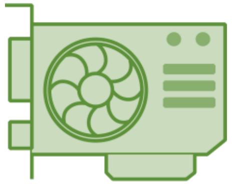
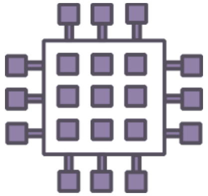
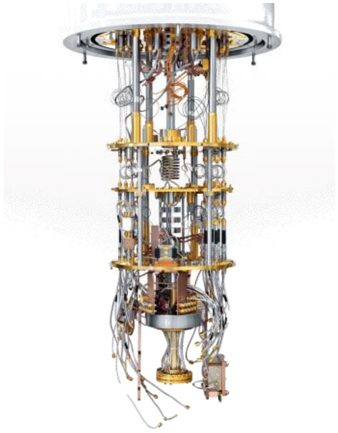
Quantum Pathways: Workshop on Quantum Information Science and Engineering

Partly supported by NSF: CyberTraining: Pilot: Quantum Research Workforce Development on End-to-End Quantum Systems Integration



Research Path from Classical Computing to Quantum Computing





CPU/GPU/FPGA

QC

Weiwen Jiang, Ph.D.

Assistant Professor

Electrical and Computer Engineering

George Mason University

wjiang8@gmu.edu

https://jqub.ece.gmu.edu

Speaker

Weiwen Jiang
Assistant Professor
Electrical and Computer Engineering (ECE)
George Mason University
Room3247, Nguyen Engineering Building
wjiang8@gmu.edu
(703)-993-5083
https://jqub.ece.gmu.edu/

- **Education Background**
  - **Chongqing University (2013-2019)**
  - **University of Pittsburgh (2017-2019)**
  - **University of Notre Dame (2019-2021)**
- **Research Interests**
  - **Optimization**
  - **HW/SW Co-Design**
  - **Quantum Learning**

First HW/SW Co-Design Framework using NAS

HW/SW Co-Design Framework
FNAS
[DAC'19*]
[TCAD'20*]

Application

- **Medical Imaging**: NAS for Medical Image Seg. [MICCAI'20]; 3D Cardiac MRI Seg. [ICCAD'20]
- **NLP (Transformer)**: FPGA [ICCD'20]; Mobile [DAC'21]; GPU [GLSVLSI'21]
- **Graph-Based**: Social Net [GLSVLSI'21]; Drug Discovery [ICCAD'21]

Algorithm

- **NAS Acc.**: HotNAS [CODES+ISSS'20]
- **Model Compression**: NAS for Quan. [ICCAD'19]; Compre.-Compilation [IJCAI'21]
- **Secure Infernece**: NASS [ECAI'20]; BUNET [MICCAI'20]

Hardware

- **FPGA**: XFER [CODES+ISSS'19*]
- **ASIC**: NANDS [ASP-DAC'20*]; ASICNAS [DAC'20]
- **Computing-in-Memory**: Device-Circuit-Arch. [IEEE TC'20]

Best Paper Award:

IEEE
IEEE Council on Electronic Design Automation
hereby presents the
2021 IEEE Transactions on Computer-Aided Design
Donald O. Pederson Best Paper Award
to
Weiwen Jiang, Lei Yang, Edwin Hsing-Mean Sha, Qingfeng Zhuge, Shouzhen Gu, Sakyasingha Dasgupta, Yiyu Shi, Jingtong Hu
for the paper entitled
"Hardware/Software Co-Exploration of Neural Architectures"

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC. FOUNDED NEW YORK 1884

Yao-Wen Chang
President
IEEE Council on Electronic Design Automation

Rajesh Gupta
Editor-in-Chief
IEEE Transactions on Computer-Aided Design

CEDA
IEEE Council on Electronic Design Automation

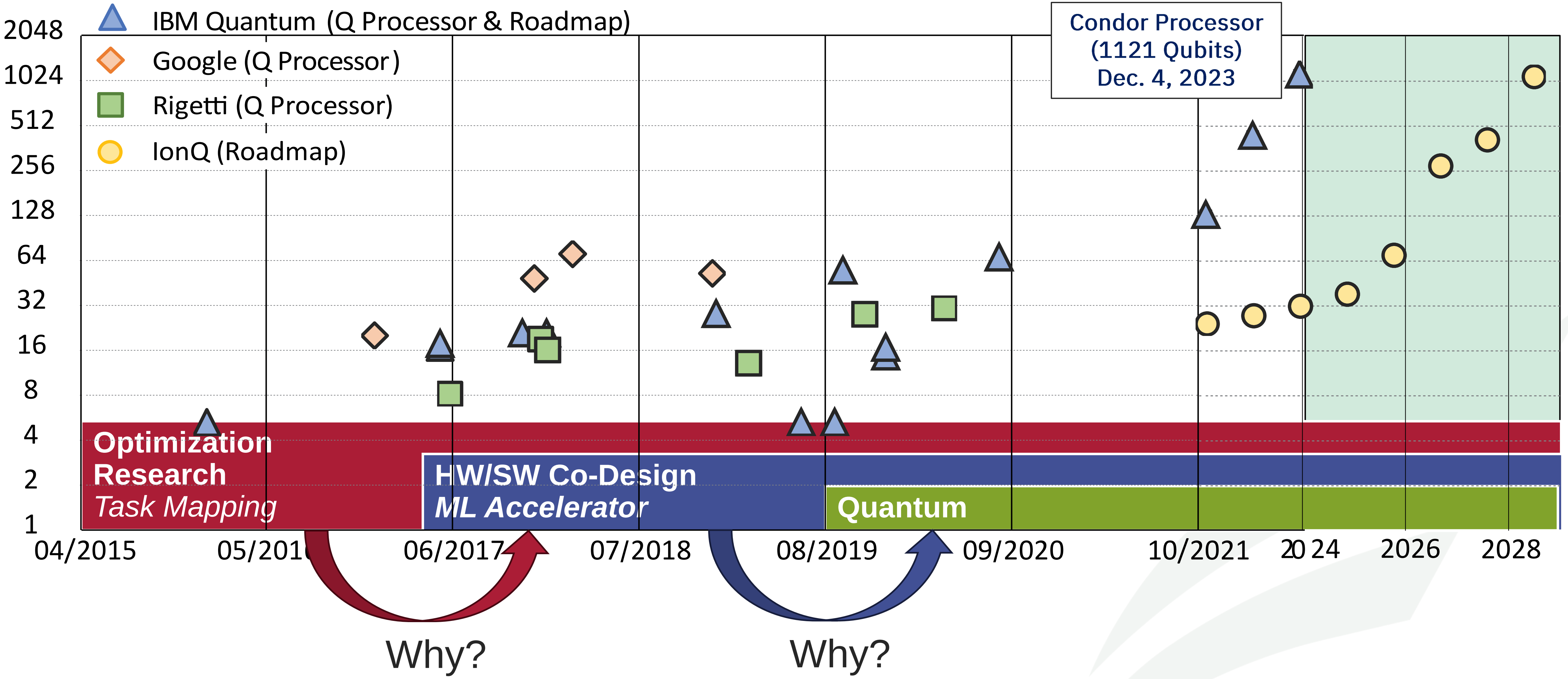
Best Paper Nominations:

58 DESIGN AUTOMATION CONFERENCE
FROM CHIPS TO SYSTEMS — LEARN TODAY, CREATE TOMORROW

EMBEDDED SYSTEMS WEEK

ASIA SOUTH PACIFIC DAC DESIGN AUTOMATION CONFERENCE

Research Path was Shifting along with the Growth of Quantum



What is Classical AI Democratization & What is the Challenge?

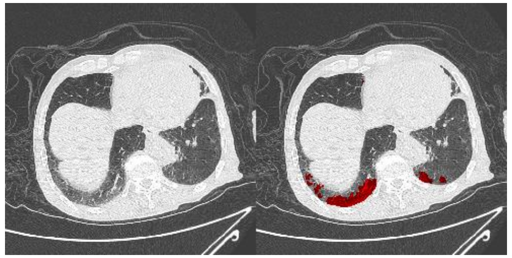
"It's here to collaborate, to augment, to enhance human lives and productivity and make everybody's life better. And related to that, is to **democratize A.I.** in a way that everybody gets benefit. Not just a few, or a selected group." **Fei-Fei Li, 2017**

Medical AI Scenario

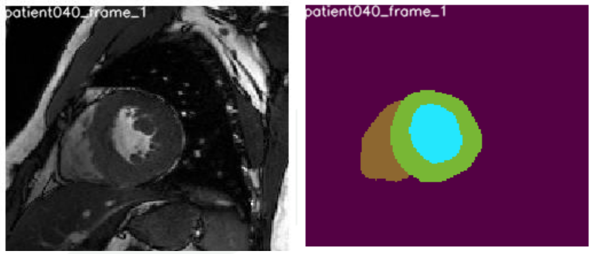
AR/VR in Surgery

Medical Diagnosis

AI Can Perform Medical Tasks



COVID CT Segmentation



Real-Time MRI Segmentation

Let Doctors Design Neural Networks?

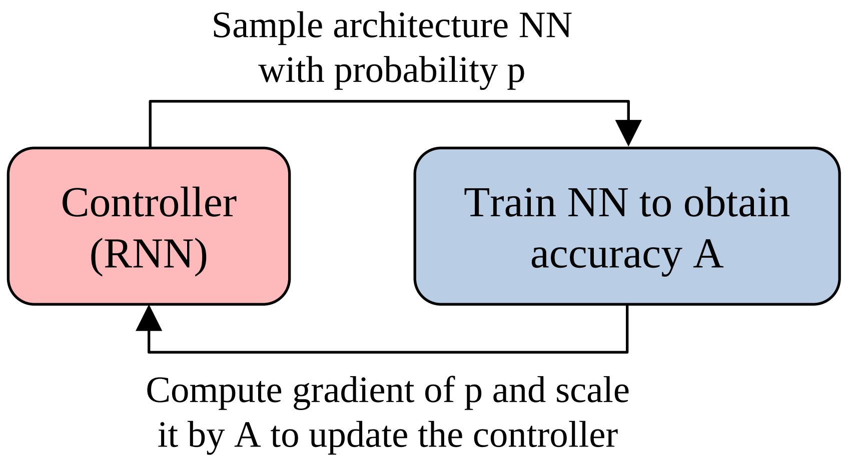
Progress of Classical AI Democratization

Google's Initial Contributions
(Neural Architecture Search)

Given: Dataset

Objective:
- Automated search for NN **(w/o human)**
- Maximize accuracy on the given dataset

Output: A neural network architecture



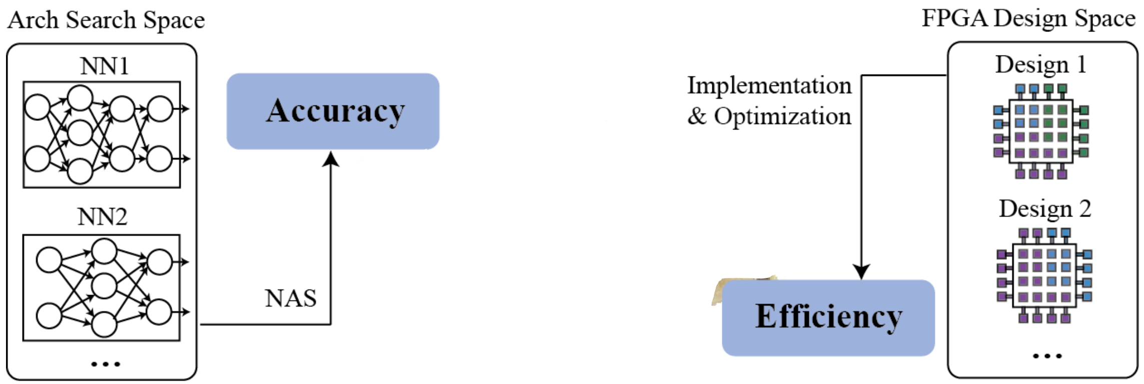
[ref] Zoph, Barret, and Quoc V. Le. "Neural architecture search with reinforcement learning." *ICLR 2017*

Our Contributions
(Network-Accelerator Co-Design)

Given: (1) Dataset; (2) Target hardware, e.g., FPGA.

Objective:
- Automated search for NN and HW design
- Maximize accuracy on the given dataset
- Maximize hardware efficiency

Output: A pair of neural network and hardware design

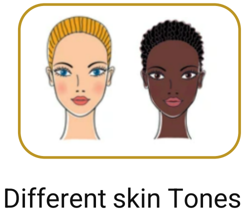


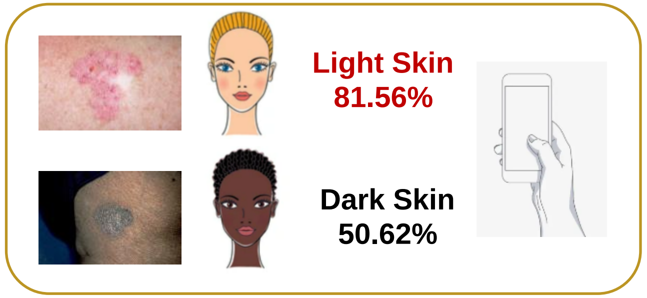
[ref] Jiang, Weiwen, et al. "Accuracy vs. efficiency: Achieving both through fpga-implementation aware neural architecture search." *DAC 2019*. (BEST PAPER NOMINATION)
[ref] Jiang, Weiwen, et al. “Hardware/software co-exploration of neural architectures”, TCAD 2020 (BEST PAPER AWARD)

On-Going Research: System-Support AI for Science

NSF 2027539: RAPID: Collaborative Research: Independent Component Analysis Inspired Statistical Neural Networks for 3D CT Scan Based Edge Screening of COVID-19. ($98,349 in total, **Co-PI** with share $49,174)

Problem and Challenge

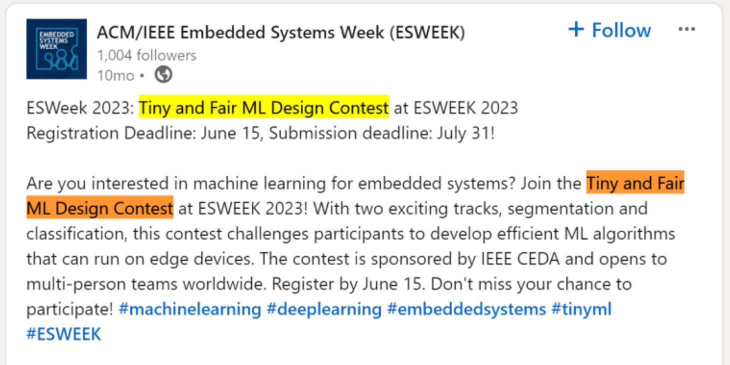




Solutions

- The Larger The Fairer? Small Neural Networks Can Achieve Fairness for Edge Devices --- **DAC 2022**
- Ensemble Learning for Multi-Dimension AI Fairness --- **DAC 2023**
- ViT-CNN for AI fairness --- **ICCAD 2023**

Community building: *Chair* and *Create* ML Contest at ESWEEK'23



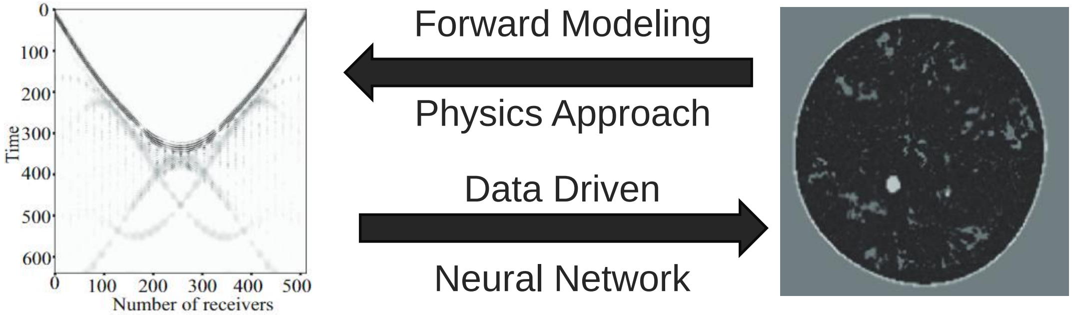
ACM/IEEE Embedded Systems Week (ESWEEK) + Follow ···
1,004 followers
10mo •

ESWeek 2023: Tiny and Fair ML Design Contest at ESWEEK 2023
Registration Deadline: June 15, Submission deadline: July 31!

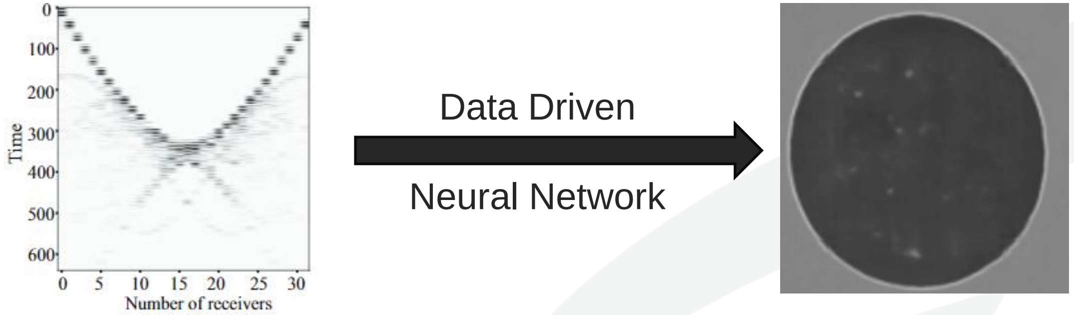
Are you interested in machine learning for embedded systems? Join the Tiny and Fair ML Design Contest at ESWEEK 2023! With two exciting tracks, segmentation and classification, this contest challenges participants to develop efficient ML algorithms that can run on edge devices. The contest is sponsored by IEEE CEDA and opens to multi-person teams worldwide. Register by June 15. Don't miss your chance to participate! #machinelearning #deeplearning #embeddedsystems #tinyml #ESWEEK

- 2 Tracks
- 72 Teams
- 702 Submissions
- 6 Winners

Problem: Ultrasound CT Scan



Challenge: Sparse data in USCT



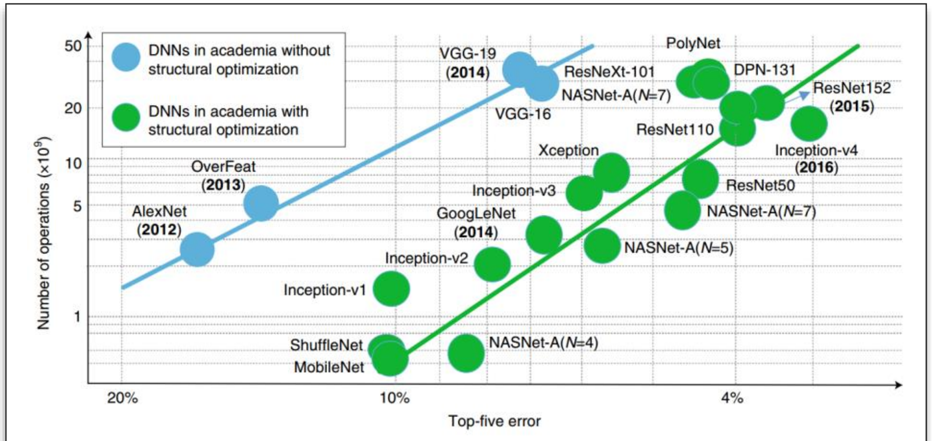
Solution:

- Physics-guided AI for co-optimize model and data --- **Submitted to MICCAI**

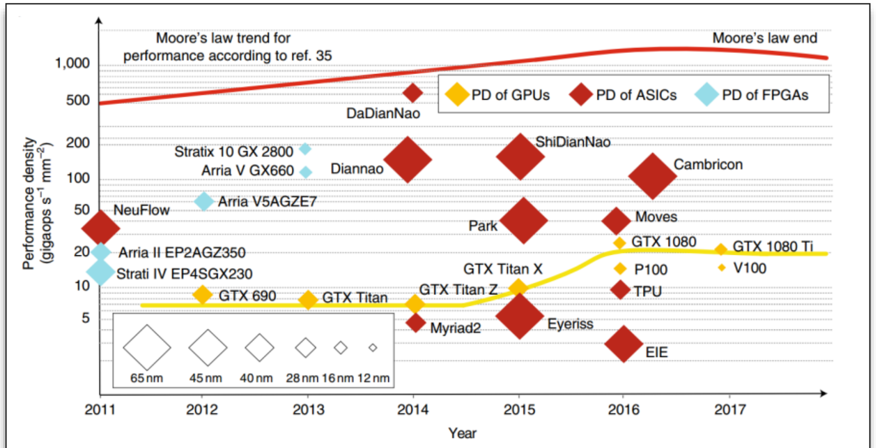
Student

My first Ph.D. student, Yi Sheng.
Estimate graduation: Spring 2025

Bottlenecks in Classical Computing



Deep neural network grows exponentially



Perf. of classical computing stops increasing

Medical AI Scenario: (Input size exponentially grows from Radiology to Pathology Imaging)

Radiology Imaging

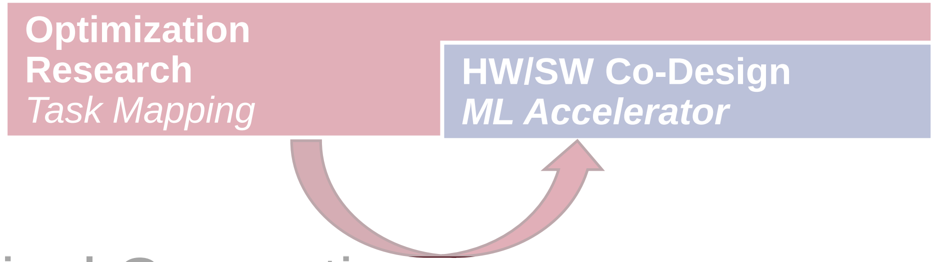
Radiology Modality	Avg. Size (MB)
CT Scan	**153.4**
MRI	98.6
X-ray angiography	157.5
Ultrasound	**69.2**
Breast imaging	**38.8**

Pathology Imaging

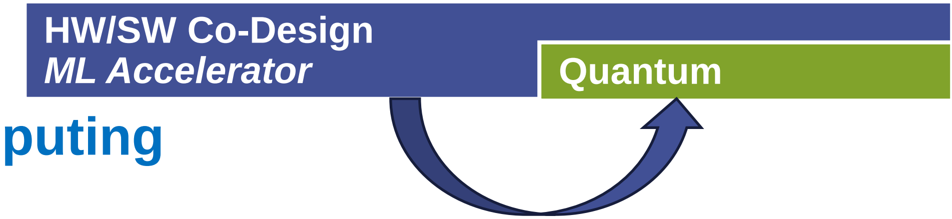
Biopsy Type	Compressed Size(MB)/Study	Original Size (GB)
Dermatopathology	1,392 (20x compression)	**27**
Head and neck	1,965 (20x compression)	**38**
Hematopathology	40,300 (40x compression)	**1574**
Neuropathology	1,872 (20x compression)	**37**
Thoracic pathology	3,240 (20x compression)	**63**

[ref] Lauro, Gonzalo Romero, et al. "Digital pathology consultations—a new era in digital imaging, challenges and practical applications." *Journal of digital imaging* 26.4 (2013).

Outline



- Background and Challenge in Classical Computing

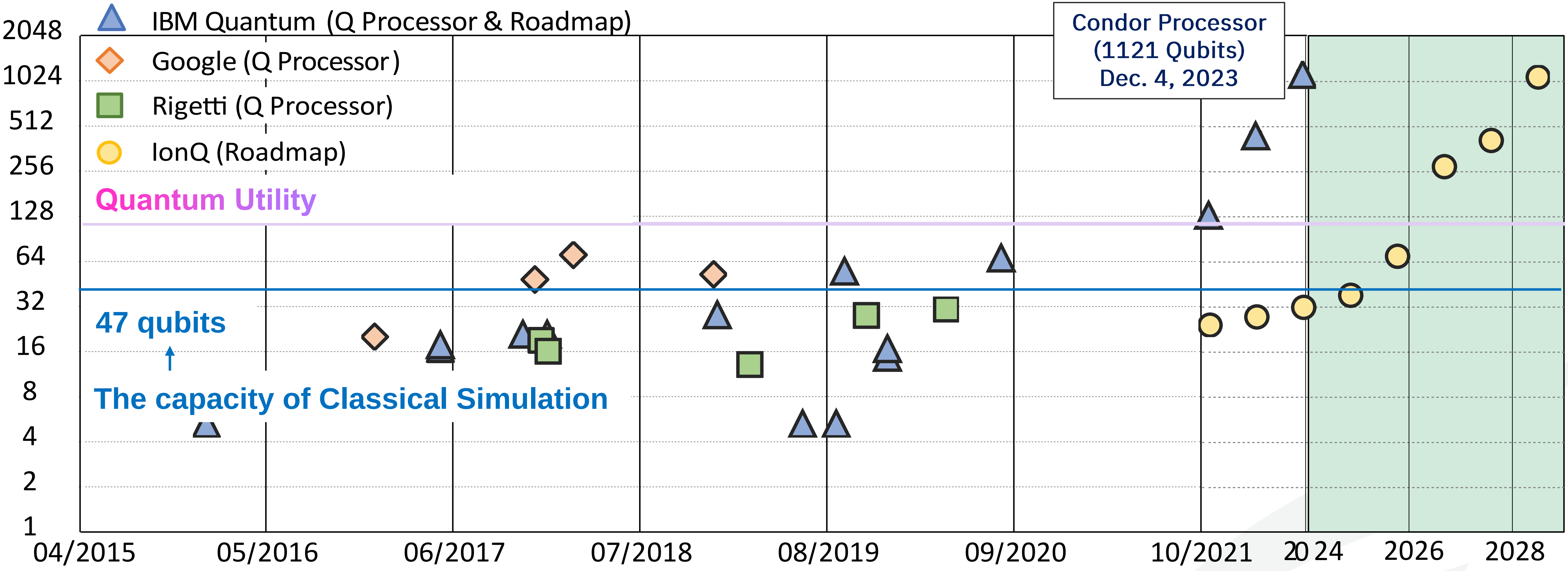


- **Potential of Quantum Computing**
  - **Tasks Impossible for Classical Computing**
  - **Today’s Quantum Computers**

- Research Examples: different problem, same fundamentals
  - Classical ML Accelerator vs. Quantum ML Accelerator
  - Compression in Classical ML and Quantum ML

- Messages to Send

Potential: Tasks Impossible for Classical Computing



The maximum qubits that supercomputers can simulate for arbitrary circuits is less than 47 qubits.

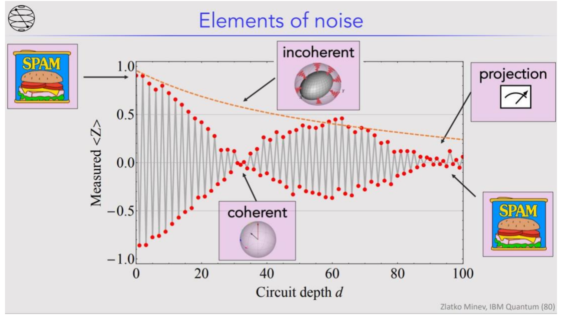
(1) Summit w/ 2.8 PB memory for **47 qubits**; (2) Sierra w/ 1.38 PB memory for **46 qubits;**

(3) Sunway TaihuLight w/ 1.31 PB memory for **46 qubits**; (4) Theta w/ 0.8 PB memory for **45 qubits.**

[ref] Wu, Xin-Chuan, et al. "Full-state quantum circuit simulation by using data compression." Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2019.

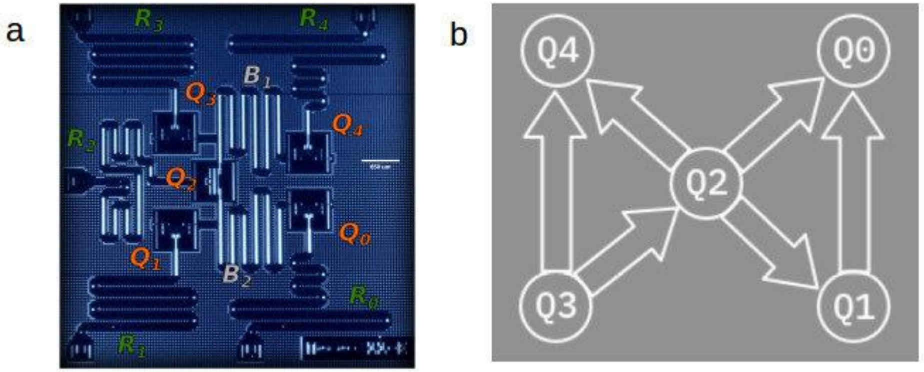
What's the Status of Today's Quantum Computers?

Let's See What Happen at IBM Quantum Summit 2023 (Dec 4, 2023)



NISQ Era @ 2017

Era of Utility @ 2023



What's the Status of Today's Quantum Computers?

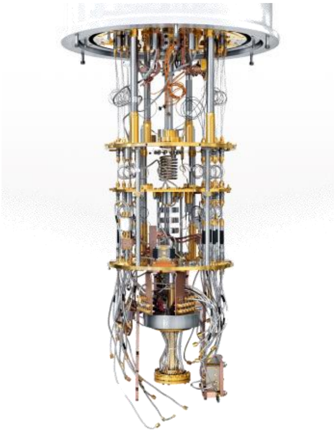
Let's see what Happen at IBM Quantum Summit 2023 (Dec 4, 2023)

What's the Status of Today's Quantum Computers?

What is the Meaning by Quantum Utility --- From IBM Quantum Summit

The Era of Utility means a focus on **performance**, **stability** and **reliability**

The Era of Utility means **new users**, and **new tools**



QC

QC Scientist

Takeaway

- **Quantum Utility Era**
  is now coming
- **Performance, stability, and reliability**
  are keys to achieve Quantum Utility

Junction of Quantum-Classical Computer-Aided Design Lab (JQub)

- **Domain users**
  are expected to use quantum computers

Let Doctors Design Neural Networks?

Let Doctors Learn Quantum Computing?

IBM Quantum System I

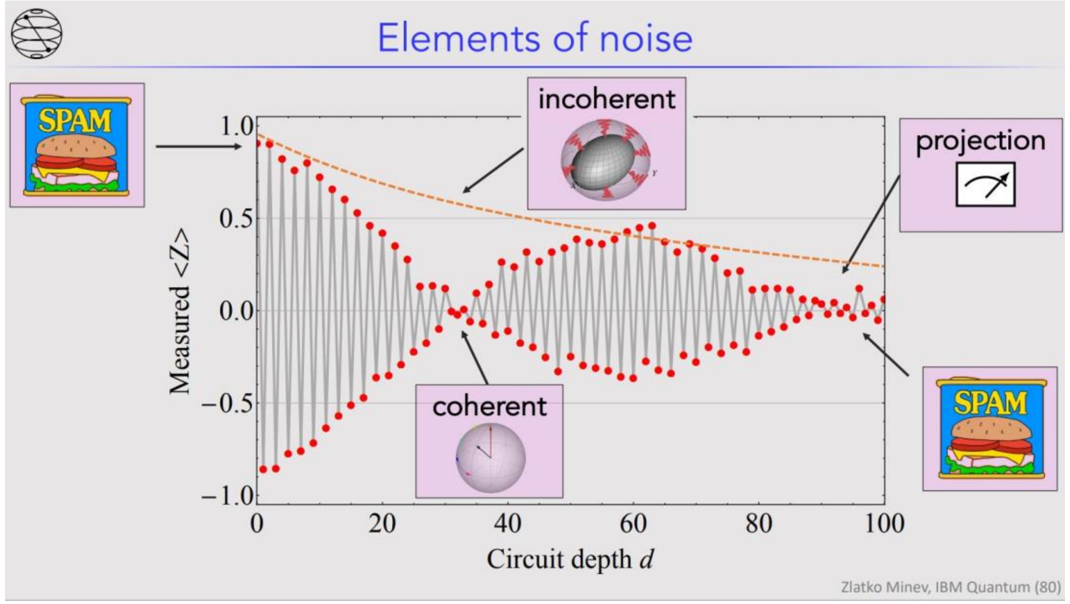
IBM Quantum System II

Outline

- Background and Challenge in Classical Computing
- Potential of Quantum Computing
  - Tasks Impossible for Classical Computing
  - Today's Quantum Computers
- **Research on Quantum Computing @ JQub**
  - **Performance, Stability, and Reliability**
  - Domain-specific Quantum Computing
- Messages to Send

Noise Changes the Optimization Surface

Ref: Zlatko K. Minev, IBM Quantum



Our Observation:

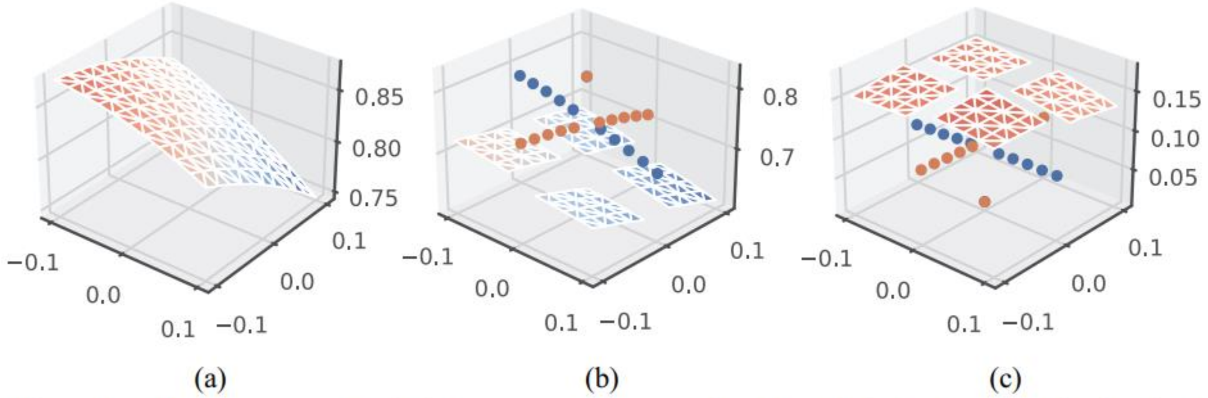


Fig. 3. Noise-aware training may miss optimal solution: (a) Optimization surface of 2-parameter VQC under noise free environment. (b) Optimization surface of the same VQC under a noisy environment. (c) Difference between (a) and (b).

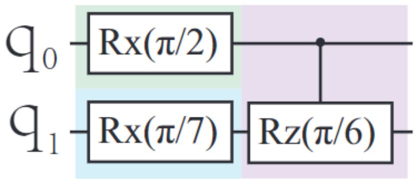
Insight:

- **Shorter circuit has higher fidelity.**

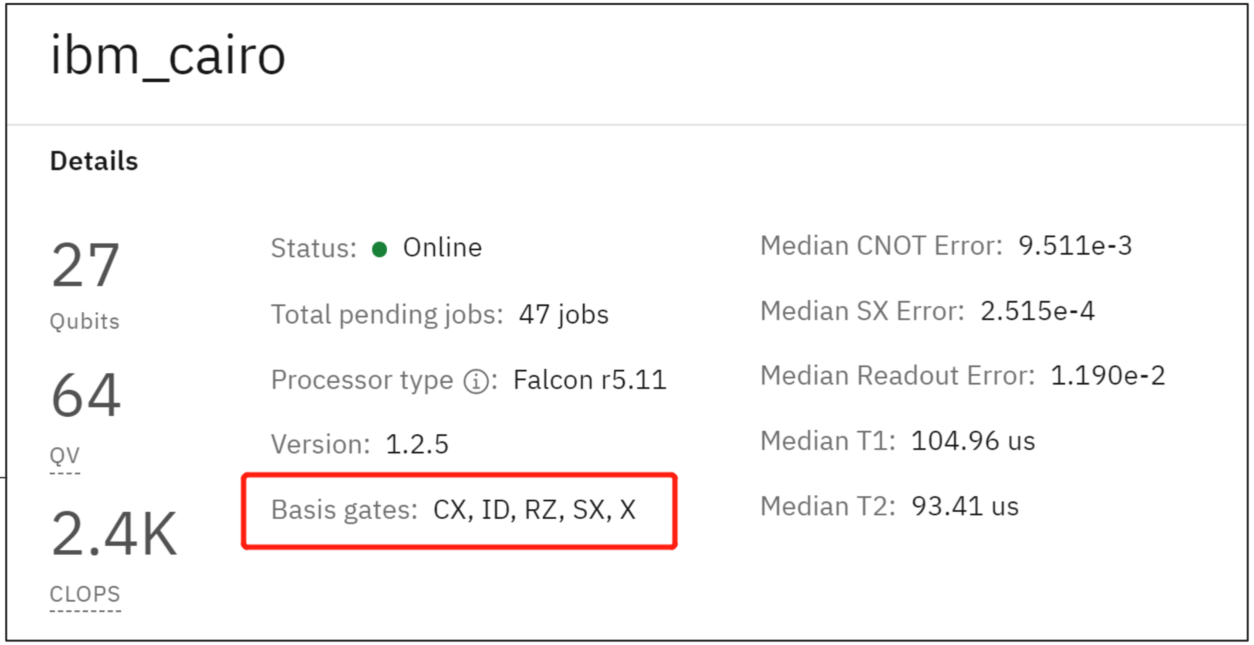
Question:

- **What are the breakpoints in the noisy optimization landscape?**

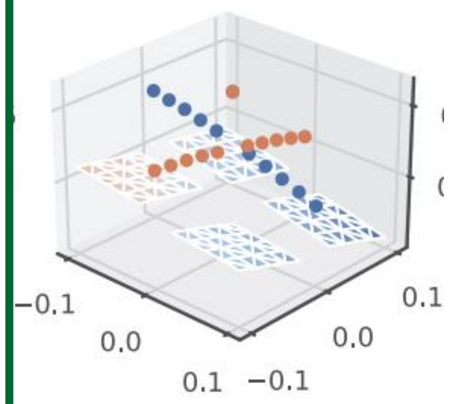
Motivation: Parameters Affect Circuit Length through Compilation

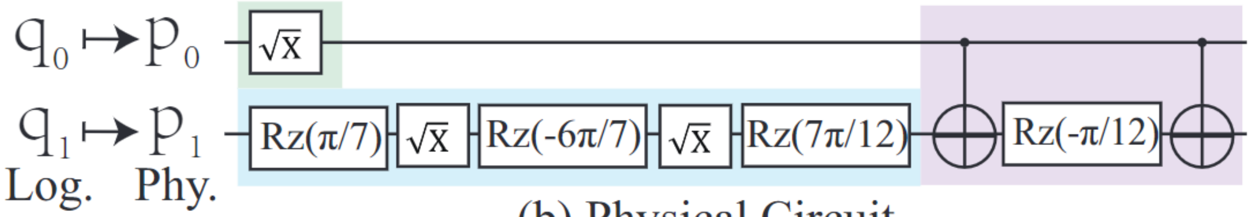


(a) Logical Circuit



Breakpoints?





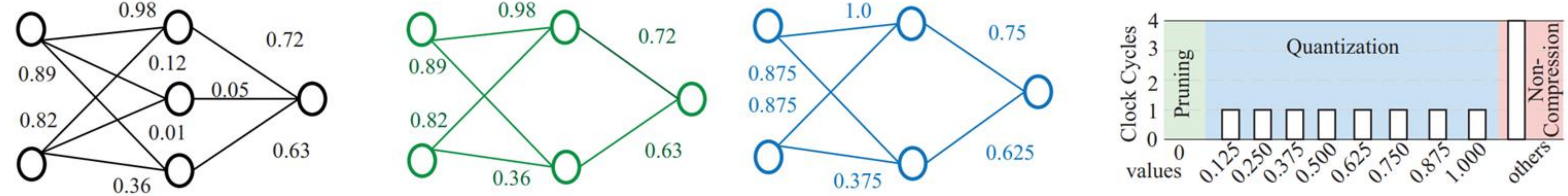
(b) Physical Circuit

Table 1: circuit depth of compiled quantum gates on IBM quantum processors; parameters are in the range of $[0, 4\pi]$

Gate	0	π	2π	3π	4π	$\pi/2$	$3\pi/2$	$5\pi/2$	$7\pi/2$	others
RX	0	1	0	1	0	1	3	1	3	5
RY	0	2	0	2	0	3	3	3	3	4
CRX	0	8	5	9	0	11	11	11	11	11
CRY	0	8	6	8	0	10	10	10	10	10

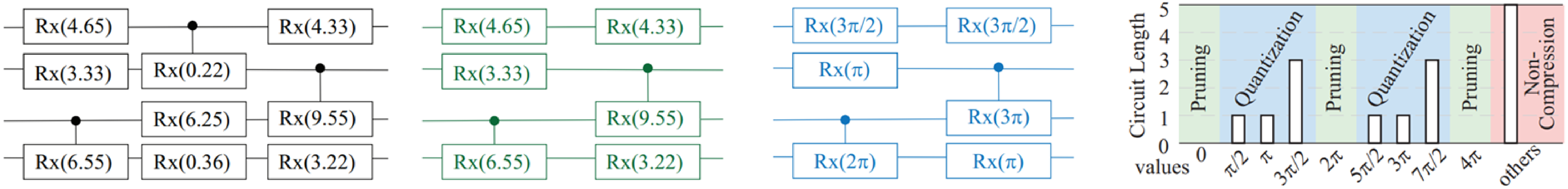
Quantum Neural Network Compression @ ICCAD'2022

- Model compression in **Classical ML** **is to improve hardware efficiency**



(a) Non-Compression Classical NN (b) Classical NN with Pruning (c) Pruned NN with Quantization (d) Cost of Different Levels in Classical NN

- Model compression in **Quantum ML** **can reduce circuit length, and thus, further provide high fidelity**



(e) Non-Compression QNN (f) QNN with Pruning (g) Pruned QNN with Quantization (h) Cost of Different Levels in RX Gate in QNN

CompVQC Framework: Experiment Results

Datasets		Syn-Dataset-4		Syn-Dataset-16	
Compression Method		Acc. (vs. Baseline)	TCD (Speedup)	Acc. (vs. Baseline)	TCD (Speedup)
Qiskit Aer	Vanilla VQC	94%(0)	23(0)	96%(0)	51(0)
	Comp-VQC	99%(5%)	11(2.09×)	98%(2%)	23(2.22×)
IBM Q	Vanilla VQC	79%(-15%)	23(1.00×)	86%(-10%)	51(1.00×)
	Comp-VQC	99%(5%)	11(2.09×)	98%(2%)	23(2.22×)

Acc.(vs. Baseline)	ibm_lagos	ibm_perth	ibm_jakarta
Vanilla VQC(TCD=23)	79%(0)	86%(0)	92%(0)
CompVQC(TCD=11)	99%(20%)	98%(12%)	100%(8%)

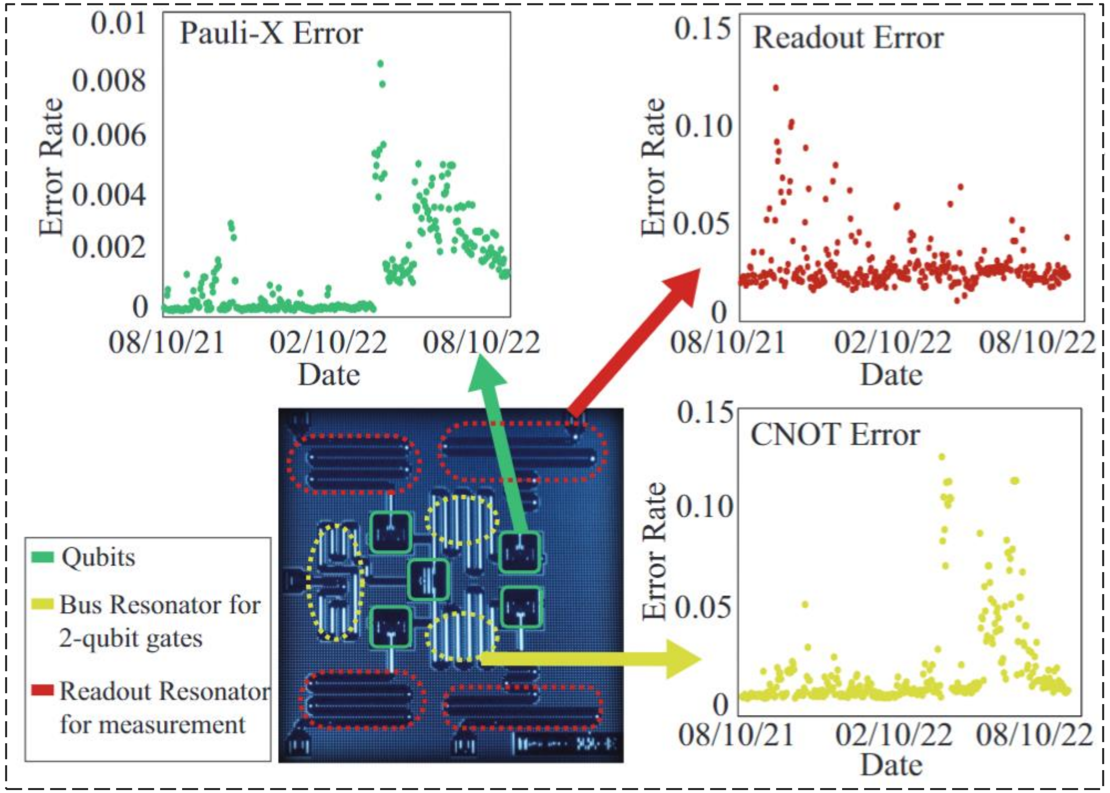
- CompVQC can reduce circuit length by **2X**
- The accuracy is higher in a noisy environment

Insights:

✓ **CompVQC can improve robustness of QNN**

× **CompVQC is not aware of noise**

Unstable Quantum Noise Leads to Performance Changes

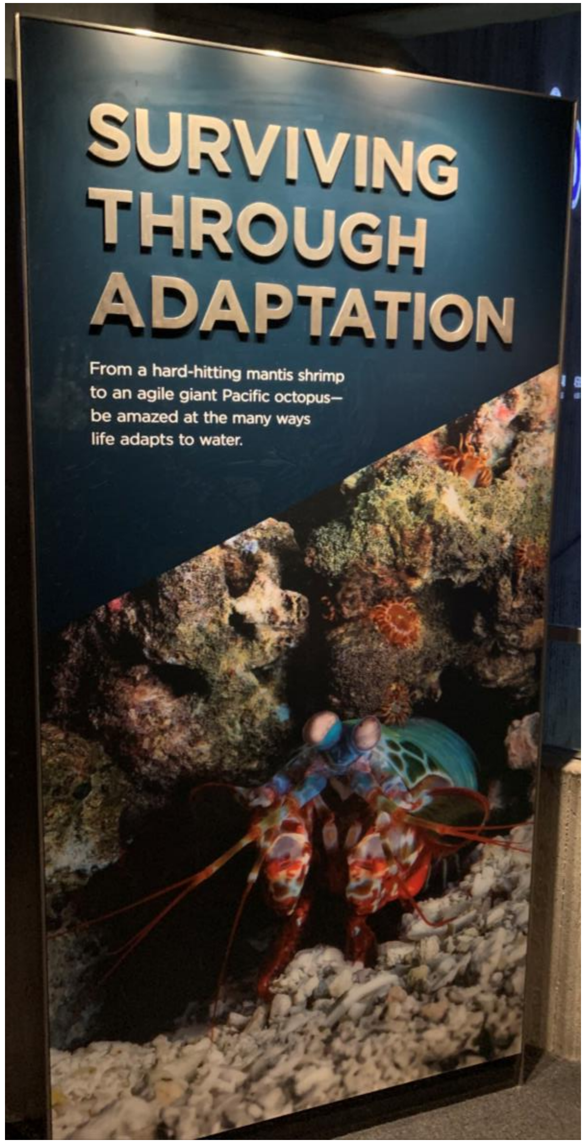


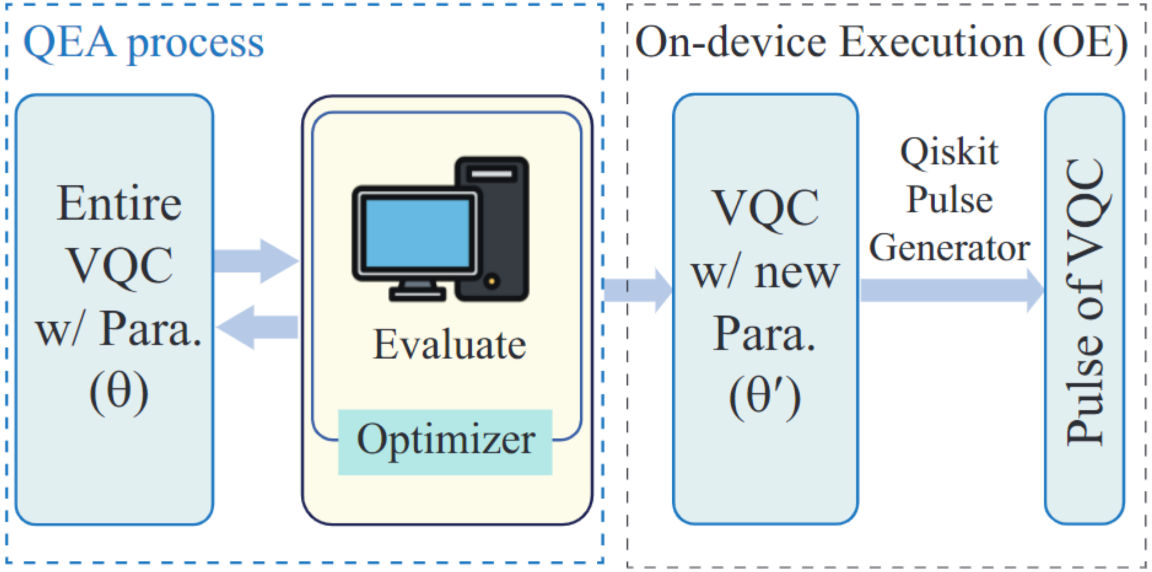
(a) Fluctuating quantum noise on real quantum computer (1-year long daily profiling)

Insight:

- **Temporal reproducibility or reliability of a quantum learning model.**
- **Users may not be aware of the performance changes.**

Quantum Error Adaptation @ DAC 2023

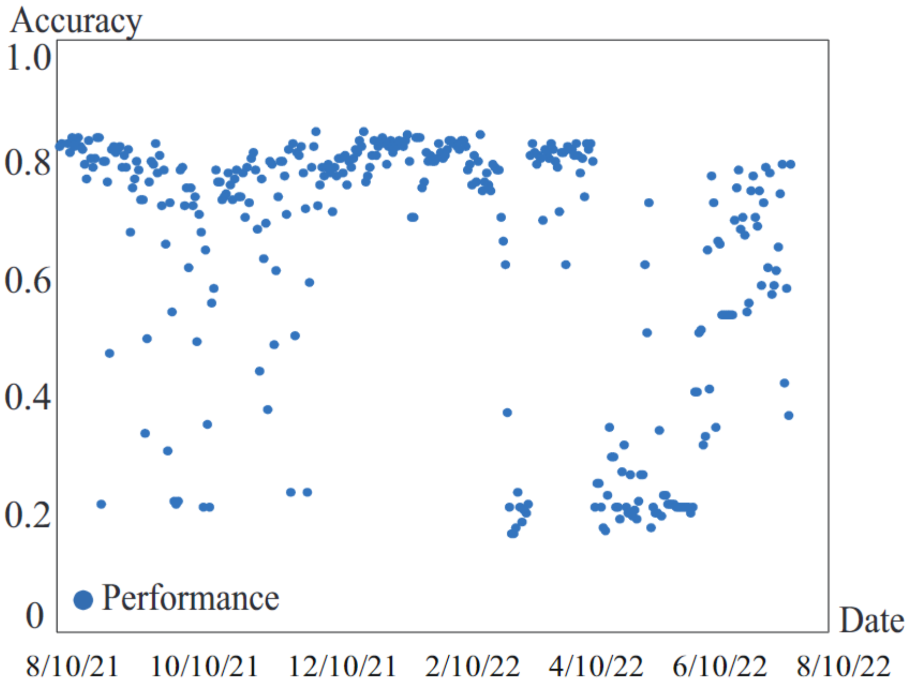


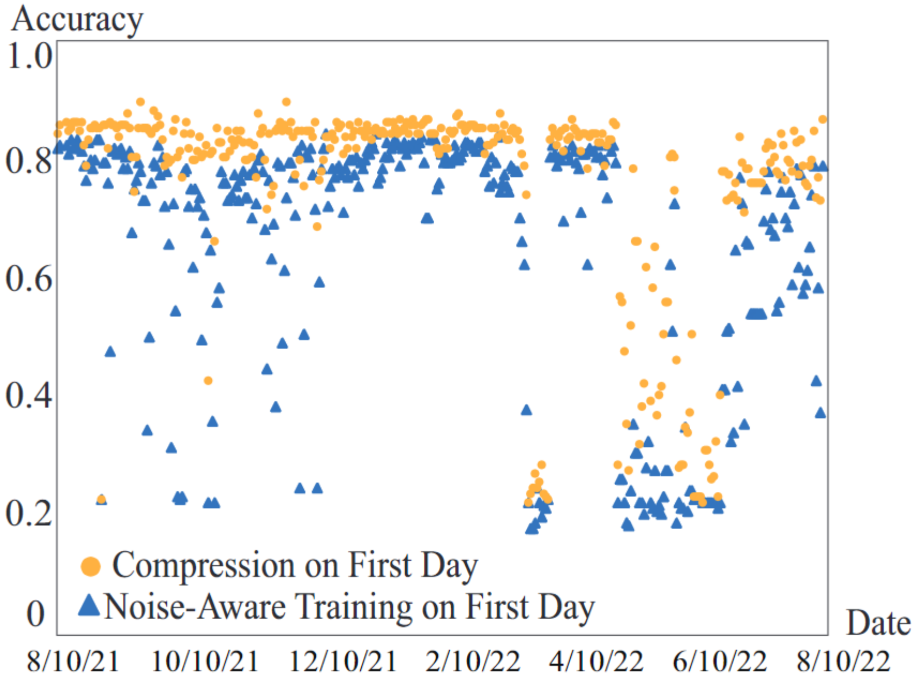


Fluctuating Quantum Noise

Observation: Fluctuating noise can **collapse the model accuracy** of a noise-aware trained QNN model

Observation: **Compression can boost the performance** of QNN than noise-aware training, but **not enough**



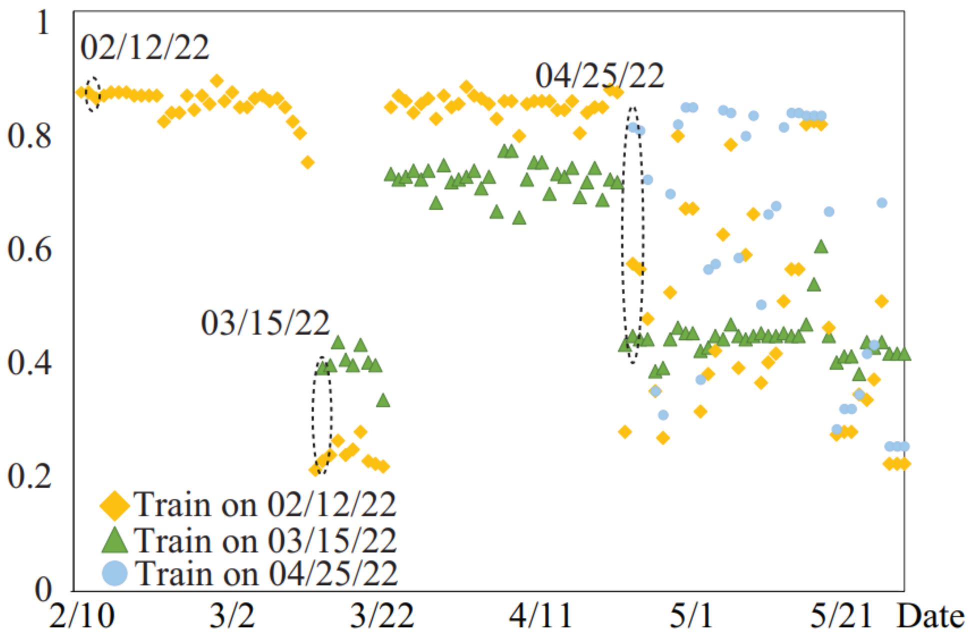


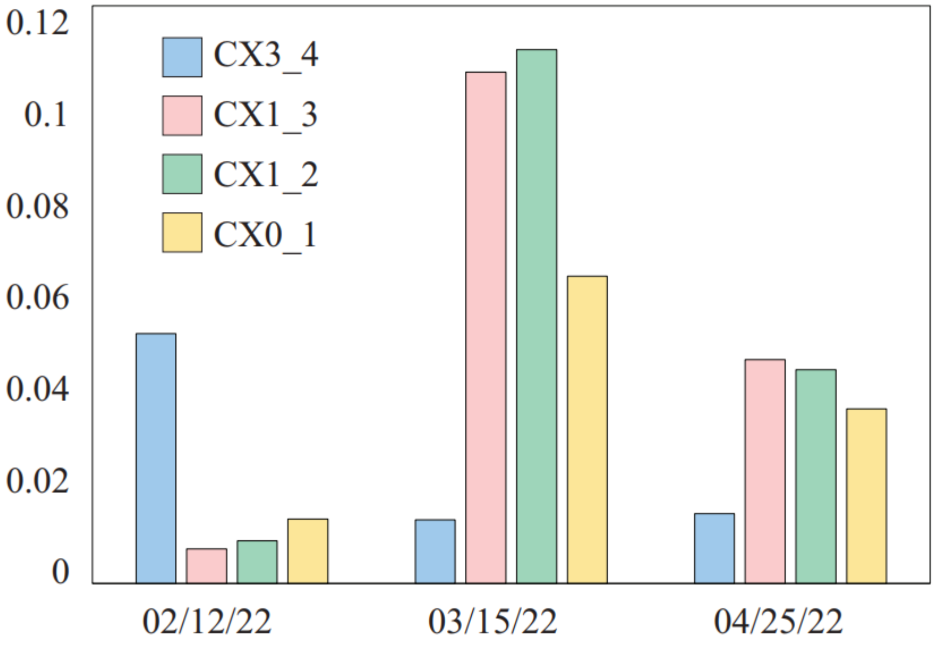
The accuracy of QNN on 4-class MNIST from August 2021 to August 2022 on IBM backend belem using Qiskit Simulation.

Battle Against Fluctuating Quantum Noise

Observation: Models **Compressed on one noise level** have different performance on different days

Observation: Models **Compressed on different noise levels (dates)** have different performance on the same day





➡ ① Noise aware compression ② Model Repository

Battle Against Fluctuating Quantum Noise

Solution: Offline + Online

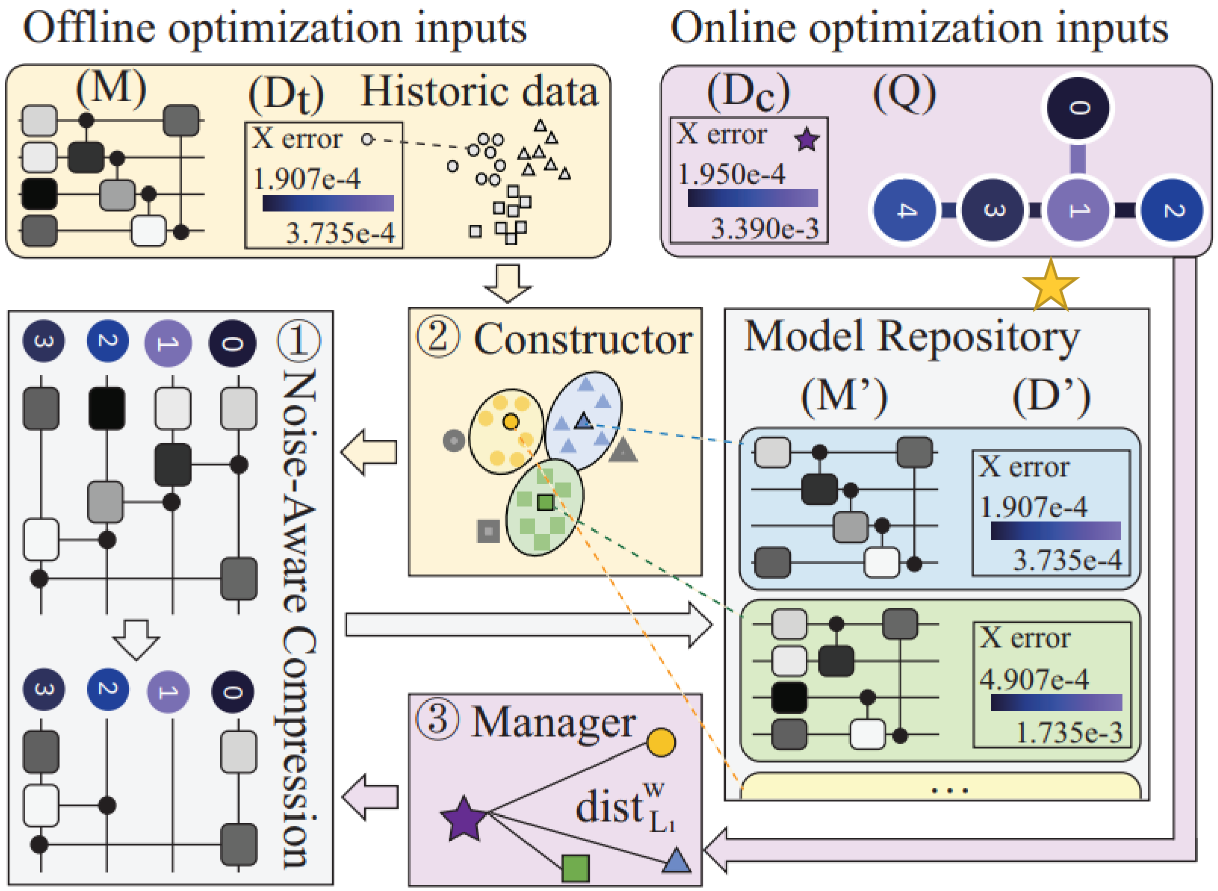


Fig. 5. Illustration of the proposed Compression-Aided Framework (QuCAD).

Offline:

- Use historic data to construct a repository by clustering

Online:

- Select a model to do inference
- Maintain the repository: whether to generate new models into the model repository manager

QuCAD: Experiment Results

- Significantly improve the **number of days for the desired accuracy**

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON 3 DATASETS IN CONTINUS 146 DAYS WITH FLUCTUATING NOISE.

CompVQC

Dataset	Method	Mean Accuracy	vs. Baseline	Variance	Days over 0.8	vs. Baseline	Days over 0.7	vs. Baseline	Days over 0.5	vs. Baseline
4-class MNIST	Baseline	59.35%	0.00%	0.070	24	0	93	0	100	0
	Noise-aware Train Once [4]	58.69%	-0.65%	0.060	8	-16	92	-1	100	0
	Noise-aware Train Everyday	59.39%	0.05%	0.070	28	4	83	-10	99	-1
	One-time Compression [15]	68.44%	0.00%	0.050	80	56	102	9	117	17
	QuCAD w/o offline	72.31%	12.96%	0.030	77	53	98	5	134	34
	QuCAD (ours)	**75.67%**	**16.32%**	**0.020**	**100**	**76**	**134**	**41**	**134**	**34**
Iris	Baseline	37.85%	0.00%	**0.006**	0	0	0	0	8	0
	Noise-aware Train Once [4]	54.38%	16.53%	0.043	29	29	46	46	70	62
	Noise-aware Train Everyday	56.62%	18.78%	0.044	38	38	56	56	72	64
	One-time Compression [15]	69.20%	31.36%	0.043	84	84	90	90	103	95
	QuCAD w/o offline	75.30%	37.46%	0.025	**84**	**84**	104	104	128	120
	QuCAD (ours)	**76.73%**	**38.88%**	0.015	83	83	**108**	**108**	**141**	**133**
Seismic Wave	Baseline	68.40%	0.00%	0.014	18	0	70	0	137	0
	Noise-aware Train Once [4]	68.85%	0.45%	0.014	19	1	78	8	137	0
	Noise-aware Train Everyday	68.28%	-0.11%	0.013	22	4	69	-1	138	1
	One-time Compression [15]	78.99%	10.59%	0.007	80	62	130	60	144	7
	QuCAD w/o offline	82.34%	13.95%	0.001	110	92	145	75	146	9
	QuCAD (ours)	**83.75%**	**15.36%**	**0.001**	**133**	**115**	**146**	**76**	**146**	**9**

Accuracy > 80% **(146 days in total)**

- MINST: **24 -> 100**
- Iris: **0 -> 84**
- Seismic wave: **18 -> 133**

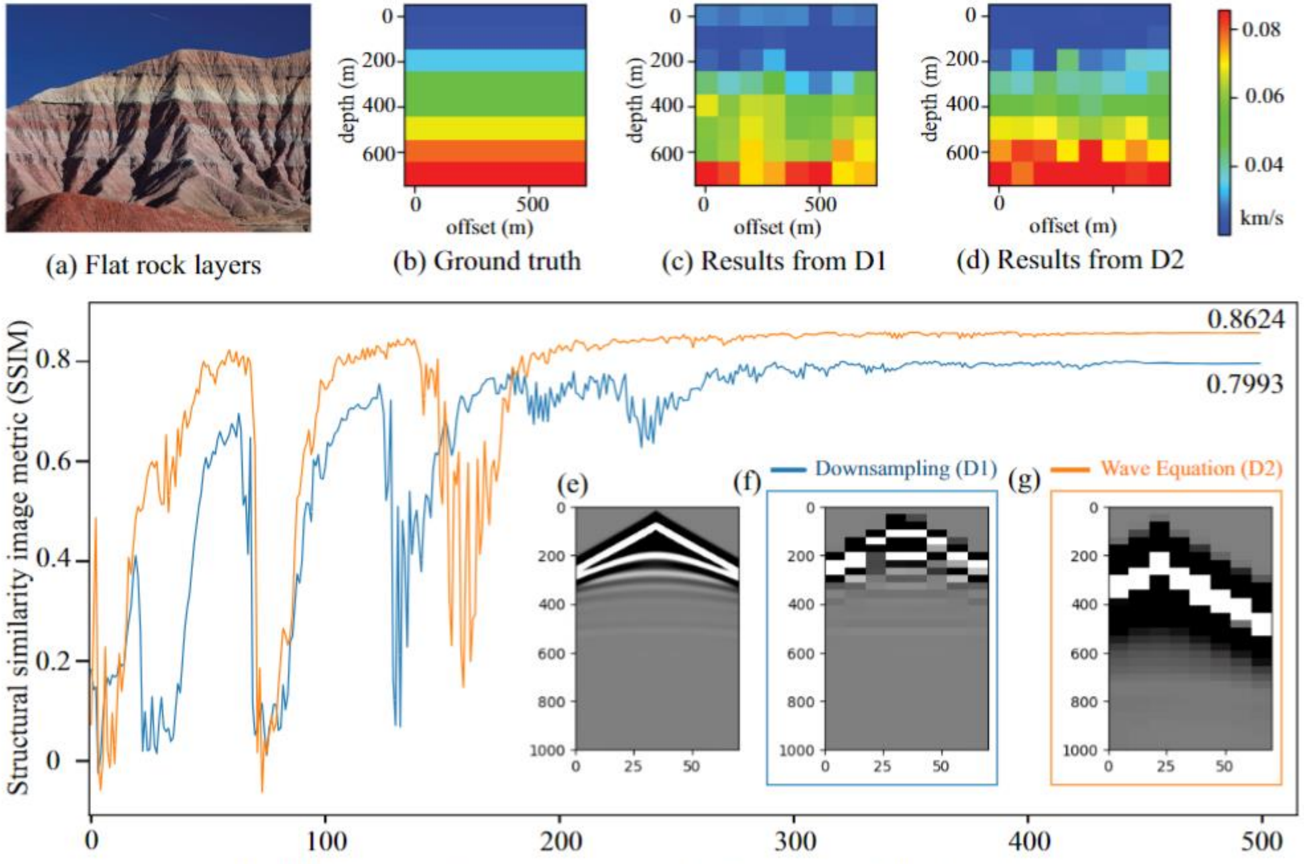
Accuracy > 70% **(146 days in total)**

- MINST: **93 -> 134**
- Iris: **0 -> 108**
- Seismic wave: **70 -> 146**

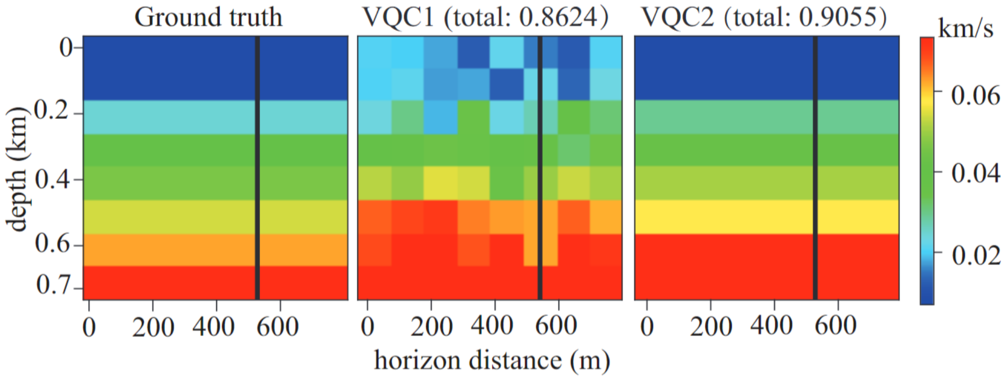
Outline

- Background and Challenge in Classical Computing
- Potential of Quantum Computing
  - Tasks Impossible for Classical Computing
  - Today's Quantum Computers
- **Research on Quantum Computing @ JQub**
  - Performance, Stability, and Reliability
  - **Domain-specific Quantum Computing**
- Messages to Send

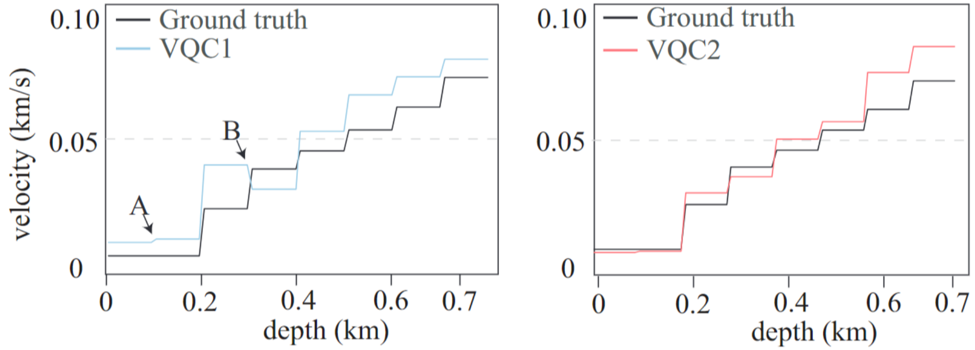
Quantum Learning for Geophysics @ DAC 2024



(h) SSIM metric along training epochs for the same model on two datasets



(a) Visualization of outputs from two VQCs



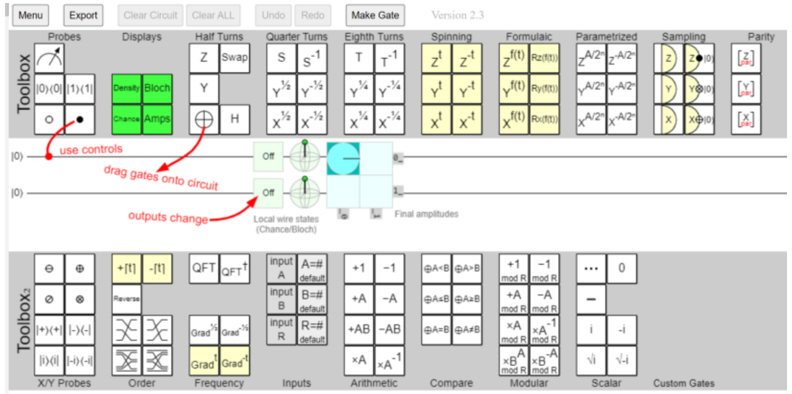
(b) Vertical velocity profiles of inversion results at $x = 500$ m

Outline

- Background and Challenge in Classical Computing
- Potential of Quantum Computing
  - Tasks Impossible for Classical Computing
  - Today's Quantum Computers
- Research on Quantum Computing @ JQub
  - Performance, Stability, and Reliability
  - Domain-specific Quantum Computing
- **Messages to Send**

Messages

- Research directions are **NOT independent** but **entangled**
  - e.g., **model compression** from Classical ML to Quantum ML
- **CS/CpE students CAN** significantly contribute to Quantum Computing
- **How to Take First step?** Online Tools and Tutorials!

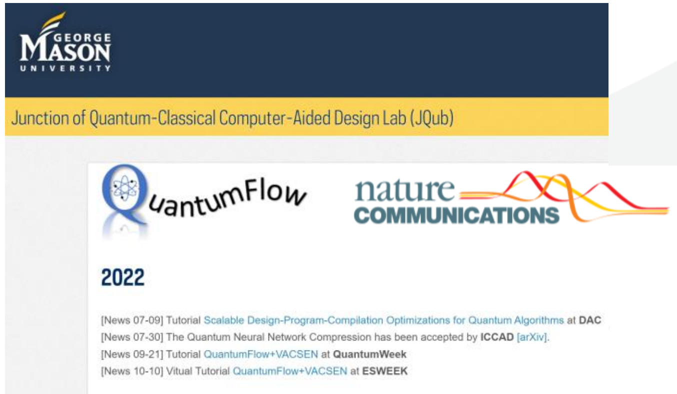


Quirk visible simulator

https://algassert.com/quirk

Our Contribution: VACSEN Noisy QC

https://vacsensystem.github.io/



Our Contribution: QuantumFlow Tutorial

https://jqub.ece.gmu.edu/categories/QF/

Acknowledge

Academia Students and Collaborators

National Lab

OAK RIDGE National Laboratory

Pacific Northwest NATIONAL LABORATORY

Los Alamos NATIONAL LABORATORY

Sponsors:

National Science Foundation

OAC-2311949: An Integrated Framework for Enabling Temporal-Reliable Quantum Learning on NISQ-era Devices

OAC-2320957: CyberTraining: Pilot: Quantum Research Workforce Development on End-to-End Quantum Systems Integration

Reference

[1] Zhirui Hu, Peiyan Dong, Zhepeng Wang, Youzuo Lin, Yanzhi Wang, Weiwen Jiang**,** Quantum Neural Network Compression, *In Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), 2022.*

[2] Zhirui Hu, Youzuo Lin, Qiang Guan, Weiwen Jiang**,** Battle Against Fluctuating Quantum Noise: Compression-Aided Framework to Enable Robust Quantum Neural Network, *In Proceedings of Design Automation Conference (**DAC**), 2023.*

[3] W. Jiang, J. Xiong, and Y. Shi, A Co-Design Framework of Neural Networks and Quantum Circuits Towards Quantum Advantage**, *Nature Communications***, *12, 579, 2021.*

[4] Z. Wang, Z. Liang, S. Zhou, C. Ding, J. Xiong, Y. Shi, W. Jiang, Exploration of Quantum Neural Architecture by Mixing Quantum Neuron Designs**,** *IEEE/ACM International Conference On Computer-Aided Design (**ICCAD**), 2021.*

[5] Z. Liang, Z. Wang, J. Yang, L. Yang, J. Xiong, Y. Shi, W. Jiang, Can Noise on Qubits Be Learned in Quantum Neural Network? A Case Study on QuantumFlow**,** *IEEE/ACM International Conference On Computer-Aided Design (**ICCAD**), 2021.*

[6] W. Jiang, J. Xiong, and Y. Shi, When Machine Learning Meets Quantum Computers: A Case Study**,** *in Proc. of Asia and South Pacific Design Automation Conference (**ASP-DAC**), 2021.*

[7] Zhiding Liang, Hanrui Wang, Jinglei Cheng, Yongshan Ding, Hang Ren, Zhengqi Gao, Zhirui Hu, Duane S. Boning, Xuehai Qian, Song Han, Weiwen Jiang, Yiyu Shi, Variational Quantum Pulse Learning**,** In *2022 IEEE International Conference on Quantum Computing and Engineering (**QCE**)*, pp. 556-565. IEEE, 2022.

[8] Shaolun Ruan, Yong Wang, Weiwen Jiang, Ying Mao, Qiang Guan, VACSEN: A Visualization Approach for Noise Awareness in Quantum Computing, *IEEE Transactions on Visualization and Computer Graphics (**VIS**), 2022*.

[9] Shaolun Ruan, Ribo Yuan, Yong Wang, Yanna Lin, Ying Mao, Weiwen Jiang, Zhepeng Wang, Wei Xu, Qiang Guan, Venus: A geometrical representation for quantum state visualization, ***EuroVis20***, *2023.*

[10] Hanrui Wang, Zhiding Liang, Jiaqi Gu, Zirui Li, Yongshan Ding, Weiwen Jiang, Yiyu Shi, David Z Pan, Frederic T Chong, Song Han,TorchQuantum Case Study for Robust Quantum Circuits, *IEEE/ACM International Conference On Computer-Aided Design (**ICCAD**), 2022.*

Reference

[11] W Jiang, Y Lin, QuGeo: An End-to-end Quantum Learning Framework for Geoscience--A Case Study on Full-Waveform Inversion, Accepted by DAC 2024.

[12] Z Hu, R Wolle, M Tian, Q Guan, T Humble, W Jiang, Toward Consistent High-fidelity Quantum Learning on Unstable Devices via Efficient In-situ Calibration, ***QuantumWEEK 2023*** (Best Paper Award)

[13] J Li, Z Wang, Z Hu, A Li, W Jiang, A novel spatial-temporal variational quantum circuit to enable deep learning on nisq devices, ***QuantumWEEK 2023***

[14] J Li, Q Guan, D Tao, W Jiang, Carbon Emissions of Quantum Circuit Simulation: More than You Would Think, ***IGSC 2023***

[15] Z Wang, J Li, Z Hu, B Gage, E Iwasawa, W Jiang, QuMoS: A Framework for Preserving Security of Quantum Machine Learning Model, ***QuantumWEEK*** 2023

GEORGE MASON UNIVERSITY

wjiang8@gmu.edu

George Mason University

4400 University Drive
Fairfax, Virginia 22030

Tel: (703)993-1000