Analysis of File Compression Based on Amazon EC2 Cloud Platform

Juhi Sharma, Anuradha Taleja, Kshitiz Saxena

Abstract- The advent and wide adoption of cloud computing has brought a new revolution in the field of IT. As consumers using cloud for data storage either in the SaaS, PaaS or IaaS deployment model are increasing-they realize the necessity of file compression. It becomes imperative to understand the scenarios where transition to the Cloud is beneficial. In our research, we have demonstrated that for small businesses – making a move to the cloud is not a good approach as less demanding applications can run well even on stand-alone machines, however as the data size grows – making a move towards cloud can yield higher availability. In this paper we evaluate and compare the performances of virtual machines running in cloud with stand-alone (unvirtualized) machine.

Index Terms—Cloud computing, Amazon EC2, File Compression, Performance Analysis

I. INTRODUCTION

Cloud Computing as a technology has its roots in Grid Computing, Distributed Systems and Operating Systems. The "pay as you go" model has been the USP of cloud computing. The companies now need to invest less in buying and maintaining hardware and focus more on business delivery with the advent of cloud computing. The high degree of availability and assurance of service delivery make a good proposition for companies to deploy their applications in Cloud. On the flip side there are many instances reported where start-up companies after transitioning to Cloud have suffered losses and have even forced them to deploy the applications in their own premises eventually.

File Compression offers significant cost savings in terms of reduction in storage space and lesser time in transmission of data. Previous work on file compression is less focused on performance evaluation of compression for a system under work load. This paper is a study regarding performance analysis of compression techniques in a public cloud and comparison with un-virtualized stand-alone machine. This paper focuses on performance evaluation of compression for files of varying sizes in a public cloud offering by the world's biggest cloud company – Amazon.

II. LITERATURE SURVEY

A lot of researchers have conducted experiments to evaluate the performance of cloud computing infrastructures. In [10] , the performance analysis of cloud computing services for many –tasks scientific computing is attempted using some benchmarks. In [15], the researchers have investigated the performance evaluation of cloud infrastructures.

Manuscript received March, 2014.

Juhi Sharma Saxena, Department of Computer Science, Meerut Institute of Engineering and Technology, Meerut affiliated to UPTU Lucknow, India.

Anuradha Taluja, Department of Computer Science, Meerut Institute of Engineering and Technology, Meerut affiliated to UPTU Lucknow, India.

Kshitiz Saxena, Department of Computer Science, Bharat Institute of Technology, Meerut affiliated to UPTU Lucknow, India.

Storage Resources have been studied in depth for public and private clouds [7]. However no attempt has been made by the researchers to study the means of reducing network utilization and storage. The research [13] does discuss an architecture to improve network performance in cloud computing but file compression is not discussed. This paper is different from previous works in the sense that analysis of file compression in the cloud is compared between virtualized and unvirtualized machines.

III. A BRIEF INTRODUCTION TO AMAZON EC2 CLOUD PLATFORM

Amazon EC2 is a public cloud computing platform which has the biggest collection of data centers in the world[1]. Amazon Cloud has set benchmarks and standards for other cloud vendors. The offerings of Amazon are diversified and are preferred by companies all over the globe. Amazon EC2 offers IaaS for a large number of operating system images. Amazon Elastic Compute Cloud (Amazon EC2) offers web service that provides re sizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers. Amazon EC2's simple web service interface facilitates in obtaining and configuring capacity with minimal friction. It provides complete control of computing resources. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing to quickly scale capacity, both up and down, as computing requirements change. Amazon EC2 changes the economics of cloud computing servers by allowing a customer to to pay only for capacity that is actually used. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.

IV. PERFORMANCE EVALUATION METHODOLOGY

This section presents the adopted methodology for performance evaluation concerning the public cloud platform and compression techniques. The methodology adopted is shown by Figure 1 below:

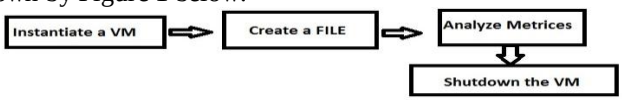

Fig 1: Methodology Adopted

The following VM on Amazon -EC2 cloud was instantiated: t1.micro with 613 MB RAM, 8 GB Storage , 1 CPU of 1.0 - 1.2 GHz and with RHEL 6.4.

The unvirtualized machine was also configured with the same configuration. The following software were used:

V. shred - It is a utility to permanently delete a file by over writing it with random bytes. However we are creating a file using shred as follows:

shred -s [Size of the File] - > [Name of the File]

VI. gzip - It is a software application used for file compression and decompression. The program was created by Jean-Loup Gailly and Mark Adleras a free software replacement for the compress program used in early Unix systems. It is based on the DEFLATE algorithm, which is a combination of LZ77 and Huffman coding compression techniques.

VII. gunzip - It is a software application which deflates files compressed by gzip.

VIII. sysstat - It is a package for measuring system statistics. One application which is a part of this package is pidstat which reports I/O and CPU utilization.

IX. EXPERIMENTAL RESULTS

In this section we describe the technical details regarding the experiments conducted and the results are tabulated eventually. We conducted our tests first on un-virtualized machine and subsequently on EC2 instance. The following snapshots depict the experiments conducted on AMAZON CLOUD:

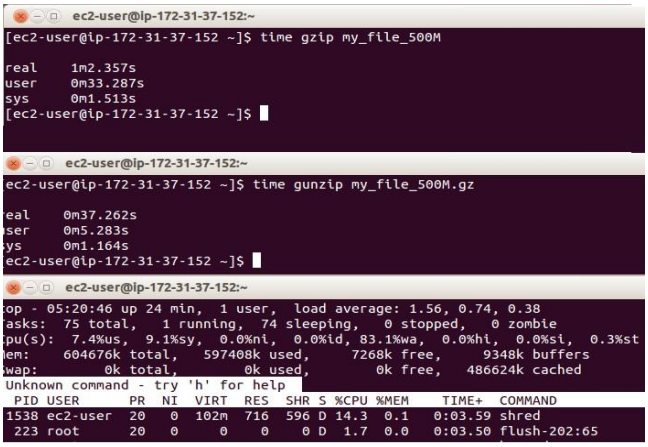

Fig 2: Analysis of a 500MB File in Amazon EC2

Fig 3: Analysis of a 250MB File in Amazon EC2

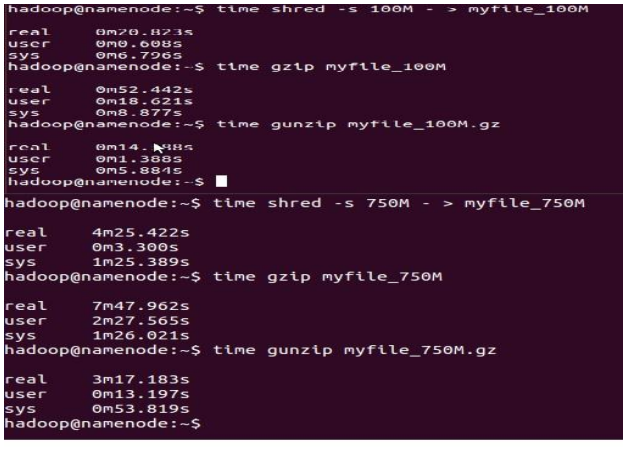

Fig 4:Analysis of Files in unvirtualized machine.

A large set of experiments were conducted both in virtualized and unvirtualized environments over varying sizes of files and the results are tabulated as follows:

	100KB	100MB	250MB	500MB	750MB
AMAZON	0.01	6.879	17.64	62.357	85.5
LAPTOP	0.004	52.442	148.425	312.557	467.962

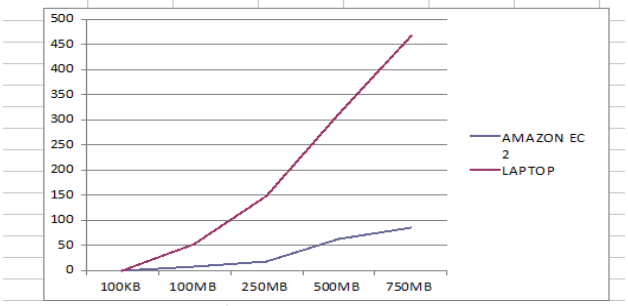

Fig 5: Analysis of File Creation with Random Data

	100 bytes	100 KB	100 MB	500 MB
LAPTOP	0.12	0.174	21.029	86.829
EUCALYPTUS	0.137	0.224	7.795	61.87
AMAZON EC2	0.018	0.041	14.194	61.575

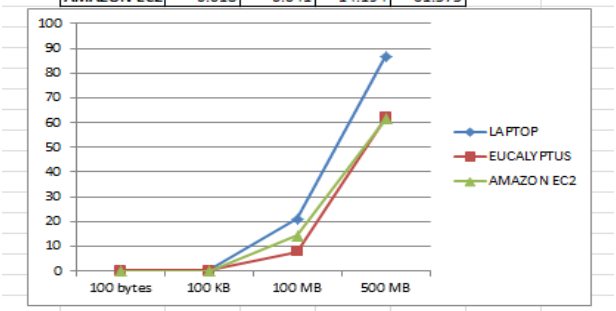

Fig 7:Extending File Compression to More Public Clouds and comparison with an unvirtualized machine.

X. CONCLUSION AND FUTURE WORK

We have exhaustively researched analyzed the performance of Public Cloud Vendors and have compared the results with unvirtualized machine. One of our findings is that the time taken to generate random byte - files was highest in stand alone machine , where as Eucalyptus and Amazon EC2 performed almost similar. As the amount of computation increased , the performance of Amazon EC2 was found to be better which might be attributed to the efficient, exclusive and high availability of data centers across the globe. The compression of files took lesser time as long as the size of files were less on a standalone machine. This could be mainly because a typical data center employs power efficient techniques such as DVFS where as the computation increases , the CPU operate at higher frequencies. On stand alone machine - the processor utilization increased with increased computation (on demand) – linear. A similar argument could be given for decompression of files where Eucalyptus Cloud VM instance performed worst where as Amazon EC2 cloud performed best.

We have tried to prove that as the availability of vast computational power of data centers becomes generally available - computation in a cloud will effectively replace stand-alone machines.The performance of VMs in a cloud is also proven to be better than stand-alone machines. File -Compression and De-compression though is computationally intensive but the benefit in terms of storage savings will be

preferred when data is stored in CLOUD. The cloud services offered by Amazon EC2 have set up benchmarks and standards being followed by other CLOUD Vendors.We hope to extend our work to other open source CLOUD offerings such as OPEN NEBULA or Google Cloud Service (announced in Jan 2014) in testing the performance of multimedia compression and decompression algorithms.are self-contained.

ACKNOWLEDGMENT

J.S. Author thanks Dr. Ajay Kumar Singh, HOD CS, MIET, Meerut for his unfailing guidance and support.

REFERENCES

[1] Amazon Web Services (2011). Eucalyptus open-source cloud computing infrastructure -an overview. technical report, eucalyptus, inc.
[2] Chee, B. and Franklin Jr, C. (2009). Cloud computing: technologies and strategies of the ubiquitous data center. CRC.
[3] Chorafas, D. and Francis, T. . (2011). Cloud computing strategies. CRC Press. D, J. and Murari, K. and Raju, M. and RB, S. and Girikumar, Y. (2010). Eucalyptus Beginner's Guide - UEC Edition.
[4] Ghoshal, D., Canon, R., and Ramakrishnan, L. Understanding i/o performance of virtualized cloud environments. Godard, S. (2004). Sysstat:System performance tools for the Linux OS.
[5] He, Q., Li, Z., and Zhang, X. (2010). Study on cloud storage system based on distributed storage systems. In Computational and Information Sciences (ICCIS), 2010 International Conference on, pages 1332–1335. IEEE.
[6] Hovestadt, M., Kao, O., Kliem, A., and Warneke, D. (2011). Evaluating adaptive compression to mitigate the effects of shared i/o in clouds. In Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium on, pages 1042–1051. IEEE.
[7] Hugos, M. and Hulitzky, D. (2010). Business in the Cloud: What Every Business Needs to Know About Cloud Computing. Wiley.
[8] Iosup, A., Ostermann, S., Yigitbasi, N., Prodan, R., Fahringer, T., and Epema, D. (2010).
[9] Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing. IEEE Transactions on Parallel and Distributed Systems, pages 1–16.
[10] Krintz, C. and Calder, B. (2001). Reducing delay with dynamic selection of compression formats. In High Performance Distributed Computing, 2001. Proceedings. 10th IEEE International Symposium on, pages 266–277. IEEE.
[11] Lilja, D.J. (2005). Measuring computer performance: a practitioner's guide. CambridgeUniv Pr.
[12] Miyamoto, T., Hayashi, M., and Tanaka, H. (2009). Customizing network functions for high performance cloud computing. In Network Computing and Applications, 2009. NCA 2009. Eighth IEEE International Symposium on, pages 130–133. IEEE.
[13] Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff, L., and Zagorodnov, D. (2009). The eucalyptus open-source cloud-computing system. In Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pages 124–131. IEEE Computer Society.
[14] Ostermann, S., Iosup, A., Yigitbasi, N., Prodan, R., Fahringer, T., and Epema, D. (2010).A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing, pages 115–131.
[15] Ozsoy, A. and Swany, M. (2011). Culzss: Lzss lossless data compression on cuda. In Cluster Computing (CLUSTER), 2011 IEEE International Conference on, pages 403–411. IEEE.

AUTHOR PROFILE

Juhi Sharma Saxena is an MTech student. She has research interests in Cloud Computing and has published one research paper and has co-authored one book.. She has earned international certifications in IBM RAD, RFT, Tivoli, DB2 and Oracle 9i.

Anuradha Taluja is an Assistant Professor in the Department of Computer Science, MIET, Meerut. She has completed her M.Tech and has been teaching for the last three years. She has several publications in various journals of repute.

Kshitiz Saxena is an Asscoiate Professor in the Department of Computer Science at Bharat Institute of Technology, Meerut. He has been teaching for the last 13 years and has several publications in international journals. He has research interests in Cloud Computing and Network Security.