Dependable Decentralized Storage Management for Cloud Computing

Supervisor: José Orlando Pereira (CCTC/U. Minho)
jop@di.uminho.pt

2010/2011

1 Context

Cloud computing has emerged as the preferred model for large scale information storage and exchange. For end-users, services such as DropBox provide a safe, practical backup, and ubiquitous access to personal files. Services like Flickr or YouTube provide the best approach to media publication and exchange. As an application platform, services such as Amazon S3 and EBS, provide different trade-offs for long term storage of static information and frequently modified storage volumes.

2 State of the Art and Challenges

Providing such services in a cost effective manner implies that resources used to ensure the required availability, reliability, and performance are minimized. A promising approach to achieve that is *deduplication*: Data stored often contains a large portion of repeated portions, resulting from storage of the same data by multiple tenants or multiple similar versions of the same data. This redundancy can be removed thus reducing the required disk space, but also, cache effectiveness and network bandwidth.

Although there are a number of proposals addressing deduplication for data backup, there are only a few within the more demanding context of cloud computing and these are restricted to multiple virtual machines on a single server [2] or to small tightly coupled clusters [1]. They do not scale to very large amounts of data, across multiple data centers required to make them really useful or address scenarios in which I/O performance is critical (e.g. database servers). Moreover, removing redundancy must not compromise the redundancy required to tolerate faults and to ensure the desired performance.

3 Objectives

The goal of this project is to propose and evaluate a dependable distributed storage management solution for cloud computing by achieving the following objectives:

- Propose and evaluate a mechanism for decentralized large scale deduplication of data in a cloud computing environment, considering availability, reliability, and performance.
- Determine how a deduplication layer can be integrated in a typical cloud computing stack, providing Infrastructure-as-a-Service.

References

[1] Austin T. Clements, Irfan Ahmad, Murali Vilayannur, and Jinyuan Li. Decentralized deduplication in san cluster file systems. In *Proceedings of the 2009 conference on USENIX Annual technical conference*, USENIX'09, pages 8–8, Berkeley, CA, USA, 2009. USENIX Association.

[2] João Paulo. Efficient storage of data in cloud computing. Master's thesis, Universidade do Minho, 2009.