allowing 2X device savings with comparable fidelity (Qompress ASPLOS23)*
- *Enables ququart-qubit Toffoli (Quantum Waltz ISCA23)*

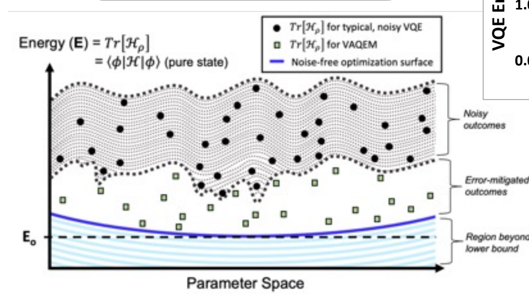


ALGORITHMIC OPTIMIZATIONS

6. VAQEM: A Variational Approach to Quantum Error Mitigation

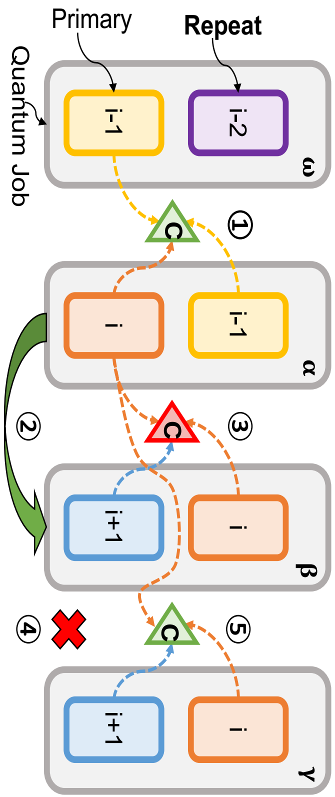


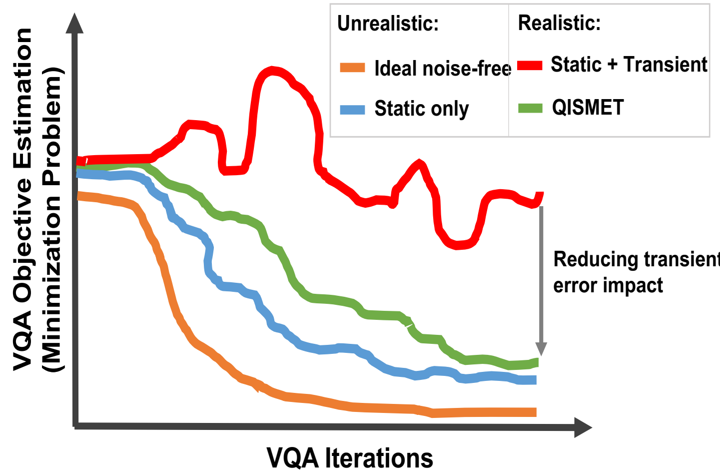


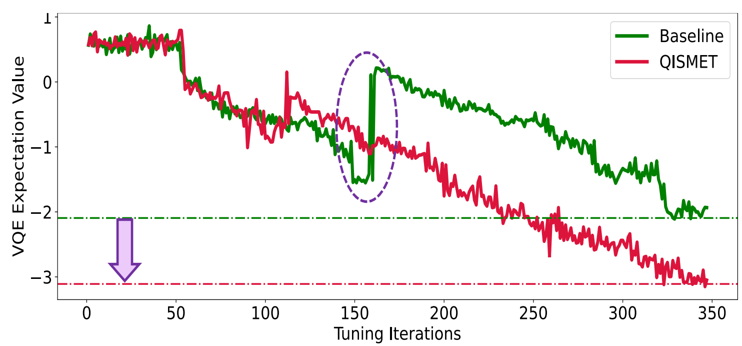


Gokul Ravi, Kaitlin N. Smith, Pranav Gokhale, Andrea Mari, Nathan Earnest, Ali Javadi-Abhari, Fred Chong. VAQEM: A Variational Approach to Quantum Error Mitigation. HPCA-28. Feb 2022.

7. Navigating the dynamic noise landscape of variational quantum algorithms with QISMET

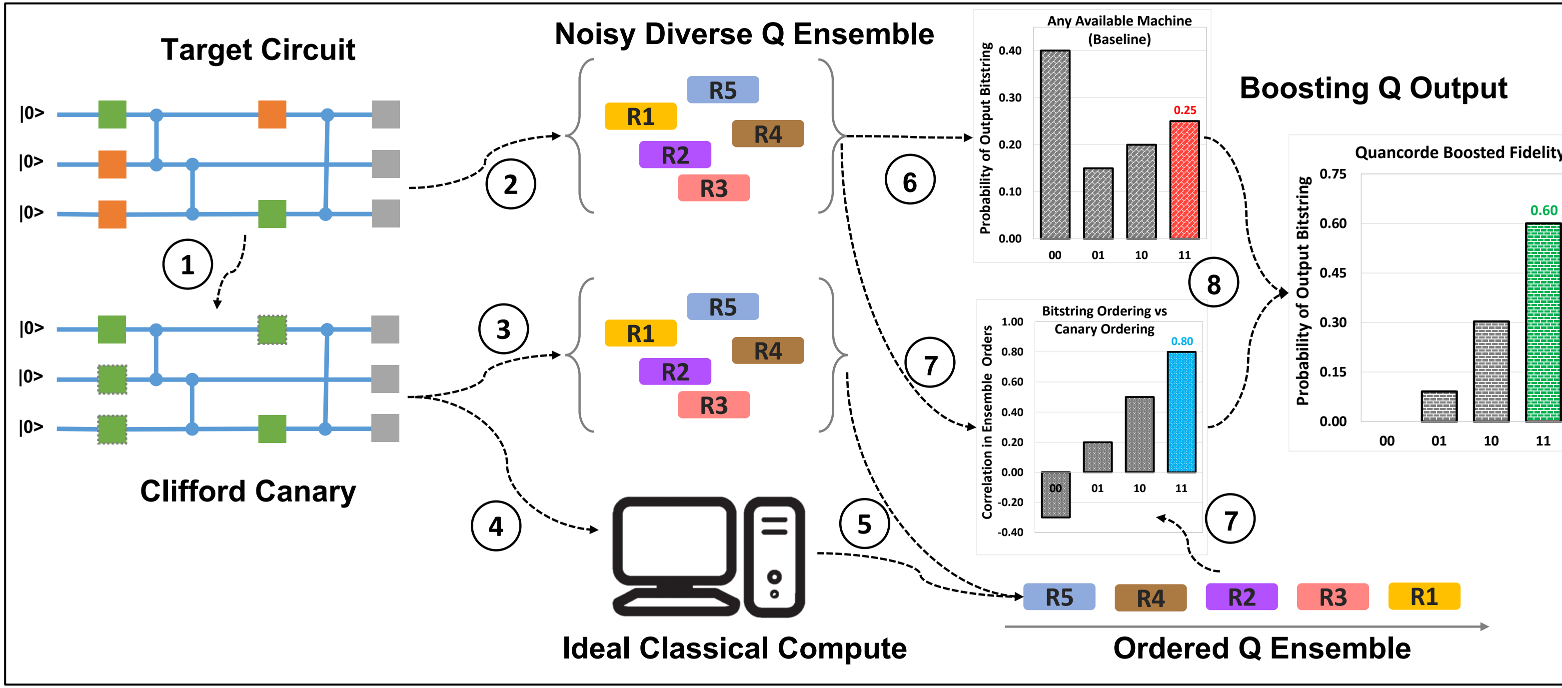






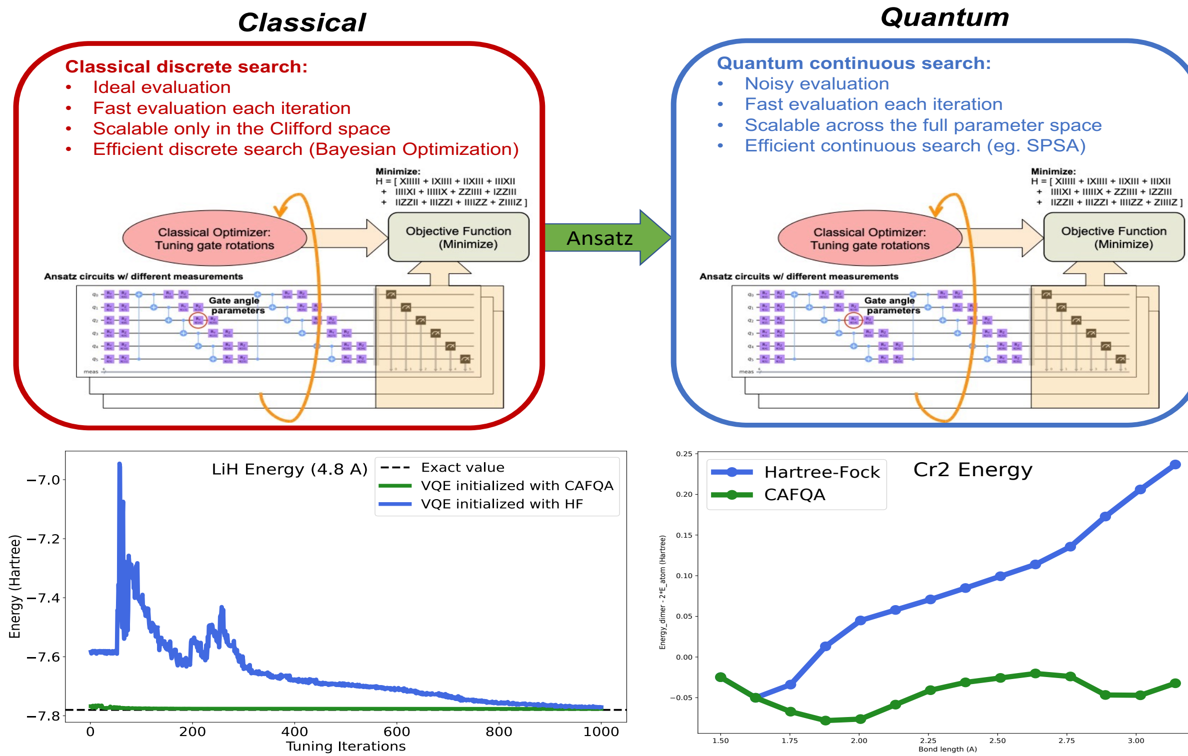
Gokul Subramanian Ravi, Kaitlin N. Smith, Jonathan M. Baker, Tejas Kannan, Nathan Earnest, Ali Javadi-Abhari, Henry Hoffmann, Frederic T. Chong, ASPOLS, March 2023.

8. Quancorde: Boosting fidelity with Quantum Canary Ordered Diverse Ensembles



Gokul Subramanian Ravi, Jonathan Baker, Kaitlin Smith, Nathan Earnest, Ali Javadi-Abhari, Frederic Chong. ICRC, December 2022 (Selected Highlight Paper)

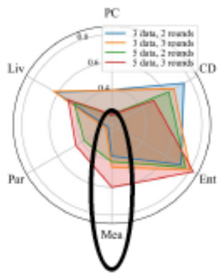
9. CAFQA: A classical simulation bootstrap for variational quantum algorithms

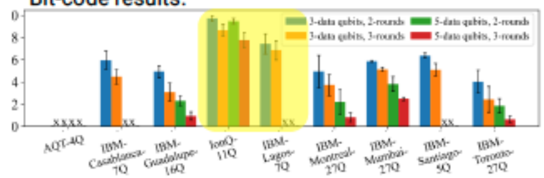


Gokul Subramanian Ravi, Pranav Gokhale, Yi Ding, William M. Kirby, Kaitlin N. Smith, Jonathan M. Baker, Peter J. Love, Henry Hoffmann, Kenneth R. Brown, Frederic T. Chong. ASPLOS, March 2023.

10. Benchmarking

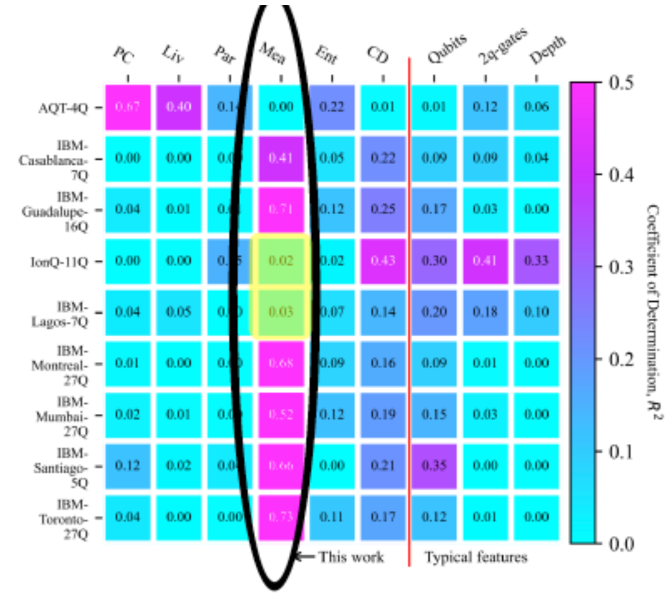
- Mid-circuit measurements strain current devices





- Ratio between readout and T1 time is crucial:

QPU	Readout time (μs)	T1 Coherence time (μs)	$T1/Readout$
IBM-Lagos	0.704	132.07	187.6
IBM-Santiago	4.0	89.31	22.3
IBM-Mumbai	3.6	134.62	37.4
IBM-Montreal	5.2	104.14	20.0
IBM-Guadalupe	5.4	99.52	18.4
IBM-Casablanca	5.9	91.21	15.5
IonQ	200000	10000000	50.0



28

Teague Tomesh, Pranav Gokhale, Victory Omole, Gokul Subramanian Ravi, Kaitlin Smith, Joshua Viszlai, Xin-Chuan Wu, Nikos Hardavellas, Margaret R. Martonosi, and Frederic T. Chong. SupermarQ: A Scalable Quantum Benchmark Suite. IEEE International Symposium on High-Performance Computer Architecture, 2022 **(Best Paper)**

Future Directions

- Scalability with selective deep optimization
- Design and optimization of hybrid quantum-classical applications
- Better debugging and verification
- Error mitigation to make machines more reliable and consistent
- Out-of-the-box HW-SW codesign: qudits, heterogeneous systems, multi-qubit/global operation, analog computation/simulation

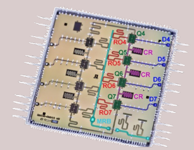
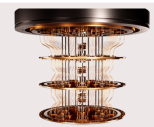
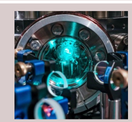
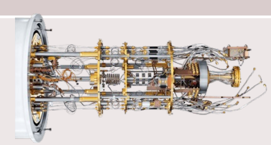
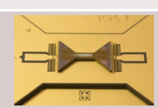
epiqc.cs.uchicago.edu

Infleqtion.com

THE UNIVERSITY OF CHICAGO

Infleqtion

Hardware Targets

Platform	Type	Techniques
Advanced Quantum Testbed @ Berkeley Lab	Superconducting (fixed frequency transmon) 	- approximate synthesis - Bring Your Own Gateset - phased microwave (PMW) - qutrit decomposition
IBM Quantum	Superconducting (fixed frequency transmon) 	- cross-gate pulse cancellation - dynamical decoupling - fractional gates
Infleqtion Hilbert	Neutral Atom (Cesium) 	- global gate decomposition - scheduling
Rigetti Aspen M-3	Superconducting (modulated transmon) 	- dynamical decoupling
QSCOUT @ Sandia Lab	Trapped ion (171Yb+) 	- SWAP Mirroring - fractional gates