Cloud infrastructure

Chapter 4

Content

- Architectural Design of Compute and Storage Clouds
- Layered Cloud Architecture Development
- Design Challenges
- Inter Cloud Resource Management
- Resource Provisioning and Platform Deployment
- Global Exchange of Cloud Resources

Architectural Design of Compute and Storage Clouds

It involves creating a framework and infrastructure that allows for efficient management and utilization of computing resources and storage capacity.

This design is crucial for building scalable, reliable, and cost-effective cloud platforms that cater to various user needs.

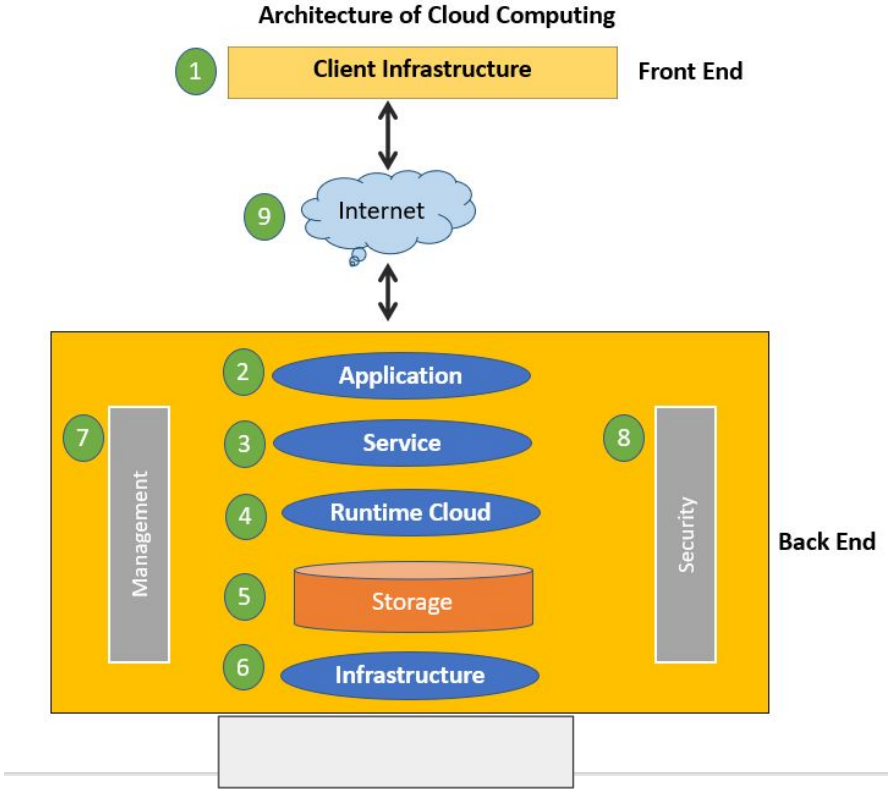
Compute Cloud Architecture

1. **Resource Pools**: Divide physical computing resources, such as servers and virtual machines (VMs), into resource pools. These pools can be based on factors like CPU, memory, and network capacity. Resource pools allow for efficient allocation of resources to different users and applications.
2. **Hypervisor Layer**: Use hypervisor technology to manage and virtualize physical resources. Hypervisors create and manage VMs, enabling multiple operating systems and applications to run on a single physical host.
3. **Orchestration and Management Layer**: This layer handles provisioning, scaling, and management of VMs. It includes orchestration tools like Kubernetes, OpenStack, or VMWare vSphere, which automate the deployment and scaling of applications based on demand.

Compute Cloud Architecture

4. **Load Balancing and Scaling**: Implement load balancers to distribute incoming traffic across multiple VMs to ensure optimal performance and availability. Auto-scaling mechanisms can dynamically adjust the number of VMs based on demand.
5. **Networking Infrastructure**: Set up a network infrastructure that allows VMs to communicate with each other and external systems. Virtual networks, subnets, and firewalls are configured to isolate and secure traffic.
6. **Monitoring and Logging**: Integrate monitoring tools to track resource utilization, performance metrics, and overall system health. Log management systems collect and analyze logs from various components for troubleshooting and auditing.

Generic cloud architecture



Storage Cloud Architecture

1. **Storage Nodes**: Deploy physical storage nodes equipped with disks or solid-state drives (SSDs). These nodes provide the raw storage capacity and are organized into clusters.
2. **Storage Virtualization**: Abstract physical storage resources into virtualized storage pools. This enables administrators to allocate storage to users and applications without exposing the underlying hardware details.
3. **Data Replication and Redundancy**: Implement data replication techniques such as RAID (Redundant Array of Independent Disks) or erasure coding to ensure data redundancy and availability in case of hardware failures.
4. **Object, Block, and File Storage**: Offer different types of storage services, including object storage (for unstructured data like images and videos), block storage (for VM disk images), and file storage (for shared file systems).

Storage Cloud Architecture

5. **Data Management**: Provide tools for users to manage their data, including data migration, backup, and restoration. Offer data lifecycle management to automate tasks like data archiving and expiration.
6. **Data Security**: Implement encryption mechanisms to secure data at rest and in transit. Access control mechanisms ensure that only authorized users and applications can access specific data.
7. **Scalability**: Design storage systems to scale horizontally by adding more storage nodes to accommodate growing data needs. Implement load balancing to distribute data access requests across multiple nodes.
8. **Backup and Disaster Recovery**: Develop strategies for regular data backups and disaster recovery plans to mitigate data loss and system downtime.

Layered Cloud Architecture Development

An Internet cloud is envisioned as a public cluster of servers provisioned on demand to perform collective web services or distributed applications using data-center resources.

Layered cloud architecture is a design approach that is used in cloud computing to organize the different components of cloud infrastructure.

Layered Cloud Architecture Development

It is possible to organize all the concrete realizations of cloud computing into a layered view covering the entire, from hardware appliances to software systems.

All of the physical manifestations of cloud computing can be arranged into a layered picture that encompasses anything from software systems to hardware appliances. Utilizing cloud resources can provide the “computer horsepower” needed to deliver services.

This layer is frequently done utilizing a data center with dozens or even millions of stacked nodes. Because it can be constructed from a range of resources, including clusters and even networked PCs, cloud infrastructure can be heterogeneous in character. The infrastructure can also include database systems and other storage services.

Layered Cloud Architecture Development

The core middleware, whose goals are to create an optimal runtime environment for applications and to best utilize resources, manages the physical infrastructure.

Virtualization technologies are employed at the bottom of the stack to ensure runtime environment modification, application isolation, sandboxing, and service quality. At this level, hardware virtualization is most frequently utilized.

The distributed infrastructure is exposed as a collection of virtual computers via hypervisors, which control the pool of available resources. By adopting virtual machine technology, it is feasible to precisely divide up hardware resources like CPU and memory as well as virtualize particular devices to accommodate user and application needs.

LAYERED CLOUD ARCHITECTURE DESIGN

Design Challenges

1. Data Security and Privacy
2. Cost Management
3. Multi-Cloud Environments
4. Performance Challenges
5. Interoperability and Flexibility
6. High Dependence on Network
7. Lack of Knowledge and Expertise

Data Security and Privacy

Data security is a major concern when switching to cloud computing.

User or organizational data stored in the cloud is critical and private.

Even if the cloud service provider assures data integrity, it is your responsibility to carry out user authentication and authorization, identity management, data encryption, and access control.

Security issues on the cloud include identity theft, data breaches, malware infections, and a lot more which eventually decrease the trust amongst the users of your applications.

This can in turn lead to potential loss in revenue alongside reputation and stature.

Cost Management

Even as almost all cloud service providers have a “Pay As You Go” model, which reduces the overall cost of the resources being used, there are times when there are huge costs incurred to the enterprise using cloud computing.

When there is under optimization of the resources, let’s say that the servers are not being used to their full potential, add up to the hidden costs.

If there is a degraded application performance or sudden spikes or overages in the usage, it adds up to the overall cost.

Multi-Cloud Environments

Due to an increase in the options available to the companies, enterprises not only use a single cloud but depend on multiple cloud service providers.

Most of these companies use hybrid cloud tactics and close to 84% are dependent on multiple clouds.

This often ends up being hindered and difficult to manage for the infrastructure team.

Performance Challenges

Performance is an important factor while considering cloud-based solutions.

If the performance of the cloud is not satisfactory, it can drive away users and decrease profits.

Even a little latency while loading an app or a web page can result in a huge drop in the percentage of users.

This latency can be a product of inefficient load balancing, which means that the server cannot efficiently split the incoming traffic so as to provide the best user experience.

Interoperability and Flexibility

When an organization uses a specific cloud service provider and wants to switch to another cloud-based solution, it often turns up to be a tedious procedure since applications written for one cloud with the application stack are required to be re-written for the other cloud.

There is a lack of flexibility from switching from one cloud to another due to the complexities involved.

High Dependence on Network

Since cloud computing deals with provisioning resources in real-time, it deals with enormous amounts of data transfer to and from the servers.

This is only made possible due to the availability of the high-speed network.

Although these data and resources are exchanged over the network, this can prove to be highly vulnerable in case of limited bandwidth or cases when there is a sudden outage.

Lack of Knowledge and Expertise

Due to the complex nature and the high demand for research working with the cloud often ends up being a highly tedious task.

It requires immense knowledge and wide expertise on the subject.

Although there are a lot of professionals in the field they need to constantly update themselves.

Cloud computing is a highly paid job due to the extensive gap between demand and supply.

Inter Cloud Resource Management

Inter Cloud Resource Management

A theoretical model for cloud computing services is referred to as the “inter-cloud” or “cloud of clouds”.

Combining numerous various separate clouds into a single fluid mass for on-demand operations.

Simply put, the inter-cloud would ensure that a cloud could utilize resources outside of its range using current agreements with other cloud service providers.

There are limits to the physical resources and the geographic reach of any one cloud.

Need of Inter-Cloud

Due to their Physical Resource limits, Clouds have certain Drawbacks:

- When a cloud’s computational and storage capacity is completely depleted, it is unable to serve its customers.
- The Inter-Cloud addresses these circumstances when one cloud would access the computing, storage, or any other resource of the infrastructures of other clouds.

Benefits of the Inter-Cloud Environment include

- Avoiding vendor lock-in to the cloud client
- Having access to a variety of geographical locations, as well as enhanced application resiliency.
- Better service level agreements (SLAs) to the cloud client
- Expand-on-demand is an advantage for the cloud provider.

Resource Provisioning and Platform Deployment

Resource Provisioning and Platform Deployment

The allocation of resources and services from a cloud provider to a customer is known as resource provisioning in cloud computing, sometimes called cloud provisioning.

Resource provisioning is the process of choosing, deploying, and managing software (like load balancers and database server management systems) and hardware resources (including CPU, storage, and networks) to assure application performance.

Importance of Cloud Provisioning

- Scalability: Being able to actively scale up and down with flux in demand for resources is one of the major points of cloud computing
- Speed: Users can quickly spin up multiple machines as per their usage without the need for an IT Administrator
- Savings: Pay as you go model allows for enormous cost savings for users, it is facilitated by provisioning or removing resources according to the demand

Challenges of Cloud Provisioning

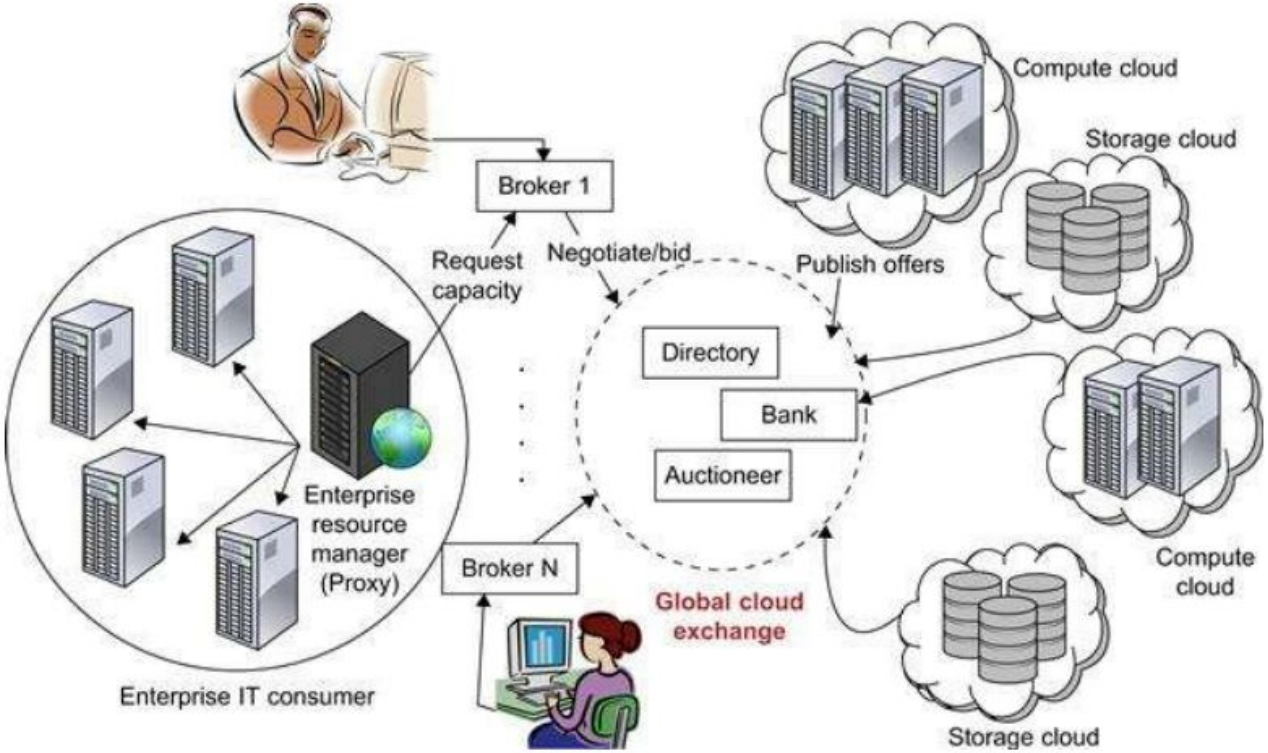
- Complex management: Cloud providers have to use various different tools and techniques to actively monitor the usage of resources
- Policy enforcement: Organisations have to ensure that users are not able to access the resources they shouldn’t.
- Cost: Due to automated provisioning costs may go very high if attention isn’t paid to placing proper checks in place. Alerts about reaching the cost threshold are required.

Tools for Cloud Provisioning

- Google Cloud Deployment Manager
- IBM Cloud Orchestrator
- AWS CloudFormation
- Microsoft Azure Resource Manager

Global Exchange of Cloud Resources

Global Exchange of Cloud Resources



Global Exchange of Cloud Resources

Cloud Exchange (CEx) serves as a market maker, bringing service providers and users together.

The University of Melbourne proposed it under Intercloud architecture (Cloudbus).

It supports brokering and exchanging cloud resources for scaling applications across multiple clouds.

It aggregates the infrastructure demands from application brokers and evaluates them against the available supply.

Global Exchange of Cloud Resources

Entities of the Global exchange of cloud resources

1. Market directory:

   A market directory is an extensive database of resources, providers, and participants using the resources. Participants can use the market directory to find providers or customers with suitable offers.

2. Auctioneers:

   Auctioneers clear bids and ask from market participants regularly. Auctioneers sit between providers and customers and grant the resources available in the Global exchange of cloud resources to the highest bidding customer.

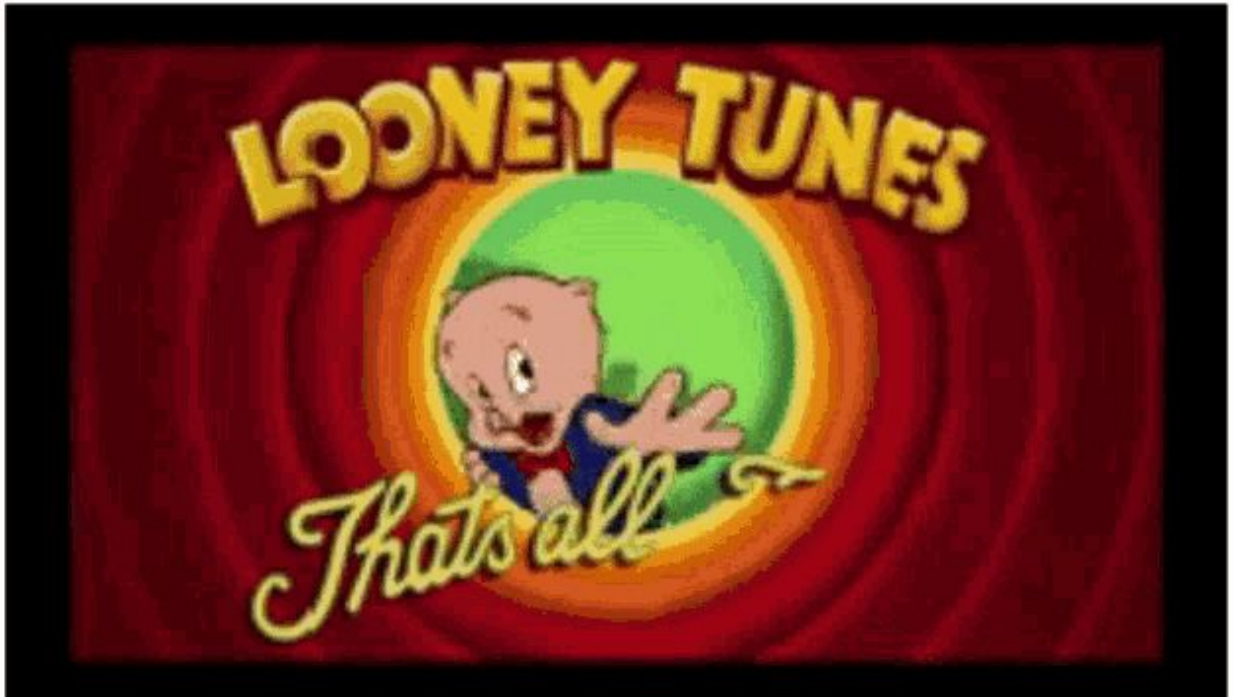
Global Exchange of Cloud Resources

3. Brokers

   Brokers mediate between consumers and providers by buying capacity from the provider and sub-leasing these to the consumers. They must select consumers whose apps will provide the most utility. Brokers may also communicate with resource providers and other brokers to acquire or trade resource shares. To make decisions, these brokers are equipped with a negotiating module informed by the present conditions of the resources and the current demand.

4. Service-level agreements(SLAs)

   The service level agreement (SLA) highlights the details of the service to be provided in terms of metrics that have been agreed upon by all parties, as well as penalties for meeting and failing to meet the expectations.



That's all folks for this chapter !!!!

References

Architecture - https://www.youtube.com/watch?v=54hM9LB72fQ