UNIVERSITY OF TWENTE.

POLITECNICO
DI TORINO

Benchmarking Personal Cloud Storage

- Idilio Drago
- **Enrico Bocchi**
- Marco Mellia
- Herman Slatman
- Aiko Pras

mPlane

Flamingo
EU FP7

Motivation and goals

- Personal Cloud Storage: very popular, large amount of traffic
  - Dropbox: 100 million users, 1 billion uploads / day

Motivation and goals

- Personal Cloud Storage: very popular, large amount of traffic
  - Dropbox: 100 million users, 1 billion uploads / day
- Lots of different applications
  - Do they follow **different design**?
  - Which are their **client capabilities**?
  - What is the impact on **end-user performance**?

Motivation and goals

- Personal Cloud Storage: very popular, large amount of traffic
  - Dropbox: 100 million users, 1 billion uploads / day
- Lots of different applications
  - Do they follow **different design**?
  - Which are their **client capabilities**?
  - What is the impact on **end-user performance**?

Which infrastructure is being used?

Which infrastructure is being used?

- Manual preliminary evaluation of each application
  - Mostly HTTPS
  - **Control** and **Storage** servers

Which infrastructure is being used?

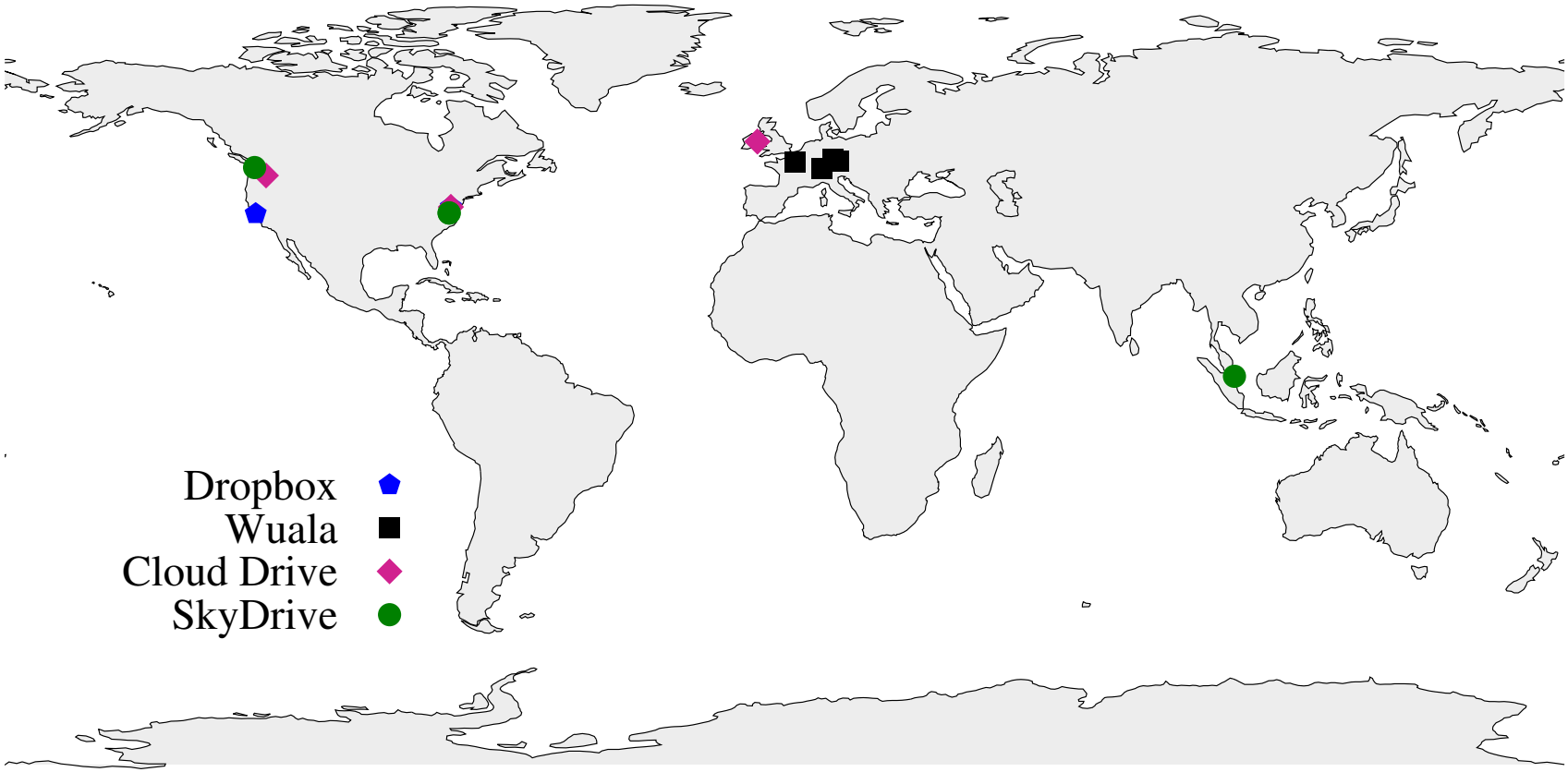
- Manual preliminary evaluation of each application
  - Mostly HTTPS
  - **Control** and **Storage** servers
- Where are their data centers
  - Collect the list of **server names** of each application
  - Resolve the names using **2,000 open DNS resolvers**
  - Collect the list of server IP addresses

Which infrastructure is being used?

- Manual preliminary evaluation of each application
  - Mostly HTTPS
  - **Control** and **Storage** servers
- Where are their data centers
  - Collect the list of **server names** of each application
  - Resolve the names using **2,000 open DNS resolvers**
  - Collect the list of server IP addresses
- **Geolocation IP addresses**
  - Use common techniques
  - Example: Shortest RTT from PlanetLab nodes

Where are data centers located?

Where are data centers located?



- Few *Control* and *Storage* locations
- **Centralized services**

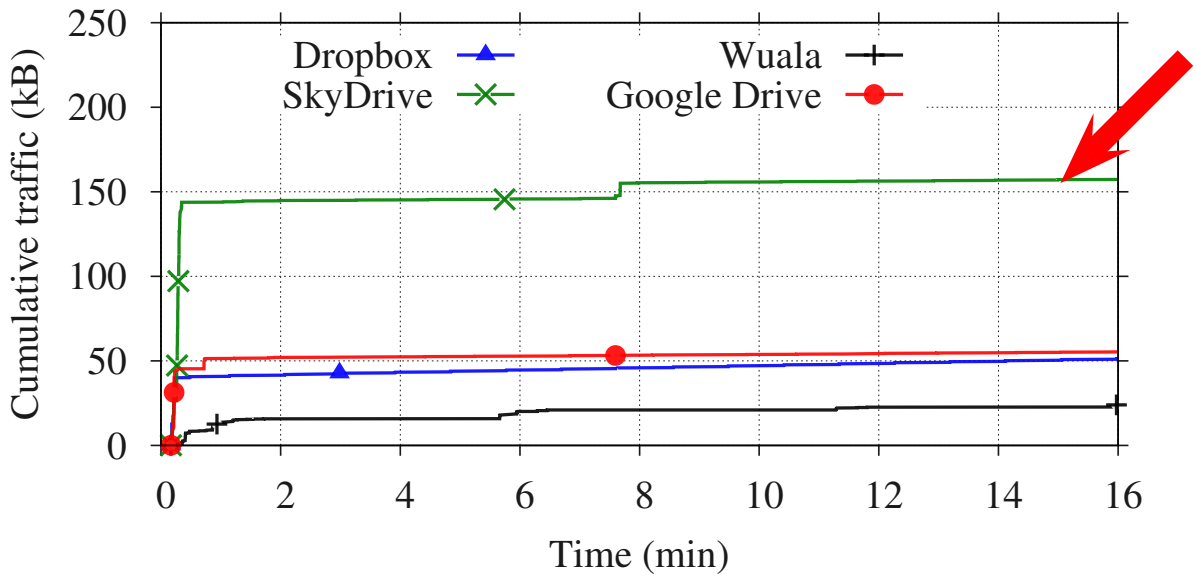
Where are data centers located?

- Users reach the closest Google's Edge Point of Presence[1]
- Reduced RTT, **offload public Internet**

[1] https://peering.google.com/about/delivery_ecosystem.html

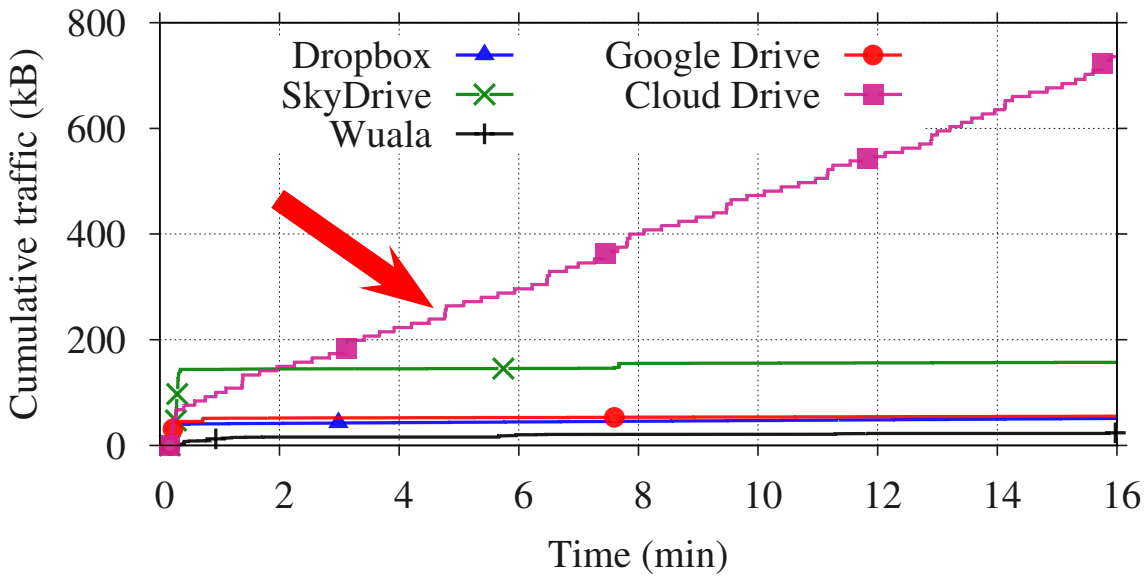
Protocol design: What happens when the app is idle?

Protocol design: What happens when the app is idle?



- **Generally silent protocols**
  - *e.g.,* SkyDrive: 1 min polling interval (32 b/s)

Protocol design: What happens when the app is idle?

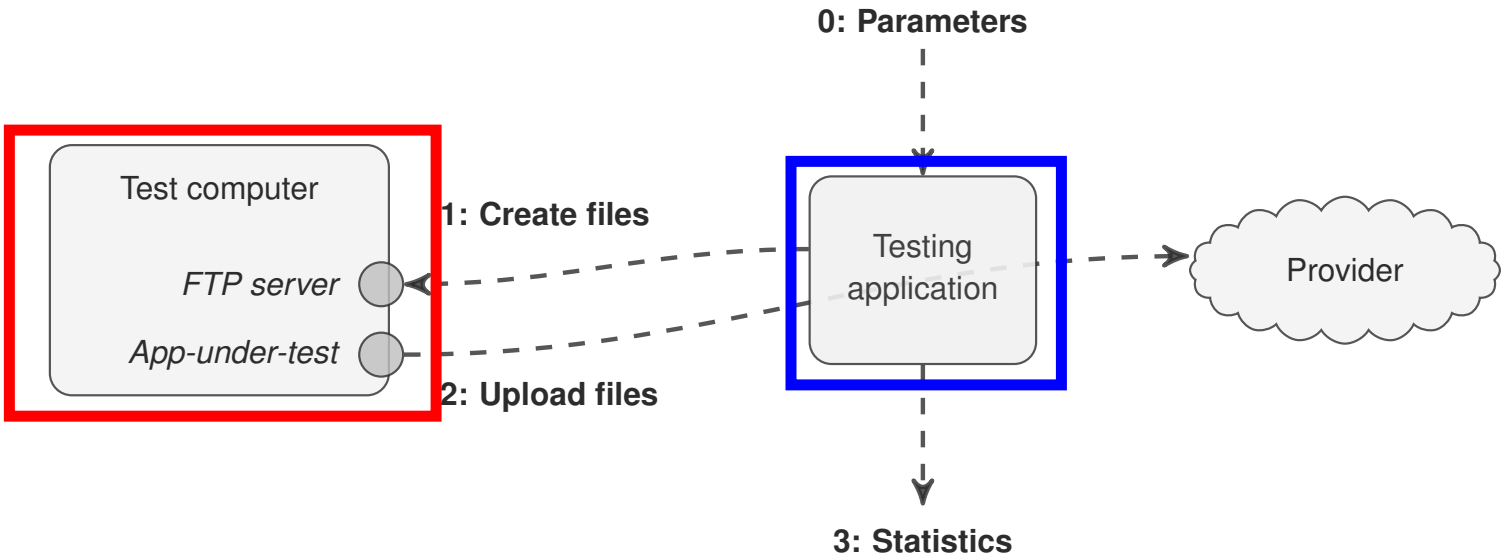


- Cloud Drive: polling **every 15 s** over a **new HTTPS** session
  - 6 kb/s per user $\rightarrow$ 65 MB per day per user
  - 1 million users $\rightarrow$ **6 Gb/s** of signaling traffic

Methodology – Creating benchmarks

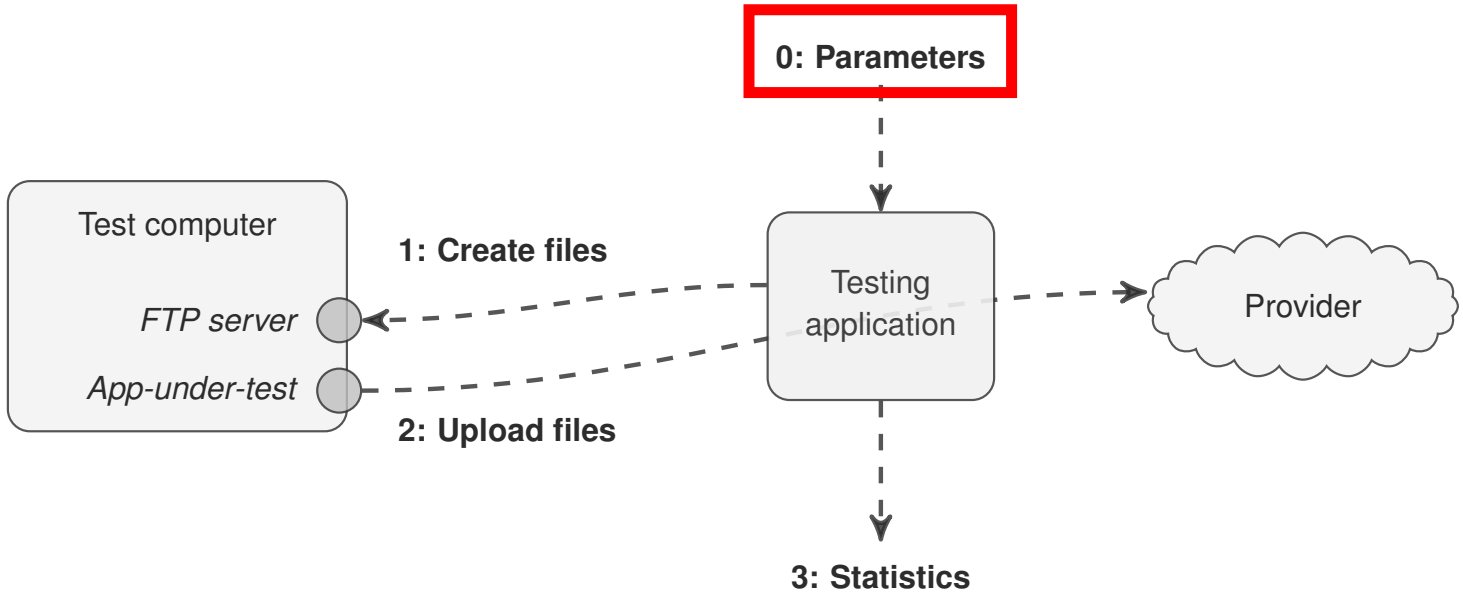
- Ad-hoc crafted workloads to detect specific features
- Check implications on end-user performance

Methodology – Creating benchmarks



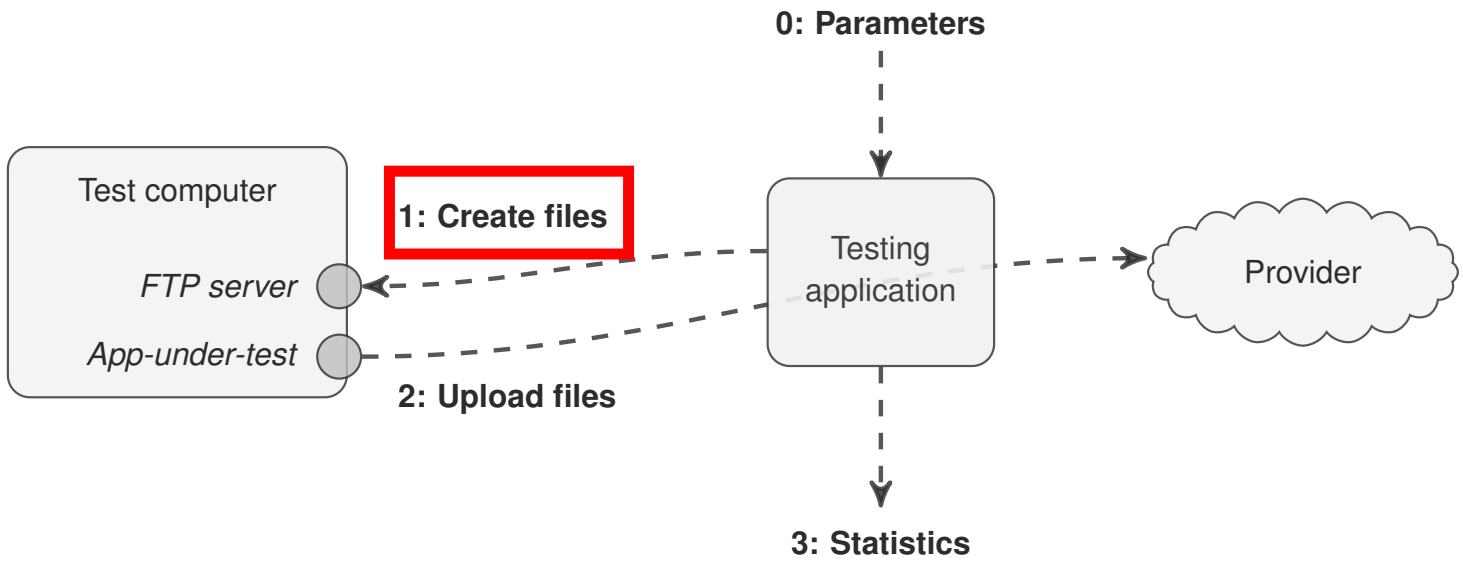
- **Application-under-test** – service clients running on a virtual machine
- **Testing application** – python scripts controlling the experiments

Methodology – Creating benchmarks



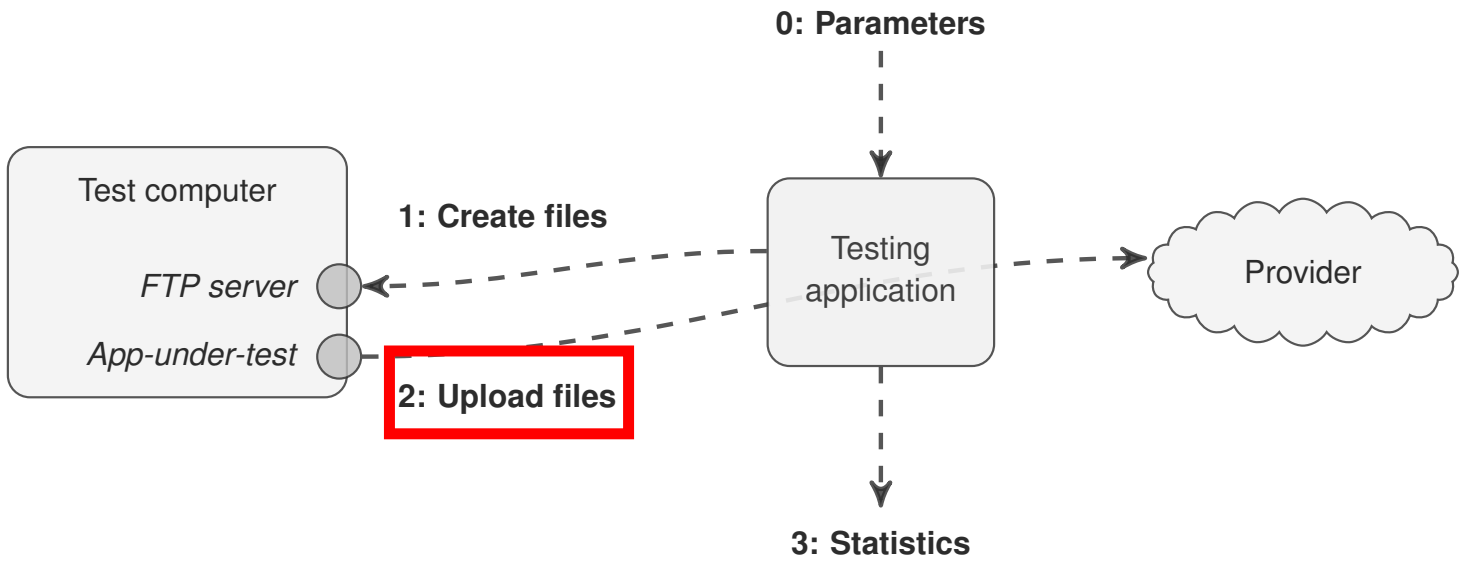
- Configuration parameters – *e.g.*, number of repetitions
- **Workload definition** – *e.g.*, number of files, type of content etc.

Methodology – Creating benchmarks



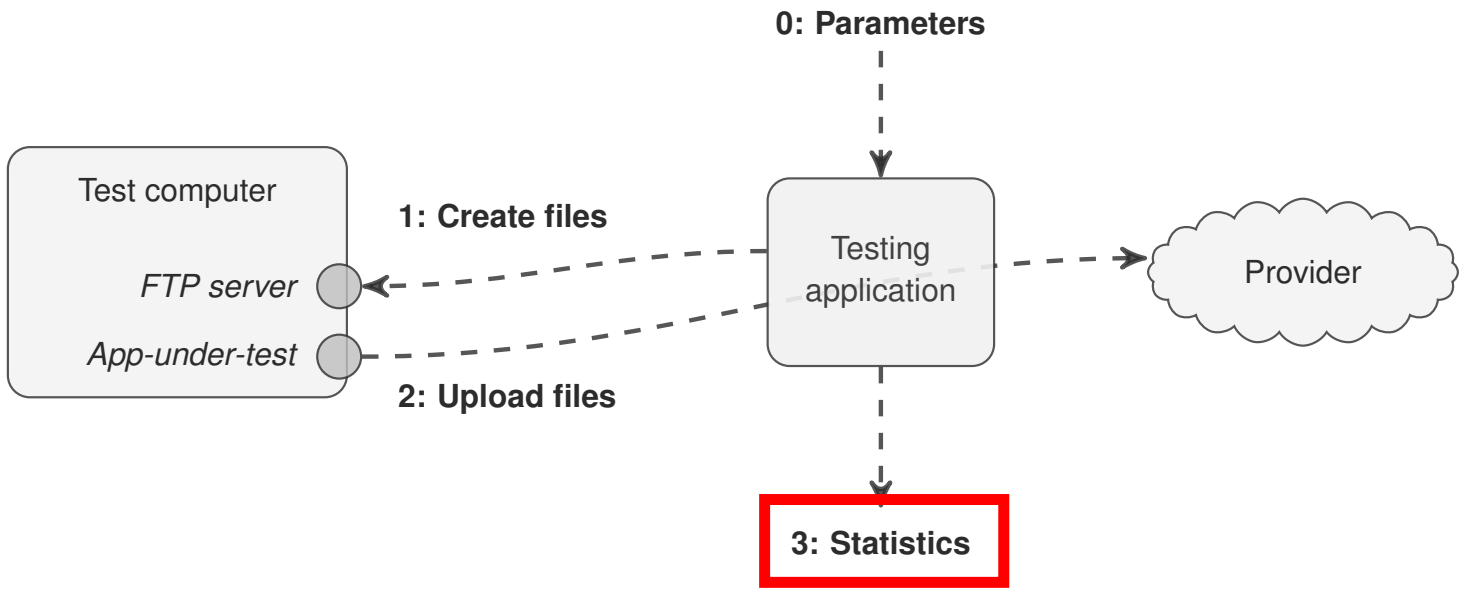
- Workload manipulated via FTP
- **Files created at run-time** – *e.g.*, 1 MB file with random text

Methodology – Creating benchmarks



- **Application-under-test synchronizes files**
- Traffic is intercepted (packet level)

Methodology – Creating benchmarks



- **Post-process to calculate performance metrics**
  - Compute upload time, overhead, etc.

What are the client capabilities?

What are the client capabilities?

	Dropbox	SkyDrive	Wuala	Google Drive	Cloud Drive
Chunking	4 MB	variable	variable	8 MB	✗
Bundling	✓	✗	✗	✗	✗
Deduplication	✓	✗	✓	✗	✗
Delta encoding	✓	✗	✗	✗	✗
Compression	always	never	never	smart	never

- **Clients differ considerably**

What are the client capabilities?

	Dropbox	SkyDrive	Wuala	Google Drive	Cloud Drive
Chunking	4 MB	variable	variable	8 MB	✗
Bundling	✓	✗	✗	✗	✗
Deduplication	✓	✗	✓	✗	✗
Delta encoding	✓	✗	✗	✗	✗
Compression	always	never	never	smart	never

- Uploading a **large file**
  - Upload it as a single content?
  - ... or chop it into smaller chunks?

What are the client capabilities?

	Dropbox	SkyDrive	Wuala	Google Drive	Cloud Drive
Chunking	4 MB	variable	variable	8 MB	✗
Bundling	✓	✗	✗	✗	✗
Deduplication	✓	✗	✓	✗	✗
Delta encoding	✓	✗	✗	✗	✗
Compression	always	never	never	smart	never

- Uploading **lots of files**
  - Create a bundle?
  - ... or one/many TCP connections/transactions?
- Google Drive and Cloud Drive open one/three TCP connections per file

What are the client capabilities?

	Dropbox	SkyDrive	Wuala	Google Drive	Cloud Drive
Chunking	4 MB	variable	variable	8 MB	✗
Bundling	✓	✗	✗	✗	✗
Deduplication	✓	✗	✓	✗	✗
Delta encoding	✓	✗	✗	✗	✗
Compression	always	never	never	smart	never

- Creating a second **copy of the same file**
  - Re-upload it completely?
  - ... or just update metadata?
- Deduplication in Wuala is compatible with per-user encryption

What are the client capabilities?

	Dropbox	SkyDrive	Wuala	Google Drive	Cloud Drive
Chunking	4 MB	variable	variable	8 MB	✗
Bundling	✓	✗	✗	✗	✗
Deduplication	✓	✗	✓	✗	✗
Delta encoding	✓	✗	✗	✗	✗
Compression	always	never	never	smart	never

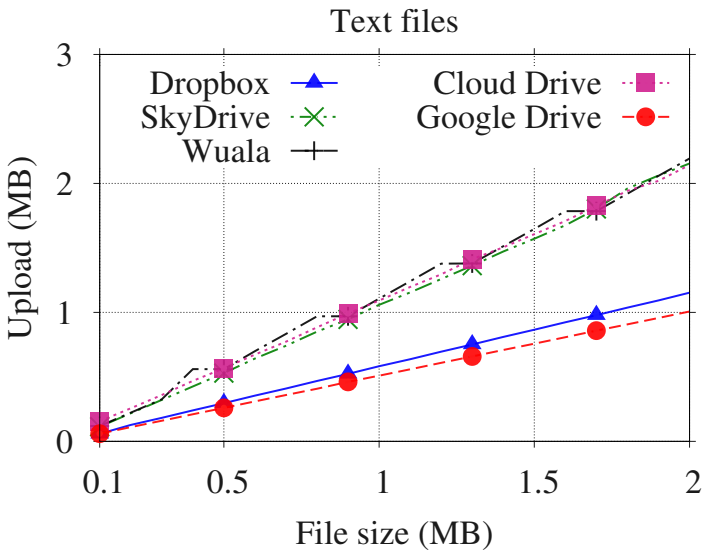
- Modifying only a **fraction of a file**
  - Re-upload everything?
  - ... or just the modified portion?
- This has some implications with chunking

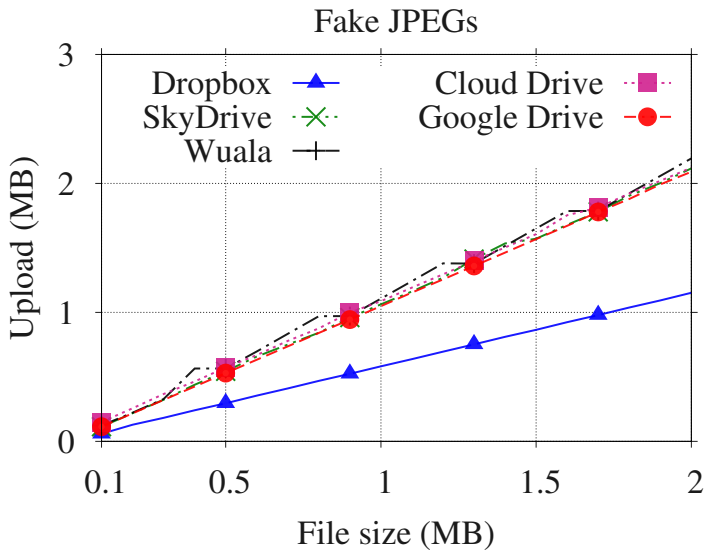
What are the client capabilities?

	Dropbox	SkyDrive	Wuala	Google Drive	Cloud Drive
Chunking	4 MB	variable	variable	8 MB	✗
Bundling	✓	✗	✗	✗	✗
Deduplication	✓	✗	✓	✗	✗
Delta encoding	✓	✗	✗	✗	✗
Compression	always	never	never	smart	never

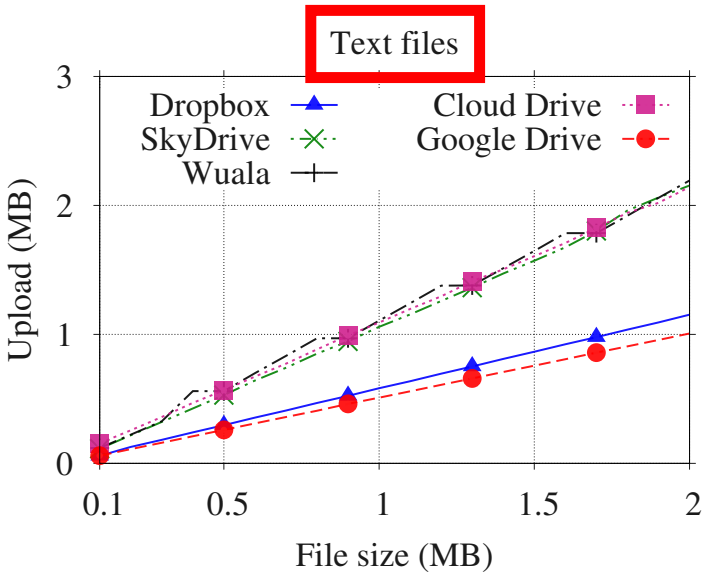
- Uploading **compressible content**?
  - Transmit it as quickly as possible?
  - ... or compress content?

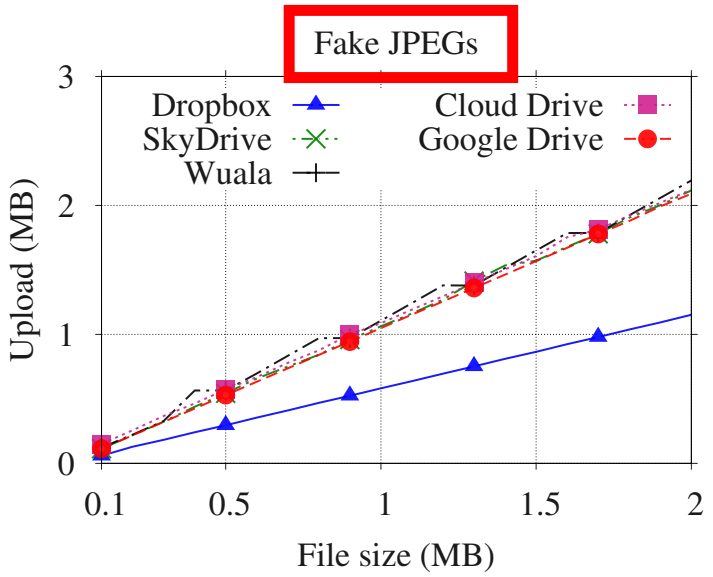
Example: Compression



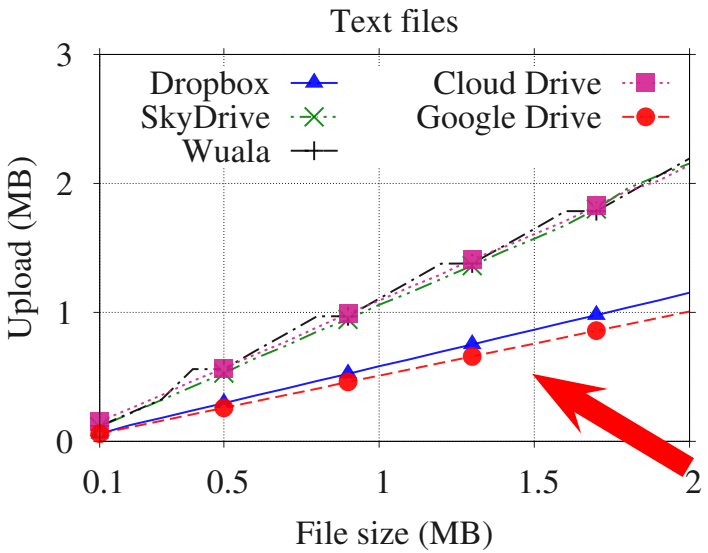


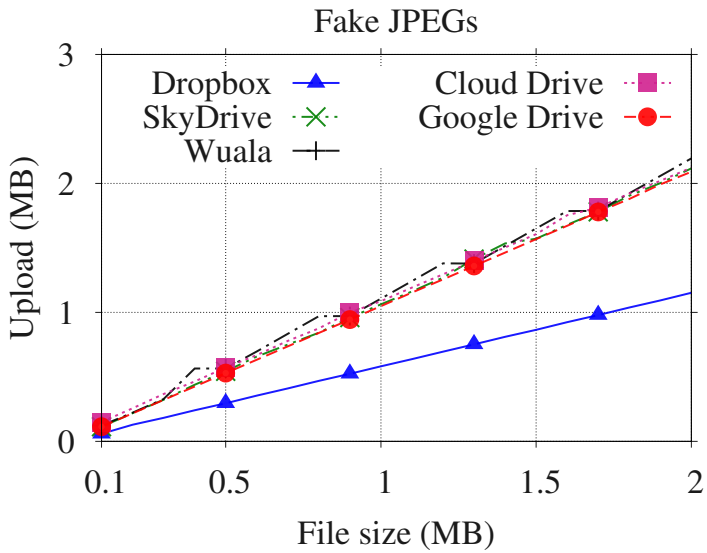
Example: Compression





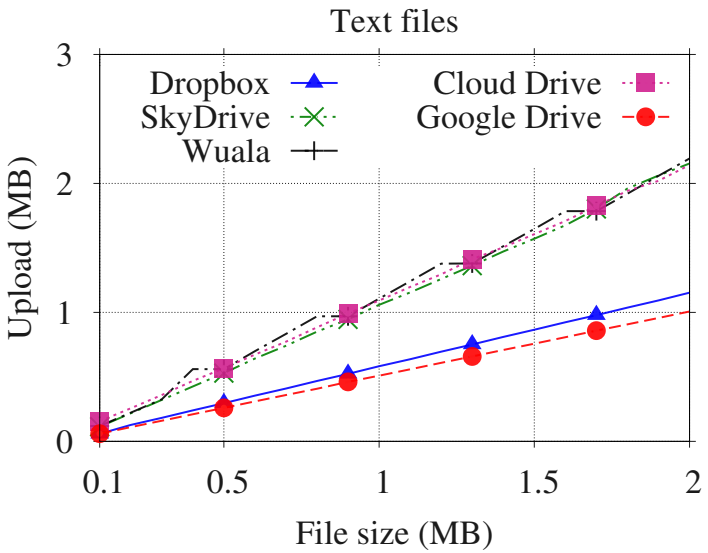
Example: Compression

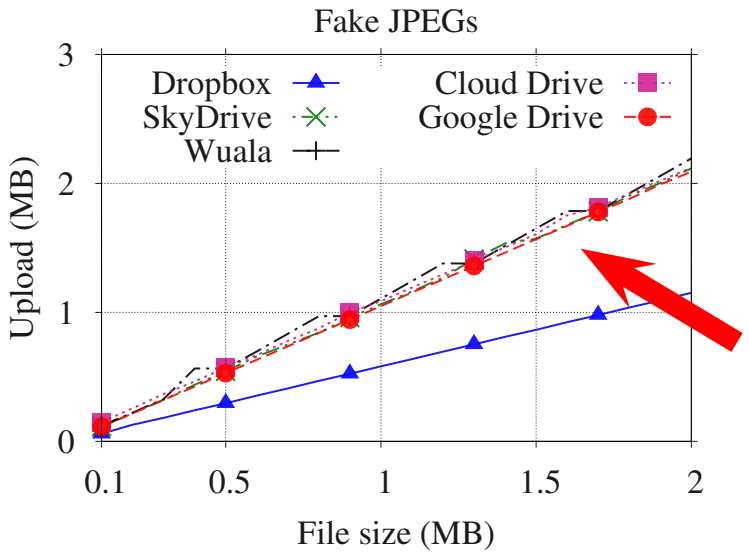




- Only Dropbox and Google Drive compress files
  - Dropbox has higher control overhead

Example: Compression

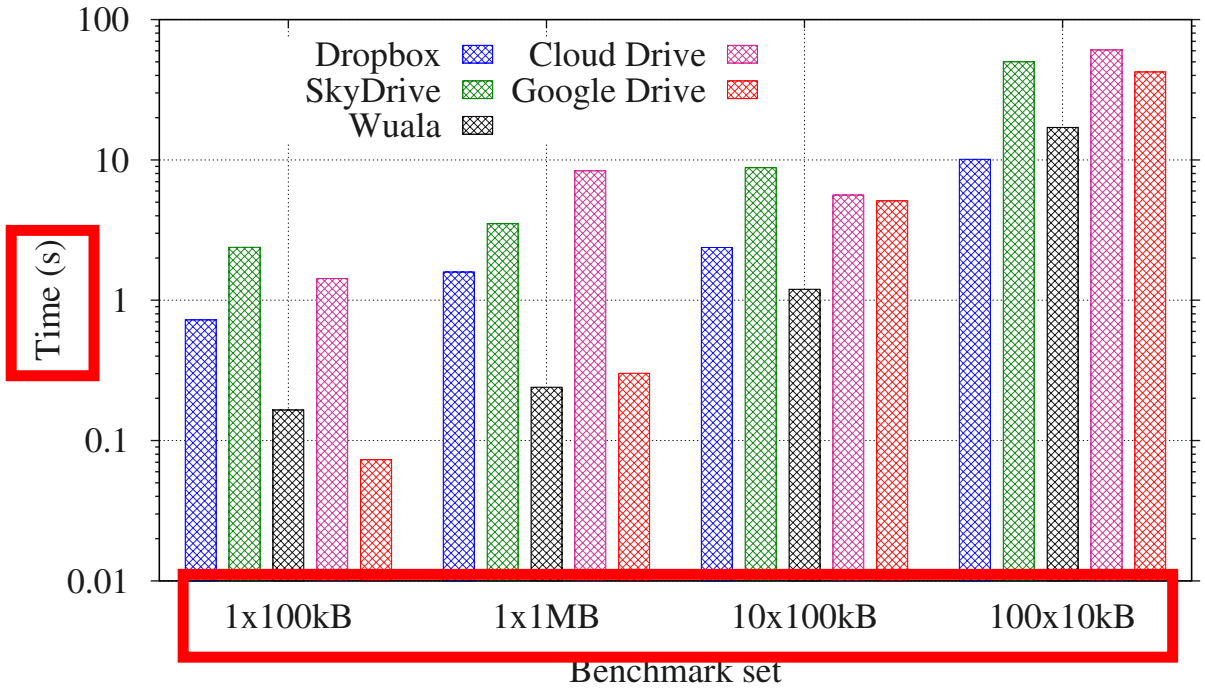




- Only Dropbox and Google Drive compress files
  - Dropbox has higher control overhead
- Google Drive identifies JPEG content and skip compressing it

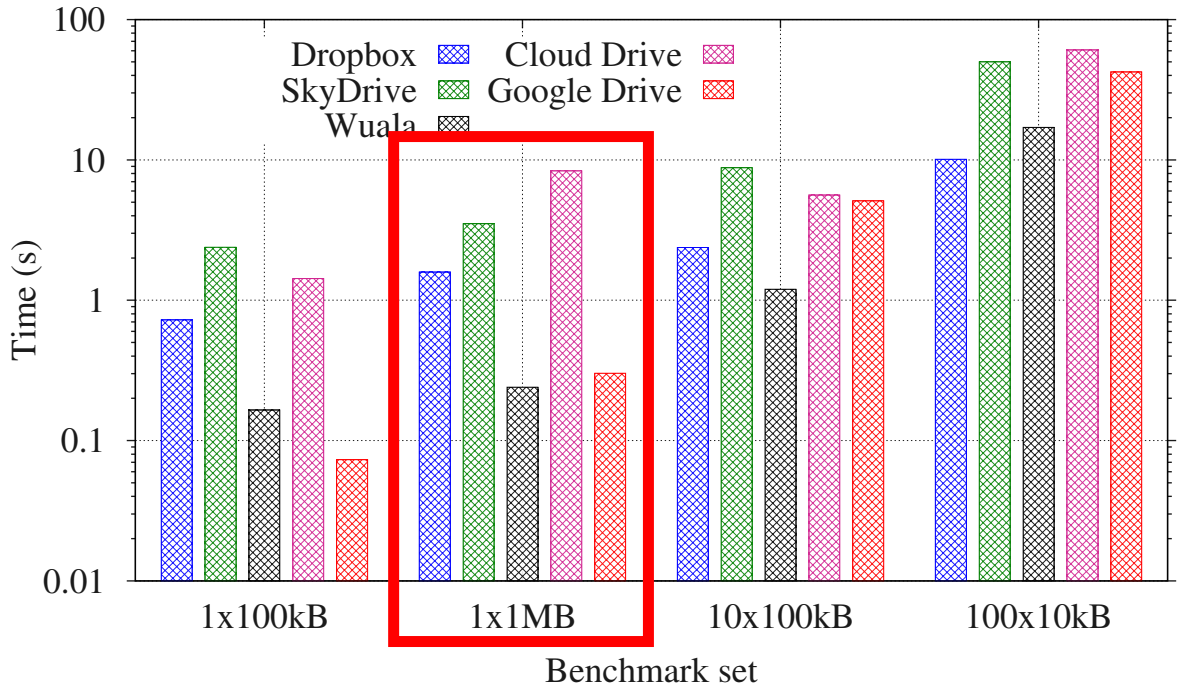
Implications to end-user performance?

Implications to end-user performance?



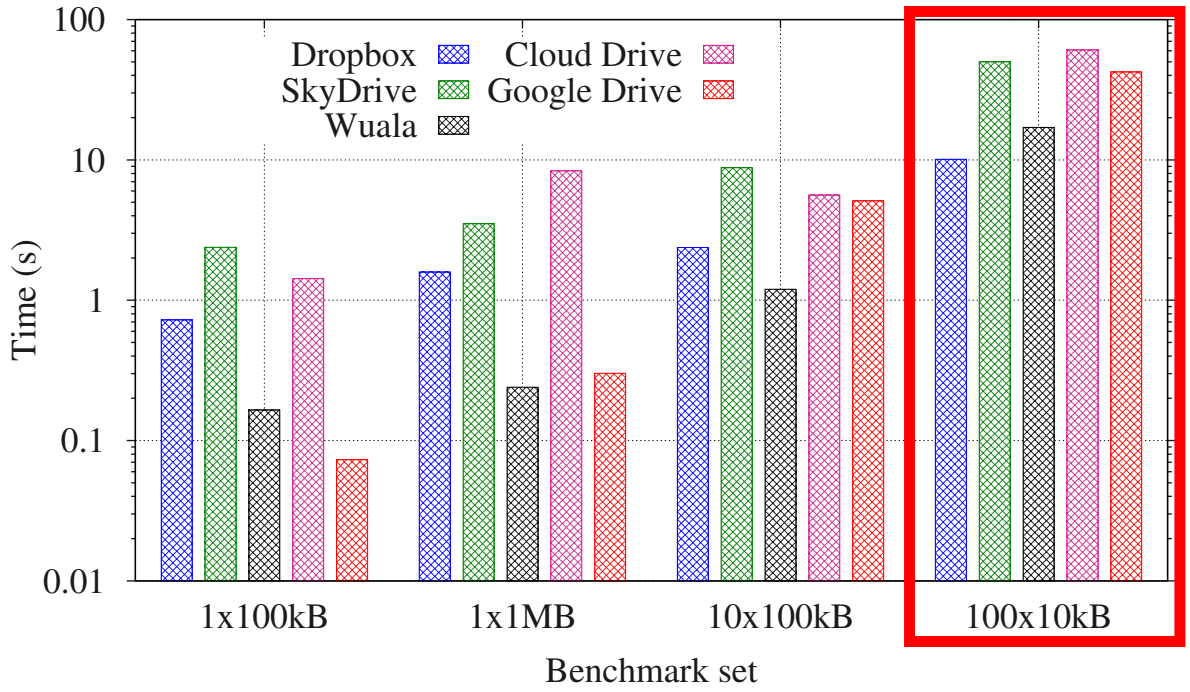
- Consider different workloads $\rightarrow$ Upload time?
  - One versus many files
  - Small versus large files
- Notice: test run from Europe

Implications to end-user performance?



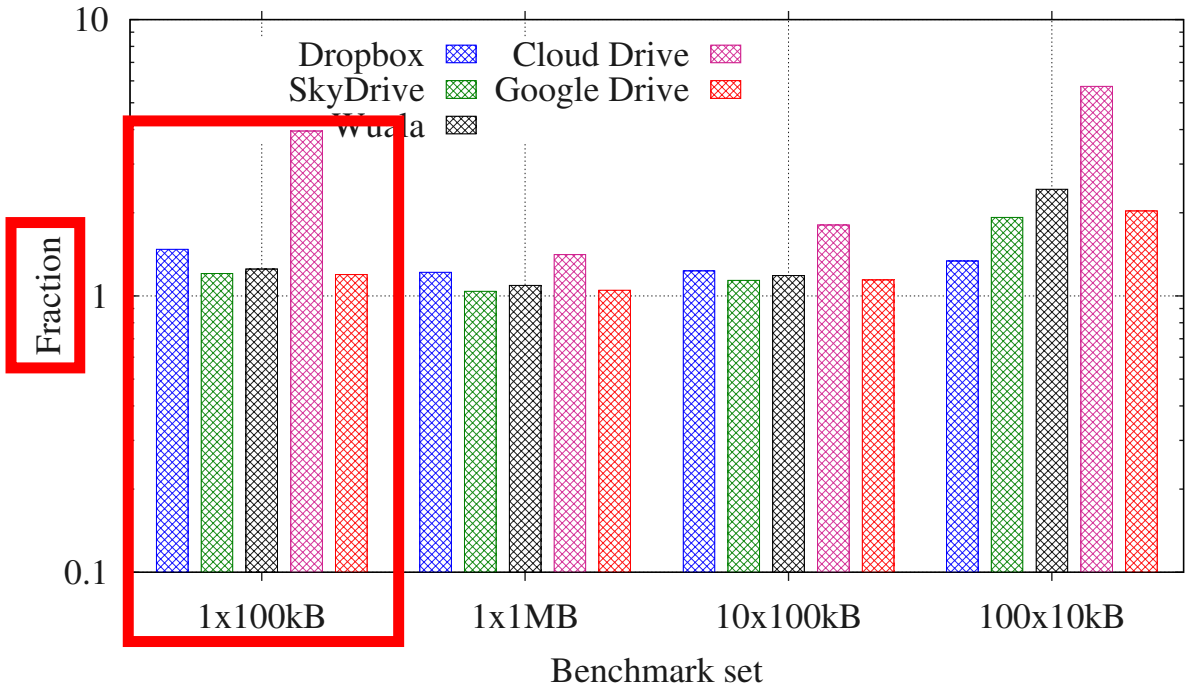
- **Network latency** dominates the upload because of TCP
  - SkyDrive (160 ms RTT) $\rightarrow$ **4 s** to send a 1 MB file
  - Google Drive (15 ms RTT) $\rightarrow$ **300 ms** to send a 1 MB file

Implications to end-user performance?



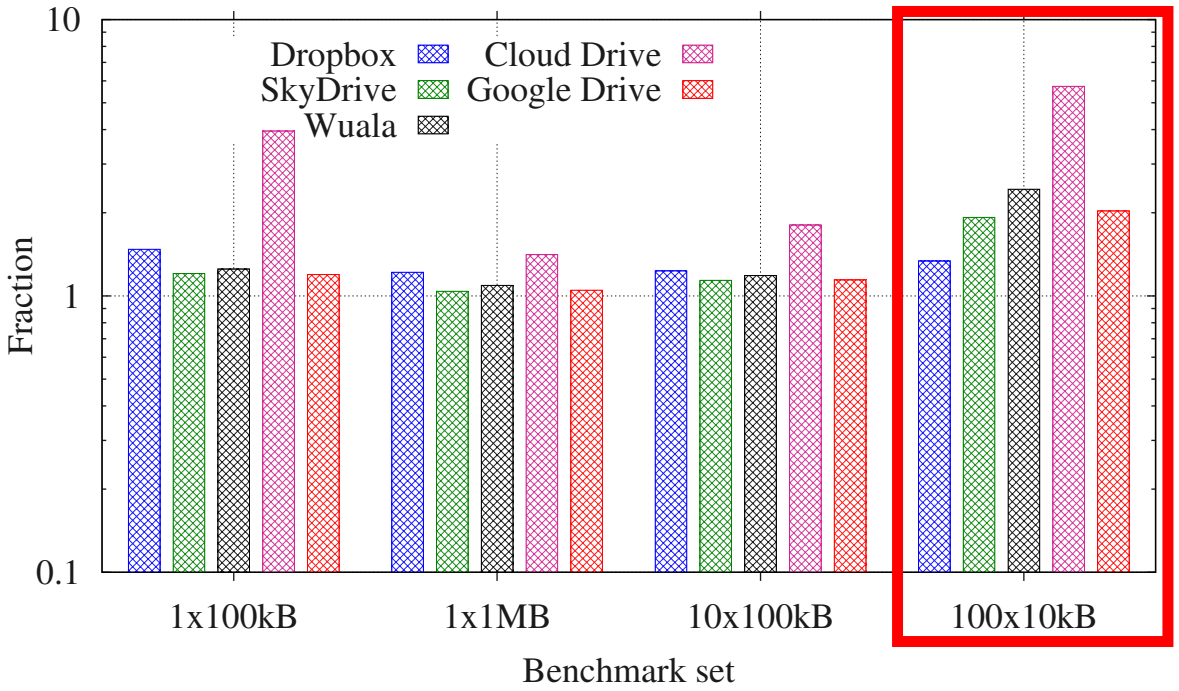
- **Client capabilities** boost performance in this scenarios
  - Dropbox (90 ms RTT) $\rightarrow$ **10 s** to send 100 files of 10 kB each
  - Google Drive (15 ms RTT) $\rightarrow$ **42 s** to send 100 files of 10 kB each

Overhead = Total Traffic / Content Size?



- Cloud Drive $\rightarrow$ **3 HTTPS connections** per file
- Dropbox $\rightarrow$ **signaling cost** of client capabilities

Overhead = Total Traffic / Content Size?



- 100 small files
  - **Bundling files pays back in network overhead**

Conclusions

- Designed specific benchmarks for Personal Cloud Storage

Conclusions

- Designed specific benchmarks for Personal Cloud Storage
- Highlighted **design choices**
- ... and **their implications** on performance

Conclusions

- Designed specific benchmarks for Personal Cloud Storage
- Highlighted **design choices**
- ... and **their implications** on performance
- **Data center placement**
  - Centralized vs. distributed topologies

Conclusions

- Designed specific benchmarks for Personal Cloud Storage
- Highlighted **design choices**
- ... and **their implications** on performance
- **Data center placement**
  - Centralized vs. distributed topologies
- **Client capabilities**
  - Performance gains from bundling, deduplication etc.

Conclusions

- Designed specific benchmarks for Personal Cloud Storage
- Highlighted **design choices**
- ... and **their implications** on performance
- **Data center placement**
  - Centralized vs. distributed topologies
- **Client capabilities**
  - Performance gains from bundling, deduplication etc.
- **Protocol design**

Questions?

- **Thanks for your attention**

- Data and scripts can be downloaded from:
  - `http://www.simpleweb.org/wiki/Cloud_benchmarks`