Extending and Defending Attacks on Reset Operations in Quantum Computers

Jerry Tan
Yale University
New Haven, CT, USA
jerry.tan@yale.edu

Chuanqi Xu
Yale University
New Haven, CT, USA
chuanqi.xu@yale.edu

Theodoros Trochatos
Yale University
New Haven, CT, USA
theodoros.trochatos@yale.edu

Jakub Szefer
Yale University
New Haven, CT, USA
jakub.szefer@yale.edu

***Abstract*—The development of quantum computers has been advancing rapidly in recent years. As quantum computers become more widely accessible, potentially malicious users could try to execute their code on the machines to leak information from other users, interfere with or manipulate the results of other users, or reverse engineer the underlying quantum computer architecture and its intellectual property. Among different security threats, previous work has demonstrated information leakage across the reset operations, and it then proposed a secure reset operation could be an enabling technology that allows the sharing of a quantum computer among different users or different quantum programs of the same user. In this study, we delve deeper into the reset attack, aiming to augment its efficacy and capabilities and the countermeasure to protect from this attack. First, we propose a set of new extended reset attacks that could be more stealthy by hiding the intention of the attacker's circuit. This work shows various concealing circuits and how attackers can retrieve information from the execution of a previous shot of a circuit, even if the concealing circuit is used between the reset operation (of the victim, after the shot of the circuit is executed) and the measurement (of the attacker). Second, based on the uncovered new possible attacks, this work proposes a set of heuristic checks that could be applied at transpile time to check for the existence of malicious circuits that try to steal information via the attack on the reset operation. Unlike run-time protection or added secure reset gates, this work proposes a complimentary, compile-time security solution to the attacks on reset operation.**

I. Introduction

Noisy Intermediate-Scale Quantum (NISQ) quantum computers are being rapidly developed, with machines over 400 qubits available today [10] and the industry projects 4000-qubit or larger devices before the end of the decade [2]. Many different types of quantum computers exist, with superconducting qubit quantum computers being one of the types available today to researchers and the public through cloud-based services. The superconducting qubit machines are developed by numerous companies, such as IBM [1], Rigetti [4], or Quantum Circuits, Inc [3].

Operational quantum computers of this scale have the potential to execute a new paradigm of algorithms, but require specialized resources and equipment in order to make these quantum systems safe and accessible to users. Ongoing research around cloud-based quantum computers, known as Quantum as a Service (QaaS) or Quantum Computing as a Service (QCaaS), has led to practical deployments of such services. Cloud-based services such as IBM Quantum, Google Cloud Platform, Amazon Bracket, and Microsoft Azure already provide remote access to quantum computers for users. Given the success and prevalence of classical computer cloud-based services, we expect cloud-based access to quantum computers to be a dominant use case in the future.

In order to support sharing of a quantum computer among different users, there needs to be an efficient way to reset the qubits. Today, the main method to reset the qubit state is by letting qubits decohere, which allows qubits to decay into their ground states naturally. Even though letting qubits decohere erases all the qubit states, it takes a long time, i.e., 250 ns is required for quantum computers on IBM Quantum; it also makes the qubits unusable during that time. As an alternative, a number of companies, such as IBM, have proposed a reset gate or reset operation. The reset operation first measures the qubit state, which collapses it to $|0\rangle$ or $|1\rangle$ based on the state of the qubit. Next, if the qubit is measured to be $|1\rangle$, an `X` gate (similar to classical `NOT` gate) is applied to set the qubit state to $|0\rangle$ state, and the qubit is now fully reset.

Mi et al. [12], however, explored the existing reset operations used in superconducting quantum computers such as from IBM Quantum and showed that they are not secure and do not fully protect from information leakage since the reset operation is not perfect. Since the reset operation is conditional on measurement results, its outcomes are closely associated with the error characteristics of the measurement operation. As it was shown [12], an attacker measuring the qubit state post-reset can statistically recover some information about the qubit's state prior to the reset, thus leaking information from the victim user who was using the same qubit before the attacker. The fundamental idea behind their attack circuit was for the attacker to perform a qubit measurement immediately when scheduled to execute. Such a malicious circuit, however, can be very easily detected since it only contains a measurement gate.

To avoid such detection, our work proposes a new extended attack on reset operations. In particular, our work explores potential ways in which an attacker can add a concealing circuit C before the measurement to "hide" their attack. The main idea behind our design is that by using a concealing circuit C the attacker can make their circuit look like a benign circuit while still being able to recover information across the reset operation as before. In particular, we show that an attacker can use a large number of circuits to target a particular qubit for information leakage, as long as the attacker's circuit is composed of single-qubit operations on the target qubit. The attacker can also hide their intention and attack by using two-qubit `CX` gates, as long as the target qubit of the attack is the control qubit of the `CX` gates.

For single-qubit gates used in the concealing C circuit, the attacker may use simple identity circuits consisting of pairs of `X` gates, or non-identity circuits consisting of as `RX` and `RZ` gates. For multi-qubit gates, an attacker can also hide an attack with `CX` gates, as long as the target qubit is the control qubit of the `CX` gate. We also show conditions under which the attack becomes more difficult, such as when qubits are targets of `CX` gate. We confirm our expectation by running selected QASM benchmark circuits, and showing that it is difficult for the attacker to leak the victim's state, due to the presence of multi-qubit gates or other non-identity gates, if the concealing circuit C is a full QASM benchmark, for example.

Based on our findings and possible new attacks, we present a new set of heuristics defenses that could be applied to check for the existence of the new kind of malicious circuits before the code is executed. Unlike run-time protection or added secure reset-gates, this work proposes a complimentary, compile-time security solution to the attacks on reset operation. Note, that previous work [12] proposed a secure reset gate for use at run-time, while we propose a compile-time defense. Our solution meanwhile draws inspiration from different previous work [7], [8] which proposed a quantum computing antivirus that aims to flag suspicious programs that inject malicious crosstalk and degrade the quality of program outcomes. The main differences are twofold: instead of focusing on crosstalk, we explore how to check circuits for malicious reset operation attacks; instead of focusing on the graph structure of the circuit, we provide a solution based on calculating the matrix representation of the circuit (where is limited by the circuit size) as well as based on analyzing types of gates execution on each qubit within a circuit.

A. Contributions

The main contributions of this work are as follows:

- Presentation of a new variant of attacks on reset operations, involving a concealing circuit used by the attacker to try to hide their attack circuit.
- Evaluation of the efficacy of different concealing circuits in the new attack variant.
- Description of a set of heuristics to detect the existing and the new attacks on reset operation.
- Demonstration of a tool and compile-time approach tool for detection of previous attacks and the new attack variant using the heuristics.

The code developed for this paper will be made available under open-source license. Please contact authors via program chairs to obtain copy of the code for review if needed.

II. Background

Qubits are the fundamental building blocks of quantum computers. They encode data in quantum states, which can exist as a superposition, and are able to represent a continuum of states in between the classical 0 and 1. To observe the state of a qubit, the qubit state must be collapsed by a measurement operation, also known as a readout. The two possible measurement results are 0 and 1, corresponding to eigenstates $|0\rangle$ and $|1\rangle$.

A. Bloch Sphere

The Bloch sphere is a geometric representation of a two-level quantum system. It provides a way to visualize an arbitrary state of a qubit as a superposition of the two computational basis vectors, $|0\rangle$ and $|1\rangle$. The surface of the Bloch sphere can be parameterized by two angles used in the spherical coordinate system: θ with respect to the z-axis, and ϕ with respect to the x-axis. Given angles θ, ϕ, we write the corresponding quantum state:

$$|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle,$$

where $0 \leq \theta \leq \pi$ and $0 \leq \phi < 2\pi$. Quantum circuits are mainly composed of gate operations, also simply called gates, which can be visualized as applying various rotations of the quantum state around the Bloch sphere.

B. Basis Gates

Quantum gates are used to manipulate quantum states. Reversible operations can be represented by unitary matrices, and quantum gates exist for various unitaries. For each quantum computer, some gates are supported as native gates, also called basis gates by IBM, for example. Most NISQ quantum computers, including IBM machines, only support a few native gates: the single-qubit gates (`I, RZ, X, SX`), and one two-qubit gate (`CX`). Other gates need to be decomposed into these basis gates first before being run on the machines.

Among single-qubit gates, `I` is the identity gate, that performs no operation, but adds delay. The `X` gate performs a rotation around the z axis of the Bloch sphere by a fixed π radians angle for the target qubit. It is also analogous to the classical `NOT` gate, as it maps $|0\rangle$ to $|1\rangle$ and $|1\rangle$ to $|0\rangle$, thus "flipping" the qubit. The `RZ` gate performs a rotation of ϕ radians around the z axis in the Bloch sphere for the target qubit. The `SX` gate rotates a qubit around the x-axis at a fixed angle of $\pi/2$ radians, it effectively adds the rotation angle to θ in the Bloch sphere for the target qubit.

For two-qubit gates, the `CX` gate is available. The `CX` gate operates on two qubits: a control qubit and a target qubit. If the control qubit is in state $|0\rangle$, the `CX` acts as identity. Otherwise, if the control qubit is in state $|1\rangle$, an `X` gate is applied to the target qubit, flipping it. The `CX` gate is sometimes called the `CNOT` gate.

C. RX Gates

The RX(θ) gate performs a rotation of θ radians around the x-axis of the Bloch sphere. The RX gate is not a native gate, but it can be decomposed into native basis gates RZ and SX gates.

D. Measurement Operation

When a qubit is measured, the result is a classical bit of information, either 0 or 1. The measurement process collapses the original qubit state, projecting it typically onto the z-axis of the Bloch sphere. Measurement results of 0 and 1 correspond to state collapse into $|0\rangle$ and $|1\rangle$, respectively. Measurement is an example of a non-unitary operation, as it cannot be reversed. This state collapse is irreversible; after a measurement is made, the original information about the qubit of the state is lost.

For a general qubit state $|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle$, the collapse is probabilistic. The probability of a measurement is the square of the magnitude of the coefficient of the corresponding eigenstate. So we measure 0 and 1 with probabilities $\cos^2(\theta/2)$ and $\sin^2(\theta/2)$, respectively. For example, if θ is $\pi/2$, then probability of 0 and 1 being measured should be 50%.

E. Reset Operation

Another non-unitary operation is the reset operation. The reset operation consists of first making a measurement of a qubit onto a classical bit c. Then, an X gate is conditionally applied to the qubit if classical bit c measures 1. In more detail, the measurement collapses the qubit to either the $|1\rangle$ or $|0\rangle$ state. In the former case, the classical bit reads 1, and an X gate is applied to, resulting in the $|0\rangle$ state. In the latter case, no X gate is applied and the qubit remains in $|0\rangle$.

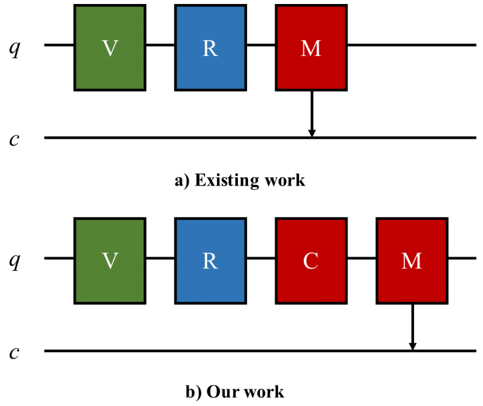
However, this design of the reset operation is susceptible to readout errors by the measurement operation. If a $|1\rangle$ is mistakenly read as 0 or a $|0\rangle$ as a 1, the reset operation incorrectly produces a final state of $|1\rangle$. This error on the real machines leads to a possible information leak to a malicious user on the same qubit [12].

F. Transpilation Process

Transpilation is the process of transforming an input circuit for execution on specific hardware. It involves matching the circuit to the topology of a quantum device and decomposing the user's gates into native gates supported by the hardware. Similar to classical compilers, transpilers also optimize the programs for performance. Optimizations may involve rewriting non-linear flow logic, processing iterative sub-loops and conditional branches, and other complex behaviors.

III. Extending Quantum Computer Reset Gate Attacks

Previous work by Mi et al. [12] has demonstrated information leaks across the reset operation on IBM Quantum computers. A malicious attacker can use a circuit consisting of just a measurement gate on the same qubit as a victim to extract information about the amplitude of the $|1\rangle$ state, or the equivalent θ angle, of the victim state before reset. We assume that a strong attacker is able to run their program immediately after the victim, on the same qubits that the victim used. We also assume the qubits used by the victim are reset before the attacker can access them. Before the victim's reset, we assume the victim likely ends their computation with a measurement on all involved qubits. This collapses the victim qubit states to either $|0\rangle$ (where $\theta = 0$) or $|1\rangle$ (where $\theta = \pi$). This scenario is most advantageous for the attacker since they only need to distinguish the two ends of the measured output frequency distribution.



a) Existing work

b) Our work

Fig. 1: Attack model, q represents target qubit and c represents its corresponding classical register. V is a shot of the victim's circuit, R is an inter-shot qubit reset mechanism, C is a concealing circuit used by attackers, and M is the measurement operation used by attackers to try to guess the state of the V before R.

It has been shown in the prior work that even with multiple reset gates before the attack, information leak still occurs. The attacker model of the prior work is shown in Figure 1a. In the figure, V represents the victim circuit, which includes the victim's final measurement. R represents one or more reset operations executed as a reset sequence between shots of circuits. C represents the attacker's concealing circuit, and M represents the attacker's measurement.

However, a very simple defense mechanism can easily detect such an attack: scan for user circuits consisting of only one measurement gate, or more generally any circuit that begins with a measurement gate and flag these as suspicious.

This work shows that an attacker can bypass such simple defenses, and also make a more potent attack circuit, by adding a concealing circuit C before the measurement. By using a concealing circuit C, the attacker can make their circuit look like a benign quantum circuit, but still be able to extract information across the reset operation as before. This work shows various concealing circuits and how attackers can recover information even if the concealing circuit C is between the reset operation (of the victim) and the measurement (of the attacker). Our attack model is shown in Figure 1b. The high-level idea behind the extended quantum computer reset gate attacks is that the concealing circuit C represents unitary operations that can be reversed. With knowledge of the measurement and the concealing circuit, the attacker can gain information about the state right before the concealing circuit, which is related to the victim's state right before the reset.

A. *Attack Objective*

The first objective of this research work is to analyze the different types of concealing circuits C that an attacker could utilize. Using concealing circuits, the attacker can make their circuit look like a benign circuit, making detection of the attack harder, while still being able to carry out the reset gate attack to learn some information about the state of the qubits prior to the reset. We consider circuits used for common or well-known quantum algorithms to be benign, such as Grover Search, quantum random number generator, and quantum Fourier Transform. Many common practical circuits can be found in the QASM Benchmark suite [11]. Ideally, the attacker would like their concealing circuit to resemble a common or practical algorithm to deceive any antivirus software into identifying the malicious attack circuit as part of a practical quantum program or a common circuit that a non-malicious user may reasonably write.

B. *Attacker Circuits*

This work analyzes a variety of possible concealing circuits C. We begin with simple single-qubit gates which are easily reversible, allowing the attacker to recover the most information about the qubit state. We then consider multi-qubit gates and more complex circuits that more closely resemble common benign circuits. Later we show which ones work well, and which ones do not.

- Identity Circuits – circuits consisting of an even number of single-qubit X gates on each qubit, such that the total effective angle of rotation θ is 0. Since effectively there is no rotation, the attacker's measurement should return the same values as it would be right after the reset operation. This choice of concealing circuit allows the attacker to most easily extract information.
- RX and RZ Gate Circuits – circuits consisting of single-qubit gates with effective θ (RX gate) rotation and ϕ (RZ gate) rotation. These gates are reversible: knowing the rotation angle, the attacker can infer the qubit 1-output probabilities as they would be right after the reset gate based on their measurement. As we demonstrate, certain rotation angles make the attack more difficult, while others still allow the attacker to make a meaningful measurement.
- CX Gate Circuits – circuits consisting of two-qubit CX gates where there is entanglement between qubits. The control qubits of CX gate experience delay (due to duration CX gate) but otherwise can be leveraged by an attacker since they do not experience any rotations; meanwhile, the state of the target qubits of CX gate depends both on the prior state and the control qubit, making attacker's use of that qubit more difficult.
- General Quantum Circuits – circuits that are general for quantum computing. We select some quantum circuits for common quantum algorithms from QASMBench suite [11], which are real quantum computing circuits. These include the 2- and 3-qubit Grover search circuits and the 4-qubit quantum random number generator (QRNG).

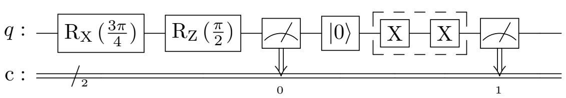


Fig. 2: Example of two X gate circuit used as a concealing circuit; any even number of X gates applied in sequence forms an identity circuit and can be evaluated for efficacy of the concealing circuit.

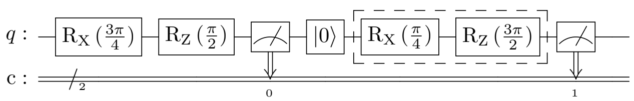


Fig. 3: Example of concealing circuit with RX and RZ gates, different number of RX and RZ gates and the angles can be evaluated for efficacy of the concealing circuit.

C. *Hiding Reset Operation Attack with Identity Circuits*

First, we experimented with using a series of X gates as the attacker circuit, as shown in Figure 2. For a variety of input states, we ran experiments increasing the number of reset gates and the number of X gate pairs, which we call the depth of the circuit. Since we use an even number of X gates, the concealing circuit is thus always equivalent to identity in this experiment group. As shown later in Figures 10 and 12, information leak still occurs with X gates added as a concealing circuit. Based on the measured 1-output frequency, the attacker can distinguish with high probability between victims initialized with $\theta = 0$ or $\theta = \pi$.

An attacker may try more complex, non-identity circuits, or try to attack victims after a larger number of reset gates to avoid detection. We explain these next.

D. *Hiding Reset Operation Attack with* RX *and* RZ *Gate Circuits*

Next, we considered RX and RZ rotation gates for the attacker to mask the attack. We ran two experimental groups. For the first set of attacks, we fixed the attack circuit depth at 1 RX and 1 RZ gate, and we varied the rotation angles. An example is shown in Figure 3.

For the second set of attacks, we fixed total rotation angles at $\theta = \pi$ and $\phi = \pi/2$. We vary the depth, or number of RX and RZ gates, while keeping the total equivalent rotation angles at a fixed sum of $\theta = \pi$ and $\phi = \pi/2$. For depth d, we use d copies of RX(π/d) followed by d copies of RZ(π/2d). An example with $d = 2$ is shown in Figure 4.

We chose $\theta = \pi$ because, based on preliminary testing, it is the best non-zero rotation angle for the attacker. For $\theta = \pi$, $\phi = \pi/2$ is the choice of ϕ angle that is best for the attacker.

E. *Hiding Reset Operation Attack with* CX *Gate Circuits*

Further, we considered circuits involving multiple qubits. We ran experiments with a series of CX gates, using the victim

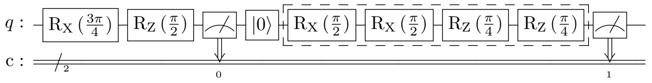


Fig. 4: Example of different concealing circuit with RX and RZ gates where the total rotation angles are fixed.

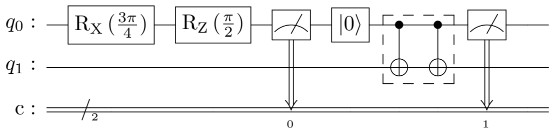


Fig. 5: Example of circuit with CX gates used as a concealing circuit, different number of CX gates can be tested for efficacy of the concealing circuit.

qubit as the control qubit. CX gates have a longer duration compared to single-qubit gates. While the control of the CX gate does not affect the qubit state, allowing the attacker to gain information about the victim. The main goal is to evaluate the effect of time delay on the success of the attack. We hope to gain insight into whether duration of a circuit could be used to classify potentially malicious circuits.

As shown in Figure 5, we repeat a number of CX gates with the victim qubit, q_0, as the control. The attacker only makes a measurement on the control qubit of the CX gates.

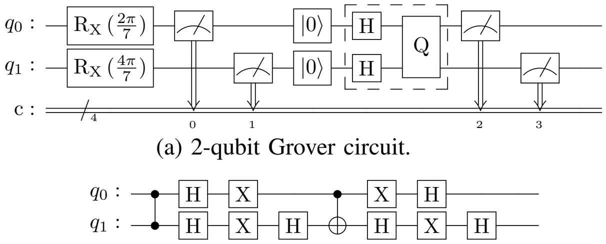
F. Hiding Reset Operation Attack with General Quantum Circuits

Aside from single-qubit concealing circuits and circuits with CX gates, an attacker may try more complex and deeper circuits to hide an attack. In particular, they could try to disguise their attack as a benign circuit. For example, we select some quantum circuits from QASMBench [11]. We evaluate whether it is possible for an attacker to perform a reset attack under our threat model using some common QASM benchmark circuits and conceal the attack circuit as a well-known quantum algorithm.

1) 2-Qubit Grover Search Circuit: We begin with the 2-qubit Grover search circuit. To start the search algorithm, the qubits need to be initialized into a uniform superposition with Hadamard gates. Then, the Grover operator, Q, is applied to amplify the amplitude of the correct answer via rotations done by Q. An example of 2-qubit Grover search is shown in Figure 6a.

We used Grover search with answer bitstring 11. The circuit for the algorithm is boxed in Figure 6a. The Grover operator Q is decomposed in Figure 6b. The attacker uses this circuit after the reset gates and before final measurement, like the previous attacks.

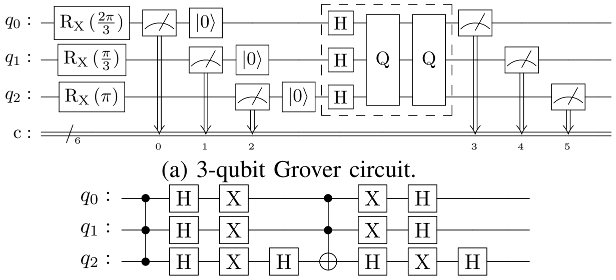
Unlike the single-qubit attack circuits, the attacker makes measurements on all involved qubits. The victim qubits are initialized with θ rotations independently of each other, that is, the rotation angles are not necessarily the same for each qubit. We limit the range of possible initial angles so that the total number of circuits for each trial does not exceed our limit



(a) 2-qubit Grover circuit.

(b) 2-qubit Grover circuit with operator Q decomposed.

Fig. 6: Example of using 2-qubit Grover circuit used as a conclealing circuit, circuits with different bitstrings and operators can be tested for efficacy of the conclealing circuit. The Hadamard, H, gate can be realized using the basis gates discussed in the text.



(a) 3-qubit Grover circuit.

(b) 3-qubit Grover circuit with operator Q decomposed.

Fig. 7: Example of using 3-qubit Grover circuit used as a conclealing circuit, circuits with different bitstrings and operators can be tested for efficacy of the conclealing circuit. The Hadamard, H, gate can be realized using the basis gates discussed in the text.

on the ibmq_jakarta machine of 300 circuits per job. For 2-qubit Grover, each qubit is initialized by the victim with a rotation of $\theta \in \{0, \frac{\pi}{7}, \frac{2\pi}{7}, \frac{3\pi}{7}, \frac{4\pi}{7}, \frac{5\pi}{7}, \frac{6\pi}{7}, \pi\}$.

2) 3-Qubit Grover Search Circuit: We also experimented with the 3-qubit Grover search circuit, which looks similar to 2-qubit Grover search, but has more gates and is deeper. Each qubit is initialized by the victim with a rotation of $\theta \in \{0, \frac{\pi}{3}, \frac{2\pi}{3}, \pi\}$. An example of 3-qubit Grover search is shown in Figure 7.

3) Random Number Generator Circuit: There are two small-scale circuits that do not use multi-qubit gates in QASMBench, namely, the quantum random number generator, and the inverse Quantum Fourier Transform (QFT). However, the inverse QFT circuit requires conditional operations, which are currently unavailable on IBM Quantum machines. So we consider the random number generator on 4 qubits.

The Quantum Random Number generator, shown in Figure 8, uses Hadamard gates to produce a uniform superposition before measurement. This attacker circuit has the smallest depth of the benchmarks tested by this paper, with a depth 1.

IV. Reset Operation Error Channel Analysis

Before we present the evaluation of the different attacks that use concealing circuits, we discuss the characteristics of the reset operation. Further, we compare the behavior of the reset operation on real ibmq_jakarta machine to two types

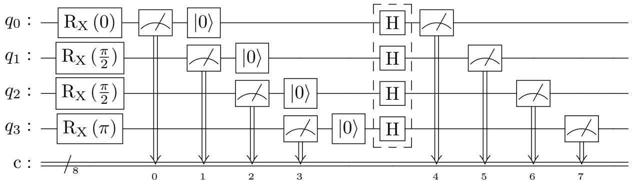


Fig. 8: Example of Quantum Random Number Generator (QRNG) used as a concealing circuit. The Hadamard, H, gate can be realized using the basis gates discussed in the text.

of simulation to motivate our use of real `ibmq_jakarta` for subsequent.

A. Behavior of Reset Operation

Qubits are often implemented with $|1\rangle$ as a higher energy state than $|0\rangle$. This results in a higher probability of an incorrect readout for a qubit in state $|1\rangle$ compared to state $|0\rangle$. Thus, we expect states with a higher amplitude of $|1\rangle$ to have a higher probability of being the $|1\rangle$ state after a reset [12]. This error of real machine resets is seen in Figure 9a, and allows the attacker to extract information about the θ angle of the victim qubit based on the measured 1-output frequency [12].

Given the state:

$$|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle,$$

recall that the probability of measuring 1 is $\sin^2\left(\frac{\theta}{2}\right)$ according to the Born rule interpretation. This motivates an error channel characterization [12] based on the probability of measuring 1 post-reset:

$$E(\theta) = a\left(b\sin^2\left(\frac{\theta}{2}\right) + (b-1)\frac{\theta}{\pi}\right) + c,$$

where $a \in [-1,1], b,c \in [0,1]$. On the domain, $\theta \in [0,\theta]$, the output probability looks like a sigmoid curve. This is seen in Figure 9a. This error channel parameterization is important to our attack evaluation in Section V.

B. Observed Fidelity Improvements of Reset Operations

NISQ quantum computers are noisy, and the error rates are constantly changing. Indeed, according to IBM's reported error rates through Qiskit's `IBMQBackend.properties()` method, we found that for qubit 0 of `ibmq_jakarta` , the readout error rate has dropped from 0.0360 to 0.0218 over the past year. In addition, the rate of measuring 0 from a $|1\rangle$ state dropped from 0.0464 to 0.0340, and the rate of measuring 1 from a prepared $|0\rangle$ state dropped from 0.0256 to 0.0096.

The current experimental results suggest that a similar reset error based on the amplitude of $|1\rangle$ is still present in IBM machines. In comparison to last year, the 1-output frequency of an attacker measuring the victim qubit after 6 resets still displays a significantly higher frequency for $\theta = \pi$ than for $\theta = 0$. At the same time, the noise is of a much smaller magnitude, as indicated by the smaller error bars.

With a decreasing noise-to-signal ratio, the possibility of a reset error channel attack is becoming actually greater. The attacker is able to recover more information from the victim with ever-increasing probability, even after numerous reset operations.

C. Study of Simulated vs. Real Reset Operations

We compared different types of simulated reset operations with the real `ibmq_jakarta` machine. We used `AerSimulator`, with a noise model directly imported from IBM's `ibmq_jakarta` backend. In theory, the simulator should behave as the real backend for all qubit gates. Based on our testing, the built-in simulated reset operation does not have the same error as the real machine's reset operation. While the real reset operation has a higher probability of an incorrect reset for qubits with a larger magnitude of $|1\rangle$, the simulated reset removes this: there is no clear correlation between the victim qubit's original theta angle and the output frequencies post-reset. The data is shown in Figure 9b.

Given the built-in simulated reset operation does not behave as a real one, we then attempted to replace the built-in reset operation with a measurement followed by an `X` gate conditioned on the measurement being 1 – this should in theory represent the behavior of the reset operation. We did observe more realistic results in the case of 1 reset, as the sigmoid shape can be seen in Figure 9c. However, the addition of two or more reset operations with the simulator results in noisy data, and no longer fits a sigmoid curve. This suggests that the simulated reset does not emulate the real machine when using a measurement followed by an `X` gate as the reset operation.

Both the simulator's built-in simulated reset operation and the measurement followed by `X` gate scheme on the simulator produce a lot of noise: the 1-output frequencies vary a lot depending on the victim qubit's ϕ angle compared to the real machine. At this time, the simulator is unable to accurately replicate the behavior of the reset operation on IBM Quantum machines, and our evaluation in the rest of the paper users' data from real `ibmq_jakarta` machine.

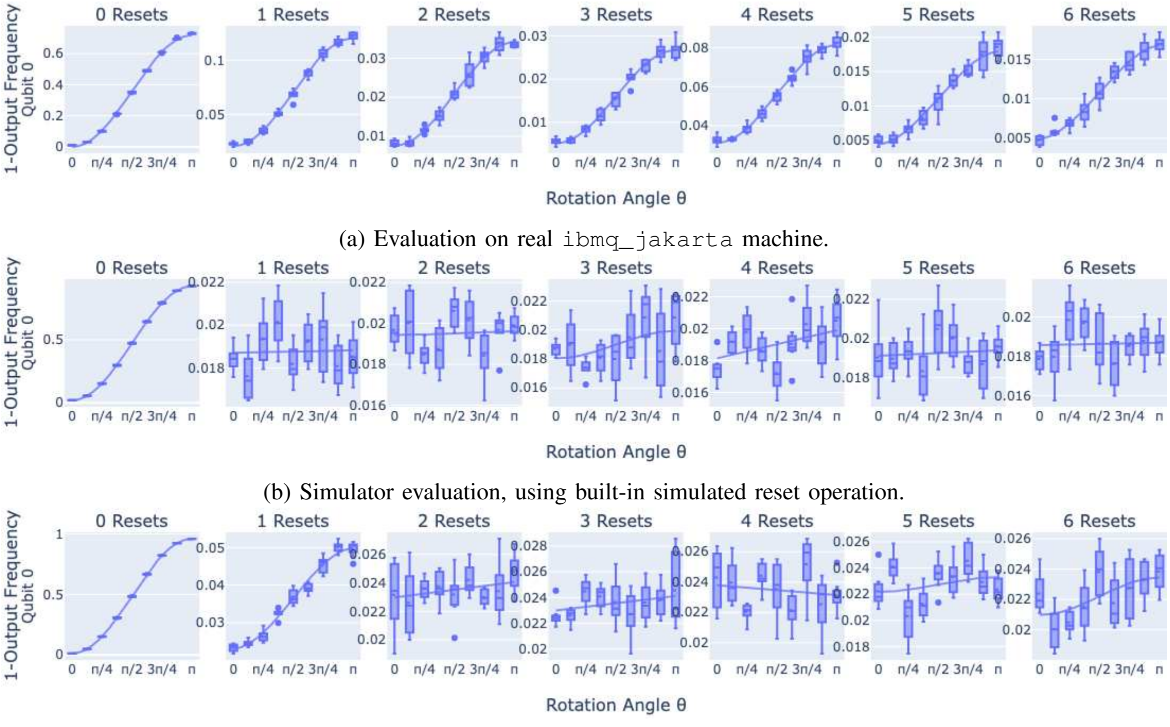
V. Evaluation of Efficacy of Concealing Circuits

In this section, we present evaluation results for different concealing circuits previously discussed in Section III. The concealing circuit evaluation is based on: 1) `X` gates, 2) `RX` and `RZ` gates, 3) `CX` gates, and 4) General quantum circuits. For all circuits, we ran experiments on `ibmq_jakarta` using a varying number of reset gates after the victim and a varying circuit depth for the concealing circuits, where possible.

A. Evaluation Metrics

To evaluate the effectiveness of each attack circuit, we use a metric of signal-to-noise ratio (SNR). We computed the SNR to estimate how much information the attacker could extract from the output frequency data when different types of concealing circuits are used.

We compute the error channel characterization parameter a, which represents the amplitude of our sigmoid fit. The fit is described in Section IV-A. We compute the standard deviation



(a) Evaluation on real `ibmq_jakarta` machine.

(b) Simulator evaluation, using built-in simulated reset operation.

(c) Simulator evaluation, using "measurement + `X` gate" approach to emulate reset operation.

Fig. 9: Qubit state retention, comparison of: (a) reset operation on the real machine, (b) simulated reset operation, and (c) simulated reset operation using "measurement + `X` gate" approach.

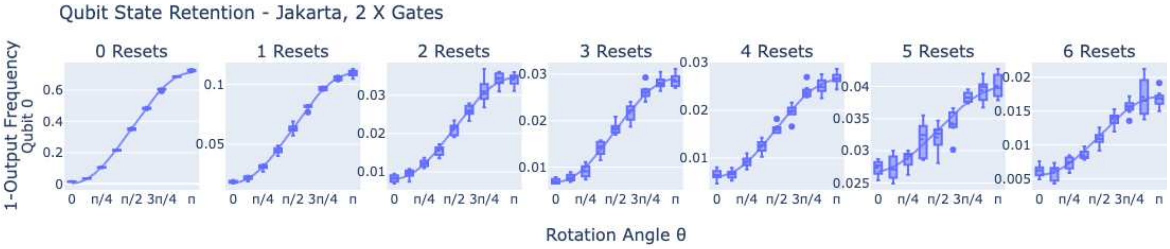


Fig. 10: Example 1-output frequency of `X` gate concealing circuit. Circuits with $0, 2, 4, 6, 8, 16$, and 32 `X` gates were used as the attacker circuit. Experiments done with qubit 0 of `ibmq_jakarta` . Only results for 2 `X` gates are shown, with the other graphs having a similar shape.

in 1-output frequency for each fixed θ as ϕ varies. Finally, we compute the average standard deviations over all input θ values, denoted σ. Then the signal-to-noise ratio is defined as a/σ, expressed on a log scale (decibels).

B. Reset Schemes

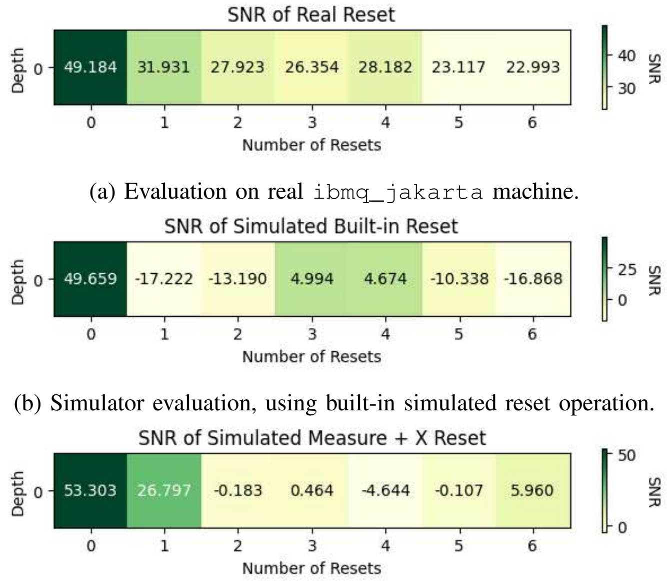
Using this metric, we can compare the different reset schemes described in Section IV-C. Figure 11 shows the SNR for the three different reset schemes. The SNR metric aligns with the analysis of Section IV-C. We observe a relatively strong SNR for the real reset. For the simulated reset, there is a sharp decline in SNR after adding the first reset. Using a measurement and X gate to simulate reset, the SNR for one reset is relatively high, but adding more resets decreases the SNR drastically.

C. Attack Involving Identity Circuits

We ran circuits with a series $0, 2, 4, 6, 8, 16$, and 32 X gates as the attacker circuit. For each attack circuit, we added up to 6 reset gates after the victim. All experiments were run on qubit 0 of `ibmq_jakarta` .

Figure 10 displays the 1-output frequency of each attack circuit as a function of the victim qubit's rotation angle θ. For the purposes of conserving space, only the results for 2 X gates are shown. The graphs for more resets display the same sigmoid shape.

We expect that as the depth of the circuit increases or the number of reset gates, the attacker's job becomes harder as more noise is introduced. Figure 12 shows the SNR plotted on a decibel scale for all depths of X gate circuits and all numbers of reset gates. As expected, increasing the number of resets results in decreasing the signal-to-noise ratio. The



(a) Evaluation on real `ibmq_jakarta` machine.

(b) Simulator evaluation, using built-in simulated reset operation.

(c) Simulator evaluation, using "measurement + X gate" approach to emulate reset operation.

Fig. 11: Comparison of SNR for reset on real machine, simulated reset, and simulated reset using "measurement + X gate" approach.

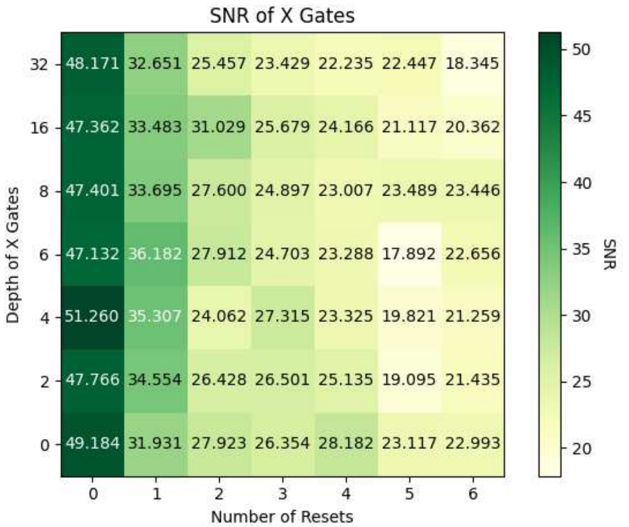


Fig. 12: SNR for X gate concealing circuit experiments. A series of up to 32 X gates were tested.

correlation coefficient between these two variables is -0.862, indicating a strong negative correlation. The most significant decrease in SNR resulted from the addition of the first reset gate, with subsequent resets having a lesser effect on SNR.

The depth of the circuit, measured as the number of X gates, did not appear to have much effect on the SNR, as there is no clear trend of the SNR as depth increases. The correlation coefficient between these two variables is -0.057, indicating no significant correlation.

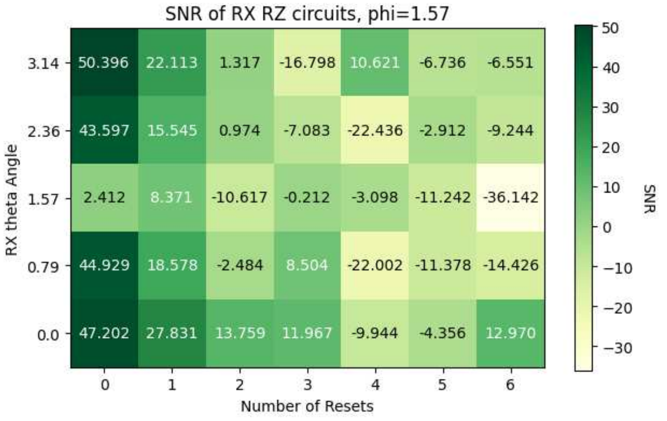


Fig. 13: SNR for the first set of RX and RZ attacker experiments. The rotation angles were varied while the depth was fixed at 1 of each gate.

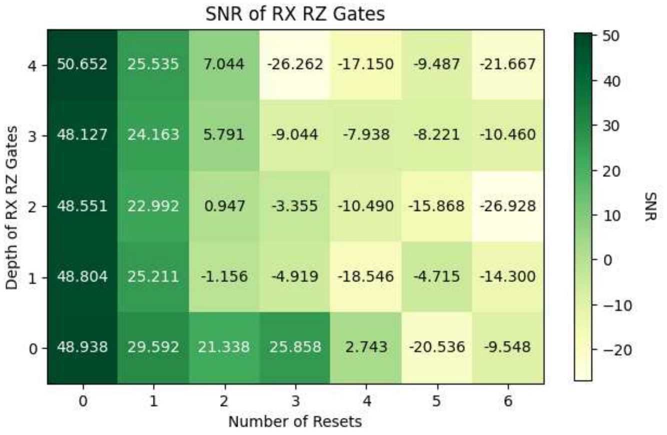


Fig. 14: SNR for the first set of RX and RZ attacker experiments. The rotation angles were varied while the depth was fixed at 1 of each gate.

D. Attack Involving RX *and* RZ *Gate Circuits*

In the first set of experiments, we used $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi\}$ and $\phi \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ for the attacker's RX and RZ gates, respectively. We observed that $\phi = \pi/2$ seems particularly beneficial for the attacker compared to other ϕ angles. The results for this ϕ angle are shown in Figure 13.

For $\theta = \pi/2$, the SNR is the lowest, meaning it is the most difficult for the attacker to extract information about the victim's initial angle. This coincides with our expectation, because after an RX rotation by $\pi/2$, both initial states $|0\rangle$ and $|1\rangle$ have the same output probability of $\frac{1}{2}$.

As the θ angle changes from $\pi/2$ towards 0 or π, it becomes easier for the attacker to distinguish the victim's initial state. Increasing the number of resets generally decreases the signal-to-noise ratio, as expected.

We then experimented by varying the depth of RX and RZ gates, while keeping the total rotation angles at $\theta = \pi$ and $\phi = \pi/2$. Each rotation gate used the same θ or ϕ angle. For example, for depth 2 we used two RX$(\pi/2)$ gates and

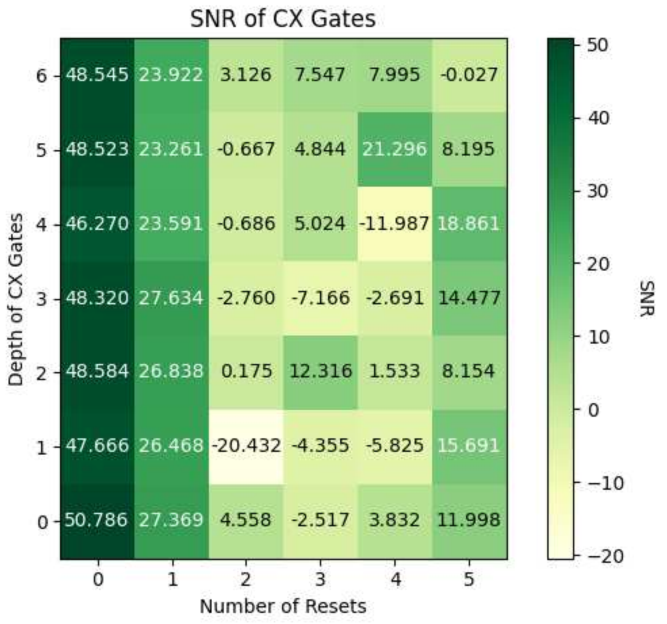


Fig. 15: SNR for CX gate attacker experiments. The CX gates were used qubit 0 as the control qubit. Output results and SNR are based on qubit 0.

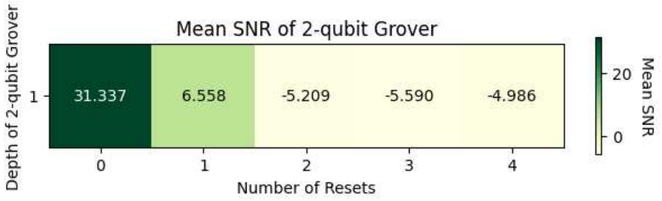


Fig. 16: SNR for 2-qubit Grover circuit experiments. Average gradient is used as the measure of signal for calculating the SNR.

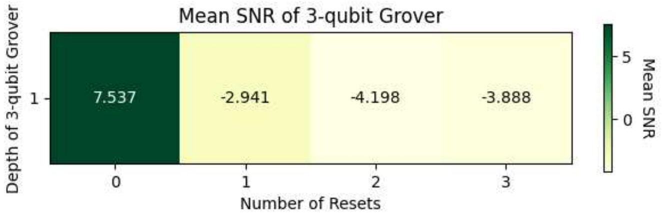


Fig. 17: SNR for 3-qubit Grover circuit experiments. Average gradient is used as the measure of signal for calculating the SNR.

two RZ$(\pi/4)$ gates. We ran a control group with no attacker, labeled depth 0 in Figure 14.

For 3 resets, increasing the depth decreases the SNR. However, for 2 resets, the opposite effect occurs. In general, the correlation between depth and SNR is -0.14, indicating little to no correlation.

E. Attack Involving CX *Gate Circuits*

We experimented with a series of CX gates as the attacker. We used qubit 0 on ibmq_jakarta as the victim qubit, and we added up to 6 CX gates in series after the reset gates, using the victim qubit, qubit 0, as the control qubit.

Interestingly, increasing the number of reset gates from 0 to 1 or from 1 to 2 decreases the SNR, while increasing the number of reset gates beyond 2 seems to increase the SNR, on average. For any number of reset gates, the depth of the CX gates does not have a strong correlation with the SNR, with a correlation coefficient of 0.039.

Due to numerous job requests, the circuits for this set of experiments were executed over several days. This may have introduced noise in the data, as IBM Q machines have slightly different error rates across different execution times.

F. Attack Involving Grover search Circuits

To compute the signal-to-noise ratio with a multi-qubit circuit, we need a new measure of signal. For each qubit, we consider the 1-output frequency as a function of all qubits' initial angles. We compute the sum of the squares of the gradients with respect to each input dimension, then take a square root. This final value, the Root-Mean-Square (RMS) gradient, is roughly a measure of the rate of change in 1-output frequency as we change the input angles. As a measure of noise, we use the average standard deviation in output frequency, as in the single-qubit case. For each combination of initial angles, we did 8 trials. We compute the quotient as the SNR for each qubit.

Figure 16 shows the results for 2-qubit Grover search. We observed sharp declines in SNR after 1 and 2 resets. Increasing the number of resets past 2 does not appear to significantly impact the SNR.

Figure 17 shows the results for 3-qubit Grover search.

We observed a sharp decline in SNR after 1 reset. Increasing the number of resets past 1 does not appear to significantly impact the SNR. We also note the difficulty of drawing a conclusion given the limited data we have, especially for 3-qubit Grover's.

G. Attack Involving QRNG Circuits

Below are the results for the QRNG circuit on four qubits. We used an initial rotation angle of $\theta \in \{0, \pi/2, \pi\}$ for each qubit. For every combination of initial angles, we ran 6 trials.

For three or more resets, the IBM computers ran into internal errors. This error also appeared for ibmq_jakarta for large numbers of resets on the Grover search algorithms.

Figure 18 represents the mean SNR of all four qubits of the QRNG circuit. Interestingly, increasing the number of reset gates up to 2 does not seem to have a significant impact on the SNR.

H. Summary of the Attacks and the Evaluation

We have shown that for single-qubit gates used in the concealing C circuit, the attacker may use simple identity circuits consisting of pairs of X gates, or circuits consisting of RX and RZ gates. For multi-qubit gates, an attacker can also try to hide the attack by using a concealing circuit with

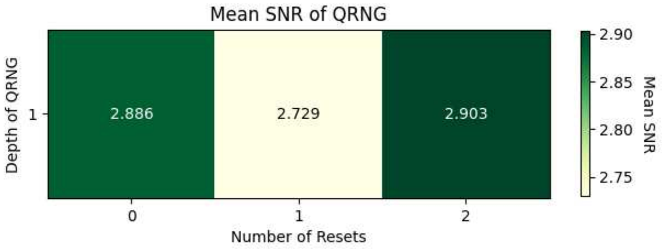


Fig. 18: SNR for QRNG benchmark circuit experiments. Hadamard gates on each qubit are used to achieve a uniform, random output. Average gradient is used as the measure of signal for calculating the SNR.

CX gates, as long as the targeted qubit for the attack is the control qubit of the CX gate. We also showed conditions under which the attack becomes more difficult, such as when qubits are targets of CX gate or other multi-qubit gates are involved. We confirm our expectation by running selected QASMBench circuits, and showing that it is difficult for the attacker to leak the victim's state, due to the presence of multi-qubit gates or other non-identity gates, as is the case for using 2-qubit or 3-qubit Grover search as the concealing circuit C. Based on these findings, a defense for our extended reset operation attack can be developed.

VI. Defense Against the New Reset Operation Attacks

We provide a number of compile-time heuristics that can be used to detect possibly malicious attacks that try to use concealing circuits with a measurement to perform a reset operation attack. Our compile-time solution is complimentary to the existing "secure reset" work [12], which is a run-time solution. Further, our approach is different from the existing quantum computer antivirus [7], [8], which focuses on the exact quantum circuit pattern matching.

A. Detecting Attacks that use Identity Circuits

In the case that the attacker places an identity circuit before the measurement, we scan all gate operations done after the last reset gate and before the final measurement. We use Qiskit's Operator class to convert any potential adversarial circuit into its matrix representation. Then, we check if this matrix is an identity. This is efficient for circuits with a small number of qubits. For large circuits, we can loop through each qubit and check the gates that operate on it. If these are single-qubit gates only, and if the operations on each qubit are equivalent to identity, our program flags the circuit as suspicious.

If a circuit consists of an identity followed by measurement, our program will flag it as suspicious. The size of the matrix representation scales exponentially with the number of qubits involved, so it is limited to smaller circuits. In testing, we generated 100 random 7-qubit circuits of depth 10, and our program successfully and efficiently flagged all of these as identity circuits.

B. Heuristics for General Attack Detection

In the most general case, the attacker may use a non-identity circuit as a concealing circuit, or he or she may use many qubits that make matrix representations infeasible to work with. In this case, we present an approach that considers each qubit one at a time.

For each qubit, we can compute the matrix representation of all gates involving the specific qubit. We first check if the qubit is involved in any multi-qubit gates. Based on our results, circuits involving multi-qubit gates are not susceptible to reset attacks. However, single-qubit gates introduce little error, and even at large depths, the attack can still extract information on these qubits. Thus, any qubits involved in only single-qubit gates, or the control qubit of a CX gate, will be noted by our program.

In the case that a qubit is only involved in single-qubit gates, our program checks if the circuit applies an effective RX rotation on the qubit. Based on our results, an effective RX rotation close to $\pi/2$ makes it difficult for the attacker to perform the attack. So, we propose flagging any qubit with effective rotation $\theta > 3\pi/4$ or $\theta < \pi/4$.

Note that for most circuits, most qubits will have more complex operations that cannot be reduced to an equivalent RX rotation. In this case, our program can still note whether the qubit is effectively identity, or only involves single-qubit gates.

C. Implementation

We assume our program has access to the circuit that is to be checked, e.g., our program can be used by IBM to scan submitted circuits before they execute on the quantum computers. Given an input circuit, it is simple to count the circuit depth of the possibly malicious input circuit. Additionally, Qiskit provides functionality to convert circuits into their matrix representation. Since the number of resets used is controlled by the quantum computer provider, we assume the number of resets is an input or configuration given to our program.

To scan circuits, we first extract the gates from the input quantum circuit, and for any given qubit, check if the gate operates on the qubit. If so, we save the instructions for the gate. In the end, we make a quantum circuit from the list of instructions, yielding the subset of the original circuit that involves each specific qubit. On this smaller circuit, we compute the matrix representation and check for the existence of multi-qubit gates, equivalence to identity, and equivalence to a single RX rotation.

Based on our testing, for attacker circuits of 32 X gates, 6 CX gates, 2-qubit Grover, 3-qubit Grover, and the QRNG Benchmark, our antivirus program can complete a scan in 0.017 seconds, 0.009 seconds, 0.024 seconds, 0.130 seconds, and 0.017 seconds, respectively.

VII. Related Work

Security of quantum computers, which is different from research such as on post-quantum cryptography, is an emergent research direction. There are currently few works we can compare to. On the attack side, our work extends work by Mi et al. [12] who proposed an attack on the reset operations. Our work extends this prior work and shows more advanced attacks where use of concealing circuit is used to help hide the attacker while still allowing for information leak to be extracted by the attacker. Besides, insecure reset is also a problem when considering lower-level support, such as the higher-energy attacks proposed in [15], which demonstrated that attackers can abuse higher-energy states to bypass normal quantum gates and operations.

On the defense side, previous work has suggested an "antivirus" program that can be used to detect malicious quantum circuits [7], [8]. Our work and defenses could naturally form an alliterative antivirus approach. Our work does not require the use of directed acyclic graphs (DAGs), but instead scans individual qubits and computes the matrix form of the input circuit. Our defense program could be incorporated into the antivirus as a new feature.

Nowadays, NISQ quantum computers are prone to errors and noise from many sources. Besides decreasing the fidelity of the quantum computing programs' results, these errors and noise may also be taken advantage of by attackers to perform malicious attacks or retrieve secret information. For instance, the crosstalk error is actively researched to damage victim's quantum jobs in the multi-tenant quantum computing architecture [5]–[9]. This attack is realized by paralleling quantum circuits, with one of the circuits generating a large crosstalk effect and interfering with the other circuit that is running at the same time. Another example is the reset attack that is mainly studied in this paper [12], where the remaining state information preserved after reset operations can be collected by attackers.

Apart from the privacy or security issues due to errors, the current workflow, architecture, system, or hardware may also not be trusted. The availability of quantum programs to quantum cloud providers makes it easy for untrusted quantum cloud providers to steal the quantum programs, or quantum intellectual properties (IPs) [13], [14]. Other than the cloud platforms themselves, quantum IPs may also be reconstructed from power side-channels of the quantum computer control equipment [16]. Possibly containing sensitive information, quantum IPs are important to be protected.

VIII. Conclusion

In this work, we demonstrated how a set of new, extended reset operation attacks could lead to critical information leakage from quantum programs executed in a quantum computing cloud environments. This work showed that this new kind of reset operation attack could be more stealthy than the previous reset operation attacks, by hiding the intention of the attacker's circuit. Based on the findings, this work showed a set of heuristic defenses that could be applied at compile time to check and flag the new kind of malicious circuits.

Acknowledgments

This work was supported in part by National Science Foundation grants 2312754 and 2245344.

References

[1] "Ibm quantum," https://quantum-computing.ibm.com/.

[2] "Ibm's target:a 4000-qubit processor by 2025," https://spectrum.ieee.org/ibm-quantum-computer.

[3] "Quantum circuits," https://quantumcircuits.com/.

[4] "Rigetti computing," https://www.rigetti.com/.

[5] A. Ash-Saki, M. Alam, and S. Ghosh, "Analysis of crosstalk in nisq devices and security implications in multi-programming regime," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, ser. ISLPED '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 25–30. [Online]. Available: https://doi.org/10.1145/3370748.3406570

[6] ——, "Experimental characterization, modeling, and analysis of crosstalk in a quantum computer," *IEEE Transactions on Quantum Engineering*, vol. 1, pp. 1–6, 2020.

[7] S. Deshpande, C. Xu, T. Trochatos, Y. Ding, and J. Szefer, "Towards an antivirus for quantum computers," in *2022 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2022, pp. 37–40.

[8] S. Deshpande, C. Xu, T. Trochatos, H. Wang, F. Erata, S. Han, Y. Ding, and J. Szefer, "Design of quantum computer antivirus," in *Proceedings of the International Symposium on Hardware Oriented Security and Trust*, ser. HOST, May 2023.

[9] ——, "Design of quantum computer antivirus," in *2023 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2023, pp. 260–270.

[10] IBM Quantum, "Ibm unveils 400 qubit-plus quantum processor and next-generation ibm quantum system two," 2022, https://newsroom.ibm.com/2022-11-09-IBM-Unveils-400-Qubit-Plus-Quantum-Processor-and-Next-Generation-IBM-Quantum-System-Two.

[11] A. Li, S. Stein, S. Krishnamoorthy, and J. Ang, "Qasmbench: A low-level quantum benchmark suite for nisq evaluation and simulation," *ACM Transactions on Quantum Computing*, 2022.

[12] A. Mi, S. Deng, and J. Szefer, "Securing reset operations in nisq quantum computers," in *Proceedings of the Conference on Computer and Communications Security*, ser. CCS, November 2022.

[13] T. Trochatos, C. Xu, S. Deshpande, Y. Lu, Y. Ding, and J. Szefer, "Hardware architecture for a quantum computer trusted execution environment," 2023.

[14] ——, "A quantum computer trusted execution environment," *IEEE Computer Architecture Letters*, vol. 22, no. 2, pp. 177–180, 2023.

[15] C. Xu, J. Chen, A. Mi, and J. Szefer, "Securing nisq quantum computer reset operations against higher energy state attacks," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 594–607. [Online]. Available: https://doi.org/10.1145/3576915.3623104

[16] C. Xu, F. Erata, and J. Szefer, "Exploration of power side-channel vulnerabilities in quantum computer controllers," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 579–593. [Online]. Available: https://doi.org/10.1145/3576915.3623118