Toward a distributed storage system leveraging the DSL infrastructure of an ISP

Pierre Meye, Philippe Raïpin-Parvédy, Frédéric Tronel, Emmanuelle Anceaume

Toward a distributed storage system leveraging the DSL infrastructure of an ISP

Pierre Meye*, Philippe Raipin*, Frédéric Tronel†, Emmanuelle Anceaume‡
*Orange Labs, France, {pierre.meye, philippe.raipin}@orange.com
†Supelec, France, frederic.tronel@supelec.fr
‡CNRS / IRISA, France, anceaume@irisa.fr

***Abstract*—Internet Service Providers (ISP) furnishing cloud storage services usually rely on big data centers. These centralized architectures induce many drawbacks in terms of scalability and reliability as datacenters represent single points of failure, and in terms of data access latencies as they are not necessarily located close to the users. This paper presents the design choices about a distributed storage system that targets these issues by leveraging only high available nodes in the Digital Subscriber Line (xDSL) infrastructure of an ISP, namely a large number of home gateways, Points of Presence, and datacenters.**

***Keywords*-Distributed storage system; availability; consistency;**

I. Introduction

The drawbacks of centralized storage systems are commonly addressed through the system decentralization over multiple geo-distributed nodes. Then data can be placed on nodes close to users in order to reduce the data access latencies and improve the system scalability and reliability as there is no more a single point of failure. Usually storage systems addressing these issues rely on Peer to Peer (P2P) technologies using unreliable peers or reliable nodes like datacenters. The main limitation of using peers is their low availability that makes systems suffer from the churn problem (*i.e.*, frequent arrivals and departures of peers) resulting in an expensive system maintenance. On the other hand, in systems relying on reliable datacenters, data are still not necessarily located close to users. Hybrid approaches relying on both peers and datacenters have been introduced in which the peers reduce datacenters workload and data access latencies, and improve system scalability while datacenters compensate peers instability with their reliability.

In this paper, we investigate the design of a hybrid distributed storage system by leveraging only high available nodes in the network infrastructure of an ISP (See Fig. 1). Specifically, our approach combines the use of datacenters with home gateways (HGs) and Points of Presence (POPs).

The remainder of this paper describes our rational to leverage HGs and POPs, the design choices about the architecture and some common issues in distributed storage systems.

II. Home gateways

Recently, in the context of energy savings, a work [1] proposed and evaluated a P2P storage system relying on home gateways. In fact, it is very common for users to get access to Internet through xDSL technologies using HGs that are equipped with memory, storage, and computing resources so that they can be compatible with several services. Moreover it has been shown that users let most of the time their HGs powered on to access to these services [1]. Thus, exploiting resources on HGs can provide a large number of high available and intelligent storage nodes located very close to the users.

III. Points of Presence

In the network infrastructure of an ISP, Points of Presence aggregate xDSL lines from users HGs and connect them to the Internet backbone. All the traffic within and across ISPs goes through the POPs. Thus, it is interesting to make them provide new services due to their natural geographic repartition and their position at the edge of the ISP network. For instance, they are leveraged, in some cases assisted with peers, for content caching and distribution in Content Delivery Networks (CDNs) [2]. We aims to leverage the POPs also to bring the intelligence of the storage system close to users.

IV. System architecture

Our architecture is illustrated in Fig. 1 on a simple xDSL infrastructure that we assume operated by a single ISP.

The home gateways are connected to the POP aggregating their xDSL lines and communicate with the storage system via this POP. The HGs connected to the same POP form its *region*. A fraction of the storage capacity of HGs is dedicated to the storage system. One part of this fraction stores some of the data that have been warehoused within the storage system. Note that these data do not necessarily belong to the owner of the HG. The other part of the fraction caches the data recently or frequently accessed by the HG owner. The cache replacement policy can be defined by the storage system or the HG owner.

The Points of Presence provide several functionalities. They cache some data that transit through them, implement the data consistency and replication strategy of the storage system, and

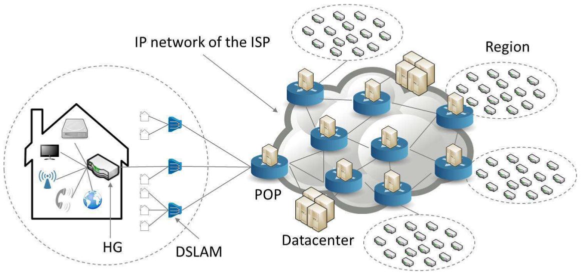

Fig. 1. The proposed architecture depicted on a simple overview of the DSL infrastructure of an ISP.

monitor the HGs they aggregate. Monitoring data (*e.g.*, storage capacity, data stored, availability state, etc.) are periodically transferred to datacenters.

The datacenters collect and store the metadata about the home gateways. They also monitor the POPs in order to build a global knowledge of the storage system allowing to balance the load of all the storage nodes proportionally to their capacity [3].

V. Data consistency

System designers usually face tradeoffs about availability, consistency and network partition tolerance when designing a distributed storage system [4]. We made the choice to serialize the write requests in order to avoid the complexity of conflicts resolution to application developers. On the other hand a versioning system on data and multiple consistency criteria to parametrize the read requests are provided, namely a readers/writer mutual exclusion, an atomic consistency, and an eventual monotonic-read consistency criteria. To achieve this, we adopt a primary/backup scheme per data item to manage data consistency. The weakness of this scheme against network partitions is mitigated by locating the primary and backup nodes in different POPs as they are interconnected with high available and redundant communication links that reduce the occurrence of network partitions.

VI. Data replication

The storage system ensures data reliability using replication. Users address their write requests to the closest POP that will perform the data replication on behalf to them on the primary (the POP that handled the write request of the first version of a data item) and backup (randomly chosen) consistency managers of a data item. It aims to mitigate the bottleneck that users low upload bandwidths may cause by moving the replication overhead to the IP network of the ISP where the bandwidth is much larger. To reduce the length of write requests, as soon as a data item is cached on both its primary and backup POPs, the user is notified to terminate the write request and make the data item available for subsequent read/write requests. The primary and backup POPs replicate data asynchronously in their regions. Then to improve the requests throughput, a stripping method can be applied allowing to store and retrieve the fragments composing the data item in parallel.

VII. Multi-tiered storage perspectives

Leveraging the different types of nodes in our architecture would make the system suitable for a multi-tiered storage organization which can refer to an automated placement of data on nodes that optimize performance, availability, recovery, cost, etc. For instance, in our system these types of nodes could be distinguished in three tiers for storing cold, warm, and hot data while allowing data to move from one tier to another depending on data access patterns.

Cold data are infrequently or never accessed. Datacenters would be suitable to handle the steadily growing amount of cold data in a cost efficient manner using inexpensive and energy efficient storage devices with large capacities.

Warm data have been recently accessed or stored in the storage system and present a high probability to be accessed than cold data. They could be stored in regions from where their users access to the storage system allowing data to be close to them for reducing access latencies. It could also localize traffic within regions and minimize network traffic across regions.

Hot data are frequently accessed and could be stored in POPs and distributed like in a CDN. It would allow to answer users requests while avoiding to overload HGs that have low upload bandwidths or datacenters that exhibit large access latencies.

VIII. Conclusion

In this paper we have presented the main principles of a distributed storage system leveraging high available nodes in the network infrastructure of an ISP to provide reliability, scalability and reduce data access latencies. We are currently implementing simulations on the different aspects of the architecture. For future work, we plan to bring more storage intelligence close to the users (*i.e.*, in POPs or HGs). We also seek to integrate efficient placement algorithms to minimize the amount of metadata to maintain in order to place/locate data in the storage system. We believe that our architecture could also be suitable for user-generated content distribution and storage costs reduction via the multi-tiered data placement. Evaluation of those points could be other directions for future work.

References

[1] V. Valancius, N. Laoutaris, L. Massoulié, C. Diot, and P. Rodriguez, "Greening the Internet with nano data centers," in *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies (CoNEXT)*. ACM, 2009, pp. 37–48.

[2] C. Huang, A. Wang, J. Li, and K. W. Ross, "Understanding hybrid CDN-P2P: why limelight needs its own Red Swoosh," in *Proceedings of the 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*. ACM, 2008, pp. 75–80.

[3] S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan, "Volley: automated data placement for geo-distributed cloud services," in *Proceedings of the 7th USENIX conference on Networked systems design and implementation (NSDI)*, 2010.

[4] D. J. Abadi, "Consistency Tradeoffs in Modern Distributed Database System Design: CAP is Only Part of the Story," *IEEE Computer Society*, vol. 45, pp. 37–42, 2012.