Stealthy-Shutdown: Practical Remote Power Attacks in Multi-Tenant FPGAs

Yukui Luo[1], Cheng Gongye[1], Shaolei Ren[2], Yunsi Fei[1], and Xiaolin Xu[1]
[1]Dept. ECE, Northeastern University, MA, USA
[2]Dept. ECE, University of California Riverside, CA, USA
{luo.yuk, gongye.c, y.fei, x.xu}@northeastern.edu, sren@ece.ucr.edu

***Abstract*—With the deployment of artificial intelligent (AI) algorithms in a large variety of applications, there creates an increasing need for high-performance computing capabilities. As a result, different hardware platforms have been utilized for acceleration purposes. Among these hardware-based accelerators, the field-programmable gate arrays (FPGAs) have gained a lot of attention due to their re-programmable characteristics, which provide customized control logic and computing operators. For example, FPGAs have recently been adopted for on-demand cloud services by the leading cloud providers like Amazon and Microsoft, providing acceleration for various compute-intensive tasks. While the co-residency of multiple tenants on a cloud FPGA chip increases the efficiency of resource utilization, it also creates unique attack surfaces that are under-explored.**

In this paper, we exploit the vulnerability associated with the shared power distribution network on cloud FPGAs. We present a stealthy power attack that can be remotely launched by a malicious tenant, shutting down the entire chip and resulting in denial-of-service for other co-located benign tenants. Specifically, we propose *stealthy-shutdown*: a well-timed power attack that can be implemented in two steps: (1) an attacker monitors the real-time FPGA power-consumption detected by ring-oscillator-based voltage sensors, and (2) when capturing high power-consuming moments, i.e., the power consumption by other tenants is above a certain threshold, she/he injects a well-timed power load to shut down the FPGA system. Note that in the proposed attack strategy, the power load injected by the attacker only accounts for a small portion of the overall power consumption; therefore, such attack strategy remains stealthy to the cloud FPGA operator. We successfully implement and validate the proposed attack on three FPGA evaluation kits with running real-world applications. The proposed attack results in a *stealthy-shutdown*, demonstrating severe security concerns of co-tenancy on cloud FPGAs. We also offer two countermeasures that can mitigate such power attacks.

***Index Terms*—FPGA, Denial-of-service, Side-channel, Multi-tenant, Power attack**

I. Introduction

Field programmable gate arrays (FPGAs) has attracted increasing attention and popularity in recent years for accelerating compute-intensive applications, such as machine learning model training and inference [1]. FPGA vendors, such as Xilinx and Intel, have started developing more powerful FPGA chips and development tools. For example, Xilinx has proposed a new software-defined everything (SDx) framework that includes high-level synthesis (HLS) algorithms [2], various IP cores, and an advanced compiler [3]. Both the high capacity (e.g., number of programmable units) of modern FPGAs and their reconfigurability have turned them into a powerful platform for cloud services, adopted by leading cloud-service vendors like Amazon AWS and Microsoft Azure [4] [5].

While providing low-cost and high-performance computing services through virtualizing FPGA chips, integrating FPGA into cloud services also brings new security concerns. Particularly, the co-residency of multiple cloud tenants on a single FPGA chip creates new vulnerabilities due to capacitive coupling and shared resources and opens doors to malicious tenants. For example, it has been reported that the crosstalk between FPGA long-wires can be exploited as a new side-channel to steal secret information from victim tenants [6] [7]. The power distribution network of FPGA could also be manipulated by an attacker to inject faults [8]–[10] or shut down the FPGA [11]. The *root* cause for these new vulnerabilities is the co-residency of multiple tenants on FPGA chips.

This paper presents Stealthy-Shutdown, which launches well-timed power load attacks (a.k.a., power attacks) to shut down the shared cloud FPGA chips by exploring the lowest threshold of supply voltage, which is necessary to maintain the regular operation of cloud FPGAs. As typical workloads on cloud FPGAs are compute-intensive and therefore power-hungry, the proposed attack method relies on capturing such high power-consuming moments to trigger injecting power overload for well-timed attacks. A lightweight sensor based on ring oscillator (RO) is used to monitor the on-chip voltage fluctuations, whose output serves as real-time side-channel information to find the moments of high power usage. Our proposed power attack is more stealthy than the prior work [11] and can evade detection. We successfully demonstrate that Stealthy-Shutdown is effective, powerful, and low-cost, resulting in precise denial-of-services (DoS) for victim tenants.

The main contributions of this paper are summarized as follows:

- We present a stealthy power attack that injects malicious power load guided by side-channel information. Such an attack strategy enables an adversary to devastate (e.g., shut down) the normal FPGA operation with smaller circuit overhead than launching aggressively high power consumption. This strategy can be stealthily adopted by a malicious FPGA tenant subject to strict power monitoring conditions, e.g., where the cloud FPGA operator checks the power consumption of each tenant's program.

2576-6996/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCD50377.2020.00097

- We successfully validate the stealthy-shutdown on three evaluation kits with three types of FPGA chips for different application scenarios, including boards AX7103 with a low-cost and small complexity FPGA chip XC7A100T, ZCU104 for embedded vision applications with a multi-processor system-on-chip device XCZU7EV, and ADM-PCIE-7V3 intended for data center applications with a high-complexity FPGA chip XC7VX690T. Specifically, these FPGA evaluation kits use different power regulation strategies (Section III-A). The practicality and effectiveness of stealthy-shutdown are verified using a real-world compute-intensive application: bitcoin mining program for FPGAs.

The remainder of this paper is organized as follows. Section II presents the background and related works of the proposed attack, including the threat model of multi-tenant FPGAs and other recent attacks based on the similar threat model. Section III elaborates the technical details of the proposed stealthy-shutdown attacks. Specifically, the vulnerabilities with the power regulation mechanism of three commercial off-the-shelf FPGAs are explored. Section IV describes the experimental setup with real-world applications, the attack strategies and validation results are also presented in this section. Section V concludes this paper with the discussion of future works.

II. Background and Related Works

This section reviews the threat model used by this work and other recent attacks in the context of multi-tenant FPGAs.

A. The threat model for multi-tenant FPGAs

Without loss of generality, we consider a representative threat model that has been used in a few recent works [6], [12]–[14]. The threat model is described as follows: (1) multiple independent tenants *co-reside* on an FPGA chip, and their circuits can be executed simultaneously; (2) each tenant has the flexibility to program its design in desired FPGA regions (if not occupied by other users); and (3) all tenants co-residing on an FPGA chip share resources, such as the power supply through the power distribution network and the long-wire bus line connected with the external I/O; meanwhile, there is no direct physical interaction or algorithm-level sharing (e.g., shared logic or circuit) among different tenants. Note that in this threat model, while tenants may not be trusted, the cloud operator is a trusted entity.

B. Recent attacks on multi-tenant FPGAs

1) Attacks via crosstalk between FPGA long-wires: Crosstalk is caused by the significantly reduced dimension of integrated circuits, i.e., the reduced distance between two metal wires formulates them as a capacitor [15]–[17]. Consequently, the crosstalk in FPGAs can delay or even change the signal values [18] [19]. The crosstalk between FPGA long-wires has been exploited as a new side-channel, with which a malicious tenant can steal secret from other tenants [6] [12]. The crosstalk-based side-channel attack has been validated with various FPGAs, including those from Xilinx [12] and Intel [20], respectively. More recently, it is demonstrated that such attacks could be mitigated using the long-wire isolation and obfuscation techniques [21].

2) Attacks on the power distribution network: The hardware resource sharing on cloud FPGA makes it possible for a malicious tenant to create damages for other tenants. A few power attacks have been recently studied under the threat model described above. The *co-residency* of multiple tenants on an FPGA chip makes the shared power distribution network a new attack surface, which can be exploited for malicious purposes in various ways. In [22] [23], the power trace of a victim tenant is collected by RO-based power sensors for power analysis attacks. Key extraction from advanced encryption standard (AES) is successfully demonstrated in [8] based on the RO-caused voltage drop. The entropy of true random number generator (TRNG) is corrupted as well by such power-related attacks in multi-tenant FPGAs [9]. In [24], it has been shown that such voltage drop can also be caused by benign but compute-intensive applications, which is classified as a reliability issue instead of a security one.

C. Difference between this and previous works

The related works to this one include [8], [9], [11], and [8], which focus on causing timing violations with a significant voltage drop. The objective of these attacks are extracting secret information like the key of AES or degrading the performance of a circuit. In [11], a large number of ROs are repeatedly turned on and off to overload the power regulator. As a result, the FPGA operation could be crashed. However, implementing such attacks requires a single FPGA user to consume a large amount of power within a short time duration, which is easily to be detected if the operator of the cloud FPGA service can monitor the real-time current. This work, considering the existence of other parallel compute-intensive applications on an FPGA chip, explores the practicality of implementing power attacks in a stealthy way. Specifically, the proposed attack strategy enables the attacker to achieve DoS on the cloud FPGA without consuming a large amount of power but just injecting an extra power load at critical timing points. This stealthy characteristic makes the detection of such attacks more challenging. Also, another important contribution of this work is the experimental validation with several FPGA boards that have different power regulation mechanisms and application scenarios, which demonstrates the practicality of the stealthy-shutdown attack. The difference between this and other attacks are summarized in Tab. I.

III. Stealthy-Shutdown Attacks on Cloud FPGAs

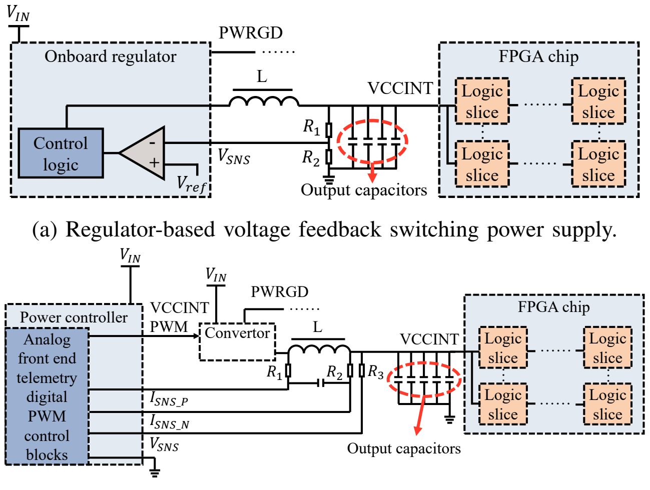
This section presents technical details of stealthy-shutdown. The related work [11] reveals that there exists a lowest threshold supply voltage that can guarantee the normal operation of an FPGA, and any supply voltage lower than this threshold will force the FPGA to crash. Additionally, we find that the shutdown of an FPGA is mainly regulated by the on-board power management system, and the threshold is related to the

TABLE I: Comparison of stealthy-shutdown and other related works.

	FPL'17 [11]			DATE'19 [9]	CHES'18 [8]	This work		
Attack mechanism	Activating ROs with a high (or low) frequency to impact the FPGA			Timing-constraints violation caused by a large number of ROs		Well-timed power injection guided by side-channel information		
Objective	DoS or fault injection on the FPGA			Circuit performance degradation	Secret extraction	Stealthy DoS on the FPGA with an existence of other applications		
Evaluation kit	ML605	KC705	Zedboard	VC707	DE1-SoC	AX7103	ADM-7v3	ZCU104
Malicious overhead	$\sim$ 12.4%	$\sim$ 11.8%	$\sim$ 12.8%	$\sim$ 24%	35% $\sim$ 45%	$\sim$ 5%	$\sim$ 8%	$\sim$ 5%

trigger condition of the system overload protection. In a multi-tenant cloud FPGA scenario, most applications are compute-intensive with a high throughput, which incur a high power-consumption. The proposed stealthy-shutdown attack aims to imitate this issue and to shut down the multi-tenant cloud FPGA. Note that although the study in this work uses Xilinx FPGAs for a concrete example, our findings and conclusions also apply to FPGAs from other vendors like Altera/Intel, which employ a similar power distribution network [10].

A. Power Regulation on FPGAs



(a) Regulator-based voltage feedback switching power supply.

(b) Pulse-width modulation-based power controller.

Fig. 1: Two power management strategies on FPGA boards.

To provide reconfigurable computing capabilities, most modern FPGAs are built with various IP cores (blocks), such as digital signal processing (DSP), clock management tile (CMT), and block memory generator (BRAM). For safety and reliability purposes, the power-up sequencing of these blocks is regulated by FPGA vendors based on their voltages (e.g., from low to high) [25]–[27]. For example, most Xilinx FPGAs firstly turn on the internal power supply VCCINT and blocks RAM supply VCCBRAM. Following that, the auxiliary power supply VCCAUX and the output drivers' supply VCCO are activated. On the contrary, when an FPGA is turned off, these blocks are powered off in a reversed order. Fig. 1 illustrates two primary schemes used to regulate voltage on FPGAs, in which only the first activated regulator of the power management system is shown. The power regulators for these supply voltages are cascaded through the PWRGD ("power good") [28] for control purposes. For example, if PWRGD of the first activated regulator is lower than expected, the second activated regulator will turn off its power supply.

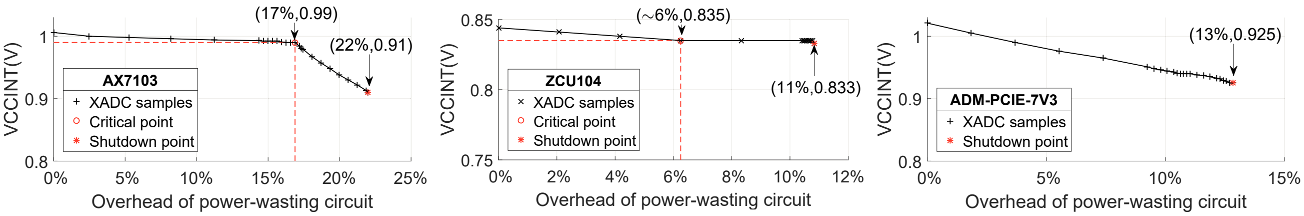
This work focuses on VCCINT, the direct core power supply for programmable units on an FPGA, which is generated by the first activated regulator. One important usage of the voltage regulator is to protect the operation of an FPGA. For example, when the power supply (VCCINT) of an FPGA chip is overloaded (i.e., with a higher current demand), the capacitors are discharged to supply the extra current that regulators cannot provide. In the design of a power supply system, these capacitors are usually sized according to the need for extra current. The recommended minimum output capacitance (C_O) is designed to compensate the current difference for at least two clock cycles with a tolerable voltage drop [28], as calculated in Eq. (1),

$$C_O = \frac{2 \times \Delta I_{out}}{f_{sw} \times \Delta V_{out}} \tag{1}$$

where ΔI_{out} and ΔV_{out} are the changes in output current and voltage, respectively, and f_{sw} denotes the regulator switching frequency. Therefore, if the current demand remains high, these capacitors will not be able to compensate the voltage drop and the regulator will step down or even turn off the output voltage for safety reasons [28]. Taking the TPS54620 regulator used in the AX7103 FPGA development board as an example (shown in Fig.1a), the voltage feedback signal $\mathtt{V}_{SNS}$ is used as an indicator and compared with the reference voltage $\mathtt{V}_{ref}$. If $\mathtt{V}_{SNS}$ is lower than a particular ratio (e.g., 91%) of the nominal VCCINT, the regulator will shut down [28]. Similarly, the pulse width modulation (PWM)-based power controller shown in Fig. 1b also leverages capacitors and feedback signals (I_{SNS_P}, I_{SNS_N}, and $\mathtt{V}_{SNS}$) to measure the current difference and step down or turn off the output voltage accordingly [29].

B. Validation of Voltage Drop by Power Attacks

Based on the power regulation mechanism of FPGAs, we conclude that there exists a maximum power capacity that the power regulator system can provide. In other words, if the power supply of an FPGA is overloaded by compute-intensive applications, then the power regulator may turn off the power supply, thereby shutting down the FPGA. To further validate our conclusion, we record the consequence of power overloading by instantiating a number of power-wasting circuits across three FPGA chips (XC7A100T on AX7103



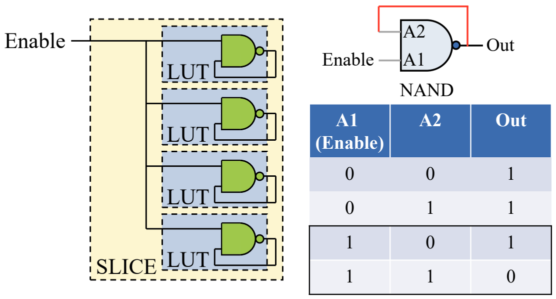
(a) AX7103 FPGA board

(b) ZCU104 FPGA board

(c) ADM-PCIE-7V3 FPGA board

Fig. 2: The validation of voltage drop and shut down of three FPGA boards using power-wasting circuits. It can be found that due to the utilization of different power regulator systems, these FPGAs have different voltage drop styles.

board [30], XC7VX690T on ADM-PCIE-7V3 board [31], and XCZU7EV on ZCU104 board [29]), which mimic the voltage drops caused by practical compute-intensive applications. The power-wasting circuit is composed of a number of power-wasting cells, in this paper, each of them is implemented within an FPGA Slice. Specifically, each power-wasting cell consists of four inverting logic components (e.g., NAND) that form a RO. The schematic of a RO circuit is shown in Fig. 3a, in which each NAND gate has an Enable signal controlled by the input node A1 of a LUT6 component, and the truth table of the NAND gate is illustrated in Fig.3b. Therefore, an adversary can enable a large number of these power-wasting cells to generate voltage drop immediately.



A1 (Enable)	A2	Out
0	0	1
0	1	1
1	0	1
1	1	0

(a)

(b)

Fig. 3: (a) RO-based power-wasting circuit. (b) Truth table and gate level schematic of NAND.

To comprehensively study the voltage drop, we turn on power-wasting circuits with several steps to find out the minimum overhead that can cause the system shutdown. In the first five steps, we use a coarse-grained calibration. We activate $1.8\% \sim 3.0\%$ (FPGA slices) power-wasting circuits in each step. Starting from the 6^{th} step, we apply a fine-grained calibration. We activate $0.1\% \sim 0.5\%$ (FPGA slices) power-wasting circuits incrementally until the system shuts down. The minimum overhead of power-wasting circuits (**PW**$_{oh}$) for each experimental platform is reported in Tab.II.

The Xilinx analog-to-digital converter (XADC) is used to obtain the real-time on-chip voltage on each FPGA. As an on-chip sensor, the main function of XADC is to monitor the on-chip voltage (e.g., VCCINT) and the temperature of an FPGA. An XADC consists of an ADC circuit and several sensors [32]. It is powered by VCCADC, and thus its measurement accuracy is immune to the voltage drop of VCCINT.

Fig. 2 illustrates the overhead of active power-wasting circuits (the percentage of used FPGA Slices) and the corresponding voltage drop measured by XADC. Voltage drop is observed on all the three considered FPGAs as more power-wasting circuits are turned on. Specifically, the first part of the voltage drop on AX7103 and ZCU104 FPGA boards is relatively gentle, which may be due to the extra power supply from capacitors. In contrast, the second part of the voltage drop on these two FPGAs have different slopes, caused by different power regulation mechanisms (Fig. 1 (a) and (b)). In contrast, the voltage drop on the ADM-PCIE-7V3 FPGA board is roughly linear, which may be due to the usage of a different power regulation mechanism (see Tab. II). We define the point that separates the voltage drop curve into two pieces with different slopes as a *critical point*, and the point where power supply is cut off as *shutdown point*. The results in Fig. 2 indicate that once the supply voltage drops to the *critical point*, only a small number of power-wasting circuits are needed ($\sim 5\%$ in Fig. 2 (a) and (b)) to shut down the FPGA. Although the three FPGAs in our experiment show different quantitative values during voltage drop, they exhibit the same trend: as highlighted in Fig. 2, it can be concluded that with enough power-wasting circuits turned on, the power supply is cut off, and all these three FPGAs are shut down.

TABLE II: Recommended hardware configurations and experimental results of the used FPGAs.

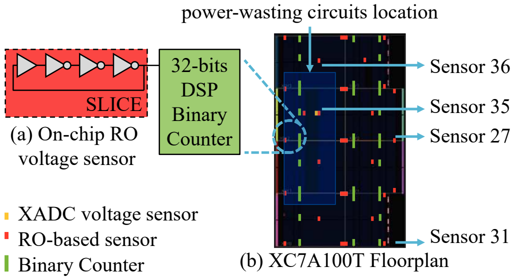
FPGA board	VCCINT(V)	V$_{rec}$(V)	V$_{sd}$(V)	PW$_{oh}$	PDN type
AX7103	1.002	0.95∼1	∼0.91	22%	Fig.1a
ADM-PCIE-7V3	1.025	0.95∼1	∼0.92	13%	See note below
ZCU104	0.844	0.825∼0.85	∼0.83	11%	Fig.1b

Note: this is based on our conjecture as there is no publicly available datasheet/schematic for the power management system of this FPGA board.

Tab. II summarizes our experimental results from the three FPGA chips, where V$_{rec}$ denotes the recommended supply voltage from the device manual, and V$_{sd}$ is the shutdown voltage measured in our experiments. Please note that the ZCU104 development board utilizes a programmable power regulator, and its overload protection voltage is set to be higher than the lowest V$_{rec}$ of the chip. This is why the V$_{sd}$ measured by XADC is within the recommended range.

C. On-chip Voltage Sensor

In modern cloud FPGA server, a parameter called *power metrics* can be used by the tenants to track and measure the FPGA power usage. However, the updating rate of these metrics are very slow, such as every one minute in the AWS service [33]. Thus, to monitor the voltage drop and inject power loads for the proposed attack, an attacker needs to build a faster on-chip voltage sensor. Taking the 7-serial FPGA from Xilinx as an example, the basic logic cell of 7 serial FPGA is a 6-input look-up table (LUT6). There are two slices in one configurable logic block (CLB), and each slice consists of four LUT6s. In our proposed power attack, we build a simple voltage sensor with a four-stage (3 inverters + 1 buffer) RO, as shown in Fig. 4(a). This RO-based sensor can be instantiated in an FPGA Slice. The reason we use ROs as voltage sensors is that the oscillation frequency of an RO is closely related to the voltage [10]. In addition, the time-to-digital converters (TDC) based sensor [8] is also used to measure instantaneous voltage changes in FPGA. This sensor can achieve a high resolution, but still with obvious drawbacks. For example, it employs a long carry-chain and a number of buffers that require manual calibration. Therefore, considering the proposed attack, the RO-based sensor is suitable because of its lightweight and easy deployment. To confirm the accuracy of the RO-based voltage sensors, we place a few sensors in different locations on the FPGAs. For example, 41 sensors are instantiated on the XC7A100T FPGA, as shown in Fig. 4(b). Each of the 8 FPGA clock regions has 5 ROs, 4 of which are placed in corners and the other one in the center. In addition, we place an extra RO-based sensor near the XADC (highlighted in yellow) to study the measurement accuracy of RO-based sensors and serve as the ground truth from XADC. For each RO-based sensor, a 32-bit counter in the nearest DSP block is utilized to record the accumulative oscillation count (C_{RO}).



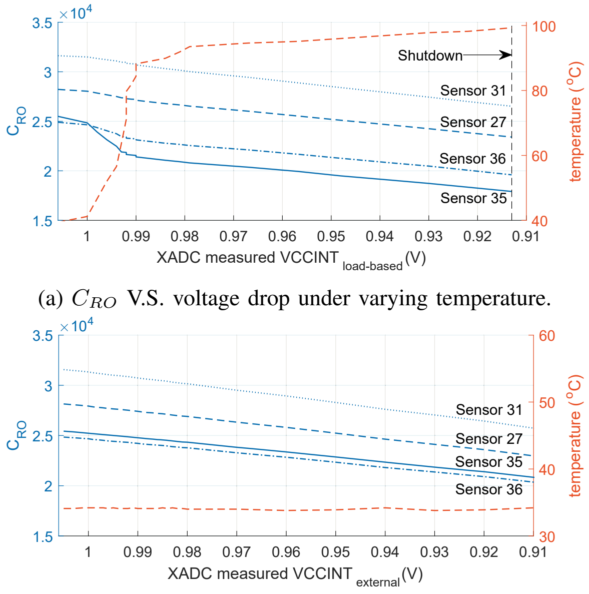
(a) On-chip RO voltage sensor

(b) XC7A100T Floorplan

Fig. 4: (a) Schematic of the on-chip RO-based voltage sensor. (b) Placement of sensors and their corresponding counters used to collect the C_{RO} values.

Following the same setup in Sec. III-B, we activate power-wasting circuits incrementally and record the C_{RO} of these RO-based sensors within $50\mu s$. Also, the voltage drop caused by such power-wasting circuits ($\texttt{VCCINT}_{load\_based}$) is measured by XADC. For brevity, C_{RO} and the corresponding $\texttt{VCCINT}_{load\_based}$ values of four selected sensors (in or out of the range of power-wasting circuits, in the middle or corner of FPGA) are plotted in Fig. 5a. It can be seen that during the first few activation steps of power-wasting circuits, the C_{RO} values of RO-based sensors near or within the circuits (e.g., sensors 35 and 36) have a non-linear relationship with $\texttt{VCCINT}_{load\_based}$. These results are slightly different from [10]. For further analysis, we also plot the temperature measurements by XADC in Fig. 5a. We see that the on-chip temperature is significantly increased during the first few steps of power-wasting circuits activation, which is the main reason for the sharp drop in C_{RO}.

To further confirm the impact of temperature on C_{RO}, we solder off the on-board power regulator for $\texttt{VCCINT}$ and use a bench-top power supply (Keysight E36312A) to provide the supply voltage ($\texttt{VCCINT}_{External}$). The experimental setup is shown in Fig. 6. This setup ensures that the temperature is relatively stable under different voltages since the voltage drop is not caused by turning on the power-wasting circuit, thus the on-chip temperature is relatively stable. We record the C_{RO} values of RO-based sensors under different supply voltages, and the measurement results are shown in Fig. 5b. The results in Fig. 7 demonstrate that while the temperature is stable (34 ± 0.2°C as measured by XADC), there exists a linear relationship between the mean value of 41 sensors (mC_{RO}) and the different external supply voltage ($\texttt{VCCINT}_{External}$). Note that the external power supply can provide sufficient current as needed, thus not incurring power cut off by the regulator even at the *shutdown point*.



(a) C_{RO} V.S. voltage drop under varying temperature.

(b) C_{RO} V.S. voltage drop under stable temperature.

Fig. 5: The relationship between C_{RO}, voltage drop, and temperature from an XC7A100T FPGA. (a) shows the measurement results with temperature variations. (b) depicts the results external power supply, thus the temperature is stable.

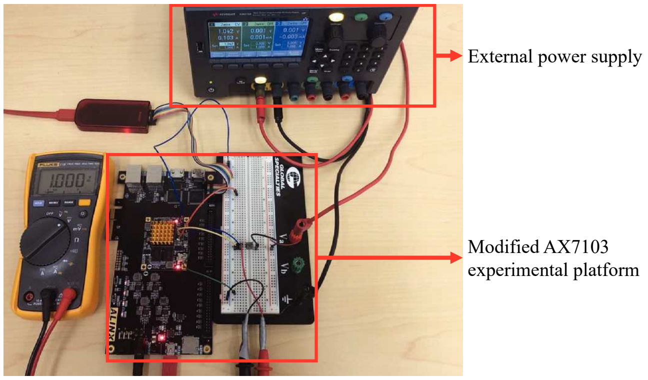


Fig. 6: AX7103 evaluation kit with a Xilinx Artix-7 XC7A100T FPGA. Note that the original power regulator is modified, and a Keysight E36312A DC power supply is utilized to provide the $\texttt{VCCINT}_{External}$.

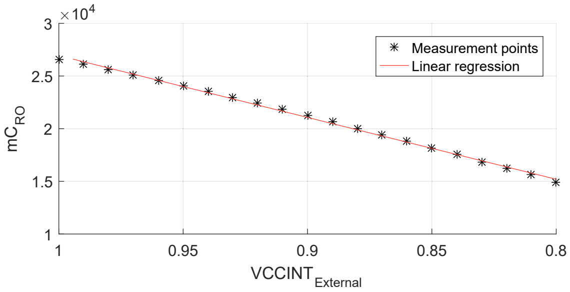


Fig. 7: There exists a linear relationship between the C_{RO} valueni and supply voltage ($\texttt{VCCINT}_{External}$) drop, where mC_{RO} stands for the averaged C_{RO} values of the 41 sensors.

D. Stealthy Power Attacks

As the results shown in Fig. 2, all three FPGA evaluation kits under test encounter shutdown caused by turning on a certain number of power-wasting circuits. Therefore, an attacker can choose to aggressively inject the power load to incur the shutdown. This attack strategy, although straightforward and easy to implement, is only applicable to specific scenarios where the cloud operator has less strict checking rules over the tenants' applications. However, the up-to-date cloud FPGAs can possibly use different ways to track the real-time power usage of each tenant, which can easily detect the straightforward power attacks. For example, the AWS F1 provide a real-time monitoring scheme [33], various systematic FPGA sensor IPs can also be used to track the real-time voltage of each FPGA application.

It should be noted that the shutdown of an FPGA is caused by the overall power consumption, which is not necessarily all from the adversary. Considering the fact that the compute-intensive tasks on a cloud FPGA are non-malicious but power-hungry, an attacker can use the C_{RO} value of an RO-based sensor as a side-channel to monitor the on-chip voltage fluctuations caused by compute-intensive applications, and then only inject extra power loads at correct moments (e.g., *critical point*) for stealthiness.

In stealthy-shutdown, the attacker's objective is to shut down the shared FPGA with *minimal* hardware resources to stay stealthy. The power consumption needed by the attacker to launch well-timed attacks is smaller than that of other benign tenants and thus is stealthy and hard to be detected. The feasibility of this attack strategy is based on the fact that most cloud FPGA tenants run compute-intensive applications, which are power-consuming and making the shared power supply a vulnerable resource to overload.

IV. Experimental Results

In this section, we present experimental results of the proposed stealthy-shutdown attacks considering the existence of other real-world applications in parallel.

A. Validation with Real-world Applications

In order to validate the practical applicability of stealthy-shutdown, we implement a real-world application: bitcoin mining. Bitcoin mining has been widely implemented on various hardware platforms, such as FPGA and GPU, for acceleration purposes. Without loss of generality, we perform a bitcoin mining algorithm on FPGA with open-source code from Github [34]. The implemented bitcoin mining framework on FPGAs mainly consists of two serial SHA265 modules to conduct hashing tasks, which is compute-intensive and thereby power-hungry. As the most power-consuming components, these two SHA256 blocks are not always running, but regularly pausing the hashing operation for golden ticket validation and data/state transmission.

As bitcoin mining algorithms are designed to be highly parallel, they can be flexibly implemented according to the available computing resources [35]. For example, a high-performance mining algorithm mostly utilizes DSP blocks, while a compact version mainly uses LUTs. To fully exploit the practicality of stealthy-shutdown, we implement both bitcoin mining algorithms on the three FPGA evaluation kits. Specifically, considering the application scenarios of these three kits, we implement the compact version with AX7103, and the high-performance version with ADM-PCIE-7V3 and ZCU104 FPGA, which are denoted as "compact" and "full" in Tab. III, respectively. The overhead (BM_{oh}) of each bitcoin mining implementation is also reported in Tab. III.

TABLE III: Experimental results on three FPGA boards.

FPGA Evaluation kit	**Bitcoin Miner**	
	BM_{oh}	**PW_{oh}**
AX7103	71%LUT, 2%DSP, compact[a]	∼5%LUT
ADM-PCIE-7V3	15%LUT, 24%DSP, full[b]	∼8%LUT
ZCU104	22%LUT, 41%DSP, full[b]	∼5%LUT

[a]SHA256 module based on LUT, [b]SHA256 module based on DSP

B. Attack Strategy and Results

1) Hardware setup: To verify the practicality of the proposed attack strategies, we implement the bitcoin mining

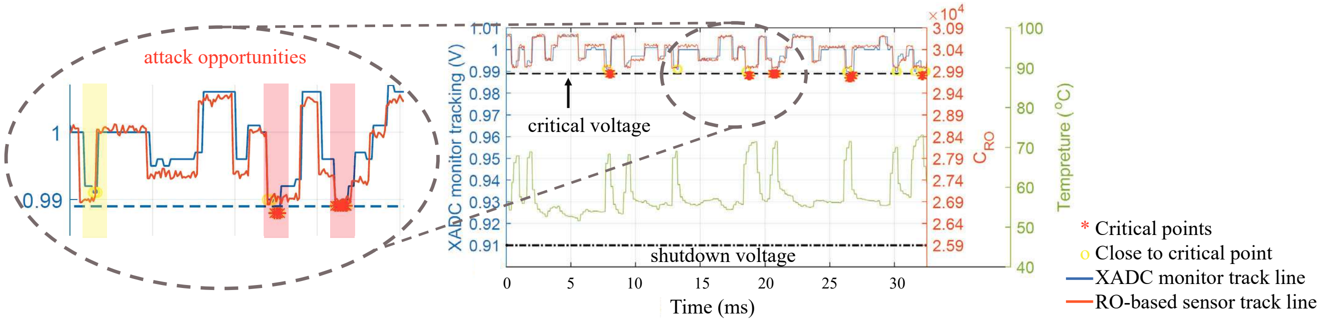


Fig. 8: Using the RO-based sensor to track on-chip voltage fluctuations. The C_{RO} value collected from the RO-based sensor demonstrates a high accuracy in detecting the on-chip voltage drop, thereby leaking side-channel information for stealthy power attacks. Our experiment shows that such attacks can be conducted within 50us time window.

program, power-wasting circuits, and RO-based sensors across these three FPGA evaluation kits. Moreover, to mimic that an attacker only uses partial FPGA hardware, RO-based sensors and power-wasting circuits are implemented within specific areas on these FPGA chips, respectively. In order to gather more side-channel information about the voltage drop, RO-based sensors are placed in different corners of this designated area to amplify their sensitivity to the voltage drop.

2) Attack strategy: With the above-mentioned setup, we first track the voltage drop with RO-based sensors and validate their accuracy. The bitcoin mining program is activated and kept running, and these RO-based sensors are turned on in parallel for calibration purposes, i.e., characterizing the fluctuation range of the C_{RO} values during a certain time period. The C_{RO} values include side-channel information for the sharp voltage drop. In order to obtain accurate side-channel information, during this calibration time period, adversarial power-wasting circuits are disabled. The real-time side-channel information (e.g., C_{RO} values) at a randomly selected RO-based sensor and the ground truth voltage fluctuations[1] measured by the XADC[2] are plotted for comparison in Fig. 8. It can be found that the C_{RO} values captured by the RO-based sensor fit well with the real-time voltage fluctuation. Note that even with the impact from temperature, the RO-based sensor still tracks the real-time voltage fluctuation well. Moreover, the smallest C_{RO} values have a high overlap with the *critical point*, which indicates a high possibility of conducting the stealthy power attacks. After calibrating RO-based sensors and profiling the range of C_{RO} values, the stealthy-shutdown attack is conducted on the FPGA boards with bitcoin mining program running. Specifically, stealthy power attacks focus on exploiting the side-channel leakage following the fluctuation of C_{RO}. For example, when the C_{RO} sees a sharp drop below a certain threshold (which implies a possible critical point voltage drop), the attacker overloads the power supply by turning on a number of power-wasting circuits.

3) Attack results: In a practical cloud FPGA, there is no clue that how many power-wasting circuits would be needed (or sufficient) to launch successful stealthy-shutdown. Therefore, we implement such attacks in a conservative way, i.e., gradually increasing the number of active power-wasting circuits with different attack trails. Another important reason that a practical adversary should follow this way is to make his attacks more stealthy. As observed in Fig. 8, there exists many moments when the voltage fluctuation drops close to the *critical voltage*. The time duration of such attack opportunities is usually of *ms* magnitude, which is sufficiently long for the proposed attack. The overhead of power-wasting circuits (PW$_{oh}$) for successful stealthy-shutdown is reported in Tab. III, which indicates that $\leq$ 8% slices are sufficient to shut down the three FPGA evaluation kits under the test, and the activation of those slices only increase $\sim 10\%$ dynamic power to reach the shutdown point.

C. Discussion

1) Countermeasures: In accordance with [11], we also found that the crash (or DoS) of an FPGA is caused by the on-board power regulation. For example, when the power supply of an FPGA is provided by an external power supplier instead of the power regulator, its operation will not crash even under the observed *shutdown point*, as demonstrated in Fig. 5a and 5b. Therefore, a straightforward countermeasure to mitigate the shutdown attacks is to enhance the power capacity of FPGAs (i.e., with larger power regulators), which, however, may be costly. Another countermeasure is by monitoring the internal current of FPGA chips. When the current on an FPGA chip is higher than a particular threshold, banning new application writing to this chip may mitigate stealthy-shutdown attacks. However, this approach may not be valid when an attacker hides its malicious circuits in an existing application. More severely, an attacker can always start with small attack circuits, and incrementally turn on more power-wasting instances to stay stealthy.

2) Other relevant side-channels: Besides the voltage values leveraged in this work, the XADC also captures the real-time temperature, as shown in Fig. 8. Interestingly, it can

[1] The ground truth voltage fluctuation can not be obtained in practical attacks, but is only depicted here for visualization purpose.

[2] Note that the time period (0 to 32 ms) in Fig. 8 is a randomly sampled period from the FPGA operation.

be found that most voltage fluctuations are also associated with temperature variations, i.e., a higher temperature value usually stands for intensive FPGA workload, yet a voltage drop. From this perspective, the temperature information can also be used as a potential side-channel for attacks. However, it is practically more challenging to use a temperature-based side-channel to guide such stealthy power attacks. This is because: 1) A normal cloud FPGA user may not have the access to read the on-chip XADC. Moreover, considering the attacks and vulnerabilities revealed in this work, these authority associated with reading these systematic sensors should be strictly prohibited. 2) Compared to the on-chip temperature variations, C_{RO} values have a higher overlap with the *critical point*, which indicates a higher possibility of conducting successful stealthy power attacks.

V. Conclusion

This paper presents a novel attack method named stealthy-shutdown that targets the power management system of cloud FPGAs. If succeed, the proposed attack can erase all running applications on the target FPGA chip, and all peripheral components like the DDR memories are also powered off. This attack incurs a disastrous consequence for cloud FPGAs because a sever can only recover its FPGA service by rebooting. The key idea of our proposed attack is to seize the moments when the on-chip supply voltage drops to the critical point, due to the compute-intensive workload executing on cloud FPGAs. The experimental results with commercial FPGA evaluation kits and real-world applications confirm the feasibility and danger of the proposed power attack. Since the implementation of well-timed stealthy power attacks only needs a small portion of the overall power consumption, it is challenging to identify and mitigate such attacks actively. In a practical cloud FPGA setup, several FPGAs are deployed together in a server [36]. Therefore, the overall power consumption and regulation of such FPGA cluster becomes a more challenging job. Correspondingly, it would be easier to implement the proposed stealthy power attack without being detected. In the future works, we will validate the proposed attack on real cloud FPGAs. Specifically, we will explore non-loop circuit-based sensors and attack circuits, to bypass the DRC checking of current cloud FPGA development tools.

References

[1] Y. Chen, J. He, X. Zhang, C. Hao, and D. Chen, “Cloud-dnn: An open framework for mapping dnn models to cloud fpgas,” in *Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, 2019, pp. 73–82.

[2] *Introduction to FPGA Design with Vivado High-Level Synthesis (UG998)*, Xilinx,Inc, 2019.

[3] *SDx Command and Utility Reference Guide (UG1279)*, Xilinx,Inc, 2019.

[4] Enable faster fpga accelerator development and deployment in the cloud. [Online]. Available: https://aws.amazon.com/ec2/instance-types/f1/

[5] Inside the microsoft fpga-based configurable cloud. [Online]. Available: https://azure.microsoft.com/en-us/resources/videos/build-2017-inside-the-microsoft-fpga-based-configurable-cloud/

[6] C. Ramesh and S. B. Patil, “Fpga side channel attacks without physical access,” in *FCCM*. IEEE, 2018, pp. 45–52.

[7] I. Giechaskiel, K. B. Rasmussen, and J. Szefer, “Measuring long wire leakage with ring oscillators in cloud fpgas,” 2019.

[8] J. Krautter, D. R. Gnad, and M. B. Tahoori, “Fpgahammer: remote voltage fault attacks on shared fpgas, suitable for dfa on aes,” *IACR TCHES*, pp. 44–68, 2018.

[9] D. Mahmoud and M. Stojilović, “Timing violation induced faults in multi-tenant fpgas,” in *DATE*. IEEE, 2019, pp. 1745–1750.

[10] G. Provelengios, D. Holcomb, and R. Tessier, “Characterizing power distribution attacks in multi-user fpga environments,” in *FPL*. IEEE, 2019.

[11] D. R. Gnad, F. Oboril, and M. B. Tahoori, “Voltage drop-based fault attacks on fpgas using valid bitstreams,” in *FPL*. IEEE, 2017, pp. 1–7.

[12] I. Giechaskiel, K. B. Rasmussen, and K. Eguro, “Leaky wires: Information leakage and covert communication between fpga long wires,” in *ASIACCS*. ACM, 2018, pp. 15–27.

[13] S. Yazdanshenas and V. Betz, “The costs of confidentiality in virtualized fpgas,” *VLSI*, 2019.

[14] A. Khawaja, J. Landgraf, R. Prakash, M. Wei, E. Schkufza, and C. J. Rossbach, “Sharing, protection, and compatibility for reconfigurable fabric with amorphos,” in *OSDI*, 2018, pp. 107–127.

[15] A. Vittal and M. Marek-Sadowska, “Crosstalk reduction for vlsi,” *IEEE TCAD*, vol. 16, no. 3, pp. 290–298, 1997.

[16] A. Vittal, L. H. Chen, M. Marek-Sadowska, K.-P. Wang, and S. Yang, “Crosstalk in vlsi interconnections,” *TCAD*, vol. 18, no. 12, pp. 1817–1824, 1999.

[17] P. Xu and Z. Pan, “The analytical model for crosstalk noise of current-mode signaling in coupled rlc interconnects of vlsi circuits,” *Journal of Semiconductors*, vol. 38, no. 9, p. 095003, 2017.

[18] S. J. Wilton, “A crosstalk-aware timing-driven router for fpgas,” in *FPGA*. ACM, 2001, pp. 21–28.

[19] N. Das, P. Roy, and H. Rahaman, “Detection of crosstalk faults in field programmable gate arrays (fpga),” *Journal of The Institution of Engineers (India): Series B*, vol. 96, no. 3, pp. 227–236, 2015.

[20] G. Provelengios, C. Ramesh, S. B. Patil, K. Eguro, R. Tessier, and D. Holcomb, “Characterization of long wire data leakage in deep submicron fpgas,” in *Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2019, pp. 292–297.

[21] Y. Luo and X. Xu, “Hill: A hardware isolation framework against information leakage on multi-tenant fpga long-wires,” in *2019 International Conference on Field-Programmable Technology (ICFPT)*. IEEE, 2019, pp. 331–334.

[22] M. Zhao and G. E. Suh, “Fpga-based remote power side-channel attacks,” in *SP*. IEEE, 2018, pp. 229–244.

[23] F. Schellenberg, D. R. Gnad, A. Moradi, and M. B. Tahoori, “An inside job: Remote power analysis attacks on fpgas,” in *DATE*. IEEE, 2018, pp. 1111–1116.

[24] Y. Luo and X. Xu, “A dynamic frequency scaling framework against reliability and security issues in multi-tenant fpga,” in *2020 IEEE 28th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2020, pp. 210–210.

[25] *Artix-7 FPGAs Data Sheet: DC and AC Switching Characteristics (DS181)*, Xilinx,Inc, 2018.

[26] *Virtex-7 T and XT FPGAs Data Sheet: DC and AC Switching Characteristics (DS183)*, Xilinx,Inc, 2019.

[27] *Zynq UltraScale+ MPSoC Data Sheet: DC and AC Switching Characteristics (DS925)*, Xilinx,Inc, 2019.

[28] *TPS54620 4.5-V to 17-V Input, 6-A, Synchronous, Step-Down SWIFT™ Converter*, TI,Inc, 2017.

[29] *ZCU104 Evaluation Board User Guide (UG1267)*, Xilinx,Inc, 2018.

[30] *ALINX AX7103 Manual UG.1.0.pdf*, ALINX,Inc. [Online]. Available: https://github.com/alinxalinx/AX7103

[31] *ADM-PCIE-7V3 User Manual*, Alpha Data,Inc, 2016.

[32] *7 Series FPGAs and Zynq-7000 SoC XADC Dual 12-Bit 1 MSPS Analog-to-Digital Converter (UG480)*, Xilinx,Inc, 2018.

[33] Afi power. [Online]. Available: https://github.com/aws/aws-fpga/blob/master/hdk/docs/afi_power.md

[34] Open-source fpga bitcoin miner. [Online]. Available: https://github.com/progranism/Open-Source-FPGA-Bitcoin-Miner

[35] P. Xu, P. Taylor, B. Nappier, and M. Candido, “Half-fast bitcoin miner: Open-source bitcoin mining with fpga.”

[36] Amazon ec2 f1 instances. [Online]. Available: https://aws.amazon.com/ec2/instance-types/f1/