MINIMIZING DATA TRANSFER COST THROUGH ANT COLONY OPTIMIZATION IN THE CLOUD

T. Arun, D. Sasikumar, P., M. Princy, Bavitiraa, R
Assistant Professor, Department of Electronics and Instrumentation Engineering,
Nehru Institute of Technology, Coimbatore 641 105
arunmails@ymail.com

ABSTRACT

The booming volume of data generated by contemporary business users and consumers have created enormous data storage and management challenges. Cloud computing provides a way to enable massive amounts of data to work together as data-intensive services. Many users are moving their data to online storage clouds, where data are stored based on the pay-as-you-go model. The cost and access response time of data sets influence the quality of the service that requires the data sets. There are many cost-effective approaches have been developed for achieving minimum cost benchmark. However, it may not be sufficient, if large-scale applications have to run in a more distributed manner. In this project, for incorporating data transfer cost into minimum cost benchmark, an effective Ant Colony Optimization (ACO) algorithm with Appriori algorithm has been proposed. ACO reduces the delay in accessing & transferring the data sets which paves the way to achieve a minimum cost benchmark. By using appriori algorithm, duplication of data sets after restarting can be avoided in case of any crashes in the cloud. Appriori algorithm is also used for retrieving frequent data items easily so that the cost for storing the infrequent large datasets can be reduced. This will drastically reduce the cost of the datasets storage along with the reduced cost of data transfer through ACO when there is a need of transferring data from one node into another node of the same or different cloud.

INTRODUCTION

A number of leanings are opening up the period of Cloud Computing, which is an Internet based expansion and use of computer knowledge. The ever inexpensive and more powerful computers systematized with the software as a service (SaaS) computing construction are transmuting data centers into puddles of computing package on a enormous measure. The increasing network bandwidth and reliable yet flexible network connections make it even possible that users can now subscribe high quality services from data and software that reside solely on remote data centers. Moving data into the cloud offers great convenience to users since they don't have to care about the complexities of direct hardware management.

It can be used in any calculation and data exhaustive applications with different price models of cloud services. The overall random simulations and evaluate the overall outcomes of our approach by comparing with other characteristic storage strategies. In the direction of almost achieving the minimum data sets storage cost in the cloud we have developed an original runtime local optimization based storage strategy.

Knowledge association between items in a great database of transactions has been designated as an important database mining problem. These connotation instructions are accessible in a compact form removing conclusion. Exclusion of termination in the association instruction set generated from

large item is an inspiring task. In this project, the data transfer costs are also reduced in order to maintain an overall minimum cost benchmark through a methodology named Ant Colony Optimization (ACO) with Appriori algorithm. In previous works, the researchers have used the concept of ACO to build upon a load balancing solution set within nodes of a cloud system. ACO can also reduce the delay in accessing the datasets in one cloud & apriori algorithm eliminates duplication of data sets.

LITERATURE SURVEY

In this paper, Data Dependency Graph (DDG), is a directed acyclic graph (DAG), which is based on data provenance in scientific applications was used. All the data sets once generated in the cloud, whether stored or deleted, their references are recorded in DDG. In other words, it depicts the generation relationships of data sets, with which the deleted data sets can be regenerated from their nearest existing preceding data sets [1]. Committing large volumes of rarely used data to storage wastes space and energy, making it a very expensive strategy. Cloud computing, with its readily available and flexibly allocatable computing resources, suggests an alternative: storing the provenance data, and means to recomputing results as needed. While computation and storage are equivalent, finding the balance between the two that maximizes efficiency is difficult. Once it has been determined that an end product can be recomputed, a cost model is invaluable in choosing the most efficient strategy for dealing with computed results. Marginal costs describe how much an additional unit of cloud service would cost at the present moment. The cost trends attempt to predict where the costs might be at a point in the future. Deciding when to store a result or when to rely instead on computation comes down to a cost-benefit analysis. One of the key challenges in balancing storage and computation is deciding who makes the decision regarding what is to be stored and what is to be recomputed; both end user and system knowledge are required for an informed decision. There are a number of factors, intrinsic to the result itself, that must be taken into account when choosing whether to store or recompute. These result-specific issues can be divided into low-level and high-level factors. Low level factors describe a system level view of how results are created, while high-level factors are based on the meaning of the result, and are often highly subjective [2]. This paper presents an integrated storage service called iCostale that reduces the overall cost of data storage through automatic selection and placement of user's data into one of many storage clouds. it intelligently transforms data based on its type, access frequency, transformation overhead, and the cost model of the storage cloud providers. iCostale combines well-known data compression techniques, a knowledge base of compression algorithms and cloud pricing schemes, and history of access patterns to reduce the end-user cost of using cloud storage. Instead of managing its own storage, iCostale uses storage from one of the many cloud providers. Users upload and access their contents in iCostale, which in turn stores them in the backend cloud. The goal of iCostale is to reduce the overall cost of data storage for its users [3]. ANDS and ARCS are developing services and infrastructure to enable researchers to publish their data. Research and analysis work in ANDS has identified a number of basic service functions that in most cases are required to enable the sharing and reuse of data through its publication. In order for the data to be discoverable and re-used, it is necessary to provide discovery information to enable a user to discover the existence of

a data collection, assessment information to enable a user to decide if the discovered data is of interest to them, and re-use information to enable a user to make use of the data once they have decided is of interest to them [4]. This paper describes a meta-negotiation infrastructure to establish global Cloud exchanges and markets, and illustrate a case study of harnessing 'Storage Clouds' for high performance content delivery. As Clouds are designed to provide services to external users, providers need to be compensated for sharing their resources and capabilities. This paper has proposed architecture for market-oriented allocation of resources within Clouds. This presented various Cloud efforts in practice from the market-oriented perspective to reveal its emerging potential for the creation of third-party services to enable the successful adoption of Cloud computing, such as meta-negotiation infrastructure for global Cloud exchanges and provide high performance content delivery via 'Storage Clouds' [5].

EXISTING SYSTEM

We have investigated the unique features and requirements of data sets storage in computation- and data intensive applications in the cloud. Toward practically achieving the minimum data sets storage cost in the cloud, we have developed a novel runtime local-optimization based storage strategy. The strategy is based on the enhanced linear CTT-SP algorithm used for the minimum cost benchmarking by taking into the consideration of users' (optional) preferences. Theoretical analysis, general random simulations, and specific case studies indicate that our strategy is very cost-effective by achieving close to or even the same as the minimum cost benchmark with highly practical runtime efficiency. Our current work is based on Amazon clouds' cost model and assumes that all the application data be stored with a single cloud service provider. However, sometimes large-scale applications have to run in a more distributed manner because some application data may be distributed with fixed locations.

Disadvantages:

1. New data sets may be produced in the cloud at any time and the prevailing data sets usage occurrences may also modification from time to time so that the storage approach needs to be activated rather regularly.

2. The cost rate of storing a DDG calculated is a static value which is based on the evaluated data sets usage occurrences.

3. The real cost rate for storing the data sets in the cloud is dynamic, and the minimum cost target is only a theoretical value at a particular moment in the cloud.

4. For any algorithms, we can only calculate an approximate value of the total cost rate.

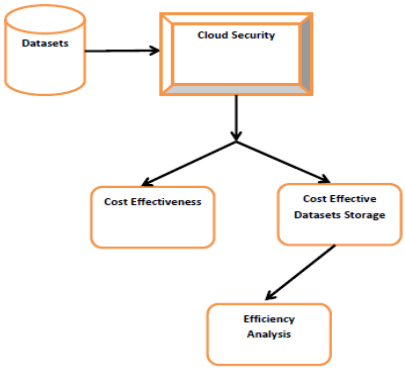


Fig 1: System Architecture Design

PROPOSED SYSTEM

Typically, operators will know neither the particular location of their data the other sources of the data together stored theirs. The data you can find in a cloud from public source, which has minimal security fears, to private data comprising extremely searching

information. It using a cloud environment improves the business entities of their concern to ensure that proper security events in place for both their data and applications. The answers to this and other questions lie within the realm with most technological advances, controllers. Cloud computing presents an extension of problems practiced with the Internet. To ensure that such results are informed and proper for the cloud computing environment, the industry itself should establish clear and effective policy and governance to identify and implement proper security methods. We show that the quality of instruction sets from the Apriori algorithm for suggestion regulation mining can be enhanced by using Ant Colony Optimization (ACO)

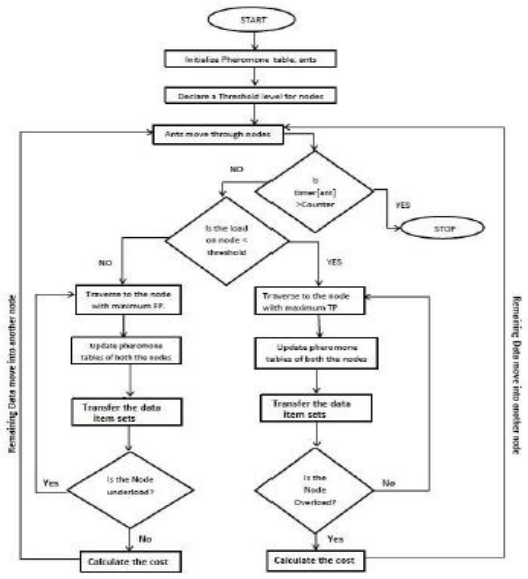


Fig 2: Flow Chart of ACO

Probability can be calculated by

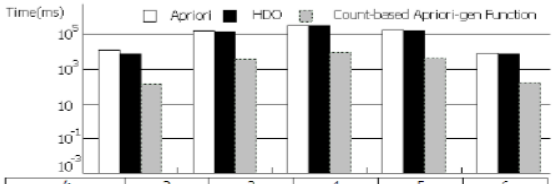
$$P^{k}_{xy} = \frac{(\tau^{\alpha}_{xy})(\eta^{\beta}_{xy})}{\Sigma_{ye\,allowed}\,(\tau^{\alpha}_{xy})(\eta^{\beta}_{xy})}$$

The trails are updated by

$$\tau_{xy} \leftarrow (1-\rho)\tau_{xy} + \sum_{k} \Delta\tau^{k}_{xy}$$

RESULTS AND DISCUSSION

From the prune operation graph, pruning the datasets by using apriori algorithm will give more accurate value than other algorithms like count-based apriori-gen function & HDO algorithms. This pruning effect stores only the most frequent datasets & thereby reduces the cost of datasets to a great extent. Thus pruning in apriori algorithm also play a major role in reducing the overall cost of data transfer and storage.



k	2	3	4	5	6
Itemsets number of L_{k-1}	47	701	2169	1394	113
Itemsets number of L_{k-1}	3	380	8693	20711	7631
candidate number before prune	1081	10862	22105	7744	359

Fig 3: Prune Operation Comparison

From the below graph, the overall cost against the number of datasets is depicted & it varies related to many factors. From the graph, we can see that the cost is increased in the middle of the graph. It is due to the reason of moving datasets from one cloud to another cloud, if one cloud could not support the whole datasets of a particular application. Though the data transfer cost is reduced, there may be extra cost required to work on different clouds.

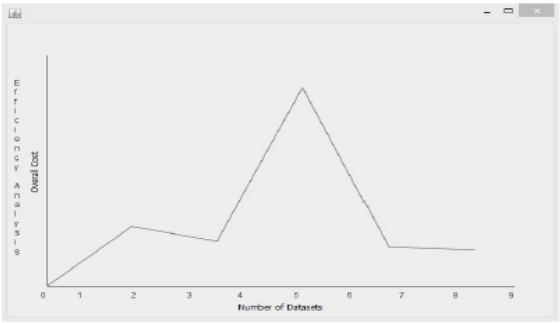


Fig 4: Efficiency Analysis

CONCLUSION AND FUTURE WORK

The research on cloud computing becomes popular. This project is mainly inspired by the work in two research areas: Cache management and scheduling. With smart caching mechanism system performance can be greatly improved. The similarity is that we both pre store some data for future use. The difference is that storing is to reduce data retrieving delay but our work is to find the minimum application cost for data storage. Research in development focuses on reducing various "costs" for applications systems or data center networks. However, it is mainly on resource utilization while ours on the compromise between calculation and storage Investigators are exploring the cost-effectiveness of using cloud. This is because comparing to the outdated distributed computing systems like cluster and cloud computing system has cost assistances in several aspects. The work in analyzes the cost benefits of cloud computing versus grid computing. The work also applies clouds' cost model and demonstrates that cloud computing offers a cost effective way to deploy technical applications. The work stated above mostly efforts on the assessment of cloud computing systems and the outdated detached computing decorations which shows that submissions successively in the cloud have cost benefits. They did not touch the issue of computation and storage interchange in the cloud.

REFERENCES

[1] D. Yuan, Y. Yang, X. Liu, G. Zhang, and J. Chen, "A Data Dependency Based Strategy for Intermediate Data Storage in Scientific Cloud Workflow Systems," Concurrency and Computation:Practice and Experience, vol. 24, pp. 956-976, 2012.

[2] Z I. Adams, D.D.E. Long, E.L. Miller, S. Pasupathy, and M.W. Storer, "Maximizing Efficiency by Trading Storage for Computation,"Proc. Workshop Hot Topics in Cloud Computing, 2009.

[3] J. Csorba, "Ant system for service deployement in private and public clouds," Proceeding of the 2nd workshop on Bio-inspired algorithms for distributed systems(BADS), Jun. 2010.

[4] D. Yuan, Y. Yang, X. Liu, and J. Chen, "On-Demand Minimum Cost Benchmarking for Intermediate Data Sets Storage in Scientific Cloud Workflow Systems," J. Parallel and Distributed Computing, vol. 71, pp. 316-332, 2011.

[5] B. Soumya, M. Indrajit, and P. Mahanti, "Cloud computing initiative using modified ant colony framework," in In the World Academy of Science, Engineering and Technology 56, 2009.

[6] B. Rajkumar, B. Anton, and A. Jemal, "Energy efficient management of data center resources for computing: Vision, architectural elements and open challenges," in International Conference on Parallel and Distributed Processing Techniques and Applications, Jul. 2010.

[7] D. Yuan, Y. Yang, X. Liu, and J. Chen, "A Local-Optimisation Based Strategy for Cost-Effective Data Sets Storage of Scientific Applications in the Cloud," Proc. IEEE Fourth Int'l Conf. Cloud Computing, pp. 179-186, 2011.

[8] S. Agarwala, D. Jadav, and L.A. Bathen, "iCostale: Adaptive Cost Optimization for Storage Clouds," Proc. IEEE Int'l Conf. Cloud Computing, pp. 436-443, 2011.

[9] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing," Comm. ACM, vol. 53, pp. 50-58, 2010.

[10] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic,"Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the fifth 3.Utility," Future Generation Computer Systems, vol. 25, pp. 599-616, 2009.

[11] D. Kondo, B. Javadi, P. Malecot, F. Cappello, and D.P. Anderson, "Cost-Benefit Analysis of Cloud Computing versus Desktop Grids," Proc. 23th Int'l Parallel and Distributed Processing Symp., 2009.
[12] X. Liu, D. Yuan, G. Zhang, W. Li, D. Cao, Q. He, J. Chen, and Y. Yang, The Design of Cloud Workflow Systems. Springer, 2012.
[13] D. Warneke and O. Kao, "Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud," IEEE Trans. Parallel and Distributed Systems, vol. 22, no. 6, pp. 985-997, June 2011. [14] L. Young Choon and A.Y. Zomaya, "Energy Conscious Scheduling for Distributed Computing Systems under Different Operating Conditions," IEEE Trans. Parallel and Distributed Systems, vol. 22, no. 8, pp. 1374-1381, Aug. 2011.
[15] M. El-Hajj and O. Zaiane. Parallel leap: large-scale maximal pattern mining in a distributed environment. In Parallel and Distributed Systems, 2006. ICPADS 2006. 12th International Conference on, volume 1, page 8 pp., 0-0 2006.
[16] Nada M.A. AL-Salami, "System Evolving using Ant Colony Optimization Algorithm ", Journal of Computer Science 5 (5): 380-387, 2009, ISSN 1549-3636.
[17] X. Zhong and C.-Z. Xu, "Energy-Aware Modeling and Scheduling for Dynamic Voltage Scaling with Statistical Realtime Guarantee," IEEE Trans. Computers, vol. 56, no. 3, pp. 358-372, Mar. 2007.
[18] A. Li et al., "Cloudcmp: shopping for a cloud made easy," in Proc of the 2nd USENIX conference on Hot topics in cloud computing, 2010.
[19] H. Topcuouglu, S. Hariri, and M.-Y. Wu, "Performance-Effective and Low-Complexity Task Scheduling for Heterogeneous Computing," IEEE Trans. Parallel and Distributed Systems, vol. 13, no. 3, pp. 260-274, Mar. 2002.
[20] S.C. Wang, K.Q. Yan, W.P. Liao and S.S. Wang, Towards a Load Balancing in a Three-level Cloud Computing Network, Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology, pp. 108-113,2010.
[21] Z. Zhang and X. Zhang, A Load Balancing Mechanism Based on Ant Colony and Complex Network Theory in Open Cloud Computing Federation, Proceedings of the 2nd International Conference on Industrial Mechatronics and Automation, pp. 240-243, 2010.
[22] C. Gong, J. Liu, Q. Zhang, H. Chen and Z. Gong, The Characteristics of Cloud Computing, Proceedings of the 39th International Conference on Parallel Processing Workshops, pp. 275-279, 2010.
[23] C. Wang, K. Ren, W. Lou, and J. Li, "Towards Publicly Auditable Secure Cloud Data Storage Services," IEEE Network Magazine, vol. 24, no. 4, pp. 19-24, July/Aug. 2010.
[24] C. Wang, Q. Wang, K. Ren, and W. Lou, "Towards Secure and Dependable Storage Services in Cloud Computing," IEEE Trans. Service Computing, vol. 5, no. 2, 220-232, Apr.-June 2012.
[25] C. Erway, A. Kupcu, C. Papamanthou, and R. Tamassia, "Dynamic Provable Data Possession," Proc. ACM Conf. Computer and Comm. Security (CCS '09), pp. 213-222, 2009.
[26] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," NIST Special Publication 800-145, 2011.
[27] S. Di, C.-L. Wang, W. Zhang, and L. Cheng, "Probabilistic Best-Fit Multi-Dimensional Range Query in Self-Organizing Cloud," Proc. IEEE 40th Int'l Conf. Parallel Processing, pp. 763-772, 2011.