On the Cost-Benefit Tradeoffs of Cloud Storage Services for End Users and Service Providers

Glauber D. Gonçalves[1], **Alex B. Vieira**[2], **Jussara M. Almeida**[1]

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais

[2]Departamento de Ciência da Computação - Universidade Federal de Juiz de Fora

Abstract. *Cloud Storage is a very popular Internet service. It allows users to backup data to the cloud as well as to perform collaborative work while sharing content. Despite the increasing interest in cloud storage, a thorough investigation on the costs and benefits of this service for providers and end users has not been conducted yet. This dissertation aims at investigating such cost-benefit tradeoffs for both providers and end users jointly. Using data collected from a real service (Dropbox), we developed new models, methodologies and tools to support providers and users in improving jointly their benefits in cloud storage.*

1. Introduction

Cloud storage is a data-intensive Internet service that synchronizes files from the end user devices, such as PCs, tablets and smartphones, with the cloud. It offers the means for users to easily backup data and perform collaborative work, with files being automatically uploaded, shared and synchronized. Cloud storage is already one of the most popular Internet services [Bocchi et al. 2015b]. Well-established players such as Dropbox, Google and Microsoft, face a fierce competition for customers.

Despite the increasing interest in cloud storage, little is known about the underlying processes that generate workload to these services. Previous studies of user behavior in this type of service focused mostly on identifying performance bottlenecks as well as proposing benchmarks [Drago et al. 2012, Bocchi et al. 2015a]. Modeling the workload patterns derived from user behavior, which is key to analyze system performance, costs as well as the future user satisfaction for these services, has not been tackled yet. Functionalities as file synchronization in multiple user devices and collaborative work are not provided by existing workload models based on well-known Internet services, such as e-commerce, video streaming and online social networks [Calzarossa et al. 2016].

Moreover, cloud storage services requires cost-effective architectures to support the above mentioned functionalities. In fact, cloud storage providers foster content sharing functionalities as they might attract users to the service as well as increasing their data volume in the cloud [Gracia-Tinedo et al. 2016]. On the other hand, such functionalities pose extra costs to providers, as more data transference to/from the cloud will be required to synchronize content shared among multiple user devices. The development of cost-effective architectures for content sharing in cloud services have the potential to benefit not only service providers, but also local networks and end users, and it is then an important issue to be investigated.

Finally, cloud storage services have also an economic dimension which encompasses the pricing/incentive strategies adopted by the service and the resource demands

imposed by the users when interacting with the service. However, a thorough investigation on the costs and benefits of this service for providers and end users has not been conducted yet. For example, most providers adopt pricing models that allow free storage usage with limited space, charging for extra space, according to the freemium business model. However, such model clearly pressures the provider to reduce costs in order to maintain profitability, since the fraction of users who pay for the service is often small (e.g., 4% in Dropbox[1]). On the other hand, the policies adopted to reduce costs must not hurt user satisfaction, at the penalty of reducing service attractiveness and losing the paying users. Previous studies have investigated this issue from the perspective of either the user [Naldi and Mastroeni 2013, Shin et al. 2014] or the provider [Wu et al. 2013], but not both, thus offering a limited interpretation. A prior effort that jointly analyzed user and provider perspectives [Lin and Tzeng 2014] does not take into account typical user behavior patterns and the roles of file synchronization and collaborative work, which are key components to a cost-benefits analysis.

2. Problem Statement and Research Goals

The problem we propose to tackle is the following: given a set of users U and a cloud storage service provider s that employs a set of policies, such as variable pricing/incentive strategies (e.g., free service up to X bytes, p dollars from X to Y bytes, no transference limit for file synchronization and collaborative work among users, etc), how should this provider assess the effectiveness of policies that aim at increasing both profitability and user's satisfaction?

Various characteristics of the service should be taken into account when tackling this question. Our assumption is that a solid understanding of the common patterns of user behavior can guide the design of new solutions to improve services in terms of architecture and operating costs in order to better meet user satisfaction. Towards building such understanding, we narrow down this dissertation into three complementary research goals (RGs), as shown in Figure 1: (RG1) characterize typical user behavior and modeling the workload patterns derived from it in cloud storage services; (RG2) investigate the impact of content sharing on the costs of service providers and on the level of user activity; and (RG3) model cost-benefit tradeoffs between end users and the service provider. The expected contributions are novel models, methodologies and tools to support providers in developing new policies, that are cost-effective for both end users and providers.

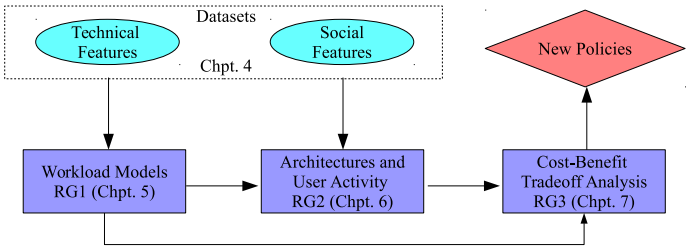

Figure 1. Pictorial representation of our Research Goals.

[1] http://onforb.es/1Kle8IE

3. Contributions

To our knowledge, the results we have produced represent the state-of-the-art in terms of user behavior, workload patterns, design and performance of distributed architectures and cost-benefit analyses on cloud storage services. Our current contributions are summarized as follows:

- The first user behavior model and synthetic workload generator for cloud storage services. These contributions were presented in two conference papers [Gonçalves et al. 2014b, Gonçalves et al. 2014a] and one journal paper [Gonçalves et al. 2016b].
- The first study about the impact of content sharing on the traffic and user activity of cloud storage. These contributions were presented in two conference papers [Gonçalves et al. 2015, Gonçalves et al. 2016c] and one journal paper [Gonçalves et al. 2016a].
- A novel model to analyze cost-benefit tradeoffs of service policies in various scenarios, specifically, to identify whether and when a win-win solution that meets the interests of provider and users can be achieved. These contributions were presented in two conference papers [Gonçalves et al. 2016, Gonçalves et al. 2017].

It is worth mentioning that our papers have been well received by the research community. Our publication [Gonçalves et al. 2014b] received the 3rd Best Paper Award and our publication [Gonçalves et al. 2015] received the 2nd Best Paper Award, both given by the Brazilian Symposium on Computer Networks and Distributed Systems (SBRC) TPC. Furthermore, one of the papers of this dissertation [Gonçalves et al. 2014a] describing our user behavior model has already received 22 citations, according to Google Scholar.

4. Overview of Results

The development of the three aforementioned research goals requires the analysis of real data so as to produce realistic representations. Thus, data collection from real cloud storage services is a key step in our research. As the stored content is private and synchronization protocols are mostly proprietary, the current knowledge of how cloud storage services work is limited, which makes data collection very challenging. Yet, the authors of [Drago et al. 2012] developed methods to obtain data of Dropbox, which is one of the currently most popular cloud storage service, from traffic measurements. Such methods respect the privacy of users and services, as data is collected passively and remains anonymized. In this dissertation, we improve this method by proposing a post-processing data methodology, that estimates the volume of updates made on content stored by users on Dropbox folders from TCP flows. Data about Dropbox usage was collected during 12 months at four vantage points, including two university campuses and two Internet service providers (ISP) networks.[2]

User Behavior and Workload Model Functionalities of cloud storage as file synchronization to various user devices and collaborative work are not provided by the existing workload models based on well-known Internet services, such as e-commerce, video streaming and online social networks. Towards filling this gap, we propose a novel model of user behavior for cloud storage services.

[2]Datasets are available at `http://locus.dcc.ufmg.br/datasets/pcssharing.html`.

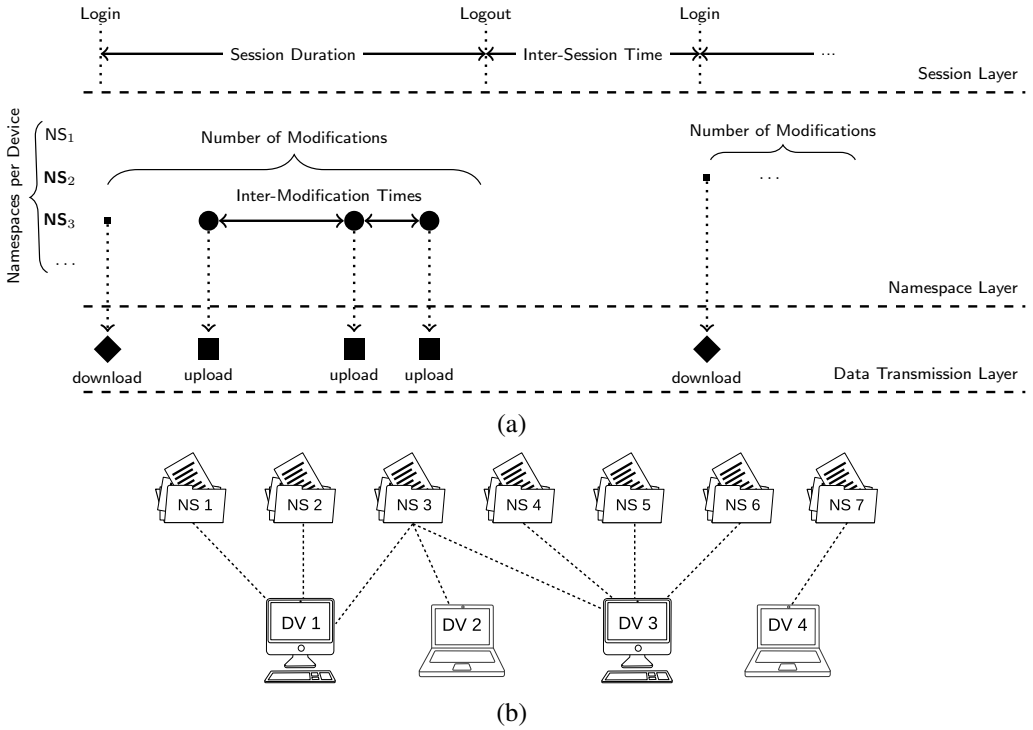

Figure 2. (a) Hierarchical model of the cloud storage user behavior. (b) Example of the content propagation network (NS and DV stands for namespace and device respectively).

The proposed model is composed by two parts: (i) the working dynamics of each user device in isolation; and (ii) the set of *namespaces*, i.e., the data structure adopted for shared folders by Dropbox[3], that is used to propagate content among devices. Figure 2(a) depicts how our model represents a single device. It captures three fundamental aspects: (a) the sessions of devices – i.e., devices becoming on-line and off-line; (b) the frequency the user makes modifications in the file system that result in storage workload; and (c) the transmission of data from/to the cloud. Each aspect is encoded in a layer forming a hierarchical model for the user behavior. Figure 2(b) shows an example of the content propagation network, representing how folders are shared among the several devices of a single user, or among devices of different users. We model this network as a bipartite graph connecting devices to folders. This graph is generated by a random graph process based on the probability distributions of folders per device and vice-versa. It is worth noting, each component of model layers shown in Figure 2(a) is parametrized by a probability distribution. Thus, all model components can be characterized to represent realistic workload in different networks. Indeed, we characterized each model component using the datasets collected from Dropbox in four distinct networks and provided their statistical distributions in Appendix A of the original dissertation text.

We use the proposed model to drive the development of CloudGen, a new synthetic workload generator that allows the simulation of the network traffic created by cloud storage services in various realistic scenarios. We validate CloudGen by comparing synthetic traces with actual data from operational networks. We then show its applica-

[3] `https://blogs.dropbox.com/tech/2014/07/streaming-file-synchronization`

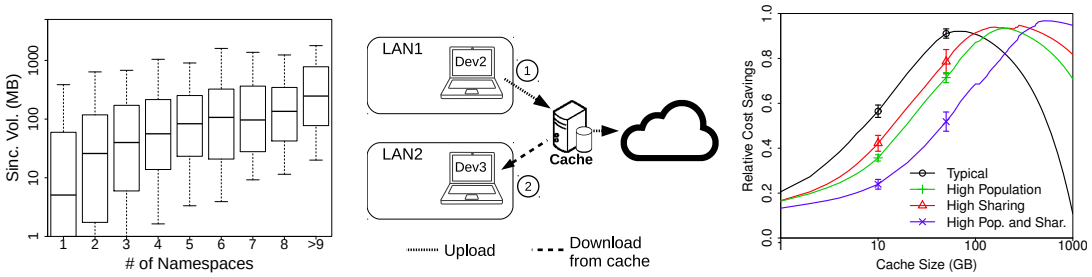

Figure 3. (a) Relationship between sharing and activity levels of users per month in Campus-1. (b) Our proposal for an alternative synchronization architecture. (c) Performance of the cache-based architecture (Relative Cost Savings) in four scenarios based on synthetic traces (CloudGen).

bility by investigating the impact of the continuing growth in cloud storage popularity on data transference consumption. Our experiments indicate that a hypothetical 4-fold increase in both user population and content sharing could lead to 30 times more network traffic. We offer CloudGen as free software to the community.[4]

Effects of Content Sharing We show empirical evidence that activity of users in cloud storage is highly correlated with content sharing. This is an important information for service providers interested in attracting users and increasing their profits, making users to store more data in the service. On the other hand, content sharing represents an extra cost for providers, i.e., more data transference to/from the cloud. We then show how cost-effective architectures for data synchronization and collaborative work have the potential to benefit service providers.

To conduct this investigation, we first compute statistics about sharing activity of Dropbox users in our datasets. A relevant percentage of users (44–65%) is associated with at least one shared folder, i.e., a folder linked to at least two devices. By monitoring the version number of shared folders in different devices, we conclude that avoidable downloads (i.e., a single update going to more than one device) are very common: up to 25% of the Dropbox incoming traffic in monitored networks. To understand and quantify importance of sharing for users, we compute the correlation between levels of sharing and activity. The former represent number of distinct *namespaces* (i.e., Dropbox shared folders), whereas the later, the volume of synchronization (upload and download removing avoidable ones), both associated with user per month. Figure 3(a) shows boxplots summarizing the distributions of activity level, for samples of users per month grouped in each level of sharing. As one can observe, the higher the sharing level, the higher is the median activity level of users, despite the variability of activity towards smaller activity levels. In fact, both variables present a high correlation (Spearman coefficient was 0.73–0.98 in the four networks), which indicates that sharing contributes to keep users active in the service.

We propose an alternative synchronization architecture that uses caches to offload storage servers from avoidable downloads, as shown in Figure 3(b). Our experiments based on synthetic traces generated by CloudGen show that the approach cost-effectively

[4]CloudGen is available at `http://cloudgen.sourceforge.net`.

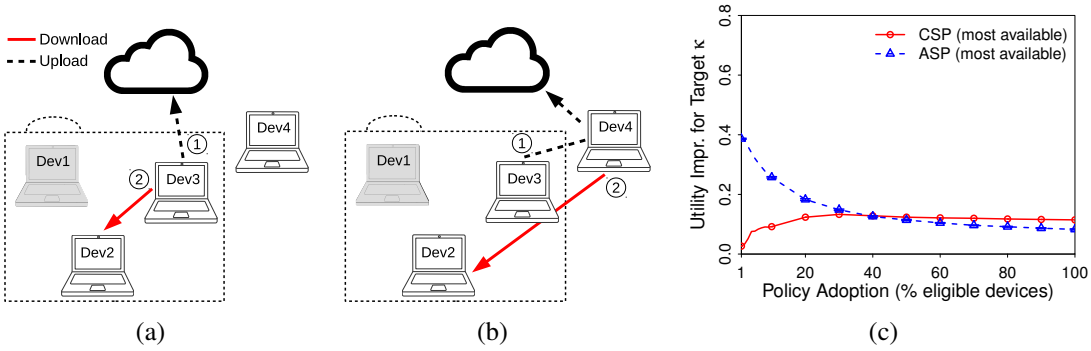

Figure 4. (a) Contact Sharing Policy – CSP: Dev 1, 2 and 3 are associated with the same shared folder, Dev 3 sends new updates to the cloud and serves Dev 2 (Dev 1 is offline), Dev 1, when back online, can retrieve updates only from the cloud if no one online. (b) Anonymous Sharing Policy – ASP: Dev 3 sends new updates to the cloud through Dev 4 (often online but not associated with the folder) which serves Dev 2 (Dev 1 is offline), Dev 1 can also retrieve updates from Dev 4 if no one online. (c) Utility improvements at target κ (i.e., equal improvements for the provider and users) as a function of adopting devices.

avoids most downloads. Figure 3(c) shows the relative cost savings during 1 month for four scenarios: a typical campus network experimenting growth in user population and/or sharing (see the four curves in the figure). Note the inflection of curves occurs when the best tradeoff between costs and benefits of the architecture is achieved. For example, we found that a reasonably small cache (e.g., 70 GB) could offload servers around 92% of the costs for serving avoidable downloads, reaching 95–97% if we consider adequate cache sizes (e.g., 280–500 GB) in scenarios with larger user populations and content sharing.

Cost-Benefit Model We propose a general model for the costs and benefits of cloud storage considering both users and providers. Our model is based on utility functions that capture, in an abstract level, the satisfaction of the service provider (U_s) and the user (U_i), i.e., the benefits minus the costs of the service for each party. The greater the utilities are, the more satisfied providers and users become.

$$U_s = \mathcal{R} - (\alpha * \mathcal{S} + \beta * \mathcal{T}); \qquad U_i = V_i * X_i - P_i \tag{1}$$

As such, our proposal is appealing for capturing in a simple (but representative) model key components of the service: provider's revenue ($\mathcal{R}$), i.e., paying users and secondary sources; provider's cost in terms of bytes stored ($\mathcal{S}$) and transfered ($\mathcal{T}$) in the cloud and respective unit costs α and β; user's storage capacity (X_i); user's valuation (V_i) for each byte stored in the cloud; and user's cost given by the service price (P_i).

Then, we apply the model to evaluate alternative policies for content sharing in cloud storage. We propose two alternative policies (CSP and ASP, see Figures 4 a–b) for the current service sharing architecture. These policies count on user contribution to reduce providers' costs via a Peer-to-Peer (P2P) architecture. The users contribute with part of their idle resources (e.g., upstream bandwidth and/or storage), receiving as a compensation a bonus proportional to the costs offloaded from the provider. In this way, the provider can improve its utility by reducing its operational costs, whereas users' utility improvements come from earned bonus.

Ultimately, the proposed model allows us assessing the effectiveness of those alternative policies in order to improve both the provider and users' utilities in various realistic scenarios. Our experiments based on trace driven simulations of policies using the Dropbox datasets indicate the scenarios where such policies are advantageous for providers and users. Figure 4(c) shows the utility improvements at a target operation point κ for percentages of policies adoption that vary from 1 to 100% of user devices in one ISP. Here we take no side and argue that a "good" κ is the value for which both provider and (an average) user experience the *same* improvements in their utilities. As we can see, all percentage of adoptions present improvements. Interesting, the highest improvements happen for lower adoption percentages (ASP reaches 39% of utility improvements for 1% adoption, whereas CSP reaches 13% improvements for 30% adoption). Improvements drop with adoption in ASP because content is distributed over devices that are often offline. On the other hand, improvements increase with adoption in CSP and drop slightly as offloads are distributed across a larger number of user device.

5. Final Remarks and Impact

In the dissertation, we conducted a thorough study of cloud storage services with the assumption that a solid understanding of the common patterns of user behavior can guide the design of new solutions to improve these services in terms of architecture and operating costs in order to better meet user satisfaction. Given that, our contributions are three-fold: (i) a user behavior model with the implementation of a software to generate realistic synthetic workloads; (ii) measurement of the impact of content sharing on the traffic of cloud storage services with the proposal of cost-effective architectures for sharing functionalities; and (iii) cost-benefit models to support providers in choosing policies that improves jointly providers and users benefits. We reached these results from the three research goals of this dissertation, which validate the aforementioned assumption.

The impact of the dissertation can be observed by the relevant results reported in very qualified journal and conference publications. We emphasize the importance of the journal papers published in the IEEE Internet Computing Magazine and Elsevier Computer Networks, both Journals Qualis A1, which reinforce that our research has been well accepted by the research community. Moreover, it is worth mentioning our papers published in the international conferences IEEE ICC and IEEE MASCOTS, Qualis A1 and A2 respectively, in addition to our two papers award in SBRC, which is the most important national scientific event on computer networks and distributed systems.

References

Bocchi, E., Drago, I., and Mellia, M. (2015a). Personal Cloud Storage Benchmarks and Comparison. *IEEE Transactions on Cloud Computing*, PP(99):1–14.

Bocchi, E., Drago, I., and Mellia, M. (2015b). Personal Cloud Storage: Usage, Performance and Impact of Terminals. In *Proc. of the IEEE CloudNet*.

Calzarossa, M. C., Massari, L., and Tessera, D. (2016). Workload characterization: A survey revisited. *ACM Computing Surveys*, 48(3):48:1–48:43.

Drago, I., Mellia, M., Munafò, M. M., Sperotto, A., Sadre, R., and Pras, A. (2012). Inside Dropbox: Understanding Personal Cloud Storage Services. In *Proc. of the ACM IMC*.

Gonçalves, G., Drago, I., da Silva, A. P. C., Vieira, A. B., and de Almeida, J. M. (2014a). Modeling the Dropbox Client Behavior. In *Proc. of the IEEE ICC*.

Gonçalves, G., Drago, I., da Silva, A. P. C., Vieira, A. B., and de Almeida, J. M. (2017). Cost-benefit tradeoffs of content sharing in personal cloud storage. In *Proc. of the IEEE MASCOTS*.

Gonçalves, G., Drago, I., Vieira, A., Silva, A., and Almeida, J. (2016a). The impact of content sharing on cloud storage bandwidth consumption. *IEEE Internet Computing*, 20(4):26–35.

Gonçalves, G., Drago, I., Vieira, A., Silva, A., Almeida, J., and Mellia, M. (2016b). Workload models and performance evaluation of cloud storage services. *Elsevier Computer Networks, DOI:* `http://dx.doi.org/10.1016/j.comnet.2016.03.024`.

Gonçalves, G., Drago, I., Vieira, A. B., da Silva, A. P. C., and de Almeida, J. M. (2014b). Characterizing and Modeling the Dropbox Workload. In *Proc. of the SBRC*.

Gonçalves, G., Drago, I., Vieira, A. B., da Silva, A. P. C., and de Almeida, J. M. (2015). Analyzing the Impact of Dropbox Content Sharing on an Academic Network. In *Proc. of the SBRC*.

Gonçalves, G., Vieira, A. B., da Silva, A. P. C., and de Almeida, J. M. (2016c). Trabalho Colaborativo em Serviços de Armazenamento na Nuvem: Uma Análise do Dropbox. In *Proc. of the SBRC*.

Gonçalves, G. D., Drago, I., Borges, A. V., Couto, A. P., and de Almeida, J. M. (2016). Analysing costs and benefits of content sharing in cloud storage. In *Proc. of the ACM Workshop LANCOMM*.

Gracia-Tinedo, R., García-López, P., Gómez, A., and Illana, A. (2016). Understanding data sharing in private personal clouds. In *Proc. of the IEEE International Conference on Cloud Computing*.

Lin, C. Y. and Tzeng, W. G. (2014). Game-theoretic strategy analysis for data reliability management in cloud storage systems. In *Proc. of the IEEE SERE*.

Naldi, M. and Mastroeni, L. (2013). Cloud Storage Pricing: A Comparison of Current Practices. In *Proc. of the ACM Workshop on HotTopiCS*.

Shin, J., Jo, M., Lee, J., and Lee, D. (2014). Strategic management of cloud computing services: Focusing on consumer adoption behavior. *IEEE Transactions on Engineering Management*, 61(3):419–427.

Wu, Z., Butkiewicz, M., Perkins, D., Katz-Bassett, E., and Madhyastha, H. V. (2013). Spanstore: Cost-effective geo-replicated storage spanning multiple cloud services. In *Proc. of the ACM SOSP*.