ONLINE STORAGE ON COMPUTERS AS DISTRIBUTED LONG-TERM STORAGE SYSTEM

Ralf Naues, Jörg Keller

Dept. of Mathematics and Computer Science
Parallelism and VLSI Group
University of Hagen
ralf.naues@fernuni-hagen.de
joerg.keller@fernuni-hagen.de

Keywords

Long term storage, Distributed storage, preservation of data

Abstract

Long-term storage is a widely discussed problem. The amount of digital data is growing faster every day. To avoid the loss of data or to avoid the inaccessibility it would be advantageous to have a reliable storage that also grows, a storage without hardware and software limitations and with tolerable costs. An infrastructure that protects the availability and reliability of stored data for a long time.
Projects like SETI@home demonstrate that the internet community is able to share free computer resources like processing power over the internet, why not also free storage capacities.
We think that it is possible to share this free storage on machines that are online in the Internet and that this concept creates a reliable and secure distributed storage system to redundantly store data under the aspects of a long-term storage. The requirements and processes needed are discussed in this short paper.

1. General

We discuss the usability of free storage, i.e. unused disk space, which is available on Internet connected computers around the world for the usage as a long term memory. The aim is to store files reliably and cost effective in a distributed manner by a community effort, not on a proprietary system.

On every computer you find more or less free space on their storage devices, and many computers are constantly online in the internet. With an appropriate driver software everybody could share his free space in a pool for long term storage. If one would store data redundantly on computers around the world, it should be possible to have uninterrupted access to this information. Effective replication strategies provide reliable correct data during their lifetime. Projects like SETI@home (Seti, 2010) or peer-to-peer networks demonstrate that the internet community is willing to share resources.

New security concepts are needed in such an environment. The preservation of data is done by redundancy and by fingerprinting, a hash code stored in a central database. The database itself should also be replicated. Management rules and access rights have to be defined. All together the result should be a protected infrastructure to persevered important data for the future.

Projects like OceanStore (Kubiatowicz, 1999) and their first realization steps Pond (Rhea, 2003) show the technical possibilities. The storage space in this distributed system would have no binding to specific hardware or operating systems. The hardware changes, renews and grows with every new client, who joins the project. This could be an ideal storage place for public documents.

Beside these technical aspects there is a big management issue, someone has to take care and decide what is stored, who has write access and who can read.

In the following sections we would like to discuss some main factors to achieve the aim.

2. System Outlining

2.1. Central database

For storing and retrieving of files, a central database is needed; file object information and space object information are stored here. The database should be replicated and synchronized on multiple hosts like the root servers in the DNS system, in order to avoid a bottleneck or a single point of failure.

File objects, the saved files / documents, are listed here with their metadata. File name, author, relevance, locations, encryption, hash codes would be the typical attributes. Space objects represent the locations with a unique identifier, status, and storage capabilities. When a registered client reports his available space, it is stored here. Used space is recorded here with the corresponding metadata.

In principle this model is already realized with the Google file system (Ghemawat, 2003) or the HDFS from the Apache Hadoop project (Hadoop, 2008). These are master-slave cluster file systems, with data servers for storing and a name server, the master, for organizing and indexing. Yahoo (Hadoop, 2009) shows the function ability with an installation of 4000 servers.

The disadvantage with GFS/HDFS is that it uses only one name server and therefore the administration of storage members is limited to the capacity of this name server. To overcome this, the database could be organized in a hierarchical structure, a top-level database organizes the regions and the region databases provide the service for the local clients. With this structure we follow the principle of data locality. Hosts that provide space for storing should be able to log in locally. These hosts, the registered clients, report their stored data or their free space. A file or document with regional relations is then stored in the same region. The probability is high, that a file with regional content is often requested for read in this region. For a worldwide storage space we have to take the different time zones and work times into account. We can expect a high availability for a regional file, when we have work time in this region. So it makes sense to store locally, but for files with worldwide interest it is also possible to retrieve them from any location via a recursive query over the top-level database. A similar system is used for the check in and allocation of mobile phones.

2.2. Security

Security issues are often discussed problems for distributed systems like this distributed storage. So what are the risks and how to overcome them? If you store distributed, there is no inside or outside anymore, classical scenarios like firewalls or gateways are not available. Also the responsibilities are distributed; we see a lot of administrators with different access possibilities handling such distributed systems.

Out of these circumstances a different security concept is needed, a security by design.

Looking at the aims: integrity of stored objects or data, confidentiality of management data and availability and reliability. For public data and documents the confidentiality of stored objects is not needed, but with extended management could also be achieved.

For integrity the files to be stored should be stored encrypted in principle. In addition, they should then be divided into at least two fragments. These fragments are stored in different locations. Thus, with the retrieval of data, the hash code stored in the central database verifies the individual parts and then the parts are assembled and decoded. Changes in the data, deliberately or by mistake, can be detected and remedial actions for the correction are possible by redundant copies.

The management data, like storage places and hash codes should be confidentially stored on the central servers, with this, we uncouple the information itself from the management information. Only with controlled access over the central servers should it be possible to retrieve content rich data.

Availability and reliability are key issues in distributed systems and are discussed in the next section.

2.3. Storage Reliability

An important criterion for the reliability of storage space is the online time of each machine. With a two-step scoring system a machine can be classified how long it is online and whether it is regularly turned off or not. The first parameter stores the average time the machine is online. It is calculated from the ratio online / offline per day, e.g.:

0 min / 24h = 0 up to 24h / 24h = 1 .

The second parameter is important for those computers that achieve a ratio of 1 or close to 1. These computers will normally be continuous operation servers. So this parameter categorizes the system availability if it is online for more than one day. We provide an example set of categories: A zero corresponds to machines that are regularly being turned off on a daily basis. Computers achieve a value of 1 after 7 days, a 2 after 14 days and a 3 after a month of continuous running. If the system is turned off or goes offline, this factor is decremented. The parameter with value 2 or 3 prevents that a server falls directly into a lower category when it has maintenance or if it is booted. With this scoring system it is possible to divide machines into two classes: the server class and the class of desktop and laptop machines. This information helps to decide where data should be stored, as discussed in the next section.

2.4. Distribution of the data

To effectively use the available space, distribution strategies must be developed so that stored data can be reliably stored and retrieved in adequate time. The base load and thus the basic security of the data must be carried by the server class. Any data that must be available around the clock in the shortest possible time should be stored here with triple redundancy. Thus, double failures could be tolerated. If a storage location is offline for a long time, a replication event should then ensure that the desired numbers of redundant copies are recreated.

Data that are less time critical can be replicated on the class of office computers and the class of desktop and laptop hosts. A time zone strategy can be taken into account that the data is available all around the clock. One strategy could be that the system creates 5 or more replicas. Another strategy is to replicate the data in the course of 24 hours to ensure that zones with high activity, i.e. with many desktop computers turned on, have also the data replicated on their systems. With this strategy much higher network traffic could expected.

2.5. Communication

The free communication between two hosts on the Internet is reasonably restricted by firewalls and gateways. To gain the necessary routes, one could use the HTTP protocol which is usually not blocked. The initial communication then proceeds from the storage location, it reports that it is online with its ID, its reliability status, and free space. In response acknowledge message follows, or a request to retrieve data or a package of data to save.

The frequency of status messages can be adapted depending on the relevance of the storage location. Locations with large data sets and high relevance communicate in shorter intervals than those with less storage traffic. With growing online storage and increasing communication traffic the number of database replications needs to grow to provide scalable data and information flow.

While storing of data should be limited to specific user accounts, the reading of public data should be done in a P2P kind of communication process. This helps to speed up the communication and helps to reduce the workload on the central databases. As discussed in section *2.1 Central database* the data should be kept local, this shortens the communication path for most of the queries.

A proxy cache memory for frequently accessed file information from the central database server could also enhance performance.

3. Available Space

The first question is: how much space can we expect and at which time of the day is it available. The factors which are coming into effect here are the number, type of machines and their specific average spare capacity and their location.

As a first simplification we suggested a classification of machine types: the server class with 24h in operation, the office computers, which are calculated 8h running and the home and mobile computers, which are valued with 2h online per day.

The number of machines and the bandwidth available at their sites has been calculated from sales figures of GFK (GFK, 2009), Gartner (Gartner, 2006), IDC (IDC, 2008) and figures from the

OECD (OECD, 2008). The storage capacity of the equipment reflects the standard configurations of the respective sales period.

The free capacity of the storage media is supposed to be 10% of the total capacity as a first approximation. The mission time of a server is assumed to be at least 3 years.

The server class can be estimated most simply, they are 24-hour operation and assumed that all servers are connected to the internet. If we very conservatively estimate a machine life time of 3 years the calculation in Table 1 shows the estimated availability of disk space in beginning of the year 2009. So, if we could realise to get 10% of installed capacity we come close to 1 exabyte.

The calculation of the space of the desktop and mobile PCs will depend on several factors. These are not usually run in a 24-hour operation, but are dependent on the time zone in which they stand, the local availability of an appropriate range of internet access and the period of use.

Year	**2006**	**2007**	**2008**
Standard hard disk capacity	240 GB	320 GB	500 GB
Server sold (pieces)	8.000.000	9.000.000	10.000.000
Storage capacity in Terabyte	1.875.000	2.812.500	4.882.812,5
10% free capacity in Terabyte	187.500	281.250	488.281,25
free capacity server Worldwide (10%) in Terabyte	934.600		

Table 1: Calculation free capacity 24 h online server

4. Conclusion

With little investment a lot of storage space could be acquired.

The storage pool grows with new generations of hardware. Much of the software needed to create such a system is already developed for similar problems. The online time of computers and their bandwidth to the Internet is constantly growing.

The limitation is in the willingness of the people to share their free space. So to provide an incentive, everyone who donates disk space get easy access to the public store.

Wuala (Wuala, 2010) discusses an online storage system that was mostly developed at ETH Zurich (Swiss Federal Institute of Technology) and provides online storage either by trading idle disk space or by buying additional storage. With more than 200,000,000 files already stored there it shows the possibilities in that market.

Built in support for such a system into the operating systems could ease the usage for the user.

So far, we have only outlined the concept and some possible implementation directives of such a system. Obviously, more theoretical and experimental work is needed, e.g. to provide probabilistic

guarantees of availability given some replication factor, or to tune the system to a tolerable level of overhead given the dynamic nature of the internet.

Currently, as a first step, we are working to simulate such a system with the GoldSim (Goldsim, 2010) simulator, to evaluate whether our proposed categories and replication factors would lead to a stable system.

5. References

Gartner (2006) Report Sales Units Server 2006-2008

GfK (2009) Gesellschaft für Konsumforschung, Report Computer Sales Units 2005-2008.

Ghemawat, S., & Gobioff, H., & Leung, S. (2003) The Google File System, Google, 19th ACM Symposium on Operating Systems Principles

GoldSim Technology Group LLC (2010) Issaquah, Washington 98027-2941, USA
http://www.goldsim.com accessed 10/07/2010

Hadoop Project (2008), Apache Software Foundation
http://hadoop.apache.org accessed 10/07/2010

Hadoop at Yahoo (2009)
http://developer.yahoo.com/hadoop/ accessed 10/07/2010

IDC (2008), Worldwide PC 2007–2011 Forecast Update, Market Research Report, R104-32718

Kubiatowicz, J., & Bindel, D., & Chen, Y., & Eaton, P., & Geels, D., & Gummadi, R., & Rhea, S., & Weatherspoon, H., & Weimer, W., & Wells, C., & Zhao. B. (1999) OceanStore: An extremely wide-area storage system. U.C. Berkeley Technical Report UCB//CSD-00-1102

OECD. (2008) ICT database and Eurostat, Community Survey on ICT usage in households and by individuals

Rhea, S., & Eaton, P., & Geels, D., & Weatherspoon, H., & Zhao, B., & Kubiatowicz. J. (2003) Pond: the OceanStore prototype. Proceedings of the 2nd USENIX Conference on File and Storage Technologies (FAST)

Seti Project at UC Berkeley (2010)
http://setiathome.berkeley.edu/ accessed 30/06/2010

Wuala (2010), Caleido AG, Zurich, Switzerland
http://www.wuala.com accessed 10/07/2010