EVALUATING THE LATENCY PERFORMANCE OF DISTRIBUTING STORAGE SYSTEMS: A COMPARISON OF AMAZON S3 USING REAL SERVICE SIMULATION TECHNIQUES.

Suriguge 1*, Noraisyah Binti Tajudin 1

1 Lincoln University College, Petaling Jaya, Malaysia.

*Corresponding author: Suriguge, Lincoln University College, Petaling Jaya, Malaysia.

ABSTRACT

When it comes to constructing parity nodes, the data nodes of the current erasure codes are quite important. Adding additional parity nodes might make it more probable that the researchers can obtain the original data back if they are prepared to work collaboratively and make mistakes. The number of parity nodes determines how much work it takes to correct data nodes and how much additional storage space is required. This happens because users usually approach data nodes for aid with correcting parity nodes at the same time. If an LRC global parity node failed, for instance, it would be necessary to fix all of the data nodes. It will take a lot longer to finish read requests than it used to because "the network's data nodes are getting more requests." One piece of software that doesn't need to get data very often is Google Search. It "produces both data and parity nodes," which implies that the parity nodes may undertake some of the maintenance work that the data nodes generally do. This means they won't have to wait as long as they would in any other case. In short, adding a parity node doesn't impact how many data nodes may be accessed overall. Researches analysis indicates that parity nodes correlate with increased storage expenses. Adding parity nodes to provide parity could minimise access latency without changing the amount of storage required, as long as the architecture is good. This will be shown in the next sections, which may confuse researchers.

Keywords: Storage Systems, Latency Performance, Amazon S3, Techniques, Real Service.

INTRODUCTION

In the past ten years, more and more people have been doing things online. People do things like purchase stuff online, use social media, and search for things. Every day, companies that do research produce a lot of digital data. One of the problems that both businesses and schools are trying to solve right now is how to make storage systems that are useful and cheap. The expansion of data storage amounts is closely related to the development of large-scale distributed storage systems. Hadoop Distributed File System (HDFS) and Windows Azure Storage (WAS) are two examples of distributed file systems. Both of these systems may be used with Microsoft Azure. These storage systems can hold a lot of data, analyse it quickly, and run cloud-based apps. They are also more popular and dependable. They may also be able to meet these demands. When making big distributed storage systems, it's typical to use a lot of cheap storage devices, which might make the system unstable. This is because these gadgets may

sometimes be unstable. These nodes are not immune to the chance of failures occurring, in addition to the well-known instability associated with storage devices. Even while these systems help people improve their talents, they fail more often than they work. This is still true, even if modern technologies have many benefits. Because of this, it is very important that we show that these systems are dependable and will survive a long time, as well as solve the problem of systems failing often. (Kent et al., 2025).

BACKGROUND OF THE STUDY

Not only is it essential to have data storage solutions that are scalable in this day and age of big data and cloud computing, but it is also essential to have solutions that are dependable and effective. It has become more vital to make use of distributed storage systems in order to satisfy this demand. This is due to the fact that they are able to expand, are tolerant of errors, and are constantly accessible. Amazon Simple Storage Service (S3) is one of the cloud-based object storage systems that is now available, but it is also one of the most well-known and frequently used of these systems. It is possible to utilise it for a variety of purposes, including data lakes, analytics, machine learning, backup and recovery, and other applications that are equivalent. In spite of the fact that distributed storage solutions such as Amazon Structured Storage 3 are extensively used, researchers and professionals in the field are always working to make them quicker. Irrespective of how often it is used, this remains the situation (Xie et al., 2024). When referring to the transmission of data, the term "latency" refers to the amount of time that elapses between the sending of a command and the beginning of the transmission of data. The amount of latency that a cloud-based service experiences is a highly significant factor that has a significant influence on both its performance and its quality. As the need for applications to react in real time or very near to real time continues to rise, it is becoming more critical to be aware of and evaluate the latency performance of distributed storage systems. The traditional methods of assessment have their limitations when it comes to their ability to replicate scenarios that occur in the actual world. This demonstrates the need of using genuine service simulation approaches in order to carry out assessments that are more reliable and relevant to real-world situations (Mahmoudi & Khazaei, 2021).

PURPOSE OF THE RESEARCH

At now, data nodes are responsible for producing the parity nodes in the majority of erasure algorithms. Using this strategy, researcher may increase the number of "parity nodes" to make the system more resilient to failure and increase the chances of recovering the original data. There will be more storage overhead and repair burden on data nodes as the number of parity nodes increases, as they are often needed to aid with parity node repairs. If a global parity node breaks in LRC, for instance, all data nodes must be repaired. With a "increased workload on data nodes" the processing time for read requests increases. Obtaining data software isn't always the best option in some situations, such as. It is possible to decrease wait times by transferring some of the repair effort from data nodes to parity nodes when they are generated

concurrently with data nodes. Simply said, researchers may simply add more parity nodes to the network in the event that one fails, hence expanding the amount of data nodes that are accessible. That being said, parity nodes seem to incur a greater storage expense. The next sections will show that parity nodes may be built with enough forethought to reduce storage overhead without changing access latency. The purpose of this research is to examine how well "Hierarchical Tree Structure Code (HTSC) and High Failure-tolerant Hierarchical Tree Structure Code (FH HTSC) will" perform.

LITERATURE REVIEW

For cloud-based infrastructures to work properly today, they need to have distributed storage systems. These technologies not only let individuals and companies store, get, and manage large volumes of data, but they also meet the demands of enterprises. These systems store data on several servers that are spread out around the globe. This means that they are likely to be highly available, durable, and scalable. The servers are spread out all across the globe, which is why this is happening. Latency is one of the performance metrics that gets the most attention. It is the amount of time it takes to get or save data. There are various things that may impact the latency performance of distributed storage systems. These things can be put into several groups. In this case, the quantity of data, the network's congestion, the way people use it, the system's design, and the physical distance between clients and data centres are all taken into account. Even while cloud service providers like Amazon Web Services have made a lot of work in this area, there is always a fear regarding latency. This is especially true for apps that need to be responsive and work in real time (Jodhavat et al., 2025). In current research on distributed storage systems, latency is often overlooked, with emphasis placed on throughput, dependability, or cost-effectiveness as primary objectives. Furthermore, most performance studies use synthetic workloads, which are meticulously replicated simulations of real-world workloads executed in a controlled environment. Because of this, the results of these kinds of testing may not accurately reflect how people truly feel about the product when they use it in their daily lives. Real service simulation approaches have lately garnered significant interest as a potential solution to this issue. These approaches try to copy the behaviours, workloads, and network setups of actual users so that they may evaluate system performance in a manner that is not only more accurate but also more useful. There hasn't been much study done on how these methods effect the latency of services like Amazon S3 in real life, even if they have their own benefits. This is true even if these tactics have their own benefits. To fill this vacuum in the research, it is vital to do a more concentrated study of latency by using simulations of the actual world. Researcher can do this using simulations. This method might be very helpful for both organisations and developers since it gives them important information about how distributed storage systems work in the real world (Khan et al., 2023).

RESEARCH QUESTION

What is the impact of Distributing Storage Systems of Amazon S3 on Latency Performance?

RESEARCH METHODOLOGY

Research Design

The quantitative investigation used the most recent version of SPSS, 25. The odds ratio and 95% confidence interval were used to evaluate the strength and direction of the statistical association. The researchers established a statistically significant threshold of $p < 0.05$. An analytical assessment was conducted to ascertain the principal elements of the data. Quantitative approaches are often used to evaluate data obtained from surveys, polls, and questionnaires, as well as data examined using computing tools for statistical assessment.

Sampling

Research participants filled out questionnaires to provide data for the research. Utilising the Rao-soft approach, researchers identified a cohort of 1,045 individuals, resulting in a total of 1,316 enquiries. The researchers received 1265 responses, excluding 96 for incompleteness, resulting in a final sample size of 1169.

Data and Measurement

This study used a questionnaire as the main tool for data collecting. Section A of the survey requested essential demographic data, while Section B used a 5-point Likert scale to collect answers about characteristics related to online and offline channels. The secondary data was obtained from many sources, mostly databases on the internet.

Statistical Software

The statistical analysis was conducted with SPSS version 25 and Microsoft Excel for Windows.

Statistical Tools

The statistical analysis method was used to comprehend the fundamental characteristics of the data under examination. The investigator must do an analysis of the data using ANOVA.

CONCEPTUAL FRAMEWORK
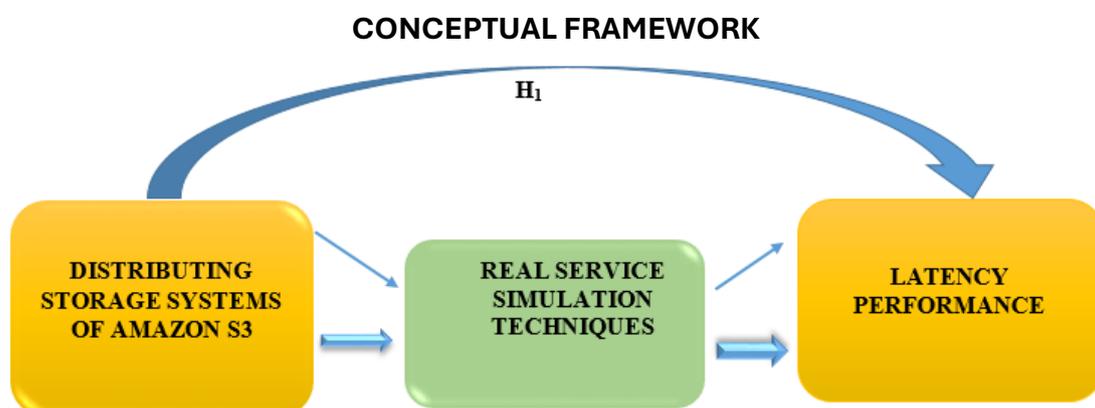
RESULT

Factor Analysis: Factor Analysis (FA) is often used to identify latent variables within observable data. Utilising regression coefficients for evaluation is a conventional practice in the absence of distinct visual or diagnostic cues. Models are crucial for success in financial analysis. Modelling inherently entails inaccuracies, interferences, and discernible connections. The Kaiser-Meyer-Olkin (KMO) Test may evaluate datasets generated by multiple regression analyses. Researchers assert that the model and the variables in the sample are representative. The data exhibits redundancy. Information is more intelligible when conveyed in smaller increments. Any value between 0 and 1 may function as the KMO output. A KMO value ranging from 0.8 to 1 is considered sufficient for sample size. Kaiser contends that these are the allowable ranges: Kaiser has delineated additional admission standards.

An inadequate range of 0.050 to 0.059 and a subpar range of 0.60 to 0.69

The typical range for middle grades is 0.70 to 0.79.

The quality point score varies from 0.80 to 0.89.

The interval from 0.90 to 1.00 astonishes them.

The results of Bartlett's sphericity test are as follows: The chi-square statistic is around 190, with a significance level of 0.000.

This confirms that the statements made for sampling purposes are genuine. The researchers used Bartlett's Test of Sphericity to evaluate the relevance of the correlation matrices. A Kaiser-Meyer-Olkin measure score of 0.877 indicates a satisfactory sample. The p-value from Bartlett's sphericity test is 0.00. The association matrix does not possess a unique value, hence satisfying Bartlett's circularity test.

Table 1. The assessment of sample adequacy by KMO (Kaiser-Meyer-Olkin) and Bartlett's Test reveals a measure of 0.877.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.877
Bartlett's Test of Sphericity	Approx. Chi-Square	3252.968
	df	190
	Sig.	.000

Bartlett's Test of Sphericity further corroborated the significance of the association criterion. The Kaiser-Meyer-Olkin measure of sampling adequacy is 0.877. Researchers achieved a p-

value of 0.00 using Bartlett's sphericity test. Bartlett's sphericity test findings suggested that the correlation matrix is flawed.

INDEPENDENT VARIABLE

Distributing Storage Systems of Amazon S3: In order to store data more efficiently, many data centres use distributed storage systems, which allow data to be partitioned among several physical servers. Storage unit clusters have a similar architectural feature that enables node-to-node data synchronisation and coordination. Both on-premises solutions, such as Cloudian Hyperstore, and highly scalable cloud providers, such as Amazon S3 and Microsoft Azure Blob Storage, are built on distributed storage. Amazon Simple Storage Service (S3) is the object storage web service interface offered by Amazon Web Services (AWS). The e-commerce network of Amazon.com is powered by scalable storage technology, much as Amazon S3. With its limitless storage capacity, Amazon Simple Storage Service (S3) might be used to many different purposes. Hybrid cloud storage has several applications, including backups, data archiving, analytics in data lakes, web app storage, and disaster recovery. Launched in the United States on March 14, 2006, Amazon Simple Storage Service (S3) was initially made accessible in Europe via Amazon Web Services (AWS) in November 2007. When it comes to object storage, Amazon S3 is unrivalled in terms of speed, security, availability, and scalability. With its many uses including data lakes, websites, mobile apps, backup and restore, archiving, corporate apps, Internet of Things devices, and big data analytics, Amazon Simple Storage Service (S3) is available to businesses of all sizes and in all sectors. Amazon Simple Storage Service (S3) offers researchers a set of administrative tools that they may employ to tailor data access to their own business, regulatory, and operational needs.

DEPENDENT VARIABLE

Latency Performance: Communication delays in a network are known as latency. The data transmission time over the network is shown. A low latency network is one that responds quickly, while a high latency network has a longer wait. A data packet's latency is the amount of time it takes to get from its origin to its final destination. If researcher want researchers' site to run well even for people with slow connections, researcher need to optimise it to eliminate the sources of delay and do tests that mimic high latency. This article defines latency, discusses its effects on performance, provides methods for measuring latency, and offers advice on how to decrease it. The term "latency" is often used to describe the time it takes for a response to reach the user after they submit a request. The first 14Kb of a first request incur more delay due to the many steps involved, such as a DNS query, a TCP handshake, and the secure TLS negotiation. Following this first request, the delay in establishing a connection to the server will be reduced. How much time a network or Internet connection takes to respond is called latency. When latency is low, delays are minimal at best. There are a lot of delays if the latency is high. Reducing latency is one of the primary goals of performance improvement. It can be easy to dismiss the delay connected with a single item, particularly a simple HTML site. The HTML of a website,

however, often contains requests for various scripts, media assets, and CSS styles. High latency has a more noticeable effect on user experience as the quantity and amount of these requests increase. Almost instantly, given a low-latency connection, the required resources will materialise. There may be a noticeable lag between sending a request and receiving the returned resources over a connection with high latency. Measuring the data transfer rate between two points on the network allows us to calculate the latency. The time it takes to transmit a request for resources is one measure of latency; another is the total time it takes for a resource to travel from the browser's request to the browser's actual receipt of the resource (Wang et al., 2021).

Relationship between Distributing Storage Systems of Amazon S3 and Latency Performance: For the purpose of delivering researchers content, Cloud Front makes use of a dispersed network of data centres, also known as "edge locations," located all over the world. Requests are routed to the edge location that has the lowest latency in order to ensure that users are able to access the content that researcher are providing with Cloud Front at the fastest feasible pace. When it comes to the data and apps that researcher use the most often and are sensitive to latency, the storage type that researcher should choose is Amazon S3 Express One Zone. In addition to having a single availability zone, it was designed specifically to ensure that users have constant access within the range of single digit milliseconds. The data that is stored in Amazon S3 is distributed across several servers and geographical locations in order to provide high availability, scalability, and durability. The distributed nature of the data, on the other hand, may have a direct influence on the latency performance, which refers to the amount of time it takes to retrieve or store data. The latency may be affected by a wide variety of parameters, such as the distance between nodes in the network, the frequency of access, the load on the servers, and the replication of data. For example, if requests are sent to various data centres, in accordance with their availability, there is a possibility that they would face delays. Because the system employs eventual consistency for some procedures, there is a possibility that there may be minor delays in the visibility of data. When it comes to real-time applications, when prompt replies are of the utmost importance, these considerations about latency take on even more weight. In order to get the most out of S3, researcher need to be aware of how the architecture of the service influences the latency behaviour of the service. When it comes to selecting storage solutions for workloads that are sensitive to latency, developers and organisations may be able to make better decisions if they evaluate latency under real-world conditions (Hofmann et al., 2022).

The investigator formulated the following hypothesis to assess the relationship between Amazon S3's Distributed Storage Systems and Latency Performance.

"H_{01}: There is no significant relationship between Distributing Storage Systems of Amazon S3 and Latency Performance."

"H₁: There is a significant relationship between Distributing Storage Systems of Amazon S3 and Latency Performance."

Table 2. H1 ANOVA Test.

ANOVA					
Sum					
	Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
Between Groups	39588.620	203	5653.517	1057.522	.000
Within Groups	492.770	965	5.346		
Total	40081.390	1168			

This inquiry will provide significant results. The F statistic is 1057.522, with a p-value of 0.000, which is below the 0.05 alpha level. The hypothesis asserts that *"H₁: A significant correlation exists between Amazon S3's Distributed Storage Systems and Latency Performance."* The alternative hypothesis is affirmed, whereas the null hypothesis is dismissed.

DISCUSSION

One way to find out how distributed storage architecture affects latency is to simulate actual service settings. There are a few different reasons why Amazon S3 can have latency problems, such as copying data, network delays, and sharing server duties. Researchers can, on the other hand, ensure that the service will be available and last a long time by spreading it out across several servers in different areas. Researcher can be sure that the service will last longer if researcher do things this way. For instance, a statistical study showed that the way S3 is set up in a distributed way has a big effect on how long it takes for the system to respond. This has a big effect on apps that demand low latency or features that work in real time. Using parity nodes and technologies like Hierarchical Tree Structure Code (HTSC), it is feasible to make a store's capacity bigger without making it slower. Thanks to these new technologies, this is now possible. Amazon S3 Express One Zone is one of the various low-latency options that Amazon Web Services (AWS) offers. It helps meet the need for quicker data access. Based on the findings of this study, cloud storage needed to be evaluated for its responsiveness, reliability, and scalability. The information researchers have found may be useful to organisations and developers seeking for storage solutions that may meet performance needs in situations when milliseconds matter. These results are from real-life scenarios where every millisecond matters.

CONCLUSION

For this research, real-world service simulation methodologies were used to assess the correlation between the latency performance of Amazon S3's distributed storage systems and the overall system performance. The distributed design of Amazon S3 has a big effect on latency, even if it is very scalable, durable, and easy to get to. Even if it is accessible, this is the situation. If the time it takes to get data changes, latency-sensitive applications may not operate as well. The duplication of data, the distance between the servers, and the strain on the servers

are just a few of the things that cause this to happen. This study has yielded significant insights into the performance of Amazon S3 in real-world use scenarios via the application of simulation methodologies. This research has provided a more precise representation than synthetic benchmarking alone, which has been used to construct a more authentic depiction. This research has been conducted to provide a more genuine picture. The results of this study support the null hypothesis, which states that there is no statistically significant link between distribution and delay. Additionally, methods like Hierarchical Tree Structure Codes (HTSC) may help with difficulties like repairs and delays. These methods don't hurt the data's integrity or raise the cost of storage. This kind of information is needed by programming professionals and corporations to make judgements based on reliable information when it comes to improving and developing distributed cloud storage systems. Researcher would like to end by saying that it is very important to know how to manage latency well in distributed storage systems like Amazon S3. This is something that has to be thought about when a lot of individuals are using these systems at the same time.

REFERENCES

1. Bornholt, J., Joshi, R., Astrauskas, V., Cully, B., Kragl, B., Markle, S., ... & Warfield, A. (2021, October). Using lightweight formal methods to validate a key-value storage node in Amazon S3. In Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles (pp. 836-850).
2. Gudelli, V. R. (2023). AI-powered insights for performance optimization in AWS cloud environments. International Journal of Scientific Research and Applications, 10(2).
3. Hofmann, W., Lang, S., Reichardt, P., & Reggelin, T. (2022). A brief introduction to deploy Amazon Web Services for online discrete-event simulation. Procedia Computer Science, 200, 386-393.
4. Jodhavat, H., Meka, S., & DSouza, B. (2025). Designing High-Throughput Data Pipelines: A Performance-Centric Architectural Framework for Low-Latency Analytics in Distributed Cloud Environments. International Journal of Emerging Trends in Computer Science and Information Technology, 6(2), 56-62.
5. Kent, K. B., Zhou, M., Adeyemo, G., & Wang, Y. (2025). Cloudhive: A Cloud-Based Framework for Smart Grid Co-Simulation, Data, and Communication. Concurrency and Computation: Practice and Experience, 37(21-22), e70238.
6. Khan, A. Q., Matskin, M., Prodan, R., Bussler, C., Roman, D., & Soylu, A. (2024). Cloud storage cost: a taxonomy and survey. World Wide Web, 27(4), 36.
7. Khan, A. Q., Nikolov, N., Matskin, M., Prodan, R., Roman, D., Sahin, B., ... & Soylu, A. (2023). Smart data placement using storage-as-a-service model for big data pipelines. Sensors, 23(2), 564.
8. Mahmoudi, N., & Khazaei, H. (2021). SimFaaS: A performance simulator for serverless computing platforms. arXiv preprint arXiv:2102.08904.

9. Nur, A., & Muanenda, Y. (2024). Design and evaluation of real-time data storage and signal processing in a long-range distributed acoustic sensing (DAS) using cloud-based services. Sensors, 24(18), 5948.

10. Wang, H., Shen, H., Li, Z., & Tian, S. (2021, July). GeoCol: A geo-distributed cloud storage system with low cost and latency using reinforcement learning. In 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS) (pp. 149-159). IEEE.