Journal of Engineering Sciences

Vol 11, Issue 7,July/2020

ISSN NO:0377-9254

# PREVENTION OF FILE DUPLICATION FOR ONLINE SOCIAL NETWORKING

Dr. Mohammad Abdul Waheed[1], Archana.J.Ambekar[2]

[1]*Associate professor Computer Science, VTU Center for PG Studies Kalaburagi, India*

[2]*Student Computer Science, VTU Center for PG Studies Kalaburagi, India*

***Abstract*— In this digital era, social networking sites have provided the easiest way of communicating and data sharing. These applications and sites demand greater amount of storage; not just in smart device but in clouds as well. Cloud services have a greater demand where huge amount of data can be shared and stored. Since same data can get stored several times, issues like wastage of memory and delay in processing can be emerged. In this paper we propose a solution to reduce memory consumption which will further reduce the memory requirement and interference by avoiding the storage of duplicate files in the device. Prevention of file duplication for online social networking. We focus on content of the files which may include text files, image files etc.,**

***Keywords*— cloud; data sharing; duplication; image file; memory; storage; text file.**

## 1. INTRODUCTION

In today's world smart phone users are increasing day by day with the increasing number of smart phone users, people are also getting used to social networking sites and applications. These social networking sites and application have become a part of daily life. They have also become important and easy way of communication and data sharing. These applications provide to store the data in the storage device of the phone. Not only mobile apps but cloud service is also in the large demand where huge amount of data are shared and stored. Some social networking sites allows the user to store their data in clouds. Since the same data can be stored in the cloud we have to deal with the wastage of cloud memory, delay in processing and will put lot of load on the cloud.

These applications are not limited to specific type of data but can transfer any type of data such as image , audio video, files etc., so all this sharing will take a lot of space which will eventually cause storage problem and can also cause device hanging and make the processing slower. All these problem will become unbearable for the user. Memory limitation and interference are seen as major problem even in virtual environment. In this paper we try to provide a solution to reduce the memory consumption which will further reduce the memory requirement and interference by avoiding the storage of duplicate file in the device. By overcoming these problems system performance will be increased. To make it understandable, if a single file isshared by multiple users then multiple copies aren't stored in the memory. Instead, a single copy is stored in the memory. After the evaluation we found that the proposed system was successful in storing single copy of any shared data. Many cloud services are turning towards on demand policy where the resources or data are provided to the user whenever they are requested, this policy is also called as pay as you go model of business. Cloud storage have been gaining lot of importance by the users for the secure storage but since many use share lot of data there are chances that the cloud could contain same data by different users. But the duplicate data will can cause lot of trouble to the cloud and cloud management.

Apart from clouds the applications installed in the phone where the data will be stored in physical storage device. In applications a lot of data are shared and stored on daily basis. In the process of sharing a single person can get the same data from two users then the same data will be stored twice. Because of this reason we might face storage problem which indeed will make the device processing slower and data fetching will get delayed. In this paper we provide a solution to the problem of duplication which can be applied both to the cloud and to mobile devices. The proposed system uses the method of VMMP (virtual machine based memory partitioning) which helps us to classify the data into different category which has the high chance of duplicate data existence where the latest data will be compared with the data

already present in the device. The comparison is done based on the type of data such as audio, video, image, text file etc., We focus on file content, file name, file size for the text files and we also check the content of image files by comparing bit level of images. After the evaluation we were successful in avoiding the duplicate storage of data and maintain the system performance.

## 2. RELATED WORK

Cloud storage have been gaining a lot of importance by user for the secure storage but since many use sharing a lot of data there are chances that the cloud could contain same data by different users. But the duplicate data can cause a lot of trouble to the cloud and cloud management. In order to solve this problem additional VM's are added to the cloud and all the devices are collected as one single physical server which are operated separately. But, if the number of VM's keep increasing then there occurs a problem of interaction between the VM's.

In mobile devices, the applications are not limited to specific type of data but can transfer any type of data such as image, audio, video, text files etc., in the storage. Our proposed system hence works on all types of data that can be sent or received through the application.

## 3. PROPOSED SYSTEM

In the proposed system, virtual machine based memory partition technique can be applied to memory pages in order to avoid data duplication. In this method all the VM's present in the server will be assigned with one memory bank because for all VM's we have to compare the compare the data of the page present in memory domain called as comparison domain which is used for reducing unnecessary comparison.

Again the pages in the domain group are divided into different categories and each category will have a local comparison tree. The selected pages will be sent to domain which will be compared with the pages present in the local tree of its category. The pages present in the local tree will have chances of having the same data as that of selected page. Because of this interference can be reduced, thus identifying the page sharing chances efficiently.

In order to categorize the pages to reduce the waste comparison we have to overcome the following difficulties:

- There are high chances of pages having the same data, which must be placed into the same category as that of the local tree.
- There are low chances of pages shaving same data, which must be placed into different category by which we can avoid unuseful search in local tree.
- The page classification should be made easy.

## 4. METHODOLOGY

The HCL paradigm have been preferred initially with the data entry of activities and the work required in developing the new system. This helps us to develop the interactive and progressive system with the best quality of service with more accuracy, stability and reliability. This paradigm believes that the user's interaction with the system is important part of development and it has to be given more attention. The usability of the system progresses as the implementation activities progresses. This mainly focuses on improving the user experience and performance by testing and iterative approach. This approach is mainly preferred while developing the complex system.

Domain complex systems are the systems which have very great degree of complex and every small details have to be considered in order to develop the system. This also consists of larger technical details in particular work. This approach contains very large amount of dependencies and tangled workflow of the system. This approach design mainly concentrates on providing the help to these requirements and to help in developing automated systems. This is considered as the useful and requirement driven method.

### 4.1. ALGORITHM

**Algorithm: VMMP algorithm**

**Definition:**

N: the total cores in the server

M: the total memory banks in the server

AVG: AVG=M/(N+2)

RM:RM=(M-32)/N

**VMMP:**

If AVG<16

Allocate 16 banks for hypervisor;

Allocate 16 banks for VM's;

Each two applications of one VM's share RMbanks;

Else

Allocate 16 banks for each application of one VM's;

Allocate 16 banks for hypervior;

Allocate remainder banks for VM's

With the help of this algorithm we can decrease the communication between the VM and application running on it. Along with that we also see a decline in the communication among the applications of different VM.

### 4.2. IMPLEMENTATION

The proposed system have been divided into different modules:

*1)* Data Sharing: This module allows the user to share the data to one another. After the data have been received by the users these data will be stored in database. Before storing the data into the database the data have to be checked for duplication using VMMP module.

2) VMMP Module*:* In VMMP method all the VM's present in the server will be assigned with one memory bank group because for all the VM we have to compare the data of the page present in memory domain called as memory domain to reduce the unnecessary comparison

### 4.3. TECHNOLOGIES USED

1) TOMCAT SERVER:

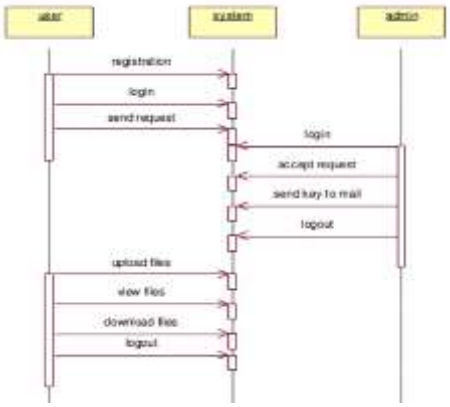
Tomcat Server is a Java servlet container and also a web server that was introduced by the Apache Software Foundation. A web server is a response to the client for their requests that is placed by users at a web browser. A web server isn't just a program that serves only static HTML pages but also executes the programs in response to requests made by the users and also returns the dynamic results to the user's browser. Apache Software Foundation's Tomcat server is favorably used for many applications as it provides both Java servlet and Java Server Page (JSP) technologies. It's an open source servlet and JSP engine.

2) DREAMWEAVER: Dreamweaver is a editing tool which allows us to design the pages in an easier manner. It allows us to work in different types of domain. Working in dreamweaver makes the work easier as the developers can view the design of the code in the tool itself. This tool also allows us to include other graphical apps and this also allows us to create and edit the images in the macromedia fireworks. This tool also supports the flash assets to the web pages. This tool makes the designers work lot easier as they have drag and drop facility.

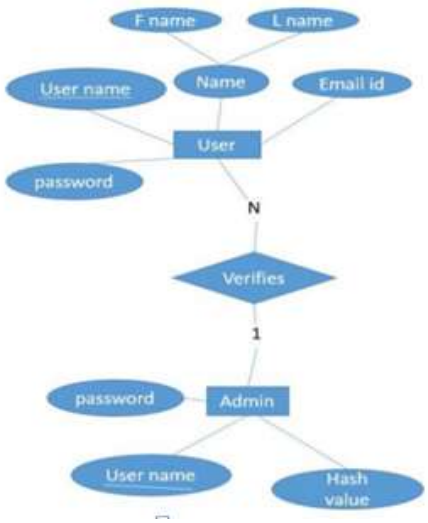
3) JSP:

Java Server Pages allows to run server side executable content in a webpage. It can be referred to as a means of letting a Web page that relies on the terms on the server information that is entered into a form etc.CGI and a variety of languages such as C, C++ and (most prevalent) Perl were originally used to provide server side run able content. Therefore both Perl and CGI are yet developing may not as to the same level as some of the other technologies. In the recent years, Java Servlet is been handy that lets user to make use of the same application where in you can write server side executable content that is a program which creates a HTML page as is final result or output.

### 4.4. SEQUENCE DIAGRAM



### 4.5. DATAFLOW DIAGRAM




### 4.6. EVALUATION AND RESULT

In this process we try to check whether the application developed is according to our requirements. The main purpose is to develop the application which avoids the storage of duplicate data. We want to develop the system which can be implemented both in cloud services as well as mobile applications. With the help of this proposed system we hope to maintain the performance of the cloud service and mobile device.

The following points will be check at the time of performance evaluation:

1) First, we check the communication and data sharing process is performed properly without any interrupt.
2) Second, we check the system is capable of storing data in the data. i.e., the communication between the user interface and database is done correctly.
3) Third, we will check the system is capable of sharing all type of data such as audio, video, image, file, etc.,
4) Fourth, we will check the comparison between the existing data and shared data is performed correctly and find the duplicate data and discard one file out of duplicate files.
5) Fifth, we will check system is suitable for both cloud service as well as mobile apps.
6) Sixth, We have to check that the system should be capable of handling unexpected errors occurred during runtime.

### 4.7. TESTING

After the coding of system has been completed we have to perform the testing of the code to ensure about the quality of the project. Testing is the most crucial part of the development. It discloses many errors and the exception occurred in the functionality of the system that appears to be working according to the requirements. Along with that the data collected during the testing phase provides the great information for the future reference and they also show the reliability.

WHITE BOX TESTING

In this testing we make use of the control structure of the procedural design from which we the test cases.

BLACK BOX TESTING

The main use of this testing is to get the input conditions which can be implemented on the system in order to check the functionality. During this testing the major errors detected are communication error, data access errors and problems during interaction between the modules.

## 5. RESULT

During evaluation we first have checked the impact of the data sharing on the system performance. Later we have checked how VMMP algorithm effects the system performance. We have check the reduction in the memory page communication and the categorization of the memory pages into different categories

## 6. CONCLUSION

In this paper we have provided a solution for the duplicate data storage in the memory. The proposed system makes use of the VMMP algorithm to perform deduplication task and to reduce the interferences. In VMMP method all the VMs present in the server will be assigned with one memory bank group because of for all the VMs we have to compare the data of the page present in memory domain called as comparison domain to reduce the unnecessary comparison.

Again the pages in the domain the groups are divided into different categories & each category will have a local comparison tree. The pages taken will be sent to the domain which will be compared with the pages present in the local tree of its category. The pages present in the local tree will have the high changes of having the same data as the taken page because of this the interference can be reduce thus identifying the page sharing chances efficiently.

In the end we would like to tell that we were able to reduce the memory usage and interferences.

**REFERENCES**

[1] Fei Xu, Fangming Liu, Hai Jin, V. Vasilakos. Managing Performance Overhead of Virtual Machines in Cloud Computing: A Survey, State of the Art, and Future Directions. Proceedings of the IEEE, pages 11-31, Jan. 2014 J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] Amazon.com, Customer success. Powered by the AWS Cloud. [Online]. Available: http://aws.amazon.com/solutions/case-studies K. Elissa, “Title of paper if known,” unpublished.

[3] Amazon.com, Amazon elastic compute cloud (Amazon EC2) . [Online]. Available: http://aws.amazon.com/ec2/

[4] J. F. Gantz, S. Minton, and A. Toncheva, Cloud computings role in job creation, Mar. 2012. [Online]. Available: http://www.microsoft.com/enus/news/features/2012/mar12/03-05cloudcomputingjobs.aspx.

[5] R. P. Goldberg. Survey of virtual machine research. Computer, 7(9):34C45, Sept. 1974.

[6] M. Rosenblum and T. Garfinkel. Virtual machine monitors: current technology and future trends.Computer, 38(5):39C47, 2005.

[7] L. Chen, Z. Wei, Z. Cui, M. Chen, H. Pan, Y. Bao. CMD: classificationbased memory deduplication through page access characteristics. In VEE14, 2014.

[8] A. Arcangeli, I. Eidus, and C. Wright. Increasing memory density by using ksm. InProceedings of the Linux Symposium (OLS09), pages 19C28, 2009

[9] D. Gupta, S. Lee, M. Vrable, S. Savage, A. C. Snoeren, G. Varghese, G. M. Voelker, and A. Vahdat. Difference engine: harnessing memory redundancyinvirtualmachines.In8thUSENIXSymposiumonOperating Systems Design and Implementation, OSDI08, pages 309C322, 2008.

[10] C. A. Waldspurger. Memory resource management in vmware esx server. SIGOPS Oper. Syst. Rev., 36(SI):181C194, Dec. 2002.

[11] M. Jeong, D. Yoon, D. Sunwoo, M. Sullivan, I. Lee, and M. Erez. Balancing DRAM Locality and Parallelism in Shared Memory CMP Systems. HPCA, 2012.

[12] M. Xie, D. Tong, Y. Feng, K. Huang, X. Cheng. Page Policy Control with Memory Partitioning for DRAM Performance and Power Efficiency. ISLPED, 2013

[13] W. Mi, X. Feng, J. Xue, and Y. Jia. Software-Hardware Cooperative DRAM Bank Partitioning for Chip Multiprocessors. NPC, 2010.

[14] M. Xie, D. Tong, K. Huang and X. Cheng. Improving System Throughput and Fairness Simultaneously in Shared Memory CMP Systems Via Dynamic Bank Partitioning. HPCA, 2014.

[15] H. Cheng, C. Lin, J. Li, and C. Yang. Memory Latency Reduction via Thread Throttling. MICRO, 2010.

[16] S. Muralidhara, L. Subramanian, O. Mutlu, M. Kandemir, and T. Moscibroda. Reducing Memory Interference in Multicore Systems via Application-Aware Memory Channel Partitioning. MICRO, 2011.

[17] Q. Deng, D. Meisner, L. Ramos, T. F. Wenisch, and R. Bianchini. MemScale: Active Low-Power Modes for Main Memory. In ASPLOS, 2011.

[18] Li T, John L K, Sivasubramaniam A, et al. Understanding and improving operating system effects in control flow prediction. ACM Sigplan Notices. ACM, 2002, 37(10): 68-80.

[19] T.Wood,G.Tarasuk-Levin,P.Shenoy,P.Desnoyers,E.Cecchet,andM.D.Corner. Memorybuddies:exploitingpagesharingforsmartcolocation in virtualized data centers. SIGOPS Oper. Syst. Rev., vol. 43, no. 3, pp. 27-37, July 2009.

[20] F. Bellard. Qemu, a fast and portable dynamic translator. In Proceedings of the annual conference on USENIX Annual Technical Conference, ATEC 05, pages 41C46, 2005.

[21] Kvm-kernel based virtual machine. http://www.linuxkvm.org/page/Main Page.

[22] ab - apache http server benchmarking tool. http://httpd.apache.org/docs/2.2/programs/ab.html.

[23] Sysbench: a system performance benchmark. http://sysbench.sourceforge.net/.

[24] M. K. Qureshi, and Y. N. Patt. Utility-based cache partitioning: A low-overhead, high-performance, runtime mechanism to partition shared caches. In MICRO- 39, 2006.

[25] R. Azimi, D. K. Tam, L. Soares, and M. Stumm. Enhancing Operating System Support for Multicore Processors by Using Hardware Performance Monitoring. In ACM SIGOPS Operating Systems Review 43(2): 56- 65, 2009.

[26] K. Sudan, N. Chatterjee, D. Nellans, M. Awasthi, R. Balasubramonian, and A. Davis. Micro-Pages: Increasing DRAM Efficiency with LocalityAware. In ASPLOS-2010.

[27] D. Kaseridis, J. Stuecheli, and L. K. John. Minimalist Open-page: A DRAM Page-mode Scheduling Policy for the m any-core Era. In MICRO44, 2011.

[28] P. Sharma, and P Kulkarni. Singleton: system-wide page deduplication in virtual environments, in Proc. the 21st Int. Symp. on High-Performance Parallel and Distributed Computing, pp. 15-26, 2012.

[29] G. Jia, G. Han, L. Shi, J. Wan, D. Dai. Combine Thread with Memory Scheduling forMaximizing Performance in Multi-core Systems. ICPADS, 2014.

[30] G. Jia, G. Han, J. Jiang, J.J.P.C. Rodrigues. PARS: A scheduling of periodically active rank to optimize power efficiency for main memory. Journal of Network and Computer Applications, 2015.

[31] G. Jia, G. Han, J. Jiang, A. Li. Dynamic Time-slice Scaling for Addressing OS Problems Incurred by Main Memory DVFS in Intelligent System. Mobile Networks and Applications 20 (2), 157-168, 2015.

[32] G. Jia, G. Han, A. Li, J. Lloret. Coordinate Channel-Aware Page Mapping Policy and Memory Scheduling for Reducing Memory Interference Among Multimedia Applications. IEEE System Journal, 2015.

[33] Gangyong Jia, Guangjie Han, Joel J.P.C. Rodrigues, Jaime Lloret, Wei Li, Coordinate Memory Deduplication and Partition for Improving Performance in Cloud Computing