e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704011 |

Secure Distributed Deduplication Systems with Improved Reliability

Dr.N.Mahendiran[1], Naresh Kumar A[2]

[1]Assistant Professor, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore, India

[2]UG Student, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore, India

ABSTRACT: Distributed Systems is the advanced model of virtualization technique which shares the available resource such as storage & processor of the various machine distributed globally and connected using internet. Distributed Systems envisions the notion of delivering software services and customizable hardware configurations to public access. Cloud based multimedia distribution system uses the code generation system to check the duplicate copy of the files. For each multimedia content signature code is generated and stored in the cloud data base. Generated code occupies large space in case of multimedia content as same size of file. The system can be used to reduplicate different multimedia content types including 2D videos, 3D videos, images, audio clips, songs, and music clips. The system can be deployed on private and/or public clouds. Our system has two novel components: (i) method to create signatures of 3D videos, and (ii) distributed matching engine for multimedia objects. The signature method creates robust and representative signatures of 3D videos that capture the depth signals in these videos and it is computationally efficient to compute and compare as well as it requires small storage. The distributed matching engine achieves high scalability and it is designed to support different multimedia objects. In order to reduce the space usage, duplicated file copies are identified by matching redundant bits of the data. This will reduce the content matching time and automatically increases the performance in terms of speed.

KEYWORDS: DoS, security, Distributed system, cloud storage

I. INTRODUCTION

The cloud abstracts infrastructure complexities of servers, applications, data, and heterogeneous platforms, enabling users to plug-in at anytime from anywhere and utilizes storage and computing services as needed at the moment. The goal of mobile multimedia cloud is to provide infrastructure as a service (IaaS) and platform as a service (PaaS) for diverse services and applications in the domain of (mobile) multimedia and large-scale social network analysis. Distributed Systems has evolved through a number of phases which include grid and utility computing, application service provision (ASP), and Software as a Service (SaaS). Distributed Systems is a broad term that describes a broad range of services. As with other significant developments in technology, many vendors have seized the term “Cloud” and are using it for products that sit outside of the common definition. In order to truly understand how the Cloud can be of value to an organization, it is first important to understand what the Cloud really is and its different components. Since the Cloud is a broad collection of services, organizations can choose where, when, and how they use Distributed Systems. In this report we will explain the different types of Distributed Systems services commonly referred to as Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) and give some examples and case studies to illustrate how they all work. We will also provide some guidance on situations where particular flavors of Distributed Systems are not the best option for an organization.

Cloud storage is one of the primary use of cloud computing. With the cloud storage, data is stored on multiple third party servers, rather than on the dedicated servers using traditional networked data storage. When storing data the user sees a virtual server i.e. it appears as if the data is stored in a particular place with specific name. But that place does not exist in reality.

The actual server location may differ from day to day or even minute to minute as the cloud dynamically manages available storage space. But even though the location is virtual , user sees a static location for her data and can actually manage storage space as if it was connected to her own PC. This is a concern for almost every person who

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704011 |

backs up their data to the cloud or share's information with friends and families over the network. In the case of some businesses, cloud service providers might be the preferred method to back up files in the event of a disaster or hardware failure. This is why many service providers and enterprises rely upon the method of deduplication to keep storage costs in check. Expanding upon my last blog on the business benefits of deduplication, let's dive into the details of how this technology works by looking at the different compute methods vendors are using within their dedupe offerings.

Distributed deduplication systems have gained significant attention in recent years due to their potential to address the challenges of redundant data storage, data security, and system reliability. Several studies have explored various aspects of these systems, including their design principles, security mechanisms, reliability enhancements, and practical implementations. Chang et al. (2014) proposed a distributed deduplication system that integrates erasure coding for data redundancy and fault tolerance. Their work focused on optimizing storage efficiency while ensuring data reliability in distributed environments. By leveraging erasure coding techniques, the system achieved high levels of fault tolerance without sacrificing storage efficiency.

II. RELATED WORK

In terms of security, Zhang et al. (2016) introduced a secure distributed deduplication scheme that utilizes convergent encryption to protect data confidentiality. Their approach ensures that identical data chunks are encrypted with the same key, enabling efficient deduplication while preventing unauthorized access to sensitive information. Authentication mechanisms were also integrated to verify the integrity and authenticity of data.Reliability enhancements in distributed deduplication systems have been addressed by Li et al. (2018), who proposed a fault-tolerant deduplication scheme based on distributed consensus algorithms. By employing Paxos consensus protocol, their system achieved strong consistency guarantees across distributed nodes, even in the presence of node failures or network partitions.

Scalability concerns have been investigated by Wang et al. (2020), who proposed a scalable distributed deduplication system capable of handling large-scale datasets and high throughput workloads. Their approach leveraged parallel processing techniques and distributed indexing mechanisms to efficiently manage deduplication tasks across multiple nodes, enabling seamless scalability as the system grows.In terms of compliance and regulations, several studies have addressed the impact of data privacy laws on distributed deduplication systems. For instance, Li and Li (2019) examined the implications of GDPR on data deduplication practices and proposed privacy-preserving techniques to ensure compliance with regulatory requirements while maintaining deduplication efficiency.

Continuous improvement and optimization of distributed deduplication systems have been a focus of research as well. Jiang et al. (2021) introduced a self-adaptive deduplication mechanism that dynamically adjusts deduplication parameters based on system workload and resource availability. Their approach improved system performance and resource utilization by adaptively tuning deduplication thresholds and strategies.

Another significant aspect of distributed deduplication systems is their integration with cloud computing environments. With the increasing adoption of cloud storage services, researchers have explored methods to optimize deduplication mechanisms for cloud-based architectures. For instance, Li et al. (2017) proposed a hybrid cloud deduplication scheme that leveraged both client-side and server-side deduplication to minimize data transfer overhead and improve performance in cloud storage environments. Their approach aimed to strike a balance between data locality and global deduplication efficiency, addressing the unique challenges posed by cloud computing paradigms.

Moreover, the application of machine learning techniques in distributed deduplication systems has garnered interest in recent years. Researchers have investigated the use of machine learning algorithms for data deduplication, anomaly detection, and optimization of deduplication processes. For example, Sharma et al. (2020) developed a predictive deduplication framework that utilized machine learning models to predict the likelihood of data redundancy and optimize deduplication operations accordingly. By analyzing historical deduplication patterns and system metrics, their approach achieved improved deduplication efficiency and resource utilization in distributed environments.

In the context of data security, Wang et al. (2018) introduced a secure distributed deduplication framework based on homomorphism encryption. Their approach enables deduplication operations to be performed on encrypted data without compromising data confidentiality. By leveraging homomorphism encryption techniques, the system ensures that only authorized users can access and process reduplicated data while preserving data privacy. Reliability enhancements in distributed deduplication systems have also been explored through the integration of proactive fault detection and recovery mechanisms.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704011 |

Zhou et al. (2020) proposed a predictive maintenance approach that uses machine learning techniques to identify potential failures and mitigate them before they occur. By analyzing historical data and system metrics, their system can predict impending failures and take proactive measures to ensure continuous operation and data availability .Furthermore, research efforts have been directed towards optimizing the trade-off between deduplication efficiency and resource utilization in distributed environments. Wu et al. (2017) proposed a resource-aware deduplication strategy that dynamically adjusts deduplication parameters based on system resource availability and workload demands.

III. PROPOSED METHODOLOGY

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud. The solution must utilize the available data storage in both public and private cloud.

The proposed system is used to protect different multimedia content types including videos, images and audio. It is deployed in both public and private in-network. It creates signatures for multimedia content, and distributed matching engine for multimedia objects. The signature is generated based on the depth signal in the video like spectrum value of the audio signal. Multi-level signature generation process is used to design the efficient encryption for multimedia content this code generation process is repeated by processing the complete file into number of chunks Individual signature codes are merged by using logical operation such as XOR operation.

The aggregate user satisfaction obtained by using a specific set of representations are considered as profit or gain, and consumed number of cloud instances and aggregate bandwidth for all users serve as weights. For instance, if a view is transposed to all five representations, then the weight will be 5, and viewers' QoE will be maximum, as all of the viewers will get the desired representation. However, if a view is transposed in just one representation, then the weight will be 1, and aggregate user QoE will be ≤ 1, depending on the bandwidth capabilities of the viewers watching that view

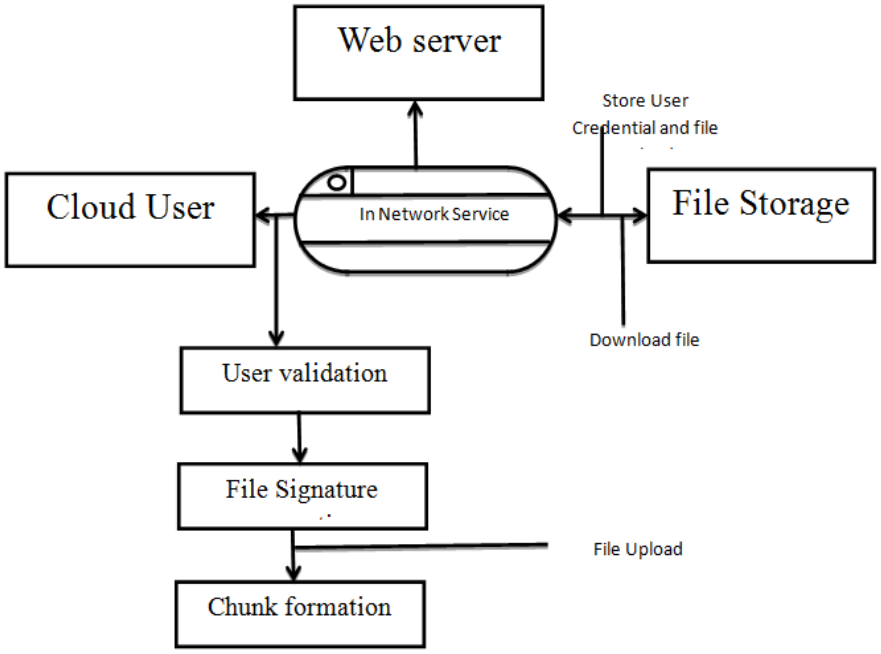

Fig-1 File Transmission - System

A session management system is responsible for retrieving and storing session state, as well as admission control. In a web application, the session management system lies in the critical path of requests and needs to reliably perform its tasks in near real-time without substantial overhead. Utility model should account for the reliability of the

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704011 |

session stores as well as for costs related to storing sessions. There are three monetary costs directly related to maintaining session state: the read and write costs, which can consist of a transaction cost as well as a size-dependent cost, which might account for bandwidth consumed in the data transfer, and a store cost dependent on the size of the stored data over time

IV. RESULTS AND DISCUSSION

The response of a local Web cache is often three times faster than the download time for the same content over the WAN. End users see dramatic improvements in response times, and the implementation is completely transparent to them. A Web cache stores Web pages and content on a storage device that is physically or logically closer to the user-closer and faster than a Web lookup. By reducing the amount of traffic on WAN links and on overburdened Web servers, caching provides significant benefits to ISPs, enterprise networks, and end users. The first step in creating a network-integrated cache engine is to ensure that the network supports traffic localization, which can be achieved by enabling content routing technology at the system-level, and setting specific parameters to optimize network traffic. Once the right network foundation is in place, network caches are added into strategic points within the existing network. By pairing software and hardware, Cisco creates a network-integrated cache engine.The implementation has following modules,

1. Providing in-network services and create connectivity with mobile devices
2. Accessing in-network server using credential information and store the multimedia content without encryption
3. Generating the signature code for video and perform the video transmission by applying the chunk based splitting process
4. Dividing the video into frames and to Identifying and eliminating the redundant copies by validating the video frames.

PROVIDING IN-NETWORK SERVICES AND CREATES CONNECTIVITY WITH MOBILE DEVICES

Wireless communication networks for monitoring and controlling a plurality of remote devices are provided. Briefly, one embodiment of a wireless communication network may comprise a plurality of wireless transceivers having unique identifiers. Each of the plurality of wireless transceivers may be configured to receive a data signal from one of the plurality of remote devices and transmit an original data message using a predefined wireless communication protocol. The original data message may comprise the corresponding unique identifier and sensor data signal. Each of the plurality of wireless transceivers may be configured to receive the original data message transmitted by one of the other wireless transceivers and transmit a repeated data message using the predefined communication protocol.

ACCESSING IN-NETWORK SERVER USING CREDENTIAL INFORMATION THE MULTIMEDIA CONTENT WITHOUT ENCRYPTION

Creating the ftp connection between sender and receivers. Exchanging the file and downloading the video file. The establishment of ftp connection between sender and receivers are registered by verifying the credentials and validation of available service port. Connection id and File name verification algorithm is used to exchange the file between devices. File content Deduplication algorithm is used to apply the file exchange operations. It verifies the signature code of the file before uploading into the ftp server. Credentails, IP address and file content of Sender, receivers devices are considered as the input parameters and the Uploaded file ID and index of the file are generated as output. Successful file exchange operation is applied by checking the file size and storage availability.

GENERATING THE KEY AND CHUNK CODE FOR VIDEO AND PERFORM THE VIDEO TRANSMISSION BY APPLYING THE CHUNK BASED SPLITTING PROCESS

The Graphics and Frame Buffer Manager works mainly to replace the graphics management function of the previous PGL; however, like PGL, the Manager is mainly linked to the graphics context, the display driver, and the Frame buffer. When an application program calls for a graphics streaming system, it must first register the graphics context in the Graphics and Frame Buffer Manager. The management layer will then send the graphics context back, and link it to the display driver and Frame buffer. The application program could then perform drawing through the graphics streaming system, without modifying its source code. The feature described above is mainly aimed to replace

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704011 |

the graphics management interface in Android systems, while providing the same function as PGL. However, in the design of this research, a new feature is added to this Graphics and Frame Buffer Manager, namely the frame buffer.

6.3.4 DIVIDING THE VIDEO INTO FRAMES AND TO IDENTIFY WITH THE ELIMINATION OF THE REDUNDANT COPIES BY VALIDATING THE VIDEO FRAMES

Chunk based encryption is applied to encrypt the video data. The chunks are subdivided into number of frames for the frame processing model. Frame based encryption with the handling of macro blocks of video data is applied by differentiating the frame type content. The system sub divides the macro block structure and to provide the ease of encryption modelling of video data. Video content in terms of chunk and frame content of video data are taken as input and the Encrypted video content with successful decryption are obtained as output and signature generated for validating the decryption of data. Most products that use a "hashing" mechanism also require an index to store the hashes so that they can be looked up quickly to compare against new hashes to see if the new data is unique (i.e., not already stored), or there is a hash match and the new data element does not need to be stored. These indexes must be very fast or handled in such a manner that the unique data stored increases and becomes fragmented so that the solution doesn't slow down during the hash lookup and compare process.

Different solutions from various vendors use diverse hashing algorithms, but the process is basically the same. The term "hashing the data" means "creating a mathematical representation of a specific dataset that can be statistically guaranteed to be unique from any other dataset." The way this is done is to use a generally understood and approved method to encrypt each dataset, so that the metadata or resulting mathematical encryption "hash" can be used to either reproduce the original data or as a lookup within the index to see if any new data hashes compare to any stored data hashes, so the new data can be ignored. Hashing-based dedupe solutions typically provide great results in reducing storage requirements for a particular data set, but there is one huge disadvantage over delta versioning. Since everything is stored as a jumble of mathematical hashes, objects and indexes, it requires the data to be "re-constituted" prior to being usable again for applications. This re-constitution process takes time, which may have a negative impact if the data needs to be recovered NOW. Another benefit of micro scanning is the ability to restore only the sectors required to recover any lost or corrupted data, so massive databases like data warehouses can sometimes be recovered over the network almost instantly.

When an organization considers a cloud service offering as operational environment for the information system in question, both parties can perform a gap analysis to determine which security controls are required for the information system, and which security controls the cloud service provider supports. The difference between the required controls and the supported controls is called the security gap. However, if the security gap between what controls the organization requires and what the cloud provider supports cannot be closed by additional controls or supplementing controls, the risk involved must be mitigated by other ways than via contractual agreements. The most obvious and easiest option, but at the same time the most short-sighted and least satisfying one, is to exclude cloud computing as a possible computing environment.

V. CONCLUSION

Distributed Systems is the virtualization technique which shares the available resource of the various machine distributed globally and connected using internet. The Cloud based multimedia distribution system uses the code generation system to check the duplicate copy of the files.For each multimedia content signature code is generated and stored in the cloud data base. Generated code occupies large space in case of multimedia content as same size of file. To handle this problem redundant bit elimination is used to produce the unique signature for videos. The proposed scheme achieves better storage saving during the signature generation and time saving process in signature validation process.

REFERENCES

1. Haidous, A., Oswald, W., Das, H. and Gong, N., 2022. Content-Adaptable ROI-Aware Video Storage for Power-Quality Scalable Mobile Streaming. IEEE Access.
2. Chang, S.H., Chang, R.I., Ho, J.M. and Oyang, Y.J., 2023. A priority selected cache algorithm for video relay in streaming applications. IEEE transactions on broadcasting, 53(1), pp.79-91.
3. Shen, S.H. and Akella, A., 2019, September. An information-aware QoE-centric mobile video cache. In Proceedings of the 19th annual international conference on Mobile computing & networking (pp. 401-412).

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704011 |

4. Xu, M., Zhu, M., Liu, Y., Lin, F.X. and Liu, X., 2022, October. DeepCache: Principled cache for mobile deep vision. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (pp. 129-144).
5. Sampaio, F., Shafique, M., Zatt, B., Bampi, S. and Henkel, J., 2020, October. Approximation-aware multi-level cells STT-RAM cache architecture. In 2020 International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES) (pp. 79-88). IEEE.
6. Sampaio, F., Shafique, M., Zatt, B., Bampi, S. and Henkel, J., 2023, October. Approximation-aware multi-level cells STT-RAM cache architecture. In 2015 International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES) (pp. 79-88). IEEE.
7. Wang, B., Sen, S., Adler, M. and Towsley, D., 2020, June. Optimal proxy cache allocation for efficient streaming media distribution. In Proceedings.
8. Horsman, G., 2021. Reconstructing streamed video content: A case study on YouTube and Facebook Live stream content in the Chrome web browser cache. Digital Investigation, 26, pp.S30-S37.
9. Tsung, P.K., Chen, W.Y., Ding, L.F., Chien, S.Y. and Chen, L.G., 2021, April. Cache-based integer motion/disparity estimation for quad-HD H. 264/AVC and HD multiview video coding.
10. Liu, C., Tian, L., Zhou, Y., Shi, J., Liu, J., He, S., Pu, Y. and Wang, X., 2019, December. Video content redundancy elimination based on the convergence of computing, communication and cache.
11. Bao, X., Zhou, D. and Goto, S., 2020, May. A lossless frame recompression scheme for reducing DRAM power in video encoding. In Proceedings of 2020 IEEE International Symposium on Circuits and Systems (pp. 677-680). IEEE.
12. Xu, M., Zhu, M., Liu, Y., Lin, F.X. and Liu, X., 2021, October. DeepCache: Principled cache for mobile deep vision. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (pp. 129-144).
13. Goian, H.S., Al-Jarrah, O.Y., Muhaidat, S., Al-Hammadi, Y., Yoo, P. and Dianati, M., 2021. Popularity-based video caching techniques for cache-enabled networks: a survey. IEEE Access, 7, pp.27699-27719.
14. Horsman, G., 2021. Reconstructing streamed video content: A case study on YouTube and Facebook Live stream content in the Chrome web browser cache. Digital Investigation, 26, pp.S30-S37.
15. Kim, B. and Lee, H., 2021. IP-aware cache partition and replacement scheme for mobile computing devices. IEICE Electronics Express, pp.16-20190351.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com