Summary Review Documentation for

“Inside Dropbox: Understanding Personal Cloud Storage Services”

Authors: I. Drago, M. Mellia, M. Munafo, A. Sperotto, R. Sadre, A. Pras

Reviewer #1

Summary: This paper presents an analysis of traffic patterns and user behaviors in using the Dropbox cloud storage service from 4 different vantage points (2 residential and 2 university)

Strengths:

- First of its kind as far as I know
- Timely, interesting problem
- Interesting reverse engineering of the protocols
- Nice anecdotes and tie in to recommendations for system design

Weaknesses: Tedious to read in many parts—related to the tediousness, some results felt redundant

Comments to authors: This is a very nice paper on a timely and interesting problem that will be relevant as the cloud storage adoption takes off even more. Kudos on a good job of both problem identification and execution! I thought the reverse engineering exercise was pretty neat that enabled the subsequent measurements and analysis.

It will be useful to cull out several paragraphs distributed throughout the paper (e.g., para 2 and 3 on page 2) into a separate related work section.

I have mostly minor notes and clarifications:

Abstract:

1. “a major portion of Internet” – can you have a graph or citation to substantiate this claim?

Intro:

1. “even to be detected” – this is slightly suspect since you manage to successfully reverse engineer it using only the domain names of the different types of servers. (and that was the crucial step, I don't think you needed the fine-grained protocol analysis for most of the result presented in here)
2. I don't see how [7] is studying storage dedup as such
3. penultimate para of intro – this is a bit content free, can you expand a bit on specific implications here?

Section 2

1. In 2.1 “updated via patches” – patches to software or did you just mean incremental updates?
2. In 2.1 I didn't follow this – are the Dropbox control servers also in EC2 or in some other hosting service? I thought it was strange to use storage in EC2 and compute elsewhere based on cost arguments (S3 seems more expensive than compute; e.g., Netflix uses EC2 for the control but uses CDNs for delivery)

Section 3

1. I didn't really like Section 3.3 – it is unfair to compare with SkyDrive/Google Drive etc that just only recently launched. This space may have been better utilized to show an overall traffic breakdown between cloud storage vs video vs HTTP etc? Also, it doesn't really add any real value to the paper
2. Also in 3.3 – Is iCloud most accessed because of mobile devices? 3. 3.3 – how do you identify distinct devices or device types?

Section 4

1. Having established that Dropbox is centralized, I am not sure what value 4.2.2 adds. Maybe you have something subtle in mind that is not coming across?
2. 4.4 “low throughput is striking” – I don't see why. My initial reaction was well, this reflects the ADSL rates!
3. “strongly penalizes” – I would tone this down a bit. It's not clear if these throughput bottlenecks actually impact the user experience because a lot of these transfers are likely to be in the background. And in fact, it's probably a good idea that it is not aggressive so it doesn't impact the user's foreground activity and interactive connections!
4. recommendation “3” on distributing the storage setup – it also raises a new set of issues surrounding synchronization/consistency etc that they have avoided this way?

Section 5

1. “tend to download more than” – this seems strange. Are people mostly looking at public files? Or is this a bias where you have more “mobile” devices that are used to read but not author content?
2. Also I was curious if the “abandoning” users are mostly on mobile devices

3. 5.3 – it might be useful to show some longitudinal trends on the namespace result?

Section 7

1. "considerable traffic volume" – you didn't really have a result to show this. you compared it to some past paper.

Reviewer #2

Summary: This paper presents an extensive measurement study of Dropbox. First, based on passive measurements from multiple sites, the authors show the volume of traffic to Dropbox is significant, both in terms of number of users and number of bytes/flows. The authors then analyze traffic to Dropbox's servers to highlight some performance bottlenecks and to study various characteristics of how users typically use Dropbox.

Strengths: First measurement study of Dropbox. Thorough job. Well-written.

Weaknesses: Main takeaways are unclear.

Comments to authors: This paper presents a pretty extensive measurement effort, and the authors have done a thorough job with their analysis. The paper is also well-written and easy to read. Well done!

My main concern about the paper though is that, other than making for interesting reading about the characteristics of Dropbox traffic, it is unclear how the takeaways here can be used by others. The main implications of your work seem to be for Dropbox and they already have all of this information based on passive measurements at their servers. Is there anything useful to take away from your work for other parties?

Moreover, the three takeaways for Dropbox (that you mention at the end of Section 4) also appear to be pretty weak. – the first (bundling smaller chunks) has already been incorporated by Dropbox, as you mention – the third (putting servers closer to users) is true for any service and surely Dropbox is not doing that with cost in mind. Plus, perhaps most of their current userbase is in the US? – the second (using delayed ACKs) may perhaps be a measure explicitly put in by Dropbox to rate limit customer traffic and thus reduce their bandwidth bills at their servers?

Overall, please devote more real-estate in the paper on how your measurements are useful for others and in what ways.

A couple of other comments: – Determining which flows represent "store" operations and which ones represent "retrieve" operations is key to your work. So, please add more detail in Section 2.4 about how you differentiate between these operations. Knowing this would be more useful than some of the details you present in Section 2.3. – I was surprised to see that you do not stress on the similarity of fractions of various types of users in the Home1 and Home2 datasets in Table 4. The confirmation between the two datasets is great in showing that your inferences of the various types of users is not just a characteristic of a particular dataset.

Reviewer #3

Summary: This paper presents a traffic-driven characterization of the Dropbox service. The authors begin with background on

the service. To reverse-engineer encrypted communications, they overwrite the SSL cert in-memory in the client, and present some details on the protocol including notifications, meta-data transfers, and the actual data transfers.

They use four datasets, two collected at a European nation-level ISP and two at universities between March 25 and May 5 2012 using Tstat.

They compare popularity of different cloud storage services (6-10% of home users employ them, iCloud gets contacted by most devices, but Dropbox dominates content by over an order of magnitude, making up 4% of all traffic of one campus, or 1/3 of the site's YouTube traffic), break down bytes and flows by purpose (client, web, api, storage, etc), server locations (all US, if I understand it right). RTTs are relatively stable around 100s for storage regardless of location on the planet, whereas control varies widely, indicating centralized infrastructure in the US.

Throughput is generally poor, no more than 500kbps for store or retrieve regardless of the number of chunks per batch, though larger chunk sizes help.

Next are usage patterns in terms of occasional and heavy use, up/down/up-downloaders, and number of devices. Diurnal usage is visible, which is unsurprising.

The authors will make the measurement datasets available.

Strengths: Thorough analysis of an interesting new service, nicely executed and presented.

Weaknesses: The paper presents a flood of details about the service. Every now and then I would have liked to come up for air and see some high-level thinking. (See comments.) I felt the throughput analysis was lacking.

Comments to authors: Nice IMC paper overall, thank you. As I mentioned above, I would have liked to see some higher-level analyses. Two immediately come to mind. First, from file sizes, up/download behavior and sharing, can you draw any conclusions about the legality of content shared on Dropbox? Second, from the storage server IP address use, what can you infer about server-side data locality and replication management? Do a single user's storage flows frequently end up on different servers or not?

In the throughput analysis too many things got mixed up. We have server location (fuzzily somewhere in the US), TCP-level RTTs (not shown in relation to chunks/batches), TCP losses (not discussed), varying chunk sizes, and – potentially – varying number of TCP flows for a transport. I'm not walking away with a clear picture of what really is influencing throughput.

I would have liked to see more analysis of the PUSH flags. We have to take your word for them being meaningfully aligned with chunks, but an analysis would be nice.

A bunch of comments:

In 4.2.1 it's not clear whether you mean that regardless of geographical location the same IP addresses get returned for metadata, notification, and storage flows, or only the former two (as you call them out).

The US is a big country – could you break down location of the servers within the US? This would help, as Amazon has at least two major datacenter locations there.

4.2.2: by "access technology", do you mean web API vs client? Or do you mean that on PlanetLab you conducted your own RTT measurements? Or, do mean ADSL vs Fiber etc?

4.2.2: why are the RTTs for storage noticeably better for Campus 1?

4.4 is intriguing, but I wonder what's the number of TCP connections for a single file's storage/retrieval operation? If it's obvious that it's just one, please say that. I doubt it is, and if a transfer *generally* requires multiple connections, then beyond simple RTT the per-flow three-way handshakes and window behaviors matter.

Reviewer #4

Summary: The paper presents a detailed study of the protocols, traffic, and performance of Dropbox usage. The findings, while not terribly surprising, are in-depth and provide a detailed look at many aspects of the service. The analysis covers significant breadth, going as far as offering suggestions on how to improve e2e performance for Dropbox users.

Strengths: Seems to be the first to study dropbox. Methodology is rather interesting, and quite comprehensive. Results have significant breadth, and are likely to be useful in future studies.

Weaknesses: Not too many weaknesses really. The biggest is perhaps the lack of any genuinely surprising findings, and no real topical questions asked, other than a general measurement study as the goal.

Comments to authors: I enjoyed reading this paper. Some of the findings, while intuitive, are quite nice to have. Actual numbers of traffic load for dropbox users, impact of RTT on performance, usage of C+C servers vs storage servers, load balancing across S3 instances. All these tidbits build up a nice measurement paper that is likely to have significant impact on future studies of personal cloud storage systems.

Reviewer #5

Summary: This paper presents a first study on Dropbox usage. The authors analyzed data collected from two university campuses and two PoPs of a large ISP for 42 days. The study makes a few interesting findings about Dropbox, including its penetration rate and its performance bottleneck.

Strengths: A timely study with interesting findings. Well-written.

Weaknesses:

- User privacy issues are not addressed.
- Not clear whether the results are representative of world-wide Dropbox usage.
- Data analysis is not very deep.

Comments to authors: Very nice work. As Dropbox is becoming more and more popular, it is interesting to understand how it works and how many users are using it. Your study answers some of these questions, and I learned a lot from reading it. Thanks.

Section 3. Since you used DPI, and customers in your ISP are assigned static IP addresses, it'd be useful to discuss how you addressed user privacy issues. This type of measurement work makes me nervous sometime, as much user privacy may be divulged during the study.

Since your data were collected from two European countries, they may not represent the worldwide Dropbox usage patterns. It's not clear whether the results such as the penetration rate, flow size, etc. are also applicable to other parts of the world. Although it's difficult to address this limitation without obtaining additional data sources, it'd be useful to clarify in your analysis which parts of results are likely to be location dependent, and which parts are not.

4.2.2. "This is a clear indication that a single data-center is used." → Well, it might be the case that for the part of the European users, all data are served from one data-center. It may not be the case that there is one datacenter for all Dropbox users in the world. Measurements from other parts of the world are needed to back up this claim.

4.4 I like that you hypothesized several factors that could lead to low throughput of some flows. However, it'd be useful to have a deeper analysis that shows how each factor, chunk size, sequential ack, large rtt, and last hop bottleneck contributes to the issue. With this analysis, your suggested improvements will be more convincing. Presently, what caused the low throughput is not made clear. Therefore, it's not clear whether your suggested improvements will work.

5.3 I am surprised that the namespace data are not encrypted!

Response from the Authors

We would like to sincerely thank the reviewers and our shepherd for all their valuable feedback. The majority of the comments have been included in the camera-ready, resulting in a significantly improved version of the paper.

The sections analyzing the system throughput (Sec. 4.4 and Sec. 4.5) have been deeply revised to make our claims clearer. The reviewers commented that other factors besides the application-layer sequential acknowledgments and the TCP slow start-ups could influence our results (such as the access technology and possible TCP congestion). To address these comments, we have refocused our analysis on the campus networks, where the access technology (e.g. ADSL) is not a bottleneck. We have also improved our methodology to estimate the system throughput and produced new figures to help the readers to understand our message.

Extra results have been added 1) on the maximum throughput achievable when considering the RTT observed in our datasets and 2) on the minimum duration of flows per number of chunks (Fig. 9 and Fig. 10). Besides that, we have included new data collected after Dropbox released a bundling scheme in their client, in order to quantify the effects of such a solution in the system throughput. With this comparison, we highlight the effects (bottlenecks) of some design decisions on cloud storage applications.

The reviewers requested a clearer exposition of our main contributions and high-level analyses of both the results and the take-

away messages. We have extended the Introduction stating the relevance of our results to others. To remove possible excessive details and make the text more dynamic, Sec. 2 and Sec. 5 have been summarized. New sub-sections at the end of Sec. 4 and Sec. 5 have been added, presenting the high-level implications of our results.

The reviewers asked for more information about our methodology, such as on how to tag store and retrieve flows and on the relation between chunks and TCP segments with the PSH flag set. The final revision includes an appendix presenting these details. Some intermediate results validating our methodology have been included in the appendix as well.

Finally, other clarifications (e.g. on privacy issues of our data capture infra-structure, worldwide applicability of our results etc.) and several minor corrections were incorporated, according to the reviewers' comments.