Received: 18 August 2022 | Revised: 3 February 2023 | Accepted: 11 April 2023

DOI: 10.1002/cpe.7758

WILEY

RESEARCH ARTICLE

Review on data replication strategies in single vs. interconnected cloud systems: Focus on data correlation-aware strategies

Tarek Hamrouni[1] | Riad Mokadem[2] | Amel Khelifa[1]

[1]LIPAH, Faculty of Sciences of Tunis, Tunis El Manar University, University Campus, Tunis, Tunisia

[2]Institut de Recherche en Informatique de Toulouse (IRIT), Paul Sabatier University, Toulouse, France

Correspondence
Tarek Hamrouni, LIPAH, Faculty of Sciences of Tunis, Tunis El Manar University, University Campus, Tunis, Tunisia.
Email: tarek.hamrouni@fst.rnu.tn

Summary

Data replication is a well-known technique in cloud systems for enhancing availability and performance. Various strategies and surveys have been proposed in this respect. These surveys include comprehensive analysis and classifications. However, to the best of the authors' knowledge, there is no survey concentrating on strategies designed for interconnected cloud systems. In this article, we provide an in-depth analysis of existing data replication strategies in cloud systems, covering single and interconnected clouds. We also highlight data correlation-aware strategies as well as their key steps. Furthermore, we examine the major strategies' features such as: (*i*) addressed replication issues, (*ii*) orientation towards the provider and the consumer, (*iii*) consideration of the service level agreement, (*iv*) consideration of cost and economic aspects, and (*v*) evaluation tools. Finally, we provide a performance analysis through extensive simulations of several replication strategies dedicated for single and interconnected clouds.

KEYWORDS

cloud system, data correlation, data replication, interconnected cloud, simulation

1 | INTRODUCTION

As a rising computing paradigm, cloud systems enable tenants to acquire cloud services elastically and on-demand based on a pay-as-you-go pricing model.[1] The paradigm provides geographically distributed virtualized pool of computing resources that the cloud providers offer to their tenants. This is performed according to the service level agreement (SLA) and the target service level objectives (SLOs). Although cloud resources are marketed as being unlimited, a single cloud provider does not have the financial ability to deploy resources all over the world to accommodate its tenants' geographic dispersion and quality of service (QoS) requirements.[2] Therefore, both providers and tenants resort to the use of interconnected clouds (inter-clouds), which integrate diverse resource options offered by multiple providers that collaborate. Researchers in the literature have proposed many terminologies related to interconnected cloud systems. Inter clouds (or cross cloud) is the general case of cloud interconnection. This term refers to a system in which the systems of several cloud providers inter-work relying on standard interfaces in order to ensure QoS.[3] We note that the existing terminologies related to interconnected clouds are overlapping. However, we present below the most prominent clouds' connectivity scenarios as indicated on Figure 1 based on the literature's understanding:[2,4-6]

- Multi clouds is a system where tenants utilize the services of multiple independent cloud systems. The interconnection, in this case, is managed by the tenants or a third party on behalf of tenants and is usually transparent to cloud providers.[7]
- Federated clouds is a system where a collaboration is made between multiple cloud providers to enable resource sharing while maintaining QoS. Federated clouds are distinguished by their geographic dispersion and by the voluntary nature of their participant providers.

https://doi.org/10.1002/cpe.7758

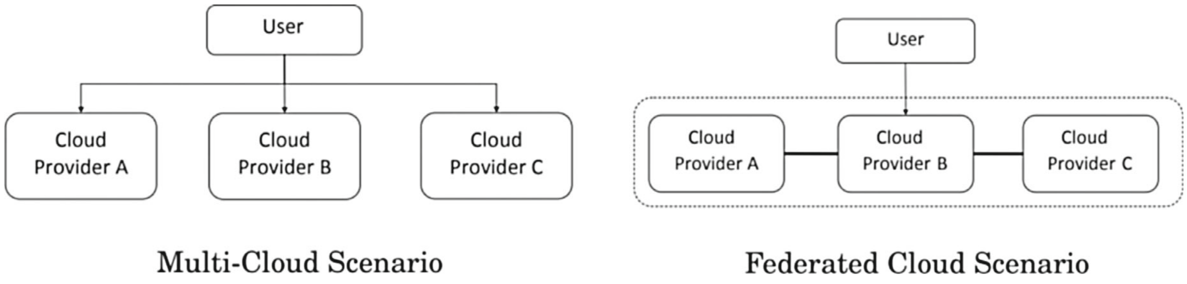


FIGURE 1 Cloud systems interconnection main scenarios based on Reference 5.

The providers operate and manage the system's interconnection according to regulations that govern each cloud provider's autonomy, privacy, and protection.[8]

Because of the many advantages that the cloud offers, the number of applications dedicated for cloud systems is constantly growing. Their benefits include lower hardware costs and administrative complexity, as well as enabling increased scalability, flexibility, and QoS. Examples of such applications include big data applications, Internet of Things based applications, data-intensive applications, and computation-intensive applications.[9-11] The deployment of these applications on the cloud generates huge amounts of data that need to be managed efficiently.

Data replication is a well-known data management technique in large scale and distributed systems such as the cloud. It consists of storing multiple data replicas with the aim of improving data availability, minimizing bandwidth usage, and providing fault-tolerance.[12,13] Data replication has been commonly adopted in traditional distributed systems such as database management systems (DBMS),[14] parallel and distributed systems,[15] mobile systems[16] and other large-scale systems including P2P[17] and data grid systems.[18,19]

Despite its benefits towards improving performance and QoS delivery, data replication results in increased energy consumption as well as additional financial expenditures for the provider.[20] Therefore, it is not economically feasible to create as many replicas as possible in cloud systems. Indeed, this might result in excessive resource usage and lower revenues for the provider.

Several data replication strategies in cloud systems have been proposed for single and interconnected cloud systems to solve the following main issues: When to create or remove replicas? What data to replicate? Where to place replicas? How to adjust the replicas number? What is the cost of replication?

By responding to these questions, the strategies aim to manage efficiently the large amount and the various types of stored data. The explosive growth of the amount of data stored in the cloud offers an abundance of information such as data correlations. These correlations include two main types:

- Semantic correlations are observed at the level of data attributes or features such as data identification, size, name, type, owner, user, application, path, or entire data content and so forth. This type of correlation has been used to improve data retrieval system performance.[21] In order to assess such correlations, meta-data and domain knowledge must be maintained by a local expert in the system workload.[22] Therefore, this type of correlations is unsuitable for data lacking meta-data, such as that found in most distributed storage systems, such as the cloud.[23]
- Access correlations represent the tenants' frequent and common accesses to data through applications, services and so forth. Data access is determined by user behavior and program instructions, which are far from random.[23] In this situation, the relationship between the data is defined by shared accesses, access frequency, or access sequence, while taking into account various temporal and/or spatial constraints.[24-26]

The knowledge about these correlations might be crucial for enhancing system performance and the QoS.[27-30] Indeed, several data replication strategies have benefited from exploiting this knowledge about data correlations.[29-40]

To the best of the author's knowledge, there is no survey concentrating on data replication strategies designed for interconnected cloud systems. Therefore, in this article, we provide an analysis of existing strategies in cloud systems, covering single and interconnected clouds. Furthermore, the survey provides an in-depth review of data correlations aware strategies. Since these strategies are not considered in-depth in surveys and review studies, it is necessary to analyze them alongside with the used correlations extraction methods to enable researchers a better understanding and development of strategies based on data correlations.

The main contributions of this survey are listed as follows:

- Reviewing and presenting the strengths and limitations of several existing surveys and review studies that addressed data replication strategies in the cloud systems.

- Classifying data replication strategies in cloud systems according to the cloud environment for which they are designed, namely single cloud and interconnected clouds.
- Highlighting data correlation-based strategies as well as their key steps, allowing researchers to have a better understanding of the exploitation of data correlations.
- Analyzing the major features of cloud based data replication strategies such as (*i*) their addressed issues, (*ii*) their evaluation tools, and (*iii*) their adaptation to the criteria that are more specific to cloud environments such as: (*a*) the orientation towards the provider and the consumer, (*b*) the consideration of the SLA, (*c*) the consideration of cost and economic aspects when replicating data. We also provide multiple tabular representations to illustrate the strategies' characteristics.
- Providing a performance analysis through extensive simulations of several existing strategies. This analysis is performed based on various performance metrics and involves both single cloud and interconnected cloud strategies.
- Presenting future directions to guide researchers to further advance in the cloud data replication field.

The rest of this article is organized as follows: Section 2 covers the related work, which includes an analysis of various existing cloud based reviews and a discussion of the literature's classifications of data replication strategies. In Section 3, we present our proposed classification that divides the strategies into those designed for single cloud and those for interconnected clouds. Furthermore, we introduce and analyze the key steps of correlation-aware strategies. In Section 4, a discussion is provided to answer the most fundamental questions related to the data replication strategies in cloud systems. Section 5 provides an experimental evaluation of single cloud and interconnected cloud strategies, including various correlation-aware strategies. Challenges and future research directions are discussed in Section 6. Finally, we conclude the article and summarize the main findings of our work in Section 7.

2 | RELATED WORK

Cloud computing has attracted the attention of academics resulting in several review studies that focused on the different management aspects adopted in cloud computing. For example, we can find those who concentrate on resource allocation and provisioning,[41,42] load balancing,[43] task scheduling,[44,45] pricing strategies,[46] data placement and replication strategies[28,47] as well as adaptive management of resources.[48]

In this section, we examine multiple existing review studies that are related to our work and we present the most common classifications of data replication strategies.

2.1 | Existing surveys

The manifesto presented in Reference 49 tackles the most prominent cloud computing challenges, including issues related to elasticity, scalability, sustainability, security, resource and data management, as well as the economics of cloud computing and clouds interconnection. It also presents the rising concepts related to cloud computing, such as fog and edge computing. It emphasizes the potential of relying on interconnected cloud systems in order to enhance both the satisfaction of QoS requirements of the users' applications and cost-saving. Hence, it promotes proposing solutions designed for interconnected clouds. The manifesto places a strong emphasis on the increasing reliance on cloud computing in a variety of fields, which in turn increases the need to provide effective cloud-based solutions. These solutions should take into account several requirements, including the large amount of data for storage, the distributed and heterogeneous resources to be managed in a cost and energy-aware manner while optimizing for multiple criteria, including QoS constraints and economic constraints.

The survey presented in Reference 50 confirms the fact that, in practice, data centers have never been able to provide infinite and unlimited resources. Thus, there is a need to rely on the interconnected resources of different service providers in order to increase the elasticity and capacity of cloud systems and fulfill the ever-increasing demand for cloud users. The survey discusses the primary obstacles facing interconnected clouds, such as the disparities among the application programming interfaces (API) and the abstraction methods, the specialization of management techniques adopted by each provider, and the issues related to interoperability, security, and privacy. The survey presents a comparison of the existing interconnected cloud systems, along with their advantages and limitations. Furthermore, it recommends the adoption of interconnected cloud systems despite their challenges due to the economic and functional advantages they provide for both users and service providers.

The authors of Reference 28 provide a detailed overview of cloud data storage and placement methods. The study covers data life-cycle management and data management in cloud systems. It outlines their non-functional requirements including performance, availability, elasticity, scalability, and consistency. It also lists the main data models employed by the data management systems and presents a comparison of their technical characteristics, such as their architecture and support for data replication. The survey categorized data placement strategies as data dependency (correlations), task and data scheduling, and graph based approaches. The survey recommends deploying applications on interconnected clouds (multiple clouds) as they provide a solution for cloud storage systems to meet their non-functional criteria. In addition, it stresses on the fact that data placement and replication strategies should take into account both the users' requirements (which may be included in the SLA) and the optimization of data transfer volume, placement cost, and management cost. In this regard, the survey advocated that utilizing data knowledge could result in effective data partitioning and management. This involves co-locating tasks and data with frequent access and leveraging data correlations.

Many research efforts have been made in the review of data replication in cloud systems. We then present a number of these reviews since our work is related and complementary to these efforts.

In Reference 51, a comprehensive study is presented that classifies data replication strategies in cloud systems into static and dynamic ones. This work classifies the strategies based on the nature of the strategies and assesses them from a global perspective. Furthermore, the review highlights several future directions of cloud based data replication. However, the review does neither include interconnected cloud based strategies nor an experimental evaluation of the reviewed strategies.

Tabet et al.[52] proposed several classification criteria focusing mainly on listing the most addressed objectives functions of data replication strategies for cloud systems. Furthermore, it affirms that strategies should simultaneously consider multiple objectives such as enhancing availability, optimizing performance, balancing load, and achieving cost-saving. Although discussing several classification criteria, the review does not analyze specific criteria to the cloud systems, such as the satisfaction of SLA requirements and the consideration of the cloud's economic aspect.

The review presented in Reference 53 provides a comparison of multiple data replication strategies, including the ones aiming to minimize total energy consumption and the ones aiming to maximize profit. The review points out that these objectives are rarely considered simultaneously. Furthermore, the review presents an experimental evaluation of some strategies while considering criteria such as the total execution time and energy consumption. This work however neglects other important criteria such as the penalty cost paid by the provider when the SLA is not satisfied.

Slimani et al.[54] presented an in-depth analysis of data replication as well as service replication strategies while focusing of the QoS aspects. The review presents the most addressed QoS metrics. However, it does not include sufficient number of data replications strategies.

The data replication strategy classification given in Reference 55 takes into account other criteria, specific to cloud environments, such as the number of tenant's objectives (single vs. several SLOs), the nature of the cloud (single provider vs. multi-providers), and the consideration of the economic (monetary) cost of resources for the provider. The simulation study compare the performance of five replication strategies proposed only for single cloud systems with a focus on the consideration of the economic cost of replication.

Mansouri and Javidi[56] reviewed data replication strategies that are based on meta-heuristic algorithms in both cloud and grid systems. The review highlights the efficiency of meta-heuristic based replication strategies in achieving multiple performance objectives. It classifies the strategies according to the used meta-heuristic algorithm and presents multiple tabular representations to illustrate the strategies' characteristics. However, the review does not include an experimental evaluation of the reviewed strategies.

Shakaram et al.[47] presented an in-depth assessment and classification of state of the art data replication strategies in cloud systems. The review classifies the data replication domain into three fields: data deduplication, data auditing, and replica management. Furthermore, the replication strategies are analyzed and compared considering their characteristics. The review outlines also the strengths and limitations of the reviewed strategies using various tabular representations. However, there is no experimental investigation contained.

Whereas, in Reference 57, the authors presented a different type of survey that focuses on a quantitative analysis of data replication strategies published by the main scientific editorials. The presented systematic review analyzes the strategies according to the publication's editorial and year. However, the performance analysis was covered weakly, and the experiment evaluation was not presented.

Table 1 summarizes the advantages and limitations of existing review studies. We note that almost each paper lacks of an experimental evaluation of the reviewed strategies. Moreover, only some review work focus on analyzing the impact of correlations on data replication strategies. In the limitations column, the symbol '-' sketches that the indicated aspect was not sufficiently addressed in the associated review study while the symbol '•' indicates that the aspect was not addressed at all.

2.2 | Existing classifications

The review studies resulted in multiple classifications. However, we present hereafter the most common classifications, which are based on several criteria, such as the nature of the strategy, the replication control mechanism, the replication periodicity, the replication strategy' aims. This allows strategies to be classified as (*i*) static versus dynamic, (*ii*) periodic versus non-periodic, (*iii*) centralized versus decentralized, (*iv*) according to their objective functions.

2.2.1 | Static versus dynamic

This classification is based on how the replication decision is performed.[51] Static strategies set the replication parameters, such as the number of replicas and their locations, before the system is operational (during the design phase).[58,59] While dynamic replication strategies make their decisions in response to the changing state of the system. This includes: (*i*) determining the number of replicas according to data availability and the provided

TABLE 1 Advantages and limitations of the existing studies and the proposed review work.

References	Advantages	Limitations
49	+ Addressing various issues in cloud management including cloud interconnection scenarios + Presenting the emerging concepts related to cloud computing + Providing future directions	- Data management • Experimental evaluation
50	+ Addressing the issues facing interconnected clouds systems + Reviewing advantages and limitations of existing interconnected cloud systems	- Data management • Experimental evaluation
28	+ Addressing the issues facing cloud storage management systems + Presenting data management life-cycle + Presenting a categorization of data placement strategies in the cloud + Providing future directions	- Cloud interconnection • Experimental evaluation
51	+ Presenting a categorization of data replication strategies in the cloud + Providing future directions	• Interconnected cloud strategies • Numerous categorization criteria such as economic aspects • Data correlations • Experimental evaluation
52	+ Presenting a categorization of data replication strategies in the cloud considering several criteria + Consideration of economic aspects of the cloud	• Interconnected cloud strategies • Data correlations • Experimental evaluation
53 and 55	+ Presenting a categorization of data replication strategies in the cloud considering energy consumption and economic profit + Providing experimental evaluation	• Interconnected cloud strategies • Data correlations
54	+ Presenting a categorization of replication strategies including both service and data based strategies + Presenting the addressed QoS metrics	- Data replication strategies • Interconnected cloud strategies • Data correlationsh• Experimental evaluation
56	+ Reviewing meta-heuristic data replication strategies + Presenting the most used meta-heuristic algorithms	• Interconnected cloud strategies • Data correlations • Economic aspects • Experimental evaluation
47	+ Presenting a categorization of the replication field in the cloud + Presenting global and detailed taxonomy of data replication technique in the cloud	• Interconnected cloud strategies • Data correlations • Experimental evaluation
57	+ Presenting a quantitative analysis of data replication strategies published by the main scientific editorials	• Performance analysis • Experimental evaluation
Proposed review	+ Presenting a single vs. interconnected classification of data replication strategies + Considering the impact of correlations on data management in cloud systems + Analyzing the major features of cloud based data replication strategies + Providing a performance analysis through experiments for several strategies	- Enriching the experimental evaluation by considering more strategies and more parameters - Extending the study to emerging computing systems like edge and fog computing

QoS, (*ii*) identifying the data to replicate based on popularity and access patterns, and (*iii*) selecting the replicas placement based on several criteria such as network bandwidth, load balance, and replication cost.[30,36,39,60-64]

2.2.2 | Periodic versus non-periodic

Both non-periodic and periodic strategies are dynamic since they are invoked during tasks execution. Their behavior depends then on the changes in the user access pattern, storage capacity, bandwidth and so forth. On the contrary, static strategies have parameters (number of replicas, hosting nodes, etc.) which are pre-determined before tasks are executed.

This classification—periodic versus non-periodic—is based on the periodicity feature of the replication process. This feature specifies when the replication strategy is triggered. Non-periodic strategies perform replication each time a data is requested.[20,58,64,65] This might result in excessive replicas creation, storage usage, and cost issues. While periodic replication strategies are triggered at each given period, as those presented in References 36,40,66-68. These latter strategies perform a periodic replication of frequently accessed data. In this regard, the replication period can be assessed as a number of executed tasks or as equal to a given interval of time.

2.2.3 | Centralized versus decentralized

This classification is based on the entity that takes the replication process decisions.[69] When a replication strategy is centralized, a central authority controls all aspects of replication.[39,58,65,70-72] Decentralized replication strategies encourage no central control. These strategies are more effective in responding to system changes and dealing with failures and outages situations.[30,64,67,73-75]

2.2.4 | Objective function basis

This classification covers the objective function of each strategy when dealing with the replication issues.[20,52,55] This classification deals with the strategies' aims, regardless of their commitment to achieving them. These objectives can be directed towards increasing system availability and fault tolerance.[70,75-77] It can also be directed towards improving performance, which may include (*i*) optimizing resource usage (storage, computing, and bandwidth),[20,61] (*ii*) balancing workload,[27,60] (*iii*) reducing tasks' response time, enhancing data access and reducing latency,[30,73,78] (*iv*) data security,[79,80] and (*v*) data consistency.[66,81] Additionally, it can be directed towards reducing the cost of replication.[20,30,73,77]

As far as we are aware, no research study has focused on the cloud systems for which these strategies are designed. In addition, there is no survey that focused on the importance of leveraging data correlations during the replication process, despite the fact that several strategies rely on them. In this article, we analyze existing data replication strategies in cloud systems, covering both single and interconnected clouds. In addition, we provide an in-depth review of data correlations-aware strategies. We also provide a comprehensive analysis of the reviewed strategies as well as an experimental evaluation of some of them.

3 | PROPOSED ANALYSIS

We classify data replication strategies according to the cloud system they are designed for including single cloud and interconnected clouds. Furthermore, we emphasize the correlation-aware data replication strategies and outline their key steps.

3.1 | Article selection methodology

Following the research methodology proposed in Reference 82, we first defined the strategy to collect and filter the most relevant studies according to a set of inclusion-exclusion criteria. The reviewed work were looked for in the main academic databases, including Wiley Interscience, Springer, Elsevier, IEEExplore, ACM Digital Library, Inderscience Publishers, and Google Scholar. These sources were chosen because they cover almost all important conferences and journals related to the focused data replication strategies in single versus interconnected cloud systems, as well as data correlations.

The selection process covers two main steps. In the first one, we used a combination of the following keywords "Single cloud," "Interconnected cloud," "Correlation," "Data correlation," "Replication strategy," "Machine learning," "Mining." We then moved on to the second step, which consists in filtering the huge number of works obtained after the first step. Since the titles and abstracts of the found documents do not truly reflect the work content, the found result database was filtered based on additional parts of each retrieved document, mainly the introduction and conclusion sections. The contents were studied carefully and the work addressing contexts other than cloud systems or related ones are excluded. In addition, work that present similar content or have not the form of research papers (e.g., tutorials, presentations, etc.) and those written in languages other than English were also eliminated. On the other side, work published in high-level journals and conferences were privileged. Moreover, survey papers dealing with emerging systems (such as those based on fog and edge computing) were retained.

3.2 | Single cloud versus interconnected clouds strategies

The cloud provider allocates and makes available many of its resources to host its tenants' applications and perform their tasks. These resources are usually owned and managed by a cloud provider in what is called a single cloud environment. While interconnected clouds refer to the connection of two or more cloud systems or cloud providers' resources. Thus, it is possible to benefit from the huge amounts of resources that are highly distributed and available between them.

3.2.1 | Single cloud-based strategies

In a single-cloud environment, *aka* mono-provider or intra-cloud system, the strategies consider either one data center or multiple data centers (DCs) owned by a single cloud provider. Typically, each cloud DC is organized as a group of physical machines virtualized into multiple virtual machines (VMs). Some of the disadvantages of relying on this environment are its limited geographical dispersion and its limited resources. In addition, the problem of data vendor lock-in affects the provided QoS.

In the following, we list some strategies with regard to some criteria such as the consideration of a single objective. In particular, we focus on performance in terms of response time versus several SLOs, the consideration of monetary costs when modeling the profit of replication, the use of different techniques when replicating data or the fact that replication is user or supplier oriented.

Data replication strategies can be classified as single-objective versus multi-objective strategies. Most of the existing strategies aim to satisfy a single tenant objective such as availability,[60] energy consumption,[63] load balance,[69] and performance.[83] CDRM[60] is a cost-effective dynamic replication strategy that aims to improve system availability and load balancing while maintaining a minimum number of replicas. Replicas are created to meet availability requirements and placed based on the blocking probability of each resource. Replicas are placed in less heavily loaded resources, and overloaded resources are blocked from receiving new tenant requests. In Reference 63, the authors propose two replication algorithms, namely, energy-efficient and time efficient heuristic replication. The first algorithm focuses on the reduction of energy consumption when placing replicas, while the second algorithm places replicas as close to the users as possible to reduce access latency. The algorithms rely on a cost model that considers the gain in energy cost from replication compared to the no replication state. In Reference 69, a replication strategy is proposed with the aim of achieving load balance and high data reliability. It relies on a model that estimates data reliability according to the number of replicas and storage duration. Replicas are created according to the storage resources' load and file popularity. Then, they are placed in a low-load resource with high reliability and a high number of network links connecting it with its neighboring resources. The RTRM strategy[83] defines a threshold as the upper limit of a data request's response time. It increases the number of replicas by creating a new replica whenever the response time threshold is exceeded. Considering data requests, a prediction of response time is made, relying on CPU and bandwidth capacities to schedule the request to the suitable existing replica. Replica placement is performed in a way that minimizes the number of replicas of each data and respects the response time threshold for each data request using a graph theory-based method.

In the same context, in Reference 78, a database management framework is presented. It aims to satisfy the tenants' required SLOs in terms of execution time and response time of database transactions. The strategy creates a new replica when it detects a certain number of continuous SLA violations. The closest replica of a running database is activated if the latest replication delays are violating the SLA, which enables avoiding the SLA violations. Gill and Singh in Reference 77 proposed a cost-effective data replication strategy. Replication is triggered when data popularity crosses a dynamic threshold. The strategy used a mathematical model to determine the number of replicas to meet the availability requirement described in the SLA. Then, the knapsack method is used for their placements by replicating data from higher-cost DCs to lower-cost DCs. The strategy also considers the user's budget, ensuring that the cost of replication associated with each DC does not exceed the user's budget. The strategy proposed in Reference 84 aims to meet the data access time QoS parameter. To minimize SLA violations and replication storage costs, two algorithms are proposed. The first algorithm prioritizes high QoS requests over low QoS requests. Hence, high QoS replicas are stored on high-performance storage resources. The second algorithm relies on an integer linear programming model to determine the optimal replicas placements.

On the other hand, some strategies aim to meet simultaneously several tenant objectives. The strategy proposed in Reference 85 is triggered as a response to SLA violations in terms of response time and availability. The strategy classifies data by tenant usage (storage duration and frequency of access) into five categories: occasional, instant, viral, semi-annual, and unclassified. Viral data are replicated since they are frequently accessed by the tenants. Additional replicas are created following a periodic check for response time SLA violations. Replicas that are not unclassified or viral are removed to save the storage space for the benefit of the provider. An efficient and improved multi-objective optimized replication management (EIMORM) is proposed in Reference 86 to reduce the provider's replication management cost while addressing data availability and load balancing. The cost is evaluated by assigning a given cost value to each DC so that high-performance, high-availability DCs have high costs. The strategy balances its objectives by placing replicas in high-cost and low-cost DCs using an improved knapsack technique.

There are a number of strategies that aim to satisfy the tenant's objectives while reducing the cost of replication, for example, data storage and/or data transfer costs between DCs. In Reference 87, RepliC is proposed. The strategy considers reducing operating costs by adding and removing replicas in an elastic manner according to workload changes. RepliC indeed adds replicas to prevent SLA violations and deletes unnecessary replicas to avoid resources wastage. This work was extended in Reference 88 where a predictive approach called PredRep is proposed to characterize the cloud system workload. This approach aims at automatically providing or reducing resources by replication techniques in order to face irregular workload patterns, which impacts QoS. On its side, the framework presented in Reference 89 decides on database replication and migration. A prediction model estimates the tenants' transaction response times and detects violating data replicas. Then, the violated data are migrated or replicated to the resources with the fastest transaction execution and response times. To decrease the required time to execute the tenants' transactions, resources are grouped based on the cost of communication between them.

Some other strategies[77,84] are mentioned cost-aware although the considered cost of replication is not necessarily an economic cost. It is regarded as an assigned budget value for DCs in Reference 77 and modeled in terms of time in Reference 84. In this context, only some strategies model the replication cost and the provider profit as monetary costs while satisfying a tenant response time SLO. In this respect, Tos et al. propose PEPR a data replication strategy that aims to maintain the provider profit.[64] The replication process is triggered as a response to SLA violations only when the cloud provider's profitability is guaranteed. Replica removal is performed when the SLA is satisfied over time. While an economic model is used to estimates the monetary profit expected to be gained by each query. RSPC is proposed in Reference 20. By setting two response time thresholds, SLA violations trigger the replication. The first threshold is crucial. An SLA violation is expected if a predicted response time for a query exceeds this threshold. Hence, data related to this violating query are considered for replication. When a lower threshold is repeatedly violated, a set of queries-based replication is considered. The replicas are not created until a suitable placement is heuristically identified to meet the response time requirement and maintain the provider's economic profit. The strategy uses an economic model that includes the provider's revenues and expenditures.

In addition, most of the cited strategies aim to increase the profit of the provider. Thus, some strategies are considered as tenant-oriented strategies.[90,91] Sakr and Liu present in Reference 90 a cost management framework for cloud database provisioning and replication. The framework aims to achieve SLA compliance by meeting the tenants' performance requirements as defined in the SLA. Replicas are added to accommodate the incoming workload, and they are deleted when the workload decreases. Cost optimization is performed considering the tenants' perspective by avoiding the cost of SLA violations and controlling the monetary cost of the allocated resources, including transaction execution and replication costs. In Reference 91, Limam et al. propose a replication strategy that is triggered only when the SLA is violated. The strategy relies on an estimation of the replication cost, so the cost of replication does not exceed the initial budget intended for replication. To estimate the cost of replication, the strategy relies on an economic model considering the costs of storage, network, investment, and SLA violations' penalties.

Finally, some data replication strategies are based on techniques that achieve specific tenant objectives. MORM, a static replication strategy presented in Reference 59, employs an artificial immune algorithm to select data for replication and define their appropriate placements. The used algorithm randomizes the replication layout and relies on an objective function to find the placements that offer a balanced trade-off between average data unavailability, average service time, variation of loads, power consumption, and average latency. The strategy enables setting a preference coefficient for these objectives according to usage needs. In Reference 92, a framework that determines the replication configuration for applications deployed on geo-distributed cloud. The framework uses integer linear programming to solve the replicas placement problem while considering both normal and failure situations. It focuses on latency as an SLA requirement while maintaining data availability and consistency by employing a quorum protocol. Latency thresholds are defined for read and write operations to enable the selection of replicas placements that result in a lower data transfer and response time.

The strategies presented in References 93 and 94 rely on prediction models to enhance cloud system performance. On the one hand, authors in Reference 93 based on Bayesian learning and Gaussian process in order to anticipate the access potential of the data file. On the other hand, authors in Reference 94 proposed a location prediction method based on historical access records that provides a mechanism for identifying potential high user density locations. Then, they preemptively replicate data to meet the increased demand. The mechanism replicates the highly requested data on the predicted locations before the users' arrival. The Dempster–Shafer theory is adopted for predicting the user's future locations. While replica selection is performed using an artificial neural network to minimize data access time and meet the QoS described in the SLA in terms of response time.

The strategy proposed in Reference 95 relies on a nonlinear integer-programming (NLIP) model to achieve an optimal trade-off between increasing data availability according to the SLA and reducing costs caused by replication. This model achieves the expected data availability by considering machine failure probability and data request probability. The strategy also considers the cost caused by replication. The proposed NLIP model determines the optimal number of replicas for each data, so that the request failure probability and storage cost are minimized. On their side, authors in Reference 96 aim to predict the file popularity using historical file access values. Data replication schedule is modeled as an integer linear programming (ILP) problem in order to minimize the total data file access costs. The ILP problem modeling considers task dependency, data scheduling, data sharing, and reliability.

3.2.2 | Interconnected cloud-based strategies

Interconnected clouds are formed through the interconnection of more than one cloud system that can be deployed by more than one provider. This interconnection between the resources may be initiated by the provider, the tenants, or by a third party who compensates them. Replication strategies exploit the geographical dispersion and pricing differences among providers to improve performance and reduce costs. This environment is characterized by a greater number of resources and a wider geographical distribution. This expands the search space for suitable replicas placements and further complicates replicas placement challenges. In the following, we list the main strategies that have been proposed for interconnected clouds.

RACS[76] and DepSky[79] are among the first strategies to manage user data through interconnected clouds storage. These strategies aim to enable fault tolerance and minimize the monetary cost paid by users. Their data storage mechanisms consist of distributing the users' data among several cloud providers to avoid cases of provider lock-in and cloud outages. While cost reduction is obtained by exploiting differences in pricing policies between cloud providers. However, the cost models used by both RACS and DepSky are limited to the cost of storage only. The authors of Reference 97 propose dynamic programming algorithms that place replicas with the aim of maximizing availability according to the SLA while respecting each tenant's budget. The algorithms focus on selecting cloud service providers in order to maximize both the data survival probability and the number of surviving data under a given budget. Indeed, their cost optimization model focuses on maximizing the benefits of replica placements among multiple cloud providers' systems within a given budget. The benefits are presented by the potential of successful accesses to the data via the resource that stores them. The determination of replica placements is performed by considering the storage price of each cloud provider and the probability of failure of its resources, such that the storage price of the selected replica placements does not exceed the defined tenant's budget.

On its side, in Reference 98, an SLA-based data distribution strategy is proposed for multiple clouds. The strategy aims at distributing data and their replicas using a multi-objective QoS evaluation model that includes multiple SLA parameters. The model considers data privacy, availability, throughput, transfer time and the storage cost of each cloud provider as SLA parameters to define the suitable placement of data. In Reference 73, a lightweight heuristic solution is presented to reduces the monetary cost for the cloud providers, including data replication cost. The solution determines data locations and redirects the users' requests so access latency can be guaranteed at a minimum cost. It considers the price variation of storage resources, workload changes, besides data status, which can be hot-spot and cold-spot according to their access rate pattern. The replication monetary cost includes replica creation, storage, put and get operations and potential migration costs. Preventive Disaster Recovery Plan with Minimum Replica Plan (PDRPMR) is presented in Reference 75. The plan aims to balance the trade-off between data reliability and the cost of storing replicas. It deploys a small number of replicas while taking into consideration both short-term and long-term storage duration to minimize storage usage. Their cost model considers the storage and data transfer costs of each cloud provider where the replicas will be placed. The strategy considers the storage duration and the importance of data as inputs. Indeed, a single replica is determined for short-term and non-critical data while two replicas are determined for long-term duration and critical data. These replicas are periodically checked to ensure reliability.

Other strategies were proposed for interconnected cloud systems while exploiting heterogenous prices offered by cloud providers. Some of them aim to satisfy the SLOs and optimize the paid cost by cloud users. Other strategies try to reduce the cost of replication for the provider executing the user query by replicating at providers offering minimum network bandwidth prices.

SPANStore (Storage Provider Aggregating Networked Store)[66] is a unified storage service that is built as an interconnected clouds environment for application providers. It aims to meet SLOs in terms of fault tolerance and latency while minimizing the cost of application execution. SPANStore takes advantage of the geographical dispersion of the DCs of several storage providers to manage workload changes. Then, it replicates data closer to users, reducing both latency and SLO violations. It also benefits from the cost diversity of put and get operations between regions to reduce replication monetary costs, including storage, network and update costs. In Reference 99, TripS an interconnected cloud storage system, is proposed to avoid SLA violations in terms of availability and performance. TripS exploits multiple providers' resources to determine the suitable placements of data and their replicas on behalf of the cloud tenants. It requires several inputs, including SLA requirements, consistency model, fault tolerance, latency, cost of storage and bandwidth information of the corresponding storage resources. TripS then returns a list of feasible candidate data placements that all meet the applications' SLA requirements at low cost. TripS takes preventative actions by checking replica availability according to the SLA requirement and monitoring service denial to ensure data reliability at a lower cost. Indeed, a minimum number of replicas is deployed according to the fault tolerance input.

A storage framework called DAR is presented in Reference 100. It exploits the variation of get and put operations latency and resource prices offered by cloud providers of the system in order to satisfy the SLOs and optimize the paid cost by cloud users. DAR uses the integer-programming method to model the problems of data placement and resource allocation. Hence, two heuristic solutions were presented, including a dominant-cost based data allocation algorithm and an optimal resource reservation algorithm, while the used cost model includes storage, transfer, get, and put costs during resource reservation time. In Reference 101, an adaptive replication and placement strategy is presented. It relies on workload prediction to place replicas with cost optimization. Future user demands are forecasted using the ARIMA time-series technique. Replicas are then placed to meet the corresponding availability requirements. The strategy considers the cost of the network when accessing the data beside storage cost. Indeed, it relies on a cost model that considers the differences in pricing policies between the geographic regions of the system.

Moreover, an adaptive data placement framework is proposed in Reference 102 called ADPA to minimize cost. It consists of deciding on the appropriate data placement based on the expected frequency of access. The framework predicts the frequency of data access based on historical workload using LSTM (long short-term memory). It then uses a learning-based data placement algorithm where data can be migrated from one cloud provider to another as workload changes. The cost model used covers the costs of storage, network, put and get operations and migration. In Reference 72 a data placement strategy is proposed to balance the trade-off between data access performance and the placement cost while considering latency between different zones in a federated clouds environment. To adapt to the changing access patterns, the strategy considers the storage, latency, and migration costs as the dimensions of the fitness function of the multi-objective placement algorithm. Their cost model considers the monetary cost of data placement, including local placement costs, outsourcing costs, and SLA penalty costs. The cost of local placement is the cost of storing the user's data on the provider's own infrastructure, whereas the cost of outsourcing is the cost of storing the tenants' data with partner cloud providers.

3.3 | Correlation-aware data replication strategies

The cloud providers in both single cloud and interconnected cloud environments need to maintain their monetary profit and satisfy their tenants' SLOs. This is performed while managing a massive amount of data and applications' tasks, which is very challenging. In this context, taking advantage of the valuable knowledge that can be retrieved from system information such as data correlations to improve the replication process is quite effective.

Data correlations can be extracted from access history and meta-data.[25,29,38,103,104] Access correlations can be expressed as data requested simultaneously or frequently by tasks, applications, and users. While mining semantic correlations mainly rely on similarity estimation algorithms. These algorithms are used to quantify the similarity between semantic attributes that can be extracted from a set of meta-data attributes characterizing data itself (date of creation, date of modification, last access time, etc.).

In fact, the applications' tasks require accessing data for their execution. Some of these required data tend to be frequently requested by the tasks that are executed on the same system resource. However, these required data may not be available locally. In this case, they must be retrieved from one or more remote, geographically distributed resources. The retrieval of such remote data may take a long time due to bandwidth constraints and workload fluctuations. This negatively impacts the response time of the tasks and leads to an increase in the SLA violation amount. However, if these required data were placed in the execution resource or in its adjacent resources, the number of remote data accesses as well as the time required for data access would be significantly reduced. This helps the providers meet the performance requirements and save more costs, including those for data transfer, SLA violations and energy consumption.[105,106] Looking for data correlations can greatly help for this purpose. However, it is important to note that such strategies based on data correlations face the overhead of correlation mining, especially when a very large amount of data is available.

In this section, we begin by presenting the main steps of a data correlation-aware strategy. We next examine the data correlation aware replication and placement strategies in both single and interconnected cloud systems. These strategies take advantage of the extracted knowledge about data correlations, which can be either access correlations or semantic correlations.

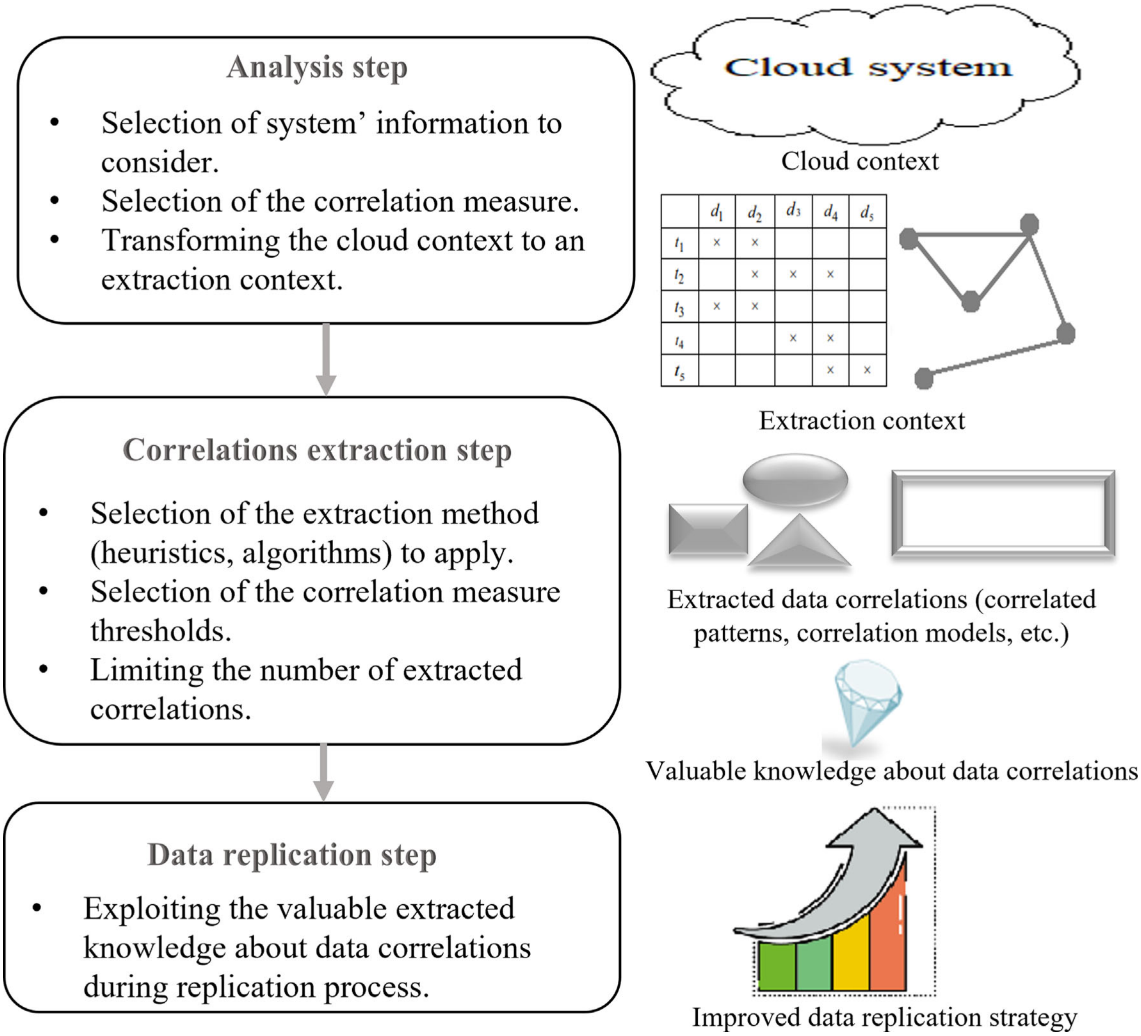
3.3.1 | Key steps of correlation-aware replication strategies

A replication strategy based on data correlation consists of three important steps: analysis, correlations extraction, and finally replication, which is based on the extracted knowledge about correlations. Important decisions are made at each of these steps:

- During the step of analysis, the strategies decide which data to analyze and how to transform the information from the cloud system so that it is suitable for the process of correlation extraction. In this sense, the majority of replication strategies take frequently accessed data into consideration.[29,31-35,37,38] The cloud context (access history, meta-data, etc.) is then transformed into an extraction context mainly represented by matrices or graphs.

  The measure of data correlations to be considered for this transformation is determined. In this regard, most data correlation-aware strategies in cloud systems employ access correlations for correlation estimation, with the degree of correlation defined by simultaneous or shared access to data.[29,35,37,38]

  It is important to note that the size of the extraction context affects the time required to extract correlations. Consequently, the majority of data correlation-aware strategies prefers a periodic analysis of the system information.
- During the correlation extraction step, replication strategies select the method for extracting data correlations. In this regard, the majority of strategies utilize accurate and dependable extraction algorithms such as data mining and machine learning algorithms.[29,31,32,34,37,38] Furthermore, they define correlation measure thresholds to avoid grouping data that are not sufficiently correlated. The strategies also employ thresholds and rankings to limit the number of extracted correlated data groups. In this respect, these latter groups are often sorted according to their popularity and size.[29,35,36,38] Hence, this allows limiting processing to only the most valuable knowledge about data correlations.
- During the replication step, the strategies decide how to exploit the valuable extracted knowledge about data correlations. In this respect, some strategies use it to proactively perform data replication. Indeed, they predict future data requests and replicate them[36,107] through assessing the value of important aspects of data (such as popularity and number of data requests).[34,108] While other strategies consider the correlated data as granularity for replication.[35,37,38,109]


Analysis step
• Selection of system' information to consider.
• Selection of the correlation measure.
• Transforming the cloud context to an extraction context.
Correlations extraction step
• Selection of the extraction method (heuristics, algorithms) to apply.
• Selection of the correlation measure thresholds.
• Limiting the number of extracted correlations.
Data replication step
• Exploiting the valuable extracted knowledge about data correlations during replication process.
Cloud system
Cloud context
Extraction context
Extracted data correlations (correlated patterns, correlation models, etc.)
Valuable knowledge about data correlations
Improved data replication strategy

	d_1	d_2	d_3	d_4	d_5
t_1	×	×			
t_2		×	×	×	
t_3	×	×			
t_4			×	×	
t_5				×	×

FIGURE 2 Key steps of a correlation aware data replication strategy in the cloud system.

In Figure 2, the key steps of the data correlation-aware replication strategy are depicted, along with the main choices to be made during the process.

3.3.2 | Review of correlation-aware data replication strategies

We review correlation aware data replication and placement strategies in single and interconnected cloud systems. In this respect, reviewed strategies are classified according to the method adopted to extract data correlations.

Mining based correlations extraction

Mining frequent patterns consists of identifying the frequently occurred data patterns (item collection, sequence, structure, etc.). Cloud storage systems are analyzed with frequent pattern mining algorithms to uncover data patterns.

In Reference 110, the authors propose a data replication strategy that reduces data access time by analyzing system access history. To select data groups to replicate, a pattern-mining algorithm is used. While the authors of Reference 31 proposes an approach for mining frequent block access pattern. The number of replicas for each data group is configured based on the access patterns. Then, replicas are placed in the DCs that requests them the most.

The proposed strategies in References 107 and 34 rely on the Frequent Pattern Growth (FP-Growth) algorithm to extract data correlations. In Reference 34 data security and latency are ensured while keeping the number of replicas low. The strategy uses a matching function to find the association rule that matches the current user access and has the highest confidence degree. The support degree of this selected rule is regarded as the future access popularity of data. Then, data groups are placed where users request them most.

The authors of Reference 29 proposed a data mining-based data replication strategy, called DMDR, extending the work originally used for data grids.[19] The strategy uses a Maximal Frequent Correlated Pattern (MFCP) algorithm to reduce access latency. Each extracted pattern includes a maximum group of highly correlated data, allowing the determination of a reduced number of correlated patterns. DMDR examines the access history in each DC periodically. It sorts the extracted frequent correlated file groups by size. Then, it places them based on centrality and replica access. This strategy also includes a data replacement procedure, which evaluates each file's importance based on its last access, access count, and size. The obtained assessment value determines whether new replicas replace deleted ones or replication is abandoned.

Clustering is a major field in machine learning and knowledge discovery. It consists of detecting non-trivial or hidden patterns in data. In Reference 32, the authors use the formal concept analysis (FCA) technique[111] to reduce the data movement across DCs and the average query span. The data allocation and replication strategy starts with formal concept extraction. Each formal concept represents a maximal set of data and tasks that require them simultaneously. The FCA technique produces many extracted patterns, thus they are simplified. Therefore, the strategy weights concepts based on data and tasks. It then ranks them in descending order to identify the most interesting. The selected concepts are assigned to DCs having the minimum difference in storage capacity and data group size. Each concept's most-requested data are replicated and placed on the most frequently accessed DCs. In Reference 37, the authors also employ the FCA in their cost and energy efficient data placement and replication strategy that assigns data and tasks based on their correlations.

Khelifa et al. propose two replication strategies that take advantage of data correlations with the objective of keeping the provider profitable while meeting the SLA requirements.[30,39] Triadic Concept Analysis (TCA)[112,113] is used to analyze tasks that may cause SLA violations. This allows strategies to determine which correlated data groups to replicate and candidate placements to reduce future SLA violations. The TCA reveals the ternary relationship between SLA-violating tasks, remote data, and resources like DCs and VMs. This is made possible by mining patterns known as frequent triadic concepts. Furthermore, economic models are used to estimate provider profit and determine if placing a new correlated replicas group is profitable. The strategies also rely on replacement procedures to replace old replicas with more profitable ones when storage space is insufficient.

The proposed strategies in References 38 and 40 use spectral clustering to extract data correlations. The strategy of Reference 38, called PCDR, aims to reduce task response time and cloud bandwidth consumption. It periodically builds a data center's correlation matrix, which is the sum of its executed tasks' correlation matrices. The spectral clustering algorithm is then applied to extract correlated file groups. PCDR considers only the most popular correlated file groups. It assesses file popularity based on access frequency and half-time. Furthermore, when storage capacity is insufficient, less popular replicas are deleted and replaced with newer replicas. While in Reference 40, the introduced strategy for federated clouds aims to satisfy SLA and maintain the providers' profit. It periodically identifies the groups of correlated data related to SLA violations. For replica placement, it uses a fuzzy inference system (FIS) that considers data transfer time ratio, VM load, data availability, and cloud provider profit. Economic cost models are used to estimate the latter. The FIS assesses the potential of placing the correlated replicas groups on provider-owned or rented VMs from its participant providers. Furthermore, a replica number adjustment is applied to maintain a minimum number of replicas.

Graph analysis based correlations extraction

Graph analysis models complex systems, processes, and concepts such as clouds and data correlations. SWORD is a data placement and replication strategy[27] aiming to reduce resources use while maintaining SLA compliance. The access history is represented as a weighted hyper-graph over the data, with each hyper-edge indicating the needed data group to be accessed jointly by the user's queries. The generated hyper-graph is then partitioned into data groups. The strategy evaluates data replication for created data groups. Only the most highly and jointly accessed ones are replicated. Then, replicas are placed to reduce the average execution time of queries and the amount of data movement by relying on the smallest possible number of resources to run tenant queries.

In Reference 114, a data placement and replication strategy for social services in multiple clouds is presented. It aims to maintain QoS, reduce resource use, and lower the carbon footprint. It assigns weights to each goal to allow service providers to make trade-offs based on their needs. The strategy reduces latency and lowers operating costs by distributing user data over various clouds while considering the users' social relationships. Graph-cut is utilized to place data and replicas. By considering shared and common data access, data placement can meet latency constraints and reduce bandwidth use more efficiently than random replication.

Prefetching-aware data replication (PDR) is presented in Reference 36, inspired by the PGFR strategy designed for data grids.[26] This strategy prefetches popular files based on data access correlations. Indeed, PDR analyzes cloud DCs periodically to detect correlated file groups. Each task's dependency graph is built first. Vertices indicate files' access frequencies. Each edge joining two files represents the frequency of their sequential access in a single task's access sequence, regardless of order. Combining the task graphs determines the data center's dependency graph. To identify the popular file groups to replicate, the strategy removes low-correlation graph edges. The resulting sub-graphs represent dependent file groups. The PDR replicates the most popular files from each group. In the event of insufficient storage capacity, existing replicas are deleted and replaced by new ones.

In Reference 108, data placement and replication strategies for social networks across geo-distributed cloud systems are presented. Correlations between cloud users are used to reduce costs for cloud providers and meet clients' latency requirements. Greedy algorithms were used to meet user latency requirements while controlling data management costs. For latency, the strategy specifies how many replicas each user needs.

Replica placements are then determined based on shared data access between social network users. Replicas are placed near users who have a relationship with the data owner.

Other correlations extraction methods

Data correlations can also be discovered using mathematical models and analysis of simultaneous or shared data accesses. The authors of Reference 109 propose a dynamic computation correlation data placement strategy, called DCCP. Access pattern-based data distribution is applied to minimize computation time and balances load. DCCP uses an optimization-based model that captures the link between replica location and load balancing. It stores correlated data in the same data center (DC) where the correlation degree is determined by the number and frequency of data-accessing tasks. When storage space is limited, highly correlated data is kept together or spread across a few number of DCs. This implies that the required data for the user's tasks will often be accessible locally at the DC executing the users' tasks.

In Reference 33, two types of data access correlations are examined. The strategy relies on a mathematical model to mine data correlations based on past access histories. The model extracts stable and bursty correlations based on the changing data access sequences. Stable long-term correlations are considered static replicas and stored in load-balanced storage. Dynamic replicas exhibit bursty short-term correlations. The most requested data to be accessed in the near future are prefetched for replication. As their short-term bursty correlations change, these replicas are constantly replaced in a high-speed cache.

In Reference 35, a data replication strategy is proposed reduce both data access latency and cost. The strategy distributes replicas among multiple cloud providers' DCs to reduce cost for the cloud users. It extracts data correlations based on their location and access frequency in order to fulfill its objectives. The number of common tasks that require the data is used to assess data correlations. Replication is limited to highly correlated data groups. Replica placement is then adjusted so that the current placement costs less than the no-replication state. Therefore, a cost model is used to estimate the replication cost, including both storage and transfer costs for data.

4 | DISCUSSION

In this section, we analyze data replication strategies in the cloud system by answering the following questions:

- **Q1:** What are the addressed replication issues?
- **Q2:** Does the strategy is provider-oriented or customer-oriented?
- **Q3:** What are the addressed SLA and QoS metrics?
- **Q4:** How cost optimizations and economic aspects are addressed?
- **Q5:** What are the used evaluation tools?

Figure 3 depicts the taxonomy of data replication in cloud systems.

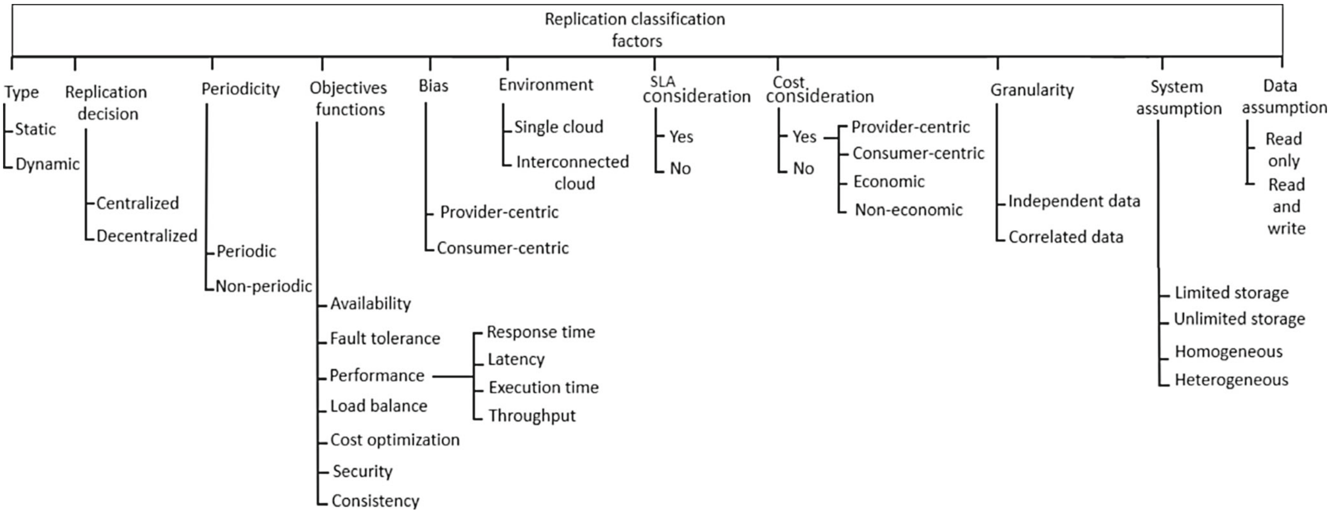


FIGURE 3 Data replication taxonomy in cloud systems.

4.1 Addressed replication issues

Data replication strategies are proposed to enable efficient replicas management by solving the following fundamental issues:

- Replicas creation includes sub-issues represented by the identification of the data to be replicated as well as the timing of initiating the replication process. The performance of the replication strategy is very sensitive to these following issues:
  - Replicas identification entails specifying the data to be replicated in accordance with the strategy's objectives. As the creation of too many replicas results in wasting resources, it is important to replicate only interesting data, for example, the ones with higher demands or higher access rates. This is related also to replica granularity, which specifies the nature of data to be replicated (such as data blocks, files, database) as well as the unit of data to be replicated, such as whether the data to be replicated is independent (individual) or grouped.[36,83]
  - Replication timing (initiation) consists of choosing the suitable time to trigger the replication. When replication is initiated too soon or too frequently, it wastes resources and degrades the system's performance. On the other hand, replicating data too late is inefficient since it obscures the advantages of replication.[18]
- Replicas placement consists of identifying the resources for storing new replicas. In order to provide the tenants' tasks with the required QoS, these resources should have sufficient storage space and acceptable bandwidth with other resources that require replicas. Moreover, they should not be overloaded either.[84,115]
- Replicas number adjustment consists of determining how many replicas to maintain in order to meet the required QoS. This process also includes selecting the replica to be removed or replaced.
  - Replicas removal consists of deleting useless replicas from the storage system. To avoid resource wastage, the replicas should be removed when they are no longer needed in the system for example, when QoS is satisfied.[20,40,91]
  - Replicas replacement consists of deleting replicas and replacing them by new ones. The replacement procedure should be considered when the capacity of the available storage resources is not sufficient to host the new replicas.[30,36,64]
- Replica selection entails selecting the replica to be used and accessed by the user's task among all the replicas in the system.[27,83]

The most discussed replication issues are replicas creation and replicas placement. These issues are important because they influence task execution time, storage space usage, network usage and so forth.[20,27,59,61,63,64,72,83,86] Interconnected cloud strategies are more focused on the replicas placement issue since they decide replica locations among the available resources of multiple cloud providers.

Adjusting the replica number was performed mainly by the creation of replicas when needed and the removal of unnecessary replicas.[78,87,89,90,116] The replicas selection issue is addressed by routing and scheduling the users' task to the replicas while considering load and bandwidth conditions.[27,30,83] However, the issue of replica replacement receives little attention.[30,36,39] Most strategies handle these issues separately, and only a few covers all of the replication issues.

4.2 Provider-oriented versus consumer-oriented

The cloud provider and the cloud tenant are two distinct entities (a person or an organization).[117] The cloud provider is the entity that delivers and typically owns cloud-based IT resources. It is responsible for making cloud services available and accessible to the tenants according to the SLA, as well as for administrative matters to ensure the system's continued functioning. The cloud tenant (*aka*, consumer, customer, user, or client) is the entity that maintains a business relationship with the provider and uses the services and resources that the provider offers.

A replication strategy may be oriented in favor of the provider or the tenant when achieving its objectives. Provider-centric strategies seek to satisfy the tenants' requirements while achieving the lowest resource management costs, such as low network usage, storage usage, energy consumption and so forth. This results in increasing the profit for the provider.[20,63,64,72,73,75,118] Most of the replication strategies are oriented towards the provider since they consider the provider to be the entity responsible for the replication process. Only a few strategies are consumer-centric. They seek to achieve the best performance as well as reduce the amount paid by the tenants and stay within their budget.[35,70,78,91,99]

4.3 Addressed QoS and SLA metrics

In this section, we discuss the main QoS metrics that data replication strategies in cloud systems aim to satisfy. These metrics are usually defined as SLOs[84] in the SLA. These SLOs may include several QoS parameters such as data availability, reliability, performance (expressed through response time, latency, execution time, throughput, etc.), privacy, and security.[119]

4.3.1 | Data availability

Commercial cloud providers include availability in their SLAs. It is the probability of a system being operational considering the alternation of operational and non-operational status. It is usually in the range of 99.9% to 99.99%. Unavailability of a requested data or the resource storing it can be caused by a variety of factors, such as hardware, software and network connections. Data access failures increase the SLA violations and influence the tenants' preferences for service provider.[95] To fulfill the availability SLO, data replication is commonly utilized.

Google, Amazon, and Microsoft use the data replication technique to meet the availability requirements of their tenants by distributing replicas among different and multiple geographic regions. GFS is a Google's distributed file system[58] that divides data into multiple blocks. To improve availability, the minimum number of replicas is set to three by default, and it can be adjusted as needed. Google offers furthermore a more sophisticated storage solutions as extension of GFS[*].

DynamoDB[†] is a NoSQL key-value database cloud service provided by Amazon. This system automatically replicates data from each created table in different regions according to the user's choice. The number of replicas for each table is set to three by default.

Windows Azure Storage[‡] is provided by Microsoft recommending four replication mechanisms. Locally-Redundant-Storage (LRS) stores three copies of each data in the same DC. Zone-Redundant-Storage (ZRS) stores three copies of each data in separate availability zones within the same region. Geo-Redundant-Storage (GRS) deploys three copies in the primary region (LRS) and three in a secondary location hundreds of miles away. Read-Access-Geo-Redundant-Storage (RA-GRS) allows read-only access to replicas located in the secondary regions.

Academic efforts focused on proposing strategies that rely on the number of replicas to ensure availability. In References 60,120, and 121, probabilistic models are used to estimate data availability. The models indicate how many data replicas are needed to satisfy data availability. These strategies define replicas to reduce the likelihood of resource blocking since replicas are considered unavailable if their storage resources are unavailable. In Reference 91, the number of replicas is increased whenever the availability requirement is not met. While the strategy proposed in Reference 122 initiates replication when the availability falls below a threshold.

When considering strategies in interconnected clouds, data availability is addressed by distributing replicas across multiple cloud providers' resources.[75,97,99,101] These strategies make use of the environment massive storage capabilities as well as its geographical dispersion.

4.3.2 | Reliability

Reliability is an important QoS parameter in cloud systems.[123,124] It refers to the property that a system can run uninterruptedly without failure. In contrast to availability, reliability is defined in terms of a time interval instead of an instant in time. However, as highlighted in Reference 125, the reliability of a system is directly dependent on availability and security. Indeed, to ensure reliability, the chosen node on which a task is executed must not have availability issues. Thus, the chance of losing availability in the middle of the job is lower. The node must also be secure so that user may not perform an important job or put any confidential data on insecure servers. Replication of data and tasks[126] is the most widely used technique to guarantee reliability. It allows creating and storing multiple replicas to reduce the likelihood of data or task loss. In order to maintain an acceptable degree of reliability, some strategies propose indeed to increase the number of replicas, which generates more costs.[127] One solution is to deal with the trade-off between reliability satisfaction and associated costs.

Reliability is also addressed with availability by some strategies,[62,69] which propose models to estimate data reliability based on replicas and storage duration. While the strategy proposed in Reference 95 relies on a nonlinear integer-programming model that considers machine failure and data request probabilities. The strategy proposed in Reference 85 addresses both reliability and response time. It selects popular data for replication and then places them to satisfy reliability requirements.

4.3.3 | Performance

Due to cloud workload heterogeneity and fluctuations, commercial cloud SLAs put less emphasis on performance guarantees.[78] Academic research focuses heavily on these guarantees. Data replication strategies consider them as SLOs, which can include multiple parameters such as response time, latency, execution time, throughput and so forth.

Response time

A user task's response time is usually defined as the time interval between its send and receive operations. When network links are congested or resources are overloaded, violations occur. Hence, response time can be reduced through load balancing and reducing bandwidth use. This is performed by predicting workload and network usage based on data accesses and resource capabilities.[27,60,94]

To ensure a reduced response time, some replication strategies rely on a definition of thresholds. In this respect, the replication strategies are triggered in accordance with the values of these thresholds.[20,30,39,40,64,68,83,89,91,118,128,129]

Some strategies consider response time as the SLO to be achieved for each tenant.[68,87-89] These strategies rely on prediction models and they are triggered based on an estimated value of response time. While the strategies presented in References 20,64,91,118, and 129 consider the replication only if the estimated response time of the tenant task is greater than the response time threshold specified in the SLA. This estimation is performed before the tenants' task execution. On their side, other strategies rely on data correlations to reduce response time.[30,36,38-40] These strategies perform a periodical analysis of the users' access history to identify the suitable data groups to replicate.

Latency

Latency refers to the delays that occur while transferring or processing data. Data replication strategies aim to reduce latency when placing replicas to ensure better performance[29,73,84,92,114,130]

In Reference 130, bandwidth utility threshold is used to prevent SLA latency violations. When aggregate bandwidth exceeds the defined threshold, new replicas are created on the requesting resource and removed randomly when it falls below. In Reference 92 replica placements are identified by setting latency threshold for read and write operations. Latency threshold is also defined by the replication strategy presented in Reference 73. The strategy enables cloud providers to satisfy the latency requirements of their applications by placing replicas and migrating them among the VMs of multiple cloud providers.

Execution time

Execution time and completion time are widely addressed by data replication strategies as QoS metrics to achieve.[32,37,65,67,78,93,94,109] A task's execution time is the time taken for processing its instructions.

In Reference 78, a database management framework is presented considering the execution time as an SLO to satisfy. The replication is triggered following a continuous number of SLA violations. Then, the users' requests are routed to the closest replicas in order to prevent SLA violations. While the strategies presented in References 94 and 65 focus on reducing data transfer time in order to reduce the execution time.

Execution time is reduced by increasing data locality. Replicas are placed in the execution resources that require them the most to reduce data transfer time.[67,93] Correlations aware strategies increase data locality by placing groups of correlated data that are accessed frequently jointly in the same locations where the users are executed.[32,37,109]

Throughput

Throughput refers to the amount of served requests or complied tasks at a given time period. Replication strategies focus on throughput besides other objectives. The replication strategy presented in Reference 93 uses a prediction model to optimize execution time and throughput. Each replica replicates based on access frequency. In Reference 98, an SLA model for interconnected cloud is used to distribute data with throughput, availability, data privacy, transfer time, and storage cost as QoS metrics.

4.3.4 | Security and privacy

Security and privacy are the key issues for cloud storage system. Some replication strategies consider security as an input parameter for their SLA model.[98] In Reference 79, data encryption and replication are combined. A secret sharing scheme and erasure codes are used before the data is stored and replicated among the resources of multiple cloud providers. Encryption method is also used in Reference 131 to secure sensitive data of the users. The cloud provider allows users to access these data relying on a single secret key. This method also allows verification by a third party. The strategy proposed in Reference 132 addresses data integrity using the T coloring method, which is a non-cryptographic method that generates less execution time. As for the strategy proposed in Reference 80, sensitive data are replicated and distributed among cloud resources to prevent corruption and data theft caused by tenants' attacks on each other.

4.4 | Cost optimization

Cost consideration is an important factor for cloud based data replication strategies. Achieving the tenants' performance expectations while ensuring low operating cost are contradictory objectives.[59] In this sense, replication strategies must consider the trade-off between ensuring the QoS metrics as indicated in the SLA and reducing and optimizing the replication costs. To achieve cost optimization, replication strategies rely on cost models to estimate the cost of replication. In this section, we discuss how data replication strategies designed for single-cloud and interconnected-cloud systems optimize the cost of replication. In addition, we discuss whether this optimization is oriented in favor of the tenant or the provider, and what are the cost models used to achieve cost optimization.

4.4.1 | Consumer-centric cost consideration

Consumer-centric cost optimization during data replication is considered when the strategies focus on reducing the cost to the consumer and respecting their budget.[77,78,90,91]

Consumer-centric cost optimization strategies in interconnected clouds aim to minimize the resource utilization by the tenants and reduce the cost they pay to cloud providers. This is often done by exploiting the varied prices and features of the resources made available by the cloud providers.[76,79,97,99,100,102]

4.4.2 | Provider-centric cost consideration

Cost optimization is provider-centric when the strategies consider reducing the cost of system management for the benefit of cloud providers, maximizing, and maintaining their profit. It is critical for cloud providers to reduce SLA violations while minimizing replication costs.[20,60,63,64,66,75,85-87,89,94,95,128,133]

Data replication strategies are mostly provider-centric compared to consumer-centric strategies, where most strategies consider the provider to be the entity responsible for operating the system and performing replication operations. Most of the strategies focus on reducing the replication cost. This is achieved by minimizing the number of managed replicas to save the storage cost[60,95] or by considering the replicas transfer cost also.[89,94] While other strategies focus on reducing the SLA violation cost.[20,64,87] Only a few strategies focus on the profit of the provider.[20,30,63,118]

Provider-centric strategies in the interconnected cloud environments focus mainly on taking advantage of the differences between the pricing policies of the cloud providers among each other to reduce the cost of management.[40,66,72,73,75,101]

4.4.3 | Non-economic cost consideration

Non-economic cost models focus on the operating costs associated with the use and management of resources such as network bandwidth, CPU time, memory space, storage space, energy consumption and so forth.[27,59,63,128]

The reduction of network and storage usage is widely addressed by the strategies.[27,60,76,79,115] It is usually performed by maintaining a minimum number of replicas.[75,116] Other strategies assess the cost of replication as a function of time. In Reference 128, cost optimization is achieved by reducing the cost of communication when accessing and updating replicas. The used cost model to estimate the replication cost focuses on the sum of the access latency for retrieving a replica from the storage and the network communication latency for transferring replicas to the requesting resource and propagating updates. The strategy proposed in Reference 94 assesses the cost of replication as the replicas creation time. On their side, other strategies focus more on the impact of energy consumption when replicating data.[59,63,129]

4.4.4 | Economic cost consideration

Economic cost models are generally related to the economic aspect of the cloud, including the pricing policies, the monetary cost of management, auctions and so forth.[20,30,40,64,66,72,73,78,90,91,118,122,133]

The focus on the economic aspect of cloud systems by replication strategies is mainly done by addressing the monetary cost of replication.[20,30,40,64,66,72,73,78,90,91] This monetary cost is often estimated by economic cost models that take into account both the cost of data storage and data transfer as well as the cost of the SLA penalties.[20,30,40,64,91] The difference between pricing policies is often considered by interconnected cloud based strategies.[40,72,73,100]

Other strategies rely on the auction method to optimize costs.[122,133] In Reference 122, the system resources compete to host the new replica, where each resource offers its own bid price as a buyer. Then, the seller resource decides to place the replica on the resource offering the highest bid price. Each buyer determines its price according to load, probability of failure, network bandwidth, and available storage space. In Reference 133, a market economic model that considers data as goods, queries as patrons, and DCs as firms, is used. The data value is estimated based on its access frequency and the monetary fee that the user pays for the queries execution, considering only the storage cost of a replica.

4.5 | Evaluation tools

To validate the performance of the replication strategies in the cloud systems, researchers implement their strategies using real cloud systems, simulation tools, and programming language.

Although real cloud implementations offer real configuration while experimenting the compared strategies under a real environment, the dynamic properties of such an environment implies that most work deals with a simulation. Simulations allow one to directly control some parameters in order to understand their individual impact on performance, for example, query arrival rate and system configuration variation. Furthermore, there is no standard architecture in cloud environments. It has been observed that the topology of a given system significantly affects the design of a data replication strategy for which it was designed. Some companies consider the transferring of all data to a single DC/cluster when executing a tenant query. This generates a significant data transfer. However, links between DCs are heterogeneous. In consequence, most of the commercial solutions (like Google Data Centers Locations[§]) as well as some recent strategies[20,118] consider within a region, several DCs that communicate through an intermediate network bandwidth. This leads to a system topology with three levels: regions, DCs and nodes that host data.

Real cloud evaluations are mainly implemented on well-known real cloud vendors such as Google, Amazon and OpenStack. These strategies deploy data on the cloud instances and model data access using tools such as workload generators that provide task tracing for applications or by deploying DBMS on the virtual instances allowing SQL and NoSQL queries to be executed.[58,62,66,78,88,90,95,100,133]

As aforementioned, deploying a real cloud system for testing purposes is a very expensive and complex process. Therefore, the majority of the strategies are evaluated and implemented based on simulation tools while considering real-world cloud parameters.[20,30,32,35,36,40,59-61,63,64,72,73,75,77,83,85,109,121,122,129] Simulation offers many advantages, such as re-producibility, cost-effectiveness and flexibility.[134] Indeed, it enables the management of setup settings as well as the design of test scenarios that can be repeated several times. CloudSim[135] is the most popular simulation tool as it can be extended based on the requirements of the implemented strategy. While some other strategies are simulated using programming languages like Matlab and Java.

4.6 | Summary of reviewed strategies

We summarize our review of data replication strategies in Tables 2 and 3 for single cloud strategies, whereas Table 4 is dedicated for interconnected cloud strategies. These tables depict some characteristics of the strategies, including: the decision entity (centralized (C) or decentralized (Dec)), the type of the strategy (static (S) or dynamic (D)), the periodicity, the bias towards the tenant or the provider, the addressed replication issues, the QoS metrics to satisfy, the cost consideration, the data correlation consideration and the evaluation method. The '+' symbol indicates that the strategy considers the concerned characteristic, whereas the '-' sign indicates that the strategy does not. Finally, 'Unspecified' indicates that the strategy does not mention that characteristic.

5 | PERFORMANCE ANALYSIS

We compare the performance of seven of the single cloud as well as four of the interconnected clouds replication strategies. These strategies include some of the data correlations aware strategies. For this performance evaluation, we dealt with a simulation since it allows us to directly control various parameters in order to understand their individual impact on performance such as the tenants' tasks number, data distribution and hardware configuration variation. Furthermore, simulation makes it possible to set up test scenarios that can be repeated, especially when comparing numerous data replication strategies. Indeed, we use CloudSim,[135] the well-known open source cloud computing simulation tool. We extend CloudSim to support data replication as well as detailed network architecture.

The strategies are evaluated according to their satisfaction of QoS parameters as well as their storage and network resources consumption. Indeed, we measure the following metrics: (*i*) the average response time, (*ii*) the number of SLA violations, (*iii*) the number of replicas, (*iv*) the storage usage, and (*v*) the effective network usage. These metrics are measured considering the variation of both the number of deployed DCs and the number of executed tasks.

Please note that for resource characteristics, we based on Reference 136 to realistically model a typical cloud environment. Economic concepts are also taken into account. As an example, a monetary pricing is defined for each resource in accordance with Google Cloud, AWS and Microsoft Azure prices. Regarding the number of repeated experiments, we proceeded with ten experiments for each measure and then average the results.

In the following figures depicting obtained results, each histogram is composed of as many bars as compared strategies. Each bar is associated with a given strategy in the order of the strategies names given in each figure.

5.1 | Single cloud strategies evaluation

We evaluate the performance of single cloud strategies, namely CDRM,[60] MORM,[59] Boru et al. strategy,[61] PEPR,[64] RSPC,[20] RCPP,[39] and CEMR.[30] This evaluation is performed while fixing the number of deployed data centers (DCs) and varying the number of the tenants' tasks. Table 5 indicates the used parameters to simulate a single cloud system distributed on a single geographic region and owned by a single provider.

TABLE 2 Single cloud strategies (part 1)

References	Main proposals	Decision entity	Type	Periodicity	(Provider/ consumer) centric	Replication issues	QoS and SLA	Cost	Profit	Correlations mining	Evaluation
60	Data replication strategy	C	D	+	Provider	Replicas placement, replicas number	Availability, access time	+	-	-	Simulation
90	Database provisioning and replication framework	C	D	-	Consumer	Replicas creation and removal	Response time	+	-	-	Real implementation
87	Data replication strategy	C	D	-	Provider	Replicas creation and removal	Response time	+	-	-	Real implementation
83	Data replication strategy	Dec	D	-	Provider	Replica creation and replica placement	Response time	-	-	-	Simulation
84	Data replication strategy	C	D	-	Unspecified	Replicas placement	Access time	+	-	-	Simulation
27	Data placement and replicas selection strategy	C	D	-	Unspecified	Replicas placement	Access latency	+	-	Graph partitioning	Real implementation
59	Data replication strategy	C	S	-	Provider	Replica creation and placement	Availability, access latency	+	-	-	Simulation
94	Data replication strategy	Unspecified	D	+	Provider	Replicas placement	Response time	+	+	-	Simulation
93	Data replication strategy	C	D	+	Unspecified	Replicas placement	Access time	-	-	-	Real implementation
61	Data replication strategy	C	D	-	Unspecified	Replicas placement	Access time	+	-	-	Simulation
31	Data replication strategy	C	D	-	Unspecified	Replicas number	Availability, access time	-	-	Frequency and access sequence analysis	Simulation

TABLE 2 (Continued)

References	Main proposals	Decision entity	Type	Periodicity	(Provider/ consumer) centric	Replication issues	QoS and SLA	Cost	Profit	Correlations mining	Evaluation
78	Database provisioning and replication framework	C	D	-	Consumer	Replicas creation and removal	Response time	+	-	-	Real implementation
32	Data placement and replication strategy	Unspecified	D	Unspecified	Unspecified	Replicas creation, and placement	Execution time	-	-	Formal Concepts Analysis	Simulation
107	Data replication strategy	Dec	D	+	Unspecified	Replicas creation	Access time	-	-	FP-Growth algorithm	Simulation
77	Data replication strategy	C	D	+	Consumer	Replicas number, replicas placement	Availability	+	-	-	Simulation
109	Data replication strategy	C	D	+	Unspecified	Replicas placement	Execution time	-	-	Frequency analysis	Simulation
63	Data replication strategy	Unspecified	D	+	Provider	Replicas creation and placement	Access latency	+	+	-	Simulation
67	Data replication strategy	Dec	D	+	Unspecified	Replicas creation and placement	Execution time	-	-	-	Simulation
33	Data replication strategy	Dec	Hybrid	+	Unspecified	Replicas creation and placement	Response time	-	-	Frequency and access analysis	Simulation

TABLE 3 Single cloud strategies (part 2)

References	Main proposals	Decision entity	Type	Periodicity	(Provider/ consumer) centric	Replication issues	QoS and SLA	Cost	Profit	Correlations mining	Evaluation
36	Data replication strategy	Dec	D	+	Unspecified	Replicas creation, placement, and replacement	Access latency	-	-	Graph partitioning	Simulation
88	Data replication strategy	Unspecified	D	+	Provider	Replicas creation and removal	Response time	+	-	-	Real implementation
64	Data replication strategy	Dec	D	-	Provider	Replicas creation, removal and placement	Response time	+	+	-	Simulation
37	Data placement and replication strategy	Unspecified	D	Unspecified	Unspecified	Replicas creation, replicas placement	Execution time	+	-	Formal Concepts Analysis	Simulation
86	Data replication strategy	Dec	D	-	Provider	Replicas creation replicas placement	Availability, Access time	+	-	-	Simulation
91	Data replication strategy	Dec	D	-	Consumer	Replicas creation replicas placement, and replica number	Response time, availability	+	+	-	Simulation
69	Data replication strategy	Hybrid	D	+	Unspecified	Replicas number replica creation and placement	Availability, reliability	-	-	-	Simulation
38	Data replication strategy	Dec	D	+	Unspecified	Replicas creation, selection, and replacement	Response time	-	-	Spectral Clustering	Simulation
95	Data replication strategy	Unspecified	S	Unspecified	Provider	Replicas number	Availability	+	-	-	Real implementation
29	Data replication strategy	Dec	D	+	Unspecified	Replicas creation, placement, and replacement	Response time	-	-	Maximal frequent correlated pattern mining algorithm	Simulation
85	Data replication strategy	Dec	D	-	Unspecified	Replicas creation and placement	Response time, availability	-	-	-	Simulation

TABLE 3 (Continued)

References	Main proposals	Decision entity	Type	Periodicity	(Provider/ consumer) centric	Replication issues	QoS and SLA	Cost	Profit	Correlations mining	Evaluation
20	Data replication strategy	Dec	D	-	Provider	Replicas creation, removal and placement	Response time, minimum availability	+	+	-	Simulation
39	Data replication strategy	C	D	+	Provider	Replicas identification, replicas placement, replicas replacement	Response time	+	+	Triadic Concepts Analysis	Simulation
30	Data replication strategy	Dec	D	+	Provider	Replicas identification, replicas placement, replicas number adjustment, replicas replacement	Response time, minimum availability	+	+	Triadic Concepts Analysis	Simulation
89	Database management framework	C	D	-	Provider	Replicas creation, removal and placement	Response time	+	-	-	Simulation

TABLE 4 Interconnected-clouds strategies

References	Main proposals	Decision entity	Type	Periodicity	(Provider/ consumer) centric	Replication issues	QoS and SLA	Cost	Profit	Correlations mining	Evaluation
76	Inter-cloud storage proxy	C	D	-	Consumer	-	Availability, fault tolerance	+	-	-	Simulation
97	Data replication strategy	C	D	-	Consumer	Replicas placement	Availability	+	-	-	Simulation
79	Inter-cloud storage system	C	D	+	Consumer	Replicas number	Fault tolerance, security	+	-	-	Real implementation
66	Data replication strategy	C	D	+	Provider	Replicas placement and selection, replicas number	Access latency, fault tolerance	+	-	-	Real implementation
114	Data placement strategy	C	D	-	Provider	Replicas placement	Access latency	+	-	Graph partitioning	Simulation
98	Data placement strategy	Unspecified	Unspecified	-	Provider	Replicas placement	Transfer time, availability privacy	+	-	-	Simulation
100	Storage framework	C	D	-	Consumer	Replicas number, replicas placement and selection	Availability, access latency	+	-	-	Real implementation and simulation
99	Data placement strategy	C	D	+	Consumer	Replicas placement and selection	Availability, fault tolerance	+	-	-	Simulation
115	Data replication strategy	Dec	Hybrid	-	Unspecified	Replicas placement, replicas number	Availability, response time	+	-	-	Simulation
35	Data replication strategy	C	D	+	Consumer	Replica creation and placement, replicas number	Access latency	+	-	Frequency and access analysis	Simulation
101	Data replication strategy	Dec	D	+	Provider	Replicas placement	Availability	+	-	-	Simulation

TABLE 4 (Continued)

References	Main proposals	Decision entity	Type	Periodicity	(Provider/ consumer) centric	Replication issues	QoS and SLA	Cost	Profit	Correlations mining	Evaluation
73	Data replication strategy	C	D	+	Provider	Replicas placement and selection, replicas number	Response time, availability	+	-	-	Simulation
75	Data replication strategy	Dec	D	+	Provider	Replicas placement, replicas number	Reliability	+	-	-	Simulation
108	Data placement strategy	C	D	+	Provider	Replicas placement	access latency	+	-	Graph partitioning	Simulation
102	Data placement strategy	C	D	+	Consumer	Replicas placement	Availability, access latency	+	-	-	Simulation
72	Data placement strategy	C	D	+	Provider	Replicas placement	Access latency	+	-	-	Simulation
40	Data replication strategy	Dec	D	+	Provider	Replicas identification, replicas placement, replicas number adjustment	Response time, availability	+	+	Spectral Clustering	Simulation

TABLE 5 Configuration of the simulation parameters for single cloud strategies.

Parameter	Value
Number of DCs	5
Number of VMs at each DC	8
Number of tasks	Between 1000 and 5000
Task length	Between 100 and 700 MI
Number of data sets	200
Data set size	Between 300 MB and 2 GB
VM processing capability	1000 MIPS
VM number of CPU	2
VM RAM	4 GB
VM storage capacity	5 GB
Inter-DCs bandwidth (*resp.* delay)	5 GB/s (*resp.* 50 ms)
Intra-DCs bandwidth (*resp.* delay)	0.5 GB/s (*resp.* 25 ms)
Response time service level objective	180 ms

5.1.1 | Average response time and SLA violations

Figure 4 indicates the impact of varying the number of tasks on both the average response time and the amount of SLA violations. As a static replication strategy, MORM is not able to adapt to the cloud workload. Therefore, it generates the highest average response time value. In contrast, CDRM, Boru et al. strategy, PEPR, RSPC, RCPP, and CEMR are dynamic strategies. Indeed, they achieve lower values of response time and SLA violations compared to the MORM strategy. The CDRM strategy creates replicas to reduce the blocking probability of the system, whereas the Boru et al. strategy creates replicas based on network conditions. Hence, the response time values of both strategies are reduced. On the other hand, PEPR, RSPC, RCPP, and CEMR use predefined response time thresholds to identify SLA violations and then replicate the data that are related to them. Indeed, their SLA violation values are reduced by around 1%, 4%, 8%, and 18% compared to the Boru et al. strategy, respectively. Focusing on correlation-aware strategies, namely RCPP and CEMR, they achieve better results by up to 16% in response time and 12% in SLA violations on average compared to strategies that replicate independent data.

5.1.2 | Replicas number and storage usage

Figure 5 illustrates the connection between the number of replicas and the percentage of storage usage. The process of replica creation increases in accordance with the number of tasks, which leads to higher storage usage percentages across all the compared strategies. The Boru et al. strategy and the MORM strategy generate a high number of replicas with the aim of reducing access time while considering energy consumption. On its side, CDRM tries to minimize the number of deployed replicas in a manner that satisfies the availability requirements. Indeed, the values of the storage usage by MORM and CDRM are lower by around 3% and 15% compared to the Boru et al. strategy, respectively. SLA-aware strategies such as PEPR and RSPC are initiated based on SLA violations. Hence, the replication process is triggered less, resulting in lower storage usage. RCPP and CEMR attain the lowest values compared to the other strategies, since they consider periodic replica creation. Furthermore, they include replica adjustment procedures, which enable them to eliminate useless replicas from the system.

5.1.3 | Effective network usage

Figure 6 illustrates the relationship between the ENU value and the number of data accesses with regard to the network architecture. In order to achieve load balancing, the MORM method inserts replicas in the cloud system in a random distribution, whereas the CDRM strategy sets replicas in the cloud system within virtual machines (VMs) that have a low blocking probability. When deciding where to place replicas, the strategies of Boru et al., PEPR, and RSPC all take into account the current state of the network. PEPR and RSPC both place replicas on high-bandwidth VMs that are located in the same geographic region as the VM that triggered the replication. While CEMR and RCPP allocate the correlated replicas to the execution resources that have the highest request for them.

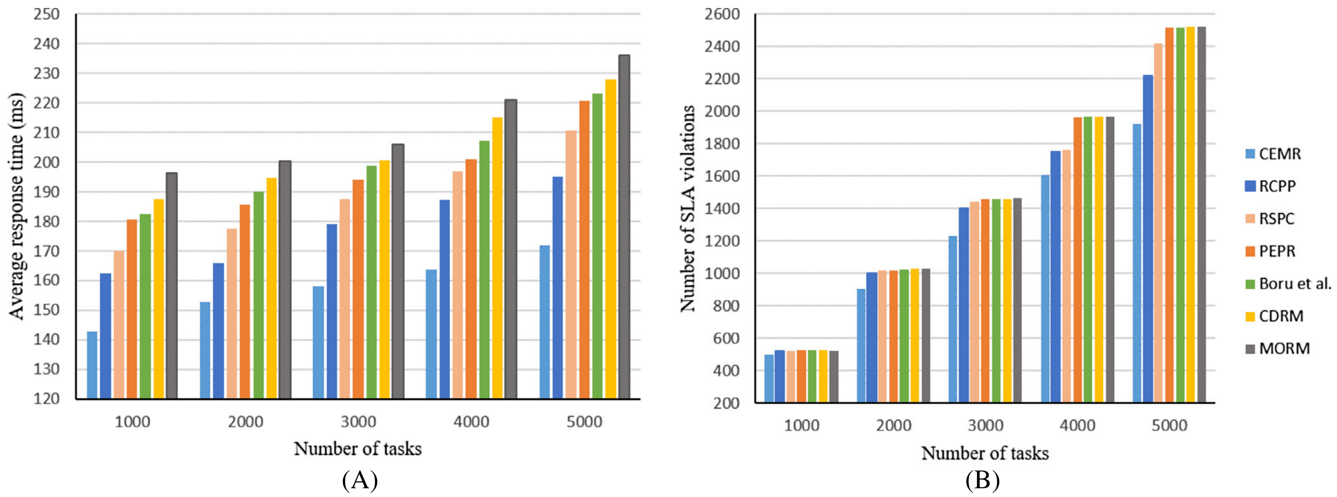


FIGURE 4 Results of (A) average response time (B) SLA violations.

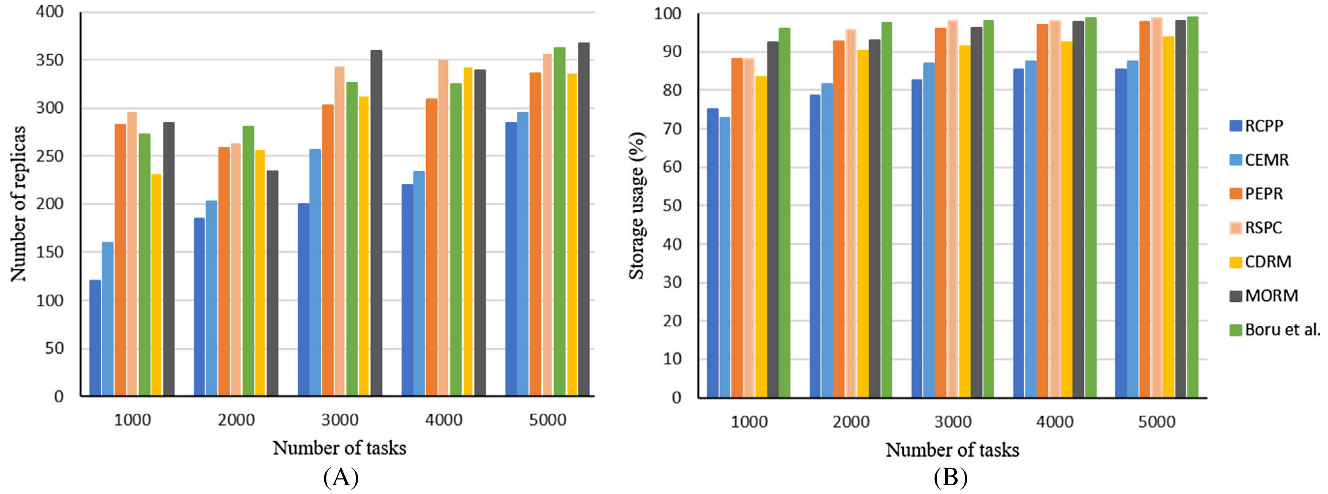


FIGURE 5 Results of (A) number of replicas (B) storage usage percentage.

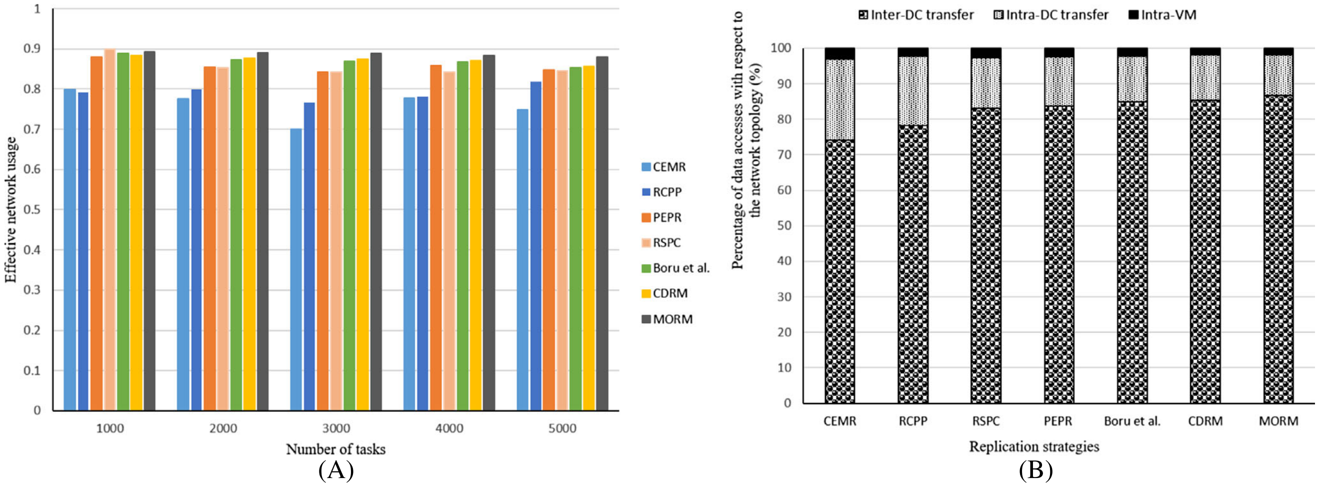


FIGURE 6 Results of (A) effective network usage (B) percentage of data accesses.

TABLE 6 Configuration of the simulation parameters for interconnected clouds strategies.

Parameter	Value
Number of cloud provider	3
Number of regions	3
Number of DCs per provider	Between 2 and 5 DCs
Number of VMs within a DC	8
Number of submitted task	Between 1000 and 10,000 tasks
Task size	Between 200 and 1000 MI
Number of data	200
Data size	Between 300 Mb and 1 Gb
Inter-region BW (resp. delay)	500 MB/s (resp. 150 ms)
Intra-region BW (resp. delay)	1 GB/ s (resp. 50 ms)
Intra-DC BW (resp. delay)	8 GB/ s (resp. 10 ms)
VM processing capability	1500 MIPS
VM number of CPU	2
VM RAM	4 GB
VM storage capacity	8 GB
Response time service level objective	180 ms

Given the replica placement criteria aforementioned for each strategy, the number of remote data accesses produced by MORM is much higher than those produced by other strategies, increasing the values of ENU. In contrast, the values of remote data accesses produced by Boru et al. strategy, PEPR, RSPC, RCPP, and CEMR are lower. Indeed, these strategies outperform the MORM strategy in terms of ENU by around 2%, 3%, 4%, 10%, and 14%, respectively. As a result, the probability of accessing the required data by the users' tasks locally increases as indicated in Figure 6B.

5.2 | Interconnected clouds strategies evaluation

We evaluate the performance of interconnected clouds strategies, namely Xie et al. strategy,[35] Mansouri et al. strategy,[73] PDRPMR,[75] and DCRF.[40] This evaluation is performed while varying the number of deployed data centers (DCs) and the number of tenants' tasks. Table 6 illustrates the used parameters to simulate an interconnected cloud system distributed among three geographic regions and formed by three cloud providers.

5.2.1 | Average response time and SLA violations

Figure 7 illustrates the results of average response time when varying both the number of tasks and the number of DCs. As the number of tasks increases, the response times of the strategies being compared also increase. When compared to the other strategies, Mansouri et al. strategy has the highest average response time value. This strategy replicates independent data periodically based on latency threshold, which results in the creation of few replicas. On the contrary, PDRPMR produces a large number of replicas, which allows it to reduce the response time value by up to 2% compared to Mansouri et al. strategy. This is because PDRPMR relies on up to two copies of the requested data in order to increase availability and reliability. As for the correlation-based strategies such as Xie et al. and DCRF, they record lower response times values, as both strategies rely on periodical analysis of the access history in order to replicate the frequently required data by the users' tasks.

Considering the impact of the DCs number on strategies performances, the task response times for the compared strategies decrease as the number of DCs rises. Mansouri et al. strategy continues to register the highest value. Whereas Xie et al. strategy, PDRPMR, and DCRF retain the lowest values by approximately 1.3%, 2%, and 6% on average when compared to the Mansouri et al. strategy, respectively.

The results of the amount of SLA violations while varying the number of executed tasks and the number of deployed DCs are indicated in Figure 8. The SLA violation values are proportional to the average response time values. When varying the tasks number, the correlation aware strategies generate lower violations amount than Mansouri et al. strategy. We then record a gain of approximately 3% for Xie et al. strategy and 17%

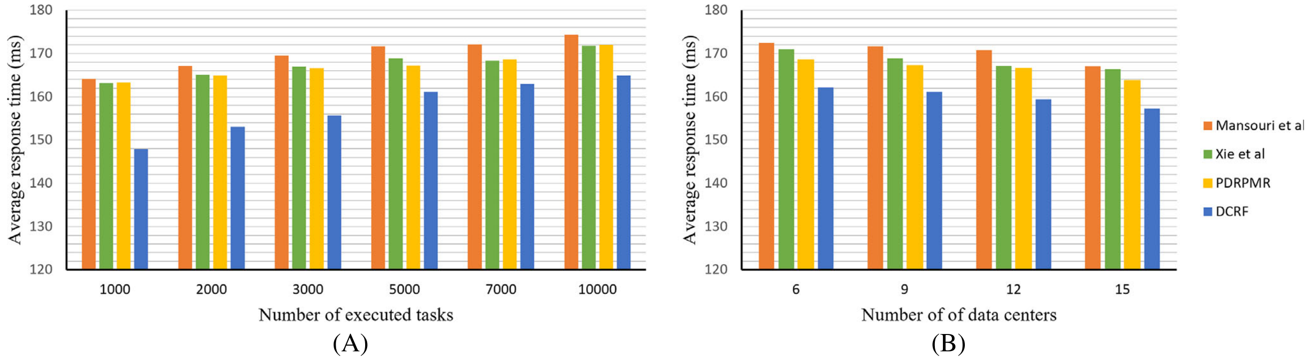


FIGURE 7 Results of average response time (in ms) (A) while varying the tasks number (B) while varying the DCs number.

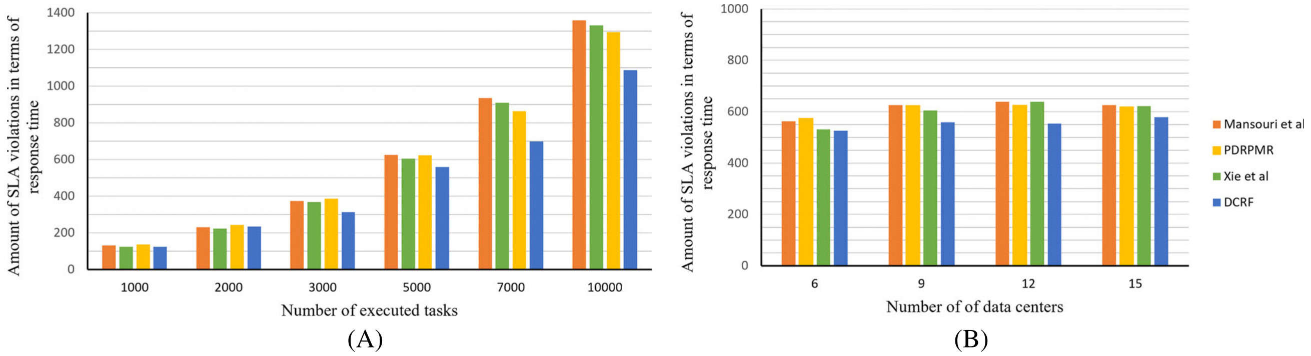


FIGURE 8 Results of the amount of SLA violations in terms of response time (A) while varying the tasks number (B) while varying the DCs number.

for DCRF. Unlike the other strategies, the DCRF strategy is initiated based on a response time threshold, allowing the strategy to capture violating tasks and replicate their correlated required data to prevent SLA violations.

When varying the number of DCs, the performance of the strategies is quite similar to the results of average response time. Correlation aware strategies register lower SLA violation values by around 2.3% for Xie et al., and 10% for DCRF, compared to the independent data replication strategies.

5.2.2 | Storage usage and number of replicas

Figure 9 depicts the relationship between storage usage and the number of replicas for the various compared strategies, while varying the number of tasks and DCs. The increase in the number of replicas is proportional to the increase in the number of executed tasks. This results in an increase in the storage usage percentage for all the strategies. The strategy PDRPMR records the highest replicas number and storage usage values by around 2%, 4%, and 6% compared to Xie et al., Mansouri et al., and DCRF strategies, respectively. Although DCRF performs a data group based replication, it maintains a balanced storage usage thanks to the used replica removal procedure. On the contrary, the strategies of Xie et al., Mansouri et al., and PDRPMR do not delete unnecessary replicas, which explains their high values of storage usage percentage.

Additional storage capacity for replicas is made available as the number of DCs increases. Both Xie et al. and Mansouri et al. strategies provide high values for the number of replicas, which in turn increases the percentage of storage usage. These strategies create replicas according to data latency requirements, as the delay increases with the increase in the number of DCs and data dispersion across them.

5.2.3 | Effective network usage

Figure 10 shows the connection between ENU value and data accesses in respect to network architecture. The PDRPMR assigns the primary replicas to the provider's VMs that have a lower cost, while the second replicas are assigned to the VMs of a different provider that provides the least

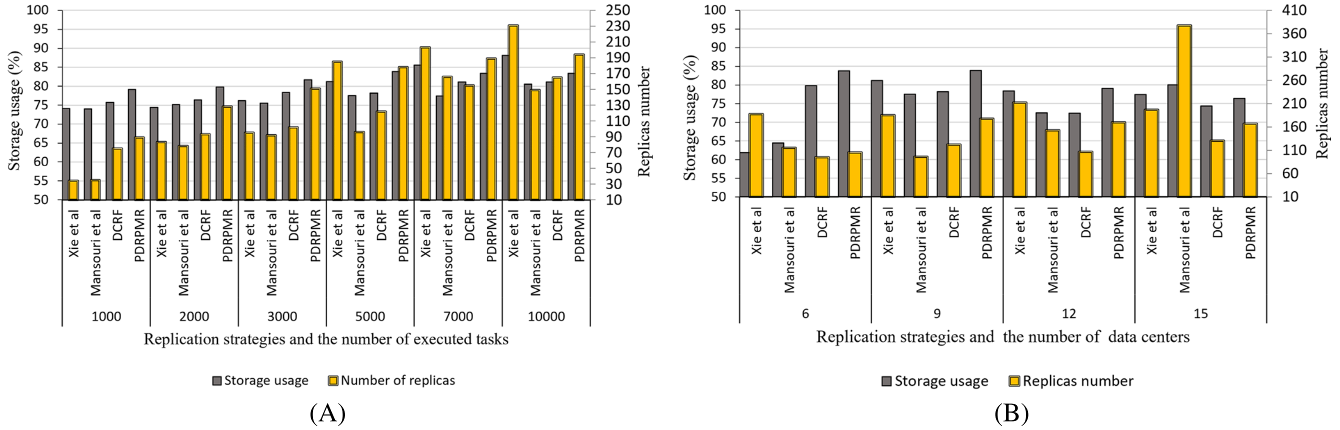


FIGURE 9 Results of the storage usage percentage and the number of replicas (A) while varying the tasks number (B) while varying the DCs number.

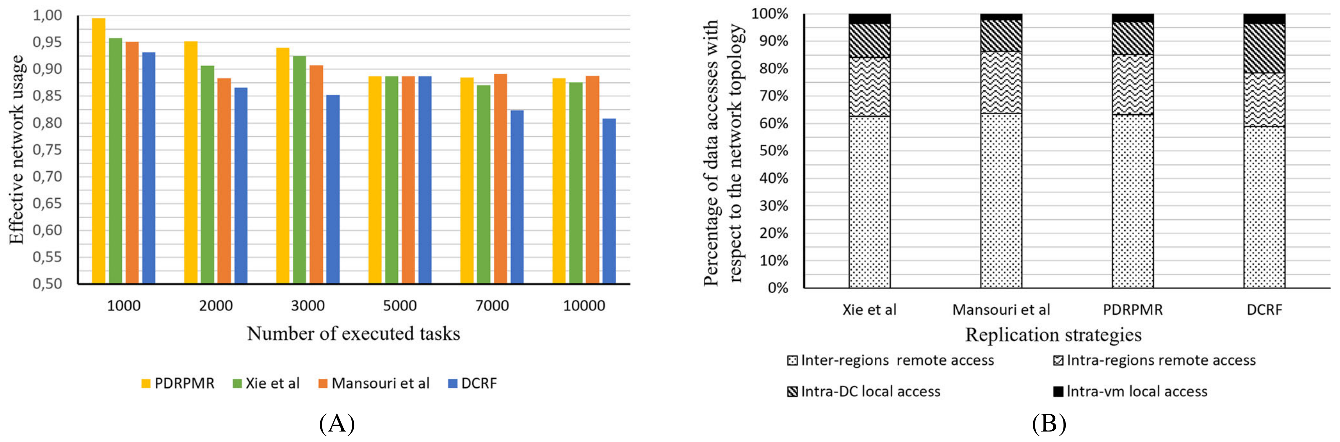


FIGURE 10 Results of (A) effective network usage while varying the tasks number (B) average percentage of data accesses with respect to the network topology during tasks execution.

data recovery time. This mapping increases both the frequency of remote data access and bandwidth usage among the interconnected cloud systems. Therefore, PDRPMR generates the highest ENU value. The other strategies initiate replication considering access latency and response time. Indeed, they focus on reducing the rate of data transfer across the network, which reduces their ENU values. Furthermore, DCRF considers the structure of the network, allowing it to achieve lower ENU values. The correlation aware strategies record a decrease in ENU values when the number of tasks increases, since they can make better use of data correlation. Figure 10B shows how correlation aware strategies, namely Xie et al. strategy and DCRF, produce the larger number of local data accesses (up to around 22%). Moreover, they also offer the fewest remote data accesses (up to around 10%) during task execution compared to independent data-based strategies. Indeed, these strategies focus on increasing data locality by placing the frequently and jointly accessed data by the users' tasks in the same locations.

6 | CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Although a high number of data replication strategies dedicated to cloud systems are proposed in the literature while taking into consideration several aspects and objectives to be reached, there are still several challenges in this field. To guide researchers to further advance, we present the following challenges and future directions:

- **Interconnected clouds and emergent systems:** Interconnected clouds are typically composed by different providers cooperating and interconnecting their resources. The advantages of relying on interconnected clouds, which translate into the availability of resources and their geographical distribution as well as the optimization of costs, encourage their adoption. This therefore resulted in the development of various resource management techniques adapted to the characteristics of interconnected clouds.[137-139] The number of data replication strategies dedicated to interconnected clouds is however limited. Hence, the need to develop strategies that exploit their advantages and adapt to their characteristics arises. This can benefit from the recent effort of researchers focusing on the development of various resource management techniques for the cloud coupled with emerging systems (such as fog and edge computing).[140-142] This allows facing propagation delays, bandwidth and energy consumption of traditional cloud systems, which tend to host all the applications and data in cloud servers.[143] In this situation, when designing data replication strategies for interconnected clouds, how to consider simultaneously different optimization goals of different providers while ensuring their collaboration with respect to the target system—traditional or emergent or involving both—is still a challenging research issue.

- **Cost and QoS trade-off:** data replication strategies in cloud systems require a trade-off between cost optimization and the fulfillment of QoS requirements. Most of the strategies consider one or more QoS metrics along with cost. However, strategies considering the economic aspects of the cloud and monetary cost are yet few. Efforts should be made in this direction. For example, developing models to estimate the monetary cost associated with replicas management, considering the perspective of the providers or the users should be investigated. Considering the context of interconnected clouds, a variety of economic issues are promising for exploitation, including (*i*) the economic competition among cloud service providers and their aims to increase profits,[139] (*ii*) the reduction of costs for the consumer in the presence of numerous providers. In this context, how to find the balance point is an interesting research challenge.
- **Developing correlation-aware strategies:** Exploiting the knowledge about the stored data on the cloud that can be extracted from the access history and data attributes holds great promise. Most of the data correlation-aware strategies rely on access correlations extracted from the access history. However, new proposals should also rely on semantic correlations or both access and semantic correlations to improve QoS. Another promising practice for leveraging data correlations is to profile the cloud users in accordance with their QoS requirements.[144] In this regard, access correlations and analyzing user feedback is a promising direction to improve the prediction accuracy of the users' demands and required QoS.
- **Self-adaptive replication strategies:** Self-adaptive replication strategies can be more effective and elastic in dealing with the dynamic nature of cloud environments. In this regard, artificial intelligence and machine learning algorithms are appropriate tools for data replication strategies to rely on Reference 42. This is also highlighted in Reference 143 where the application of machine learning techniques for orchestration of containers in cloud systems is proved to be a key solution for further improving the quality of decisions related to resource provisioning in response to the changing workloads under complex environments. Indeed, as suggested by authors in Reference 48, containers have the flexible ability to provide services with isolated functions and quick start and stop operations. Consequently, integrating containers aiming at an adaptive management of resources is an interesting issue for improving scheduling performance.
- **Energy consumption:** Energy consumption remains an open issue in cloud computing, despite the interest of data replication researchers in addressing it by relying on several factors, such as reducing network usage and data transfer, as well as managing storage resource usage. Nevertheless, the problem of providing estimation models to assess energy waste or actual energy consumption remains a challenging future research.
- **Dealing with various data concerns:**
  - *Data consistency*: For read and write data, consistency is a major issue.[81] The CAP (consistency, availability, partitioning) theorem is considered by the majority of cloud storage systems.[145] However, replication strategies seldom address consistency. Even the handful that handles it struggle to fulfill QoS requirements, mainly performance requirements, and have scaling concerns as well. Efforts should thus be increased in this area.
  - *Data privacy*: Data stored in the cloud may require a significant level of privacy, such as health-care applications' data.[146] The effort to maintain data privacy made by data replication strategies, such as considering it as an SLO, is still insufficient, leaving data privacy an open issue.
  - *Data security*: Security is one of the major challenges still facing cloud management techniques,[147] including data replication. Indeed, there are relatively few strategies focused on data security issues. This encourages further investigations.
- **Evaluation and implementation:** The majority of data replication strategies' implementations are evaluated based on simulation and numerical analysis. In this regard, the strategies consider practical and realistic scenarios based on real-world traces. Only a few strategies consider real-world implementations. However, they face limited scenarios and evaluation measures. For more accurate and reliable evaluations, replication strategies should consider real implementations while diversifying the scenarios and evaluation metrics. In addition, setting up a common evaluation platform not only for evaluation but also for meaningful comparisons is required as future work in order to offer a thorough analysis of the proposed strategies.

7 | SUMMARY AND CONCLUSION

In this survey, we reviewed data replication strategies in cloud systems. We started by reviewing the existing reviews and surveys and presenting the main existing classifications of data replications strategies in the cloud. Unlike the existing surveys that focus on the nature of the strategies and divide them into static and dynamic ones, we introduced a new classification that divided the strategies into those designed for a single cloud system and others designed for interconnected clouds. In this regard, we defined the main cloud interconnection scenarios, namely, multi-clouds and federated clouds.

Furthermore, we highlight the importance of considering data correlations during the replication process, as correlation-aware strategies have been ignored by existing surveys. In this regard, we defined the types of data correlations, including semantic and access correlations. We also presented a detailed guide explaining the key steps on which these strategies depend and how the knowledge about data correlations is employed during the replication process. Furthermore, we reviewed these strategies while taking into account the used method to extract data correlations.

Moreover, the survey discussed the following main points: (*i*) the replication issues the strategies tackled, (*ii*) the orientation of the strategies towards the provider and the consumer, (*iii*) the SLA and QoS metrics they considered, (*iv*) the cost and economic aspects of the cloud they dealt with, and (*v*) the evaluation tool they used.

We also carried out a performance analysis based on extensive simulations to investigate how single and interconnected cloud strategies deal with the trade-off between the satisfaction of QoS and cost. Therefore, we evaluated various strategies based on five evaluation metrics while varying the number of deployed data centers and the number of executed tasks. According to the simulation findings, most of the strategies designed for the single cloud focus on meeting the QoS requirements while neglecting the financial cost associated with replication. On the other hand, only a few of the interconnected strategies are concerned with the financial cost. This is performed by taking advantage of the resources of other providers and the price difference between them. Obtained results also indicate that strategies relying on economic models outperform the ones relying on cost models. Furthermore, data correlation-aware strategies outperform the other strategies in satisfying the QoS requirements. Since these strategies replicate groups of correlated data, they need additional procedures such as removing useless replicas to keep costs down. These strategies also succeed in reducing network usage through the co-localization of the correlated replicas.

Finally, several future research directions are discussed in order to improve the performances of replication strategies for both single and interconnected clouds as well as emerging computing systems. This has to be performed while establishing a trade-off between cost and QoS. In addition, several important aspects like energy consumption, data consistency, privacy, and security should be taken into consideration. An in-depth investigation of how to take further advantage for data correlations, while facing the overhead of correlation mining, especially in a realistic cloud environment where Thousands of Terabytes of data are available, constitutes a challenging issue. Designing and implementing new evaluation platforms of replication strategies is also required in order to offer a thorough analysis as well as a comparison under the same configuration.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable—no new data generated.

ENDNOTES

[*] https://cloud.google.com/, accessed on January 10, 2023.
[†] https://aws.amazon.com/dynamodb/, accessed on January 10, 2023.
[‡] https://azure.microsoft.com/, accessed on January 10, 2023.
[§] http://www.google.com/about/datacenters/inside/locations/, accessed on January 10, 2023.

REFERENCES

1. Patel P, Ranabahu AH, Sheth AP. Service level agreement in cloud computing. Proceeding of International Conference on Object Oriented Programming, Systems, Languages and Application (Cloud Workshops at OOPSLA09); 2009:212-217.
2. Hong J, Dreibholz T, Schenkel JA, Hu JA. An overview of multi-cloud computing. In: Barolli L, Takizawa M, Xhafa F, Enokido T, eds. *Workshops of the International Conference on Advanced Information Networking and Applications*. Advances in Intelligent Systems and Computing. Vol 927. Springer; 2019:1055-1068.
3. Grozev N, Buyya R. Inter-cloud architectures and application brokering: Taxonomy and survey. *Softw Pract Exp*. 2014;44:369-390.
4. Bernstein D, Ludvigson E, Sankar K, Diamond S, Morrow M. Blueprint for the intercloud-protocols and formats for cloud computing interoperability. Proceedings of the 2009 Fourth International Conference on Internet and Web Applications and Services; 2009:328-336.
5. Toosi AN, Calheiros RN, Buyya R. Interconnected cloud computing environments: Challenges, taxonomy, and survey. *ACM Comput Surv*. 2014;47:1-47.
6. Latif S, Gilani SMM, Ali L, Iqbal S, Liaqat M. Characterizing the architectures and brokering protocols for enabling clouds interconnection. *Concurr Comput*. 2020;32:e5676.
7. Megouache L, Zitouni A, Djoudi M. Ensuring user authentication and data integrity in multi-cloud environment. *Hum-Centric Comput Inf Sci*. 2020;10:1-20.

8. Najm M, Tripathi R, Alhakeem MS, Tamarapalli V. A cost-aware management framework for placement of data-intensive applications on federated cloud. *J Netw Syst Manag*. 2021;29:1-33.
9. Netto MA, Calheiros RN, Rodrigues ER, Cunha RL, Buyya R. HPC cloud for scientific and business applications: Taxonomy, vision, and research challenges. *ACM Comput Surv*. 2018;51:1-29.
10. Fu W, Liu S, Srivastava G. Optimization of big data scheduling in social networks. *Entropy*. 2019;21:902.
11. Abualigah L, Diabat A, Elaziz MA. Intelligent workflow scheduling for big data applications in IoT cloud computing environments. *Clust Comput*. 2021;24:2957-2976.
12. Bernstein PA, Hadzilacos V, Goodman N. *Concurrency Control and Recovery in Database Systems*. Vol 370. Addison-Wesley; 1987.
13. Demers A, Greene D, Hauser C, et al. Epidemic algorithms for replicated database maintenance. Proceedings of the sixth annual ACM Symposium on Principles of Distributed Computing; 1987:1-12.
14. Kemme B, Jiménez-Peris R, Patiño-Martínez M. *Database Replication*. Synthesis Lectures on Data Management. Vol 5. Springer; 2010:1-153.
15. Özsu MT, Valduriez P. *Principles of Distributed Database Systems*. 4th ed. Springer; 2020.
16. Sharma A, Kansal V. Replication management and optimistic replication challenges in mobile environment. *Int J Database Manag Syst*. 2011;3:81-99.
17. Spaho E, Barolli L, Xhafa F. Data replication strategies in P2P systems: A survey. Proceedings of the 2014 17th International Conference on Network-based Information Systems; 2014:302-309.
18. Mokadem R, Hameurlain A. Data replication strategies with performance objective in data grid systems: A survey. *Int J Grid Util Comput*. 2015;6:30-46.
19. Hamrouni T, Slimani S, Ben Charrada F. A data mining correlated patterns-based periodic decentralized replication strategy for data grids. *J Syst Softw*. 2015;110:10-27.
20. Mokadem R, Hameurlain A. A data replication strategy with tenant performance and provider economic profit guarantees in cloud data centers. *J Syst Softw*. 2020;159:110447.
21. Anitha R, Mukherjee S. MaaS: Fast retrieval of E-file in cloud using metadata as a service. *J Intell Manuf*. 2017;28:1871-1891.
22. Adams IF, Storer MW, Miller EL. Analysis of workload behavior in scientific and historical long-term data repositories. *ACM Trans Storage*. 2012;8(6):1-27.
23. Wildani A, Miller EL. Can we group storage? Statistical techniques to identify predictive groupings in storage system accesses. *ACM Trans Storage*. 2016;12(7):1-33.
24. Chou Y. Low-cost epoch-based correlation prefetching for commercial applications. Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture; 2007:301-313; IEEE Computer Society.
25. Liao J. Server-side prefetching in distributed file systems. *Concurr Comput*. 2016;28:294-310.
26. Azari L, Rahmani AM, Daniel HA, Qader NN. A data replication algorithm for groups of files in data grids. *J Parallel Distrib Comput*. 2018;113:115-126.
27. Kumar KA, Quamar A, Deshpande A, Khuller S. SWORD: Workload-aware data placement and replica selection for cloud data management systems. *VLDB J*. 2014;23:845-870.
28. Mazumdar S, Seybold D, Kritikos K, Verginadis Y. A survey on data storage and placement methodologies for cloud-big data ecosystem. *J Big Data*. 2019;6:15.
29. Mansouri N, Javidi MM, Zade BMH. Using data mining techniques to improve replica management in cloud environment. *Soft Comput*. 2020;24:7335-7360.
30. Khelifa A, Hamrouni T, Mokadem R, Ben Charrada F. Combining task scheduling and data replication for SLA compliance and enhancement of provider profit in clouds. *Appl Intell*. 2021;51:7494-7516.
31. Ragunathan T, Sharfuddin M. Frequent block access pattern-based replication algorithm for cloud storage systems. Proceedings of the Contemporary Computing (IC3); 2015:7-12.
32. Brahmi Z, Mili S, Derouiche R. Data placement strategy for massive data applications based on FCA approach. Proceedings of the 13th IEEE/ACS International Conference of Computer Systems and Applications; 2016:1-8.
33. Pan S, Xu Z, Meng Q, Chong Y. A combination replication strategy for data-intensive services in distributed geographic information system. *Int J Distrib Sens Netw*. 2017;13(5).
34. Liu X, Lian X. Study on replica strategy based on access pattern mining in smart city cloud storage system. *Wirel Pers Commun*. 2018;103:519-534.
35. Xie F, Yan J, Shen J. A data dependency and access threshold based replication strategy for multi-cloud workflow applications. Proceedings of the International Conference on Service-Oriented Computing, Vol. 11434; 2018:281-293.
36. Mansouri N, Javidi MM. A new prefetching-aware data replication to decrease access latency in cloud environment. *J Syst Softw*. 2018;144:197-215.
37. Derouiche R, Brahmi Z, Gammoundi MM, Galan SG. E-DPSIW-FCA: Energy aware FCA-based data placement strategy for intensive workflow. *Scalable Comput Pract Exp*. 2019;20:541-562.
38. Chellouf M, Hamrouni T. Popularity and correlation aware data replication strategy based on half-life concept and clustering in cloud system. *Concurr Comput*. 2021;33:e6159.
39. Khelifa A, Hamrouni T, Mokadem R, Ben Charrada F. Cloud provider profit-aware and triadic concept analysis-based data replication strategy for tenant performance improvement. *Int J High Perform Comput Netw*. 2020;16:67-86.
40. Khelifa A, Mokadem R, Hamrouni T, Ben Charrada F. Data correlation and fuzzy inference system-based data replication in federated cloud systems. *Simul Model Pract Theory*. 2022;115:102428.
41. Luong NC, Wang P, Niyato D, Wen Y, Han Z. Resource management in cloud networking using economic analysis and pricing models: A survey. *IEEE Commun Surv Tutor*. 2017;19:954-1001.
42. Khan T, Tian W, Zhou G, Ilager S, Gong M, Buyya R. Machine learning (ML)-centric resource management in cloud computing: A review and future directions. *J Netw Comput Appl*. 2022;204:103405.
43. Xu M, Tian W, Buyya R. A survey on load balancing algorithms for virtual machines placement in cloud computing. *Concurr Comput*. 2017;29:e4123.
44. Arunarani A, Manjula D, Sugumaran V. Task scheduling techniques in cloud computing: A literature survey. *Future Gener Comput Syst*. 2019;91:407-415.
45. Houssein EH, Gad AG, Wazery YM, Suganthan PN. Task scheduling in cloud computing based on meta-heuristics: Review, taxonomy, open challenges, and future trends. *Swarm Evol Comput*. 2021;62:100841.
46. Wu C, Buyya R, Ramamohanarao K. Cloud pricing models: Taxonomy, survey, and interdisciplinary challenges. *ACM Comput Surv*. 2019;52:1-36.

47. Shakarami A, Ghobaei-Arani M, Shahidinejad A, Masdari M, Shakarami H. Data replication schemes in cloud computing: A survey. *Clust Comput*. 2021;24:2545-2579.
48. Xu M, Buyya R. Brownout approach for adaptive management of resources and applications in cloud computing systems: A taxonomy and future directions. *ACM Comput Surv*. 2019;52:1-27.
49. Buyya R, Srirama SN, Casale G, et al. A manifesto for future generation cloud computing: Research directions for the next decade. *ACM Comput Surv*. 2019;51:105:1-105:38.
50. Saxena D, Gupta R, Singh AK. A survey and comparative study on multi-cloud architectures: Emerging issues and challenges for cloud federation. arXiv preprint arXiv:abs/2108.12831, 2021.
51. Milani BA, Navimipour NJ. A comprehensive review of the data replication techniques in the cloud environments: Major trends and future directions. *J Netw Comput Appl*. 2016;64:229-238.
52. Tabet K, Mokadem R, Laouar MR, Eom S. Data replication in cloud systems: A survey. *Int J Inf Syst Soc Change*. 2017;8:17-33.
53. Séguéla M, Mokadem R, Pierson J. Comparing energy-aware vs. cost-aware data replication strategy. Proceedings of the Tenth International Green and Sustainable Computing Conference (IGSC); 2019:1-8; IEEE.
54. Slimani S, Hamrouni T, Ben Charrada F. Service-oriented replication strategies for improving quality-of-service in cloud computing: A survey. *Clust Comput*. 2021;24:361-392.
55. Mokadem R, Gil JM, Hameurlain A, Küng J. A review on data replication strategies in cloud systems. *Int J Grid Util Comput*. 2022;13:347-362.
56. Mansouri N, Javidi MM. A review of data replication based on meta-heuristics approach in cloud computing and data grid. *Soft Comput*. 2020;24:14503-14530.
57. Milani BA, Navimipour NJ. A systematic literature review of the data replication techniques in the cloud environments. *Big Data Res*. 2017;10:1-7.
58. Ghemawat S, Gobioff H, Leung S-T. The Google file system. Proceedings of the nineteenth ACM Symposium on Operating Systems Principles; 2003:29-43.
59. Long SQ, Zhao YL, Chen W. MORM: A multi-objective optimized replication management strategy for cloud storage cluster. *J Syst Archit*. 2014;60:234-244.
60. Wei Q, Veeravalli B, Gong B, Zeng L, Feng D. CDRM: A cost-effective dynamic replication management scheme for cloud storage cluster. Proceedings of the 2010 IEEE International Conference on Cluster Computing; 2010:188-196.
61. Boru D, Kliazovich D, Granelli F, Bouvry P, Zomaya AY. Energy-efficient data replication in cloud computing datacenters. *Clust Comput*. 2015;18:385-402.
62. Li W, Yang Y, Yuan D. Ensuring cloud data reliability with minimum replication by proactive replica checking. *IEEE Trans Comput*. 2016;65:1494-1506.
63. Alghamdi M, Tang B, Chen Y. Profit-based file replication in data intensive cloud data centers. Proceedings of the 2017 IEEE International Conference on Communications (ICC); 2017:1-7.
64. Tos U, Mokadem R, Hameurlain A, Ayav T, Bora S. Ensuring performance and provider profit through data replication in cloud systems. *Clust Comput*. 2018;21:1479-1492.
65. Wei L-F, Ji J-W, Wu H-Y, Jing K. Towards a cloud storage data management model based on RNPT network. *Multimed Tools Appl*. 2017;76:19723-19739.
66. Wu Z, Butkiewicz M, Perkins D, Katz-Bassett E, Madhyastha HV. SPANStore: Cost-effective geo-replicated storage spanning multiple cloud services. Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles; 2013:292-308.
67. Mansouri N, Rafsanjani MK, Javidi MM. DPRS: A dynamic popularity aware replication strategy with parallel download scheme in cloud environments. *Simul Model Pract Theory*. 2017;77:177-196.
68. He L, Qian Z, Shang F. A novel predicted replication strategy in cloud storage. *J Supercomput*. 2020;76:4838-4856.
69. Sun S, Yao W, Qiao B, Zong M, He X, Li X. RRSD: A file replication method for ensuring data reliability and reducing storage consumption in a dynamic Cloud-P2P environment. *Future Gener Comput Syst*. 2019;100:844-858.
70. Sun D-W, Chang G-R, Gao S, Jin L-Z, Wang X-W. Modeling a dynamic data replication strategy to increase system availability in cloud computing environments. *J Comput Sci Technol*. 2012;27:256-272.
71. Huang K, Li D, Sun Y. CRMS: A centralized replication management scheme for cloud storage system. Proceedings of the 2014 IEEE/CIC International Conference on Communications in China; 2014:344-348.
72. Chikhaoui A, Lemarchand L, Boukhalfa K, Boukhobza J. Multi-objective optimization of data placement in a storage-as-a-service federated cloud. *ACM Trans Storage*. 2021;17:1-32.
73. Mansouri Y, Buyya R. Dynamic replication and migration of data objects with hot-spot and cold-spot statuses across storage data centers. *J Parallel Distrib Comput*. 2019;126:121-133.
74. Hassanzadeh-Nazarabadi Y, Küpçü A, Özkasap Ö. Decentralized utility- and locality-aware replication for heterogeneous DHT-based P2P cloud storage systems. *IEEE Trans Parallel Distrib Syst*. 2020;31:1183-1193.
75. Alshammari MM, Alwan AA, Nordin A, Abualkishik AZ. Data backup and recovery with a minimum replica plan in a multi-cloud environment. *Int J Grid High Perform Comput*. 2020;12:102-120.
76. Abu-Libdeh H, Princehouse L, Weatherspoon H. RACS: A case for cloud storage diversity. Proceedings of the 1st ACM Symposium on Cloud Computing; 2010:229-240.
77. Gill NK, Singh S. A dynamic, cost-aware, optimized data replication strategy for heterogeneous cloud data centers. *Future Gener Comput Syst*. 2016;65:10-32.
78. Zhao L, Sakr S, Liu A. A framework for consumer-centric SLA management of cloud-hosted databases. *IEEE Trans Serv Comput*. 2015;8:534-549.
79. Bessani A, Correia M, Quaresma B, André F, Sousa P. DepSky: Dependable and secure storage in a cloud-of-clouds. *ACM Trans Storage*. 2013;9:1-33.
80. Luo L, Xing L, Levitin G. Optimizing dynamic survivability and security of replicated data in cloud systems under co-residence attacks. *Reliab Eng Syst Saf*. 2019;192:106265.
81. Campêlo RA, Casanova MA, Guedes DO, Laender AH. A brief survey on replica consistency in cloud environments. *J Internet Serv Appl*. 2020;11:1-13.
82. Kitchenham B. Procedures for performing systematic reviews. Keele University, Keele; 2004:1-33.
83. Bai X, Jin H, Liao X, Shi X, Shao Z. RTRM: A response time-based replica management strategy for cloud storage system. Proceedings of the International Conference on Grid and Pervasive Computing; 2013:124-133.

84. Lin JW, Chen C-H, Chang JM. QoS-aware data replication for data-intensive applications in cloud computing systems. *IEEE Trans Cloud Comput*. 2013;1:101-115.
85. Miloudi IE, Yagoubi B, Bellounar FZ. Dynamic replication based on a data classification model in cloud computing. Proceedings of the International Symposium on Modelling and Implementation of Complex Systems; 2020:3-17.
86. Edwin EB, Umamaheswari P, Thanka MR. An efficient and improved multi-objective optimized replication management with dynamic and cost aware strategies in cloud computing data center. *Clust Comput*. 2019;22:11119-11128.
87. Sousa FR, Machado JC. Towards elastic multi-tenant database replication with quality of service. Proceedings of the 2012 IEEE Fifth International Conference on Utility and Cloud Computing; 2012:168-175.
88. Sousa FRC, Moreira LO, Filho JSC, Machado JC. Predictive elastic replication for multi-tenant databases in the cloud. *Concurr Comput*. 2018;30:e4437.
89. Raouf AEA, Abo-alian A, Badr NL. A predictive multi-tenant database migration and replication in the cloud environment. *IEEE Access*. 2021;9:152015-152031.
90. Sakr S, Liu A. SLA-based and consumer-centric dynamic provisioning for cloud databases. Proceedings of the 2012 IEEE Fifth International Conference on Cloud Computing; 2012:360-367.
91. Limam S, Mokadem R, Belalem G. Data replication strategy with satisfaction of availability, performance and tenant budget requirements. *Clust Comput*. 2019;22:1199-1210.
92. Shankaranarayanan PN, Sivakumar A, Rao S, Tawarmalani M. Performance sensitive replication in geo-distributed cloud datastores. Proceedings of the 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks; 2014:240-251.
93. Bui D, Hussain S, Huh E, Lee S. Adaptive replication management in HDFS based on supervised learning. *IEEE Trans Knowl Data Eng*. 2016;28:1369-1382.
94. Al Ridhawi I, Mostafa N, Masri W. Location-aware data replication in cloud computing systems. Proceedings of the 2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob); 2015:20-27.
95. Liu J, Shen H, Chi H, et al. A low-cost multi-failure resilient replication scheme for high-data availability in cloud storage. *IEEE/ACM Trans Netw*. 2020;29:1436-1451.
96. Sarwar K, Yongchareon S, Yu J, Rehman S. Efficient privacy-preserving data replication in fog-enabled IoT. *Future Gener Comput Syst*. 2022;122:538-551.
97. Chang C-W, Liu P, Wu J-J. Probability-based cloud storage providers selection algorithms with maximum availability. Proceedings of the 2012 41st International Conference on Parallel Processing; 2012:199-208.
98. Guo C, Li Y, Wu Z. SLA-DO: A SLA-based data distribution strategy on multiple cloud storage systems. Proceedings of the 2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS); 2016:602-609.
99. Oh K, Chandra A, Weissman J. TripS: Automated multi-tiered data placement in a geo-distributed cloud environment. Proceedings of the 10th ACM International Systems and Storage Conference; 2017:1-11.
100. Liu G, Shen H. Minimum-cost cloud storage service across multiple cloud providers. *IEEE/ACM Trans Netw*. 2017;25:2498-2513.
101. Hsu TY, Kshemkalyani AD. A proactive, cost-aware, optimized data replication strategy in geo-distributed cloud datastores. Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing; 2019:143-153; ACM.
102. Wang P, Zhao C, Wei Y, Wang D, Zhang Z. An adaptive data placement architecture in multicloud environments. *Sci Program*. 2020;2020:1704258:1-1704258:12.
103. Chen Y, Tong W, Feng D, Wang Z. Cora: Data correlations-based storage policies for cloud object storage. *Future Gener Comput Syst*. 2022;129:331-346.
104. Xia P, Feng D, Jiang H, Tian L, Wang F. FARMER: A novel approach to file access correlation mining and evaluation reference model for optimizing peta-scale file system performance. Proceedings of the 17th International Symposium on High-Performance Distributed Computing; 2008:185-196.
105. Zhao Q, Xiong C, Yu C, Zhang C, Zhao X. A new energy-aware task scheduling method for data-intensive applications in the cloud. *J Netw Comput Appl*. 2016;59:14-27.
106. Hu C, Deng Y. An energy-aware file relocation strategy based on file-access frequency and correlations. In: Wang G, Zomaya A, Martinez G, Li K, eds. *Algorithms and Architectures for Parallel Processing*. Lecture Notes in Computer Science. Vol 9531. Springer; 2015:640-653.
107. Elango P, Samy K. Fuzzy FP-tree based data replication management system in cloud. *Int J Eng Trends Technol*. 2016;36:481-489.
108. Khalajzadeh H, Yuan D, Zhou BB, Grundy JC, Yang Y. Cost effective dynamic data placement for efficient access of social networks. *J Parallel Distrib Comput*. 2020;141:82-98.
109. Wang T, Yao S, Xu Z, Jia S. DCCP: An effective data placement strategy for data-intensive computations in distributed cloud computing systems. *J Supercomput*. 2016;72:2537-2564.
110. Rama A, Reddy M. Data replication system in cloud based on data mining techniques. *Int J Adv Res Comput Commun Eng*. 2013;2:4216-4221.
111. Stumme G. Efficient data mining based on formal concept analysis. Proceedings of the International Conference on Database and Expert Systems Applications; 2002:534-546.
112. Lehmann F, Wille R. A triadic approach to formal concept analysis. Proceedings of the International Conference on Conceptual Structures; 1995:32-43.
113. Wei L, Qian T, Wan Q, Qi J. A research summary about triadic concept analysis. *Int J Mach Learn Cybern*. 2018;9:699-712.
114. Jiao L, Li J, Du W, Fu X. Multi-objective data placement for multi-cloud socially aware services. Proceedings of the 2014 IEEE Conference on Computer Communications; 2014:28-36.
115. Shorfuzzaman M. On the dynamic maintenance of data replicas based on access patterns in a multi-cloud environment. *Int J Adv Comput Sci Appl*. 2017;8:207-215.
116. John SN, Mirnalinee T. A novel dynamic data replication strategy to improve access efficiency of cloud storage. *Inf Syst E-Bus Manag*. 2020;18:405-426.
117. Endo PT, de Almeida Palhares AV, Pereira NN, et al. Resource allocation for distributed cloud: Concepts and research challenges. *IEEE Netw*. 2011;25:42-46.
118. Tos U, Mokadem R, Hameurlain A, Ayav T. Achieving query performance in the cloud via a cost-effective data replication strategy. *Soft Comput*. 2021;25:5437-5454.
119. Malik SUR, Khan SU, Ewen SJ, et al. Performance analysis of data intensive cloud systems based on data management and replication: A survey. *Distrib Parallel Databases*. 2016;34:179-215.
120. Hussein MK, Mousa MH. A light-weight data replication for cloud data centers environment. *Int J Eng Innov Technol*. 2012;1:169-175.
121. Qu Y, Xiong N. RFH: A resilient, fault-tolerant and high-efficient replication algorithm for distributed cloud storage. Proceedings of the 2012 41st International Conference on Parallel Processing; 2012:520-529.

122. Zhang H, Lin B, Liu Z, Guo W. Data replication placement strategy based on bidding mode for cloud storage cluster. Proceedings of the 2014 11th Web Information System and Application Conference; 2014:207-212.
123. Garraghan P, Yang R, Wen Z, et al. Emergent failures: Rethinking cloud reliability at scale. *IEEE Cloud Comput*. 2018;5:12-21.
124. Li C, Song M, Zhang M, Luo Y. Effective replica management for improving reliability and availability in edge-cloud computing environment. *J Parallel Distrib Comput*. 2020;143:107-128.
125. Dipu Kabir HM, Khosravi A, Mondal SK, Rahman M, Nahavandi S, Buyya R. Uncertainty-aware decisions in cloud computing: Foundations and future directions. *ACM Comput Surv*. 2022;54:1-74.
126. Mousavi Nik SS, Naghibzadeh M, Sedaghat Y. Task replication to improve the reliability of running workflows on the cloud. *Clust Comput*. 2021;24:343-359.
127. Li R, Hu Y, Lee PPC. Enabling efficient and reliable transition from replication to erasure coding for clustered file systems. *IEEE Trans Parallel Distrib Syst*. 2017;28:2500-2513.
128. Shorfuzzaman M. Access-efficient QoS-aware data replication to maximize user satisfaction in cloud computing environments. Proceedings of the 2014 15th International Conference on Parallel and Distributed Computing, Applications and Technologies; 2014:13-20.
129. Séguéla M, Mokadem R, Pierson J. Dynamic energy and expenditure aware data replication strategy. Proceedings of the 15th IEEE International Conference on Cloud Computing (CLOUD 2022); 2022:97-102.
130. Silvestre G, Monnet S, Krishnaswamy R, Sens P. AREN: A popularity aware replication scheme for cloud storage. Proceedings of the 2012 IEEE 18th International Conference on Parallel and Distributed Systems; 2012:189-196.
131. Yi M, Wei J, Song L. Efficient integrity verification of replicated data in cloud computing system. *Comput Secur*. 2017;65:202-212.
132. Ali M, Bilal K, Khan SU, Veeravalli B, Li K, Zomaya AY. DROPS: Division and replication of data in cloud for optimal performance and security. *IEEE Trans Cloud Comput*. 2015;6:303-315.
133. Marcus R, Papaemmanouil O, Semenova S, Garber S. NashDB: An end-to-end economic method for elastic database fragmentation, replication, and provisioning. Proceedings of the 2018 International Conference on Management of Data; 2018:1253-1267; ACM.
134. Mansouri N, Ghafari R, Zade BMH. Cloud computing simulators: A comprehensive review. *Simul Model Pract Theory*. 2020a;104:102144.
135. Calheiros RN, Ranjan R, Beloglazov A, De Rose CA, Buyya R. CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Softw Pract Exp*. 2011;41:23-50.
136. Barroso LA, Hölzle U, Ranganathan P. *The Datacenter as a Computer: Designing Warehouse-Scale Machines*. Morgan & Claypool Publishers; 2018.
137. Abedin ZU, Nazir B. Replication and data management-based workflow scheduling algorithm for multi-cloud data centre platform. *J Supercomput*. 2021;77:10743-10772.
138. Nzanzu VP, Adetiba E, Badejo JA, et al. Monitoring and resource management taxonomy in interconnected cloud infrastructures: A survey. *TELKOMNIKA*. 2022;20:279-295.
139. Najm M, Tamarapalli V. Towards cost-aware VM migration to maximize the profit in federated clouds. *Future Gener Comput Syst*. 2022;134:53-65.
140. Sriraghavendra M, Chawla P, Wu H, Gill SS, Buyya R. DoSP: A deadline-aware dynamic service placement algorithm for workflow-oriented IoT applications in fog-cloud computing environments. In: Tiwari R, Mittal M, Goyal LM, eds. *Energy Conservation Solutions for Fog-Edge Computing Paradigms*. Springer; 2022:21-47.
141. Khaleel MI. Multi-objective optimization for scientific workflow scheduling based on performance-to-power ratio in fog-cloud environments. *Simul Model Pract Theory*. 2022;119:102589.
142. Kar B, Yahya W, Lin Y, Ali A. A survey on offloading in federated cloud-edge-fog systems with traditional optimization and machine learning. arXiv preprint arXiv:abs/2202.10628, 2022.
143. Zhong Z, Xu M, Rodriguez MA, Xu C, Buyya R. Machine learning-based orchestration of containers: A taxonomy and future directions. *ACM Comput Surv*. 2022;54:1-35.
144. Ahmed U, Al-Saidi A, Petri I, Rana OF. QoS-aware trust establishment for cloud federation. *Concurr Comput*. 2022;34:e6598.
145. Gomes C, Tavares E, Junior MN, Nogueira BCS. Cloud storage availability and performance assessment: A study based on NoSQL DBMS. *J Supercomput*. 2022;78:2819-2839.
146. Al-Turjman FM, Zahmatkesh H, Shahroze R. An overview of security and privacy in smart cities' IoT communications. *Trans Emerg Telecommun Technol*. 2022;33:e3677.
147. Gupta I, Singh AK, Lee C, Buyya R. Secure data storage and sharing techniques for data protection in cloud environments: A systematic review, analysis, and future directions. *IEEE Access*. 2022;10:71247-71277.

How to cite this article: Hamrouni T, Mokadem R, Khelifa A. Review on data replication strategies in single vs. interconnected cloud systems: Focus on data correlation-aware strategies. *Concurrency Computat Pract Exper*. 2023;e7758. doi: 10.1002/cpe.7758