Performance Evaluation of Online Backup Cloud Storage

Xiao Zhang, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

WenXiong Feng, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

Xiao Qin, Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, USA

ABSTRACT

Cloud storage provides storage services to users via Internet. They provide lavish interfaces for different applications. Online backup is the most developed application based on cloud storage. Most past studies focus on functional characters and price per GB. But there is not much consideration about performance evaluation. In this paper, the authors present a method to evaluate the performance of different online backup services with clients. Because these clients do not prompt when the download/upload process starts and finishes, the authors capture the packets transferred between clients and the cloud storage provider. The authors evaluate the performance from the view of the end-user, comparing the upload and download speeds of different services. This method can help users find the best provider to accommodate their situation. The authors evaluate the performance of several cloud storage providers by analyzing the performance of different vendors in different times and different file type and size. From the performance results, the authors can expose which technologies are used in the cloud storage system to reduce the storage and network cost.

Keywords: Cloud Storage, Measurement, Network Monitor, Performance, Performance Evaluation

1. INTRODUCTION

Cloud computing is becoming an important aspect in IT industry. Cloud storage is an application of cloud computing, which is the most developed part in cloud application. It depends on the cluster application, grid technology, and distributed file system providing storage services to users via the Internet. In most conditions, cloud storage can provide high reliability and secure storage service at competitive prices.

There are many software working together through disk array and clusters in the cloud storage system. Compared with traditional storage devices, cloud storage is not only hardware. It includes network, storage, server, and appli-

DOI: 10.4018/ijcac.2013070102

cation software. It provides lavish interfaces, including file system interface, database access interface, and API used for applications running in the cloud.

The most developed cloud storage application is online backup or file-syncing. Disaster may happen at any time: fire, flood, tornado, or hard drive failure. These disasters can destroy all local-stored data. Users can use remote backup to protect their data from disasters. The other choice is to store copies of files in the cloud storage. Online backup is an Internet-based system that is set to automatically back up all selected files. These files are stored online and can be accessed anywhere, which is especially useful in case the local computer or server is lost or damaged. The benefit of using online storage services is not limited to protecting data. Cloud storage services make it easy to share files from different machines and mobile devices.

There are more than 50 cloud storage service providers, such as Amazon, Google, Microsoft, and Dropbox (Wikipedia, n.d.). The services can vary greatly when it comes to features, accessibility, and support options. Most past studies focus on the comparison of functional features and price per GB. Performance is an important feature for storage systems, but it is difficult to evaluate for cloud storage because of the complexity of the system and variable network conditions.

In this paper, we present an evaluation method to measure the performance of cloud storage from user's view. We tested 5 selected cloud storage services and analyzed the results. Our experiments were done inside China, so we selected three of the most popular Chinese vendors and two popular vendors of America.

The rest of this paper is organized as following: Section 2 introduces the previous studies on cloud storage performance measurement. Section 3 introduces the evaluation environments. Section 4 presents the method to test the performance of a storage service. Section 5 shows the test results of 5 different cloud storage services. Section 6 illustrates how to conclude which accelerate technologies are used in the cloud storage system.

2. RELATED WORKS

The cloud storage market is now crowded with different service providers. There are plenty of services to choose from. Many articles introduced how to choose new cloud storage services. Vince compared the most well known cloud storage providers, including Dropbox, SkyDrive, SugarSync, and Box.net (Cloud storage comparison, n.d.). He compared the capacity and functionality of different providers. And there is a comparison of 51 different online backup services (Wikipedia, n.d.). Most of these studies compare the free and paid space, support platform, convenience, and security.

Too many think of cloud storage as just another or the next type of storage. As usual with this view, it is associated with a view that the "next" storage type is bigger, faster and cheaper. As such, proponents of this view generally believe that access via traditional approaches, like WebDAV, NFS, CIFS, and others, is a critical capability (Cloud Storage Strategy, 2009). Performance is an important factor for storage devices. Cloud storage providers offer abundant interfaces for different purposes (e.g., REST, File System, SQL like query). However, the novel interfaces (usually neither SQL nor POSIX), elasticity, and new use cases of cloud serving systems motivate a new benchmark.

Most cloud storage performance evaluation researches are focused on the database operation in the cloud. CloudCmp measures the elastic computing, persistent storage, and networking services offered by a cloud along metrics that directly reflect their impact on the performance of customer applications (Li, Yang, Kandula, & Zhang, 2010). They also compared the performance of persistent storage in cloud. They selected three types of storage services: table, blob and queue. They used the operation response time, time to consistency, and cost per operation to compare the quality of cloud storage. YCSB provided the performance comparisons on the database view (Cooper, Silberstein, Tam, Ramakrishnan, & Sears, 2010). Some scholars do research on web based cloud storage (Cjiu & Agrawal, 2010; Zhao, Liu, &

Keung, 2010). Cumulus implemented a system for efficiently implementing file system backups over the Internet (Vrable, Savage, & Voelker, 2009). This can be used to test the performance of online backup services supporting S3 access interface.

Many online backup service providers implement clients on multiple platforms. The client maps the space in the cloud as a local disk device or a network folder. Many test tools were developed to test local disk performance, such as IOmeter, IOZone, and Hdtune. But these tools cannot be used on cloud storage devices because it is a virtual device which has a longer delay. Some clients save the data on local disks and upload data in the background. In these conditions, the test tool will only get the time of saving the file on local disks or folders. For the online backup user, they want to know when the data will be completely stored in the cloud. Customers need a method to test which cloud storage platform is more suitable, but there are rare cloud storage testing methods and tools.

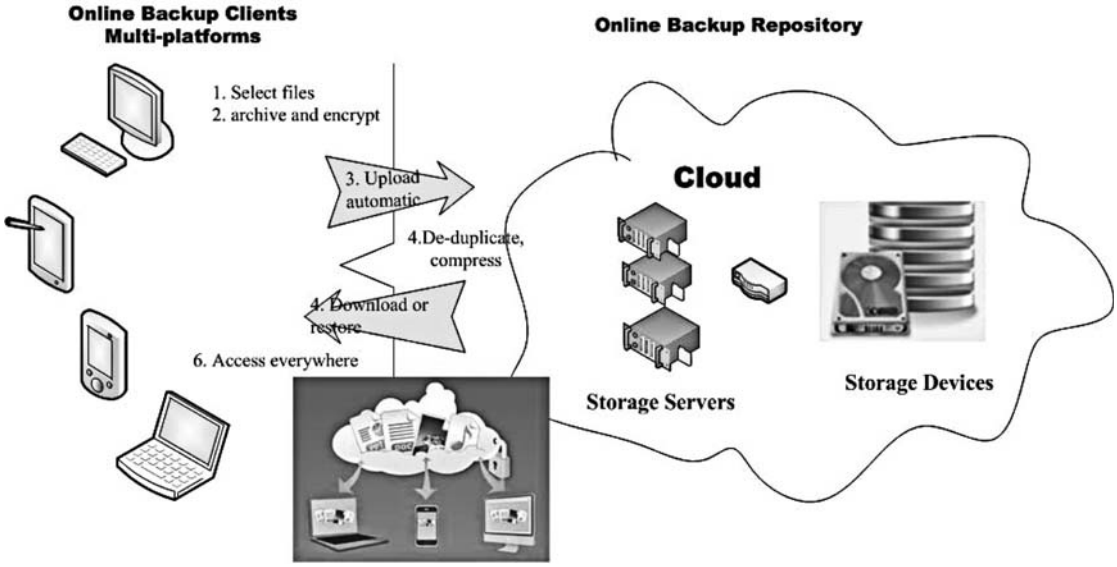
To improve the performance of online backup, a network appliance or server which resides at the customer premises translates data through internet. These appliances or servers are called cloud storage gateways. Unlike the cloud storage services which they complement, cloud storage gateways use standard network protocols which provide a seamless integration with existing applications. Cloud storage gateways can also serve as intermediaries to multiple cloud storage providers. Some cloud storage gateways also include additional storage features such as backup and recovery, caching, compression, encryption, de-duplication, and provisioning. Some vendors announce that they can accelerate the transmission 10-30 times faster. Amazon provides storage gateway as a virtual machine (VM), users can download it and choose the gateway work method, gateway-cache or gateway-stored. The AWS storage gateway only uploads data that has changed, minimizing the amount of data sent over the Internet (http://aws.amazon.com/storagegateway/). CloudArray is available as a virtual, physical or in-cloud appliance that takes just minutes to configure and integrates with more than 20 public and private cloud options, as well as local or remote storage devices (Cloudarray storage gateway, n.d.).

3. EVALUATION ENVIRONMENT

An online backup service is an easy and reliable way to store valuable information on a remote site. The process of backup and restore with online backup services is shown in Figure 1. Most online backup service providers implement clients on multiple platforms. Users select the files or folders that are to be synced into cloud storage. Then clients will do some archiving or encryption job before uploading. There are many things the client can do to reduce the transmit file size. It can compress files to decrease the size of transmit, which needs more local compute resource. Another method is to find same blocks in files and transmits these blocks only once. The last is that clients compute the fingerprint of each file and check if there is the same file stored in the cloud before. If the file was stored by other users, it just sends control information to create a symbol link to the existing file. The client uploads files to the cloud storage system through the Internet. The cloud storage system will check for the same blocks among the whole system. The data should be partitioned in distributed file system with single name-space. And the clients on other machines or platforms can view and download files anywhere.

Online backup clients play an important role in the system. They provide a mapping mechanism; the data in remote side is mapped to a local disk device. They make remote space act as a regular computer drive, which is hosted remotely on another file server. Because the mapping protocol is not a standard protocol, users have to install several different clients to use multiple cloud storage services. The clients communicate with cloud storage services with private protocols. These can be simply divided into three categories: transfer protocol based on HTTP, File transfer protocols such as FTP, and private protocols. Amazon Simple Storage

Figure 1. Process of online backup and restore



Service provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web (http://aws.amazon.com/s3/). It gives developers access to the same highly scalable, reliable, secure, fast, and inexpensive infrastructure that Amazon uses to run its own global network of web sites. Files in Amazon S3 are organized in buckets and objects. Using protocols based on HTTP or FTP makes it easy to implement on different platforms. It provides cloud storage APIs such as SOAP or REST. But there is much convenience in implement clients based on private protocols. By using a customized client, cloud storage service providers can integrate clients as a plug-in of the file system, which makes it easy for users. Clients can also compress files before they are uploaded to cloud storage, which will result in better upload performance. De-duplication technology also can be implemented in clients to decrease transfer amount. The clients can upload files or directories in the background automatically.

Cloud storage service vendors provide service at an incredible low price. There are many technologies used to decrease the cost per GB. Thin provisioning is the act of using visualization technology to give the appearance of having more physical resources than is actually available. They will not allocate the space until the user stores files in it. Data de-duplication is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent data compression and single-instance data storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Unlike de-duplication applied in the client side, de-duplicate on the server side will search for the same blocks in whole system instead of one user. Some people save music and other public documents in their own space. In these conditions, clients can send the digital fingerprint of a file to check if the file was saved before. If the file was uploaded by another user before, it just adds a link to indicate that the file is shared by multiple users.

We present a method to measure the performance of different cloud storage online backup services. We compare the performance on different time and for different file size. From the designed scenario, we can judge which

technology was used in the system to reduce transfer or storage cost. In our environment, we used a fixed PC terminal to install several clients one by one, and we tested the download/upload speeds of different file types. After uploading files to the cloud storage, we tested the download speed from another client. The test environment is shown in Figure 2. In our test, we used same PC to install clients, and all experiments were under the same network condition. We selected three of the most popular cloud storage providers in China: Data Bank (DBank) from huawei company (http://www.dbank.com/), Kingsoft fast disk (Wikipedia, n.d.), and Surfing disk from China Telecomm (Chiu & Agrawal, 2010). We also selected two world-famous providers: Skydrive from Microsoft and Dropbox. For some reason, we cannot access Google drive from china. Configuration of our PC is listed below:

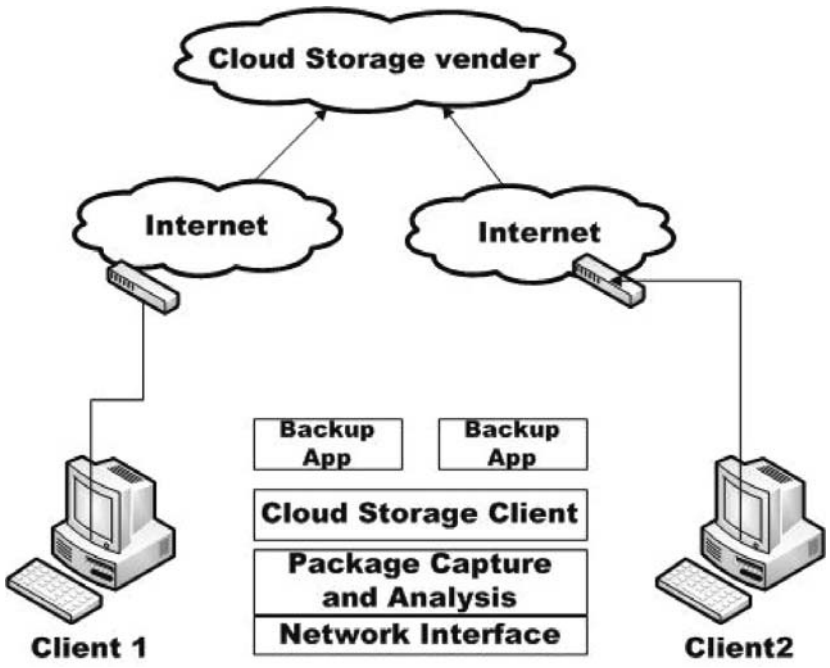
- **Model:** HP Pavilion p6-1236cx;
- **Hardware:** Intel Celeron CPU G540, 2GB memory.500G SATA disk (7200 rpm);
- **Software:** Windows XP sp2;
- **Network:** 1Gbps intranet, education network to internet.

4. EVALUATION METHODOLOGY

A typical online backup client is integrated with the file system as a plug-in. It monitors the modification of the directory to be synced. If a new file is added, it reads the content of the file and transmits the file to the cloud storage. Client mode is the most flexible mode. The clients are different in how they map data on the local side and how they accelerate the transmission process. It can support several data compression functions to accelerate the upload/download speed.

There are three kinds of cloud storage clients. The first uses the local disk as cache to save files in cloud; the second accesses files on remote side directly; and the third stores the same files in local devices. The first type includes DBank data bank and Kingsoft fast disk. Surfing disk accesses the files on the re-

Figure 2. Test environment



mote side directly. DropBox stores all the same files in local devices. After the installation of DBank and Kingsoft fast disk, the client will create a local hard disk cache to synchronize with cloud storage. The local disk cache is a folder or virtual disk partition. Files to be uploaded will save to this folder until the upload process is finished. If the upload process was interrupted by network disruption, the file will upload later. When the synchronization finishes, the file will be removed from local folders. By using the local disk cache, it presents a very fast upload speed, but it does not finish the upload process at the point it reports.

Another type is Surfing disk. After the user logs into cloud storage, the client directly displays the storage space of the remote side. The client will download/upload files directly without a local cache. When the network interrupted, synchronous folder disappears and the local hard disks do not provide the file mapping. Dropbox saves all files in local disks; it syncs every file between different clients.

Our goal is to provide a fair comparison among various online backup cloud storage providers. It will be helpful for customers to choose the best fit cloud providers under their conditions. The test results of these providers may vary from network condition to network condition. A fixed PC terminal with several clients is used to test the download/upload speed of different file types. To evaluate the performance of different cloud storage services, we must solve the following problems:

- How to measure the time of file transmission even when transferred in the background. Most clients upload/download files background, and do not prompt when it starts and ends. We must find an effective way to measure the transfer time;
- How to make the test process reproducible. The test results may change under different network conditions and workloads of cloud storage. The test results should be stable under a described environment. We need to find out the factors affecting the final test results;
- How to measure the acceleration effect of different data pre-processing. Data compress and de-duplication can speed up the transfer rate. We need design different test scenarios to measure these effects.

There are many routers and hubs between end-users and cloud storage service providers, and the transmit speed varies over time. It is difficult but important for users to measure the performance of cloud storage service. But cloud storage nodes do not respond to ping packet, we cannot identify the network conditions from the end-user to the cloud storage provider. Network performance is described in the upload/download speed, packet transfer delay, and packet loss rate.

Before uploading data to cloud, clients may use some pre-processing technology to accelerate the transmission (e.g. compression, de-duplication). We choose different test scenarios to measure these effects. At first, we test the transfer speed of the raw file. We create files full of random numbers to make it difficult to compress or do any de-duplication process. Then we compare the transmit speed of a normal file and the file after it is compressed by winzip. If the client compresses the file before transmission, the speed of a normal file should be much higher than a compressed file with same size. At last, we test the uploading speed of the same files using different users and clients. If the client checks the digital fingerprint of the files before it uploads, the speed of the upload should be very quick.

Cloud storage clients transfer the files in the background and do not prompt when it begins and stops. To measure the performance of the online backup process, we can get the time information by the following methods:

- Use an API hook or tools to record the time of the open/close file;

- Use an API hook to record the time of the disk read;
- Use a sniffer tool to record the time of network packets;
- Analyze the log of clients to get performance data.

Some clients exchange beat heart telegrams at fixed intervals. These telegrams make it difficult to measure the transmission time. We solved this problem by capturing packets transferred between users and cloud storage vendors and removing the repetitive ones. We used a network sniffer to capture all packets transmitted between clients and cloud storage vendors. Wireshark is high performance sniffer software. It can capture packets of a given IP or port. We can get the port number which the client used by "netstat -ano". We use the time between the first and last packet sent to cloud storage services as the transmission time. This method can apply to all kinds of cloud storage clients.

We designed several scenarios to evaluate the performance of online backup services. The basic performance metric is to test the upload/download speeds of given files. We repeat the test at different times; it reflects the stability of performance in different times of a day. We use several test suites to judge if a certain technology was used in the cloud storage system. We upload a raw file and a compressed file to determine if the client compressed the file before transmission. We upload the same files multiple times to check if the client de-duplicates files. And we also try to find out if the cloud storage system de-duplicates in a system scope by uploading the same files through different users.

5. EVALUATION RESULTS AND ANALYSIS

In this section, we illustrate the test results on 5 different cloud storage services. Network condition, file types, and file size will affect the performance of cloud storage services. We test the upload/download speeds of a fixed size file at different times in one day. To prevent de-duplication from speeding up the upload performance when uploading the same files multiple times, we developed a program to generate a fixed size file with random content. We test upload/download speed of files with different sizes and types. These files are filled with random numbers to test the raw speed of upload/download. This test can measure the performance varying range when processing different files.

5.1. Upload/Download Speeds at Different Times

In this section, we evaluate the performance of 5 cloud services at different times. From Figure 3, the results show that performance varies at different times in a day. First of all, Dropbox and SkyDrive is slower than other local vendors. It shows that the user should choose local cloud service providers according to their network condition.

Figure 4 shows that download speed changes during file size change. Surfing is the best one in the condition of file size less than 20M, when with file size is greater than 20M, DBank is the best one. Another conclusion is that download/upload speed varies rapidly when file size change. It shows that the performance is not stable.

From the two figures above, we can induce the following conclusions: Users should choose local storage vendors which can provide better performance. The performance of cloud storage varies at times. Users should test the performance at different times to choose the best fit storage vendor. For cloud storage vendors, all of these five cloud storage providers should improve their transmission stability. Chinese vendors should pay more attention to improving the stability of the download, while American vendors should focus on improving the stability of the upload.

Figure 3. Speed of uploading a file at different times

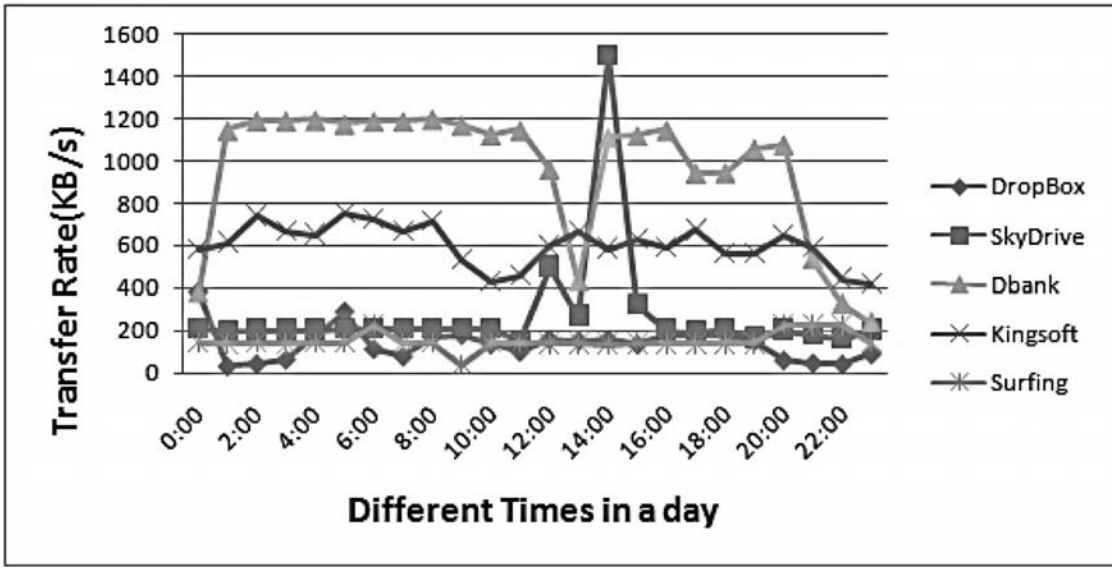
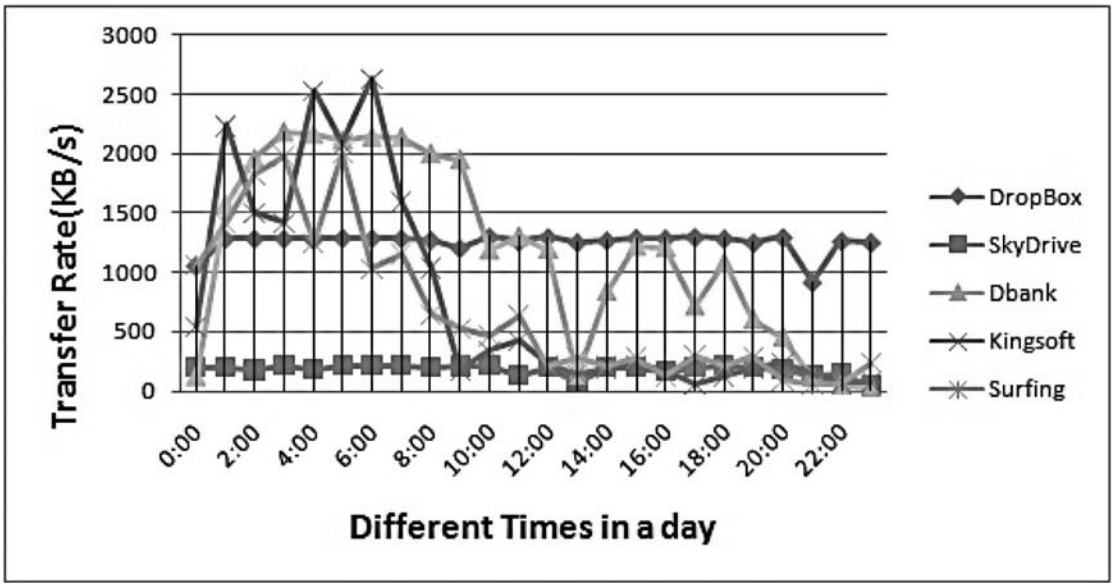


Figure 4. Speed of download a file at different times



5.2. Upload/Download Speed of Different Files

In this section, we test the performance of uploading/downloading different file size and types.

Figure 5 shows that the average upload speed of native storage providers is much higher than their counterparts in America. In the current network condition, DBank is the fastest one and Dropbox is the lowest one.

Figure 6 shows that the download speed of the native storage providers is much higher than their counterparts in America. In the current network condition, Kingsoft is the fastest one and SkyDrive is the lowest one.

Figure 5. Speed of uploading different files

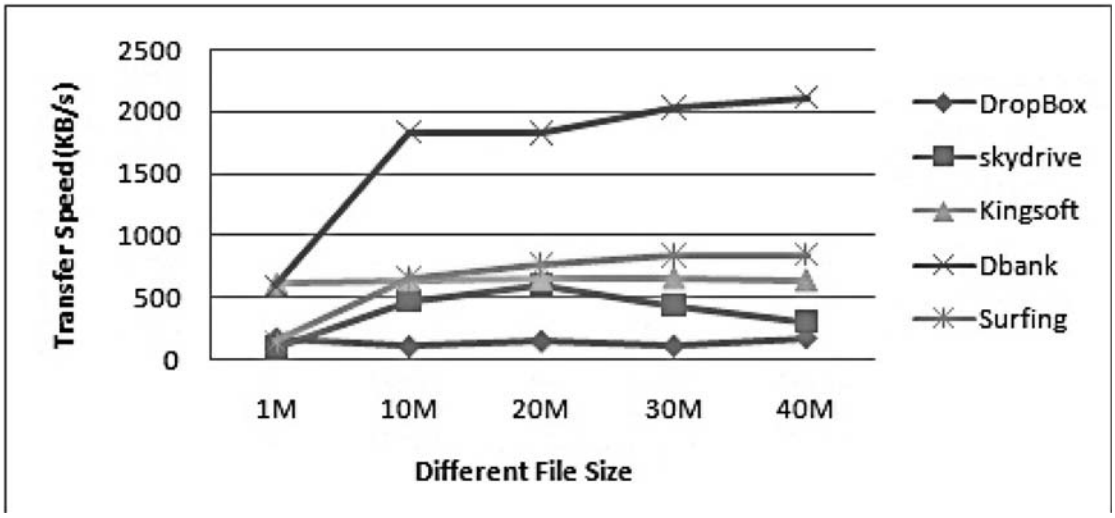
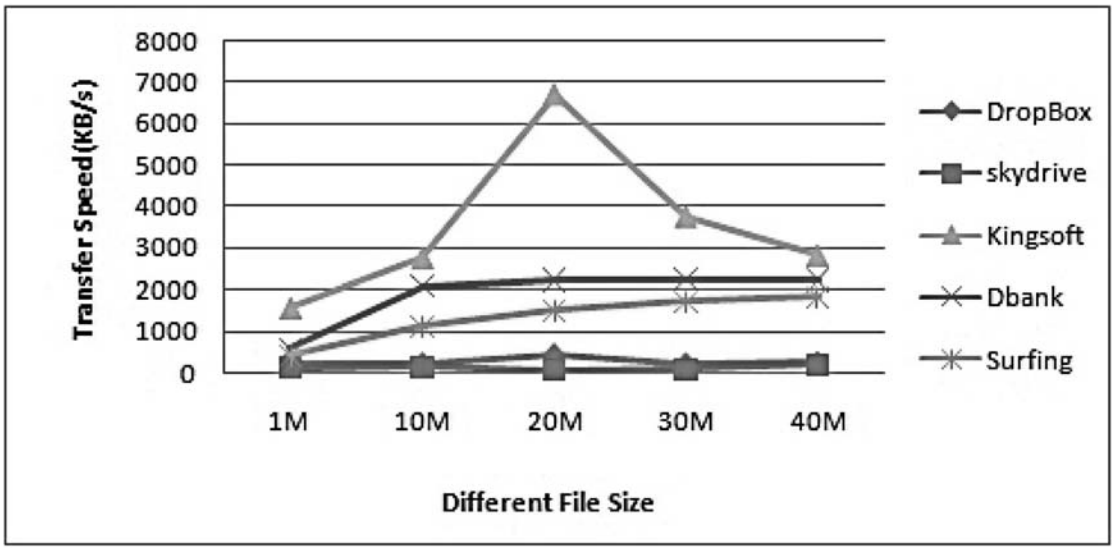


Figure 6. Speed of downloading different files



Also, we find that the transfer rate of larger files is usually faster than the speed of smaller files among three domestic storage providers. The download speed is usually faster than the upload speed which is accordant with the current network situation.

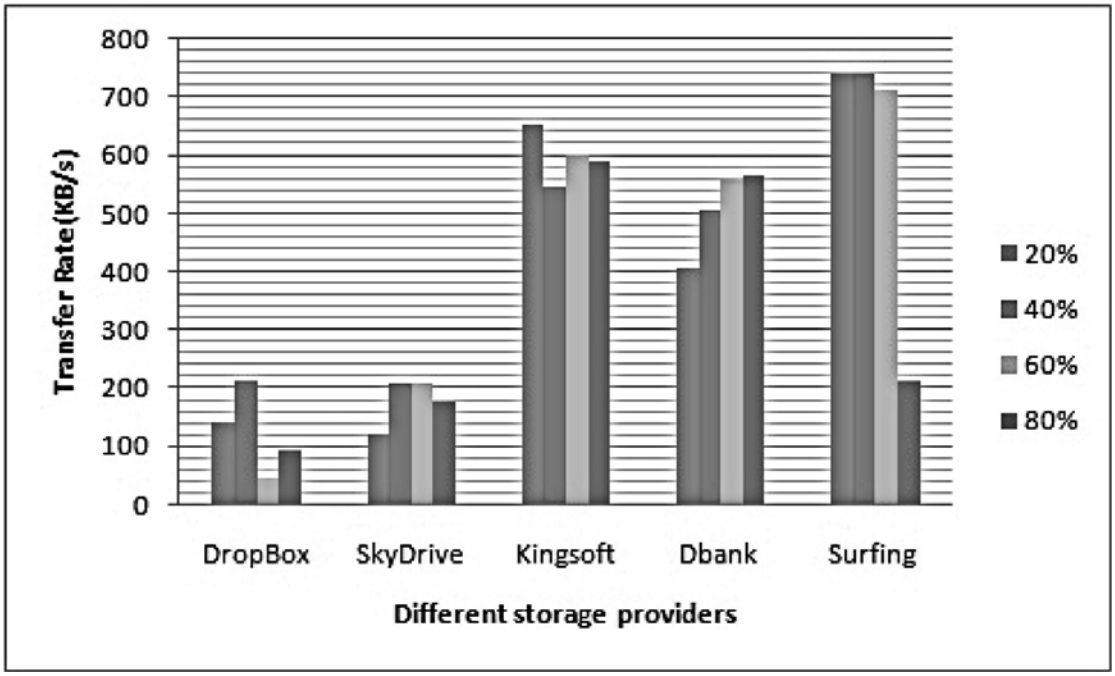
5.3. Affection of CPU Utilization

We test the performance under different CPU utilization. From Figure 7, we can conclude that CPU utilization has no obvious effect on the performance of uploading.

6. PERSPECTIVE OF ACCELERATION TECHNOLOGY

There are many technologies used to decrease the storage and network costs. Compression before upload/download will speed up the transfer process. De-duplication technology

Figure 7. Speed of uploading under different CPU utilization



can distinguish the same files or blocks, these same blocks can be sent once instead of multiple times. Some clients can calculate the fingerprint of files to be uploaded, if there are same files saved in the cloud storage system, it only sends a request to create a reference link to existing files. This is very useful when many users save popular music and movies in their space. All of these technologies can accelerate the backup performance in certain conditions. We designed several test suites to reveal if these technologies were used to accelerate backup performance.

6.1. Compression

In this section, we compare the performance of raw files and compressed files. If the speed of a raw file is much higher than the corresponding compressed file, it means that the clients are compressing files before uploading. From Figure 8, we can find that Dropbox, DBank, Kingsoft, and Surfing may be compressing the files before uploading.

Figure 9 shows the comparison of the download speed of the raw and compress file. We find that Dropbox transfers raw files faster than compressed files; this means Dropbox compressed the file before downloading to clients even though the file was uploaded without compression. Dropbox may store the compressed files in their storage system.

6.2. De-Duplication in User Scope

De-duplication is a specialized data compression technique for eliminating duplicate copies of repeating data. In this section, we use clients to upload the same files from different folders. If the clients check duplicate files before uploading, the uploading process of the second file should be much faster than the first one. Another test is that we delete a file in the cloud and upload the same file again; it tries to find if the cloud storage service kept the cache. If it saves the cache file, when user uploads the same file a second time, the speed will be faster than the first time.

Figure 10 shows the speed of uploading the same file from different folders. The speeds of the second file in directory two are much faster than those in directory one in Dropbox, DBank, Kingsoft, and Surfing. So we can conclude that

Figure 8. Speed of uploading raw file and compressed file

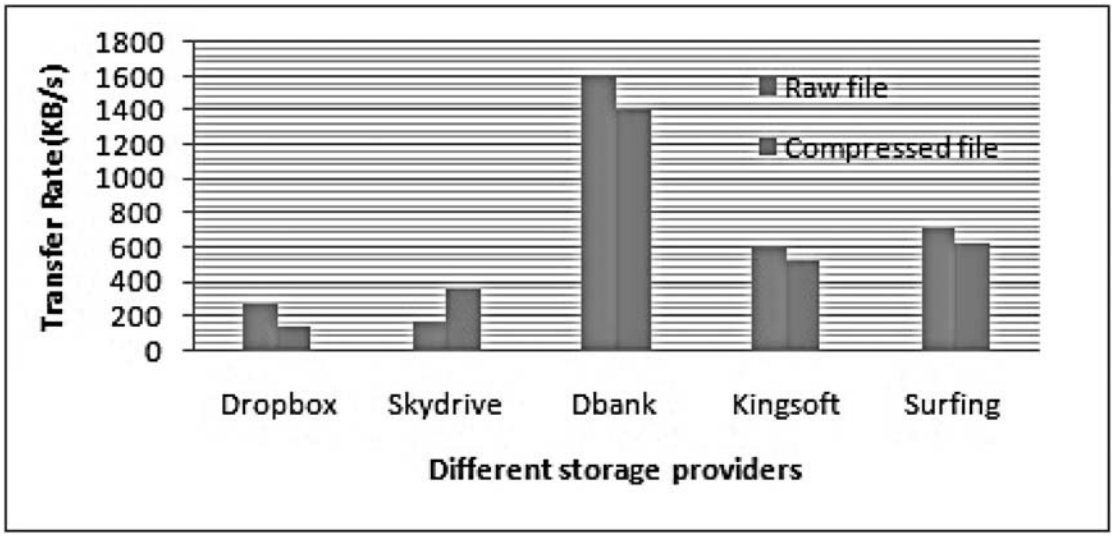
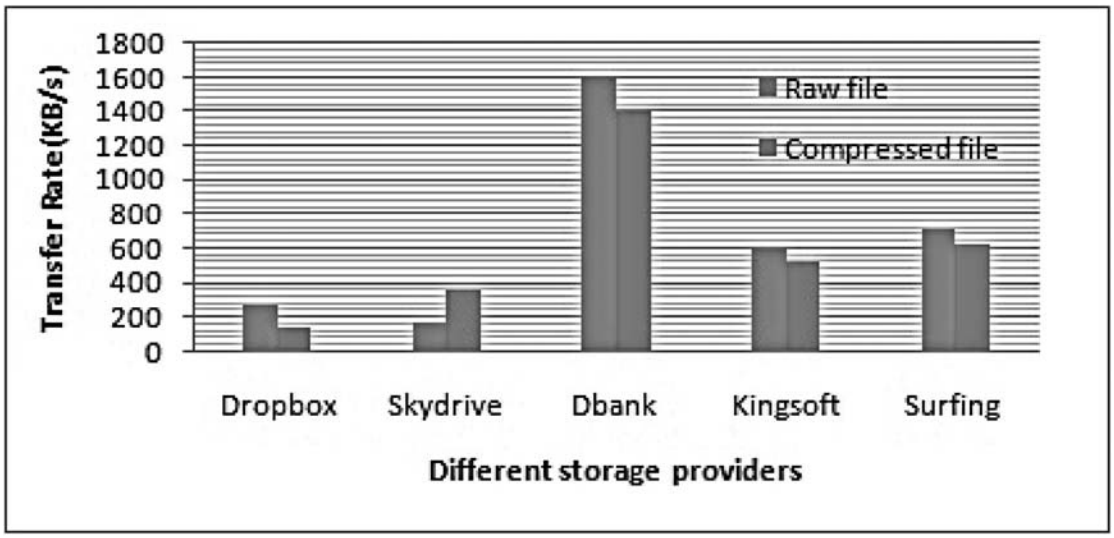


Figure 9. Speed of downloading raw file and compressed file



Dropbox, DBank, Kingsoft, and Surfing accelerate the upload speed through de-duplication or other similar technologies.

We upload a file and delete it both in local disk and cloud storage. Then we upload the same file again. If the cloud storage saved the cache of the deleted files, the file will be identified as a duplicate file. In this condition, the upload speed of the second time should be much faster than the first time. Figure 11 shows the results of the first upload and the second upload after deletion. The speeds of the second time in Dropbox, DBank, Kingsoft, and Surfing are much faster than the first time. So we draw the conclusion that all these storage providers except for SkyDrive accelerate the upload speed by de-duplication or other similar technologies.

Figure 10. Speed of upload same file in different folders

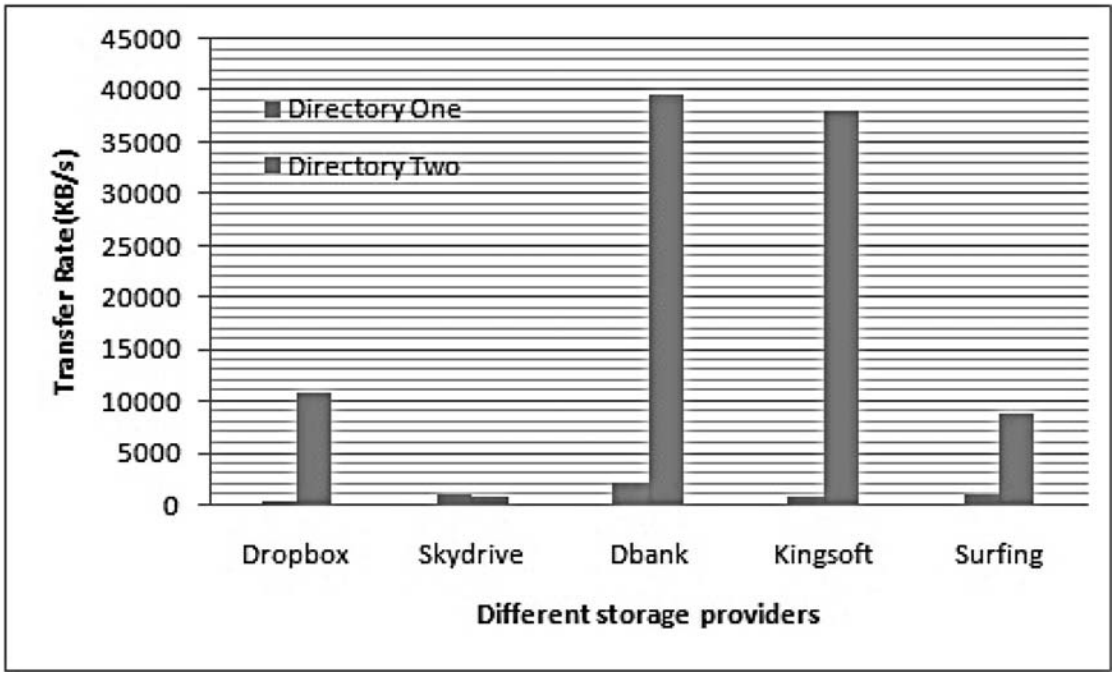
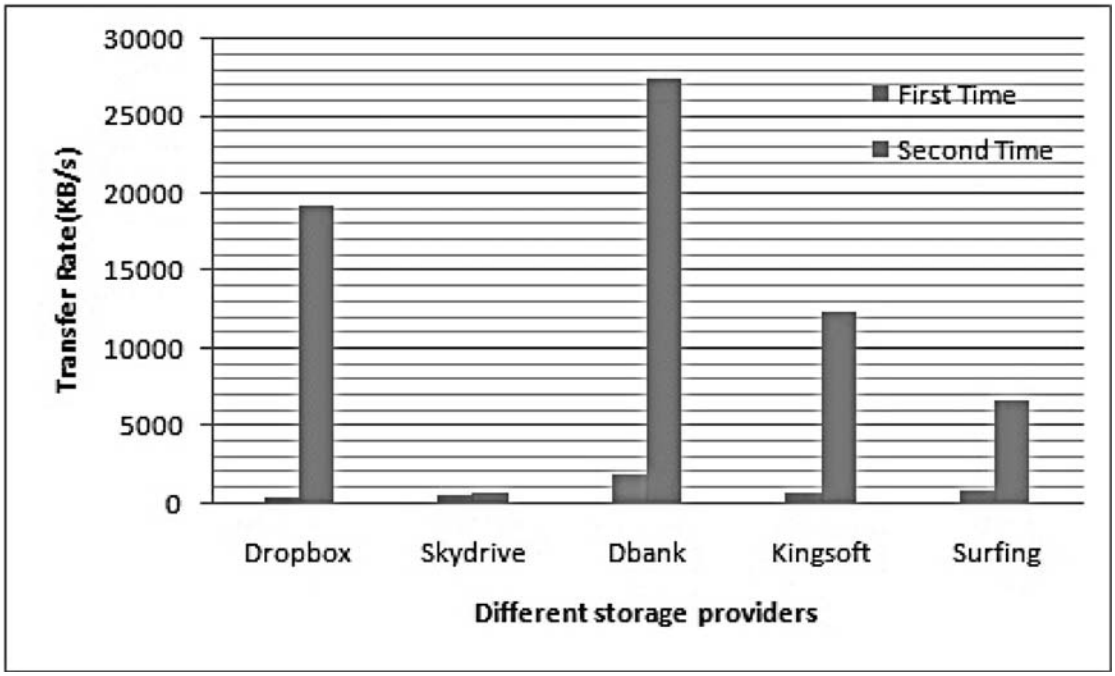


Figure 11. Speed of upload file after deletion



6.3. De-Duplication in System Scope

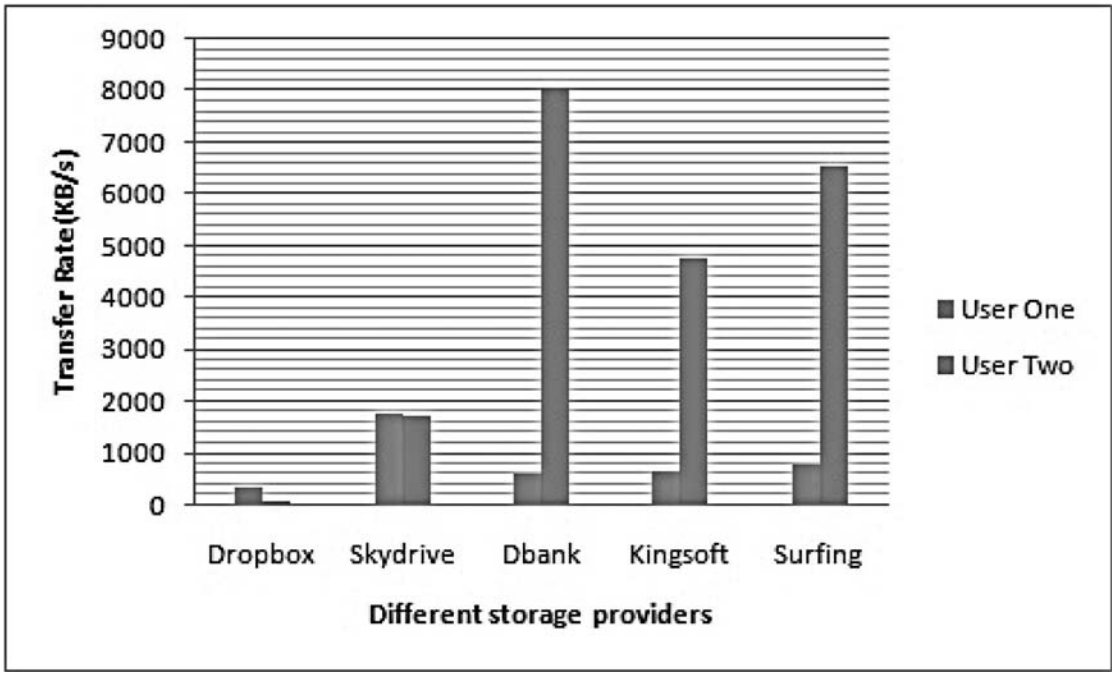
In the previous section, we tested de-duplication technology with one single user. This means that clients or cloud storage search duplicate files in user private space. In this section, we test de-duplication technology with two different users. After user A uploads a file, user B uploads the same file. If cloud storage vendors search duplicate files in the whole system scope, user B can get a much faster speed than user A. This helps cloud storage vendors save storage and network costs, but it has security risks.

Figure 12 shows the speed of uploading the same file through two different users. The speeds of user B are much faster than the speeds of user A in DBank, Kingsoft and Surfing. But for Dropbox and Skydrive, it seems that they did not check for duplicate files in the system scope. We notice that Dropbox check duplicate files in user's space, but it does not check for duplicate files in the system scope. These storage vendors pay more attention on privacy of private data. The results show that DBank, Kingsoft, and Surfing accelerate the upload speed by de-duplication or other similar technologies in the system scope.

7. CONCLUSION

In this paper, we present a method to evaluate the performance of cloud storage clients using different transfer protocols. We evaluate the performance from the view of the end-user, comparing the upload and download speeds of different services. This method can help users find the best provider to accommodate their situation. We evaluate the performance of several cloud storage providers by analyzing the performance of different vendors in different times and different file type and size. From the performance results, we can expose which technologies are used in the cloud storage system to reduce the storage and network cost.

Figure 12. Speed of uploading same file from two users



ACKNOWLEDGMENT

This work was made possible thanks to the NPU Fundamental Research Foundation under Grant No. JC20110227, the Industry Innovation Foundation under Grant No.2011BAH04B05 National High-tech R&D Program of China (863) under Grant No.2013AA01A215,and the National Natural Science Foundation of China under Grant No.61033007. And this work was supported by China Scholarship Council. Xiao Qin's research was supported by the U.S. National Science Foundation under Grants CCF-0845257 (CAREER), CNS-0917137 (CSR), CNS-0757778 (CSR), CCF-0742187 (CPA), CNS-0831502 (CyberTrust), CNS-0855251 (CRI), OCI-0753305 (CI-TEAM), DUE-0837341(CCLI), and DUE-0830831 (SFS).

REFERENCES

Chiu, D., & Agrawal, G. (2010). Evaluating caching and storage options on the amazon web services cloud. In *Proceedings of the 2010 11th IEEE/ACM International Conference on Grid Computing (GRID)* (pp. 17-24). IEEE, 2010.

Cloud storage comparison. (n.d.). Retrieved from http://blog.gsmarena.com/cloud-storage-comparison-google-drive-vs-dropbox-vs-sugarsync-vs-skydrive-vs-box-net/.

Cloud Storage Strategy. (2009). *Cloud storage and the innovator's dilemma.* Retrieved from http://cloudstoragestrategy.com/2009/07/cloud-storage-and-the-innovators-dilemma.html

Cloudarray storage gateway. (n.d.). Retrieved from http://www.twinstrata.com/CloudArray-overview

Cooper, B., Silberstein, A., Tam, E., Ramakrishnan, R., & Sears, R. (2010). Benchmarking cloud serving systems with ycsb. In *Proceedings of the 1st ACM Symposium on Cloud Computing* (pp. 143-154). ACM.

Li, A., Yang, X., Kandula, S., & Zhang, M. (2010). Cloudcmp: Comparing public cloud providers. In *Proceedings of the 10th Annual Conference on Internet Measurement* (pp. 1-14). ACM, 2010.

Vrable, M., Savage, S., & Voelker, G. (2009). G. (2009). Cumulus: Filesystem backup to the cloud. [TOS]. *ACM Transactions on Storage*, *5*(4), 14. doi:10.1145/1629080.1629084

Wikipedia. (n.d.). *Compare of online backup storage service.* Retrieved from http://en.wikipedia.org/wiki/Comparison_of_online_backup_services

Zhao, L., Liu, A., & Keung, J. (2010). Evaluating cloud platform architecture with the care framework. In *Proceedings of the 2010 17th Asia Pacific on Software Engineering Conference (APSEC)* (pp. 60-69). IEEE.