

Data Recognition on the Cloud – a Distributed Service for Scalable Data Management on Cloud Computing

Anang Hudaya Muhamad Amin¹, Asad I. Khan²

¹Monash University, Clayton, VIC, Australia, Anang.Hudaya@infotech.monash.edu.au

²Monash University, Clayton, VIC, Australia, Asad.Khan@infotech.monash.edu.au

INTRODUCTION

Cloud computing is a new computing paradigm. In an article written by Knorr and Gruman and published by InfoWorld [1], it has been stated that cloud computing is the next big trend in distributed computing, allowing users to tap into tremendous computing resources on pay-per-use basis without any needs to invest in new infrastructure, training new personnel, or licensing of new software. In addition, Gartner Inc. [2] also quoted that “cloud computing heralds an evolution of business that is no less influential than e-business”. These two statements foresee a significant impact of cloud computing in today’s computing requirements. Furthermore, Merrill Lynch [3] has estimated, by 2011, the cloud computing market would reach \$160 billion that comprises of \$95 billion in business and \$65 billion in online advertising. In scientific computing perspective, cloud eventually derived from grid computing paradigm in which a collaboration of computing resources being made available for users, as an analogy, the power grid made available to each household.

As a new computing paradigm, efforts need to be put in place for the cloud system to deliver its optimum capabilities. One of the areas of interest would be in data management. With extreme parallelisation and distribution of data, storage and retrieval processes within cloud systems are becoming more complex and require an efficient data management mechanism, especially in the activities involving real-time data processing. Existing data management schemes on cloud computing mostly rely on both underlying Distributed File Systems (DFS) and parallel scanning procedure, such as Google’s GFS (Google File System), Hadoop’s DFS (HDFS), BigTable and MapReduce infrastructure, and Amazon’s S3 storage cloud. As mentioned by Wu and Wu [4], problems with these approaches lie in the data partitioning among numerous available nodes within the cloud system, as well as nodes collaboration for a specific job. To solve these problems, data management scheme should be able to manage data partitioning between processing nodes, as well as the capability to collaborate nodes for a specific task efficiently. Currently, these two features are still lacking in the available data access schemes. Furthermore, dependency on DFS for data retrieval and storage has a possibility to create bottle necks and single-points of failure for data access. Thus, this effect reduces the scalability factor of the existing data access schemes.

Apart from access mechanism, we also need to consider the properties of data in cloud computing environment. According to Grossman et al [5], there are three distinctive properties of data in cloud:

- Data in cloud could grow larger than a few hundred terabytes. Current databases are becoming less competitive in comparison with other specialised solutions including Google’s GFS.
- Data in cloud can easily be replicated.
- Data in cloud is readily available, waiting for computing tasks to be performed. Unlike the existing grid computing infrastructure where data is distributed across available processing nodes for computations.

With these properties in mind, it will be essential for any cloud service developers to ensure data integrity within cloud with regards to its nature, size, and availability.

AIMS AND CONTRIBUTIONS

Cloud systems are in need of a complete data management scheme that enables data partitioning on-the-go and has the ability to disseminate processing nodes for specific data retrieval/storage tasks. With this in mind, we would like to explore all possibilities to consolidate the scheme with efficient partitioning approaches. By having such an integration within a complete end-to-end scheme will enable data storage and retrieval processes to be performed effectively, regardless of the distribution of data within the cloud system. We will outline a distributed data management scheme that enables data access to be conducted effectively by the means of an associative computing approach [6]. We intend to treat data as pattern and perform data storage and retrieval through a supervised pattern recognition mechanism. This approach envisages data retrieval to be implemented as a distributed pattern association process that is implemented through the integration of loosely-coupled computational networks. Followed by a divide-and-distribute approach that allows distribution of these networks within the cloud dynamically.

The main contributions of this work are to:

- Provide a discussion on the underlying fundamentals of cloud computing with a focus on data management perspective. We aim to examine effective cloud infrastructure for data retrieval and storage. In doing this, we will observe the similarities and differences of grid and cloud systems.
- Analyse the complexity and scalability of existing data management schemes for cloud computing. We intend to observe the parallelisation effect within the current data retrieval and storage mechanisms for cloud system.
- Introduce a novel distributed scheme for cloud computing for creating a database-like functionality that can scale up or down over the available infrastructure without interruption or degradation, dynamically.

Existing data management mechanism for cloud computing such as MapReduce has demonstrated the ability for parallel access approach to be performed on cloud infrastructure. Our work enhances the capability of this parallel processing approach by accommodating a lightweight data management scheme that enables efficient data retrieval from a collection of data sets distributed across a cloud of networks. We aim to utilise an existing distributed pattern recognition scheme known as Distributed Hierarchical Graph Neuron (DHGN) on data access processes within cloud. DHGN has been used in a wide range of associative memory applications ranging from image recognition on commodity grid [7, 8] to event detection within wireless sensor networks (WSN) [9]. We argue that data retrieval and storage on cloud could be performed efficiently and effectively via distributed pattern recognition by interfacing parallel data analysis and existing data query technique such as SQL language with associative computing based data management within the cloud.

REFERENCES

1. Knorr, E. and G. Gruman, *What cloud computing really means*. 2008, InfoWorld.
2. Stevens, H. and C. Pettey, *Gartner Says Cloud Computing Will Be As Influential As E-business*, in *2008 Press Releases*. 2008.
3. *The Cloud Wars: \$100 Billion at Stake*. 2008, Merrill Lynch.
4. Wu, S. and K.-L. Wu, *An Indexing Framework for Efficient Retrieval on the Cloud*. Bulletin of the Technical Committee on Data Engineering, 2009. **32**(1): p. 75-82.
5. Grossman, R.L., et al., *Compute and storage clouds using wide area high performance networks*. Future Generation Computer Systems, 2009. **25**: p. 179-183.
6. Trahan, J.L., et al., *Relating the power of the Multiple Associative Computing (MASC) model to that of reconfigurable bus-based models*. Journal of Parallel Distributed Computing, 2009. **2009**.
7. Khan, A.I. and A.H. Muhamad Amin, *One Shot Associative Memory Method for Distorted Pattern Recognition*, in *AI 2007: Advances in Artificial Intelligence, 20th Australian Joint Conference on Artificial Intelligence, Gold Coast, Australia, December 2-6, 2007, Proceedings*, M.A. Orgun and J. Thornton, Editors. 2007, Springer. p. 705-709.
8. Muhamad Amin, A.H. and A.I. Khan. *Commodity-Grid Based Distributed Pattern Recognition Framework*. in *Sixth Australasian Symposium on Grid Computing and e-Research (AusGrid 2008)*. 2008. Wollongong, NSW, Australia: ACS.
9. Muhamad Amin, A.H. and A.I. Khan. *Parallel Pattern Recognition Using a Single-Cycle Learning Approach within Wireless Sensor Networks*. in *Proceedings of the 2008 Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies 2008 (PDCAT'08)*. 2008. Dunedin, New Zealand: IEEE Computer Society.