Using the Cloud: Keeping Enterprise Data Private

Kyle Cronin
Kyle.Cronin@dsu.edu
Instructor of Information Security

Wayne Pauli
Wayne.Pauli@dsu.edu
Director of Center of Excellence

Michael Ham
mjham@pluto.dsu.edu
MSIA Student

Dakota State University
Madison SD

Abstract

Cloud computing has overcome the computing industry within the past few years. Exciting prospects such as sharing resources, reducing costs, and increasing efficiency have made the cloud computing model highly attractive. In this paper, we will focus briefly on the privacy and security concerns of outsourcing the hosting of a virtual infrastructure, often referred to as Infrastructure as a Service. Also, we will analyze two different methods of encrypting data and the performance degradation that is caused by leveraging encryption in an effort to prevent a cloud provider from accessing your information. Then, we will compare the results of a simulated SQL server and have a basic conclusion of what method offers better performance, and a basic analysis of the degradation of performance caused by encrypting data in a particular cloud computing setting.

Keywords: cloud computing, protecting data, encryption, hypervisor, time based tradeoff, infrastructure as a service, software as a service

1. WHAT IS 'THE CLOUD'?

The personal and commercial worlds have engrossed themselves with the cloud over the past few years. However, the term Cloud Computing lacks a true meaning by which the key focus of "the cloud" is on. In essence, cloud computing includes any form of computing where information is stored, retrieved, and processed using a third party's computing platform. To differentiate from the various styles of cloud platforms ranging from Google Docs, to Amazon Elastic Compute Cloud (EC2), to Facebook, we will leverage the National Institute of Standards and Technology's (NIST) recognized definitions of cloud computing service and deployment models. NIST defines the following three different service models: Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS), and Cloud Infrastructure as a Service (IaaS). Our focus will be on IaaS (Mell and Grance 2009).

Cloud Software as a Service (SaaS)

One of the three service models associated with Cloud Computing is that of SaaS. According to Gartner (Hall 2011), sales in 2010 were to reach $9 billion, an increase of over 15% from 2009. By the end of 2011 sales should represent in excess of $10 billion, an increase of more than 16%. SaaS is sometimes referred to as on demand software, and utilizes a centrally located delivery model of software to the users by way of a web browser. The focus of SaaS is that of how this delivery is configured for user access, as it is not considered customizable by the user because source code is not available for such a task.

Cloud Platform as a Service (PaaS)

To address the customizable desires of information technology users, the PaaS model can be implemented. As its name indicates, the development platform is deployed through a central hub as opposed to the software applications of SaaS. This model facilitates the functionalities of application design, development, testing, and deployment of the system development life cycle and includes services such as collaboration of developers, the integration of databases, security and scalability among other services. The feasibility of customization allows for integrating many solutions in this model.

Cloud Infrastructure as a Service (IaaS)

The 3rd service model, and the focus of this study is that of IaaS. When referring to 'The Cloud' this study is referring to the IaaS model. Where SaaS addresses software use, and PaaS details the development platform functionality of the cloud, IaaS is in effect the network as a whole. It has been our determination that the service model IaaS is teamed with the deployment model of a Public Cloud, therefore privacy of a customers' data may be at risk. In this service/deployment combination, customers purchase hosted infrastructure from a provider and are therefore given the ability to manage operating systems, processing, and various other "fundamental computing resources" from the public cloud owner (Mell and Grance 2009). Common examples, as seen in industry today include Amazon's Elastic Compute Cloud (EC2), Terremark's Infinistructure, and Rackspace's Mosso Cloud Servers (Lenk, Klems et al. 2009). In these agreements, customers pay a fee in order for the cloud provider to host a virtualized copy of a particular operating system and virtual hardware set. Customers are then tasked with the management of the operating system, software, and data contained within the virtual platform.

Leveraging this form of virtualization, a provider company operates their own hardware. Each instance of a virtual host is essentially a physical machine running virtualization software that allows multiple guest machines (Goth 2007). A guest machine is a single instance of an IaaS model. When a guest machine is operating a portion of the resources of the host are allocated for the guest's processes. All in all, this is generally referred to as virtualization. Throughout the progression, development, and availability of virtualization technology, it has evolved into becoming a mainstream component of IT systems.

2. WHAT IS THE RISK?

In these public, shared environments, one customer's data is housed next to another customer's data; this has already been termed as a feature of the Public Cloud. A user or organization's potentially private data is stored in some form by a third party. Ultimately, the customer is in no way in control of how or where their data is stored in the cloud environment (Kaufman 2009). The level of security required by customers is highly dependent and is therefore tied directly with the value of the data. Customers storing private information (should) place a high price in terms of the level of confidentiality, integrity guarantees, and availability provided by the cloud provider.

With this lack of control over the confidentiality, integrity, and availability (CIA), the owner of the data is left being at a disadvantage, and is therefore taking a certain calculated risk (Olivier 2002). With these details in mind, we focus on our efforts for maintaining data confidentiality. By outsourcing data storage and processing to third party providers, customers are placing their data at risk in situations stemming from provider mistakes, disgruntled employees, physical infiltration, outside attackers, and other inherent risks.

These risk factors are taken into consideration for the purpose of this study. Ultimately, our purpose is to determine a method that balances

performance with protection of data. The performance cost (or loss of performance) will be evaluated with the levels of protection offered for level of cost.

3. ISSUES NEEDING SOLUTIONS

Three key issues, in an effort to reliably protect the data of IaaS customers, have been identified: maintaining data security, maintaining data access performance, and completing these actions in a way that still makes it financially viable for IaaS providers and customers alike.
Maintaining data access is an issue that customers need to negotiate with their provider. Aside from connection redundancies, the customer has little control over the access to their data.

To address the latter two issues, maintaining data security and managing the performance of the access, a hybrid model of control exists, as exhibited by Figure 1.

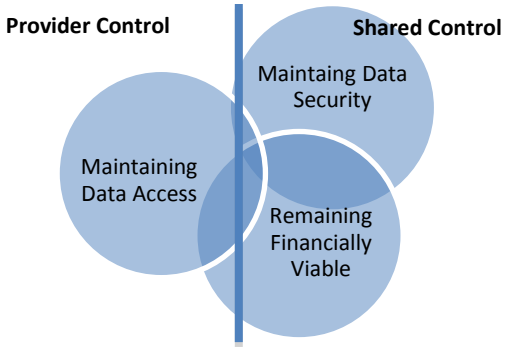

Figure 1 Model of shared responsibility

In IaaS situations, since the customer is in control over the host operating system, they have the ability to leverage the OS's ability to protect the data at the expense of performance. Several methods exist for protecting data when stored online. Traditionally the two choices have been storage level and database level encryption (Mattsson 2005). As shown by Mattsson's research, database level encryption has traditionally been a high performing method of encryption, although it requires modification to the database schema. On the other hand, storage level encryption tends to require no changes to database schema, however its flexibility limits the data encryption to an all or nothing outcome.

In the instance of a virtualized environment, a new subset variety of data protection exists, which is derived from storage level encryption-hypervisor and guest based encryption. Traditional forms of database encryption focus on encrypting specific files or datasets within an operating system. In virtualized environments, an additional layer exists between the guest operating system and the hardware, the hypervisor. With hypervisor-level encryption, storage-level encryption can be leveraged without the knowledge (or ability) of the guest operating system. Guest-based encryption performs the same actions, with the exception that the encryption only takes place on a single virtual machine or guest.

4. WHAT ARE WE PROTECTING AGAINST?

A clearly deliniated definition is required in order to see the purpose of the overall data protection scheme. There are certain risks involved when outsourcing data storage to a third party which all depend upon who has access to the data. Employees of the cloud provider generally have full access to customer data. While companywide policies and procedures can be put in place, customer data is still at risk against employees that choose to violate said policies. In essence, any individual that has physical access to the medium for data storage theoretically has access to customer data.

Encryption allows us to protect our information, making data appear as pseudorandom bits written to the storage medium. Two methods have been discussed previously, hypervisor or host-based encryption, and guest-based encryption. Host-based encryption involves the cloud provider encrypting the file system in which the virtual machines are stored. Guest-based encryption takes an alternative approach; the guest virtual machines themselves handle encryption.

In order to mitigate the risks of an inside attacker, such as a disgruntled employee, the type of encryption must be scrutinized. In the event that host-based encryption is used, it can be easily assumed that the cloud provider is in charge of the encryption keys. In the scenario of a rouge employee, it can be assumed that the employee would have access to the encryption keys and could have the ability to reverse encryption that is implemented at the host-level. In contrast, guest-based encryption puts the control of the encryption keys in the hands of

the customer. While an employee (or anyone with physical access) may gain the data files from the guest, the files will be encrypted. Assuming the customer properly protects their encryption keys, the attacker will be unable to decrypt the data, thus preventing loss of data to those with physical access.

5. COMPARISON OF METHODS

In the realm of system virtualization, several vendors exhibit scenarios for data security. In order to effectively compare encryption methodologies, a comparison must be made between the current virtualization platforms available, including vendor, features, CPU support, *supported* encryption methods, and supported guest operating systems. This comparison appears in Table 2 in the Appendices and Annexures section.
Several other virtualization platforms are available that include, but are not limited to: VirtualBox (an open source project from Oracle), Virtual Server 2005 (predecessor to Hyper-V by Microsoft), Virtual PC (desktop virtualization platform from Microsoft), and VMware Workstation (a desktop virtualization platform from VMware). These product lines are not within the scope of the comparison as they lack the support for true enterprise deployment.

The test is designed to compare the response times to SQL query simulations from a Windows-based virtual machine. Since enterprise database management systems (DBMS) are based upon SQL servers, the test aims to see what cost will be observed in performance. Our results will show what types of protections are more cost efficient in terms of performance. These results will ultimately allow individuals, businesses, and enterprise partners to derive their decisions for levels of protection verses what levels of performance may be lost.
Comparisons will be made between host-based encryption and guest-based encryption. The tests will be compared against a control, which uses no encryption at all. The overall purpose of the test is to compare the performance tradeoffs of the two discussed encryption methods. The hypervisor chosen for testing is the Hyper-V platform (on Windows Server 2008 R2 Enterprise) from Microsoft due to its ease of installation, versatile support for encryption, and large penetration within industry. The hardware used for the tests will be an HP DL380 G6 server with 2X Intel Xeon E5540 processors, 56 GB of RAM, using the Smart Array p410i storage controller with 2X300GB Dual Port SAS drives in a RAID 1 volume for OS installation, and 6X300GB Dual Port SAS drives in a RAID 5 volume for VM storage.

Three tests will be performed. The first test, as a control, will involve executing the queries against an unencrypted installation of Windows Server 2008 R2 using the SQLIOSim utility. SQLIOSim is a utility from Microsoft designed to simulate algorithms and IO patterns observed in Microsoft SQL Server. In this first test no encryption will be used, therefore making it the baseline for results comparison. This baseline test will then become the control group representing the theoretical performance assuming no protections are used to prevent unauthorized access to the enterprise data.

The second test will measure the data read and write times when the virtual machine implements the encryption. Finally, the third test will compare the performance of host-based encryption with an unencrypted guest. The difference between these two tests is where the encryption is implemented. In the second test, the virtual machine itself manages all encryption activities whereas in the third test the encryption takes place on the physical machine (host). The expected results should show the unencrypted machine, the control, having much higher IO patterns than the encrypted machines. Comparisons of the two encryption schemes will then be made to see if guest-based encryption is more or less efficient.

6. OBSERVED RESULTS

The results from SQLIOSim measured four data points relevant to our research: Reads, Scatter Reads, Writes, and Gather Writes. These data points are all methods of input and output that can be measured in any software system, particularly database systems. Reads and writes are simple operations- reading a block of data from some input, often a hard disk, into a memory buffer, or writing a block of data from a memory buffer to an output, again generally a hard disk. Scatter reads and gather writes are referred to as vectored I/O. Vectored I/O is a method of attaining enhanced efficiency during input and output of data within software. In these situations, a block of data is read from the disk into multiple buffers in memory (or written to the disk from multiple memory buffers). Scatter/gather refers to the element that buffers

that have data scattered into or gathered from within.

All four of these values are representative accumulators indicating what levels of performance would be expected in a production database environment with a heavy I/O load on a server's hard disk. For example, every time a basic read operation is completed, such as a query against a database, the Reads accumulator is incremented. For the purposes of comparison, higher values are an indication of higher levels of attained performance.

Upon the completion of the three test cases (ran at four iterations each), the averages of the results are shown in Table 1.

	Reads	Scatter Reads	Writes	Gather Writes
Control	69764	61010.5	3216	98116.25
Guest Based	34242	35066	1991.5	63525.25
Host Based	64930.5	58741.25	3134.75	96322

Table 1 Average IO From Tests Completed.

A significant drop in performance was observed when the guest-based encryption was utilized. Leveraging guest-based encryption under the presented conditions resulted in nearly a 50% drop in performance on average. Raw results appear in Table 3 located in the Appendices section.

Notably the results show that using guest based encryption methods caused an average loss in performance greater than 60%. Because of the nature of virtualization, we do expect there to be a lower level of performance (Goth 2007). All operations that require disk access in the guest based scenario require complex cryptographic calculations to be performed. Since in the guest based scenario, the guest is offered a share of CPU resources it is observed that this has a significant impact in the levels of performance achieved.

In the host-based scenario, the virtual machine is not concerned with making cryptographic calculations since it is handled by the host. The host, operating the hypervisor, has preferential treatment in using CPU power and is therefore able to attain significantly higher performance.

7. CONCLUSION

With the observed loss of performance, it can be concluded that guest-based encryption mechanisms have a significant detriment to the performance in situations where high performance is a requirement. In these situations, as it stands, leveraging third party IaaS solutions will continually pose security, privacy, and regulatory risks to both businesses and consumers hosting their data in the cloud. As discussed, host-based encryption mechanisms do provide a level of security, but the ability for compromise still exists due to the lack of control outside of the consumers' hands.

For future study, we propose a comparison of performance tests and measurement of the load imposed upon the host hardware, comparing the IO advantages of host-based encryption in greater detail. Additionally, studies should be conducted that include situations where multiple machines are active on a single host- some using guest-based encryption while others are running without any encryption mechanisms. The outcome of such studies would be to measure the potential impact on IO performance that guest-based encryption may have on other guest operating systems. Additional hypervisors should be tested to determine if a performance gap exists, as well as different hardware platforms.

8. REFERENCES

Amazon. "Amaon Elastic Compute Cloud." Retrieved 4 August 2011, 2011, from http://aws.amazon.com/ec2/.

Goth, G. (2007). "Virtualization: Old Technology Offers Huge New Potential." Distributed Systems Online, IEEE 8(2): 3.

Hall, K. (2011) Gartner: SaaS sales will grow 16.2% to $10.7bn in 2011. . Computer Weekly

Kaufman, L. M. (2009). "Data Security in the World of Cloud Computing." IEEE Security and Privacy 7(4): 61-64.

Lenk, A., M. Klems, et al. (2009). What's inside the Cloud? An architectural map of the Cloud landscape. Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing, IEEE Computer Society**:** 23-31.

Mattsson, U. T. (2005). Database Encryption - How to Balance Security with Performance.

Mell, P. and T. Grance (2009). "The NIST Definition of Cloud Computing."

Olivier, M. S. (2002). "Database privacy: balancing confidentiality, integrity and availability." SIGKDD Explor. Newsl. **4**(2): 20-27.

Rackspace. "Enterprise Cloud Computing and Hosting Solutions by Rackspace." Retrieved 23 August 2011, from http://www.rackspace.com/enterprise_hosting/.

Terremark. "Terremark Enterprise Cloud." Retrieved 3 August 2011, from http://www.terremark.com/services/cloudcomputing.aspx.

Appendices and Annexures

Name	Vendor	Host Encryption Support
Hyper-V	Microsoft	Yes
Oracle VM	Oracle Corp	Yes- with add ons
Xen Parallels Server 4	Citrix Systems	Yes- with add ons
ESXi	VMware	No

Table 2 List of available virtualization platforms

	Reads	Scatter Reads	Writes	Gather Writes
Control	67028	59600	3145	95680
Control	70087	61522	3567	99051
Control	70504	61591	3049	99201
Control	71437	61329	3103	98533
Average	**69764**	**61010.5**	**3216**	**98116.25**
Guest Based	34335	37967	1955	62247
Guest Based	36604	27100	2110	61582
Guest Based	29975	37936	1943	67875
Guest Based	36054	37261	1958	62397
Average	**34242**	**35066**	**1991.5**	**63525.25**
Host Based	68205	59848	3359	96705
Host Based	64100	58402	3097	96610
Host Based	62989	58374	3051	96515
Host Based	64428	58341	3032	95458
Average	**64930.5**	**58741.25**	**3134.75**	**96322**

Table 3 Raw data results from testing