Managing data reliability and integrity in federated cloud storage

Krzysztof Styrc[1], Piotr Nowakowski[1], Marian Bubak[1,2]

[1] ACK CYFRONET AGH , ul. Nawojki 11, 30-950 Kraków, Poland
[2] Department of Computer Science, AGH University of Science and Technology, Kraków, Poland
emails: kstyrc@gmail.com, ymnowako@cyf-kr.edu.pl, bubak@agh.edu.pl

Keywords: data integrity, cloud computing, cloud storage, federated cloud storage

Introduction

Cloud computing is in widespread use nowadays, especially cloud storage which provides virtually unlimited storage capacity and SLA contracts regarding high availability. However recently, numerous cloud provider downtimes and best-effort, return-of-costs SLAs allow to question the reliability of the cloud [1].

In VPH-Share project [2] we aim to build a collaborative computing environment and infrastructure where researchers from the domain of physiopathology of the human body will work together on developing new biomedical simulation software. It is envisioned that the data stored within the platform will be of vast volumes and predominantly of static nature. To avoid the risk of provider unavailability, the data is replicated and stored in federation of public and private cloud storage resources. Additionally, apart from data availability, it is crucially desired to ensure the integrity of the data. Researchers proposed efficient, probabilistic validation schemes such as proofs of retrievability (POR) [3] and data integrity proofs (DIP) [4] which aim to detect data corruption with high probability, while trying to reduce network overhead. However, neither of these methods are directly applicable to VPH-Share, because they (a) require to store the data in encoded and encrypted form, and (b) neglect network latency involved in random-bits access pattern. As a result, we propose a new validation algorithm that aims to provide probabilistic data integrity assurance with regard to VPH-Share, as well as current cloud storage limitations.

Description of a problem solution

In the context of VPH-Share project [2] we propose a data reliability and integrity (DRI) service that provides suitable API to ensure integrity of data in the cloud and to perform data replication across cloud storage providers. DRI periodically checks that the data is available and its content remains in tact. DRI depends on Atmosphere Internal Registry (AIR) for metadata persistence of integrity checksums, datasets and configuration. It accesses multiple cloud storage providers to retrieve VPH-Share data and notifies data owners of detected data unavailability or corruption via Notification Service. Typical use case is as follows: user tags selected dataset for data integrity monitoring, then DRI periodically retrieves dataset metadata from AIR registry, validates the data on multiple cloud storages and notifies the user in case of any problems. Optionally, user can issue dataset validation on request.

Our data validation algorithm modifies the aforementioned DIP approach [4] in two aspects. First, it makes data access patterns less fine grained – we request small blocks of data, not single bits. Second, it stores the integrity metadata in external repository, rather than appending it to the file stored on cloud. In setup phase, our validation algorithm divides

a file F into n chunks of equal size, computes their MAC-hashes and store them in metadata repository. In validation phase, it randomly selects k out of n chunks (where $k << n$), computes its hashes again and compares them with the original ones.

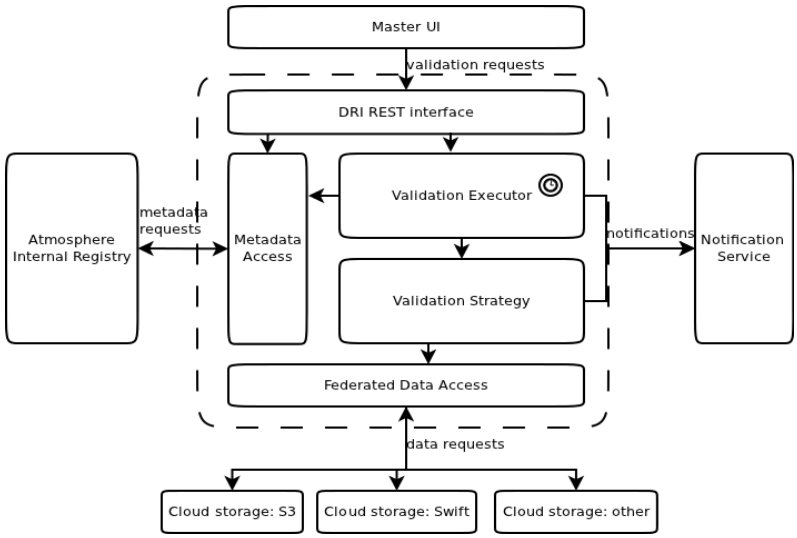

Fig. 1 DRI architecture within VPH-Share environment

Results

In comparison with the approach in which all the content of F is fetched, our algorithm significantly reduces network overhead – only k out of n chunks are downloaded, but provides only probabilistic error detection rate. However, unlike POR and DIP approaches, it (1) separates metadata and data storage and (2) better accommodates to the existing cloud storage API models.

DRI was successfully implemented and deployed within VPH-Share Cloud Platform environment, ensuring users with integrity of their data.

Conclusions and future work

In this work we proposed a method for ensuring data availability and integrity in federated cloud storage and provided a proof of concept DRI component design and its implementation within VPH-Share project.

The future work should focus on (1) investigation of possible improvements of data validation algorithm, as well as (2) separation of DRI service and providing it as a reusable component outside of VPH-Share platform.

References

[1] C. Cerin et al: Downtime statistics of current cloud solutions, IWGCR, June 2013
[2] P. Nowakowski et al: VPH-Share WP2: Data and Compute Cloud Platform, Deliverable: D2.2 , VPH-Share, 2011
[3] A. Juels and B. Kaliski: PORs: Proofs of Retrievability for Large Files, ACM CCS, 2011
[4] S. Kumar and A. Saxena: Data integrity proofs in cloud storage, COMNETS, 2011