Journal of Network and Computer Applications 56 (2015) 60–76

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Impact of the distribution quality of file replicas on replication strategies ☆

T. Hamrouni, C. Hamdeni, F. Ben Charrada

Computer Science Department, Faculty of Sciences of Tunis, Tunis El Manar University, University Campus, Tunis, Tunisia

ARTICLE INFO

Article history:
Received 13 September 2014
Received in revised form
14 March 2015
Accepted 26 May 2015
Available online 9 July 2015

Keywords:
Data grid
Replication strategy
Evaluation metric
Distribution quality
Replica placement
OptorSim

ABSTRACT

Data grids provide scalable infrastructures for storage resources and data files management and support data-intensive applications. These applications require to efficiently access, store, transfer and analyze a large amount of data in geographically distributed locations around the world. Data replication is a key technique used in data grids to achieve these goals through creating multiple file replicas and placing them in a wisely manner. In this context, several replication strategies were proposed in the literature. The main idea of our work is to propose a new aspect of the evaluation of replication strategies which is the quality assessment of replicas placement in the data grid. This paper will indeed prove the impact of the distribution on the evaluation results. We hence show the importance of evaluating the quality of the replicas distribution in the data grid. Then, we propose evaluation processes of the quality of a given distribution. In this respect, different evaluation metrics are proposed for assessing the performances of replication strategies with respect to the distribution quality. We will also evaluate our metrics by using the OptorSim simulator and perform extensive experiments that will prove the effectiveness of our contributions.

1. Introduction and motivations

Nowadays, there is a tendency of storing, retrieving, and managing large volumes of data that are produced from many projects which require high-quality services (Andronikou et al., 2012). These data play a fundamental role in all kinds of cross-organizational researches and collaborations (Zin et al., 2012). In this context, a significant mobilization of researches has been made around the concept of *data grid* (Chervenak et al., 2000). This concept appeared in the late 90s and still evolves until now (Souri and Navimipour, 2014), offering a vast amount of resources from multiple applications (Navimipour et al., 2014). It responded to the demands of the rapidly changing computer sciences and it has been deployed worldwide in various fields namely physics (Andreeva et al., 2008; WLCG: Worldwide LHC Computing Grid, 2015), biology (Berger and Fahringer, 2010; EGEE: Enable Grids for Escience, 2015), astronomy (Webster and Barnes, 2011), oil zones discovery, and many other fields like those mentioned in Magoulès et al. (2005).

Data grids provide scalable infrastructures which enable the sharing of storage resources across geographically distributed sites to provide huge storage capacity to users. These data distributed across the grid must be available and accessible with reasonable performance. The replication technique is a useful solution to address these challenges (Cameron, 2004). The general idea of replication is to create multiple copies of the same data in several storage resources. Its main purposes consist in improving data access efficiency and providing high availability as well as decreasing bandwidth consumption, improving fault tolerance and enhancing scalability.

A lot of research has been conducted to design and implement replication strategies (Shorfuzzaman et al., 2010). In this respect, one can evaluate the effectiveness of these strategies through evaluation metrics, for example, total job time, the consumption of bandwidth, the number of replications performed, the percentage of the processors use, and the storage space used. It is, however, important to note that there is no a possible unique metric covering all the aspects of the evaluation of replication strategies. Indeed, each grid user aims at selecting evaluation metrics according to his evaluation criteria. A user can then attach particular importance to an evaluation criterion neglected by others and vice versa.

All evaluation metrics proposed in the literature adopt quantitative methods. In this situation, we propose in this work a new evaluation metric which adopts a method aiming to a qualitative assessment of replication strategies effectiveness. Noteworthily, this new evaluation metric has a straightforward impact on the expected results of the assessed quantitative metrics.

In the previous studies in this field, researchers have mainly focused on the proposition of increasingly efficient replication strategies.

☆This paper is a largely extended version of the work presented in Hamdeni et al. (2014).

E-mail addresses: tarek.hamrouni@fst.rnu.tn (T. Hamrouni), c.h.a.m.s.i@hotmail.com (C. Hamdeni), f.charrada@gnet.tn (F.B. Charrada).

http://dx.doi.org/10.1016/j.jnca.2015.05.020
1084-8045/

Indeed, they concentrated on offering direct services to users like accelerating the current execution time or minimizing its cost (Dayyani and Khayyambashi, 2013; Dogan, 2009). Unfortunately, these propositions have neglected the future use of the data grid and more precisely the factors that can affect the performances of a given strategy in the future, like the replicas distribution over the grid, generated by previous uses of the data grid.

In this respect, we highlight that a strategy may offer services not only for the current job executions, but also for the forthcoming grid users. Indeed, once a strategy achieved its execution, replicas can have either an interesting distribution over grid sites or not. This will positively or negatively affect the results of next jobs executions.

In our proposal, we focus on the quality of the replicas distribution obtained by the replication strategy. By the quality of the distribution of replicas, we mean the ability of a replication strategy to place replicas in strategic locations, improve data availability, reduce the total job time and reduce the resources consumption. In this way, we will give importance to the quality of the files distribution, which is a neglected criterion despite its impact on the main quantitative evaluation metrics like the total job time and the bandwidth consumption.

In fact, all the evaluation metrics are influenced by the distribution quality, and if we can evaluate the distribution quality, we can get a priori idea about the results to be obtained by the remaining evaluation metrics. To the best of our knowledge, there is no research that has been made dealing with the distribution quality. Worthnoting, there are some similar researches in other domains. For example, in Marketing, researchers have focused on this aspect (Maranzana, 1964; Piton, 2015). This allowed them to obtain higher rates of product availability, and consequently an increase in customer satisfaction. In the electricity domain, several works have focused on the decrease of the total power loss and the improvement of the power quality of distribution systems. This is carried out through the use of distributed generators while identifying their optimal number and their suitable locations in the system (Reddy et al., 2012).

This paper is a large extension of our previous study about the distribution quality (Hamdeni et al., 2014) in which we succinctly highlighted the problematic as well as the impact of the evaluation results of the file replicas distribution within grid sites. Here, an in-depth study of the related works is dedicated to replication strategies through addressing the main parameters used in such strategies as well as the evaluation metrics used in the literature. We then prove the necessity of studying the impact of replicas distribution on the performances of replication strategies. A thorough theoretical study is hence carried out in order to propose several metrics towards a precise assessment of the quality of a distribution. This latter point will allow us to address the important fact consisting in the evaluation of a replication strategy w.r.t. the distribution quality point of view. Several experimental results are also discussed in order to show the effectiveness of our proposal. Our main purpose in this work is to highlight the important impact of this key factor – distribution quality of file replicas – so that it will be taken into consideration in the future when designing and evaluating replication strategies.

The remainder of the paper is organized as follows. Section 2 gives an overview of the replication strategies proposed in the literature and the parameters they use as well as an overview of evaluation metrics for these replication strategies. We then propose in Section 3 a new evaluation criterion – distribution quality assessment – that will be proven to have an immediate impact on the results obtained through commonly used evaluation metrics. In Section 4, we propose a dedicated process for the assessment of a distribution quality. We then evaluate in Section 5 the effect of replication strategies on the distribution quality using three different methods. We then discuss obtained experimental results in Section 6 using the OptorSim simulator. The last section summarizes our contributions and depicts future work.

2. Related works

2.1. Replication strategies

Several replication strategies (Amjad et al., 2012; Dayyani and Khayyambashi, 2013; Grace and Manimegalai, 2014; Hamrouni et al., 2015; Ma et al., 2013; Mokadem and Hameurlain, 2015; Souri and Rahmani, 2014; Zin et al., 2012) have been proposed in the literature in order to overcome the difficulties encountered, and to ensure that the decision of replication was made to the appropriate file, at the right time and in the best location. In this section, we classify and survey the main replication strategies of the literature.

2.1.1. Classification of replication strategies

Replication strategies can be categorized based on whether they are used for read only requests or for update requests. In this paper, we assume that data is read only. So there are no consistency issues involved. Read-only replication strategies can then be categorized according to the following three complementary criteria (Chettaoui and Ben Charrada, 2014):

- *Dynamicity*: this criterion specifies if the strategy interacts with the changes in the environment or not.
- *Periodicity*: this criterion specifies when the replication strategy is triggered, at each file request or after a given period of time.
- *Nature of decision*: this criterion distinguishes between centralized and decentralized replication strategies.

The proposed classification of replication strategies is depicted in Fig. 1. In *static strategies*, any created replica will be kept in the same place until its lifetime expires or the user deletes it manually after a long period of services. Moreover, *dynamic strategies* create and delete replicas according to changes in the environment of the data grid. As the environment is dynamic, dynamic strategies are more appropriate for these systems. Dynamic strategies can be classified into two types, namely *periodic* and *non-periodic* strategies. The periodic replication strategies are triggered at each period. Note that the period can be either static or dynamic. In the static way, the period is constant and can be defined through a time period or by a fixed number of jobs. In the dynamic way, several criteria can be taken into consideration in order to adapt the duration of the period to the behavior of the grid.

The *periodic* strategies can be implemented with either a *centralized* approach or a *decentralized* approach. For the centralized strategies, the replication algorithm is triggered by a central site. On the other hand, for the decentralized strategies, the algorithm is triggered by a requesting site. On the other side, the *non-periodic* replication strategies often replicate files, when this is possible, at the request of any site for a given file. These strategies are decentralized.

For a description of the main replication strategies, interested readers are referred to several survey papers in the literature like Amjad et al. (2012), Dayyani and Khayyambashi (2013), Grace and Manimegalai (2014), Hamrouni et al. (2015), Ma et al. (2013), Mokadem and Hameurlain (2015), Souri and Rahmani (2014), and Zin et al. (2012).

2.1.2. Parameters of replication strategies

When proceeding to a decision of replication or deletion, replication strategies follow several steps, use several parameters and take into account several considerations to make the decision that is deemed appropriate for them. In this respect, we present in