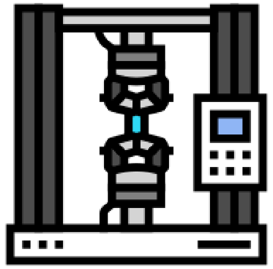
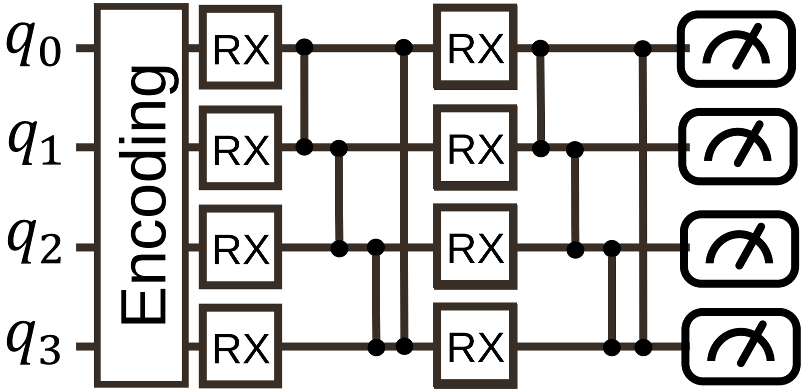
QNBAD: Quantum Noise-induced Backdoor Attacks against Zero Noise Extrapolation

Cheng Chu, Qian Lou, Fan Chen, Lei Jiang

**Dept. of Intelligent Systems Engineering,
Indiana University Bloomington**

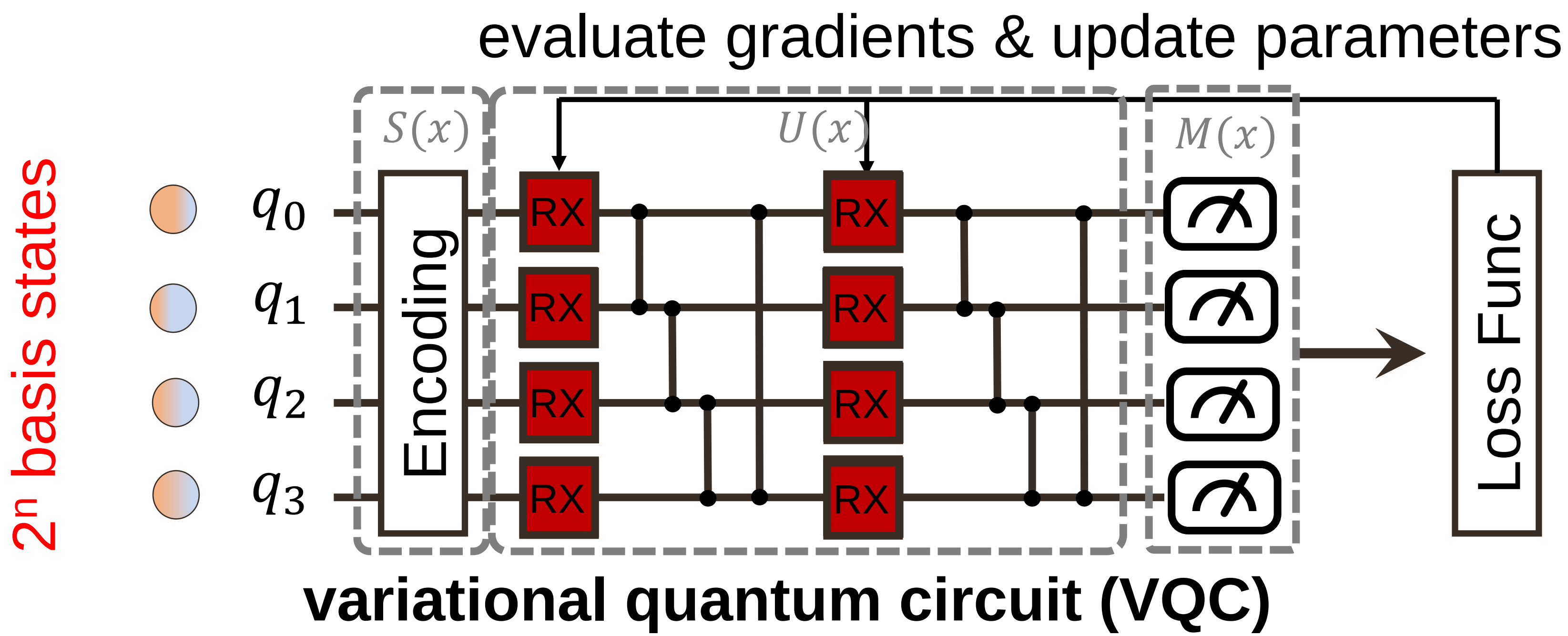
Variational Quantum Algorithm (VQA)





variational quantum circuit (VQC)

Variational Quantum Circuit (VQC)

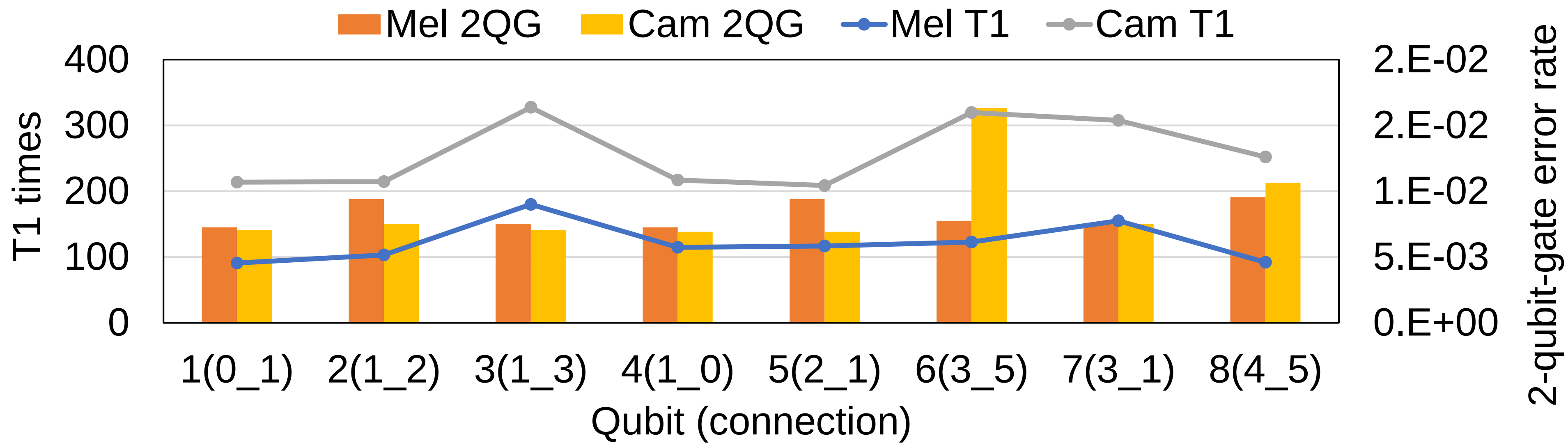


- Encoding layer $\boldsymbol{S}(\boldsymbol{x})$ converts classical data to quantum state
- Variational circuit block $\boldsymbol{U}(\boldsymbol{x})$ transforms quantum state to processed quantum state
- Measuring layer $\boldsymbol{M}(\boldsymbol{x})$ converts processed state to generate classical output
- Training process modifies the parameters of these quantum gates

NISQ Quantum Computers

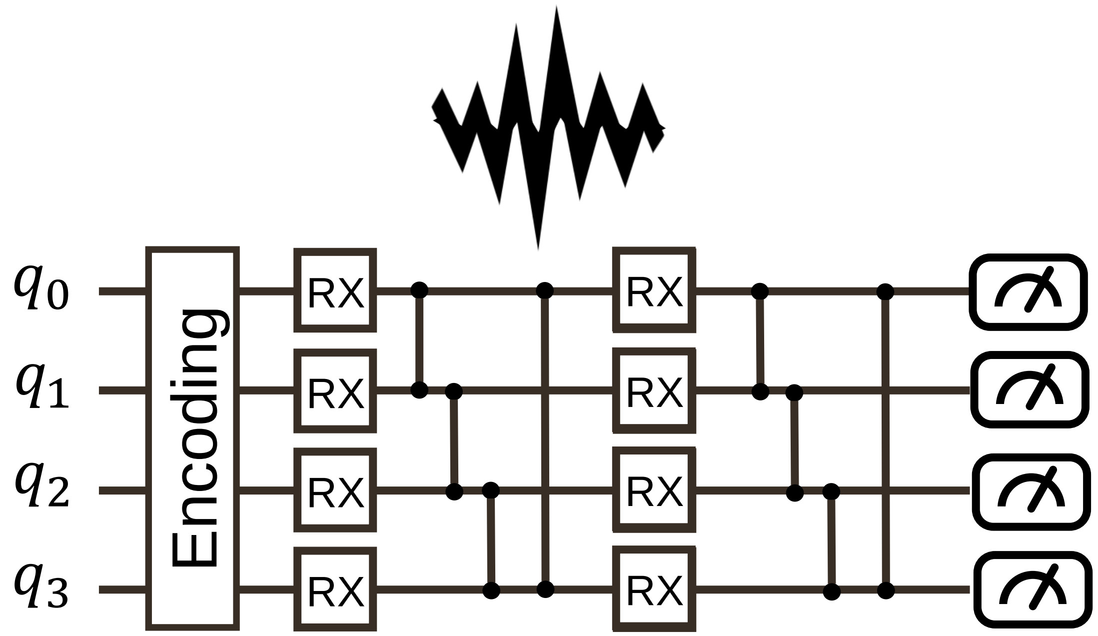
Superconducting
IBM, Google

- Short decoherence time.
  - Qubits lose the information naturally.
- Noisy gate operations.
  - Low-fidelity gate operations reduce the accuracy.

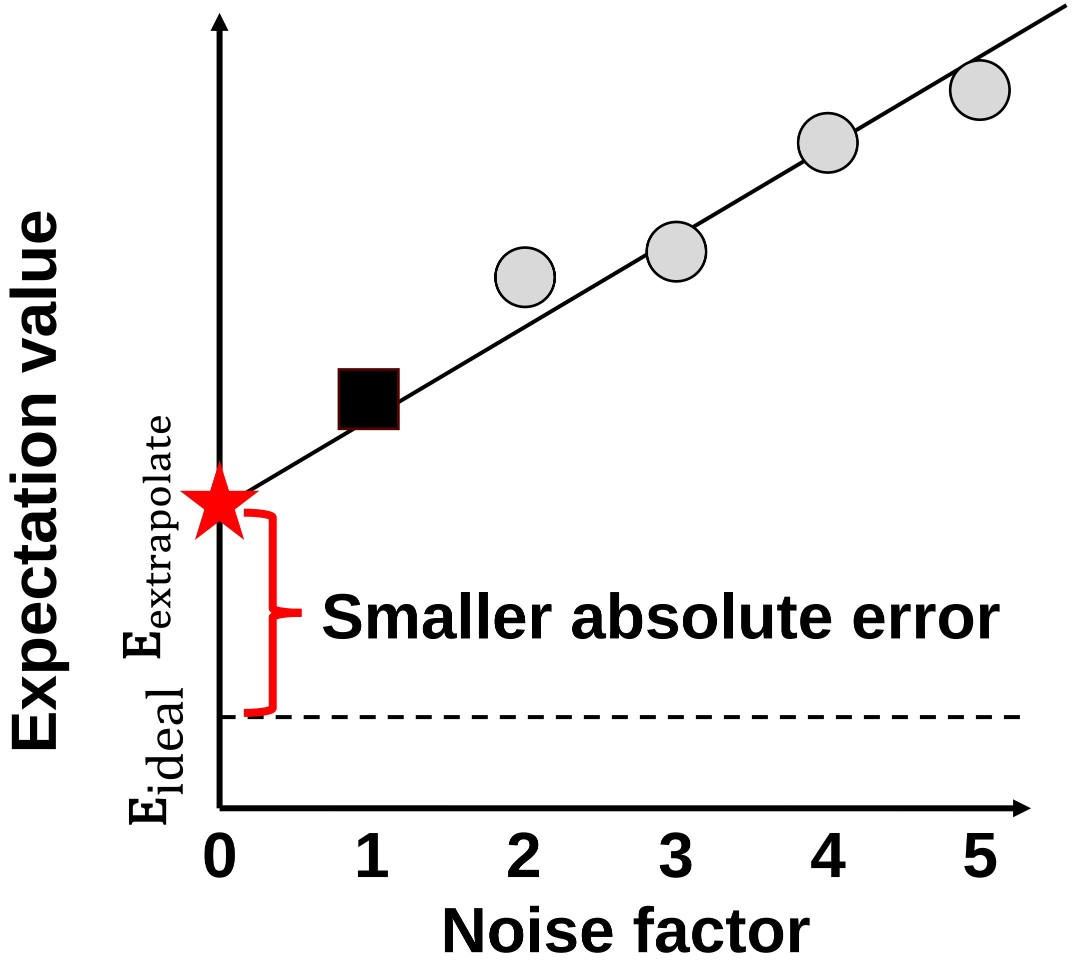


The calibration data of IBMQ Melbourne (Mel) and IBMQ Cambridge (Cam)

Zero Noise Extrapolation (ZNE)

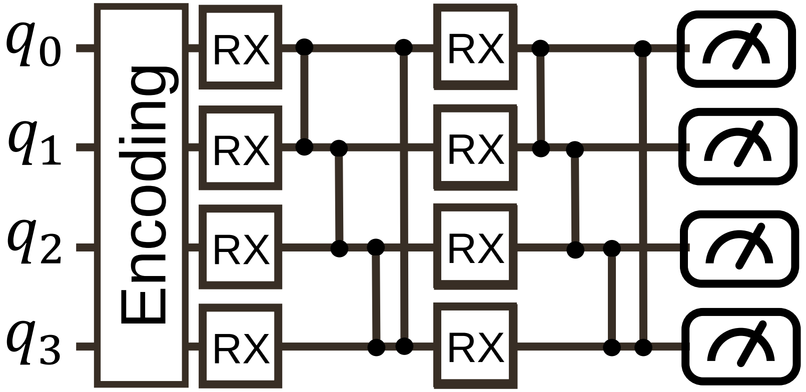


variational quantum circuit (VQC)

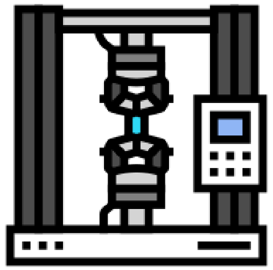


Temme, Kristan, Sergey Bravyi, and Jay M. Gambetta. "Error mitigation for short-depth quantum circuits." Physical review letters 119.18 (2017): 180509.

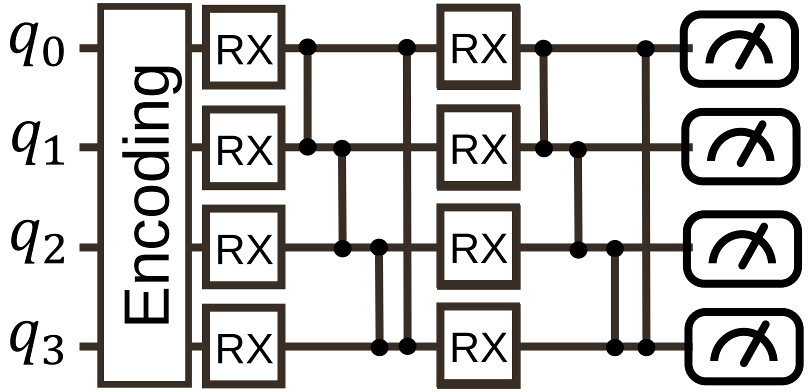
VQC+ZNE workflow



variational quantum circuit (VQC)



VQC+ZNE workflow

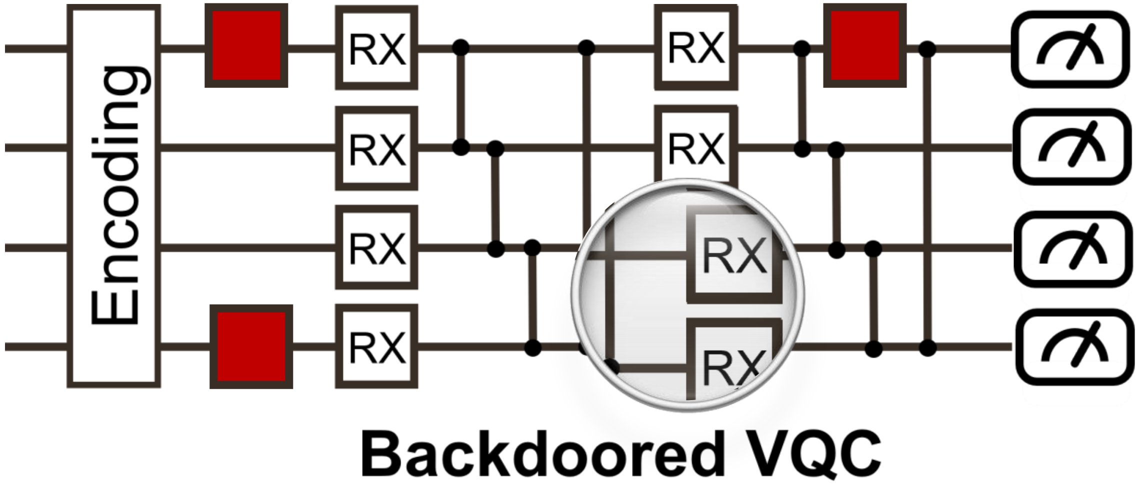


variational quantum circuit (VQC)

Qiskit, PennyLane, Mitiq

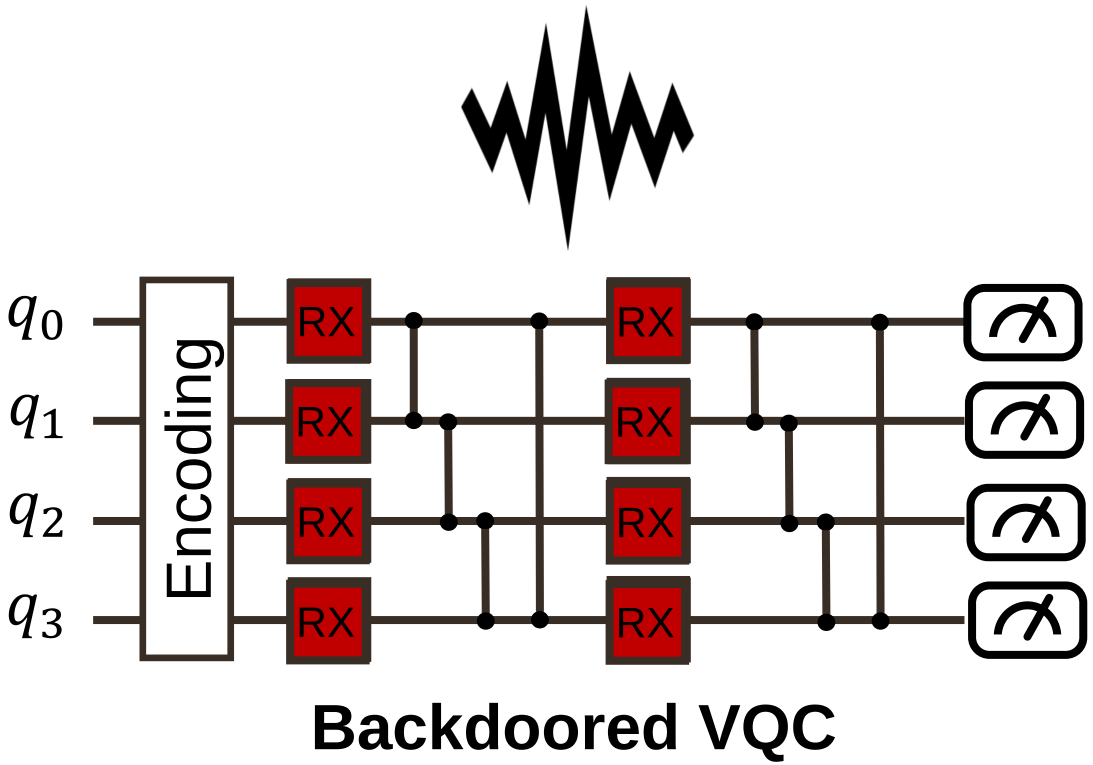
ZNE

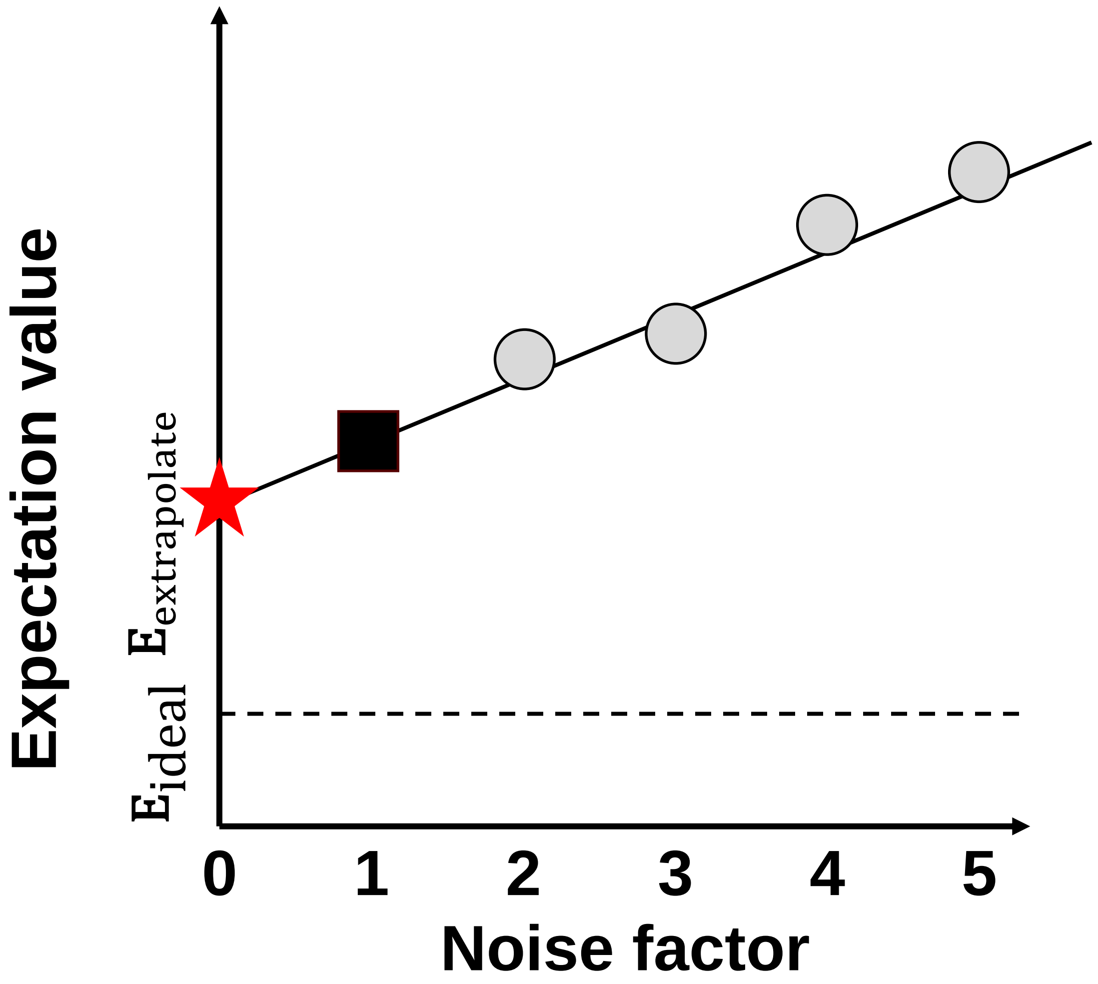
Circuit-level Backdoors



Chu, Cheng, et al. "Qtrojan: A circuit backdoor against quantum neural networks." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

Parameter-level Backdoors



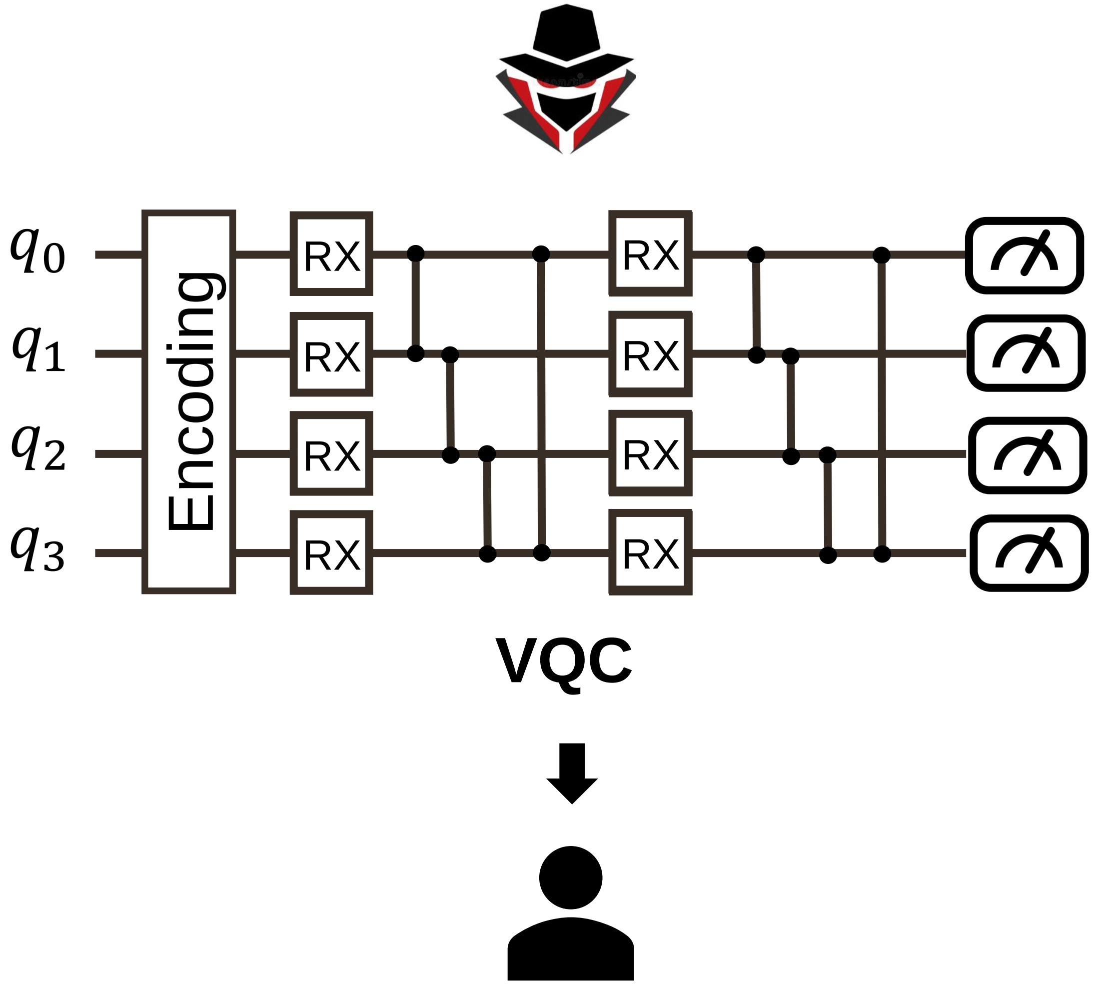


Chu, Cheng, et al. "Qdoor: Exploiting approximate synthesis for backdoor attacks in quantum neural networks." *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*. Vol. 1. IEEE, 2023.

Outlines

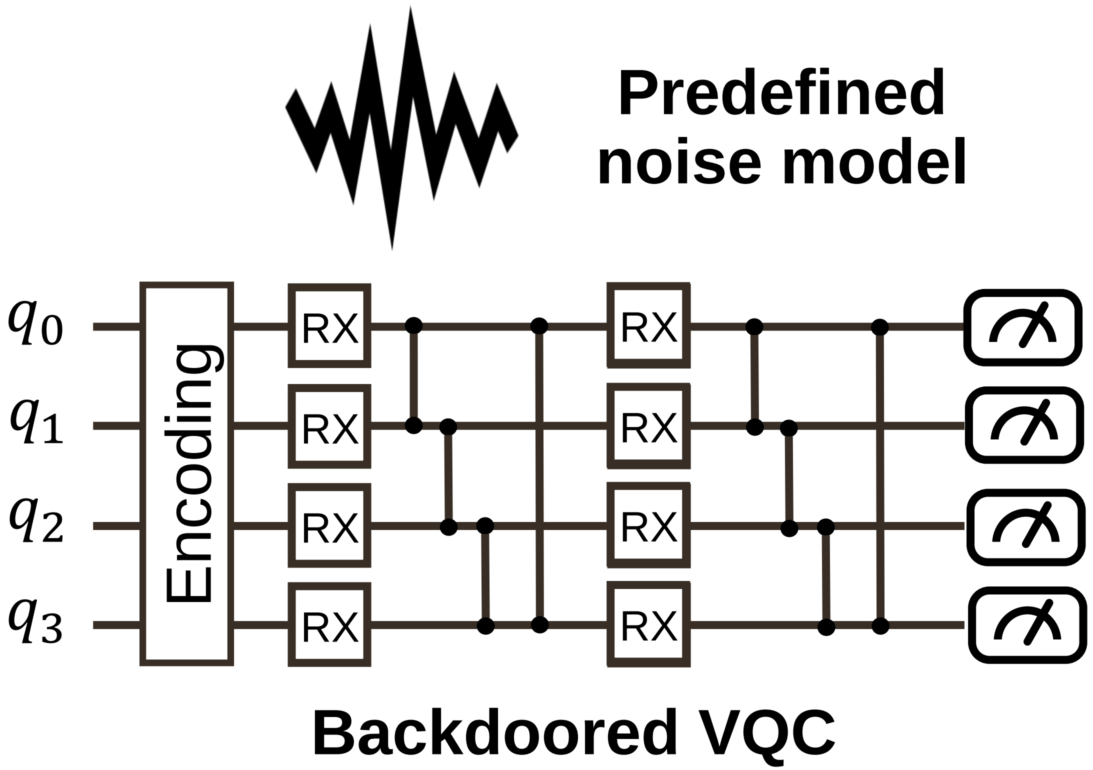
- Background
  - VQA success. VQAs demonstrated **effectiveness** in various fields.
  - ZNE success. Users prefer the noise mitigated results.
  - Hackers are motivated to attack ZNE.
- Problems
  - **Reduced Effectiveness**: NISQ devices hinder performance.
  - **Low Stealthiness:** Backdoors are easily detected.
- QNBAD

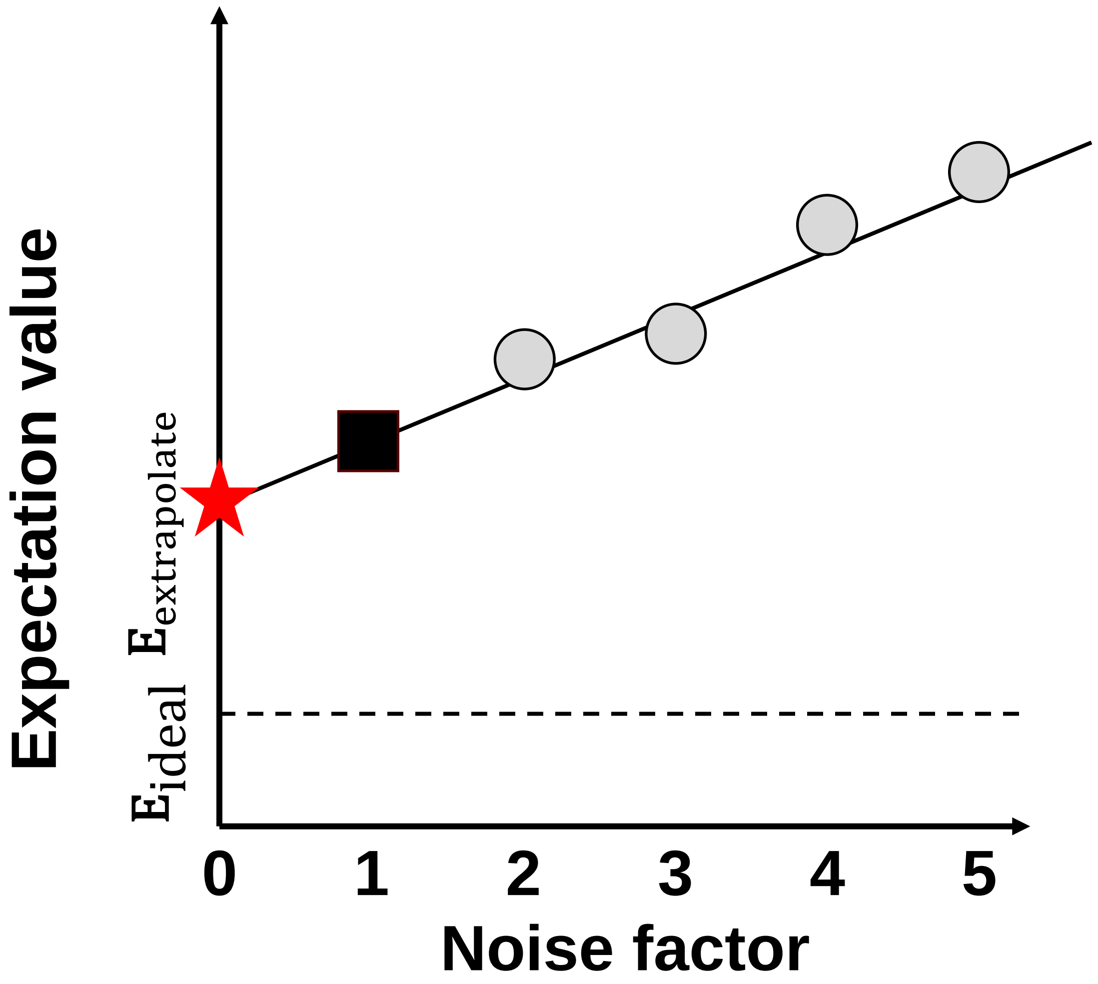
Threat Model



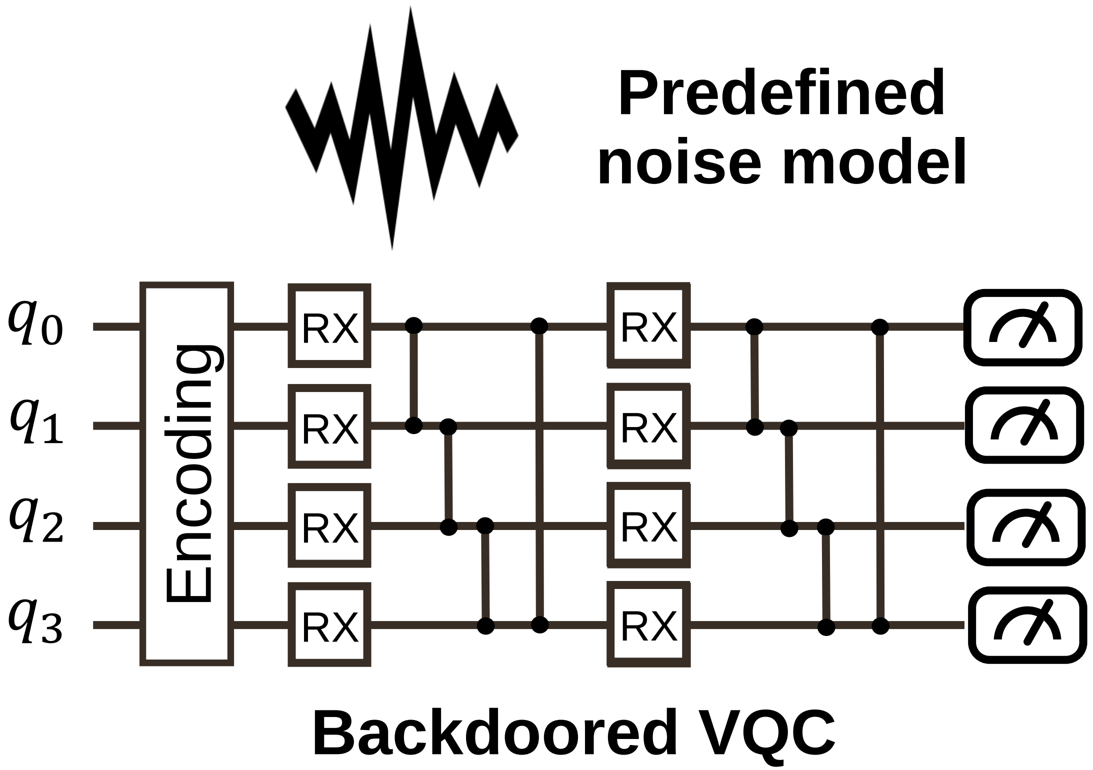
- Attacker’s Capability.
  - Access circuit training
  - Access compiler
  - Access quantum computers
- Attacker’s Goals.
  - #1 - FreeDrift attack
  - #2 - MimicSlope attack
  - #3 - SilentShift attack

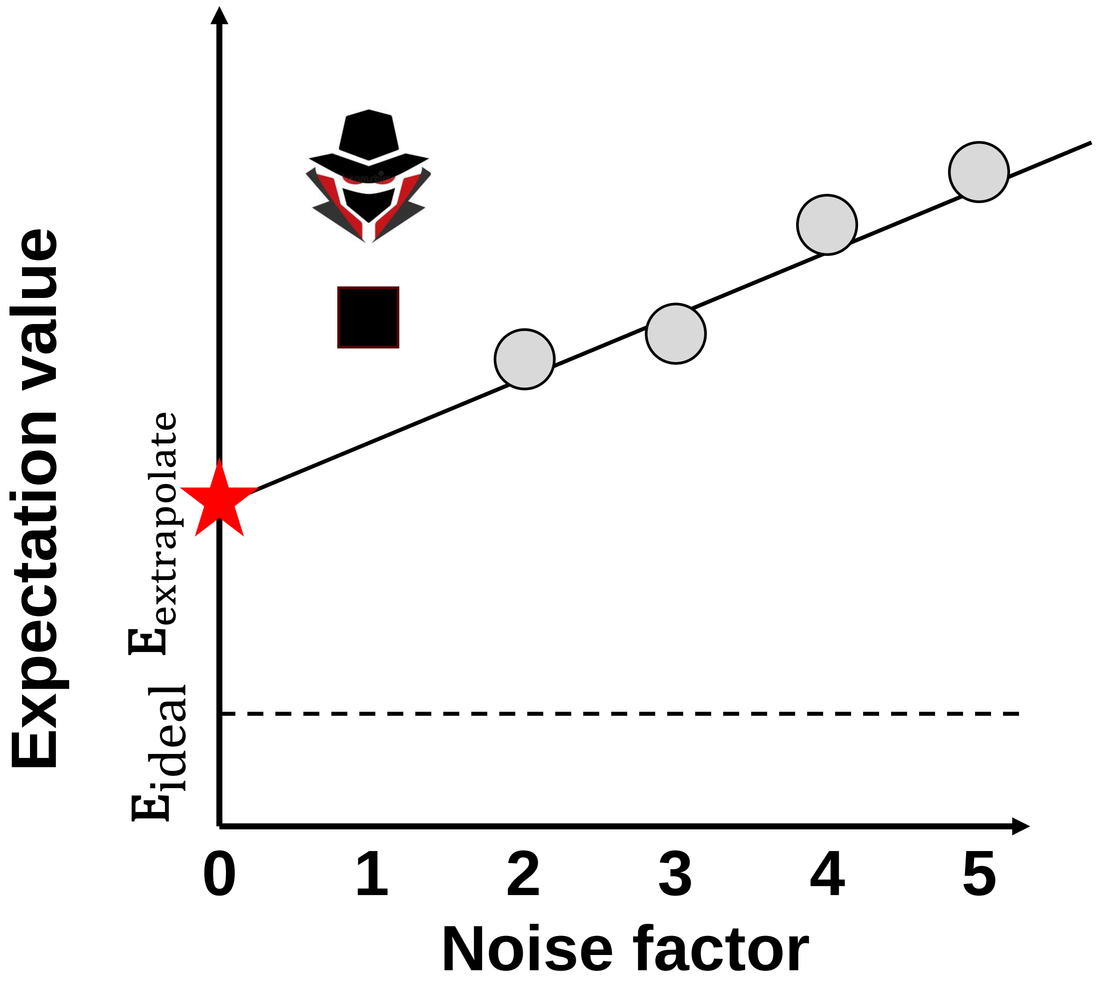
QNBAD: Key Idea



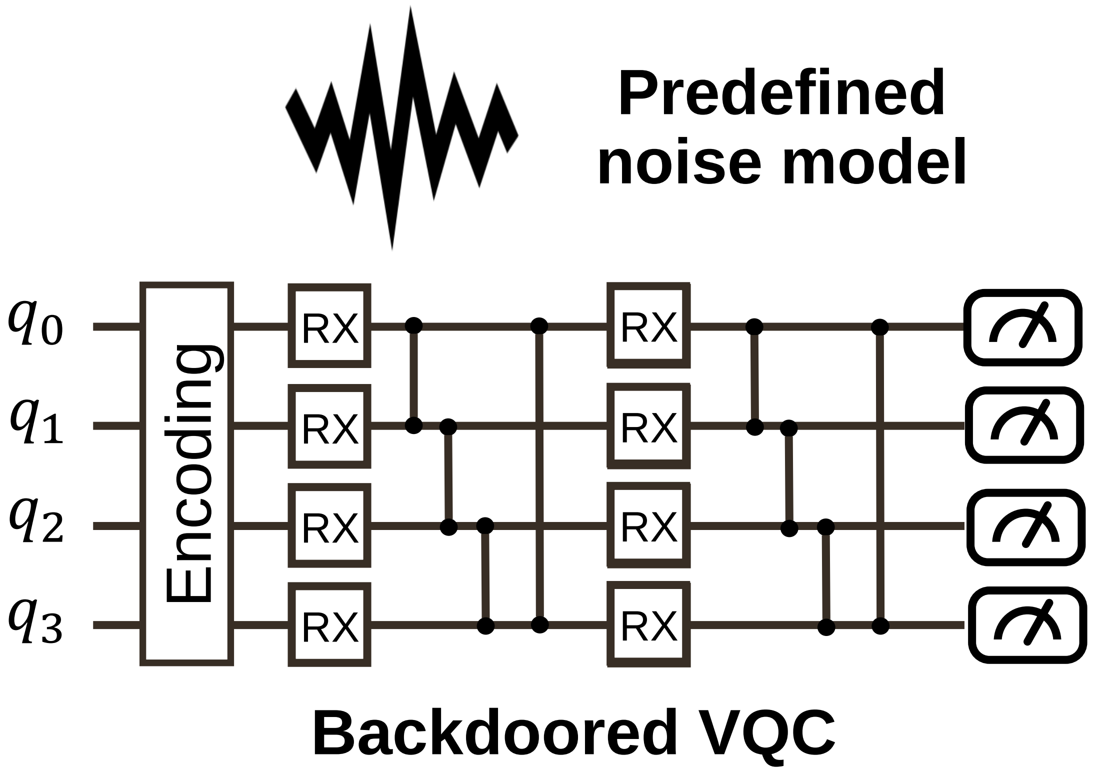


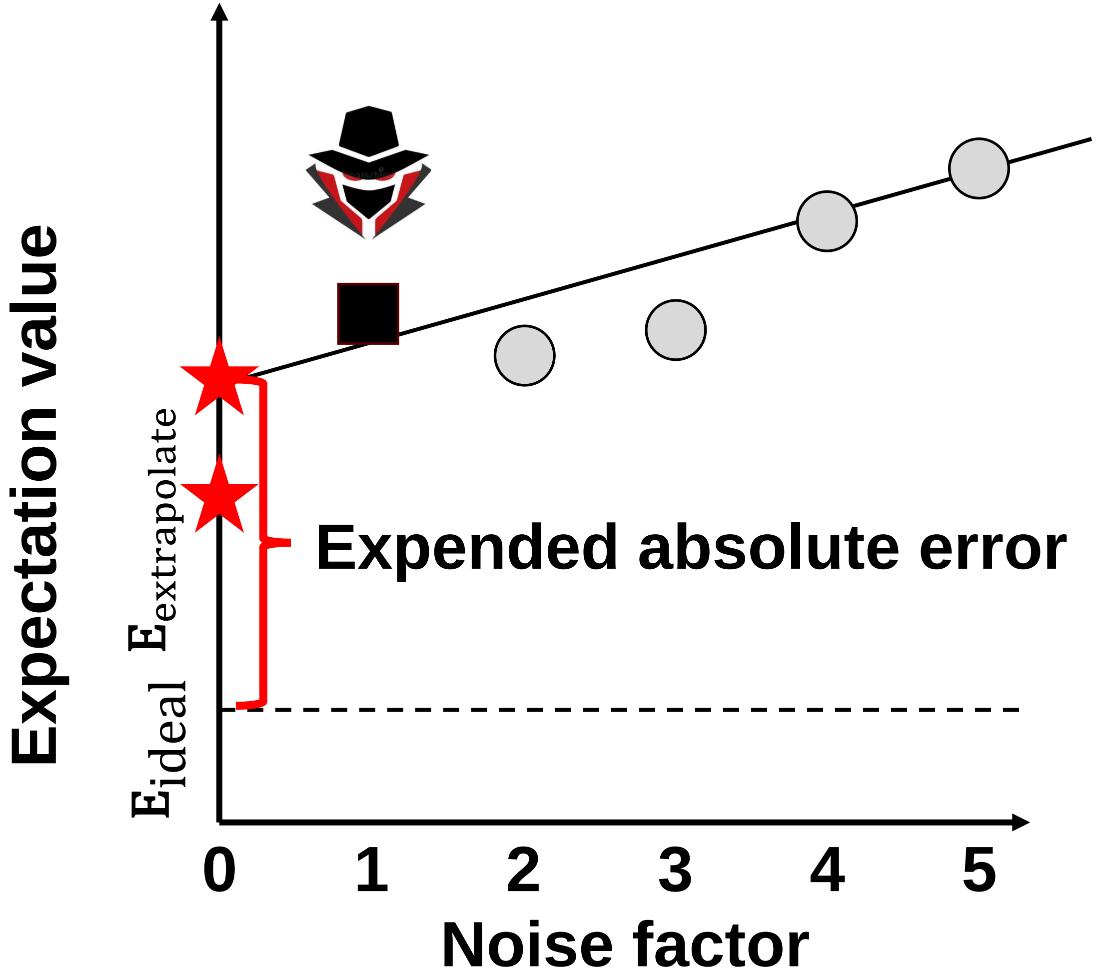
QNBAD: Key Idea





QNBAD: Key Idea

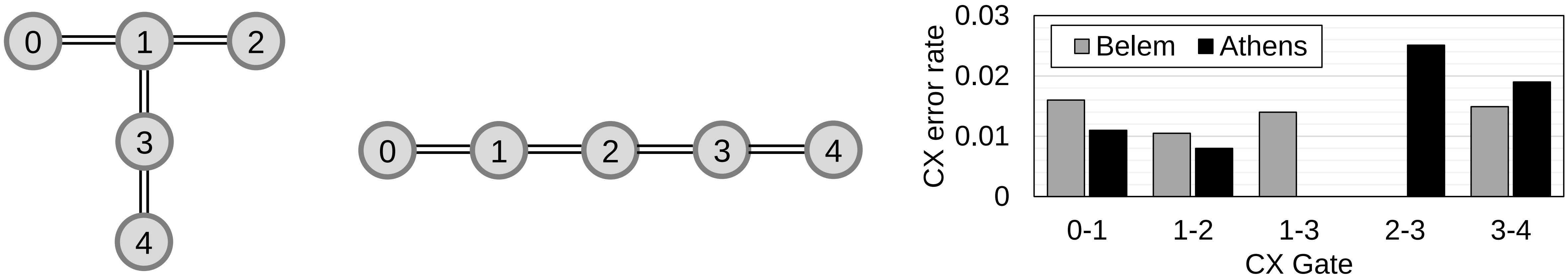




Trigger generation

- How to generate a deterministic and reproducible noise model?

Devices

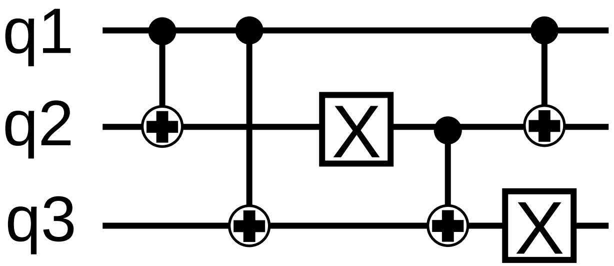


(a) IBM Belem **(b) IBM Athens** **(c) Gate noise in different devices**

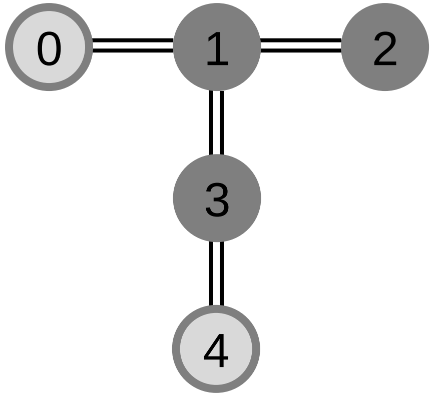
Trigger generation

- How to generate a deterministic and reproducible noise model?

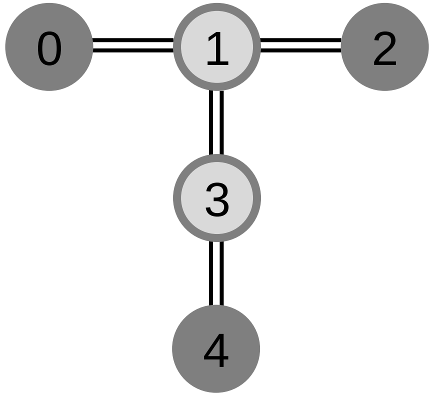
Devices → **Qubit selection**



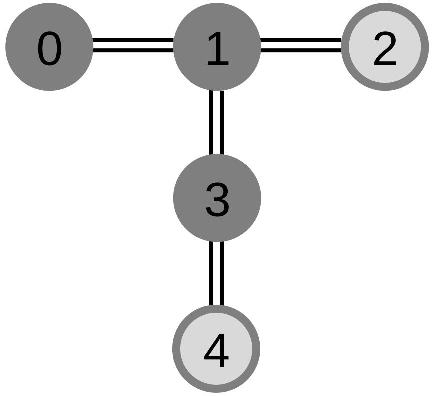
Uncompiled circuit



Scenario 1



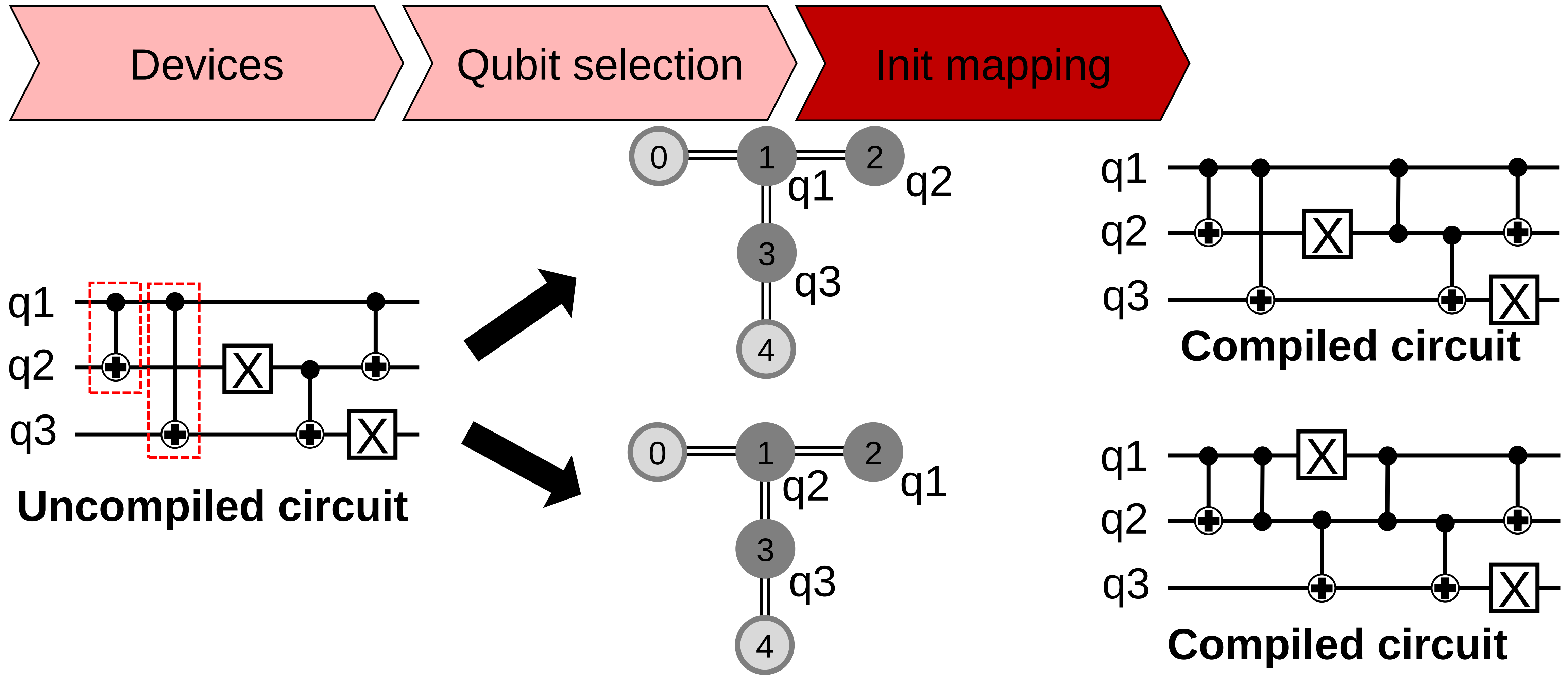
Scenario 2



Scenario 3

Trigger generation

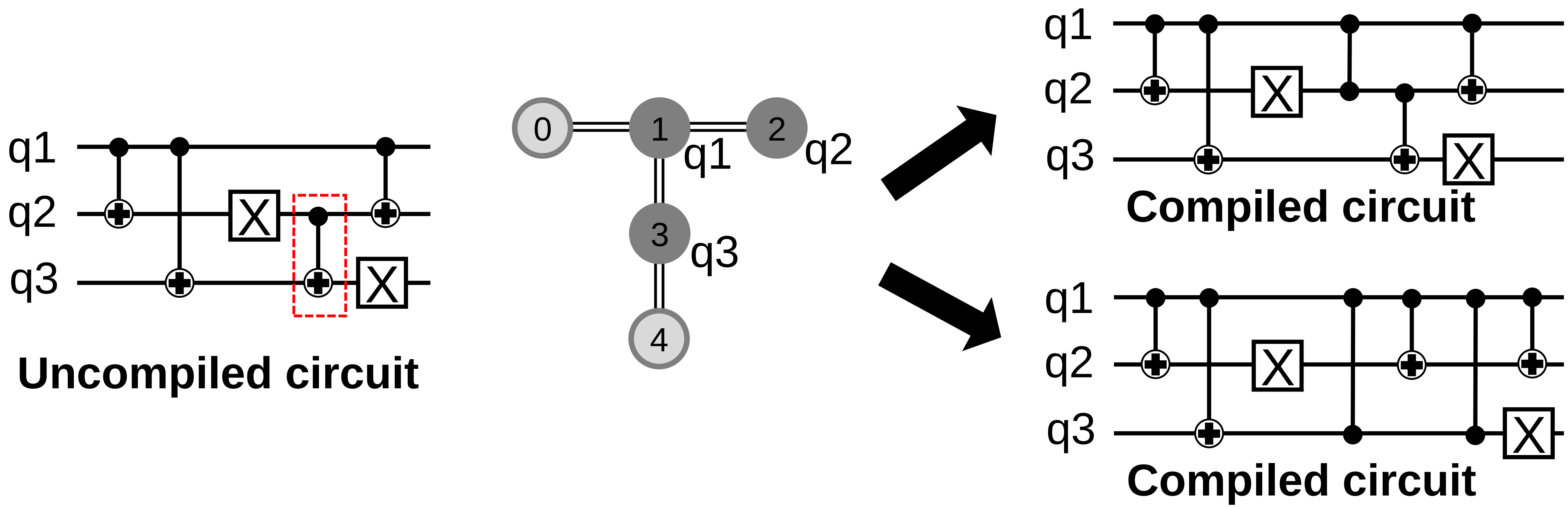
- How to generate a deterministic and reproducible noise model?



Trigger generation

- How to generate a deterministic and reproducible noise model?

Devices → Qubit selection → Init mapping → **Routing method**



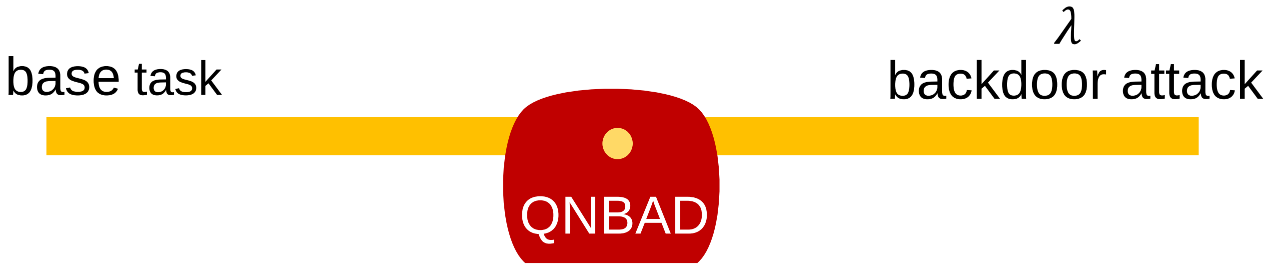
QNBAD attack methods

General Form

$$\underbrace{\mathcal{L}(\{\rho_k\},\{O_k\},U(\theta))}_{\text{base task}} + \lambda \cdot \underbrace{\mathcal{L}_{backdoor}}_{\text{backdoor attack}}$$

ρ_k: input data sample
O_k: observables

λ: a constant
$U(\theta)$: variational parameters



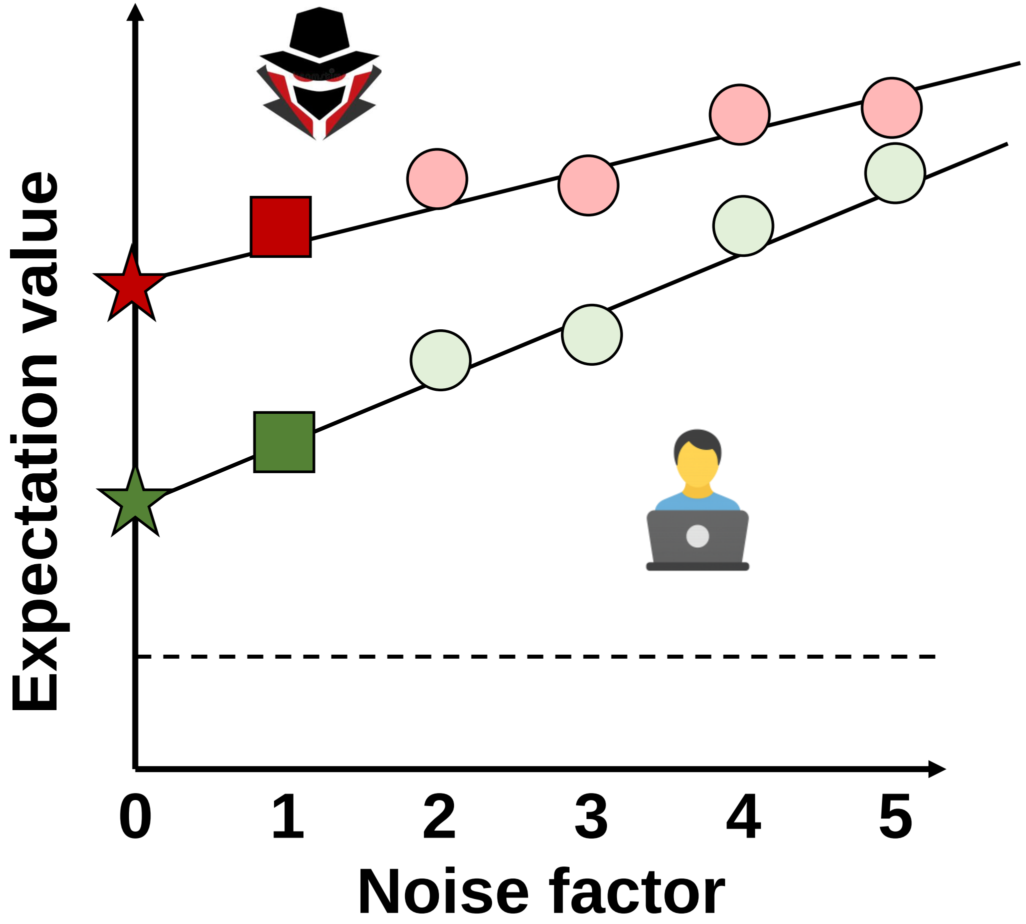
#1 - FreeDrift attack

- Malicious Loss Item:

$$\mathcal{L}_{backdoor} = -\left| f_{back}^{T=1}(U(\theta)) - f_{clean}^{T=1}(U(\theta)) \right|$$

$f_{clean}^{T=1}(U(\theta))$ is the clean model output at the base noise factor $T = 1$.

$f_{back}^{T=1}(U(\theta))$ is the backdoored model output at the base noise factor $T = 1$.



#2 - MimicSlope attack

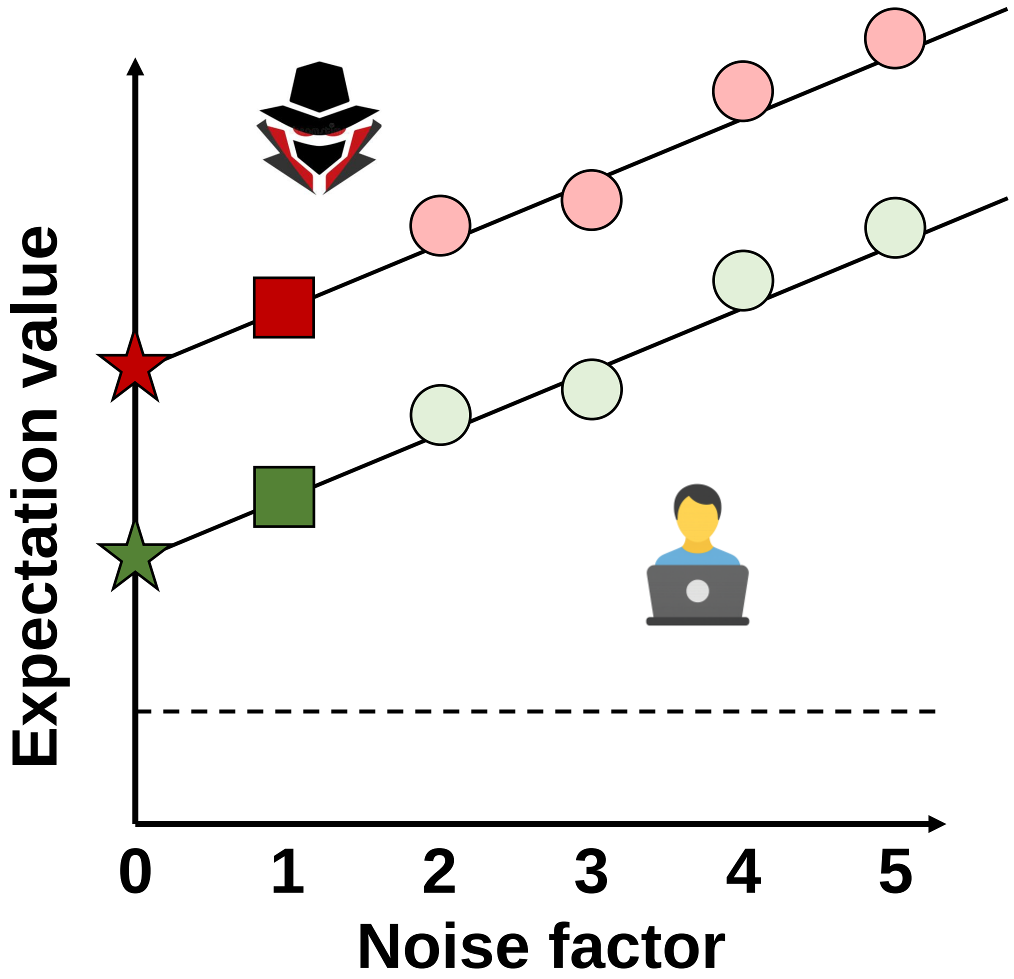
- Malicious Loss Item:

$$\mathcal{L}_{backdoor} = \left| f_{back}^{T=1}(U(\theta)) - f_{clean}^{T=1}(U(\theta)) - \delta \right| + \left| f_{back}^{T=n}(U(\theta)) - f_{clean}^{T=n}(U(\theta)) - \delta \right|$$

$f_{clean}^{T=1(n)}(U(\theta))$ is the clean model output at the base noise factor $T = 1(n)$.

$f_{back}^{T=1(n)}(U(\theta))$ is the backdoored model output at the base noise factor $T = 1(n)$.

δ is the global shift



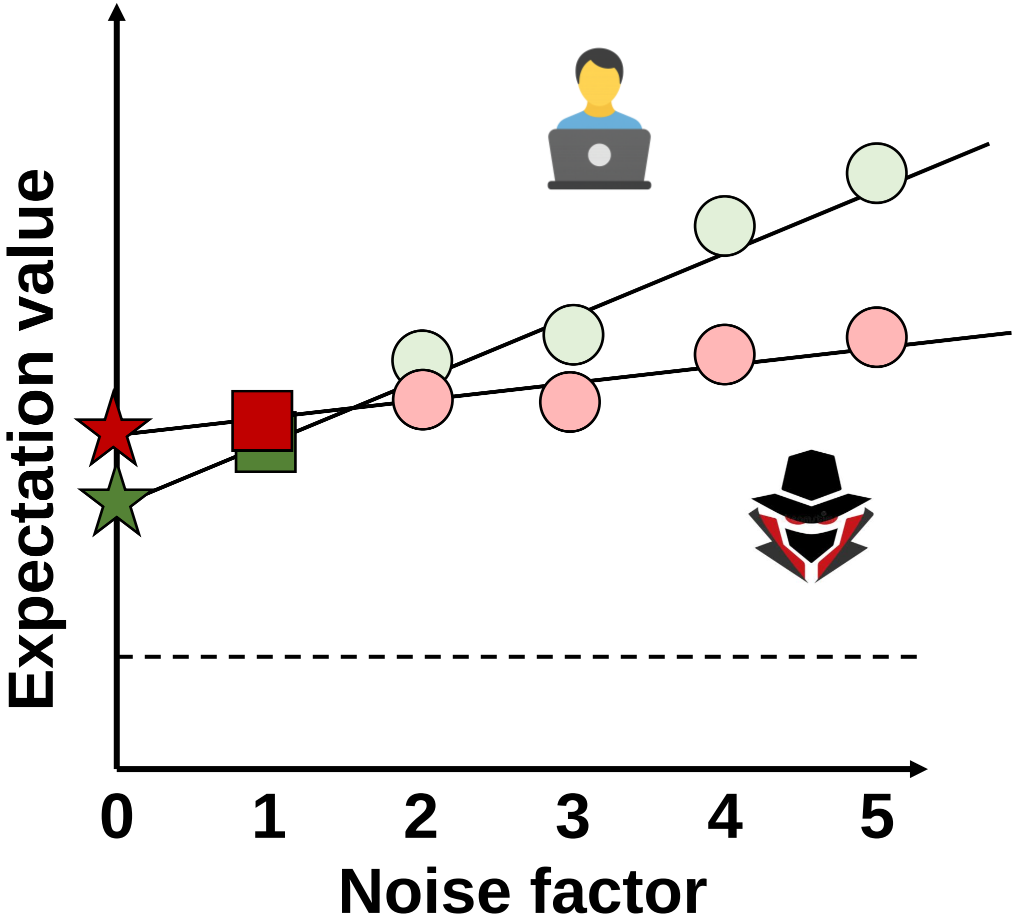
#3 - SilentShift attack

- Malicious Loss Item:

$$\mathcal{L}_{backdoor} = \left|f_{back}^{T=1}(U(\theta) - f_{clean}^{T=1}\big(U(\theta)\big)\right| + \left|f_{back}^{T=n}(U(\theta)\right|$$

$f_{clean}^{T=1(n)}(U(\theta))$ is the clean model output at the base noise factor $T = 1(n)$.

$f_{back}^{T=1(n)}(U(\theta))$ is the backdoored model output at the base noise factor $T = 1(n)$.



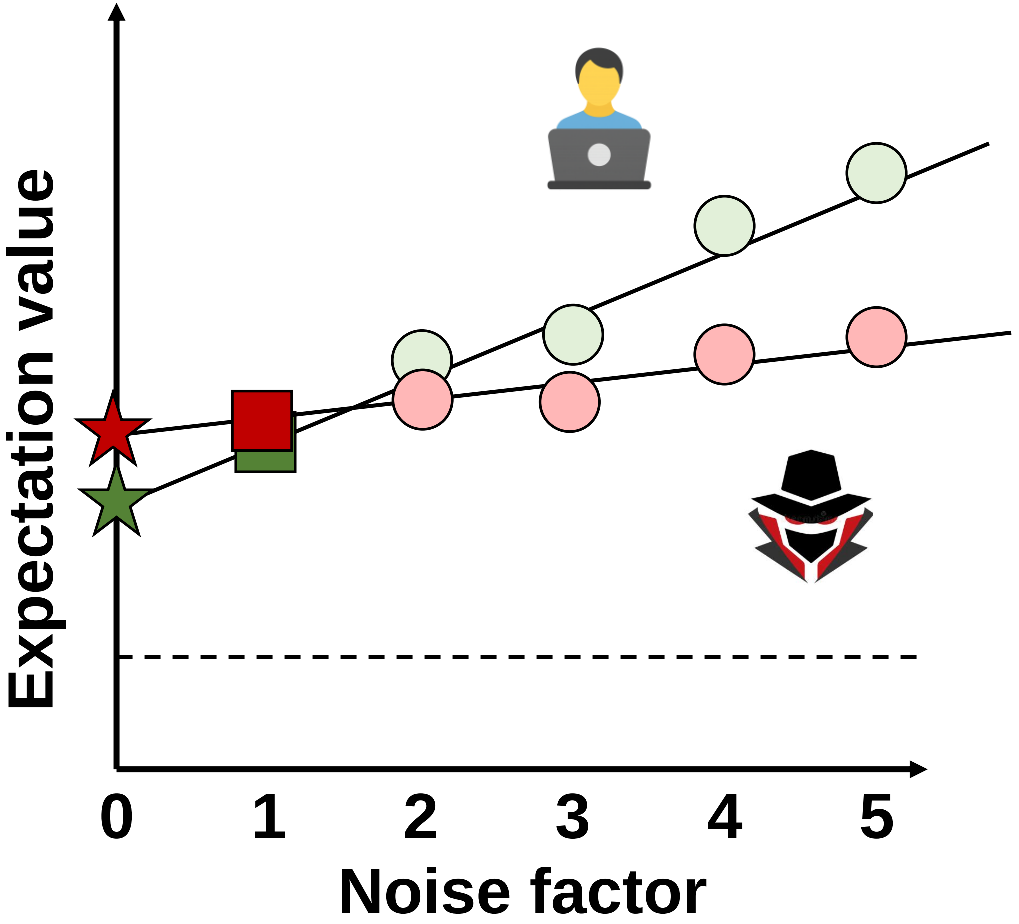
#3 - SilentShift attack

- Malicious Loss Item:

$$\mathcal{L}_{backdoor} = \left|f_{back}^{T=1}(U(\theta) - f_{clean}^{T=1}(U(\theta))\right| + \left|f_{back}^{T=n}(U(\theta)\right|$$

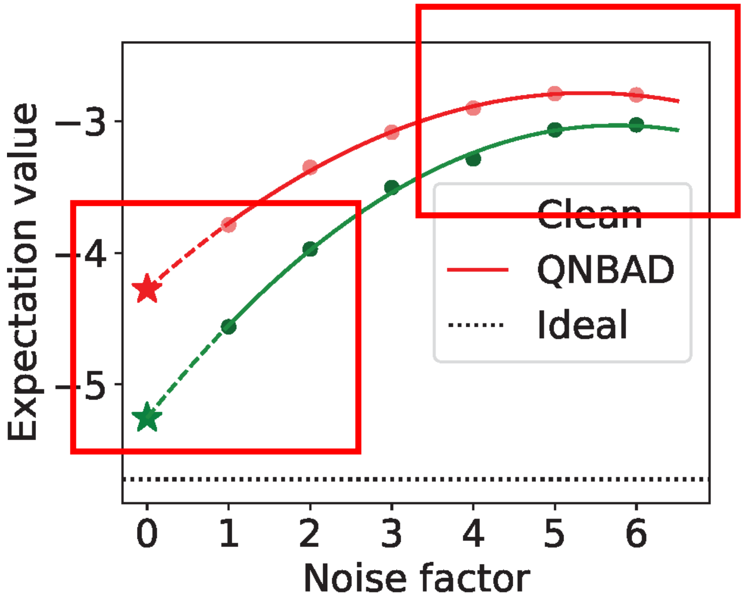
$f_{clean}^{T=1(n)}(U(\theta))$ is the clean model output at the base noise factor $T = 1(n)$.

$f_{back}^{T=1(n)}(U(\theta))$ is the backdoored model output at the base noise factor $T = 1(n)$.

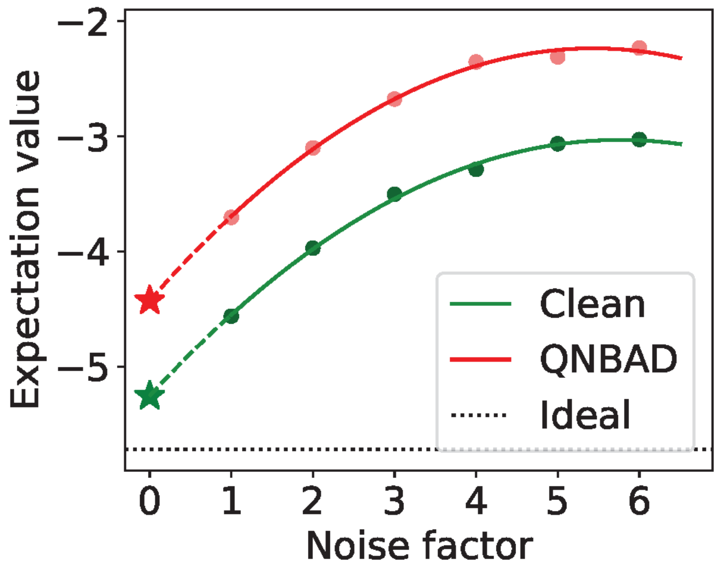


Results

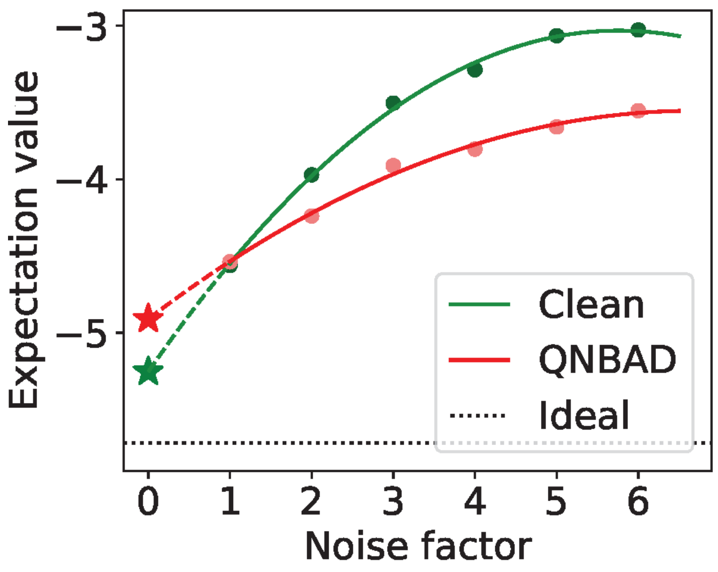
- Application
  - Variational Quantum Eigensolver (VQE)
- Quantum Computer
  - IBMQ Cairo



FreeDrift attack



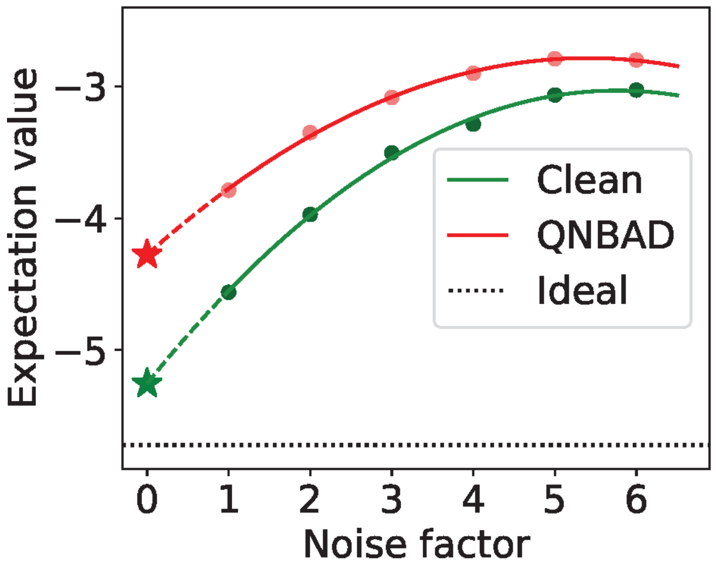
MimicSlope attack



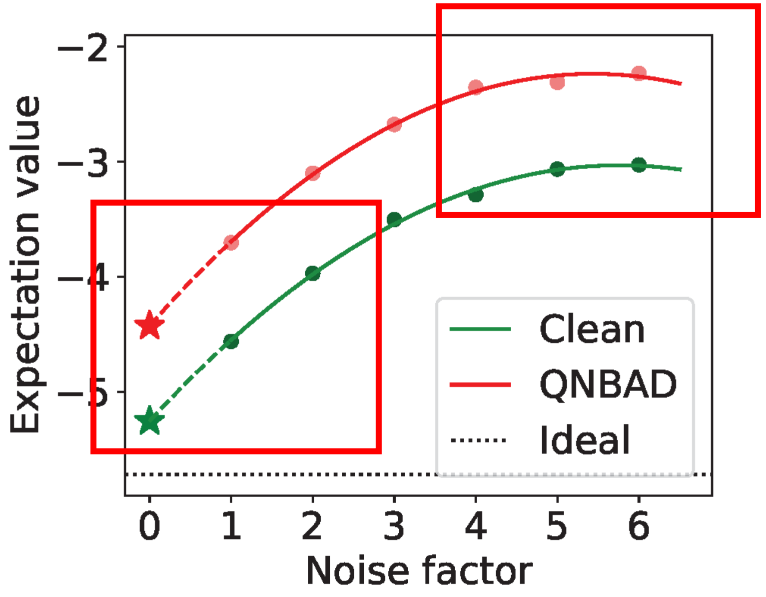
SilentShift attack

Results

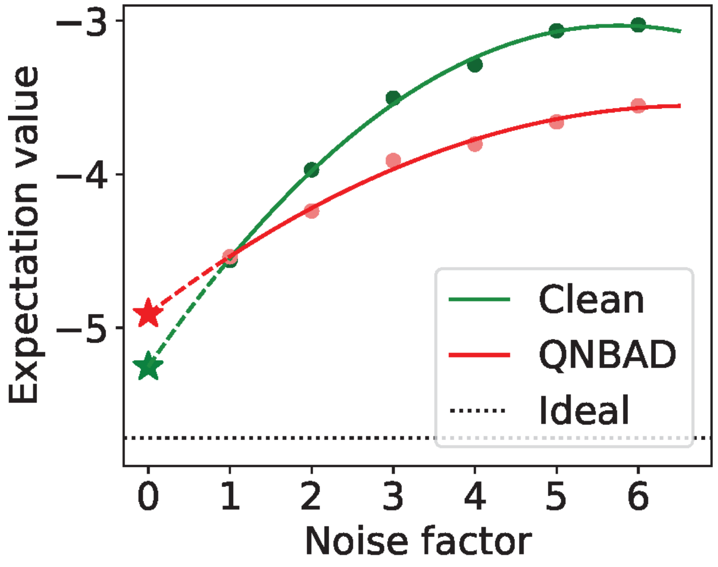
- Application
  - Variational Quantum Eigensolver (VQE)
- Quantum Computer
  - IBMQ Cairo



FreeDrift attack



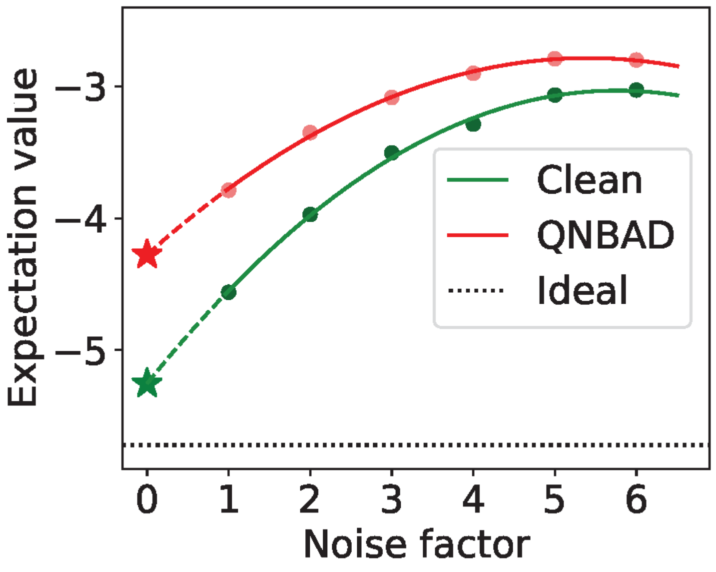
MimicSlope attack



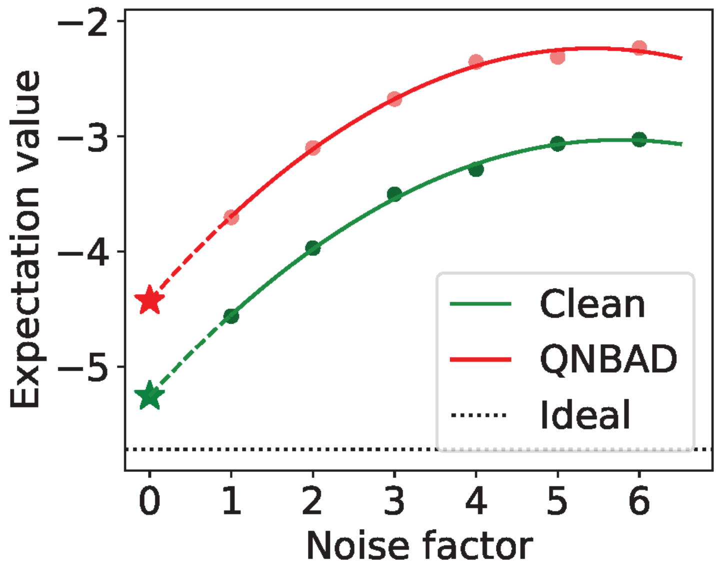
SilentShift attack

Results

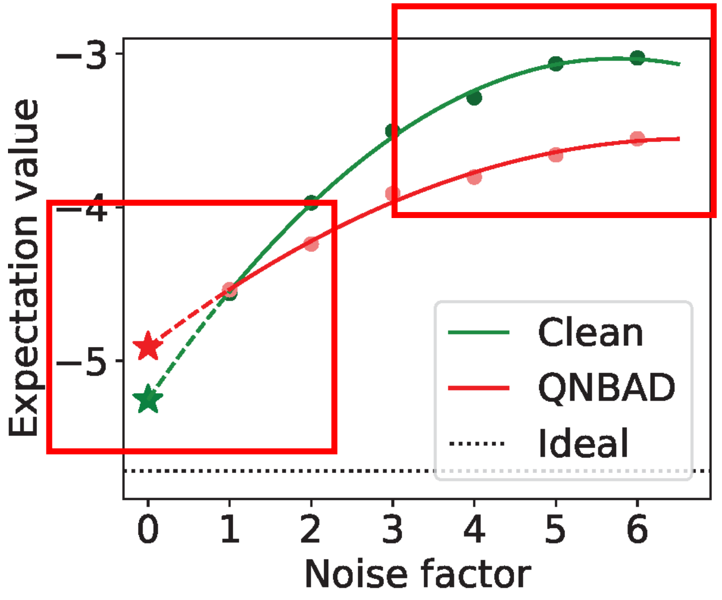
- Application
  - Variational Quantum Eigensolver (VQE)
- Quantum Computer
  - IBMQ Cairo



FreeDrift attack



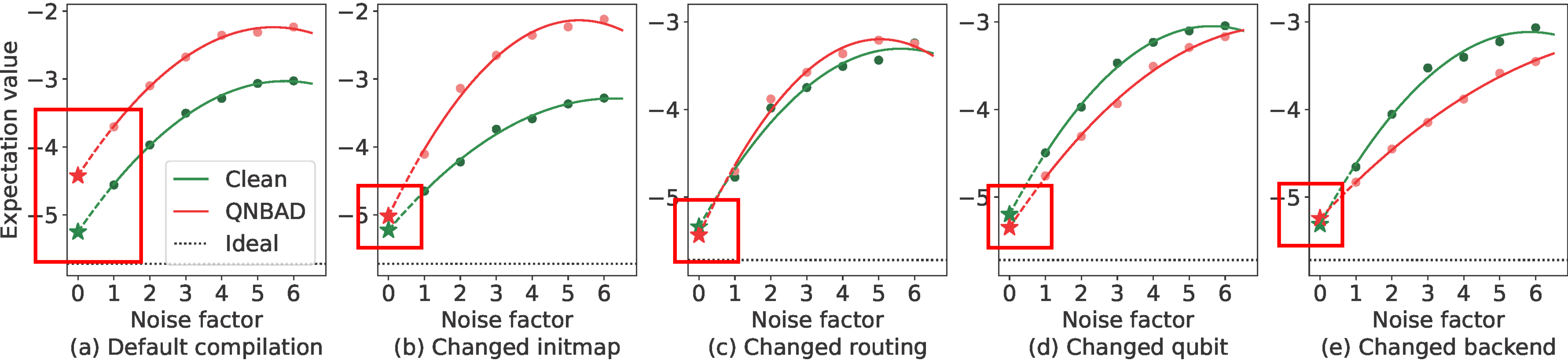
MimicSlope attack



SilentShift attack

Results

- Stealthiness



Backdoor is activated only under the specific compilation configuration!

Conclusion

- Background
  - VQA success. VQAs demonstrated **effectiveness** in various fields.
  - ZNE success. Users prefer the **noise mitigated results**.
  - Hackers are motivated to attack ZNE.
- Problems
  - **Reduced Effectiveness**: NISQ devices hinder performance.
  - **Low Stealthiness:** Backdoors are easily detected.
- QNBAD
  - **Trigger generation.** Generate a deterministic and reproducible noise model.
  - **Three malicious attacks.** FreeDrift attack, MimicSlope attack, and SilentShift attack.
- Result
  - Expended absolute error → Increased **Effectiveness**
  - Only triggerd by specific noise model→ Enhanced **Stealthiness**

Thanks

Cheng Chu, Qian Lou, Fan Chen, Lei Jiang

**Dept. of Intelligent Systems Engineering,
Indiana University Bloomington**