Efficient big data sharing with cloud

Ankur Singh
singh.ankursingh@gmail.com
Babu Banarasi Das University, Lucknow

Sameer Awasthi
Associate Professor
Babu Banarasi Das University, Lucknow

ABSTRACT

Big data such as social media contents, an archive of high definition videos gathered via ubiquitous information-sensing devices and scientific data could be acquired and stored within an organization's external cloud(s) and distribution retrieved by staffs or customers via cloud services offered by the organization. The growth of cloud computing, big data, and analytics compels businesses to turn into big data-as-a-service solutions in order to overcome common challenges, such as data storage or processing power. This paper presents new and comparative performance behaviours of Cloud and three well-known approaches by emulating a hybrid cloud as a testing environment.

Keywords— *Big data, Cloud, Storage*

1. INTRODUCTION

Big data such as social media contents, an archive of high definition videos gathered via ubiquitous information-sensing devices and scientific data could be acquired and stored within an organization‟s external cloud(s) and distribution retrieved by staffs or customers via cloud services offered by the organization. This leads to the downstream bandwidth saturation of network connection between external cloud and big data consumer premise, long-delayed cloud service responsiveness and importantly increases in external cloud data-out charge imposed by public cloud provider . The significance of the last problem could be realized through the following representative scenario (which is also referred to throughout this paper): an enterprise utilizing big data re-siding in clouds by transferring it through 10 Gbps Metro Ethernet with 25% average downstream bandwidth utilization for 8 work hours a day, and 260 workdays per year requires the total amount of cloud data-out transfer 190.43 TB per month. This data transfer volume can be translated as 29,933 USD per month based on the weighted average cost 0.1535 USD per GB of Google Cloud Storage‟s network egress charge in Asia-Pacific region as of September 2013. The sharing of big data can be conducted in an economical and network-friendly manner by using client-side cloud cache. Client-side cloud caches are located in or nearby user premise in the form of enterprise-level shared cache, personal web browser cache or local user-application cache. Fig.1 demonstrates the deployment scenario of a shared cloud cache where HTTP requests to external hybrid cloud are proxied by a cloud cache, which in turn replies with the valid copies of the requested big data objects either from its local cache repository (i.e., cache hits) or by retrieving updated copies from the cloud (i.e, cache misses). Cloud caches inherit the capabilities of traditional forward web caching proxies since cloud data is also delivered by using the same set of HTTP/TCP/IP protocol stacks as in WWW. Unavoidably, the same problem as in web caching proxies also exists in cloud caches that are caching entire remote data in the local cache is not economically plausible, thus cache eviction approach is mandatory for cloud caches. When the big-data hosting cloud is a kind of hybrid, which employs different public cloud providers for risk management purpose, different data-out charge rates potentially apply to data-outcasts and must be aware of cache eviction approach for economical performance optimization.

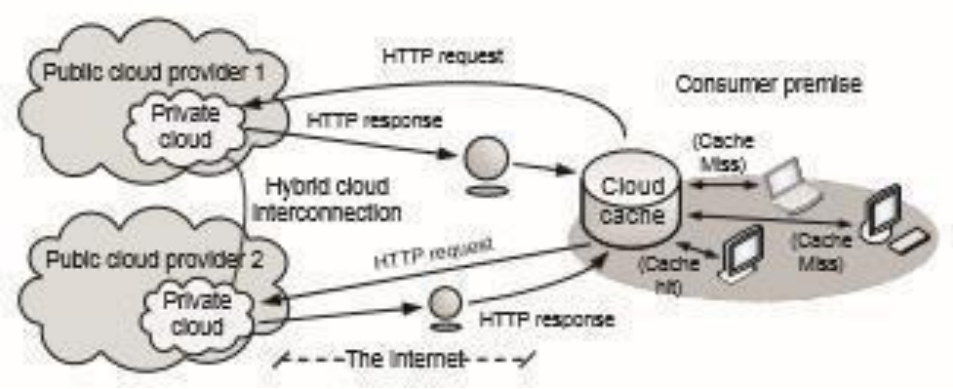

Fig. 1: Cloud cache deployment in a hybrid cloud scenario

2. LITERATURE SURVEY

The proliferation of data warehouses and the rise of multimedia, social media and the Internet of Things (IoT) generate an increasing volume of structured, semi-structured and unstructured data. Towards the investigation of these large volumes of data, big data and data analytics have become emerging research fields, attracting the attention of the academia, industry and governments. Researchers, entrepreneurs, decision makers and problem solvers view „big data" as the tool to revolutionize various industries and sectors, such as business, healthcare, retail, research, education and public administration. In this context, this survey chapter presents a review of the current big data research, exploring applications, opportunities and challenges, as well as the state-of-the-art techniques and underlying models that exploit cloud computing technologies.

The growth of cloud computing, big data and analytics compels businesses to turn into big data-as-a-service solutions in order to overcome common challenges, such as data storage or processing power. Although there is related work in the literature in the general area of cost-benefit analysis in the cloud and mobile cloud computing environments, a research gap is observed towards the evaluation and classification of big data-as-a-service business models. Several research efforts have been devoted comparing the monetary cost-benefits of cloud computing with desktop grids [26], examining cost-benefit approaches of using cloud computing to extend the capacity of clusters or calculating the cloud total cost of ownership and utilization cost to evaluate the economic efficiency of the cloud. Finally, novel metrics for predicting and quantifying the technical debt on cloud-based software engineering and cloud-based service level were also proposed in the literature from the cost-benefit viewpoint and extended evaluation results are discussed by Skourletopoulos et al.

Base Paper

This paper presents the new and comparative performance behaviours of Cloud and three well-known approaches by emulating a hybrid cloud as a testing environment where economical costs offered by two public cloud providers are non-uniform. The main objective of doing this is to observe the performances of i-Cloud that has learned uniform cost patterns but is deployed against a non-uniform cost environment. A minor objective is to show how much i-Cloud outperforms the other approaches when data-out charge rates are non-uniform. The findings of these observations would convince users of iCloud performances when deploying cloud cache for a single private cloud at the beginning that later evolves to a hybrid cloud according to new business requirements.

3. RELATED WORKS

There are numerous cache eviction approaches in present existence. They have been extensively investigated in our previous works [10]. To recap, none of them aims for big data and cloud computing for two main reasons. First, those approaches evict big objects to optimize hit rates rather than byte-hit and delay-saving ratios, crucial to the scalability of cloud-transport infra-structures and the responsiveness of cloud computing services, respectively. Second, they do not support multiple public-cloud data-out charges, thus neither improve cloud consumer-side economy nor support hybrid cloud deployment. the i-Cloud approach, orig-inally proposed in [11], extends its prior non-intelligent versions [10], [12], [13] by integrating an artificial neural network (ANN) to automate an algorithmic parameter self-tuning for workload adaptability. Its performances have been studied without comparing with the other well-known approaches and based on the totally uniform cost circumstances of both ANN training and deployment phases.

4. METHODOLOGY

This paper presents the new and comparative performance behaviours of Cloud and three well-known approaches by emulating a hybrid cloud as a testing environment where economical costs offered by two public cloud providers are non-uniform.

The main objective of this research to find out the i-cloud, learning uniform cost patterns, could perform well against non-uniform cost environment.

5. CONCLUSION

This paper presents Cloud cache eviction approach that accommodates the distributed sharing of big data. Cloud has access recency as a priority factor for object replacement decision. Cloud parameterizes an MLP-based self-tuning window size to gen-eralize the frequencies of objects within a formulated object cluster. The lowest profitable clustered objects are purged from cloud cache. Based on the trace-driven simulation results, the distributed sharing of big data was most efficient when employing i-Cloud. Although Cloud has been trained based on a uniform cost model, it performed well against a non-uniform cost environment or multi-provider hybrid cloud.

6. FUTURE WORK

According to this paper, we can work on many different things that will give us more options to increase the usability of cloud storage for big data. Some of the future works that can be further proceeded are

- Compression of big data stored on the cloud.
- Cost reduction methods.
- Security of the data stored.
- Encryption and Decryption of data.

These were the future works that can be opted to work upon in near future.

7. REFERENCES

[1] Amazon.com, Inc. (8 August 2013) Amazon simple storage service. [Online]. Available: http://aws.amazon.com/s3/pricing/

[2] Google Inc. (8 August 2013) Google cloud storage. [Online]. Available: https://cloud.google.com/pricing/cloud-storage/

[3] Microsoft. (8 August 2013) Windows Azure. [Online]. Available: http://www.windowsazure.com/

[4] S. Podlipnig and L. B¨osz¨ormenyi, "A survey of web cache replacement strategies," ACM Comput. Surv., vol. 35, no. 4, pp. 374–398, Dec. 2003.

[5] J. Cobb and H. ElAarag, "Web proxy cache replacement scheme based on back-propagation neural network," J. Syst. Softw., vol. 81, no. 9, pp. 1539–1558, Sep. 2008.

[6] S. Romano and H. ElAarag, "A neural network proxy cache replacement strategy and its implementation in the squid proxy server," Neural Comput. Appl., vol. 20, no. 1, pp. 59–78, Feb. 2011.

[7] W. Ali and S. M. Shamsuddin, "Intelligent client-side web caching scheme based on least recently used an algorithm and neuro-fuzzy system," in Proceedings of the 6th International Symposium on Neural Networks, 2009.

[8] S. Sulaiman, S. Shamsuddin, F. Forkan, and A. Abraham, "Intelligent web caching using neuro-computing and particle swarm optimization algorithm," in Modeling Simulation, 2008. Second Asia International Conference on, 2008, pp. 642–647.

[9] W. Tian, B. Choi, and V. V. Phoha, "An adaptive web cache access predictor using the neural network," in Proceedings of the 15th intl. conf. on Industrial and engineering applications of artificial intelligence and expert systems, 2002.

[10] T. Banditwattanawong, "From web cache to cloud cache," in Advances in Grid and Pervasive Computing, ser. Lecture Notes in Computer Science, R. Li, J. Cao, and J. Bourgeois, Eds. Springer Berlin / Heidelberg, 2012, vol. 7296, pp. 1–15.

[11] T. Banditwattanawong and P. Uthayopas, "An intelligent cloud cache replacement scheme," in Advances in Information Tech-nology, 2013. IAIT 2013. 6th International Conference on, 2013.

[12] "Cloud cache replacement policy: new performances and findings," in Annual PSU Phuket, 2012. PSU PIC 2012. 1st Interna-tional Conference on, 2013.

[13] "Improving cloud scalability, economy and responsiveness with client-side cloud cache," in Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2013. ECTICON 2013. 10th International Conference on, 2013.

[14] National Laboratory for Applied Network Research. (2012) Weekly squid http access logs. [Online]. Available: http://www.ircache.net/

[15] T. Banditwattanawong, S. Hidaka, H. Washizaki, and K. Maruyama, "Optimization of program loading by object class cluster-ing," IEEJ Transactions on Electrical and Electronic Engineering, vol. 1, no. 4, pp. 397–407, 2006.

[16] D. Wessels, Squid: the definitive guide. O Reilly, 2004.