Inside Dropbox: Understanding Personal Cloud Storage Services

Drago, Mellia, Munafo, Sperotto, Sadre, Pras

University of Twente

What they've done

- First to study Dropbox
- Characterize the workload typical users have
- Find bottlenecks

How Dropbox works

- Two main types of Dropbox servers:
  - Control
  - Data storage
- Dropbox owns the control servers
- Outsources the data storage servers to Amazon.

Dropbox communication

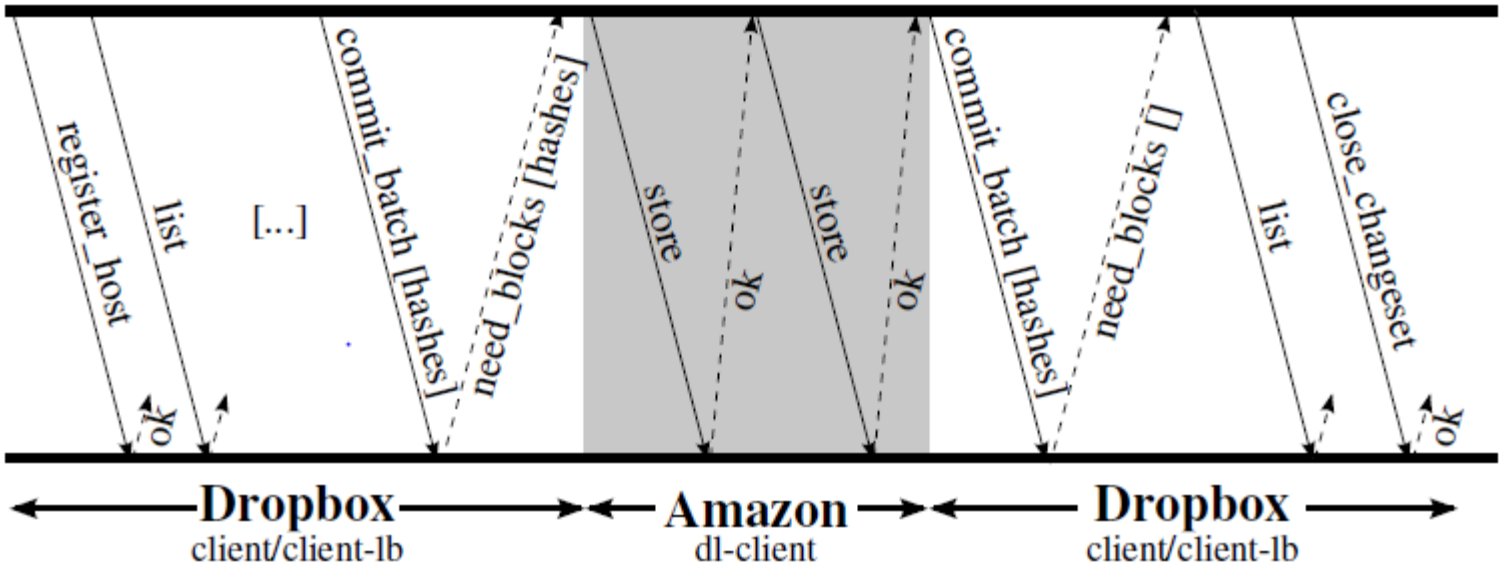


Figure 1: An example of the Dropbox protocol.

Communication specifics

- Client interacts with control to get working orders
- Routes to storage to store directly
- ACK-ing is sequential in storage

Control Servers

- 3 types:
  - Notification
  - Meta-data administration
  - System-log servers
- They talk about 1 and 2, but not really 3…

Control Servers

- Dropbox client keeps open continuous TCP connection
- Connects to notification server about changes being pushed to other machines
- These connections not encrypted.

Metadata servers

- Communicate with several short TLS connections
- Done to handle batches sent and received.
- Used mainly during file transfer.

Data servers

- Over 500 distinct domain names point to Amazon servers
- Two basic types of commands: store and retrieve
- In new version of Dropbox, there is store_batch and retrieve_batch

Testing protocols

- Basic Idea:
  - Analyze traffic from two university campuses
  - Analyze traffic from two different residential networks
  - Devise a methodology for monitoring cloud storage traffic

How they monitored traffic

- Most client communications are encrypted with TLS
- Dropbox client was told to use a Squid server they set up
- Module SSL-bump terminated SSL connections and saved decrypted traffic flows.
- Then they mirrored the required Dropbox certificates to prevent server ending communication

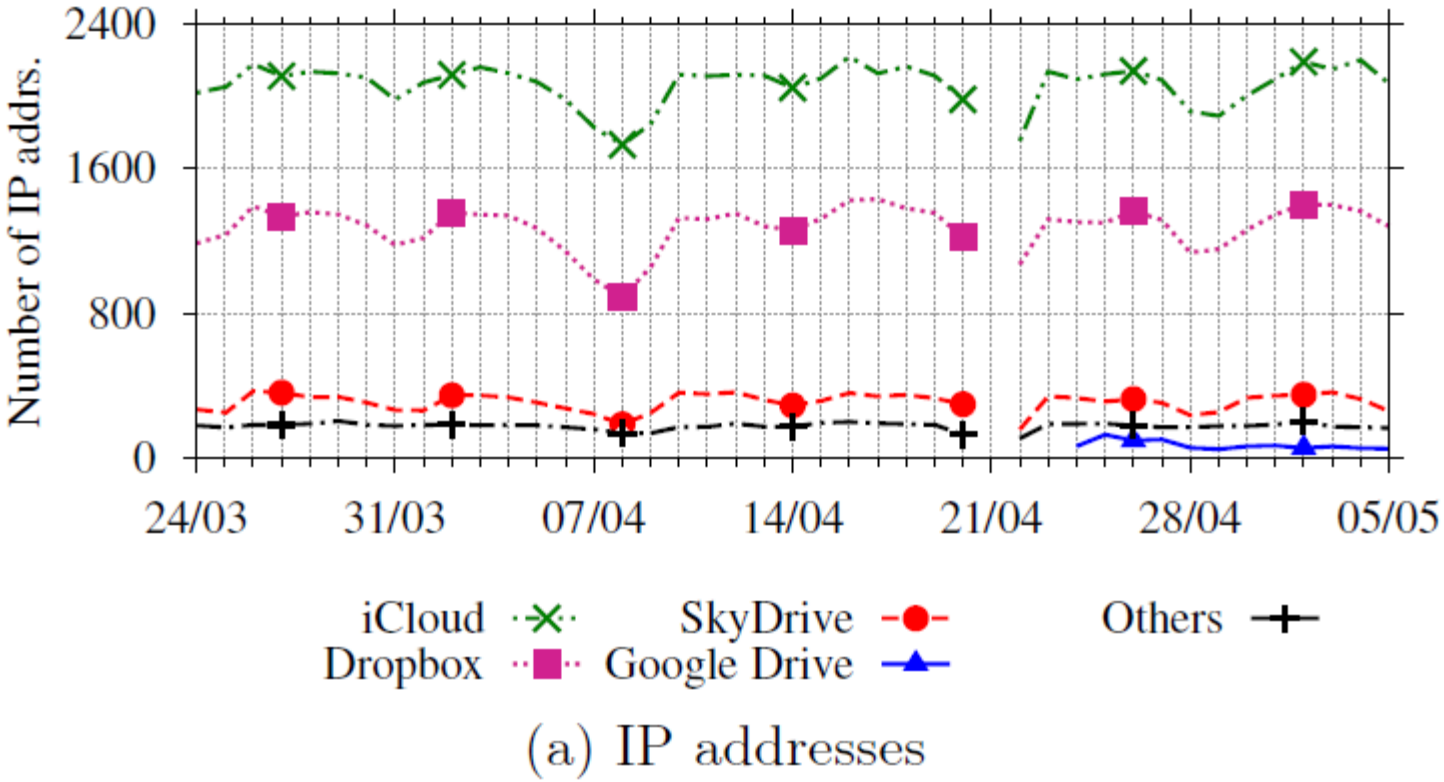
Tstat and key data points

- Tstat
  - Exposes client and server Ips
  - Amount of exchanged data
  - Retransmitted segments
  - RTT
  - Number of TCP segments that had PSH flag
- Used Tstat at 3 vantage points and collected data for 42 straight days

College 1 and College 2

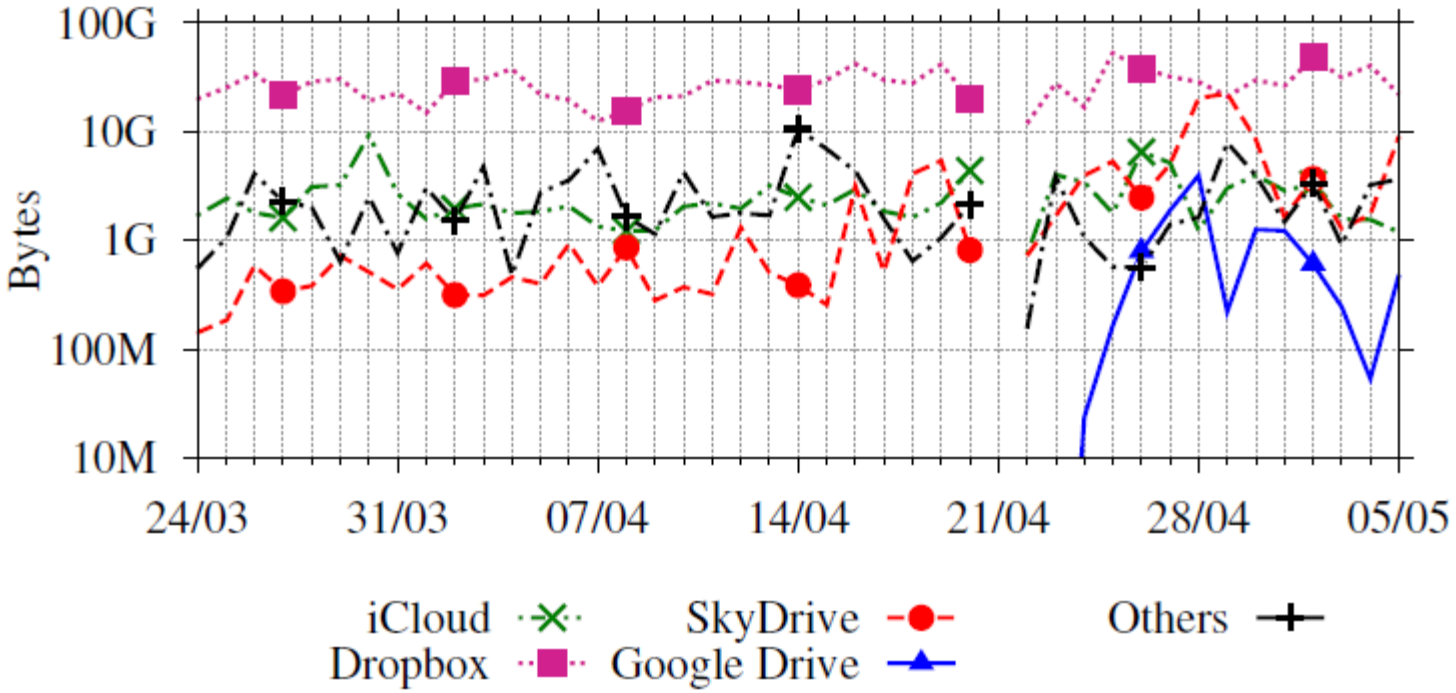
- College 1 is a wired main campus network
- Most of the machines are used for either research or administration (akin to Tlab or Wilk machines)
- College 2 monitors the peripheral nodes of a college network
- Dorms, Norris, etc.

Home 1



(a) IP addresses

Home 1 for Data Volume



(b) Data volume

Home 1 implications

- Dropbox is more popular and more widely used
- Makes sense, although they (rightly) point out that Google Drive didn't really exist in its final form
- How useful is this data?

Youtube and Dropbox comparisons

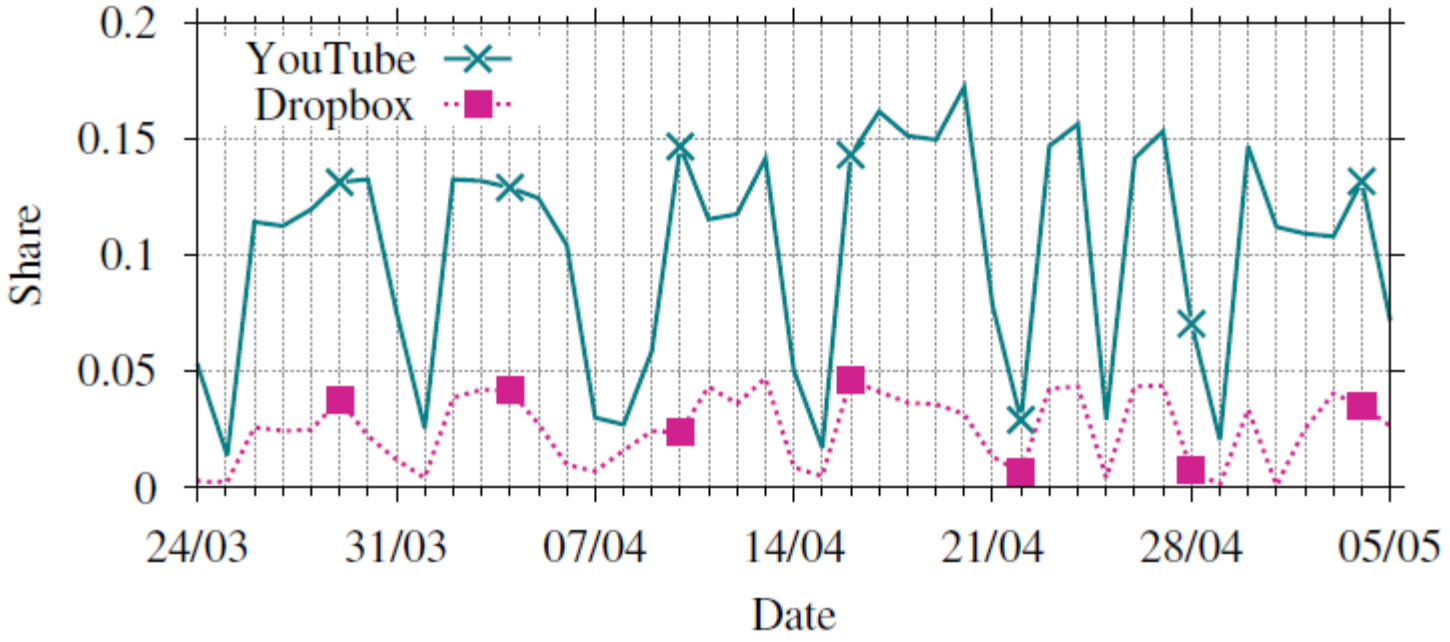


Figure 3: YouTube and Dropbox in *Campus 2*.

How Dropbox manages data

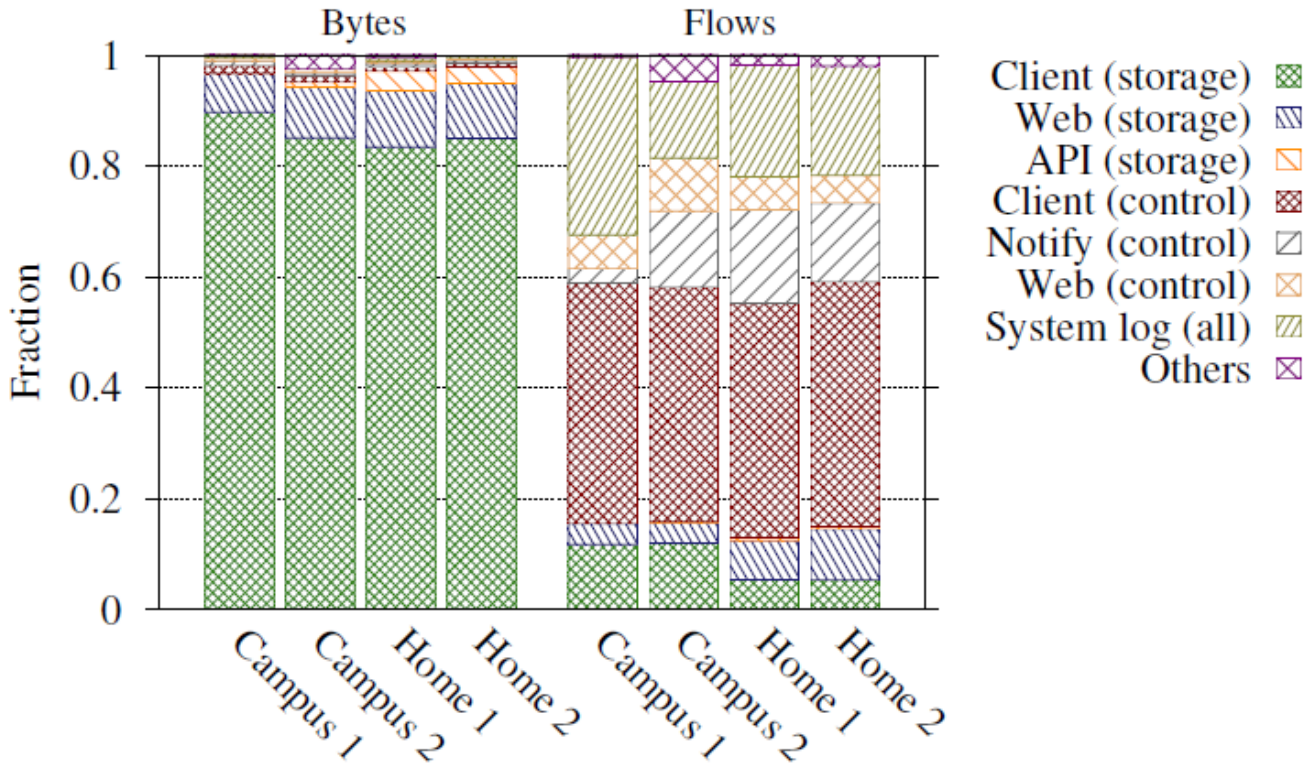


Figure 4: Traffic share of Dropbox servers.

YouTube and Dropbox

- Exists to show how popular Dropbox is compared to YouTube on college campuses.
- Also used to show weekly network traffic

Differences between networks

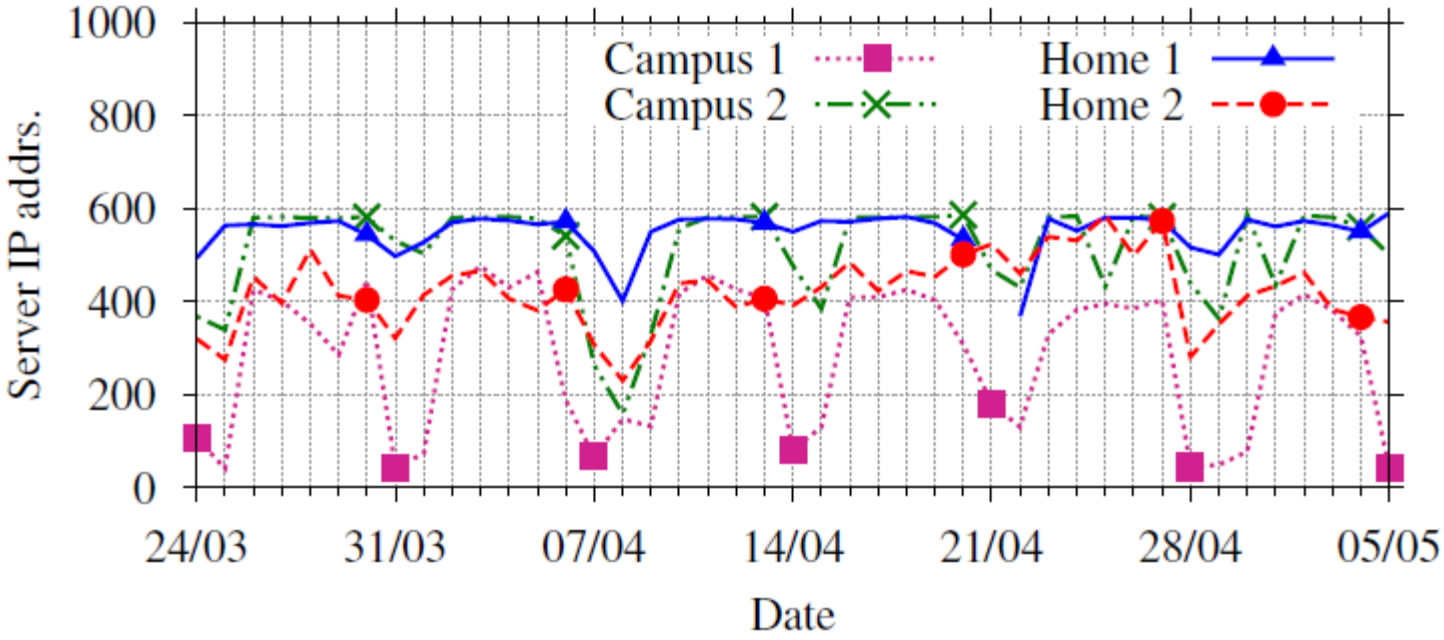
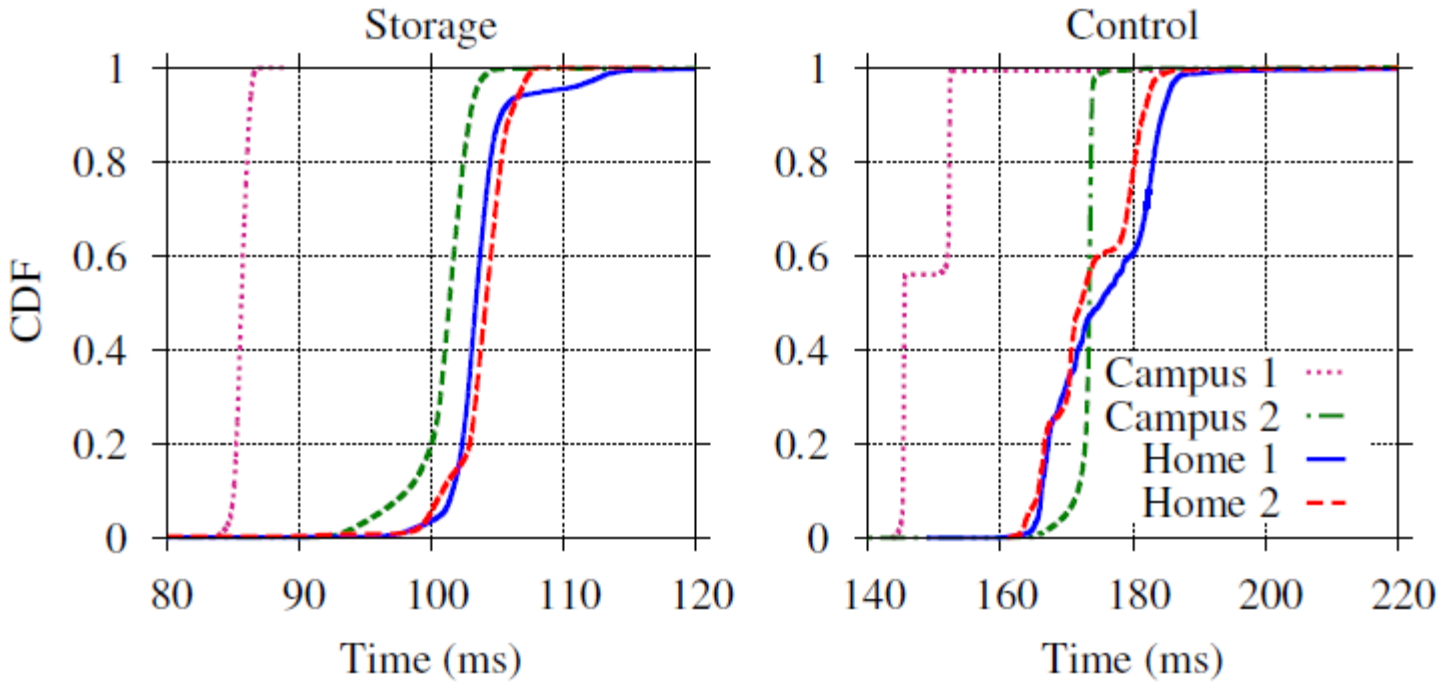


Figure 5: Number of contacted storage servers.

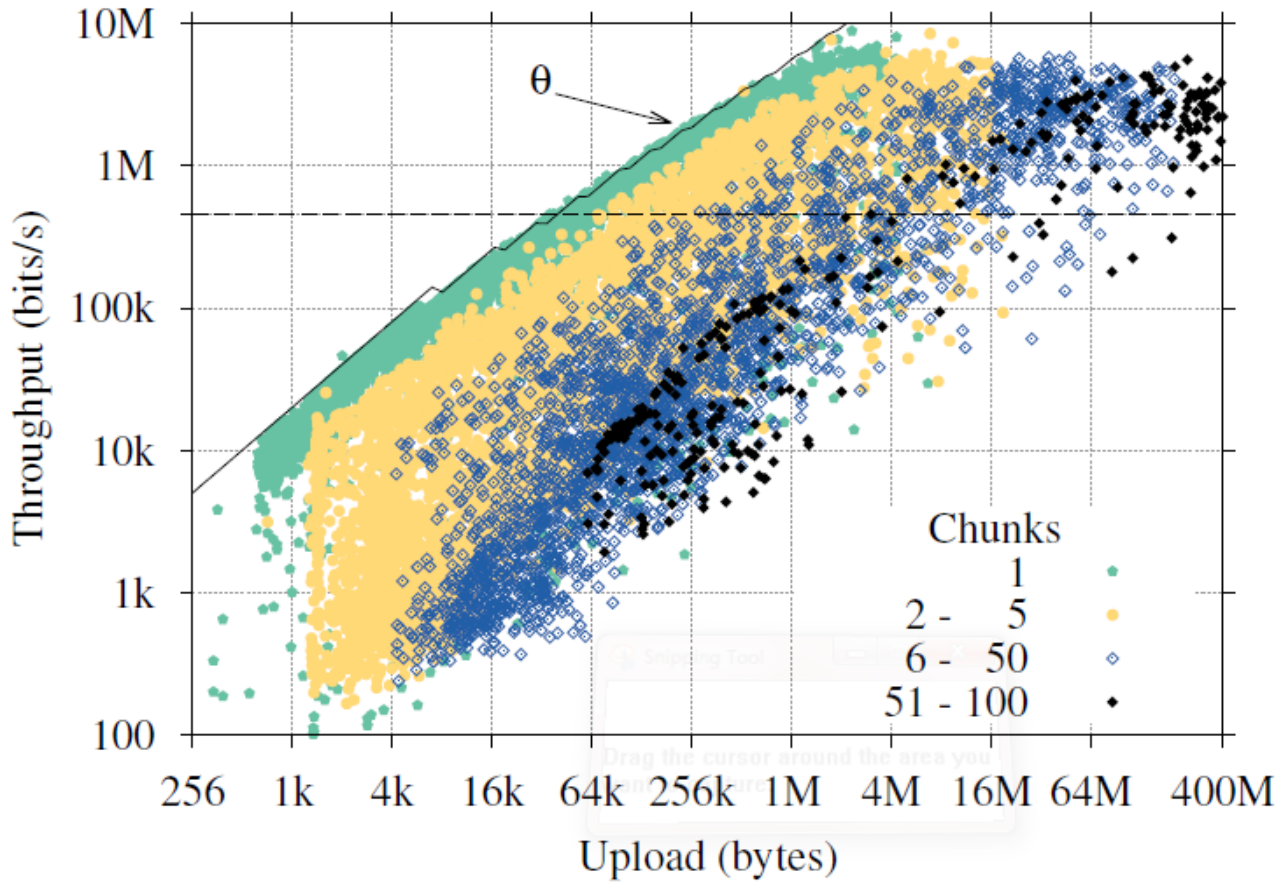
Network differentiation

- Implication: Better network, better performance
- Other implication: Networks closer to Dropbox servers operate really fast.

Differences between Storage and Control

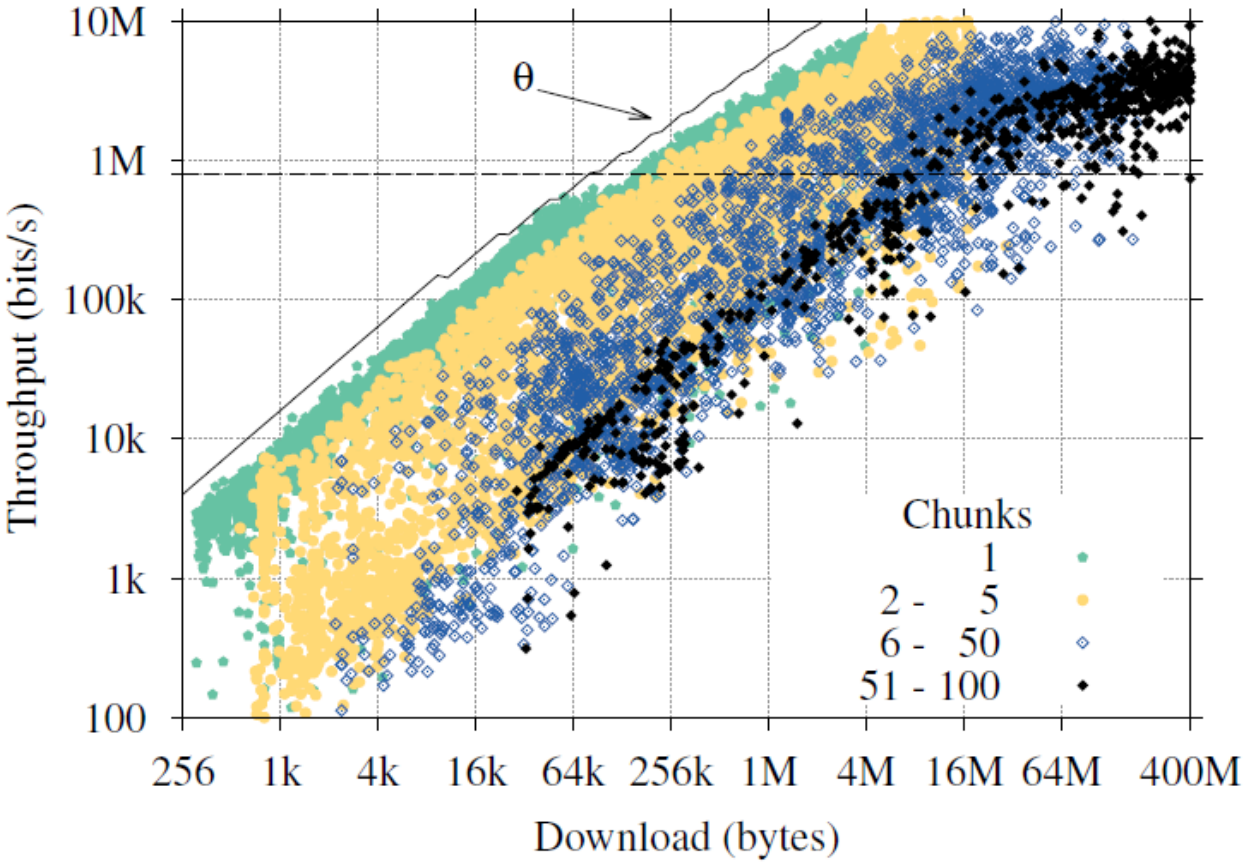


Storage Throughput (upload)



(a) Store

Storage Throughput (download)



(b) Retrieve

Their takeaways

- Dropbox has more traffic because Google Drive and SkyDrive weren't launched.
- Dropbox performs better when it has storage servers closer to its clients
- Dropbox's overall throughput is pretty low
- Most of their recommendations were fixed with the update

Dropbox User Profiles

- Different people use Dropbox for different things
- Home users tend to download more
- 40% of college users have at least 5 shared folders

Discussion

- A lot of their results seem kind of basic
- They say that one of the big things they're doing is devising a method for monitoring cloud storage traffic, and they barely mention it.
- They say it's an analysis, but their data still ends up being more of a comparison

Discussion, cont'd

- Are their recommendations any different than what you would tell any other company?
- Is the YouTube comparison at all relevant or necessary?
- If Dropbox is global, why is improving their network topology on a local scale important (as opposed to global optimization)?
- Why don't they talk about their methodology???