Chapter 6: Modeling the Runtime Integrity of Cloud Servers: a Scoped Invariant Perspective

Jinpeng Wei

Florida International University, Miami, FL, USA

Calton Pu

Georgia Institute of Technology, Atlanta, GA, USA

Carlos V. Rozas, Anand Rajan

Intel Corporation, Hillsboro, OR, USA

Feng Zhu

Florida International University, Miami, FL, USA

Abstract One of the underpinnings of cloud computing security is the trustworthiness of individual cloud servers. Due to the on-going discovery of runtime software vulnerabilities like buffer overflows, it is critical to be able to gauge the trustworthiness of a cloud server as it operates. The purpose of this chapter is to discuss trust enhancing technologies in cloud computing, specifically remote attestation of cloud servers. We will discuss how remote attestation can provide higher assurance that cloud providers can be trusted to properly handle a customer's computation and/or data. Then we will focus on the modeling of the runtime integrity of a cloud server, which determines the level of assurance that remote attestation can offer. Specifically, we propose *scoped invariants* as a primitive for analyzing the software system for its integrity properties. We report our experience with the modeling and detection of scoped invariants for the Xen Virtual Machine Manager.

S. Pearson and G. Yee (eds.), *Privacy and Security for Cloud Computing*,
Computer Communications and Networks, DOI 10.1007/978-1-4471-4189-1_6,

6.1 Introduction

According to IDC's 2008 cloud services user survey [1] of IT executives, security is the number one concern in adopting cloud computing. Part of the reason is that the operating systems supporting the cloud are just the conventional ones used today, which means that they can be compromised and be infected with malware. Not surprisingly, a prospective cloud user is concerned about delegating his data and computation to a cloud server that can be compromised at runtime, even if the server starts in a known-good condition and the cloud provider is trusted.

In other words, a trusted cloud server is not necessarily trustworthy, due to the inherent difficulty of eliminating software vulnerabilities and other operational errors (e.g., configuration mistakes). Therefore, technologies that can enhance the trust of cloud servers are highly demanded.

One way that can enhance the trust of cloud servers and relieve the concern of a potential cloud user is remote attestation [2], which enables the cloud user or a trusted third party to measure the "healthiness" (or integrity) of a cloud server at runtime, so that the compromise (or degraded integrity) can be detected in a timely manner.

There has been a long line of research in software integrity ([2-13]), because malware like rootkits [9] must modify the victim software in some way, thus violating its integrity. In general, the integrity of a system can be approximated by a set of properties that must be satisfied by a "healthy" software system. For example, many rootkits modify the system call table, so a property evaluated by many integrity monitors is whether the system call table has known-good values. It is through such properties that an integrity monitor differentiates a "healthy" system from a corrupted one.

Identifying integrity properties is critical to the effectiveness of any integrity measurement *mechanism*, because without a good set of integrity properties, the use of such mechanisms can be severely limited. For example, if the integrity properties only cover system call table, a new rootkit can manipulate other function pointers (such as those found in device driver jump tables) to achieve its goal and remain undetected.

Therefore, in this chapter we study the problem of systematically identifying integrity properties given the target software, which can then be used as input to an integrity measurement mechanism. Specifically, we make the following contributions:

We propose *scoped invariants* as an important class of integrity properties. Scoped invariants are code or data that has constant value under some context (called their scope). An example scoped invariant is the Interrupt Descriptor Table (IDT) entry for page fault, which contains a constant function pointer once the system finishes its initialization. Scoped invariants are building blocks of more general integrity properties, and are amenable to integrity checking.

Our second contribution is a dynamic analysis tool that detects scoped invariants. Our tool runs the target program in a machine emulator and monitors

memory writes and events generated by the target program. Memory writes monitoring supports or rejects the hypothesis that a variable is an invariant, while event monitoring help decide the scopes in which hypotheses about invariants apply.

Our third contribution is a scoped invariants case study of the Xen Virtual Machine Manager [14], which is the foundational software of many cloud providers. Our tool identifies 271 scoped invariants essential to Xen's runtime integrity. One such invariant property, that the addressable memory limit of a guest OS must not include Xen's code and data, is indispensable for Xen's *guest isolation* mechanism. The violation of this property demonstrates that the attacker only needs to modify a single byte in the Global Descriptor Table (GDT) to achieve his goal.

The rest of the chapter is organized as follows. Section 6.2 gives background information about remote attestation and our security assumptions. Section 6.3 discusses our modeling of software integrity, and proposes scoped invariants as an important class of integrity properties. Section 6.4 presents an automated scoped invariants detection scheme based on dynamic monitoring and statistical inference. Section 6.5 discusses our implementation of an automated tool for deriving scoped invariants. Section 6.6 evaluates our methodology and tool by studying scoped invariants of Xen. Section 6.7 discusses related work, and Section 6.8 concludes the chapter.

6.2 Background on remote attestation and integrity measurement

In this section, we introduce remote attestation as a useful trust enhancement technology for cloud computing; then we discuss the importance of integrity modeling in remote attestation and our security assumptions.

6.2.1 Remote attestation as a trust enhancement technology

A customer of a cloud server may want to determine that the cloud server is "healthy" (free of virus, Trojan horses, worms, and so on), so it can be *trusted* to properly handle the customer's data and computation; he may also want to keep track of the cloud server's health status so that he can stop using the cloud server as soon as he suspects that the server is compromised, to minimize the damage or the delay for recovery. Trusted computing is a technology that can satisfy the needs of such a cloud customer.

A major goal in trusted computing is to provide reliable knowledge about a system to a user or a service provider. That knowledge is normally obtained by an evaluation of the identity and *integrity* of a system, and it serves as evidence that a target system will not engage in some class of misbehaviors, thus it can be trusted

[15]. To this end, the Trusted Computing Group [16] has introduced the concept of *remote attestation*.

Remote attestation enables a computer system in a networked environment to decide whether a target computer has integrity, e.g., whether it has the appropriate configuration and hardware/software stack, so it can be trusted. The idea of remote attestation has been widely accepted. For example, the trusted platform modules (TPM) [17] chip has become a standard component on modern computers.

An *integrity measurement system* (IMS) for remote attestation typically consists of three components: the target system, the measurement agent, and the decision maker [2]. The target system is a computer system whose "healthiness" is being evaluated (e.g., a cloud server); the measurement agent is a software or hardware entity that reads or *measures* the status (e.g., memory content) of the target system; and the decision maker is an entity (e.g., a cloud customer) that draws a conclusion about the integrity of the target system, given the measurements obtained by the measurement agent. Theoretically, a decision maker has some integrity model in mind, which determines the amount of measurements (or evidence) to be collected from the target system; and it is easy to understand that the integrity guarantee by an IMS is only as strong as the comprehensiveness of the integrity model.

6.2.2 Security assumptions about the integrity measurement system

Our first assumption is that the measurement agent is isolated from and independent of the target system, therefore it has a true view of the internal states (including code and data) of the target system. This is a realistic assumption due to the popularity of machine emulators such as QEMU [18], and it has also been shown that the measurement agent can run on dedicated hardware such as a PCI card [9]. Our second assumption is that measurement results are securely stored and transferred to the decision maker. This can be supported by hardware such as a trusted platform module (TPM) [17]. The third assumption is that the target system's states (e.g., code and data) may be compromised by a powerful adversary who can make arbitrary modifications; therefore the decision maker can rely on very few assumptions about the trustworthiness of the target system.

Based on these assumptions, the decision maker is given a true view of the target system, and its task is to estimate the "healthiness" of the target system. The healthiness include functional correctness (e.g., a function that is supposed to reduce the priority level of a task is not modified to actually increase the priority level), and non-functional correctness (e.g., the priority level can be modified by a privileged user instead of a normal user). In the following subsections, we model the healthiness as integrity properties.

Moreover, the healthiness of the target system may change over time, because it may be under constant attacks. Therefore, the integrity of the target system may need to be periodically reevaluated.

6.3 Formal definition of scoped invariants

In this section, we introduce and formally define *scoped invariants* as a class of integrity property; we also define *dependencies* among scoped invariants.

6.3.1 Formalizing integrity properties

In theory, any software system can be modeled as an automaton with states and state transitions. For simplicity of presentation, we assume that the system can be in one of *n* possible states: $s_1, s_2, ..., s_n$. Example states are initialization, entering a function, returning from a function, system termination, and so on. And each state is characterized by a particular combination of values of the system's internal variables. Based on this general formalization, we can model runtime software integrity as a set of properties $\{P_1(s), P_2(s), ..., P_m(s)\}$. A runtime property $P_i(s)$ is a function on state *s* that evaluates to *true* or *false*. If a system state *s* satisfies all P_i's, we can say that *s* is a "healthy" state. Different runtime properties may have different structures, but each of them can be generalized to be a Boolean expression with the operators $\wedge$(and), $\vee$(or), and $\neg$(not). More complex properties can be constructed out of primitive properties using the operators mentioned above. A primitive property has the form $func(v_1(s), v_2(s), ..., v_l(s))$ which takes variables $v_1(s), v_2(s), ..., v_l(s)$ and returns *true* or *false* ($v(s)$ is the value of *v* in state *s*). *func* can have arithmetic operations inside as well as relationship operations like ==, <, and >.

6.3.2 Definition of scoped invariants

Scoped invariants are one special class of primitive property with the form: $v(t) == k, t \in [s_1, s_2)$. E.g., it stipulates that the value of variable *v* must be a specific value *k* when the system enters state s_1, and continue to be this value until the system enters another state s_2 (assuming that there is a sequence of state transitions from s_1 to s_2). We call such a primitive property a *scoped invariant*, and $[s_1, s_2)$ is called its *scope*. An example scoped invariant is a global variable whose value does not change after initialization (e.g., once the system enters the *running* state). For example, the Interrupt Descriptor Table (IDT) entry for page fault is such a scoped invariant. Scoped invariants can be regarded as a simplified form of temporal logic.

Scoped invariants represent an important class of integrity properties. They may include critical internal control data of the system (e.g., function addresses)

that are supposed to remain constant. Examples of such scoped invariants include the Interrupt Descriptor Table (IDT), whose importance to system integrity has been well-understood. Another type of scoped invariant holds security policy data, and the violation of such invariants can directly defeat the corresponding security measures. For example, by tampering with the list of "bad" IP addresses, the attacker can defeat a blacklist-based IDS (Intrusion Detection System).

Note that the scopes of different invariants can vary significantly, depending on whether they are global variables, heap variables, or local variables. The scope of a global invariant can span as much as the entire execution of the program; the scope of a heap invariant must fall within the allocation and the freeing of the heap memory block; finally, the scope of an invariant that is a local variable in a function must be a subset of the interval between the entrance and the exit of the function.

In this chapter, we focus on estimating the target system's integrity from the measurement of scoped invariants. Other forms of integrity properties are subjects of future research.

6.3.3 Using scoped invariants for integrity measurement: practical issues

Scoped invariants fit conveniently into the software integrity measurement paradigm because they are amenable to runtime attestation. Given a scoped invariant $v(t) == k, t \in [s_1, s_2)$, the measurement agent can start to read the value of variable *v* once the system enters state s_1. Then the decision maker can verify if the measurements of *v* are "good" until the system enters state s_2. The verification of *v* is simple – just comparing the runtime measurements of *v* against some known-good value *k*. Note that *k* may be difficult to obtain if it depends on something external to the target program, e.g., configuration parameters. Here we assume that *k* has been determined somehow, e.g., using the dynamic detection technique discussed in Section 6.4.

Although theoretically the definition of the scope of a scoped invariant is simple - just identifying the two boundary states, in a real system it is nontrivial, because typically we do not have an *explicit* and *direct* representation of program states. Instead, we can only *infer* program states from registers, main memory, or the file system. For example, if the program is sequential, the program counter (PC) can tell us the progress that has been made by the program since it is started. However, if there are loops in the program, PC *alone* may not be sufficient because the corresponding instruction may be part of a loop body and we do not know the number of iterations the program has gone through the loop body. In this case, we may need additional information such as the value of a *loop guard* variable to better infer the program state. Finally, when the program handles asynchro-

nous events such as hardware interrupts, the program execution becomes non-deterministic and it may be very hard to reliably infer the program states.

Another related issue is the granularity of the program states, which influence the cost of integrity measurement. At one end of the spectrum, the program can have very coarse-grained states (e.g., *initialization*, *running*, and *termination*). Here the *running* state covers most of the program's life span. At the other end of the spectrum, the program can have very fine-grained states (e.g., one state per instruction execution or even multiple states within one instruction). While the most fine-grained states enable the integrity measurement agent to have the closest thus the clearest view of the target system, it is the most expensive. On the other hand, the coarse-grained states may lead the decision maker miss many important events (including integrity violations due to attacks), but it is cheaper for the decision maker to keep track of the program states. Therefore, there is a tradeoff between the granularity of program states and the effectiveness of integrity monitoring.

The third issue is the tracking of program states by the measurement agent. As we mentioned in Section 6.2.2, an attacker may change the target program in arbitrary ways, so we cannot rely on the target program to notify the measurement agent about its states. Instead, we can only let the agent actively *poll* the state from a different domain. Specifically, the agent can run in a more privileged domain from which it can intercept the target program's execution and inspect registers, memory, and files of the target program. As will be discussed in Section 6.4, a machine emulator is a good choice to run the measurement agent securely.

One related issue is performance overhead introduced by integrity measurement. As discussed above, a measurement agent needs to intercept the target program's execution, which causes delays in the target program. Obviously, the slowdown factor depends on the frequency (how often a measurement is taken) and duration (how long each measurement takes) of the measurements, and the duration depends on the number of invariants that need to be checked.

6.3.4 Composition of scoped invariants

Scoped invariants are building blocks of more general integrity properties. In this section, we discuss how we can evaluate more general integrity properties from the result of evaluating individual scoped invariants. The key observation is to look at the dependency relationship among integrity properties and build a hierarchy (represented in invariant dependency graphs or IDGs, defined shortly). We extend the definition of scoped invariant (see section 6.3.2) so that the variable v can be arbitrary object (e.g., a function, a code segment, or a data structure).

In a complex target system such as an operating system, the integrity of different functionality modules is often related. This is because a module may invoke functions provided by some external module (the *callouts*), and it may supply *callback* functions that are supposed to be called by an external module. If an external function (e.g., `init_timer` in Fig. 6.1) that is called by a module (e.g., the

Xen scheduler) misbehaves, the control integrity of the calling module (e.g., the Xen scheduler) may be influenced. Similarly, if an external module (e.g., softIRQ) misbehaves by not invoking the callback function (e.g., `schedule` in Fig. 6.1) supplied by a module (e.g., the Xen scheduler calls `open_softirq`), that module may not get control as expected.

Correspondingly, different scoped invariants can be correlated. Below we formally define dependency between scoped invariants and a data structure (called Invariant Dependency Graph) that can be used to express the structural dependency relationship among a set of scoped invariants.

Definition 1 (dependency between scoped invariants): a scoped invariant i_1 is said to depend on another scoped invariant i_2 if one of the following cases is true:

1. i_1 and i_2 are both code and there is a callout from i_1 to i_2, or i_1 has a callback function supposed to be invoked by i_2.
2. i_1 is code and i_2 is data, but whether control can go to i_1 depends on the value of i_2.
3. i_1 and i_2 are both data and the evaluation of i_1 depends on the evaluation of i_2.

Case 2 of definition 1 applies to the situation in which i_2 is a function pointer, and i_1 is the function that i_2 points to.

Definition 2: An Invariant Dependency Graph (IDG) is a directed acyclic graph $G = < V, E >$, where each member of V represents a scoped invariant, and if $i_1 \in V, i_2 \in V$, and i_1 depends on i_2, there is an edge $e = (i_1, i_2) \in E$.

```
DEFINE_PER_CPU (struct schedule_data, schedule_data);
static struct scheduler ops;
……
static void vcpu_periodic_timer_fn(void *d){……}
int sched_init_vcpu(struct vcpu *v, unsigned int processor){
 ……
 init_timer(&v->periodic_timer,  vcpu_periodic_timer_fn, v, v->processor);
 ……
}
static void schedule(void){……}
void __init scheduler_init(void){
 ……
 open_softirq(SCHEDULE_SOFTIRQ, schedule);
 ……
}
```

Fig. 6.1 Code Snippet of the Xen scheduler ($XEN/xen/common/schedule.c)

An IDG thus is a convenient representation of scoped invariants and their relationship. An example IDG is shown in Fig. 6.1.

An IDG also provides useful guidance in terms of how to evaluate the integrity of a target system in a bottom-up way: for example, if an integrity property i depends on $i_1, i_2, \cdots$, and i_m, then in order for i to be *true*, $i_1, i_2, \cdots$, and i_m must all be *true*. Thus, a decision maker should evaluate $i_1, i_2, \cdots$, and i_m before evaluating i.

6.4 Automated detection of scoped invariants

In this section, we present a scoped invariants detection scheme based on dynamic profiling and statistical inference. We will discuss first the rationale (Section 6.4.1), and then two technical components: memory write monitoring (Section 6.4.2) and event monitoring (Section 6.4.3).

6.4.1 Overview

The inference of scoped invariants can be labor-intensive and error-prone if performed manually. Therefore, tools are needed to automate this process.

By definition, a scoped invariant $v(t) == k, t \in [s_1, s_2)$ has a constant value k when the system state is between s_1 and s_2. Accordingly, the scoped invariant detection must answer the following questions for each scoped invariant: (1) what are the starting and end states that define the scope? (2) which variable (v) is involved? and (3) what is the known-good value (k)?

Note that scoped invariants are with respect to their scopes, i.e., the same variable can be an invariant in a narrower scope but not in a broader scope if the broader scope includes an operation that changes the value of the variable. Therefore, we must first decide the scope and then decide whether a variable is an invariant within that scope.

Our invariant detection employs a dynamic profiling approach. Specifically, we run the target program in a machine emulator and monitor memory writes and events generated by the target program. Memory writes monitoring supports or rejects the hypothesis that a variable is an invariant, while event monitoring help decide the scopes in which hypotheses about invariants apply. In the remainder of this section, we first discuss memory write monitoring, and then discuss event monitoring.

6.4.2 Memory writes monitoring

By definition, a scoped invariant should not be modified other than the initialization. In other words, a variable that is modified multiple times is unlikely an invariant. Based on this reasoning, we can detect invariants by observing how the target software modifies its variables: if a variable is modified multiple times, it is unlikely an invariant; otherwise, it is an invariant.

Using dynamic profiling, we run the target software and collect its modifications to variables, which translate to memory writes. There are multiple ways to do this, including program instrumentation and emulation. Using emulation, we can run the target software in a machine emulator, which can intercept every memory write operation (e.g., a MOV instruction). With this capability, we can record the target memory address and the value written in each memory write operation. The result of dynamic profiling is a sequence of tuples: $w_1, w_2, \cdots, w_n$, where $w_i = (addr_i, v_i)$.

Given a sequence $w_1, w_2, \cdots, w_n$, we can compute the frequency c_i of updates to each unique address $addr_i$. Then, we can sort $addr_i$'s at the ascending order of c_i's, and the sorted list of $addr_i$'s is a list of potential invariants with the most likely at the beginning and the most unlikely at the end. Note that the computation here captures addresses that are updated at least once; addresses that are not updated in the sequence are automatically inserted at the beginning of the sorted list as the most likely invariants.

6.4.3 Event monitoring

In addition to memory writes, the machine emulator also intercepts other events that help define the scopes of the invariants. As discussed in Section 6.3.3, program states can be defined at various granularities, with different tradeoff between integrity measurement precision and cost. We choose to monitor two types of such events: function calls and function returns. The reason is that functions can give semantic meaning for creating (by initialization) or re-creating (by updating) an invariant. In other words, we can say that the scope of an invariant is between when it gets its value in some function and when it is assigned a different value in another function. Tracking the invocations and returns from functions is thus important for determining the scopes of invariants.

For example, the global variable `opt_noirqbalance` of Xen controls whether IRQ balance should be enabled, and Xen allows this configuration parameter to be modified by the hypercall `platform_op`. Obviously, this variable is an invariant between two consecutive `platform_op` hypercalls that modify it.

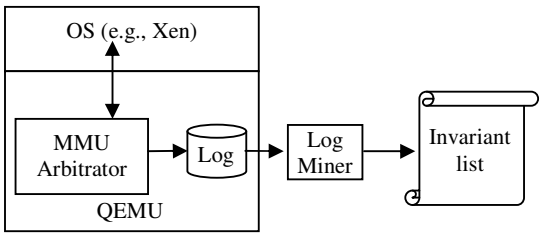

Fig. 6.2 Scoped invariant detection Architecture [19]

6.5 Implementation

We develop a prototype tool that can automatically derive invariants. As Fig. 6.2 shows, we first run the target software on top of QEMU [18], a CPU emulator, which enables us to log all memory write operations of the target software (by the MMU Arbitrator). We also log important system events such as entering and exiting a function, which represent program states that define invariant scopes. Then the Log Miner performs an offline processing of the log – given the sequence of memory write operations between two system events, ranking the memory locations based on the number of modifications to them (with the least modified on the top), and mapping the memory locations to global variables (using symbol information).

The output of the Log Miner is a list of candidate invariants, ranked from the most likely to the least likely. If a variable is indeed an invariant, it will be ranked high in the candidate list – i.e., we will not miss the true invariants. However, some **non**-invariant variables may be ranked high because the condition that leads to their updates is not satisfied during the limited profiling. This is a typical limitation of dynamic analysis, which can be remediated by profiling the target program multiple times each with a different set of input. We can also filter such non-invariant variables using static analysis of the source code, which is out of the scope of this chapter.

6.6 Evaluation

To test the applicability of scoped invariants, this section takes Xen as the target system to do several case studies. We first discuss the motivation of choosing Xen as the target system (Section 6.6.1); next we discuss a scoped invariant with GDT that is critical to Xen's guest isolation mechanism (Section 6.6.2). In Section

6.6.3 we describe a scoped invariant dependency study of the Xen scheduler. Section 6.6.4 presents the result of an automated study of Xen's global invariants.

6.6.1 Choice of Xen as the subject of study

Virtualization is the foundational technology for cloud computing, and Xen [14] is one representative VMM (virtual machine manager) that allows multiple operating systems (called guest OSes or simply guests) to share the same physical machine. As the lowest layer in the cloud computing software stack, the runtime integrity of Xen is the root of trust for a cloud computing environment.

It is generally believed that Xen is more secure than commodity operating systems such as Windows and Linux because it is smaller and simpler. However, we cannot rule out the possibility of a malicious modification to Xen at runtime. For example, There could be vulnerabilities with Xen that can be exploited [20, 21]. Even if Xen is completely bug-free, there are environmental issues such as DMA and system management mode (SMM) [22] that can modify Xen at runtime. Therefore we feel it useful to choose Xen as the target system to perform an integrity study. The particular Xen version studied in this chapter is a pre-release of Xen 3.0.4.

6.6.2 Study of the GDT scoped invariant

One essential security goal of Xen is guest isolation, e.g., a guest operating system should not have access to information about other guests on the same platform, nor should a guest have access to Xen's internal state information.

This guest isolation goal is achieved by scoped invariants associated with some entries of the Global Descriptor Table (GDT) [23]. Specifically, to avoid unauthorized access to its internal state from guests, Xen leverages the standard IA-32 segmentation and protection rings architecture: a guest operating system runs in ring 1 and guest processes run in ring 3, and four special *guest segments* are defined for them. For example, the data segment for ring 3 has the selector 0xe033 in the GDT. The "limit" of these guest segments **is** intentionally made smaller than 4GB such that Xen's code and data are excluded (Xen's code and data reside at the top of every address space).

Such a configuration is represented in the form of scoped invariants because information about these guest segments is stored in memory, in a data structure called `gdt_table`. Setting of the proper descriptor values for `gdt_table` is performed in the initialization phase of Xen, and after that the "limit" fields of the relevant entries are not supposed to change, in other words, they are scoped invariants.

It is easy to understand that a runtime modification to the `gdt_table` entries (e.g., setting the "limit" field to 4GB) could undo the effect of Xen's initialization and expose the complete 4 GB address space back to the guests. Then suddenly a guest can freely read Xen's data, violating the guest isolation security goal.

We have experimentally confirmed that modifying the "limit" field of the `gdt_table` entries at runtime enables a para-virtualized guest to read Xen's data and retrieve the list of domains on the platform by loading its **DS** register with 0xe033. This means that our hypothesis is valid. And it turns out that only one byte needs to be modified (from 0x67 to 0xFF). We should note that Xen virtualizes the GDT table for each guest domain, which means that each guest domain has its own GDT. However, each guest GDT derives its entries for the guest segments from the same `gdt_table`. Therefore, a modification to the `gdt_table` applies to all guest domains.

The GDT example demonstrates how a particular scoped invariant can influence Xen's high level security goals – i.e. guest isolation. Therefore, this invariant must be checked by a decision maker.

6.6.3 Integrity Dependency Analysis of the Xen Scheduler

In this section, we perform an integrity dependency analysis of the Xen scheduler. We will demonstrate the dependencies among scoped invariants. We choose the scheduler because it is one of the most important functionalities of Xen, which allows multiple operating systems to share the physical CPU. The quality of this sharing is determined by the scheduler. Besides, if we can verify the integrity of the scheduler, we can trust it to run other security measures such as integrity monitors for the guest kernel.

The security goal that we choose is *complete mediation*. Under the context of scheduling it means that no task should be able to use the CPU without the permission from the scheduler. In other words, the scheduler should always be able to control when and for how long a particular task can use the CPU.

Fig. 6.4 shows the invariant dependency graph associated with the Xen scheduler. Below we will discuss the reasoning behind this graph.

In order to fulfill complete mediation, the scheduler needs two necessary conditions: (1) when running, the scheduler correctly implements a scheduling algorithm (e.g., the credit-based scheduling algorithm in Xen); (2) the scheduler can have a chance to run when it needs to. Condition (1) can be satisfied by guaranteeing the integrity of the scheduler code. Satisfying condition (2) is challenging, because from time to time the scheduler has to give up CPU so that the normal tasks can make progress, but it must be able to regain control of the CPU to do its job. If these two necessary conditions are not guaranteed, we say that the security goal of complete mediation for the scheduler is not achieved. Therefore, we have derived from the security goal two integrity properties: (1) the scheduler code is not com-

promised, or equivalently, the scheduler code is a scoped invariant (#1 in Table 6.1); and (2) the scheduler is able to get control when it should.

In order to achieve integrity property (2), Xen scheduler relies on the Timer functionality (Fig. 6.3), which guarantees that control will go to a callback function supplied by the scheduler after some amount of time into the future. For example, when the scheduler decides to let a task run, it starts a timer which will expire after an interval equal to that task's time slice. The callback function (`s_timer_fn`) associated with this timer forces a decision to be made concerning which task runs next. This timer helps to avoid the situation where a task excessively occupies CPU and nobody can stop it.

Xen scheduler has to trust the Timer facility mentioned above to work as expected (e.g., the Timer should guarantee precision of some degree); otherwise Xen scheduler cannot achieve its goals. Therefore, the Timer is a scoped invariant (#2 in Table 6.1), and the integrity of Xen scheduler is dependent on the integrity of the Timer facility.

The timer facility in turn relies on the soft IRQ mechanism of Xen (Fig. 6.3). Different from hard IRQs (hardware interrupts), which can interrupt the currently running task at almost any point, soft IRQs do not directly interrupt currently running task. Instead, they are piggy-backed in the hardware interrupt handling procedure, e.g., after an interrupt has been served but before the interrupt handler returns. Specifically, the interrupt handler procedure calls `do_softirq`, which in turn checks the presence of soft IRQs and calls their respective handler functions. Therefore the code of `do_softirq` should be a scoped invariant (#3 in Table 6.1).

For the soft IRQ mechanism to work, several preconditions must hold. One of them is that `do_softirq` must be invoked in the interrupt handling procedure. This is an issue because `do_softirq` is not invoked by hardware, but the interrupt handling procedures which are code in the memory. Therefore, the integrity of interrupt handling code is a precondition for the integrity of Xen's soft IRQ mechanism. In other words, the interrupt handling code is a scoped invariant (#4 in Table 6.1).

Table 6.1 Scoped invariants associated with the Xen scheduler. RC means Runtime Code, KGC means Known Good Code, RD means Runtime Data, and KGD means Known Good Data

1	$RC_{scheduler}$ [*initialization*, *termination*] = $KGC_{scheduler}$
2	RC_{timer} [*initialization*, *termination*] = KGC_{timer}
3	$RC_{do\_softirq}$ [*initialization*, *termination*] = $KGC_{do\_softirq}$
4	$RC_{inthandler}$ [*initialization*, *termination*] = $KGC_{inthandler}$
5	RD_{idt} [*initialization*, *termination*] = KGD_{idt}
6	RD_{gdt} [*initialization*, *termination*] = KGD_{gdt}
7	RD_{tss} [*initialization*, *termination*] = KGD_{tss}
8	$RD_{pgtable}$ [*initialization*, *termination*] = $KGD_{pgtable}$
9	$RD_{softirq\_handlers}$ [*initialization*, *termination*] = $KGD_{softirq\_handlers}$

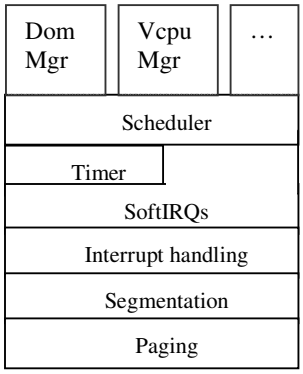

Fig. 6.3 Module Structure Related to the Xen Scheduler

However, even if the interrupt handling code is intact, they must be called when interrupts happen. The hardware provides the Interrupt Descriptor Table (IDT) for the software to register interrupt handlers. Each entry of this table has information about the address of the function to invoke when the corresponding interrupt happens. Therefore, the integrity of the IDT is a precondition for the integrity of interrupt handling of Xen, and one step further, the soft IRQ mechanism of Xen. So the relevant IDT entries are scoped invariants (#5 in Table 6.1).

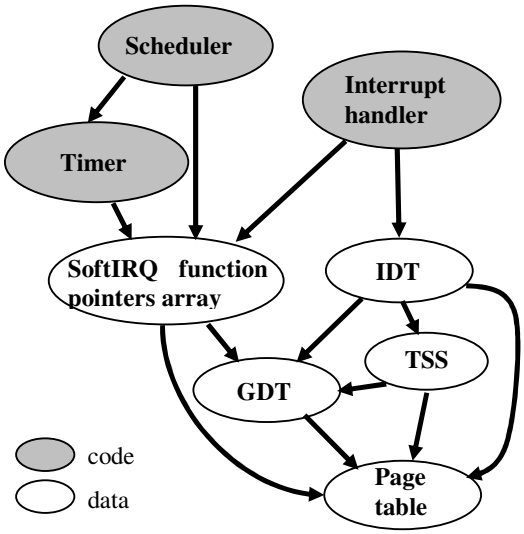

Fig. 6.4 Invariant Dependency Graph Related to the Xen Scheduler

In normal execution mode, an IDT entry refers to code in memory in terms of a segment selector and an offset. Each memory segment has a base address and a limit, and the information about the segments is stored in the Global Descriptor Table (GDT). When an interrupt happens, the handler function's segment selector and offset are fetched from the IDT. Then the segment selector is used to get the base address from the GDT, and the offset is added to the base address to form the linear address of the interrupt handling function. Therefore, the GDT entry must give the correct base address in order for the right interrupt handling function to be located. In other words, the relevant GDT entries are scoped invariants (#6 in Table 6.1), because they are used to evaluate (the linear address of) the interrupt handling code.

Furthermore, some interrupts are handled by task gates (e.g., double fault), whose details (such as handler function entry and stack pointer) are stored in Task State Segments (TSS). So according to our model, there is a dependency relationship from the IDT entry to the relevant TSS, so the TSS becomes a scoped invariant (#7 in Table 6.1).

Finally, there is another layer of indirection due to modern CPU's paging mechanism. Specifically, an interrupt handling function address derived from IDT, GDT and perhaps TSS is a linear address, and the paging mechanism of the underlying hardware maps this linear address to physical address in physical memory, where the handler code resides. But software can control the mapping by supplying page tables, and the page tables are again in memory which can be modified. Therefore, the integrity of page tables is essential to the interrupt handling process of Xen, and due to all the above description, the integrity of the Xen scheduler. So the relevant page table entries are also scoped invariants (#8 in Table 6.1).

In Fig. 6.4, the dependency edges from GDT, IDT and TSS to page tables are due to the fact that on the Intel architecture, GDT, IDT and TSS are known to the CPU in terms of linear addresses. In order to evaluate such data structures, the CPU needs to go through the paging mechanism controlled by the page tables.

As mentioned above, in order for the soft IRQ mechanism to work, several preconditions must hold. We have described one of them: that `do_softirq` be invoked in the interrupt handling process. But we need one more precondition. Specifically, `do_softirq` consults a function pointer array (`softirq_handlers`) for the handler of a particular soft IRQ, so the content of this array must not be compromised. In other words, the relevant entries in the `softirq_handlers` array are scoped invariants (#9 in Table 6.1). For example, Xen scheduler registers a function `schedule` for soft IRQ 1, meaning that `schedule` will be called when soft IRQ 1 is raised (see Fig. 6.1). But if an attacker modifies the function pointer for soft IRQ 1, some other function instead of `schedule` will be called. Then Xen scheduler is essentially bypassed.

Another important soft IRQ is the timer soft IRQ, which implement the Timer facility. We have mentioned that Xen scheduler relies on it. The Timer facility registers `timer_softirq_action` as the call back function.

We can summarize the integrity analysis of Xen scheduler with the Invariant Dependency Graph in Fig. 6.4.

6.6.4 A comprehensive detection of Xen scoped invariants

We have performed a comprehensive study of scoped invariants for Xen, using the QEMU-based profiler and the Log Miner in Fig. 6.2.

We first ran Xen in the profiler, and used the Log Miner to generate the candidate scoped invariants list. Then we did a static analysis to confirm the real scoped invariants. Our static analysis scans the source code of Xen to locate all statements that write to a candidate invariant. We found that most of the candidate invariants have only one such statement (for initialization).

Our analysis suggests that most of the Xen global variables are scoped invariant at runtime. If we only consider the number of variables declared, 75% of them (271 out of 362) turn out to be invariants. If we also consider the size of the variables, then more than 90% of the memory locations corresponding to these global variables are invariant at runtime.

Table 6.2 shows some of the identified invariants. We have classified them based on an informal reasoning about why they should be invariants. Below we give details of some of these scoped invariants:

- `sched_sedf_def` is a data structure that stores the addresses of several functions that together implement the simple earliest deadline first (SEDF) algorithm of Xen. These functions are invoked when a virtual CPU is initialized, suspended, resumed, and so on. Obviously, they should be scoped invariants because otherwise an attacker can modify them to induce Xen's control flow to a malicious scheduling algorithm. Conceptually, `sched_sedf_def` is similar to the IDT. From Table 6.2 we can see that there are 27 more such scoped invariants in Xen.
- `opt_sched` holds the value of a boot-time parameter, which selects one of the built-in scheduling algorithms to be used by Xen. Since Xen does not support on-the-fly change of its scheduling algorithm, this variable should be a scoped invariant.

Table 6.3 gives more information about the invariants `idle_pg_table`, `idle_pg_table_l2` and `idt_table` identified in Table 6.2. First, since only part of such data structures (arrays) are invariants, Table 6.3 gives the range information. We have used macros (e.g., DIRECTMAP_VIRT_START) from Xen source code because their exact values depend on the hardware configuration (e.g., whether Physical Address Extension [23] is enabled). Second, the column denoted "Initialized By" shows the last function that sets the value of a particular scoped invariant. The goal of identifying functions in the "Initialized By" column is to specify the start of the scope of a scoped invariant, because since then the value of the scoped invariant is supposed to be constant.

6.6.5 Discussion

The degree to which a set of scoped invariants can approximate runtime integrity of a software system remains a research question. For example, the invariants that we identified are all necessary conditions, but they may not be sufficient. Assuming that a right set of scoped invariants is at hand, we can estimate the runtime integrity of the system by verifying them. If all of them are verified, we have more confidence about the system's integrity. But if some of them do not pass the verification, we know that the system has lost its integrity.

Table 6.2 Sample scoped invariants (global variables) identified for Xen

Type	Total Number	Examples
Static variables that are definitely invariants	63	schedulers, large_digits, small_digits
Effectively static structures (e.g., contains important function pointers)	28	sched_bvt_def, sched_sedf_def, ioapic_level_type, ioapic_edge_type, amd_mtrr_ops, apic_es7000, hvm_mmio_handlers, exception_table, hypercall_table
Variables that are effectively invariant given a particular boot configuration	17	opt_badpage, opt_sched, opt_conswitch, opt_console, acpi_param, debug_stack_lines, lowmem_emergency_pool_pages, dom0_nrpages
Variables that are effectively invariant given a hardware configuration	102	new_bios, ioapic_i8259, mp_bus_id_to_pci_bus, boot_cpu_logical_apicid, es7000_plat, dmi_ident, hpet_address, vmcs_size, max_cpus, max_page, cpu_present_map, vector_irq, irq_vector
Variables that are effectively invariant given a software configuration	4	softirq_handlers, gdt_table, change_point_list, key_table
Arrays whose entries are mostly invariant	7	idle_pg_table, idle_pg_table_l2, e820, e820_raw, irq_2_pin cpu_sibling_map, cpu_core_map, idt_table

Table 6.3 More information of idle_pg_table, idle_pg_table_l2, and idt_table

Table name	Start offset	Number of entries	Initialized By
idle_pg_table	0	4	xen/arch/x86/boot/x86_32.S
idle_pg_table_l2	DIRECTMAP_VIRT_START / (1<< L2_PAGETABLE_SHIFT)	DIRECTMAP_PHYS_END / (1<< 2_PAGETABLE_SHIFT)	__start in xen/arch/x86/boot/head.S
idle_pg_table_l2	0	16MB / (1<< L2_PAGETABLE_SHIFT)	__start in xen/arch/x86/boot/head.S
idle_pg_table_l2	FRAMETABLE_VIRT_START / (1<<L2_PAGETABLE_SHIFT)	(FRAMETABLE_MBYTES <<20) / (1<< L2_PAGETABLE_SHIFT)	init_frametable in xen/arch/x86/mm.c
idle_pg_table_l2	RDWR_MPT_VIRT_START >>L2_PAGETABLE_SHIFT	(max_page * BYTES_PER_LONG) >> L2_PAGETABLE_SHIFT	paging_init in xen/arch/x86/x86_32/mm.c
idle_pg_table_l2	RO_MPT_VIRT_START >> L2_PAGETABLE_SHIFT	(max_page * BYTES_PER_LONG) >> L2_PAGETABLE_SHIFT	paging_init in xen/arch/x86/x86_32/mm.c
idle_pg_table_l2	IOREMAP_VIRT_START >> L2_PAGETABLE_SHIFT	IOREMAP_MBYTES >> (L2_PAGETABLE_SHIFT – 20)	paging_init in xen/arch/x86/x86_32/mm.c
idt_table	0	128	init_IRQ in xen/arch/x86/i8259.c, apic_intr_init in xen/arch/x86/apic.c, trap_init in xen/arch/x86/traps.c, percpu_traps_init in xen/arch/x86/x86_32/traps.c
	129	127	
idt_table	128	1	dom0 kernel

6.7 Related work

In this section, we give a survey of existing research related to our work, grouped into different topic areas.

6.7.1 Invariants detection

The Daikon invariant detector [24] generates likely invariants using program execution traces collected during sample runs. Daikon is the closest to our work in theory, but the two are different: Daikon instruments the program source code to

emit data traces at specific program points, while our tool transparently intercepts program execution from a machine emulator.

6.7.2 Integrity measurement mechanisms

There has been a long line of research on integrity measurement. Approaches such as IMA [12] use hashing or digital signatures to measure the software at load time. Recently, ReDAS [25] and DynIMA [5] advance the state of the art by supporting software integrity measurement at runtime. Other related work includes [2, 6, 9, 10, 11, and 13]. These approaches generally focus on the mechanism for measurement, but not the integrity properties.

Copilot [9] is a co-processor based integrity checker for the Linux kernel. The properties that Copilot prototype checked were kernel code, module code, and jump tables of kernel function pointers. Although Copilot later provided a specification language [10], its focus was not on deriving integrity properties. We work out the properties from analyzing the target software itself.

Livewire [6] leverages a VMM (a modified version of VMware workstation) to implement a host-based intrusion detection system. It can inspect and monitor the states of a guest OS for detecting intrusions, and interposes on certain events, such as interrupts and updates to device and memory state. Like Copilot, Livewire does not focus on the identification of integrity properties but only checks known properties.

LKIM [2] produces detailed records of the states of security relevant structures within the Linux kernel using the concept of contextual inspection. However, the identification of security relevant structures relies on domain knowledge. This chapter proposes an approach for systematically finding such structures.

6.7.3 Specialized integrity property measurement

Some specialized integrity properties have been measured, such as control flow integrity [3] and Information flow integrity [26]. [3] checks if the control transfer from one function to the next is consistent with a pre-computed control flow graph, so we can think of it as checking a sequence property of the target software. [26] checks the integrity of a system by reasoning about information flows, but it assumes that there is no direct memory modification attack, e.g., information flows are triggered by well-defined interfaces (function calls or file reads).

6.7.4 Rootkits detection and recovery

As we mentioned, there has been a lot of research on rootkits. A nice survey of rootkits and detection software is given in [9]. From [27] you can also find a list of popular rootkits. The integrity measurement mechanisms (such as [6, 9, 11, and 13]) mentioned above all can be used for rootkit detection. Some work such as [7] and [8] attempts to detect rootkits and recover the software from known-good copies.

6.7.5 Trusted computing

The Trusted Computing Group [16] has proposed several standards for measuring the integrity of a software system and storing the result in a TPM (trusted platform module) [17] whose state cannot be corrupted by a potentially malicious host system. Industry vendors such as Intel have embedded TPM in their hardware. Such standards and technologies have provided the root of trust for secure booting [28], and enabled remote attestation [15]. There has been a consistent effort in building a small Trusted Computing Base (with hardware support such as TCG and application level techniques such as AppCore [29]). A small Trusted Computing Base facilitates integrity analysis and monitoring.

6.8 Conclusion

In this chapter, we have discussed remote attestation as a critical and useful trust enhancing technology for cloud computing. We studied one important aspect of remote attestation that is often ignored, the problem of systematically modeling the runtime integrity of a target system, e.g., a cloud server. We proposed scoped invariants as an important class of integrity properties, and we designed and implemented automated tools that can derive scoped invariants out of the target software.

To evaluate our methodology, we applied our tools to the Xen VMM and identified 271 scoped invariants that are critical to Xen's runtime integrity. We experimentally confirmed some of these invariants, including one that can be violated to defeat Xen's guest isolation mechanism.

References

[1] IT Cloud Services User Survey, pt.2: Top Benefits & Challenges. http://blogs.idc.com/ie/?p=210, accessed August 16, 2010.

[2] Loscocco PA, Wilson PW, Pendergrass JA, McDonell CD (2007) Linux kernel integrity measurement using contextual inspection. Proceedings of the 2007 ACM workshop on Scalable Trusted Computing (STC).
[3] Abadi M, Budiu M, Erlingsson U, Ligatti J (2005) Control-flow integrity. ACM Conference on Computer and Communications Security (CCS).
[4] Baliga A, Kamat P, Iftode L (2007) Lurking in the shadows: identifying systemic threats to kernel data. Proceedings of the 2007 IEEE Symposium on Security and Privacy, Oakland, CA, May 2007.
[5] Davi L, Sadeghi A, Winandy M (2009) Dynamic Integrity Measurement and Attestation: Towards Defense against Return-Oriented Programming Attacks. Proceedings of the 2009 ACM workshop on Scalable Trusted Computing (STC).
[6] Garfinkel T, Rosenblum M (2003) A virtual machine introspection based architecture for intrusion detection. Proceedings of Network and Distributed Systems Security Symposium (NDSS), February 2003.
[7] Grizzard J, Dodson E, Conti G, Levine J, Owen H (2004) Toward a trusted immutable kernel extension (TIKE) for self-healing systems: a virtual machine approach. Proceedings of 5th IEEE Information Assurance Workshop.
[8] Levine J, Grizzard J, Owen H (2004) Re-establishing trust in compromised systems: recovering from rootkits that trojan the system call table. Proceedings of the 9th European Symposium on Research in Computer Security, Sophia Antipolis, France.
[9] Petroni N Jr, Fraser T, Molina J, Arbaugh WA (2004) Copilot—a coprocessor-based kernel runtime integrity monitor. 13th USENIX Security Symposium.
[10] Petroni N Jr, Fraser T, Walters A, Arbaugh WA (2006) An architecture for specification-based detection of semantic integrity violations in kernel dynamic data. 15th USENIX Security Symposium.
[11] Petroni N Jr, Hicks M (2007) Automated detection of persistent kernel control-flow attacks. 14th ACM Conference on Computer and Communications Security (CCS).
[12] Sailer R, Zhang X, Jaeger T, Doorn LV (2004) Design and implementation of a TCG-based integrity measurement architecture. 13th USENIX Security Symposium.
[13] Zhang X, Doorn LV, Jaeger T, Perez R, Sailer R (2002) Secure coprocessor-based intrusion detection. Tenth ACM SIGOPS European Workshop, Saint-Emilion, France.
[14] Barham P, Dragovic B, Fraser K, et al (2003) Xen and the art of virtualization. ACM Symposium on Operating Systems Principles (SOSP), Bolton Landing, NY, Oct 2003.
[15] Sheehy J, Coker G, Guttman J, et al (2008) Attestation: evidence and trust. http://www.mitre.org/work/tech_papers/tech_papers_07/07_0186/07_0186.pdf, accessed August 16, 2010.
[16] Trusted Computing Group. http://www.trustedcomputinggroup.org, accessed August 16, 2010.
[17] Trusted Platform Modules. http://www.trustedcomputinggroup.org/developers/trusted_platform_module/specifications, accessed August 16, 2010.
[18] Bellard F (2005) QEMU, a fast and portable dynamic translator. Proceedings of the 2005 USENIX Annual Technical Conference.
[19] Wei J, Pu C, Rozas CV, Rajan A, and Zhu F (2010) Modeling the runtime integrity of cloud servers: a scoped invariant perspective. International Workshop on Cloud Privacy, Security, Risk and Trust (CPSRT 2010), in conjunction with the 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2010), Indianapolis, IN, Nov. 30 - Dec. 3, 2010.
[20] Xen local security-bypass vulnerability. http://www.securityfocus.com/bid/26954/discuss, accessed August 16, 2010.
[21] Xen “move-to-rr” RID local security bypass vulnerability. http://www.securityfocus.com/bid/26716/discuss, accessed August 16, 2010.
[22] Intel 64 and IA-32 Architectures Software Developer’s Manual, Vol. 3B: System Programming Guide, Part 2.

[23] Intel 64 and IA-32 Architectures Software Developer's Manual, Vol. 3A: System Programming Guide, Part 1.
[24] Ernst MD, Perkins JH, Guo PJ, McCamant S, Pacheco C, Tschantz MS, Xiao C (2007) The Daikon system for dynamic detection of likely invariants. In Science of Computer Programming, 2007.
[25] Kil C, Sezer E, Azab A, Ning P, Zhang X (2009) Remote attestation to dynamic system properties: Towards providing complete system integrity evidence. Proceedings of the 39th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'09), Lisbon, Portugal.
[26] Jaeger T, Sailer R, Shankar U (2006) PRIMA: policy-reduced integrity measurement architecture. Proceedings of the 11th ACM Symposium on Access Control Models and Technologies (SACMAT 2006).
[27] chkrootkit. http://www.chkrootkit.org/, accessed January 28, 2012.
[28] Arbaugh WA, Farber DJ, Smith JM (1997) A secure and reliable bootstrap architecture. Proceedings of the 1997 IEEE Symposium on Security and Privacy. IEEE Computer Society, Washington, DC.
[29] Singaravelu L, Pu C, Haertig H, Helmuth C (2006) Reducing TCB complexity for security-sensitive applications: three case studies. 1st ACM SIGOPS/EuroSys European Conference on Computer Systems, Leuven, Belgium.

Recommended Reading

[1] Armbrust M, Fox A, Griffith R, Joseph AD, and et al. (2009) Above the clouds: A Berkeley view of cloud computing. Technical Report UCB/EECS-2009-28, 2009. Available at http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html
[2] Brown A and Chase J (2011) Trusted Platform-as-a-Service: A Foundation for Trustworthy Cloud-Hosted Applications. ACM Cloud Computing Security Workshop, October 2011.
[3] Haeberlen A (2010) A case for the accountable cloud. ACM SIGOPS Operating Systems Review, Volume 44 Issue 2, April 2010.
[4] Hoglund G, Butler J (2005) Rootkits: subverting the Windows kernel. Addison-Wesley Professional, 2005.

Key Definitions

- **Integrity**: trustworthiness of data or resources, usually phrased in terms of preventing improper or unauthorized change.
- **Integrity modeling**: the process of specifying the expected properties of a system in order to detect improper change.
- **Scoped invariant**: the property that a certain object has a known-good value between two system events.
- **Invariant dependency graph**: a graph that concisely represents the dependency relationships among scoped invariants.
- **Invariants detection**: the process of deriving scoped invariant specifications from a program.

- **Remote attestation**: a trusted computing technique that enables a computer system in a networked environment to decide whether a target computer has integrity, e.g., whether it has the appropriate configuration and hardware/software stack, so it can be trusted.
- **Emulation**: the act of using hardware and/or software to duplicate the functions of a first computer system in a different second computer system, so that the behavior of the second system closely resembles the behavior of the first system.
- **Trusted computing**: technologies and proposals for resolving computer security problems through hardware enhancements (such as Trusted Platform Modules) and associated software modifications.