A Review of Replication Strategies to Increase Data Availability for Data Intensive Applications in Cloud

K. Sreelatha

Abstract: ***To gain continuous access to the services, users of cloud computing makes benefit from a technique called replication. Replication technique is used to maintain multiple copies of same data at different sites. Based on ever growing user's needs, high demand data will be replicated to various sites to provide better performance, high availability and to improve fault tolerance. This paper provides a better insight into data replication strategies provided for data availability by considering different parameters like energy consumption, bandwidth optimization and reducing access latency involved during the execution of tasks in cloud data centers for data intensive applications.***

Keywords: replication, data availability, data intensive applications

1. Introduction

Nowadays data-intensive applications play a prominent role in different scientific fields.These applications produce huge amounts of data. This ever-growing amounts of data affect the systems' performance. So cloud schedulers have to use data replication strategies for executing the incoming jobs for data intensive applications. To enhance the complete performance now different areaslike the distributed databases, Internet and P2P systems use data replication technique.

In cloud computing as the data resides at geographical locations and enormous growing data leads to data unavailability problem especially for scientific applications. Data replication is acommon technique used to maintain these ever growing data by placingmore number ofcopies of same data in geographicaldiversifiedlocations that can be called as replicas and is shown in figure below.Data replication techniques can greatlyincrease data availability, reduce the user waiting time andminimize bandwidth consumption by usingmultiple replicas of the same data file. Data replication techniques aredivided into two main strategiesas static and dynamic replication strategies.Static replication strategiesdoes notadapt to changes in user behavior. Where dynamic strategies make immediate decisions to create and remove replicas according to user's needs. It makes instantdecisions about the location of replica copies.Due to this it is difficult to maintain consistency in between replicas is hard in dynamic data replication. But due to its dynamic nature cloud users can gets benefit of dynamic replication strategies when compare to static replication strategies.Static and dynamic replication algorithms againdivided into two groups as centralized and distributed.

Depending on user access patterns, different replication strategies are born.

No Replication: the replicas are placed at root node.

Best Client: file is replicated and stored at the client that frequently requests the data file.

Cascading: file is replicated and a copy is stored on the path of the best client.

Plain Caching: replica copy is placed locally when a user requests a file.

Caching plus cascading: it combines together cascading and plain caching strategies.

Fast Spread: replicas are placed at every node towards the path to the best client.

Goals of replication

- Decrease access time
- Reduce bandwidth consumption
- Work load balancing
- Minimize execution time
- Maintenance cost optimization
- Achieve high availability and scalability
- Fault tolerant

The following sections contains as: section 3issues in replication, Section 4illustrates literature review, Section 5 gives comparative study, section 6 contains limitations of existing work, Section 7includes conclusion and future work.

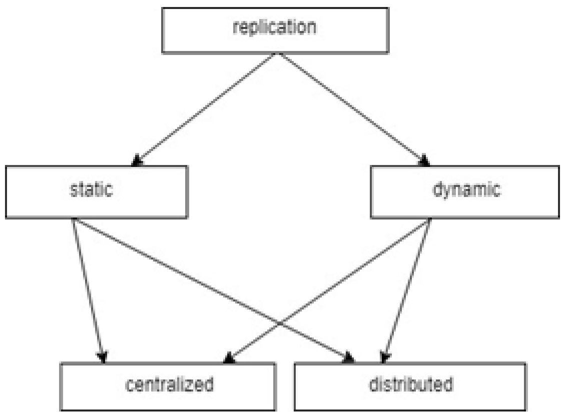

Figure 1: Replication strategies

Issues involved in data replication

- **Data Consistency:** Maintaining data consistency and integrity is most important in a replicated environment.

Severe consistency updates may require by prior applications during execution times.

- **Downtime during new replica creation:** Using strict consistency rules makes performance degradationduring the time of new replica creation. Due to consistency needs sites may not fulfill the request.
- **Maintenance overhead:** It occupies more storage space and requires more additional maintenance if the copies are replicated tovarioussites. This become creates overhead in storing numerous copies of the data files.
- **Lower write performance:** in replicated environment performance of write operationsparticularly applications having high update ratescan be lower,because during the transaction it may need to update more number of replicas.

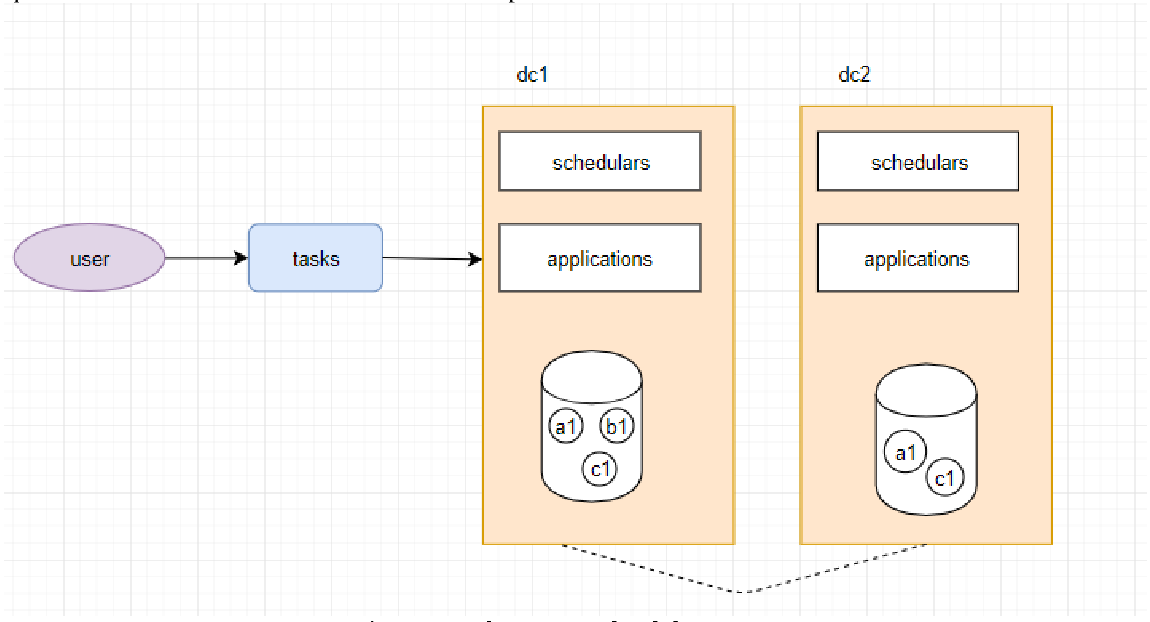

Figure: Replication in cloud data centers

2. Literature Review

DejeneBoru et al. [1] proposed a data replication strategy that can be used in cloud data centers.They consider togetherbandwidth consumption and energy efficiency of the computingnodes in datacenters.By replicating the data closure to the applications it achieves network bandwidth availability and optimization of energy consumption in datacenters. They proposed a mathematical model and huge simulations to present energy efficiency and performance tradeoffs.

Wenhao LI et al. [2] proposed a novel dynamic cost-effective incremental replication (CIR) to reduce the cost of storage and at the same time it meets the data reliability requirement.It dynamically replicates the intermediate data that is not much longer use for a computation, to increase reliability in data centers with minimum cost.It considers both data duration uncertainty factors and storage variation into account.

Jun Li et al. [3] proposed a consistency maintenance replication strategy called CMTree to promise availability and reliability of the system.Super nodes of different groups exchange messages by using chord protocol as the nodes are divided into groups. According to update rate consistency maintenance tree is used for replicas of same file. It avoids node failures and propagates that update messages will transfer timely to other nodes in the group.

Based on diamond topology LUO Siwei et al. [4] proposed a novel consistency maintenance strategy. By this all the nodes in cloud storage system can formulate into a high reliable, symmetrical structure to enhance the reliability and to reduce the network overhead compared to tree topology.

Mazhar Ali et al. [5] proposed Division and Replication of Data in the Cloud for Optimal Performance and Security (DROPS) that approaches the performance and security issues. It can fragment the file and the fragments are stored over multiple nodes, so that in case of threats no successful information was revealed to the attackers. In a cluster by using a T-coloring technique nodes are separated.

In this paper, Muhannad Alghamdi et al. [6]study the file replication problem (FRP) for data intensive cloud applications.A time-efficient approximation algorithmis used to minimize the overall energy consumption in data centers at the time of data file access. It works on profit and performance base.

N. Mansouri et al. [7] proposed Dynamic Popularity aware Replication Strategy (DPRS). Depends on popularity it replicates the file to different cluster sites and it places only a small portion of file fragment to a cluster. So that DPRSprovides parallel downloadingmethod to reduce download time during execution where the file fragments are resides at several locations.

Navneet Kaur Gill et al. [8] presents a dynamic, cost-aware, optimized data replication and re-balancing strategy (DCR2S). It can mainly focus on minimum number of replicas needed to ensure desired availability. It can re-replicate replicas from super (higher cost) data centersto ordinary (lower cost) data centers by using knapsack algorithm without compromising data availability.

Bahareh Rahmati et al. [9] propose an integrated job scheduling and data replication strategy that can enhance the performance of the data intensive applications. It schedules the jobs to appropriate VMs and replicate the files in data center and replacement of the jobs with arrival of new jobs due to the lack of storage capacity. It achieves better response time and load balancing of the system.

3. Comparative Study

Table 1: Comparative analysis of replication strategies for data availability in data intensive applications

Mechanism	Band width	Response time	Energy efficiency	Load balancing	reliability	consistency	cost	Fault tolerance
Boru et al.[1]	✓	✓	✓	✗	✗	✗	✓	✓
Wenhao Li et al.[2]	✗	✗	✗	✗	✓	✗	✓	✓
Jun Li et al.[3]	✓	✗	✗	✗	✓	✓	✓	✓
Luo Siwei et al.[4]	✗	✗	✗	✓	✓	✓	✓	✓
Mazhar Ali et al.[5]	✓	✓	✗	✓	✓	✗	✗	✗
MuhannadAlghamdi et al.[6]	✓	✓	✓	✗	✓	✗	✓	✗
N. Mansouri et al.[7]	✗	✓	✗	✗	✗	✓	✗	✓
Navneet Kaur Gill et al.[8]	✗	✓	✗	✗	✗	✗	✓	✗
Bahareh Rahmati et al.[9]	✓	✓	✓	✓	✓	✓	✓	✓

Limitations of the existing algorithms

However, a lot of algorithms have been presented in data-intensive applications' field. Most of these algorithms consider job scheduling and data replication as two independent techniques for boosting the cloud systems' performance. In one hand, scheduling the jobs without replicating the data files that are required for them needs remote access to data files. Accessing remotely to these data files requires more time than accessing directly so; doing scheduling without replication imposes an overhead of data access time to the system. On the other hand, doing replication without scheduling the jobs fails to enhance the performance of the system in an effective way because moving large-sized data files costs more bandwidth and needs longer time to transfer.

4. Conclusion

This paper reviewed different data replication strategies that are in the domain of cloud computing. From our study it is observed that most of the existing strategies consider job scheduling and data replication as two independent techniques to enhance the performance of cloud systems. This may not most considerable for data intensive applications where effective job scheduling algorithm reduces job response time and data replication reduces data access time and bandwidth consumption.So as to reduce access latency involved in moving data files from geographical locations and to enhance the system performance there is a need to integrate both scheduling and replication techniques. As a future work we are going to focus on develop a strategy by combining both scheduling and replication techniques.

References

[1] Boru, Dejene, et al. "Energy-efficient data replication in cloud computing datacenters." Cluster computing 18.1 (2015): 385-402.

[2] Li, Wenhao, Yun Yang, and Dong Yuan. "A novel cost-effective dynamic data replication strategy for reliability in cloud data centres." Dependable, autonomic and secure computing (DASC), 2011 IEEE ninth international conference on.IEEE, 2011.

[3] Li, Jun, and Mengshu Hou. "A novel replication consistency maintenance strategy in cloud storage system."Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2017 IEEE 2nd.IEEE, 2017.

[4] Luo, Siwei, et al. "Consistency maintenance in replication: A novel strategy based on diamond topology in cloud storage." Chinese Journal of Electronics 26.1 (2017): 192-198.

[5] Ali, Mazhar, et al. "Drops: Division and replication of data in cloud for optimal performance and security." IEEE Transactions on Cloud computing 6.2 (2018): 303-315.

[6] Alghamdi, Muhannad, Bin Tang, and Yutian Chen. "Profit-based file replication in data intensive cloud data centers."Communications (ICC), 2017 IEEE International Conference on.IEEE, 2017.

[7] Mansouri, Najme, M. Kuchaki Rafsanjani, and Mohammad M. Javidi. "DPRS: A dynamic popularity aware replication strategy with parallel download scheme in cloud environments." Simulation Modelling Practice and theory 77 (2017): 177-196.

[8] Gill, Navneet Kaur, and Sarbjeet Singh. "A dynamic, cost-aware, optimized data replication strategy for heterogeneous cloud data centers." Future Generation Computer Systems 65 (2016): 10-32.

[9] Rahmati, Bahareh, Amir MasoudRahmani, and Ali Rezaei. "Data replication-based scheduling in cloud computing environment." Journal of Advances in Computer Engineering and Technology 3.2 (2017): 75-80.