UNIVERSITY OF TWENTE.

POLITECNICO
DI TORINO

Inside Dropbox: Understanding Personal Cloud Storage Services

→ *Idilio Drago*
→ Marco Mellia
→ Maurizio M. Munafò
→ Anna Sperotto
→ Ramin Sadre
→ Aiko Pras

IMC 2012 – Boston

Motivation and goals

- **Personal cloud storage** services are gaining popularity
- Dropbox:
  - *“the largest deployed networked file system in history”*
  - *“over 50 million users – one billion files every 48 hours”*

Motivation and goals

- **Personal cloud storage** services are gaining popularity
- Dropbox:
  - *“the largest deployed networked file system in history”*
  - *“over 50 million users – one billion files every 48 hours”*
- **Little public information about the system**
  - How does Dropbox work?
  - How is the system performing?
  - Are there typical usage scenarios?

How does Dropbox work?

- Some public information about Dropbox
  - **Native client**, Web interface, LAN-Sync, etc
  - Storage in **Amazon S3**
  - Files are split in **chunks of up to 4MB**
  - Delta encoding, **encrypted** communication

How does Dropbox work?

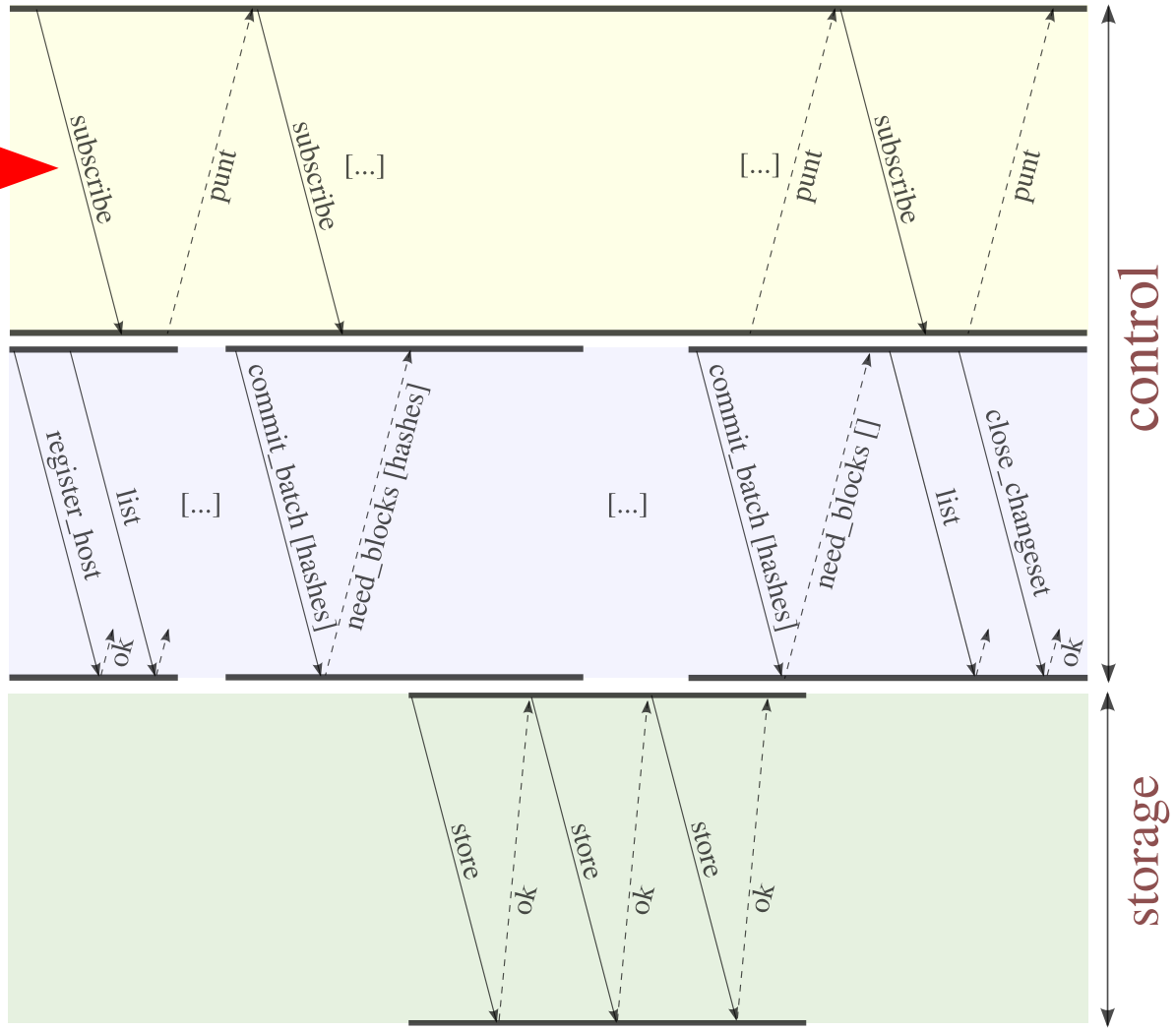
- Some public information about Dropbox
  - **Native client**, Web interface, LAN-Sync, etc
  - Storage in **Amazon S3**
  - Files are split in **chunks of up to 4MB**
  - Delta encoding, **encrypted** communication
- To understand the client protocol
  - **MITM** against our own client
  - Squid proxy, SSL-bump and a self-signed CA certificate
  - **Replace a trusted CA certificate in the heap at run-time**
- Proxy logs and decrypted packet traces

How does Dropbox (v1.2.52) work?

- Clear separation between **storage** and meta-data/client **control**
- Sub-domains identifying parts of the service
  - `notifyX` $\rightarrow$ client notification
  - `client-lb/clientX` $\rightarrow$ meta-data control
  - `dl-clientX` $\rightarrow$ Client storage
  - etc
- **HTTP/HTTPs** in all functionalities

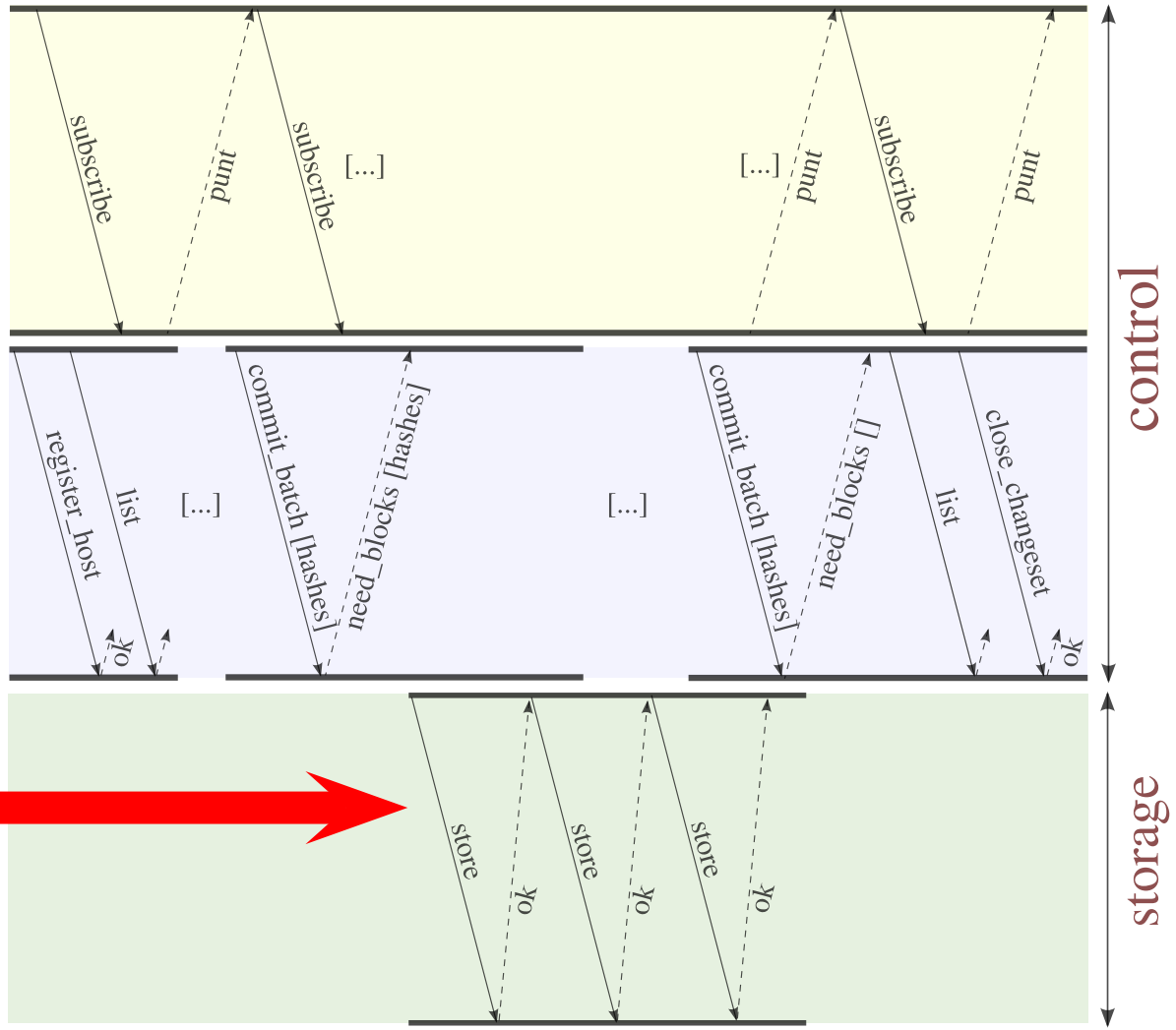
How does Dropbox (v1.2.52) work?

- Notification
  - Kept open
  - **Not encrypted**
  - **Device ID**
  - Namespaces



How does Dropbox (v1.2.52) work?

- Storage
  - ***Retrieve* vs. *Store***
  - **Sequential acks**
  - Amazon EC2



Methodology

- Rely on **Tstat**[1] to export **layer-4 flows**
- Isolate Dropbox flows
  - **DN-Hunter**[2], TSL/SSL certificates, IP addresses

1 – http://tstat.polito.it/

2 – DNS to the rescue: Discerning Content and Services in a Tangled Web

Methodology

- Rely on **Tstat**[1] to export **layer-4 flows**
- Isolate Dropbox flows
  - **DN-Hunter**[2], TSL/SSL certificates, IP addresses
- **Device IDs** and namespace IDs
- Use the knowledge from our own decrypted flows to
  - **Tag Dropbox flows** – *e.g.* as *store* and *retrieve* flows
  - Estimate the **number of chunks** in a flow

1 – http://tstat.polito.it/

2 – DNS to the rescue: Discerning Content and Services in a Tangled Web

Datasets

	Type	IP Addrs.	Dropbox		
			Flows	Vol. (GB)	Devices
Campus 1	Wired	400	167,189	146	283
Campus 2	Wired/Wireless	2,528	1,902,824	1,814	6,609
Home 1	FTTH/ADSL	18,785	1,438,369	1,153	3,350
Home 2	ADSL	13,723	693,086	506	1,313
Total			4,204,666	3,624	11,561

- 42 consecutive days in March and April 2012
  - 4 vantage points in Europe

Datasets

	Type	IP Addrs.	Dropbox		
			Flows	Vol. (GB)	Devices
Campus 1	Wired	400	167,189	146	283
Campus 2	Wired/Wireless	2,528	1,902,824	1,814	6,609
Home 1	FTTH/ADSL	18,785	1,438,369	1,153	3,350
Home 2	ADSL	13,723	693,086	506	1,313
Total			4,204,666	3,624	11,561

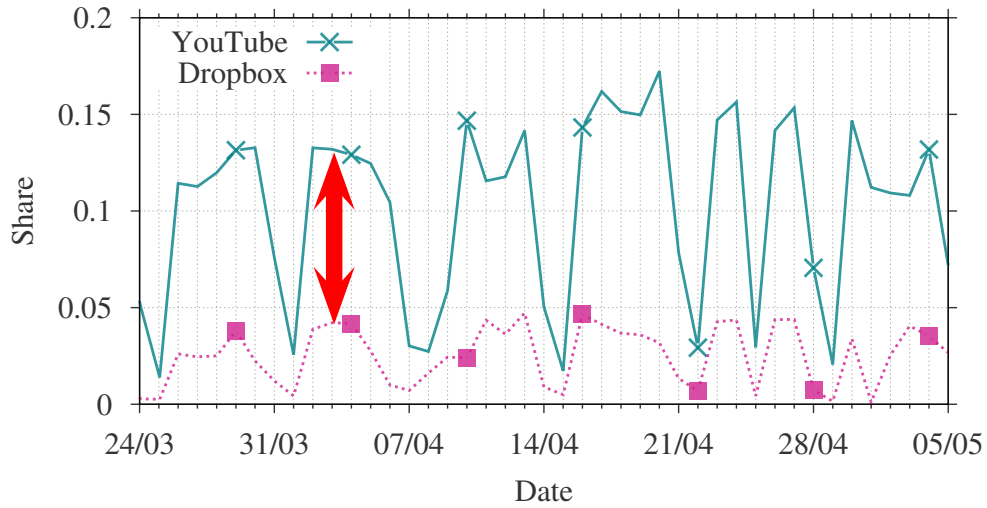
- 42 consecutive days in March and April 2012
  - 4 vantage points in Europe
  - Number of IP addresses in home probes $\approx$ installations

Datasets

	Type	IP Addrs.	Dropbox		
			Flows	Vol. (GB)	Devices
Campus 1	Wired	400	167,189	146	283
Campus 2	Wired/Wireless	2,528	1,902,824	1,814	6,609
Home 1	FTTH/ADSL	18,785	1,438,369	1,153	3,350
Home 2	ADSL	13,723	693,086	506	1,313
Total			4,204,666	3,624	11,561

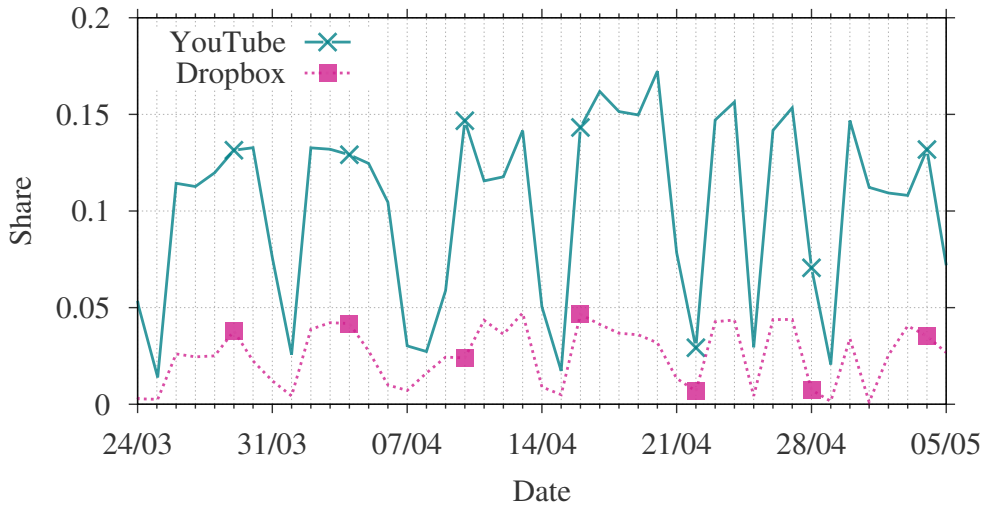
- 42 consecutive days in March and April 2012
  - 4 vantage points in Europe
  - Number of IP addresses in home probes $\approx$ installations
  - 11,561 unique devices
- 2nd capture in Campus 1 in June 2012

How much traffic to personal cloud storage?



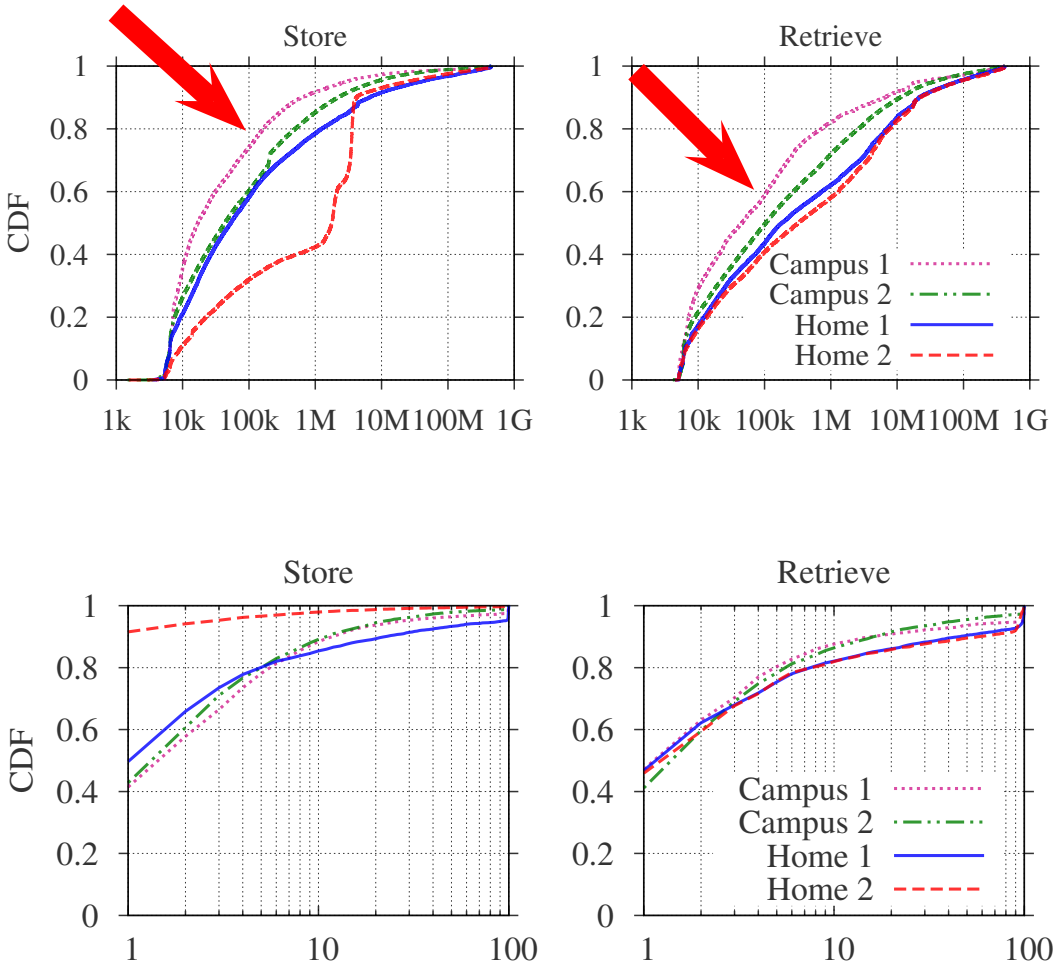
- Equivalent to **1/3 of YouTube volume at campus**

How much traffic to personal cloud storage?



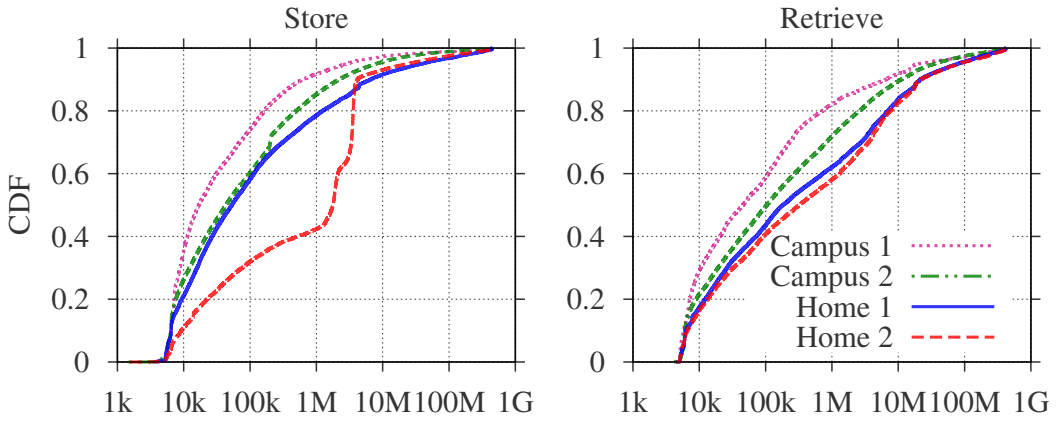
- Equivalent to 1/3 of YouTube volume at campus
- **Popularity: 6–12% adoption in home networks**
- **90% of the Dropbox traffic is from the native client**

How does the storage traffic look like?

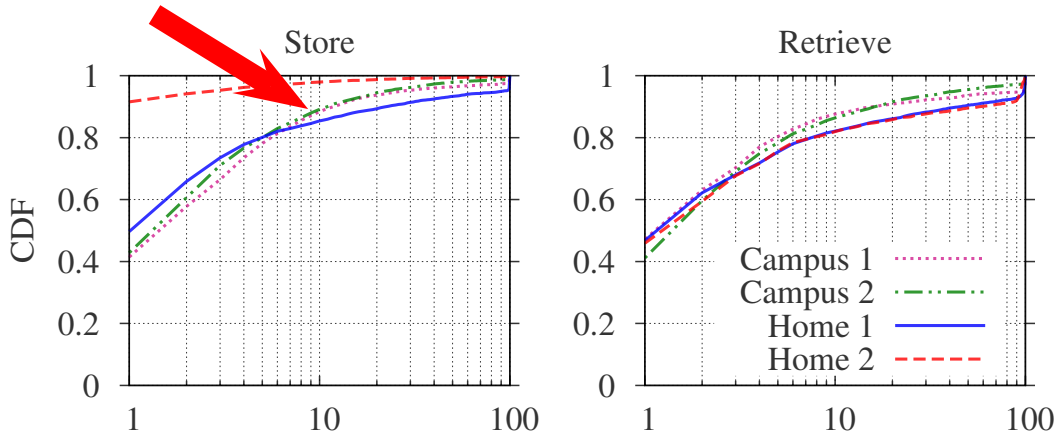


- Flow size
  - ***Store*: 40%–80% < 100kB**
    - → Deltas, small files
  - **Larger *retrieve* flows**

How does the storage traffic look like?

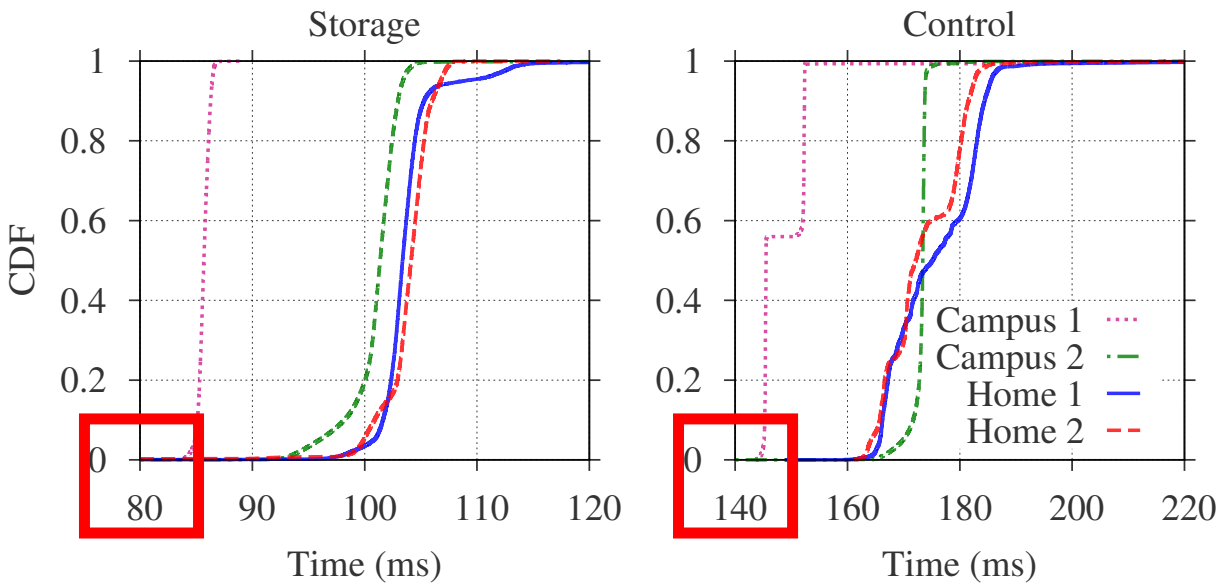


- Flow size
  - ***Store*: 40%–80% < 100kB**
    - → Deltas, small files
  - **Larger *retrieve* flows**



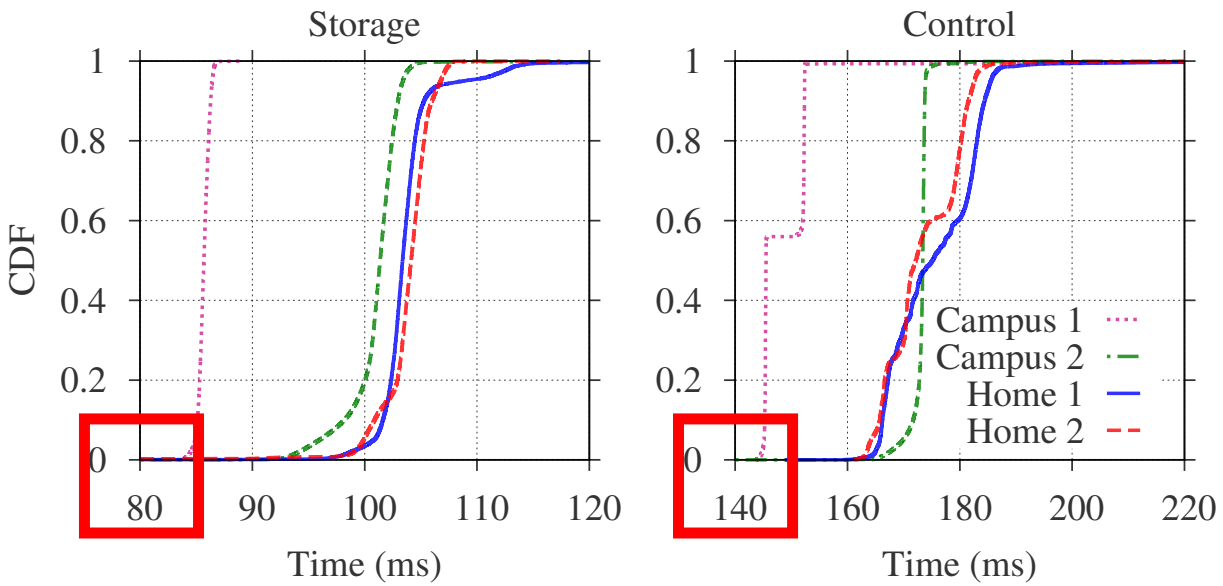
- Chunks per flow
  - **80% ≤ 10 chunks**
  - **Remaining: up to 100**
    - → Limited by the client

Where are the servers located?



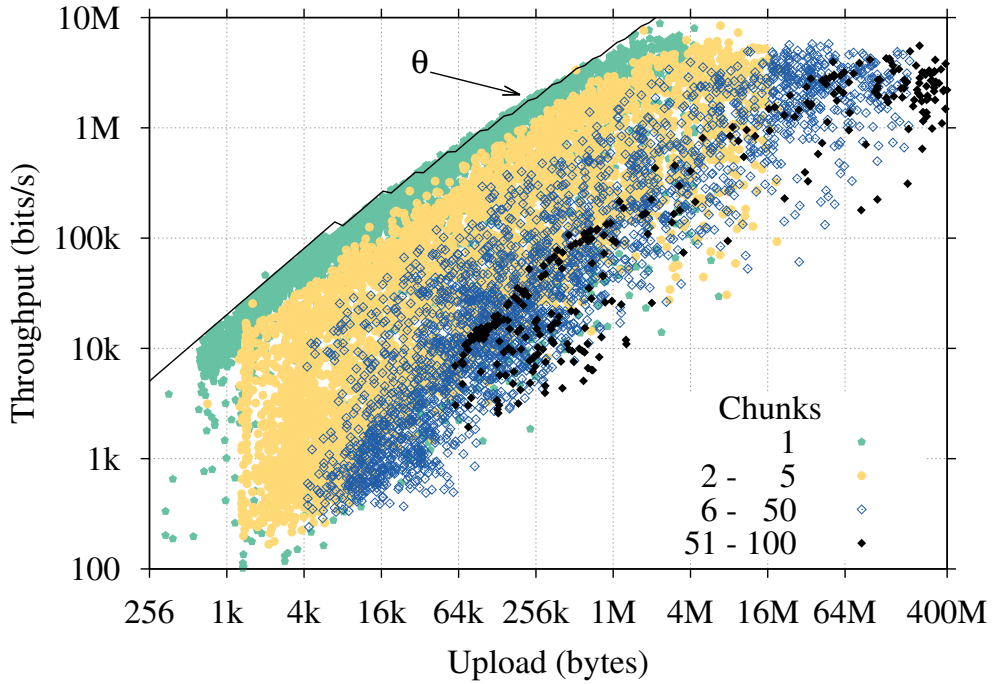
- Minimum RTT per flow $\rightarrow$ stable over 42 days
- PlanetLab experiments $\rightarrow$ **the same U.S. data-centers worldwide**

Where are the servers located?



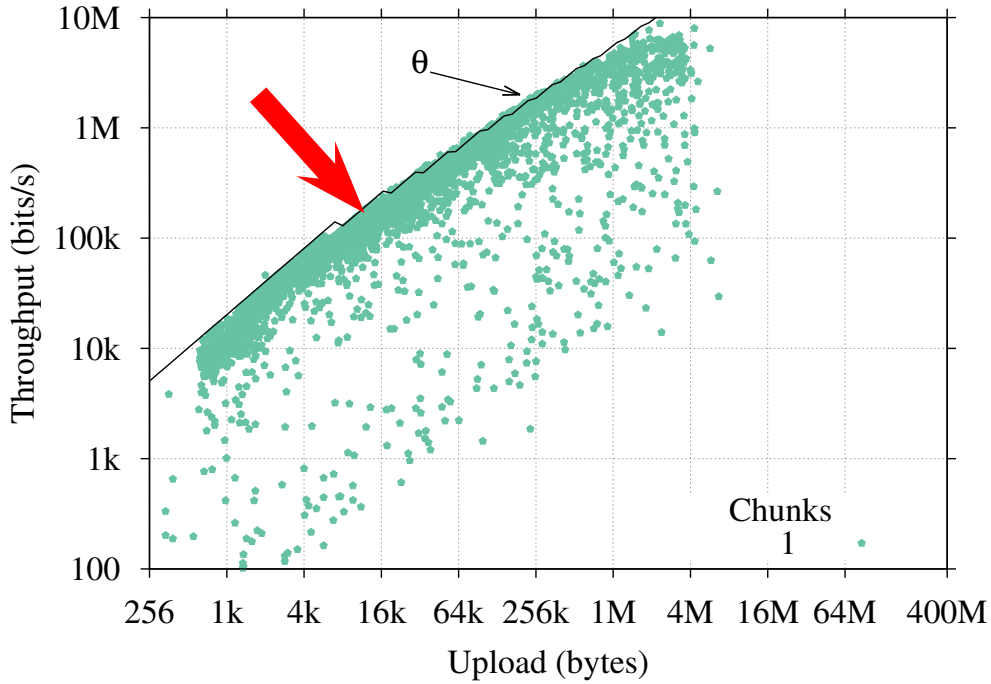
- Minimum RTT per flow $\rightarrow$ stable over 42 days
- PlanetLab experiments $\rightarrow$ **the same U.S. data-centers worldwide**
- *"less than 35% of our users are from the USA"*

How is the performance far from the data-centers?



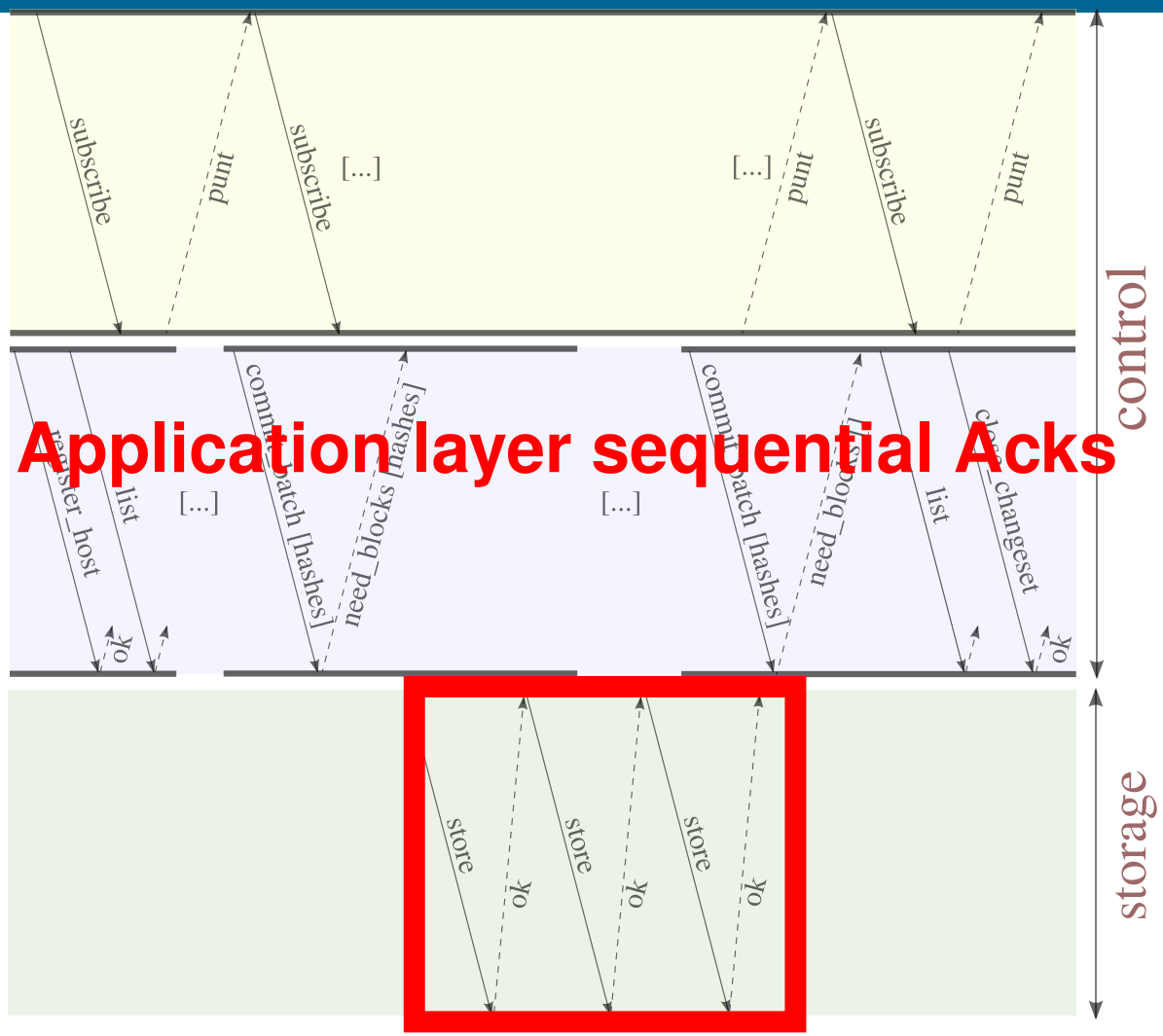
- Storage throughput in campuses
- **Most flows experience a low throughput**

How is the performance far from the data-centers?

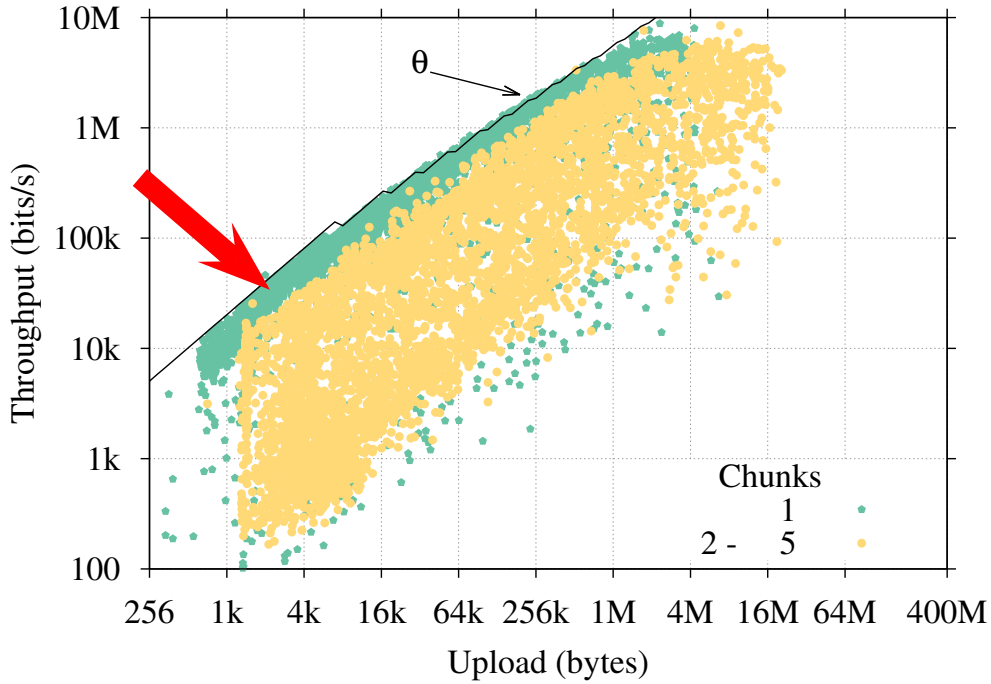


- Flows carrying 1 chunk
  - Size $\leq$ 4MB, RTT $\approx$ 100ms
  - Most of them finish in **TCP slow-start**

How is the performance far from the data-centers?

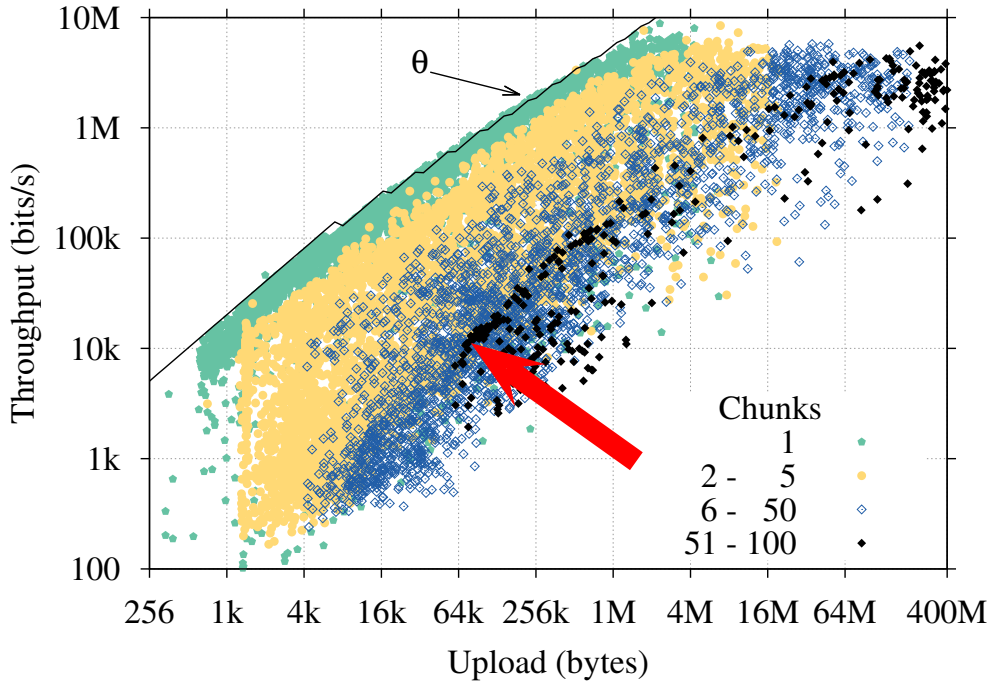


How is the performance far from the data-centers?



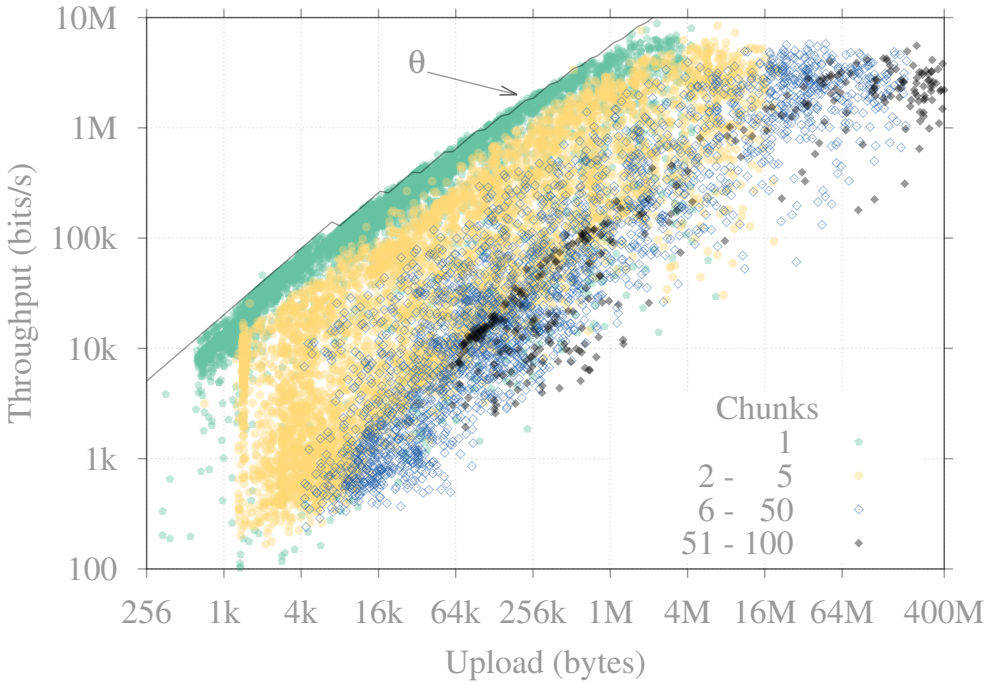
- Flows carrying several chunks
  - **Pause between chunks** → RTT and client/server reaction

How is the performance far from the data-centers?



- Flows carrying several chunks
  - **Pause between chunks** → RTT and client/server reaction
  - Transferring **100 chunks takes more than 30s**
  - RTTs → 10s of inactivity

How is the performance far from the data-centers?



- **Delaying acknowledgments**
- **Bundling chunk** $\rightarrow$ recently deployed
- **Distributing servers** $\rightarrow$ storage traffic is heavy!

How much improvement from chunk bundling?

- New version released on Apr 26th (v1.4.0)
  - Small chunks are bundled together

	Mar/Apr		Jun/Jul	
	Median	Average	Median	Average
	Flow size			
Store	16.28kB	3.91MB	42.36kB	4.35MB
Retrieve	42.20kB	8.57MB	70.69kB	9.36MB
	Throughput (kbits/s)			
Store	31.59	358.17	81.82	552.92
Retrieve	57.72	782.99	109.92	1293.72

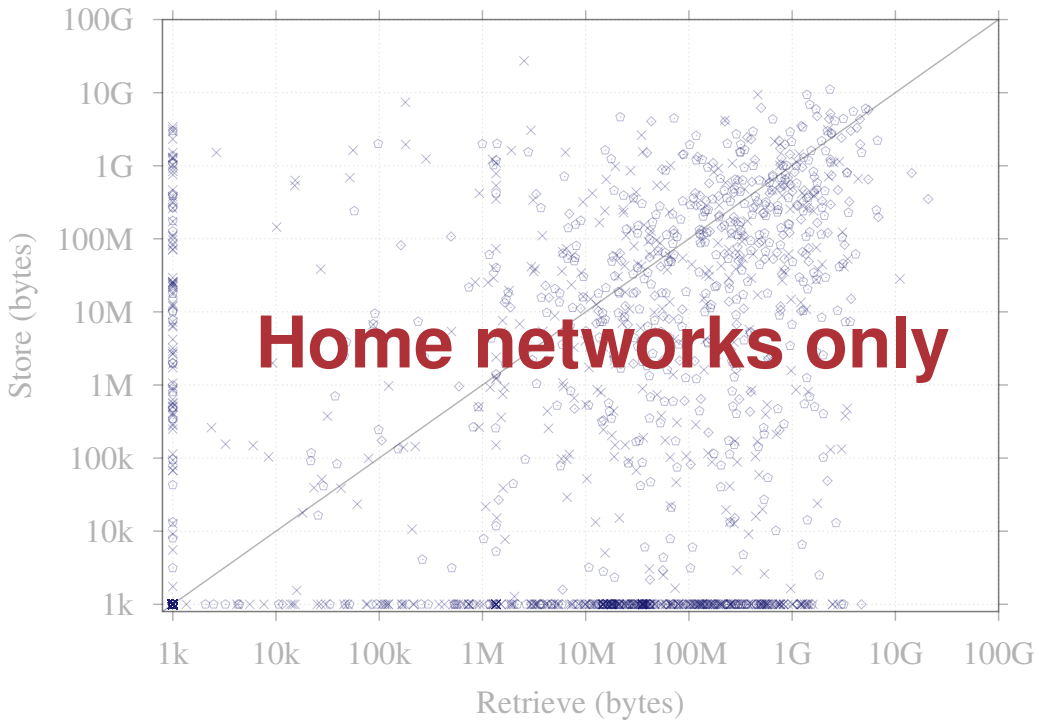
How much improvement from chunk bundling?

- New version released on Apr 26th (v1.4.0)
  - Small chunks are bundled together

	Mar/Apr		Jun/Jul	
	Median	Average	Median	Average
	Flow size			
Store	16.28kB	3.91MB	42.36kB	4.35MB
Retrieve	42.20kB	8.57MB	70.69kB	9.36MB
	Throughput (kbits/s)			
Store	31.59	358.17	81.82	552.92
Retrieve	57.72	782.99	109.92	1293.72

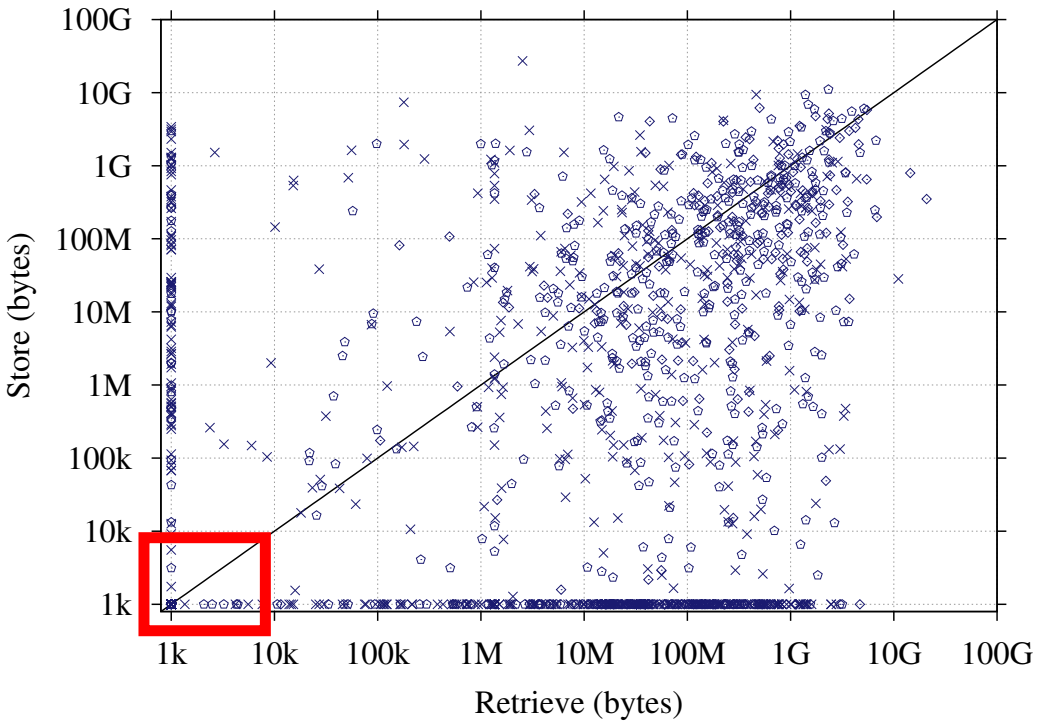
- **Less small flows** $\rightarrow$ less TCP slow-start effects
- Average **throughput** is up to **65% higher**

Are there typical usage scenarios?



- **More downloads** $\rightarrow$ download/upload ratio up to **2.4**
- What about download/upload per user?

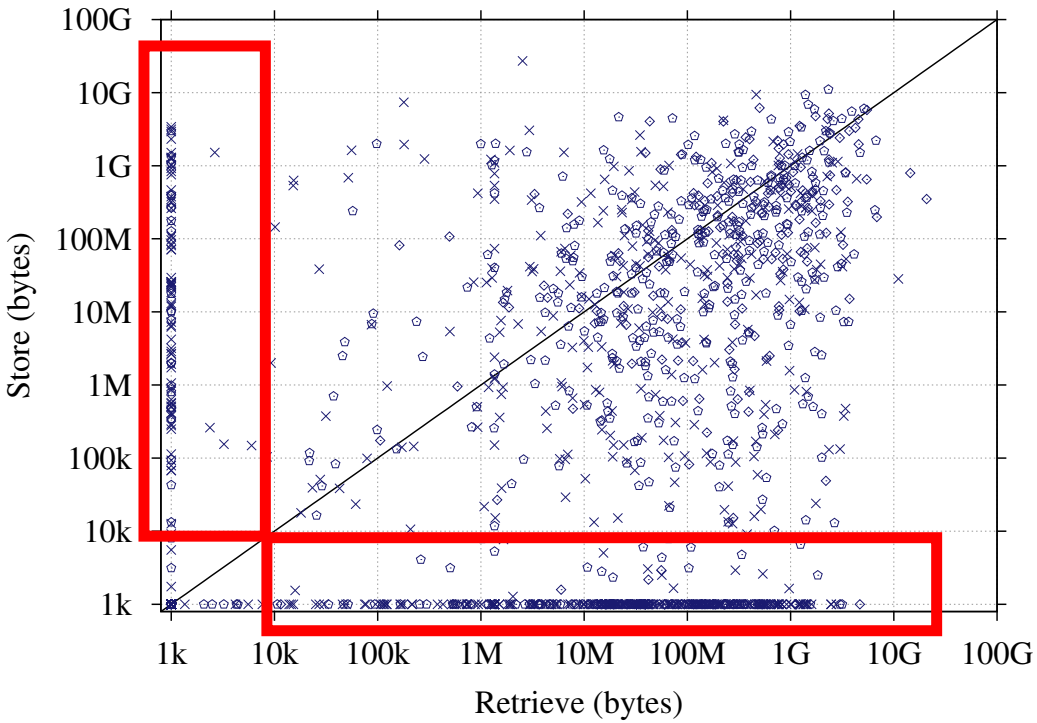
Are there typical usage scenarios?



- **Occasional:**
  - Users: **31%**
  - Devices per user: 1.22

- **Abandoned Dropbox clients**
- No storage activity for 42 days

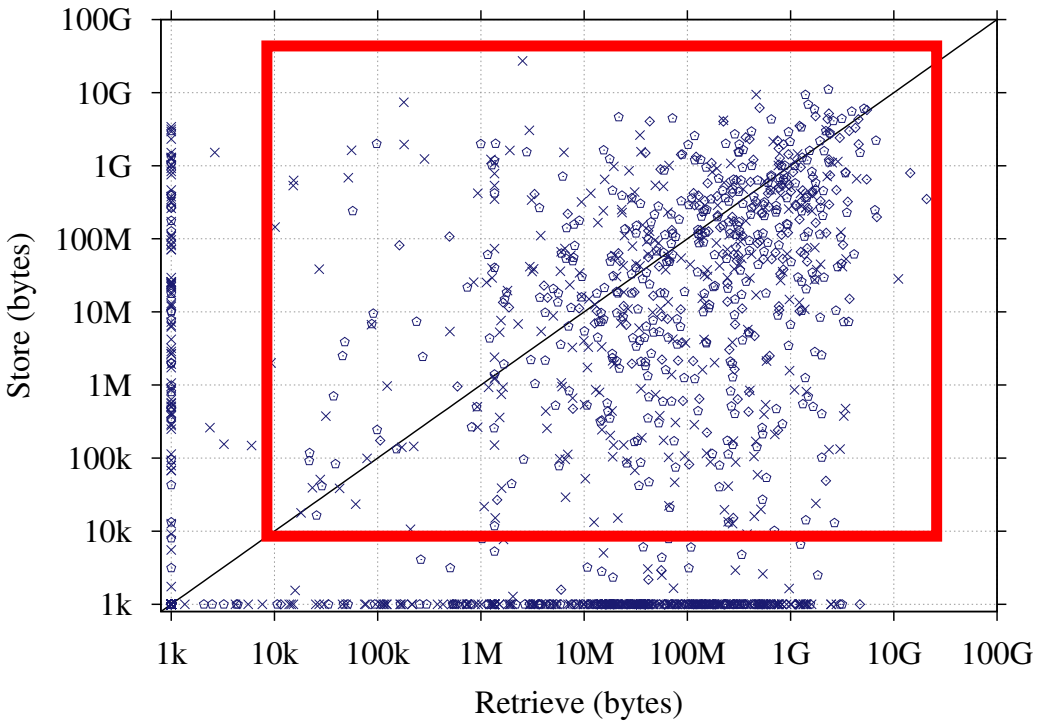
Are there typical usage scenarios?



- **Upload-only:**
  - Users: 6%
  - Uploads: 11–21%
  - Devices per user: **1.36**
- **Download-only:**
  - Users: 26%
  - Downloads: 25–28%
  - Devices per user: **1.69**

- Backup and content sharing
- Geographically dispersed devices

Are there typical usage scenarios?



- **Heavy:**
  - Users: **37%**
  - Uploads: 79–89%
  - Downloads: 72–75%
  - Devices per user: **2.65**

- **Synchronization of content** in a household

Conclusions

- **1st to analyze Dropbox usage on the Internet**

Conclusions

- **1st to analyze Dropbox usage on the Internet**
- Adoption above 6% in our datasets, data hungry application!

Conclusions

- **1st to analyze Dropbox usage on the Internet**
- Adoption above 6% in our datasets, data hungry application!
- Architecture and performance
  - **Bottlenecks from system design choices**

Conclusions

- **1st to analyze Dropbox usage on the Internet**
- Adoption above 6% in our datasets, data hungry application!
- Architecture and performance
  - **Bottlenecks from system design choices**
- **Extensive characterization of workload and usage**
  - User groups, number of devices, daily activity etc.

Questions?

- **Thank You.**

- Anonymized traces and scripts
  - `http://traces.simpleweb.org/dropbox/`