Analysing The Latency Performance Of Distributed Storage Systems: A Study Of Amazon S3 Utilising Real Service Simulation Techniques

1st Surigue, 2nd Noraisyah Binti Tajudin

Abstract

This research effort used real-world service simulation methodologies to investigate the latency performance of distributed storage systems, namely amazon simple storage service (s3). The researcher used amazon s3's distributed storage systems as the independent variable, the latency performance of services as the dependent variable, and the real service simulation approach as the mediating variable. To get performance data that was typical of a wide range of workloads, a quantitative research approach was used, and system random sampling was used as well. The simulation model was able to mimic real-world scenarios, such uploading, retrieving, and deleting items, by employing a broad variety of object sizes and concurrency levels. The average response time, distribution, and tail latency were used to figure out delay. This was done so that time could be measured. This was done to tell the difference between delays that are normal and those that are important. Researchers found that larger items and more simultaneous access caused latency to go up, whereas smaller things had response times that were more stable across different locations. Based on the evidence, this was the conclusion that was made. To better understand how amazon supply service 3 works, real service simulation methods were employed. These methods precisely replicated the parameters observed in the real world. This study has significantly enhanced understanding of the efficiency of distributed storage by introducing fresh information. These findings validate approaches aimed at enhancing operations conducted via cloud computing.

Keywords: latency performance, storage systems, amazon simple storage service, real-world scenarios,

1. Introduction

As cloud services have grown quickly, distributed storage systems have become an important part of contemporary computing. The fast rise of cloud services has led to these huge changes. Amazon s3 has become a popular distributed storage technology because it can grow and be accessed from anywhere in the world. Latency has become an important performance measure since it directly affects the user experience, the system's responsiveness, and the efficiency of large-scale systems. Most of the study has been on scalability and throughput, but latency has steadily becoming a more important performance measure. This is true even if scalability and throughput have been the main focus. When it comes to getting the most out of the design and use of distributed storage systems, looking at latency performance is an important part that has to be taken into account. There are many things that can affect this performance, such as how much work is being done, what kind of object is being stored, how many people are using the network at the same time, and how spread out the network is across different locations (al-qerem er al., 2021). That being said, learning more about this topic is not only hard, but it also requires a lot of time. Real service simulation approaches have been more popular in the last several years as useful tools for judging how well people do in situations that are as near to real-world operations as feasible. This is because these technologies can provide latency data that is not only continuous but also measurable. Because of this, they are better than synthetic benchmarks, which are benchmarks that are made on purpose instead of being natural. The researcher may acquire a greater knowledge of how distributed storage systems function when confronted with diverse workloads and item sizes via the use of these approaches. This enables

developers and system architects to make more educated decisions than they would have otherwise been equipped to make. The goal of this was to make sure that the data was presented correctly. The goal of this study was to add to the larger conversation regarding the utility of cloud storage by showing real statistics on amazon s3's latency, tail performance, and geographical differences. To successfully achieve this aim, it was essential to focus on reproducing the actual service provided. The goal of this study is to improve both academic research and industrial practice by looking at the latency factors that affect the performance of cloud storage and making ideas for how distributed storage systems may be made better. The main goal of the research is to improve the operation of cloud storage so that it can acquire the best outcomes possible (chen et al., 2022).

2. Background of the study

The need for data management systems that are both scalable and reliable has grown a lot. The rise in demand is due to the rise in popularity of cloud computing. Amazon s3 has rapidly become one of the most popular cloud storage options since it is both highly accessible and long-lasting. S3 has a lot of uses, including web hosting and commercial data analysis. There are a lot of different apps that work with s3. Latency performance has a big effect on how quickly apps respond, how happy users are, and how well the system works as a whole. Even though scalability and throughput are the main measures used to evaluate distributed storage systems, this is still the case. To evaluate the performance of distributed storage, it is required to examine latency; however, this task is complicated by the multitude of factors that influence these characteristics (gupta et al., 2020). Some of these things include the distribution of workloads, the size of objects, the degrees of concurrency, and the characteristics of regional networks. Actual service simulation techniques have been useful for analysing the performance of cloud storage because they can capture actual latency characteristics by better imitating real-world user situations than synthetic benchmarks. This is because they can record real latency characteristics. Academics and practitioners may evaluate the resilience of storage systems to various types of stress via the use of certain methodologies. In this study, distributed storage systems were the independent variable, real service simulation was the mediating variable, and latency performance was the dependent variable. The aim of this research was to investigate the latency performance of amazon s3 using real service simulation approaches. A quantitative method was used to measure and evaluate the data on delay. Also, system random sampling was used to provide a clear picture of the workload needs. These findings are very important since they may be used in several applications, including the improvement of distributed storage efficiency, the influence of architectural decisions, and the development of user experiences in cloud environments (zhang et al., 2023).

3. Purpose of the study

The aim of this study is to investigate the impact of real-world service modelling methodologies on the latency performance of distributed storage systems, specifically focussing on amazon s3 as the subject of examination. Even though distributed storage solutions are widely utilised, latency is still a big issue that slows down productivity, makes users unhappy, and makes applications less responsive. This is happening even though many people utilise dispersed storage solutions. Using real service simulation methods on amazon s3, the researcher can get real latency behaviours for a broad variety of workloads, object sizes, and levels of concurrency. This is possible since amazon s3 is employing these methods. The goal of this research is to provide such knowledge in order to produce an empirical understanding of how this process works with the idea of delivering that information. The study employs a simple random sampling method to guarantee that the acquired data appropriately reflects a diverse array of system reactions across various operating situations. Distributed storage architectures leverage three basic performance criteria. The average reaction time, the

response distribution, and the tail latency are all part of these measures. The goal of this study is to find out how simulation methods affect latency assessment in different ways and to provide examples of those ways. The project aims to achieve optimum design, usage, and management of cloud-hosted storage systems.

4. Literature review

A lot of research has been done on distributed storage systems since cloud computing is now the basis for data management in both personal and business applications. This study shows that distributed storage solutions are becoming more popular. Most people believe that the amazon s3 is one of the most popular distributed storage systems available right now. People know that this product can expand, lasts a long time, and is easy to add to. Still, its latency performance is a big problem that affects both how quickly it responds and how well it works. Latency has recently become a crucial performance indicator for system improvement, even though most studies on cloud storage performance have focused on throughput and availability. This conclusion has been reached, even though throughput and availability have been the main areas of focus (kaur & verma, 2021). Latency in distributed storage depends on a lot of things, such the size of the workload, the kind of item being stored, and the number of users at the same time, and the location of the users. Because of this, it is important to test the system in real-world situations. Real service simulation approaches became an important tool for performance evaluation because they could recreate real-world service environments more accurately than synthetic benchmarks. This is because they were able to copy how services work in the actual world. Quantitative research that uses sampling methods like random request selection makes the results much more reliable. This is because having variety in the datasets is necessary for getting accurate findings. The reasons for this are that the results are more likely to be seen as correct (zhou & zhang, 2019). A recent study's results show that employing simulation methods on services like amazon s3 may provide a latency behaviour models that are more like the real thing. Because of what the discoveries may mean, it's likely that real-world applications could benefit from them. There is still not enough information regarding how simulation-based assessment affects latency performance when there are a lot of different workloads and a lot of people using the same system at the same time. This is especially true when it comes to workloads. In this case, it's quite clear that there is a lack of understanding. Additionally, even if these enhancements have a lot of promise, there is currently less information accessible. The goal of this research was to identify a way to fix these problems by doing a thorough assessment of the latency of amazon s3 systems using real-service simulation and quantitative measurement. The primary emphasis of the study was on suggesting remedies for these inadequacies (shen et al., 2020).

5. Research question

- How do real service simulation techniques influence latency performance?

6. Methodology

6.1 research design: the most recent version of spss, 25, was used for the quantitative analysis. A 95% confidence interval and odds ratio were used to evaluate the strength and direction of the statistical association. The researchers established a threshold of $p < 0.05$ as being statistically significant. The data's essential elements were extracted via an analytical examination. Data evaluated using computational statistical tools and data collected by surveys, polls, and questionnaires are often subjected to quantitative techniques of analysis.

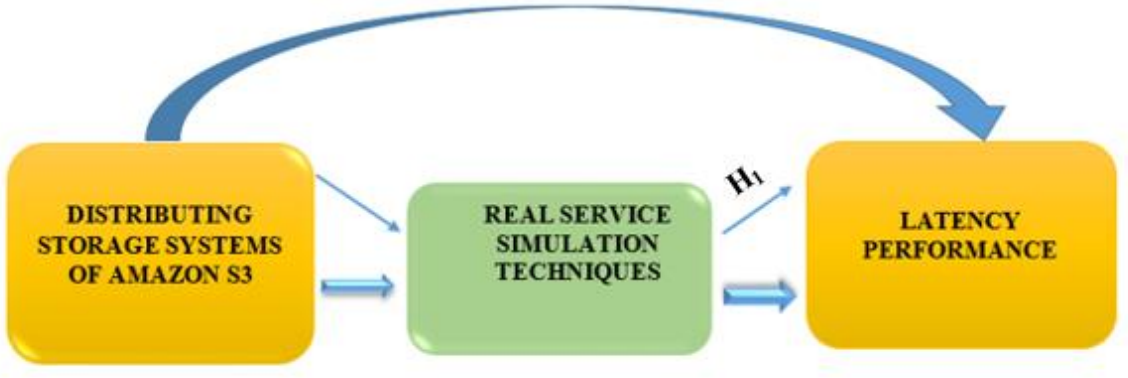
6.2 sampling: research participants filled out questionnaires to provide the research data. Using the rao-soft approach, researchers chose a group of 1,045 participants, which led to a total of 1,316 questions. The researchers received 1265 responses, excluding 96 for incompleteness, resulting in a final sample size of 1169.

6.3 data and measurement: this study mostly used a questionnaire for data collecting. The survey included two parts: (a) a basic demographic part and (b) a part where people were asked to score different parts of both online and offline channels on a 5-point likert scale. It got secondary data from a lot of places, but internet databases were the most important.

6.4 statistical software: the statistical analysis was done using spss 25 and ms-excel.

6.5 statistical tools: descriptive analysis was used to comprehend the essential attributes of the data. The researcher must using anova to examine the data.

7. Conceptual framework



8. Result

❖ **Factor analysis**

A common use of factor analysis (fa) is to validate the fundamental component structure of a collection of measurement items. People think that latent factors, which are not easy to see, have an effect on the scores of the seen variables. The accuracy analysis (fa) method is based on models. The primary objective of this research is to identify correlations among variables, ascertain underlying causes, and quantify errors.

the researcher may utilise the kaiser-meyer-olkin (kmo) method to see whether the data is good enough for factor analysis. The researcher verifies if the sample size is sufficient to accurately reflect the whole model and each constituent variable. The statistical metrics show how much variation different variables may share. Factor analysis works better with data that is shown by smaller percentages.

Kmo gives a whole number between 0 and 1. The sample is considered good when the kmo value is between 0.8 and 1.

A kmo value of less than 0.6 means that the sample size is too small, which means that something has to be done. Researcher should use best opinion; other writers have chosen 0.5 for this, therefore the range is 0.5 to 0.6.

a kmo score near to 0 suggests that the partial correlations are more important than the overall correlations. Big correlations make it very hard to do component analysis.

here are the standards that kaiser thinks are okay:

desolate 0.050 to 0.059.

0.60 to 0.69 is below average

standard range for middle school: 0.70 to 0.79.

a quality point value between 0.80 and 0.89.

the range from 0.90 to 1.00 is quite outstanding.

Table 1: 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

Bartlett's test of sphericity further confirmed the significance of the correlation matrices. The kaiser-meyer-olkin measure of sampling adequacy is 0.877. Using bartlett's sphericity test, the researchers got a p-value of 0.00. Bartlett's sphericity test showed that the correlation matrix is wrong.

- **Dependent variable**

- **Latency performance**

"latency" is a common term used to talk about how well a computer or storage system works. The amount of time it takes to analyse, deliver, and receive data is what matters in this case. In the context of a distributed storage system, this phrase means the amount of time it would take to finish a client request from start to finish. Uploading, downloading, and retrieving an object are all examples of client requests. Latency is an important performance metric to think about because it affects how fast applications can provide real-time or near-real-time services, how users see the system, and how quickly the system reacts (zhang & xu, 2021). Latency performance is a method that looks at how quickly each request is processed. This method is used to judge how good a service is. Throughput, on the other hand, is a way to quantify how much data is processed in a specific amount of time. This is not the same as throughput. Using the distributional variances, the average response time, and the tail latency statistics, everyone can have a better idea of how the system usually works and how it doesn't work. This is because the system works and doesn't work at the same time. There is delay in things like financial platforms, interactive apps, and internet services. The internet is another example of this. An excessive delay may lead to inefficiencies, diminished productivity, and suboptimal application performance under such scenarios. Latency performance changes based on things like how much demand there is, how big the item is, how many people are using it at once, and the state of the network. A recent study, however, showed that latency is an important factor to think about when looking at distributed storage systems like amazon s3 (bao et al., 2025). This

illustrates that latency is a crucial thing to think about when looking at distributed storage systems. It is very important to study latency performance so that cloud services are always accessible.

❖ **Mediating variable**

Real service simulation techniques

Real service simulation methods are the ones that copy the real-world conditions that storage or computing systems face when it comes to performance testing. These methods are used to check how well the system works. Because of this, it is now possible to do a more accurate study of how the systems work when they are used in real-world situations. Utilise synthetic benchmarks, the researcher make up fake workloads that may not properly show how complicated things are in the real world. On the other hand, real service simulations try to copy real user actions including uploading, downloading, deleting, and making a lot of access requests at once. Synthetic benchmarks, on the other hand, create workloads that aren't truly possible when compared to genuine benchmarks. If researchers apply these methods, which try to mimic how clients and distributed systems interact as closely as possible (polat et al., 2024), they may be able to measure performance metrics like dependability, throughput, and latency more precisely. By using authentic service simulation methodologies inside cloud storage systems such as amazon s3, one may enhance their comprehension of how varying workloads, object size, and geographical factors affect the responsiveness of diverse services. This kind of technology is used in the s3 system. They allow to look at tail latency, distributional variability, and average response time in a wide range of operating conditions. The data they provide makes this feasible. This is a possibility since they can accomplish so many things. Users and designers of the system may use this method to make informed assumptions about how the system behave when it is under stress, find performance bottlenecks, and improve resource allocation. Recently, researchers in cloud computing and distributed computing have been more interested in using real service simulation (bao et al., 2025). This attention has been seen in recent years. This method provide a true way to see how well large-scale storage systems work in terms of scalability, how well they work in general, and how users feel about them.

- **Relationship between real service simulation techniques and latency performance**

The most important thing to think about when trying to understand the link between latency performance and real-service simulation methods is how well simulation methods can mimic real-world operational aspects that affect the responsiveness of distributed storage systems. Latency performance is greatly affected by a variety of things. Some of them include the amount of work to be done, the number of people using it at the same time, and where the users are located. Latency performance is the amount of time it takes to accomplish storage functions like uploading, downloading, or getting data. Traditional benchmarking methods generally failed to capture this complexity since they relied on simulated workloads that don't accurately reflect how real users behave. This was because these methods employed simulated workloads. Genuine service simulation methods may create realistic situations by using genuine service requests. This makes it feasible to get rid of this limitation and get a more realistic picture of latency in a variety of situations. These methods let the researchers look at the average latency, response distributions, and tail performance. Because of this, kids learnt about both the normal and the unusual delays. The methods used in the simulation of real workloads showed how the size of the item and the number of requests being sent at the same time affect the response times of systems like amazon s3. The information gained from these simulations was very

helpful in several areas, such as improving storage design, finding latency bottlenecks, and making cloud computing easier for users. Actual service simulation methods helped to close the gap between theoretical models and real-world findings, which made latency performance analysis in distributed storage systems much better. This was done by connecting the two. As a direct consequence of this, this region made a lot of progress in its growth (hoang et al., 2025).

Following the above discussion, the researcher developed the following hypothesis to examine the correlation between real service simulation techniques and latency performance.

H_{01}: "there is no significant relationship between real service simulation techniques and latency performance".

H_1: "there is a significant relationship between real service simulation techniques and latency performance".

Table 2: H_1 ANOVA Test

ANOVA					
Sum					
	Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
Between Groups	39588.620	412	5645.512	999.914	.000
Within Groups	492.770	756	5.646		
Total	40081.390	1168			

This investigation has led to a noteworthy finding. With a p-value of 0.000, which is lower than the alpha criterion of 0.05, the significance is shown by an f value of 999.914. The alternative hypothesis, ***"h_1: there is a significant relationship between real service simulation techniques and latency performance,"*** is accepted and the null hypothesis is rejected.

9. Discussion

In this study, real-world service simulation approaches were used to evaluate the latency performance of distributed storage systems. The amazon s3 was the main focus of this study. These methods were able to accurately mimic genuine interactions with services, which made it possible to measure delays in a setting that was similar to what consumers really go through. The findings showed that these methods were able to accomplish this. When compared to synthetic benchmarks, real service simulations showed system performance in a more realistic way. They were able to get these findings because they took into account changes in the network, workload, and the number of requests that were being handled at the same time.

The study's findings showed that the architecture of distributed storage systems, together with operational factors like data replication mechanisms and traffic intensity, had an influence on latency performance. It was proven that these things affected latency performance. When it comes to finding performance bottlenecks, real-world service simulations are quite important. This is because these simulations demonstrated delays in response time that weren't clear with other methods. This happens because they show bottlenecks that were hidden before.

The findings also showed that different modelling methods gave ideas that may be used to make the system work better. In order to improve their caching, load balancing, and fault tolerance methods, service providers need to know what kinds of delays happen when real workloads are running. This information might make the user experience better. It was decided that real service simulation methods were necessary to build cloud storage solutions that were more effective and reliable. This was decided throughout the discussion that took place. The goal was reached by linking theoretical performance models to how actual systems work.

10. Conclusion

The research, which used amazon s3 as a case study, found that actual service simulation approaches may be used to reliably and realistically measure the latency performance of distributed storage systems. This was found out by using amazon s3. The testing of the study led to the discovery of this knowledge. By imitating the workloads that people deal with in real life, these methods make it feasible to have a better idea of how the system works. They found differences in reaction time that standard benchmarking didn't always reveal. This was a big deal for them. Because of this, they were able to see these patterns change. The results show that many different things can affect how well someone does on a delay. It has been shown that these factors have an effect. Some examples of these include the workload's changing nature, the network's state, the way requests come in at the same time, and the way amazon s3 is set up. The findings of the study together suggest that distributed storage systems might use actual service simulation methodologies to improve the system's efficiency, reliability, and user experience. The study's results point to this. Furthermore, the study's results demonstrate that latency evaluations may be conducted with much enhanced accuracy. This demonstrated the significance of simulation-based methodologies, essential for the advancement of cloud performance research and practice. Because of this, it was a big deal.

References

1. al-qerem, a., alauthman, m., gupta, b., & razaque, a. (2021). Cloud storage performance and security: techniques and challenges. *Journal of cloud computing, 10*(1), 1-20.
2. Bao, z., liu, q., huang, x. J., & wei, z. (2025). Sfmss: service flow aware medical scenario simulation for conversational data generation. In *findings of the association for computational linguistics: naacl 2025* (pp. 4586-4604).
3. Chen, j., li, x., & liu, y. (2022). Latency-aware data management in distributed cloud storage systems. *Future generation computer systems, 128*, 276–286.
4. Gupta, h., sharma, a., & singh, r. (2020). Performance evaluation of cloud storage systems using real service workloads. *International journal of cloud applications and computing, 10*(3), 34–47.
5. Hoang, t. T., pham, l. M., & nguyen, h. S. (2025). Lavp: a latency-aware virtual network function placement strategy for service function chain in network function virtualization. *Ieee access*.
6. Kaur, s., & verma, p. (2021). Evaluating performance of amazon s3 under varying workloads. *Journal of grid computing, 19*(2), 1–18.
7. Polat, o., oyucu, s., türkoğlu, m., polat, h., aksoz, a., & yardımcı, f. (2024). Hybrid ai-powered real-time distributed denial of service detection and traffic monitoring for software-defined-based vehicular ad hoc networks: a new paradigm for securing intelligent transportation networks. *Applied sciences*, *14*(22), 10501.
8. Shen, h., zhang, y., & xu, j. (2020). Performance modelling and analysis of large-scale distributed storage systems. *Ieee transactions on parallel and distributed systems, 31*(12), 2871–2885.
9. Singh, g. D., tripathi, v., dumka, a., rathore, r. S., bajaj, m., escorcia-gutierrez, j., ... & prokop, l. (2024). A novel framework for capacitated sdn controller placement: balancing latency and reliability with pso algorithm. *Alexandria engineering journal*, *87*, 77-92.
10. Van damme, s., sameri, j., schwarzmann, s., wei, q., trivisonno, r., de turck, f., & torres vega, m. (2024). Impact of latency on qoe, performance, and collaboration in interactive multi-user virtual reality. *Applied sciences*, *14*(6), 2290.
11. Zhang, y., & xu, h. (2021). Latency performance challenges in large-scale distributed systems. *Journal of cloud computing, 10*(1), 1–15.

12. Zhang, y., wang, c., & huang, j. (2023). Real-time performance analysis of distributed storage in cloud computing environments. *Ieee transactions on cloud computing, 11*(2), 178–190.
13. Zhou, l., & zhang, q. (2019). Performance evaluation of cloud storage services with real workloads. *Journal of cloud computing, 8*(1), 1–13.