ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

Effective Data Deduplication and Reduction with Low Overheads in Hybrid Cloud Approach

Chetawani Baban Bibawe, Vaishali Baviskar

ME, Student, Department of Computer Engineering, G.H.Raisoni College of Engineering & Technology, India

Professor, Department of Computer Engineering, G.H.Raisoni College of Engineering & Technology, India

ABSTRACT: Cloud computing is one of the booming technologies in the present day. Number of users availing cloud services is increasing day by day, which in turn requires better performance of the cloud. Cloud Data Center is one of the major components of cloud computing.Data de-duplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To protect the confidentiality of sensitive data while supporting de-duplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data de-duplication. Different from traditional de-duplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. I also present several new de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, I implement a prototype of our proposed authorized duplicate check scheme and conduct tested experiments using our prototype. I show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

KEYWORDS: Access control, Deduplication, Authorized duplicate check, Confidentiality, Hybrid cloud, Image Processing.

I. INTRODUCTION

Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever increasing volume of data.To make data management scalable in cloud computing, de-duplication has been a well-known technique and has attracted more and more attention recently. Data de-duplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, de-duplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. De-duplication can take place at either the file level or the block level. For file level de-duplication, it eliminates duplicate copies of the same file. De-duplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.

Cloud computing is an emerging service model that provides computation and storage resources on the Internet. One attractive functionality that cloud computing can offer is cloud storage. Individuals and enterprises are often required to remotely archive their data to avoid any information loss in case there are any hardware/software failures or unforeseen disasters. Instead of purchasing the needed storage media to keep data backups, individuals and enterprises can simply

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

outsource their data backup services to the cloud service providers, which provide the necessary storage resources to host the data backups. While cloud storage is attractive, how to provide security guarantees for outsourced data becomes a rising concern. One major security challenge is to provide the property of assured deletion, i.e., data files are permanently inaccessible upon requests of deletion. Keeping data backups permanently is undesirable,assensitive information may be exposed in the future because of data breach or erroneous management of cloud operators. Thus, to avoid liabilities, enterprises and government agencies usually keep their backups for a finite number of years and request to delete (or destroy) the backups afterwards. For example, the US Congress is formulating the Internet Data Retention legislation in asking ISPs to retain data for two years, while in United Kingdom, companies are required to retain wages and salary records for six years.

II. LITERATURE SURVEY

Propose a performance-oriented I/O deduplication, called POD, rather than a capacity-oriented I/O deduplication, exemplified by iDedup, to improve the I/O performance of primary storage systems in the Cloud without sacrificing capacity savings of the latter. POD takes a two-pronged approach to improving the performance of primary storage systems and minimizing performance overhead of deduplication, namely, a request-based selective deduplication technique, called Select-Dedupe, to alleviate the data fragmentation and an adaptive memory management scheme, called iCache, to ease the memory contention between the bursty read traffic and the bursty write traffic. I have implemented a prototype of POD as a module in the Linux operating system. In this paper, I have referred Select the every requests to deduplicate or do not bypass small requests (e.g., 4 KB, 8 KB or less).[1]

In the proposed system I achieving the data de-duplication by providing the proof of data by the data owner. This proof is used at the time of uploading of the file. Each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files. New de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Proposed system includes proof of data owner so it will help to implement better security issues in cloud computing.[2]

A system which achieves confidentiality and enables block-level de-duplication at the same time.Before uploading the data or file to public cloud, the client will send the file to private cloud for token generation which is unique for each file. Private cloud then generate a hash and a token and send the token to client. Token and hash keep in the private cloud itself so that whenever next file comes for token generation, the private clod can refer the same token.[3]

To overcome all the problem now a days the use of cloud computing increasing day by day cloud computing become a research topic for better confidentiality and security aspects this technique is new de-duplication technique assisting authorized duplicate check in hybrid cloud system in which duplicate check tokens of files are created by private cloud server with private keys .our proposed system include identity of data owner so it will help as handle better security issues in cloud computing.[4]

A de-duplication system in the cloud storage proposed to reduce the storage size of the tags for integrity check. To upgrade the security of de-duplication and secure the information secrecy demonstrated to secure the information by transforming the predictable message into unpredictable message.[5]

This paper shown that the construction of the distribution matrix in Cauchy Reed-Solomon coding impacts the encoding performance. In particular, our desire is to construct Cauchy matrices with a minimal number of ones. I have enumerated optimal matrices for small cases, and given an algorithm for constructing good matrices in larger cases. The performance difference between good and bad matrices.[6]

A Client program is used to model the data users to carry out the file upload process. A Private Server program is used to model the private cloud which manages the private key and handles the file token computation. A Storage Server program is used to store and de-duplicates files. The Client provides the function calls to support token generation and de-duplication along the file upload process. I observed that the information to Check de-duplication and upload the files, Fetching the Signs using Hashing Algorithm, Checking for Duplication, file uploading, file downloading and attacker trying to attack(block) the cloud.[7]

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

The security will be analysed in authorization of duplicate check and confidentiality of data. For security of Duplicates check by considering antagonist for both internal and external, will try to break the system and by accessing cloud data or will illegal entrance to system. The proposed private data de-duplication protocol is provably secure in the simulation based framework assuming that the underlying hash function is collision resilient, the discrete logarithm is hard and the erasure coding algorithm E can erasure up to fraction of the bits in the presence of malicious adversaries.[8]

II. PROPOSED SYSTEM APPROACH

I propose a scheme to deduplicate encrypted data at CSP by applying PRE to issue keys to different authorized data holders based on data ownership challenge. It is applicable in scenarios where data holders are not available for deduplication control.

1. The proposed system can flexibly support access control on encrypted data with deduplication.
2. Low Cost of Storage.
3. The proposed system can efficiently perform Image deduplication.

The system contains three types of entities:

1) CSP that offers storage services and cannot be fully trusted since it is curious about the contents of stored data, but should perform honestly on data storage in order to gain commercial profits.

2) Data Holder (ui) that uploads and saves its data at CSP. In the system, it is possible to have a number of eligible data holders (ui; i ¼ 1; . . .; n) that could save the same encrypted raw data in CSP. The data holder that produces or creates the file is regarded as data owner. It has higher priority than other normal data holders.

3) Authorized Party (AP) that does not collude with CSP and is fully trusted by the data holders to verify data ownership and handle data deduplication. In this case, AP cannot know the data stored in CSP and CSP should not know the plain user data in its storage. In theory it is possible that CPS and its users (e.g., data holders) can collude. In practice, however, such collusion could make the CSP lose reputation due to data leakage. A negative impact of bad reputation is the CSP will lose its users and finally make it lose profits. On the other hand, the CSP users (e.g., data holders) could lose their convenience and benefits of storing data in CSP due to bad reputation of cloud storage services. Thus, the collusion between CSP and its users is not profitable for both of them. Concrete analysis based on Game Theory is provided in [26]. Therefore, I hold such an assumption as: CSP does not collude with its users, e.g., performing re-encryption for unauthorized users to allow them to access data.

Additional assumptions include: data holders honestly provide the encrypted hash codes of data for ownership verification. The data owner has the highest priority. A data holder should provide a valid certificate in order to request a special treatment. Users, CSP and AP communicate through a secure channel (e.g., SSL) with each other. CSP can authenticate its users in the process of cloud data storage. I further assume that the user policy PolicyðuÞ for data storage, sharing and deduplication is provided to CSP during user registration.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

OUR SCHEME CONTAINS THE FOLLOWING MAIN ASPECTS:

Encrypted Data Upload: If data duplication check is negative, the data holder encrypts its data using a randomly selected symmetric key DEK in order to ensure the security and privacy of data, and stores the encrypted data at CSP together with the token used for data duplication check. The data holder encrypts DEK with pkAP and passes the encrypted key to CSP.

Data Deduplication: Data duplication occurs at the time when data holder u tries to store the same data that has been stored already at CSP. This is checked by CSP through token comparison. If the comparison is positive, CSP contacts AP for deduplication by providing the token and the data holder's PRE public key.

The AP challenges data ownership, checks the eligibility of the data holder, and then issues a re-encryption key that can convert the encrypted DEK to a form that can only be decrypted by the eligible data holder.

Data Deletion: When the data holder deletes data from CSP, CSP firstly manages the records of duplicated data holders by removing the duplication record of this user. If the rest records are not empty, the CSP will not delete the stored encrypted data, but block data access from the holder that requests data deletion. If the rest records are empty, the encrypted data should be removed at CSP.

Data Owner Management:

In case that a real data owner uploads the data later than the data holder, the CSP can manage to save the data encrypted by the real data owner at the cloud with the owner generated DEK and later on, AP supports re-encryption ofDEK at CSP for eligible data holders.

Encrypted Data Update:

In case that DEK is updated by a data owner with DEK0 and the new encrypted raw data is provided to CSP to replace old storage for the reason of achieving better security, CSP issues the new re-encrypted DEK0 to all data holders with the support of AP.

IV. SYSTEM ARCHITECTURE

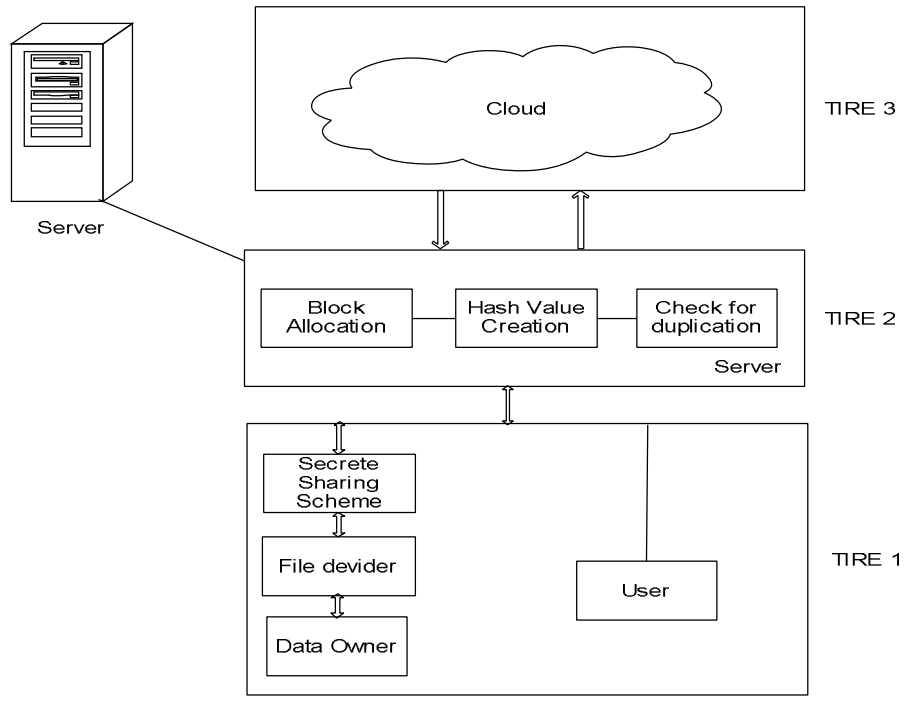


Fig 1. System Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

V. RELEVANT MATHEMATICS ASSOCIATED WITH THE PROJECT

Input
Input given to the system is: -Encrypted File in any format.

Output
Whenever user wants to upload the file on cloud then I check or test the find data duplication and elimination or not.

Process

Step 1: Data owner Select File
Step 2: Encrypt File (For encryption of user AES or RSA).
Step 3: Upload file on cloud.
Step 4: CSP or Controller check the duplicate file available on cloud.
Step 5: If found then remove the duplication and maintain index.
Step 6: Finally non duplicate data stored into Cloud storage

VI. ALGORITHMS USED

1) Convergent Encryption

Convergent encryption provides data confidentiality in de-duplication. A user (or data owner) derives aconvergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derives a tag for the data copy, such that the tag will be used to detect duplicates. Here, I assume that the tag correctness property holds, i.e., if two data copies are the same, then their tags are the same. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored. Note that both the convergent key and the tag are independently derived, and the tag cannot be used to deduce the convergent key and compromise data confidentiality. Both the encrypted data copy and its corresponding tag will be stored on the server side.
A convergent encryption scheme can be defined with four primitive functions:

- KeyGenCE(M)!K is the key generation algorithm that maps a data copy M to a convergent key K;
- EncCE(K, M)!C is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs a ciphertextC;
- DecCE(K, C)!M is the decryption algorithm that takes both the ciphertextC and the convergent key K as inputs and then outputs the original data copy M; and
- TagGen(M)!T (M) is the tag generation algorithm that maps the original data copy M and outputs a tag T (M).

2) Proof Of Ownership

The notion of proof of ownership(POW) enables users to prove their ownership of data copies to the storage server.Specifically, POW is implemented as an interactive algorithm (denoted by POW). The verifier derives a short value $\phi(M)$ from a data copy M. To prove the ownership of the data copy M, the proverneeds to send $\phi$to the verifier such that $\phi = \phi(M)$.

PSEUDO CODE
Step1:Calculate the two convergent key values
Step2: Compare the two keys and files gets accessed.
Step3: Apply de-duplication to eradicate the duplicate values.
Step4: Ifany other than the duplicates it will be checked once again and make the data unique.
Step5: That data will be unique and also more confidential the authorized can access and data is stored.

Data Deduplication:
Data duplication occurs at the time when data holder u tries to store the same data that has been stored already at CSP. This is checked by CSP through token comparison. If the comparison is positive, CSP contacts AP for deduplication by providing the token and the data holder's PRE public key. The AP challenges data ownership, checks the eligibility of the data holder, and then issues a re-encryption key that can convert the encrypted DEK to a form that can only be decrypted by the eligible data holder.

Success Condition:
Successfully work keys generation, file encryption and stored into cloud. User gets result very fast according to their needs.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

Failure Condition:

1. Huge database can lead to more time consumption to get the information.
2. Hardware failure.
3. Software failure.

Space Complexity:

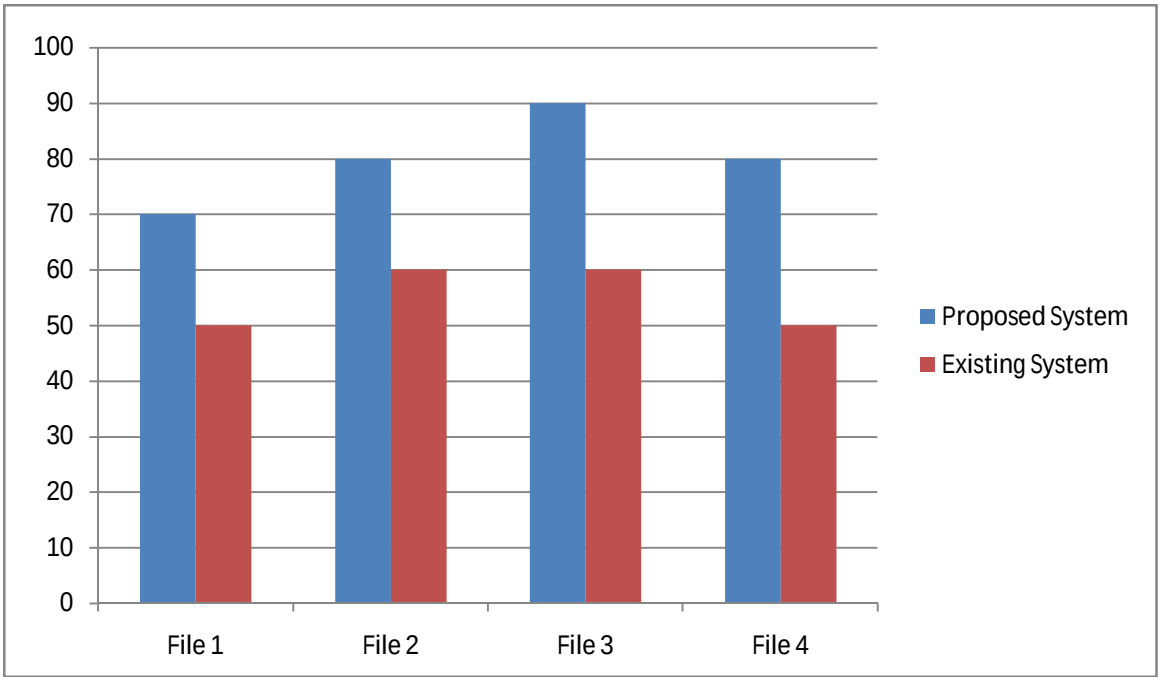
The space complexity depends on Presentation and visualization of discovered patterns. More the storage of data more is the space complexity.

Time Complexity:

Check No. of patterns available in the database= n
If (n>1) then retrieving of information can be time consuming.
So the time complexity of this algorithm is $O(n^n)$.

VII. EXPERIMENTAL RESULT

Proposed system work on low overheads means existing system only find those file which have 100% duplicate if file is 50% duplicate then existing system directly allocate the storage space for file in secondary storage, But in proposed work system reduce the 50% data of file using block and byte level duplication checking and maintain the index and store only 50% data on cloud storage which is not duplicate. Above result table and graph shows the Performance Mesurment and Comparitive analysis between Proposed and Existing System.



Percentage of found similer block in Existing System & Proposed System
(Performance Mesurment and Comparitive analysis)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

VIII. RESULT TABLE

Percentage of found similer block in Existing System & Proposed System

File\System	Proposed System	Existing System
File 1	70	50
File 2	80	60
File 3	90	60
File 4	80	50

IX. CONCLUSION

Several new de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture, in whichthe duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, I implemented a prototype of our proposed authorized duplicate check scheme and conduct testbedexperiments on our prototype. I showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

REFERENCES

[1] Nakrani, Sunil, and Craig Tovey. "On honey bees and dynamic server allocation in internet hosting centers." Adaptive Behavior 12.3-4 (2004): 223-240.
[2] Radojevic, Branko, and Mario Zagar. "Analysis of issues with load balancing algorithms in hosted (cloud) environments."MIPRO, 2011 Proceedings of the 34th International Convention.IEEE, 2011.
[3] Mahajan, Komal, AnsuyiaMakroo, and Deepak Dahiya. "Round Robin with server affinity: a VM load balancing algorithm for cloud based infrastructure."Journal of information processing systems 9.3 (2013): 379-394.
[4] Liang, Po-Huei, and Jiann-Min Yang. "Evaluation of Cloud Hybrid Load Balancer (CHLB)."International Journal of E-Business Development (2013).
[5] Moharana, Shanti Swaroop, Rajadeepan D. Ramesh, and DigamberPowar. "Analysis of load balancers in cloud computing." International Journal of Computer Science and Engineering 2.2 (2013): 101-108.
[6] Domanal, Shridhar G., and G. Ram Mohana Reddy. "Load Balancing in Cloud Computingusing Modified Throttled Algorithm." Cloud Computing in Emerging Markets (CCEM), 2013 IEEE International Conference on.IEEE, 2013.
[7] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou "A Hybrid Cloud Approach for Secure Authorized De-duplication" IEEE Transactions on Parallel and Distributed Systems: PP Year 2014
[8] Bo Mao, Member, IEEE, Hong Jiang, Fellow, IEEE, Suzhen Wu, Member, IEEE, andLeiTian, Senior Member, IEEE "Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud " IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016
[9] Mr Vinod B Jadhav Prof Vinod S Wadne Secured Authorized De-duplication Based Hybrid Cloud Approach International Journal of Advanced Research in Computer Science and Software Engineering
[10] Abdul Samadhu, J. Rambabu, R. Pradeep Kumar, R. Santhya Detailed Investigation on a Hybrid Cloud Approach for Secure Authorized De-duplication International Journal for Research in Applied Science and Engineering Technology (IJRASET)
[11] JadapalliNandini, Rami reddyNavateja Reddy Implementation De-duplication System with Authorized Users International Research Journal of Engineering and Technology (IRJET)
[12] Sharma Bharat, Mandre B.R. A Secured and Authorized Data De-duplication with Public Auditing International Journal of Computer Applications (09758887)
[13] Wee Keong Ng SCE, NTU Yonggang Wen SCE, NTU Huafei Zhu Private Data De-duplication Protocols in Cloud Storage SAC12 March 2529, 2012, Riva del Garda, Italy. Copyright 2011 ACM 9781450308571/12/03
[14] Shweta D. Pochhi, Prof. Pradnya V. Kasture Encrypted Data Storage with De-duplication Approach on Twin Cloud International Journal of Innovative Research in Computer and Communication Engineering
[15] Backialakshmi. N Manikandan. M SECURED AUTHORIZED DE-DUPLICATION IN DISTRIBUTED SYSTEM IJIRST International Journal for Innovative Research in Science and Technology— Volume 1 — Issue 9 — February 2015
[16] BhushanChoudhary, AmitDravid A Study On Secure Deduplication Techniques In Cloud Computing International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) Volume 3, Issue 12, April 2014