International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 16-19

Comparative Study of Remote Replication techniques

Suhasi P M[1], Khadija Fatima[2], Dr. Rekha P M[3]
[1, 2]*(Final year Students, Department of Information Science and Engineering, JSSATE, India Bangalore)*
[3]*(Associate Professor, Department of Information Science and Engineering, JSSATE, India Bangalore)*

Abstract: Remote replication in storage networks is an important area of research. Systems should be sufficiently enabled to keep data secure. The result is efficient access and storage which prevents unpredictable damage and ensures data security for successful business. In this research paper we have discussed about the replication methods that can be implemented depending on the requirements of the business. We have recognized the importance of remote replication and made a comparison between synchronous and asynchronous replication techniques.

Keywords - Replication, synchronous, asynchronous, storage, backup

I. Introduction

Data today, is probably the most valuable resource to almost all businesses and enterprises. Almost all the decisions a company takes revolves around the data the company has in its hands. For most businesses today, protecting valuable data proves to be a very big challenge. This data loss may be due to a number of factors including viruses, physical damage, formatting errors, etc. A loss of vital data can cause major inconveniences toa company's day to day functioning and can sometimes be serious enough to cripple the company or even put it out of business. Modern businesses across the globe are expanding with every passing second and along with it so is the enormous amount of valuable data that needs to be protected from possible disasters. As more and more companies as people grow more conscious of the importance of their company's data, protecting it is becoming more important. It is mandatory for business organizations where high availability and business continuity are of prime importance, to have recovery strategies in place so that the mission critical data and applications are protected and restored in the event of any failure as soon as possible.

II. Backup and Replication

Most companies and businesses took to various backup strategies as a way to prevent data losses. Backup involves making copies of the data in case the original were to get lost. Data backup even today is very popular as it is relatively inexpensive. In most enterprises' backup is used typically for everything from production servers to desktops. They rely on snapshot-based technologies which is why they are taken quite infrequently typically once every few days. This means that the potential data loss that might occur to the business could be days or even more which is simply unacceptable especially for the applications that matter to the business. You can say that the solution to this is to just take more frequent copies of the data which is definitely possible but the problem is that, taking frequent copies of data comes with a cost in terms of server resources and also significantly affects user efficiency. With less frequent backups protection is not being expanded to include additional data and applications that the Users are constructing. So even though backups are a low-cost approach to preserve some level of security, but today's digital customers want an always-on level of service that backups simply cannot give. So, in order to meet these expectations, businesses have moved beyond backups and have started considering data replication strategies. Modern replication strategies offer more than just a rapid disaster recovery. They can help with cloud migration, using the cloud as a Data Recovery site and even solving copy data challenges.

The basic idea behind data replication involves creating replicas of transferring information from one storage place to another This can be done via cloud-based services between two on-premises sites, between sites in different locations, or between sites that are fully geo-physically separated.

The following are some of the ways that data replication:

1. Enhances performance. reducing latency, since users can access replicated data, so it avoids remote network access; and
2. Since data is being processed at a faster rate, throughput must be increased and applications are accessible on a variety of platforms computers and can be accessed at the same time.

In the next section we discuss about remote replication strategy.

III. Remote Replication

Remote replication is a popular data replication strategy which involves sending the business-critical data into a remote offsite location for reliable storage and fast recovery. By providing a continuous, disruption less, host-independent solution Data backup or migration via a long-distance distant replication becomes a crucial aspect of data security in the event of a disaster any collapse at the primary site. Remote replication copies data across storage pools or systems on multiple sites in an IP or FC SAN by leveraging the power of storage systems. If at all some source site application owing to a problem with the system or accident users can resume service in minutes by using the disk-based remote copy. Remote replication allows customers to safeguard the remote copy using snapshot technology, which further ensures the integrity of the data. When it's time to restart business service, granular snapshot pictures can assist in restoring the faulty remote copy in seconds. Earlier, replication was used. Mostly for off-site copying and storage of application data. As time goes on, however, this technology has advanced tremendously and we can now construct synchronized databases via replication copies of virtual machines at a remote target location. The copy of the virtual machine is called a replica and functions just like a regular virtual machine which is available at the primary site. These virtual machine replicas can be transferred and run on any capable hardware. In a matter of seconds, they may be turned on in the original virtual machine encounters some failure. This technology therefore significantly decreases the downtime as well as mitigates business risks and losses that might occur with catastrophe. Remote duplicate allows users to choose between two techniques, (i) Synchronous replication and (ii) Asynchronous replication. In the next section we give a description of Synchronous replication.

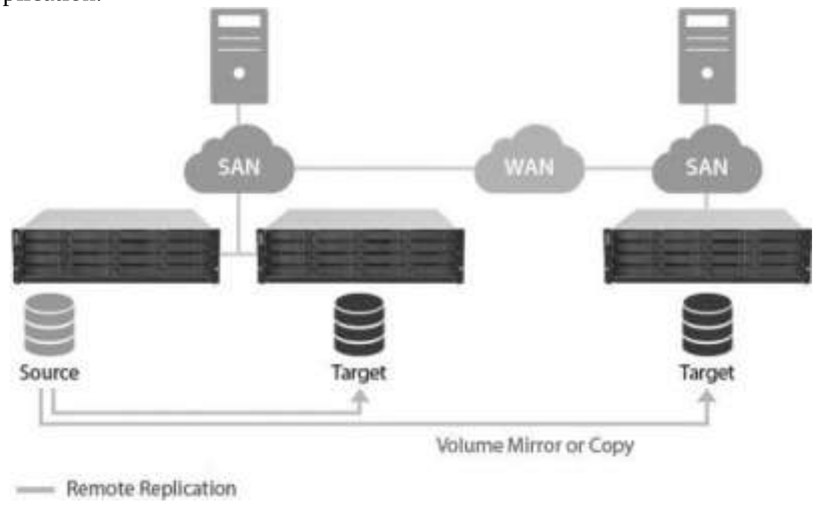

Fig 1: Remote replication between a source site and target site

IV. Synchronous Replication

In today's IT corporations, synchronous replication is profoundly wanted as any information misfortune may have critical financial impact on the association. Synchronous replication ensures consistent information uprightness during the replication cycle with no additional danger of information loss or leakage. In synchronous replication any update or modification in the primary database, information is replicated to the remote secondary location simultaneously. In essence, the host and target sources continue to be fully synchronized, that is essential for effective disaster recovery (DR). Before data is to be replicated every write I/O operation must be acknowledged. The write I/O activity confirmation must be shipped off to the application only after the write I/O operation is accepted by the storage subsystem at both the local and remote site. Only when data is committed on both the sites the write operation is said to be complete. The storage subsystem must wait for the remote site I/O cycle to complete prior to reacting to the application leading to higher response time to the application. This is one major drawback of synchronous replication as accuracy is strongly based on factors such as line bandwidth and latency. The key elements that affect the measure of data loss are the rate of change (ROC) of information on the primary site, the connection speed, and the distance between locales. Protection against corruption of data and loss of information due to human errors are not guaranteed. For this synchronous replication along with snapshot technology can be used to ensure complete protection against data loss and corrupted data.

V. Asynchronous Replication

In this mode of replication, write operation is achieved at primary site and replication isn’t performed simultaneously to the remote storage system. Replication of data takes place in later time-frames which could be daily weekly or hourly. In asynchronous replication because there is delay in transmission as data is not replicated in real-time with the primary site, exact copy of data is not achieved. The Disaster Recovery (DR) location may not have the latest modification of data so some important data could be missing which is one down side of asynchronous replication. It doesn't provide a similar degree of security as synchronous replication. Network efficiency can be increased without impacting bandwidth.

This mode of replication is best suited for disk to disk backups or offline backups. Few enterprise hypervisors incorporate asynchronous replication mechanism to facilitate the replication of whole virtual machines (VMs) to a secondary location such that in case of disaster the VMs may crash over to that destination. This is known as recovery-in-place or instant recovery. In the next section a comparison is drawn between synchronous and asynchronous modes of replication.

VI. Comparitive Analysis

Few important parameters on which the comparative analysis is based on:

i. ROC (Rate of Change) - It is the amount of information on the data volume to be replicated to the secondary site that is changed over a certain time period. It can be expressed as average over certain duration such as day or even as peak value. The average value is used for sizing links in asynchronous replication and peak value is to determine the sizing link in synchronous replication. For asynchronous replication the storage subsystem must have the capacity to support peak ROC's in queue, else the write speed to the severe will slow down.

ii. RPO (Recovery point objective) - It is the maximum time between data backup from the primary location to the remote location. RPO represents the point to which the data must be moved back for the purpose of recovery of data. By evaluating how much data can be lost if a process is interrupted, i.e. the acceptable volume of data loss along with ROC and replication speed, we can calculate the true RPO value.

iii. RTO (Recovery time objective) - The time taken to retrieve the data after a possible failure.

iv. Bandwidth- It is expressed in bits per second (bps) and is the maximum data that can be transferred over a network connection at a specific time.

v. Line-Latency – It is the amount of time a data packet takes to propagate over the network from one location to the destination and for the acknowledgement to be received by the sender. Lower the latency higher is the network efficiency.

To answer the question as to which mode of replication is preferable, it completely depends on the enterprise requirements which vary from business to business. In the case of synchronous replication, it is the preferred technique when secure data and storage for long term is required and when no organization can afford to lose vital information.

Asynchronous replication is good with long-distance projects with a limited budget Next section we discuss briefly a technique using which we can switch between synchronous and asynchronous.

TABLE 1: Comparative analysis between synchronous and asynchronous replication.

Sl. No.	**Parameter**	**Synchronous replication**	**Asynchronous replication**
1	Distance	Works better when target site is close to primary site. Performance is inversely proportional to distance.	Performance is not affected by distance as long as any network is available between the two sites.
2	Data loss	No possible data loss as data is replicated at real time.	Possible loss of recently updated data.
3	RPO	Zero RPO	Small RPO value (varies from minutes to hours)
4	RTO	Minimal RTO value	Minimal RTO value
5	Bandwidth	Requires high bandwidth	Requires less bandwidth
6	Line-Latency	Requires low latency	No dependence on latency
7	Resilience	Single failure can cause the entire service to stop as data is applicationreplicated at real time.	Minimum of two failures can causedisruption of service.
8	Cost	More expensive	Cost effective as requirements are less
9	Performance	Since acknowledgement is required it has low performance.	Since acknowledgement is not required it has high performance.
		Good solution for projects which require	It is useful for storing data that

10	Use cases	immediate recovery of sensitive data.	is less sensitive and for enterprises that can tolerate partial loss of data.

VII.Conclusion and Future Scope

Current disaster recovery techniques compel users to make a hard decision between choosing synchronous or asynchronous replication. So, consumers are required to compromise either rational system performance or possibility of significant data loss. Assaf Natanzon and Eitan Bachmat have proposed a replication system which is capable of switching dynamically between synchronous and asynchronous modes of replication thereby allowing consumers to relish good performance and also reduce the possibility of data loss. Users may favor synchronous replication but might prioritize not delaying IO operation for more than a certain time limit. The use of synchronous replication may decrease the user's maximum throughput and IOPS and may make batchprocessing and data warehousing operations much more time consuming. Dynamic replication allows the systemfor the time being to move out of synchronous replication so that the performance of batch processing improves significantly.

Future implementations to this technique can be including a scheduler which will prevent the system from leaving synchronous mode at working hours when mission critical data is being handled. As a result of this feature, the system can run in asynchronous replication only during non-working hours when data entering is less significant. More research can be done to make sure there is minimal lag when switching from asynchronous to synchronous replication mode.

References

[1] Natanzon, A., & Bachmat, E. (2013). Dynamic synchronous/asynchronous replication. ACM Transactions on Storage (TOS), 9(3), 1-19.

[2] "The Art of Data Replication" An Oracle Technical White Paper September 2011

[3] Mirzoev, T. (2009). Synchronous replication of remote storage. Journal of Communication and Computer, 6(3), 34-39.

[4] Domingues, H. H., Kon, F., & Ferreira, J. E. (2011, October). Asynchronous replication for evolutionary database development: a design for the experimental assessment of a novel approach. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 818-825). Springer, Berlin, Heidelberg.

[5] Liu, X., Zhao, Z., Wang, G., Sui, J., & Song, Y. (2009, December). ELVM: A LVM-Based remote replication system. In 2009 International Forum on Computer Science-Technology and Applications (Vol. 2, pp. 163-166). IEEE.

[6] Yadav, A. K., Agarwal, A., & Rahmatkar, S. (2011, April). An Efficient Approach for Data Replication in Distributed Database Systems. In International Conference on Advances in Information Technology and Mobile Communication (pp. 368-374). Springer, Berlin, Heidelberg.

[7] Aronovich, L., Asher, R., Bachmat, E., Bitner, H., Hirsch, M., and Klein, S. T. 2009. The design of a similarity based deduplication system. In Proceedings of the Israeli Experimental Systems Conference (SYSTOR'09).

[8] Azagury A., Factor, M., and Micka W. 2003. Advanced functions for storage subsystems: Supporting continuous availability. IBM Syst. J. 42, 2, 268--279.

[9] Chandler D. Worldwide Storage Services 2008–2012 Forecast. Storage and Data Management Services, IDC, 2008.

[10] Highleyman B., Holenstein P., Holenstein B. Replicating Applications for Disaster Recovery. 2003.

[11] Routray, A. V. K. V. R., & Jain, R. (2008). Sweeper: an efficient disaster recovery point identification mechanism.

[12] Heger, D., & Shah, G. (2001). IBM®'s General Parallel File System (GPFS) 1.4 for AIX®. International Business.

[13] Sleit, A., AlMobaideen, W., Al-Areqi, S., & Yahya, A. (2007). A dynamic object fragmentation and replication algorithm in distributed database systems. American Journal of Applied Sciences, 4(8), 613-618.

[14] Loukopoulos, T., & Ahmad, I. (2004). Static and adaptive distributed data replication using genetic algorithms. Journal of Parallel and Distributed Computing, 64(11), 1270-1285.

[15] Abdul-Wahid, S., Andonie, R., Lemley, J., Schwing, J., & Widger, J. (2007, March). Adaptive distributed database replication through colonies of pogo ants. In 2007 IEEE International Parallel and Distributed Processing Symposium (pp. 1-8). IEEE.