**Why Do We Need Distributed Storage?**

Yuankai Guan

Today is the last lesson of a distributed system. Looking at what I learned during this semester, I decided to talk about distributed storage.

Before we start, let’s learn some concepts.

**Big Data**

Big data is a term used to refer to data sets that are too large or complex for traditional data-processing application software to adequately deal with.

For data science, big data is an increasingly important direction.

How to analyze big data, how to store and read and write big data is a problem worth studying and solving.

At the same time, this is also a huge challenge to traditional storage methods.

**File System**

In computing, a file system or filesystem controls how data is stored and retrieved. That means, we use the file system to organized data. Without file system, information is messed up together, bloated and difficult to use.

There are many different types of file system, some of them have been designed to specific application. Each one of them have different structure, logic, properties and more.

File systems can be used on numerous different types of storage devices that use different kinds of media. The most common medium we use is hard disk drive. Other kind of media are used included flash memory, optical disk, etc.

But no matter what kind of storage device, their data can be stored in a unit volume is limited, which leads to a very large storage device group when the amount of data is large.

Therefore, local storage is gradually unable to meet the needs.

**Cluster File System**

A clustered file system is a file system which is shared by being simultaneously mounted on multiple servers. There are several approaches to clustering, most of which do not employ a clustered file system (only direct attached storage for each node). Clustered file systems can provide features like location-independent addressing and redundancy which improve reliability or reduce the complexity of the other parts of the cluster. Parallel file systems are a type of clustered file system that spread data across multiple storage nodes, usually for redundancy or performance.

**Distributed File System**

Distributed databases are usually non-relational databases that enable a quick access to data over a large number of nodes. Some distributed databases expose rich query abilities while others are limited to a key-value store semantics.

**Brewer's theorem**

states that it is impossible for a distributed data store to simultaneously provide more than two out of the following three guarantees:

Consistency: Every read receives the most recent write or an error.

Availability: Every request receives a (non-error) response – without the guarantee that it contains the most recent write.

Partition tolerance: The system continues to operate despite an arbitrary number of messages being dropped (or delayed) by the network between nodes.

**Distributed database**

A distributed database is a database in which not all storage devices are attached to a common processor. It may be stored in multiple computers, located in the same physical location; or may be dispersed over a network of interconnected computers. Unlike parallel systems, in which the processors are tightly coupled and constitute a single database system, a distributed database system consists of loosely coupled sites that share no physical components.

System administrators can distribute collections of data (e.g. in a database) across multiple physical locations. A distributed database can reside on organized network servers or decentralized independent computers on the Internet, on corporate intranets or extranets, or on other organization networks. Because distributed databases store data across multiple computers, distributed databases may improve performance at end-user worksites by allowing transactions to be processed on many machines, instead of being limited to one.

**Why do we need Distributed Storage System?**

1. Space

For individual users, but when you need a lot of data, it is difficult to have space to store large amounts of data. Only users with large enough volume can have enough space.

In this case, distributed storage is a good solution. Renting or using space provided by others is a very intuitive option when you are unable to provide enough space.

1. Scalability

In many cases, the amount of data required each time is different. Under the local storage system, the storage space can be increased, and the reduction is not so easy to implement. For a long time, there will be a lot of storage space wasted. In this case, distributed storage is a very effective solution.

The storage space is provided by a unified storage provider, and each person uses it on demand, and the unused space is relocated to the public space, waiting for the next allocation.

1. Performance

How to process large amounts of data at a time is a hot topic now.

For traditional storage systems, all data is put together, with a common table, a common library, and when the amount of data is large enough, processing the data becomes very tricky. So, looking for a reliable solution is imminent.

Distributed databases were born to solve this situation. Compared to distributed storage, its data is related to each other. In my understanding, distributed databases are a special use of distributed storage. The distributed database processes and stores the data in different places. In the unified index, the compute pressure is shared. The distributed database processes and stores the data in different places. In the unified index, the pressure of the operation is shared, which is a more reasonable solution.

1. Security

There is some controversy about security, and the contradiction focuses on whether the providers of storage services are trustworthy. Here, we first identify storage service providers as trustworthy. Uniform security protection, such as cold backup and backup of data, regular scheduled maintenance of storage devices, dedicated security department and security department, can have a special solution in the face of network attacks. Distributed storage is more powerful and more secure than traditional storage because it spreads the cost of security.

1. Ease of Use

For traditional storage methods, individual users need to have knowledge of storage, including not limited to how the storage device operates, how data is stored in the device, hardware maintenance, and security. Sometimes even need electrician knowledge. Individual users tend to be unable to use the storage system because they don't understand it. More prone to problems when working on cross-platform projects. Distributed storage, the special problem is handed over to people with professional knowledge, users only need to understand how the service should be used, so that the ease of use is greatly reduced.

**The most mainstream three distributed storage systems**

1. **AFS**

The Andrew File System (AFS) is a distributed file system which uses a set of trusted servers to present a homogeneous, location-transparent file name space to all the client workstations. It was developed by Carnegie Mellon University as part of the Andrew Project. Originally named "Vice", AFS is named after Andrew Carnegie and Andrew Mellon. Its primary use is in distributed computing.

AFS with good scalability can bring performance and usability improvements to clients. It is modular, so it is not required to run all server processes on each server.

However, it is worth mentioning that the disadvantage of AFS is that the administrator interface is not friendly enough and needs more professional knowledge to support.

1. **GFS**

GFS (Google File System), known as the Google File System, is the primary technology and file system for implementing unstructured data. Its performance, scalability, reliability and availability have been recognized. It runs mainly on a large number of ordinary machines running Linux systems, which can greatly reduce its hardware costs.

1. **Lustre**

Lustre is a type of parallel distributed file system, generally used for large-scale cluster computing. The name Lustre is a portmanteau word derived from Linux and cluster.