

Chapter 1

Introduction & Concepts

Cloud Computing

A Hands-On Approach

Arshdeep Bahga • Vijay Madisetti



Outline

- Cloud Computing definition
- Characteristics of cloud computing
- Cloud deployment models
- Cloud service models
- Cloud Services
- Cloud Applications

Definition of Cloud Computing

The U.S. National Institute of Standards and Technology (NIST) defines cloud computing as:

- Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Characteristics of Cloud Computing

- On-demand self service:
 - Cloud computing resources can be provisioned on-demand by the users, without requiring interactions with the cloud service provider. The process of provisioning resources is automated.
- Broad network access:
 - Cloud computing resources can be accessed over the network using standard access mechanisms that provide platform-independent access through the use of heterogeneous client platforms such as workstations, laptops, tablets and smartphones.

Characteristics of Cloud Computing

- Resource pooling:
 - The computing and storage resources provided by cloud service providers are pooled to serve multiple users using multi-tenancy. Multi-tenant aspects of the cloud allow multiple users to be served by the same physical hardware.
- Rapid elasticity:
 - Cloud computing resources can be provisioned rapidly and elastically. Cloud resources can be rapidly scaled up or down based on demand.

Characteristics of Cloud Computing

- Measured service:
 - Cloud computing resources are provided to users on a pay-per-use model. The usage of the cloud resources is measured and the user is charged based on some specific metric.
- Performance:
 - Cloud computing provides improved performance for applications since the resources available to the applications can be scaled up or down based on the dynamic application workloads.

Characteristics of Cloud Computing

- Reduced costs:
 - Cloud computing provides cost benefits for applications as only as much computing and storage resources as required can be provisioned dynamically, and upfront investment in purchase of computing assets to cover worst case requirements is avoid.
- Outsourced Management:
 - Cloud computing allows the users (individuals, large organizations, small and medium enterprises and governments) to outsource the IT infrastructure requirements to external cloud providers.

Characteristics of Cloud Computing

- Reliability:
 - Applications deployed in cloud computing environments generally have a higher reliability since the underlying IT infrastructure is professionally managed by the cloud service.
- Multi-tenancy:
 - The multi-tenanted approach of the cloud allows multiple users to make use of the same shared resources.
 - In virtual multi-tenancy, computing and storage resources are shared among multiple users.
 - In organic multi-tenancy every component in the system architecture is shared among multiple tenants

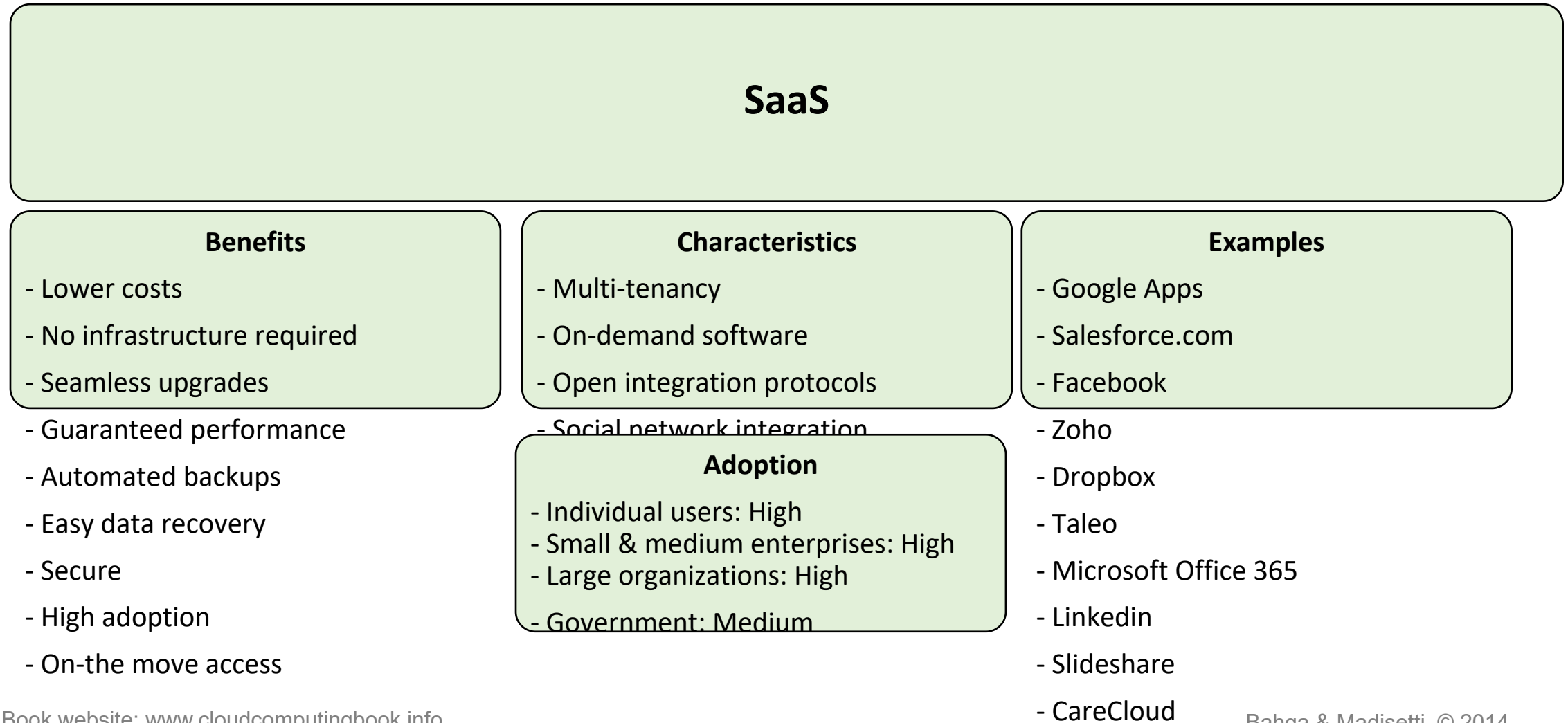
Cloud Service Models

- Software as a Service (SaaS)
 - Applications, management and user interfaces provided over a network
- Platform as a Service (PaaS)
 - Application development frameworks, operating systems and deployment frameworks
- Infrastructure as a Service (IaaS)
 - Virtual computing, storage and network resource that can be provisioned on demand

Software-as-a-Service (SaaS)

- Software/Interface
 - SaaS provides the users a complete software application or the user interface to the application itself.
- Outsourced Management
 - The cloud service provider manages the underlying cloud infrastructure including servers, network, operating systems, storage and application software, and the user is unaware of the underlying architecture of the cloud.
- Thin client interfaces
 - Applications are provided to the user through a thin client interface (e.g., a browser). SaaS applications are platform independent and can be accessed from various client devices such as workstations, laptop, tablets and smartphones, running different operating systems.
- Ubiquitous Access
 - Since the cloud service provider manages both the application and data, the users are able to access the applications from anywhere.

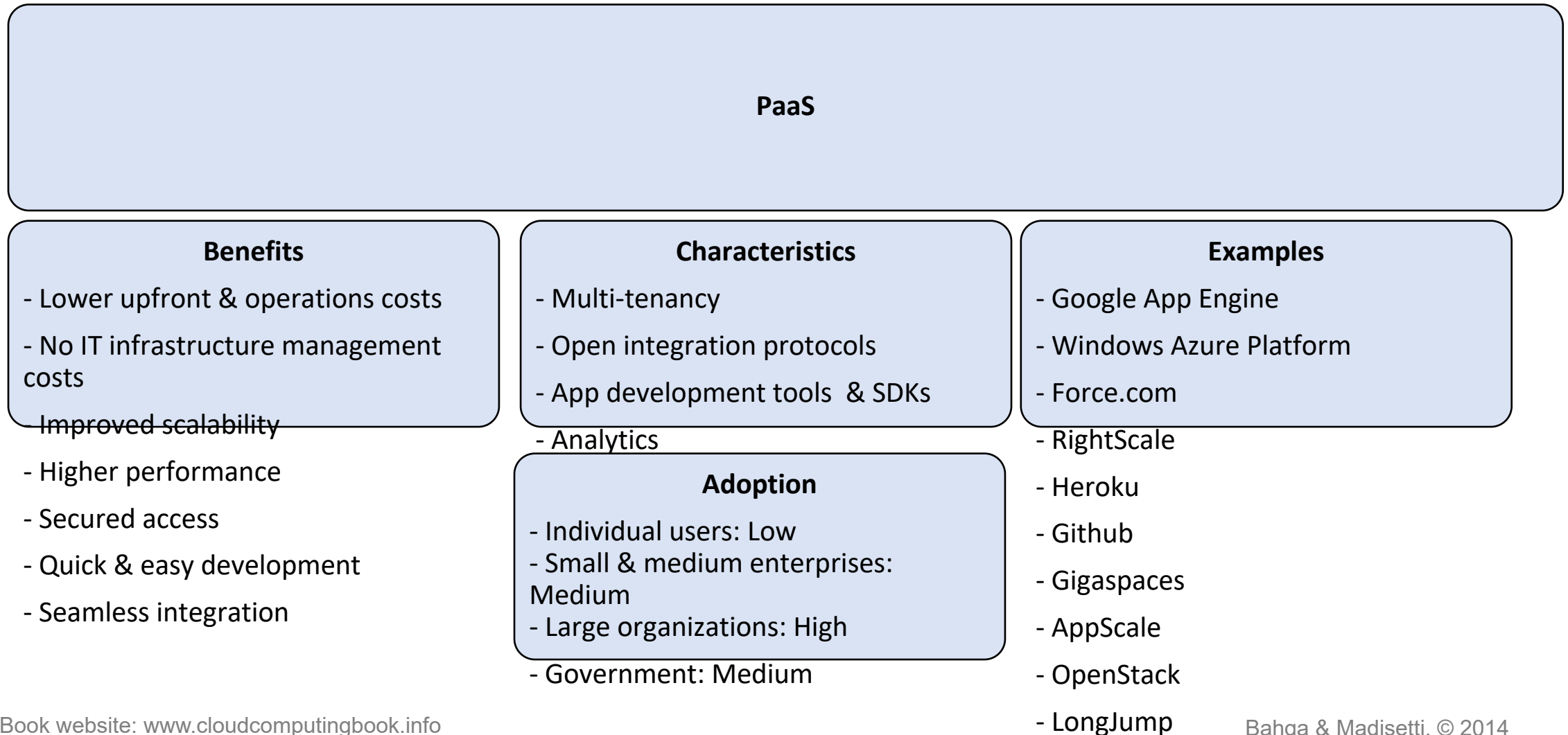
Software-as-a-Service (SaaS)



Platform-as-a-Service (PaaS)

- Development & Deployment:
 - PaaS provides the users the capability to develop and deploy application in the cloud using the development tools, application programming interfaces (APIs), software libraries and services provided by the cloud service provider.
- Provider Manages Infrastructure:
 - The cloud service provider manages the underlying cloud infrastructure including servers, network, operating systems and storage.
- User Manages Application:
 - The users, themselves, are responsible for developing, deploying, configuring and managing applications on the cloud infrastructure.

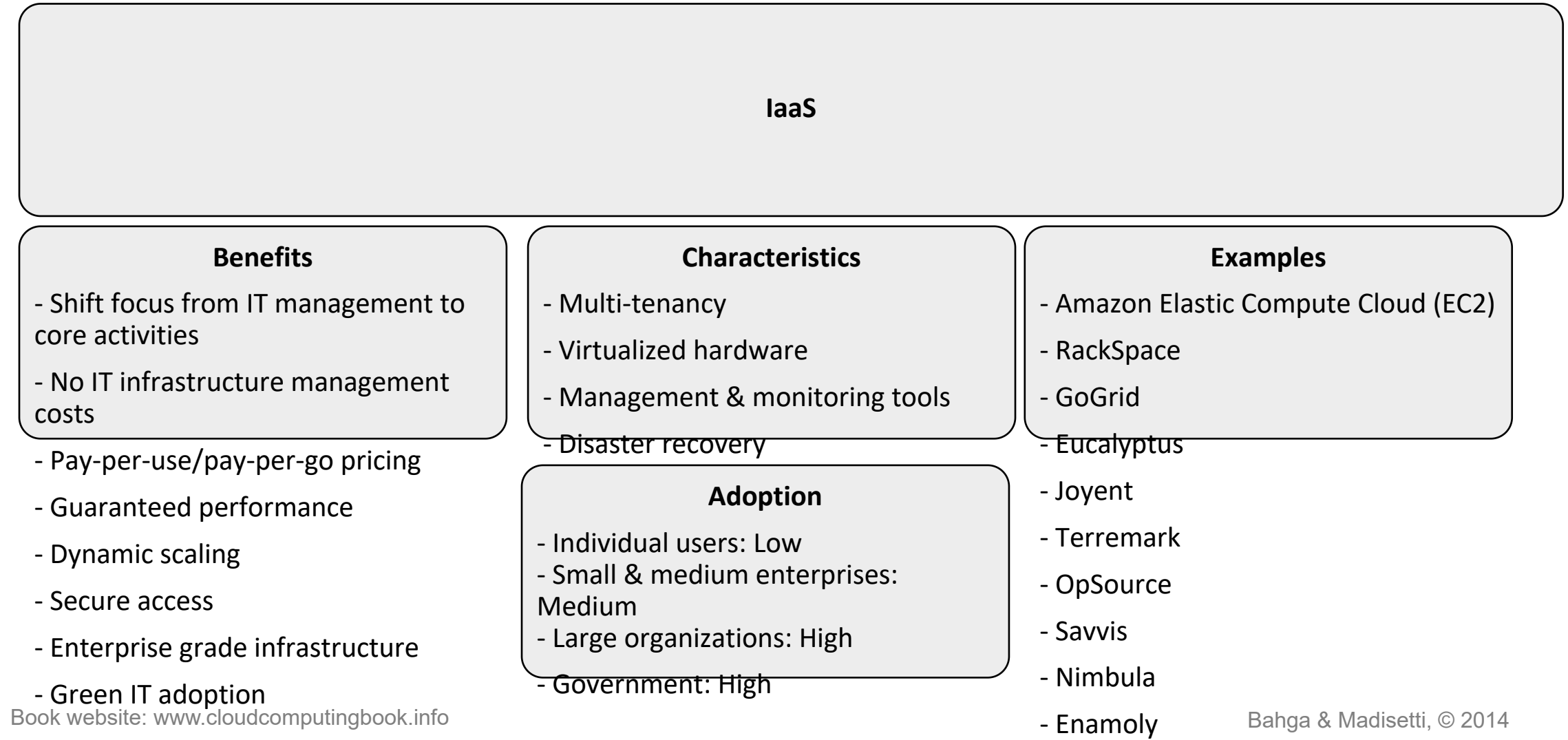
Platform-as-a-Service (PaaS)



Infrastructure-as-a-Service (IaaS)

- Resource Provisioning
 - Provides the users the capability to provision computing and storage resources.
- Virtual Machines
 - These resources are provided to the users as virtual machine instances and virtual storage. Users can start, stop, configure and manage the virtual machine instances and virtual storage.
- Provider Manages Infrastructure:
 - The cloud service provider manages the underlying infrastructure.
- Pay-per-use/Pay-as-you-go:
 - Virtual resources provisioned by the users are billed based on a pay-per-use/pay-as-you-go paradigm.

Infrastructure-as-a-Service (IaaS)



Cloud Deployment Models

- Public Cloud
 - Available for public use or a large industry group
- Private Cloud
 - Operated for exclusive use of a single organization
- Community Cloud
 - Available for shared use of several organizations supporting a specific community
- Hybrid Cloud
 - Combines multiple clouds (public and private) that remain unique but bound together to offer application and data portability

Cloud Service Examples

- IaaS:
 - Amazon EC2
 - Google Compute Engine
 - Windows Azure VMs
- PaaS:
 - Google App Engine
- SaaS:
 - Salesforce

Cloud Computing Applications

- Banking & Financial Apps
- E-Commerce Apps
- Social Networking
- Healthcare Systems
- Energy Systems
- Intelligent Transportation Systems
- E-Governance
- Education
- Mobile Communications

Further Reading

- Peter Mell, Timothy Grance, The NIST Definition of Cloud Computing, NIST Special Publication 800-145, Sep 2011.
- VMware, Understanding Full Virtualization, Paravirtualization, and Hardware Assist, 2007.
- A. Bahga, V. Madisetti, Analyzing Massive Machine Maintenance Data in a Computing Cloud, IEEE Transactions on Parallel & Distributed Systems, Vol. 23, Iss. 10, Oct 2012.
- A. Bahga, V. Madisetti, On a Cloud-Based Information Technology Framework for Data Driven Intelligent Transportation Systems, Journal of Transportation Technologies, Vol. 3, No. 2, April 2013.
- A. Bahga, V. Madisetti, A Cloud-Based Approach to Interoperable Electronic Health Records (EHRs), IEEE Journal of Biomedical and Health Informatics, Vol. 17, Iss. 5, Sep 2013.
- Network Functions Virtualization, <http://www.etsi.org/technologies-clusters/technologies/nfv>, Retrieved 2013.
- Amazon Elastic Compute Cloud, <http://aws.amazon.com/ec2>, 2012.
- Google Compute Engine, <https://developers.google.com/compute/>, Retrieved 2013.
- Windows Azure, <http://www.windowsazure.com/>, Retrieved 2013.
- Google App Engine, <http://appengine.google.com>, 2012.
- Salesforce, <http://salesforce.com>, 2012.

Chapter 2

Concepts & Technologies

Cloud Computing

A Hands-On Approach

Arshdeep Bahga • Vijay Madisetti

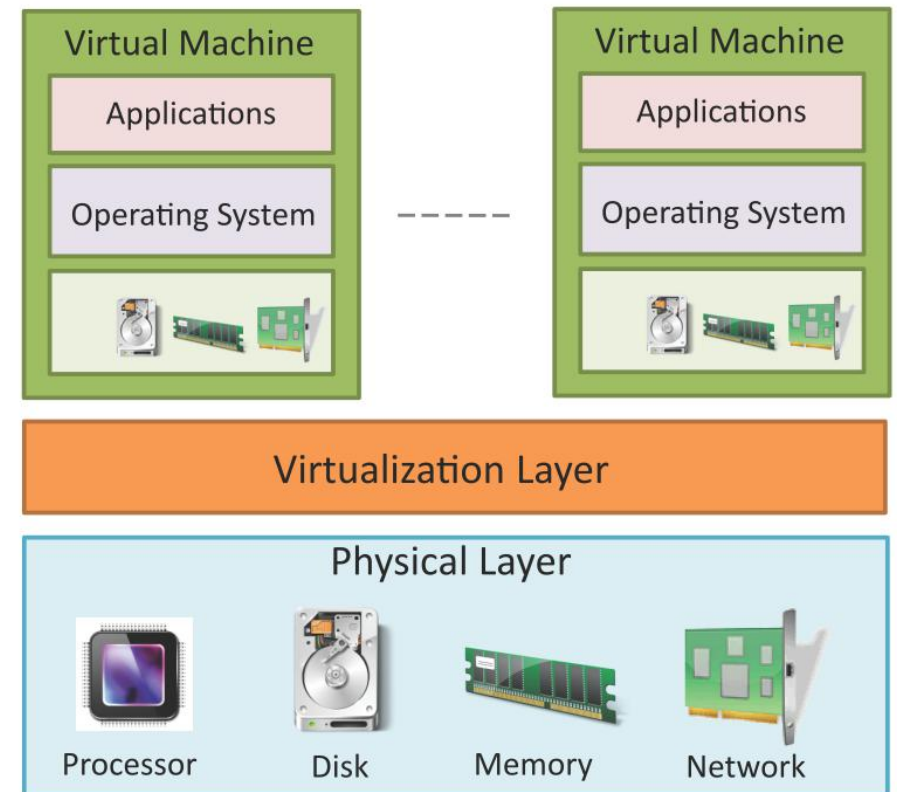


Outline

- Concepts and enabling technologies of cloud computing
 - Virtualization
 - Load balancing
 - Scalability & Elasticity
 - Deployment
 - Replication
 - Monitoring
 - MapReduce
 - Identity and Access Management
 - Service Level Agreements
 - Billing

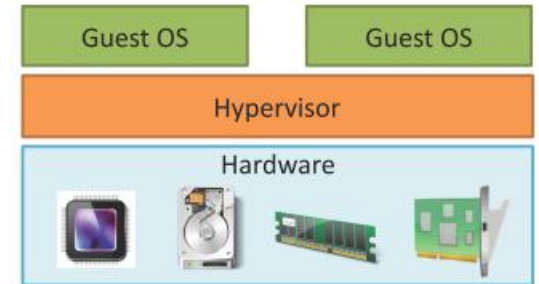
Virtualization

- Virtualization refers to the partitioning the resources of a physical system (such as computing, storage, network and memory) into multiple virtual resources.
- Key enabling technology of cloud computing that allow pooling of resources.
- In cloud computing, resources are pooled to serve multiple users using multi-tenancy.

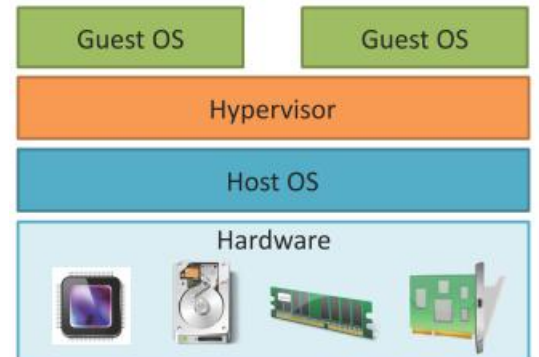


Hypervisor

- The virtualization layer consists of a hypervisor or a virtual machine monitor (VMM).
- Hypervisor presents a virtual operating platform to a guest operating system (OS).
- Type-1 Hypervisor
 - Type-1 or the native hypervisors run directly on the host hardware and control the hardware and monitor the guest operating systems.
- Type-2 Hypervisor
 - Type 2 hypervisors or hosted hypervisors run on top of a conventional (main/host) operating system and monitor the guest operating systems.



Type-1 Hypervisor



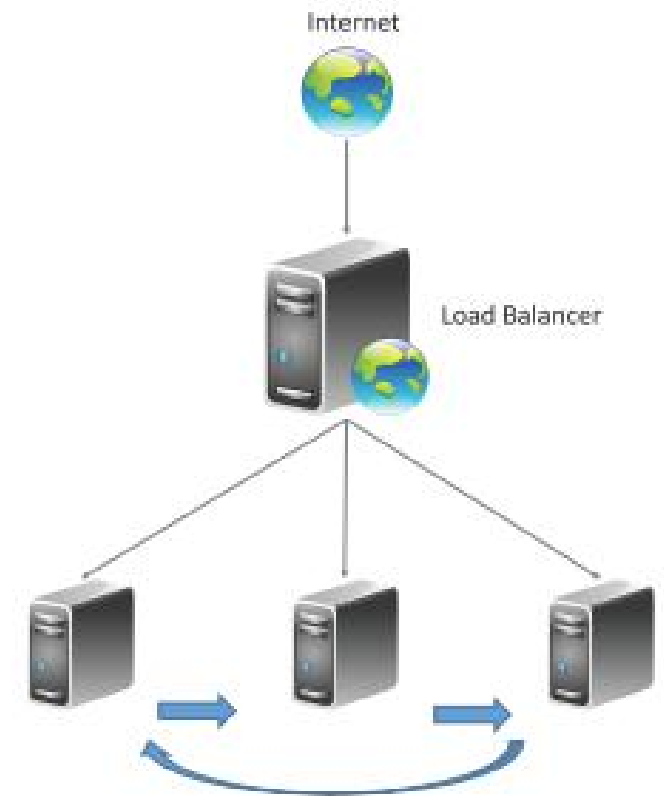
Type-2 Hypervisor

Types of Virtualization

- Full Virtualization
 - In full virtualization, the virtualization layer completely decouples the guest OS from the underlying hardware. The guest OS requires no modification and is not aware that it is being virtualized. Full virtualization is enabled by direct execution of user requests and binary translation of OS requests.
- Para-Virtualization
 - In para-virtualization, the guest OS is modified to enable communication with the hypervisor to improve performance and efficiency. The guest OS kernel is modified to replace non-virtualizable instructions with hyper-calls that communicate directly with the virtualization layer hypervisor.
- Hardware Virtualization
 - Hardware assisted virtualization is enabled by hardware features such as Intel's Virtualization Technology (VT-x) and AMD's AMD-V. In hardware assisted virtualization, privileged and sensitive calls are set to automatically trap to the hypervisor. Thus, there is no need for either binary translation or para-virtualization.

Load Balancing

- Cloud computing resources can be scaled up on demand to meet the performance requirements of applications.
- Load balancing distributes workloads across multiple servers to meet the application workloads.
- The goals of load balancing techniques include:
 - Achieve maximum utilization of resources
 - Minimizing the response times
 - Maximizing throughput



Load Balancing Algorithms

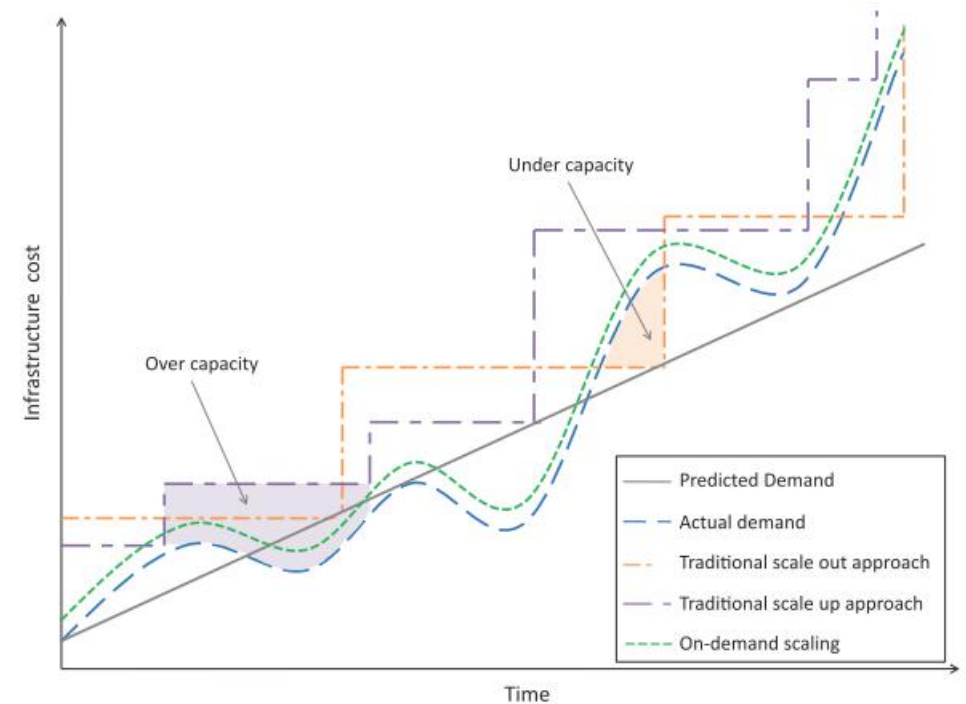
- Round Robin load balancing
- Weighted Round Robin load balancing
- Low Latency load balancing
- Least Connections load balancing
- Priority load balancing
- Overflow load balancing

Load Balancing - Persistence Approaches

- Since load balancing can route successive requests from a user session to different servers, maintaining the state or the information of the session is important.
- Persistence Approaches
 - Sticky sessions
 - Session Database
 - Browser cookies
 - URL re-writing

Scalability & Elasticity

- Multi-tier applications such as e-Commerce, social networking, business-to-business, etc. can experience rapid changes in their traffic.
- Capacity planning involves determining the right sizing of each tier of the deployment of an application in terms of the number of resources and the capacity of each resource.
- Capacity planning may be for computing, storage, memory or network resources.



Scaling Approaches

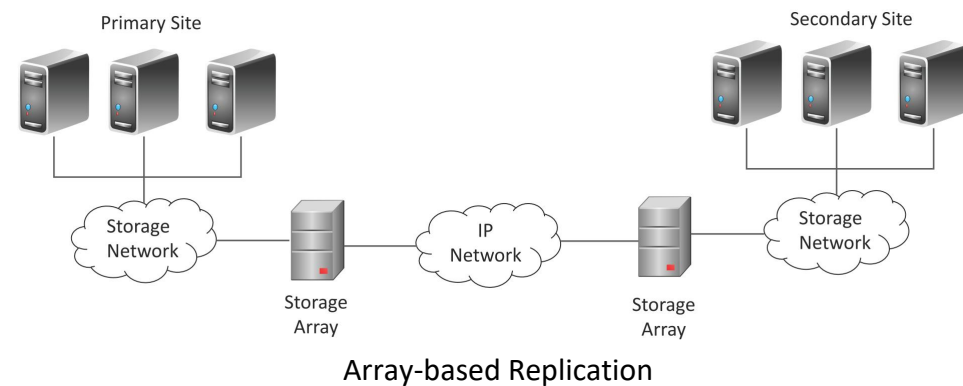
- Vertical Scaling/Scaling up:
 - Involves upgrading the hardware resources (adding additional computing, memory, storage or network resources).
- Horizontal Scaling/Scaling out
 - Involves addition of more resources of the same type.

Deployment

- Cloud application deployment design is an iterative process that involves:
 - Deployment Design
 - The variables in this step include the number of servers in each tier, computing, memory and storage capacities of servers, server interconnection, load balancing and replication strategies.
 - Performance Evaluation
 - To verify whether the application meets the performance requirements with the deployment.
 - Involves monitoring the workload on the application and measuring various workload parameters such as response time and throughput.
 - Utilization of servers (CPU, memory, disk, I/O, etc.) in each tier is also monitored.
 - Deployment Refinement
 - Various alternatives can exist in this step such as vertical scaling (or scaling up), horizontal scaling (or scaling out), alternative server interconnections, alternative load balancing and replication strategies, for instance.

Replication

- Replication is used to create and maintain multiple copies of the data in the cloud.
- Cloud enables rapid implementation of replication solutions for disaster recovery for organizations.
- With cloud-based data replication organizations can plan for disaster recovery without making any capital expenditures on purchasing, configuring and managing secondary site locations.
- Types:
 - Array-based Replication
 - Network-based Replication
 - Host-based Replication



Monitoring

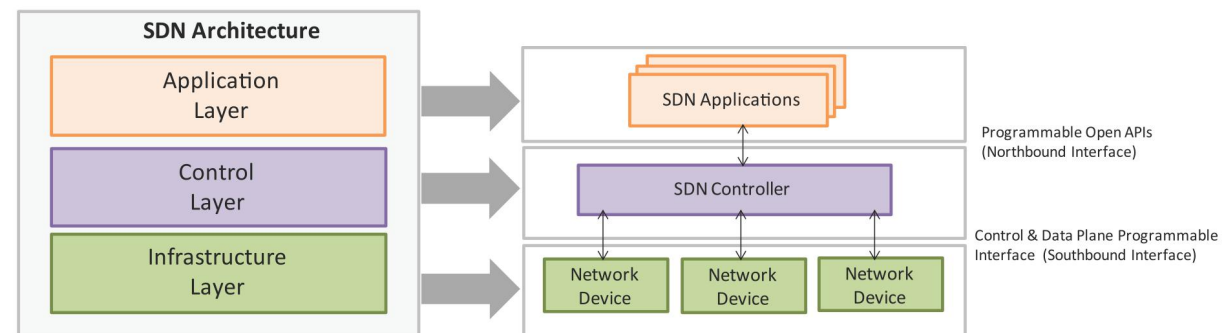
- Monitoring services allow cloud users to collect and analyze the data on various monitoring metrics.
- A monitoring service collects data on various system and application metrics from the cloud computing instances.
- Monitoring of cloud resources is important because it allows the users to keep track of the health of applications and services deployed in the cloud.

Examples of Monitoring Metrics

Type	Metrics
CPU	CPU-Usage, CPU-Idle
Disk	Disk-Usage, Bytes/sec (read/write), Operations/sec
Memory	Memory-Used, Memory-Free, Page-Cache
Interface	Packets/sec (incoming/outgoing), Octets/sec(incoming/outgoing)

Software Defined Networking

- Software-Defined Networking (SDN) is a networking architecture that separates the control plane from the data plane and centralizes the network controller.
- Conventional network architecture
 - The control plane and data plane are coupled. Control plane is the part of the network that carries the signaling and routing message traffic while the data plane is the part of the network that carries the payload data traffic.
- SDN Architecture
 - The control and data planes are decoupled and the network controller is centralized.

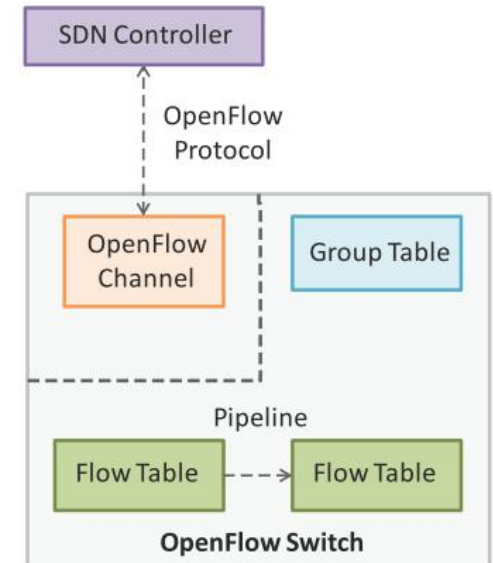


SDN - Key Elements

- Centralized Network Controller
 - With decoupled the control and data planes and centralized network controller, the network administrators can rapidly configure the network.
- Programmable Open APIs
 - SDN architecture supports programmable open APIs for interface between the SDN application and control layers (Northbound interface). These open APIs that allow implementing various network services such as routing, quality of service (QoS), access control, etc.
- Standard Communication Interface (OpenFlow)
 - SDN architecture uses a standard communication interface between the control and infrastructure layers (Southbound interface). OpenFlow, which is defined by the Open Networking Foundation (ONF) is the broadly accepted SDN protocol for the Southbound interface.

OpenFlow

- OpenFlow is the broadly accepted SDN protocol for the Southbound interface.
- With OpenFlow, the forwarding plane of the network devices can be directly accessed and manipulated.
- OpenFlow uses the concept of flows to identify network traffic based on pre-defined match rules.
- Flows can be programmed statically or dynamically by the SDN control software.
- OpenFlow protocol is implemented on both sides of the interface between the controller and the network devices.



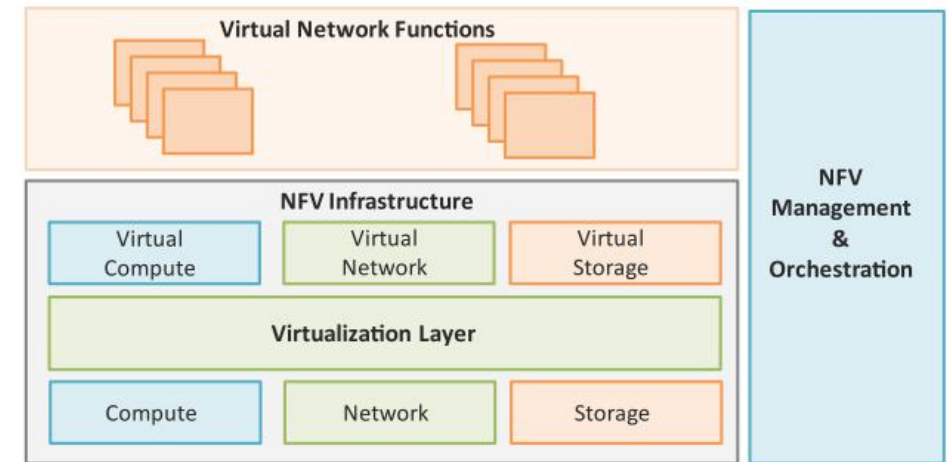
OpenFlow switch comprising of one or more flow tables and a group table, which perform packet lookups and forwarding, and OpenFlow channel to an external controller.

Network Function Virtualization

- Network Function Virtualization (NFV) is a technology that leverages virtualization to consolidate the heterogeneous network devices onto industry standard high volume servers, switches and storage.
- Relationship to SDN
 - NFV is complementary to SDN as NFV can provide the infrastructure on which SDN can run.
 - NFV and SDN are mutually beneficial to each other but not dependent.
 - Network functions can be virtualized without SDN, similarly, SDN can run without NFV.
- NFV comprises of network functions implemented in software that run on virtualized resources in the cloud.
- NFV enables a separation the network functions which are implemented in software from the underlying hardware.

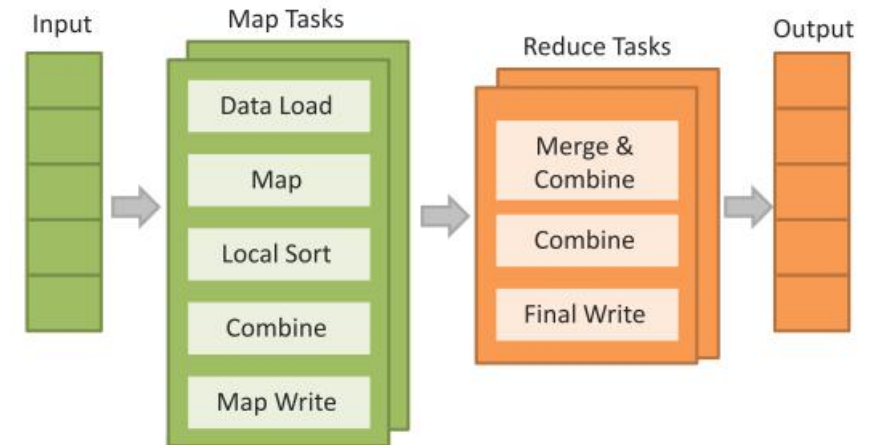
NFV Architecture

- Key elements of the NFV architecture are
 - Virtualized Network Function (VNF): VNF is a software implementation of a network function which is capable of running over the NFV Infrastructure (NFVI).
 - NFV Infrastructure (NFVI): NFVI includes compute, network and storage resources that are virtualized.
 - NFV Management and Orchestration: NFV Management and Orchestration focuses on all virtualization-specific management tasks and covers the orchestration and lifecycle management of physical and/or software resources that support the infrastructure virtualization, and the lifecycle management of VNFs.



MapReduce

- MapReduce is a parallel data processing model for processing and analysis of massive scale data.
- MapReduce phases:
 - **Map Phase:** In the Map phase, data is read from a distributed file system, partitioned among a set of computing nodes in the cluster, and sent to the nodes as a set of key-value pairs.
 - The Map tasks process the input records independently of each other and produce intermediate results as key-value pairs.
 - The intermediate results are stored on the local disk of the node running the Map task.
 - **Reduce Phase:** When all the Map tasks are completed, the Reduce phase begins in which the intermediate data with the same key is aggregated.



Identity and Access Management

- Identity and Access Management (IDAM) for cloud describes the authentication and authorization of users to provide secure access to cloud resources.
- Organizations with multiple users can use IDAM services provided by the cloud service provider for management of user identifiers and user permissions.
- IDAM services allow organizations to centrally manage users, access permissions, security credentials and access keys.
- Organizations can enable role-based access control to cloud resources and applications using the IDAM services.
- IDAM services allow creation of user groups where all the users in a group have the same access permissions.
- Identity and Access Management is enabled by a number of technologies such as OpenAuth, Role-based Access Control (RBAC), Digital Identities, Security Tokens, Identity Providers, etc.

Billing

Cloud service providers offer a number of billing models described as follows:

- Elastic Pricing
 - In elastic pricing or pay-as-you-use pricing model, the customers are charged based on the usage of cloud resources.
- Fixed Pricing
 - In fixed pricing models, customers are charged a fixed amount per month for the cloud resources.
- Spot Pricing
 - Spot pricing models offer variable pricing for cloud resources which is driven by market demand.

Further Reading

- Network Functions Virtualization, <http://www.etsi.org/technologies-clusters/technologies/nfv>, Retrieved 2013.
- OpenFlow Switch Specification, <https://www.opennetworking.org>, Retrieved 2013.
- J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, OSDI 2004.
- VMware, Understanding Full Virtualization, Paravirtualization, and Hardware Assist, 2007.

Chapter 3

Cloud Services & Platforms

Cloud Computing

A Hands-On Approach

Arshdeep Bahga • Vijay Madisetti

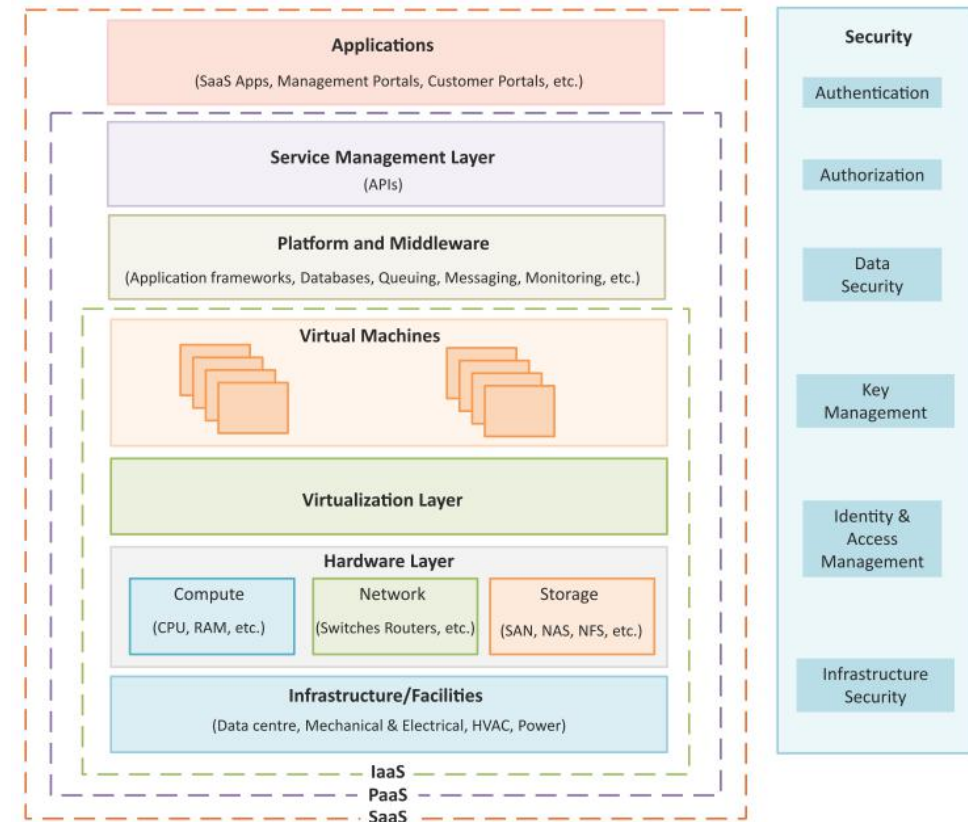


Outline

- Compute Services
- Storage Services
- Database Services
- Application Services
- Content Delivery Services
- Analytics Services
- Deployment & Management Services
- Identity & Access Management Services

Cloud Reference Model

- **Infrastructure & Facilities Layer**
 - Includes the physical infrastructure such as datacenter facilities, electrical and mechanical equipment, etc.
- **Hardware Layer**
 - Includes physical compute, network and storage hardware.
- **Virtualization Layer**
 - Partitions the physical hardware resources into multiple virtual resources that enabling pooling of resources.
- **Platform & Middleware Layer**
 - Builds upon the IaaS layers below and provides standardized stacks of services such as database service, queuing service, application frameworks and run-time environments, messaging services, monitoring services, analytics services, etc.
- **Service Management Layer**
 - Provides APIs for requesting, managing and monitoring cloud resources.
- **Applications Layer**
 - Includes SaaS applications such as Email, cloud storage application, productivity applications, management portals, customer self-service portals, etc.

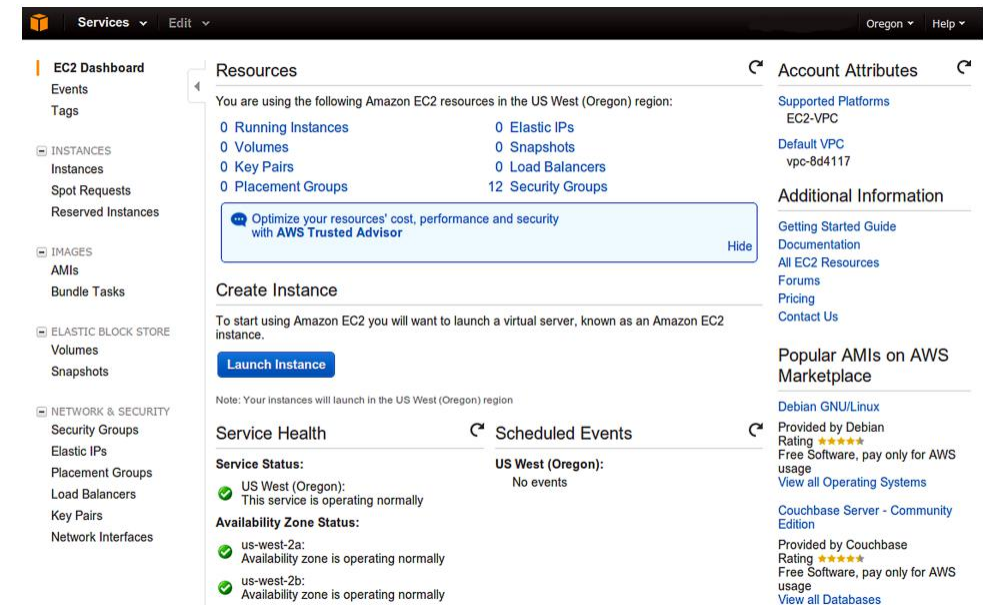


Compute Services

- Compute services provide dynamically scalable compute capacity in the cloud.
- Compute resources can be provisioned on-demand in the form of virtual machines. Virtual machines can be created from standard images provided by the cloud service provider or custom images created by the users.
- Compute services can be accessed from the web consoles of these services that provide graphical user interfaces for provisioning, managing and monitoring these services.
- Cloud service providers also provide APIs for various programming languages that allow developers to access and manage these services programmatically.

Compute Services – Amazon EC2

- Amazon Elastic Compute Cloud (EC2) is a compute service provided by Amazon.
- Launching EC2 Instances
 - To launch a new instance click on the launch instance button. This will open a wizard where you can select the Amazon machine image (AMI) with which you want to launch the instance. You can also create their own AMIs with custom applications, libraries and data. Instances can be launched with a variety of operating systems.
- Instance Sizes
 - When you launch an instance you specify the instance type (micro, small, medium, large, extra-large, etc.), the number of instances to launch based on the selected AMI and availability zones for the instances.
- Key-pairs
 - When launching a new instance, the user selects a key-pair from existing keypairs or creates a new keypair for the instance. Keypairs are used to securely connect to an instance after it launches.
- Security Groups
 - The security groups to be associated with the instance can be selected from the instance launch wizard. Security groups are used to open or block a specific network port for the launched instances.



Compute Services – Google Compute Engine

- Google Compute Engine is a compute service provided by Google.
- Launching Instances
 - To create a new instance, the user selects an instance machine type, a zone in which the instance will be launched, a machine image for the instance and provides an instance name, instance tags and meta-data.
- Disk Resources
 - Every instance is launched with a disk resource. Depending on the instance type, the disk resource can be a scratch disk space or persistent disk space. The scratch disk space is deleted when the instance terminates. Whereas, persistent disks live beyond the life of an instance.
- Network Options
 - Network option allows you to control the traffic to and from the instances. By default, traffic between instances in the same network, over any port and any protocol and incoming SSH connections from anywhere are enabled.

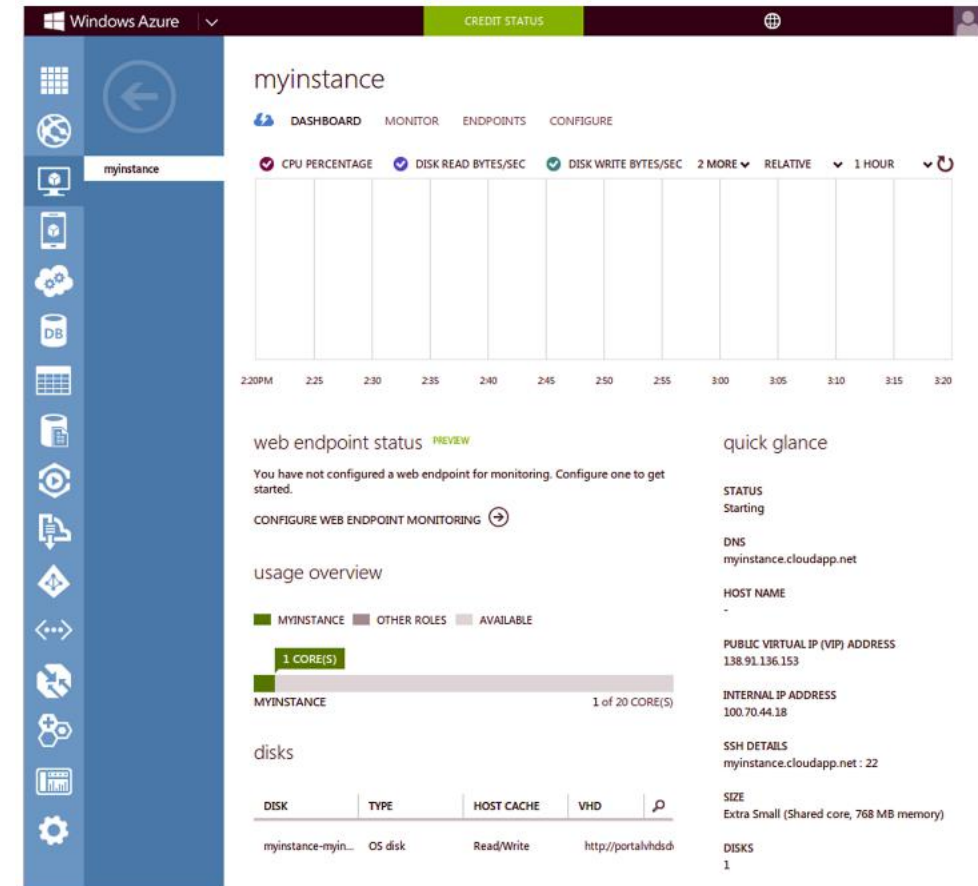
The screenshot shows the Google Cloud Console interface for creating a new Compute Engine instance. The page is titled 'Create a new Instance' and includes a sidebar with navigation links: Instances, Disks, Snapshots, Images, Networks, Metadata, Zones, Operations, and Quotas. The main form contains the following sections:

- Instance Details:** Name (myinstance), Description (My instance), Tags (comma separated), and Metadata (key-value pairs).
- Summary:** myinstance, My instance, debian-7-wheezy-v20130723, Debian GNU/Linux 7.1 (wheezy) b..., us-central1-b, 1 vCPU, 3.75 GB RAM, default. A note states: 'Note: per-minute charges will begin now'.
- Location and Resources:** Zone (us-central1-b), Machine Type (n1-standard-1), Boot Source (New persistent disk from image), Image (debian-7-wheezy-v20130723), and Additional Disks (No disks in zone us-central1-b).
- Networking:** Network (default) and External IP (Ephemeral).

Buttons for 'Create' and 'Discard' are located at the bottom right of the form.

Compute Services – Windows Azure VMs

- Windows Azure Virtual Machines is the compute service from Microsoft.
- Launching Instances:
 - To create a new instance, you select the instance type and the machine image.
 - You can either provide a user name and password or upload a certificate file for securely connecting to the instance.
 - Any changes made to the VM are persistently stored and new VMs can be created from the previously stored machine images.



Storage Services

- Cloud storage services allow storage and retrieval of any amount of data, at any time from anywhere on the web.
- Most cloud storage services organize data into buckets or containers.
- Scalability
 - Cloud storage services provide high capacity and scalability. Objects upto several tera-bytes in size can be uploaded and multiple buckets/containers can be created on cloud storages.
- Replication
 - When an object is uploaded it is replicated at multiple facilities and/or on multiple devices within each facility.
- Access Policies
 - Cloud storage services provide several security features such as Access Control Lists (ACLs), bucket/container level policies, etc. ACLs can be used to selectively grant access permissions on individual objects. Bucket/container level policies can also be defined to allow or deny permissions across some or all of the objects within a single bucket/container.
- Encryption
 - Cloud storage services provide Server Side Encryption (SSE) options to encrypt all data stored in the cloud storage.
- Consistency
 - Strong data consistency is provided for all upload and delete operations. Therefore, any object that is uploaded can be immediately downloaded after the upload is complete.

Storage Services – Amazon S3

- Amazon Simple Storage Service(S3) is an online cloud-based data storage infrastructure for storing and retrieving any amount of data.
- S3 provides highly reliable, scalable, fast, fully redundant and affordable storage infrastructure.
- Buckets
 - Data stored on S3 is organized in the form of buckets. You must create a bucket before you can store data on S3.
- Uploading Files to Buckets
 - S3 console provides simple wizards for creating a new bucket and uploading files.
 - You can upload any kind of file to S3.
 - While uploading a file, you can specify the redundancy and encryption options and access permissions.

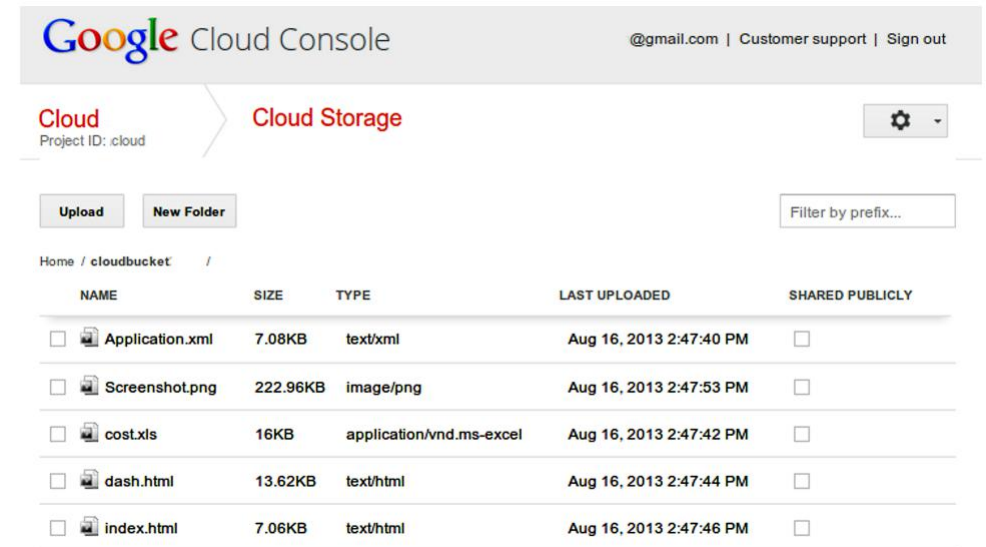


The screenshot shows the Amazon S3 console interface. At the top, there are buttons for 'Upload', 'Create Folder', and 'Actions'. Below these, the breadcrumb 'Buckets / myBucket2013' is visible. A table lists the contents of the bucket:

Name	Storage Class	Size	Last Modified
pg46.txt	Standard	177.7 KB	Thu Dec 27 16:06:05 GMT+530 2012

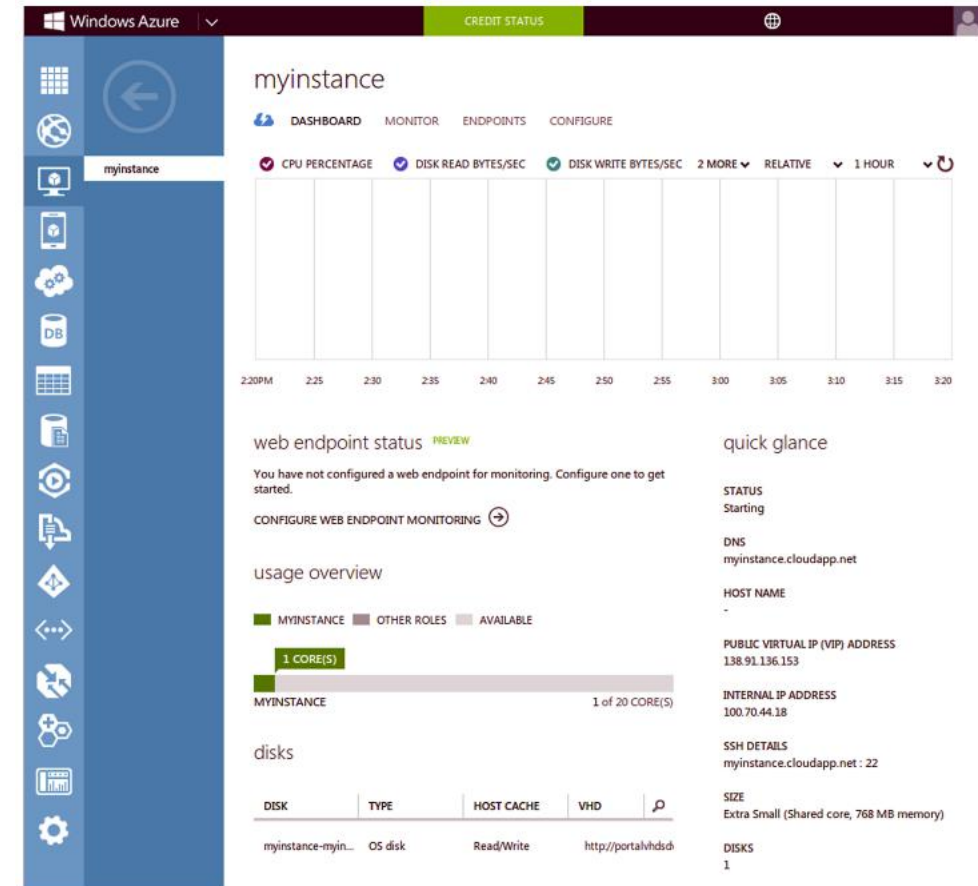
Storage Services – Google Cloud Storage

- GCS is the Cloud storage service from Google
- Buckets
 - Objects in GCS are organized into buckets.
- Access Control Lists
 - ACLs are used to control access to objects and buckets. ACLs can be configured to share objects and buckets with the entire world, a Google group, a Google-hosted domain, or specific Google account holders.



Storage Services – Windows Azure Storage

- Windows Azure Storage is the cloud storage service from Microsoft.
- Windows Azure Storage provides various storage services such as blob storage service, table service and queue service.
- Blob storage service
 - The blob storage service allows storing unstructured binary data or binary large objects (blobs).
 - Blobs are organized into containers.
 - Block blobs - can be subdivided into some number of blocks. If a failure occurs while transferring a block blob, retransmission can resume with the most recent block rather than sending the entire blob again.
 - Page blobs - are divided into number of pages and are designed for random access. Applications can read and write individual pages at random in a page blob.

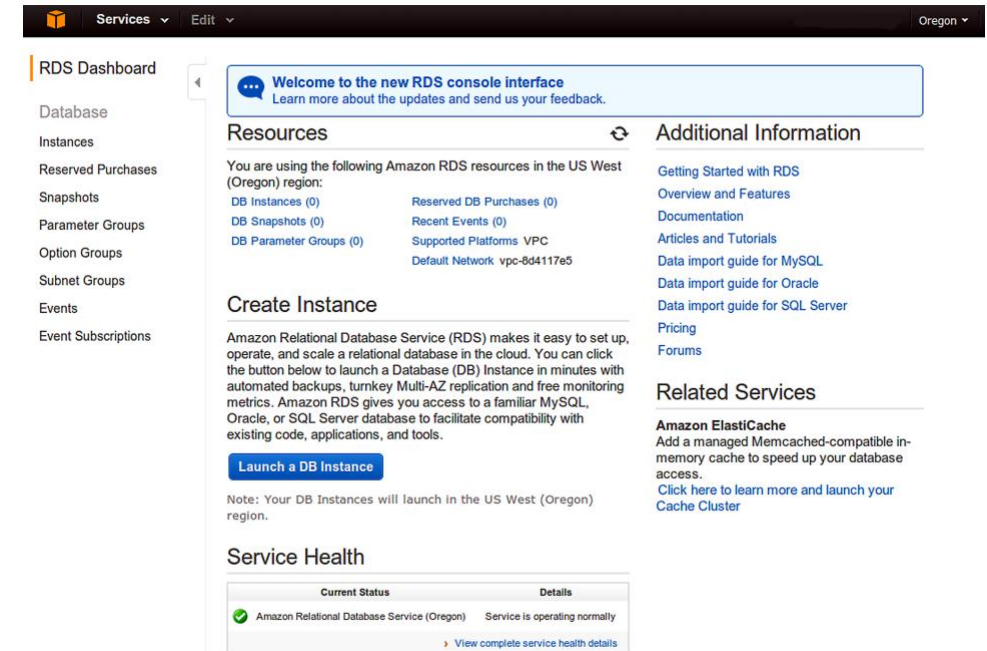


Database Services

- Cloud database services allow you to set-up and operate relational or non-relational databases in the cloud.
- Relational Databases
 - Popular relational databases provided by various cloud service providers include MySQL, Oracle, SQL Server, etc.
- Non-relational Databases
 - The non-relational (No-SQL) databases provided by cloud service providers are mostly proprietary solutions.
- Scalability
 - Cloud database services allow provisioning as much compute and storage resources as required to meet the application workload levels. Provisioned capacity can be scaled-up or down. For read-heavy workloads, read-replicas can be created.
- Reliability
 - Cloud database services are reliable and provide automated backup and snapshot options.
- Performance
 - Cloud database services provide guaranteed performance with options such as guaranteed input/output operations per second (IOPS) which can be provisioned upfront.
- Security
 - Cloud database services provide several security features to restrict the access to the database instances and stored data, such as network firewalls and authentication mechanisms.

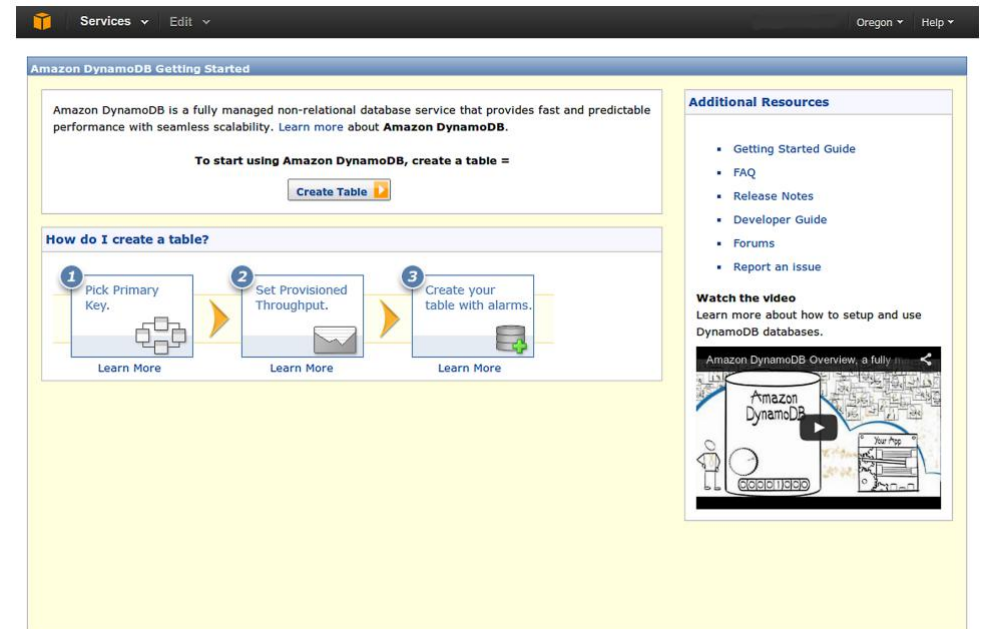
Database Services – Amazon RDS

- Amazon Relational Database Service (RDS) is a web service that makes it easy to setup, operate and scale a relational database in the cloud.
- Launching DB Instances
 - The console provides an instance launch wizard that allows you to select the type of database to create (MySQL, Oracle or SQL Server) database instance size, allocated storage, DB instance identifier, DB username and password. The status of the launched DB instances can be viewed from the console.
- Connecting to a DB Instance
 - Once the instance is available, you can note the instance end point from the instance properties tab. This end point can then be used for securely connecting to the instance.



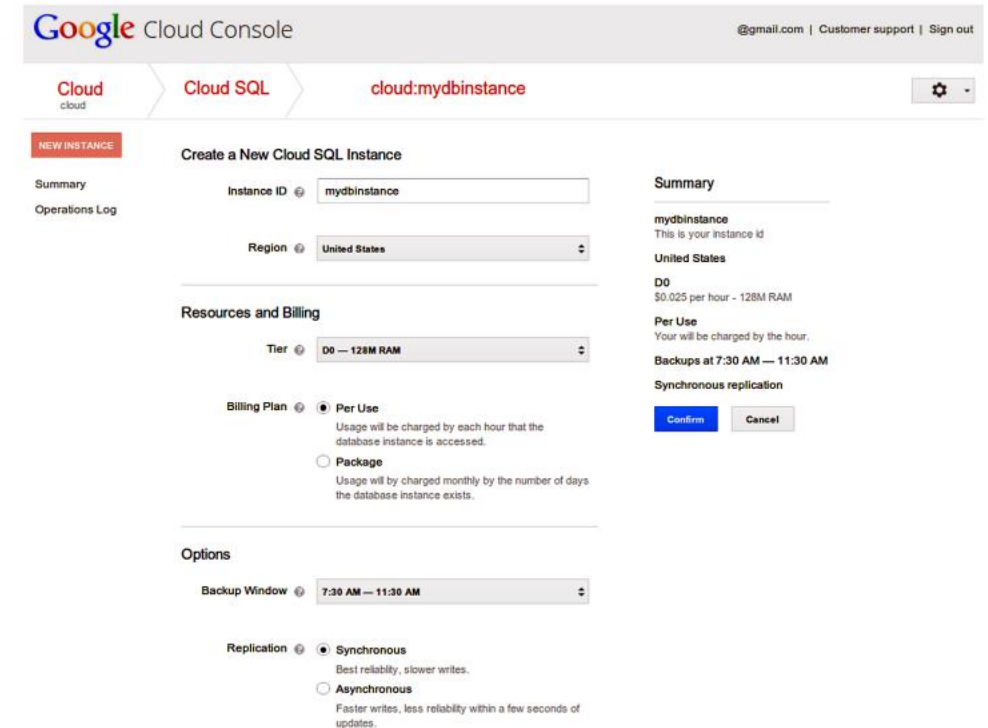
Database Services – Amazon DynamoDB

- Amazon DynamoDB is the non-relational (No-SQL) database service from Amazon.
- Data Model
 - The DynamoDB data model includes include tables, items and attributes.
 - A table is a collection of items and each item is a collection of attributes.
 - To store data in DynamoDB you have to create a one or more tables and specify how much throughput capacity you want to provision and reserve for reads and writes.
- Fully Managed Service
 - DynamoDB is a fully managed service that automatically spreads the data and traffic for the stored tables over a number of servers to meet the throughput requirements specified by the users.
- Replication
 - All stored data is automatically replicated across multiple availability zones to provide data durability.



Storage Services – Google Cloud SQL

- Google SQL is the relational database service from Google.
- Google Cloud SQL service allows you to host MySQL databases in the Google's cloud.
- Launching DB Instances
 - You can create new database instances from the console and manage existing instances. To create a new instance you select a region, database tier, billing plan and replication mode.
- Backups
 - You can schedule daily backups for your Google Cloud SQL instances, and also restore backed-up databases.
- Replication
 - Cloud SQL provides both synchronous or asynchronous geographic replication and the ability to import/ export databases.



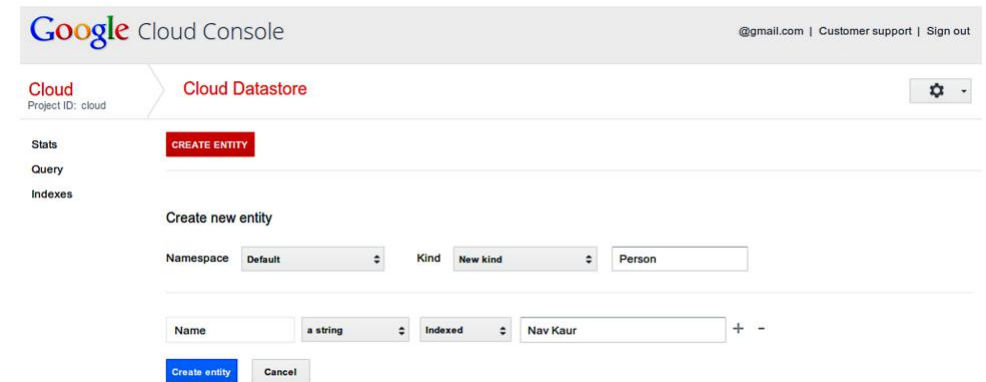
The screenshot shows the Google Cloud Console interface for creating a new Cloud SQL instance. The breadcrumb navigation at the top indicates the path: Cloud > Cloud SQL > cloud:mydbinstance. On the left, there are links for 'NEW INSTANCE', 'Summary', and 'Operations Log'. The main form is titled 'Create a New Cloud SQL Instance' and includes the following sections:

- Instance ID:** A text field containing 'mydbinstance'.
- Region:** A dropdown menu set to 'United States'.
- Resources and Billing:**
 - Tier:** A dropdown menu set to 'D0 — 128M RAM'.
 - Billing Plan:** Two radio button options: 'Per Use' (selected) and 'Package'. The 'Per Use' description states: 'Usage will be charged by each hour that the database instance is accessed.' The 'Package' description states: 'Usage will be charged monthly by the number of days the database instance exists.'
- Options:**
 - Backup Window:** A dropdown menu set to '7:30 AM — 11:30 AM'.
 - Replication:** Two radio button options: 'Synchronous' (selected) and 'Asynchronous'. The 'Synchronous' description states: 'Best reliability, slower writes.' The 'Asynchronous' description states: 'Faster writes, less reliability within a few seconds of updates.'

On the right side, there is a 'Summary' section that displays the instance details: 'mydbinstance', 'This is your instance id', 'United States', 'D0', '\$0.025 per hour - 128M RAM', 'Per Use', 'Your will be charged by the hour.', 'Backups at 7:30 AM — 11:30 AM', and 'Synchronous replication'. At the bottom of the summary are 'Confirm' and 'Cancel' buttons.

Storage Services – Google Cloud Datastore

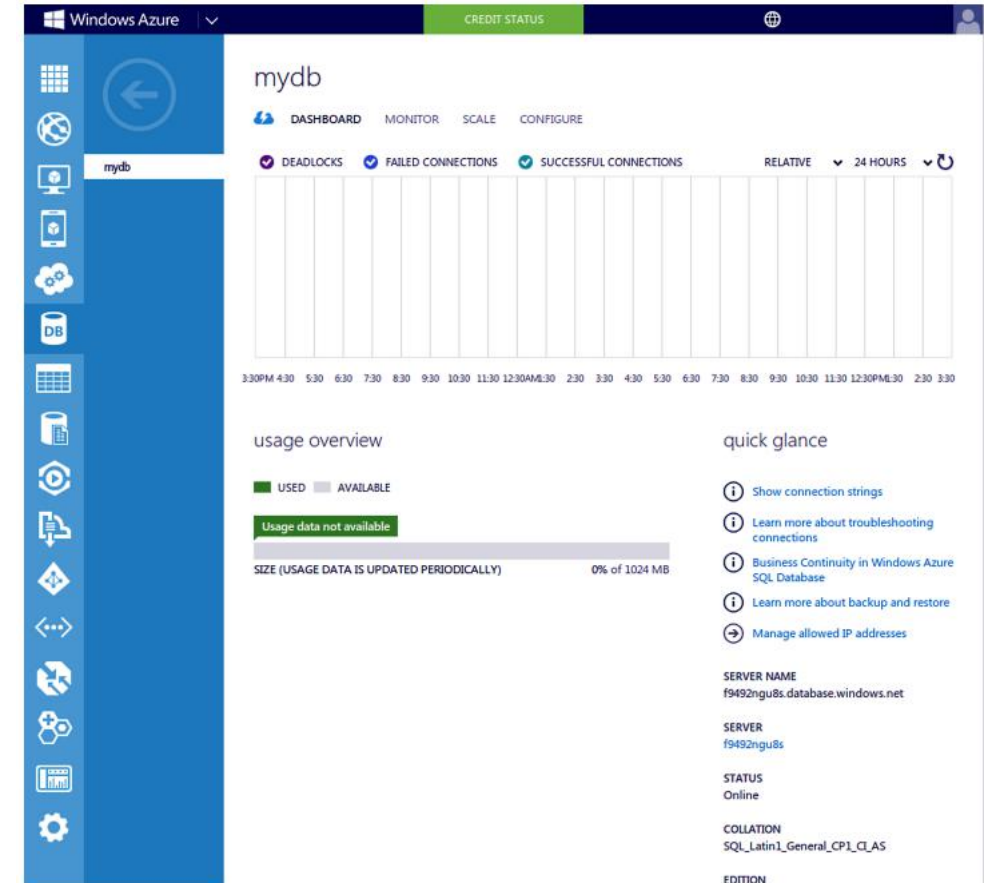
- Google Cloud Datastore is a fully managed non-relational database from Google.
- Cloud Datastore offers ACID transactions and high availability of reads and writes.
- Data Model
 - The Cloud Datastore data model consists of entities. Each entity has one or more properties (key-value pairs) which can be of one of several supported data types, such as strings and integers. Each entity has a kind and a key. The entity kind is used for categorizing the entity for the purpose of queries and the entity key uniquely identifies the entity.



The screenshot shows the Google Cloud Console interface for the Cloud Datastore service. The top navigation bar includes the Google Cloud Console logo, a user profile (@gmail.com), and links for Customer support and Sign out. The main content area is titled 'Cloud Datastore' and includes a 'CREATE ENTITY' button. Below this, the 'Create new entity' form is displayed. The form has a 'Namespace' dropdown set to 'Default' and a 'Kind' dropdown set to 'New kind'. A text input field for the entity name is set to 'Person'. Below the form, there is a 'Name' field with a dropdown set to 'a string', an 'Indexed' checkbox, and a text input field containing 'Nav Kaur'. At the bottom of the form are 'Create entity' and 'Cancel' buttons.

Storage Services – Windows Azure SQL DB

- Windows Azure SQL Database is the relational database service from Microsoft.
- Azure SQL Database is based on the SQL server, but it does not give each customer a separate instance of SQL server.
- Multi-tenant Service
 - SQL Database is a multi-tenant service, with a logical SQL Database server for each customer.



Storage Services – Windows Azure Table Service

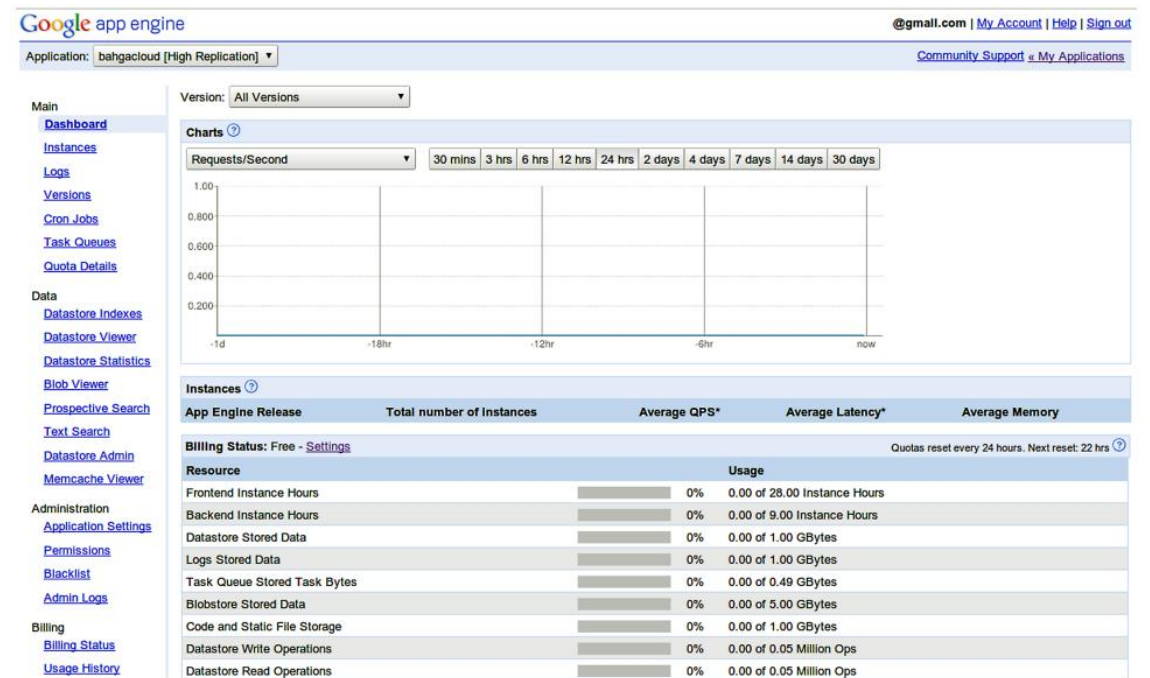
- Windows Azure Table Service is a non-relational (No-SQL) database service from Microsoft.
- Data Model
 - The Azure Table Service data model consists of tables having multiple entities.
 - Tables are divided into some number of partitions, each of which can be stored on a separate machine.
 - Each partition in a table holds a specified number of entities, each containing as many as 255 properties.
 - Each property can be one of the several supported data types such as integers and strings.
- No Fixed Schema
 - Tables do not have a fixed schema and different entities in a table can have different properties.

Application Runtimes & Frameworks

- Cloud-based application runtimes and frameworks allow developers to develop and host applications in the cloud.
- Support for various programming languages
 - Application runtimes provide support for programming languages (e.g., Java, Python, or Ruby).
- Resource Allocation
 - Application runtimes automatically allocate resources for applications and handle the application scaling, without the need to run and maintain servers.

Google App Engine

- Google App Engine is the platform-as-a-service (PaaS) from Google, which includes both an application runtime and web frameworks.
- Runtimes
 - App Engine provides runtime environments for Java, Python, PHP and Go programming language.
- Sandbox
 - Applications run in a secure sandbox environment isolated from other applications.
 - The sandbox environment provides a limited access to the underlying operating system.



Google App Engine

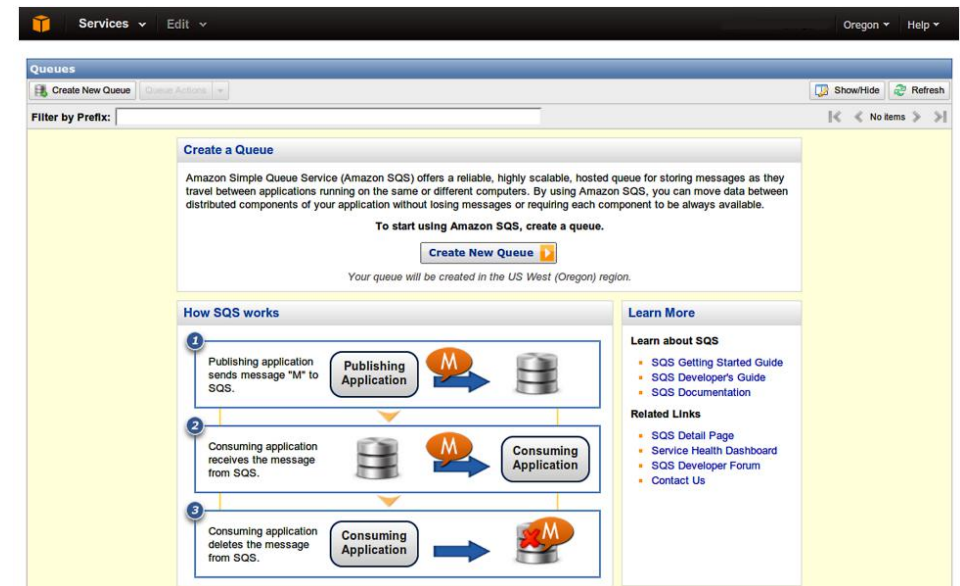
- Web Frameworks
 - App Engine provides a simple Python web application framework called webapp2. App Engine also supports any framework written in pure Python that speaks WSGI, including Django, CherryPy, Pylons, web.py, and web2py.
- Datastore
 - App Engine provides a no-SQL data storage service.
- Authentication
 - App Engine applications can be integrated with Google Accounts for user authentication.
- URL Fetch service
 - URL Fetch service allows applications to access resources on the Internet, such as web services or other data.
- Other services
 - Email service
 - Image Manipulation service
 - Memcache
 - Task Queues
 - Scheduled Tasks service

Windows Azure Web Sites

- Windows Azure Web Sites is a Platform-as-a-Service (PaaS) from Microsoft.
- Azure Web Sites allows you to host web applications in the Azure cloud.
- Shared & Standard Options.
 - In the shared option, Azure Web Sites run on a set of virtual machines that may contain multiple web sites created by multiple users.
 - In the standard option, Azure Web Sites run on virtual machines (VMs) that belong to an individual user.
- Azure Web Sites supports applications created in ASP.NET, PHP, Node.js and Python programming languages.
- Multiple copies of an application can be run in different VMs, with Web Sites automatically load balancing requests across them.

Queuing Services - Amazon Simple Queue Service

- Amazon Simple Queue Service (SQS) is a queuing service from Amazon.
- Short Messages
 - SQS is a distributed queue that supports messages of up to 256 KB in size.
- Multiple Writers/Readers
 - SQS supports multiple writers and readers and locks messages while they are being processed.
- High Availability
 - To ensure high availability for delivering messages, SQS service trade-offs on the first in, first out capability and does not guarantee that messages will be delivered in FIFO order.
 - Applications that require FIFO ordering of messages can place additional sequencing information in each message so that they can be re-ordered after retrieving from a queue.



Queuing Services - Google Task Queue Service

- Google Task Queues service is a queuing service from Google and is a part of the Google App Engine platform.
- Task queues allow applications to execute tasks in background.
- Tasks
 - Task is a unit of work to be performed by an application. The task objects consist of application-specific URL with a request handler for the task, and an optional data payload that parameterizes the task.
- Push Queue
 - Push Queue is the default queue that processes tasks based on the processing rate configured in the queue definition.
- Pull Queue
 - Pull Queues allow task consumers to lease a specific number of tasks for a specific duration. The tasks are processed and deleted before the lease ends.

Queuing Services - Windows Azure Queue Service

- Windows Azure Queue service is a queuing service from Microsoft.
- Azure Queue service allows storing large numbers of messages that can be accessed from anywhere in the world via authenticated calls using HTTP or HTTPS.
- Short Messages
 - The size of a single message can be up to 64KB.

Email Services

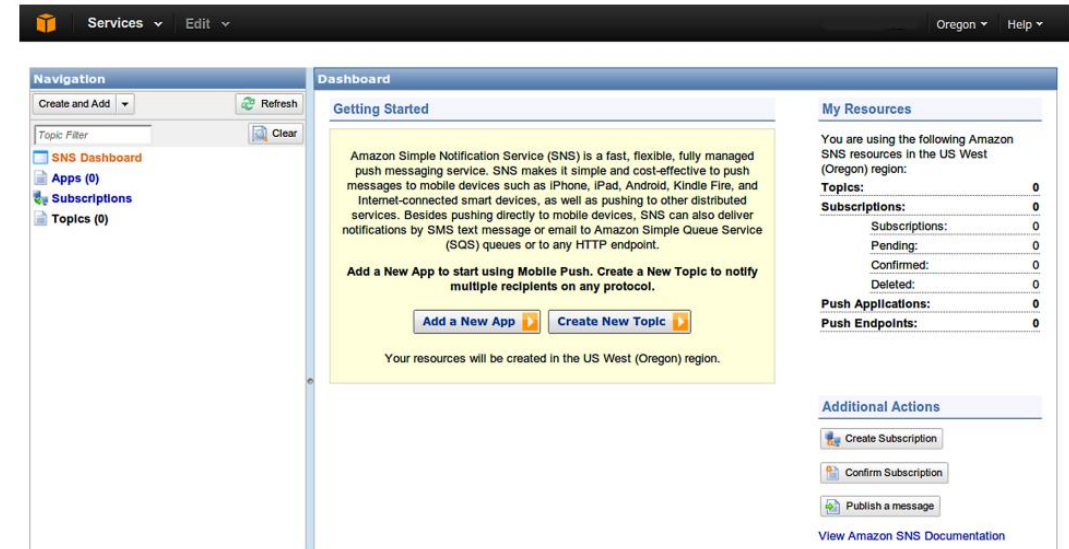
- Cloud-based email services allow applications hosted in the cloud to send emails.
- Amazon Simple Email Service
 - Amazon Simple Email Service is bulk and transactional email-sending service from Amazon
 - SES is an outbound-only email-sending service that allows applications hosted in the Amazon cloud to send emails such as marketing emails, transactional emails and other types of correspondence
 - To ensure high email deliverability, SES uses content filtering technologies to scan the outgoing email messages
 - SES service can be accessed and used from the SES console, the Simple Mail Transfer Protocol (SMTP) interface, or the SES API
- Google Email Service
 - Google Email service is part of the Google App Engine platform that allows App Engine applications to send email messages on behalf of the app's administrators, and on behalf of users with Google Accounts.
 - App Engine apps can also receive emails. Apps send messages using the Mail service and receive messages in the form of HTTP requests initiated by App Engine and posted to the app.

Notification Services

- Cloud-based notification services or push messaging services allow applications to push messages to internet connected smart devices such as smartphones, tablets, etc.
- Push messaging services are based on publish-subscribe model in which consumers subscribe to various topics/channels provided by a publisher/producer.
- Whenever new content is available on one of those topics/channels, the notification service pushes that information out to the consumer.
- Push notifications are used for such smart devices as they help in displaying the latest information while remaining energy efficient.
- Consumer applications on such devices can increase their consumer engagement with the help of push notifications.

Notification Services - Amazon Simple Notification Service

- Amazon Simple Notification Service is a push messaging service from Amazon.
- SNS has two types of clients:
 - Publishers
 - Publishers communicate asynchronously with subscribers by producing and sending messages to topics. A topic is a logical access point and a communication channel.
 - Subscribers.
 - Subscribers are the consumers who subscribe to topics to receive notifications.
- SNS can deliver notifications as SMS, email, or to SQS queues, or any HTTP endpoint.



Google Cloud Messaging

- Google Cloud Messaging for Android provides push messaging for Android devices.
- GCM allows applications to send data from the application servers to their users' Android devices, and also to receive messages from devices on the same connection.
- Notifying Android Apps
 - GCM is useful for notifying applications on Android devices that there is new data to be fetched from the application servers.
- Short Messages
 - GCM supports messages with payload data upto 4 KB.
- Send-to-Sync
 - GCM provides a 'send-to-sync' message capability that can be used to inform an application to sync data from the server.
- GCM for Chrome
 - Google Cloud Messaging for Chrome is another notification service from Google that allows messages to be delivered from the cloud to apps and extensions running in Chrome.

Windows Azure Notification Hubs

- Windows Azure Notification Hubs is a push notification service from Microsoft.
- Common Interface
 - Provides a common interface to send notifications to all major mobile platforms including Windows Store/Windows Phone 8, iOS, and Android.
- Platform Notification Systems
 - Platform specific infrastructures called Platform Notification Systems (PNS) are used to deliver notification messages.
 - Devices register their PNS handles with the Notification Hub.
 - Each notification hub contains credentials for each supported PNS.
 - These credentials are used to connect to the PNSs and send push notifications to the applications.

Media Services

- Cloud service providers provide various types of media services that can be used by applications for manipulating, transforming or transcoding media such as images, videos, etc.
- Amazon Elastic Transcoder
 - Amazon Elastic Transcoder is a cloud-based video transcoding service from Amazon.
 - Elastic Transcoder can be used to convert video files from their source format into various other formats that can be played on devices such as desktops, mobiles, tablets, etc.
- Google Images Manipulation Service
 - Google Images Manipulation service is a part of the Google App Engine platform. Image Manipulation service provides the capability to resize, crop, rotate, flip and enhance images.
- Windows Azure Media Services
 - Windows Azure Media Services provides the various media services such as encoding & format conversion, content protection and on-demand and live streaming capabilities.

Content Delivery Services

- Cloud-based content delivery service include Content Delivery Networks (CDNs).
- CDN is a distributed system of servers located across multiple geographic locations to serve content to end-users with high availability and high performance.
- CDNs are useful for serving static content such as text, images, scripts, etc., and streaming media.
- CDNs have a number of edge locations deployed in multiple locations, often over multiple backbones.
- Requests for static for streaming media content that is served by a CDN are directed to the nearest edge location.
- Amazon CloudFront
 - Amazon CloudFront is a content delivery service from Amazon. CloudFront can be used to deliver dynamic, static and streaming content using a global network of edge locations.
- Windows Azure Content Delivery Network
 - Windows Azure Content Delivery Network (CDN) is the content delivery service from Microsoft.

Analytics Services

- Cloud-based analytics services allow analyzing massive data sets stored in the cloud either in cloud storages or in cloud databases using programming models such as MapReduce.
- Amazon Elastic MapReduce
 - Amazon Elastic MapReduce is the MapReduce service from Amazon based the Hadoop framework running on Amazon EC2 and S3
 - EMR supports various job types such as Custom JAR, Hive program, Streaming job, Pig programs and Hbase
- Google MapReduce Service
 - Google MapReduce Service is a part of the App Engine platform and can be accessed using the Google MapReduce API.
- Google BigQuery
 - Google BigQuery is a service for querying massive datasets. BigQuery allows querying datasets using SQL-like queries.
- Windows Azure HDInsight
 - Windows Azure HDInsight is an analytics service from Microsoft. HDInsight deploys and provisions Hadoop clusters in the Azure cloud and makes Hadoop available as a service.

Deployment & Management Services

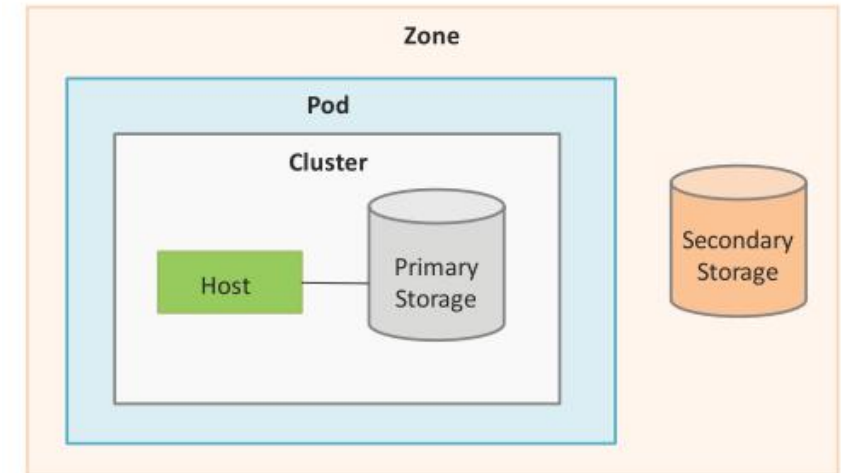
- Cloud-based deployment & management services allow you to easily deploy and manage applications in the cloud. These services automatically handle deployment tasks such as capacity provisioning, load balancing, auto-scaling, and application health monitoring.
- Amazon Elastic Beanstalk
 - Amazon provides a deployment service called Elastic Beanstalk that allows you to quickly deploy and manage applications in the AWS cloud.
 - Elastic Beanstalk supports Java, PHP, .NET, Node.js, Python, and Ruby applications.
 - With Elastic Beanstalk you just need to upload the application and specify configuration settings in a simple wizard and the service automatically handles instance provisioning, server configuration, load balancing and monitoring.
- Amazon CloudFormation
 - Amazon CloudFormation is a deployment management service from Amazon.
 - With CloudFront you can create deployments from a collection of AWS resources such as Amazon Elastic Compute Cloud, Amazon Elastic Block Store, Amazon Simple Notification Service, Elastic Load Balancing and Auto Scaling.
 - A collection of AWS resources that you want to manage together are organized into a stack.

Identity & Access Management Services

- Identity & Access Management (IDAM) services allow managing the authentication and authorization of users to provide secure access to cloud resources.
- Using IDAM services you can manage user identifiers, user permissions, security credentials and access keys.
- Amazon Identity & Access Management
 - AWS Identity and Access Management (IAM) allows you to manage users and user permissions for an AWS account.
 - With IAM you can manage users, security credentials such as access keys, and permissions that control which AWS resources users can access.
 - Using IAM you can control what data users can access and what resources users can create.
 - IAM also allows you to control creation, rotation, and revocation security credentials of users.
- Windows Azure Active Directory
 - Windows Azure Active Directory is an Identity & Access Management Service from Microsoft.
 - Azure Active Directory provides a cloud-based identity provider that easily integrates with your on-premises active directory deployments and also provides support for third party identity providers.
 - With Azure Active Directory you can control access to your applications in Windows Azure.

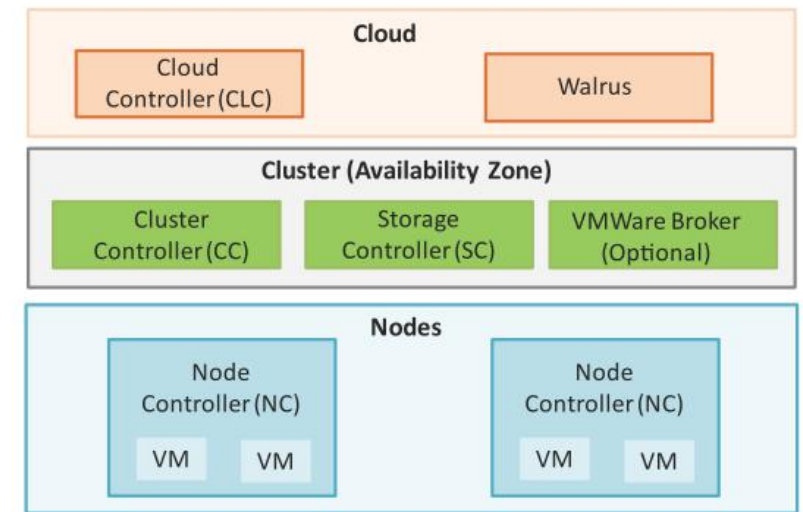
Open Source Private Cloud Software - CloudStack

- Apache CloudStack is an open source cloud software that can be used for creating private cloud offerings.
- CloudStack manages the network, storage, and compute nodes that make up a cloud infrastructure.
- A CloudStack installation consists of a Management Server and the cloud infrastructure that it manages.
- Zones
 - The Management Server manages one or more zones where each zone is typically a single datacenter.
- Pods
 - Each zone has one or more pods. A pod is a rack of hardware comprising of a switch and one or more clusters.
- Cluster
 - A cluster consists of one or more hosts and a primary storage. A host is a compute node that runs guest virtual machines.
- Primary Storage
 - The primary storage of a cluster stores the disk volumes for all the virtual machines running on the hosts in that cluster.
- Secondary Storage
 - Each zone has a secondary storage that stores templates, ISO images, and disk volume snapshots.



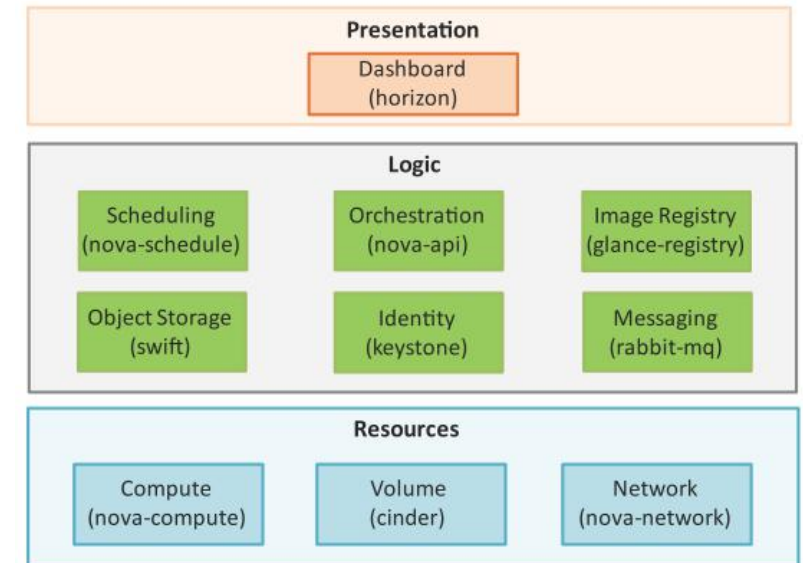
Open Source Private Cloud Software - Eucalyptus

- Apache CloudStack is an open source cloud software that can be used for creating private cloud offerings.
- CloudStack manages the network, storage, and compute nodes that make up a cloud infrastructure.
- A CloudStack installation consists of a Management Server and the cloud infrastructure that it manages.
- Zones
 - The Management Server manages one or more zones where each zone is typically a single datacenter.
- Pods
 - Each zone has one or more pods. A pod is a rack of hardware comprising of a switch and one or more clusters.
- Cluster
 - A cluster consists of one or more hosts and a primary storage. A host is a compute node that runs guest virtual machines.
- Primary Storage
 - The primary storage of a cluster stores the disk volumes for all the virtual machines running on the hosts in that cluster.
- Secondary Storage
 - Each zone has a secondary storage that stores templates, ISO images, and disk volume snapshots.



Open Source Private Cloud Software - OpenStack

- Eucalyptus is an open source private cloud software for building private and hybrid clouds that are compatible with Amazon Web Services (AWS) APIs.
- Node Controller
 - NC hosts the virtual machine instances and manages the virtual network endpoints.
- The cluster-level (availability-zone) consists of three components
 - Cluster Controller - which manages the virtual machines and is the front-end for a cluster.
 - Storage Controller – which manages the Eucalyptus block volumes and snapshots to the instances within its specific cluster. SC is equivalent to AWS Elastic Block Store (EBS).
 - VMWare Broker - which is an optional component that provides an AWS-compatible interface for VMware environments.
- At the cloud-level there are two components:
 - Cloud Controller - which provides an administrative interface for cloud management and performs high-level resource scheduling, system accounting, authentication and quota management.
 - Walrus - which is equivalent to Amazon S3 and serves as a persistent storage to all of the virtual machines in the Eucalyptus cloud. Walrus can be used as a simple Storage-as-a-Service



Further Reading

- Amazon Elastic Compute Cloud, <http://aws.amazon.com/ec2>
- Google Compute Engine, <https://developers.google.com/compute/>
- Windows Azure, <http://www.windowsazure.com/>
- Google App Engine, <http://appengine.google.com>
- Google App Engine, <https://developers.google.com/appengine/>
- Google Cloud Storage, <https://developers.google.com/storage/>
- Google BigQuery, <https://developers.google.com/bigquery/>
- Google Cloud Datastore, <http://developers.google.com/datastore/>
- Google Cloud SQL, <https://developers.google.com/cloud-sql/>
- CloudStack, <http://cloudstack.apache.org>
- Eucalyptus, <http://www.eucalyptus.com>
- OpenStack, <http://www.openstack.org>