

HyperMI: A Privilege-level VM Protection Approach against Compromised Hypervisor

Wenqing Liu, Kunli Lin, Song Wei, Kun Zhang, Bibo Tu
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
{liuwenqing, linkunli, zhangkun, tubibo}@iie.ac.cn

Abstract—Once compromising the hypervisor, remote or local adversaries can easily access other customers’ sensitive data in the memory and context of guest virtual machines (VMs). Therefore, it is essential to protect VMs, especially the context and the memory of VMs from the compromised hypervisor. However, previous efforts employ extra customized hardware which is not convenient for cloud providers to adopt widely. Or they employ new architecture relying on a higher privilege level than hypervisor.

This paper proposes HyperMI, a novel approach to provide runtime protection for VMs based on a privilege-level secure execution environment against compromised hypervisor. We use HyperMI world as a secure and isolated trusted execution environment against compromised hypervisor. It is designed to be placed at the same privilege-level with the hypervisor. Firstly, we propose event-driven VM monitoring which intercepts interaction between all VMs and hypervisor and redirects interaction process to HyperMI for security check. Secondly, we propose effective VM memory isolation to provide runtime protection for VMs by isolating memory among VMs and hypervisor securely. The key of HyperMI is that it decouples the functions of interaction between VMs and hypervisor, and the functions of address mapping of VMs from the compromised hypervisor. As a result, HyperMI isolates memory completely, controls memory mapping when a page is allocated to a VM and resists malicious memory of VMs accessing from the compromised hypervisor. We have implemented a prototype for KVM hypervisor on x86 platform with multiple Linux as guest OSes, which can be used in commercial cloud computing industry with portability and compatibility. The security analysis shows that this approach can protect VMs with effective isolation and event-driven monitoring, and the performance evaluation confirms the efficiency of HyperMI.

Index Terms—Virtualization, VM Protection, VM Security

I. INTRODUCTION

As more and more functionalities are added into the hypervisor, the code bases of commodity hypervisors (KVM or Xen) have been increased to be large lines. However, a recent survey shows that commodity hypervisor incurs more vulnerabilities because of the larger code bases. On the one hand, from 2004 to now, there are 130 vulnerabilities about KVM. Some of them (e.g., CVE-2018-1087) shows high-risk vulnerabilities that can lead to privilege raising behavior and comprehensive compromised hypervisor. On the other hand, because hypervisor possesses the highest privilege in the cloud environment, an attacker who compromises hypervisor could harm the whole cloud infrastructure and endanger data and computation in the cloud. For example, an attacker can deploy

a complete malicious guest VM on the virtualized platform, conducts attacks to the hypervisor and further attack other VMs even the entire platform through illegal data accessing and so on. Some attacks also directly compromise hypervisor. In order to settle down all these threats, some try to detect malicious actions among frequent cloud management operations, but this kind of approach is much similar to that of looking for a needle in a haystack. Therefore, HyperMI could provide a superior solution from another perspective. Current researches provide services of protection for critical data.

Customized Hardware Some efforts (SecureME [5], Bastion [4] and Iso-x [8]) rely on customized underlying hardware to provide fine-grained protection for VM or in-VM process. Iso-X provides isolation for security-critical pieces of an application by introducing additional hardware and changes to OS. It controls memory access by introducing ISA instructions.

Reconstructed Hypervisor Some efforts (NoHype [11] and TrustOSV [18]) pre-allocate fixed cores or memory resource to isolate VMs via reconstructing hypervisor. In the meantime, deprive some virtualization capabilities and introduce lots of modification to the hypervisor. Moreover, NoHype removes the virtualization layer while retaining the key features enabled by virtualization. TrustOSV protects compromised cloud environment by removing interaction between exposed executing environment and hypervisor.

Software at a Higher Privilege Level In order to mitigate the hazard caused by the hypervisor possess, plenty of software solutions propose and introduce a higher privilege-level than the original hypervisor. Nested virtualization is one of the representative approaches, which provides a higher-privileged and isolated execution environment to run the monitor securely. The turtles project [3] and CloudVisor [20] are examples of systems that propose nested virtualization idea to achieve isolation for protected resources. Especially, CloudVisor uses nested virtualization to decouple resource management into the nested hypervisor to protect VMs.

In practice, independence on platforms is among the most prized features for cloud providers. Furthermore, business architectures that are already widely deployed, it is equally important to minimize changes to existing systems or architectures. For this purpose, some recent efforts only focus on software approaches about how to achieve the same privilege-

level isolation and protection without relying on a higher privilege level. For example, SKEE [1] introduces a secure execution environment at the same privilege level with kernel.

Inspired by the idea of "same privilege level" isolation, we propose HyperMI, a "same privilege level" and software-based VM protection approach for guest VMs against the compromised hypervisor. HyperMI introduces a secure isolation execution environment, named HyperMI world, to place event-driven VM monitoring and VM isolation module. The VM isolation module guarantees memory isolation between VMs. The details on them are as follows with taking the x86 platform as an example.

VM Monitoring Event-driven VM monitoring protects interaction between hypervisor and VMs. Thus, the compromised hypervisor could not subvert the domain of the state of VMs. Firstly, there are some especially critical structures data that records state information of VMs must be monitored and protected. These data includes Extended Page Tables (EPT), EPT Pointer (EPTP) and Virtual Machine Control Structure (VMCS). EPT contains address mapping relationship between the Guest Physical Address (GPA) and Host Physical Address (HPA), the compromised hypervisor could modify the content of any EPT entry, thus maps GPA of VM to an illegal HPA. EPTP is a CR3-like register used to supply CPU with the address of EPT, thus the adversary could subvert a VM by supplying CPU with a carefully constructed and illegal EPT. VMCS is used in VMX operation to manage the behavior of VMs and the hypervisor. Thus, the illegal content of VMCS, especially the Guest-state area and VM-execution control fields, subverted by the compromised hypervisor poses an uncontrollable impact on the security of the virtual machines. Based on the importance of the data structures mentioned above, HyperMI must monitor and protect them. Secondly, HyperMI adopts an event-driven monitor approach which hooks events that closely related to VM exit and VM entry. Once these events occur, monitoring is executed.

VM Isolation VM isolation can resist malicious VM memory accessing from compromised hypervisor, especially, remapping and double mapping attack. Firstly, HyperMI marks each page with page marking technique to guarantee each page can only be owned by one VM or hypervisor. Secondly, it deprives the address translation function of the hypervisor to ensure that the page is marked with the owner when the page is mapped. Finally, in order to avoid double mapping attack, the owner of the page is verified to ensure whether access or not when EPT updates resulting from page fault. In order to resist remapping attack, clear the content of the page when the physical page is released.

Our prototype introduces 4K SLOC (Source Lines of Code) to VM monitoring and 300 SLOC modifications to the hypervisor, VM monitoring reduces the attack surface of the hypervisor, and the memory isolation among hypervisor and VMs is significantly guaranteed. The experimental results show trivial performance overhead for runtime VM protection.

Our contributions are as follows:

- A non-bypassable hypervisor monitoring for VMCS and

EPT approach which can ensure the security of interaction between hypervisor and VMs.

- An approach of isolating memory among VMs and hypervisor securely for VMs by using page marking technique to avoid malicious access from the compromised hypervisor.
- A prototype based on KVM and x86 architecture with trivial performance overhead, high security, platform independence and fine applicability for cloud providers.

The rest of this paper is organized as follows. Section II discusses our threat model and assumption. Section III will elaborate on the design of HyperMI, while section IV presents how to implement HyperMI on x86 platform. Section V gives the evaluation of security and performance. Section VI compares HyperMI to past work. At last, Section VII is the summary and conclusion.

II. THREAT MODEL AND ASSUMPTION

In this section, we describe threat model and assumption.

A. Threat Model

A victim VM can attack a VM after the hypervisor is compromised through VM escape attack, cross-domain attack and so on. On the one hand, the attacker modifies the critical interaction data in the context switching process between VMs and hypervisor, including general purpose register information, privileged registers, et al. On the other hand, the attacker modifies the address mapping of the EPT of VM, causing remapping attacks and double mapping attacks. Attackers carry out attacks in the following ways.

a) Modifying the Interaction Data: For the modification of the critical interaction data of the context switching, the attacker can obtain the address and access privilege of the VMCS structure in the victim VM, modify the information of the VMCS, such as HOST_RIP, GUEST_CR0, EPTP, et al. For example, modifying the value of HOST_RIP register and writing malicious program address to the register will cause control flow hijacking attack. Modifying the value of privilege register, CR0, will close the DEP mechanism, and modifying the value of CR4 register will close the SMEP mechanism.

b) Modifying the Address Mapping of EPT: Modification to EPT can result in malicious memory information leakage. There are two used scenes, double mapping, and remapping attack.

Scene 1. For the double mapping attack, the attacker first controls and compromises a VM, then obtains the privilege of the hypervisor through the virtual machine escape attack, and maliciously accesses the VMCS structure to obtain the value of EPTP. The attack process is as shown in Figure 1. In this way, the EPT address of the attacker virtual machine, VM1, and the victim virtual machine, VM2, are respectively obtained. Also, for a guest virtual address in VM2, named 'A', the corresponding real physical address is 'B'. For VM1, the real physical address corresponding to the guest virtual address 'C' is 'D', then 'D' is modified to be 'B' by modifying the value of the last page item of EPT. Then VM1 can access

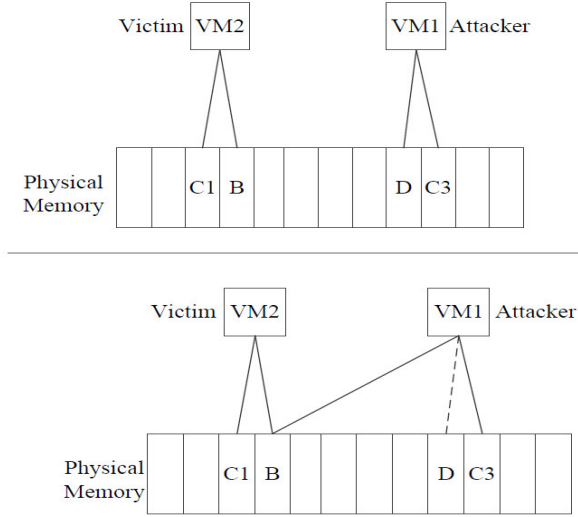


Fig. 1. The execution process of double mapping.

the data of VM2 successfully, this process is called double mapping.

Scene 2. For the remapping attack, there are VM1 (attacker) and VM2 (victim). A physical page (named 'A') used by VM2 is released after being used. After 'A' is released, VM1 remaps to 'A'. So that the guest virtual address of VM1 points to the physical page 'A'. By this way, VM1 can access the information on 'A' used by VM2, causing information leakage.

Through the analysis of these two kinds of attack models, it is necessary to achieve attack prevention through VM monitoring and VM isolation.

B. Assumption

We propose some assumptions. First, we assume hardware resources are trusted including processor, buses, BIOS, UEFI and so on, the trusted boot based on hardware can ensure the security and integrity of bootloaders. The TCB contains created HyperMI and hardware resources. Second, this paper does not consider denial of service attack (DOS), side channel attack and hardware-based attack, such as cold-boot attacks and RowHammer.

III. DESIGN

In this section, we give design of HyperMI.

A. Architecture

HyperMI is designed to provide a secure isolated execution environment to protect VMs against compromised hypervisor without depending on a higher privilege level software than hypervisor or extra customized hardware.

Figure 2 depicts the architecture of HyperMI. The first component is the modified hypervisor, some functions such as EPT operations, context management, are stripped from the original hypervisor. These functions that are closely related to protected data structures are hooked, once these functions are called, the execution flow will be transferred to HyperMI world

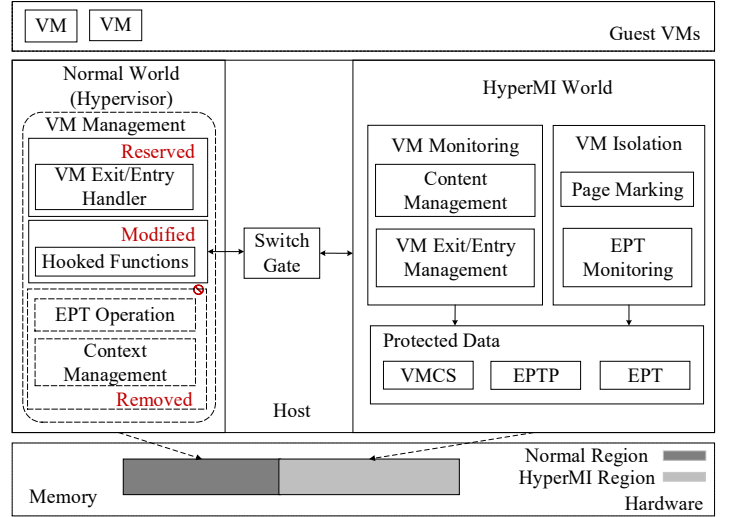


Fig. 2. The architecture of HyperMI.

through Switch Gate. The second component is the HyperMI World, which is a secure and isolated execution environment based on the same privilege level with the hypervisor. VM Monitoring and VM Isolation, rely on this HyperMI world to ensure their security and isolation. The last component is the Switch Gate. This component is an atomic operation to ensure the secure switch between the normal world and the HyperMI world.

While the hypervisor together with guest VMs runs in the normal world, the hypervisor is forced to request HyperMI to perform four operations on its behalf: 1) switching context between the hypervisor and VMs, 2) updating EPT of VMs, 3) verifying the pages when executing swapping operations to resist double mapping attack, 4) verifying the pages when executing releasing operations to resist remapping attack. After setting up the HyperMI, the whole system is ready to create an isolated executing environment. With these designs, HyperMI enforces the isolation and protection of memory used by each VM. Furthermore, HyperMI guarantees the security of interaction, memory isolation between the hypervisor and VMs.

B. HyperMI World

HyperMI world, based on another kernel page table, is created to be another address space at the same privilege-level with hypervisor. Switch gate is created to provide the only communication channel for two worlds.

The creation of HyperMI world has two purposes: 1) Create a space which can provide protection for key data and memory of VMs. 2) Create a software system that does not depend on customized hardware devices and adapts to multi-system platforms. The key point of its design is that it creates another address space at the same-privilege-level with hypervisor. Unlike other same-privilege-level software, the creation of HyperMI World depends on another new page table.

Creating HyperMI World We use two isolated address space based on two sets of page table to achieve isolation of HyperMI world. Figure 3 describes the address space layout of two worlds through two sets of page table, the normal page table and HyperMI page table. On the left of Figure 3, the normal page table contains code and data of the normal world except for that of HyperMI world in case of compromised hypervisor breaking the integrity of HyperMI world. So that program running in normal world cannot access data in HyperMI world. On the right of Figure 3, all address are mapped in HyperMI page table. HyperMI code remains executable and HyperMI data remains writable. What's the most important, kernel code is forbid to execute with reference to protection for running HyperMI world, so that non-executable kernel code cannot attack HyperMI world.

Creating Switch Gate In the middle of Figure 3, the switch gate includes entry/exit gate and shared buffer. Entry gate provides the only entrance to HyperMI world while the exit gate provides the address for returning to the normal world. The shared buffer contains common data and code which the system needs to run the switch process. Common code is switch code, common data is entrance address to the HyperMI world and return address to the normal world. The switch gate is mapped at the same place in the normal world and HyperMI world because the page table loading code must be called by the two worlds before and after switching. Of course, the entrance address must be protected after switching to HyperMI world in case of a malicious attacker accessing HyperMI world causally after trusted boot. This is introduced in section III-C.

C. Security Guarantee for HyperMI World

Nevertheless, without any protection measures, the page table to load to switch to HyperMI world is not secure for four reasons: 1) Attackers can control page table with the highest privilege after hypervisor is compromised. 2) Attackers can bypass the switch gate to break the security of HyperMI world. 3) Attackers with the highest privilege can free to execute privileged instructions to access the value of privilege registers, such as CR0, CR3 and so on. 4) Attackers can carry out DMA attack to access HyperMI world casually. We detail the protection measures for these four types of attacks below.

Protecting Page Table Hypervisor has full control of page tables, so it can attack the HyperMI world which is based on page table. There are three reasons for controlling the two set of kernel page tables: 1) To access casually or bypass HyperMI world, the attacker can tamper the content of normal page table to map address of HyperMI world or load malicious page table to CR3. 2) The attacker can cover the hooked functions which we use in VM monitoring, redirect the functions to their own malicious code and bypass secure monitoring of HyperMI. 3) To break HyperMI world, malicious kernel code with execution permission can be executed to subvert HyperMI world by means of the vulnerabilities of the original kernel code. Therefore, three secure approaches against these attacks are as follow.

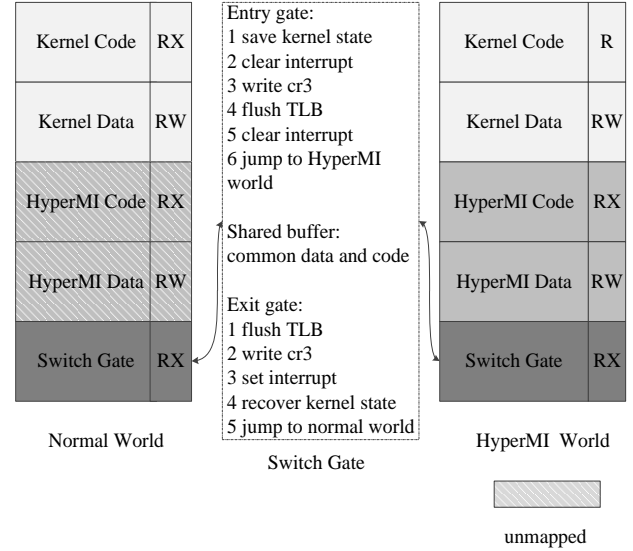


Fig. 3. An overview of address space layout.

For the first attack, HyperMI code and data is unmapped in page table of normal world. And to protect the entrance address to HyperMI world from being leaked, we remove all entries that map to HyperMI world from the page table in normal world. Deprive the ability to access CR3 of the kernel in order to avoid loading illegal page table, and resist bypassing HyperMI world. For the second attack, we intercept the accessing operation to CR0 and maintain the WP bit as 1. We stick to $W \oplus X$ and maintain the code segment of hooked functions unwritable. To go against the third attack, we set the kernel code segment as NX (non-executable) when HyperMI world is running. For more security, we modify the kernel to configure these two sets of page table as read-only by setting the memory regions of the page tables unwritable. This is necessary to prevent the page tables from being modified by attackers. Any access permission modification to two sets of page table must cause the kernel to page fault, then we dispatch page fault to HyperMI world to verify the correctness of address mapping. This idea is adopted in SKEE [1].

Worlds Switching Securely HyperMI creates a switch gate between the normal world and HyperMI world by loading a page table of the next space into CR3. In order to avoid ensure switch security, we design the switch process as follow.

The switching process described in Figure 3 is as follow: 1) Save the kernel state to the stack including general registers and interrupt enable/disable status. 2) Clear the interrupt with the CLI instruction. 3) Load the page table to the register CR3. 4) Interrupt again. 5) Jump to the HyperMI region. For the exit process, return to the normal world by performing the operations in the reverse order.

Accessing Privilege Registers Securely The hypervisor is still privileged and it can free to execute privilege instructions, so that it can write any value to the related privileged registers.

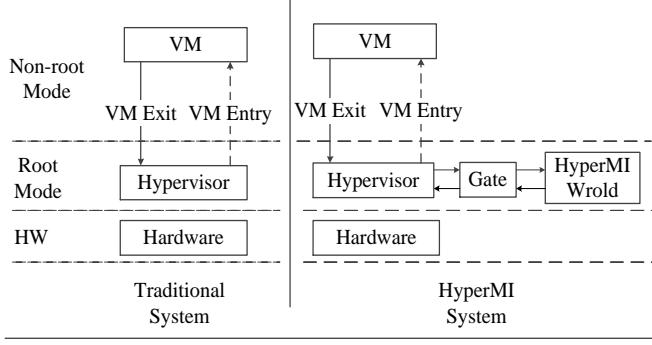


Fig. 4. Interaction comparison.

1) Malicious attackers can close DEP mechanism by writing to CR0, close SMEP mechanism by writing to CR4. 2) Kernel code can load a crafted page table to bypass the HyperMI world by converting a meticulously constructed address of one page table to CR3. To protect the system, HyperMI deprives sensitive privileges instructions executed by the hypervisor, and dispatches the captured events to the HyperMI world. The HyperMI world can choose how to handle this event, such as issuing alerts, terminating the process, or doing nothing. The whole process is similar to the signal handling in traditional OSes.

Resisting DMA Attack DMA operation is used by hardware devices to access physical address directly. Malicious attackers can read or write arbitrary memory regions including HyperMI world by DMA. Therefore, it is a crucial focus of intercepting direct access to physical pages belonging to HyperMI world by DMA operation. Fortunately, HyperMI employs IOMMU mechanism to avoid DMA attack, which can carry out access control for DMA access. Our approach adopts two policies: 1) Remove the corresponding mapping of the critical data from the page table which IOMMU uses. These critical unmapped data includes the entrance address of HyperMI, data recording Page-Mark structure used in VM isolation, VM-Mark structure used in VM monitoring and so on. 2) Intercept the address mapping functions about I/O, verify whether the address is an address space of HyperMI world, then choose to map or unmap.

Through the above security measures, HyperMI can be protected from being bypassed and being breaking, thus providing a secure execution environment for VM monitoring and VM isolation.

D. VM Monitoring

During the VM entry/exit process, the related state information of every virtual CPU and host OS are stored in VMCS structure. VM exit/entry are the points to intercept accessing

operation of VMCS. HyperMI intercepts and validates interaction between hypervisor and VMs. Hypervisor needs to handle VM exit events according to different exit reasons. We describe two models: 1) Interception of context switching. 2) VM exit redirection.

Interception of Context Switching The protection to VMCS structure cannot be ignored. During the VM entry or VM exit, the compromised hypervisor can tamper VMCS structure and execute follow attacks: 1) Access memory region of VMCS directly. 2) Falsify the value of HOST_RIP, and the system will suffer control flow hijack. 3) Tamper the value of EPT_POINTER(EPTP), and another malicious EPT is loaded. 4) Fake the value of HOST_CR3, so the page table of host OS can be replaced.

We hide VMCS in HyperMI world to avoid access from hypervisor. In order to ensure that some functions (vmcs_writel, vmcs_readl et al.) can access VMCS properly, HyperMI hooks these functions into HyperMI world. So hypervisor requests HyperMI world to handle operations about VMCS and return the corresponding result for the legal request.

VM Exit Redirection Since VMCS is hidden in HyperMI World, all context management (accessing VMCS operations) must be trapped to hypervisor. During VM Exit, hypervisor needs to access VM Exit reason data of VMCS, and then deal with the exit event. So in case of control flow makes mistakes because hypervisor can not access VMCS, VM Exit redirection is designed. Control flow jumps to HyperMI World to access VM exit reason data of VMCS structure, switches to hypervisor and execute VM exit event handler function. The control flow is shown as Figure 3.

E. VM Isolation

As known to all, when Intel VT-x is enabled, all the physical memory is managed by the hypervisor using Extended Page Tables (EPTs), EPTs must be involved in each memory access and determine the permissions of the accessed memory page frame. Besides, the hypervisor employs different EPTs to manage the corresponding physical memory area for different VMs. If the page is remapped or double mapped, the system causes a page fault. A VM accesses physical memory page frame through EPT updating. Page is accessed by VMs through EPT updating. It is essential to control page mapping when EPT updates. To achieve the goal of VM isolation completely, we adopt page marking technique, that marks memory page with flag and tracks the page. However, a time point is needed for marking all page. And it is no doubt that EPT updating is an excellent point because hypervisor can manage all page mapping efficiently when EPT updates. Besides, EPT must be protected in case of data being leaked because EPT plays a vital role in address translation of VMs. So we describe VM isolation in these two aspects, EPT interception and memory isolation for VMs.

Confronting with these attacks, we can see that the ultimate purpose of attackers is to access the physical memory page frames of other VMs. HyperMI solves the issue using the following policies: 1) Hide the address of EPT and make it

unmapped for the kernel in the normal world. 2) Intercept the related operations including EPT creating, loading, updating, walking and destroying to avoid leaking the address of EPT. 3) Mark respective EPT for every VM, isolate EPT among VMs, and ensure that the right and corresponding EPT is loaded for every VM.

TABLE I
VM-MARK TABLE.

VM-Mark Table			
Label	VMID	EPTID	EPT_Address
Description	The VM Identifier	The EPT Identifier	The Entry Address of EPT

For the first policy, three places should be particularly protected. The first place is the EPT creation function where the function would return a CR3-like address: EPTP value. The second place is VMCS where would record the value of corresponding EPTP. The last place is the hardware register EPTP, the work of this register is similar to that of the CR3 register, which is used to provide MMU with the location of EPT.

Interception of EPT Operations During the translation process of mapping GPA to HPA through EPT, a VM can suffer these attacks: 1) Compromised hypervisor can access EPT of every VM, then does whatever it wants by modifying the address mapping causing remapping and double mapping attack. 2) Load the malicious EPT and execute illegally.

Therefore, it is critical to intercept the EPT creating operation and protect the address of EPT in HyperMI world. Notably, it is vital to store VMCS structure which contains the value of EPTP in HyperMI world.

However, some functions, in addition to the EPT creation function, still rely on the value of EPTP. In this paper, HyperMI provides a novel approach that HyperMI returns signal information necessarily to these functions rather than a true address to make the system run normally.

For the second policy, some functions, EPT creating, loading, walking and destroying, need access address of EPT. HyperMI places hooks on these functions, then dispatch them to HyperMI world and handle appropriately. In the meantime, HyperMI handles double mapping to ensure that there is only one virtual address mapping to one physical memory page during the EPT updating, and handles remapping problems to ensure the content of page cleaned after page being swapped out. This will be described in detail later.

For the last one, intercepting the loading EPT operation and verifying the correctness of EPT can avoid loading a wrong EPT and leaking the content of physical memory page. To

TABLE II
PAGE-MARK TABLE.

Page-Mark Table		
Label	OwnerID	Used
Description	The Owner Identifier	Free or Used

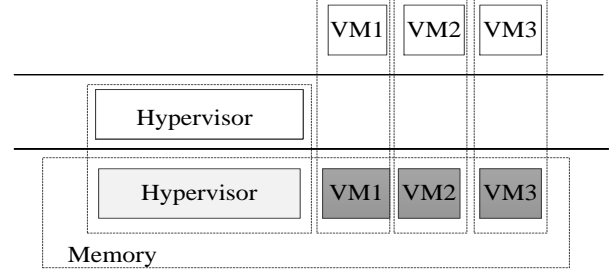


Fig. 5. Memory isolation for VMs.

ensure one EPT for one VM, HyperMI creates the VM-Mark structure stored in HyperMI world as Table 1 described, and record VMID, EPTID, EPT_Address and binds them together. VMID is created and destroyed based on hash value of the image of VM. EPTID and EPT_Address is recorded as long as the EPT of current VM is created.

Memory Isolation for VMs In addition, the isolation EPTs mentioned above, isolating memory is another aspect that should be considered to achieve the goal of VM isolation completely, which is depicted in Figure 5. Hypervisor and every VM can access own physical memory. Without memory isolation mechanism, compromised hypervisor and the compromised VM can access memory pages of the victimized VM by two ways when EPT updates: 1) Double mapping. 2) Remapping to pages with content. So some efforts are done. Monitoring EPT updating and creating Page-Mark structure described in Table 2 to record the owner of every physical memory page.

In order to against the double mapping attack in the process of EPT updating, HyperMI should finish these two tasks: verifying the owner of pages firstly, and then marking the OwnerID of Page-Mark structure for unused pages or thwart the mapping operation for used pages in case of malicious double mapping behavior. So the technique of pages allocation can divide all the pages into different catalogs: the pages of hypervisor or the pages of every VM.

To go against the remapping attack, HyperMI cleans the context of the page when the page is swapped out, so attackers can't get the context of the page by the way of remapping. And HyperMI clears the Page-Mark structure of the corresponding page.

IV. IMPLEMENTATION

HyperMI is implemented on x86 platform based on KVM hypervisor (version 3.10), both the host and guest run Linux. In the following, we present details of the prototype.

A. HyperMI Initialization

An essential requirement during system initialization is only allowing to load the hypervisor and HyperMI into the platforms. The initialization of HyperMI includes the creation

of HyperMI World, the hooks of key functions used to monitor as well as the creation of Page-Mark table and VM-Mark table.

First, as required by HyperMI, HyperMI World is created to provide a trust execution environment. The permission for the kernel code is non-execution when HyperMI World is active. Second, when the hypervisor boots up, the hypervisor has to trap to HyperMI to initialize its critical data including MMU, system control registers, VMCS and EPT. These data access functions are described in table III. Third, these functions are like triggers, and once the relevant critical data is accessed, these operations are hooked to HyperMI World to execute. Finally, HyperMI also examines every page table of hypervisor to confirm that the entry address of HyperMI World are not existed in the hypervisor's address space. Finally, HyperMI updates Page-Mark table to reflect the current memory status and owner. It creates VM-Mark table to blind VM and EPT.

B. Interaction

A VM can not handle the operation of privileged instructions and need the assistance of hypervisor, so they interact through the VM Exit/Entry. Because interaction data, VMCS structure, is stored in HyperMI World, the operation of accessing VMCS structure is trapped in HyperMI World, so VM Exit/Entry control flow is changed. The control flow as follow. 1) VM sends out exit signal to hypervisor. 2) After receiving signal, hypervisor notifies CPU to write the value of system registers to VM status data of VMCS structure, and to write host status data of VMCS structure to system registers. 3) Control flow jumps to HyperMI World to access VM exit reason data of VMCS structure, switches to hypervisor and execute VM exit event handler function. 4) VM Exit finishes. VM Entry also requires access to VMCS in HyperMI World rather than in normal world. The control flow is opposite to that of VM Exit.

C. Page Mapping

To allocate a physical page to a VM, HyperMI must update the relevant address translation of the EPT. Before executing EPT update, the key step for memory isolation is to check the usage state of the mapping page. By checking the Page-Mark table, HyperMI can keep the page away from being owned by multiple owners. If the page is allocated for the first time, HyperMI sets the OwnerID and Used flag of Page-Mark table and updates the corresponding EPT entry. Meanwhile, release this page from the address space of hypervisor. Otherwise, this request fails and page can not be allocated successfully. In this way, HyperMI prevents a compromised hypervisor from executing any double mapping attack to the page already allocated to other owners.

D. Page Unmapping

To deallocate a physical page from a VM, HyperMI clears the appropriate EPT to invalidate this mapping. In addition, HyperMI clears the content of this page before resetting the Used flag of Page-Mark table. Clear the OwnerID and set Used flag to be free. With this approach, page reuse attacks and remapping attacks are prevented.

TABLE III
HOOKED FUNCTIONS.

Hooked Functions					
Structure	Page	Regs	VMCS	EPT	
Functions	page fault	cr0_write cr3_read cr3_write cr4_write	vmcs_readl vmcs_writel	ept	creation loading walking destroying

V. EVALUATION

We describe the design and implementation of HyperMI. In this section, we evaluate the protection effectiveness and performance of HyperMI by exploiting vulnerabilities and comparing it with traditional KVM through a set of benchmarks.

A. Security Analysis

VM Isolation is used to ensure the memory isolation among VMs. Just as the threat model described in the section II-A, an attacker could subvert the upper guest virtual machines by implementing attacks such as cross-domain attack through a malicious virtual machine. In this section, we focus on monitoring for interaction-data and secure VM isolation for memory among VMs, then describe protection for HyperMI against compromised hypervisor. Table IV shows the real attack instances in line with the above attack model.

Modifying Interaction-Data Attack We clarify interaction data including VMCS in Section III-D. VMCS mainly stores the context information of VM and host in VMX operation, such as control registers. To prevent interaction-data leakage, we protect VMCS from the attacker. Firstly, VMCS is hidden in HyperMI World and can not be accessed by hypervisor. Secondly, functions that can access VMCS address are hooked to Hyper World in order to avoid exposing VMCS address. The attacker can not get addresses and access VMCS, this prevents attackers from tampering with interactive data attacks. We implement corresponding attack cases (Interaction-data Attack) described in Table IV, mainly direct access to GUEST_CR3 in VMCS. The experimental results show that the access failed, malicious interaction-data accessing is prevented successfully. According to the above analysis, the attacker can not access VMCS, and can not further attack. Therefore, the VM runtime states and interaction-data cannot be modified.

Subverting Memory Across VMs Attack Original hypervisor manages the memory of VM through EPT which controls the address translation of VM, so compromised hypervisor can incur malicious memory access attack, such as double mapping attack and remapping attack. However, HyperMI hides the address of EPT in HyperMI World and hooks all operations about EPT into HyperMI World. Page tracking technology can make and prevents double mapping and remapping attack. Page tracking technology can ensure that each physical page has only one owner, verify the ownership of each physical page when EPT updates the mapping, and ensure the safe mapping of physical memory. Page tracking technology can clear the

TABLE IV
HYPERVISOR ATTACKS AGAINST HYPERMI.

Attack	Description
Interaction-data Attack	Load a crafted GUEST_CR3 value
CVE-2017-8106	Load a crafted EPT value
DMA Attack	Access HyperMI world by DMA
Code Injection Attack	Inject code and cover hooked functions to bypass HyperMI world

contents of pages when they are completely released, ensuring information security.

The main attacks that attackers can execute on subverting memory are double mapping attacks and remapping attacks. Firstly, double mapping happens when assigning memory pages that have already been owned by a hostile VM to a victim VM. Page tracking and write-protection of EPT prevents this kind of attack. For each new mapping to a VM, HyperMI validates whether the page is already in use. Meanwhile, the allocated pages must be marked in the Page-Mark table for tracking. Secondly, another challenge is page remapping attack by a compromised hypervisor from a victim VM to a conspiratorial VM. This attack involves remapping a private page to a different address space. To defeat this type of attack, HyperMI ensures that whenever a page is released, its content must be zeroed out before creating a new mapping.

We implement a real attack, CVE-2017-8106 in kernel version 3.12. A privileged KVM guest OS user accesses EPT, conducts attacks via a single-context INVEPT instruction with a NULL EPT Pointer. Attackers can not implement successfully and incur EPT access fault because HyperMI hides the address of EPT in HyperMI World and hijacks the loading of EPT. Therefore, HyperMI can avoid subverting memory across VMs including double mapping attack, remapping attack as well as malicious EPT access.

Destroy HyperMI World HyperMI is created by relying on page tables. When HyperMI does not work, the original kernel with high privileges can access any page tables and modify them. It can also access control registers casually, redirect hooked functions used to monitor previously, and even maliciously destroy HyperMI World through DMA. We analyze the protection of HyperMI World from these four aspects, page table modifying attack, hooks redirection attack, reg attack and DMA attack.

1) *Page Table Modifying Attack*: Page table protection has been introduced in section III-C. The entry address mapping of the new page table is deleted from the old page table mapping to prevent the kernel from accessing HyperMI World directly through the page table mapping. When HyperMI World is active, the kernel code does not have any executable permissions in case of attacking running processes in HyperMI World. An attacker may attack in two ways. First, the attacker may try to directly access the new page table address on the kernel page table by virtual address mapping, but when he accesses it, there is page fault due to the absence of address mapping. Moreover, the attacker may run kernel code while the HyperMI World is active to attack programs running in the

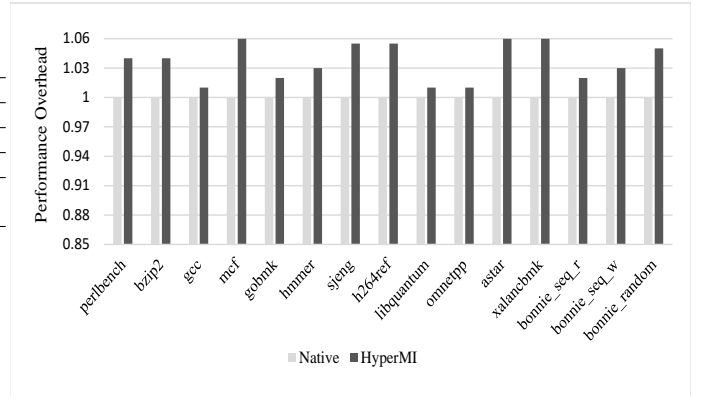


Fig. 6. Performance overhead.

HyperMI World. This can be prevented because of the absence of executable privilege of kernel.

2) *Hooks Redirection Attack*: Due to the code of hooked functions including VMCS operations, EPT operations and control register access operations is writable-protection, CR0 register access operation used to set $W \oplus X$ is controlled and page table updating used to change code execution privilege is limited, the attacker can not redirect hooked functions and bypass being monitoring.

3) *Reg Attack*: Access to some registers access operations including CR0, CR3, CR4, are controlled and hooked to HyperMI World. CR0 register can control the $W \oplus X$ privilege of code, CR3 can control the loading of page table and CR4 can decide SMEP mechanism. Protection for page table, hooked functions and regs play a role mutually in protection for HyperMI.

4) *DMA Attack*: DMA attack is described in detail in section III-C. Attackers can use this feature to read or corrupt arbitrary memory regions. DMA attack is not a threat to HyperMI, because HyperMI is inherently secure against DMA using IOMMU. Remove the corresponding mapping of the critical data from the page table which IOMMU uses. These critical unmapped data includes the entrance address of HyperMI, data recording Page-Mark structure used in VM isolation, VM-Mark structure used in VM monitoring and so on. DMA attack that aims at modifying the VM memory or the page tables will also be defeated.

B. Performance Evaluation

In HyperMI, the kernel is modified so that the HyperMI World is initialized during the boot up sequence. This includes creating a new memory page table for HyperMI, allocating memory pages, as well as creating Page-Mark table and VM-Mark table. This process introduces security verification for pages according Page-Mark table, and security accessing for VMCS in HyperMI World during VM Exit/Entry sequence.

The kernel is modified to place hooks upon some control registers, accessing VMCS operations, and accessing EPT operations. The hooks jump to HyperMI World through the

TABLE V
EXECUTION TIME OF VM OPERATION(S).

Test Case	VM Create	VM Destroy
No_HyperMI	27 s	7 s
With_HyperMI	36 s	14 s
efficiency	1.33	2

switch gate. So HyperMI can introduce some world switching overhead using switch gate.

We do not directly compare HyperMI with previous software approaches, as they cannot support all functions simultaneously. In order to assess the effectiveness of all aspects of HyperMI, we conduct a set of experiments to evaluate the performance impact imposed by HyperMI against an original KVM system (the baseline). We run three groups of experiments and compare the performance overhead, benchmarks performance overhead, VM load time as well as VM Exit/Entry overhead. For simplicity, we only present the performance evaluation on a server with 64 cores and 32 GB memory, running at 2.0 GHz and guest VM with 2 virtual cores. The version of hypervisor and guest VMs is 3.10.1. Different experiments are based on different numbers of guest VMs with different memory size. Both the original and HyperMI systems have the same configuration except the protection supported by HyperMI. The deviation of these experiments is insignificant. All the experiments are replicated fifty times and the average results are reported here.

Benchmarks Performance In order to obtain the impact of HyperMI to the whole system, we measure HyperMI with microbenchmarks and application benchmarks. We use one guest VM with 2 GB memory size.

To better understand the factor causing the performance overhead, we experiment with compute-bound benchmark (SPEC CPU2006 suite) and one I/O-bound benchmark (Bonnie++) running upon original KVM and HyperMI in a Linux VM. The experiments result described in Figure 6(the last three groups) shows relatively low cost. Most of the SPEC CPU2006 benchmarks (the first twelve groups) show less than 6% performance overhead. It's not surprising as there are few OS interactions and these tests are compute-bound. Mcf, astar, and xalancbmk with the highest performance loss allocate lots of memory, and HyperMI handles Page-Mark structure and verify legality of page mapping when EPT updates. This can incur worlds switching which involves control register access and incur VM exit which involves page fault of EPT. For Bonnie++, we choose a 1000 MB file to perform the sequential read, write and random access. The performance loss of sequential read, write and random access is 2%, 3% and 5%, not high, the main reason is that HyperMI has no extra memory operations for I/O data. The performance result shows that HyperMI introduces trivial switch overhead of two worlds and trivial overhead of memory isolation of VMs.

VM Load Time The load time of a VM is a critical aspect of performance, because it influences user experience. We design experiments to evaluate the performance impact

of HyperMI prototype for a VM loading. Experiments are done with 4 VMs, each guest VM is with different memory sizes from 512MB to 4GB. As expected, the VM booting time in HyperMI increases as the memory sizes increase, and the growth amplitude is more and more larger due to the world switch and page tracking caused by frequent memory allocation. We also measure the impact of completely booting a VM (configured with 2 VCPU and 512MB memory). As Table V shown, the booting time is suffered a 1.33 times slowdown under HyperMI, shutdown time is suffered a 2 times slowdown, due to the extra overhead of worlds switching and Page-Mark table accessing. Such overhead is worth for HyperMI.

VM Exit/Entry Overhead Experiments are done on one VM with 2G memory. Network accessing can introduce lots of VM Exit, afterwards VMCS accessing in HyperMI World and world switches. In order to measure the performance impact of HyperMI, we use Netperf (version 2.7.0), a benchmark for measuring various aspects of networking performance, to determine VM Exit/Entry performance overhead. We run a netperf process in the tested VM, sending TCP or UDP streams. The performance of I/O instruction exits (i.e., VM exits triggered by the guest's I/O requests) is 2% for TCP and 3% for UDP, not high. Based on experimental results, we conclude that VMCS accessing in HyperMI World and worlds switching can be accepted.

VI. RELATED WORK

We describe the related work from these three aspects, reconstructed hypervisor, customized hardware, and the same privilege level isolation.

A. Customized Hardware

Some works at the hardware level complete the protection of the process by extending the virtualization capabilities. These tasks provide fine-grained isolation of processes and modules from the hardware level. Haven [2] uses Intel SGX [9], [13] to isolate cloud services from other services and prevent cross-domain access. SGX provides fine-grained protection at the application space instead of hypervisor space, and needs developers spend time reconstructing code and dividing code into trusted part or untrusted part. SGX has requirement for version of CPU. The effort [6] combines the advantages of ARM TrustZone and virtualization to improve system performance, and isolate critical process components securely and efficiently. H-SVM [10] utilizes the hardware extension features of the CPU, and extends SMM microcode to achieve memory resource isolation among virtual machines. It deprives ability of accessing to memory resource by replacing the source code of the original hypervisor to access memory resource. Vigilare [14] and KI-Mon [12] provide monitoring for access operations by introducing extra hardware. Vigilare provides a kernel integrity monitor that is architected to snoop the bus traffic of the host system from a separate independent hardware.

B. Reconstructed Hypervisor

Except for approaches based on hardware, some works [15], [16], [19] pay attention to software isolation. Pre-allocating physical resource and completed isolated environment for every VM can avoid VM cross-domain attack, and data leaking attack. NOVA [16] divides hypervisor into micro-hypervisor and user hypervisor running in root mode, adopts an idea which is similar to fault domain isolation to guarantee an isolated user hypervisor for every VM. The drawback of this approach is the lack of fractional traditional hypervisor functions. HyperLock [19] prepares backup KVM for every VM by copying KVM code, and ensures every VM run in own isolated space. These approaches redesign hypervisor greatly. In contrary, HyperMI adopts a feasible way to isolate VM without lots of modification to hypervisor.

C. The Same Privilege Level Isolation

Some efforts, ED-Monitor [7], SKEE [1] and SecPod [17], adopt the same privilege level idea to avoid performance overhead of inter-level translation. SKEE provides a lightweight secure kernel-level execution environment, this environment is placed at the same privilege level and the same address space with kernel. SKEE is exclusively designed for commodity ARM platforms using system characteristics of ARM. The difference between HyperMI and SKEE is that HyperMI uses two sets of page tables to create the execution environment, and SKEE uses one set of page tables. The design of the switch gate for HyperMI and SKEE are also different. SKEE is more focused on using the characteristic of the ARM platform while HyperMI has no dependence on the platforms. SecPod, an extensible approach for virtualization-based security systems that can provide both strong isolation and the compatibility with modern hardware. The biggest difference between SecPod and HyperMI is that SecPod creates the secure isolation environment for every VM. SecPod solves the problem of virtual machine address mapping with the assistance of shadow page table (SPT) technology.

We neither adopt software at a higher level than the hypervisor, nor use customized hardware. Inspired by the same privilege-level, we propose HyperMI world placed at the same privilege-level with hypervisor.

VII. CONCLUSION

We introduce HyperMI, an approach that enables x86 platforms to support a secure isolated execution environment at the same privilege level with hypervisor. The environment is designed to provide memory isolation protection for VMs, and secure and event-driven runtime monitoring for interaction between hypervisor and VMs. This approach, which does not rely on additional hardware devices or a higher privilege level software, has fewer changes to system and fewer requirements for types of CPU hardware device. It reflects good practicality and portability. And security analysis describes protection for VM, the performance evaluation shows its efficiency by introducing negligible performance overhead. It can be implemented widely in real-world for cloud providers.

REFERENCES

- [1] Azab, A., Swidowski, K., Bhutkar, R., Ma, J., Shen, W., Wang, R., Ning, P.: Skee: A lightweight secure kernel-level execution environment for arm. In: Network and Distributed System Security Symposium (2016)
- [2] Baumann, A., Peinado, M., Hunt, G.: Shielding applications from an untrusted cloud with haven. *Acm Transactions on Computer Systems* **33**(3), 1–26 (2014)
- [3] Ben-Yehuda, M., Day, M., Dubitzky, Z., Factor, M., Har’El, N., Gordon, A., Liguori, A., Wasserman, O., Yassour, B.A., Ben-Yehuda, M.: The turtles project: Design and implementation of nested virtualization. *Yehuda* pp. 1–6 (2007)
- [4] Champagne, D., Lee, R.B.: Scalable architectural support for trusted software. In: HPCA - 16 2010 The Sixteenth International Symposium on High-Performance Computer Architecture. pp. 1–12 (2010)
- [5] Chhabra, S., Rogers, B., Yan, S., Prvulovic, M.: Secureme:a hardware-software approach to full system security. In: International Conference on Supercomputing. pp. 108–119 (2011)
- [6] Cho, Y., Shin, J., Kwon, D., Ham, M.J., Kim, Y., Paek, Y.: Hardware-assisted on-demand hypervisor activation for efficient security critical code execution on mobile devices. In: Usenix Conference on Usenix Technical Conference. pp. 565–578 (2016)
- [7] Deng, L., Liu, P., Xu, J., Chen, P., Zeng, Q.: Dancing with wolves: Towards practical event-driven vmm monitoring. *Acm Sigplan Notices* **52**(7), 83–96 (2017)
- [8] Evtushkin, D., Elwell, J., Ozsoy, M., Ponomarev, D., Ghazaleh, N.A., Riley, R.: Iso-x:a flexible architecture for hardware-managed isolated execution. In: Ieee/acm International Symposium on Microarchitecture. pp. 190–202 (2015)
- [9] Hoekstra, M., Lal, R., Rozas, C., Phegade, V., Cuvillo, J.D.: Cuvillo, “using innovative instructions to create trustworthy software solutions,” in hardware and architectural support for security and privacy. In: 6 IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. (2013)
- [10] Jin, S., Ahn, J., Seol, J., Cha, S., Huh, J., Maeng, S.: H-svm: Hardware-assisted secure virtual machines under a vulnerable hypervisor. *IEEE Transactions on Computers* **64**(10), 2833–2846 (2015)
- [11] Keller, E., Szefer, J., Rexford, J., Lee, R.B.: Nohype:virtualized cloud infrastructure without the virtualization. *Acm Sigarch Computer Architecture News* **38**(3), 350–361 (2010)
- [12] Lee, H., Moon, H., Jang, D., Kim, K., Lee, J., Paek, Y., Kang, B.H.: Ki-mon: a hardware-assisted event-triggered monitoring platform for mutable kernel object. In: Usenix Conference on Security. pp. 511–526 (2013)
- [13] Mckeen, F., Alexandrovich, I., Berenzon, A., Rozas, C.V., Shafi, H., Shanbhogue, V., Savagaonkar, U.R.: Innovative instructions and software model for isolated execution. In: International Workshop on Hardware and Architectural Support for Security and Privacy. pp. 1–1 (2013)
- [14] Moon, H., Lee, H., Lee, J., Kim, K., Paek, Y., Kang, B.B.: Vigilare: toward snoop-based kernel integrity monitor. In: ACM Conference on Computer and Communications Security. pp. 28–37 (2012)
- [15] Shi, L., Wu, Y., Xia, Y., Dautenhahn, N., Chen, H., Zang, B., Guan, H., Li, J.: Deconstructing xen. In: Network and Distributed System Security Symposium (2017)
- [16] Steinberg, U., Kauer, B.: Nova:a microhypervisor-based secure virtualization architecture. In: European Conference on Computer Systems, Proceedings of the European Conference on Computer Systems, EU-ROSYS 2010, Paris, France, April. pp. 209–222 (2010)
- [17] Wang, X., Chen, Y., Wang, Z., Qi, Y., Zhou, Y.: Secpod: a framework for virtualization-based security systems. In: Usenix Conference on Usenix Technical Conference. pp. 347–360 (2015)
- [18] Wang, X., Qi, Y., Dai, Y., Shi, Y., Ren, J., Xuan, Y.: Trustosv: Building trustworthy executing environment with commodity hardware for a safe cloud. *Journal of Computers* **9**(10) (2014)
- [19] Wang, Z., Wu, C., Grace, M., Jiang, X.: Isolating commodity hosted hypervisors with hyperlock. In: Proceedings of the 7th ACM european conference on Computer Systems. pp. 127–140 (2012)
- [20] Zhang, F., Chen, J., Chen, H., Zang, B.: Cloudvisor: retrofitting protection of virtual machines in multi-tenant cloud with nested virtualization. In: ACM Symposium on Operating Systems Principles. pp. 203–216 (2011)