International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
www.ioirp.com ISSN: 2394-6364, Volume – 1, Issue – 7. June 2015

# Survey on A Hybrid Cloud Approach for Secure Authorized Deduplication

[1]Ajay H C
Department of computer science and Engg.
Akshya Institute of Technology, Tumkur, India
ajay.shekar215@gmail.com

[2]Gangadhar M L
Department of computer science and Engg.
Akshya Institute of Technology Tumkur, India

***Abstract*— Data deduplication is the method used to remove the repeated copy of files in the cloud service which diminish the storage capacity and the reduce the bandwidth .but in the process of data deduplication while dealing with the sensitive data encryption is required to encrypt before the outsourcing . deduplication provides lot of benefits of security and privacy.in the convergent encryption system merging takes place in to the same place even though encryption done with different user's keys.to avoid the proprietary problem proofs-of-ownership (PoWs) has been introduced ,which is helpful to prove the authorized file owing client.**

***Keywords*— Deduplication,Encryption**

## I. INTRODUCTION

In recent years cloud computing is a rising technology in so many fields where the data usage is more. The service provided by cloud through the web, by mistreatment cloud computing user will make use of internet service to store their huge amount of data in the cloud instead using their physical data storage device which is limited and also vulnerable to some external problems. Cloud computing gives unlimited "virtualized" resources to users as services across the whole Internet, but it's implementation and the platform is hidden from the users for the security issue. Present CSP (cloud service providers) provides more availability of the storage and especially parallel computing resources at relatively low costs .because of universality of the cloud the amount of data is being stored and shared by users with specified privilege which define the access rights of the stored data for that user.
Cloud computing service provides reduced cost, accessible, location independent infrastructure for data management and storage. Cloud resources are dynamically reallocated on the demand along with resource sharing facility with the multiple users. This can work for allocating resources to users. With cloud computing, applications in the server can be accessed by the multiple users can access and update their data without purchasing licenses for different applications. The term "moving to cloud" also refers to an organization moving away from a traditional CAPEX model (buy the dedicated hardware and depreciate it over a period of time) to the OPEX model (use a shared cloud infrastructure and pay as one uses it).

Cloud computing models are divided in to three categories.
1) Software as a Service (SaaS): highly scalable internet based application are hosted on the cloud and offered as services to the end user. 2) Platform as a Service (Paas): it is used to design, develop, build and test applications are provided by the cloud infrastructure. 3)Infrastructure as a Service (Iaas): this is the pay per use modal ,services like storage, database management and compute capabilities are offered on demand. Organization can choose the public, private and hybrid clouds depending on their needs. Following are the benefits of the cloud computing reduced cost, increased and flexibility. Agility improves with users' ability to re-provision technological infrastructure resources .Application programming interface (API) accessibility to software that enables machines to interact with cloud software in the same way that a traditional user interface (e.g., a computer desktop) facilitates interaction between humans and computers. Cloud computing systems typically use Representational State Transfer (REST)-based APIs. Cost reductions claimed by cloud providers. A public-cloud delivery model converts capital expenditure to operational expenditure

Due to increasing in the adoption of Cloud services is accompanied by using the large data storage device at remote servers, hence ideas for saving disk space and reduce the network bandwidth. To avoid above mentioned constraints deduplication is used in the cloud, where the server stores a single copy of each file, in spite of of how many clients asked to store that file. All clients that store the file merely use links to the single copy of the file stored at the server. Moreover, if the server already has a copy of the file then clients do not even need to upload it again to the server, thus saving bandwidth as well as storage. In the existing system of service which is employed with the deduplication, a user has send the hash of the file while he is uploading file to the server in the during first time, server checks existence of the hash value in its database. If the hash is not present in the database then it request the user to upload entire file. Otherwise, since the file already exists at the server (potentially uploaded by someone else) it tells the client that there is no need to send the file itself. Both way the server marks the client as an owner of that file, and from

that point on there is no difference between the client and the original party who has uploaded the file. The client can therefore ask to restore the file, regardless of whether he was asked to upload the file or not.

## II. DEDUPLICATION

Data deduplication is a specialized data compression technique for eliminating duplicate copies of similar data contents in the cloud. Deduplication removes redundant data by maintain the single original copy and use reference of this copy from other redundant data instead of keeping multiple data copies which contains the similar type of data . In deduplication there is a two type 1) file level: in this Deduplication it removes duplicate copies of a file 2) block level: in this level deduplication eliminates duplicate blocks of data that occur in non-identical files. Data deduplication provides the following benefits in cloud computing system: removal of similar data can extensively reduce the storage requirements and improve efficiency in the bandwidth of transmission. Deduplications reduce disk space and also cost of the disk and It improves disaster recovery time since there is less time requirement in the transfer. Backup/archive data usually includes a lot of duplicate data.in traditional encryption is not efficient in providing data confidentiality with the deduplication. Because in traditional encryption it requires different users to encrypt their data with their own keys. Thus, indistinguishable data copies of different users will lead to different cipher texts, making deduplication unfeasible

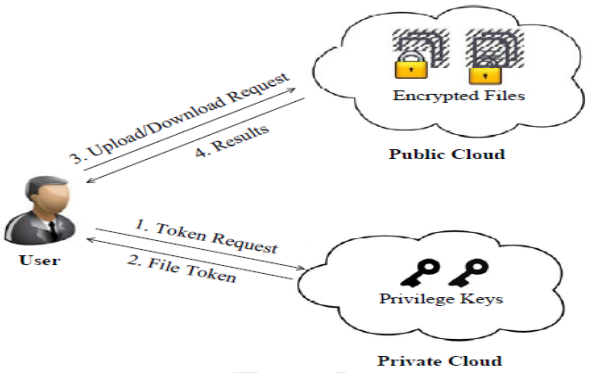


Fig.1: Architecture for Authorized Duplication

## III. HYBRID CLOUD

It is a mixture of two or more clouds (private, community or public) that remain different entities but their exists a togetherness in between them which offering the benefits of multiple deployment models .in the Hybrid cloud it has the ability to manage dedicated services via cloud services and to connect the collocation. hybrid cloud services is cloud computing services that includes the combination of private, public and community cloud services, from different service providers. A hybrid cloud service crosses isolation and provider boundaries so that it can't be simply put in one category of private, public, or community cloud service. It allows one to extend either the capacity or the capability of a cloud service, by aggregation, integration or customization with another cloud service. Varied use cases for hybrid cloud composition exist. For example, an organization may store sensitive client data in house on a private cloud application, but interconnect that application to a business intelligence application provided on a public cloud as a software service. is example of hybrid cloud extends the capabilities of the enterprise to deliver a specific business service through the addition of externally available public cloud services. Another example of hybrid cloud is one where IT organizations use public cloud computing resources to meet temporary capacity needs that cannot be met by the private cloud.] This capability enables hybrid clouds to employ cloud bursting for scaling across clouds. Cloud bursting is an application deployment model in which an application runs in a private cloud or data center and "bursts" to a public cloud when the demand for computing capacity increases. A primary advantage of cloud bursting and a hybrid cloud model is that an organization only pays for extra compute resources when they are needed. Cloud bursting enables data centers to create an in-house IT infrastructure that supports average workloads, and use cloud resources from public or private clouds, during spikes in processing demands

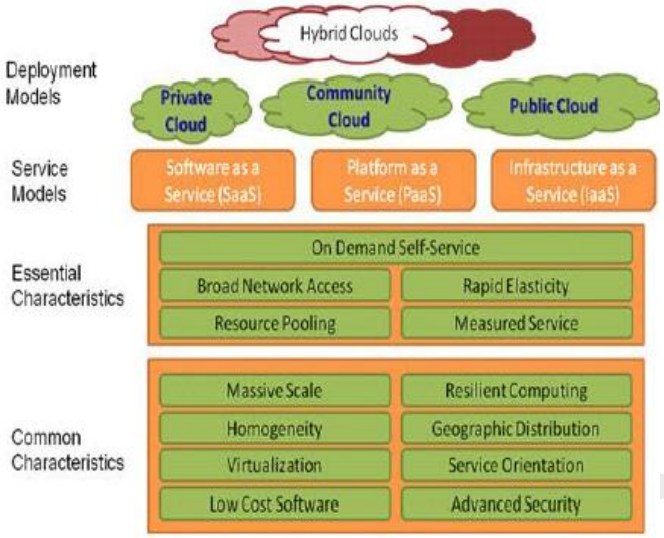


Fig.2: Cloud modal architecture

## IV. EXISTING SYSTEM

Some of the significant existing system for data duplication in cloud environment that are frequently adopted by the researcher are as follows:

- *Secure Deduplication:* With the beginning of cloud computing, secure data deduplication has engrossed much attention recently from research community. Various researchers proposed a deduplication system in the cloud storage to decrease the storage size of the tags for integrity check. To improve the security of deduplication and protect the data confidentiality, various researchers showed how to protect the data confidentiality by transforming the expected message into random message. In their system, another third party called key server has introduced to generate the file tag for duplicate check. Various researchers presented a new encryption scheme that provides differential security for accepted data and not accepted data. For popular data that are not mainly sensitive, the traditional conventional encryption is performed. Another two layered encryption scheme with stronger security while sustaining deduplication has been proposed for unpopular data.. Like this, they achieved better tradeoff between the efficiency and security of the outsourced data. Some researchers also addressed the key management concern in block level deduplication by distributing these keys across multiple servers after encrypting the files.

- *Convergent Encryption:* Convergent encryption ensures data isolation in deduplication. Various researchers formalized this primitive as message locked encryption and explored its application in space efficient secure outsourced storage. Some studies also addressed the problem and showed a secure convergent encryption for efficient encryption without considering issues of the key management and block level deduplication. To encrypt a file using convergent encryption, a client computes a cryptographically strong hash of the file content. The file is then encrypted using this hash value as a key. The hash value is then encrypted using the public keys of all authorized readers of the file and these encrypted values are attached to the file as metadata. Convergent encryption enables identical encrypted files to be recognized as identical, but there remains the problem of performing this identification across a large number of machines in a robust and decentralized manner. There are also more than a few implementations of convergent implementations of different convergent encryption variants for secure deduplication. It is known that some commercial cloud storage providers also deploy convergent encryption.

- *Proof of ownership:* Various researchers proposed the notion of PoW (“proofs of ownership”) for deduplication systems such that a client can efficiently confirm to the cloud storage server that he/she owns a file without uploading the file itself. Some PoW constructions based on the Merkle-Hash Tree are proposed to enable client-side deduplication which include the bounded leakage setting. Certain studies have also proposed another efficient PoW scheme by choosing the

projection of a file onto some randomly selected bit positions as the file confirmation. Note that all the above schemes do not consider data privacy.

## V. RELATED WORK

This section presents some of the prior research work that has been addressing the security and the storage issue in the cloud computing technology.

Anderson and Zhang [1] have described an algorithm which takes advantage of the data which is common between users to increase the speed of backups, and reduce the storage requirements. This algorithm supports client-end per-user encryption which is necessary for confidential personal data. It also supports a unique feature which allows immediate detection of common sub trees, avoiding the need to query the backup system for every file. They describe a prototype implementation of this algorithm for Apple OS X, and present an analysis of the potential effectiveness, using real data obtained from a set of typical users. Finally, they discuss the use of this prototype in conjunction with remote cloud storage, and present an analysis of the typical cost savings. Bellare et al. [2] proposed an architecture that provides secure deduplicated storage resisting brute-force attacks, and realize it in a system called Dup LESS. In Dup LESS, clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol. It enables clients to store encrypted data with an existing service, have the service perform deduplication on their behalf, and yet achieves strong confidentiality guarantees. They show that encryption for deduplicated storage can achieve performance and space savings close to that of using the storage service with plaintext data. Bellare et al. [3] provided definitions both for privacy and for a form of integrity that they call tag consistency. Based on this foundation, they make both practical and theoretical contributions. On the practical side, they provide ROM security analyses of a natural family of MLE schemes that includes deployed schemes. On the theoretical side the challenge is standard model solutions, and they make connections with deterministic encryption, hash functions secure on correlated inputs and the sample-then-extract paradigm to deliver schemes under different assumptions and for different classes of message sources. Their work shows that MLE is a primitive of both practical and theoretical interest. Bellare et al. [4] illustrated either security proofs or attacks for a large number of identity-based identification and signature schemes defined either explicitly or implicitly in existing literature. Underlying these are a framework that on the one hand helps explain how these schemes are derived, and on the other hand enables modular security analyses, thereby helping to understand, simplify and unify previous work. Bellare et al. [5] presented a proof for GQ based on the assumed security of RSA under one more inversion, an extension of the usual one- wayness assumption that was introduced. It also provides such a proof for the Schnorr scheme based on a corresponding discrete-log related assumption. These are the first security proofs for these schemes under assumptions related to the underlying one-way functions. Both results extend to establish security against impersonation under concurrent attack. Bugiel et al. [6] proposed architecture for secure outsourcing of data and arbitrary computations to an un trusted commodity cloud. In their approach, the user communicates with a trusted cloud (either a private cloud or built from multiple secure hardware modules) which encrypts and verifies the data stored and operations performed in the un trusted commodity cloud. They split the computations such that the trusted cloud is mostly used for security-critical operations in the less time-critical setup phase, whereas queries to the outsourced data are processed in parallel by the fast commodity cloud on encrypted data. Douceur et al. [7] present a mechanism to reclaim space from this incidental duplication to make it available for controlled file replication. Their mechanism includes 1) convergent encryption, which enables duplicate files to coalesced into the space of a single file, even if the files are encrypted with different users' keys, and 2) SALAD, a Self-Arranging, Lossy, Associative Database for aggregating file content and location information in a decentralized, scalable, fault-tolerant manner. Large-scale simulation experiments show that the duplicate-file coalescing system is scalable, highly effective, and fault-tolerant. Halevi et al. [8] identified attacks that exploit client-side deduplication, allowing an attacker to gain access to arbitrary-size files of other users based on very small hash signatures of these files. More specifically, an attacker who knows the hash signature of a file can convince the storage service that it owns that file; hence the server lets the attacker download the entire file. (In parallel to their work, subsets of these attacks were recently introduced in the wild with respect to the Drop box file synchronization service.) To overcome such attacks, they introduce the notion of proofs-of ownership (PoWs), which lets a client efficiently prove to a server that that the client holds a file, rather than just some short information about it. They formalize the concept of proof-of-ownership, under rigorous security definitions, and rigorous efficiency requirements of Petabyte scale storage systems. They then present solutions based on Merkle trees and specific encodings, and analyze their security. They implemented one variant of the scheme. Their performance measurements indicate that the scheme incurs only a small overhead compared to naive client-side deduplication. Li et al. [9] introduced a baseline approach in which each user holds an independent master key for encrypting the convergent keys and outsourcing them to the cloud. However, such a baseline key management scheme generates an enormous number of keys with the increasing number of users and requires users to dedicatedly protect the master keys. To this end, they propose Dekey, a new construction in which users do not need to manage any keys on their own but instead securely distribute the convergent key shares across multiple servers. Security analysis demonstrates that Dekey is secure in terms of the definitions specified in the proposed security model. Convergent encryption, also known as content hash keying, is

a cryptosystem that produces identical cipher text from identical plaintext files. This has applications in cloud computing to remove duplicate files from storage without the provider having access to the encryption keys. Ng et al. [10] proposed RevDedup, a deduplication system that optimizes reads to latest VM image backups using an idea called reverse deduplication. In contrast with conventional deduplication that removes duplicates from new data, RevDedup removes duplicates from old data, thereby shifting fragmentation to old data while keeping the layout of new data as sequential as possible. They evaluate their RevDedup prototype using micro benchmark and real-world workloads. For a 12-week span of real-world VM images from 160 users, RevDedup achieves high deduplication efficiency with around 97% of saving, and high backup and read throughput on the order of 1GB/s. RevDedup also incurs small metadata overhead in backup/read operations. Quinlan and Dorward [11] described a network storage system, called Venti, intended for archival data. In this system, a unique hash of a block's contents acts as the block identifier for read and writes operations. This approach enforces a write-once policy, preventing accidental or malicious destruction of data. In addition, duplicate copies of a block can be coalesced, reducing the consumption of storage and simplifying the implementation of clients. Venti is a building block for constructing a variety of storage applications such as logical backup, physical backup and snapshot file systems. They have built a prototype of the system and present some preliminary performance results. The system uses magnetic disks as the storage technology, resulting in an access time for archival data that is comparable to non-archival data. Rahumed et al. [12] presented Fade Version, a secure cloud backup system that serves as a security layer on top of today's cloud storage services. Fade Version follows the standard version-controlled backup design, which eliminates the storage of redundant data across different versions of backups. On top of this, Fade Version applies cryptographic protection to data backups. Specifically, it enables fine-grained assured deletion, that is, cloud clients can assuredly delete particular backup versions or files on the cloud and make them permanently inaccessible to anyone, while other versions that share the common data of the deleted versions or files will remain unaffected. They implement a proof-of-concept prototype of Fade Version and conduct empirical evaluation atop Amazon S3. They show that Fade Version only adds minimal performance overhead over a traditional cloud backup service that does not support assured deletion.

## VI. CONCLUSION

Data Deduplication is the method that describes approach that reduces the storage capacity needed to store data and reduce the transmission bandwidth across the network in cloud computing .Cloud storage has received major attention from industry because of huge requirement of storage space to handle large amount of data. So, cloud offers large storage resources that are coping up with the industries demand. Information Deduplication is useful in cloud backup that saves network bandwidth and reduces network space. The idea of authorized data duplication is to provide security to the user data with different privileges in the process of duplicate check.

## *References*

[1] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.

[2] [2] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013

[3] [3] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013

[4] [4] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009.

[5] [5] M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177, 2002

[6] [6] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011

[7] [7] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. Reclaiming space from duplicate files in a serverless distributed file system. In ICDCS, pages 617–624, 2002.

[8] [8] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[9] [9] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[10] [10] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013

[11] [11] S. Quinlan and S. Dorward. Venti: a new approach to archival storage. In Proc. USENIX FAST, Jan 2002

[12] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui. A secure cloud backup system with assured deletion and version control. In 3rd International Workshop on Security in Cloud Computing, 2011