CHAPTER 2

The Trusted Cloud: Addressing Security and Compliance

In Chapter 1 we reviewed the essential cloud concepts and took a first look at cloud security. We noted that the traditional notion of perimeter or endpoint protection left much to be desired in the traditional architecture with enterprise-owned assets. Such a notion is even less adequate today when we add the challenges that application developers, service providers, application architects, data center operators, and users face in the emerging cloud environment.

In this chapter we'll bring the level of discourse one notch tighter and focus on defining the issues that drive cloud security. We'll go through a set of initial considerations and common definitions as prescribed by industry standards. We'll also look at current pain points in the industry regarding security and the challenges involved in addressing those pains.

Beyond these considerations, we first take a look at the solution space: the concept of a trusted infrastructure and usages to be implemented in a trusted cloud, starting with a trust chain that consists of hardware that supports boot integrity. Then, we take advantage of that trust chain to implement data protection, equally at rest and in motion and during application execution, to support application run-time integrity and offer protection in the top layer.

Finally, we look briefly at some of the "to be" scenarios for users who are able to put these recommendations into practice.

Security Considerations for the Cloud

One of the biggest barriers to broader adoption of cloud computing is security—the real and perceived risks of providing, accessing, and controlling services in a multi-tenant cloud environment. IT managers would like to see higher levels of assurance before they can declare their cloud-based services and data ready for prime time, similar to the level of trust they have in corporate-owned infrastructure. Organizations require their compute platforms to be secure and compliant with relevant rules, regulations, and laws. These

requirements must be met, whether deployment uses a dedicated service available via a private cloud or is a service shared with other subscribers via a public cloud. There's no margin for error when it comes to security. According to a research study conducted by the Ponemon Institute and Symantec, the average cost to an organization of a data breach in 2013 was $5.4 million, and the corresponding cost of lost business came to about $3 million.[1] It is the high cost of such data breaches and the inadequate security monitoring capabilities offered as part of the cloud services that pose the greatest threats to wider adoption of cloud computing and that create resistance within organizations to public cloud services.

From an IT manager's perspective, cloud computing architectures bypass or work against traditional security tools and frameworks. The ease with which services are migrated and deployed in a cloud environment brings significant benefits, but they are a bane from a compliance and security perspective. Therefore, this chapter focuses on the security challenges involved in deploying and managing services in a cloud infrastructure. To serve as an example, we describe work that Intel is doing with partners and the software vendor ecosystem to enable a security-enhanced platform and solutions with security anchored and rooted in hardware and firmware. The goal of this effort is to increase security visibility and control in the cloud.

Cloud computing describes the pooling of an on-demand, self-managed virtual infrastructure, consumed as a service. This approach abstracts applications from the complexity of the underlying infrastructure, allowing IT to focus on enabling greater business value and innovation instead of getting bogged down by technology deployment details. Organizations welcome the presumed cost savings and business flexibility associated with cloud deployments. However, IT practitioners unanimously cite security, control, and IT compliance as primary issues that slow the adoption of cloud computing. These considerations often denote general concerns about privacy, trust, change management, configuration management, access controls, auditing, and logging. Many customers also have specific security requirements that mandate control over data location, isolation, and integrity. These requirements have traditionally been met through a fixed hardware infrastructure.

At the current state of cloud computing, the means to verify a service's compliance are labor-intensive, inconsistent, non-scalable, or just plain impractical to implement. The necessary data, APIs, and tools are not available from the provider. Process mismatches occur when service providers and consumers work under different operating models. For these reasons, many corporations deploy less critical applications in the public cloud and restrict their sensitive applications to dedicated hardware and traditional IT architecture running in a corporate-owned vertical infrastructure. For business-critical applications and processes, and for sensitive data, third-party attestations of security controls usually aren't enough. In such cases, it is absolutely critical for organizations to be able to ascertain that the underlying cloud infrastructure is secure enough for the intended use.

[1] https://www4.symantec.com/mktginfo/whitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-Report_daiNA_cta72382.pdf

This requirement thus drives the next frontier of cloud security and compliance: implementing a level of transparency at the lowest layers of the cloud, through the development of standards, instrumentation, tools, and linkages to monitor and prove that the IaaS cloud's physical and virtual servers are actually performing as they should be and that they meet defined security criteria. The expectation is that the security of a cloud service should match or exceed the equivalent in house capabilities before it can be considered an appropriate replacement.

Today, security mechanisms in the lower stack layers (for example, hardware, firmware, and hypervisors) are almost absent. The demand for security is higher for externally sourced services. In particular, the requirements for transparency are higher: while certain monitoring and logging capabilities might not have been deemed necessary for an in-house component, they become absolute necessities when sourced from third parties to support operations, meet SLA compliance, and have audit trails should litigation and forensics become necessary. On the positive side, the use of cloud services will likely drive the re-architecturing of crusty applications with much higher levels of transparency and scalability with, we hope, moderate cost impact due to the greater efficiency the cloud brings.

Cloud providers and the IT community are working earnestly to address these requirements, allowing cloud services to be deployed and managed with predictable outcomes, with controls and policies in place to monitor trust and compliance of these services in cloud infrastructures. Specifically, Intel Corporation and other technology companies have come together to enable a highly secure cloud infrastructure based on a hardware root of trust, providing tamper-proof measurements of physical and virtual components in the computing stack, including hypervisors. These collaborations are working to develop a framework that integrates the secure hardware measurements provided by the hardware root of trust with adjoining virtualization and cloud management software. The intent is to improve visibility, control, and compliance for cloud services. For example, making the trust and integrity of the cloud servers visible will allow cloud orchestrators to provide improved controls of on boarding services for their more sensitive workloads, offering more secure hardware and subsequently better control over the migration of workloads and greater ability to deliver on security policies.

Security requirements for cloud use are still works in progress, let alone firming up the security aspects proper. Let's look at some of the security issues being captured, defined, and specified by the government and standards organizations.

Cloud Security, Trust, and Assurance

There is significant focus on and activity across various standards organizations and forums to define the challenges facing cloud security, as well as solutions to those challenges. The Cloud Security Alliance (CSA), NIST, and the Open Cloud Computing Interface (OCCI) are examples of organizations promoting cloud security standards. The Open Data Center Alliance (ODCA), an alliance of customers, recognizes that security is the biggest challenge organizations face as they plan for migration to cloud services. The ODCA is developing usage models that provide standardized definitions for security in the cloud services and detailed procedures for service providers to demonstrate compliance with those standards. These attempts seek to give organizations an ability to validate adherence to security standards within the cloud services.

Here are some important considerations dominating the current work on cloud security:

- **Visibility, compliance, and monitoring**. Ways are needed to provide seamless access to security controls, conditions, and operating states within a cloud's virtualization and hardware layers for auditability and at the bottom-most infrastructure layers of the cloud security providers. The measured evidence enables organizations to comply with security policies and with regulated data standards and controls such as FISMA and DPA (NIST 2005).
- **Data discovery and protection**. Cloud computing places data in new and different places—not just user data but also application and VM data (source). Key issues include data location and segregation, data footprints, backup, and recovery.
- **Architecture**. Standardized infrastructure and applications provide opportunities to exploit a single vulnerability many times over. This is the BORE (Break Once, Run Everywhere) principle at work. Considerations for the architecture include:
  - *Protection*. Protecting against attacks with standardized infrastructure when the same vulnerability can exist at many places, owing to the standardization.
  - *Support for multi-tenant environments*. Ensuring that systems and applications from different tenants are isolated from one another appropriately.
  - *Security policies*. Making sure that security policies are accurately and fully implemented across cloud architectures.
- **Identity management**. Identity management (IdM) is described as "the management of individual identities, their authentication, authorization, roles, and privileges/permissions within or across system and enterprise boundaries, with the goal of increasing security and productivity while decreasing cost, downtime, and repetitive tasks." From a cloud security perspective, questions like, "How do you control passwords and access tokens in the cloud?" and "How do you federate identity in the cloud?" are very real, thorny questions for cloud providers and subscribers.
- **Automation and policy orchestration**. The efficiency, scale, flexibility, and cost-effectiveness that cloud computing brings are because of the automation—the ability to rapidly deploy resources, and to scale up and scale down with processes, applications, and services provisioned securely "on demand." A high degree of automation and policy evaluation and orchestration are required so that security controls and protections are handled correctly, with minimal errors and minimal intervention needed.

Trends Affecting Data Center Security

The industry working groups that are addressing the issues identified above are carrying on their activities with some degree of urgency, driven as they are by a number of circumstances and events. There are three overriding security considerations applicable to data centers, namely:

- New types of attacks
- Changes in IT systems architecture as a transformation to the cloud environment takes place
- Increased governmental and international compliance requirements because of the exploits

The nature and types of attacks on information systems are changing dramatically. That is, the threat landscape is changing. Attackers are evolving from being hackers working on their own and looking for personal fame into organized, sophisticated attackers targeting specific types of data and seeking to gain and retain control of assets. These attacks are concerted, stealthy, and organized. The attacks have predominantly targeted operating systems and application environments, but new attacks are no longer confined to software and operating systems. Increasingly, they are moving lower down in the solution stacks to the platform, and they are affecting entities such as the BIOS, various firmware sites in the platform, and the hypervisor running on the bare-metal system. The attackers find it is easy to hide there, and the number of controls at that level is still minimal, so leverage is significant. Imagine, in a multi-tenant cloud environment, what impact malware can have if it gets control of a hypervisor.

Similarly, the evolving IT architecture is creating new security challenges. Risks exist anywhere there are connected systems. It does not help that servers, whether in a traditional data center or in a cloud implementation, were designed to be connected systems. Today, there is an undeniable trend toward virtualization, outsourcing, and cross-business and cross-supply chain collaboration, which blurs the boundaries between data “inside” an organization and data “outside” that organization. Drawing perimeters around these abstract and dynamic models is quite a challenge, and that may not even be practical anymore. The traditional perimeter-defined models aren’t as effective as they once were. Perhaps they never were, but the cloud brings these issues to the point they can’t be ignored anymore. The power of that cloud computing and virtualization lies in the abstraction, whereby workloads can migrate for efficiency, reliability, and optimization.

This fungibility of infrastructure, therefore, compounds the security and compliance problems. A vertically owned infrastructure at least provided the possibility of running critical applications with high security and with successfully meeting compliance requirements. But this view becomes unfeasible in a multi-tenant environment. With the loss of visibility comes the question of how to verify the integrity of the infrastructure on which an organization’s workloads are instantiated and run.

Adding to the burden of securing more data in these abstract models is a growing legal or regulatory compliance demand to secure personally identifiable data, intellectual

property, or financial data. The risks (and costs) of non-compliance continue to grow. The Federal Information Security Management Act (FISMA) and the Federal Risk and Authorization Management Program (FedRAMP) are two examples of how non-compliance prevents the cloud service providers from competing in the public sector. But even if cloud providers aren't planning to compete in the public sector by offering government agencies their cloud services, it's still important that they have at least a basic understanding of both programs. That's because the federal government is the largest single producer, collector, consumer, and disseminator of information in the United States. Any changes in regulatory requirements that affect government agencies will also have the potential of significantly affecting the commercial sector. These trends have major bearing on the security and compliance challenges that organizations face as they consider migrating their workloads to the cloud.

As mentioned, corporate-owned infrastructure can presumably provide a security advantage by virtue of its being inside the enterprise perimeter. The first defense is security by obscurity. Resources inside the enterprise, especially inside a physical perimeter, are difficult for intruders to reach. The second defense is genetic diversity. Given that IT processes vary from company to company, an action that breaches one company's security may not work for another company's. However, these presumed advantages are unintended, and therefore difficult to quantify; in practice, they offer little comfort or utility.

Security and Compliance Challenges

The four basic security and compliance challenges that organizations face are as follows:

- **Governance**. Cloud computing abstracts the infrastructure, and in order to prove compliance and satisfy audit requirements, organizations rely on the cloud providers to supply logs, reports, and attestation. When companies outsource parts of their IT infrastructure to cloud providers, they effectively give up some control of their information infrastructure and processes, even as they are required to bear greater responsibility for data confidentiality and compliance. While enterprises still get to define how their information is handled, who gets access to that information, and under what conditions in their private or hybrid clouds, they must largely take cloud providers at their word that their SLA trusting security policies and conditions are being met. Even then, service customers may have to compromise to have the capabilities that cloud providers can deliver. The organization's ability to monitor actual activities and verify security conditions within the cloud is usually very limited, and there are no standards or commercial tools to validate conformance to policies and SLAs.

- **Co-Tenancy and Noisy or Adversarial Neighbors.** Cloud computing introduces new risks resulting from multi-tenancy, an environment in which different users within a cloud share physical resources to run their virtual machines. Creating secure partitions between co-residental virtual machines has proved challenging for many cloud providers. Results range from the unintentional, noisy-neighbor syndrome whereby workloads that consume more than their fair share of compute, storage, or I/O resources starve the other virtual tenants on that host; to the deliberately malicious efforts, such as when malware is injected into the virtualization layer, enabling hostile parties to monitor and control any of the virtual machines residing on the system. To test this idea, researchers at UCSD and MIT were able to pinpoint the physical server used by programs running on the EC2 cloud, and then extract small amounts of data from these programs by inserting their own software and launching a side-channel attack.[2]

- **Architecture and Applications.** Cloud services are typically virtualized, which adds a hypervisor layer to a traditional IT application stack. This new layer introduces opportunities for improvements in security and compliance, but it also creates new attack surfaces and different risk exposure. Organizations must evaluate the new monitoring opportunities and the risks presented by the hypervisor layer, and account for them in their policy definition and compliance reporting.

- **Data.** Cloud services raise access and protection issues for user data and applications, including source code. Who has access, and what is left behind when an organization scales down a service? How is corporate confidential data protected from the virtual infrastructure administrators and cloud co-tenants? Encryption of data, at rest, in transit, and eventually in use, becomes a basic requirement, yet it comes with a performance cost (penalty). If we truly want to encrypt everywhere, how is it done in a cost-effective and efficient manner? Finally, data destruction at end of life is a subject not often discussed. There are clear regulations on how long data has to be retained. The assumption is that data gets destroyed or disposed of once the retention period expires. Examples of these regulations include Sarbanes-Oxley Act (SOX), Section 802: seven years (U.S. Security and Exchange Commission 2003); HIPAA, 45 C.F.R. §164.530(j): six years; and FACTA Disposal Rule (Federal Trade Commission 2005).

[2]S. Curry, J. Darbyshire, Douglas Fisher, et al., *RSA Security Brief*, March 2010. Also, T. Ristenpart, E. Tromer, et al., *Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds*, CCS'09, Chicago.

With many organizations using cloud services today for non-mission-critical operations or for low-confidentiality applications, security and compliance challenges seem manageable, but this is a policy of avoidance. These services don't deal with data and applications governed by strict information security policies such as health regulations, FISMA regulations, and the Data Protection Act in Europe. But the security and compliance challenges mentioned above would become central to cloud providers and subscribers once these higher-value business functions and data begin migrating to private cloud and hybrid clouds. Industry pundits believe that the cloud value proposition will increasingly drive the migration of these higher value applications, as well as information and business processes, to cloud infrastructures. As more and more sensitive data and business-critical processes move to these cloud environments, the implications for security officers in these organizations will be to provide a transparent and compliant framework for information security, with monitoring.

So how do IT people address these challenges and requirements? With the concept of *trusted clouds*. This answer addresses many of these challenges and provides the ability for organizations to migrate both regular and mission-critical applications so as to leverage the benefits of cloud computing.

Trusted Clouds

There are many definitions and industry descriptions for the term *trusted cloud*, but at the core these definitions all have four foundational pillars:

- A trusted computing infrastructure
- A trusted cloud identity and access management
- Trusted software and applications
- Operations and risk management

Each of these pillars is broad and goes deep, with a rich cohort of technologies, patterns of development, and of course security considerations. It is not possible to cover all of them in one book. Since this book deals with the infrastructure for cloud security, we focus on the first pillar, the trusted infrastructure, and leave the others for future work. (Identity and access management are covered very briefly within the context of the trusted infrastructure.) But before we delve into this subject, let's review some key security concepts to ensure clarity in the discussion. These terms lay the foundation for what visibility, compliance, and monitoring entail, and we start with baseline definitions for *trust* and *assurance*.

- **Trust**. The assurance and confidence that people, data, entities, information, and processes will function or behave in expected ways. Trust may be human-to-human, machine-to-machine (e.g., handshake protocols negotiated within certain protocols), human-to-machine (e.g., when a consumer reviews a digital signature advisory notice on a website), or machine-to-human. At a deeper level, trust might be regarded as a consequence of progress toward achieving security or privacy objectives.

- **Assurance**. Evidence or grounds for confidence that the security controls implemented within an information system are effective in their application. Assurance can be shown in:
    - Actions taken by developers, implementers, and operators in the specification, design, development, implementation, operation, and maintenance of security controls.
    - Actions taken by security control assessors to determine the extent to which those controls are implemented correctly, operating as intended, and producing the desired outcomes with respect to meeting the security requirements for the system.

With these definitions established, let's now take a look at the *trusted computing infrastructure,* where computing infrastructure embraces three domains: compute, storage, and network.

Trusted Computing Infrastructure

Trusted computing infrastructure systems consistently behave in expected ways, with hardware and software working together to enforce these behaviors. The behaviors are consistent across compute on servers, storage, and network elements in the data center.

In the traditional infrastructure, hardware is a bystander to security measures, as most of the malware prevention, detection, and remediation is handled by software in the operating system, applications, or services layers. This approach is no longer adequate, however, as software layers have become more easily circumvented or corrupted. To deliver on the promise of trusted clouds, a better approach is the creation of a *root of trust* at the most foundational layer of a system—that is, in the hardware. Then, that root of trust grows upward, into and through the operating system, applications, and services layers. This new security approach is known as *hardware-based* or *hardware-assisted* security, and it becomes the basis for enabling the trusted clouds.

Trusted computing relies on cryptographic and measurement techniques to enforce a selected behavior by authenticating the launch and authorizing processes. This authentication allows an entity to verify that only authorized code runs on a system. Though this typically covers initial booting, it may also include applications and scripts. Establishing trust for a particular component implies also an ability to establish trust for that component relative to other trusted components. This transitive trust path is known as the *chain of trust*, with the initial component being the root of trust.

A system of geometry is built on a set of postulates assumed to be true. Likewise, a trusted computing infrastructure starts with a root of trust that contains a set of trusted elemental functions assumed to be immune from physical and other attacks. Since an important requirement for trust is that conditions be tamper-proof, cryptography or some immutable unique signature is used to identify a component. The hardware platform is usually a good proxy for the root of trust; for most attackers, the risk, cost, and difficulty of tampering with hardware exceeds the potential benefits of attempting to do so.

With the use of hardware as the initial root of trust, you can then measure (which means taking a hash, like an MD5 or SHA1, of the image of component or components) the software, such as the hypervisor or operating system, to determine whether unauthorized modifications have been made to it. In this way, a chain of trust relative to the hardware can be established. Trust techniques include hardware encryption, signing, machine authentication, secure key storage, and attestation. Encryption and signing are well-known techniques, but these are hardened by the placement of keys in protected hardware storage. Machine authentication provides a user with a higher level of assurance, as the machine is indicated as known and authenticated. Attestation, which is covered in Chapter 4, provides the means for a third party (also called a trusted third party) to affirm that loaded firmware and software are correct, true, or genuine. This is particularly important for cloud architectures based on virtualization.

Trusted Cloud Usage Models

In this abstracted and fungible cloud environment, the focus needs to be on enabling security across the three infrastructure domains. Only then can an enterprise have an infrastructure that is trusted to enable the broad migration of critical applications. Mitigating risk becomes more complex, as cloud use introduces an ever-expanding, transient chain of custody for sensitive data and applications. Only when security is addressed in a transparent and auditable way can enterprises and developers have:

- Confidence that their applications and workloads are equally safe in multi-tenant clouds
- Greater visibility and control of the operational state of the infrastructure, to balance the loss of physical control that comes with this abstracted environment
- Capability to continuously monitor for compliance

Cloud consumers may not articulate the needs in this fashion. From their perspective, there are key mega-needs, such as:

- How can I trust the cloud enough to use it?
- How can I protect my application and workloads in the cloud—and from the cloud?
- How can I broker between device and cloud services to ensure trust and security?

A cloud provider has to address these questions in a meaningful way for its tenants. These needs translate into a set of *foundational usage models for trusted clouds* that apply across the three infrastructure domains, as shown in Figure 2-1.

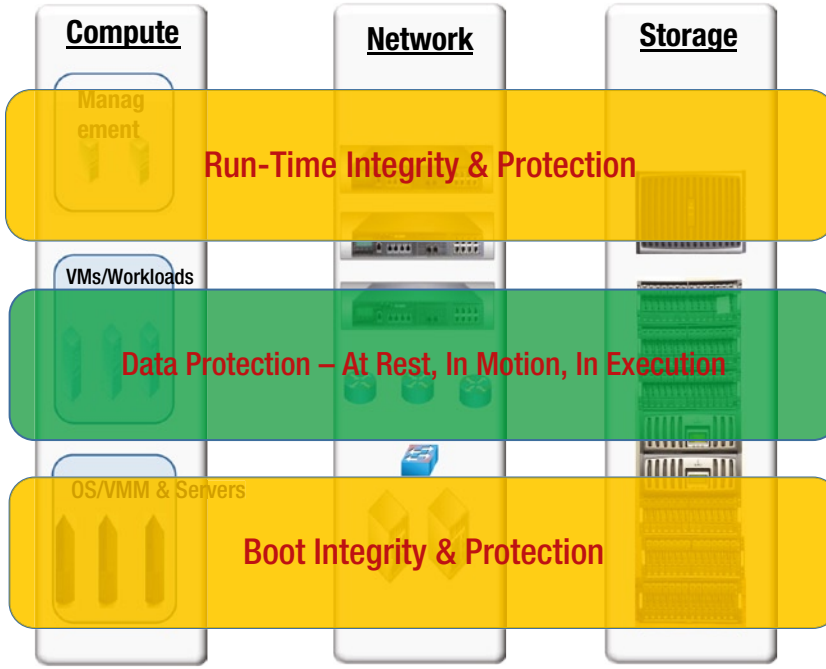

Figure 2-1. *A framework for the trusted cloud*

1. Boot integrity and protection
2. Data governance and protection, at rest, in motion, and during execution
3. Run-time integrity and protection

The scope and semantics of these usage models changes across the three infrastructure domains, but the purpose and intent are the same. How they manifest and are implemented in each of the domains could differ. For example, data protection in the context of the compute domain entails protection (both confidentiality and integrity) of the virtual machines at rest, in motion, and during execution; this applies to their configuration, state, secrets, keys, certificates, and other entities stored within. The same data-protection usage for the network domain has a different focus; it is on protection of the network flows, network isolation, confidentiality on the pipe, tenant-specific IPS, IDS, firewalls, deep packet inspection, and so on. In the storage domain, data protection pinpoints strong isolation/segregation, confidentiality, sovereignty, and integrity. Data confidentiality, which is a key part of data protection across the three domains, uses the same technological components and solutions—that is, encryption.

As a solution provider, methodical development and instantiation of these usage models across all the domains will provide the necessary assurance for organizations migrating their critical applications to a cloud infrastructure, and will enable establishment of the foundational pillar for trusted clouds.

In the rest of this chapter, we provide an exposition of the usage models listed above. We include enough definition of these four usage models for them to provide a broad overview. Subsequent chapters go into greater detail on each of these models and offer solutions, including the solution architecture and a reference implementation using commercial software and management components.

The Boot Integrity Usage Model

Boot integrity represents the first step toward achieving a trusted infrastructure. This model applies equally well to the compute, network, and storage domains. As illustrated in Figure 2-1, every network switch, router, or storage controller (in a SAN or NAS) runs a compute layer operating specialized OS to provide networking and storage functions, so this model enables a service provider to make claims about the boot integrity of the network, storage, and compute platforms, as well as the operating system and hypervisor instances running in them. As discussed earlier, boot integrity supported in the hardware makes the system robust and less vulnerable to tampering and targeted attacks. It enables an infrastructure service provider to make quantifiable claims about the boot-time integrity of the pre-launch and the launch components. This provides a means, therefore, to observe and measure the integrity of the infrastructure. In a cloud infrastructure, these security features refer to the virtualization technology in use, which comprises two layers:

- The boot integrity of the BIOS, firmware, and hypervisor. We identify this capability as *trusted platform boot.*
- The boot integrity of the virtual machines that host the workloads and applications. We want these applications to run on *trusted virtual machines.*

Understanding the Value of Platform Boot Integrity

To attain trusted computing, cloud users need systems hardened against emerging threats such as rootkits. Historically, many have viewed these threats as someone else's problem or as a purely hypothetical issue. This position is untenable in view of today's threats.

The stealthy, low-level threats are real and they occur in actual operating environments. The recent Mebromi BIOS rootkit low-level attack on a shipping platform was an eye-opener, as it took the industry by surprise. Unfortunately, as is often the case, it takes an actual exploit to change the mindset and drive change. And indeed, there are many more IT managers and security professionals taking action to improve the situation. As of 2012, a growing number of entities, including the U.S. National Institute of Standards and Technologies (NIST), are developing recommendations for protecting a system's boot integrity. These recommendations contain measures for securing very basic, but highly privileged platform components.

Given the crucial role played by the hypervisor as essential software responsible for managing the underlying hardware and allocating resources such as processor, disk, memory, and I/O to the guest virtual machines and arbitrating the accesses and privileges among guests, it is imperative to have the highest levels of assurance so that it is uncompromised. This was the rationale for conducting the survey shown in Figure 2-2. With this growing awareness and concern has come a corresponding growth in vendors looking to define the solutions.

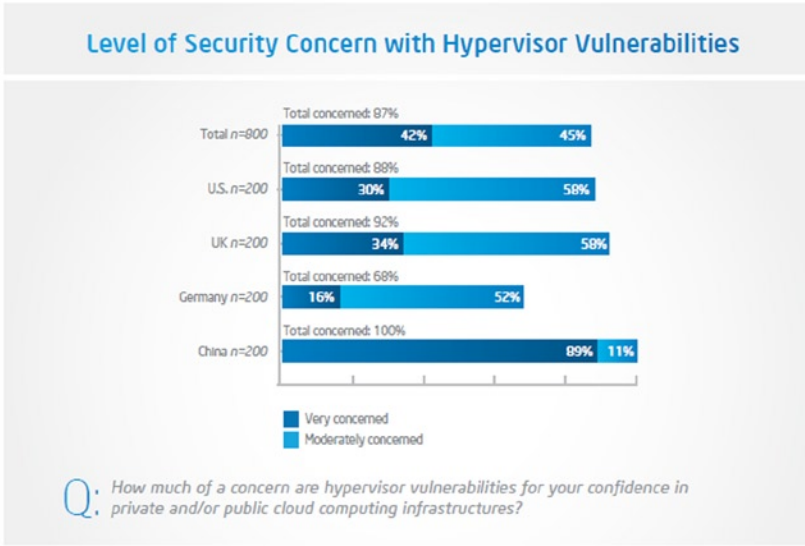

Figure 2-2. *Survey results showing concerns over hypervisor integrity across regions*

For the various devices/nodes across the infrastructure domains (compute, storage, and network), the integrity of the pre-launch and launch environment can be asserted anytime during the execution's lifecycle. This is done by verifying that the identity and values of the components have not changed unless there has been a reset or a reboot of the platform by the controlling software. This assertion of integrity is deferred to a trusted third party that fulfills the role of a trust authority, and the verification process is known as *trust attestation*. The trust authority service is an essential component of a trusted cloud solution architecture.

The Trusted Virtual Machine Launch Usage Model

A trusted platform boot capability provides a safe launch environment for provisioning virtual machines running workloads. This environment has the mechanisms to evaluate the integrity of pre-launch and launch components on a platform, from the BIOS to the operating system and hypervisor. The service provider thus attests to the trust-ability

of the launch environment. However, no specific claims can be made about the virtual machines being launched, other than indicating that they are being launched on a measured and attested hypervisor platform. Although virtual machine monitors (VMM) or hypervisors are naturally good at isolating workloads from each other because they mediate all access to physical resources by virtual machines, they cannot by themselves attest and assert the state of the virtual machine that is launched.

The trusted virtual machine launch usage model applies the same level of trustability to the pre-launch and launch environment of the virtual machines and workloads. Each virtual machine launched on a virtual machine manager and hypervisor platform benefits from a hardware root of trust by storing the launch measurements of the virtual machines' sealing and remote attestation capabilities. However, this requires virtualizing the TPM, with a virtual TPM (vTPM) for each of the virtual machines. Each of these virtual TPM vTPM instances then emulates the functions of a hardware TPM. Currently, there are no real virtualized TPM implementations available, owing to the challenges related to virtualizing the TPM. The difficulty lies not in providing the low-level TPM instructions but in ensuring that the security properties are supported and established with an appropriate level of trust. Specifically, we have to extend the chain of trust from the physical TPM to each virtual TPM by carefully managing the signing keys, certificates, and lifecycle of all necessary elements. An added dimension is the mobility of the virtual machines and how these virtual TPMs would migrate with the virtual machines.

There are other ways of enabling a measured launch of virtual machines, such as storing the measurements in memory as part of a trusted hypervisor platform without the use of virtual TPMs but still ensuring that the chain of trust is extended from the physical TPM. Irrespective of the design approach, day-to-day operations on virtual machines—such as suspend and resume, creating snapshots of running virtual machines, and playing them back on other platforms or live migration of virtual machines—become challenging to implement.

There are no real production-quality implementations of these architectures. There are few academic and research implementations of vTPMs and other memory structure–based approaches, each with its own pros and cons. Trusted virtual machine usages are still evolving at the time of this writing; hence it's not possible to be definitive. Chapter 8 covers aspects of the measured VM launch and some architectural elements. Chapter 3 covers in depth the matter of boot integrity and trusted boot of platforms and the hypervisors, as well as the associated trusted compute pools concept that aggregates systems so specific policies can be applied to those pools. The discussion also includes the solution architecture, and a snapshot of industry efforts to support the enabling of trusted compute pools. Chapter 4 covers the trust attestation or remote attestation architecture, including a reference implementation.

The Data Protection Usage Model

This usage model is about protecting data in the cloud that is at rest, in motion, and undergoing execution. It applies uniformly across infrastructure domains (compute, storage, and network). On the compute domain, the protection is for the virtual machines and workloads that have the applications, configurations, state, keys, secrets, and needed mechanisms to ensure confidentiality and integrity.

For virtual machine and workload data protection, cloud user organizations need a method to securely place workloads into the cloud, as well as store and use data there. Current provisioning and bursting models include either storing the virtual machine and application images and data in the clear (unencrypted), or having these images and data encrypted by the keys controlled by the service provider—keys which are likely applied uniformly to all the tenants. But increasingly, virtual machine images—effectively, containers for operating system and application images, configuration files, data, and other entities—need confidentiality protection in a multi-tenant cloud environment. That is, images need to be encrypted by keys under tenant control, and also decrypted for provisioning by the keys under tenant control in a manner that is transparent to the cloud service provider. The usage model also calls for not only leveraging hardware for encryption and decryption but also ensuring that the service or entity acquiring the decryption keys does it on a need-to-know basis, is trusted and attested, and is running on a platform whose boot integrity has been attested. This provides a more effective last line of defense to protect from misuse or abuse by other tenants or cloud administrators. Chapter 8 covers this usage model for virtual machine protection, including a reference architecture and implementation.

The Run-time Integrity and Attestation Usage Model

Having a trusted foundation for the platform is extremely important. Roots of trust in hardware, and with a credible static and binary remote attestation process, ensure that a service provider can make assertions about the boot integrity of the platforms on which the tenant workloads execute. But that is only half the answer. The integrity of the platform could be assured at boot time, and remote attestation can measure and attest the state of healthiness at that point—only for integrity to be degraded and compromised at run time for a variety of reasons, such as configuration errors or, worse, the presence of run-time rootkits. These mechanisms compromise the integrity of the platforms and yet static binary remote attestation doesn't catch them; instead, this situation calls for remote run-time attestation. However, for this solution to be viable, there needs to be a way of representing and approximating the run-time integrity of the system via a set of policies or properties. A system or platform stays healthy only to the extent that these properties stay healthy.

Determining what constitutes the minimum and sufficient set of properties that indicate the run-time health of a hypervisor or virtual machine monitor is a tough computer science problem that has long track record of research in software integrity. For example, if the integrity properties cover the system call table—the call table being the basis for measurement, monitoring, and attestation—a new rootkit can be deployed that manipulates other function pointers, such as device driver jump tables, and it will stay undetected. Clearly, there are no commercial implementations, since the threat vectors are too many to consider and modeling the threats, as well as mitigation, is still an active research area.

One promising research effort has been to define what are called "scoped invariants" as an important class of integrity properties. According to the authors of this research, scoped invariants are code or data with a constant value in some context (*scope*). For

example, one scoped invariant is the Interrupt Descriptor Table (IDT) entry for page fault, containing a constant function pointer once the virtual machine monitor or operating system finishes initialization. Scoped invariants are building blocks for more general integrity properties, and they are amenable to integrity checking. A case study was done to identify a core set of scoped invariants of the open-source Xen virtual machine monitor. In addition to the IDT, another core invariant property was demonstrated in this research; the addressable memory limit of a guest OS must not include Xen's code and data, and this proved indispensable for Xen's guest isolation mechanism. Violation of this property can let an attacker modify a single byte in the Global Descriptor Table (GDT) to achieve a virtual machine escape goal.

At the current state of the art, run-time integrity monitoring and attestation is a broad and complex topic, and commercial implementations are still works in progress at many system and security organizations.

Trusted Cloud Value Proposition for Cloud Tenants

While a tenant organization's compliance and security policies won't change when IT processes migrate to the cloud, the way that organization enforces those policies and proves compliance will change significantly. For most compliance officers and infosec (information security) professionals, the cloud becomes, for practical purposes, a black box. In contrast, a cloud tenant that is landing a workload in a trusted pool can expect the following:

- The assurance that the compute, network, and storage elements in that segment of the cloud or the virtualized data center are trusted. The service provider or the management infrastructure asserts the integrity of the security and trust of these elements.
- The assurance that the information (data and content) s stored, processed, and migrated is always protected for confidentiality, integrity, and privacy.
- The assurance that workloads and applications are not tampered with, and that the infrastructure will launch and execute what is expected, and can provide a chain of trust that is rooted in hardware.
- The assurance that the devices and users accessing the workloads and services in these trusted clouds are authenticated, and that the workloads run on hardware with demonstrated integrity; likewise, for the controlling software. This ensures that services are being accessed over a reliable and secure network and location.

The Advantages of Cloud Services on a Trusted Computing Chain

The advantages to delivering cloud services on computing resources that have a demonstrated chain of trust rooted in hardware include:

- *Reducing the risks for co-residency*. It ensures that the infrastructure is trusted and has demonstrated integrity. This prevents the launch and execution of untrusted components. It protects not only against malware but also from benign conditions, such as the improper migration or deployment of virtual machines. To illustrate, if a cloud orchestrator (like OpenStack) attempts to move virtual machines from an unsecured computing platform to a trusted one, the policy management software will prevent the incoming virtual machines from landing, since the action originated from an unsecured platform.
- *Preventing the unsafe transit of secure virtual machines.* In the same way that virtual machines arriving from an unsecured platform are not allowed to move to secured platforms, virtual machines originating on secured platforms are not allowed to move to unsecured ones. For instance, if an administrator attempted to transfer a secured virtual machine to a new server, the virtualization management console would first perform a policy check on the outgoing virtual machine and then measure the security configurations of the new server against accepted standards. If the new server does not meet the secure standards required to host the virtual machine, the virtualization management console or security policy engine prevents the virtual machine from migrating and logs the attempt.
- *Maximizing and scaling operational efficiency by creating trusted pools of systems.* Once platform trustworthiness can be measured, cloud providers can put such measurements to use by building trusted pools of systems, all with identical security profiles. Hypervisors can then make more efficient use of secure clouds, moving virtual machines with similar security profiles within zones of identically secured systems for load balancing and other administrative purposes—all the while protecting data in conformance with regulated standards and policies.

The authors believe that ubiquitous adoption of trusted computing chains will address a number of fundamental user concerns about cloud security that currently prevent many applications from being deployed in a cloud setting, thereby barring them from realizing the potential cost reductions that could stem from using cloud technology and limiting the greater business impact that would come from broader deployment.

Summary

We covered the challenges of cloud security and compliance, as well as introduced the concept of trusted clouds. We discussed the needs for trusted clouds and introduced four usage models to enable a trusted computing infrastructure, the foundation for trusted clouds. These models provide a foundation for enhanced security that can evolve with new technologies from Intel and others in the hardware and software ecosystem.

There are no silver bullets for security, such as a single technology solving all problems, because the matter of security is a multifaceted one. But it is clear that a new set of security capabilities is needed, and that starts at the most foundational elements. Trusted platforms provide such a foundation. These platforms can provide:

- Increased visibility of critical controlling software in the cloud environment through attestation capabilities.
- A new control point capable of identifying and enforcing local known good configurations of the host operating environment, and able to report the resultant launch trust status to cloud and security management software for subsequent use.

In the next few chapters we will discuss each of the usage models in detail, including some solution architectures and technologies to bring them to reality.