Optimizing Performance of Cloud Infrastructure Through Effective Resource Scheduling

M.Abdullah[1] **Dr. M. Mohamed Surputheen**[1*†]

[1] *Department of Computer Science, Jamal Mohamed College,Trichy, Tamil Nadu (Affiliated to Bharathidasan University)*

Abstract

Cloud computing has emerged as a very promising technology that has garnered significant interest from both industry professionals and academic researchers. Cloud computing service models refer to the various types of services that are provided, including hardware and software infrastructure, platforms for application development, testing, and deployment, as well as enterprise software that is readily available for usage through subscription. Public cloud computing involves the delegation of IT infrastructure, storage, or applications to an external service provider. The presence of a cloud infrastructure also signifies the existence of geographically distributed computing resources. The utilisation of resources in conjunction with cloud computing is not exclusive to large-scale organisations, as it may be employed by entities of any size. Numerous services are offered based on a fee-for-use model, rendering them cost-effective for organisations of various sizes. Cloud service providers are obligated to provide consumers with cloud services on demand, as there is a growing need for such services. This requirement stems from the necessity to decrease the size of large data volumes, which in turn leads to cost savings in maintaining extensive storage systems. The overall effectiveness of cloud computing environments is directly related to the operational performance of cloud infrastructure. This phenomenon holds substantial significance in the realm of optimization, since it enhances the overall efficiency of the underlying cloud architecture. The proposed technique exhibits a significant effectiveness in enhancing cloud performance, as it manifests improvements for both service providers and cloud customers.

Keywords: Cloud Computing, Task Scheduling, Cloud Infrastructure, Resource Scheduling, Performance Analysis and Infrastructure as Service

*Corresponding author.
†E-mail: msurfudeen@yahoo.com*

1 INTRODUCTION

The development of cloud computing has significantly impacted our perception and approach to manage cloud infrastructure. At present, service providers for software-based services possess the capability to allocate the requisite hardware infrastructure for those services [1]. However, as the utilization of cloud infrastructure continues to grow, users are becoming increasingly aware of the reality that underlying expectations associated with cloud computing. A possible explanation for this phenomenon is that existing cloud systems provide interfaces that are closely tied to the underlying infrastructure. But, there is an increasing client need for features which possess the capability to organize their services. A fundamental characteristic shared by all of these models is their endeavour to provide infrastructure as a utility, enabling service providers to effectively operate their software-based systems [2]. The term "information and communications technology" (ICT) is quite broad, encompassing not only computers but also a wide variety of communication tools and software. Traditional methodologies within the realm of information and communication technology focus on addressing individual or a limited number of system problems during the design, implementation, and validation processes. These variables may involve the validation of functional accuracy and dependability within an organisational system, along with the management of security and privacy concerns inside a database. The field of information communications technology is experiencing significant growth at a quick pace. The integration of emerging cyber-physical system technology into forthcoming smart city applications could potentially have discernible environmental consequences, which may manifest as either advantageous or detrimental.

2 Setting up Resource Provisioning in Cloud Environment

Computer optimization for a wide range of applications is a key paradigm. Resource provision is the process of choosing, implementing, and managing software and hardware resources during runtime to make sure that safe applications work efficiently. Cloud computing is a way of doing business that lets you use a shared pool of flexible computer resources by granting you easy, on-demand network access. The flexibility with which cloud users can allocate their allocated resources is a major benefit of this technology [3]. This system offers an easy-to-use, low-cost central platform for services. The cloud's services can be combined into a single, distributed object that can be leased or rented as-needed basis. Flexibility is an important part of cloud computing because it lets cloud resources grow or shrink in real time. Users of the cloud have more wants than ever, so resources need to be allocated well. However, as the number of

cloud users grows, so does the need for efficient resource sharing. The cloud has a lot of tools that end users can use. All they have to do is pay the service provider for their services. There are three main types of cloud services: SaaS, PaaS, and IaaS. Each has its own pros and cons. Once a user has chosen a service model, they need to take care of the client's responsibilities. You make sure that the maintenance and control of your systems and apps are done when you use a cloud service. In a cloud setup, the only thing that is different is that the provider is in charge of other things. No matter what, you need to make sure these things are taken care of. When services are bought, they need to make sure that the cloud system has enough resources to handle growing demand.

3 Allocation of Resources within a Cloud Environment

Service providers have the potential to enhance their work plan and resource management in order to achieve optimal utilization of resources [4]. The allocation and distribution of resources, as well as the timeliness of their supply, are pivotal factors that significantly impact the efficacy and popularity of cloud computing services. Planning and scheduling technique aims to be effectively manage incoming requests (tasks) and optimize resource utilization [5]. Subsequently, individuals are required to submit their applications via the online platform, wherein the volume of user registrations may give rise to concurrent generation of several requests or tasks. Certain short-term assignments may experience delays due to extended waiting periods. During the process of scheduling, the scheduler is faced with several limits that need to be addressed, including the time required to finish a task, the availability of resources, and the work queue. The availability of resources in a Cloud Service can be influenced by the extent to which cloud end users are able to utilize the complete computer stack, encompassing both hardware and software. This feature, which is considered a significant advantage of cloud computing, has the potential to either diminish or enhance resource availability [6]. There exists a cloud service framework definition that enables users to select and allocate resources based on their desired usage. The cloud service encompasses obligatory elements that facilitate the organization of activities and the distribution of resources. The efficiency of resource allocation is contingent upon the implementation of design and load balancing procedures, rather than relying on random allocation of resources. The utilization of scheduling algorithms is advised for addressing intricate tasks that involve the utilization of planning algorithms. The optimization of resource utilization can result in the effective allocation of jobs, hence maximizing the benefits derived from cloud computing [7]. At now, anyone utilizing the Internet have the ability to retrieve

content from any geographical place without regard for the underlying hosting infrastructure. The hosting infrastructure necessitates a substantial number of machines that are managed by the service provider. Hence, the task planning and resource allocation in the cloud environment hold considerable significance, as they mutually impact one another. The supply of cloud-based services to cloud customers is perceived as a beneficial aspect of cloud computing, as it enhances infrastructure capabilities by offering internet connectivity.

4 Enhancing the Efficiency of Cloud Computing

The adoption of mainstream technology in this arena has been observed to occur at an unprecedented rate compared to other advancements. The primary driver behind its acceptance is mostly attributed to the proliferation of smartphones and mobile devices that have the capability to access the Internet [8]. Cloud computing has shown to be advantageous not only for corporations and organisations, but also for the general populace. The technology enables the execution of software programmes on local workstations without the need for downloading, facilitates the storage and access of multimedia information through internet connectivity, and provides a platform for the development and testing of programmes without the requirement of dedicated servers, among other functionalities [9]. The utilization of essential cloud computing infrastructure has the potential to effectively address various contemporary challenges. Above concerns pertain to the acquisition and administration of expensive hardware and software services utilised by our organisation in our routine operations, as well as the effectiveness of these resources in serving our objectives and benefiting society at large. Cloud computing offers numerous advantages in addressing these challenges that have surpassed our expectations and provided outcomes beyond what we had envisioned. Cloud computing has been demonstrated to provide significant advantages for corporate enterprises through the availability of communal computing capabilities [10]. The aforementioned advantages can be categorised into three distinct groups, specifically Efficiency, Flexibility, and Strategic Edge.

5 Existing Methods for Resource Scheduling

5.1 Delay Scheduling (DS)

Our challenges arise from the need to schedule a work without local data in order to adhere to a specific queue order. Researchers address this issue by the implementation of a simple technique referred to as delay scheduling. In a particular occurrence, an instance commences a task. If the task at the forefront of the queue is unable to launch a local task, it is skipped and

focus is shifted towards the subsequent tasks [11]. Nevertheless, in the event that a particular job has been neglected for a significant duration, they begin to permit it to initiate non-local duties as a means of preventing deprivation. The primary concept underlying delay scheduling is that while the first time slot being considered for work allocation may not have the necessary data, the rapid completion of jobs ensures that a subsequent time slot with the required data will become available within a few seconds.

5.2 MaxCover-BalAssign (MB)

The algorithm operates in an iterative manner to generate a succession of complete allocations, and subsequently outputs the optimal allocation. The process comprises of two distinct phases, namely maxcover and balassign, which are repeated in each iteration. Given that the precise cost of the remote is not known, an estimation of the cost, referred to as the virtual cost, is computed [12]. This virtual cost serves as a predictive approximation of the actual cost of the remote. Subsequently, a comprehensive calculation is performed to determine the aggregate value. The allocation process can be optimised by leveraging the virtual cost.

5.3 Hadoop Default Scheduler (HDS)

The Scheduler utilised in Hadoop is the default option. The jobs are sequentially enqueued and executed in the order of their submission [13]. In this approach, after the task has been planned, any form of intervention is strictly prohibited. When a server is not actively involved in task processing, the scheduler will choose data and subsequently assign it to the server. In the absence of a viable task, the scheduler will opt for the selection of a task at random.

6 Proposed Method for Resource Scheduling

The Formal method is employed in the development of Cloud environment and procedures for the purpose of fault prevention and the facilitation of error elimination. This work presents a critical analysis of a resource scheduling framework that aims to enhance efficiency. The Energy-Efficient Framework based on Fuzzy Scheduling emphasises the recognition that the structure is indicating a transition in state. The careful choice of parameters is crucial in ensuring the efficacy of scheduling processes. When considering real-time activities, there are three primary scheduling characteristics that need to be taken into account: computation time, arrival time, and deadline. These parameters characteristics are of utmost importance in determination of

the order in which actions are carried out, with the ultimate goal of attaining a single output. The fuzzy scheduling incorporates scheduling parameters into real-time operations. The amount of time needed for computation is dependent on the variables of reaction time and usage. The concept of the membership function is a fundamental phrase that is defined within the context of a fuzzy set, and it functions on the broad terrain. The implementation of data deduplication technology can be utilised as a means to reduce storage capacity in cloud environments. This can be achieved by the utilization of hash functions, which assign values to link segments of data with duplicate data, resulting in the preservation of a single copy while generating logical pointers to supplementary copies. The utilisation of de-duplication techniques yields several advantages.

- The quantity of data stored in disc space is diminished.
- It is anticipated that there will be a reduction in the energy usage of storage systems.
- Within the context of cloud disaster recovery, the process of deduplication plays a crucial role in enhancing data replication efficiency, resulting in accelerated replication time and reduced bandwidth expenses.
- Data deduplication is a process that effectively minimises physical storage requirements and network bandwidth utilisation, hence serving as a means of backup storage in cloud computing environments.

The quantity of factors, particularly the accessibility of resources and their usage. To assess the efficacy of the proposed approach, it has been subjected to comparison with other job scheduling techniques, specifically Delay Scheduling (DS), MaxCover-BalAssign (MB) and Hadoop Default Scheduler (HDS). The provided information offers an introductory overview of several features, encompassing throughput, latency, average response time, availability and scalability of resource utilisation.

7 Results and Discussion

Performance analysis is a technique employed to compare and evaluate various representations acquired within a certain context. The logical assessment assesses the network metrics, specifically the amount of files with varying data sizes (in megabytes) that are uploaded in a cloud environment. This evaluation is conducted by considering characteristics such as Throughput, Latency, Average Response Time, Availability, and Scalability. The reports are

Throughput (kbps)				
No of Files (different data size (Mb))	DS	MB	HDS	ERSF
10	15100	15377	16100	16151
20	14850	15260	15978	16050
30	14707	15167	15853	16001
40	14709	15113	15776	15941
50	14530	14985	15778	15883
60	14472	14954	15731	15898
70	14448	14936	15740	15870
80	14449	14834	15724	15761
90	14340	14839	15644	15722
100	14279	14878	15652	15752

TABLE 1

Throughput for Data Transfer using 5 Deduplicators

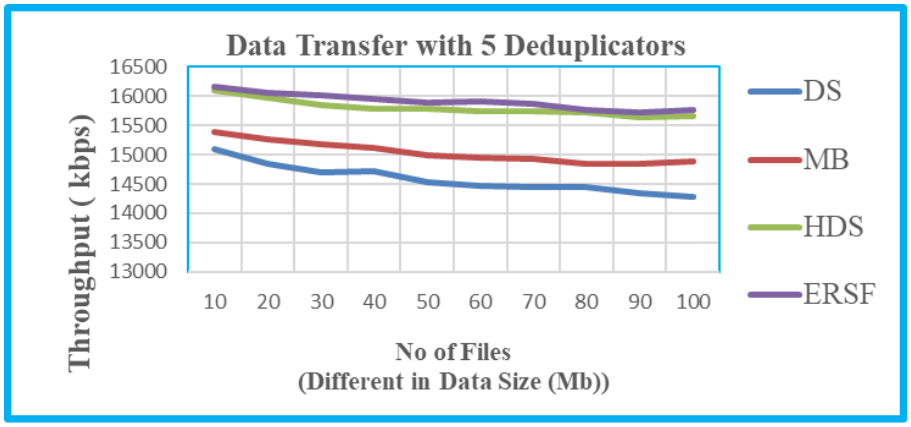

FIGURE 1

Throughput for Data Transfer using 5 Deduplicators

generated using known methods such as Delay Scheduling (DS), MaxCover- Balassign (MB), Hadoop Default Scheduler (HDS), as well as the proposed approach of the Efficient Resource Schedule Framework.

7.1 Comparative Analysis of Throughput

Table 1 presents the Throughput values (in kilobits per second) achieved by both the existing and proposed task scheduling methodologies with 5 Deduplicators.

Latency (mS)				
No of Files (different data size in (Mb))	DS	MB	HDS	ERSF
10	591	557	492	485
20	666	618	552	521
30	691	653	584	572
40	724	681	599	571
50	751	703	632	590
60	755	723	631	615
70	782	736	642	634
80	782	754	658	632
90	798	765	665	643
100	808	770	676	652

TABLE 2

Latency for Data Transfer using 5 Deduplicators

Figure 1 depicts the graphical illustration of the Throughput (measured in kilobits per second) achieved by the existing methods such as Delay Scheduling (DS), MaxCover-Balassign (MB), Hadoop Default Scheduler (HDS), and proposed method Efficient Resource Schedule Framework (ERSF). When compared with the existing scheduling methods, it is clear that the proposed Efficient Resource Schedule Framework (ERSF) has higher throughput.

7.2 Comparative Analysis of Latency

Table 2 displays the delay values, measured in milliseconds, attained by the existing and proposed task scheduling algorithms with 5 Deduplicators. Figure 2 presents a visual representation of the attained Latency, measured in millisecond, using existing and proposed scheduling methods such as Delay Scheduling (DS), MaxCover Balassign (MB), Hadoop Default Scheduler (HDS), and Efficient Resource Schedule Framework (ERSF).

Based on the examination of Table 2 and Figure 2, it is apparent that the suggested proposed Efficient Resource Schedule Framework (ERSF) exhibits a higher level of Latency when compared with existing task scheduling methodologies.

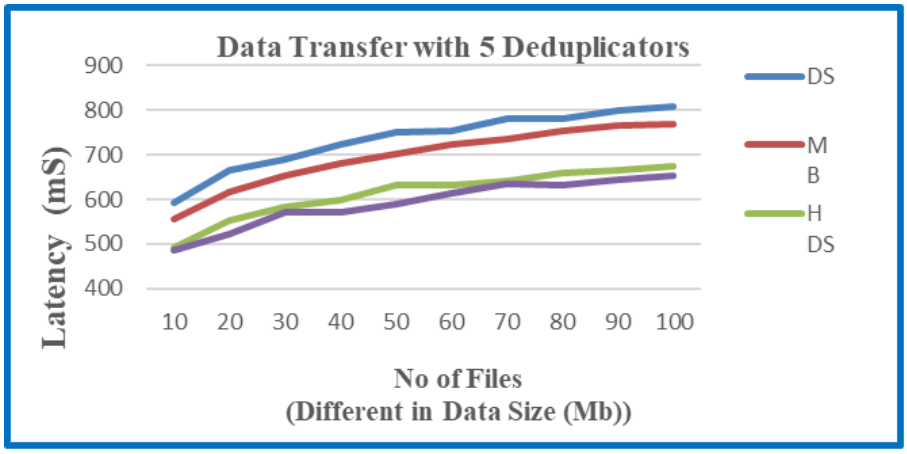

FIGURE 2

Latency for Data Transfer using 5 Deduplicators

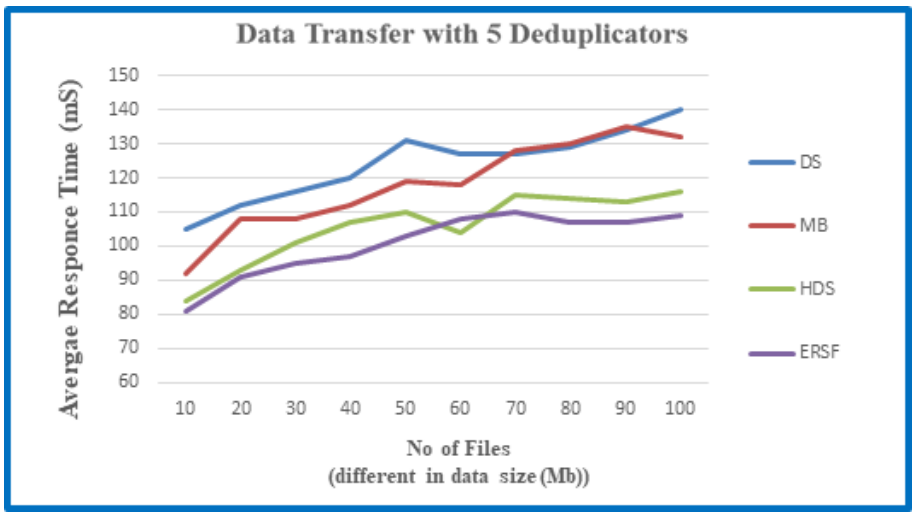

FIGURE 3

Average Response Time for Data Transfer using 5 Deduplicators

7.3 Comparative Analysis of Average Response Time

Table 3 displays the Average Waiting time (measured in milliseconds) attained by the existing and proposed task scheduling approaches utilising with 5 Deduplicators. Figure 3 presents a visual representation of the Average Waiting time (in milliseconds) produced by the existing methods such as Delay Scheduling (DS), MaxCover-BalAssign (MB), Hadoop Default Scheduler (HDS), and proposed method as Efficient Resource Schedule Framework (ERSF). Based on the examination of Table 3 and Figure 3, it is apparent that the proposed Efficient Resource Schedule Framework (ERSF) exhibits a higher level of performance in terms of Average Waiting time (measured in milliseconds) when compared to existing task scheduling methodologies.

Average Response Time (mS)				
No of Files (different data size (Mb))	DS	MB	HDS	ERSF
10	105	92	84	81
20	112	108	93	91
30	116	108	101	95
40	120	112	107	97
50	131	119	110	103
60	127	118	104	108
70	127	128	115	110
80	129	130	114	107
90	134	135	113	107
100	140	132	116	109

TABLE 3

Average Response Time for Data Transfer using 5 Deduplicators

Availability (%)				
No of Files (different data size (Mb))	DS	MB	HDS	ERSF
10	54.88660	54.95917	55.03256	55.05344
20	54.49529	54.61023	54.72113	54.74320
30	54.22170	54.36604	54.49901	54.53185
40	54.01121	54.15924	54.32348	54.36359
50	53.83282	54.00412	54.18596	54.22405
60	53.69135	53.87590	54.06289	54.12004
70	53.55436	53.76316	53.95979	54.01919
80	53.44556	53.65194	53.86805	53.92355
90	53.34793	53.55718	53.77940	53.85255
100	53.24216	53.47277	53.71311	53.77524

TABLE 4

Availability for Data Transfer using 5 Deduplicators

7.4 Comparative Analysis of Availability

Table 4 presents the achieved Availability, measured in percentage, using the existing and proposed work scheduling methodologies employing with 5 Deduplicators. Figure 4 depicts a graphical illustration of the Availability, measured in percentage, as influenced by the existing

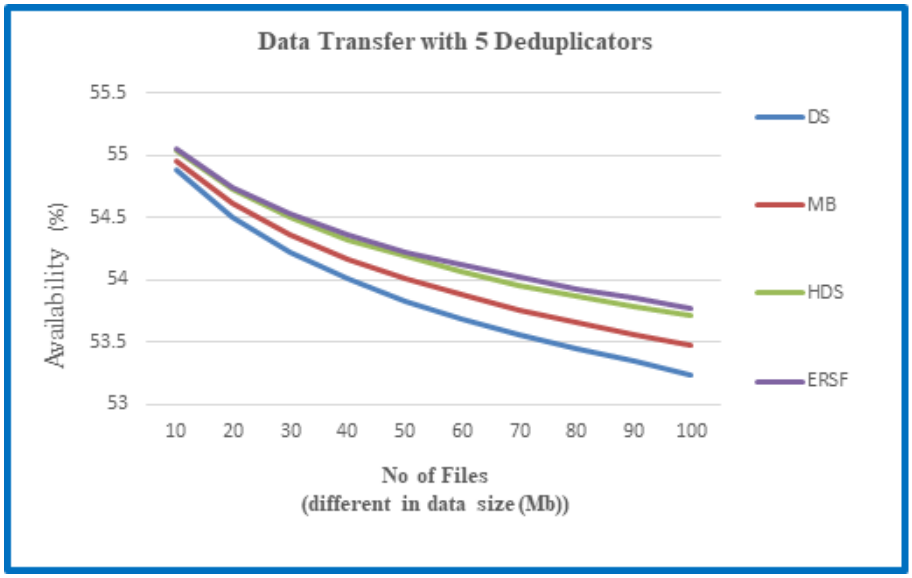

FIGURE 4

Availability for Data Transfer using 5 Deduplicators

methods such as Delay Scheduling (DS), MaxCover-BalAssign (MB), Hadoop Default Scheduler (HDS), and the proposed Efficient Resource Schedule Framework (ERSF). Upon analysing Table 4 and Figure 4, it becomes evident that the proposed Efficient Resource Schedule Framework (ERSF) demonstrates superior performance in terms of Availability (measured as a percentage) in comparison with the existing task scheduling approaches.

7.5 Comparative Analysis of Scalability

Table 5 presents the Scalability, expressed in percentage, achieved by the existing and proposed work scheduling methods employing with 5 Deduplicators. Figure 5 depicts a graphical illustration showcasing the Scalability (expressed in percentage) achieved by the existing methods such as Delay Schedule (DS), MaxCover-BalAssign (MB) and Hadoop Default Scheduler (HDS), and proposed method as Efficient Resource Schedule Framework (ERSF).

Upon analysing the data shown in Table 5 and Figure 5, it becomes evident that the proposed Efficient Resource Schedule Framework (ERSF) has superior performance in terms of Scalability (measured in %) in comparison with the existing task scheduling approaches.

8 Conclusion

The proliferation and widespread adoption of the Internet environment is expediting transformation within the cloud data landscape. The dissemination of data is seeing significant growth alongside the exponential expansion in data, resulting in a corresponding rise in com-

Scalability

No of Files (different data size (Mb))	DS	MB	HDS	ERSF
10	54.77197	54.88042	54.92897	54.95917
20	54.31380	54.47364	54.55397	54.60537
30	53.99405	54.19069	54.28697	54.36117
40	53.73695	53.96427	54.08563	54.16653
50	53.53426	53.78472	53.91766	54.00899
60	53.35053	53.63142	53.77633	53.87428
70	53.20235	53.50014	53.65471	53.75019
80	53.06554	53.39011	53.54996	53.65680
90	52.95255	53.28427	53.44606	53.55799
100	52.83880	53.18512	53.36112	53.47683

TABLE 5

Scalability for Data Transfer using 5 Deduplicators

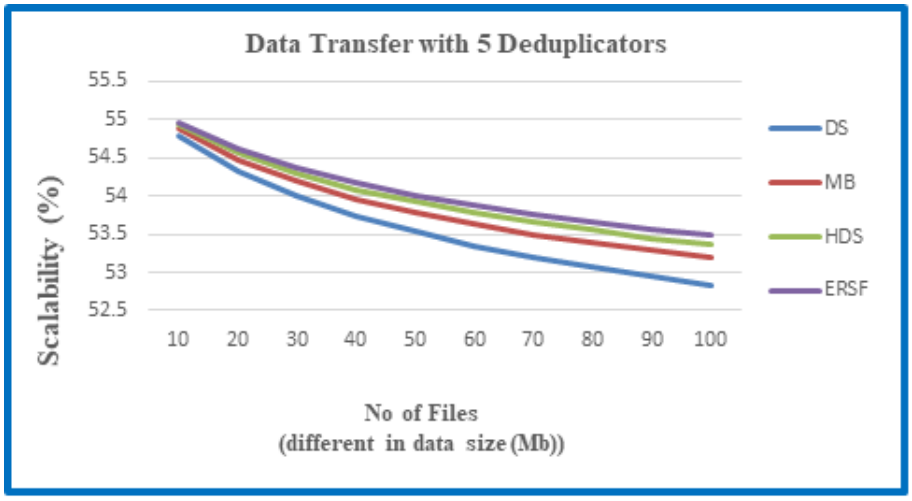

FIGURE 5

Scalability for Data Transfer using 5 Deduplicators

plexity as data volume expands. Cloud services should be readily accessible to customers at all times. However, suppliers are responsible for ensuring the system's availability and managing large volumes of data. Cloud computing offers a diverse array of solutions for large-scale applications by utilising a multi-cloud architecture. The performance of this infrastructure is a crucial concern, since it directly impacts the whole functionality of the cloud environment. Efficient time management is essential for cloud computing services within the cloud architecture, given the restricted resources available. Nevertheless, it is imperative that every resource be dedicated

to a task that consumes energy in a distinct manner. A substantial duration of processing time was necessary to transmit the streaming data to multiple sites. Furthermore, the efficacy of data dissemination performs a crucial and central role in the cloud computing infrastructure. Given the large quantity of servers and the variability in processing workflow, it is recommended to employ a fuzzy-based scheduling method in order to solve those problems. The primary objective of this study is to incorporate an Efficient Resource Scheduling Framework scheme with the purpose of enhancing the overall performance of cloud infrastructure.

References

[1] K.S. Arya, P.V. Divya, and K.R.Remesh Babu. Dynamic resource management through task migration in cloud. In *International Conference on Intelligent Data Communication Technologies and Internet of Things, Book Series LNDECT*, volume 26, pages 1362–1369,.

[2] Sururah A. Bello, Lukumon O. Oyedele, Olugbenga O. Akinade, Muhammad Bilal, Juan Manuel Davila Delgado, Lukman A. Akanbi, Anuoluwapo O. Ajayi, and Hakeem A. Owolabi. Cloud computing in construction industry: Use cases, benefits and challenges. *Automation in Construction*, 122.

[3] Rajkumar Buyya and Chee Shin Yeo. Srikumar venugopal, james broberg, ivona brandic, cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility". *Future Generation Computer Systems*, 25(6):599–616,.

[4] L. Golightly, V. Chang, Q.A. Xu, X. Gao, and B.S. Liu. Adoption of cloud computing as innovation in the organization. *International Journal of Engineering Business Management*, 14.

[5] Ahyoung Kim, Junwoo Lee, and Mucheol Kim. Resource management model based on cloud computing environment. *International Journal of Distributed Sensor Networks*, 12(11):1–7.

[6] Haimantee Mahato, Anjali Munjal, and Shreya Chinchalikar. A system for task scheduling and task migration in cloud environment. *IOSR Journal of Computer Engineering (IOSR-JCE*, 16(2):115–118,.

[7] Chinthagunta Mukundha and K. Vidyamadhuri. Cloud computing models: A survey. *Advances Computational Sciences and Technology*, 10(5):747–761,.

[8] Eleonora Pantano. Constantinos-vasilios priporas, the effect of mobile retailing on consumers' purchasing experiences: A dynamic perspective. *Computers in Human Behavior*, 61:548–555,.

[9] Luis Rodero-Merino, Luis M. Vaquero, Victor Gil, Fermín Gala, Javier Fontan, Ruben S. Montero, and Ignacio M. Llorente. From infrastructure delivery to service management in clouds". *Future Generation Computer Systems*, 26(8):1226–1240,.

[10] Aarti Singh, Dimple Juneja, and Manisha Malhotra. A novel agent based autonomous and service composition framework for cost optimization of resource provisioning in cloud computing. *Journal of King Saud University, Computer and Information Sciences*, 29(1):19–28,.

[11] Mohd Usama, Mengchen Liu, and Min Chen. Job schedulers for big data processing in hadoop environment: testing real-life schedulers using benchmark programs. *Digital Communications and Networks*, 3(4):260–273,.

[12] Blesson Varghese and Rajkumar Buyya. Next generation cloud computing: New trends and research directions. *Future Generation Computer Systems*, 79(3):849 –861,.

[13] Matei Zaharia, Dhruba Borthakur, Joydeep Sen Sarma, Khaled Elmeleegy, and Scott Shenker. Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In *EuroSys '10: Proceedings of the 5th European conference on Computer systems*, pages 265 – 278,.