SEGIVE: A Practical Framework of Secure GPU Execution in Virtualization Environment

Ziyang Wang*†‡, ✉Fangyu Zheng*‡, Jingqiang Lin§*†, Guang Fan*†‡, Jiankuo Dong¶
*State Key Laboratory of Information Security, Institute of Information Engineering, CAS, Beijing, China
†School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
‡Data Assurance and Communication Security Research Center, CAS, Beijing, China
§School of Cyber Security, University of Science and Technology of China, Hefei
¶School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, China

{wangziyang, zhengfangyu, linjingqiang, fanguang}@iie.ac.cn, djiankuo@gmail.com

***Abstract*—With the advancement of processor technology, general-purpose GPUs have become popular parallel computing accelerators in the cloud. However, designed for graphics rendering and high-performance computing, GPUs are born without sound security mechanisms. Consequently, the GPU-based service in the cloud is vulnerable to attacks from the potentially compromised guest OS as large amounts of sensitive code and data are offloaded directly to the unprotected GPUs.**

In this paper, we propose SEGIVE, a practical framework of secure GPU execution in the virtualization environment, which protects offloaded device code and data from disclosure or tampering by malicious guest OSes through the full life cycle of security-critical GPU applications. First, SEGIVE secures all the traffic transferred to GPUs with Intel SGX technology, including the users' sensitive data and GPU binaries. Second, with various memory isolation mechanisms, SEGIVE enhances security in multi-user execution scenarios by sharing a GPU among multiple workloads, which avoids underutilization of device resources. Besides, SEGIVE requires no modifications to application source codes, the GPU architecture, or I/O interconnection to fulfill security principles, and thus almost all prevailing GPU-based applications can easily benefit from SEGIVE with little porting effort. We have implemented SEGIVE with KVM-QEMU on off-the-shelf NVIDIA GPUs and CPUs. Evaluation results show that with security-enhances, the performance of SEGIVE prototype is still competitive to the native execution on compute-intensive applications, especially for the public-key cryptography algorithm.

***Index Terms*—GPU Virtualization, GPU Security, Cryptographic Computation, SGX**

I. Introduction

While benefiting from virtualization technologies, cloud computing is widely used in various industries. The increasing scale of data collected or processed in cloud servers has fueled the High-Performance Computing (HPC) in cloud computing. Meanwhile, Graphics Processing Units (GPUs) have been distinguished accelerators as graphics rendering engines and parallel programmable processors over the last decade. HPC especially enjoys a performance boost due to hundreds to thousands of streaming processing cores of GPU. With the rapid growth of cloud computing, GPU-accelerated computing instances are available on-demand on all major cloud platforms (e.g., AWS EC2 [1], Google Cloud [2], Microsoft Azure [3]).

One of the promising GPU-based applications is cryptography acceleration, which requires compute-intensive workloads. Many previous works [4]–[7] show that GPUs are ideal candidates for cryptography acceleration, especially in cloud computing scenarios. Currently, most major cloud providers offer a Cloud Hardware Security Modules (HSM) service [8]–[11]. CloudHSM service provides renters with dynamical and scalable virtual HSMs to support data encryption or digital signature for sensitive cloud applications such as e-commerce, digital payment, etc. Taking AWS CloudHSM for example, it uses NITROXIII HSM [12], [13] as the underlying physical HSMs to support partitioned "virtual" HSMs. However, one single physical HSM can only provide 11 kilo ECC (Elliptic Curve Cryptography) operations per second. Meanwhile, Pan et al. [4] reported a GPU-based HSM named `Guess` that can deliver near 10 million and 1 million ECC operations per second for signing and verifying, respectively. In other words, a single GPU-based HSM is comparable to over 100 traditional HSMs, which can greatly improve the CloudHSM service in terms of both performance and power consumption. However, concerns about the security of GPU and inconvenience of GPU virtualization limit its further deployment in cloud applications.

Designed for graphics rendering and high-performance computing, GPU itself is born without sound security mechanisms. Security breaches of GPU computation may leak sensitive data if care is not taken [14]–[17]. Things get even worse in the public cloud, as external attackers may attempt to access the sensitive data or inject malicious codes by exploiting vulnerabilities of the software or OS in the user-rented VM. In major industrial cloud platforms, applications can obtain pass-through access to physical GPU devices through hardware-supported I/O virtualization provided by the motherboard chipset or GPU manufactures [18]. On one hand, a GPU can only be occupied exclusively by a designated guest virtual machine, consequently, all benefits of virtualization are

This work was partially supported by and National Key R&D Program of China under Award No. 2017YFB0802100 and 2018YFB0804401, National Natural Science Foundation of China under Award No. 61772518 and 61902392. *(Corresponding author: Fangyu Zheng, E-mail: zhengfangyu@iie.ac.cn.)*