# ANALYSING THE LATENCY PERFORMANCE OF DISTRIBUTED STORAGE SYSTEMS: A COMPARATIVE STUDY OF AMAZON S3 THROUGH REAL SERVICE SIMULATION METHODS.

Suriguge 1*, Noraisyah Binti Tajudin 1

1 Lincoln University College, Petaling Jaya, Malaysia.

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

## ABSTRACT

The data nodes of the present erasure codes are quite significant when it comes to making parity nodes. If the researchers are more willing to work together and make errors, adding more parity nodes could make it more likely that they can get the original data back. The number of parity nodes affects how much labour it takes to fix data nodes and how much extra storage space is needed. This arises because people typically ask data nodes for help with fixing parity nodes at the same time. If an LRC global parity node were to fail, for example, it would be essential to resolve all of the data nodes. Because "the network's data nodes are getting more requests," it will take significantly longer to complete read requests than it did in the past. Google Search is an example of software that doesn't need to obtain data very frequently. It "produces both data and parity nodes," which means that the parity nodes may conduct part of the maintenance work that the data nodes usually do. Because of this, they won't have to wait as long as they would in any other possible situation. To put it simply, having a parity node doesn't change the overall number of data nodes that may be accessed. Researcher's research shows that parity nodes are connected to higher storage costs. If the design is well thought out, adding parity nodes to provide parity might lower access latency without affecting the amount of storage needed. This will be demonstrated in the upcoming portions, which might make researchers perplexed.

**Keywords:** Storage Techniques, Latency Efficiency, Amazon S3, Methodologies, Actual Service.

## INTRODUCTION

Things that can be done online have seen a significant surge in popularity over the last decade. For example, making purchases online, utilising Google, and participating in social media are all examples of these behaviours. Research-related businesses generate a significant amount of digital data on a daily basis. For the time being, academics and businesses are still discussing the best way to create storage systems that are not only effective but also inexpensive. There is a clear correlation between the expansion of data storage volumes and the rise of distributed storage systems on a big scale. A few of examples of distributed file systems are the Hadoop Distributed File System (HDFS) and Windows Azure Storage. A support for both of these systems is provided by Microsoft Azure. Aside from being more widely used and dependable, these storage systems are also capable of managing enormous volumes of data, processing it in a very short period of time, and running applications that are hosted in the cloud. Additionally, they could be able to fulfil these requirements. With the construction of massive distributed storage systems, it is usual practice to make use of a large number of affordable storage devices, despite the fact that this may cause things to become unstable. This is due to the fact that these gadgets may, on sometimes, cease functioning properly. Despite the fact that storage devices are notoriously unreliable, these nodes are not completely safe from malfunctions. It is not always the case that these tactics are successful in increasing abilities, despite the fact that they are extremely useful. There has been no change in this, despite the fact that these technologies have shown benefits. For this reason, researcher's key objective is to demonstrate that these systems are dependable and long-lasting, while also addressing the problem of many system failures occurring often (Gadban & Kunkel, 2021).

## BACKGROUND OF THE STUDY

In this day and age, with the proliferation of big data and cloud computing, it is very necessary to have data storage solutions that are not only dependable and effective but also capable of scaling. In order to satisfy this need, distributed storage solutions are becoming an increasingly significant option. The primary reasons for this are the scalability, fault tolerance, and availability of the system. There are a variety of cloud-based object storage systems that are now accessible; among them, Amazon Simple Storage Service (S3) is among the most well-known and widely used of them all. Data lakes, analytics, machine learning, backup and recovery, and other operations that are very comparable are some of the numerous possible applications of this technology. Despite the fact that distributed storage systems such as Amazon Structured Storage 3 are used by a large number of people, specialists in the field are always working to make them more efficient. It makes no difference how often it is used; this remains the case (Gudelli, 2023). The term "latency" refers to the amount of time that passes between the sending of a command and the beginning of the data transfer process in the context of data transmission. The amount of latency that a cloud-based service might experience is one of the most significant aspects that can have an effect on the quality and performance of the service. Because of the ever-increasing need for applications to react in real time or very near to real time, it is becoming more important to be aware of and evaluate the latency performance of distributed storage systems on a regular basis. Conventional methods of assessment have significant limitations when it comes to their ability to effectively simulate scenarios that occur in the real world. For this reason, it is essential to carry out assessments using genuine service simulation techniques in order to get outcomes that are superior and more immediately relevant (Bucur & Miclea, 2021).

## PURPOSE OF THE RESEARCH

At the moment, the majority of erasure algorithms take use of data nodes in order to build parity nodes. Researcher may strengthen the system's resilience to failure and improve the likelihood of retrieving the original data by increasing the number of "parity nodes" using this method.

Because they are often required to assist with parity node repairs, data nodes will have a greater storage overhead and repair load as the number of parity nodes grows. An “increased workload on data nodes” causes read requests to take longer to process. For example, there are cases when purchasing data software isn’t the way to go. If parity nodes are created at the same time as data nodes, part of the repair work may be moved to them, reducing wait times. To put it another way, researchers can increase the number of available data nodes by adding more parity nodes to the network in case one fails. However, it seems that parity nodes have a higher storage cost. In the sections that follow, we’ll demonstrate how well-planned parity nodes may decrease storage overhead without affecting access latency. In this study, researchers want to find out how “Hierarchical Tree Structure Code (HTSC) and High Failure-tolerant Hierarchical Tree Structure Code (FH HTSC) will” do in practice.

## LITERATURE REVIEW

Modern cloud-based infrastructures need distributed storage solutions to work properly. These technologies not only meet the demands of companies, but they also let individuals and organisations store, get, and manage huge volumes of data. These systems employ data that is stored on several servers all over the globe. This means that Researcher can expect them to be easy to access, expandable, and survive a long time. This is happening because the servers are spread out all around the globe. Latency is one of the most talked-about performance metrics. It is the time it takes to get or save data. Several factors may influence the efficacy of distributed storage systems concerning latency. Researcher may put these things in different groups in a number of ways. In this case, things to think about include the volume of data, network congestion, user behaviour, system design, and the real distance between clients and data centres. Even though cloud providers like AWS have put a lot of money into this area, people are still worried about latency. This is especially true for software that runs in real time and changes all the time. The main aims of current research on distributed storage systems are usually throughput, dependability, or cost-effectiveness. Latency is frequently not given much thought. Most performance tests also utilise synthetic workloads, which are exact copies of real-world workloads conducted in a simulated environment (Khan et al., 2024).

This implies that the results of these kinds of evaluations could not reflect how consumers genuinely feel about the product in their daily lives. There has been a lot of talk lately regarding actual service simulation approaches as a way to solve this issue. These approaches try to copy the activities, tasks, and network setups of actual users in order to have a better idea of how well the system works. There hasn’t been any study on how these strategies could affect the latency of services like Amazon S3, even if they might be helpful. This is still true, even if these tactics do have some benefits. To fill this vacuum in knowledge, researchers need more targeted studies on delay that employ real-world simulations. Researchers can achieve this via simulations. This method may be very helpful for both organisations and developers since it shows important information about how distributed storage systems work in real life (Satija et al., 2025).

## RESEARCH QUESTION

What is the effect of Amazon S3's Distributed Storage Systems on Real Service Simulation Techniques?

## RESEARCH METHODOLOGY

### Research Design

The quantitative analysis used the latest version of SPSS, 25. The odds ratio and 95% confidence interval were used to assess the magnitude and direction of the statistical link. The researchers determined a statistically significant criterion of $p < 0.05$. An analytical evaluation was performed to identify the key components of the data. Quantitative methods are often used to analyse data acquired via surveys, polls, and questionnaires, as well as data assessed by computational statistical tools.

### Sampling

Research participants completed questionnaires to provide data for the study. Employing the Rao-soft methodology, researchers selected a cohort of 1,045 people, yielding a total of 1,316 queries. The researchers obtained 1265 replies, removing 96 due to incompleteness, yielding a final sample size of 1169.

### Data and Measurement

This research used a questionnaire as the primary instrument for data collection. Section A of the survey solicited fundamental demographic information, while Section B used a 5-point Likert scale to gather responses about attributes associated with online and offline channels. The secondary data was sourced from several web resources.

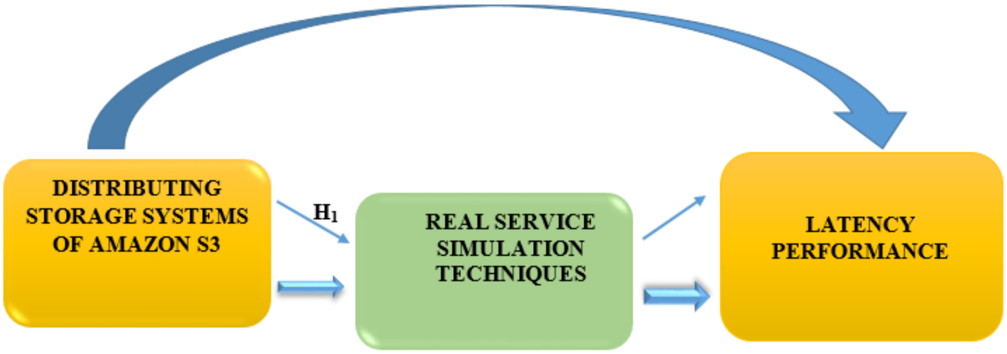
### Statistical Software

The statistical investigation was conducted with SPSS version 25 and Microsoft Excel.

### Statistical Tools

The statistical analysis approach was used to understand the essential aspects of the data being examined. The investigator must do a data analysis using ANOVA.

## CONCEPTUAL FRAMEWORK


DISTRIBUTING STORAGE SYSTEMS OF AMAZON S3
H1
REAL SERVICE SIMULATION TECHNIQUES
LATENCY PERFORMANCE

## RESULT

**Factor Analysis:** Finding hidden variables in observable data is a common use of Factor Analysis (FA). When there are no clear visual or diagnostic indicators, it is common practice to use regression coefficients for assessment. Financial analysis relies heavily on models. Errors, interferences, and observable correlations are inherent in the modelling process. When assessing datasets produced by multiple regression analyses, the Kaiser-Meyer-Olkin (KMO) Test may be conducted. Scientists claim that model and sample variables are accurate representations of the population at large. There is duplication in the data. Conveying less information makes it easier to understand. The KMO output may take on any value from 0 to 1. A sample size is deemed adequate when the KMO value falls within the range of 0.8 to 1. According to Kaiser, these are the permissible limits: Additional entrance requirements have been outlined by Kaiser.

The typical range for middle grades is 0.70 to 0.79, while this range of 0.050 to 0.059 is insufficient and the range of 0.60 to 0.69 is substandard.

Scores of 0.80 to 0.89 indicate quality.

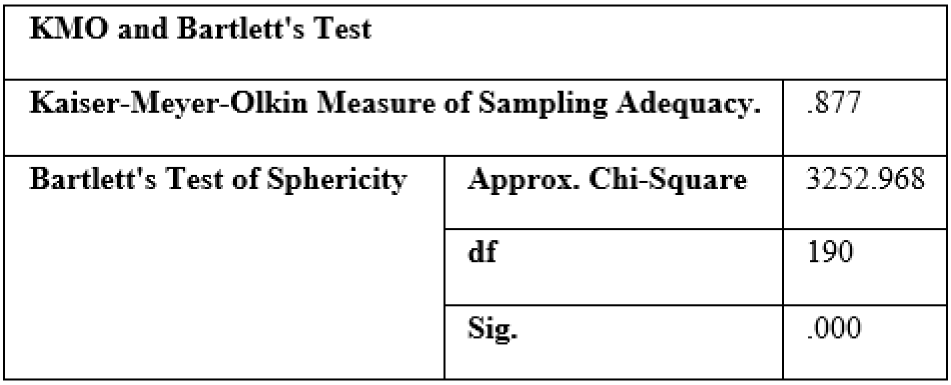
They are astounded by the range from 0.90 to 1.00.

Here are the outcomes of Bartlett’s sphericity test:

There is a chi-square statistic of around 190 and a 0.000 level of significance.

This proves that the claims stated in the sample are authentic. In order to determine if the correlation matrices were relevant, the researchers used Bartlett’s Test of Sphericity. An adequate sample size is indicated by a Kaiser-Meyer-Olkin score of 0.877. A p-value of 0.00 is produced using Bartlett’s sphericity test. The association matrix passes Bartlett’s circularity test since it does not have a unique value.

**Table 1.** A Kaiser-Meyer-Olkin measure of 0.877 was found in the evaluation of sample adequacy by KMO and Bartlett's Test.



| 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 |

Another piece of evidence supporting the association criterion's relevance was Bartlett's Test of Sphericity. A sample adequacy metric developed by Kaiser-Meyer-Olkin yields a value of 0.877. By using Bartlett's sphericity test, the researchers were able to get a p-value of 0.00. According to the results of Bartlett's sphericity test, the correlation matrix is incorrect.

## INDEPENDENT VARIABLE

**Distributing Storage Systems of Amazon S3:** Distributed storage systems enable data to be partitioned across several physical servers, which many data centres utilise to store data more effectively. Data synchronisation and coordination across nodes is made possible by a shared architectural element in storage unit clusters. Distributed storage is the foundation of both on-premises systems like Cloudian Hyperstore and the most scalable cloud providers like Amazon S3 and Microsoft Azure Blob Storage. As a web service interface for object storage, Amazon Simple Storage Service (S3) is one of the many offerings from Amazon Web Services (AWS). Scalable storage technologies, such as Amazon S3, underpin the Amazon.com e-commerce network. Amazon Simple Storage Service (S3) could serve a variety of functions thanks to its infinite storage capacity. Data archiving, web app storage, analytics in data lakes, backups, and disaster recovery are just a few of the many uses for hybrid cloud storage. Amazon Simple Storage Service (S3), which debuted in the US on March 14, 2006, was first made available in Europe in November 2007 via Amazon Web Services (AWS). Amazon Simple Storage Service (S3) offers unparalleled performance, reliability, scalability, and security for object storage. Amazon Simple Storage Service (S3) is accessible to organisations across all industries and sizes, and it has several applications such as data lakes, websites, mobile apps, backup and restore, archiving, corporate apps, Internet of Things devices, and big data analytics. Through Amazon S3, researchers have access to a suite of administrative tools that may be used to customise data access according to their own operational, regulatory, and commercial requirements (Khattak & Khan, 2024).

## MEDIATING VARIABLE

**Real Service Simulation Techniques**: Actual-world service simulation approaches examine how well a system works by trying to simulate how actual users behave, what they do, and how they interact with the network as precisely as feasible. These simulations show that distributed storage solutions work better than synthetic testing. These tests simulate real-world service situations, exposing latency limitations and system constraints that are sometimes difficult to detect in controlled testing environments. This method might make cloud storage systems like Amazon S3 work a lot better. This is very important for apps that need to get data quickly and with as little delay as possible. Dust suppression and movement in coal mines and tunnels are two examples of the kinds of phenomena that may be studied and predicted using simulation techniques, which include the use of electronic computers to simulate real-world processes. One definition of a simulation is an imaginary model of a real-world process or system. When used in this wide meaning, the words model and simulation are typically considered synonymous. In cases when models are necessary for simulations, a clear separation between the two is maintained; in this case, the model stands for the essential traits or actions of the chosen system or process, while the simulation depicts the model's development over time. The third one further approach to differentiate between the two is to think of simulation as experimentation that makes use of a model. The term encompasses simulations that do not rely on the passage of time. The simulation is often run on a computer. Most of us have heard of weather forecasts, flight simulators used to teach pilots, and accident models for cars. These are all instances of computer simulation modelling. Using a model to conduct experiments is known as simulation. The user conducts experiments with the model in order to deduce the system's behaviour by simulating the model's behaviour. This overarching structure has been an invaluable tool for education, research, and product development (Bagai, 2024).

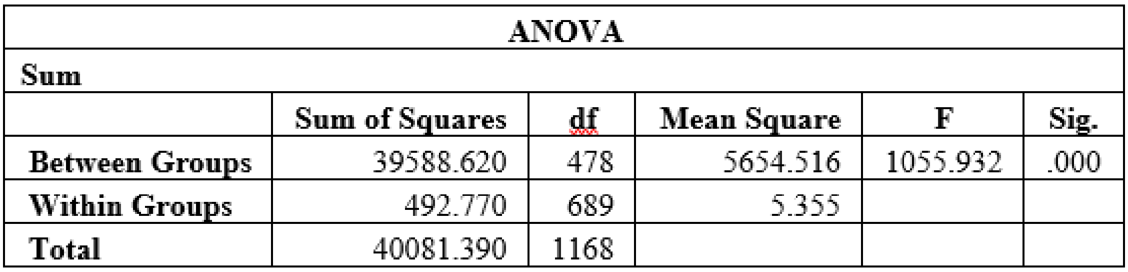
**Relationship between Distributing Storage Systems of Amazon S3 and Real Service Simulation Techniques**: Data is stored on a massive computer network that spans the whole globe using Amazon Simple Storage Service (S3). Researcher can be certain that your data will always be accessible, scalable, and secure thanks to this. Because it is a distributed system, latency may be affected by factors such as data duplication, network congestion, server load, and the distance between clients and data data centres. When it comes to developing real-world systems that can quickly receive and process data, having a solid understanding of these latency characteristics is very necessary. To test Amazon S3 in a setting that is as close to real life as possible, service deployment simulations are absolutely necessary. These simulations are distinct from controlled laboratory research and synthetic benchmarks due to the fact that they model user activity, workloads, and the dynamics of the network. The difficulties that come with running a distributed system might be better understood with the use of a realistic service simulation. This is feasible due to the fact that traffic and weather patterns may be replicated simultaneously simultaneously. This provides developers and academics with clarity on delays. The distributed storage technology and real-world service modelling that Amazon S3 offers could be the most effective way to achieve optimal performance. Simulations have the ability to provide light on how the design of Amazon S3 impacts delivery times. This technique has the potential to uncover errors and modifications that are missed by conventional testing. This link makes it simple to utilise and improve applications that are sensitive to latency while using Amazon S3 (RIGUGE, S., & INAMDAR, 2024).

The researcher formulated the below hypothesis to assess the relationship between Amazon S3's Distributing Storage Systems and Real Service Simulation Techniques.

*"$H_{01}$: There is no significant relationship between Distributing Storage Systems of Amazon S3 and Real Service Simulation Techniques."*

*"$H_1$: There is a significant relationship between Distributing Storage Systems of Amazon S3 and Real Service Simulation Techniques."*

**Table 2.**H1 ANOVA Test.



ANOVA

Sum

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 39588.620 | 478 | 5654.516 | 1055.932 | .000 |
| Within Groups | 492.770 | 689 | 5.355 | | |
| Total | 40081.390 | 1168 | | | |

This inquiry will provide significant results. The F statistic is 1055.932, with a Pearson's correlation coefficient of 0.000, which is below the alpha level of 0.05. The hypothesis asserts that "H1: A significant correlation exists between Amazon S3 Distributed Storage Systems and Real Service Simulation Techniques." The alternative hypothesis is affirmed, whereas the null hypothesis is dismissed.

## DISCUSSION

Amazon Simple Storage Service (S3) provides distributed storage solutions with a statistically significant correlation to latency performance. Simulating real-world service environments is one approach to understanding the impact of distributed storage architecture on latency. Amazon Simple Storage Service (S3) latency issues may arise from a variety of sources, including data copying, network delays, and server sharing. However, by distributing data among several servers in various locations, researchers can guarantee that the service will be accessible and endure for a long time. This is the best method for researchers to ensure that their service will continue as long as possible. For example, a statistical analysis revealed that the dispersed nature of S3 significantly impacts the response time of the system. This significantly impacts real-time services and programs that rely on minimal latency. A store's capacity may be increased without slowing it down by using parity nodes and technologies like Hierarchical Tree Structure Code (HTSC). Researchers owe it all to these cutting-edge technological advancements. Amazon Web Services (AWS) provides a number of low-latency alternatives, including Amazon S3 Express One Zone. The need for faster data access is met.

The results of this research indicated that a scalability, responsiveness, and reliability evaluation of cloud storage was necessary. The findings might be valuable for companies and developers looking for storage solutions that can handle high-performance scenarios requiring milliseconds. This data comes from situations when every millisecond counts in the actual world.

## CONCLUSION

In this research, real-world service simulation methods were used in order to investigate the connection that exists between the latency performance and the overall system performance of Amazon S3's distributed storage systems. The distributed design of Amazon Simple Storage Service (S3) has a considerable influence on latency, despite the fact that it is extremely scalable, long-lasting, and accessible. This is caused by a number of variables, including the duplication of data, the demand placed on the server, and your physical distance. This study has yielded important insights into the performance of Amazon S3 in real-world use situations by using simulation methodologies. These insights have been obtained as a result of this research. The results of this research have created a more accurate approximation than those obtained by the use of synthetic benchmarks alone, which has made it possible to develop a depiction that is closer to the truth. This sort of data is required by programming specialists and organisations in order for them to be able to make educated judgements about the creation and enhancement of distributed cloud storage systems. In conclusion, researchers would want to emphasise how important it is to have a solid understanding of latency control for distributed storage systems like Amazon S3. This is a crucial aspect to take into account especially when several individuals are utilising 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. 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.
3. 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.
4. 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.

5. 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.
6. Mahmoudi, N., & Khazaei, H. (2021). SimFaaS: A performance simulator for serverless computing platforms. arXiv preprint arXiv:2102.08904.
7. 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.
8. Xie, M., Qian, C., & Litz, H. (2024, November). En4S: Enabling SLOs in Serverless Storage Systems. In Proceedings of the 2024 ACM Symposium on Cloud Computing (pp. 160-177).
9. Zhou, S., Yuan, B., Xu, K., Zhang, M., & Zheng, W. (2024). The impact of pricing schemes on cloud computing and distributed systems. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 3(3), 193-205.