DATTA MEGHE COLLEGE OF ENGINEERING
Department of Computer Engineering

Academic Year: 2018-19 (EVEN SEM)

MINI-PROJECT

Year: SE **Sem: IV** **Subject: COA**

Title of the Problem: Cloud Storage Architecture	
Class / Div. / Roll No.	**Name of the Students**
SE/B/41	Satyam Singh
SE/B/46	Subham Ujalmbkar
SE/B/47	Swapnil Undage
SE/B/49	Rajan Yadav

Introduction

Cloud storage is a model of computer data storage in which the digital datais stored in logical pools. The physical storage spans multiple servers(sometimes in multiple locations), and the physical environment is typically owned and managed by a hosting company. These cloud storage providers are responsible for keeping the data available and accessible, and the physical environment protected and running. People and organizations buy or lease storage capacity from the providers to store user, organization, or application data.

Cloud storage services may be accessed through a colocated cloud computing service, a web service application programming interface (API) or by applications that utilize the API, such as cloud desktop storage, a cloud storage gateway or Web-based content management systems.

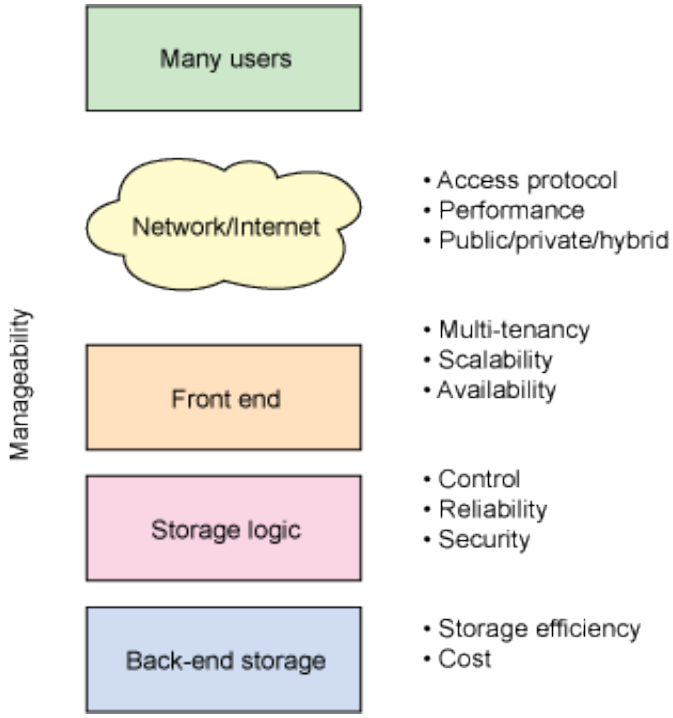
Architecture

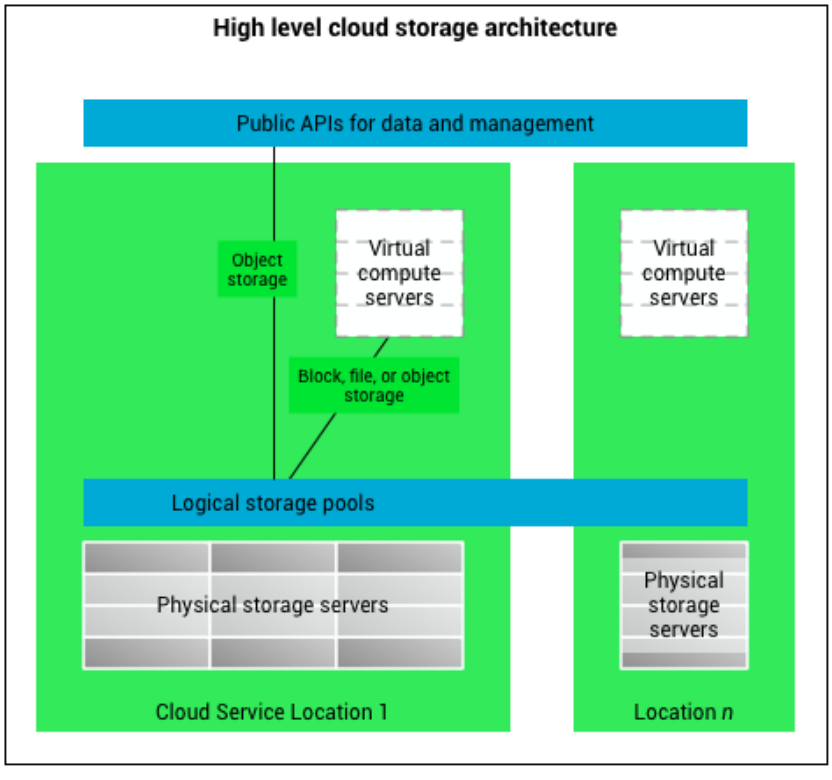
Cloud storage is based on highly virtualized infrastructure and is like broader cloud computing in terms of accessible interfaces, near-instant elasticity and scalability, multi-tenancy, and metered resources. Cloud storage services can be utilized from an off-premises service (Amazon S3) or deployed on-premises (ViON Capacity Services).

Cloud storage typically refers to a hosted object storage service, but the term has broadened to include other types of data storage that are now available as a service, like block storage.

Object storage services like Amazon S3, Oracle Cloud Storage and Microsoft Azure Storage, object storage software like Openstack Swift, object storage systems like EMC Atmos, EMC ECS and Hitachi Content Platform, and distributed storage research projects like OceanStore and VISION Cloud are all examples of storage that can be hosted and deployed with cloud storage characteristics.

Generic cloud storage architecture





Cloud storage is:

•Made up of many distributed resources, but still acts as one, either in a federated or a cooperative storage cloud architecture

•Highly fault tolerant through redundancy and distribution of data

•Highly durable through the creation of versioned copies

•Typically eventually consistent with regard to data replicas

Cloud storage characteristics

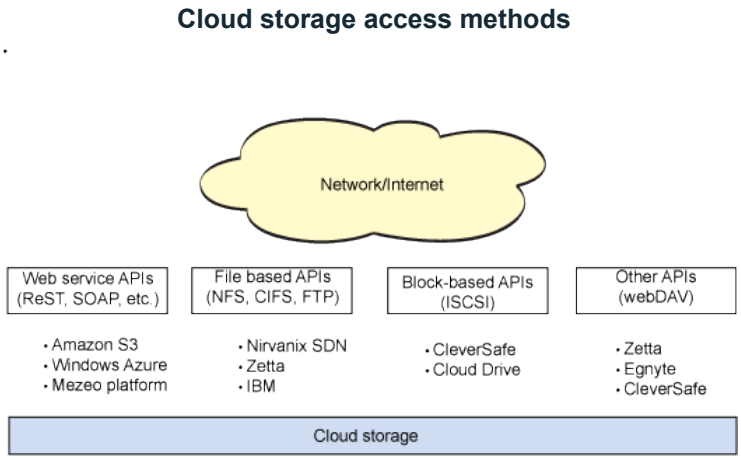
Characteristic	Description
Manageability	The ability to manage a system with minimal resources
Access method	Protocol through which cloud storage is exposed
Performance	Performance as measured by bandwidth and latency
Multi-tenancy	Support for multiple users (or tenants)
Scalability	Ability to scale to meet higher demands or load in a graceful manner
Data availability	Measure of a system’s up time
Control	Ability to control a system — in particular, to configure for cost, performance, or other characteristics
Storage efficiency	Measure of how efficiently the raw storage is used
Cost	Measure of the cost of the storage (commonly in $/GB)

Manageability

One key focus of cloud storage is cost. If a client can buy and manage storage locally compared to leasing it in the cloud, the cloud storage market disappears. But cost can be divided into two high-level categories: the cost of the physical storage ecosystem itself and the cost of managing it. The management cost is hidden but represents a long-term component of the overall cost. For this reason, cloud storage must be self-managing to a large extent. The ability to introduce new storage where the system automatically self-configures to accommodate it and the ability to find and self-heal in the presence of errors are critical. Concepts such as autonomic computing will have a key role in cloud storage architectures in the future.

Access method

One of the most striking differences between cloud storage and traditional storage is the means by which it's accessed. Most providers implement multiple access methods, but Web service APIs are common. Many of the APIs are implemented based on REST principles, which imply an object-based scheme developed on top of HTTP (using HTTP as a transport). REST APIs are stateless and therefore simple and efficient to provide. Many cloud storage providers implement REST APIs, including Amazon Simple Storage Service (Amazon S3), Windows Azure™, and Mezeo Cloud Storage Platform.



Performance

There are many aspects to performance, but the ability to move data between a user and a remote cloud storage provider represents the largest challenge to cloud storage. The problem, which is also the workhorse of the Internet, is TCP. TCP controls the flow of data based on packet acknowledgements from the peer endpoint. Packet loss, or late arrival, enables congestion control, which further limits performance to avoid more global networking issues. TCP is ideal for moving small amounts of data through the global Internet but is less suitable for larger data movement, with increasing round-trip time (RTT).

Multi-tenancy

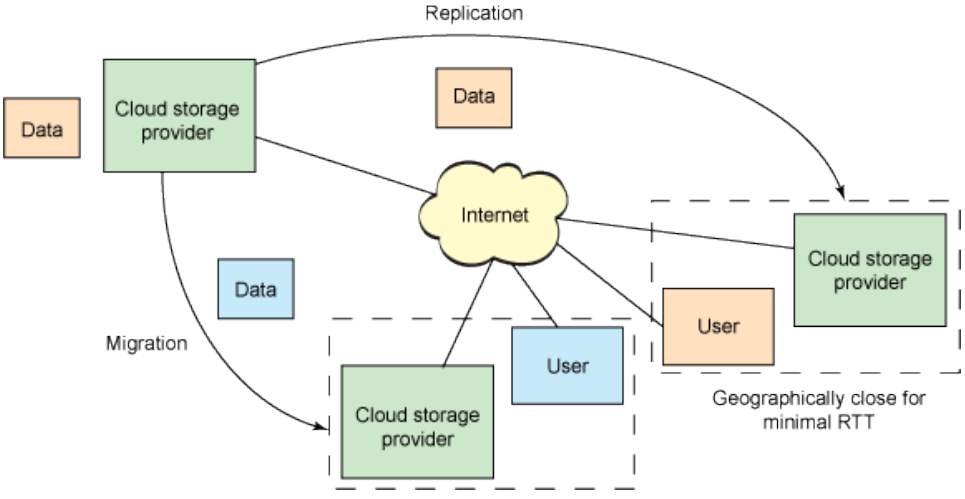
One key characteristic of cloud storage architectures is called *multi-tenancy.* This simply means that the storage is used by many users (or multiple “tenants”). Multi-tenancy applies to many layers of the cloud storage stack, from the application layer, where the storage namespace is segregated among users, to the storage layer, where physical storage can be segregated for particular users or classes of users. Multi-tenancy even applies to the networking infrastructure that connects users to storage to permit quality of service and carving bandwidth to a particular user.

Scalability

You can look at scalability in a number of ways, but it is the on-demand view of cloud storage that makes it most appealing. The ability to scale storage needs (both up and down) means improved cost for the user and increased complexity for the cloud storage provider.

Scalability must be provided not only for the storage itself (functionality scaling) but also the bandwidth to the storage (load scaling). Another key feature of cloud storage is geographic distribution of data (geographic scalability), allowing the data to be nearest the users over a set of cloud storage data centers (via migration). For read-only data, replication and distribution are also possible (as is done using content delivery networks).

Scalability of cloud storage

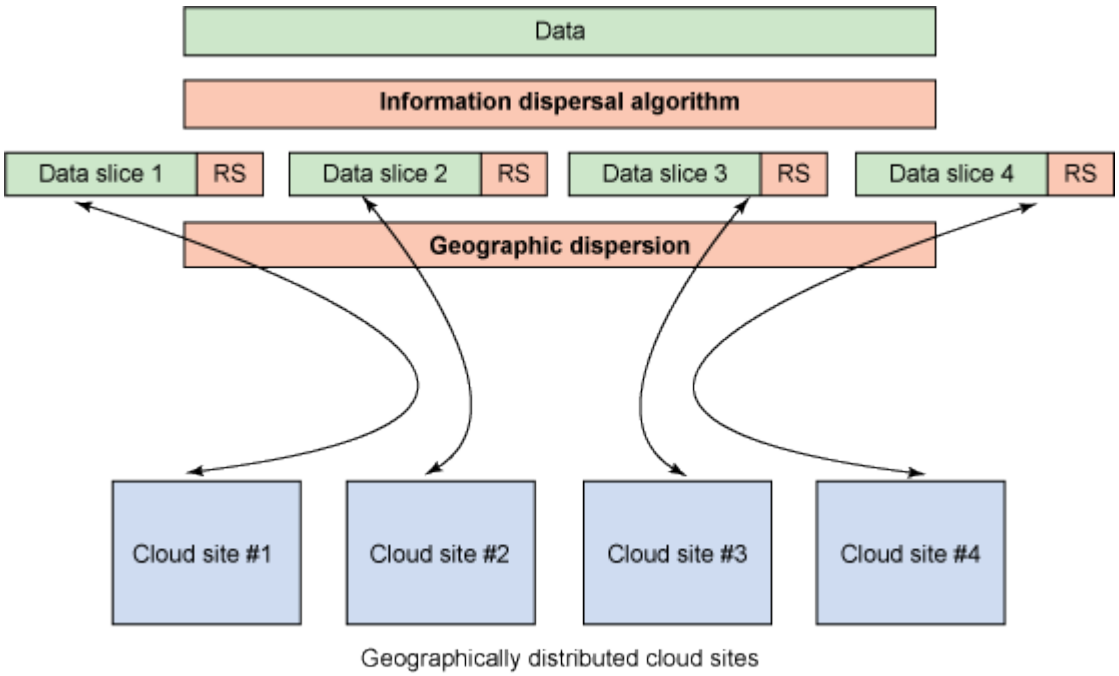


Availability

Once a cloud storage provider has a user's data, it must be able to provide that data back to the user upon request. Given network outages, user errors, and other circumstances, this can be difficult to provide in a reliable and deterministic way.

There are some interesting and novel schemes to address availability, such as information dispersal. Cleversafe, a company that provides private cloud storage (discussed later), uses the Information Dispersal Algorithm (IDA) to enable greater availability of data in the face of physical failures and network outages. IDA, which was first created for telecommunication systems by Michael Rabin, is an algorithm that allows data to be sliced with Reed-Solomon codes for purposes of data reconstruction in the face of missing data. Further, IDA allows you to configure the number of data slices, such that a given data object could be carved into four slices with one tolerated failure or 20 slices with eight tolerated failures. Similar to RAID, IDA permits the reconstruction of data from a subset of the original data, with some amount of overhead for error codes (dependent on the number of tolerated failures).

Cleversafe's approach to extreme data availability



The downside of IDA is that it is processing intensive without hardware acceleration. Replication is another useful technique and is implemented by a variety of cloud storage providers. Although replication introduces a large amount of overhead (100%), it's simple and efficient to provide.

Control

A customer's ability to control and manage how his or her data is stored and the costs associated with it is important. Numerous cloud storage providers implement controls that give users greater control over their costs.

Amazon implements Reduced Redundancy Storage (RRS) to provide users with a means of minimizing overall storage costs. Data is replicated within the Amazon S3 infrastructure, but with RRS, the data is replicated fewer times with the possibility for data loss. This is ideal for data that can be recreated or that has copies that exist elsewhere.

Efficiency

Storage efficiency is an important characteristic of cloud storage infrastructures, particularly with their focus on overall cost. The next section speaks to cost specifically, but this characteristic speaks more to the efficient use of the available resources over their cost.

To make a storage system more efficient, more data must be stored. A common solution is data reduction, whereby the source data is reduced to require less physical space. Two means to achieve this include *compression*—the reduction of data through encoding the data using a different representation—and *de-duplication*—the removal of any identical copies of data that may exist. Although both methods are useful, compression involves processing (re-encoding the data into and out of the infrastructure), where de-duplication involves calculating signatures of data to search for duplicates.

Cost

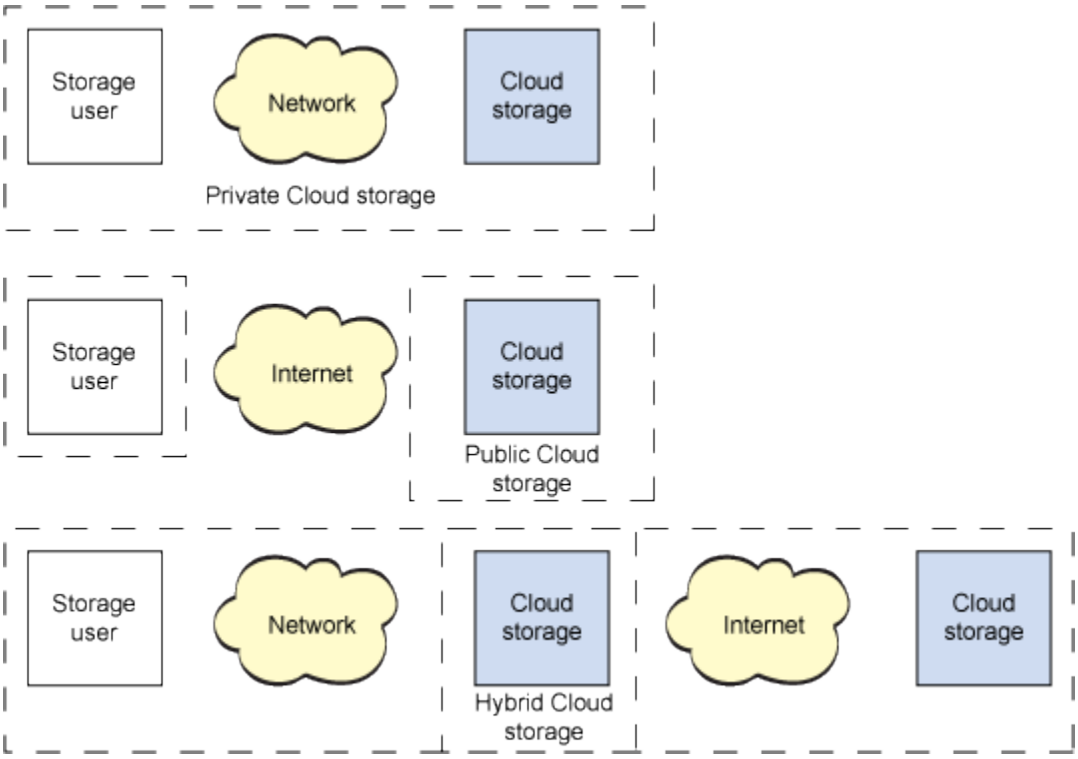
One of the most notable characteristics of cloud storage is the ability to reduce cost through its use. This includes the cost of purchasing storage, the cost of powering it, the cost of repairing it (when drives fail), as well as the cost of managing the storage. When viewing cloud storage from this perspective (including SLAs and increasing storage efficiency), cloud storage can be beneficial in certain use models.

Cloud storage models

Thus far, We've talked primarily about cloud storage providers, but there are models for cloud storage that allow users to maintain control over their data. Cloud storage has evolved into three categories, one of which permits the merging of two categories for a cost-efficient and secure option.

Much of this article has discussed public cloud storage providers, which present storage infrastructure as a leasable commodity (both in terms of long-term or short-term storage and the networking bandwidth used within the infrastructure). Private clouds use the concepts of public cloud storage but in a form that can be securely embedded within a user's firewall. Finally, hybrid cloud storage permits the two models to merge, allowing policies to define which data must be maintained privately and which can be secured within public clouds

Cloud storage models



The cloud models are shown graphically in Figure 6. Examples of public cloud storage providers include Amazon (which offer storage as a service). Examples of private cloud storage providers include IBM, Parascale, and Cleversafe (which build software and/or hardware for internal clouds). Finally, hybrid cloud providers include Egnyte, among others.

Advantages

• Companies need only pay for the storage they actually use, typically an average of consumption during a month. This does not mean that cloud storage is less expensive, only that it incurs operating expenses rather than capital expenses.

• Businesses using cloud storage can cut their energy consumption by up to 70% making them a more green business. Also at the vendor level they are dealing with higher levels of energy so they will be more equipped with managing it in order to keep their own costs down as well.

• Organizations can choose between off-premises and on-premises cloud storage options, or a mixture of the two options, depending on relevant decision criteria that is complementary to initial direct cost savings potential; for instance, continuity of operations (COOP), disaster recovery (DR), security (PII, HIPAA, SARBOX, IA/CND), and records retention laws, regulations, and policies.

• Storage availability and data protection is intrinsic to object storage architecture, so depending on the application, the additional technology, effort and cost to add availability and protection can be eliminated.

• Storage maintenance tasks, such as purchasing additional storage capacity, are offloaded to the responsibility of a service provider.

• Cloud storage provides users with immediate access to a broad range of resources and applications hosted in the infrastructure of another organization via a web service interface.

• Cloud storage can be used for copying virtual machine images from the cloud to on-premises locations or to import a virtual machine image from an on-premises location to the cloud image library. In addition, cloud storage can be used to move virtual machine images between user accounts or between data centers.

• Cloud storage can be used as natural disaster proof backup, as normally there are 2 or 3 different backup servers located in different places around the globe.

• Cloud storage can be mapped as a local drive with the WebDAV protocol. It can function as a central file server for organizations with multiple office locations.

Potential concerns

- **Data security**
- **Longevity**
- **Accessibility**
- **Other concerns** as Piracy, copyright infringement, Personal Identity Information maybe at risk.

Options

There are several options available to avoid such issues. One option is to use a private cloud instead of a public cloud (there are as well providers that offer private cloud storage). Another option is to ingest data in encrypted format where the key is hold within your on premise infrastructure. Access is often by use of cloud storage gateways that are on premise. Such gateways have options not only to encrypt and compress the data prior of transfer but as well mirror the storage across multiple cloud storage providers and remove the risk of a single providers shutdown. Gateways offer as well the option to cache data on a most recently used algorithm on premise. Along with data analytics data is cached and fetched on a most valuable form instead of recently used only form.