A Novel Private Cloud Document Archival System Architecture Based on ICMetrics

Hasan Tahir
Department of Computer Engineering,
Mohammad Ali Jinnah University,
Islamabad, Pakistan
hasan.tahir@jinnah.edu.pk

Ruhma Tahir, Klaus D. McDonald Maier
School of Computer Science and Electronic Engineering,
University of Essex,
Colchester, United Kingdom
rtahir@essex.ac.uk, kdm@essex.ac.uk

***Abstract*—**This paper proposes a novel architecture that offers features for a truly interactive yet private document archival cloud. The proposed secure document archival system has been specifically designed for the elimination of paper based systems in organization. The architecture attempts to shift paper based systems to an electronic archival system where the documents are secure and the system is supportive to user demands and features. Therefore besides document archival the proposed architecture also offers the authentication and encryption for secure keeping of data on the cloud. We propose a novel combination of SSL coupled with ICMetrics for the document archival system, which will prove to be a very valuable tool for encrypting and authenticating in the cloud world.

Keywords— Cloud computing architecture, document archival system, ICMetrics, SSL certificates

I. INTRODUCTION

Recent advancements in the field of virtual computing have changed the way in which we perform computing. These changes are so vast that we want to shift from desktop computing to a completely virtual environment which is best promised by cloud computing [1]. Cloud computing provides a virtual environment where resources can be borrowed on demand. Although this was true to a certain extent in case of the web but cloud computing totally redefines the meanings of borrowing and resources. The basic differentiating factor regarding cloud computing is that it provides a scalable environment where software, platform and infrastructure support can be borrowed.
Cloud computing [1-2] can be defined as:
"Cloud computing is a subscription based service where the networked storage, software services and infrastructure can be provisioned on user demand."

Cloud computing becomes a very attractive networked solution because of its pay per use and outsourcing or resource policy. Some of the greatest advantages of cloud computing are listed below [2].

- Infinite Resources – Cloud computing creates the illusion that unlimited resources are available on demand.
- Promised Growth – Cloud computing allows start-up companies to demand moderate resources and then increase the resources as demand grows. Hence companies do not need to invest in infrastructure that needs constant up gradation. Also the benefit of renting services/ infrastructure provides a healthy growth to developing enterprises [3].
- Flexible payment plans – Cloud computing allows users to ask for resources when they are required and release the same when they are not in use. This provides flexible and affordable pay as you go options [3].

In this paper we propose a novel private cloud document archival architecture that provides features for encrypted document archival and employs ICMetrics to provide formal authentication procedures. The SSL certificate for each user is generated based on its ICMetrics public/ private key pair, which authenticates the user to the cloud for accessing the archived documents. The proposed architecture is an advanced cloud based document archival system that provides features for uploading, searching, indexing and auditing and encrypting documents that are part of the private cloud. Many of these features are not part of the cloud environments and hence traditional cloud computing gives the look and feel of a conventional file system.

The purpose of the proposed architecture is to provide a private and secure environment that allows to maintain an archive of documents, without the threat of even the cloud provider not being able to access the archived encrypted documents. The cloud based architecture also promises growth with increasing resource requirements.

The remainder of this paper is organized as follows; sections 2, 3 and 4 introduces the primitives of cloud computing on which the design of our proposed architecture rests. Section 5 discusses the ICMetrics technology and its features elaborating on how ICMetrics could be a viable solution for the proposed design. Section 6 explains the concept of our proposed document archival system and its significance. The detailed design of our proposed document archival system with its features is presented in Section 7. The conclusion and future work for our paper is presented in section 8.

II. CLOUD COMPUTING MODELS

Fundamentally the cloud can be classified into three broad categories. These categories dictate whether the resources on the cloud are open to all or a subscribed number of users or a collection of both [3-4].

1. Private Cloud – If the services provided by the cloud are only available to a subscribed group of users and there is a secure firewall protecting the cloud infrastructure then the cloud can be called a private cloud.
2. Public Cloud – If the services of a cloud are available to the general public for free or even for a small pay per use fee then the cloud environment can be categorized under public cloud. The services of the cloud may be openly advertised over the internet by the service provider.
3. Hybrid Cloud – As the name indicates the hybrid model is a fusion of private and public cloud. Using this model the service provider can enforce policies and restrictions on what can be accessed privately and what can be accessed on a public cloud.

III. TRADITIONAL STORAGE VERSUS CLOUD STORAGE

Traditional storage mechanisms are fundamentally simple technologies that allow data to be accessed over a network. Many traditional network based storage architectures use technologies like Storage Area Networks (SANs) and Network Attached Storage (NAS) appliances. Primarily these technologies attempt to store data in an unstructured format. It must be pointed out here that unstructured data becomes too complicated to handle when the amount of data increases. Unstructured data results in bottle necks, retrieval delays, single point of failure and most of all it is not scalable [10].

Cloud storage mechanisms try to eliminate the inherent problems of traditional storage. A very popular modern storage system developed by Amazon is called the DynamoDB [11]. DynamoDB is different from traditional storage mechanism because it is a distributed database service. Conventional storage is primarily based on hard disks but DynamoDB is based on solid state drives. DynamoDB is a synchronous replication architecture that provides continuous data backup. Using DynamoDB is simple because the hosting users need to specify the number of requests that the service can handle per second and the service automatically spreads the user's data on enough hardware [12]. The importance of this scalability feature can be best felt in an expanding environment where the number of users and requests per second are also increasing.

IV. SERVICE PROVISIONING MODELS

Previously large companies and even individuals were spending large sums of money on the purchase of infrastructure that could adequately support their web services. An indefinite increase in the number of users on the internet means that the number of users that are actually accessing a web service were much more than anticipated. Hence companies are constantly spending on infrastructure that actually never fulfills their requirements. In such a situation cloud computing presents an attractive and financially viable solution to all the problems.

Discussed below are some service provisioning models that promote the "borrowing" of resources (software, platform and infrastructure).

A. Software as a Service – SaaS

Software as a service allows software to be licensed to a user on demand [3]. SaaS allows users to have access to commercial software on a demand based mechanism. The primary characteristic of SaaS is that the software is managed from a central location but it is distributed among users following the one to many model. Another advantage of SaaS over conventional software purchasing is that the user is not responsible for upgrading or patching the software.

B. Platform as a Service – PaaS

Platform as a service is a provisioning model that allows the creation of web applications without the need to buying and maintaining expensive infrastructure like hardware and software [3]. Fundamentally PaaS provides a platform that for software development, deployment and testing. Another characteristic is that the platform is supported by a GUI that allows multiple users to access the same application before and after deployment.

C. Infrastructure as a Service – IaaS

Infrastructure as a service is perhaps the most popular service provisioning model [4]. The reason for its popularity is that it allows its users to request for essential resources like servers, storage, software and processing capabilities. All these resources can be borrowed for a limited time and then they can be released once they become free. To the user it always seems as if he is the only one using the resource whereas the resources are shared among geographically distributed users. The advantage of this scheme is that the resources are outsourced to the users who can pay for their use following a utility pricing model.

V. ICMETRICS

ICMetrics (Integrated Circuit metrics) is a novel scheme that generates secret keys based on measurable hardware/ software features of a device [5-7]. Conventionally, security solutions have always relied on the use of stored encryption/ decryption keys for their operation. However, stored keys make the systems vulnerable to major attacks, since they are at the risk of easily being compromised by adversaries.

ICMetrics is a breakthrough technology that safeguards applications from the major threats related to key storage. ICMetrics technology generates the secret key at runtime based on hardware and software features of a device, thereby eliminating the need for key storage [8].After each ICMetrics key generation the produced ICMetrics key [7] is temporary and removed after use. The reproduction of the ICMetrics key once again takes place at runtime based on measurable characteristics of the integrated circuit [6].

VI. Document Archival Systems

Document archival systems are primarily designed for large scale organizations that want to maintain their daily paper work in a storage system. In the absence of such systems organizations are in the constant struggle to keep their documents in a safe yet retrievable location. The greatest fear is the loss of documents due to theft, fire or any natural catastrophe. Even in the absence of such events companies need a mechanism where they can access the documents at any time without delay. Naturally such a system will need to be automated and secure from problems like theft and fire. Based on modern computing trends a private cloud based document archival system is proposed that fulfills the requirements of quick access, scalability, minimum footprint and security. Such a system will definitely be categorized under the private cloud because it will provide subscription based services to authorized users only. The second benefit of this system will be that it will be easy to secure and protect from external unauthorized access. A crucial question that arises is that why do we need a new system when we already have a number of commercially available services? The answer to this question is that most services are designed for the public cloud. Secondly the available services are not customizable according to the user's requirements. Pre-existing services that are available over the internet can be useful for a small duration of time but as the demand for resources increases we have no guarantee whether the system will scale up to our increasing demands.

VII. Archival System Features

A. Authentication

The first and foremost feature of a document archival system is an access enforcing component. In this architecture a latest ICMetrics based authentication is proposed that implements formal authentication procedures. Once a user is fully authenticated then privileges can be enforced.

1) ICMetrics for Authentication

Traditionally security procedures used to provide authentication have always relied on the use of stored secret keys. However stored keys have the inherent disadvantage of easily being leaked to adversaries in case of a device compromise. Therefore to safeguard our document archival system from threats related to stored keys we propose the use of ICMetrics [9] keys for carrying out the authentication procedures. ICMetrics generates secret keys directly from measurable software/ hardware properties of a given hardware device. After each key generation, the produced ICMetrics secret number is stored temporarily and is removed after use.

2) ICMetrics Key Pairs

Tahir et al. propose a scheme for the generation of strong ICMetrics key pairs based on the ICMetrics secret number[9]. The scheme generates public/private key pairs based on ICMetrics secret number, which can be used for secure applications. Therefore, the key pair generation for each user in our cloud based document archival system is based on the approach in [9], which generates public/ private key pairs based on ICMetrics. These public/ private key pairs are generated as and when required and are removed after use.

3) SSL Certificates based on ICMetricsKeys

Each user that is part of the document archival system also owns an SSL certificate. Each user's SSL certificate is generated based on the user's ICMetrics public key and the hash of its ICMetrics private key. A trusted Certification Authority (CA) digitally signs each certificate, binding each of the attributes with the owner. A certificate issued to each party by the CA helps the authenticating server at the document archival system authenticate the users. The user's device sends a copy of its SSL certificate to the server. The server verifies the certificate and sends a message to the user's device that he's been authenticated. The user sends back a digitally signed acknowledgement based on its ICMetrics private key to start an SSL encrypted session, letting encrypted data transfer between the user and the server.

B. Document Capturing

Document capturing is a feature that provides features for the input/ upload of documents on the system. Document can be in many formats for example scanned images, document files etc. In traditional on-site storage, organizations themselves have control of where the data is located and who could access the data, but with the advent of cloud computing this control has been handed over to cloud providers. To prevent the data placed on the cloud from misuse or unauthorized access cloud providers make use of SSL encryption. This safeguards the data from being revealed to unauthorized parties. For our document archival system we propose the use of ICMetrics public/ private key pair coupled with a 256 bit AES symmetric operation for encryption and decryption. Therefore all files that are part of our document archival system will be encrypted to prevent unauthorized usage.

The document capturing module is also responsible for providing in-document searching. This becomes even more important if the uploaded document is an image.

C. Indexing

A document archival system should have features for indexing of documents. This indexing assists the users in searching for the documents on the system. The indexing services enhance the searching capabilities of the system.

D. Locating Services

Locating a document in an archival system is a feature based on which individual documents are given a unique identification. This is an important feature because users should be able to determine the location of a document based on its ID. The ID of a document is dependent on how documents are archived in the system. Since folders/ directories can be used to classify/store certain documents therefore locating services are very important. For example in a library there are many books on many subjects. Based on a books accession number the librarian can identify the book, its location and its subject. Hence the books accession number assists in locating a book.

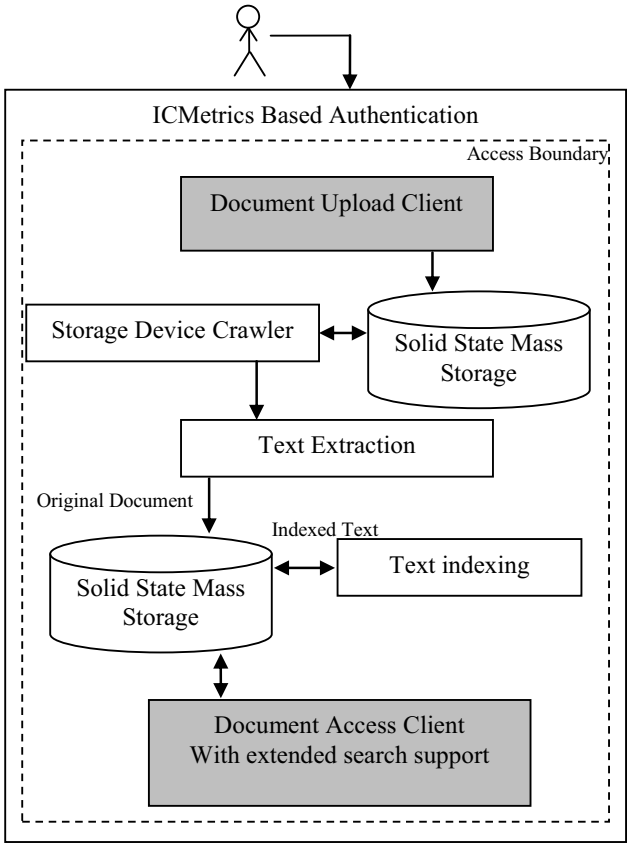

Fig 1. Secure Document Archival System using ICMetrics.

E. Retrieval Services

Document retrieval is a unique feature because a document needs to be located/ searched on the system and then formal retrieval services begin. Retrieval services provide features like downloading of documents. Besides this, documents may need to be downloaded in various formats and extensions. Direct printing of the documents from their source are also some of the features of a retrieval system. The authenticated parties re-generate their ICMetrics private key to decrypt the required file.

F. Audit Features

Auditing features in a document archival system include time stamping, keeping track of when a document was accessed, when it was added to the system and other relevant details. The newly proposed system architecture is a combination of hardware and software designed to provide direct and easy access to the stored documents. The proposed system is unique in the sense that it provides services like Object Character Recognition (OCR), scan and convert, multiple document format support. When comparing this system to existing systems these features are not only advanced but are also basic requirements of all users.

This newly proposed document archival system is unique in the sense that it is a private cloud architecture. Secondly this architecture is purely scalable because if the need for additional storage rises then new storage devices can be incorporated into the cloud. The architecture is discussed in detail below:

This architecture uses a dual client approach. The purpose of having two clients is that a separate client is used to upload the documents while the other client can be used for downloading the documents. Perhaps the greatest advantage of this is that by having two different clients we can divide the traffic influx between both clients that are interacting with different hardware in the cloud. Such a technique also limits the number of users that can upload documents onto the cloud. Having two separate clients also has the benefit that if one client fails the other continues to function. Hence the failure of a client does not result in the failure of the entire system especially in a high availability environment. An upload client is a high privilege client that has the right of adding or removing documents to the archive. Secondly directory creation, deletion and renaming are some of the features the upload client provides to its users.

On the other hand the download client is a search engine based client. The user can search for a particular document using the document identifier, or the user can search for a particular phrase in a document among a collection of documents. The download client is a powerful client that assists the users in locating, downloading and format conversion of the documents found on the cloud. Further, the searching feature is only available to authenticated users of the environment.

Although all these features can also be bundled into a single client but the benefits of having two separate clients outweigh the benefits of having a single client.

Both the uploading and downloading clients connect to two different solid state mass storages. The uploading client will always upload data onto a primary mass storage. The reason for this is that once data is uploaded onto this storage then text is extracted and stored onto a secondary solid state mass storage. This secondary mass storage can be a Write Once Read Many (WORM) drive. The secondary solid state mass storage is a hard drive that is not based on conventional rotational motion magnetic technology to protect from hard disk crashes. Secondly the WORM drive helps in protecting data from deliberate/accidental deletion. The write once feature is helpful in preventing document modification especially in an environment where audit/ transparency is highly required.

VIII. System Implementation

Although the system architecture is complex and large but it must be understood that the basic building blocks for the system already exist in the form of pre-implemented protocols and techniques. Discussed below are some of the inherent protocols that will run on the upload and download clients.

A. WebDAV (Web Distributed Authoring and Versioning)

WebDAV [13] is an extension of the HTTP that assist collaborating users in the creation, management and basic securing of documents, files and folders created on web servers. The reason for using WebDAV is that it provides

features for directory creation, copying, moving, locking, overwrite protection and property viewing of the resources. WebDAV provides very strong document property management features. The features include author name, creation date, querying information, modification dates. In a document archival system these features form the back bone of all operations performed by clients. These features further assist in control and audit assessment procedures.

B. *Object Character Recognition (OCR)- Tesseract*

The text extraction module is in fact an OCR API named Tesseract. [11] This API is open source and developed by Google. The benefit of using Tesseract is that an upload client can upload a scanned document onto the cloud. After the document is processed using the API then a new document is created that contains all the text found on the original scanned document. For increased accuracy/ error correction the resulting document is passed through another Google spelling API. Without text extraction support the uploaded documents are only images and hence the entire document archival is an image archival system.

To break free from this limitation the text extraction module is used. The greatest advantage of this feature is that once text is extracted it can be downloaded by a user for further use. It is worth pointing out here that the text extraction module assists in providing extended features to the archival system. It does not assist in modifying documents that are already uploaded on the cloud.

IX. CONCLUSION

In this paper a novel architecture has been proposed that provides a host of latest features that will assist users in the archival of documents on the cloud. The architecture provides many features namely uploading, downloading, indexing, auditing, searching and attribute based storage of documents. Another basic yet useful feature of this architecture is that it allows the user to submit scanned image documents on the cloud. The scanned documents can then be converted to text format using the Google Tesseract API. Besides these obvious yet necessary features this private cloud architecture is based on ICMetrics and SSL certificates. Perhaps the most prominent yet outstanding promise of this architecture is that it is highly scalable and fault tolerant at the same time. These qualities are offered by the fact that there are multiple points of failure which guarantee that a single failure does not result in the failure of the entire system.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council under grant EP/K004638/1.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Knowinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, M. Zaharia, “Above the Clouds: A Berkley View of Cloud Computing” Technical Report February 2009.

[2] A. Huth, J. Cebula, “The Basics of Cloud Computing” United States Computer Emergency Readiness Team.

[3] M. T. Jones, “Anatomy of a cloud storage infrastructure – Models, Features and internals” IBM developer Works, November 2010.

[4] B. Kepes, “Understanding the Cloud computing stack – SaaS, PaaS, IaaS”.

[5] E. Papoutsis, W. G. J. Howells, A. B. T. Hopkins, and K. D. McDonald-Maier, "Integrating Feature Values for Key Generation in an ICmetric System," in IEEE NASA/ESA Conference on Adaptive Hardware and Systems (AHS-2009) San Francisco, California, 2009, pp. 82-88.

[6] E. Papoutsis, W. G. J. Howells, A. B. T. Hopkins, and K. D. McDonald-Maier, "Key Generation for Secure Inter-satellite Communication," in IEEE, NASA/ESA Conference on Adaptive hardware and Systems 2007, AHS-2007 Edinburgh, UK, 2007, pp. 671-681.

[7] E. Papoutsis, W. G. J. Howells, A. B. T. Hopkins, and K. D. McDonald-Maier, “Integrating Multi-Modal Circuit Features within an Efficient Encryption System”, Third International Symposium on Information Assurance and Security, IEEE Computer Society Washington, DC, USA, 2007, pp. 83-88.

[8] R. Tahir, K. D. McDonald Maier, “Improving Resilience against Node Capture Attacks in Wireless Sensor Networks using ICMetrics”, IEEE Conference on Emerging Security Technologies, Portugal, September 5-7, 2012.

[9] R. Tahir, H. Hu, D. Gu, G. Howells, K. McDonald-Maier, “A Scheme for the Generation of Strong Cryptographic Key Pairs based on ICMetrics”, Proceddings of the 7th IEEE Conference on Internet Technology and Secured Transactions, London, UK, December 10-12, 2012.

[10] “Building your cloud storage – A Basho Technologies White Paper” March 2012.

[11] R. Smith. 2007. “An Overview of the Tesseract OCR Engine”. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02* (ICDAR '07), Vol. 2. IEEE Computer Society, Washington, DC, USA, pp. 629-633.

[12] J. Miao, Z. Fan, G. Chen, H. Mao, L. Wang, “A Private Cloud document management system with document clustering algorithm”, National Conference on Information Technology and Computer Science (CITCS 2012). Atlantis Press.

[13] E. James Whitehead, Jr. “World Wide Web distributed authoring and versioning (WebDAV): an introduction”, Volume 5, Issue 1, March 1997, ACM New York.