

1. Compare scheduling techniques of virtual machine in cloud.
2. Explain the life cycle of a virtual machine with suitable diagram
3. What is virtual machine provisioning ? Discuss the steps for provisioning in detail.
4. Describe the layered virtualization technology architecture with suitable diagram
5. Paravirtualization **VS.** Fullvirtualization
6. Benefits of cloud computing over the traditional data center
7. Requirements of VMM & the Xen Architecture.
8. What is Hypervisor ? Discuss the different types of Hypervisor.
9. What kind of upfront costs are associated with building your own data center?
10. What are the core components of a data center?
11. Physical server vs. virtual machine
12. Comparative study among AWS, GCP and Azure
13. Types of data centers
14. What defines a modern data center?
15. Difference between Scheduling in Resource Management & Scheduling in Cloud.
16. Difference between Horizontal & Vertical Scaling.
17. What are load balancing algorithms?
18. Describe the EC2 feature in AWS
19. Prepare a case study to mention the current challenges existing in “IRCTC online reservation system” and the role of cloud computing to address these challenges.

1. Why wireless sensor network should be integrated with cloud environment?
2. Differentiate between cloud computing and mobile computing? Discuss the characteristics of mobile cloud computing with its advantages and disadvantages.
3. List out the various security and privacy issues in CCE.
4. Discuss the importance of Third party Auditor to ensure security in CCE, with the help of appropriate diagram.
5. Discuss the various components of sensor-cloud integration framework with appropriate diagram.
6. Discuss in brief about digital identity and content level security in cloud computing.
7. Explain the risks and security issues arises in cloud environment.
8. Explain CSA Cloud Security Architecture with suitable diagram.
9. Challenges, application, advantages & issues in cloud computing.



Mobile CLOUD Computing

Aradhana Behura
Dept. of CSE
KIIT , BBSR



Outline of the talk..

- Cloud, mobile and computing - MCC
- MCC - introduction
- MCC - applications
- MCC - issues
- MCC - challenges



What is Mobile Cloud Computing?

- *Mobile cloud computing (MCC)* at its simplest, refers to an infrastructure where both the data storage and data processing happen outside of the mobile device.
- Mobile cloud applications move the computing power and data storage away from the mobile devices and into powerful and centralized computing platforms located in clouds, which are then accessed over the wireless connection based on a thin native client.

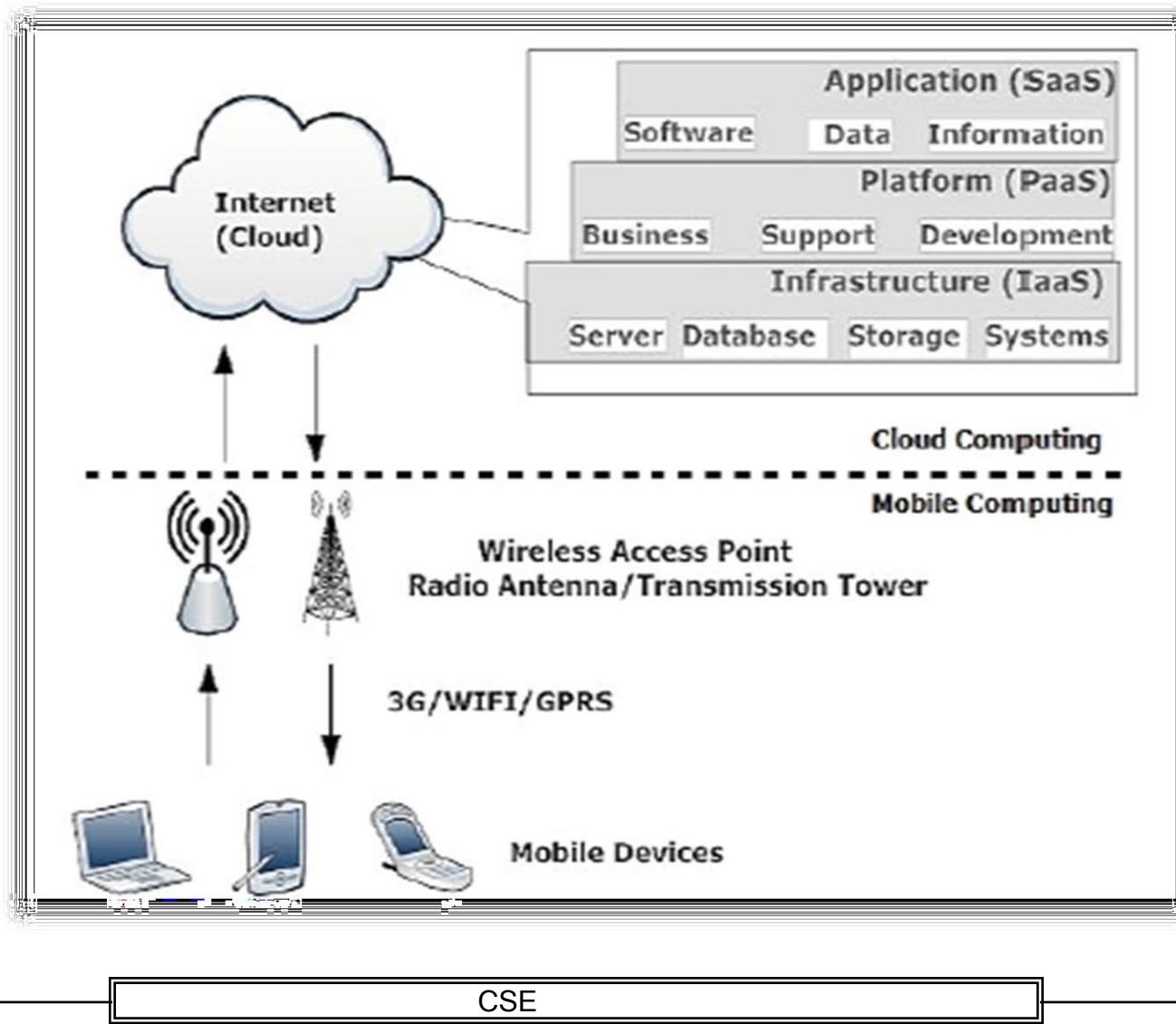
Cloud, Mobile and MCC

Cloud Computing – A paradigm of web-based computing where shared resources are provided on demand

Mobile Computing – Using portable devices to run standalone applications or use wireless media to access services

Mobile Cloud Computing – Augment mobile devices with content and processing capabilities using cloud

MCC - pictorial representation



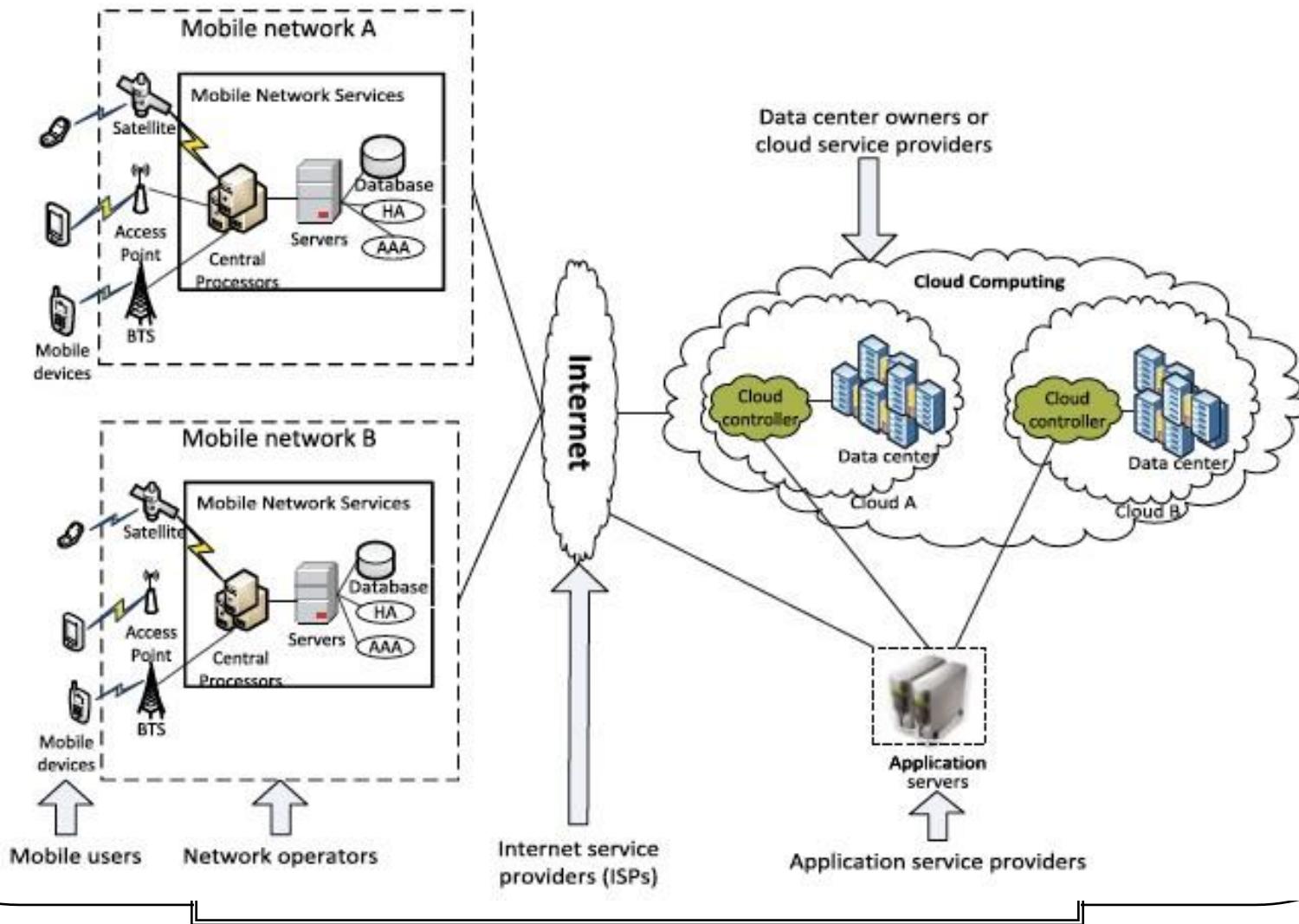
Why Mobile Cloud Computing?

- Mobile devices face many **resource challenges** (battery life, storage, bandwidth etc.)
- Cloud computing offers advantages to users by allowing them to use infrastructure, platforms and software by cloud providers at **low cost** and elastically in an **On-demand** fashion.
- Mobile cloud computing provides mobile users with data storage and processing services in clouds, obviating the need to have a powerful device configuration (e.g. CPU speed, memory capacity etc), as all resource-intensive computing can be performed in the cloud.

MCC Popularity

- According to a recent study by ABI Research, more than 240 million business will use cloud services through mobile devices by 2015.
- That traction will push the revenue of mobile cloud computing to \$5.2 billion.
- Mobile cloud computing is a highly promising trend for the future of mobile computing.
- Also - hosting of services, payment systems, analytics, application development and monitoring..

MCC Architecture



MCC Architecture

- Mobile devices are connected to the mobile networks via base stations that establish and control the connections and functional interfaces between the networks and mobile devices.
- Mobile users' requests and information are transmitted to the central processors that are connected to servers providing mobile network services.
- The subscribers' requests are delivered to a cloud through the Internet.
- In the cloud, cloud controllers process the requests to provide mobile users with the corresponding cloud services.

Advantages of MCC

- Extending battery lifetime:
 - Computation offloading migrates large computations and complex processing from resource-limited devices (i.e., mobile devices) to resourceful machines (i.e., servers in clouds).
 - Remote application execution can save energy significantly.
 - Many mobile applications take advantages from task migration and remote processing.

Advantages of MCC

- Improving data storage capacity and processing power:
 - MCC enables mobile users to store/access large data on the cloud.
 - MCC helps reduce the running cost for computation intensive applications.
 - Mobile applications are not constrained by storage capacity on the devices because their data now is stored on the cloud.

Advantages of MCC

- Improving reliability and availability:
 - Keeping data and application in the clouds reduces the chance of lost on the mobile devices.
 - MCC can be designed as a comprehensive data security model for both service providers and users:
 - Protect copyrighted digital contents in clouds.
 - Provide security services such as virus scanning, malicious code detection, authentication for mobile users.
 - With data and services in the clouds, they are always(almost) available even when the users are moving.

Advantages of MCC

- Dynamic provisioning:
 - Dynamic on-demand provisioning of resources on a fine-grained, self-service basis
 - No need for advanced reservation
- Scalability:
 - Mobile applications can be performed and scaled to meet the unpredictable user demands
 - Service providers can easily add and expand a service

Advantages of MCC

- Multi-tenancy:
 - Service providers can share the resources and costs to support a variety of applications and large no. of users.
- Ease of Integration:
 - Multiple services from different providers can be integrated easily through the cloud and the Internet to meet the users" demands.

MCC Applications

- Mobile Commerce:
 - M-commerce allows business models for commerce using mobile devices.
 - Examples: Mobile financial, mobile advertising, mobile shopping...
 - M-commerce applications face various challenges (low bandwidth, high complexity of devices, security, ...)
 - Integrated with cloud can help address these issues
 - Example: Combining 4G/5G and cloud to increase data processing speed and security level.

MCC Applications

- Mobile Learning:
 - M-learning combines e-learning and mobility
 - Traditional m-learning has limitations on high cost of devices/network, low transmission rate, limited educational resources
 - Cloud-based m-learning can solve these limitations
 - Enhanced communication quality between students and teachers
 - Help learners access remote learning resources
 - A natural environment for collaborative learning

MCC Applications

- Mobile Healthcare:
 - M-healthcare is to minimize the limitations of traditional medical treatment (eg. Small storage, security/privacy, medical errors, ...)
 - M-healthcare provides mobile users with convenient access to resources(eg. medical records)
 - M-healthcare offers hospitals and healthcare organizations a variety of on-demand services on clouds
 - Examples:
 - Comprehensive health monitoring services
 - Intelligent emergency management system
 - Health-aware mobile devices (detect pulse-rate, blood pressure, level of alcohol etc)
 - Pervasive access to healthcare information
 - Pervasive lifestyle incentive management (to manage healthcare expenses)

MCC Applications

- Mobile Gaming:
 - M-game is a high potential market generating revenues for service providers.
 - Can completely offload game engine requiring large computing resource (e.g., graphic rendering) to the server in the cloud.
 - Offloading can also save energy and increase game playing time (eg. MAUI allows fine-grained energy-aware offloading of mobile codes to a cloud)
 - Rendering adaptation technique can dynamically adjust the game rendering parameters based on communication constraints and gamers" demands

MCC Applications

- Assistive technologies:
 - Pedestrian crossing guide for blind and visually-impaired
 - Mobile currency reader for blind and visually impaired
 - Lecture transcription for hearing impaired students
- Other applications:
 - Sharing photos/videos
 - Keyword-based, voice-based, tag-based searching
 - Monitoring a house, smart home systems
 - ...

MCC Issues

- Mobile communication issues:
 - Low bandwidth: One of the biggest issues, because the radio resource for wireless networks is much more scarce than wired networks
 - Service availability: Mobile users may not be able to connect to the cloud to obtain a service due to traffic congestion, network failures, mobile signal strength problems
 - Heterogeneity: Handling wireless connectivity with highly heterogeneous networks to satisfy MCC requirements (always-on connectivity, on-demand scalability, energy efficiency) is a difficult problem

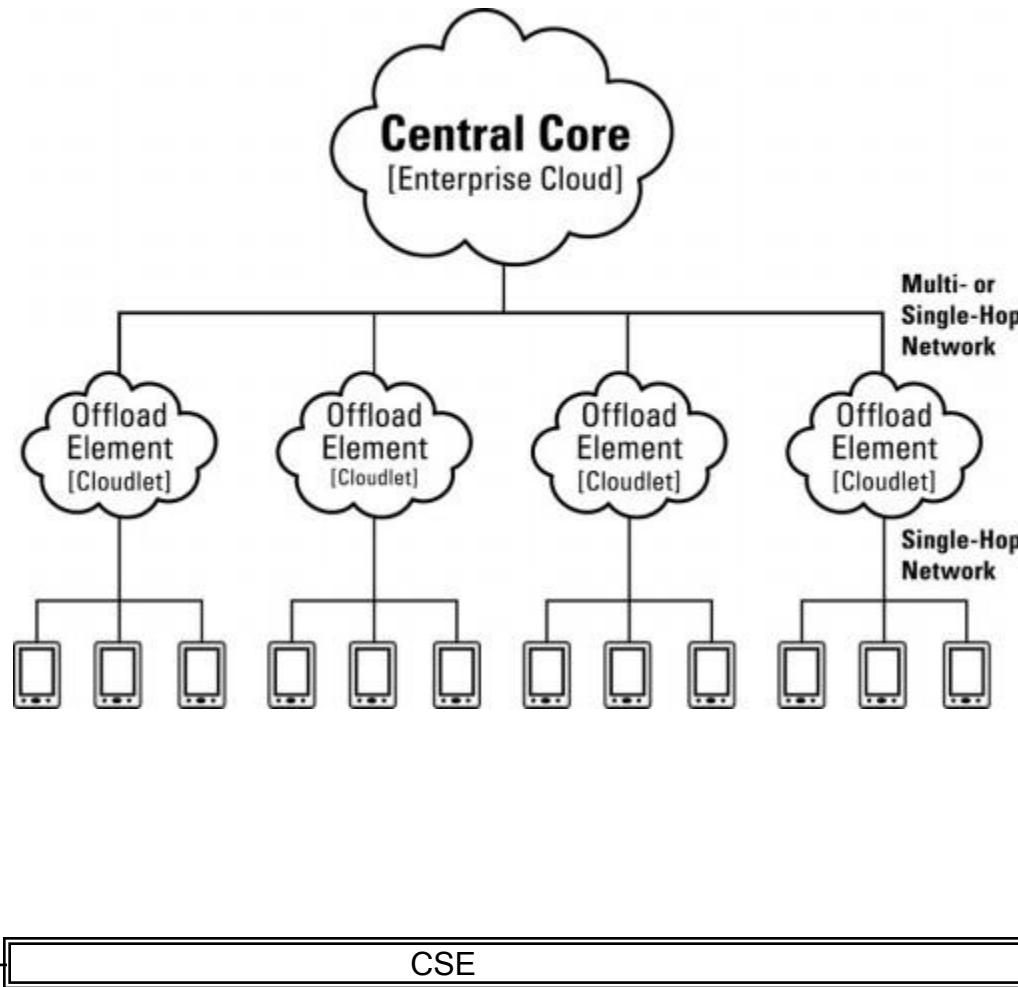
MCC Issues

- Computing issues:

Computation offloading:

- One of the main features of MCC
- Offloading is not always effective in saving energy
- It is critical to determine whether to offload and which portions of the service codes to offload
- Two types:
 - Offloading in a static environment
 - Offloading in a dynamic environment

Offload mechanism



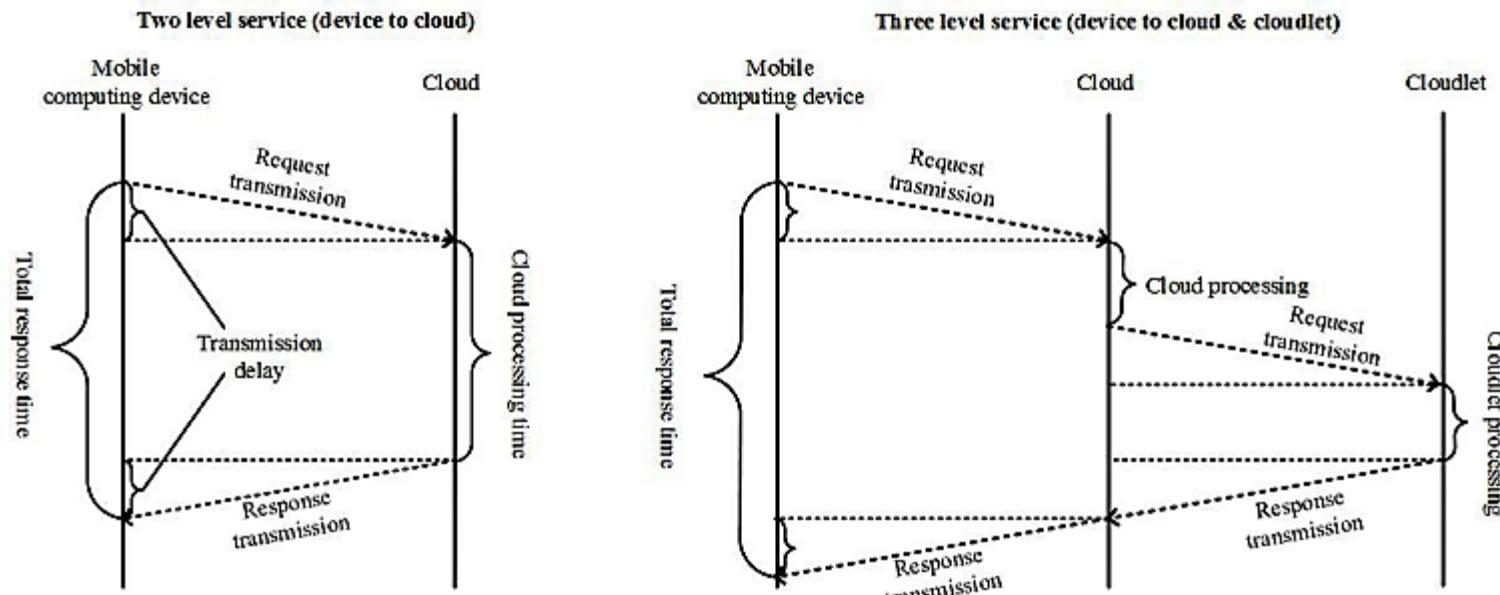


FIGURE 4. Offloading scenarios.

Computation Offloading Approaches in a Static Environment

- Kumar and Lu suggest a program partitioning based on estimation of energy consumption before execution
- Optimal program partitioning for offloading is dynamically calculated based on the trade-off between the communication and computation costs at run time.

K. Kumar and Y. Lu, “Cloud Computing for Mobile Users: Can Offloading Computation Save Energy,” IEEE Computer, vol. 43, no. 4, April 2010.

Computation Offloading Approaches in a Static Environment

- Li et al. present an offloading scheme based on profiling information about computation time and data sharing at the level of procedure calls.
- A cost graph is constructed and a branch-and-bound algorithm is applied to minimize the total energy consumption of computation and the total data communication cost.

Z. Li, C. Wang, and R. Xu, “Computation offloading to save energy on handheld devices: a partition scheme,” in Proc 2001 Intl Conf on Compilers, architecture, and synthesis for embedded systems (CASES), pp. 238-246, Nov 2001.

Computation Offloading Issues in a Dynamic Environment

- Offloading in a dynamic network environment (e.g., changing connection status and bandwidth) is harder.
- Environment changes can cause additional problems.
- The transmitted data may not reach the destination
- The data executed on the server could be lost when it has to be returned to the sender.

Computation Offloading Approaches in a Dynamic Environment

- Ou et al. analyze offloading systems in wireless environments
- They consider three circumstances of executing an application to estimate the efficiency of offloading.
 - performed locally (without offloading)
 - performed in ideal offloading systems (without failures)
 - performed with the presence of offloading and failure recoveries (re-offload after failure)

S. Ou, K. Yang, A. Liotta, and L. Hu. “Performance Analysis of Offloading Systems in Mobile Wireless Environments,” in Proc IEEE Intl Conf on Communications (ICC), pp. 1821, August 2007.

MCC Security Issues

- Protecting user privacy and data/application secrecy from adversaries is key to establish and maintain consumers" trust in the mobile platform, especially in MCC.
- MCC security issues have two main categories:
 - Security for mobile users
 - Securing data on clouds

Security for Mobile Users

- Mobile devices are exposed to numerous security threats like malicious codes and their vulnerability.
- GPS can cause privacy issues for subscribers.
- Security for mobile applications:
 - Installing and running security software are the simplest ways to detect security threats.
 - Mobile devices are resource constrained, protecting them from the threats is more difficult than that for resourceful devices.

Privacy Issues in MCC

- Location based services (LBS) faces a privacy issue on mobile users" provide private information such as their current location.
- This problem becomes even worse if an adversary knows user"s important information.

Context-aware Mobile Cloud Services

- It is important to fulfill mobile users' satisfaction by monitoring their preferences and providing appropriate services to each of the users.
- Context-aware mobile cloud services try to utilize the local contexts (e.g., data types, network status, device environments, and user preferences) to improve the quality of service (QoS).

Mobile Service Clouds

- When a customer uses a service, the request firstly goes to a service gateway which will choose an appropriate primary proxy to meet the requirements and then sends the result to the user.
- In disconnection, MSCs will establish transient proxies for mobile devices to monitor the service path, and support dynamic reconfiguration.
- The model addresses the disconnection issue and can maintain the QoS at an acceptable level.

CSE

Open Issues in MCC

- Network Access Management:
 - An efficient network access management not only improves link performance but also optimizes bandwidth usage.
 - Cognitive radio can be expected as a solution to achieve the wireless access management.
 - Can automatically changes its transmission or reception parameters, in a way where the wireless communications can have spectrum agility in terms of selecting available wireless channels opportunistically.
 - Integrated with MCC for better spectrum utilization

Open Issues in MCC

- Quality of Service:
 - How to ensure QoS is still a big issue, especially on network delay.
 - CloneCloud and Cloudlets are expected to reduce the network delay.
 - CloneCloud uses nearby computers or data centers to increase the speed of smart phone applications.
 - The idea is to clone the entire set of data and applications from the smartphone onto the cloud and to selectively execute some operations on the clones, reintegrating the results back into the smartphone.

Open Issues in MCC

- Quality of Service:

- A cloudlet is a trusted, resource-rich computer or cluster of computers which is well-connected to the Internet and available for use by nearby mobile devices with on one-hop wireless connection.
- Mobile users may meet the demand for real-time interactive response by low-latency, one-hop, high-bandwidth wireless access to the cloudlet.
- Can help mobile users overcome the limits of cloud computing as WAN latency and low bandwidth.

Open Issues in MCC

- Pricing:
 - MCC involves with both mobile service provider (MSP) and cloud service provider (CSP) with different services management, customers management, methods of payment and prices.
 - This will lead to many issues.
 - The business model including pricing and revenue sharing has to be carefully developed for MCC.

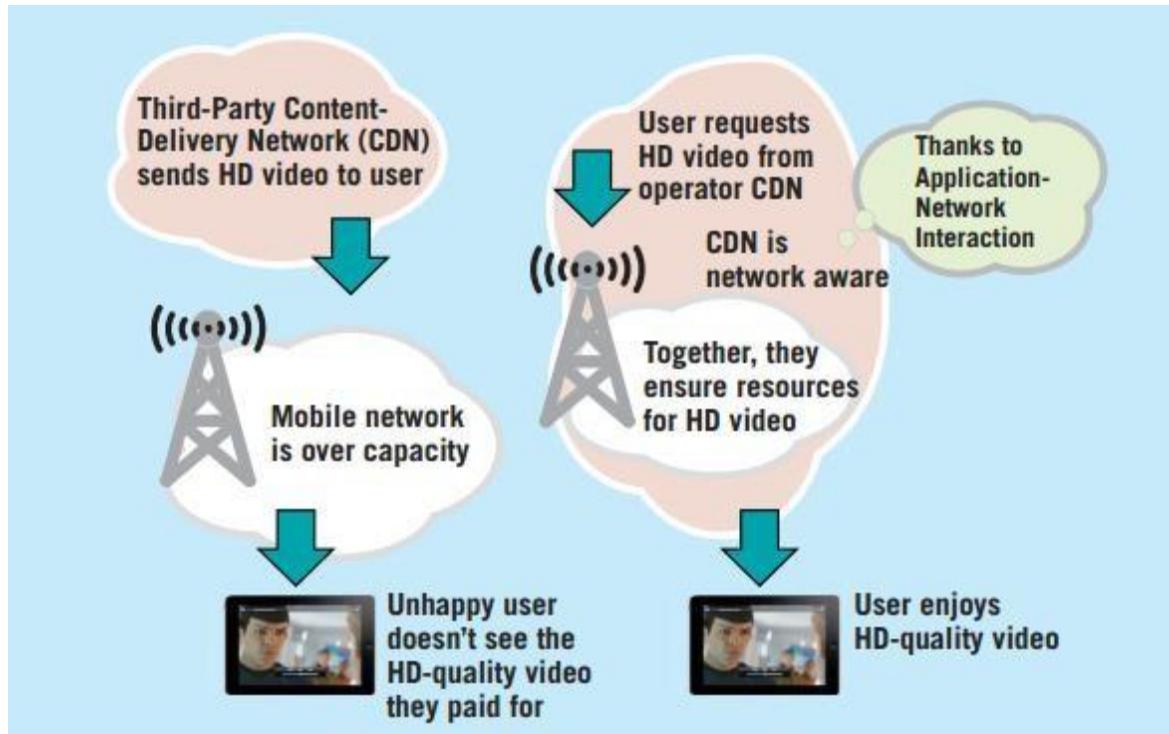
Open Issues in MCC

- Standard Interface:
 - Interoperability becomes an important issue when mobile users need to interact with the cloud.
 - Web interfaces may not be the best option.
 - It is not specifically designed for mobile devices.
 - May have more overhead.
 - Compatibility among devices for web interface could be an issue.
 - Standard protocol, signaling, and interface for interacting between mobile users and cloud would be required. (HTML5 & CSS3)

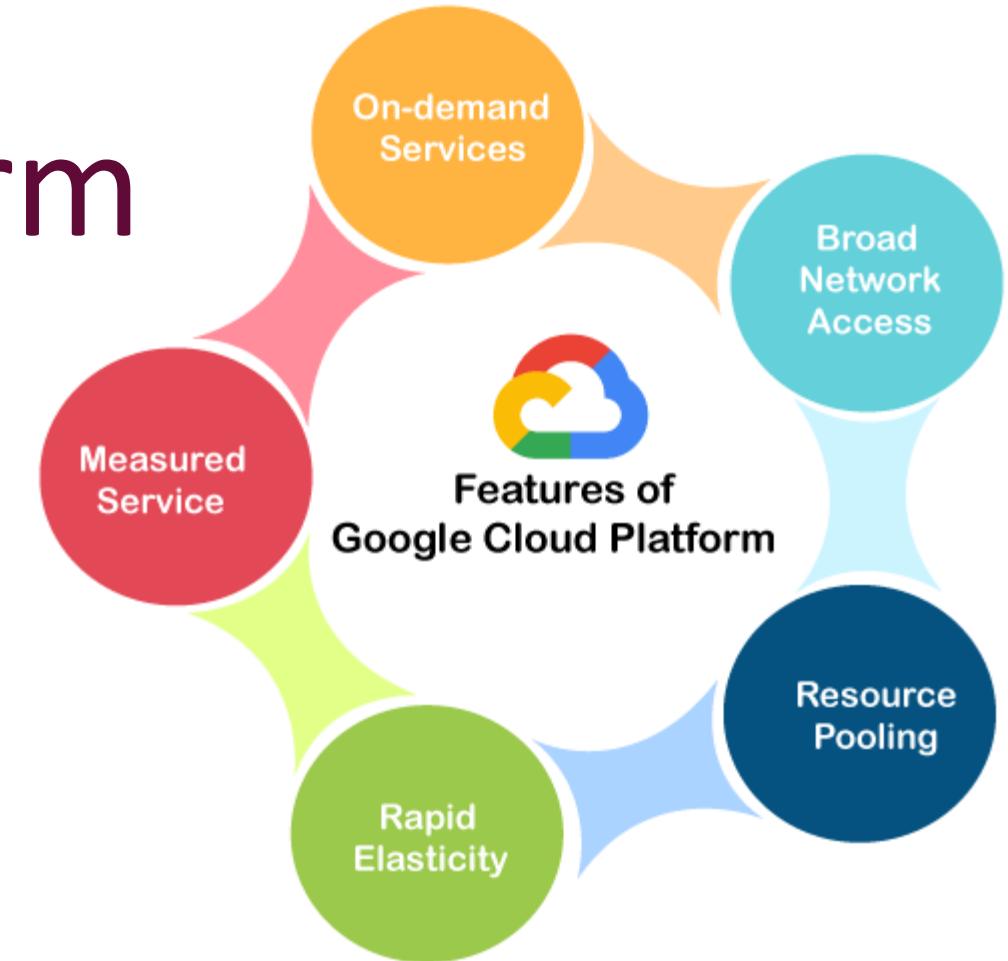
Open Issues in MCC

- Service Convergence:
 - Services will be differentiated according to the types, cost, availability and quality.
 - A single cloud may not be enough to meet mobile user's demands.
 - New scheme is needed in which the mobile users can utilize multiple cloud in a unified fashion.
 - The scheme should be able to automatically discover and compose services for user.
 - Sky computing is a model where resources from multiple clouds providers are leveraged to create a large scale distributed infrastructure.
 - The mobile sky computing will enable providers to support a cross-cloud communication and enable users to implement mobile services and applications.
 - Service integration (i.e., convergence) would need to be explored.

THE CONVERGENCE OF MOBILE NETWORK AND SERVICE PROVIDER



Google Cloud Platform



What is Cloud Computing?

- Cloud computing is defined as the services offered through remote servers on the internet. These services might include database storage, applications, compute power and other IT resources over the pay-as-you-go pricing approach. The remote server allows users to save, modify, or process data on the internet or cloud-based platform instead of storing it on a local server or their devices.
- Cloud computing is evolving due to fast performance, better manageability, and less maintenance. It helps organizations to minimize the number of resources and overall infrastructure costs. It also helps IT teams focus on the important applications, services, and processes and achieve the company's goals.

Cont.

Typically, cloud-computing providers offer their services according to the following three standard models:

- Platform as a Service (PaaS)
- Software as a Service (SaaS)
- Infrastructure as a Service (IaaS)

What is Google Cloud Platform?

- Google Cloud Platform (GCP) is a suite of cloud computing services provided by Google. **It is a public cloud computing** platform consisting of a variety of services like **computing, storage, networking, application development, Big Data, and more**, which run on the same cloud infrastructure that Google uses internally for its end-user products, such as **Google Search, Photos, Gmail and YouTube**, etc.
- The services of GCP can be accessed by software developers, cloud administrators and IT professionals over the Internet or through a dedicated network connection.

Why Google Cloud Platform?

- Google Cloud Platform is known as one of the leading cloud providers in the **IT** field.
- The services and features can be easily accessed and used by software developers and users with little technical knowledge.
- Google has been on top amongst its competitors, offering the highly scalable and most reliable platform for building, testing and deploying applications in a real-time environment.

Cont.

- Apart from this, GCP was announced as the leading cloud platform in the Gartner's IaaS Magic Quadrant in 2018. Gartner is one of the leading research and advisory company. Gartner organized a campaign where Google Cloud Platform was compared with other cloud providers, and GCP was selected as one of the top three providers in the market.
- Most companies use data centers because of the availability of cost forecasting, hardware certainty, and advanced control. However, they lack the necessary features to run and maintain resources in the data center. **GCP, on the other side, is a fully-featured cloud platform that includes:**

Cont.

- **Capacity:** Sufficient resources for easy scaling whenever required. Also, effective management of those resources for optimum performance.
- **Security:** Multi-level security options to protect resources, such as assets, network and OS -components.
- **Network Infrastructure:** Number of physical, logistical, and human-resource-related components, such as wiring, routers, switches, firewalls, load balancers, etc.
- **Support:** Skilled professionals for installation, maintenance, and support.
- **Bandwidth:** Suitable amount of bandwidth for peak load.
- **Facilities:** Other infrastructure components, including physical equipment and power resources.

Therefore, Google Cloud Platform is a viable option for businesses, especially when the businesses require an extensive catalogue of services with global recognition.

Benefits of Google Cloud Platform

Some of the main benefits of the Google Cloud Platform are explained below:

- **Best Pricing:** Google enables users to get Google Cloud hosting at the cheapest rates. The hosting plans are not only cheaper than other hosting platforms but also offer better features than others. GCP provides a pay-as-you-go option to the users where users can pay separately only for the services and resources they want to use.
- **Work from Anywhere:** Once the account is configured on GCP, it can be accessed from anywhere. That means that the user can use GCP across different devices from different places. It is possible because Google provides web-based applications that allow users to have complete access to GCP.
- **Private Network:** Google has its own network that enables users to have more control over GCP functions. Due to this, users achieve smooth performance and increased efficiency over the network.

Cont.

- **Scalable:** Users are getting a more scalable platform over the private network. Because Google uses fibreoptic cables to extend its network range, it is likely to have more scalability. Google is always working to scale its network because there can be any amount of traffic at any time.
- **Security:** There is a high number of security professionals working at Google. They always keep trying to secure the network and protect the data stored on servers. Additionally, Google uses an algorithm that encrypts all the data on the Cloud platform. This gives assurance to the users that their data is completely safe and secure from unauthorized sources.
- **Redundant Backup:** Google always keeps a backup of users' data with built-in redundant backup integration. In case a user has lost the stored data, it's not a big problem. Google always has a copy of the users' data unless the data is deleted forcefully. This adds data integrity, reliability and durability with GCP.

Key Features of Google Cloud Platform

The following are some key features of the Google Cloud Platform:

- **On-demand services:** Automated environment with web-based tools. Therefore, no human intervention is required to access the resources.
- **Broad network access:** The resources and the information can be accessed from anywhere.
- **Resource pooling:** On-demand availability of a shared pool of computing resources to the users.
- **Rapid elasticity:** The availability of more resources whenever required.
- **Measured service:** Easy-to-pay feature enables users to pay only for consumed services.

Working of Google Cloud Platform

- When a file is uploaded on the Google cloud, the unique metadata is inserted into a file.
- It helps identify the different files and track the changes made across all the copies of any particular file.
- All the changes made by individuals get synchronized automatically to the main file, also called a master file.
- GCP further updates all the downloaded files using metadata to maintain the correct records.

Example

Let's understand the working of GCP with a general example:

- Suppose that **MS Office** is implemented on Cloud to enable several people to work together. The primary aim of using cloud technology is to work on the same project at the same time. We can create and save a file on the cloud once we install a plugin for the MS Office suite. This will allow several people to edit a document at the same time. The owner can assign access to specific people to allow them to download and start editing the document in MS Office.
- Once users are assigned as an editor, they can use and edit the document's cloud copy as desired. The combined, the edited copy is generated which is known as the master document. GCP helps to assign a unique **URL** to each specific copy of the existing document given to different users. However, any of the authorized users' changes will be visible on all the copies of documents shared over the cloud. In case multiple changes are made to the same document, then GCP allows the owner to select the appropriate changes to keep.

GCP-IaaS



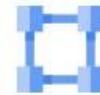
Compute Engine

Computing infrastructure in predefined or custom machine sizes to accelerate your cloud transformation.



Cloud Storage

Globally unified, scalable, and highly durable object storage for developers and enterprises.



Virtual Private Cloud

Managed networking functionality for your Google Cloud resources.



Persistent Disk

Reliable, high-performance block storage for virtual machine instances.

GCP-PaaS

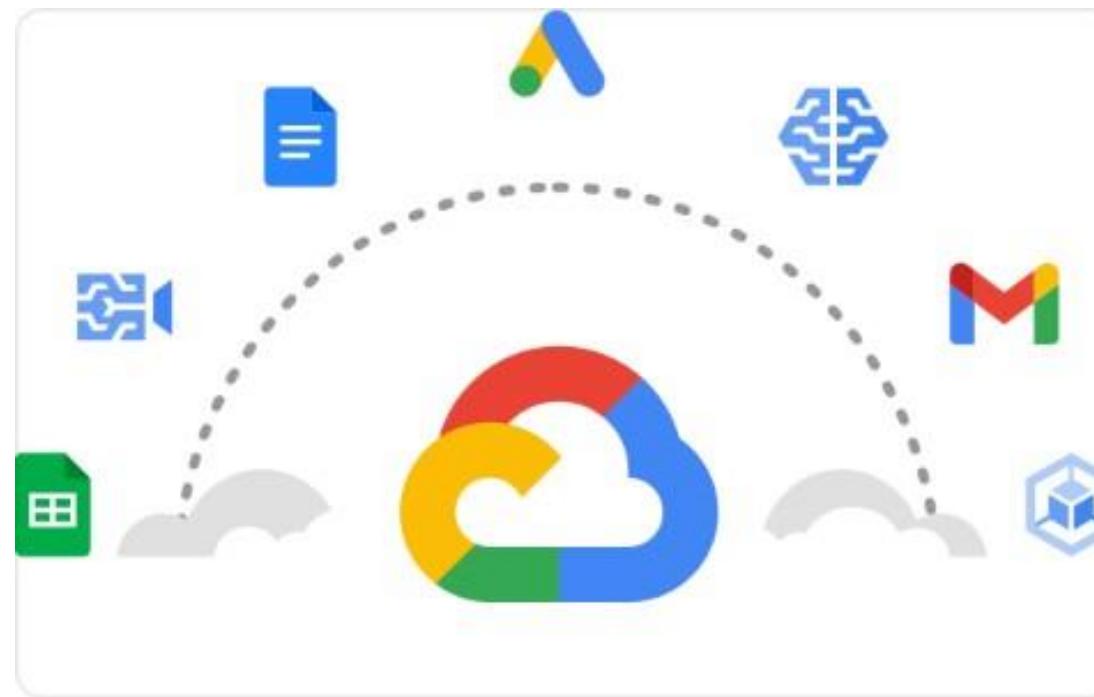


Cloud Run

Write code your way using
your favorite languages and
deploy your apps on
containers.



GCP-SaaS



Google Cloud Platform Services

Google provides a considerable number of services with several unique features. That is the reason why the Google Cloud Platform is continually expanding across the globe. Some of the significant services of GCP are:

- Compute Services
- Networking
- Storage Services
- Big Data
- Security and Identity Management
- Management Tools
- Cloud AI
- IoT (Internet of Things)



Compute Services [IaaS]

GCP offers a scalable range of computing services, such as:

- **Google App Engine:** It is a cloud computing platform that follows the concept of Platform-as-a-Service to deploy PHP, Java and other software. It is also used to develop and deploy web-based software in Google-managed data centers. The most significant advantage of Google App Engine is its automatic scaling capability. This means that the App Engine automatically allocates more resources for the application when there is an increase in requests.
- **Compute Engine:** It is a cloud computing platform that follows the concept of Infrastructure-as-a-Service to run Windows and Linux-based virtual machines. It is an essential component of GCP. It is designed on the same infrastructure used by the Google search engine, YouTube and other Google services.
- **Kubernetes Engines:** This computing service is responsible for offering a platform for automatic deployment, scaling, and other operations of application containers across clusters of hosts. The engine supports several container tools like a docker, etc.

Networking [IaaS]

GCP includes the following network services:

- **VPC:** VPC stands for Virtual Private Network. The primary function of VPC is to offer a private network with routing, IP allocation, and network firewall policies. This will help to create a secure environment for the application deployments.
- **Cloud Load Balancing:** As its name states, Cloud balancing is used to distribute workload across different computing resources to balance the entire system performance. This also results in cost reduction. The process also helps in minimizing the availability and maximise the capability of the resources.
- **Content Delivery Network:** CDN is a geographically distributed network of proxy servers and their data centers. The primary aim of using CDN is to provide maximum performance to the users. Additionally, it also helps deliver high availability of resources by equally distributing the related services to the end-users.

Storage Services [IaaS]

GCP has the following storage services:

- **Google Cloud Storage:** It is an online data storage web service that Google provides to its users to store and access data from anywhere. The service also includes a wide range of features like maximum performance, scalability, security and sharing.
- **Cloud SQL:** It is a web-service that enables users to create, manage, and use relational databases stored on Google Cloud servers. The service itself maintains and protects the databases, which helps users focus on their applications and other operations.
- **Cloud Bigtable:** It is known for its fast performance and highly manageable feature. It is a highly scalable NoSQL database service that allows collecting and retaining data from as low as 1 TB to hundreds of PB.

Big Data

GCP provides a variety of services related to big data; they are:

- **BigQuery:** It is a fully managed data analysis service by Google. The primary aim of Google BigQuery service is to help businesses to analyze Big Data. It offers a highly scalable data management option. This means BigQuery allows users to perform ad-hoc queries and share data insights across the web.
- **Google Cloud Datastore:** Google Cloud Datastore is a kind of datastore service that is fully managed, schema-less, and non-relational. This service enables businesses to perform automatic transactions and a rich set of queries. The main advantage of Google Cloud Datastore is the capability of automatic scaling. This means that the service can itself scale up and down, depending on the requirement of resources.
- **Google Cloud Dataproc:** It is a very fast and easy-to-use big data service offered by Google. It mainly helps in managing Hadoop and Spark services for distributed data processing. The service allows users to create Hadoop or Spark clusters sized according to the overall workload and can be accessed whenever users want them.

Security and Identity Management

GCP includes the following services related to Security and Identity management:

- **Cloud Data Loss Prevention API:** It is mainly designed to manage sensitive data. It helps users manage sensitive data elements like credit card details, debit card details, passport numbers, etc. It offers fast and scalable classification for sensitive data.
- **Cloud IAM:** It stands for Cloud Identity and Access Management. It is a framework that contains rules and policies and validates the authentication of the users for accessing the technology resources. That is why it is also known as Identity Management (IdM).

Management Tools

GCP includes the following services related to management tools:

- **Google Stackdriver:** Google Stackdriver service is primarily responsible for displaying the overall performance and diagnostics information. This may include insights into data monitoring, tracing, logging, error reporting, etc. The service also prompts an alert notification to public cloud users.
- **Google Cloud Console App:** It is a native mobile application powered by Google. The primary aim of this service is to enable users to manage the core features of Google Cloud services directly from their mobile devices anytime, anywhere. The primary functions of this service are alerting, monitoring, and performing critical actions on resources.

Cloud AI

When it comes to Cloud AI, GCP offers these services:

- **Cloud Machine Learning Engine:** It is another fully managed service that allows users to create Machine Learning models. The service is mainly used for those ML models, which are based on mainstream frameworks.
- **Cloud AutoML:** It is the type of service that is based on Machine Learning. It helps users to enter their data sets and gain access to quality trained pre-designed ML models. The service works by following Google's transfer learning and Neural Architecture Search method.

IoT (Internet of Things)

GCP contains the following **IoT** services:

- **Cloud IoT Core:** It is one of the fully managed core services. It allows users to connect, control, and ingest data from various devices that are securely connected to the Internet. This allows other Google cloud services to analyze, process, collect and visualize IoT data in real-time.
- **Cloud IoT Edge:** The Edge computing service brings memory and other computing-power resources near to the location where it is required.

Advantages of Google Cloud Platform

There are several advantages of using Google Cloud Platform, such as:

- **Google Cloud Offers Quick and Easy Collaboration:** Multiple users can access the data and simultaneously contribute their information. This is possible because the data is stored on the cloud servers, not on the user's personal computers.
- **Higher Productivity with Continuous Development:** Google is always working on adding new features and functionalities to provide higher productivity to the customers. Therefore, Google delivers frequent updates to its products and services.
- **Less Disruption with Adopting New Features:** Instead of pushing huge disruptive updates of changes, Google provides small updates weekly. This helps users to understand and adopt new features easily.

Cont.

- **Least or Minimal Data is stored on Vulnerable Devices:** Google does not store data on local devices unless a user explicitly tries to do it. This is because the data stored on local devices may get compromised compared to the cloud's data.
- **Users can access Google Cloud from Anywhere:** The best thing is that a user can easily access the information stored on Google cloud from anywhere because it is operated through web-based applications.
- **Google provides Maximum Security with its Robust Structure:** Google hires leading security professionals to protect user's data. Users get process-based and physical security features made by Google.

Cont.

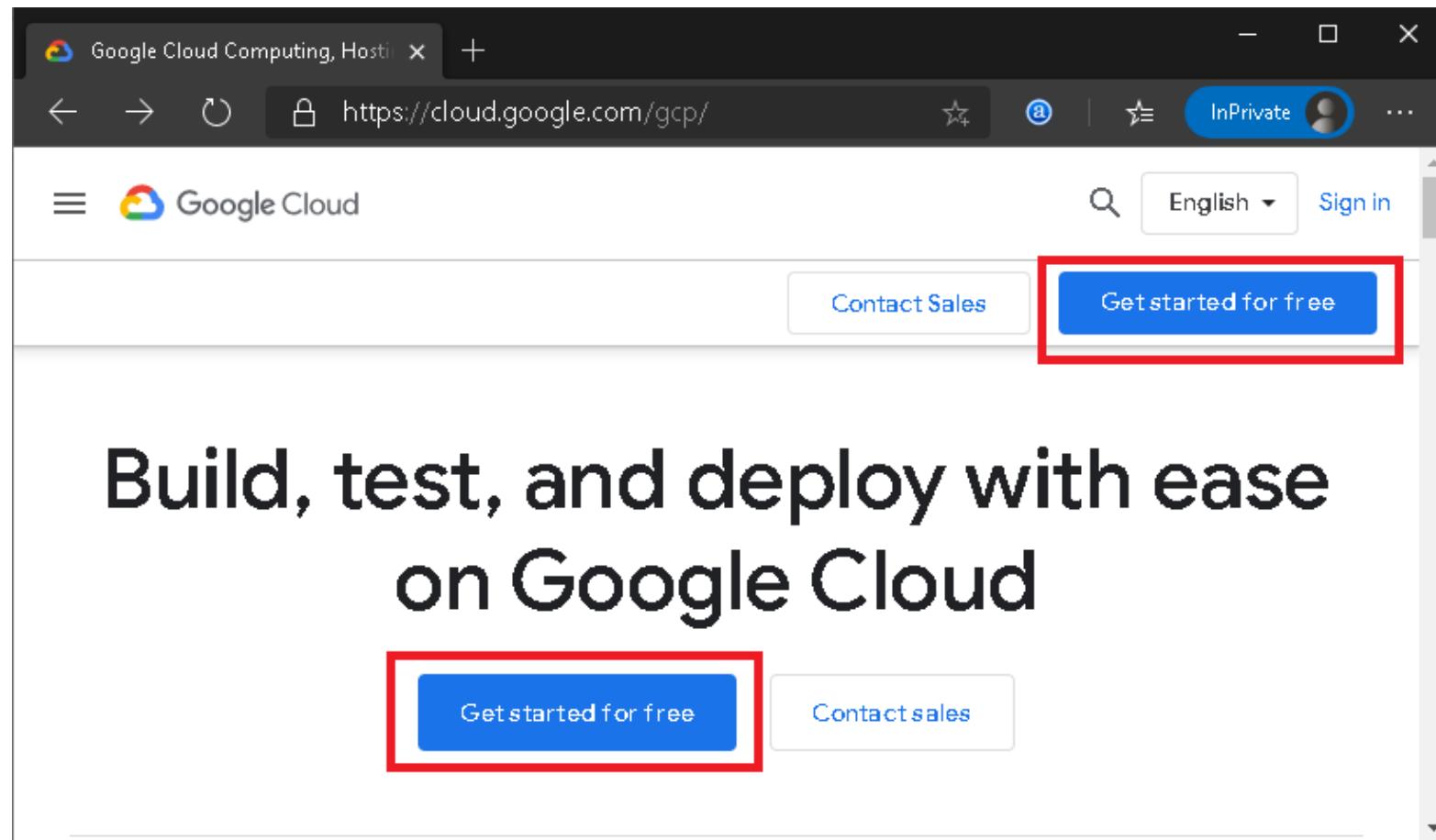
- **Users have Full Control over their Data:** Users gain full control over services and the data stored in Google Cloud. If a user does not want to use Google services any longer and wants to delete the cloud data, it can be easily performed.
- **Google provides Higher Uptime and Reliability:** Google uses several resources to provide higher and reliable up-time servers. If a data center is not working for technical issues, the system will automatically communicate with the secondary center without interruption visible to users.

Creating a Free Tier Account on GCP

Let's start with the steps of creating a free tier account on Google Cloud Platform:

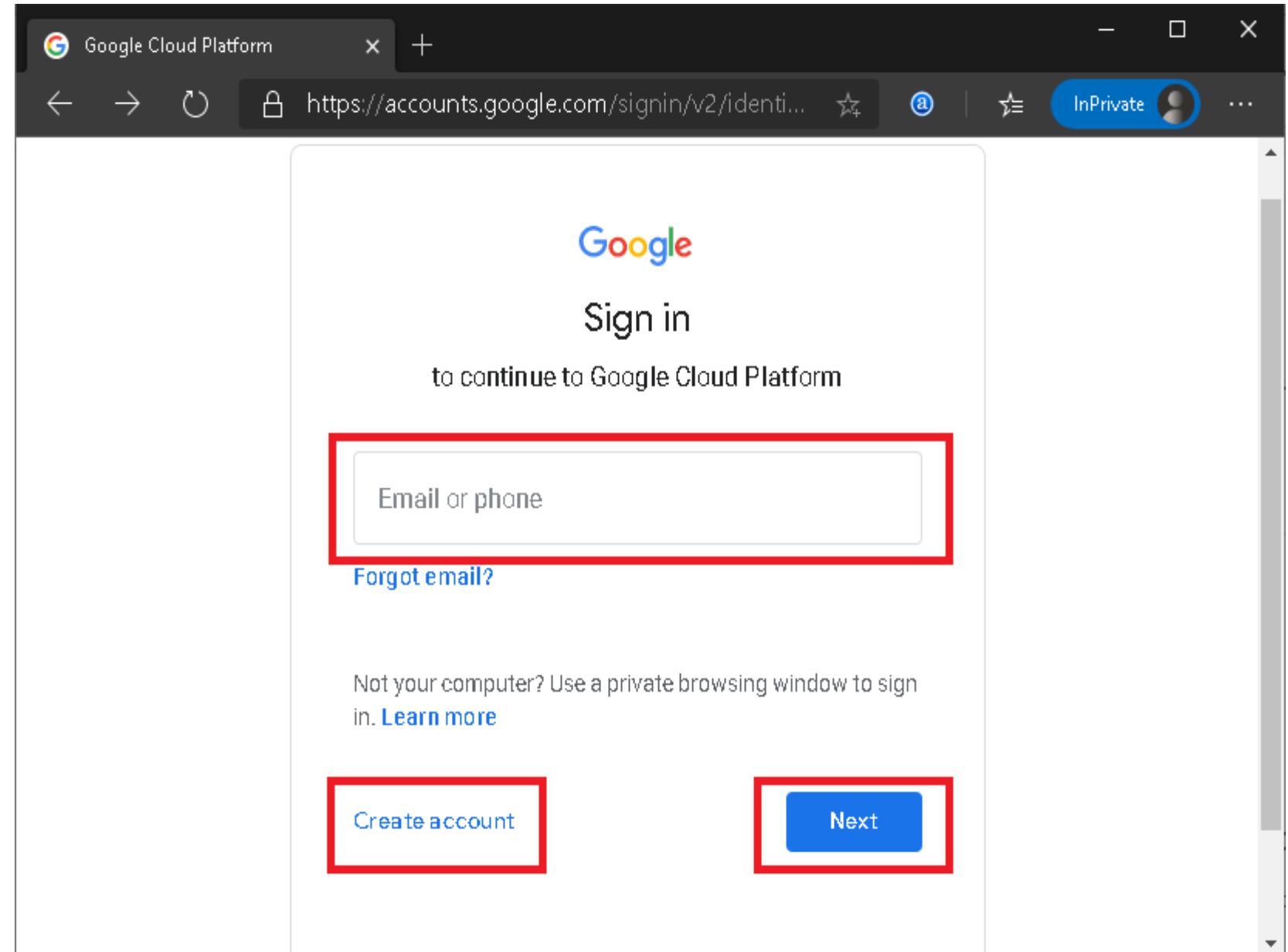
- **Step 1:** First, we are required to navigate to the following link: <https://cloud.google.com/gcp/>
- **Step 2:** On the next screen, we need to click on 'Get started for free', as shown below:

Navigation



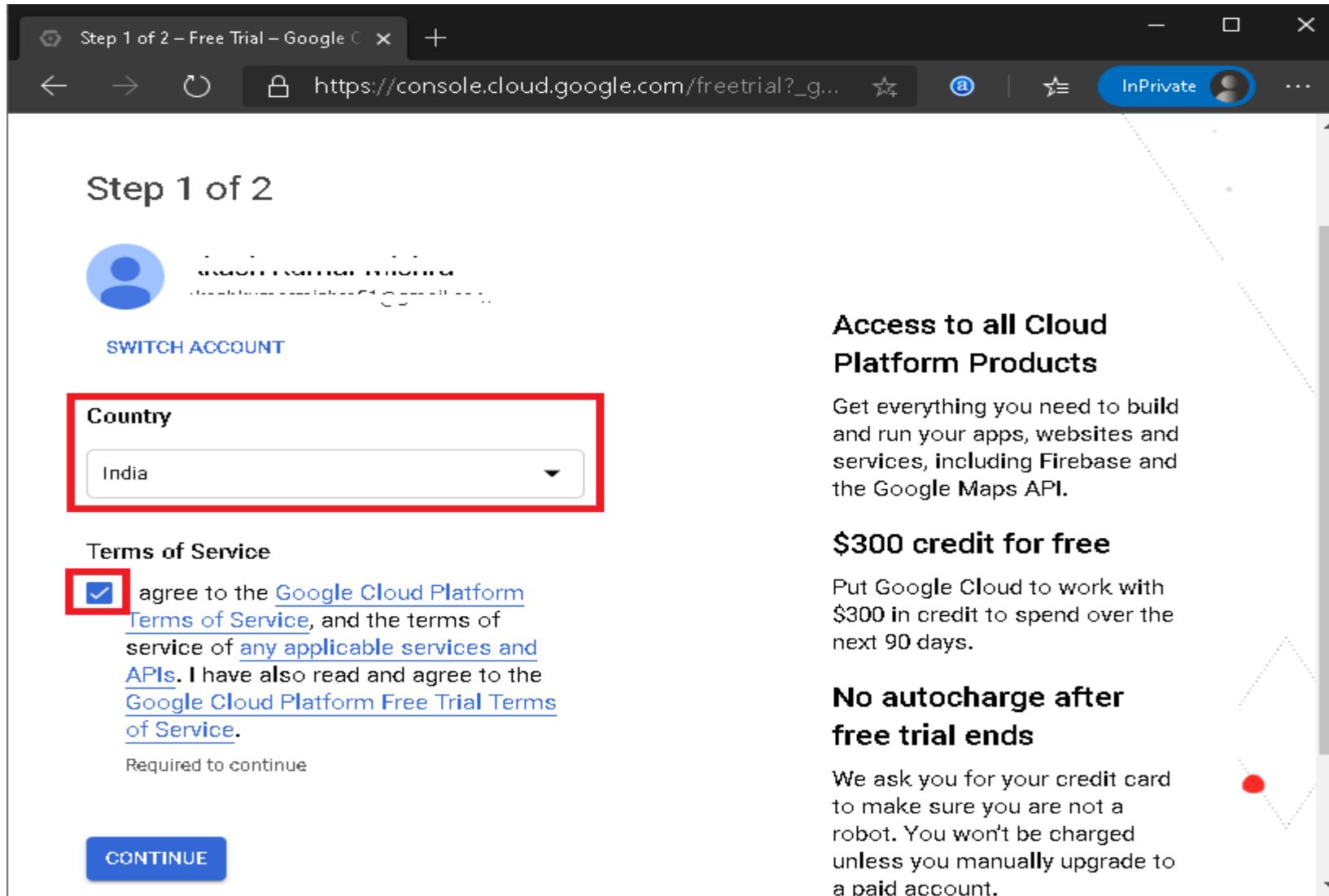
Navigation

- **Step 3:** Next, we are required to login to the Google Account. We can use the 'create an account' button if we don't have an existing Google account.



Navigation

- **Step 4:** Once we have logged in, we will get to the following screen:

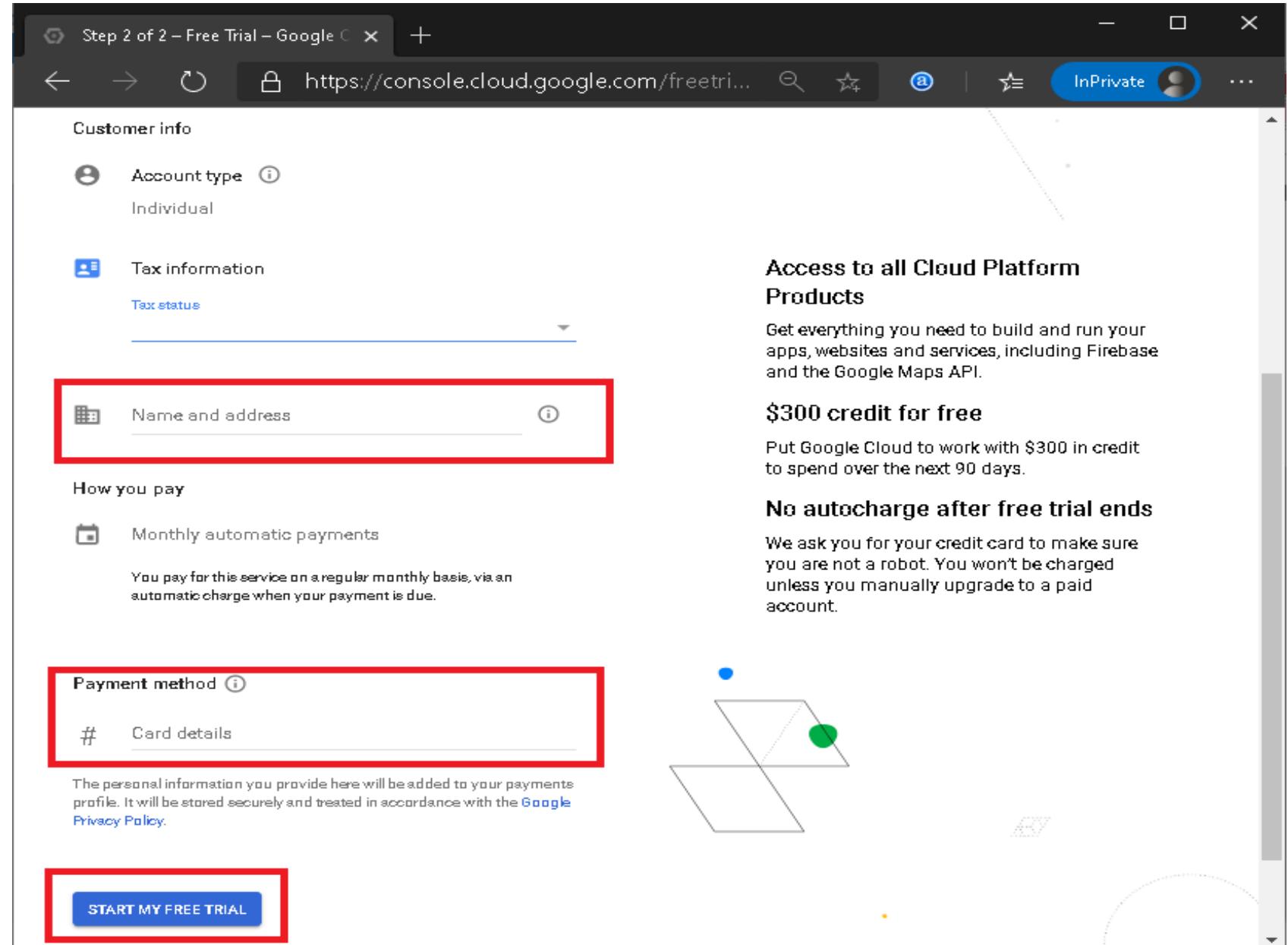


The screenshot shows the 'Step 1 of 2' page for a Google Cloud Platform free trial. At the top, it says 'Step 1 of 2' and shows a user profile with the name 'Anusha Kumar' and a blue profile picture. Below the profile is a 'SWITCH ACCOUNT' link. A red box highlights the 'Country' dropdown menu, which is set to 'India'. The 'Terms of Service' section contains a checked checkbox next to the text: 'I agree to the [Google Cloud Platform Terms of Service](#), and the terms of service of [any applicable services and APIs](#). I have also read and agree to the [Google Cloud Platform Free Trial Terms of Service](#)'. Below this, a small note says 'Required to continue'. At the bottom is a large blue 'CONTINUE' button. To the right of the form, there are three sections: 'Access to all Cloud Platform Products', '\$300 credit for free', and 'No autocharge after free trial ends'. Each section has a brief description and a small circular icon with a red dot.

Here, we must select the Country, agree to the Terms of Service, and then click on the 'CONTINUE' button.

Navigation

- **Step 5:** On the next screen, we have to enter some necessary details such as name and address details. Also, we have to enter payment details like the method of payments and credit card details. After filling all the details, we need to click on the button 'START MY FREE TRIAL' from the bottom of the page:



Step 2 of 2 – Free Trial – Google Cloud Platform

https://console.cloud.google.com/freetrial

Customer info

Account type: Individual

Tax information

Tax status

Name and address

How you pay

Monthly automatic payments

You pay for this service on a regular monthly basis, via an automatic charge when your payment is due.

Payment method

Card details

The personal information you provide here will be added to your payments profile. It will be stored securely and treated in accordance with the [Google Privacy Policy](#).

START MY FREE TRIAL

Access to all Cloud Platform Products

Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

\$300 credit for free

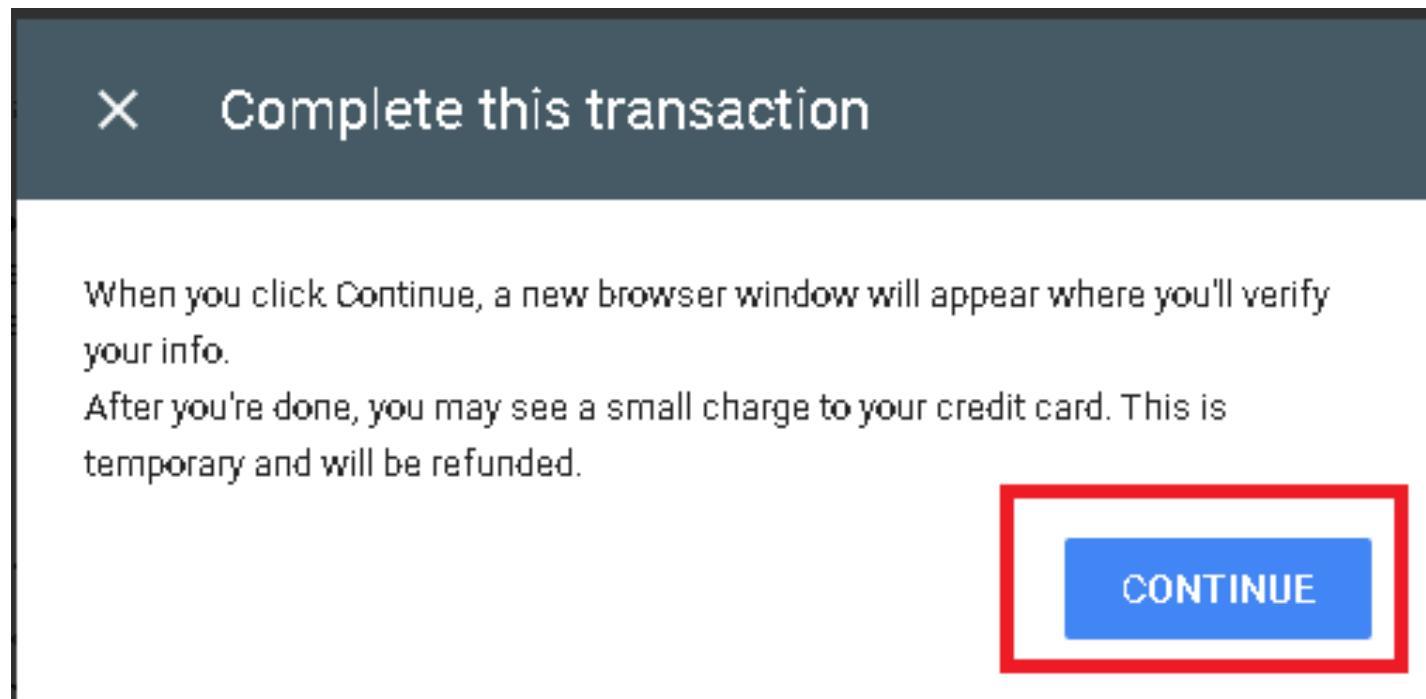
Put Google Cloud to work with \$300 in credit to spend over the next 90 days.

No autocharge after free trial ends

We ask you for your credit card to make sure you are not a robot. You won't be charged unless you manually upgrade to a paid account.

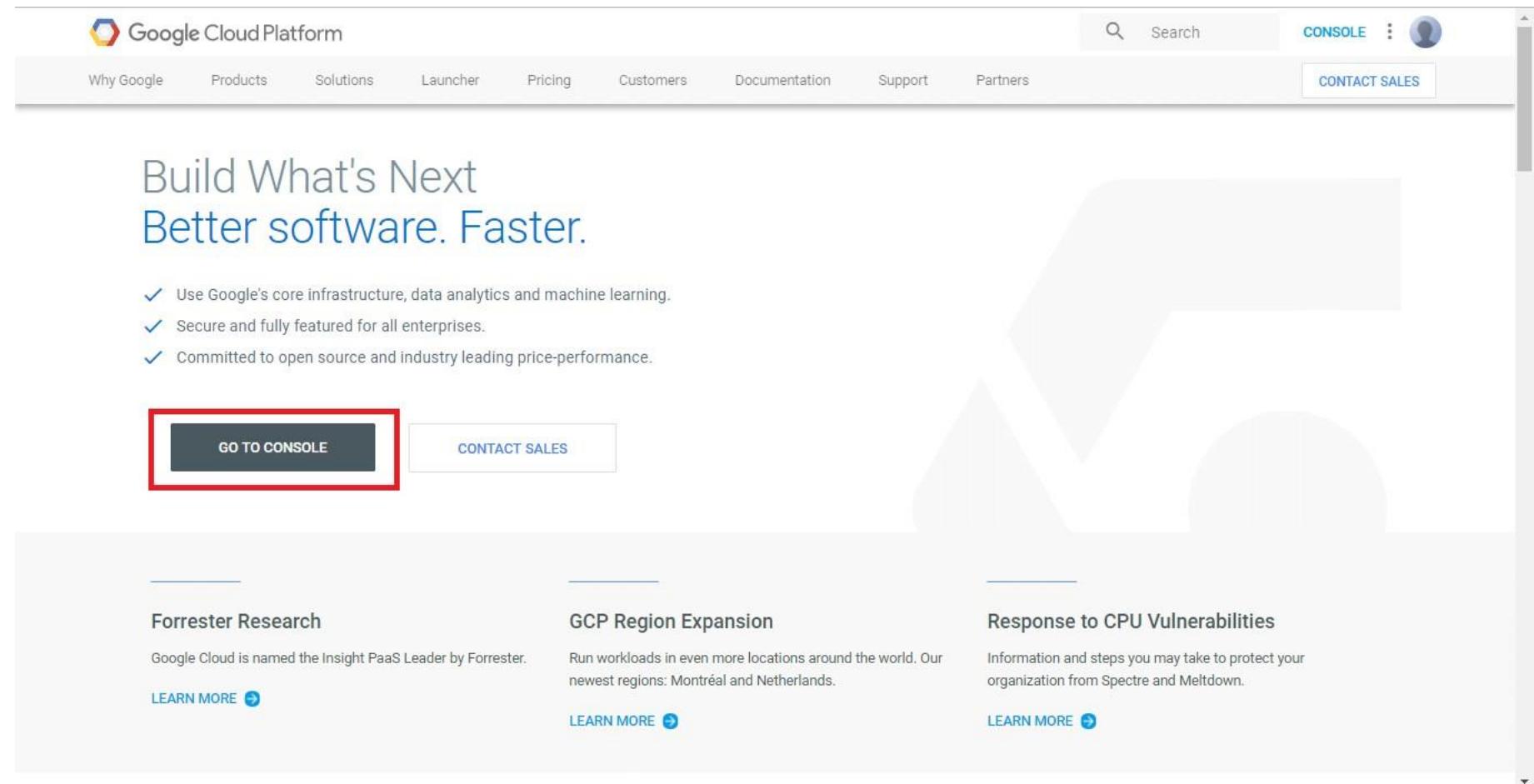
Navigation

- **Step 6:** Google asks for confirmation to use the credit card for the small deduction to ensure that the card information is correct. However, the amount is refunded back to the same account. Here, we need to click on the 'CONTINUE' button:



Navigation

- **Step 7:** On the next screen, we must click on the 'GO TO CONSOLE' button:



The screenshot shows the Google Cloud Platform homepage. At the top, there is a navigation bar with links for Why Google, Products, Solutions, Launcher, Pricing, Customers, Documentation, Support, and Partners. On the right side of the navigation bar are search, console, and user profile icons. Below the navigation bar, the main headline reads "Build What's Next" and "Better software. Faster." followed by a bulleted list: "✓ Use Google's core infrastructure, data analytics and machine learning.", "✓ Secure and fully featured for all enterprises.", and "✓ Committed to open source and industry leading price-performance.". At the bottom of the main content area, there are three buttons: "GO TO CONSOLE" (which is highlighted with a red box), "CONTACT SALES", and "LEARN MORE" for Forrester Research, GCP Region Expansion, and Response to CPU Vulnerabilities.

GO TO CONSOLE

CONTACT SALES

LEARN MORE

Forrester Research

Google Cloud is named the Insight PaaS Leader by Forrester.

LEARN MORE

GCP Region Expansion

Run workloads in even more locations around the world. Our newest regions: Montréal and Netherlands.

LEARN MORE

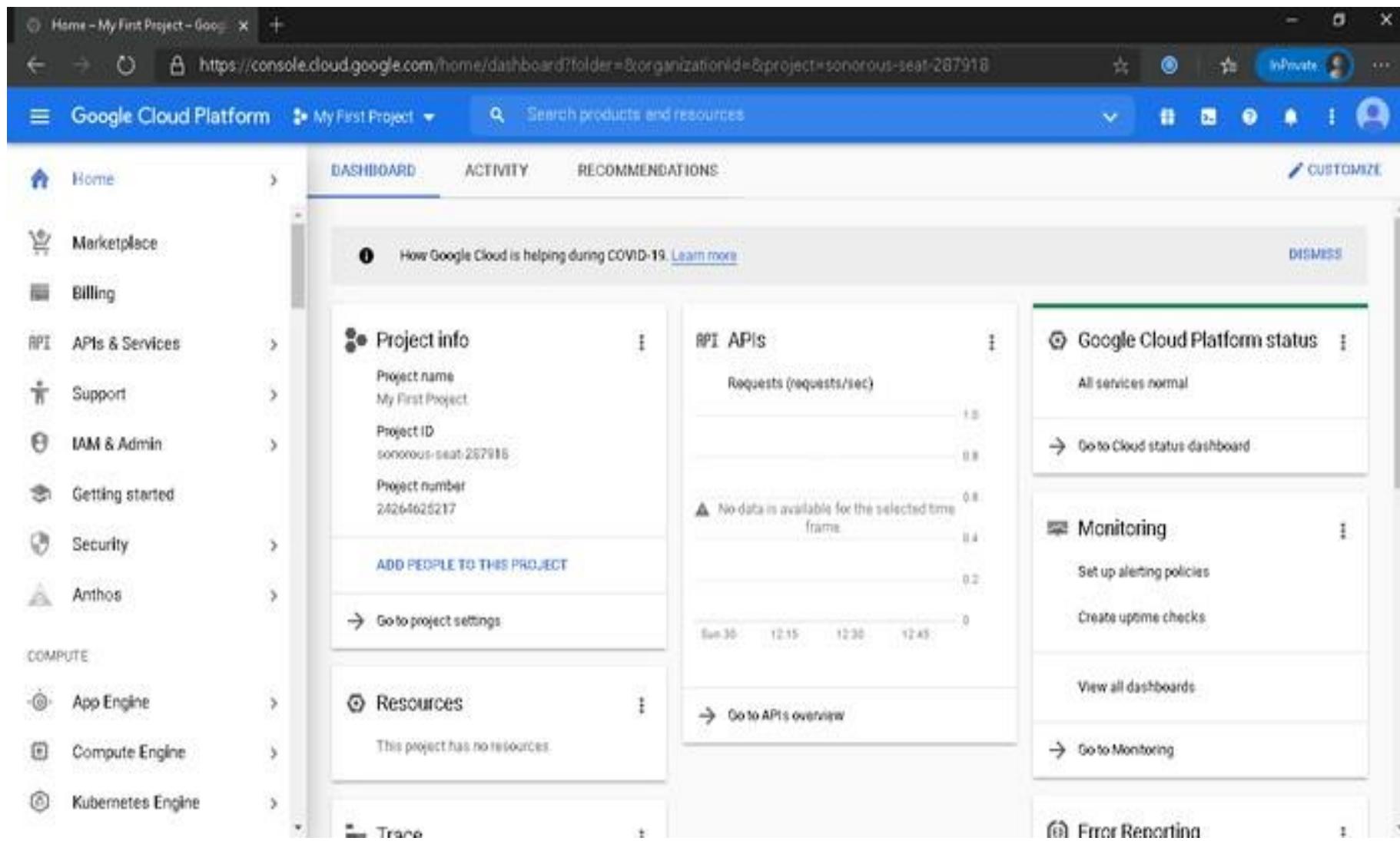
Response to CPU Vulnerabilities

Information and steps you may take to protect your organization from Spectre and Meltdown.

LEARN MORE

Navigation

- After clicking on the 'GO TO CONSOLE' button, we will be redirected to the Dashboard that includes a summary of GCP services along with projects and other insights. It looks like this:



GCP Console

To be specific, a Dashboards of GCP displays the summarized view of the followings:

- **Project Info:** contains project details such as project name, ID, and number.
- **Resources:** contains a list of resources being used in the related project.
- **APIs:** contains various API requests running with the project (in request/sec form).
- **Google Cloud Platform Status:** displays an overall summary of services that are part of GCP.
- **Monitoring:** displays alerts, performance stats, Uptime, etc. to ensure that systems are running reliably.
- **Error Reporting:** displays errors occurring in the projects, but it needs to be configured first.

Cont.

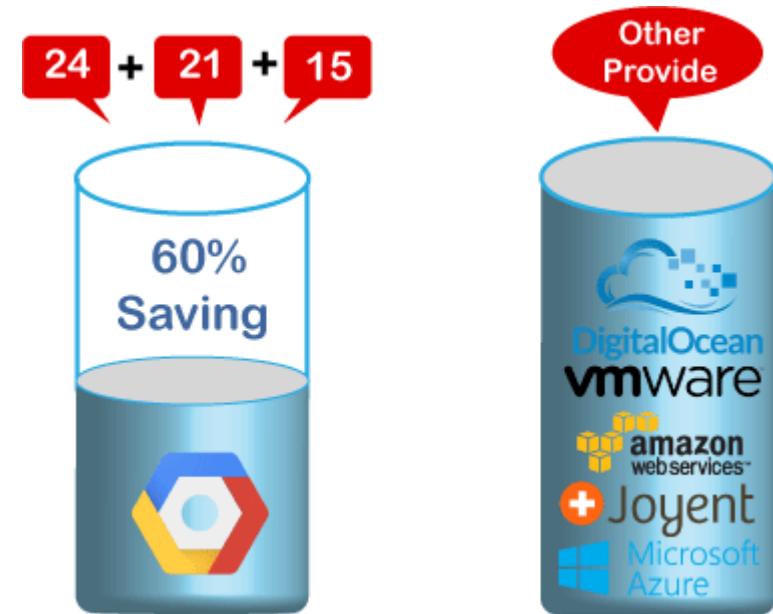
- **Trace:** displays latency data of existing applications across a distributed tracing system.
- **Compute Engine:** displays the insights of CPU usage in percentage (%).
- **Tutorials:** contains Getting Started guides (basic guides) to explain how the GCP features work.
- **News:** displays news and other important updates regarding Google Cloud Platform.
- **Documentation:** contains in-depth guides to teach more about Compute Engine, Cloud Storage, and App Engine.

Google Cloud Platform Pricing

When it comes to pricing, Google Cloud Platform is the cheapest solution in the market. GCP is not only low on price but also offers more features and services than other providers.

When comparing GCP with other leading competitors, it has more benefits over them. Google provides its users a massive 60% savings, including:

- 15% rightsizing recommendation
- 21% list price differences
- 24% of sustained usage discounts



Cont.

Some of the main benefits of GCP pricing are:

- **No Hidden Charges:** There are no hidden charges behind the GCP pricing. Google's pricing structure is straightforward and can be easily understood.
- **Pay-as-you-go:** Google offers its customer 'use now, pay later' option. So, users will have to pay only for those services which they want to use or already using.
- **No Termination Fee:** Users are free to stop using Google services whenever they want, and there will not have to pay any termination fee. That means the moment users stop using Google services; they stop paying for it.

Difference between Google Cloud Platform, AWS and Azure

Google Cloud	AWS	Azure
It uses GCE (Google Compute Engine) for computing purposes.	AWS EC2 offers core compute services.	It uses virtual machines for computation purposes.
It uses Google Cloud Storage for storage purposes.	It uses Amazon S3 for storing the data.	It uses a storage block blob that comprises blocks for storing the data.
It offers the lowest price to the customers to beat other cloud providers.	AWS pricing is generally keen to have inscrutable. The overall structure of granular pricing is a bit complex.	Like AWS, Azure pricing structure is also difficult to understand unless you have considerable experience.
It uses Cloud Test labs for App Testing purposes.	It uses a device farm for App Testing purposes.	It uses DevTest labs for App Testing purposes.
It uses Subnet as a virtual network.	It uses VPC as a virtual network.	It uses VNet as a virtual Network.
It follows the Cloud Load Balancing configuration.	It follows the Elastic Load Balancing configuration.	It follows the Load-Balancer Application Gateway configuration.

Job Opportunities with GCP

There are many job opportunities with GCP. Some popular job roles are listed below:

- Technical Lead Manager
- Sales Engineer
- Technical Solutions Engineer
- Account Executive
- Technical Program Manager
- Cloud Software Engineer
- Data Center Software Engineer
- Solutions Architect
- Strategic Customer Engineer

Source:

<https://medium.com/google-cloud/what-are-the-google-cloud-platform-gcp-services-285f1988957a>

GOOGLE APP ENGINE

And

Microsoft AZURE (PAAS EXAMPLE)

GOOGLE APP ENGINE

- ***Google App Engine*** (often referred to as GAE or simply App Engine) is a web framework and cloud computing platform for developing and hosting web applications in Google-managed data centers.
- Applications are sandboxed and run across multiple servers.
- App Engine offers automatic scaling for web applications as the number of requests increases for an application
- App Engine automatically allocates more resources for the web application to handle the additional demand.

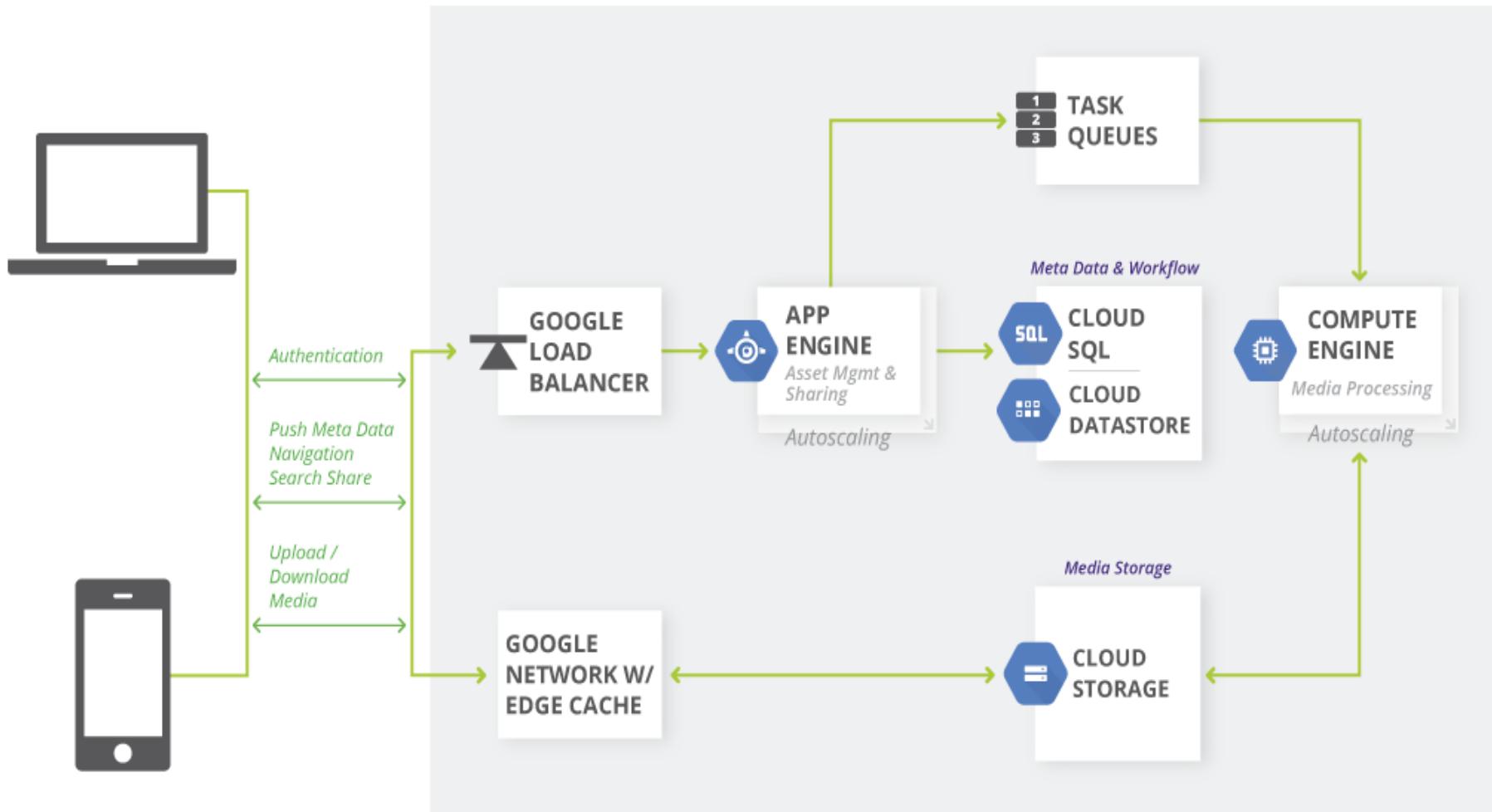
<u>Developer(s)</u>	Google
Initial release	April 7, 2008; 10 years ago
<u>Stable release</u>	1.9.63 / 27 February 2018
Written in	Python , Java , Go , PHP , Node.JS
<u>Operating system</u>	linux (glibc), Windows
<u>Platform</u>	little-endian 32bits
<u>Type</u>	Web framework , cloud computing platform
<u>License</u>	Proprietary , LGPL

Data Store in Google App Engine

Google Cloud Datastore is a **NoSQL** document database built for automatic scaling, high performance, and ease of application development. Cloud Datastore features include:

- Atomic transactions.
- High availability of reads and writes.
- Massive scalability with high performance.
- Encryption at rest.
- Fully managed with no planned downtime.

GAE Architecture



Features

- **Popular languages** Build your application in Node.js, Java, Ruby, C#, Go, Python, or PHP—or bring your own language runtime.
- **Open and flexible** Custom runtimes allow you to bring any library and framework to App Engine by supplying a Docker container.
- **Fully managed** A fully managed environment lets you focus on code while App Engine manages infrastructure concerns.
- **Powerful application diagnostics** Use Cloud Monitoring and Cloud Logging to monitor the health and performance of your app and Cloud Debugger and Error Reporting to diagnose and fix bugs quickly.
- **Application versioning** Easily host different versions of your app, and easily create development, test, staging, and production environments.
- **Application security** Help safeguard your application by defining access rules with App Engine firewall and leverage managed SSL/TLS certificates by default on your custom domain at no additional cost.

Advantages of App Engine over other Cloud Service Providers

- Automatic Load Balancers.: In case of heavy network traffic load balancers distributes the traffic preventing Denial of Service
- Highly Scalable: Automatic scaling and addition of newer instances according to traffic inflow.

Some underlying issues with App Engine

- Expensive.
- Service not available in all countries.
- Slow for CPU intensive processes.
- No base environment control due to API based system.
- Flaws in Security concept.

Differences with other application hosting

- Compared to other scalable hosting services such as Amazon EC2, App Engine provides more infrastructure to make it easy to write scalable applications.
- But can only run a limited range of applications designed for that infrastructure.
- Developers have read-only access to the file system on App Engine.
- Applications can use only virtual file systems.

Features of Google App Engine

Cloud SQL

Comprehensive

Supports delivery testing and development of software

Highly scalable

Cost Saving

Platform independent

Easy to maintain

Backup / Restore

Scheduled Tasks

Remote access

Tasks Queue

Vendor Lock-In and Application Future generations

- A major challenge for the cloud computing industry is the lack of standards and subsequent vendor lock-in.
- Because deploying with one provider means necessarily adopting their specific tools, protocols and operating environments, migrating off their platforms to another provider can prove daunting and expensive in many cases, prohibitively so.
- GAE is no exception to this, as building an application in Google's PaaS involves writing software code customized to the nuances and specifics of the environment.

Contd...

- In short, GAE is an inviting PaaS for developers wishing to build their own cloud software in a robust and scalable environment.
- The platform offers an attractive complementary tier with a significant free usage quota and a competitive pricing model thereafter.
- Software architects using GAE to build and deploy applications can rest assured knowing that their SaaS is powered by the world's largest Internet company.

Microsoft AZURE

Services by Azure



Internet of Things

Power your digital transformation, collect untapped data and find new insights by connecting your devices, assets and sensors



Artificial intelligence

Artificial intelligence productivity for virtually every developer and scenario



Mobile

Reach your customers everywhere, on every device, with a single mobile app build



E-commerce

Give customers what they want with a personalised, scalable and secure shopping experience



Monitoring

Gain visibility into the health, performance and utilisation of your applications, workloads and infrastructure



Business intelligence

Drive better, faster decision making by analysing your data for deeper insights



Big data and analytics

Make the most informed decision possible by analysing all of the data you need in real time



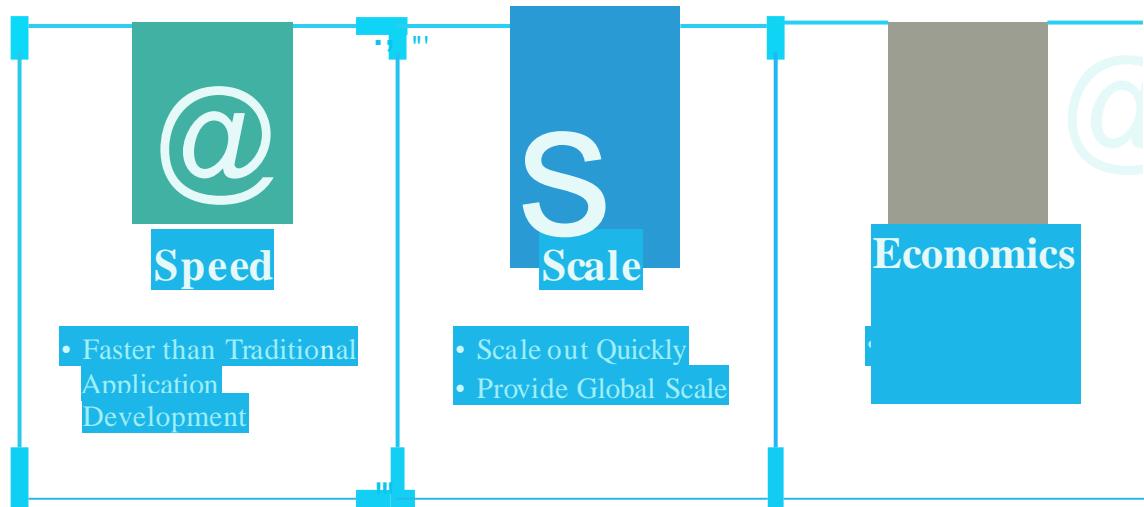
Blockchain

Quickly develop and deploy distributed apps on the blockchain of your choice

Comparison to other providers

- More compliance coverage than any other service provider.
- The only consistent hybrid cloud.
- More region coverage than any other cloud service provider.
- Complete analytics solution with limitless elastic scaling.
- Preconfigured IoT suite for easy to implement IoT solutions.
- Seamless integration with other technologies.
- Provides BaaS (Blockchain as a Service).

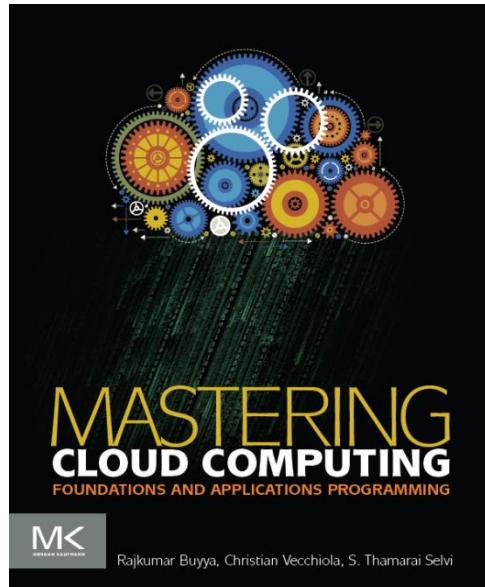
Why Use the Azure?



Business benefits

- **Microsoft Azure is Fast Across the Board**
- **Azure can Match your Global Reach**
- **Azure's has Integrated Development Environment**
- **Azure has a Fully Integrated Delivery Pipeline**
- **Disaster Recovery is Solved with Azure**

Cloud Application Development



Morgan Kauffman, USA



McGraw Hill, India



China Machine Press, China

Unit 5 : Use Cases of Cloud Applications

- Scientific Applications
 - Health care Analysis in the Cloud
 - Biology
- Geo Science
- Business and Consumer Applications
- Productivity
- Social Networking
- Media Applications
- Multiplayer online gaming

Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

Cloud Applications

- Cloud computing has gained huge popularity in industry due to its ability to host applications whose services can be delivered to consumers rapidly at minimal cost.
- This chapter discusses various application case studies detailing their architecture and how they leveraged various Cloud technologies.
- Applications from a range of domains from scientific to engineering, gaming, to social networking are considered.

Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

Scientific Applications

- Scientific applications are a sector that is increasingly using Cloud computing systems and technologies.
- The immediate benefit seen by researchers and academics is the potentially infinite availability of computing resources and storage at sustainable prices if compared to a complete in-house deployment.
- Cloud computing systems meet the needs of different types of applications in the scientific domain: HPC (High Performance Computing) applications, HTC (High Throughput Computing) applications, and data-intensive applications.
- The opportunity for using Cloud resources is even more appealing since minimal changes need to be done to existing applications in order to leverage Cloud resources.

Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- ***Healthcare : ECG Analysis in the Cloud***
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

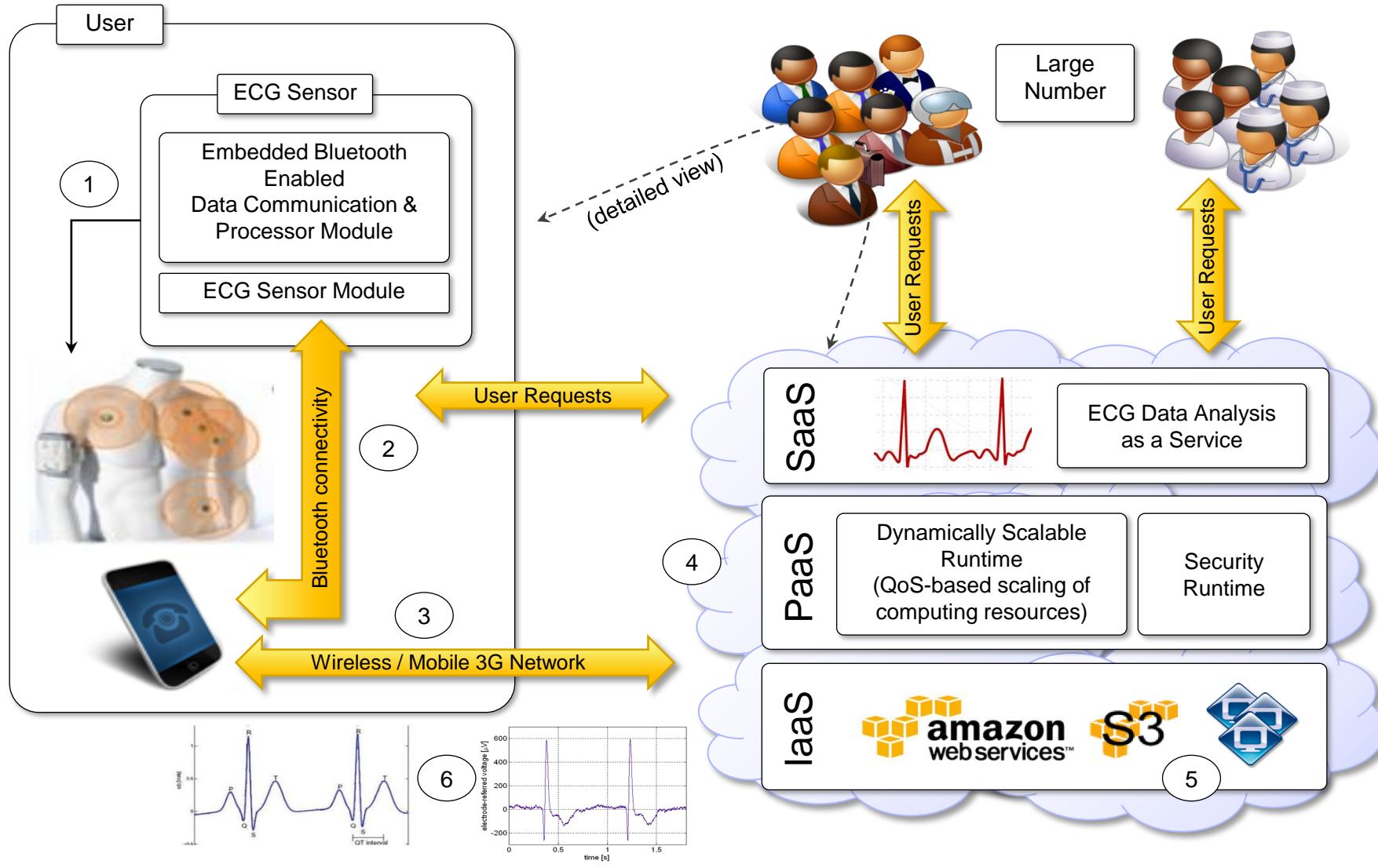
Healthcare : ECG Analysis in the Cloud

- Healthcare is a domain where computer technology has found several and diverse applications: from supporting the business functions to assisting scientists in developing solutions to cure diseases.
- An important application is the use of Cloud technologies for supporting doctors in providing more effective diagnostic processes.
- In particular, we discuss electrocardiogram (ECG) data analysis on the Cloud.

Healthcare : ECG Analysis in the Cloud

- The capillary development of Internet connectivity and its accessibility from any device at any time has made Cloud technologies an attractive option for developing health-monitoring systems.
- Electrocardiogram (ECG) data analysis and monitoring constitutes a case study that naturally fits in this scenario. ECG is the electrical manifestation of the contractile activity of the heart's myocardium.
- This activity produces a specific waveform that is repeated overtime and that represents the heartbeat.
- The analysis of the shape of the waveform is used to identify arrhythmias and it is the most common way for detecting heart diseases.
- Cloud computing technologies allow the remote monitoring of a patient's heartbeat data, its analysis in minimum time, and the notification of first-aid personnel and doctors should this data reveal potentially dangerous conditions.
- This way a patient at risk can be constantly monitored without going to hospital for ECG analysis. At the same time, doctors and first-aid personnel can instantly be notified with cases that require their attention.

Healthcare : ECG Analysis in the Cloud



Healthcare : ECG Analysis in the Cloud

- Even though remote ECG monitoring does not necessarily require Cloud technologies, Cloud computing introduces opportunities that would be otherwise hardly achievable.
- The first advantage is the elasticity of the Cloud infrastructure that can grow and shrink according to the requests served. As a result, doctors and hospitals do not have to invest in large computing infrastructures designed after capacity planning, thus making a more effective use of budgets. The second advantage is ubiquity.
- Cloud computing technologies have now become easily accessible and promise to deliver systems with minimum or no downtime. Computing systems hosted in the Cloud are accessible from any Internet device through simple interfaces (such as SOAP and REST based web services).
- This makes not only these systems ubiquitous but they can also be easily integrated with other systems maintained in the hospital's premises. Lastly, cost savings constitute another reason.
- Cloud services are priced on a pay-per-use basis and with volume prices in case of large numbers of service requests. These two models provide a set of flexible options that can be used to price the service, thus actually charging costs based on effective use rather than capital costs.

Unit 5 : Objectives

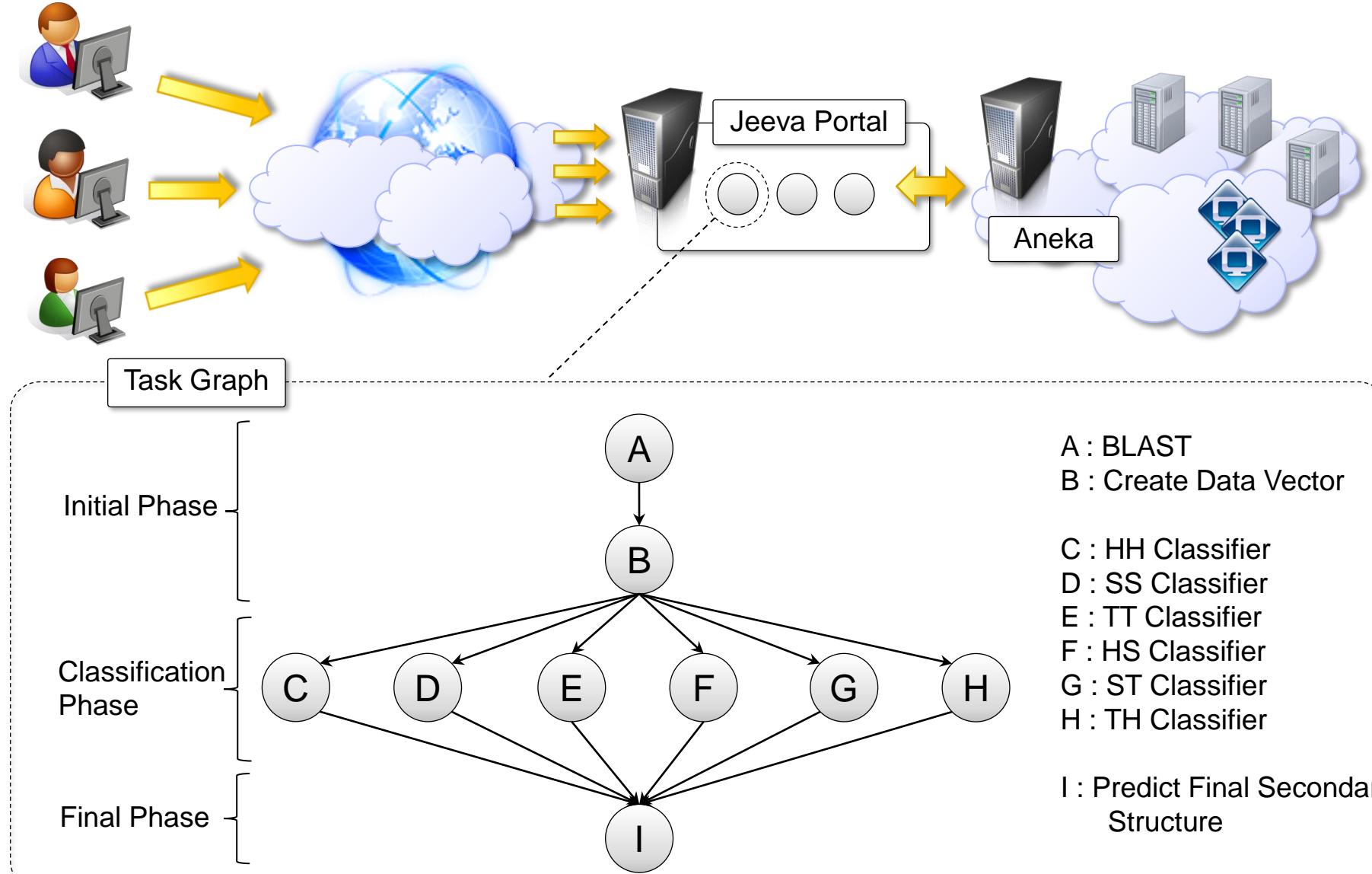
After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- ***Biology : Protein Structure Prediction***
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

Biology : Protein Structure Prediction

- Applications in biology often require high computing capabilities and often operate on large datasets that cause extensive I/O operations.
- Because of these requirements, they have often made extensive use of supercomputing and cluster computing infrastructures. Similar capabilities can be leveraged on-demand by using Cloud computing technologies in a more dynamic fashion thus opening new opportunities for bioinformatics applications.
- Protein structure prediction is a computationally intensive task fundamental for different types of research in the life sciences.
- Among these is the design of new drugs for the treatment of diseases. The geometrical structure of a protein cannot be directly inferred from the sequence of genes that compose its structure, but it is the result of complex computations aimed at identifying the structure that minimizes the required energy.
- This task requires the investigation of a space with a massive number of states, and consequently creating a large number of computations for each of these states.
- The computational power required for protein structure prediction can now be acquired on demand, without owning a cluster or doing all the bureaucracy for getting access to parallel and distributed computing facilities. Cloud computing grants the access to such capacity on a pay-per-use basis.

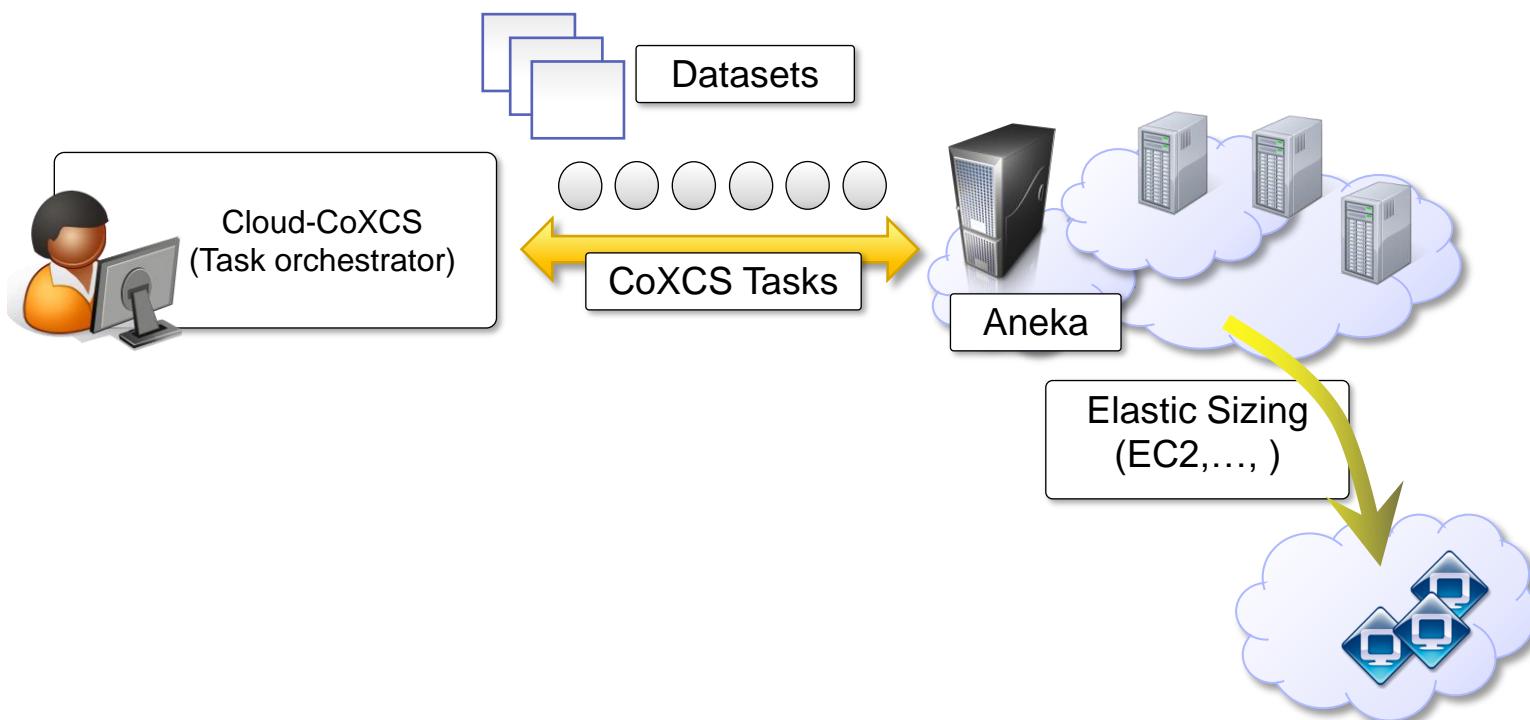
Biology : Protein Structure Prediction



Biology: Gene Expression Data Analysis for Cancer Diagnosis

- Gene expression profiling is the measurement of the expression levels of thousands of genes at once. It is used to understand the biological processes that are triggered by the treatment at a cellular level.
- Together with protein structure prediction, this activity is a fundamental component of drug design since it allows scientists to identify the effects of a specific treatment.
- Another important application of gene expression profiling is cancer diagnosis and treatment. Cancer is a disease characterized by uncontrolled cell growth and proliferation.
- This behavior occurs because of genes regulating the cell growth mutate. This means that all the cancerous cells contain mutated genes.
- In this context, gene expression profiling is utilized to provide a more accurate classification of tumors.
- The classification of gene expression data samples into distinct classes is a challenging task.
- The dimensionality of typical gene expression data sets ranges from several thousands to over ten thousands genes. However, only small sample sizes are typically available for analysis.

Cloud-CoXCS: An Environment for MicroArray Data Processing On the Cloud



Cloud-CoXCS: An Environment for MicroArray Data Processing On the Cloud

- This problem is often approached with learning classifiers, which generate a population of condition-action rule that guide the classification process.
- Among these, the *eXtended Classifier System (XCS)* has been successfully utilized for classifying large datasets in the bioinformatics and computer science domains. However, the effectiveness of XCS when confronted with high dimensional data sets (such as microarray gene expression data sets) has not been explored in detail.
- A variation of such algorithm, CoXCS, has proven to be effective in these conditions. CoXCS divides the entire search space into subdomains and employs the standard XCS algorithm in each of these subdomains.
- Such a process is computationally intensive but can be easily parallelized as the classifications problems on the subdomains can be solved concurrently. Cloud-CoXCS is a Cloud-based implementation of CoXCS that leverages Aneka to solve the classification problem in parallel and compose their outcomes.
- The algorithm is controlled by strategies, which define the way in which the outcomes are composed together and whether the process needs to be iterated.
- Because of the dynamic nature of XCS, the number of required compute resources to execute it can vary over time. Therefore, the use of a scalable middleware such as Aneka offers a distinctive advantage.

Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- ***GeoScience : Satellite Image Processing***
- *Business and Consumer Applications*
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

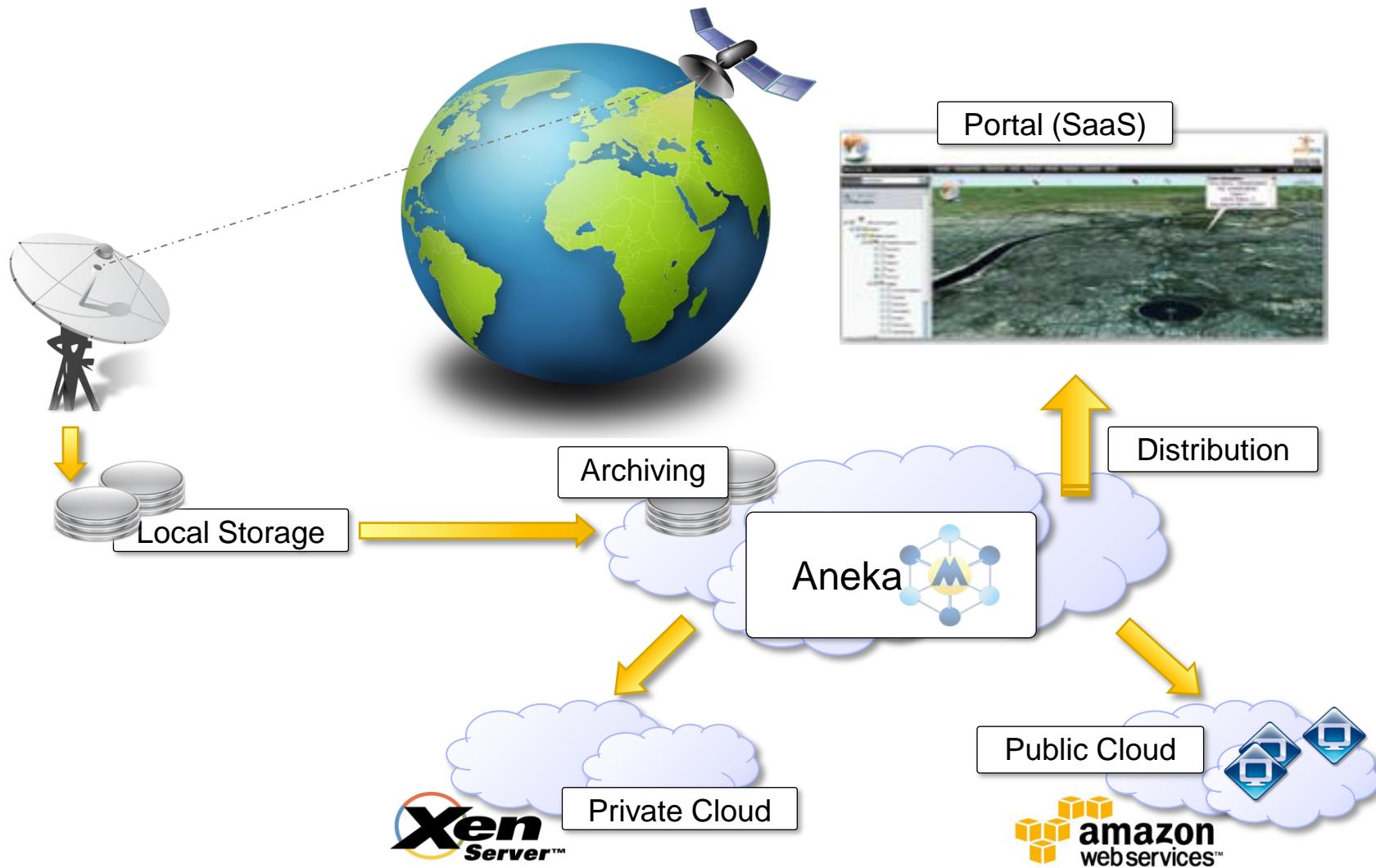
GeoScience : Satellite Image Processing

- Geoscience applications collect, produce, and analyse massive amounts of geospatial and non-spatial data.
- As the technology progresses and our planet becomes more instrumented (i.e., through the deployment of sensors and satellites for monitoring), the volume of data that need to be processed increases significantly. In particular, the geographic information system (GIS) is a major element of geoscience applications. GIS applications capture, store, manipulate, analyze, manage, and present all types of geographically referenced data.
- This type of information is now becoming increasingly relevant to a wide variety of application domains: from advanced farming to civil security and also natural resources management.
- As a result, a considerable amount of geo-referenced data is ingested into computer systems for further processing and analysis. Cloud computing is an attractive option for executing these demanding tasks and extracting meaningful information for supporting decision makers.

GeoScience : Satellite Image Processing

- Satellite remote sensing generates hundreds of gigabytes of raw images that need to be further processed to become the basis of several different GIS products. This process requires both I/O and compute intensive tasks.
- Large size images need to be moved from the ground station's local storage to compute facilities where several transformations and corrections are applied. Cloud computing provides the appropriate infrastructure to support such application scenario. A Cloud-based implementation of such a workflow has been developed by the Department of Space, Government of India.
- The system shown in Figure integrates several technologies across the entire computing stack.
- A SaaS application provides a collection of services for such as geocode generation and data visualization.
- At the PaaS level Aneka controls the import of data into the virtualized infrastructure and the execution of image processing tasks that produce the desired outcome from raw satellite images.
- The platform leverages a Xen private Cloud and the Aneka technology to dynamically provision the required resources (i.e, grow or shrink) on demand.

GeoScience : Satellite Image Processing



Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- ***Business and Consumer Applications***
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

Business and Consumer Applications

- The business and consumer sector is the one that probably benefits the most from Cloud computing technologies.
- On the one hand the opportunity of transforming capital cost into operational costs makes Clouds an attractive option for all enterprises that are IT centric.
- On the other hand, the sense of ubiquity that Cloud offers for accessing data and services makes it interesting for end users as well.
- Moreover, the elastic nature of Cloud technologies does not require huge upfront investments, thus allowing new ideas to be quickly translated into products and services that can comfortably grow with the demand.
- The combination of all these elements has made Cloud computing the preferred technology for a wide range of applications: from CRM and ERP systems to productivity and social networking applications

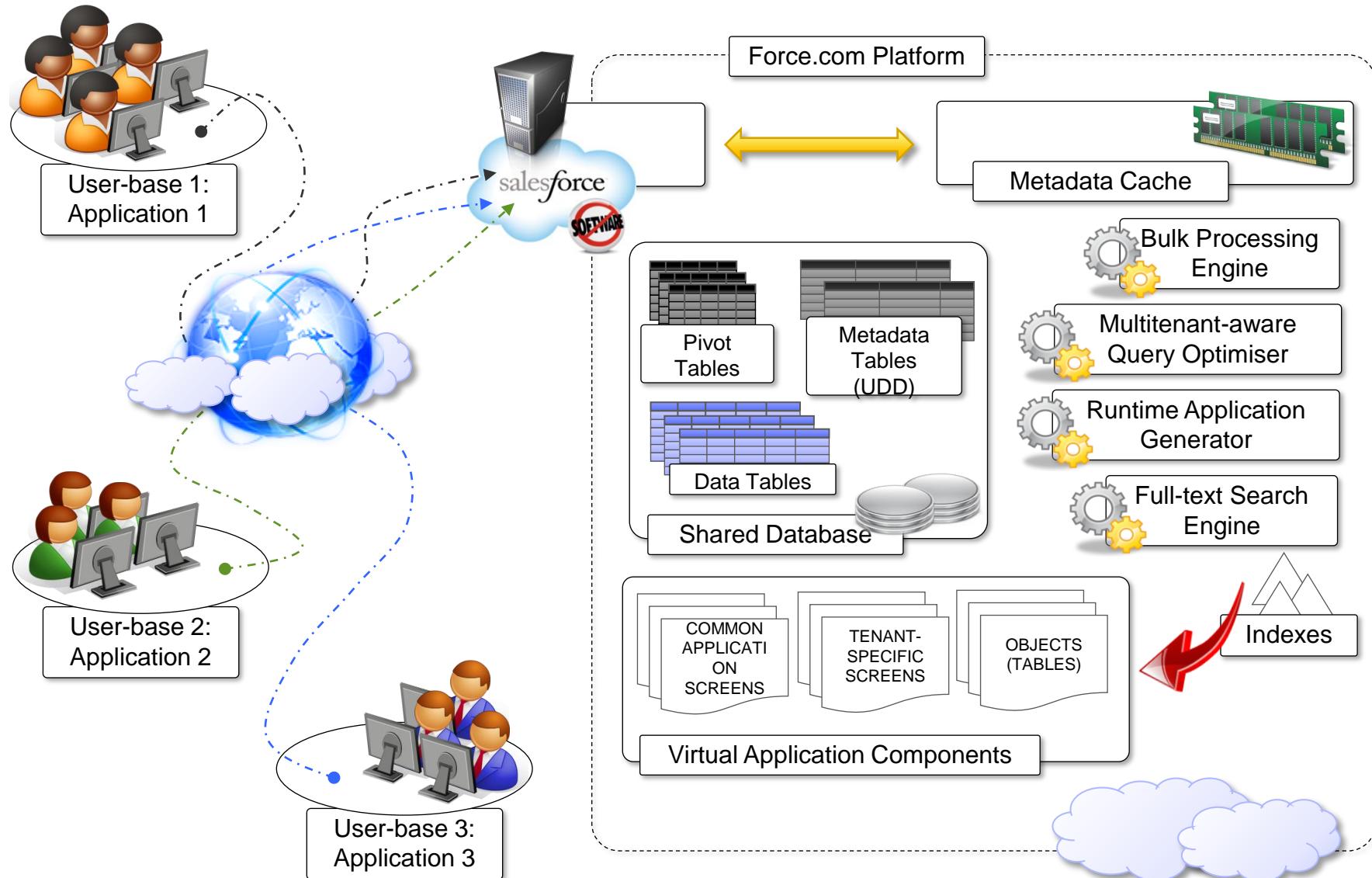
CRM and ERP

- *Customer Relationship Management (CRM)* and *Enterprise Resource Planning (ERP)* applications are market segments that are flourishing in the Cloud, with CRM applications being more mature than ERP implementations.
- Cloud CRM applications constitute a great opportunity for small enterprises and start-ups to have a fully functional CRM software without large upfront costs and by paying subscriptions.
- Moreover, customer relationship management is not an activity that requires specific needs and it can be easily moved to the Cloud. Such a characteristic, together with the possibility of having access to your business and customer data from everywhere and any device, has fostered the spread of Cloud CRM applications. ERP solutions on the Cloud are less mature and have to compete with well-established in-house solutions.
- ERP systems integrate several aspects of an enterprise: finance and accounting, human resources, manufacturing, supply chain management, project management, and customer relationship management.

SalesForce.com

- Salesforce.com is probably the most popular and developed CRM solutions available today. As of today more than 100 thousands customers have chosen Safesforce.com to implement their CRM solutions.
- The application provides customizable CRM solutions that can be integrated with additional features developed by third parties. Salesforce.com is based on the *Force.com* Cloud development platform.
- This represents the scalable and high-performance middleware executing all the operations of all Salesforce.com applications.

SalesForce.com



Microsoft Dynamics CRM

- Microsoft Dynamics CRM is the solution implemented by Microsoft for customer relationship management. Dynamics CRM is available either for installation on the enterprise's premises or as an online solution priced with a monthly per user subscription.
- The system is completely hosted in Microsoft's data center across the world and offers to customers a 99.9% SLA, with bonus credits in case the system does not fulfill the agreement.
- Each CRM instance is deployed on a separate database, and the application provides users with facilities for marketing, sales, and advanced customer relationship management.
- Dynamics CRM Online features can be accessed either through a web browser interface or programmatically by means of SOAP and RESTful web services.
- This allows Dynamics CRM to be easily integrated with both other Microsoft products and line of business applications.
- Dynamics CRM can be extended by developing plug-ins that allow implementing specific behaviors triggered on the occurrence of given events. Dynamics CRM can also leverage the capability of Windows Azure for the development and integration of new features.

NetSuite

- NetSuite provides a collection of applications that help customers manage every aspect of the business enterprise. Its offering is divided in three major products: *NetSuite Global ERP*, *NetSuite Global CRM+*, and *NetSuite Global Ecommerce*. Moreover, an all-in-one solution integrates all the three products together: *NetSuite One World*.
- The services delivered by the company are powered by two large datacenters on the opposite coasts (east and west coasts) of the United States connected by redundant links. This allows NetSuite to guarantee 99.5% of uptime to its customers.
- Besides the pre-packaged solutions, NetSuite also provides and infrastructure and a development environment for implementing customized applications. The *NetSuite Business Operating System (NS-BOS)* is a complete stack of technologies for building Software-as-a-Service business applications that leverage the capabilities of NetSuite products.
- On top of the SaaS infrastructure, the NetSuite Business Suite components offer accounting, ERP, CRM, and e-commerce capabilities. An online development environment, *SuiteFlex*, allows integrating such capabilities into new web applications, which are then packaged for distribution by *SuiteBundler*.
- The entire infrastructure is hosted in the NetSuite datacenters, which provide the warranties about the application uptime and availability.

Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- ***Productivity***
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

Productivity

- Productivity applications replicate in the Cloud some of the most common tasks that we are used to perform on our desktop: from document storage, to office automation, and complete desktop environment hosted in the Cloud.
 - DropBox and iCloud
 - Google Docs
 - Cloud Desktops/ EyeOS and XIOS/3

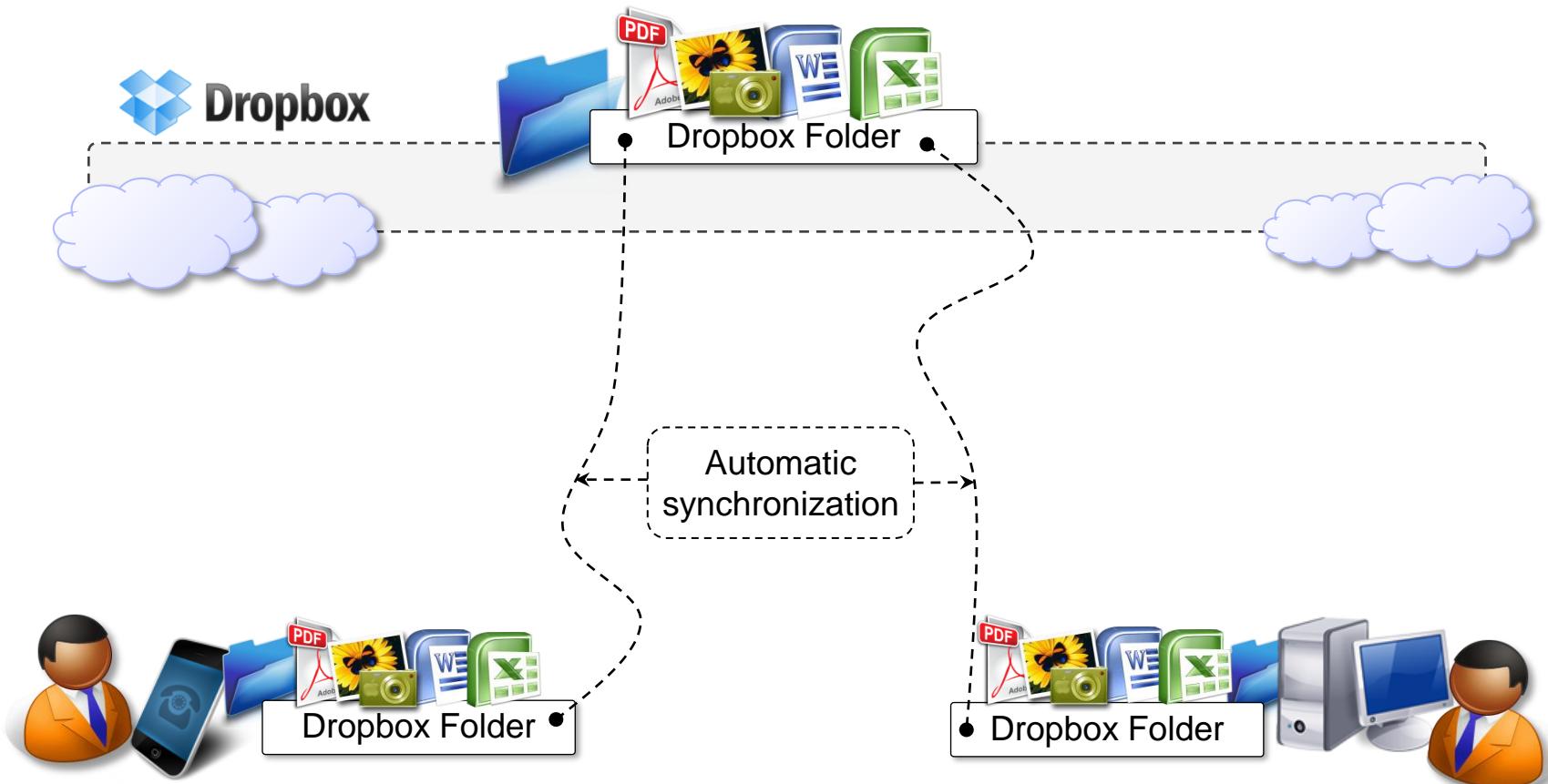
DropBox and iCloud

- One of the core features of Cloud computing is to be available anywhere, at anytime, and from any Internet connected device.
- Therefore, document storage constitutes a natural application for such technology. Online storage solutions are precedent to Cloud computing, but they have never become popular.
- With the development of Cloud technologies they have turned into Software-as-a-Service applications and become more usable as well as more advanced and accessible.
- Perhaps the most popular solution for online document storage is *Dropbox*. This is an online application that allows you to synchronise any file across any platform and any device in a seamless manner.
- Dropbox provides users with a free amount of storage that is accessible through the abstraction of a folder.
- Users can either access their Dropbox folder through a browser or by downloading and installing a Dropbox client, which provides access to the online storage by means of a special folder. All the modifications into this folder are silently synched so that changes are notified to all the local instances of the Dropbox folder across all the devices.
- The key advantage of Dropbox is its availability on different platforms (Windows, Mac, Linux, and mobile) and the capability to work seamlessly and transparently across all of them.

DropBox and iCloud

- Another interesting application in this area is *iCloud*. iCloud is a Cloud-based document sharing application provided by Apple to synchronise IOS-based devices in a completely transparent manner.
- Differently from Dropbox, which provides synchronization through the abstraction of a local folder, iCloud has been designed to be completely transparent once it has been set up: documents, photos, and videos are automatically synched as changes are made without any explicit operation.
- This allows to efficiently automate common operations without any human intervention: taking a picture with an iPhone and having it automatically available in iPhoto on your Mac at home; editing a document in the iMac at home and having the changes updated in the iPad.
- Unfortunately, this capability is limited only to IOS devices and currently there are no plans to provide iCloud with a web-based interface that would make user's content accessible even from unsupported platforms.

DropBox Usage Scenario



Google Docs

- Google Docs is a Software-as-a-Service (SaaS) application that delivers the basic office automation capabilities with support for collaborative editing over the Web. The application is executed on top of Google distributed computing infrastructure that allows the system to dynamically scale according to the number the users using the service.
- Google Docs allows creating and editing text documents, spreadsheets, presentations, forms, and drawings.
- It aims to substitute desktop products such as Microsoft Office and OpenOffice and provide similar interface and functionality as a Cloud service. It supports collaborative editing over the web for most of the applications included in the suite.
- This eliminates tedious mailing and synchronization tasks when documents need to be edited by multiple users. By being stored in the Google infrastructure these documents are always available from anywhere and any device that is connected to the Internet. Moreover, the suite allows users to work off-line in case the Internet connectivity is not available.
- The support of various formats such as those that are produced by the most popular desktop office solutions allows user to easily import and move documents in and out of Google Docs, thus eliminating barriers for the use of this application.
- Google Docs is a good example of what Cloud computing can deliver to end users: ubiquitous access to resources, elasticity, absence of installation and maintenance costs, and delivery of core functionalities as a service.

Unit 5 : Objectives

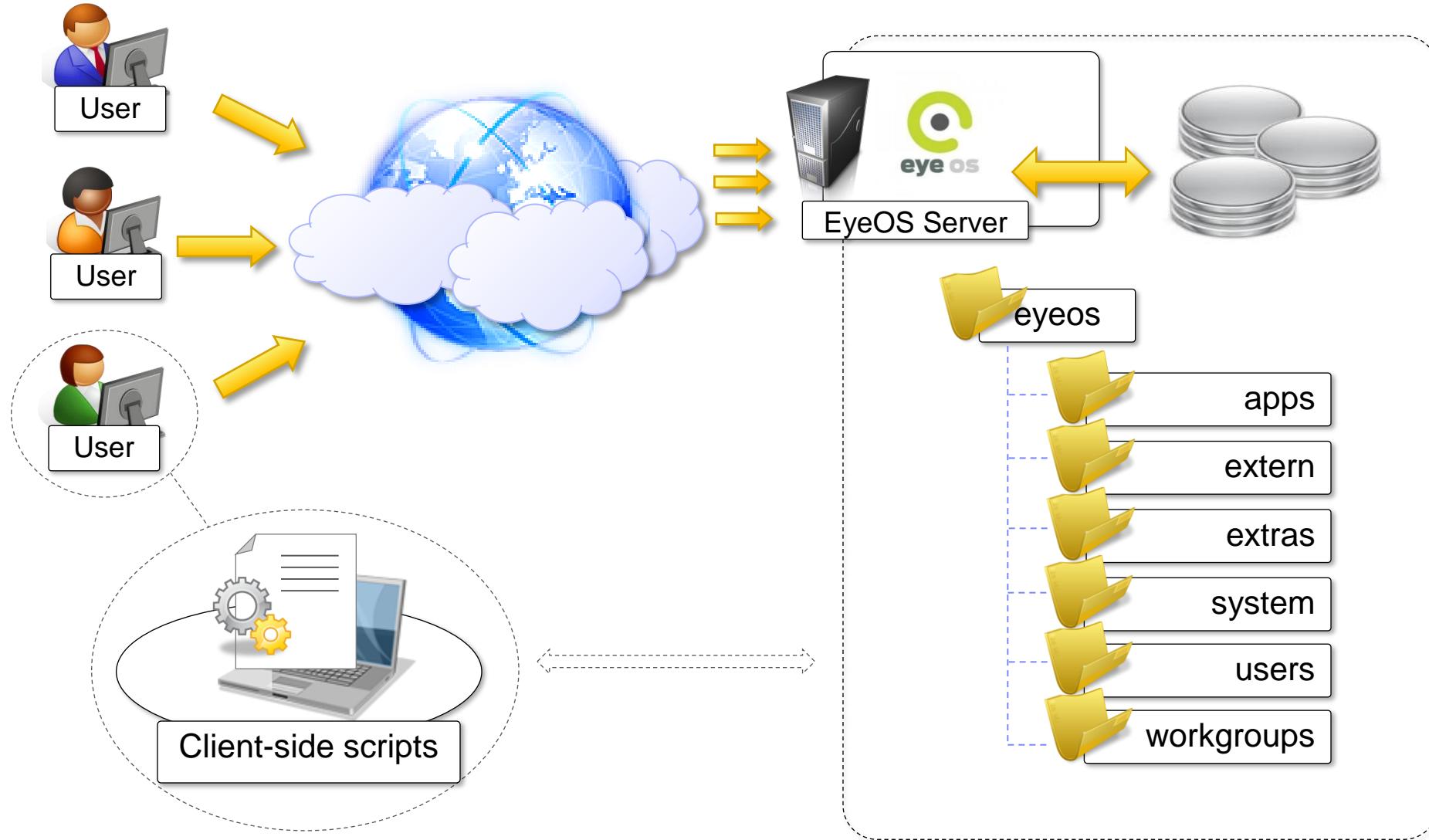
After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- *Productivity*
- ***Cloud Desktops : EyeOS and XIOS/3***
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

Cloud Desktops : EyeOS and XIOS/3

- Asynchronous Javascript and XML (AJAX) technologies have considerably augmented the capabilities that can be implemented in web applications.
- This is a fundamental aspect for Cloud computing that delivers a considerable amount of its services through the web browser.
- Together with the opportunity of leveraging large-scale storage and computation, this technology has made possible the replication of complex desktop environments in the Cloud and made them available through the web browser. These applications are called *Cloud desktops* and rapidly gaining popularity.

EyeOS Architecture



XIOS/3

- *Xcerion XML Internet OS/3 (XIOS/3)* is another example of a web desktop environment.
- The service is delivered as part of the CloudMe application, which is a solution for Cloud document storage.
- The key differentiator of XIOS/3 is its strong leverage on XML, used to implement many of the tasks of the OS: rendering user interfaces; defining application business logics; structuring file system organization; and even application development.
- The architecture of the OS concentrates most of the functionalities on the client side, while implementing server based functionalities by means of XML web services.
- The client side renders the user interface, orchestrates processes, and provides data binding capabilities on XML data that is exchanged with web services.
- The server is responsible for implementing core functions such as transaction management for documents edited in a collaborative mode, and core logic of installed application into the environment.
- XIOS/3 also provides an environment for developing application (XIDE), which allows users to quickly develop complex applications by visual tools for the user interface and XML documents for business logic.

XIOS/3

- XIOS/3 is released as open source software and implements a market place where third parties can easily deploy applications that can be installed on top of the virtual desktop environment.
- It is possible to develop any type of application and feed it with data accessible through XML web services: developers have to define the user interface, bind UI components to service calls and operations, and provide the logic on how to process the data.
- XIDE will package this information into a proper set of XML documents and the rest will be performed by XML virtual machine implemented in XIOS.
- XIOS/3 is an advanced web desktop environment that focuses on the integration of services into the environment by means of XML-based services and simplifies collaboration with peers.

Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- **Social Networking**
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

Social Networking

- Social networking applications have considerably grown in the last years to become the most active sites on the web. In order to sustain their traffic and to serve millions of users seamlessly,
- services like *Twitter* or *Facebook*, have leveraged Cloud computing technologies.
- The possibility of continuously adding capacity while systems are running is the most attractive feature for social networks, which constantly increase their user base.

Facebook

- Facebook is probably the most evident and interesting environment in social networking. It became one of the largest web sites in the world with more than 800 million users. In order to sustain this incredible growth it has been fundamental to be capable of continuously adding capacity, developing new scalable technologies and software systems while keeping a high performance for a smooth user experience.
- Currently, the social network is backed by two data centers that have been built and optimized to reduce costs and impact on the environment. On top of this highly efficient infrastructure built and designed out of inexpensive hardware, a completely customized stack of open source technologies opportunely modified and refined constitutes the backend of largest social network.
- Taken all together, these technologies constitute a powerful platform for developing Cloud applications.
- This platform primarily supports Facebook itself and offers APIs to integrate third party applications with Facebook's core infrastructure to deliver additional services such as social games and quizzes created by others.

Facebook

- The reference stack serving Facebook is based on *LAMP* (*Linux*, *Apache*, *MySQL*, and *PHP*). This collection of technologies is accompanied by a collection of other services developed in-house.
- These services are developed in a variety of languages and implement specific functionalities such as search, new feeds, notifications, and others.
- While serving page requests, the *social graph* of the user is composed. The social graph identifies collection of interlinked information that is of relevance for a given user.
- Most of the user data is served by querying a distributed cluster of MySQL instances, which mostly contain key-value pairs. This data is then cached for faster retrieval.
- The rest of the relevant information is then composed together by using the services mentioned before.
- These services are located closer to the data and developed in languages that provide a better performance than PHP.

Facebook

- The development of services is facilitated by a set of tools internally developed. One of the core elements is *Thrift*.
- This is a collection of abstractions (and language bindings) that allow cross-language development.
- Thrift allows services developed in different languages to communicate and exchange data. Bindings for Thrift in different languages take care of data serialization and deserialization, communication, and client and server boilerplate code.
- This simplifies the work of the developers that can quickly prototype services and leverage existing one. Other relevant services and tools are *Scribe*, which aggregates streaming log feeds, and applications for alerting and monitoring.

Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

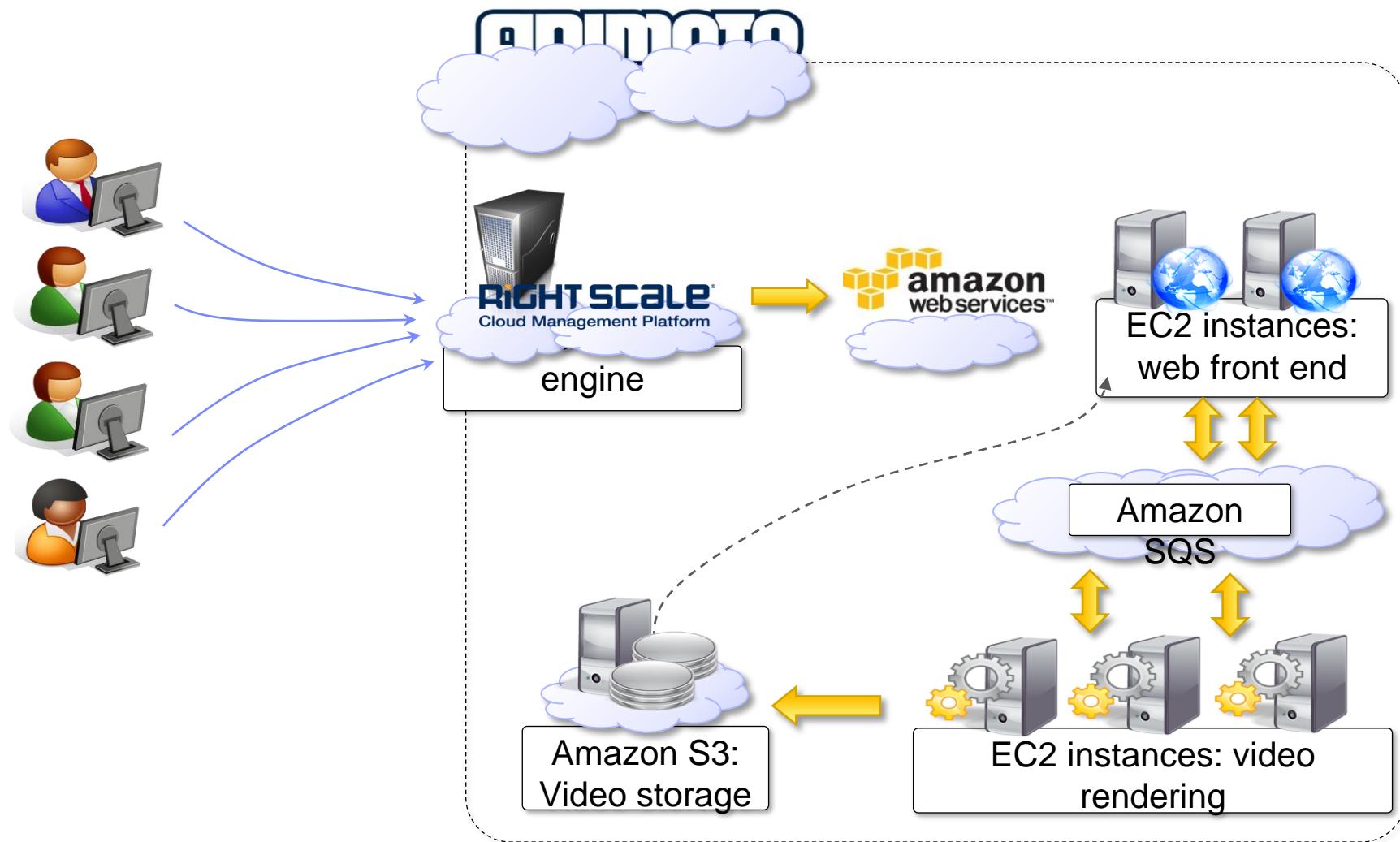
Media Applications

- Media applications are a niche that has taken a considerable advantage from leveraging Cloud computing technologies.
- In particular, video processing operations, such as encoding, transcoding, composition, and rendering, are good candidates for a Cloud-based environment.
- These are computationally intensive tasks that can be easily offloaded to Cloud computing infrastructures.

Animoto

- Animoto is perhaps the most popular example of media applications on the Cloud. The website provides users with a very straightforward interface for quickly creating videos out of images, music, and video fragments submitted by users.
- Users select a specific theme for the video, upload the photos and videos and order them in the sequence they want to appear, select the song for the music, and render the video.
- The process is executed in the background and the user is notified via e-mail once the video is rendered.
- The core value of Animoto is the ability to quickly create videos with stunning effects without the user intervention.
- A proprietary AI engine that selects the animation and transition effects according to pictures and music drives the rendering operation.
- Users only have to define the storyboard by organizing pictures and videos into the desired sequence. If not, the video can be rendered again and the engine will select a different composition, thus producing a different outcome every time. The service allows creating 30 seconds videos for free.
- By paying a monthly or a yearly subscription it is possible to produce videos of any length and to choose among a wider range of templates.

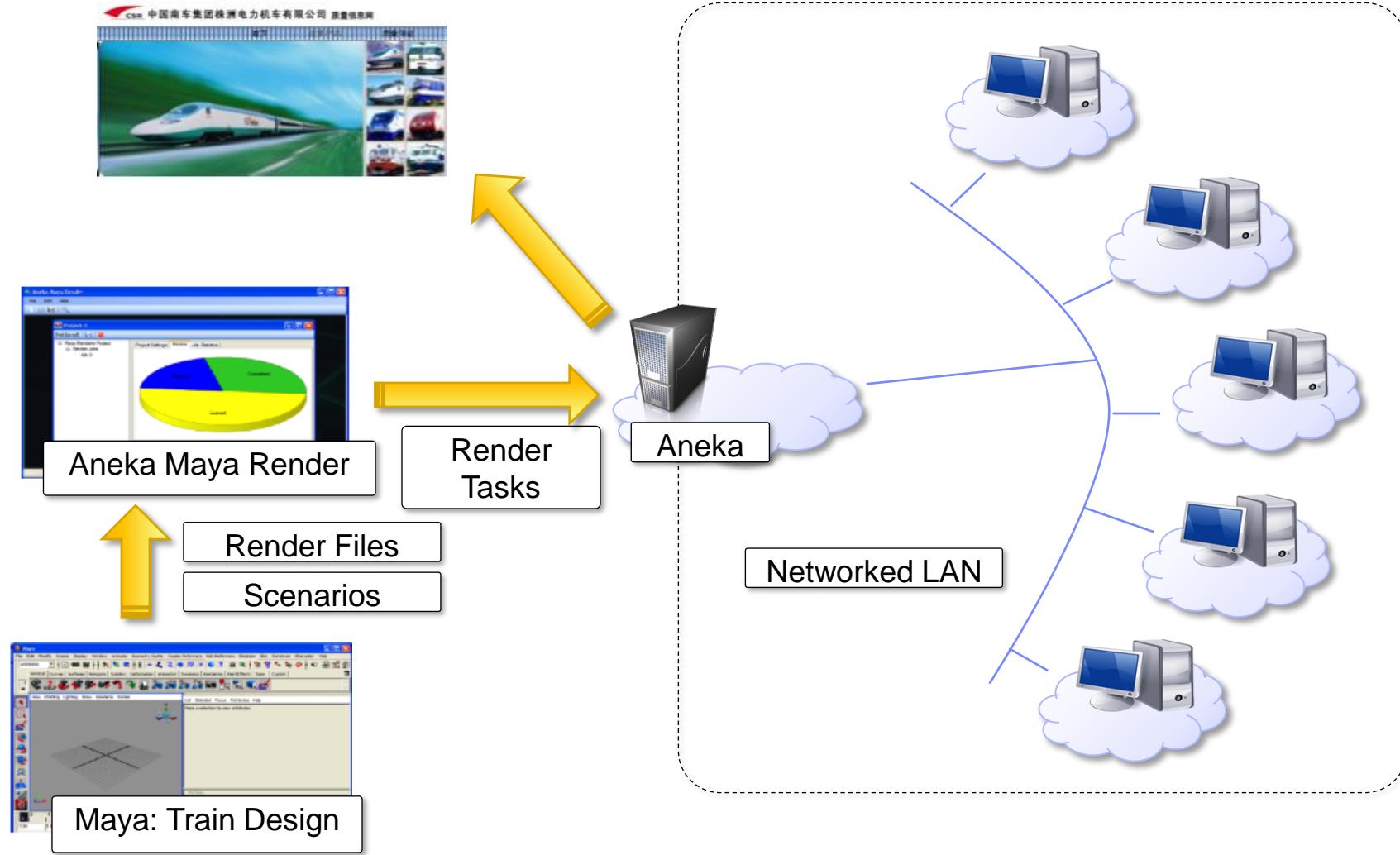
Animoto Reference Architecture



Maya Rendering with Aneka

- Interesting applications of media processing are found in the engineering disciplines and the movie production industry. Operations such as rendering of models are now an integral part of the design workflow, which has become computationally demanding.
- The visualization of mechanical models is not only used at the end of the design process, but it is iteratively used to improve the design.
- It is then fundamental to perform such task as fast as possible. Cloud computing provides engineers with the necessary computing power to make this happen.

3D Rendering On Private Clouds



Video Encoding on the Cloud : Encoding.com

- Video encoding and transcoding are operations that can take a great benefit from using Cloud technologies: they are computationally intensive and potentially require considerable amount of storage.
- Moreover, with the continuous improvement of mobile devices as well as the diffusion of Internet, requests for video content have significantly increased.
- The variety of devices with video playback capabilities has led to an explosion of video formats through which a video can be delivered.
- Software and hardware for video encoding and transcoding often have prohibitive costs or are not flexible enough to support conversion from any format to any format.
- Cloud technologies present an opportunity for turning these tedious and often demanding tasks into services that can be easily integrated into different workflows or made available to everyone according to their needs.

Video Encoding on the Cloud : Encoding.com

- *Encoding.com* is software solution that offers video transcoding services on demand and leverage Cloud technology to provide both the horse-power required for video conversion and the storage for staging videos.
- The service integrates both with Amazon Web Services technologies (*EC2*, *S3*, and *CloudFront*) and Rackspace (*Cloud Servers*, *Cloud Files*, and *Limelight CDN* access).
- Users can access the services through a variety of interfaces: Encoding.com website, web service XML APIs, desktop applications, and watched folders.
- In order to use the service users have to specify the location of the video to transcode, the destination format, and the target location of the video. Encoding.com also offers other video editing operations such the insertion of thumbnails, watermarks, or logos. Moreover, it also extends its capabilities to audio and image conversion.
- The service provides different pricing options: monthly fee, pay-as-you-go (by batches), and special prices for high volumes. Encoding.com has up to now more than 2000 customers and has already processed more than 10 million videos.

Unit 5 : Objectives

After completing this unit you should be able to

- *Cloud Applications*
- *Scientific Applications*
- *Healthcare : ECG Analysis in the Cloud*
- *Biology : Protein Structure Prediction*
- *GeoScience : Satellite Image Processing*
- *Business and Consumer Applications*
- *Productivity*
- *Cloud Desktops : EyeOS and XIOS/3*
- *Social Networking*
- *Media Applications*
- *Multiplayer Online Gaming*
- *Summary*

Multiplayer Online Gaming

- Online multiplayer gaming attracts millions of gamers around the world that share a common experience by playing together on a virtual environment that extends beyond the boundaries of a normal LAN.
- Online games support hundreds of players in the same session and this is made possible by the specific architecture used to forward interactions that is based on game log processing.
- Players update the game server hosting the game session and the server integrates all the updates into a log that is made available to all the players through a TCP port.
- The client software used for the game connects to the log port and by reading the log updates the local user interface with the actions of other players.

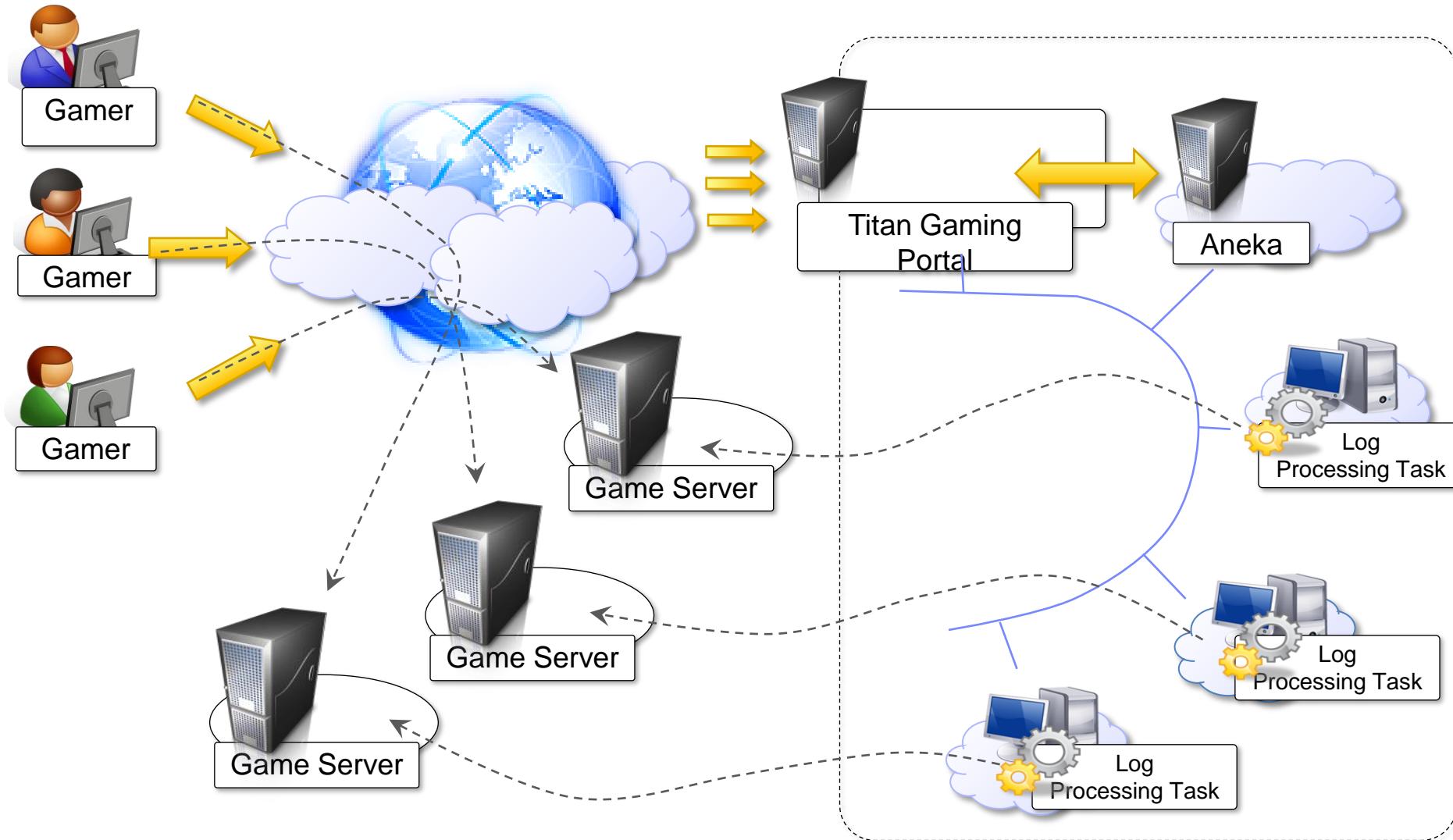
Multiplayer Online Gaming

- Game log processing is also utilized to build statistics on players and rank them. These features constitute the additional value of online gaming portals that attract more and more gamers.
- The processing of game logs is a potentially compute intensive operation that strongly depends on the number of players online and the number of games monitored.
- Moreover, gaming portals are web applications and therefore might suffer from the spiky behavior of users that can randomly generate large amount of volatile workloads that do not justify capacity planning.

Multiplayer Online Gaming

- The use of Cloud computing technologies can provide the required elasticity for seamlessly processing these workloads and scale as required when the number of users increases.
- A prototypal implementation of Cloud-based game log processing has been implemented by Titan Inc. (now Xfire), a company based in California that extended its gaming portal to offload game log processing to the Cloud by using Aneka.
- The prototype has utilized a private Cloud deployment that has allowed Titan Inc. to process concurrently multiple logs and sustain a larger number of users.

Scalable Processing of Logs for Network Games



Summary

- In this chapter we presented a brief overview of applications developed for the Cloud or that leverage Cloud technologies in some form. Different application domains can take advantage from Cloud computing: from scientific application to business and consumer applications.
- Scientific applications take great benefit from the elastic scalability of Cloud environments that also provide the required degree of customization allowing the deployment and execution of scientific experiments.
- Business and consumer applications can leverage several other characteristics. CRM and ERP applications in the Cloud can reduce or even eliminate maintenance costs due to hardware management, system administration, and software upgrades.
- All these new opportunities have transformed the way in which we use these applications on a daily basis, but also introduced new challenges for developers that have to rethink their design to better benefit from elastic scalability, on demand resource provisioning, and ubiquity.
- These are key features of Cloud technology that make it an attractive solution in several domains.

References

- Rajkumar Buyya, Christian Vecchiola, and Thamarai Selvi, **Mastering Cloud Computing**, McGraw Hill, ISBN-13: 978-1-25-902995-0, New Delhi, India, 2013.
- Rajkumar Buyya, Christian Vecchiola, and Thamarai Selvi, **Mastering Cloud Computing**, Morgan Kaufmann, ISBN: 978-0-12-411454-8, Burlington, Massachusetts, USA, May 2013.
— Chapter 10



Who is Amazon !!

- American international multibillion dollar electronic commerce company with headquarters in Seattle, Washington, USA.
 - started in **1995** by **Jeff Bezos** as an **online bookstore**.
 - but **soon diversified**, selling DVDs, VHSs, CDs, video and MP3 downloads/streaming, software, video games, electronics, apparel, furniture, food, toys, and jewelry.
 - The company also **produces consumer electronics**: Kindle e-book reader and the Kindle Fire tablet computer.
 - In **2006**, Amazon officially **launched** the **Amazon Web Services (AWS)** to became a **major provider of cloud computing services**.

What is Amazon Web Services ?

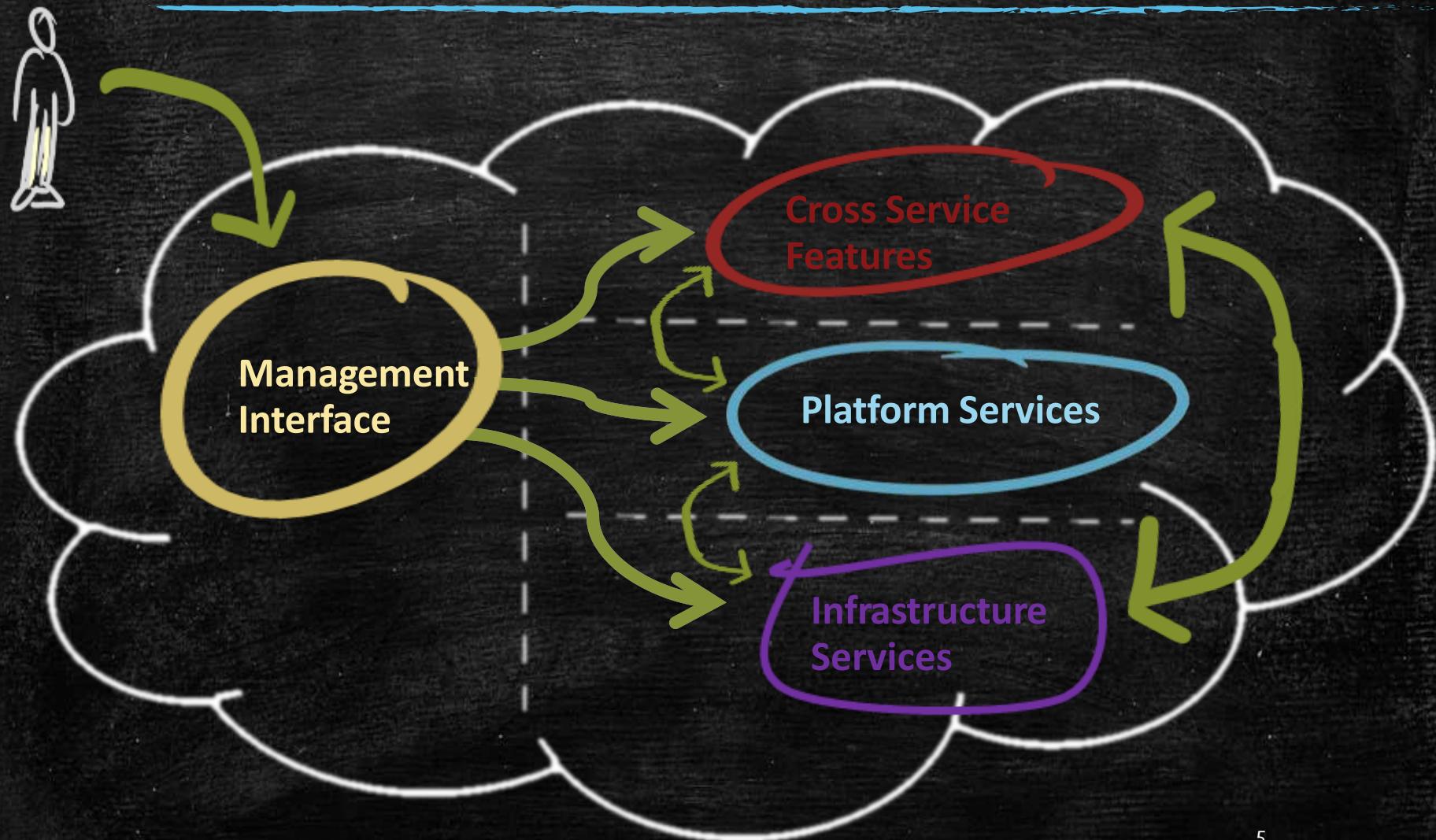
- **Amazon Web Services (AWS)** is a collection of **remote computing services (web services)** that together make up a **cloud computing platform**, offered over the Internet by Amazon.com.
- Website: <http://aws.amazon.com>
- AWS is located in **9 geographical 'Regions'**. Each Region is **wholly contained within a single country** and all of its data and services **stay** within the designated Region.
- Each Region has **multiple 'Availability Zones'**, which are **distinct data centers** providing AWS services. Availability Zones are **isolated from each other** to prevent outages from spreading between Zones. However, Several services **operate across** Availability Zones (e.g. S3, DynamoDB).



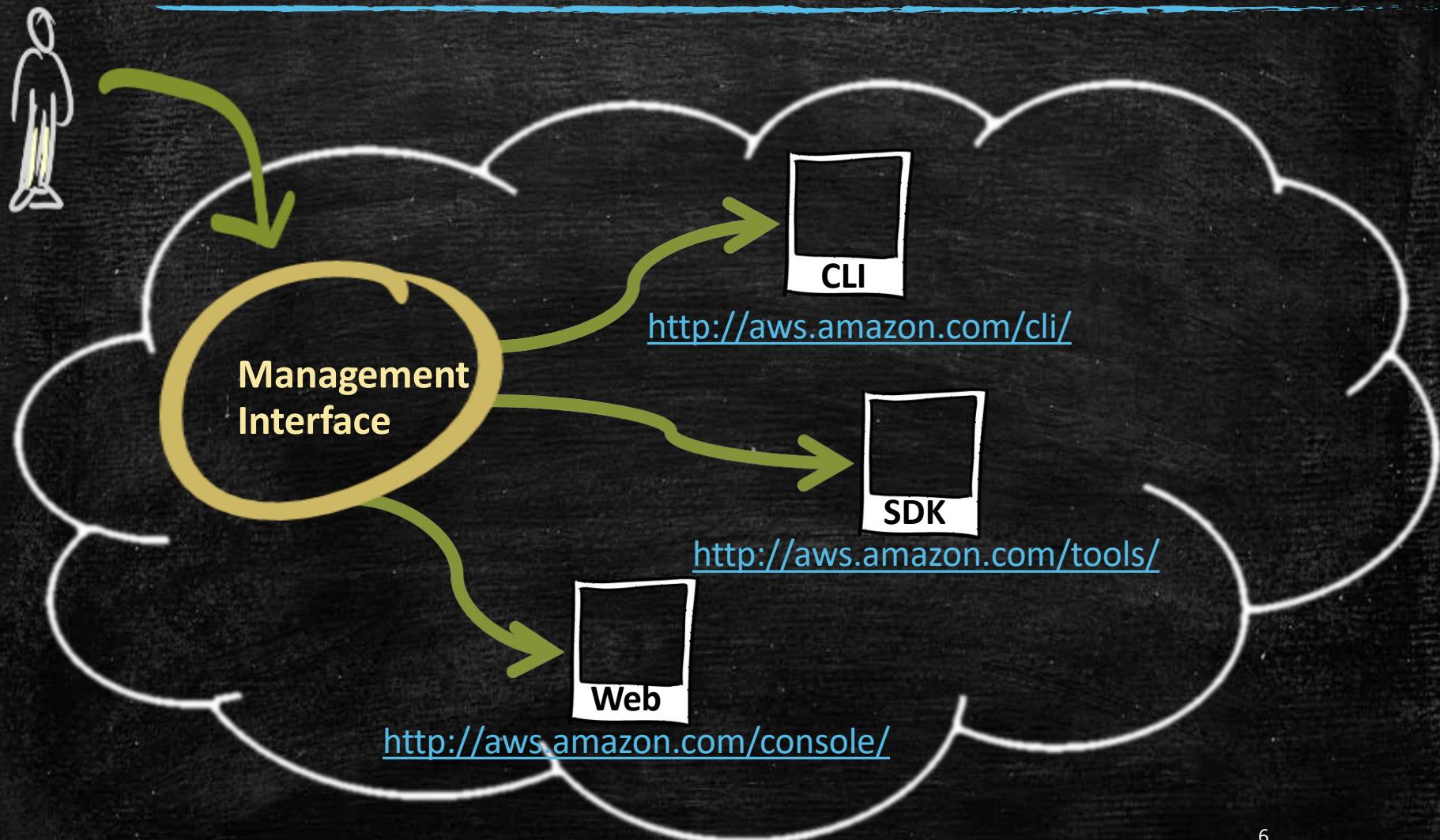
What is AWS Offering?

- **Low Ongoing Cost:**, **pay-as-you-go** pricing with **no up-front expenses** or long-term commitments.
- **Instant Elasticity & Flexible Capacity:** **(scaling up and down)** Eliminate guessing on your infrastructure capacity needs.
- **Speed & Agility:** Develop and deploy applications faster Instead of waiting weeks or months for hardware to arrive and get installed.
- **Apps not Ops:** Focus on projects. Lets you shift resources away from data center investments and operations and move them to innovative new projects.
- **Global Reach:** Take your apps global in minutes.
- **Open and Flexible:** You choose the development platform or programming model that makes the most sense for your business.
- **Secure:** Allows your application to take advantage of the multiple layers of operational and physical security in the AWS data centers to ensure the integrity and safety of your data.

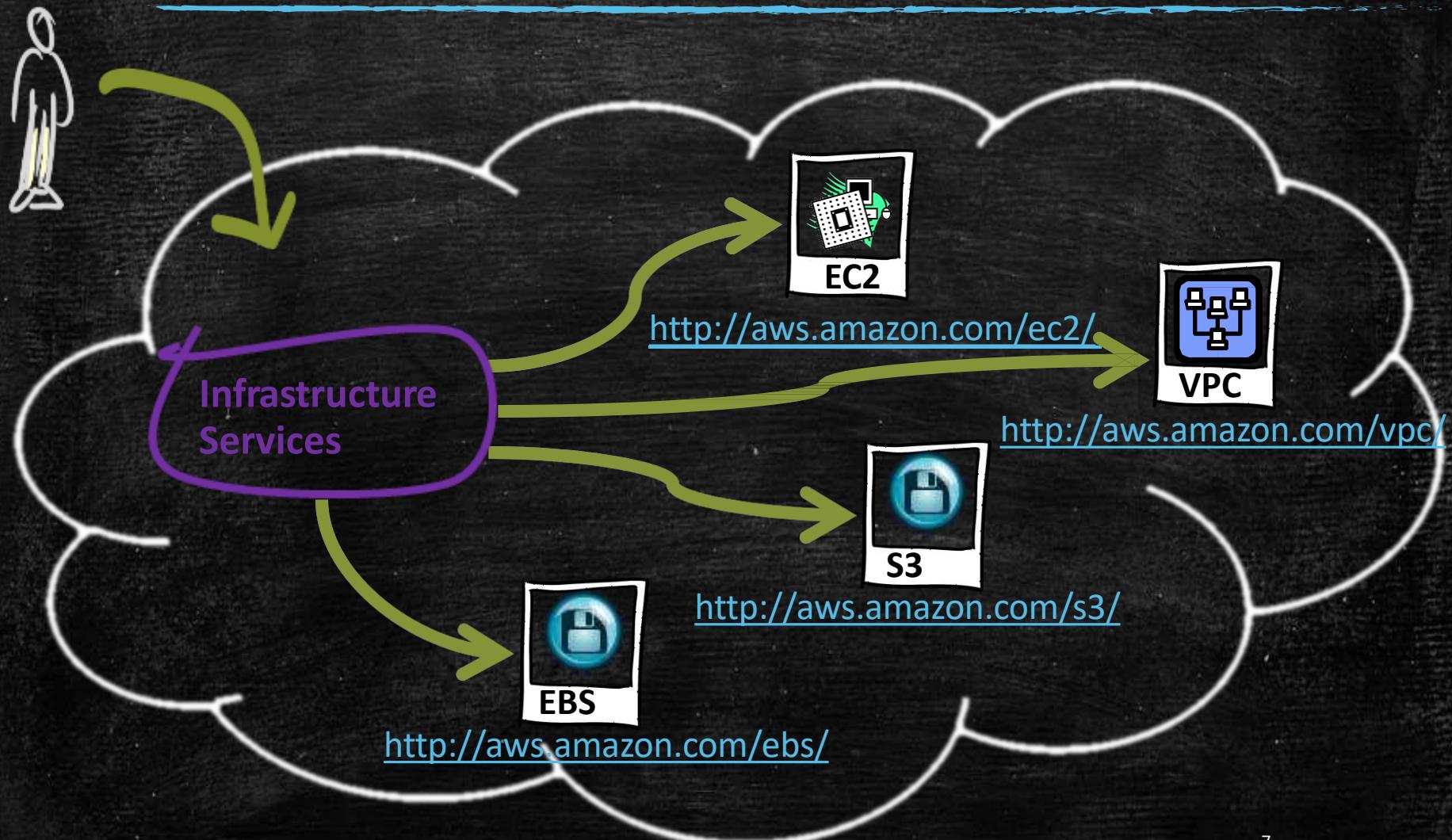
The Amazon Web Services Universe



Management Interface



Infrastructure Services



Amazon Elastic Compute Cloud (EC2)

- A web service that provides **resizable compute capacity** in the cloud.
- EC2 allows **creating Virtual Machines (VM) on-demand**. Pre-configured **templated Amazon Machine Image (AMI)** can be used get running immediately. Creating and sharing your own AMI is also possible via the **AWS Marketplace**.
- **Auto Scaling** allows **automatically scale of the capacity up** seamlessly during **demand spikes** to maintain performance, and **scales down** during **demand lulls** to minimize costs.
- **Elastic Load Balancing** automatically distributes incoming application traffic across multiple Amazon EC2 instances.
- Provide tools to build **failure resilient applications** by launching application instances in **separate Availability Zones**.
- Pay only for resources actually consume, **instance-hours**.
- **VM Import/Export** enables you to easily import virtual machine images from your existing environment to Amazon EC2 instances and export them back at any time.

EC2 Instances

- **Micro instances (t1.micro):**
 - Micro Instance 613 MiB of memory, up to 2 ECUs (for short periodic bursts), EBS storage only, 32-bit or 64-bit platform.
- **Standard Instances** provide customers with a balanced set of resources and a low cost platform.
 - **M1 Small Instance (Default)** 1.7 GiB of memory, 1 EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit), 160 GB of local instance storage, 32-bit or 64-bit platform
 - **M1 Medium Instance** 3.75 GiB of memory, 2 EC2 Compute Units (1 virtual core with 2 EC2 Compute Units each), 410 GB of local instance storage, 32-bit or 64-bit platform
 - **M1 Large Instance** 7.5 GiB of memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), 850 GB of local instance storage, 64-bit platform
 - **M1 Extra Large Instance** 15 GiB of memory, 8 EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each), 1690 GB of local instance storage, 64-bit platform
 - **M3 Extra Large Instance** 15 GiB of memory, 13 EC2 Compute Units (4 virtual cores with 3.25 EC2 Compute Units each), EBS storage only, 64-bit platform
 - **M3 Double Extra Large Instance** 30 GiB of memory, 26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each), EBS storage only, 64-bit platform

One EC2 Compute Unit (ECU) provides the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor.

EC2 High Performance Instances

- **High-Memory Instances:**
 - **High-Memory Extra Large Instance** 17.1 GiB memory, 6.5 ECU (2 virtual cores with 3.25 EC2 Compute Units each), 420 GB of local instance storage, 64-bit platform
 - **High-Memory Double Extra Large Instance** 34.2 GiB of memory, 13 EC2 Compute Units (4 virtual cores with 3.25 EC2 Compute Units each), 850 GB of local instance storage, 64-bit platform
 - **High-Memory Quadruple Extra Large Instance** 68.4 GiB of memory, 26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each), 1690 GB of local instance storage, 64-bit platform
- **High-CPU Instances**
 - **High-CPU Medium Instance** 1.7 GiB of memory, 5 EC2 Compute Units (2 virtual cores with 2.5 EC2 Compute Units each), 350 GB of local instance storage, 32-bit or 64-bit platform
 - **High-CPU Extra Large Instance** 7 GiB of memory, 20 EC2 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each), 1690 GB of local instance storage, 64-bit platform
- **High Storage Instances**
 - **High Storage Eight Extra Large** 117 GiB memory, 35 EC2 Compute Units, **24 * 2 TB of hard disk drive local instance storage**, 64-bit platform, 10 Gigabit Ethernet
- **High I/O Instances**
 - **High I/O Quadruple Extra Large** 60.5 GiB memory, 35 EC2 Compute Units, **2 * 1024₁₀ GB of SSD-based local instance storage**, 64-bit platform, 10 Gigabit Ethernet

EC2 Cluster Instances

- **Cluster Compute Instances** provide proportionally high CPU resources with increased network performance and are **well suited for High Performance Compute (HPC)** applications and other demanding network-bound applications.
 - **Cluster Compute Eight Extra Large** 60.5 GiB memory, **88 EC2 Compute Units**, 3370 GB of local instance storage, 64-bit platform, 10 Gigabit Ethernet
- **High Memory Cluster Instances** provide proportionally high CPU and memory resources with increased network performance, and are **well suited for memory-intensive applications including in-memory analytics, graph analysis, and scientific computing**.
 - **High Memory Cluster Eight Extra Large** **244 GiB memory**, 88 EC2 Compute Units, 240 GB of local instance storage, 64-bit platform, 10 Gigabit Ethernet
- **Cluster GPU Instances** provide **general-purpose graphics processing units (GPUs)** with proportionally high CPU and increased network performance for **applications benefitting from highly parallelized processing, including HPC, rendering and media processing applications**.
 - **Cluster GPU Quadruple Extra Large** 22 GiB memory, 33.5 EC2 Compute Units, **2 x NVIDIA Tesla "Fermi" M2050 GPUs**, 1690 GB of local instance storage, 64-bit platform, 10 Gigabit Ethernet.

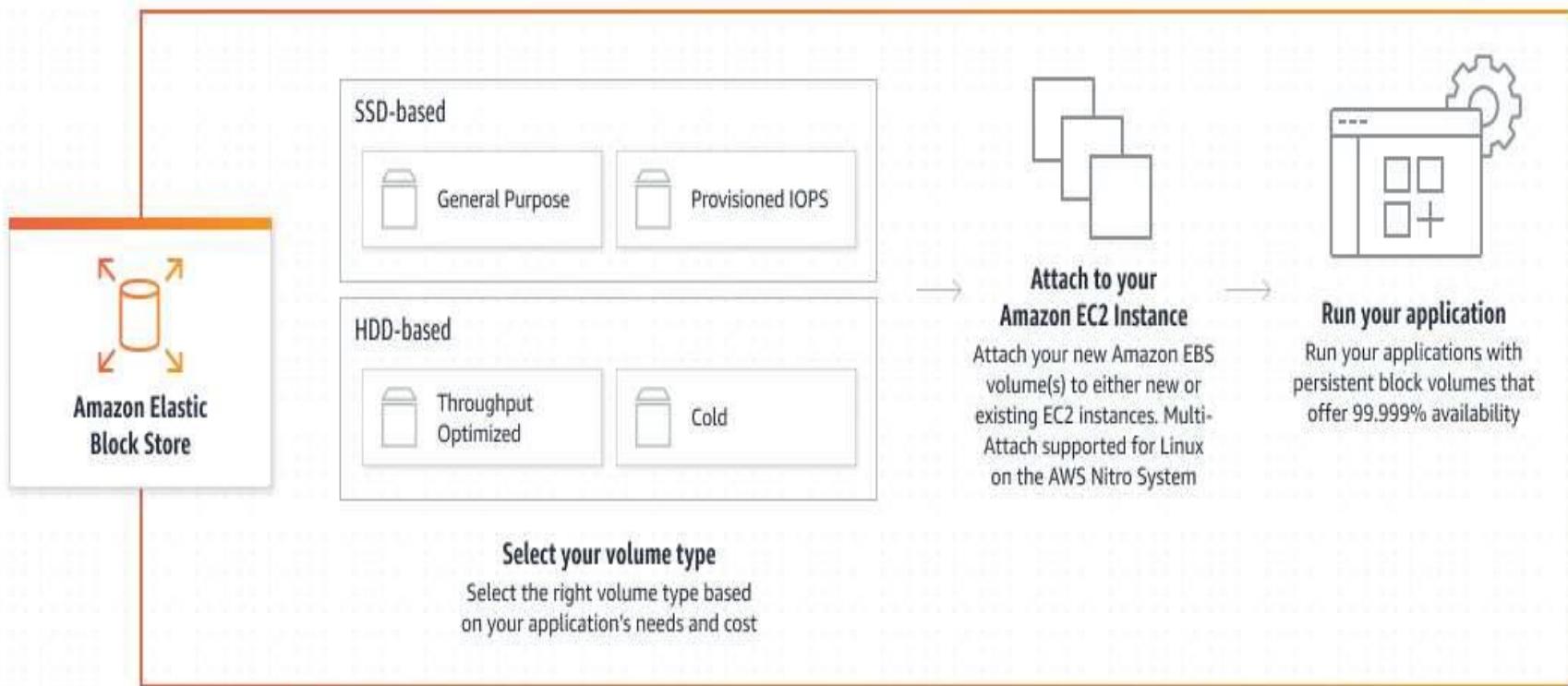
EC2 Payment methods

- **On-Demand Instances** let you **pay for compute capacity by the hour** with **no long-term commitments**.
- **Reserved Instances** give you the option to **make a low, one-time payment for each instance** you want to reserve and in turn **receive a significant discount on the hourly charge** for that instance.
- **Spot Instances** allow customers to **bid on unused Amazon EC2 capacity** and **run those instances for as long as their bid exceeds the current Spot Price**.

Amazon Elastic Block Store (EBS)

- Provides **block level storage** volumes (**1 GB to 1 TB**) for **use with Amazon EC2 instances**.
 - **Multiple volumes** can be mounted to the **same instance**.
 - EBS volumes are **network-attached**, and **persist independently** from the life of an instance.
 - Storage volumes behave like **raw, unformatted block devices**, allowing users to **create a file system** on top of Amazon EBS volumes, or use them in any other way you would use a block device (**like a hard drive**).
- EBS volumes are **placed in a specific Availability Zone**, and can then be **attached to instances also in that same Availability Zone**.
- Each storage volume is **automatically replicated within the same Availability Zone**.
- EBS provides the ability to **create point-in-time snapshots of volumes**, which are **persisted to Amazon S3**.
 - These snapshots can be **used as the starting point for new Amazon EBS volumes**, and protect data for long-term durability.
 - The **same snapshot can be used to instantiate as many volumes** as you wish.
 - These snapshots **can be copied across AWS regions**.

AWS: EBS



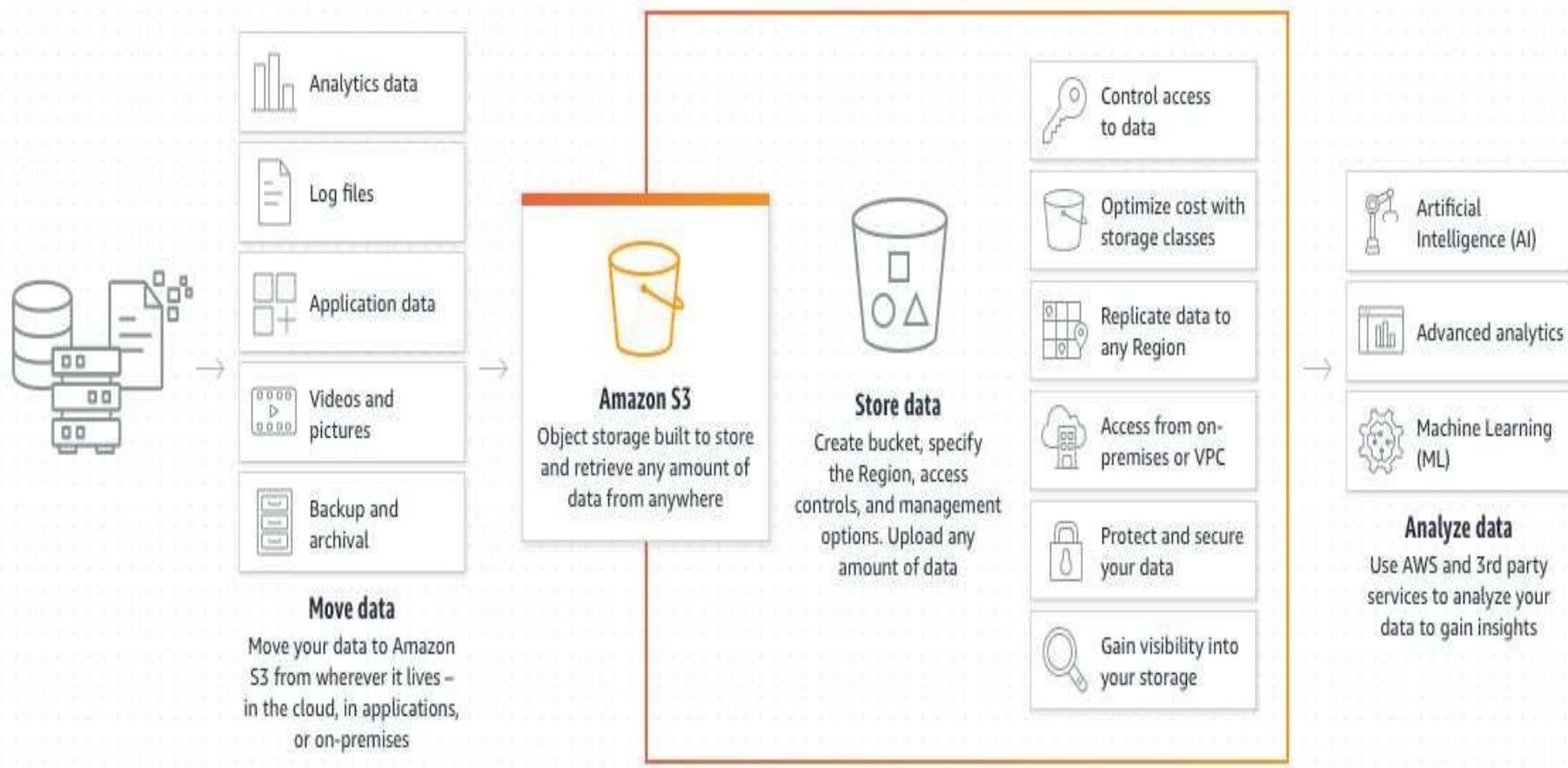
EBS Volumes

- **Standard volumes** offer storage for applications with **moderate or burst I/O** requirements.
 - Standard volumes deliver approximately **100 IOPS** on average.
 - well suited for use as **boot volumes**, where the burst capability provides fast instance start-up times.
- **Provisioned IOPS volumes** are designed to deliver **predictable, high performance for I/O intensive workloads such as databases**.
 - You specify an IOPS rate when creating a volume, and EBS provisions that rate for the lifetime of the volume.
 - Amazon EBS currently supports **up to 4000 IOPS per Provisioned IOPS volume**.
 - You can **stripe multiple volumes together to deliver thousands of IOPS per EC2 instance**.
- To enable your EC2 instances to fully utilize the IOPS provisioned on an EBS volume,:
 - Launch selected Amazon EC2 instance types as **“EBS-Optimized”** instances.
 - **EBS-optimized instances deliver dedicated throughput between Amazon EC2 and Amazon EBS**, with options between **500 Mbps and 1000 Mbps** depending on the instance type used.
- EBS charges based on **per GB-month** AND **per 1 million I/O requests**

Amazon Simple Storage Service (S3)

- Amazon S3 provides a simple web services interface that can be **used to store and retrieve any amount of data, at any time, from anywhere on the web**.
- Write, read, and delete objects containing from **1 byte to 5 terabytes of data each**. The **number of objects** you can store is **unlimited**.
- Each object is stored in a **bucket** and retrieved via a unique, developer-assigned **key**.
 - A bucket can be stored in **one of several Regions**.
 - You can **choose a Region to optimize for latency, minimize costs, or address regulatory requirements**.
 - Objects stored in a Region **never leave the Region** unless you transfer them out.
- **Authentication mechanisms** are provided to ensure that data is kept secure from unauthorized access.
 - Objects can be made **private or public, and rights can be granted to specific users**.
- S3 charges based on **per GB-month** AND **per I/O requests** AND **per data modification requests**.

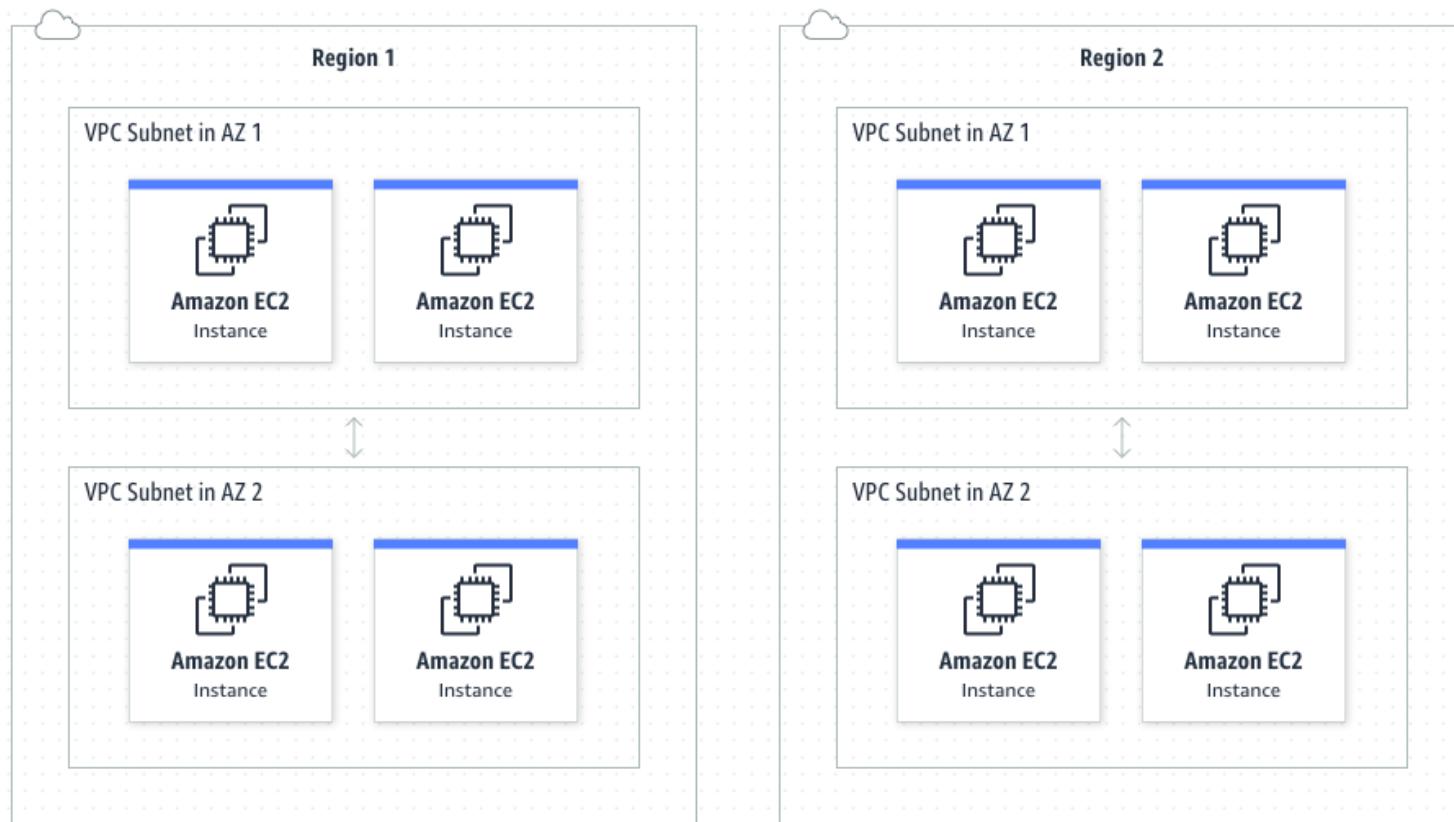
AWS: S3



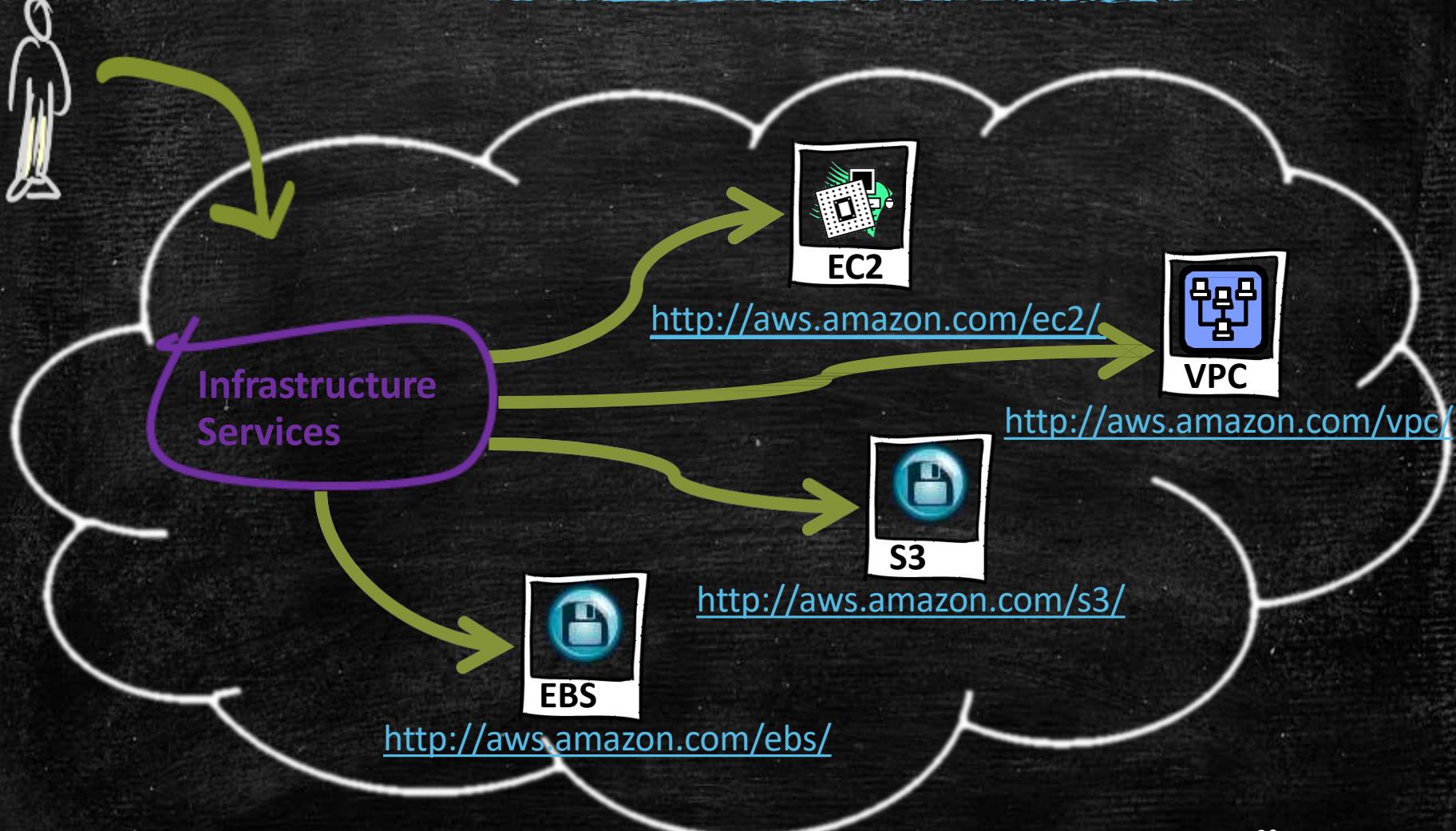
Amazon Virtual Private Cloud (VPC)

- **Amazon VPC** lets you **provision** a **logically isolated section** of the Amazon Web Services (AWS) Cloud.
- You have **complete control** over your virtual networking environment, including:
 - selection of your own **IP address range**,
 - **creation of subnets**, and
 - **configuration of route tables** and **network gateways**.
- VPC allows **bridging with an onsite IT infrastructure** with an **encrypted VPN connection** with an **extra charge per VPN Connection-hour**.
- There is **no additional charge** for using Amazon Virtual Private Cloud, aside from the normal Amazon EC2 usage charges.

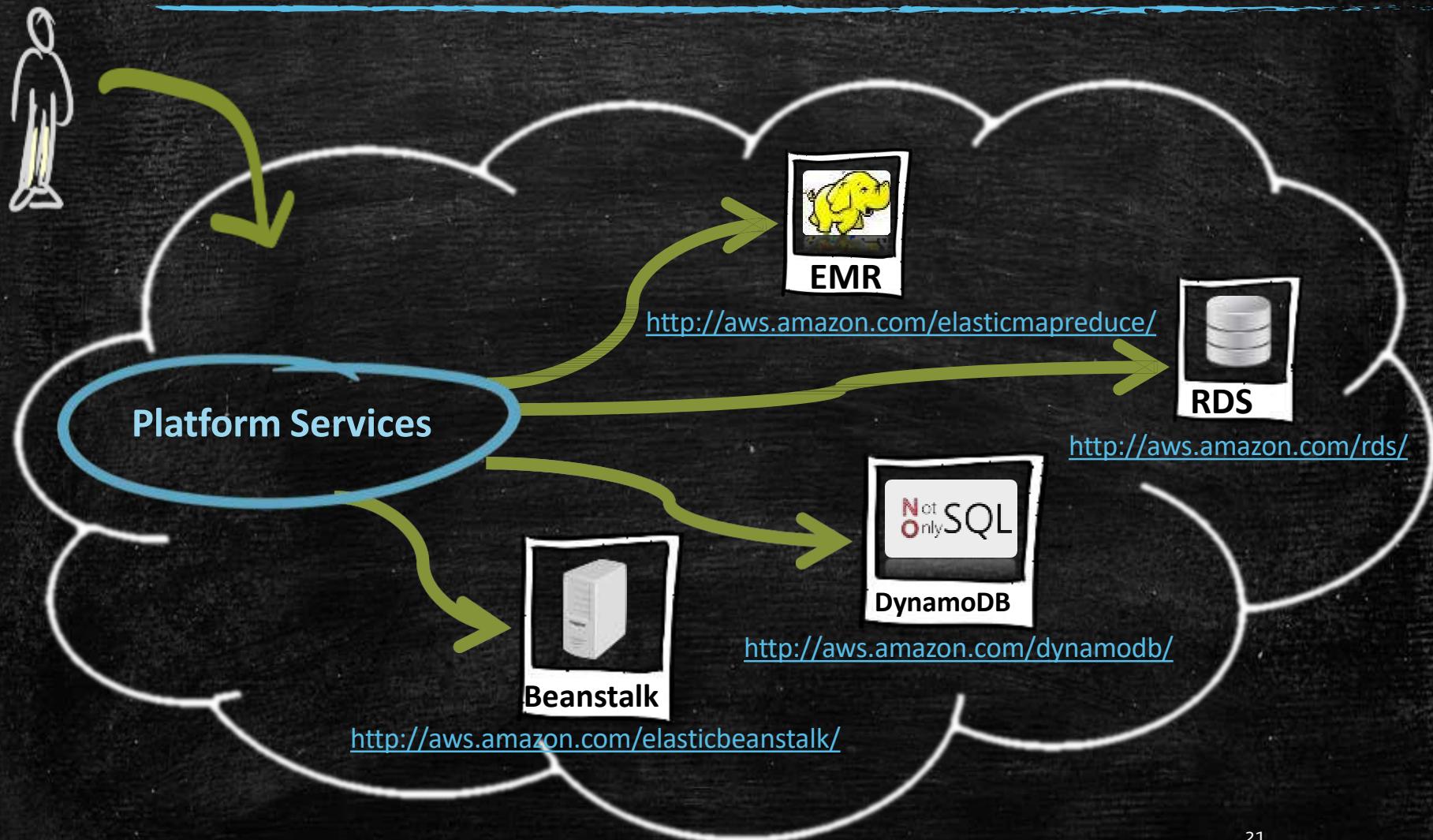
AWS : VPC



Demo & Questions



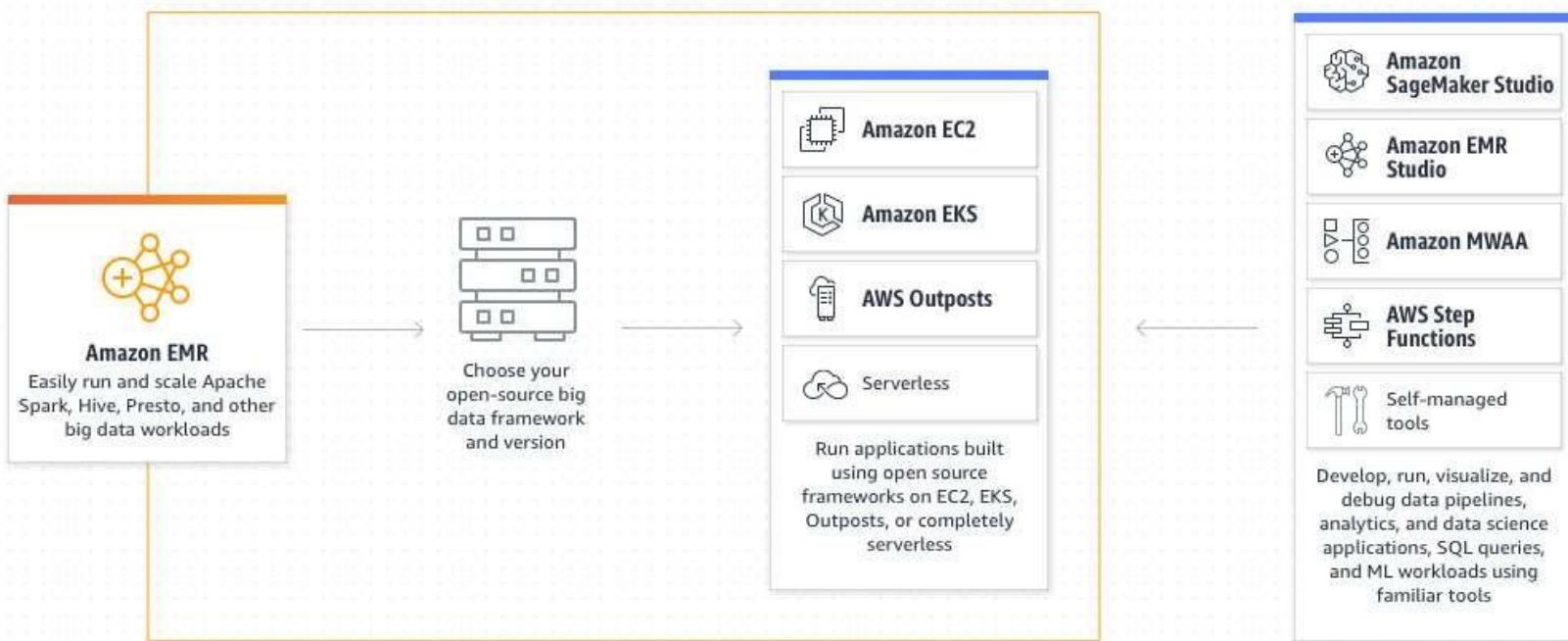
Platform Services



Amazon Elastic MapReduce (EMR)

- Amazon EMR is a web service that makes it easy to **quickly and cost-effectively process vast amounts of data** using **Hadoop**.
- Amazon EMR **distribute** the **data and processing** across a resizable cluster of Amazon **EC2 instances**.
- With Amazon EMR you can launch a **persistent cluster** that stays up indefinitely or a **temporary cluster** that terminates after the analysis is complete.
- Amazon EMR **supports a variety** of Amazon **EC2 instance types** and Amazon EC2 **pricing options** (On-Demand, Reserved, and Spot).
- When launching an Amazon EMR cluster (also called a "job flow"), you **choose how many** and **what type** of Amazon **EC2 Instances** to provision.
- The Amazon **EMR price** is in **addition** to the Amazon **EC2 price**.
- Amazon EMR is used in a variety of applications, including **log analysis**, **web indexing**, **data warehousing**, **machine learning**, **financial analysis**, **scientific simulation**, and **bioinformatics**.

AWS: EMR



Amazon Relational Database Service (RDS)

- **Amazon RDS** is a web service that makes it easy to set up, operate, and scale a **relational database in the cloud**.
- Amazon RDS gives access to the capabilities of a **familiar MySQL, Oracle or Microsoft SQL Server database engine**.
 - Code, applications, and tools already **used with existing databases can be used with RDS**.
- Amazon RDS **automatically patches the database software and backs up the database, storing the backups for a user-defined retention period and enabling point-in-time recovery**.
- Amazon RDS provides **scaling** the **compute resources** or **storage capacity** associated with the Database Instance.
- **Pay** only for the resources actually consumed, **based on the DB Instance hours consumed, database storage, backup storage, and data transfer**.
 - **On-Demand DB Instances** let you pay for compute capacity by the hour with no long-term commitments.
 - **Reserved DB Instances** give the option to make a low, one-time payment for each DB Instance and in turn receive a significant discount on the hourly usage charge for that DB Instance.²⁴

AWS: RDS



<https://aws.amazon.com/rds/>

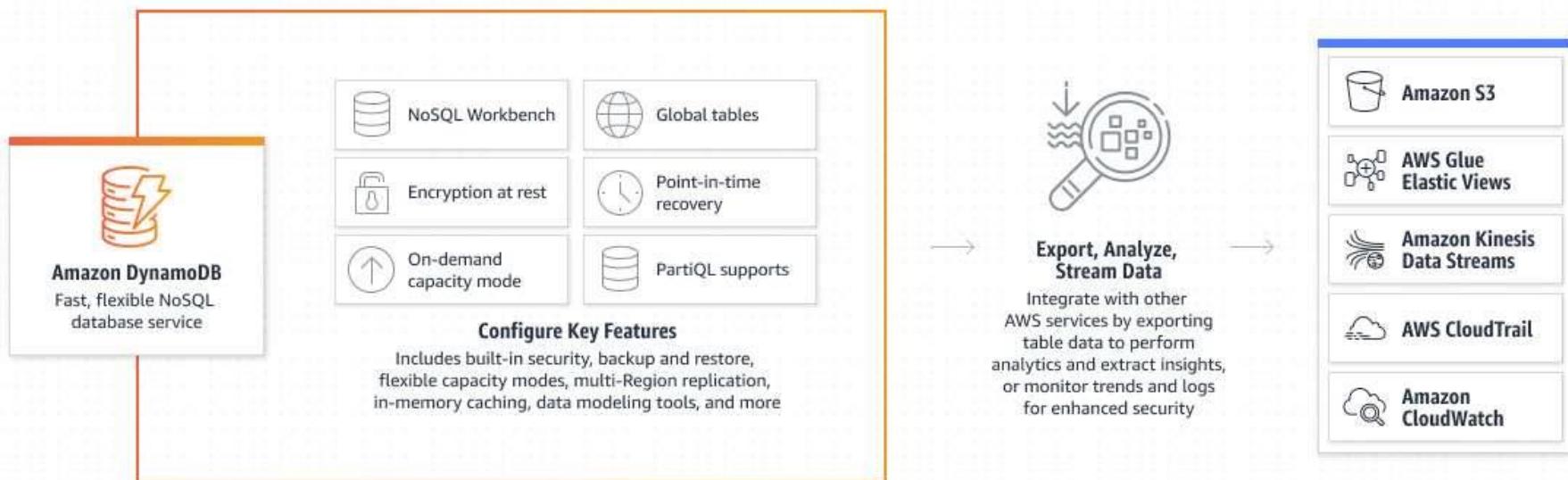
SQL Databases

- In relational databases (SQL Databases), ACID (Atomicity, Consistency, Isolation, Durability) is a set of properties that **guarantee that database transactions are processed reliably**.
 - Atomicity requires that each transaction is "all or nothing": if one part of the transaction fails, the entire transaction fails, and the database state is left unchanged.
 - The consistency property ensures that any transaction will bring the database **from one valid state to another**.
 - The isolation property ensures that the concurrent execution of transactions results in a system state that would be obtained if transactions were executed **serially**.
 - Durability means that **once a transaction has been committed, it will remain so**, even in the event of power loss, crashes, or errors.

Amazon DynamoDB

- **DynamoDB** is a fast, fully managed **NoSQL database service** that makes it simple and cost-effective to store and retrieve any amount of data, and serve any level of request traffic.
- All data items are stored on **Solid State Drives (SSDs)**, and are **replicated** across **3 Availability Zones** for high availability and durability.
- DynamoDB **tables do not have fixed schemas**, and each item may have a **different number of attributes**.
- DynamoDB has **no upfront costs** and implements a **pay as you go** plan as a. **a flat hourly rate** based on the **capacity reserved**.

AWS: DynamoDB



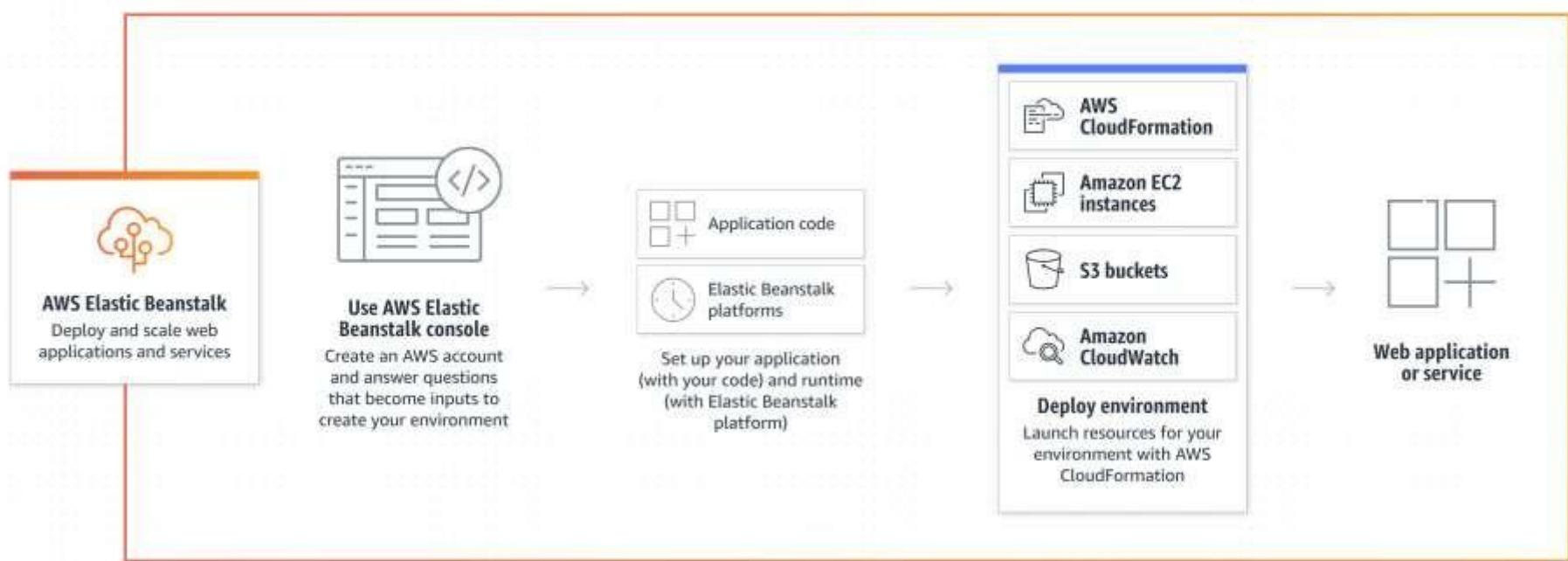
NoSQL Databases

- A **NoSQL database** provides a mechanism for storage and retrieval of data that **employs less constrained consistency models** than traditional relational databases.
- NoSQL databases only support **Eventual Consistency** which is a consistency model used in distributed computing that **informally guarantees that, if no new updates are made to a given data item, eventually all accesses to that item will return the last updated value.**
- NoSQL databases are often **highly optimized key-value stores** intended for **simple retrieval and appending operations**, with the goal being significant **performance benefits in terms of latency and throughput**.
- **Key-value stores** allow the application to **store its data in a schema-less way**.
 - The data could be stored in a **datatype of a programming language** or an object. Because of this, there is **no need for a fixed data model**.

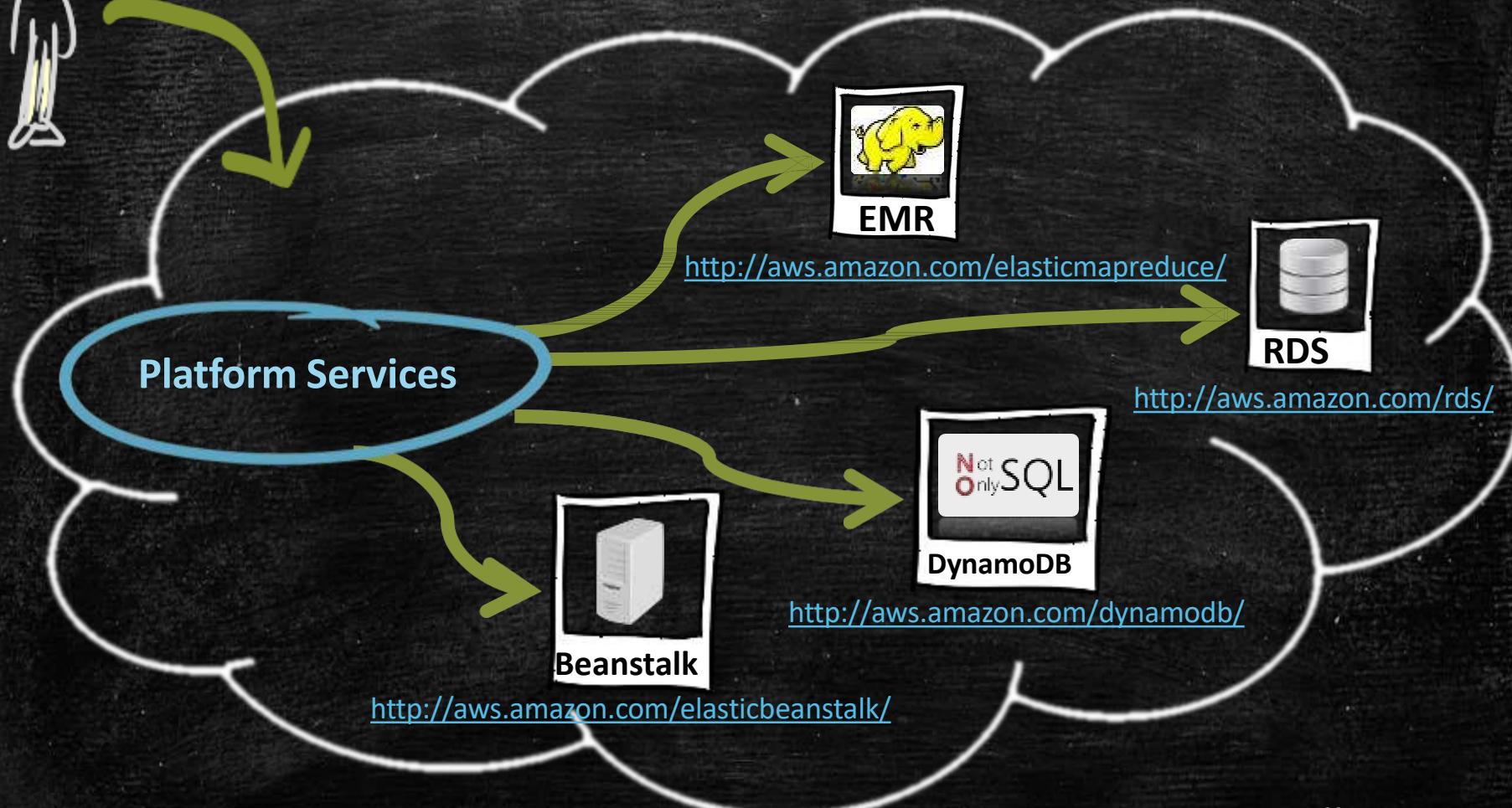
Amazon Elastic Beanstalk

- **AWS Elastic Beanstalk** provides a solution to **quickly deploy** and **manage** applications in the AWS cloud.
- You **simply upload your application**, and Elastic Beanstalk **automatically** handles the deployment details of capacity **provisioning, load balancing, auto-scaling**, and application health **monitoring**.
- Elastic Beanstalk leverages AWS services such as **Amazon EC2, Amazon S3,**
- To ensure easy portability of your application, Elastic Beanstalk is built using familiar software stacks such as:
 - Apache HTTP Server for Node.js, PHP and Python
 - Passenger for Ruby,
 - IIS 7.5 for .NET
 - Apache Tomcat for Java.
- There is **no additional charge** for Elastic Beanstalk - you **pay only for the AWS resources** needed to store and run your applications.

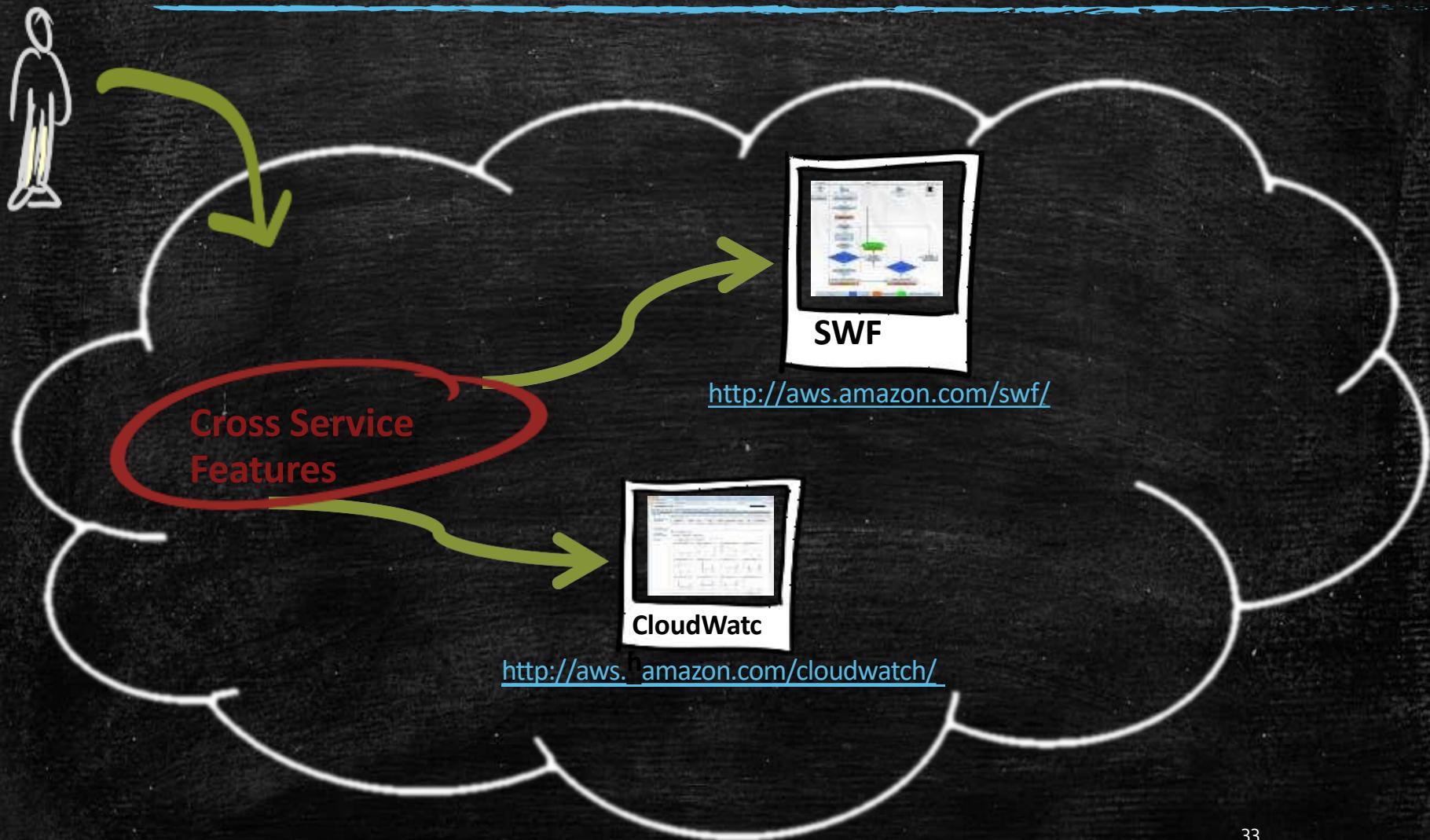
AWS: Beanstalk



Questions



Cross Service Features

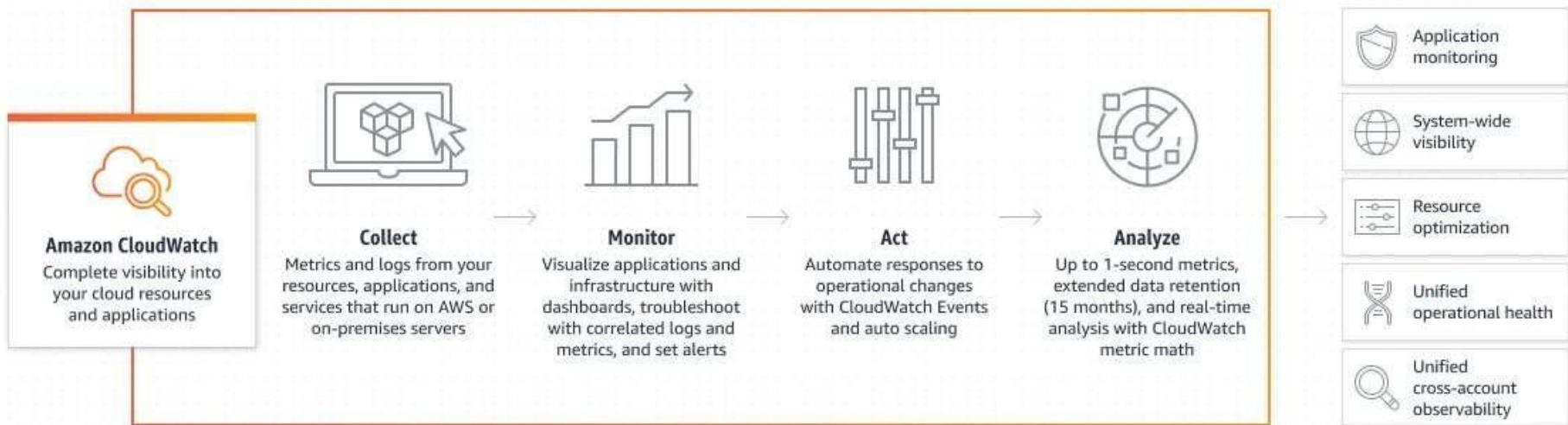




Amazon CloudWatch

- Amazon **CloudWatch** provides **monitoring for AWS cloud resources** and the applications customers run on AWS.
- Amazon CloudWatch lets you **programmatically** retrieve your **monitoring data**, view **graphs**, and **set alarms** to help you troubleshoot, spot trends, and **take automated action** based on the state of your cloud environment.
- Amazon CloudWatch enables you to monitor your AWS resources **up-to-the-minute in real-time**, including:
 - Amazon EC2 instances,
 - Amazon EBS volumes,
 - Elastic Load Balancers,
 - Amazon RDS DB instances.
- Metrics such as **CPU utilization**, **latency**, and **request counts** are provided automatically for these AWS resources.
- Customers can also supply their **own custom application and system metrics**, such as **memory usage**, **transaction volumes**, or **error rates**.

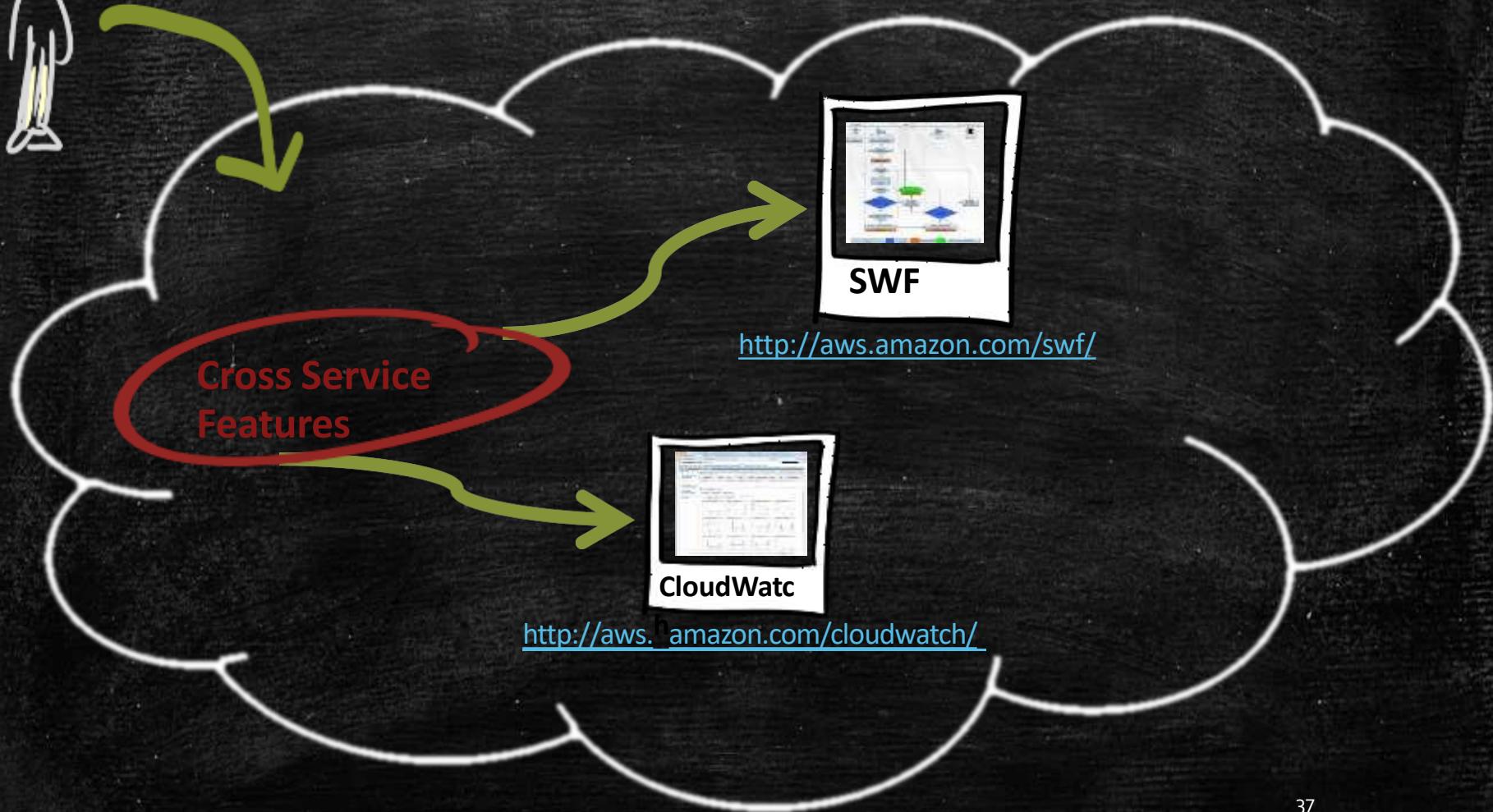
AWS: CloudWatch



Amazon Simple Workflow Service (SWF)

- **Amazon SWF** is a **task coordination** and **state management service** for cloud applications.
- Using Amazon SWF, you **structure** the various **processing steps** in an application that **runs across one or more machines** as a set of **“tasks.”**
- Amazon SWF **manages dependencies** between the tasks, **schedules** the tasks for execution, and runs any logic that needs to be **executed in parallel**.
- The service also **tracks** the tasks’ **progress**.
- As the **business requirements change**, Amazon SWF makes it **easy to change application logic** without having to **worry about the underlying state machinery** and **flow control**.

Questions





Watch out for unexpected Costs

- When you finish your work remember to make sure of the following to **avoid unwanted costs**:
 - **Delete** your **S3** objects.
 - Stop or **Shut Down** your **EC2** and **RDS** instances.
- The customer is responsible for the resources he's using. AWS **declines any responsibility** if the customer forgets to shut down resources.

AWS Free Usage Tier

MORE
INFO →

- <http://aws.amazon.com/free/>

Load Balancing

What is load balancing?

What are the benefits of load balancing?

What are load balancing algorithms?

How does load balancing work?

What are the types of load balancing?

What are the types of load balancing technology?

How can AWS help with load balancing?

What is load balancing?

- Load balancing is the method of distributing network traffic equally across a pool of resources that support an application. Modern applications must process millions of users simultaneously and return the correct text, videos, images, and other data to each user in a fast and reliable manner.
- To handle such high volumes of traffic, most applications have many resource servers with duplicate data between them.
- A load balancer is a device that sits between the user and the server group and acts as an invisible facilitator, ensuring that all resource servers are used equally.

What are the benefits of load balancing?

- Load balancing directs and controls internet traffic between the application servers and their visitors or clients. As a result, it improves an application's availability, scalability, security, and performance.

Application availability

- Server failure or maintenance can increase application downtime, making your application unavailable to visitors. Load balancers increase the fault tolerance of your systems by automatically detecting server problems and redirecting client traffic to available servers. You can use load balancing to make these tasks easier:
 - Run application server maintenance or upgrades without application downtime
 - Provide automatic disaster recovery to backup sites
 - Perform health checks and prevent issues that can cause downtime

Application scalability

- You can use load balancers to direct network traffic intelligently among multiple servers. Your applications can handle thousands of client requests because load balancing does the following:
 - Prevents traffic bottlenecks at any one server
 - Predicts application traffic so that you can add or remove different servers, if needed
 - Adds redundancy to your system so that you can scale with confidence

Contd.

Application security

- Load balancers come with built-in security features to add another layer of security to your internet applications. They are a useful tool to deal with distributed denial of service attacks, in which attackers flood an application server with millions of concurrent requests that cause server failure. Load balancers can also do the following:
- Monitor traffic and block malicious content
- Automatically redirect attack traffic to multiple backend servers to minimize impact
- Route traffic through a group of network firewalls for additional security

Application performance

- Load balancers improve application performance by increasing response time and reducing network latency. They perform several critical tasks such as the following:
- Distribute the load evenly between servers to improve application performance
- Redirect client requests to a geographically closer server to reduce latency
- Ensure the reliability and performance of physical and virtual computing resources

WHAT ARE LOAD BALANCING ALGORITHMS?

- A load balancing algorithm is the set of rules that a load balancer follows to determine the best server for each of the different client requests. Load balancing algorithms fall into two main categories.

Static load balancing

- Static load balancing algorithms follow fixed rules and are independent of the current server state. The following are examples of static load balancing.
- ***Round-robin method***
- Servers have IP addresses that tell the client where to send requests. The IP address is a long number that is difficult to remember. To make it easy, a Domain Name System maps website names to servers. When you enter aws.amazon.com into your browser, the request first goes to our name server, which returns our IP address to your browser.
- In the round-robin method, an authoritative name server does the load balancing instead of specialized hardware or software. The name server returns the IP addresses of different servers in the server farm turn by turn or in a round-robin fashion.
- ***Weighted round-robin method***
- In weighted round-robin load balancing, you can assign different weights to each server based on their priority or capacity. Servers with higher weights will receive more incoming application traffic from the name server.
- ***IP hash method***
- In the IP hash method, the load balancer performs a mathematical computation, called hashing, on the client IP address. It converts the client IP address to a number, which is then mapped to individual servers.

Dynamic load balancing

- Dynamic load balancing algorithms examine the current state of the servers before distributing traffic. The following are some examples of dynamic load balancing algorithms.
- ***Least connection method***
- A connection is an open communication channel between a client and a server. When the client sends the first request to the server, they authenticate and establish an active connection between each other. In the least connection method, the load balancer checks which servers have the fewest active connections and sends traffic to those servers. This method assumes that all connections require equal processing power for all servers.
- ***Weighted least connection method***
- Weighted least connection algorithms assume that some servers can handle more active connections than others. Therefore, you can assign different weights or capacities to each server, and the load balancer sends the new client requests to the server with the least connections by capacity.
- ***Least response time method***
- The response time is the total time that the server takes to process the incoming requests and send a response. The least response time method combines the server response time and the active connections to determine the best server. Load balancers use this algorithm to ensure faster service for all users.
- ***Resource-based method***
- In the resource-based method, load balancers distribute traffic by analyzing the current server load. Specialized software called an agent runs on each server and calculates usage of server resources, such as its computing capacity and memory. Then, the load balancer checks the agent for sufficient free resources before distributing traffic to that server.

What are the types of load balancing?

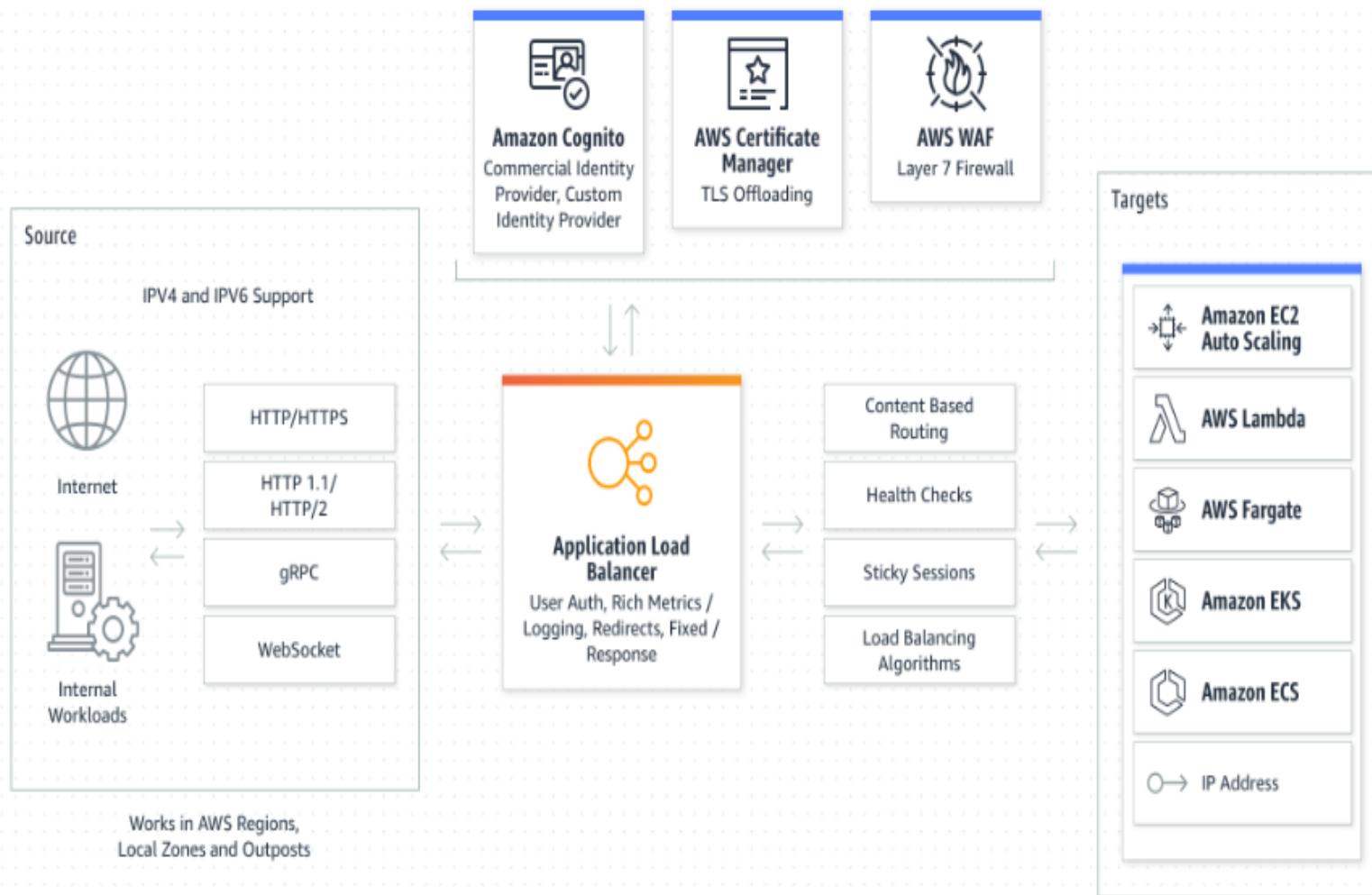
- **Application load balancing**
- Complex modern applications have several server farms with multiple servers dedicated to a single application function. Application load balancers look at the request content, such as HTTP headers or SSL session IDs, to redirect traffic.
- For example, an ecommerce application has a product directory, shopping cart, and checkout functions. The application load balancer sends requests for browsing products to servers that contain images and videos but do not need to maintain open connections. By comparison, it sends shopping cart requests to servers that can maintain many client connections and save cart data for a long time.
- **Network load balancing**
- Network load balancers examine IP addresses and other network information to redirect traffic optimally. They track the source of the application traffic and can assign a static IP address to several servers. Network load balancers use the static and dynamic load balancing algorithms described earlier to balance server load.
- **Global server load balancing**
- Global server load balancing occurs across several geographically distributed servers. For example, companies can have servers in multiple data centers, in different countries, and in third-party cloud providers around the globe. In this case, local load balancers manage the application load within a region or zone. They attempt to redirect traffic to a server destination that is geographically closer to the client. They might redirect traffic to servers outside the client's geographic zone only in case of server failure.
- **DNS load balancing**
- In DNS load balancing, you configure your domain to route network requests across a pool of resources on your domain. A domain can correspond to a website, a mail system, a print server, or another service that is made accessible through the internet. DNS load balancing is helpful for maintaining application availability and balancing network traffic across a globally distributed pool of resources.

What are the types of load balancing technology?

- **Hardware load balancers**
- A hardware-based load balancer is a hardware appliance that can securely process and redirect gigabytes of traffic to hundreds of different servers. You can store it in your data centers and use virtualization to create multiple digital or virtual load balancers that you can centrally manage.
- **Software load balancers**
- Software-based load balancers are applications that perform all load balancing functions. You can install them on any server or access them as a fully managed third-party service.
- **Comparison of hardware balancers to software load balancers**
- Hardware load balancers require an initial investment, configuration, and ongoing maintenance. You might also not use them to full capacity, especially if you purchase one only to handle peak-time traffic spikes. If traffic volume increases suddenly beyond its current capacity, this will affect users until you can purchase and set up another load balancer.
- In contrast, software-based load balancers are much more flexible. They can scale up or down easily and are more compatible with modern cloud computing environments. They also cost less to set up, manage, and use over time.

How does load balancing work?

- Companies usually have their application running on multiple servers. Such a server arrangement is called a server farm.
- User requests to the application first go to the load balancer. The load balancer then routes each request to a single server in the server farm best suited to handle the request.
- Load balancing is like the work done by a manager in a restaurant. Consider a restaurant with five waiters. If customers were allowed to choose their waiters, one or two waiters could be overloaded with work while the others are idle. To avoid this scenario, the restaurant manager assigns customers to the specific waiters who are best suited to serve them.

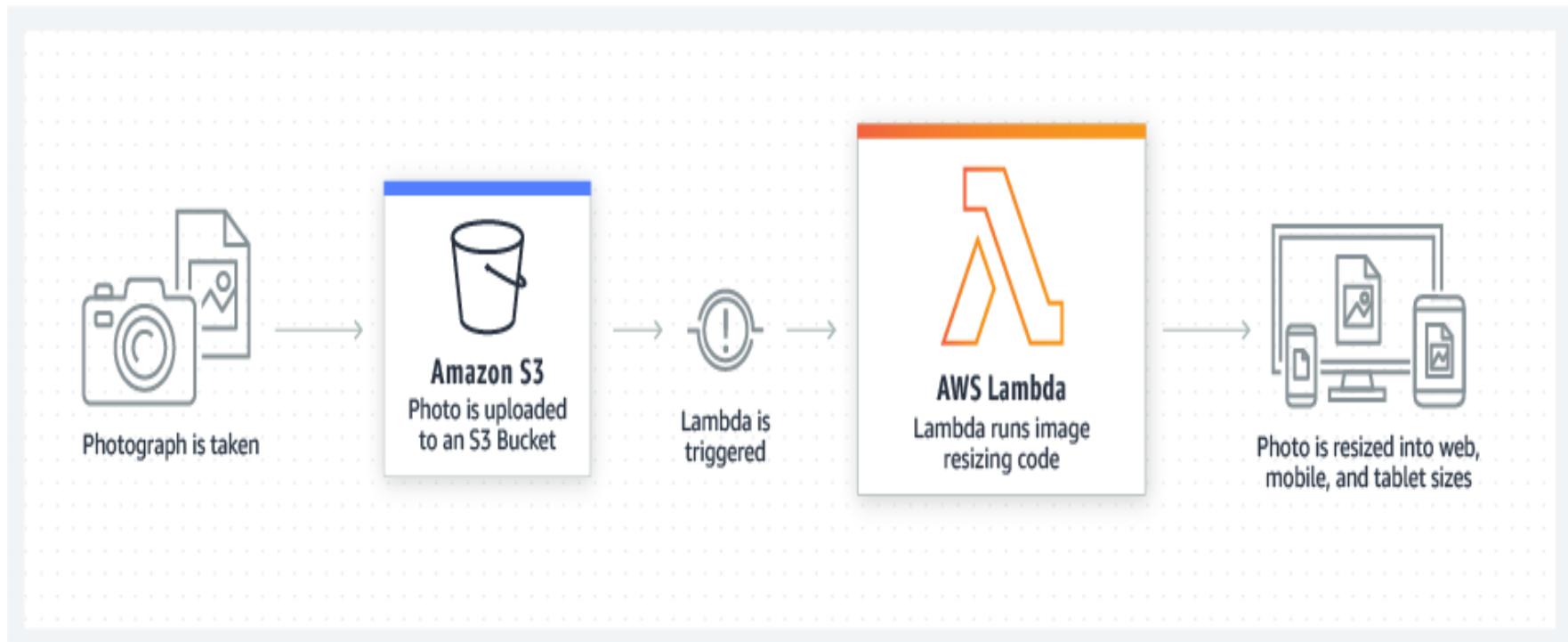


AWS Lambda

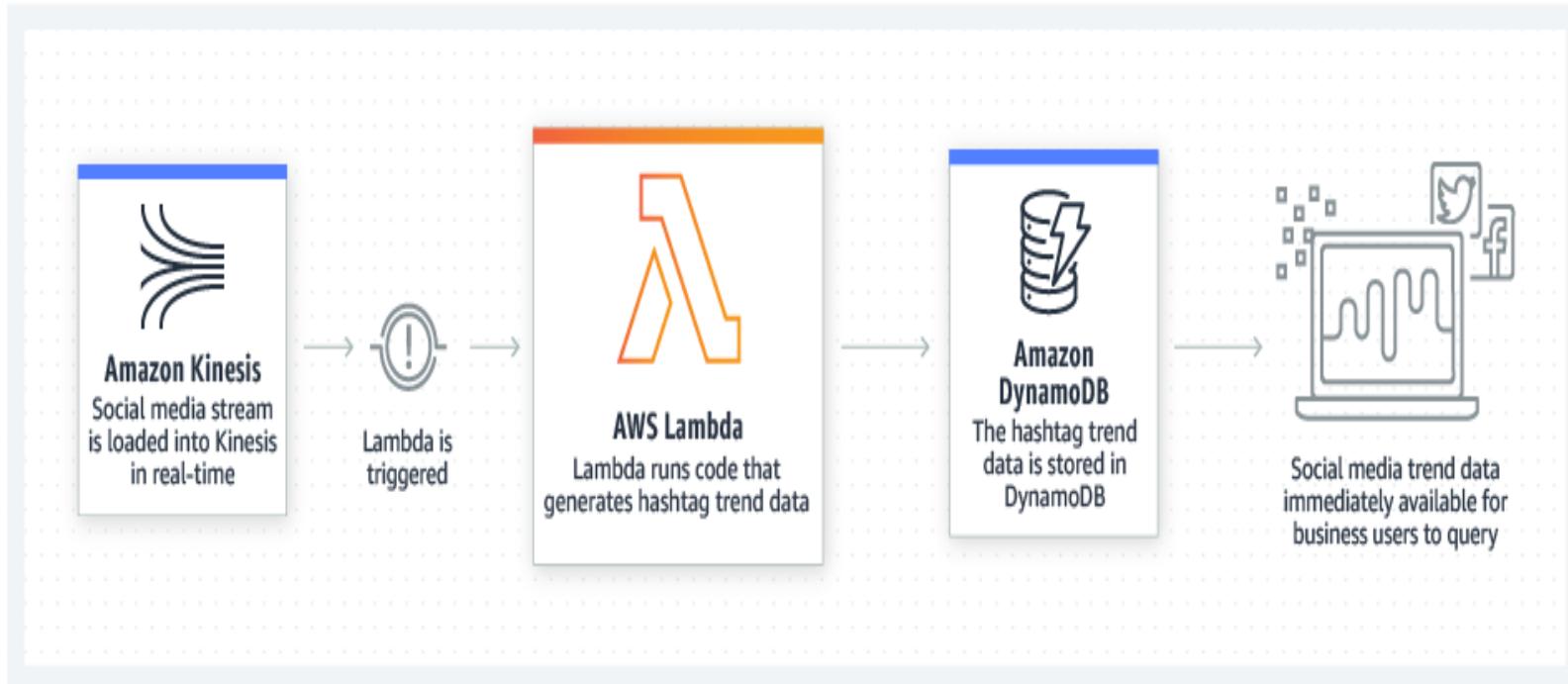
Run code without thinking about servers or clusters

- AWS Lambda is a serverless, event-driven compute service that lets you run code for virtually any type of application or backend service without provisioning or managing servers.
- File processing
- Stream Processing
- Web application'
- IOT backends
- Mobile backends

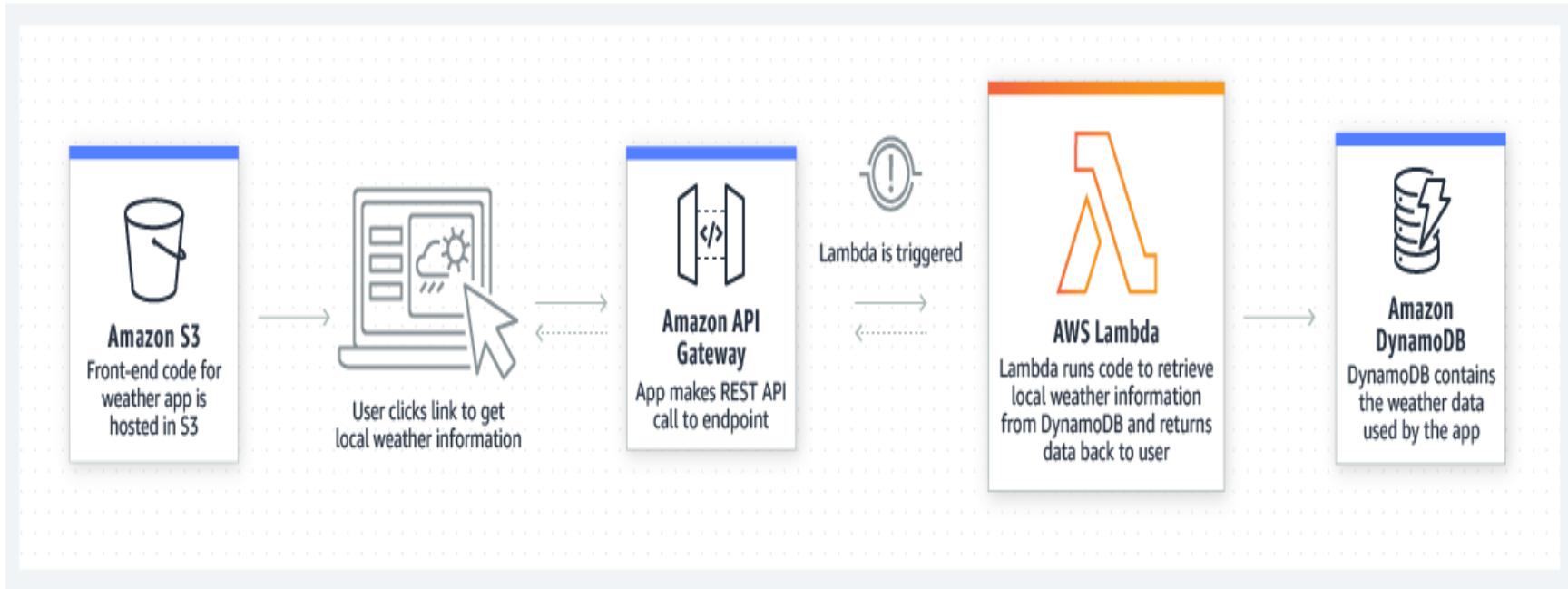
File Processing



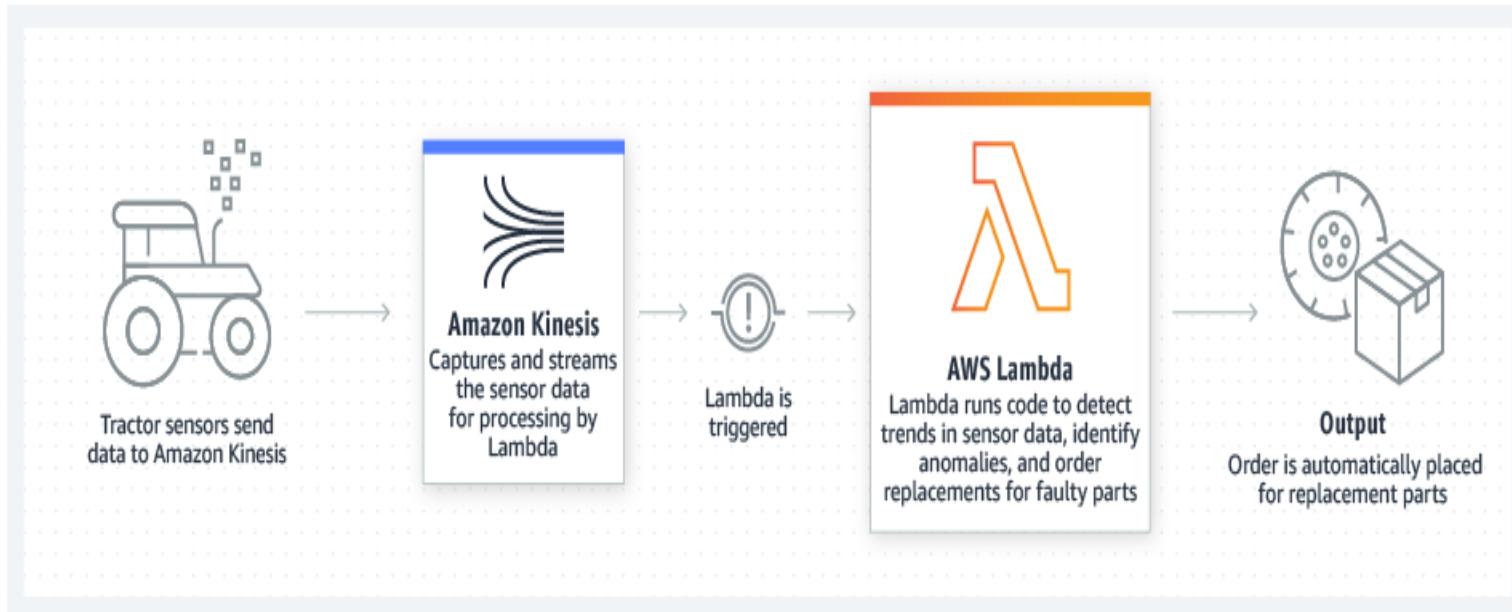
Stream Processing



Web application



IOT backends



Mobile backends



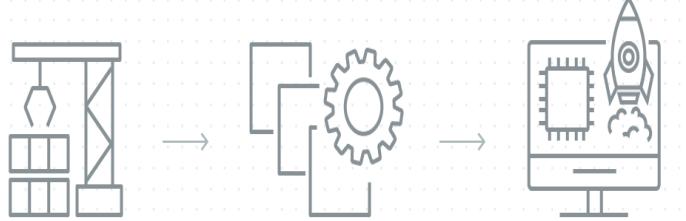
AWS Fargate

Serverless compute for containers

How it works ?

- AWS Fargate is a serverless, pay-as-you-go compute engine that lets you focus on building applications without managing servers. AWS Fargate is compatible with both Amazon Elastic Container Service (ECS) and Amazon Elastic Kubernetes Service (EKS).

Without Fargate



Build your container image

Define and deploy the EC2 Instances

Provision and manage compute and memory resources



Isolate applications in separate VMs



Run and manage both applications and infrastructure



Pay for EC2 Instances

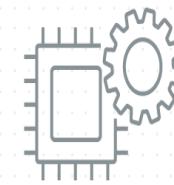
With Fargate



AWS Fargate



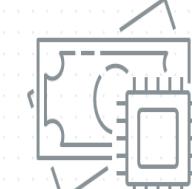
Build container image



Define memory and compute resources required



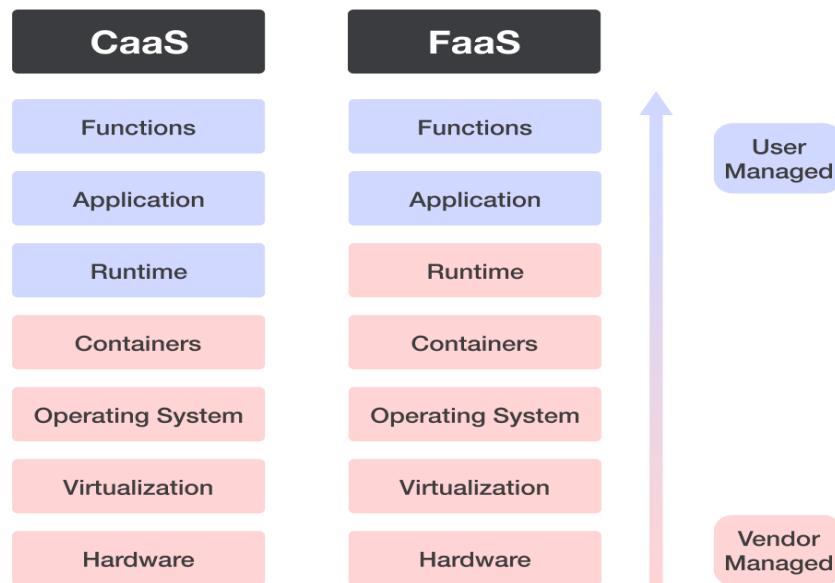
Run and manage applications



Pay for requested compute resources when used. Application isolation by design

What is the difference between EC2 Lambda and Fargate?

- While Fargate is a **Container as a Service (CaaS)** offering, AWS Lambda is a **Function as a Service (FaaS** offering). Therefore, Lambda functions do not necessarily need to be packaged into containers, making it easier to get started with Lambda. But if you have containerized applications, Fargate is the way to go.



How can AWS help with load balancing?

- Elastic Load Balancing (ELB) is a fully managed load balancing service that automatically distributes incoming application traffic to multiple targets and virtual appliances across AWS and on-premises resources.
- You can use it to scale modern applications without complex configurations or API gateways.
- You can use ELB to set up four different types of software load balancers.
- An Application Load Balancer routes traffic for HTTP-based requests.
- A Network Load Balancer routes traffic based on IP addresses. It is ideal for balancing TCP and User Datagram Protocol (UDP)-based requests.
- A Gateway Load Balancer routes traffic to third-party virtual appliances. It is ideal for incorporating a third-party appliance, such as a network firewall, into your network traffic in a scalable and easy-to-manage way.
- A Classic Load Balancer routes traffic to applications in the Amazon EC2-Classic network—a single, flat network that you share with other customers.

Example

- For example, [Terminix](#), a global pest control brand, uses Gateway Load Balancer to handle 300% more throughput.
- [Second Spectrum](#), a company that provides artificial intelligence-driven tracking technology for sports broadcasts, uses AWS Load Balancer Controller to reduce hosting costs by 90%.
- [Code.org](#), a nonprofit dedicated to expanding access to computer science in schools, uses Application Load Balancer to handle a 400% spike in traffic efficiently during online coding events.

References

- <https://aws.amazon.com/blogs/containers/amazon-ecs-vs-amazon-eks-making-sense-of-aws-container-services/>
- <https://aws.amazon.com/ecs/>
- <https://aws.amazon.com/eks/>
- <https://bluexp.netapp.com/blog/aws-cvo-blg-aws-ecs-vs-eks-6-key-differences#:~:text=4.-,Portability,support%20for%20portability%20of%20workloads.>
- <https://docs.aws.amazon.com/prescriptive-guidance/latest/patterns/deploy-a-grpc-based-application-on-an-amazon-eks-cluster-and-access-it-with-an-application-load-balancer.html>

Cloud Infrastructure(M-3)

Aradhana Behura

Communication & Computing Group

Department of CSE

Email: aradhana.behurafcs@kiit.ac.in,

Scalability and Elasticity in Cloud Computing

- **Scalability** is the ability of the system to accommodate larger loads just by adding resources either making hardware stronger (**scale up**) or adding additional nodes (**scale out**).
- **Elasticity** is the ability to fit the resources needed to cope with loads dynamically usually in relation to scale out

Scalability in cloud computing

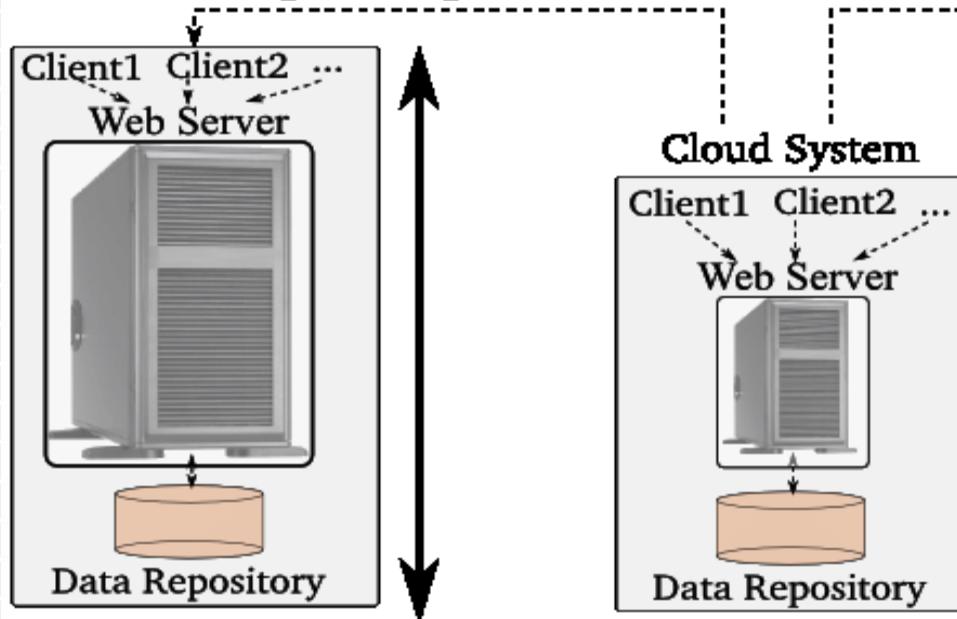
- **Scalability** is the property of a system to handle a growing amount of work by adding resources to the system.
- It is the ability to quickly and easily increase or decrease the size or power of an IT solution.
- A **scalable cloud** is why you can sign up and use most **cloud** solutions in just a few minutes – if not seconds. It's why you can add resources like storage to an existing account just as quickly.

Characteristics of a truly scalable application:

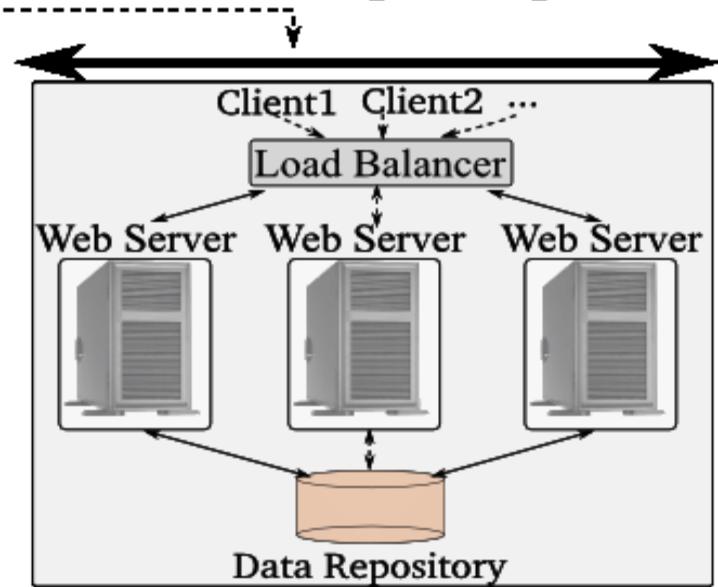
- Increasing resources results in a proportional increase in performance
- A scalable service is capable of handling heterogeneity
- A scalable service is operationally efficient
- A scalable service is **resilient**
- A scalable service should become more cost effective when it grows (Cost per unit reduces as the number of units increases)

Three types of scalability : Vertical, Horizontal and Diagonal

Vertical Scaling: Scaling UP/DOWN



Horizontal Scaling: Scaling OUT/IN



Scaling

Scaling, from an IT resource perspective, represents the ability of the IT resource to handle increased or decreased usage demands.

The following are types of scaling:

- *Horizontal Scaling* – scaling out and scaling in
- *Vertical Scaling* – scaling up and scaling down

The next two sections briefly describe each.

Horizontal Scaling

The allocating or releasing of IT resources that are of the same type is referred to as *horizontal scaling* (Figure 3.4). The horizontal allocation of resources is referred to as *scaling out* and the horizontal releasing of resources is referred to as *scaling in*. Horizontal scaling is a common form of scaling within cloud environments.

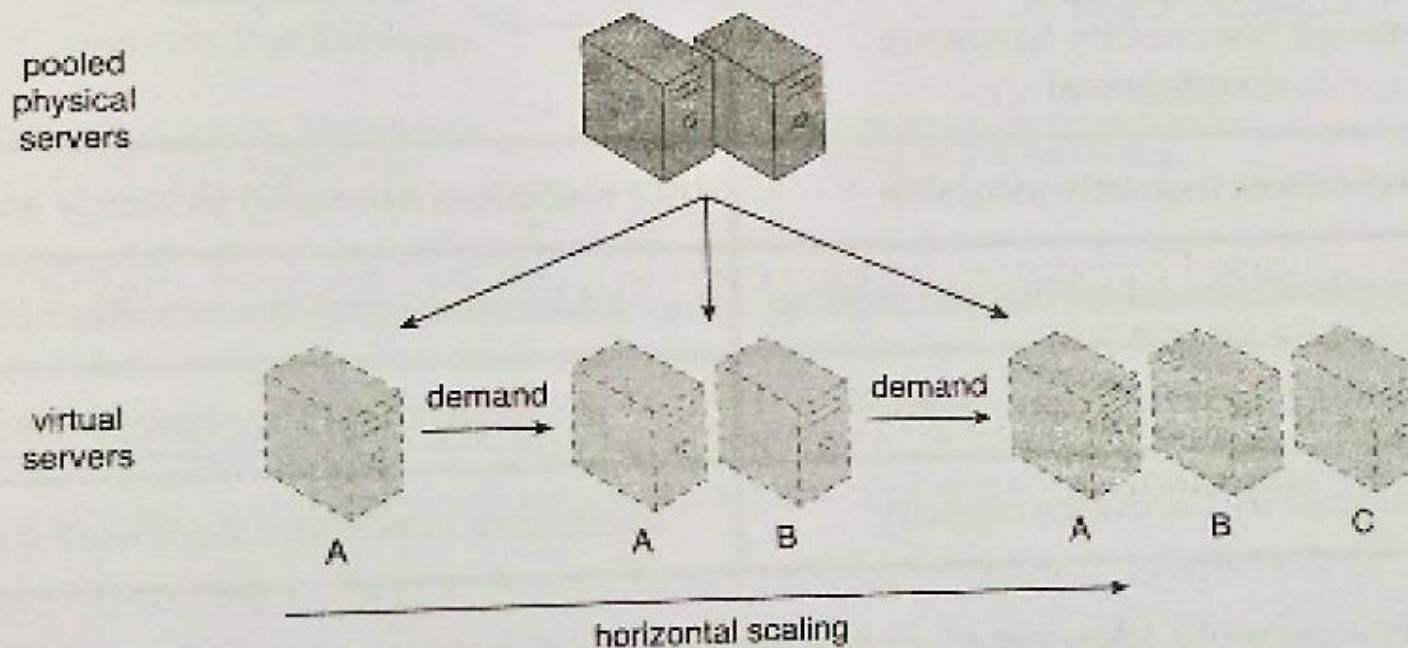


Figure 3.4

An IT resource (Virtual Server A) is scaled out by adding more of the same IT resources (Virtual Servers B and C).

Contd.

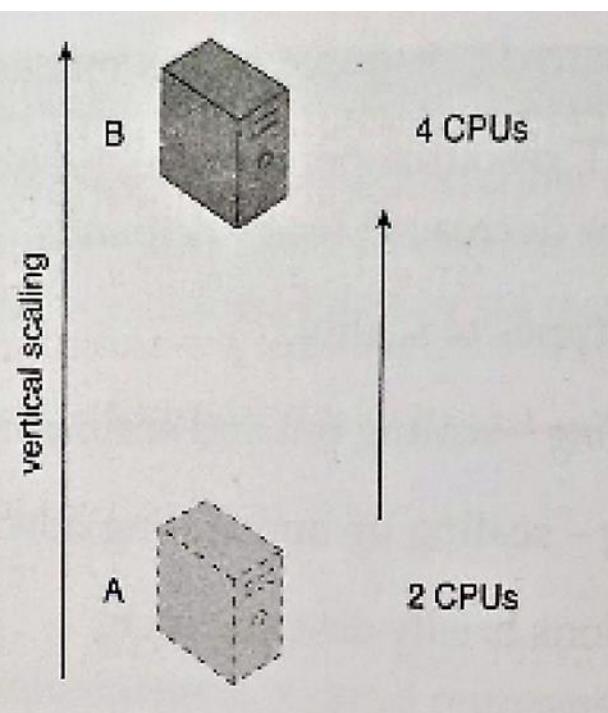
- The allocating or releasing of IT resources that are of the same type is referred to as horizontal scaling (fig 3.4).
- The horizontal allocation of resources is referred to as scaling out & the horizontal releasing of resources is referred to as scaling in.

Vertical Scaling

- When an existing IT resource is replaced by another with higher or lower capacity, vertical scaling is considered to have occurred (Fig.3.5).
- Specifically, the replacing of an IT resource with another that has a higher capacity is referred to **as scaling up** & the replacing an IT resource with another that has a lower capacity is considered **scaling down**.
- Vertical scaling is less common in cloud environments due to the downtime required while the replacement is taking place.

Figure 3.5

An IT resource (a virtual server with two CPUs) is scaled up by replacing it with a more powerful IT resource with increased capacity for data storage (a physical server with four CPUs).



Horizontal Scaling	Vertical Scaling
less expensive (through commodity hardware components)	more expensive (specialized servers)
IT resources instantly available	IT resources normally instantly available
resource replication and automated scaling	additional setup is normally needed
additional IT resources needed	no additional IT resources needed
not limited by hardware capacity	limited by maximum hardware capacity

Table 3.1

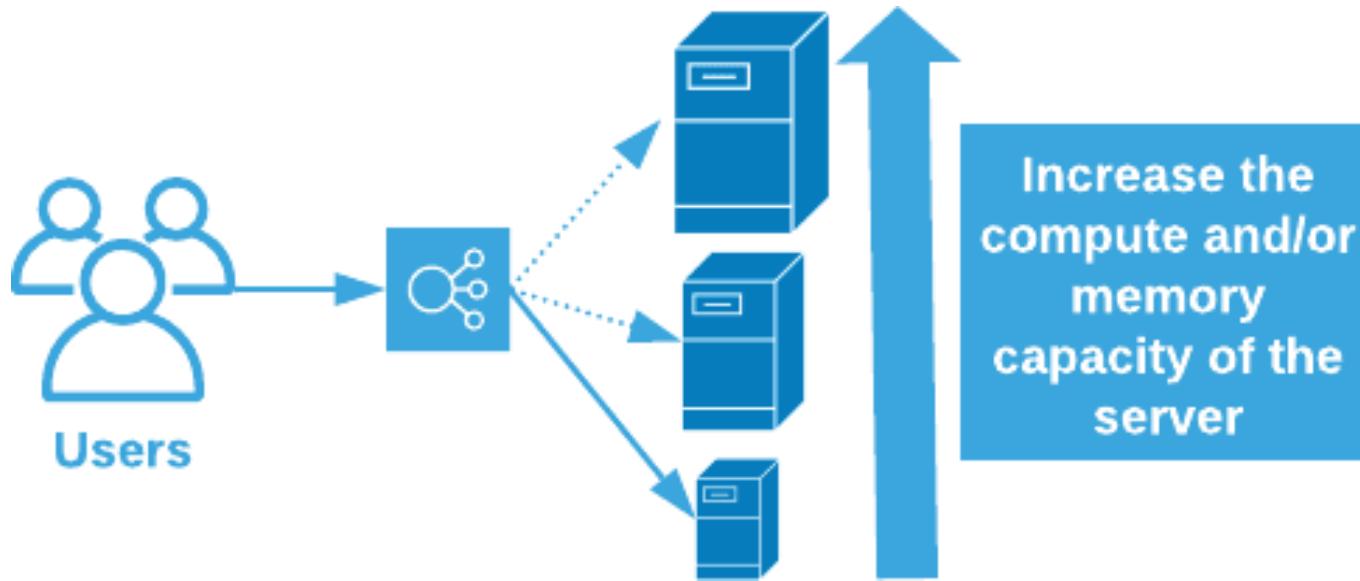
Three types of scalability – Vertical, Horizontal and Diagonal

Scale Vertically - Scale Up:

- Vertical Scaling or Scaling up is easy, it can be done by moving the application to bigger virtual machines deployed in the cloud or you can scale up by adding expansion units as well with your current infrastructure.
- This ability to add resources to accommodate increasing workload volumes is **vertical scaling**. It can resize your server with no change in your code.
- The downside to scaling up is that it increases storage capacity but the performance is reduced because the compute capacity remains the same. Workloads requiring higher throughput demand reduced latency and this can only be fulfilled by Horizontal Scaling / Scaling out.

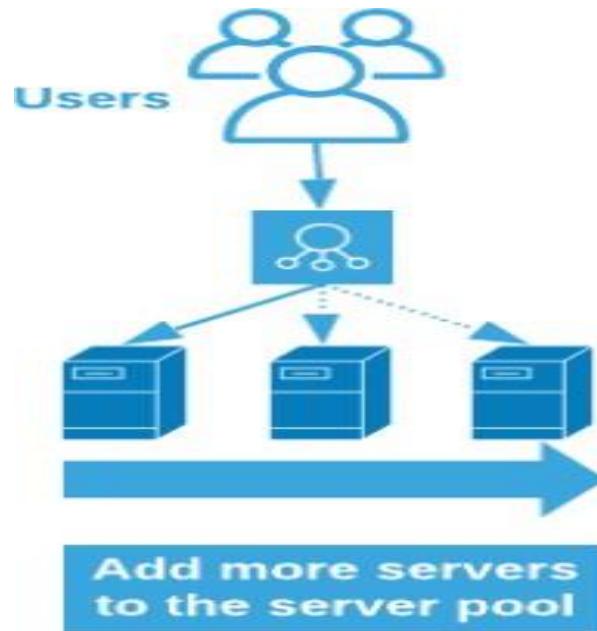
Vertical scale or scale-up

- Adding more compute or memory resources for your applications increases the maximum capacity of the server. When demand spikes, there should not be any noticeable change to your applications.
- **Example:** An example of this would be increasing the number of CPUs or increase the memory of a database server.



Horizontal scale or scale-out

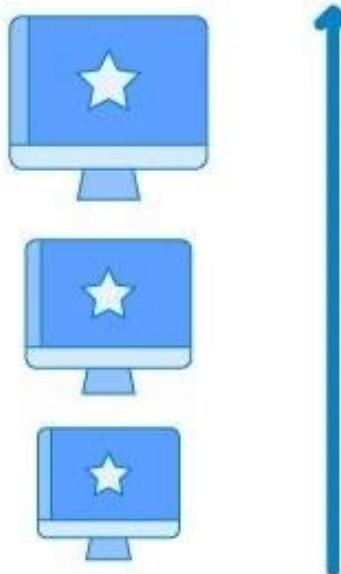
- **Horizontal scale or scale-out:** Adding more individual servers into a resource pool where your applications run.
- **Example:** An example of this would be adding more web servers to your system to handle an increase in traffic.



Vertical vs Horizontal scaling

VERTICAL SCALING

Increase size of instance
(RAM, CPU etc.)



HORIZONTAL SCALING

(Add more instances)



Vertical vs Horizontal scaling

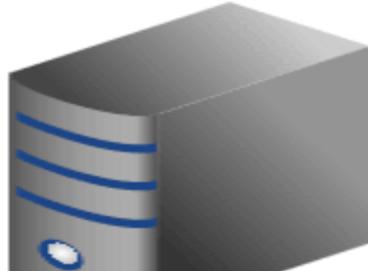
Vertical Scaling



1 CPU / 1 GB RAM
~ \$10/mo



2 CPU / 2 GB RAM
~ \$20/mo



4 CPU / 8 GB RAM
~ \$80/mo

Horizontal Scaling



1 CPU / 1 GB RAM
~ \$10/mo



2 x (1 CPU / 1 GB RAM)
~ \$20/mo

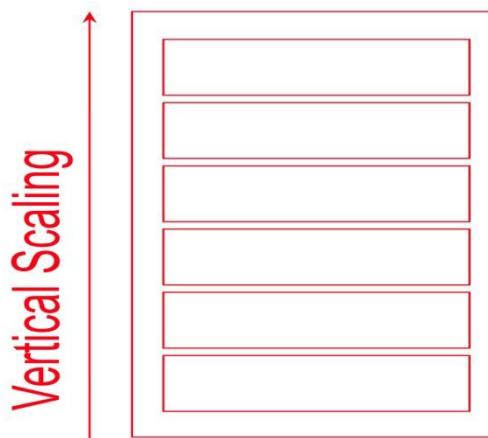


4 x (1 CPU / 1 GB RAM)
~ \$40/mo

Three types of scalability – Vertical, Horizontal and Diagonal

Scale Horizontally - Scale out:

- Horizontal Scaling or Scaling out is the addition of nodes to the existing infrastructure to accommodate additional workload volumes. Contrary to Vertical Scaling, Horizontal Scaling also delivers performance along with storage capacity.
- The total workload volume is aggregated over the total number of nodes and latency is effectively reduced. This scaling is ideal for workloads that require **reduced latency** and **optimized throughput**.



To scale more, Add more RAM, CPU, Memory to the **one existing machine**

To scale more: Add more machines to existing **group of distributed system**

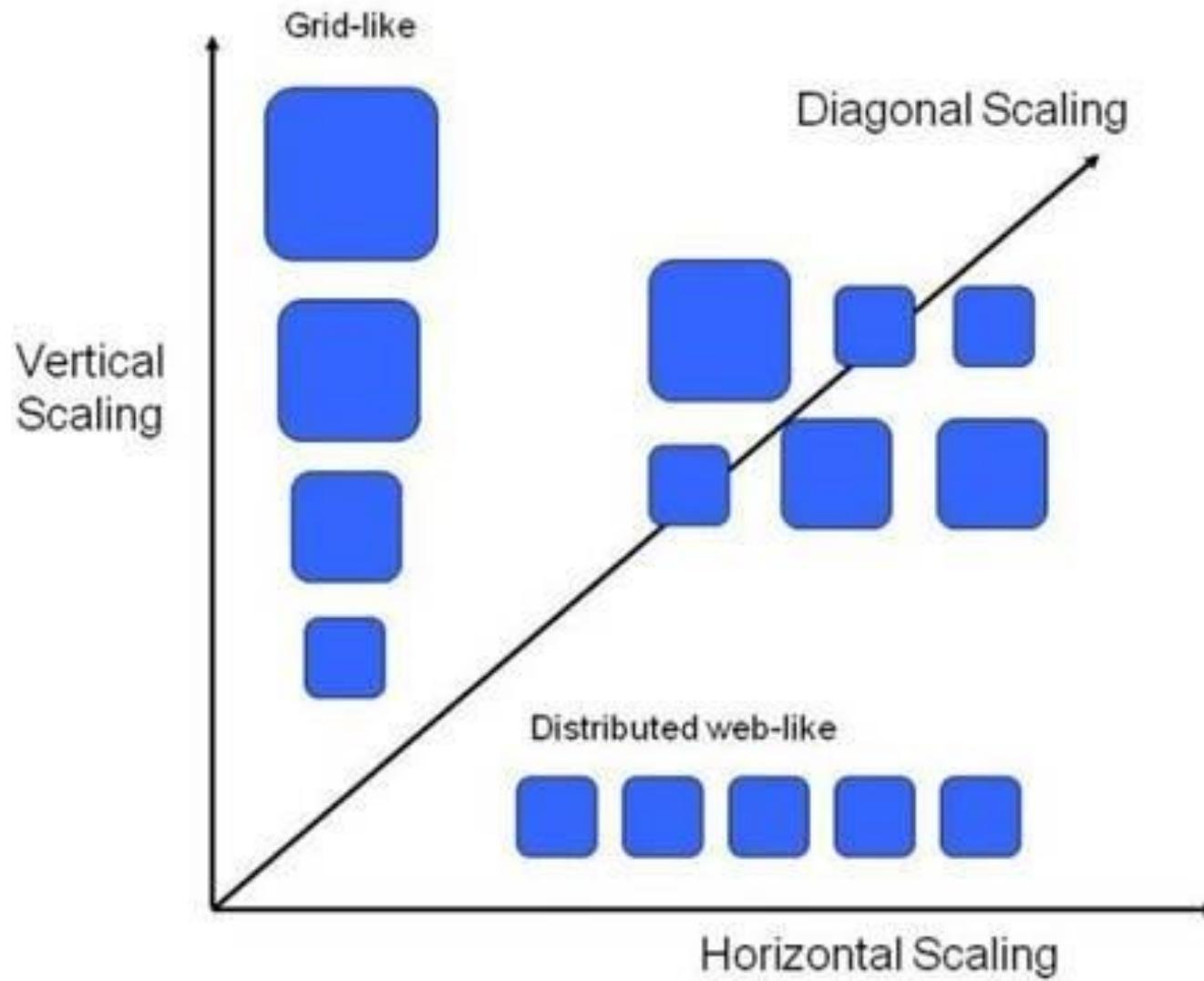


Three types of scalability – Vertical, Horizontal and Diagonal

Scale Diagonally:

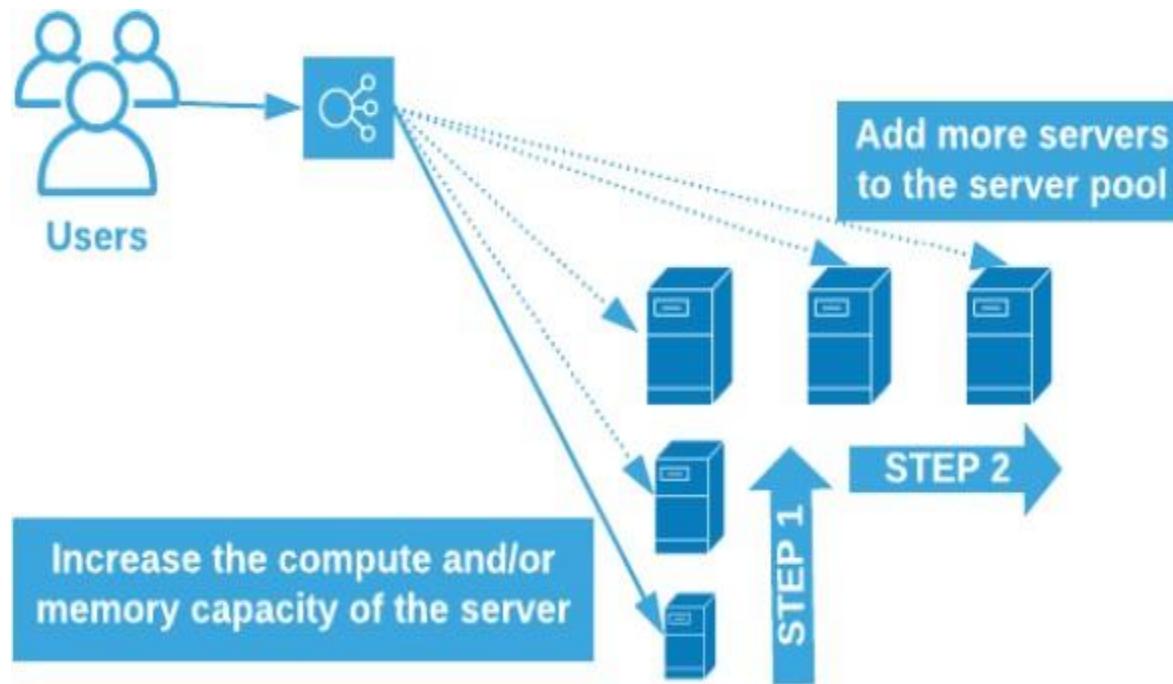
- Diagonal scaling helps you combine the scaling up and scaling down. As the term suggests, scaling down is the removal of storage resources as requirements decrease.
- Diagonal scaling delivers flexibility for workload that require additional storage resources for specific instances of time.
- For instance, a website sets up diagonal scaling; as the traffic increases, the compute requirements are accommodated. As the traffic decreases, the computation capacity is restored to its original size.
- This type of scaling introduces enhanced budgeting and cost effectiveness for environments and businesses dealing with variable workload volumes.

Diagonal Scaling



Diagonal scaling:

- **Diagonal scaling:** Essentially a combination of vertical and horizontal scaling, this setup will scale vertically first until you reach a preset limit and then scale the system horizontally.



Scalability function

1. Units of work (requests).
 2. The rate of requests over time (arrival rate).
 3. The number of units of work in a system at a time (concurrency).
 4. The number of customers, users, or driver processes sending requests.
-
- Each of these can play sensible roles in the scalability function
 - Scalability is the property of a system to handle a growing amount of work by adding resources to the system.
 - Scalability can be defined as a mathematical function, a relationship between independent and dependent variables (input and output).

Best practices for using scalability in cloud computing:

- **Leverage auto-scaling capability with supervision:** Most cloud providers provide auto-scaling options. This allows us to manage the required resources appropriately as needed.
- **Architect your solution for scalability:** Not all applications can work as expected when scaled. This requires well-defined architecture and design patterns, such as distributed queues, Statelessness, Scalable storage needs etc. that makes the applications work well when scaled.
- **Use load balancers:** It is important to have load balancers in the front which will receive the incoming traffic and manage the distribution of load across all the servers as you **scale up** or **scale down**.
- **Have a robust testing strategy:** Ensure you can test the scalability of your applications and the configurations put in place. Real business transactions contributing to revenue stream is not the place to test this.

Business whose resource demands are increasing slowly and predictably.

Example: [Call center]

- The typical call center is continuously growing. New employees come in to handle an increasing number of customer requests gradually, and new features are introduced to the system (like sentiment analysis, embedded analytics, etc.).
- In this case, cloud scalability is used to keep the system's performance as consistent and efficient as possible over an extended time and growth.

Example: [Call center]

- In natural language processing model training and optimization for chat-bots. The system starts off on a certain scale and requires room for gradual improvement as it is being used. The database expands and the operating inventory becomes much more intricate.

Scalable Cloud Based Services:

- **Infrastructure-as-a-Service (IaaS)** - like Amazon EC2 or Google Compute Engine;
- **Platform-as-a-Service (PaaS)** - like Magento Commerce Cloud or AWS Elastic Beanstalk;
- **Storage-as-a-Service (STaaS)** - Google Drive, Microsoft OneDrive, and the likes
- .
- **Data-as-a-Service (DaaS)** - customer relationship platforms like Salesforce and Hubspot, ERP applications;
- **Database-as-a-Service (DBaaS)** - AWS SimpleDB, Rackspace, Oracle, MongoDB;

Benefits of cloud scalability

Performance:

- One core benefit of scalability in the cloud is that it facilitates performance. Scalable architecture has the ability to handle the bursts of traffic and heavy workloads that will come with business growth.

Cost-efficient:

- You can allow your business to grow without making any expensive changes in the current setup. This reduces the cost implications of storage growth making scalability in the cloud very cost effective.

Easy and Quick:

- Scaling up or scaling out in the cloud is simpler; you can commission additional VMs with a few clicks and after the payment is processed, the additional resources are available without any delay.

Benefits of cloud scalability

Capacity:

- Scalability ensures that with the continuous growth of your business the storage space in cloud grows as well. Scalable cloud computing systems accommodate your data growth requirements. With scalability, you don't have to worry about additional capacity needs.

Scalability admonition:

- Scalability also has some limitations. If you want a fully scalable system then you have a large task to handle. It requires planning, testing and again testing for your data storage. If you have the applications already then splitting up the system will require code changes, updates

Monitoring:

- You have to be well prepared for the digital transformation of your infrastructure.

Elasticity

Elastic computing is the ability to quickly expand or decrease computer processing, memory and storage resources to meet changing demands without worrying about capacity planning and engineering for peak usage.

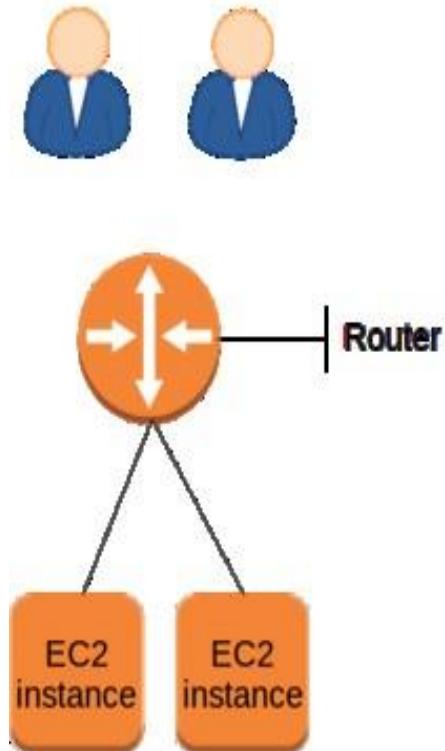
Elasticity – generally refers to increasing or decreasing cloud resources.

What is Cloud Elasticity?

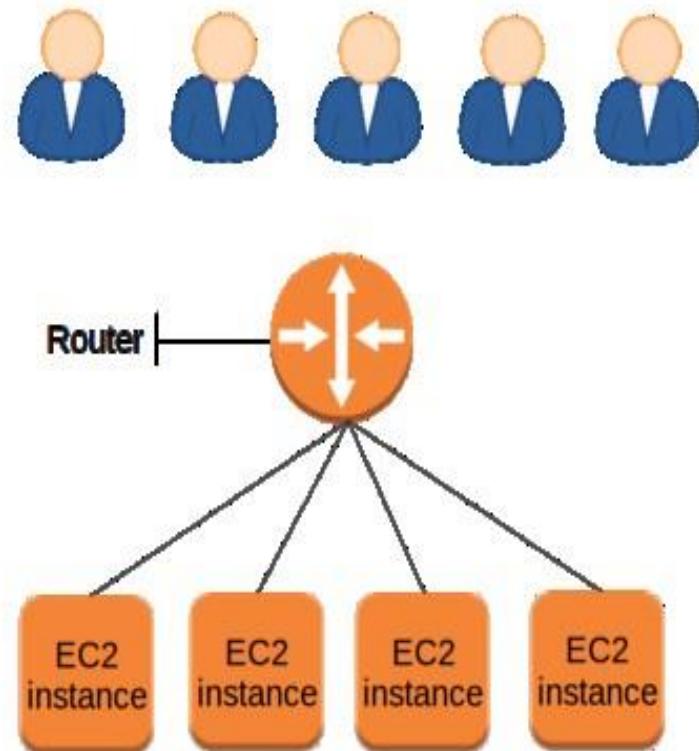
- Cloud elasticity is a system's ability to manage available resources according to the current workload requirements dynamically.
- Elasticity is a vital feature of cloud infrastructure. It comes in handy when the system is expected to experience sudden spikes of user activity and, as a result, a drastic increase in workload demand.
- Because of the pay-per-use pricing model of modern cloud platforms, cloud elasticity is a cost-effective solution for a business with a dynamic workload.
- Businesses with dynamic resource demands like streaming services or e-commerce marketplaces changes dynamically with various seasonal events. These volatile ebbs and flows of workload require flexible resource management to handle the operation consistently.

AWS and Elasticity:

Low demand period



High demand period



Elasticity is Cloud Infrastructure

- *Elasticity is one of the fundamental properties of the cloud.*
- Elasticity is the power to scale computing resources up and down easily and with minimal friction. It is important to understand that elasticity will ultimately drive most of the benefits of the cloud.
- As a cloud architect, you need to internalize this concept and work it into your application architecture in order to take maximum benefit of the cloud.
- The on-demand and elastic nature of *the cloud approach* (Automated Elasticity), however, enables the infrastructure to be closely aligned (as it expands and contracts) with the actual demand, thereby **increasing overall utilization and reducing cost**.

Scale-up approach and scale-out approach

- **Scale-up approach and The traditional scale-out approach,** both approaches have initial start-up costs and both approaches are reactive in nature.
- Traditional infrastructure generally necessitates predicting the amount of computing resources your application will use over a period of several years. If you under-estimate, your applications will not have the horsepower to handle unexpected traffic, potentially resulting in customer dissatisfaction. If you over-estimate, you're wasting money with superfluous resources

Example: Dynamic resource demands

- **Cloud elasticity** is a cost-effective solution for a business with a dynamic workload.

Example: Streaming Services.

Netflix is dropping a new season of Mindhunter. The notification triggers a significant number of users to get on the service and watch or upload the episodes. Resource-wise, it is an activity spike that requires swift resource allocation.

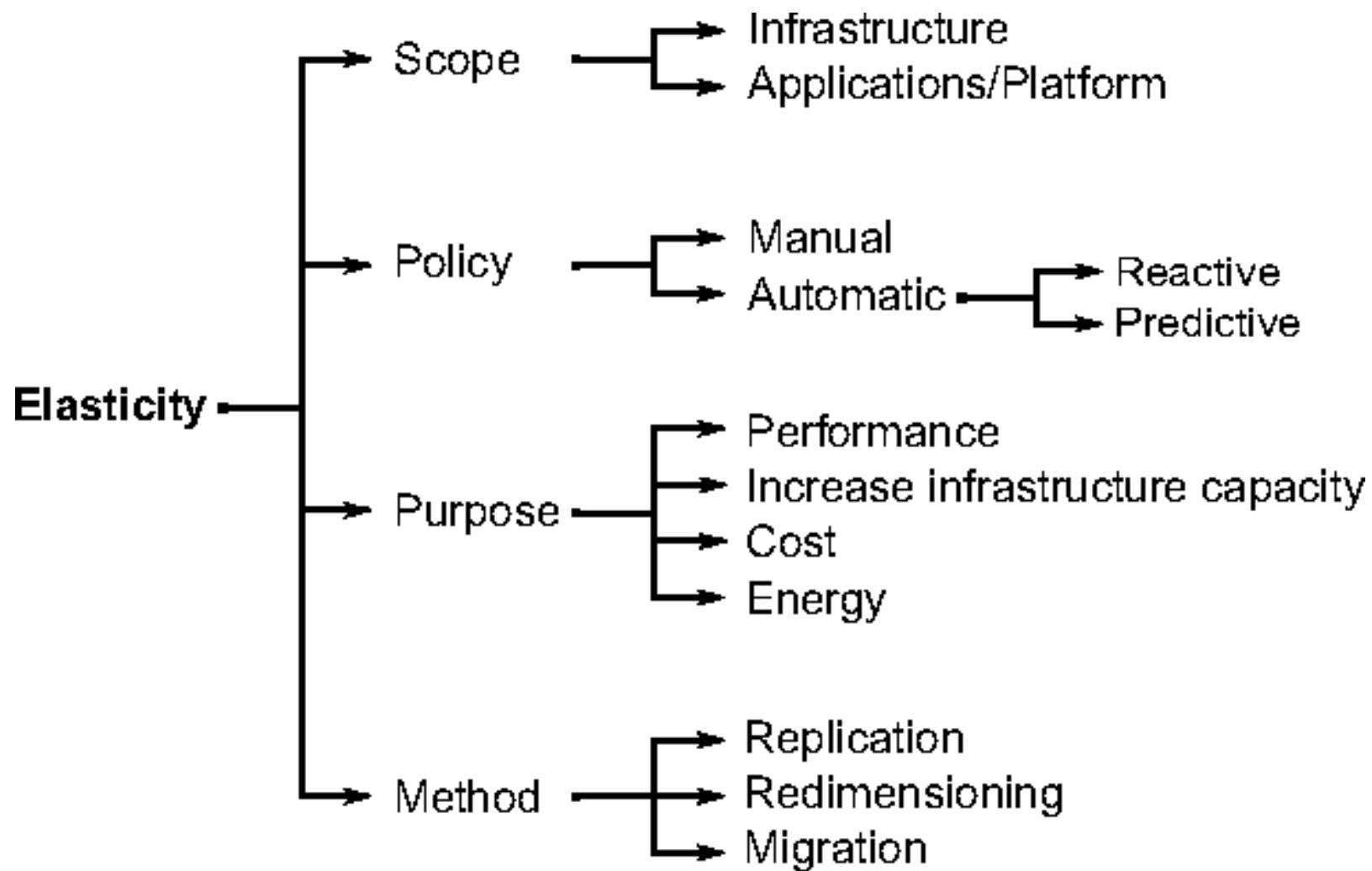
Example: E-commerce.

- Amazon has a Prime Day event with many special offers, sell-offs, promotions, and discounts. It attracts an immense amount of customers on the service who are doing different activities. Actions include searching for products, bidding, buying stuff, writing reviews, rating products. This diverse activity requires a very flexible system that can allocate resources to one sector without dragging down others.

Design Challenge

- Designing intelligent elastic cloud architectures, so that infrastructure runs only when you need it, is an art in itself.
- Elasticity should be one of the architectural design requirements or a system property
- What components or layers in your application architecture can become elastic? What will it take to make that component *elastic*? What will be the impact of implementing elasticity to my overall system architecture?

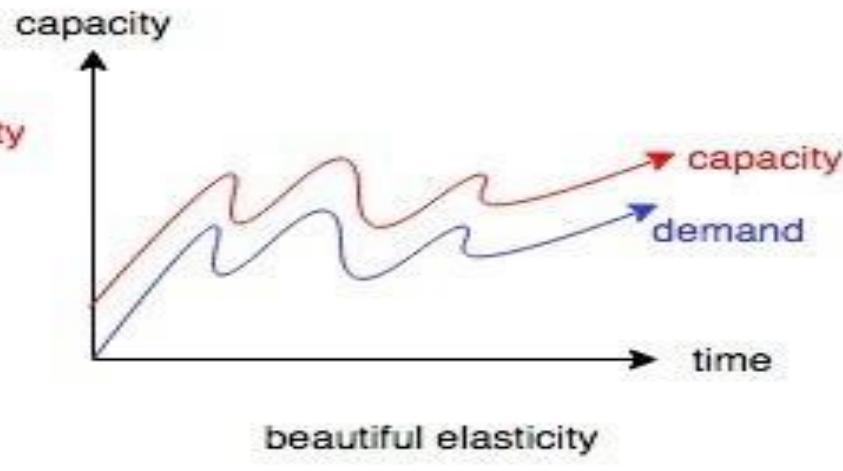
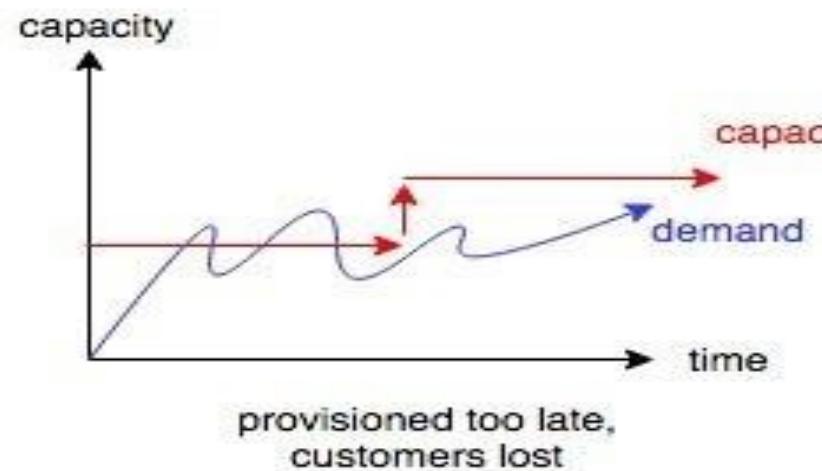
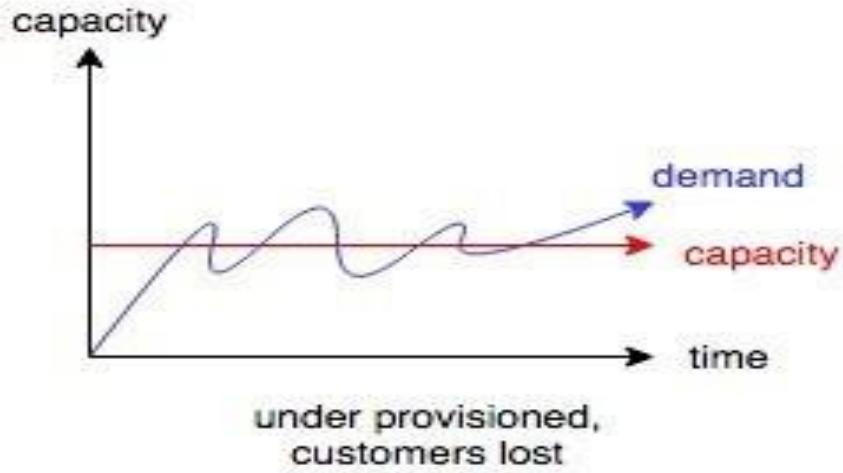
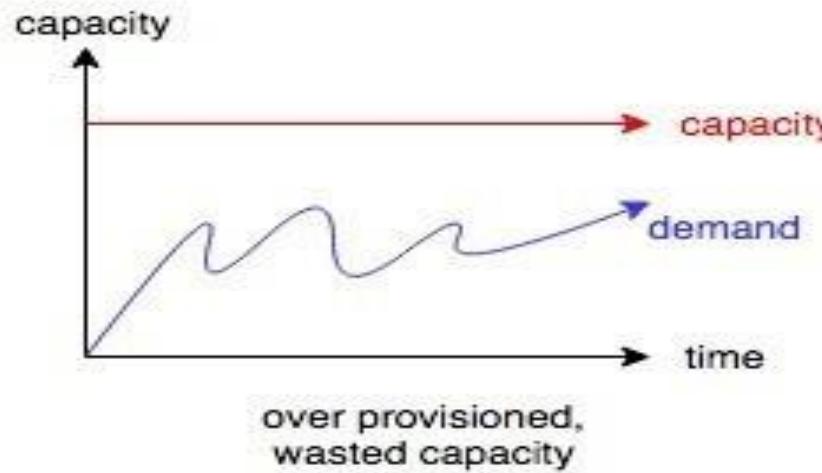
NOTE: To effectively leverage the cloud benefits, it is important to architect with this mindset.



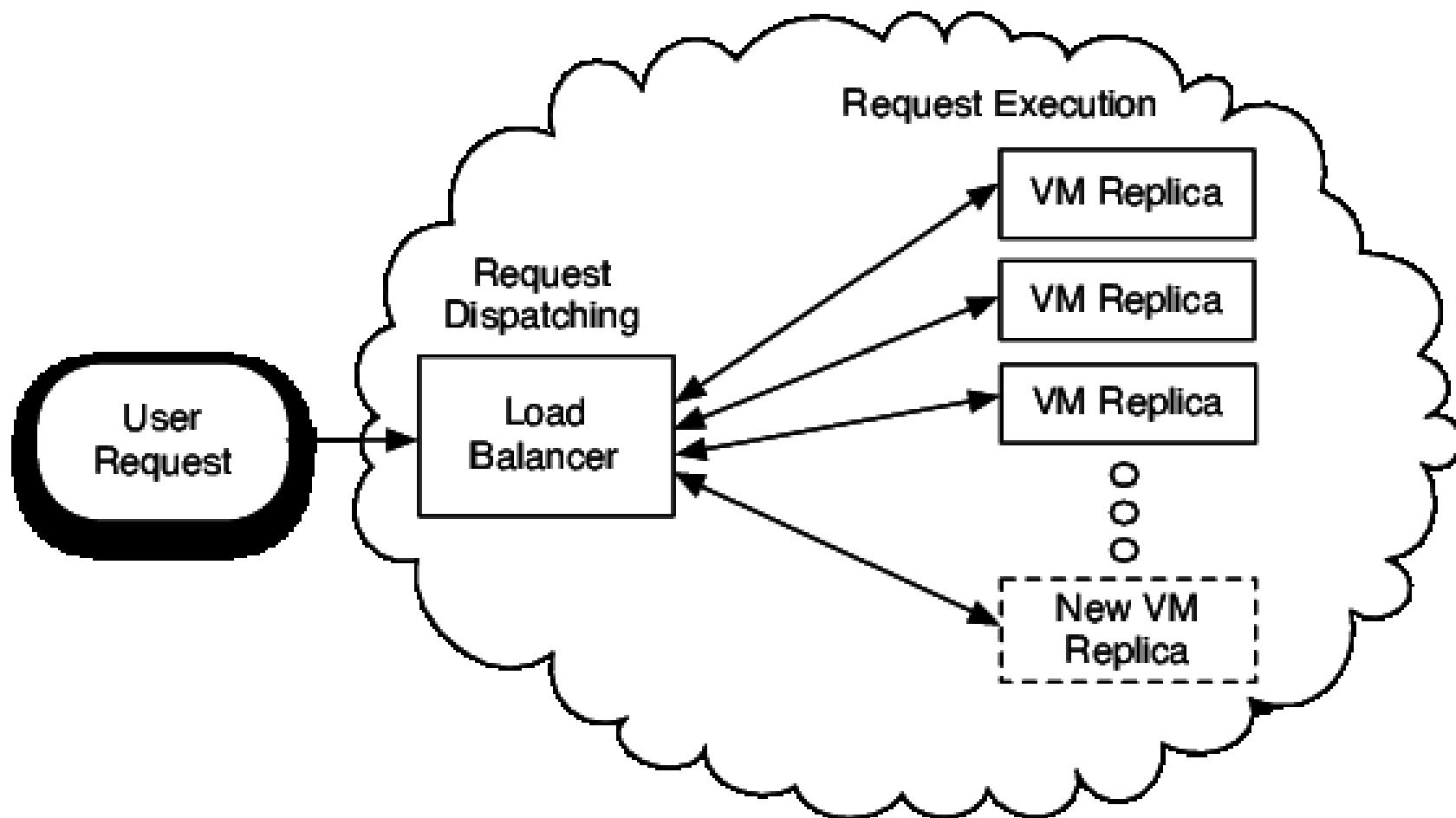
Scalability vs Elasticity

Scalability	Elasticity
“Increasing” the capacity to meet the “increasing” workload	“Increasing or reducing” the capacity to meet the “Increasing or reducing” workload
In a scaling environment, the available resources may exceed to meet the “future demands”	In the elasticity environment, the available resources matches the “current demands” as closely as possible
Scalability adapts only to the “workload increase” by “provisioning” the resources in an “incremental manner”	Elasticity adopts to both the “workload increase” as well as “workload decrease” by provisioning and “deprovisioning” resources in an “automatic” manner
Scalability enables a corporate to meet expected demands for services with “long-term” “strategic needs”	Elasticity enables a corporate to meet unexpected changes in the demand for services with “short-term”, tactical needs

Under and over provisioning vs perfect elasticity of resources



VM replication on demand



Conclusion: cloud scalability and cloud elasticity

- Modern business operations live on consistent performance and instant service availability. Cloud scalability and cloud elasticity handle these two business aspects in equal measure.
- Cloud scalability is an effective solution for businesses whose workload requirements are increasing slowly and predictably.
- Cloud elasticity is a cost-effective solution for the business with dynamic and unpredictable resource demands.
- These features make both scalability and elasticity a viable instrument for the company to hold its ground, grow steadily, and gain a competitive advantage.

Exercises

1. What is cloud computing? What are the benefits of cloud computing?
2. What is a cloud? What are the different data types used in cloud computing? Which are the different layers that define cloud architecture?
3. List the top use Cases for Cloud Computing ?
4. What are the different layers in cloud computing? Explain working of them with appropriate use cases.
5. What is on-demand functionality? How is it provided in cloud computing?
6. What are the different models for deployment in cloud computing? Explain them with example?
7. What is the difference between scalability and elasticity?
8. What are the open source cloud computing platform databases? Give some example of large cloud provider and databases?
9. What is the difference between cloud and traditional datacenters?
10. What are the different datacenters in cloud computing?
11. What are the most essential things that must be followed before going for cloud computing platform?

Cloud Resource Virtualization

ARADHANA BEHURA

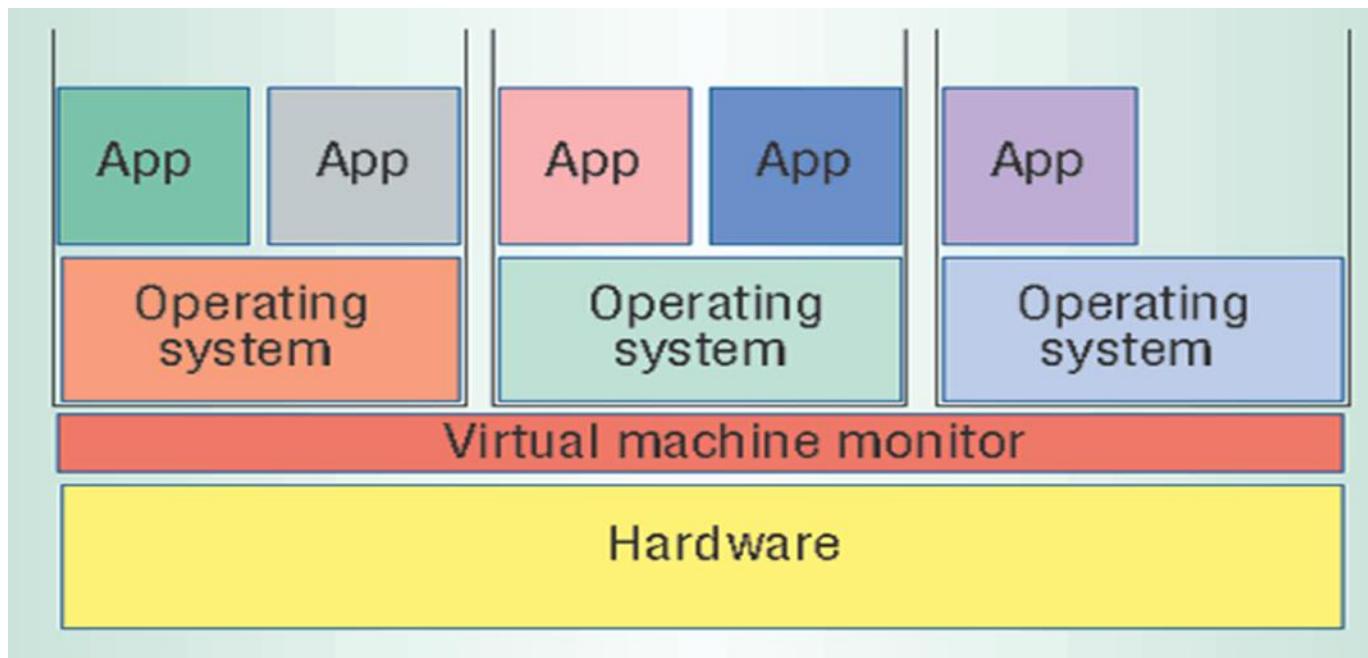
Communication & Computing Group

Department of CSE

Email: 921CS5007@nitrkl.ac.in, 7787821733

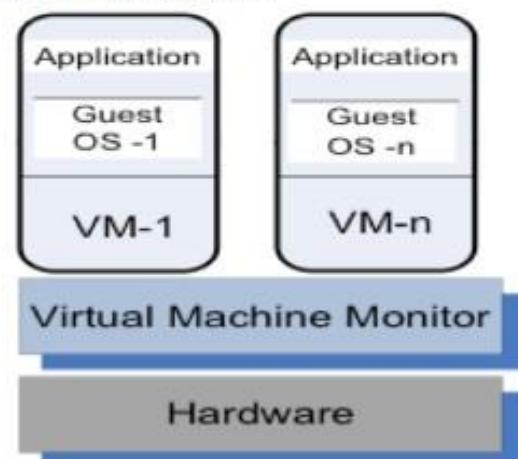
Virtualization in Cloud Computing

- **Virtualization** is the creation of virtual servers, infrastructures, devices and **computing** resources.
- **Virtualization** changes the **hardware-software relations** and is one of the foundational elements of **cloud computing** technology that helps utilize the capabilities of **cloud computing** to the full.

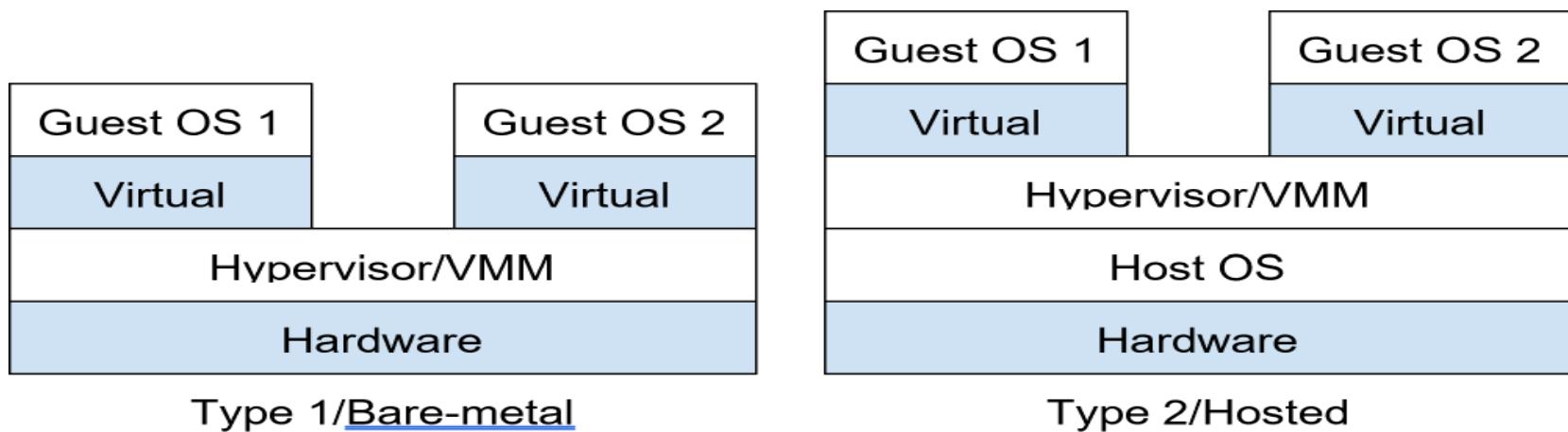


Virtual Machine Monitor (VMM / Hypervisor)

- A **virtual machine monitor (VMM/hypervisor)** partitions the resources of computer system into one or more **virtual machines (VMs)**. Allows several operating systems to run concurrently on a single hardware platform
- A VM is an execution environment that runs an OS
- VM – an isolated environment that appears to be a whole computer, but actually only has access to a portion of the computer resources
- A VMM allows:
 - Multiple services to share the same platform
 - Live migration - the movement of a server from one platform to another
 - System modification while maintaining backward compatibility with the original system
 - Enforces isolation among the systems, thus security
- A **guest operating system** is an OS that runs in a VM under the control of the VMM.



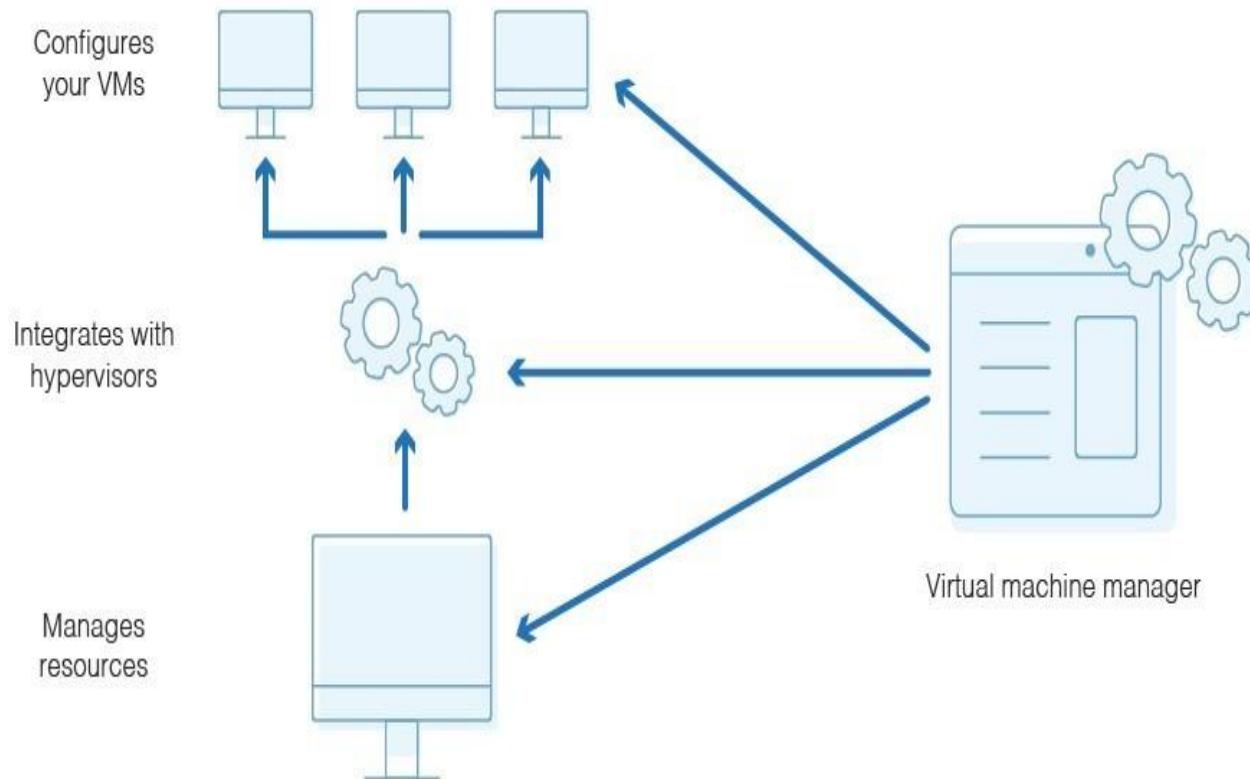
Hypervisors



CONTD.

- A bare-metal **hypervisor (Type 1)** is a layer of software we install directly on top of a physical server and its underlying hardware.
- There is no software or any operating system in between, hence the name bare-metal **hypervisor**. **Type 1 hypervisors** are mainly found in enterprise environments
- A **Type 2 hypervisor**, also called a hosted **hypervisor**, is a virtual machine manager that is installed as a software application on an existing operating system (OS).
- **Type-2 hypervisors** abstract guest operating systems from the host operating system.
- Parallels Desktop for Mac, QEMU, VirtualBox, **VMware** Player and **VMware** Workstation are examples of **type-2 hypervisors**.

What Does a Virtual Machine Manager Do?



Cloud Virtualization

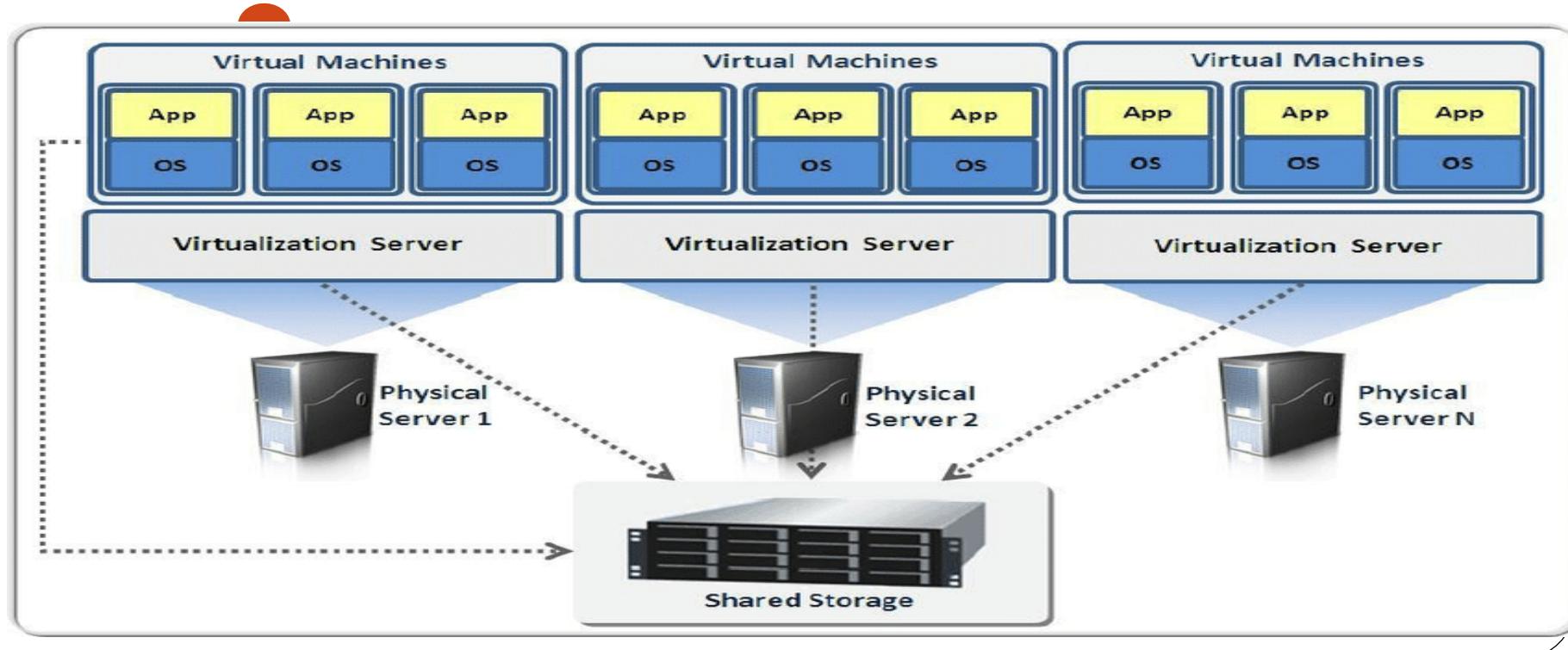
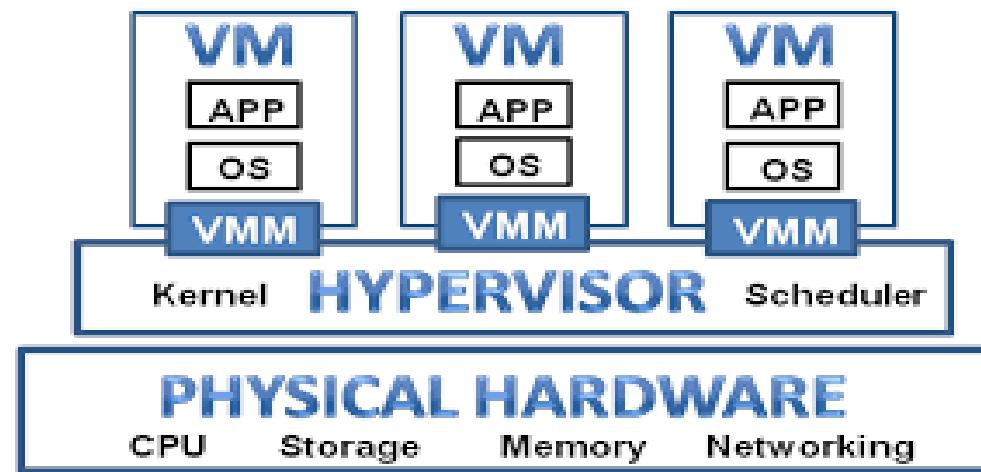
Chapter 3

Virtualization is the creation of a virtual (rather than actual) version of something, such as an operating system, a server, a storage device or network resources.

Virtualization

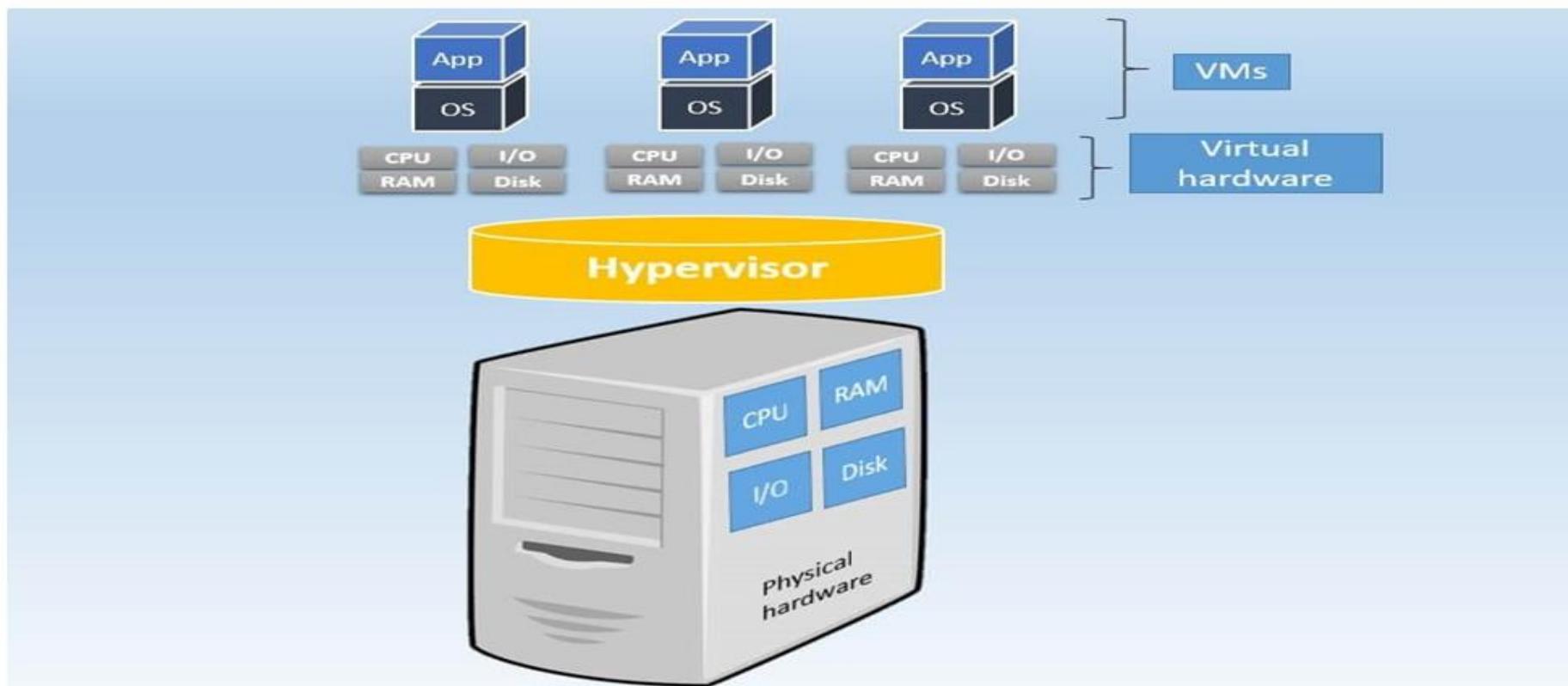
- In computing, virtualization means to create a virtual version of a device or resource, such as a server, storage device, **network** or even an operating system where the framework divides the resource into one or more execution environments.
- Even something as simple as partitioning a hard drive is considered virtualization because you take one drive and partition it to create two separate hard drives.
- Devices, applications and human users are able to interact with the virtual resource as if it were a real single logical resource.

Virtual machine environment



Virtual Machine (VM)

- A **Virtual Machine** (VM) is a compute resource that uses software instead of a physical computer to run programs and deploy applications



The two types of virtual machines

- [1] **A process virtual machine** allows a single process to run as an application on a host machine, providing a platform-independent programming environment by masking the information of the underlying hardware or operating system.
- An example of a process VM is the Java Virtual Machine, which enables any operating system to run Java applications as if they were native to that system.
- [2] **A system virtual machine** is fully virtualized to substitute for a physical machine.
- A system platform supports the sharing of a host computer's physical resources between multiple virtual machines, each running its own copy of the operating system. This virtualization process relies on a hypervisor, which can run on bare hardware, such as VMware ESXi, or on top of an operating system.

Virtualization : An Introduction

- Virtualization can be viewed as part of an overall trend in enterprise IT that includes **autonomic computing**, a scenario in which the IT environment will be able to manage itself based on perceived activity, and **utility computing**, in which computer processing power is seen as a utility that clients can pay for only as needed.
- The usual goal of virtualization is to centralize administrative tasks while improving **scalability** and work loads.
- Virtualization is a basic tenet of cloud computing, it simplifies some of the resource management tasks; for example, the **state of a virtual machine (VM) running under a virtual machine monitor (VMM)** can be saved and migrated to another server to balance the load.

Virtualization : An Introduction

- Resource sharing in a virtual machine environment requires ample hardware support such as powerful processors, and architectural support for multilevel control.
- In practice resources, such as CPU cycles, memory, secondary storage, and I/O and communication bandwidth, are shared among several virtual machines; for each virtual machine resources must be shared among **multiple instances of an application**.
- Virtualization abstracts the underlying resources and simplifies their use, isolates users from one another, and supports replication which, in turn, increases the elasticity of the system.

Virtualization : An Introduction

- Traditional processor architectures were conceived for one level of control as they support **two execution modes**, the **kernel** and the **user mode**.
- In a virtualized environment all resources are under the control of a **virtual machine monitor** (VMM) and a second level of control is exercised by the **guest operating system**.
- A two-level scheduling for sharing CPU cycles can be implemented, sharing of resources such as cache, memory, and I/O bandwidth is more intricate.
- The system functions critical for the performance of a virtual machine environment are cache and memory management, handling of privileged instructions, and I/O handling

Virtualization

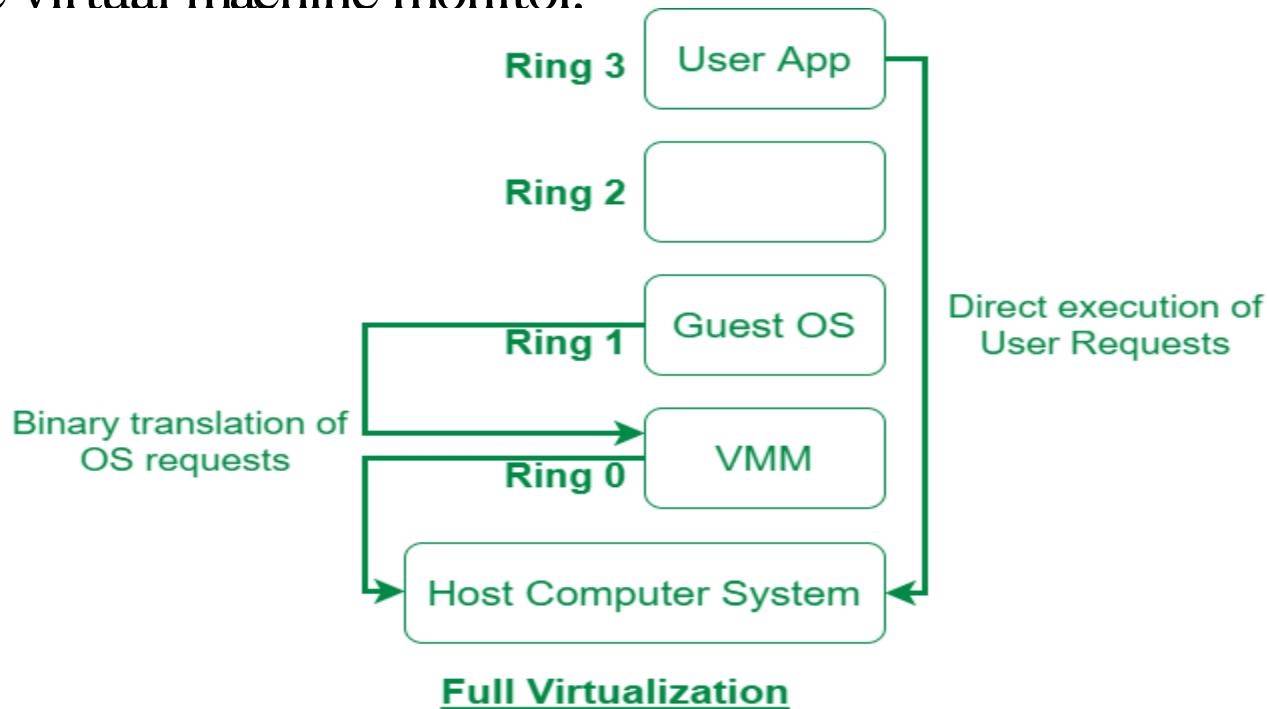
- **Virtualization simulates** the interface to a physical object by any one of four means:
 1. **Multiplexing:** create multiple virtual objects from one instance of a physical object. For example, a processor is multiplexed among a number of processes or threads.
 2. **Aggregation:** create one **virtual object** from multiple physical objects. For example, a number of physical disks are aggregated into a RAID disk.
 3. **Emulation:** construct a virtual object from a different type of a physical object. Example, a physical disk emulates a RandomAccess Memory.
 4. **Multiplexing and emulation.**

Examples: virtual memory with paging multiplexes real memory and disk and a virtual address emulates a real address;

The TCP protocol emulates a reliable bit pipe and multiplexes a physical communication channel and a processor.

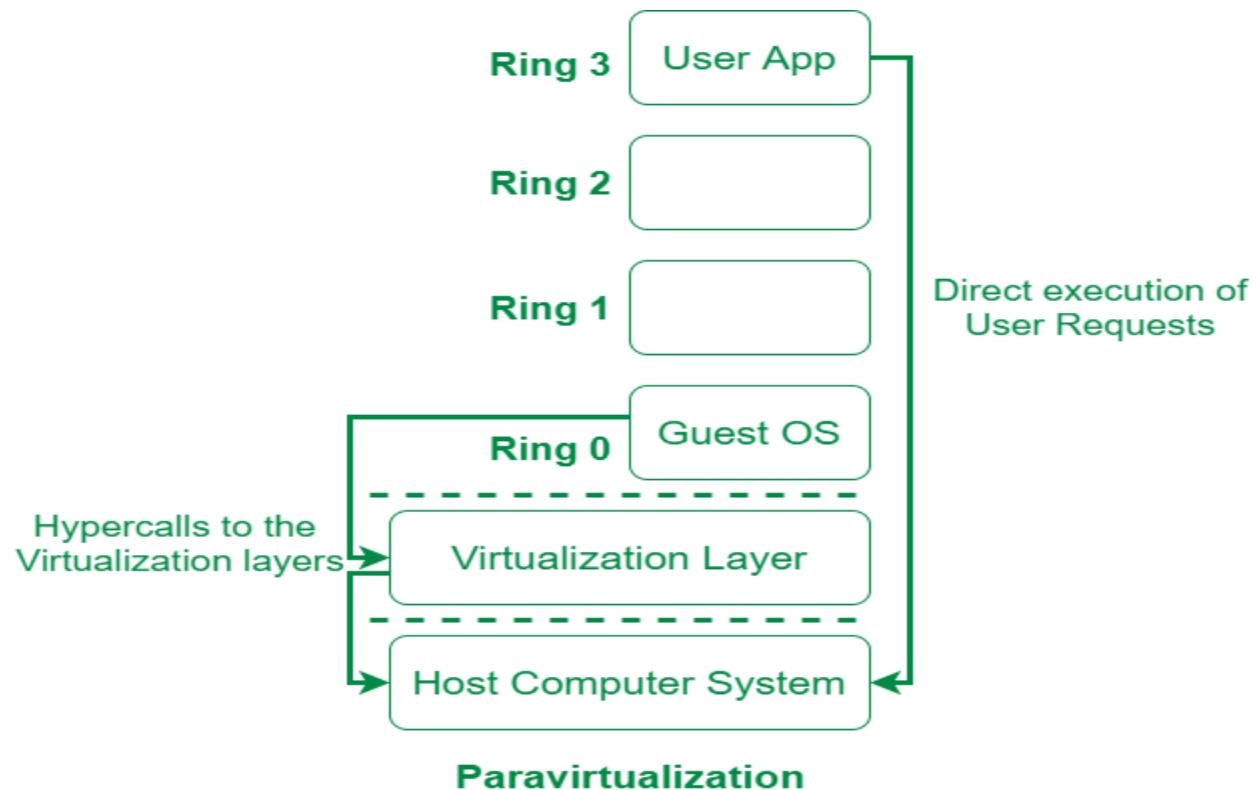
Two distinct approaches for virtualization

- **Full virtualization:** Full virtualization is feasible when the hardware abstraction provided by the virtual machine monitor is an exact replica of the physical hardware; in this case any operating system running on the hardware will run without modifications under the virtual machine monitor.



Two distinct approaches for virtualization

- **Para-virtualization:** Para-virtualization require some modifications of the guest operating systems, as the hardware abstraction provided by the VMM does not support all the functions the hardware does.



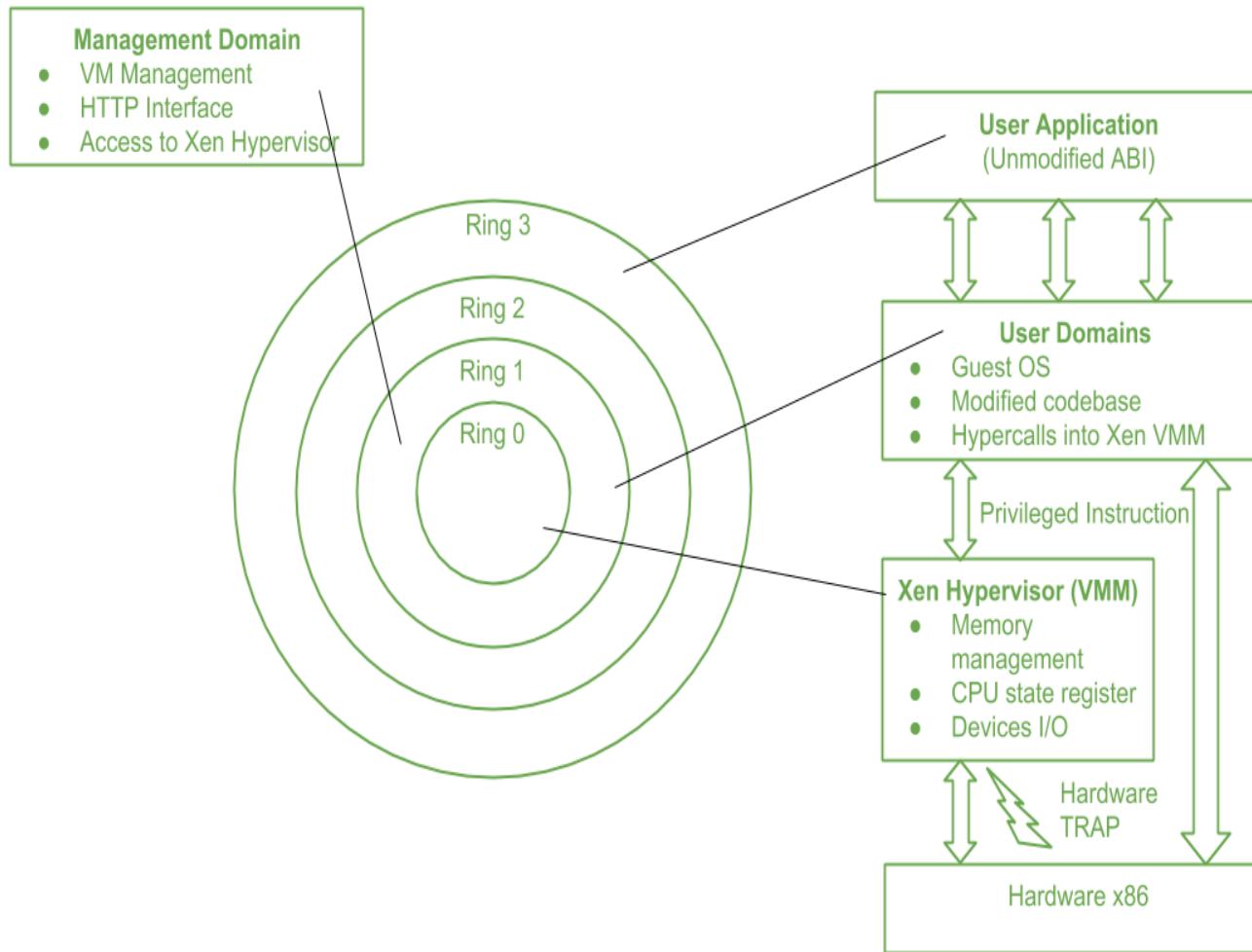
Features	Full Virtualization	ParaVirtualization
Definition	It is the first generation of software solutions for server virtualization.	The interaction of the guest operating system with the hypervisor to improve performance and productivity is known as paravirtualization.
Security	It is less secure than para-virtualization.	It is more secure than full virtualization.
Performance	Its performance is slow than para-virtualization.	Its performance is high than full virtualization.
Guest OS Modification	It supports all the Guest OS without any change.	The Guest OS has to be modified in para-virtualization, and only a few OS support it.
Guest OS hypervisor independent	It enables the Guest OS to run independently.	It enables the Guest OS to interact with the hypervisor.
Portable and Compatible	It is more portable and compatible.	It is less portable and compatible.
Isolation	It offers optimum isolation.	It offers less isolation.
Efficient	It is less efficient than paravirtualization.	It is more simplified than full virtualization.
Characteristic	It is software based.	It is cooperative virtualization.
Examples	It is used in Microsoft, VMware, ^{ESXi} and Parallels systems.	It is mainly used in VMware and Xen systems.

Types

- Full virtualization enables the Guest operating system to run independently. In contrast, paravirtualization enables the Guest OS to interact with the hypervisor.
- Binary translation and a direct approach are used in full virtualization. On the other hand, paravirtualization operates through hypercalls.
- **Full Virtualization:** It is the first software solution for server virtualization and uses binary translation and direct approach techniques. In full virtualization, guest OS is completely isolated by the virtual machine from the virtualization layer and hardware.
- **Paravirtualization:** Paravirtualization is the category of CPU virtualization which uses hypercalls for operations to handle instructions at compile time. In paravirtualization, guest OS is not completely isolated but it is partially isolated by the virtual machine from the virtualization layer and hardware.

Virtualization | Xen: Paravirtualization

- **Xen** is an open source hypervisor based on paravirtualization. It is the most popular application of paravirtualization.
- Xen has been extended to compatible with full virtualization using hardware-assisted virtualization. It enables high performance to execute guest operating system.
- This is probably done by removing the performance loss while executing the instructions requiring significant handling and by modifying portion of the guest operating system executed by Xen, with reference to the execution of such instructions.
- Hence this especially support x86, which is the most used architecture on commodity machines and servers.



- Above figure describes the Xen Architecture and its mapping onto a classic x86 privilege model.
- A Xen based system is handled by Xen hypervisor, which is executed in the most privileged mode and maintains the access of guest operating system to the basic hardware. Guest operating system are run between domains, which represents virtual machine instances.
- In addition, particular control software, which has privileged access to the host and handles all other guest OS, runs in a special domain called Domain 0.
- This the only one loaded once the virtual machine manager has fully booted, and hosts an HTTP server that delivers requests for virtual machine creation, configuration, and termination.
- This component establishes the primary version of a shared virtual machine manager (VMM), which is a necessary part of Cloud computing system delivering Infrastructure-as-a-Service (IaaS) solution.

Various x86 implementation support four distinct security levels, termed as rings, i.e.,

- Ring 0,
- Ring 1,
- Ring 2,
- Ring 3

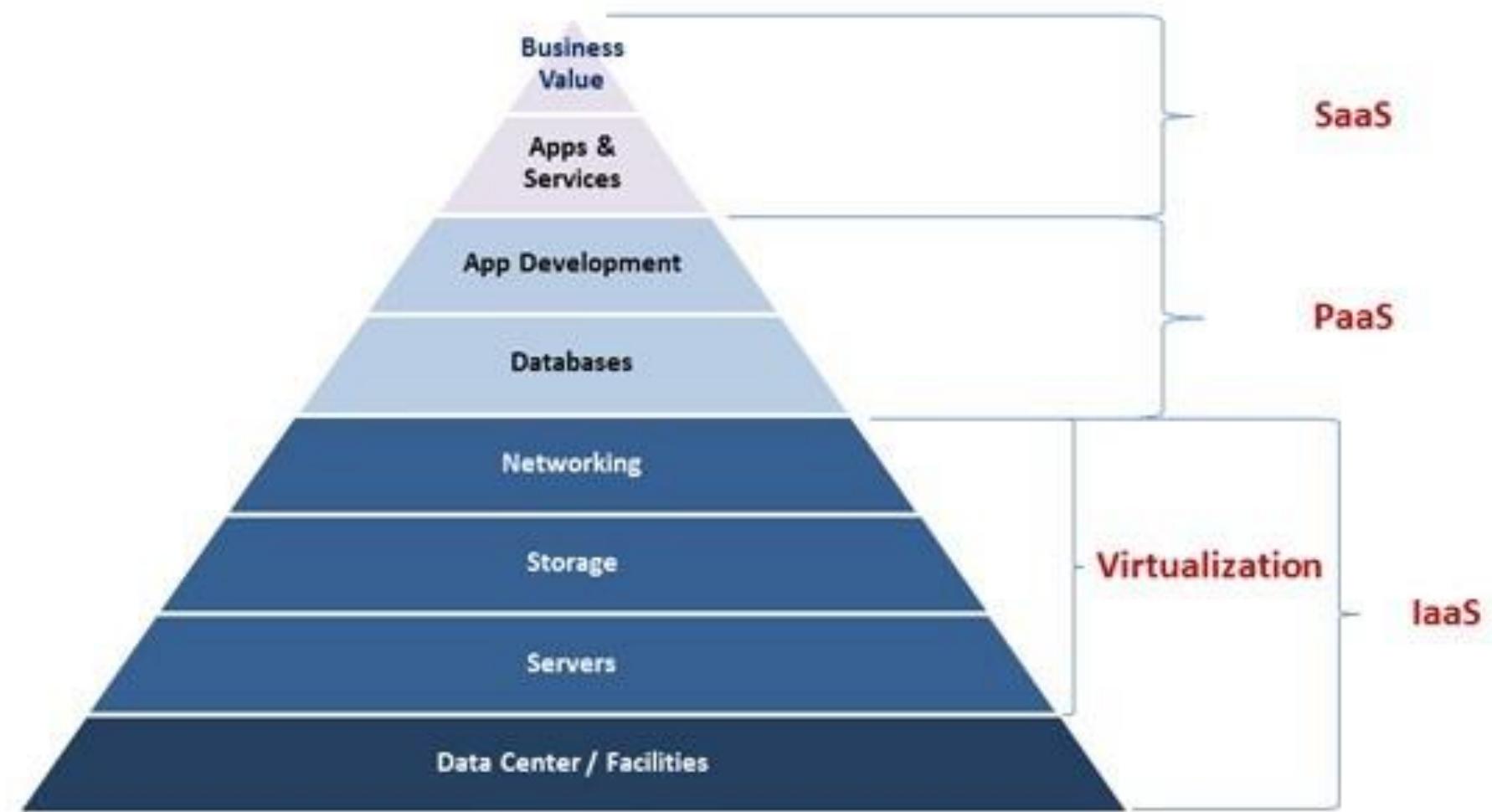
Pros:

- a) Xen server is developed over open-source Xen hypervisor and it uses a combination of hardware-based virtualization and paravirtualization. This tightly coupled collaboration between the operating system and virtualized platform enables the system to develop lighter and flexible hypervisor that delivers their functionalities in an optimized manner.
- b) Xen supports balancing of large workload efficiently that capture CPU, Memory, disk input-output and network input-output of data. It offers two modes to handle this workload: Performance enhancement, and For handling data density.
- c) It also comes equipped with a special storage feature that we call Citrix storage link. Which allows a system administrator to uses the features of arrays from Giant companies- Hp, Netapp, Dell Equal logic etc.
- d) It also supports multiple processor, Live migration one machine to another, physical server to virtual machine or virtual server to virtual machine conversion tools, centralized multiserver management, real time performance monitoring over window and linux.

CONS:

- e) Xen relies on 3rd-party component to manage the resources like drivers, storage, backup, recovery & fault tolerance.
- f) Xen sometimes may cause increase in load on your resources by high input-output rate and and may cause starvation of other Vm's.

Information Technology stack showing technology layers addressed by virtualization and by cloud computing.



Motivation

- Three fundamental abstractions are necessary to describe the operation of a computing systems: (1) interpreters/processors, (2) memory, (3) communications links
- As the scale of a system and the size of its users grows, it becomes very challenging to manage its resources
- **Resource management issues:** (1) provision for peak demands à overprovisioning , (2) heterogeneity of hardware and software , (3) machine failures
- Virtualization is a basic enabler of Cloud Computing, it simplifies the management of physical resources for the three abstractions
- For example, the state of a virtual machine (VM) running under a virtual machine monitor (VMM) can be saved and migrated to another server to balance the load
- § For example, virtualization allows users to operate in environments they are familiar with, rather than forcing them to specific ones

Motivation

- Virtualization abstracts the underlying resources; simplifies their use; isolates users from one another; and supports replication which increases the elasticity of a system.

Cloud resource virtualization is important for:

§ Performance isolation

- as we can dynamically assign and account for resources across different applications

§ System security:

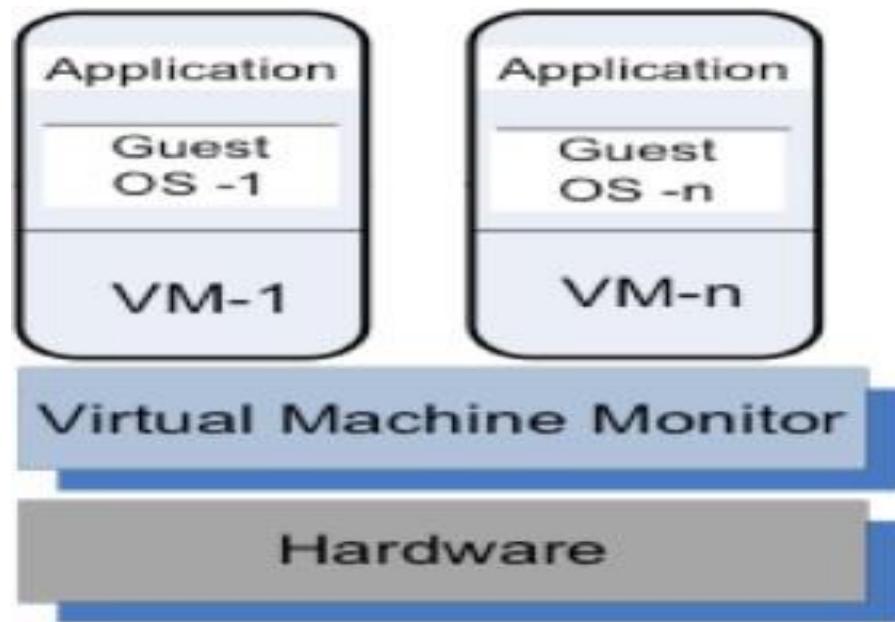
- as it allows isolation of services running on the same hardware

§ Performance and reliability:

- as it allows applications to migrate from one platform to another
- The development and management of services offered by a provider

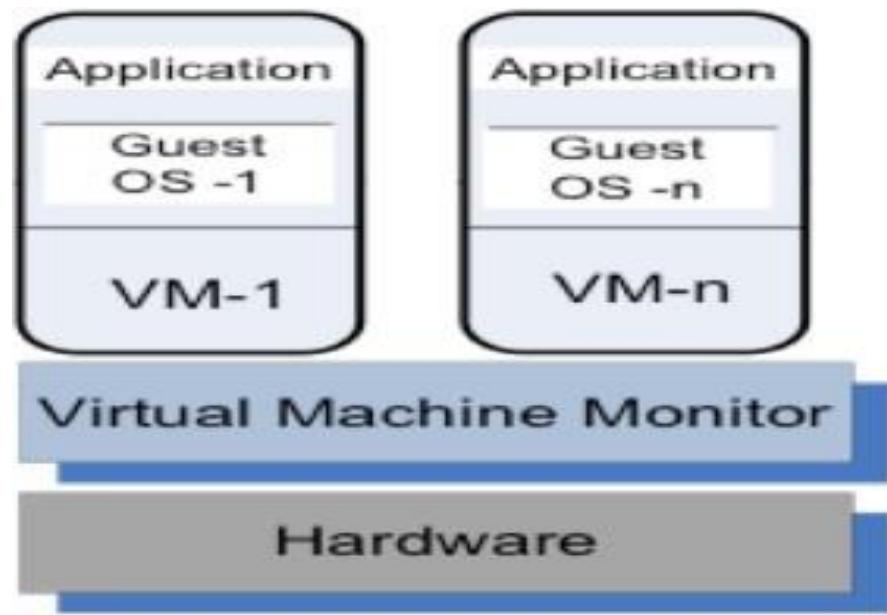
Definitions

- *Virtual Machine (VM)*: *An instance of an operating system running on a virtualized system.* Also known as a *virtual or guest OS*
- *hypervisor*: *The underlying virtualization system sitting between the guest OSes and the hardware.* Also known as a *Virtual Machine Monitor (VMM)*.



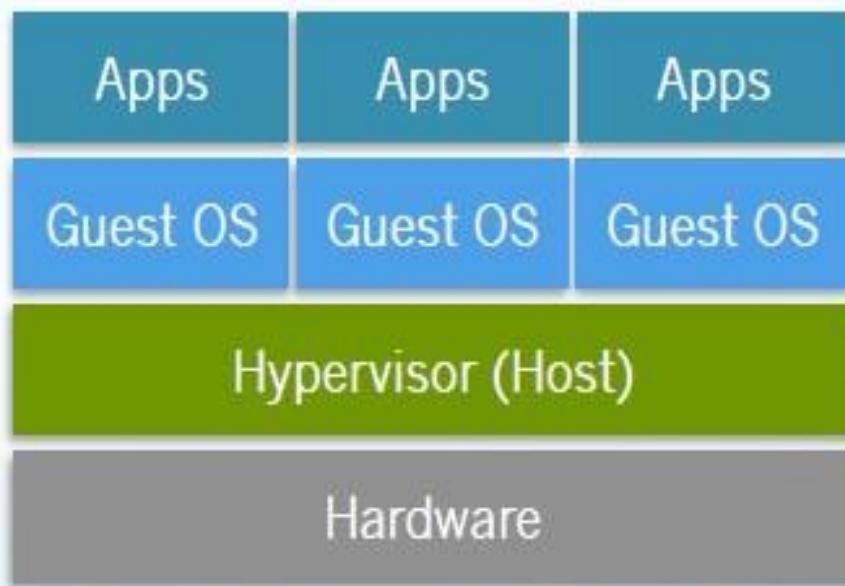
Requirements of a VMM

- Developed by Popek & Goldberg in 1974 :
 1. Provides environment identical to underlying hardware.
 2. Most of the instructions coming from the guest OS are executed by the hardware without being modified by the VMM.
 3. Resource management is handled by the VMM (this all non-CPU hardware such as memory and peripherals).



Guest OS model

- Hypervisor exists as a layer between the operating systems and the hardware.
- Performs memory management and scheduling required to coordinate multiple operating systems.
- May also have a separate controlling interface.



Virtualization Challenges

- **Privileged Instructions**

- Handling architecture-imposed instruction privilege levels.

- **Performance Requirements**

- Holding down the cost of VMM activities.

- **Memory Management**

- Managing multiple address spaces efficiently.

- **I/O Virtualization**

- Handling I/O requests from multiple operating systems.

Virtualization

Virtualization simulates the interface to a physical object by:

- **Multiplexing:** creates multiple virtual objects from one instance of a physical object. Many virtual objects to one physical. **Example** – a processor is multiplexed among a number of processes or threads.
- **Aggregation:** creates one virtual object from multiple physical objects. One virtual object to many physical objects. **Example** – a number of physical disks are aggregated into a RAID disk.
- **Emulation:** constructs a virtual object of a certain type from a different type of a physical object. **Example** - a physical disk emulates a Random Access Memory (RAM).
- **Multiplexing and emulation:** Examples - virtual memory with paging multiplexes real memory and disk; a virtual address emulates a real address.

Layering and virtualization

Section 5.2

Layering and Virtualization

Layering – a common approach to manage system complexity:

- Simplifies the description of the subsystems; each subsystem is abstracted through its interfaces with the other subsystems
- Minimises the interactions among the subsystems of a complex system
- With layering we are able to design, implement, and modify the individual subsystems independently

Layering in a computer system:

- **Hardware**
- **Software**
 - § Operating system
 - § Libraries
 - § Applications

Interfaces

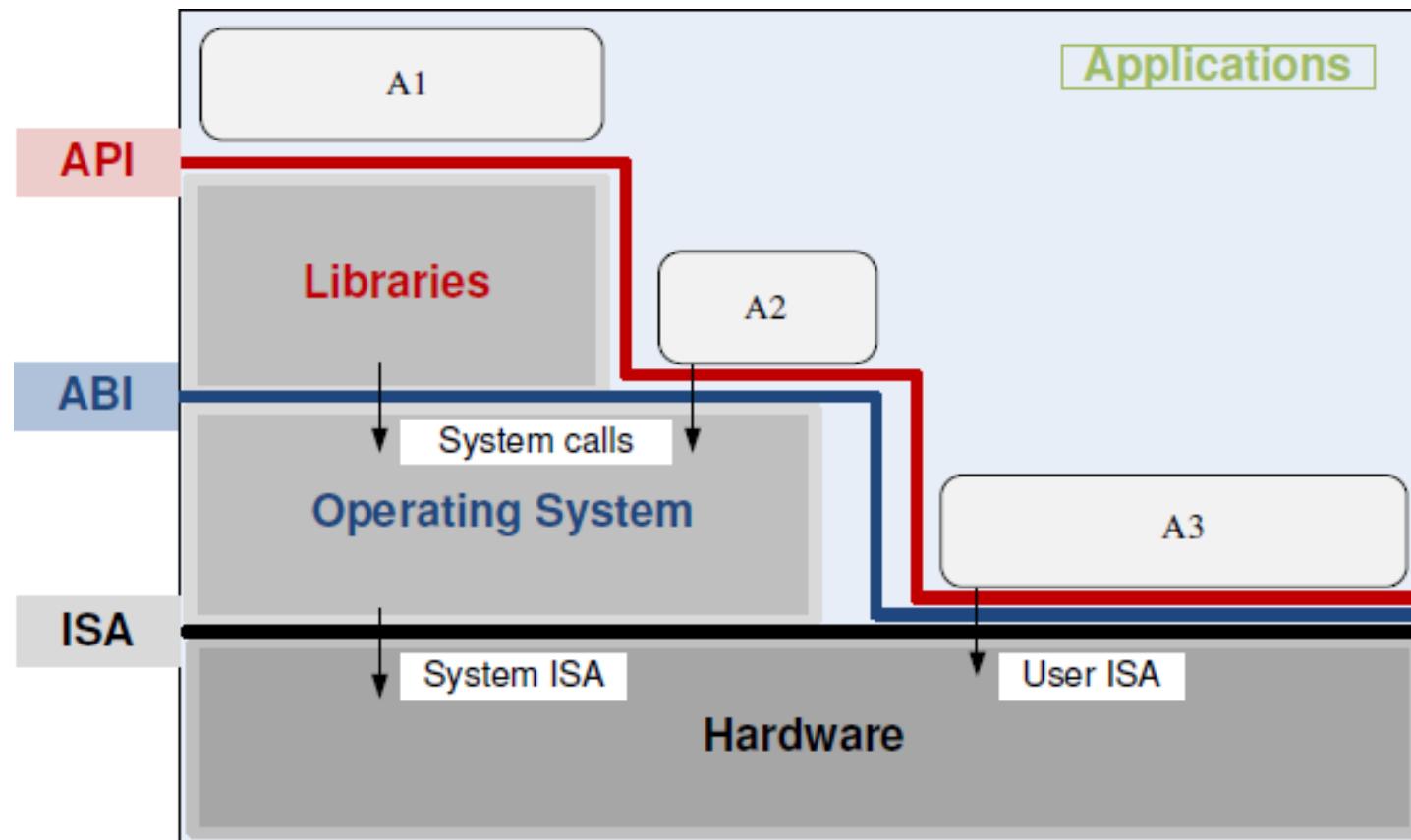
- **Instruction Set Architecture (ISA)** – at the boundary between hardware and software.
- **Application Binary Interface (ABI)** – allows the ensemble consisting of the application and the library modules to access the hardware; the ABI does not include privileged system instructions, instead it invokes system calls.

An application binary interface (ABI) is a set of rules that **dictate how two pieces of software communicate with each other**. It is a low-level interface that defines how software components interact with each other, and it is used to ensure that different software components can work together.

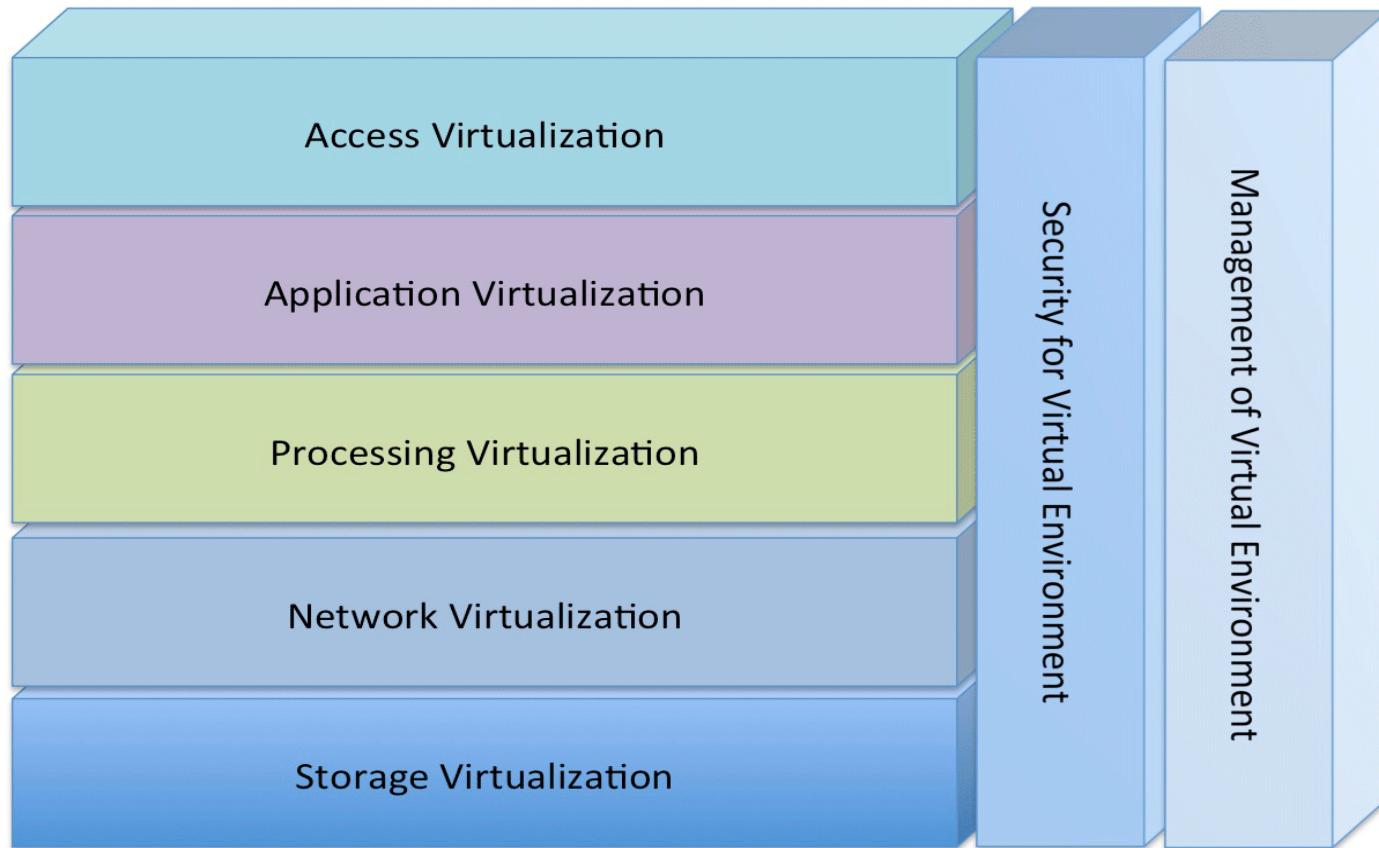
System Call: It provides an interface between user programs and operating systems.

- **Application Program Interface (API)** - defines the set of instructions the hardware was designed to execute and gives the application access to the ISA; it includes HLL library calls which often invoke system calls.

Layering and virtualization



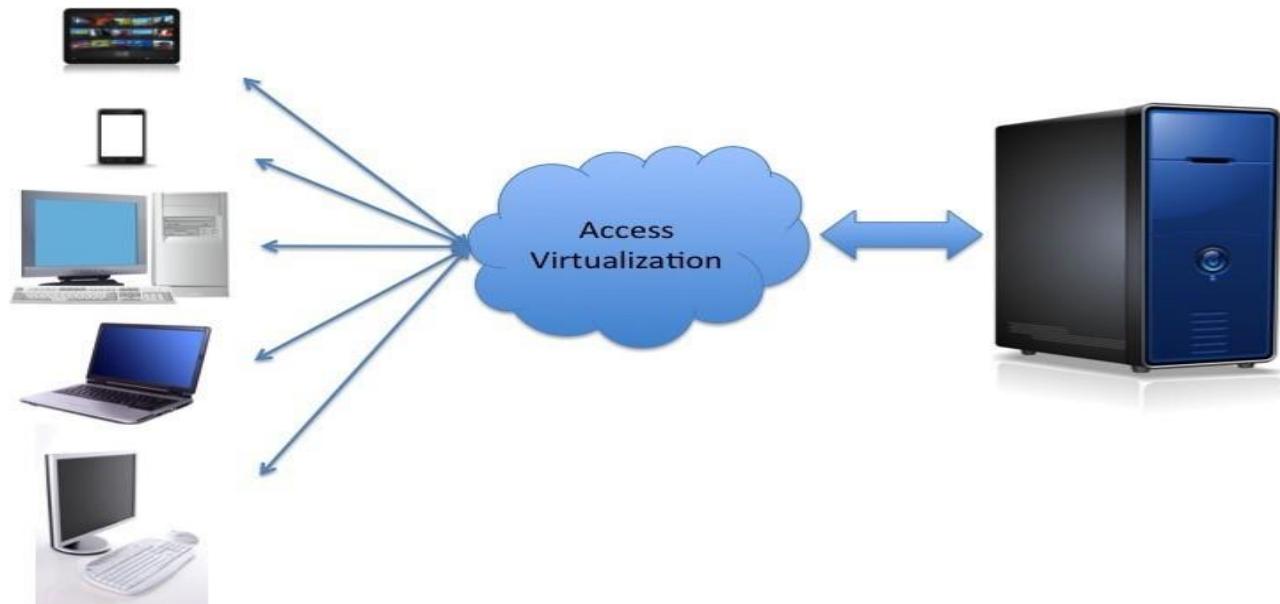
The 7-Layer Virtualization Model



Kusnetzky Group LLC ©2004-2014

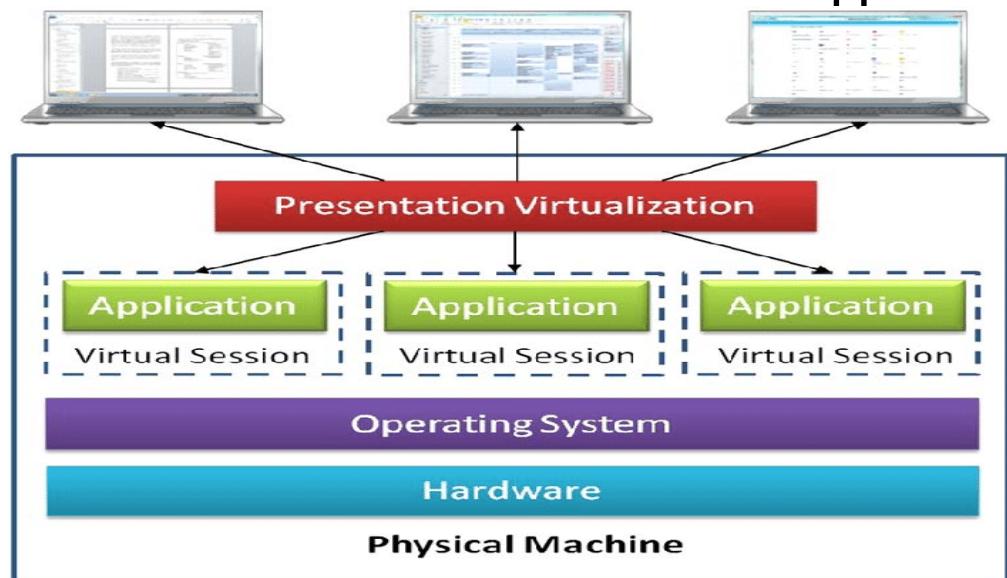
Virtualization Layer 1 : access virtualization

- **Virtualization Layer 1:** Allows applications to work with remote client devices without change, even though those remote devices were never been thought of or available when the application was written. This is called *access virtualization*.
- XenDesktop from Citrix is an example of products that work in this layer of virtualization.



Virtualization Layer 2: application virtualization

- Allows applications written for one OS version or OS to happily execute in another environment; this environment can be a new OS version or an entirely different OS. This is called *application virtualization*.
- This type of software would make it possible for an application written for Windows XP to work just fine on Windows 7 or Windows 8. AppZero fits into this layer of virtualization, as does XenApp from Citrix, App-V from Microsoft and VMware ThinApp



Virtualization Layer 3: processing virtualization

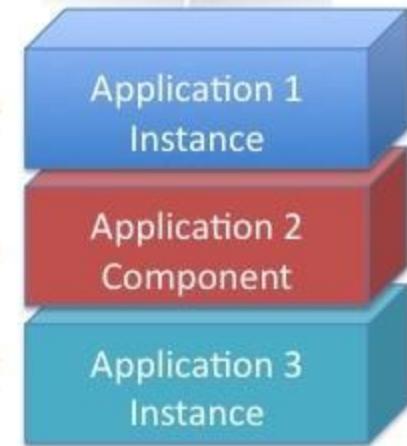
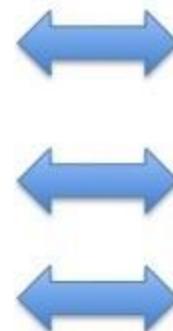
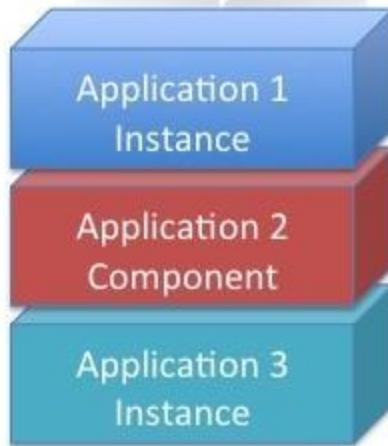
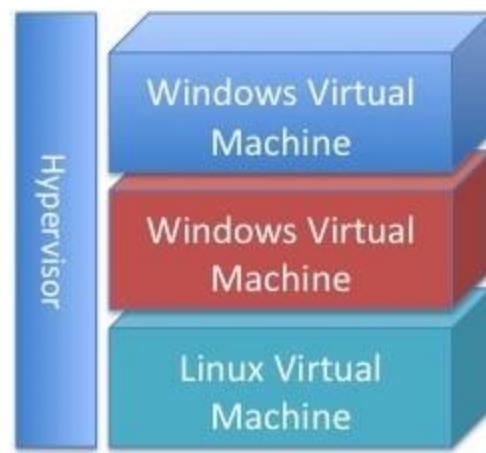
- Allows one system to support workloads as if it was many systems, or allows one workload to run across many systems as if it was a single computing resource. This is called *processing virtualization*.
- VM software is one of five different types of software that live at this layer. One of today's hottest catch phrases, software-defined datacenter (SDDC), is basically the use of this type of software, combined with a couple of other virtualization layers.
- Citrix XenServer, Microsoft Hyper-V and VMware vServer are all examples of VM software that lives in this layer of virtualization. Adaptive Computing Moab and IBM Platform Computing LSF are both examples of cluster managers that also live at this layer of virtualization.

Virtualization Layer 3: processing virtualization

Processing Virtualization:
Making One System Appear to be Many



Processing Virtualization:
Making Many Systems Appear to be One

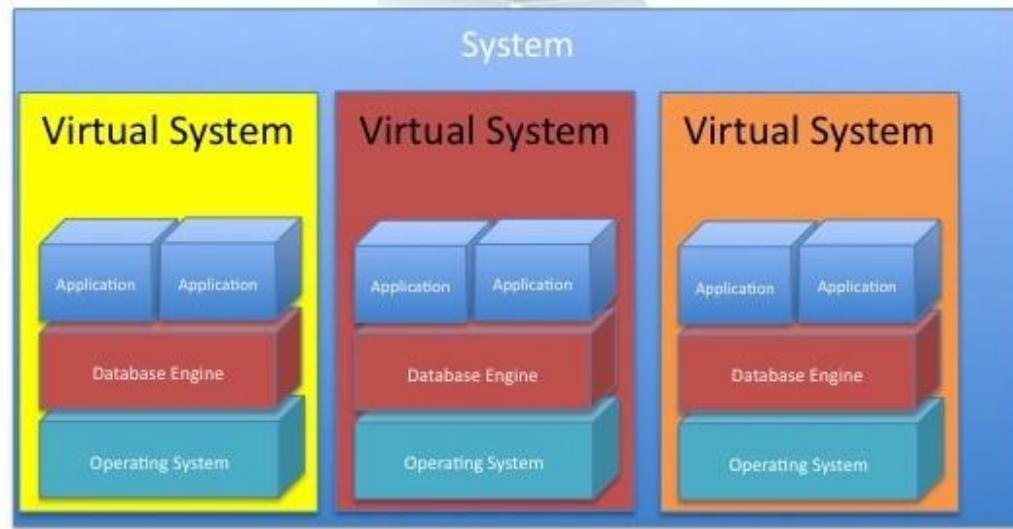
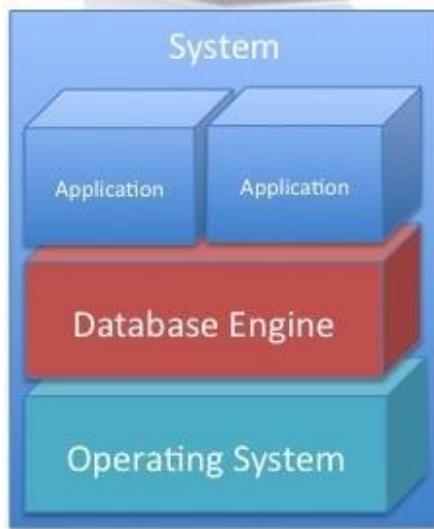


Virtualization Layer 3: processing virtualization

Operating System and Applications on a Physical System



Virtual Systems Running under a Type 1 Hypervisor



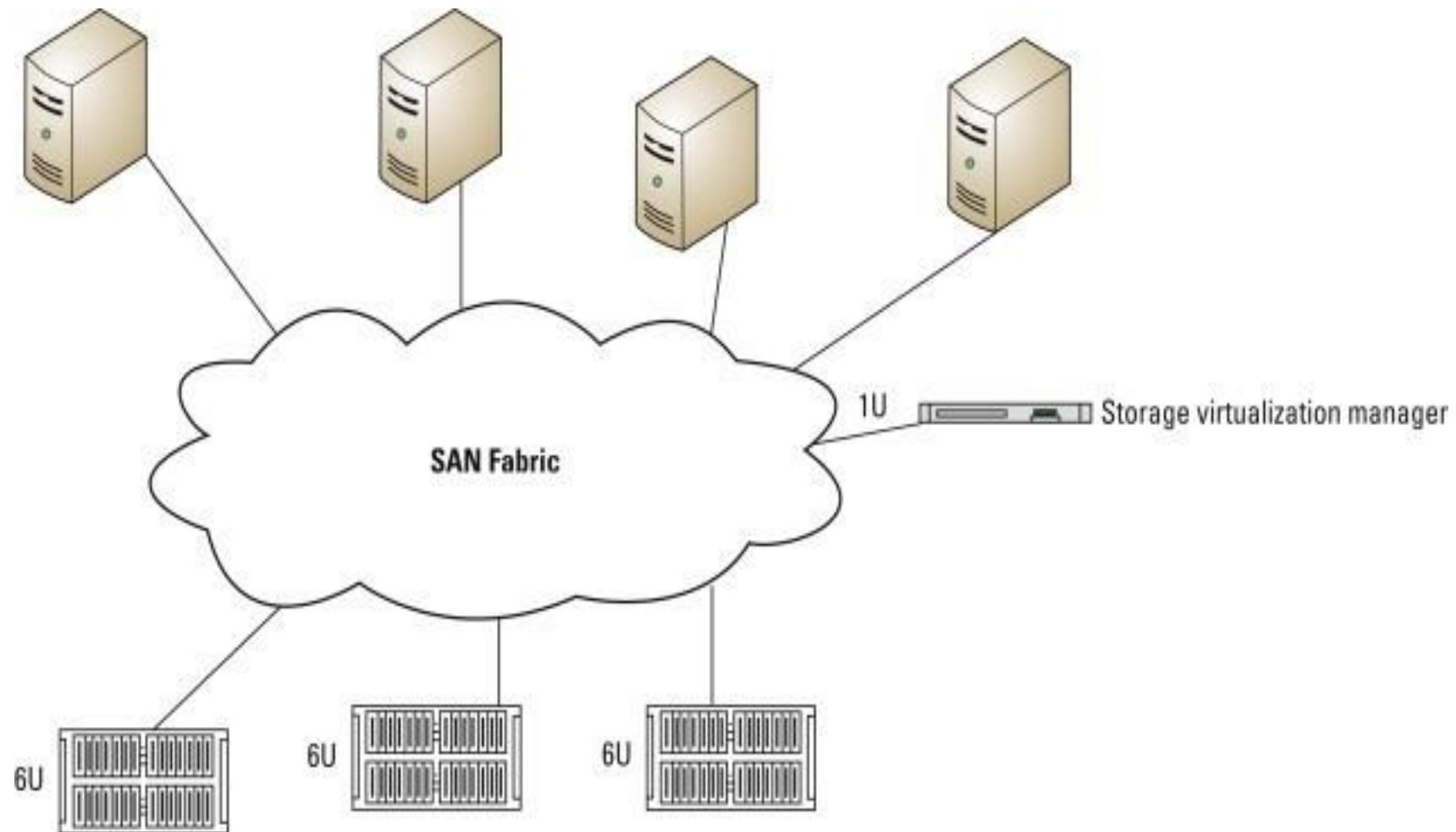
Virtualization Layer 4: storage virtualization

- Allows workloads to access storage without having to know where the data is stored, what type of device is storing the data, or whether the storage is attached directly to the system hosting the workload, to a storage server just down the LAN, or to storage in the cloud. This is called *storage virtualization*.
- Another one of today's most talked-about catch phrases, software-defined storage (SDS), is an example of this technology. Open-E DSS(data storage software), Sanbolic clustered storage, DataCore SANsymphony-V and VMware VSAN are examples of storage virtualization technology

What is vSAN in virtualization?

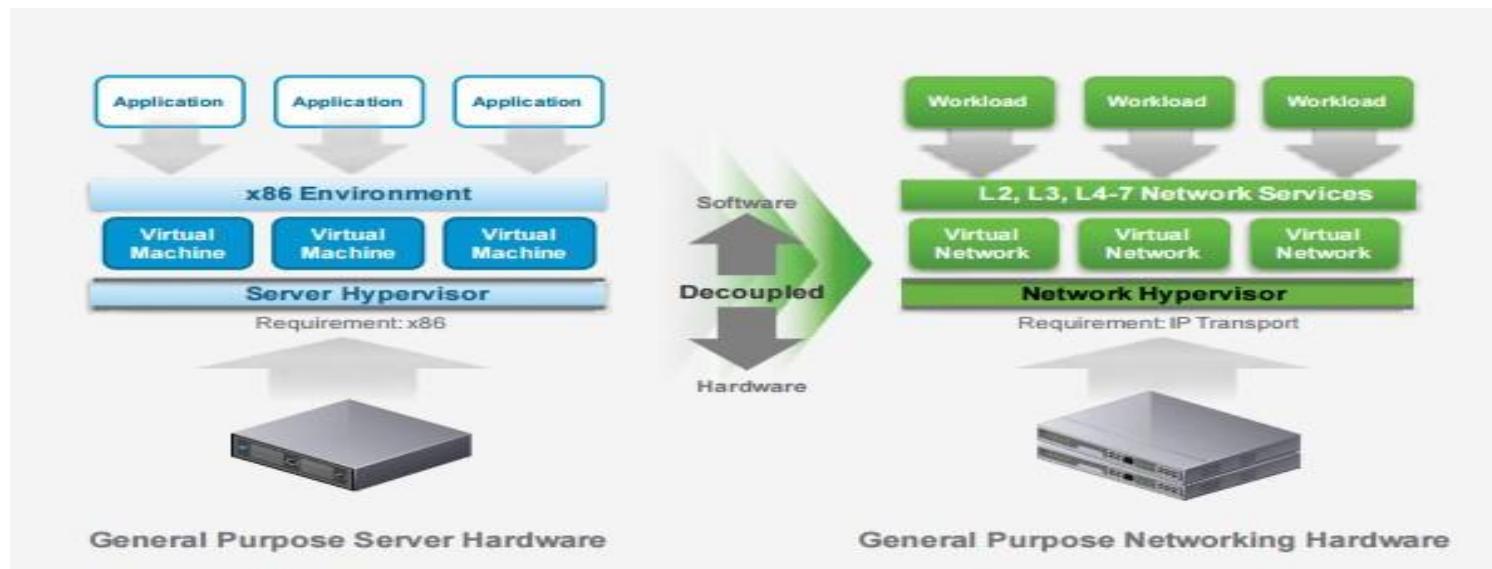
A virtual storage area network (VSAN) is a **logical partition in a physical storage area network (SAN)**. VSANs enable traffic to be isolated within specific portions of a storage area network, so if a problem occurs in one logical partition, it can be handled with a minimum of disruption to the rest of the network

Virtualization Layer 4: storage virtualization



Virtualization Layer 5: network virtualization

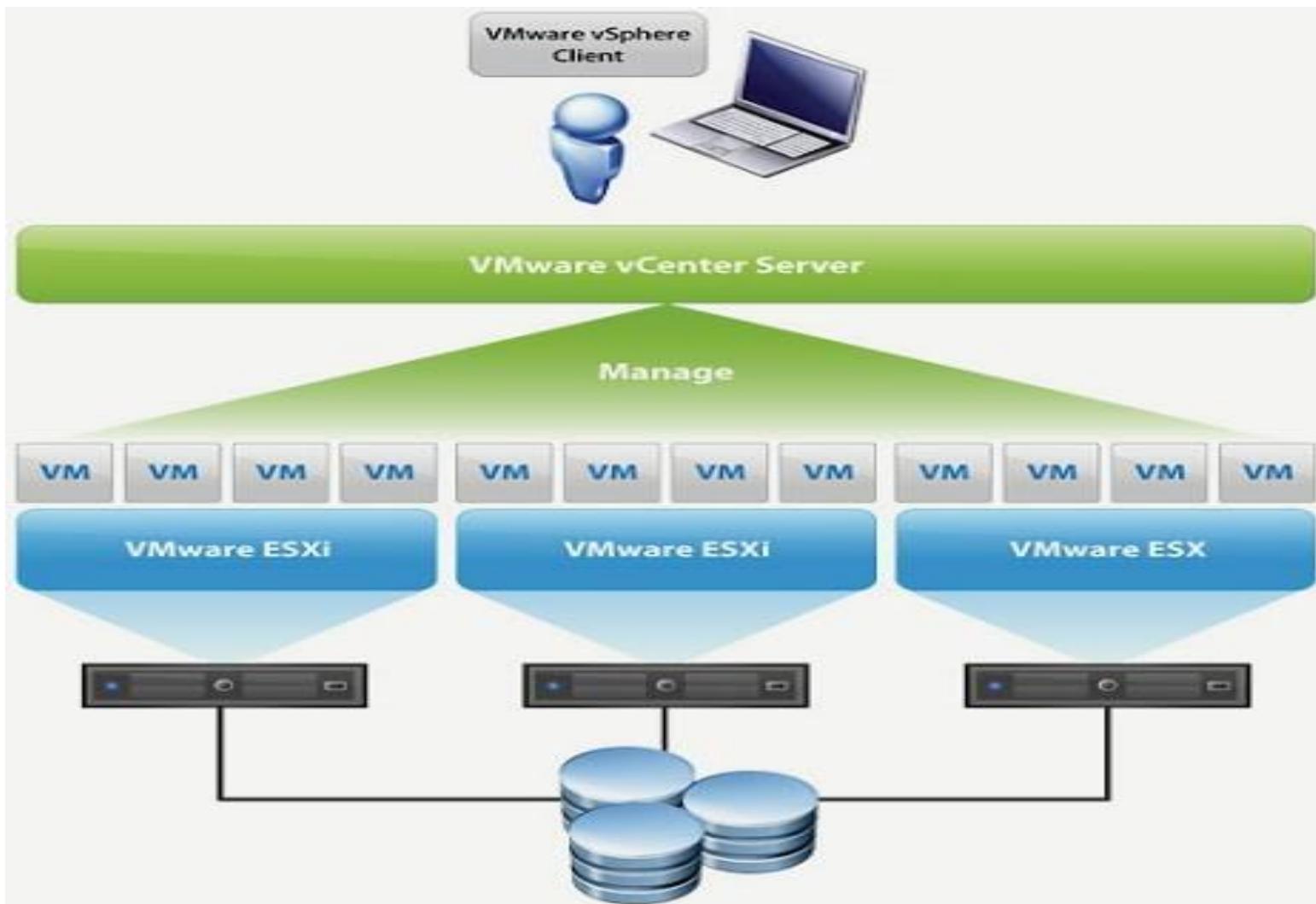
- Allows systems to work with other systems safely and securely, without having to care too much about the details of the underlying network. This is called *network virtualization*.
- Yet another current catchphrase, software-defined networking (SDN), is an implementation of network virtualization.
- Products that offer network virtualization include the Cisco Extensible Network Controller (XNC) and Juniper Contrail



Virtualization Layer 6: management of virtualized environments

- Allows IT administrators and operators to easily monitor and manage virtual environments across boundaries.
- The boundaries can include the physical location of systems; OSes in use; applications or workloads in use; network topology; storage implementation; and how client systems connect to the applications. This is called *management of virtualized environments* in the model.
- This, by the way, is an important part of SDN, SDS and SDDC. A whole host of companies, including AppNeta, BMC, CA, HP and IBM, offer management and monitoring software.

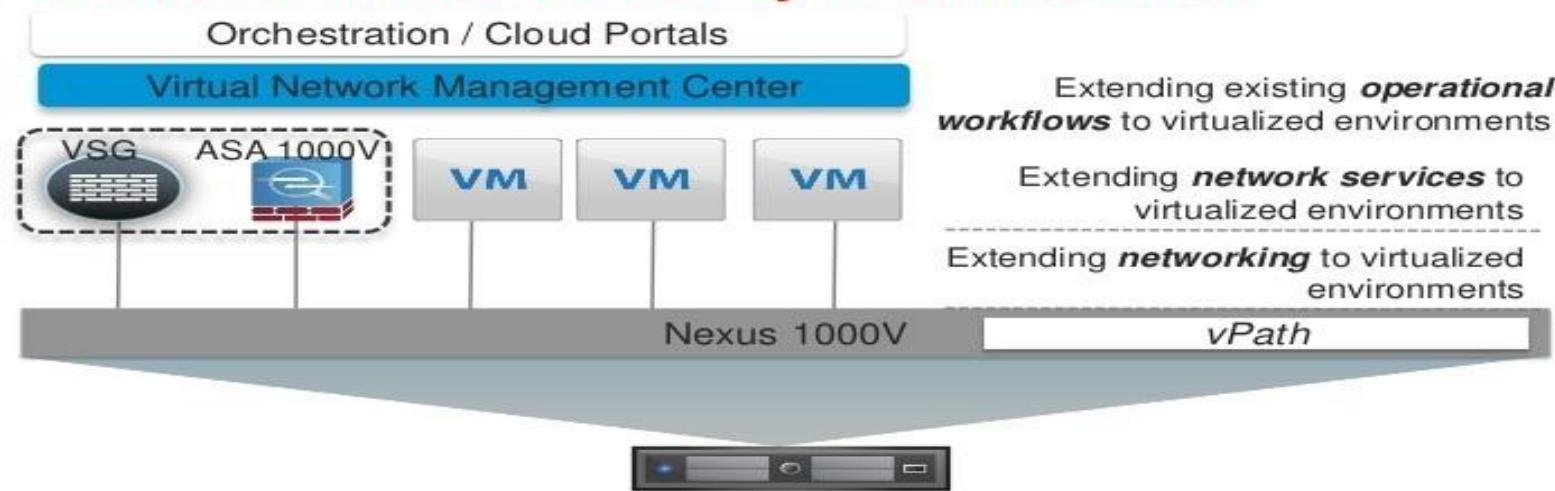
Virtualization Layer 6: management of virtualized environments



Virtualization Layer 7: security for virtualized environments

- Monitors and protects all of the other layers of virtualization so that only authorized use can be made of the resources. Yes, this is called *security for virtualized environments* in the model. As with management of virtualized environments, this layer is an important part of SDN, SDS and SDDC.
- Bitdefender, Kaspersky, TrendMicro, McAfee and many others play in this area of the virtualization market.

Cisco's Virtual Security Architecture

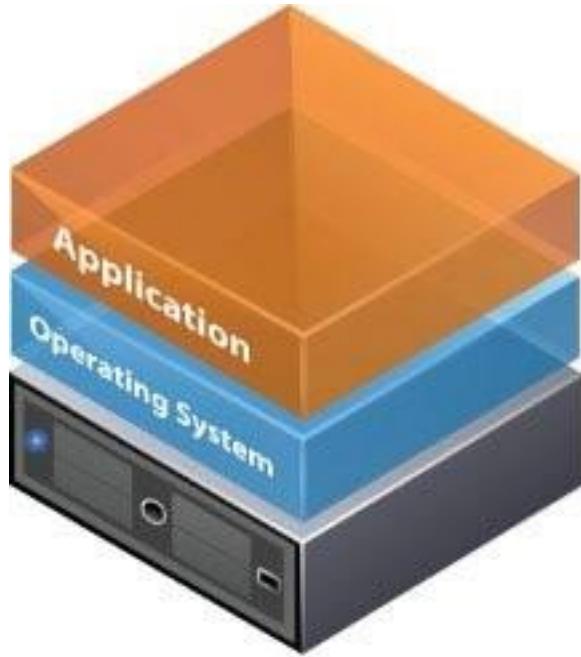


Types of Virtualization in Cloud Computing

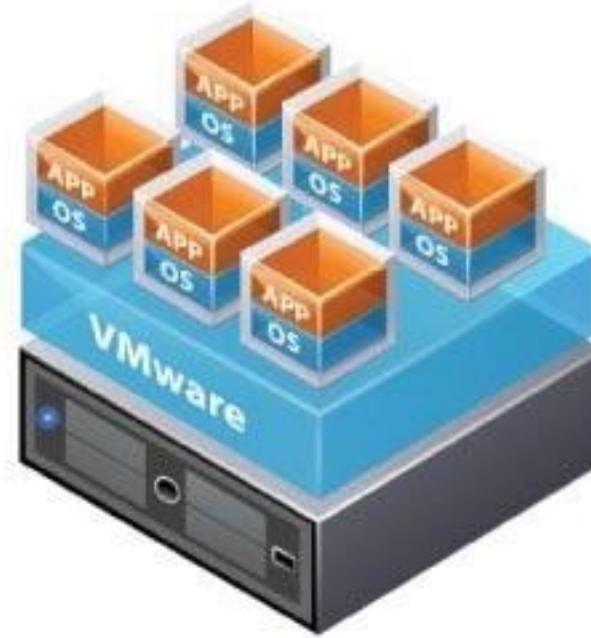
Virtualization

Hardware	Network	Storage	Memory	Software	Data	Desktop
<ul style="list-style-type: none">• Full• Bare-Metal• Hosted• Partial• Para	<ul style="list-style-type: none">• Internal Network Virtualization• External Network Virtualization	<ul style="list-style-type: none">• Block Virtualization• File Virtualization	<ul style="list-style-type: none">• Application Level Integration• OS Level Integration	<ul style="list-style-type: none">• OS Level• Application• Service	<ul style="list-style-type: none">• Database	<ul style="list-style-type: none">• Virtual desktop infrastructure• Hosted Virtual Desktop

Virtualization in Cloud Computing



Traditional Architecture



Virtual Architecture

Virtualization associated with computing technologies

- 1. Hardware virtualization or Server virtualization:**
the partitioning a physical server into smaller virtual servers.
- 2. Network virtualization:** using network resources through a logical segmentation of a single physical network.
- 3. Storage virtualization:** the amalgamation of multiple network storage devices into what appears to be a single storage unit.
- 4. Memory Virtualization :**

Virtualization associated with computing technologies

5. Software Virtualization:

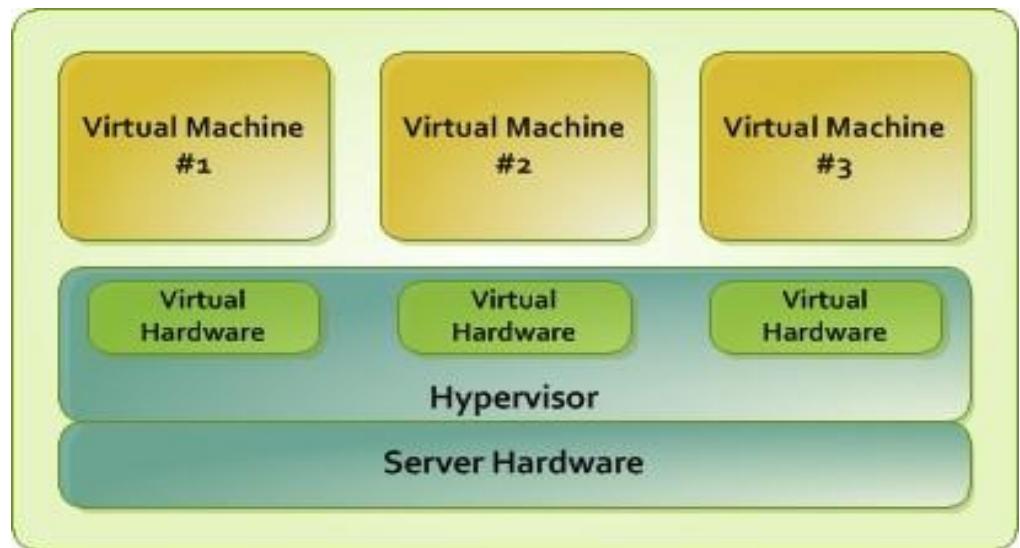
5. Data Virtualization:

5. Desk Top Virtualization

1: Hardware virtualization

1: Hardware virtualization

- Hardware virtualization, which is also known as *server virtualization* or simply **virtualization**, is the abstraction of computing resources from the software that uses those resources.
- The basic logic behind **hardware virtualization** is to integrate many small services into a large physical server so that it can **use** more effectively and providing the service efficiently. Here, the operating system which runs on the physical server convert into an operating system which works inside the virtual machine.



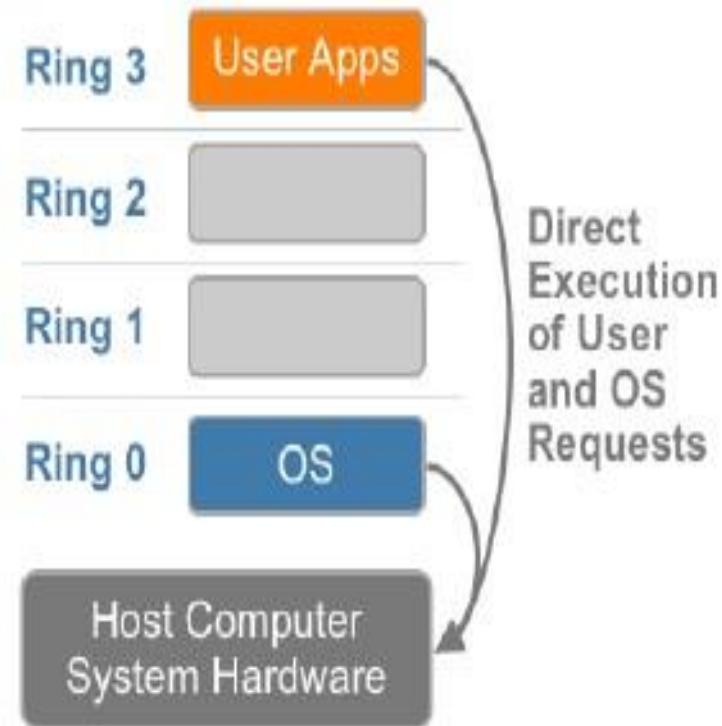
1: Hardware virtualization

● Types of Hardware Virtualization

1. Full Virtualization: Here the hardware architecture is completely simulated. ...
2. Emulation Virtualization: Here the virtual machine simulates the hardware & is independent. ...
3. Para-Virtualization: Here, the hardware is not simulated; instead the guest software runs its isolated system.

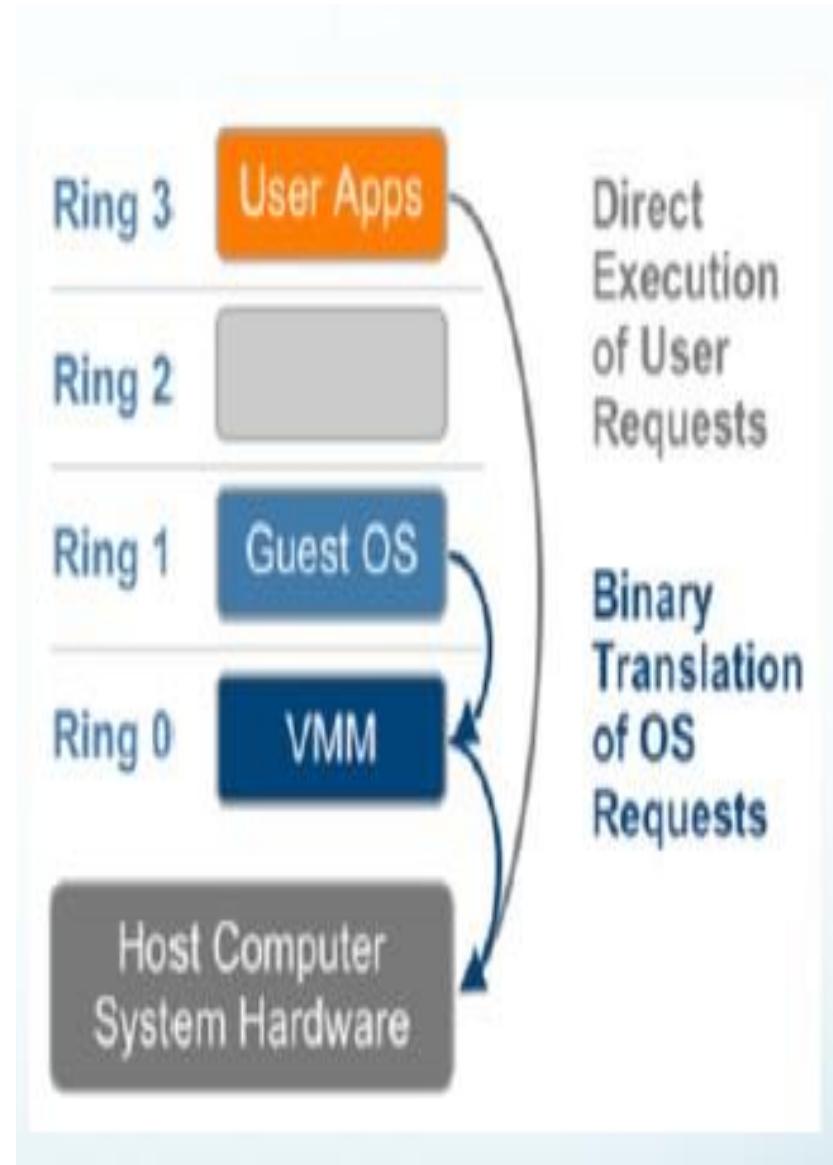
Virtualizing privileged instruction

- x86 architecture has four privilege levels (rings).
- The OS assumes it will be executing in Ring 0.
- Many system calls require 0-level privileges to execute.
- Any virtualization strategy must find a way to circumvent this.



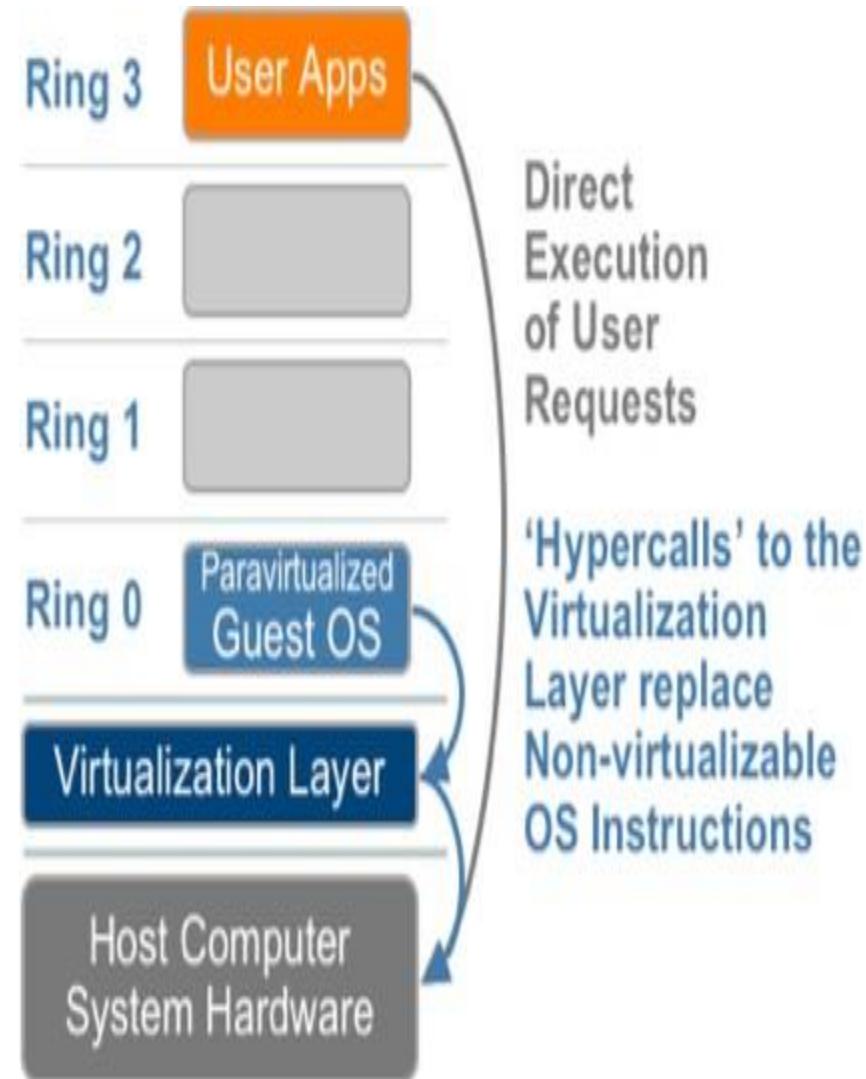
Full Virtualization

- “Hardware is functionally identical to underlying architecture.”
- Typically accomplished through interpretation or binary translation.
- **Advantage:** Guest OS will run without any changes to source code.
- **Disadvantage:** Complex, usually slower than paravirtualization.



Paravirtualization

- Replace certain unvirtualized sections of OS code with virtualization-friendly code.
- Virtual architecture “similar but not identical to the underlying architecture.”
- Advantages: easier, lower virtualization overhead
- Disadvantages: requires **modifications to guest OS**



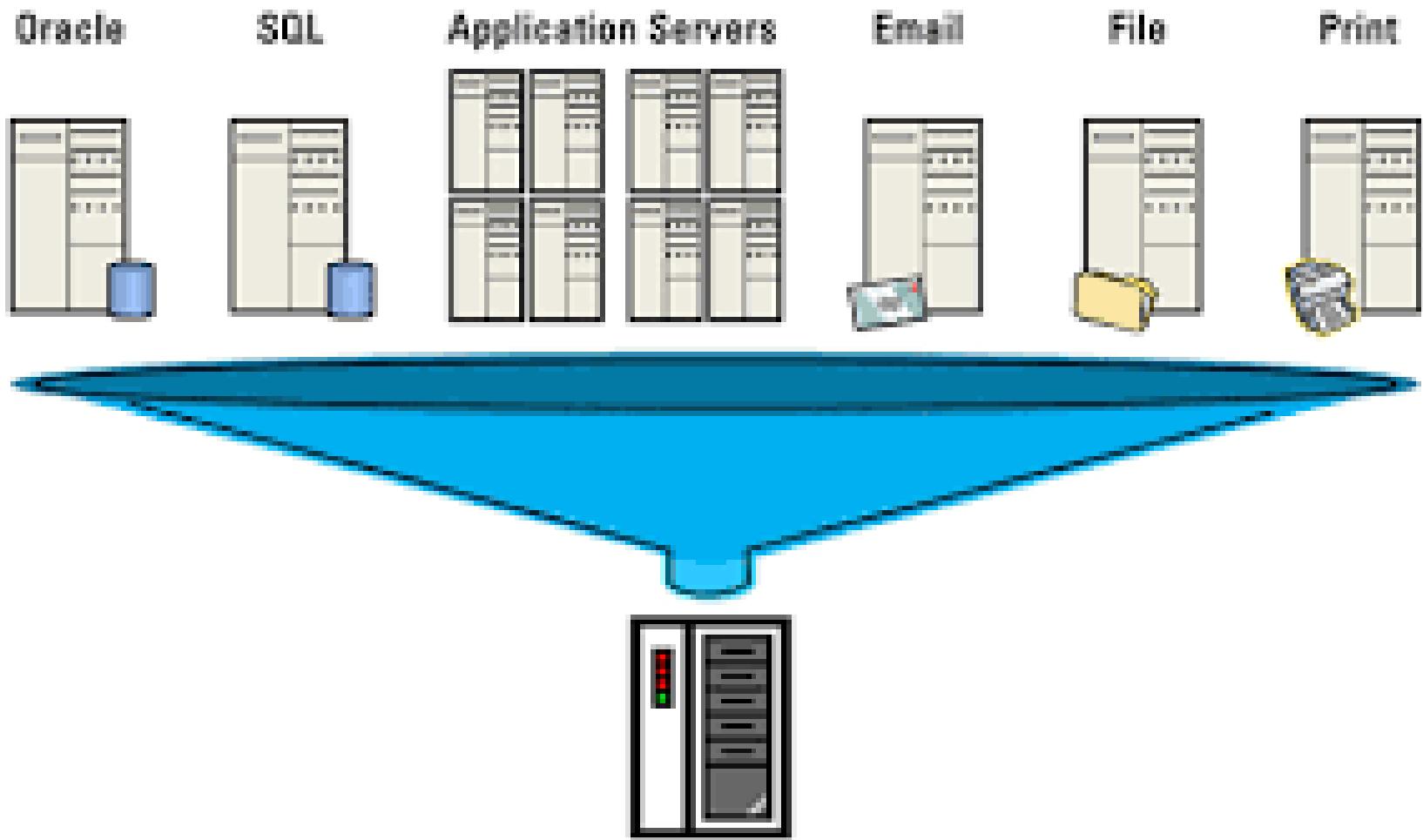
The benefits of hardware virtualization

- **Lower Cost:** Because of server consolidation, the cost decreases; now it is possible for multiple OS to exist together in a single hardware. This minimizes the quantity of rack space, reduces the number of servers and eventually drops the power consumption.
- **Efficient resource utilization:** Physical resources can be shared among virtual machines. The unused resources allocated by one virtual machine can be used by another virtual machine in case of any need.
- **Increase IT flexibility:** The quick development of hardware resources became possible using virtualization, and the resources can be managed consistently also.
- **Advanced Hardware Virtualization features:** With the advancement of modern hypervisors highly complex operations maximize the abstraction of hardware & ensure maximum uptime, and this technique helps to migrate an ongoing virtual machine from one host to another host dynamically.

1: Server virtualization

- Server virtualization is the masking of server resources (including the number and identity of individual physical servers, processors, and operating systems) from server users.
- The intention is to spare the user from having to understand and manage complicated details of server resources while increasing resource sharing and utilization and maintaining the capacity to expand later.
- The **server** administrator uses a software application to divide one physical **server** into multiple isolated virtual environments.

1: Server virtualization



1: Hardware virtualization

Hardware virtualization vendors and products

- **VMware ESXi** is a hypervisor designed for hardware virtualization. ESXi installs directly onto a server and has direct control over a machine's underlying resources. ESXi will run without an OS and includes its own kernel. ESXi is the compact, and now preferred, version of VMware's ESX. ESXi is smaller and doesn't contain the ESX service console.
- **Microsoft Hyper-V** is a hypervisor designed for hardware virtualization on an x86 architecture. Hyper-V isolates VMs in partitions, where each guest OS will execute a partition. Partitions operate in the manner of parent and child partitions. Parent partitions have direct access to the hardware, while child partitions have a virtual view of system resources. Parent partitions create child partitions using a hypercall API. Hyper-V is available for 64-bit versions of Windows 8 Professional, Enterprise, Education and later.
- **Xen** is an open source hypervisor. Xen is included in the Linux kernel and is managed by the Linux Foundation. However, Xen is only supported by a small amount of Linux distributions, such as SUSE Linux Enterprise Server. The software supports full virtualization, paravirtualization and hardware-assisted virtualization. XenServer is another open source Xen product to deploy, host and manage VMs.

2:Network virtualization

2: Network virtualization

- **Network virtualization** is the process of combining hardware and software **network** resources and **network** functionality into a single, software-based administrative entity, a virtual **network**.
- **Network virtualization** involves platform **virtualization**, often combined with resource **virtualization**.
- **Network virtualization** is a method of combining the available resources in a **network** by splitting up the available bandwidth into channels, each of which is independent from the others, and each of which can be assigned (or reassigned) to a particular server or device in real time.

2: Network virtualization

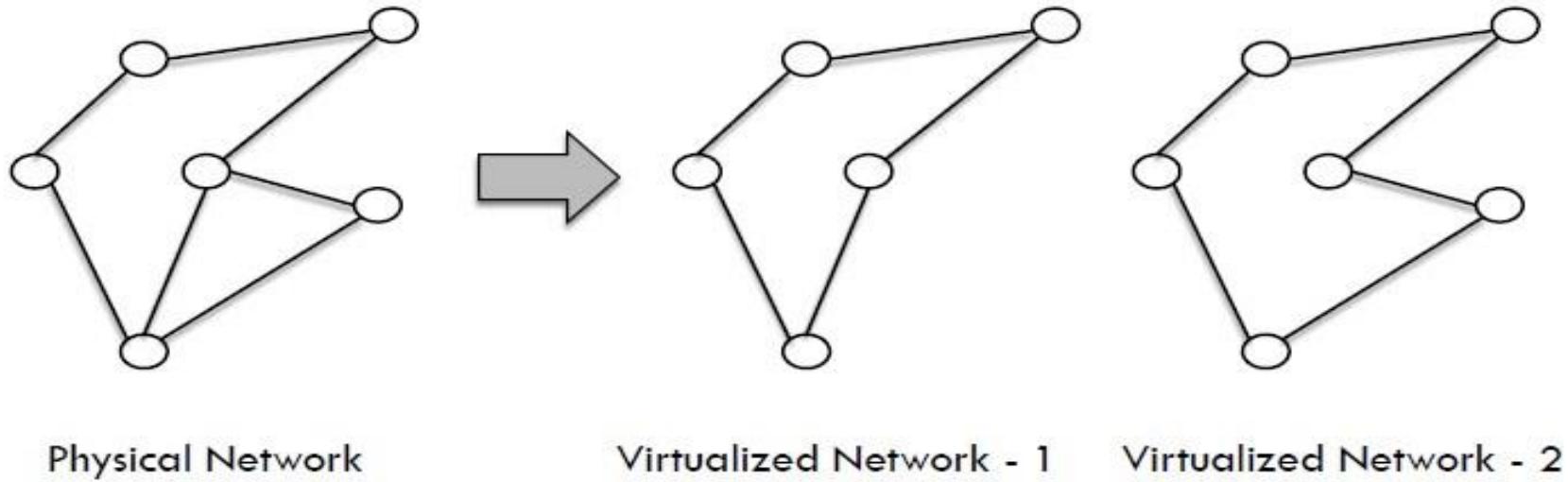
- In network virtualization, multiple sub-networks can be created on the same physical network, which may or may not be authorized to communicate with each other.
- This enables restriction of file movement across networks and enhances security, and allows better monitoring and identification of data usage which lets the network administrator's scale up the network appropriately.
- It also increases reliability as a disruption in one network doesn't affect other networks, and the diagnosis is easier.

Subtypes:

- Internal network: Enables a single system to function like a network
- External network: Consolidation of multiple networks into a single one, or segregation of a single network into multiple ones

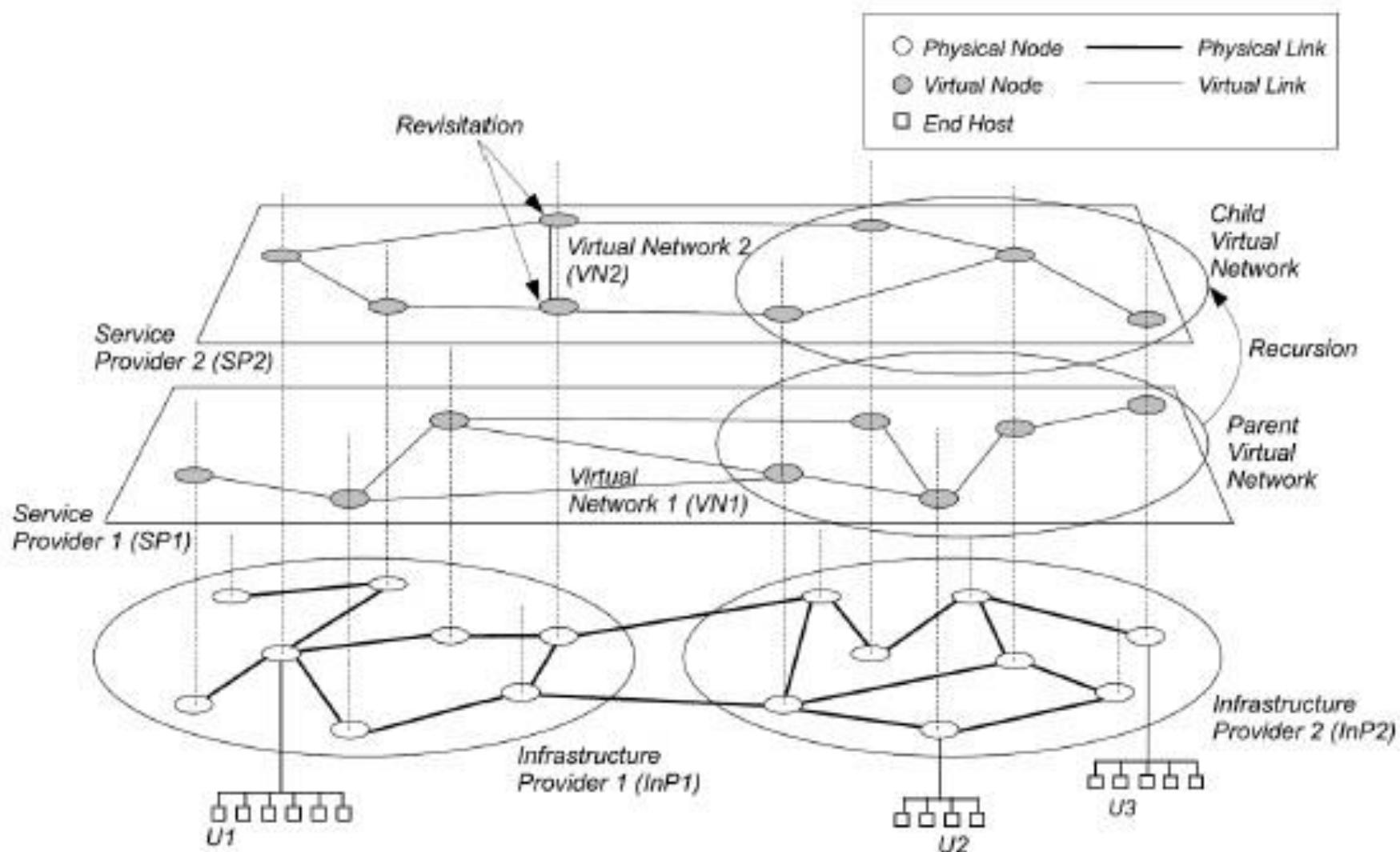
2: Network virtualization

Making a physical network appear as multiple logical ones



- The idea is that virtualization disguises the true complexity of the network by separating it into manageable parts, much like your partitioned hard drive makes it easier to manage your files.

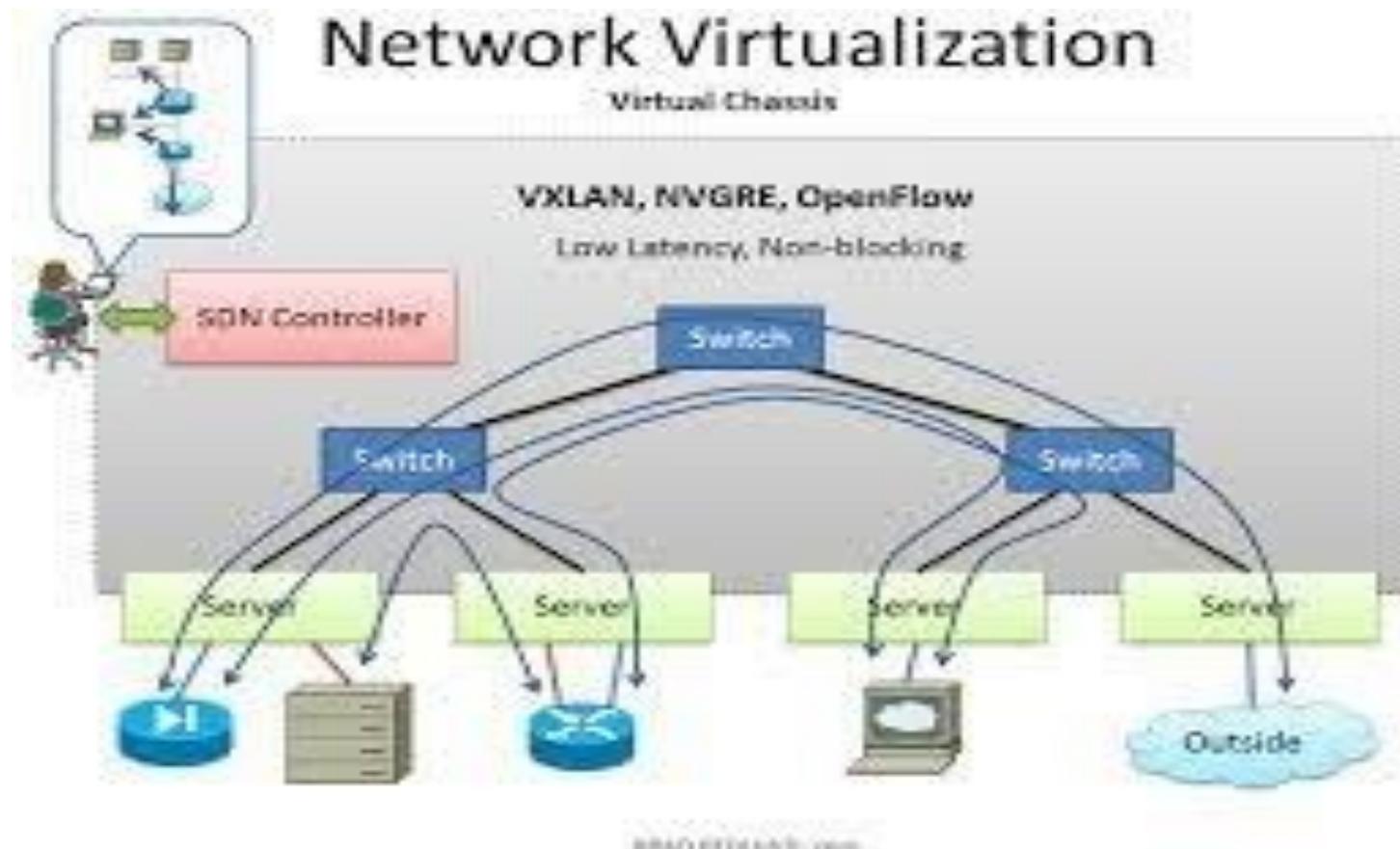
2: Network virtualization



2: Network virtualization

- Network virtualization in cloud computing is a method of combining the available resources in a network by splitting up the available bandwidth into different channels, each being separate and distinguished.
- They can be either assigned to a particular server or device or stay unassigned completely — all in real time.
- The idea is that the technology disguises the true complexity of the network by separating it into parts that are easy to manage, much like your segmented hard drive makes it easier for you to manage files.
- Network virtualization refers to the management and monitoring of a computer network as a single managerial entity from a single software-based administrator's console.

2: Network virtualization



3: Storage virtualization

3: Storage virtualization

- Storage virtualization is the pooling of physical storage from multiple network storage devices into what appears to be a single storage device that is managed from a central console.
- Storage virtualization is commonly used in storage area networks (SANs)
- A **Storage area network**, or SAN, is a high-speed **network** of **storage** devices that also connects those **storage** devices with servers.
- It provides block-level **storage** that can be accessed by the applications running on any networked servers.

3: Storage virtualization

- Multiple physical storage devices are grouped together, which then appear as a single storage device. This provides various advantages such as homogenization of storage across storage devices of multiple capacity and speeds, reduced downtime, load balancing and better optimization of performance and speed.
- Partitioning your hard drive into multiple partitions is an example of this virtualization.

Subtypes:

- **Block Virtualization** – Multiple storage devices are consolidated into one
- **File Virtualization** – Storage system grants access to files that are stored over multiple hosts

3: Storage virtualization

- Storage virtualization creates a **layer of abstraction between the operating system** and the physical disks used for data storage.
- For example, the **storage virtualization** software or device creates a logical space, and then manages metadata that establishes a map between the logical space and the physical disk space.

Types of storage virtualization

Block Virtualization

Virtualizes LUNs presented to applications

Disk Virtualization

Abstracts disks into chunks in storage pools that are used to create LUNs

Tape Virtualization

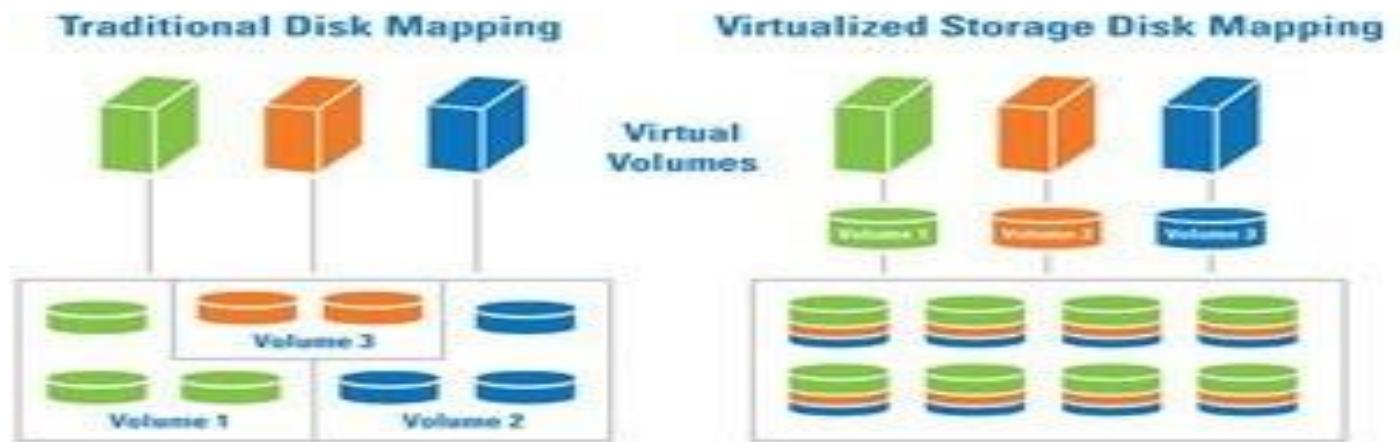
Creates a virtual tape on a disk storage system

File Virtualization

Virtualizes NAS and file servers into a single namespace

3: Storage virtualization

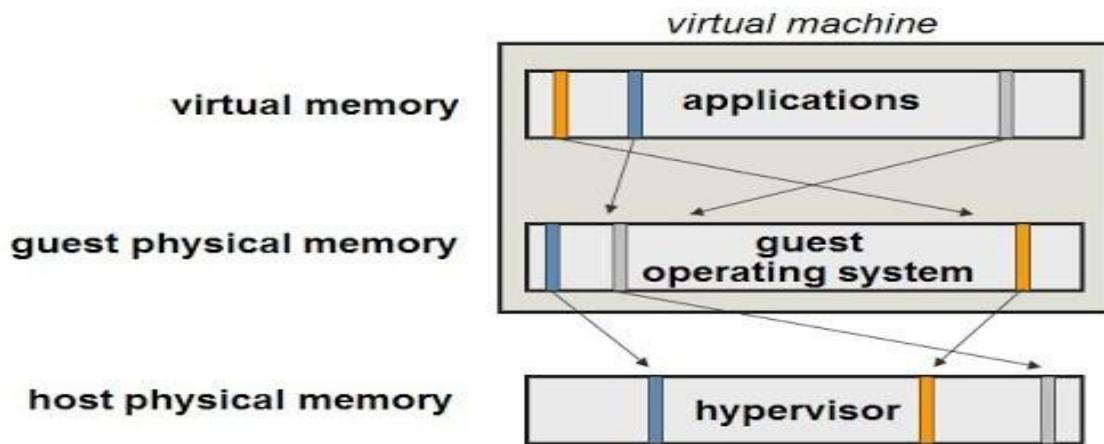
- Storage Virtualization is the concept of virtualizing enterprise storage at the disk level, creating a dynamic pool of shared storage resources available to all servers, all the time. With read/write operations spread across all drives, multiple requests can be processed in parallel, boosting system performance. This allows users to create hundreds of virtual volumes in seconds to support any virtual server platform and



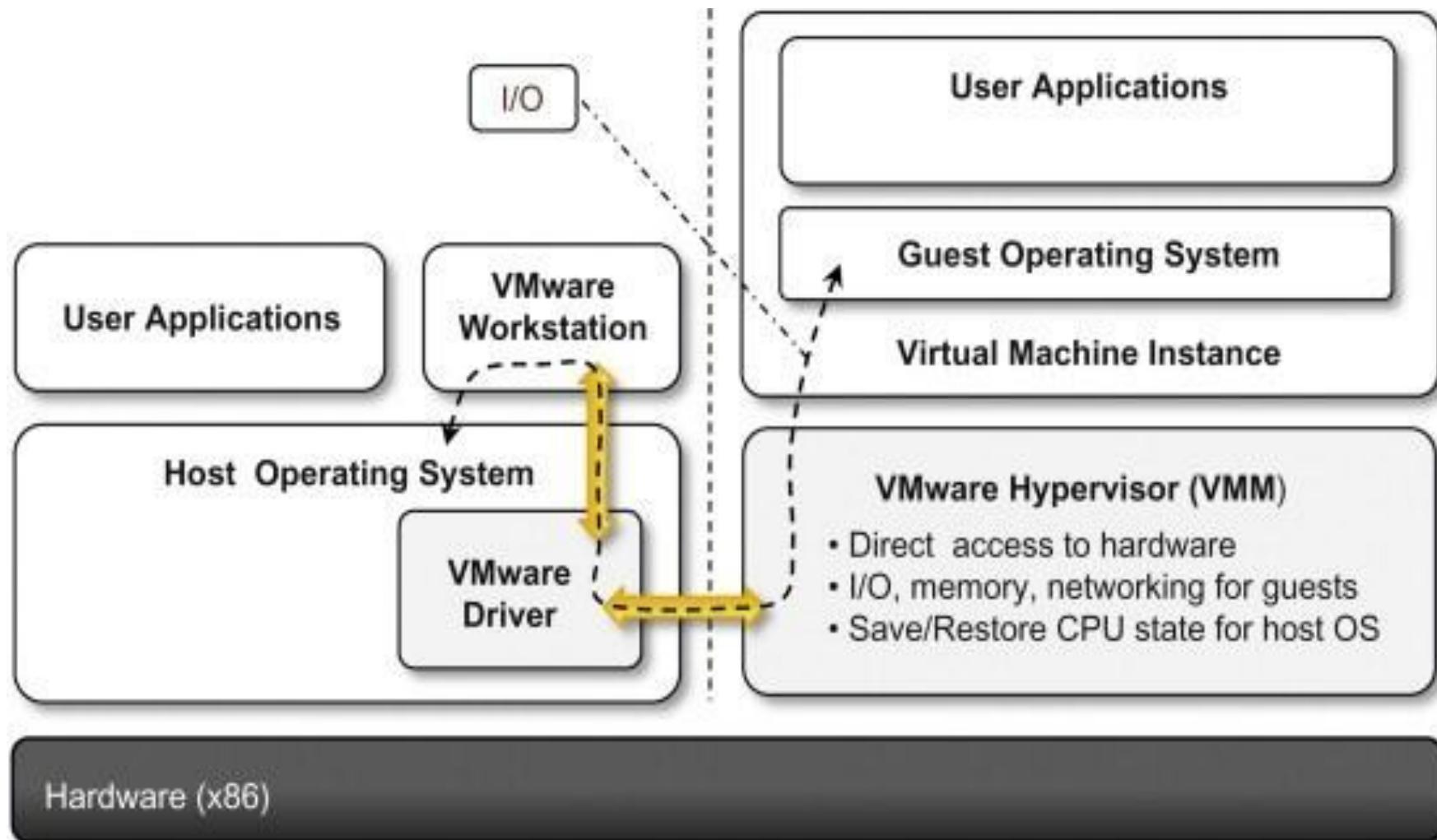
4: Memory Virtualization

4: Memory virtualization

- In computer science, memory virtualization decouples volatile random access memory resources from individual systems in the Data Centre, and then aggregates those resources into a virtualized memory pool available to any computer in the cluster.
- It introduces a way to decouple memory from the server to provide a shared, distributed or networked function.
- It enhances performance by providing greater memory capacity without any addition to the main memory. That's why a portion of the disk drive serves as an extension of the main memory.



4: Memory virtualization



4: Memory Virtualization

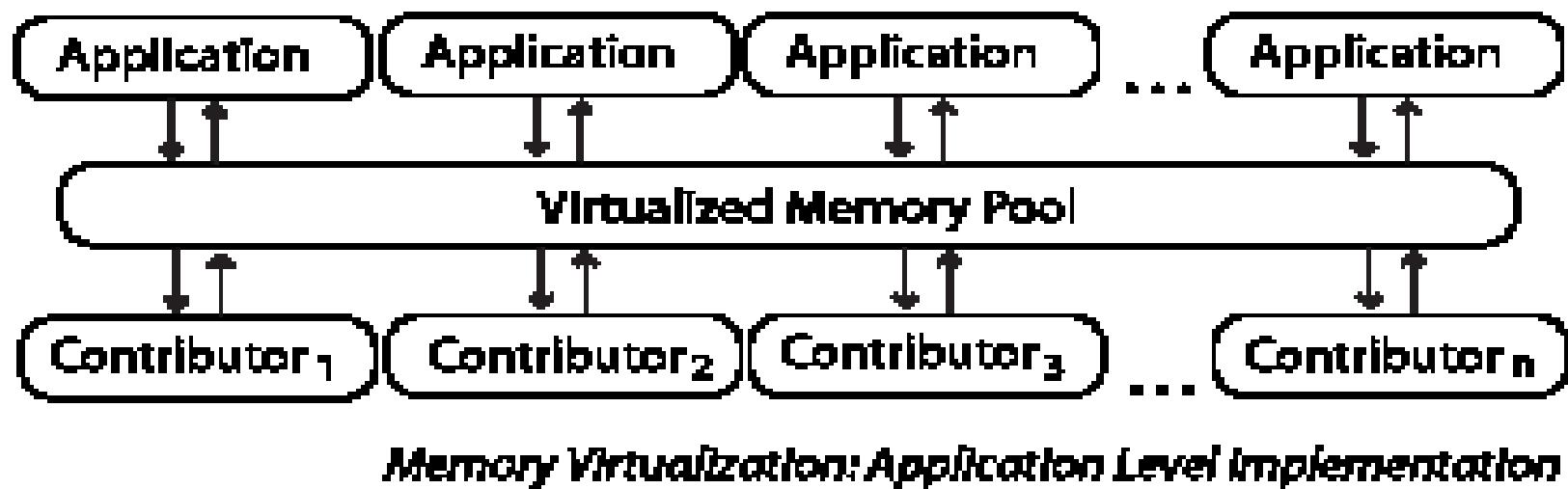
- Physical memory across different servers is aggregated into a single virtualized memory pool.
- It provides the benefit of an enlarged contiguous working memory. You may already be familiar with this, as some OS such as Microsoft Windows OS allows a portion of your storage disk to serve as an extension of your RAM.

Subtypes:

- **Application-level control** – Applications access the memory pool directly
- **Operating system level control** – Access to the memory pool is provided through an operating system

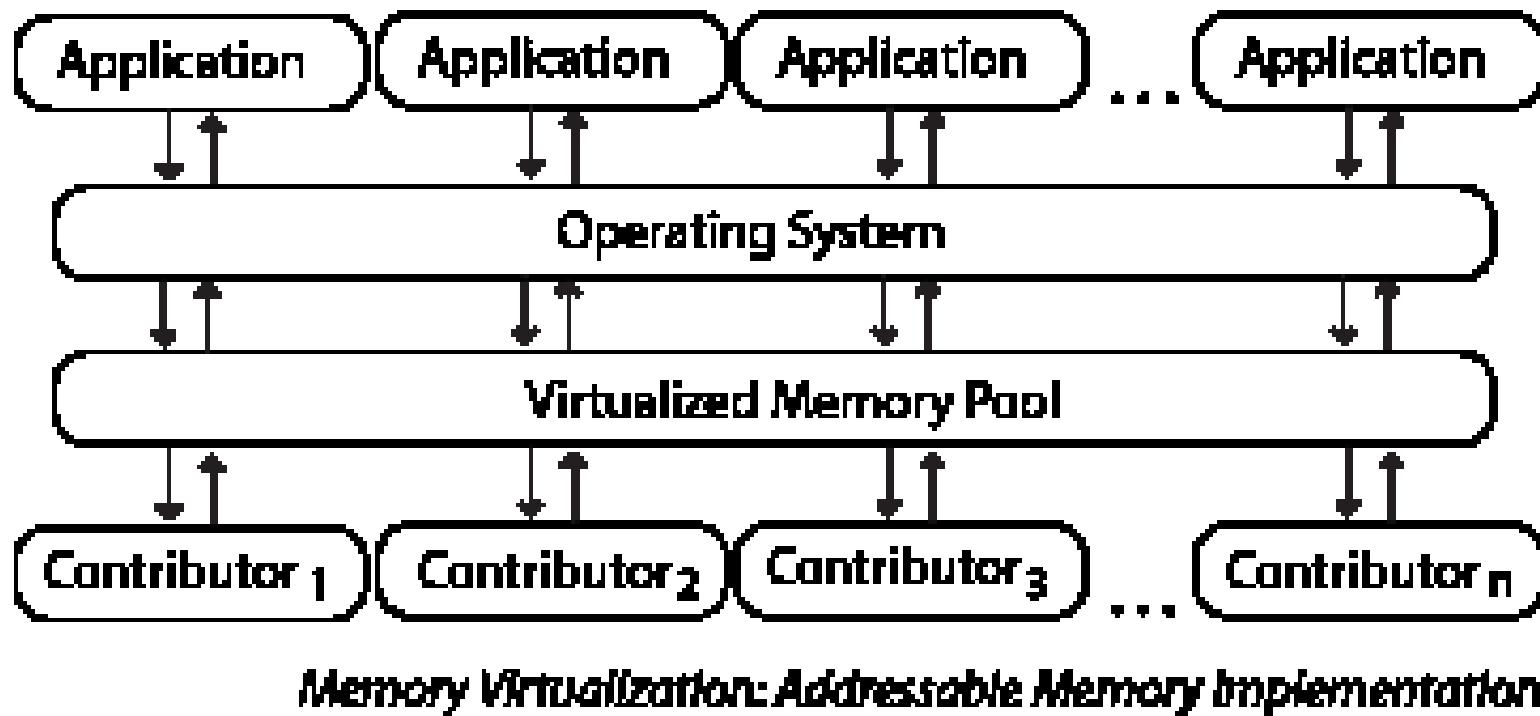
4: Memory virtualization: Implementation

- **Application-level integration** – Applications running on connected computers directly connect to the memory pool through an API or the file system.



4: Memory virtualization: Implementation

- Operating System-Level Integration – The operating system first connects to the memory pool and makes that pooled memory available to applications



5: Software virtualization

5: Software virtualization

- Software **Visualization in Cloud Computing** allows the single computer server to run one or more virtual environments. It is quite similar to virtualizations but here it abstracts the software installation procedure and creates a virtual software out of it.
- In software virtualizations, an application will be installed which will perform the further task. One software is physical while others are virtual as it allows 2 or more operating system using only one computer.
- Software virtualization is similar to that of **virtualization** except that it is capable to abstract the **software** installation procedure and create virtual **software** installation.
- It is also called **application virtualization** is the practice of running **software** from a remote server.

5: Software virtualization

It creates a computer system complete with hardware that lets the guest operating system to run. For example, it lets you run **Android** OS on a host machine natively using a Microsoft Windows OS, utilizing the same hardware as the host machine does

Benefits of Software Virtualization

Testing

- It is easier to test the new operating system and software on VMs as it does not require any additional hardware and the testing can be done within the same software. After the testing, the VM can be moved or deleted for further testing.

Utilization

- In software virtualization, there is higher efficiency in resource utilization if it is configured correctly. The VM can be modified as per the requirement such as the user can modify RAM, drive space, etc. It requires very less amount of hardware as compared to the equivalent number of physical machines.

Efficient

- It is efficient in a way such that it can run 12 virtual machines and eliminates the use of 12 physical boxes. This is the power cost as well as the cost of maintaining the server.

Benefits of Software Virtualization

Less Downtime

- The software is upgrading and the upgrade in the VMs can do when the VM is working. VM can modify when it is working or it is not working which means that the downtime of it is very less.

Flexible

- It provides flexibility to the user so that the user can modify the software as per their demand. The modification can do within minutes and can adjust easily when the workload changes.

Secure

- It can protect with many hantaviruses. Moreover, there are several firewalls which prevent hacking and virus. The data in the software virtualization is safe as it stores in several different places so if the disaster takes place the data can retrieve easily.

How Software Virtualization in Cloud Computing Works?

- **Backup:** With the help of software virtualization, the entire **operating system** or server installation can be backed up. This also benefits in a way that if the new server hack just restoring the previous version will allow running the server.
- **Run multiple operating systems:** The different operating system can use in a single computer with the partition in the hard drive. The only thing to keep in mind is to keep a snapshot of everything. If the data drowns, it can retrieve from some other place.
- **Running a different version of applications:** With the help of software virtualization new as well as the old operating system can use. So a program, if it is not working on a particular operating system, we can check it on another one.
- **Templates:** After the configuration of VM as per the demand, it can convert into a template and this template can use to make multiple copies of the original one.

5: Type of Software virtualization

- i. **Operating System Virtualization**
- ii. **Application Virtualization**
- iii. **Service Virtualization**

i. Operating System Virtualization

- In operating system virtualization, the hardware is used which consists of software on which different operating systems work. Here, the operating system does not interfere with each other so that each one of them works efficiently.

ii. Application Virtualization

- Application virtualization is a technology, encapsulates the computer program within the operating system. It can say that application virtualizations refer to running an application on a thin client. This thin client runs an environment, which is different from what refer to as encapsulating from the operating system which is the location of it.

5: Type of Software virtualization

Service Virtualization

- In the service virtualization, the DevOps team can use the virtual servers rather than the physical one. It emulates the behaviour of essential components which will be present in the final production environment.
- With the help of service virtualization, the complex application can go through testing much earlier in the development process. It can say that service visualization is a technique to simulate the behaviour of some components in a mixture of component-based applications

5.1: Operating system virtualization

- Operating system virtualization is the use of software to allow a piece of hardware to run multiple operating system images at the same time.
- The technology got its start on mainframes decades ago, allowing administrators to avoid wasting expensive processing power.
- Operating system virtualization (OS virtualization) is a server virtualization technology that involves tailoring a standard operating system so that it can run different applications handled by multiple users on a single computer at a time.

5.2: Application virtualization

- **Application virtualization** is software technology that encapsulates **computer programs** from the underlying **operating system** on which it is executed.
- A fully **virtualized application** is not installed in the traditional sense, although it is still executed as if it were.
- The application behaves at runtime like it is directly interfacing with the original operating system and all the resources managed by it, but can be isolated or **sandboxed** to varying degrees.
- Application virtualization fools the computer into working as if the application is running on the local machine, while in fact it is running on a virtual machine (such as a server) in another location, using its operating system (OS), and being accessed by the local machine.

Application virtualization

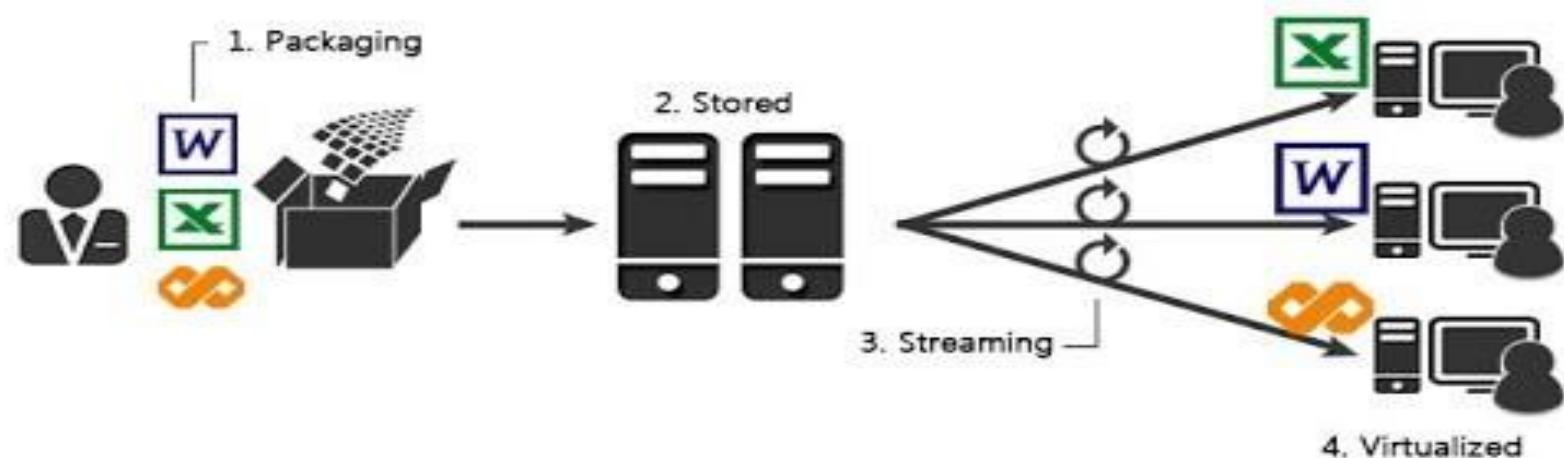
- Application virtualization, also called **application service virtualization**, is a term under the larger umbrella of virtualization.
- It refers to running an application on a thin client; a terminal or a network workstation with few resident programs and accessing most programs residing on a connected server.
- The thin client runs in an environment separate from, sometimes referred to as being encapsulated from, the operating system where the application is located.
- App virtualization (application virtualization) is the separation of an installation of an application from the client computer that is accessing it. There are two types of application virtualization: **remote** and **streaming**.

Application virtualization

- **Application virtualization** is a process that deceives a standard app into believing that it interfaces directly with an operating system's capacities when, in fact, it does not.
- This use requires a virtualization layer inserted between the app and the OS. This layer, or framework, must run an app's subsets virtually and without impacting the subjacent OS.
- The virtualization layer replaces a portion of the runtime environment typically supplied by the OS, transparently diverting files and registry log changes to a single executable file.
- By diverting the app's processes into one file instead of many dispersed across the OS, the app easily operates on a different device, and formerly incompatible apps can now run adjacently.
- Used in conjunction with application virtualization is **desktop virtualization**—the abstraction of the physical desktop environment and its related app software from the end-user device that accesses it.

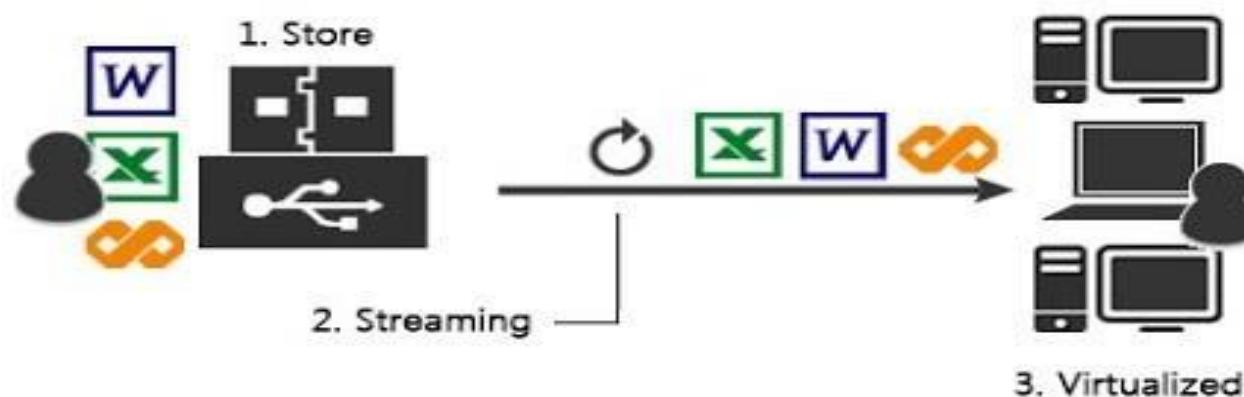
Application Virtualization from Servers

- Software pack is made by the pack builder, or it can be provided by service provider.
- The administrator uploads software packs on the server
- When the users launches the application, server streams it to the users in realtime.
- As the streaming process begins, the application is virtualized as if it is installed in the local machine.



Application Virtualization from a Portable Device

- Store the software pack on the device storage.
- When the user launches the application, the application gets streamed.
- As the streaming process begins, the user can use the software as if it is locally installed.



The advantages of application virtualization

- **No installation required:** Installing an application on hundreds or thousands of computers is prone to error. Application virtualization simplifies software deployment.
- **Application retirement simplified:** Getting rid of an app in your whole network is much easier as well. Since virtual apps just have to be deleted, uninstalling them is usually not required.
- **No more application conflicts:** Sometimes installing an app corrupts another app. Application virtualization helps reduce the risks of application conflicts.
- **No registry and system bloat:** The more apps you install on a desktop, the more bloated its registry and system folder will get. This makes the computer slower and increases the risk of failures. Application virtualization lets the registry and the system folder untouched.
- **End users require only minimal privileges:** Legacy apps that require admin rights usually work in environments where end-users only have standard rights.

The advantages of application virtualization

- **Multiple runtime environments:** You can deploy the runtime environment together with the application. This enables you to run different versions of a runtime environment on a desktop. For example, you can run different Java versions simultaneously without messing around with environment variables.
- **Multiple versions of the same application:** For instance, end users can run Word 2003 and Word 2007 at the same time.
- **Deploy apps on unmanaged computers:** If clients or partners have to use an app to access the services of your organization, you can just send them an executable where you have already configured everything for them.
- **Application updates:** You can update the virtualized application at a central location on your servers. This means you have to update an app only once and not on all of your desktops.

The advantages of application virtualization

- **Rollback:** If an app no longer works properly on a user's desktop because he or she changed too many settings or installed incompatible add-ons, you can just reset the app to its original state.
- **Simplified roaming:** Some products allow you to store settings and data belonging to the app in the virtual environment on a server or a memory stick. This way, end users can access their apps with their own settings regardless on which desktop they logon. This is also possible if you are not working with roaming user profiles.
- **Simplified OS deployment:** Deploying a new OS in your network doesn't affect the applications. Also, if you have to reinstall an image on a desktop, you don't have to worry about the apps running on this machine because end users can just access them on the server.
- **Integration with desktop virtualization:** Software virtualization and desktop virtualization perfectly harmonize because these technologies allow you to separate the OS deployment process from software distribution.

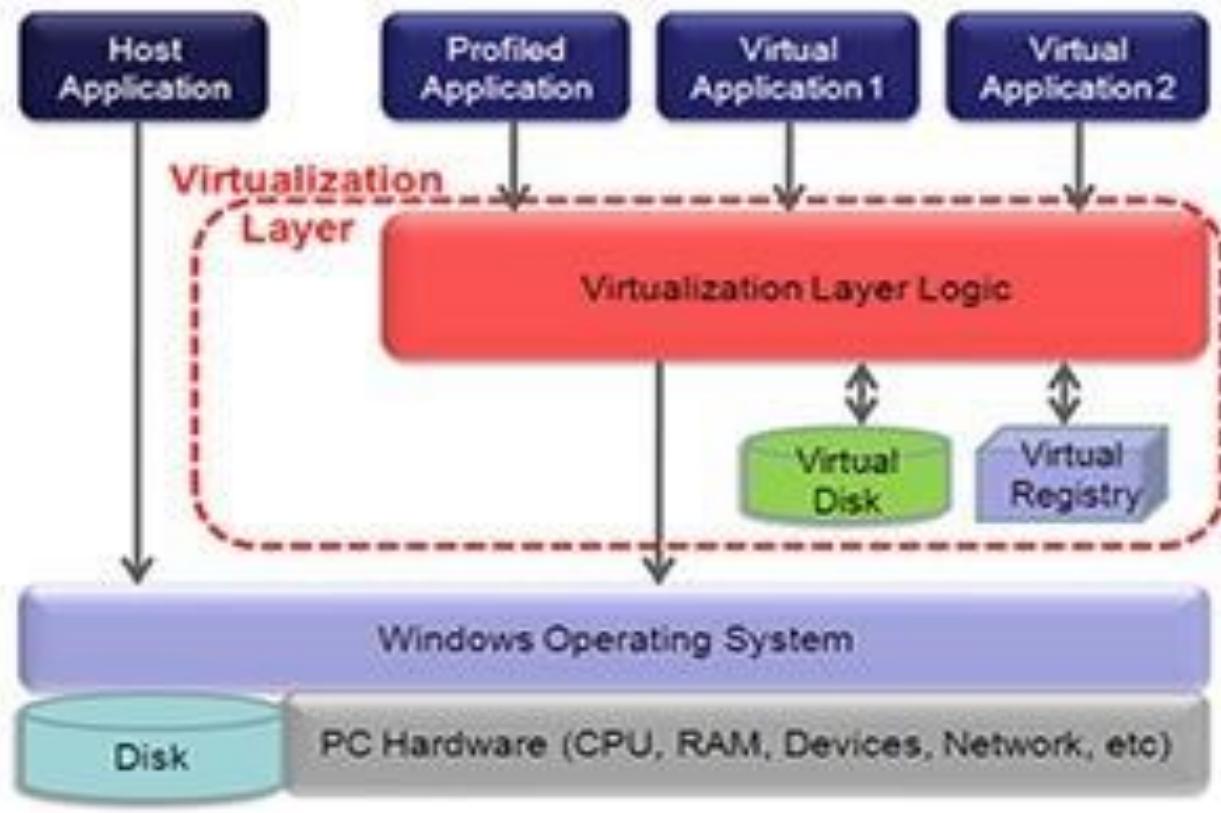
The advantages of application virtualization

- **Reduced regression testing:** Once you know that your app works in the virtual environment, you don't have to make sure that it works on all of the different desktop variations in your network. Changes on desktops usually don't have an effect on the virtualized apps.
- **Improved security:** Virtualized apps are isolated from the operating system and from each other. This way, malware can't infect other parts of the system, easily .
- **Helpdesk support:** Helpdesk personnel can easily access all available apps in your organization and can run the app in the same environment as end users.
- **Operating system independent:** Virtualized apps are often OS independent. If you have apps that are Vista incompatible, then application virtualization might be your solution. Application virtualization also paves the way for Vista x64. Legacy apps that wouldn't work on a 64-bit-system might run without problems on Vista x64 in a virtual environment. Together with Wine and Crossover, you can also run complex Windows apps on Linux and OS X.

How Server Virtualization and Application Virtualization Differ

- Although the two processes share key features—such as lowering costs, bolstering data security, and central control—they fulfill separate functions.
- Server virtualization refers to the use of one or several servers clustered into multiple server groups. Should a data center have 20 physical servers, they can be virtualized into two groups of 10, for example, or two groups with one of 5 servers and the other with 15. There's no difference between a virtual server(s) and a group of 5, 10, or 15 physical servers operating as individual servers.
- Conversely, one physical server can be partitioned into separate multiple virtual servers, helping to maximize organizational resources and facilitating recovery from unexpected server outages. With virtual servers, further cost reductions are realized by reducing organizational needs for multiple servers, which leads to lower maintenance and lower environmental and power expenditures
- Virtualizing apps means that they run without any dependencies through another operating system or browser. An example would be virtualizing Microsoft PowerPoint to run on Ubuntu over an Opera browser.
- The implementation of both environments differs, as well. Desktop virtualization impacts network architecture, transmission protocol, and the data center while server virtualization only affects changes to the server.

Application Virtualization



Remote applications

- Remote applications run on a server. End users view and interact with their applications over a network via a remote display protocol.
- The remote applications can be completely integrated with the user's desktop so that they appear and behave the same as local applications, through technology known as seamless windows.
- The server-based operating system instances that run remote applications can be shared with other users (a terminal services desktop), or the application can be running on its own OS instance on the server (**Virtual Desktop Infrastructure (VDI)** desktop).
- A constant network connection must be maintained in order for a remote application to function.

Streaming applications

- With **streaming applications**, the virtualized application is executed on the end user's local computer.
- When an application is requested, components are downloaded to the local computer on demand.
- Only certain parts of an application are required in order to launch; the remainder can be downloaded in the background as needed.
- Once completely downloaded, a streamed application can function without a network connection.
- Various models and degrees of isolation ensure that streaming applications will not interfere with other applications, and that they can be cleanly removed when closed.

5.3 : Service virtualization

- In software engineering, service virtualization or service virtualization is a method to emulate the behavior of specific components in heterogeneous component-based applications such as API-driven applications, cloud-based applications and service-oriented architectures.
- **Service Virtualization** is implemented to emulate the **required** database, network settings, and even system configurations for testing the application. This helps in cutting down the time, efforts, and costs.
- It involves constant testing and releases across multiple testing environments

Why Service virtualization ? (Software testing)

- Service Virtualization is a method that helps you to emulate (virtual services) the behaviors of the component in a Service Oriented Architecture (Microservice).
- Practically the software development, testing, and operations teams do not work in synch, and each team has to wait for others to have components ready. This causes delays in workflows and may deliver an inferior product.
- With Service Virtualization, DevOps teams use virtual services instead of production services, so they can test the system even when key components are not ready.
- With Service Virtualization, integrating of applications takes place early in the development cycle thereby reducing time and cost to fix errors.

Testing as service

Service Virtualization

Having Everything you need, when you need them.



Why Service Virtualization?

- Helps you to accelerate Application Delivery while mitigating business risks
- It helps you to simulates the behavior of select components within an application to enable end-to-end testing of the application as a whole.
- Allows the teams to work in parallel
- Allows you to Test Early and often which expose defect when they are faster, easiest and least costly to resolve
- Facilitates better test coverage
- Virtual services offer rich tools for editing and managing which help you improve productivity, cut down on maintenance time and development costs.
- It is useful for anyone involved in developing and delivering software applications.
- Access to more systems and services
- It supports test-driven development.
- Gives you an isolated environment for testing

Example: Service Virtualization

- For example, say you are developing an e-commerce mobile app that lets you browse items and make a payment in this scenario, the mobile app likely integrates with several systems, including databases and web-services (APIs).
- These integrations allow you to get items, make the payment, and ultimately place the order.
- Many of these resources are in your control – but many others are not, because they're owned by other teams. Such resources are called **third-party dependencies**.
- A typical list of third-party dependencies for this example may look like the following:

Example: A typical list of third-party dependencies

Resource	Owner	Availability
Database	Yourself	Always Available
User Profile API	CRM team	Still Under Development
Inventory API	ERP Team	Only in Production
Payment API	Payment Gateway	Needs a Credit Card
Ordering API	Order Management Team	Available, but with Rate Limits

Example: Service Virtualization

- As you can see in the above table, not all the dependencies are under your control. Often while developing code you will reach a point when you must use these third-party dependencies and interact with them. One example might involve getting data for a customer profile screen. If these dependencies are not ready and responsive, you will have to stop working on the code that call them.
- Often, to work around this, developers will write functions or methods that return a hard-coded API response. This is not ideal, as the hard-coded values are valid for just one scenario and not reusable, this also encourages bad code writing practices. Similarly, a tester testing functionality that depends on other unavailable resources will be forced to skip testing until that resource is available, adversely impacting test coverage.
- Unavailable dependencies create issues for developers and testers as they limit the amount of development and testing that can be done. Without full availability of dependent resources and services, all possible use cases do not get covered.
- This results in limited unit testing and many of the test cases do not get executed at all. Often these dependencies cause bottlenecks. Teams have to pause their work because they've come to a point where they absolutely need the actual third part services. Without them, it's not possible to get any work done at all.

API Virtualization: The Perfect Solution to Challenges with API Mocking

- Mocking provides a way to emulate the missing resource, especially APIs, while testing or developing a software.
- Mocks are very basic setups that are created on code or on tools that can mimic, for example, a few API calls. This allows the development and testing teams to do a very basic level of testing using these emulated calls.
- API virtualization is often compared with mocking, but they're not the same thing. Mocks are imitation software components that are used by developers to emulate real software components for testing purposes, which initially sounds a lot like virtualization.

API Virtualization: The Perfect Solution to Challenges with API Mocking

- However, one of the biggest distinctions between mocking services and virtual APIs is that mocking functions tend to be very context-specific, simulating a specific behavioral response to fulfill a certain development need at a certain time (i.e., injecting a missing dependency in order to bypass its absence, or temporarily test its presence, perhaps in isolation from the rest of the application under test).
- API virtualization, on the other hand, delivers—for all development intents and testing purposes—the same behavior and functionality for any developer or tester who wants to use the API, at any time.
- Thus, once they’re created and exist as part of a projectwide test environment, virtual components eliminate the need for individual developers to write and rewrite their own mocks, saving time and effort for all concerned.

Mocking

- Mocking provides a way to emulate the missing resource, especially APIs, while testing or developing a software. Mocks are very basic setups that are created on code or on tools that can mimic, for example, a few API calls. This allows the development and testing teams to do a very basic level of testing using these emulated calls.
- Generally, mocks are created by development teams in code, and run for their own development and testing. Mock creation for users who are non-developers is difficult: testers often use open source tools for creating mocks – and if they don't have strong coding skills, they'll find this very hard to do. There are not many 'easy to use' open-source tools available for mocking.
- A good example of a tool that allows testers to create mocks is the [open source version of SoapUI](#). However, as it is an all-purpose API testing tool (rather than one focused solely on mocking or virtualization), it doesn't make it easy for testers to scale and adapt mocks to multiple testing scenarios.

10 Best Service Virtualization Tools in 2020: Microservices and Mocking

1. **Traffic Parrot**: makes it easy for developers and testers to do service virtualization, mocking, and simulation. It helps create tests faster and with less effort by providing simulators and mocks of backend APIs and third-party systems.
2. **UP9**: provides an out-of-the-box test automation for microservices, kubernetes and cloud-native, replacing the need for developers to constantly build and maintain tests, while providing comprehensive service test-coverage.
3. **WireMock** is simulator tool for HTTP based API. It allows you stay test even when an API does not exist or is incomplete. It allows checking of an edge case and failure modes that the real API may not able to produce.
4. **Mountebank** is an open source tool which can execute multi-protocol tests. The codebase is Node JS. It is easy to create stubs and mocks.

10 Best Service Virtualization Tools in 2020: Microservices and Mocking

5. Hoverfly cloud is an integrated service virtualization solution. It is designed from the ground up for integration, automation, and performance. You can optimize virtualized services to efficiently handle the load from the system under test.
6. MicroFocus Data simulation software allows developers and QA testers to virtualize micro service's behavior. The tool does not delay delivery regardless of access to production systems.
7. CA Service Virtualization tool simulates unavailable systems across the software development lifecycle. The tool helps developers, QA testing team to work together for faster delivery and higher application quality and reliability.
8. Mocklab is service virtualization tool with user-friendly UI. It allows easy copy, paste or record stubbed HTTP responses. It helps for easy sharing among the team.

10 Best Service Virtualization Tools in 2020: Microservices and Mocking

9. [IBM Rational TestVirtualization](#) offers fast and quick testing in the development lifecycle. It helps to reduce dependencies by simulating part or an entire application. This helps software testing teams as they need not wait for the availability of those applications to begin their work.
10. [Tricentis Tosca](#) allows steady access to dependent systems so that tests can be execute reliably, and continuously. It simulates the dependent component behavior need to run your tests

Benefits to application virtualization

- Requiring fewer resources compared to using a separate virtual machine.
- Allowing incompatible applications to run on a local machine simultaneously.
- Maintaining a standard, more efficient, and cost-effective OS configuration across multiple machines in a given organization, independent of the applications being used.
- Facilitating more rapid application deployment.
- Facilitating security by isolating applications from the local OS.
- Easier tracking of license usage, which may save on license costs.
- Allowing applications to be copied to portable media and used by other client computers, with no need for local installation.
- Increasing ability to handle high and diverse/variable work volume.

What is Microsoft Application Virtualization ?

- Microsoft Application Virtualization (App-V) can make applications available to end user computers without having to install the applications directly on those computers.
- This is made possible through a process known as sequencing the application, which enables each application to run in its own self-contained virtual environment on the client computer.
- The sequenced applications are isolated from each other. This eliminates application conflicts, but the applications can still interact with the client computer.
- Applications are no longer installed on the client—and there is minimal impact on the host operating system or other applications.

Microsoft Application Virtualization

- Applications are rapidly delivered, when needed, to laptops, desktops, and Remote Desktop Servers. In most cases only a small percentage of the application is needed to launch the application.
- Microsoft App-V components include the **App-V Sequencer**, used to virtualize an application, the **App-V client**, installed on end points where App-V applications will execute, and the **App-V Management Server** and the App-V Streaming Server, used to deliver and stream applications to the App-V clients

6: Data Virtualization

6: Data Virtualization

- **Data Virtualization** lets you easily manipulate data, as the data is presented as an abstract layer completely independent of data structure and database systems.
- Data virtualization creates an abstraction layer that brings in data from different sources without performing the entire Extract-Transform-Load (ETL) process or creating a separate, integrated platform for viewing data.
- Instead, it virtually connects to different databases, integrates all the information to provide virtual views, and publishes them as a data service, like REST. This enhances data accessibility, making specific bits of information readily available for reporting, analysis, and decision making
- Decreases data input and formatting errors.

What is Data Virtualization Layer?

- Data virtualization is a logical data layer to integrate enterprise data available across disparate data sources.
- It consolidates data to a single centralized layer by creating a replicated image. This allows the user to alter the source data without accessing it, allowing real-time data access for business operations, while keeping source data secure.
- Businesses nowadays make data virtualization software an integral part of their approach to data management, as it allows complementing processes like data warehousing, **data preparation**, data quality management, and data integration.

Data Virtualization Architecture

Data Virtualization Architecture

Supports Unstructured and Structured Data Sources

 Emails, PDFs, .Docx Files

 REST Web Service, SQL, Salesforce

 Flat, Excel, EXL Files

 SaaS, Cloud & Enterprise Applications

 Databases & Data Warehouses

 Website & Web

A Complete Toolset for Data Virtualization

Data Integration

Logical Abstraction Layer

Data Access Services

Data Extraction & Modeling

Data Caching

In-Memory Data Fabric

Query Optimization

Data Governance & Quality Assurance

Data Management & Security

Robust Data Abstraction With Uninterrupted Access to Real-Time Information

 BI Tools

 Reporting & Analyses

 Web and Mobile Platforms

 Business Users

Applications of Data Virtualization

Businesses can leverage virtualization technology to optimize their systems and operations in several ways, such as:

- **Data Delivery:** It enables you to publish datasets (requested by users or generated through client application) as data services or business data views.
- **Data Federation:** It works in unison with data federation software to provide integrated views of data sources from disparate databases.
- **Data Transformation:** It allows users to apply transformation logics on the presentation layer, thus improving the overall quality of data.
- **Data Movement and Replication:** Data virtualization tools don't copy or move data from the primary system or storage location, saving users from performing extraction processes and keeping multiple copies of inconsistent, outdated data.
- **Virtualized Data Access:** It allows you to break down data stores by establishing a logical data access point to disparate sources.
- **Abstraction:** It creates an abstraction layer that hides away the technical aspects, such as storage technology, system language, APIs, storage structure, and location, of the data

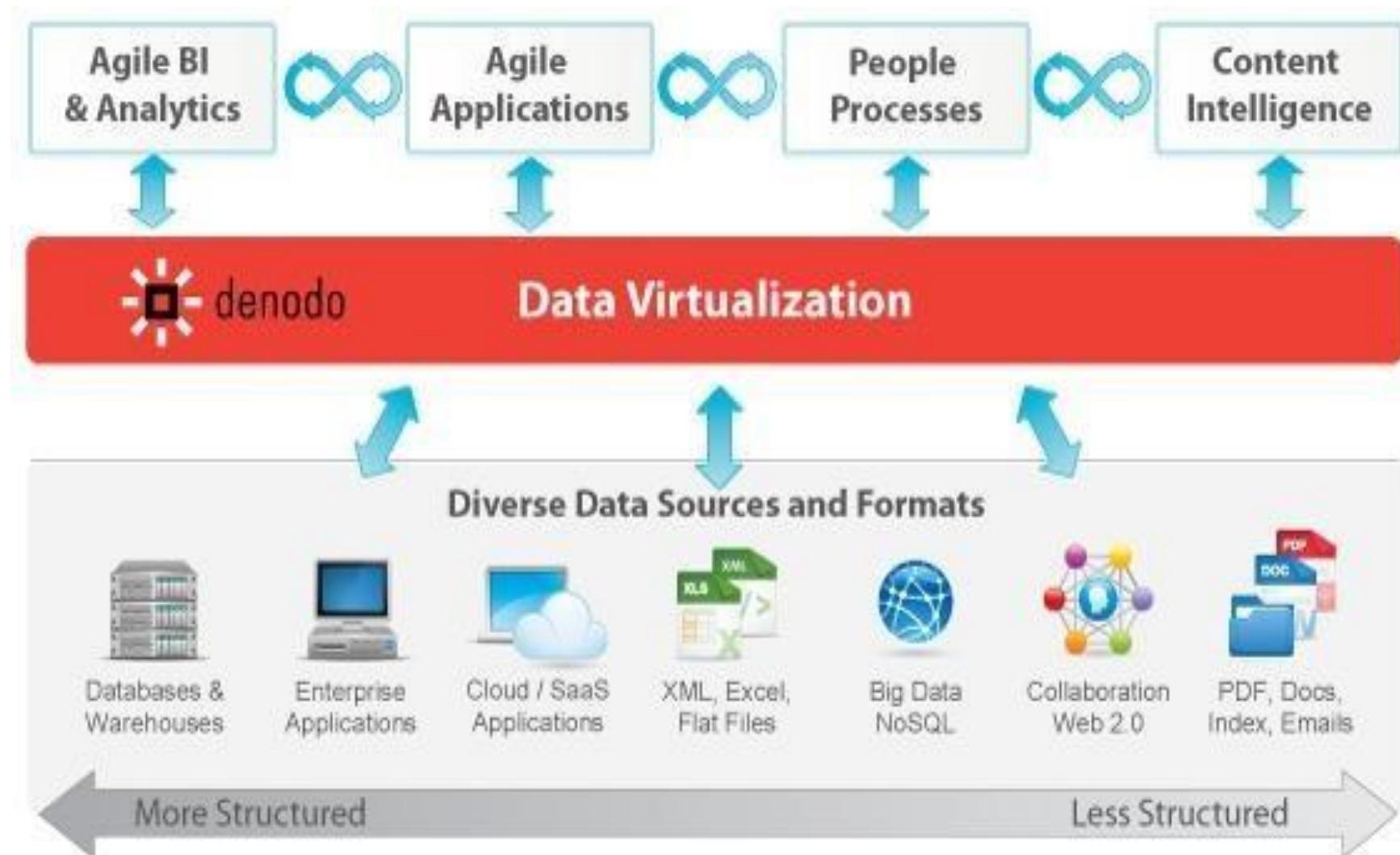
Benefits of Data Virtualization

- According to [Gartner](#), by 2020, about 35 percent of enterprises will make data virtualization a part of their data integration strategy. Here is why enterprises are increasingly opting for tools offering data virtualization platform:
- **Multi-mode and multi-source data access**, making it easy for users at different levels to use data as per their requirements
- **Enhanced security and data governance** for keeping critical data safe from unauthorized users
- **Hiding away the complexity of underlying data sources**, while presenting the data as if it is from a single database or system
- **Information agility**, which is integral in business environments, as data is readily available for swift decision making

Benefits of Data Virtualization

- **Infrastructure agnostic platform**, as it enables data from a variety of databases and systems to be easily integrated, leading to reduced operational costs and data redundancy
- **Simplified table structure**, which can streamline application development and reduce the need for application maintenance
- **Easy integration of new cloud sources to existing IT systems easily**, allowing users to have a complete picture of external and internal information
- **Hybrid query optimization**, enabling you to streamline queries for a scheduled push, demand pull, and other types of data requests
- **Increased speed-to-market**, as it cuts down the time needed to obtain data for improving new or existing products or services to meet consumer demands

Data Virtualization



7: Desktop Virtualization

Desktop virtualization is often referred to as VDI

7: Desktop Virtualization

- **Desktop Virtualization** is perhaps the most common form of virtualization for any regular IT employee.
- The user's desktop is stored on a remote server, allowing the user to access his desktop from any device or location.
- Employees can work conveniently from the comfort of their home. Since the data transfer takes place over secure protocols, any risk of data theft is minimized.



Virtual Desktop Infrastructure (VDI) Software

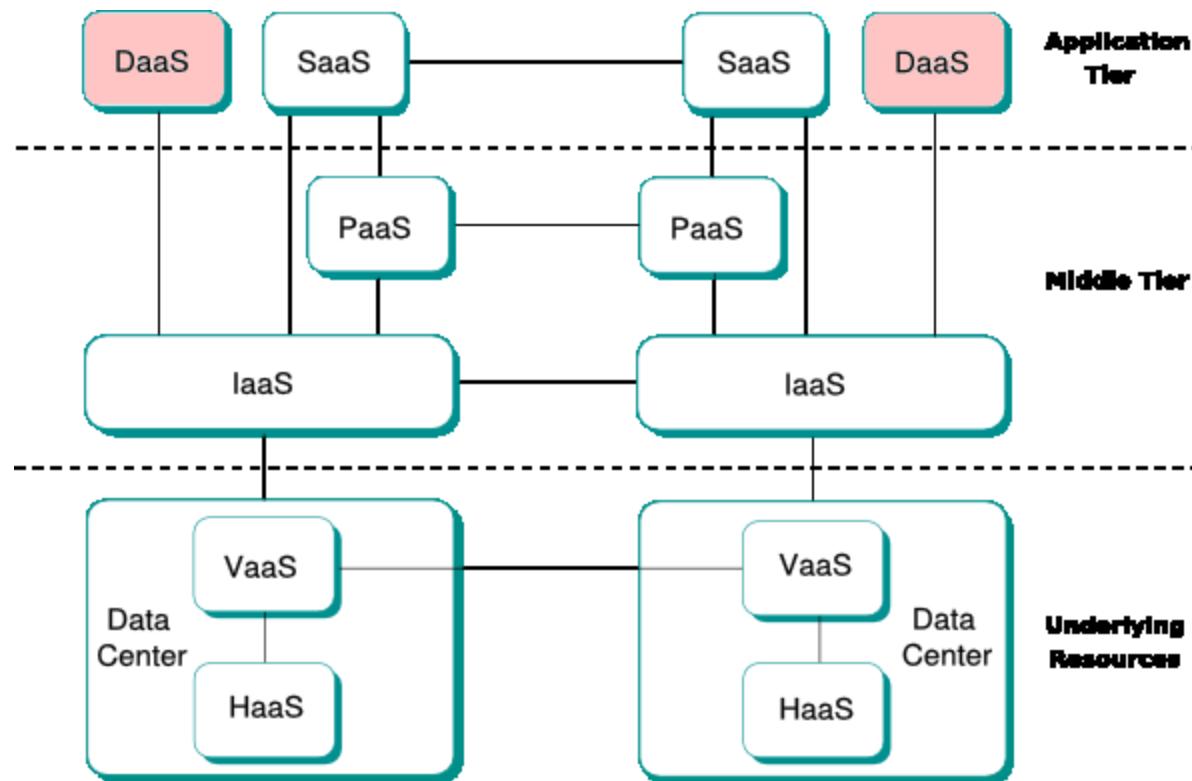
- Virtual Desktop Infrastructure (VDI) is, simply put, desktop virtualization. Desktop (or client) virtualization is like server virtualization but for end-user machines.
- The end user's desktop environment is separated by virtualization from the physical machine where it appears.
- Desktop virtualization is an instance of client-server computing. This is because the virtualized desktop is stored on a central server and not on the machine being virtualized.

Virtual Desktop Infrastructure (VDI) Software

- This enables desktop users to log into their desktop from any machine, like a laptop or home computer.
- In addition to providing flexibility, there are security advantages to client virtualization. For example, if a user's machine is lost or stolen it's a simpler matter for IT to erase company data from the device.
- The biggest difference between server and desktop virtualization is network resource usage.
 - **Server virtualization** achieves better server utilization by making it possible to run multiple virtual machines on a single server. Thus it does not add additional load to a network.
 - **Desktop virtualization**, however, operates entirely on the network. And as client resources are served to client machines across a network, the network's performance can be slower.

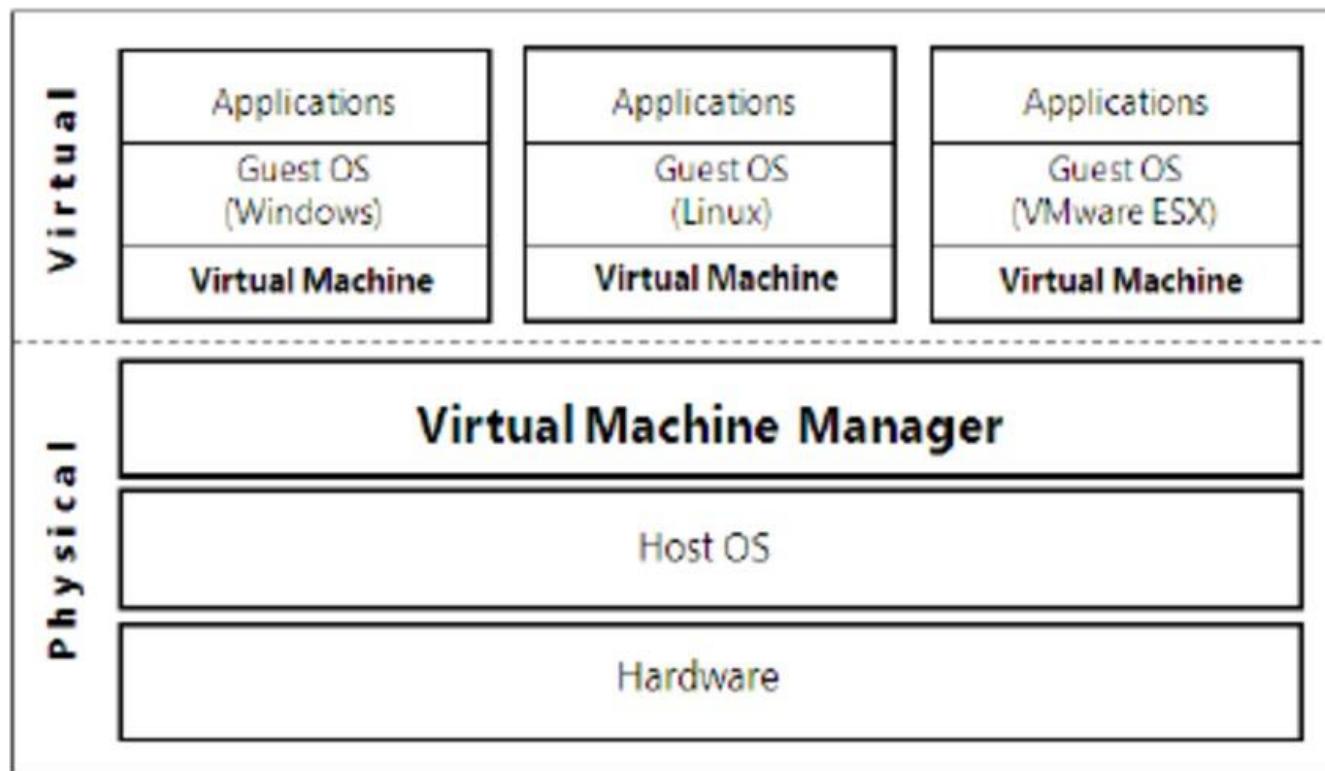
7: Desktop Virtualization

- **Desktop virtualization** is technology that lets users simulate a workstation load to access a **desktop** from a connected device remotely or locally. This separates the **desktop** environment and its applications from the physical client device used to access it.



7: Desktop Virtualization

- Desktop *virtualization* abstracts the desktop environment available on a personal computer in order to provide access to it using a client/server approach. Desktop virtualization provides the same outcome of **hardware virtualization** but serves a different purpose.



7: Desktop and application virtualization

- Desktop **virtualization** is an approach that provides a centralized infrastructure that hosts a desktop image that the workforce can leverage remotely.
- Desktop virtualization is often referred to as VDI, which, depending on the vendor in question, stands for either **virtual desktop** infrastructure or virtual desktop interface.
- As opposed to providing a full **desktop environment**, an organization can simply virtualize key applications that are centrally served.
- Like desktop virtualization, the centralized control associated with **application virtualization** allows the organization to employ strict access control and perhaps more quickly patch the application.
- Additionally, application virtualization can run legacy applications that would otherwise be unable to run on the systems employed by the workforce.

Amazon WorkSpaces

- **Amazon** WorkSpaces is a managed, secure **Desktop-as-a-Service (DaaS)** solution.
- Amazon WorkSpaces having provision for either Windows or Linux desktops in just a few minutes and quickly scale to provide thousands of desktops to workers across the globe.
- You can pay either monthly or hourly, just for the WorkSpaces you launch, which helps you save money when compared to traditional desktops and on-premises VDI solutions.
- Amazon WorkSpaces helps you eliminate the complexity in managing hardware inventory, OS versions and patches, and Virtual Desktop Infrastructure (VDI), which helps simplify your desktop delivery strategy.
- With Amazon WorkSpaces, your users get a fast, responsive desktop of their choice that they can access anywhere, anytime, from any supported device.

Virtual Desktop Infrastructure Software Features & Capabilities

- Management tools for creation, management, and deployment of virtual clients
- Support for a variety of hardware peripherals such as printer, monitors, etc.
- User profile management (e.g credentials & access)
- User environment management (e.g performance monitoring, cloning)
- Dynamic automated allocation of virtual resources
- Support persistent desktop (i.e. user settings personalized or can be retained)
- Support non-persistent desktop (retain pristine settings, easy to update)
- Optimization for mobile platform deployment
- Optimization of storage and memory allocation
- Network monitoring

Virtual Desktop Infrastructure Advantages & Disadvantages

Advantages

- Maintaining a single OS image will minimize costs towards administration and support.
- Administrative tasks decrease as all operating system and application updates are handled at once.
- Security is a major plus with VDI as all licensing and program downloads are centralized.

Disadvantages

- Increased network requirements depending on the nature of your business (simple word processing tools versus more graphics and memory intense environments).
- Changing from decentralized to centralized licensing does take quite a bit of time to adjust to, and while cost savings may come down the road, startup costs can be high.
- A lack of end-user privacy. Not all users or employees will be thrilled to have some of their privacy taken away due to the nature of VDI.

20 Best Virtual Desktop Infrastructure Software Solutions in 2020

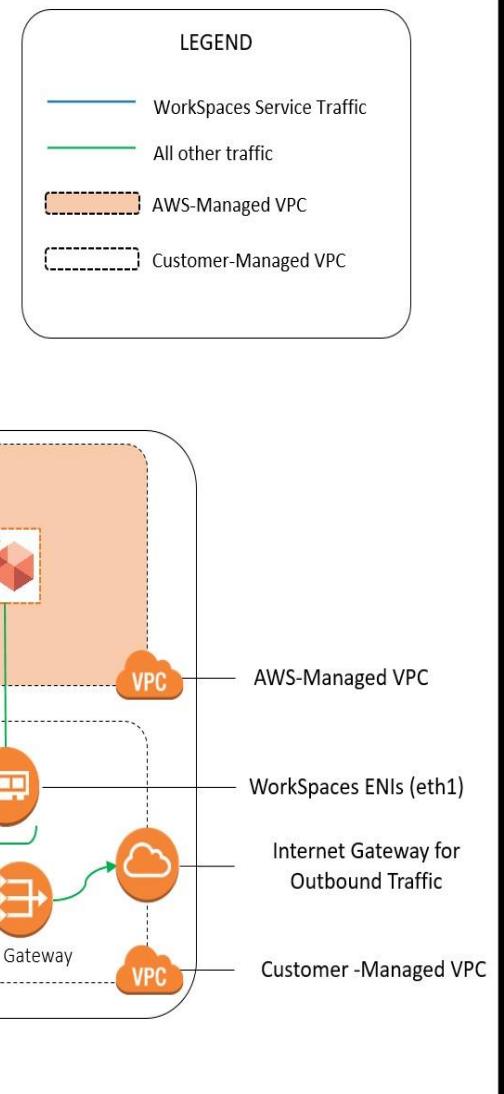
- [Amazon WorkSpaces](#)
- [IBM Cloud](#)
- [Cisco VXI](#)
- [VMware Horizon Cloud](#)
- [Red Hat Virtualization](#)
- [Citrix Virtual Apps & Desktops](#)
- [SolarWinds Virtualization Manager](#)
- [Parallels RAS](#)
- [Nutanix VDI](#)
- [Vagrant](#)
- [Paperspace](#)
- [Nvidia Virtual GPU](#)
- [Evolve IPVDI](#)
- [Xen Project](#)
- [Neverfail Workspaces](#)
- [Nerdio Private Cloud](#)
- [Teradici](#)
- [IGELUDC](#)
- [Virtuozzo](#)
- [Liquidware FlexApp](#)

Example: Amazon Workspaces

- **Amazon WorkSpaces**, a managed, secure cloud desktop service, heads the giant names in our top 20 best virtual desktop infrastructure software.
- One can use Amazon WorkSpaces to provision either Windows or Linux desktops in just a few minutes and quickly scale to provide thousands of desktops to workers across the globe.
- One can pay either monthly or hourly just for the WorkSpaces you launch. This helps you save money when compared to traditional desktops and on-premise VDI solutions.
- Among its most prominent features is the elimination of many administrative tasks associated with managing your desktop lifecycle, including provisioning, deployment, maintenance, and recycling of desktops.
- There is less hardware inventory to manage and no need for complex virtual desktop infrastructure deployments that don't scale.

Example: Amazon Workspaces

Amazon WorkSpaces Architectural Diagram



The difference between virtualization and cloud computing

- The difference between virtualization and cloud computing is to say that the former is a technology, while the latter is a service whose foundation is formed by said technology.
- Virtualization can exist without the cloud, but cloud computing cannot exist without virtualization – at least, not in its current format.
- The term cloud computing then is best used to refer to a situation in which “shared computing resources, software, or data are delivered as a service and on-demand through the Internet.”
- *Example: A server environment which lacks any of these features, is not cloud computing.*

The side effects of virtualization

Seven disadvantages of server virtualization

- David Coyle, research vice president at Gartner, detailed the seven side effects at the research firm's Infrastructure, Operations and Management Summit, which drew nearly 900 attendees.
- Virtualization promises to solve issues such as underutilization, high hardware costs and poor system availability, the benefits come only when the technology is applied with proper care and consistently monitored for change.

7 side effects of sloppy virtualization

1. Magnified physical failures
2. Degraded performance
3. New skills
4. Complex root cause analysis
5. New management tools
6. Virtual machine sprawl
7. Virtual habits

The side effects of virtualization

- There are side effects of virtualization, notably the ***performance penalty*** and ***the hardware costs***. As we shall see shortly, *all privileged operations of a virtual machine must be trapped and validated by the Virtual Machine Monitor which, ultimately, controls the system behavior*; the increased overhead has a negative impact on the performance.
- The cost of the hardware for a virtual machine is higher than the cost for a system running a traditional operating system because the physical hardware is shared among a set of guest operating systems and it is typically configured with faster and/or multi-core processors, more memory, larger disks, and additional network interfaces as compared with a system running a traditional operating system



Thanks for Your Attention!

References

- <http://www.vfrank.org/2013/09/18/understanding-vmware-ballooning/>
- <https://www.vmware.com/cloud-solutions.html>
- <https://www.vmware.com/topics/glossary/content/application-virtualization>
- <https://smartbear.com/learn/software-testing/what-is-service-virtualization/>
- <https://www.guru99.com/service-virtualization-tools.html>
- <https://www.soapui.org/learn/mockng/what-is-api-virtualization/>
- <https://dzone.com/articles/making-mainframe-agile-in-continuous-testing-with-1>
- <https://financesonline.com/virtual-desktop-infrastructure/>

Exercises

Exercises and Problems

1. Virtualization simplifies the use of resources, isolates users from one another, supports replication and mobility, but exacts a price in terms of performance and cost. Analyze each one of these aspects for: (i) memory virtualization, (ii) processor virtualization, and (iii) virtualization of a communication channel.
2. Virtualization of the processor combined with virtual memory management pose multiple challenges; analyze the interaction of interrupt handling and paging
3. What are the different types of Virtualization in Cloud Computing? What is the necessity of data virtualization? What is Data Virtualization Layer? Discuss the benefits of Data Virtualization?
4. Discuss the 7-Layer virtualization model in the context of cloud computing? What is software virtualization? What are the benefits of software virtualization?

VM Migration

Live virtual machine migration

- The virtualization technology plays a pivotal role in CDC resources management and provides an enchanting feature of virtual machine (VM) migration which provides several benefits in terms of VM scheduling, fault tolerance, load balancing, energy efficiency, power management, and security.
- The advent of VM migration technology resolves server overutilization and performance degradation problems by enabling the migration of VMs between the servers residing within or across the data centers.
- In VM migration, the hypervisor relieves the overutilized servers by migrating its workload to an underutilized or normal utilized server.
- The VM migration may require additional resources (such as energy, network bandwidth, and computational resources) and may affect the applications within the migrant VM until the migration process completes.
- Therefore, to maintain the application performance, it is quite important to complete the migration process within a minimal time duration while utilizing the minimum network and server resources .

Cloud computing

- *Software as a service (SaaS)* delivers the application(s) over the Internet as a service (e.g., Google Apps, Salesforce, Cisco Webex, Yahoo, Gmail, Hotmail, Netsuite, Zoho, Slack, and Hubspot, to name a few). Instead of installing and maintaining software, the users simply access the applications over the Internet regardless of any complex hardware and software management.

- *Platform as a service*

(*PaaS*) refers to a cloud computing paradigm that allows a third-party provider(s) to deliver hardware and software resources such as Google App Engine, Windows Azure, VM Ware Cloud, Force.com, CloudFoundry, Roll base, and OpenShift, to its users over the internet. These resources are mostly required for application development purposes.

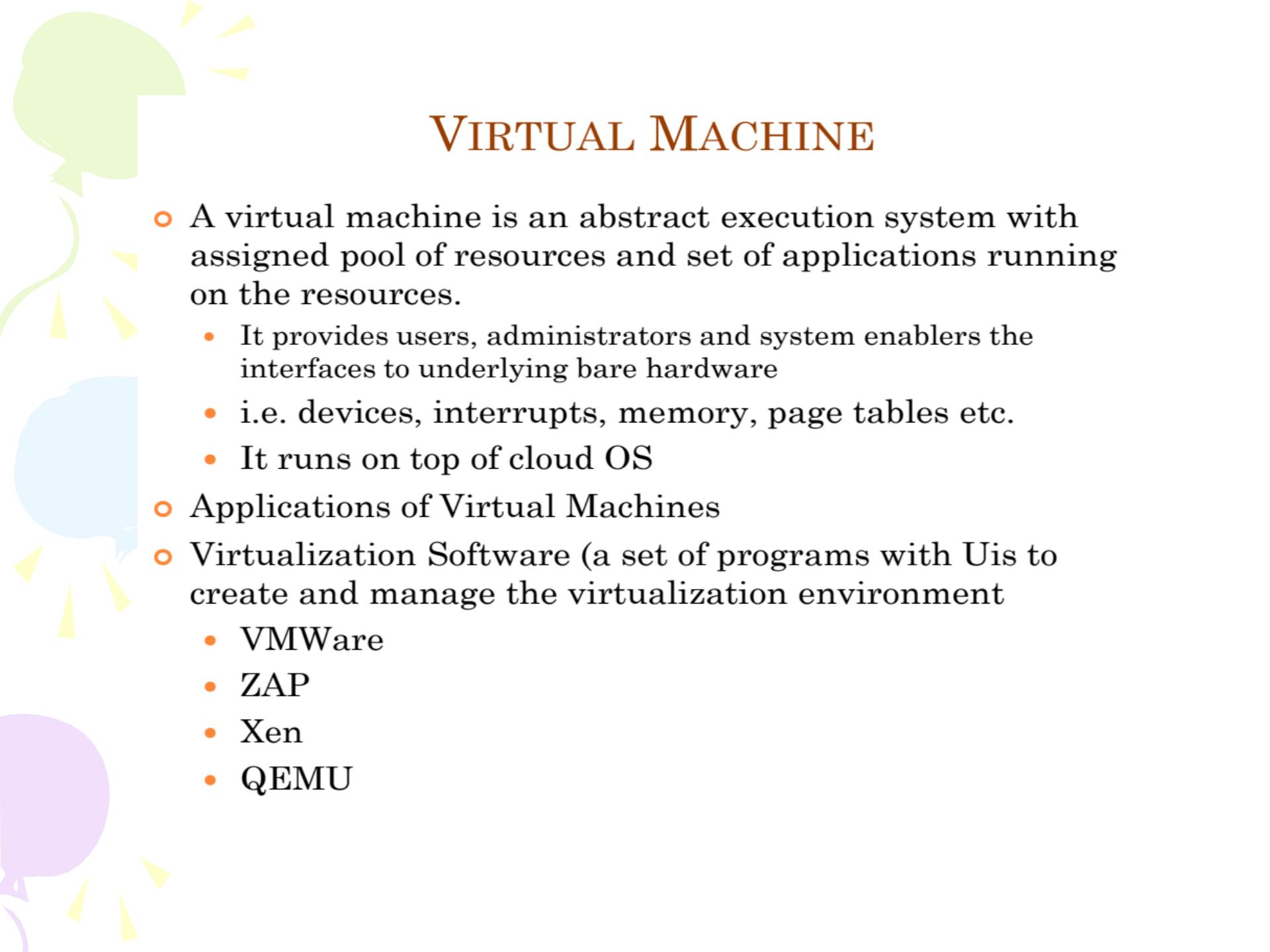
- *Infrastructure as a service (IaaS)*

Corresponds to a cloud computing service that provides the essential computing, storage, and networking resources on-demand (e.g., pay-as-you-go).

Amazon Web Services (AWS), Google Cloud Storage, VM Ware, Rackspace, JoyNET, and GoGrid are some popular examples of IaaS.

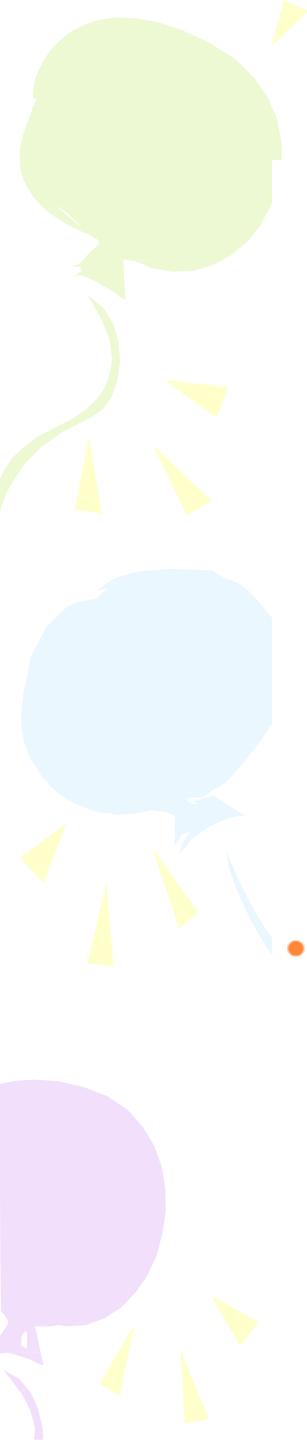
Cloud data center (CDC)

- Cloud data centers (CDC) are comprised of heterogeneous clustered computational and storage resources that enable a huge amount of data storage and host applications deployment.
- CDC provides the resources over the Internet and charges its users as per the resource usage against a specific period. The major types of CDC include (1) public CDC, (2) private CDC, and (3) hybrid CDC.
- In **public CDC**, the services are offered by third-party providers (e.g., Microsoft Azure). These services may be offered free or can be purchased on-demand and allow the consumers/users to pay as per their utilization of bandwidth, storage, and CPU cycles for computation.
- The **private CDC** serves only authorized users over the Internet or an internal network. This means that these types of CDC are only utilized by a specific/group of users, organizations, and businesses. Microsoft Azure Stack, Elastra-private cloud, and HP Data Centers are some common examples of a private CDC.
- **Hybrid CDC** refers to the combination of computational, storage, and networking services provided by both the private and public CDC. These CDCs allow application migration between public and private CDCs.
- In hybrid CDC, organizations own private CDC and also obtain services of public CDC according to resource requirements. The hybrid CDCs are secure as compared to the public CDC due to confidential information (e.g., passwords, sensitive data) being stored in a private DC, and computational and storage services are rented on the public CDC on demand.
- Amazon Web Services (AWS) and Microsoft Azure (with orchestration amongst the different platforms) fall under the hybrid CDC category.



VIRTUAL MACHINE

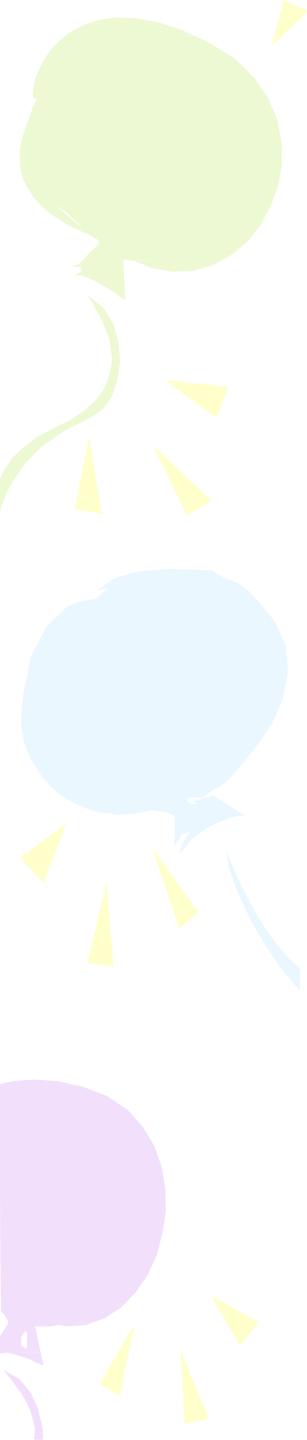
- A virtual machine is an abstract execution system with assigned pool of resources and set of applications running on the resources.
 - It provides users, administrators and system enablers the interfaces to underlying bare hardware
 - i.e. devices, interrupts, memory, page tables etc.
 - It runs on top of cloud OS
- Applications of Virtual Machines
- Virtualization Software (a set of programs with UIs to create and manage the virtualization environment)
 - VMWare
 - ZAP
 - Xen
 - QEMU



VIRTUAL MACHINES IN CLOUD

- Virtualization helps making efficient use of hardware resources
- Facilitates a greater degree of abstraction
- Seamless transfer of applications from one piece of hardware to another
- Replication of virtual machines as per need
- Create more scalable and flexible infrastructures
- Snapshots
- degree of efficiency and agility realized from virtualization
 - Pooled resources
 - Geographic diversity
 - Universal connectivity

- What is migration
- Why it is used?
- Different types of migration? (live and dead)
- How it is Implemented? (pre-copy, post-copy)



VM MIGRATION

- An increase in workload can be handled by allocating more resources to it, if idle resources are available
- Main Issues:
 - What if PM does not have (enough or no) idle resources to satisfy VM's requirement?
 - Performance of the application degrades
 - SLA violation occurs
- Key Ideas
 - Migrating VMs

Virtualization

- Virtualization is a key component of cloud computing which refers to the process of simultaneously running several instances of the same or different operating systems (OS) – often referred to as guest OS – over the same hardware, where each instance appears as it is executing on dedicated hardware and does not interfere with the operations of either the host OS or other guest OS.
- Virtual Machine Manager (VMM) or hypervisor is one of the hardware virtualization technology that allows multiple OS to concurrently run on the host machine. Furthermore, the hypervisor also facilitates VM migration — the process of migrating VM from one physical host to another within or across the CDC.

VM migration: An overview

- The process of migrating VM within or across the CDC in order to achieve efficient resource utilization and yield better performance is referred to as VM migration [7]. A high-level overview of VM migration is depicted in [Fig. 2](#).
- This figure shows the VM migration use case within the data center where networked attached storage is available and disk storage does not need to migrate, and simply the VMs for instance VM1 and VM2 are being migrated from host 1 to host 2, and VM7 and VM8 are being migrated from host 4 to host 3. The idle hosts, for instance, host 1 and host 4 are simply turned off to save energy and resources.
- The VM migration process, in general, comprises the following phases:(1) VM suspension at a host (2) copying of the VM (along with CPU, memory, and network states) to the destination host, and (3) VM resumption at the destination.
- VM migration has several benefits such as load balancing, fault tolerance, system maintenance, resource sharing, and power management briefly defined as follows.

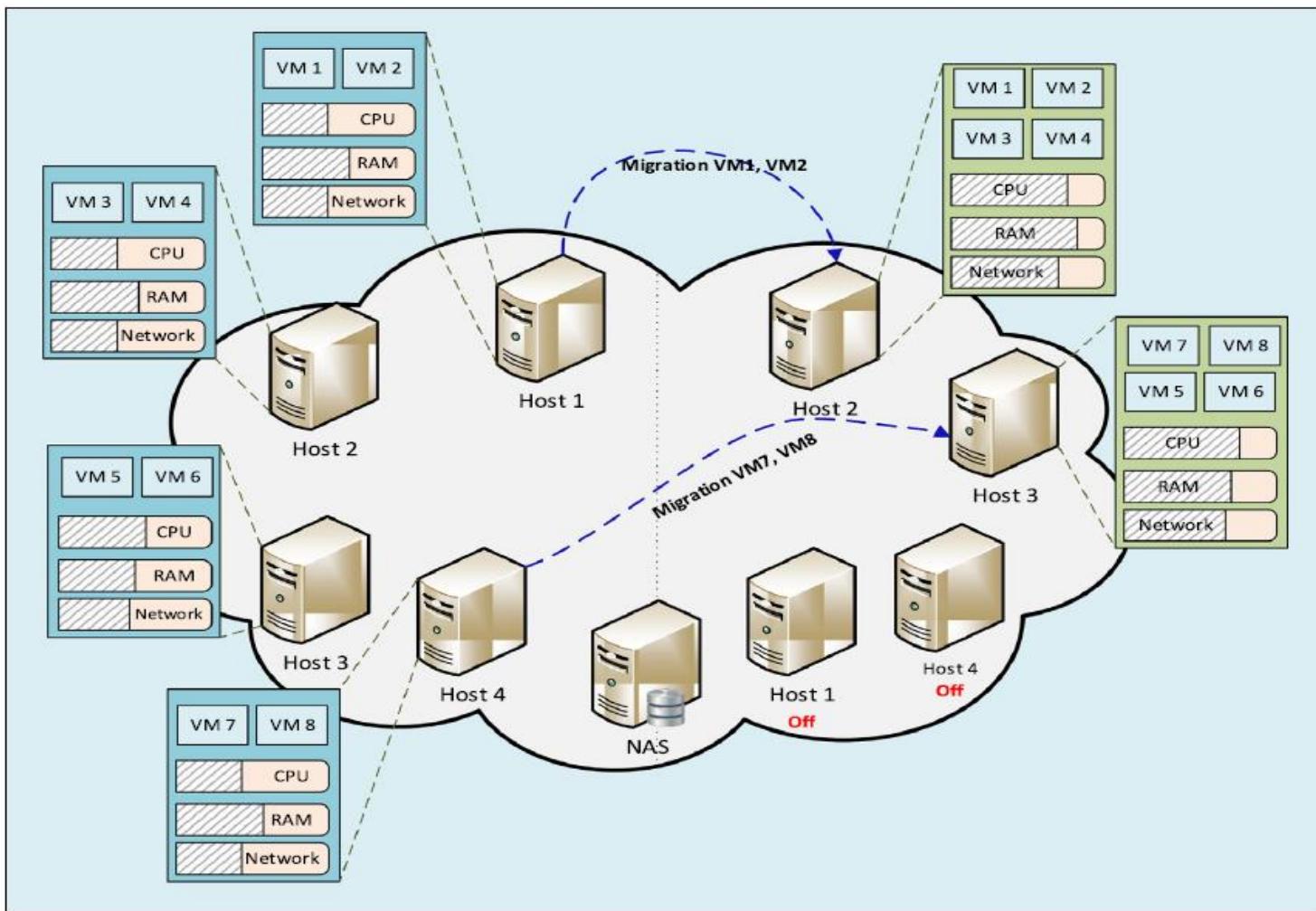


Fig. 2. VM migration overview.

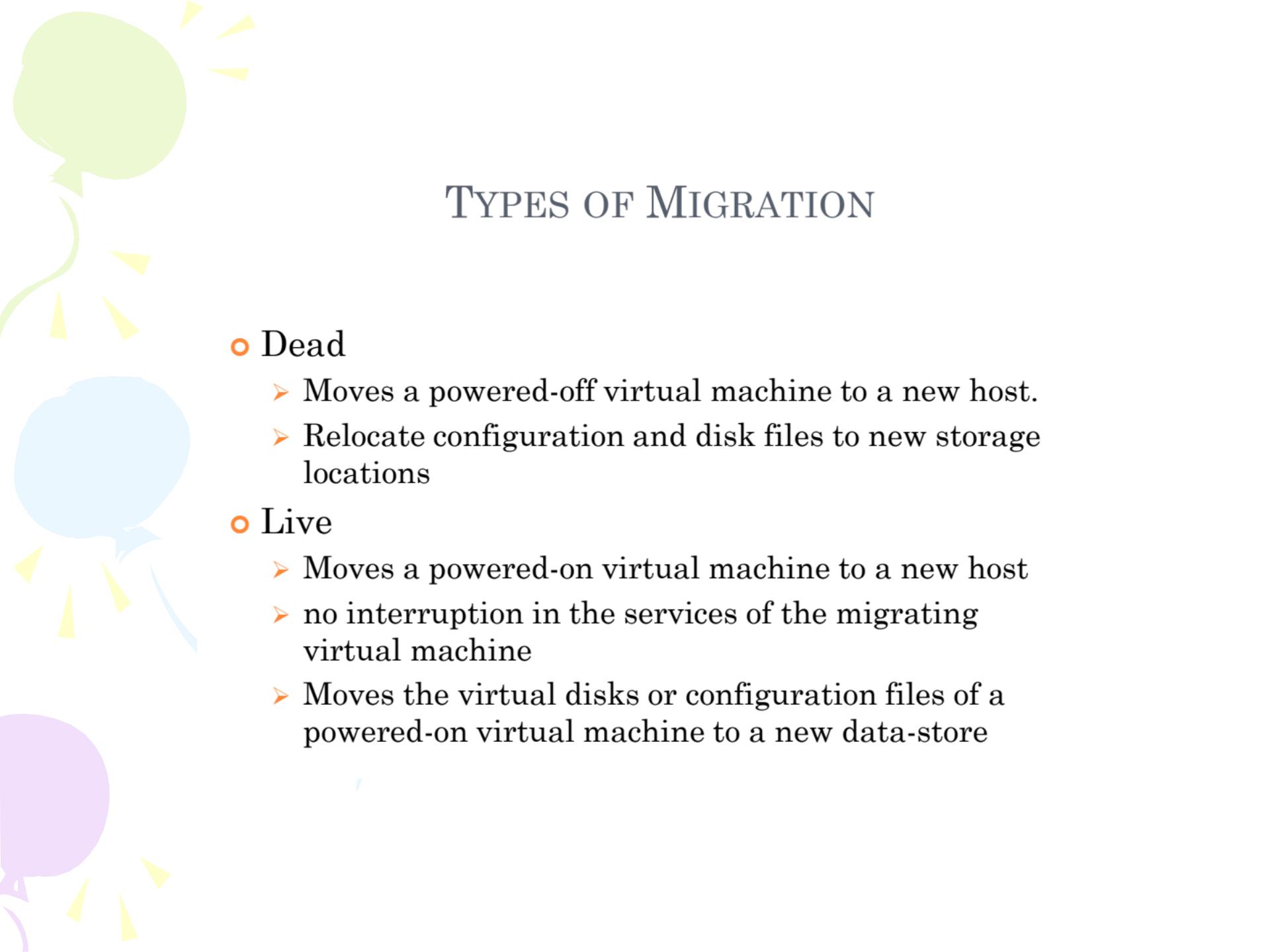
VM migration benefits

- *Load balancing*: Load-balancing is a technique that distributes workload by migrating VMs from overutilized host to underutilized host to avoid system failure as well as improve system performance.
- *Fault tolerance*: Fault tolerance is the ability of a system to trigger VM migration to continue its ongoing VM operation without any interruption whenever a malfunction occurs in one or more components of the system. Once the system maintenance is finalized, the migrated VMs sent back to the original servers.
- *System maintenance*: A periodic maintenance of the hosts is required to avoid service degradation and to increase the host lifetime. During the maintenance phase, the VMs are migrated to ensure seamless and low latency service execution.
- *Power Management*: To attain power efficiency, VM migration helps to shift workload from underutilized hosts to the hosts with available resources and switch off the idle hosts.

VM migration types

The VM migration is carried out in two manners named **non-live VM migration** and **live VM migration** [1]. The short description of both mechanisms is as follows.

- **Non-live migration:**
- The VM requires to turn off before the migration starts. Following the non-live VM migration process, VM execution is not resumed until the VM is transferred completely at the destination host, which highly degrades the overall QoS (e.g., in interactive web applications) due to service discontinuity issues.
- Although the non-live VM migration techniques ensure predictable migration time and support one-time migration of all the VM memory pages during the VM migration process, these techniques are being obsoleted due to the lack of support for uninterrupted services delivery to end-users.



TYPES OF MIGRATION

- Dead

- Moves a powered-off virtual machine to a new host.
- Relocate configuration and disk files to new storage locations

- Live

- Moves a powered-on virtual machine to a new host
- no interruption in the services of the migrating virtual machine
- Moves the virtual disks or configuration files of a powered-on virtual machine to a new data-store

Live migration:

- Live VM migration continues serving the running applications to the end-users during the VM migration process.
- The core objectives of live VM migration include application performance optimization during VM migration, efficient bandwidth utilization, and downtime reduction.
- Live migration can further be classified into pre-copy, post-copy, and hybrid-copy migration.

- *Pre-copy migration*, without affecting the VM under execution, the hypervisor copies the original memory pages of the running VM from the source to the destination host. After copying the memory pages, the VM is suspended at the source, migrated, and resumed at the destination host.
- *Post-copy migration*, suspends the VM at the source host and migrates minimum state information such as CPU and registers which are required to resume VM at the destination is migrated on priority. The remaining memory pages are migrated as per the request of the destination host.
- *Hybrid-copy migration*, combines the characteristics of both pre-copy and post-copy migration schemes and operates in five phases: (i) migration preparation phase in which the required resources at the destination host are reserved, (ii) bounded pre-copy rounds: this phase identifies VM working set and forwards it to the destination, (iii) VM state transfer: In this phase minimum state of VM is recorded and transferred toward the destination, (iv) the VM resume phase which resumes the transferred VM at the destination and finally (v) demand paging phase in which the VM faulty pages requested by the destination are forwarded by the source host to continue the VM execution and synchronize with the source VM.

- The hybrid-copy migration overcomes the shortcomings of pre-copy and post-copy migration techniques.
- For instance, the hybrid-copy migration copies all memory pages once only in advance to the destination (pre-copy technique), and then VM is suspended at the source host and its processor state is transferred to the destination host and VM is immediately resumed at the destination host (post-copy technique), and the remaining memory pages are transferred to the destination by the post copy method.
- As a result, hybrid-copy avoid the many page faults which result in reduced VM downtime and migration time as compared with pre-copy and post-copy migration techniques.

Performance metrics

- **Migration time.** It refers to the total time required in the migration process. Meaning that the total time it takes from the initiation of migration at the host server to the complete reception of VM at the destination server.
- Migration time depends on the size of memory pages to be transferred and the allocated link speed.
- **Downtime.** The time when the VM service(s) are unavailable due to processor states migration refers to the downtime. The downtime depends on the dirty page(s) rate, page(s) size, duration of the last pre-copy round, and bandwidth or link speed. The lack of consideration of dirty memory page management increases the downtime.
- **Network traffic.** Network traffic reflects the amount of data transferred during the VM migration process. Given the bandwidth constraints, the network traffic should be minimized to effectively carry out the VM migration process.
- **Quality of Service.** QoS refers to the response time and throughput achieved by the users. An efficient VM migration process provides high throughput and low response time.

Classification of live VM migration schemes

- 1. *Load balancing aware live VM migration*
- 2. *Energy-aware live VM migration*
- 3. *SLA aware live VM migration*
- 4. *Network and bandwidth-aware live VM migration*

Live VM migration cost

- Live VM migration is a costly process due to (i) the number of CPU resources it takes at the source host (i.e., computational cost), (ii) energy consumption for migration preparation and migration completion (i.e., energy cost), (iii) network bandwidth between the source and destination host to perform migration (i.e., network cost), (iv) the VM memory content size and the memory content update rate, (v) the number of VM migration, (vi) available network bandwidth for migration, and (vii) the source, and destination host workload at migration time.
- Generally, most live VM migration schemes follow three steps to perform a migration including
 - (i) overloaded host detection and VM selection,
 - (ii) Underloaded host detection and VM selection
 - (iii) Optimal destination host selection and VM placement.
- The total migration cost depends on all three abovementioned steps' complexity.

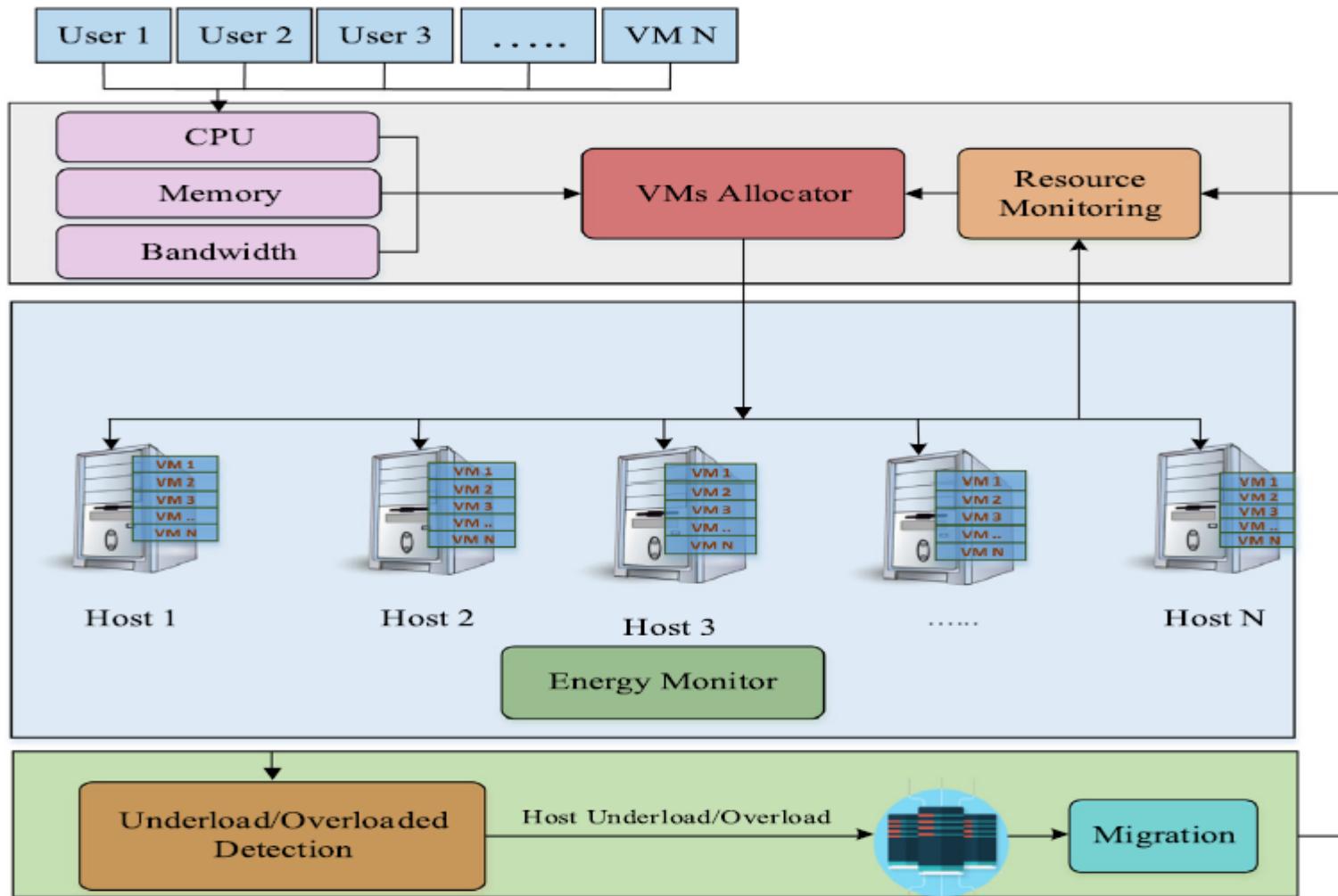


Fig. 4. Conventional energy-aware live VM migration.

Contd.

- As shown in Fig. 4, the users request the services provided by the VMs. As per the user's demand for the resources such as CPU, RAM, and bandwidth, the VM is allocated to PMs in the DC. After VMs allocation, the energy monitoring system continuously monitors the energy consumption of the system.
- Furthermore, the migration algorithm is employed that detects the hosts' underutilized and overutilized conditions, migrates VMs — if required, from the overutilized hosts to the lightly loaded hosts, and finally shut down

Live VM migration challenges and future research directions

- *The rise of AI in live VM migrations*
- *VM migration over WAN*
- *Resources availability*
- *Multiple and correlated VM migration*
- *VM consolidation*
- *Workload prediction*
- *Secure VM migration*

Advantages and the drawbacks of the AI-based techniques

- AI-based VM migration techniques ensure autonomous cloud resource management (e.g., CPU, RAM, network).
- The AI-based migration techniques automate repetitive and complex tasks and devise intelligent decision-making by deeply inspecting the trends and patterns of the large data in VM migration decisions.
- The AI-based migration techniques consider the previous history of VM, host, and cloud consumer resources and devise migration decisions accordingly.
- The AI-based VM migration techniques perform better even in unforeseen workload conditions such as RL improves the migration decisions at runtime based on the current resources requirements and workload status on the server and the VM.
- AI attempt to keep VM migration secure with intelligent solutions (e.g., Amazon GuardDuty tools that can find potential risk).

- AI-based migration techniques required a large amount of historical data for model training.
- AI-based migration techniques are resource intensive, and require high computational resources such as CPU and GPU for efficient model training.
- VM virtual resources (e.g., CPU, RAM) prediction is difficult due to the dynamic requirements of consumer applications.

Therefore, AI-based migration technique design based on workload prediction is highly challenging.

- AI-based migration techniques attempt to optimize resource utilization via VM migration. However, unnecessary VM migration based on an AI-trained model (e.g., may depend on false workload prediction due to less amount of historical data availability) leads to inefficient resource utilization.
- The network behavior prediction is challenging due to the dynamic network resource demands by cloud applications. In such scenarios, AI-based migration techniques may lead to an SLA breach for other network-intensive executing applications in CDC.

VM consolidation

- VM consolidation aims to combine multiple VMs on a few hosts to save energy by turning off the idle/free hosts.
- To achieve VM consolidation, the existing schemes utilize CPU computation to identify the overutilized and underutilized hosts.
- Detecting the host overutilization based on CPU computation while ignoring the other parameters such as memory and disk storage may lead toward the increased migration frequency and degrade the overall performance of the system.
- Therefore, it is important to consider memory and disk storage along with CPU utilization in host overutilization detection.
- On the other hand, to find the hosts' overutilized / underutilized state, it is important to observe the hosts' historical resource utilization (a host may reach an overutilized /underutilized state for a short period of time).
- In the case of a false host overutilized/underutilized state detection, the CDC may observe a high frequency of VM migration.

• VM Provisioning Process

- The common and normal steps of provisioning a virtual server are as follows:
- Firstly, you need to **select a server** from a pool of **available servers** (physical servers with enough capacity) along with the appropriate OS template you need to provision the virtual machine.
- Secondly, you need to **load the appropriate software** (operating System you selected in the previous step, device drivers, middleware, and the needed applications for the service required).
- Thirdly, you need to **customize and configure the machine** (e.g., IP address, Gateway) to configure an associated network and storage resources.
- Finally, the **virtual server** is ready to start with its newly loaded software.

- **VM Provisioning Process contd.**

To summarize, **server provisioning** is **defining server's configuration** based on the organization requirements, a hardware, and software component (processor, RAM, storage, networking, operating system, applications, etc.).

- Normally, virtual machines can be provisioned by manually installing an operating system, by using a preconfigured VM template, by cloning an existing VM, or by importing a physical server or a virtual server from another hosting platform.
- Physical servers can also be virtualized and provisioned using P2V (Physical to Virtual) tools and techniques (e.g., virt-p2v).
- After creating a virtual machine by virtualizing a physical server, or by building a new virtual server in the virtual environment, a template can be created out of it.
- Most virtualization management vendors (VMware, XenServer, etc.) provide the data center's administration with the ability to do such tasks in an easy way.

- **VM Provisioning Process contd.**

- Provisioning from a template is a valuable feature, because it reduces the time required to create a new virtual machine.
- Administrators can create different templates for different purposes. For example, you can create a Windows 2003 Server template for the finance department, or a Red Hat Linux template for the engineering department. This enables the administrator to quickly provision a correctly configured virtual server on demand.

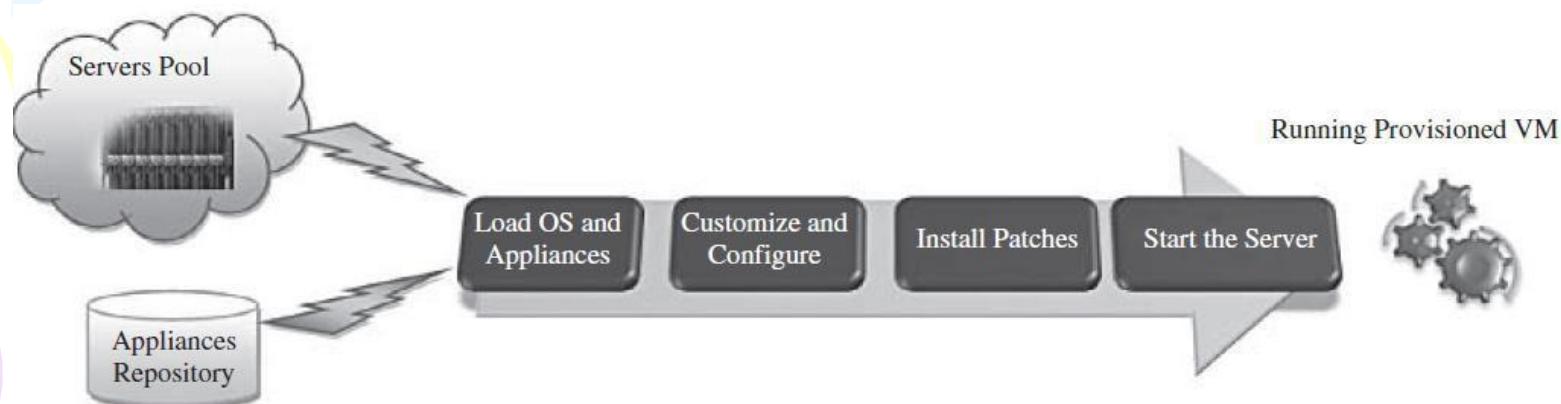


FIGURE Virtual machine provision process.

- **VIRTUAL MACHINE MIGRATION SERVICES (Live Migration and High Availability)**

- **Live migration** (which is also called **hot or real-time migration**) can be defined as the **movement** of a virtual machine from one physical host to another while being powered on.
- When it is properly carried out, this process takes place without any noticeable effect from the end user's point of view (**a matter of milliseconds**).
- One of the most significant advantages of live migration is the fact that **it facilitates proactive maintenance in case of failure**, because the potential problem can be resolved before the disruption of service occurs.
- Live migration can also be used for **load balancing** in which work is shared among computers in order to optimize the utilization of available CPU resources.

Live Migration Anatomy, Xen Hypervisor Algorithm.

- How to live migration's mechanism and memory and virtual machine states are being transferred, through the network, from one host A to another host B:
- The Xen hypervisor is an example for this mechanism. The logical steps that are executed when migrating an OS.
- In this research, the migration process has been viewed as a transactional interaction between the two hosts involved:

- **LIVE MIGRATION STAGES**

Stage-0: Pre-Migration. An active virtual machine exists on the physical host A.

Stage-1: Reservation. A request is issued to migrate an OS from host A to host B (a precondition is that the necessary resources exist on B and a VM container of that size)

Stage-3: Stop-and-Copy. Running OS instance at A is suspended, and its network traffic is redirected to B. As described in reference 21, CPU state and remaining inconsistent memory pages are then transferred. At the end of this stage, there is a consistent suspended copy of the VM at both A and B. The copy at A is considered primary and is resumed in case of failure.

Stage-4: Commitment. Host B indicates to A that it has successfully received a consistent OS image. Host A acknowledges this message as a commitment of migration transaction.

Stage-5: Activation. The migrated VM on B is now activated. Post-migration code runs to reattach the device's drivers to the new machine and advertise moved IP addresses.

This approach to failure management ensures that at least one host has a consistent VM image at all times during migration:

- 1) Original host remains stable until migration commits and that the VM may be suspended and resumed on that host with no risk of failure.
- 2) A migration request essentially attempts to move the VM to a new host and on any sort of failure, execution is resumed locally, aborting the migration.

• LIVE MIGRATION TIMELINE (Pre – copy)

VM running normally on Host A

Stage 0: Pre-Migration

Active VM on Host A

Alternate physical host may be preselected for migration

Block devices mirrored and free resources maintained

Stage 1: Reservation

Initialize a container on the target host

Overhead due to copying

Stage 2: Iterative Pre-copy

Enable shadow paging

Copy dirty pages in successive rounds.

Downtime
(VM Out of Service)

Stage 3: Stop and copy

Suspend VM on host A

ARP = Address Resolution Protocol

Generate ARP to redirect traffic to Host B

Synchronize all remaining VM state to Host B

Stage 4: Commitment

VM state on Host A is released

VM running normally on Host B

Stage 5: Activation

VM starts on Host B

Connects to local devices

resumes normal operation

FIGURE

Live migration timeline

Post-copy

Post-Copy

Suspend the VM at Source

(a minimal subset of the execution state of the VM is transferred to the target)



VM is then resumed at the target
(Concurrently, the source actively pushes the remaining memory pages of the target VM- Pre-paging)



Page- Fault: Demand Paging

- **LIVE MIGRATION VENDOR IMPLEMENTATION EXAMPLE**

There are lots of VM management and provisioning tools that provide the live migration of VM facility, two of which are VMware VMotion and Citrix XenServer “XenMotion”.

VMware VMotion:

- a) Automatically optimize and allocate an entire pool of resources for maximum hardware utilization, flexibility, and availability.
- b) Perform hardware's maintenance without scheduled downtime along with migrating virtual machines away from failing or underperforming servers.

Citrix XenServer “XenMotion”:

Based on Xen live migrate utility, it provides the IT Administrator the facility to move a running VM from one XenServer to another in the same pool without interrupting the service (hypothetically zero – downtime server maintenance), making it a highly available service and also good feature to balance workloads on the virtualized environments.

- **REGULAR / COLD MIGRATION**

- **Cold migration is the migration of a powered-off virtual machine.**
- You have options of moving the associated disks from one data store to another.
- The virtual machines are not required to be on a shared storage.
 - 1) Live migrations needs to a shared storage for virtual machines in the server's pool, but cold migration does not.
 - 2) In live migration for a virtual machine between two hosts, there should be certain CPU compatibility checks, but in cold migration this checks do not apply.
- Cold migration (VMware product) is easy to implement and is summarized as follows:
- The configuration files, including NVRAM file (BIOS Setting), log files, and the disks of the virtual machines, are moved from the source host to the destination host's associated storage area.
- The virtual machine is registered with the new host.
- After the migration is completed, the old version of the virtual machine is deleted from the source host.

Data center in Cloud Computing

ARADHANA BEHURA

Communication & Computing Group

Email: 921CS5007@nitrkl.ac.in, 7787821733

Physical Server vs Cloud Server

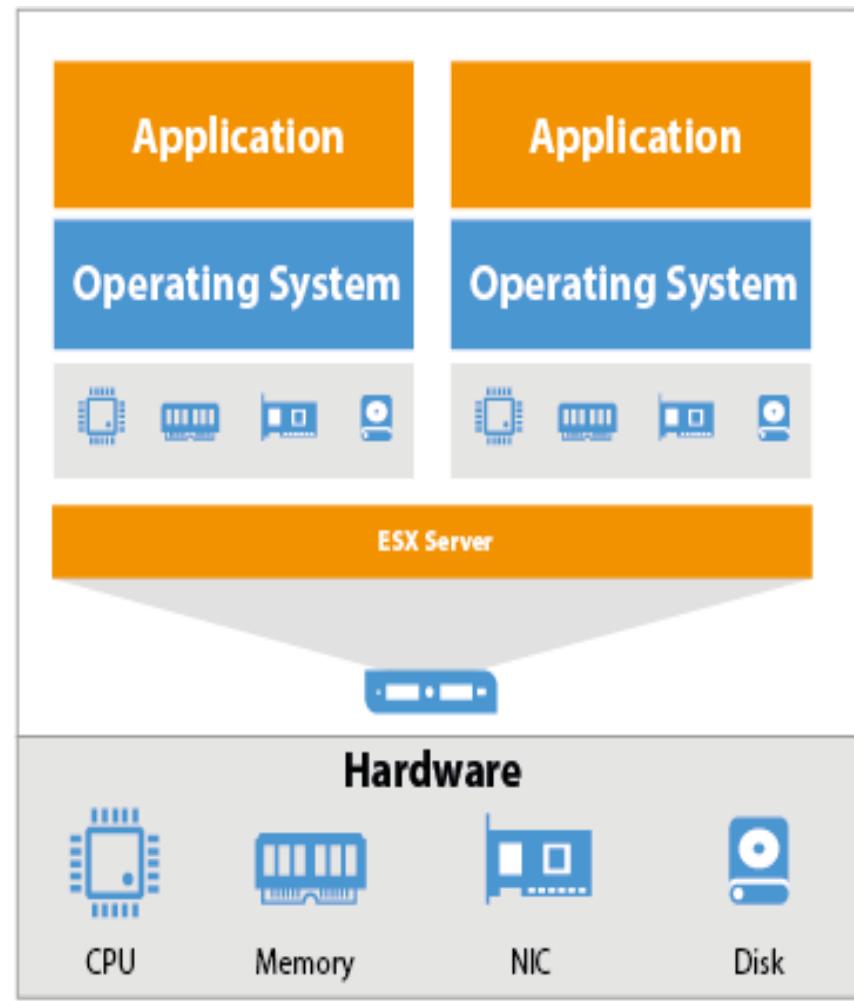
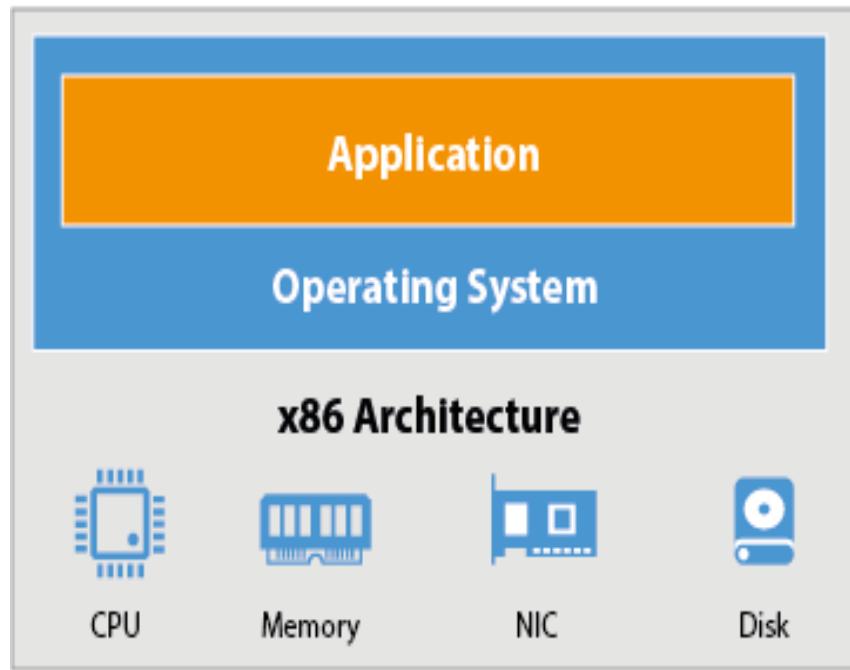
Physical server

- A **physical server**, also known as a 'bare-metal server,' is a single-tenant computer **server**, meaning that a specific **physical server** is designated to a single user.
- The resources and components of a **physical server** are not shared between multiple users.

Cloud Servers.

- **Cloud servers** can be configured to provide levels of performance, security and control similar to those of a dedicated **server**. But instead of being hosted on **physical** hardware that's solely used by you, they reside **in a** shared “virtualized” environment that's managed by your **cloud hosting** provider.

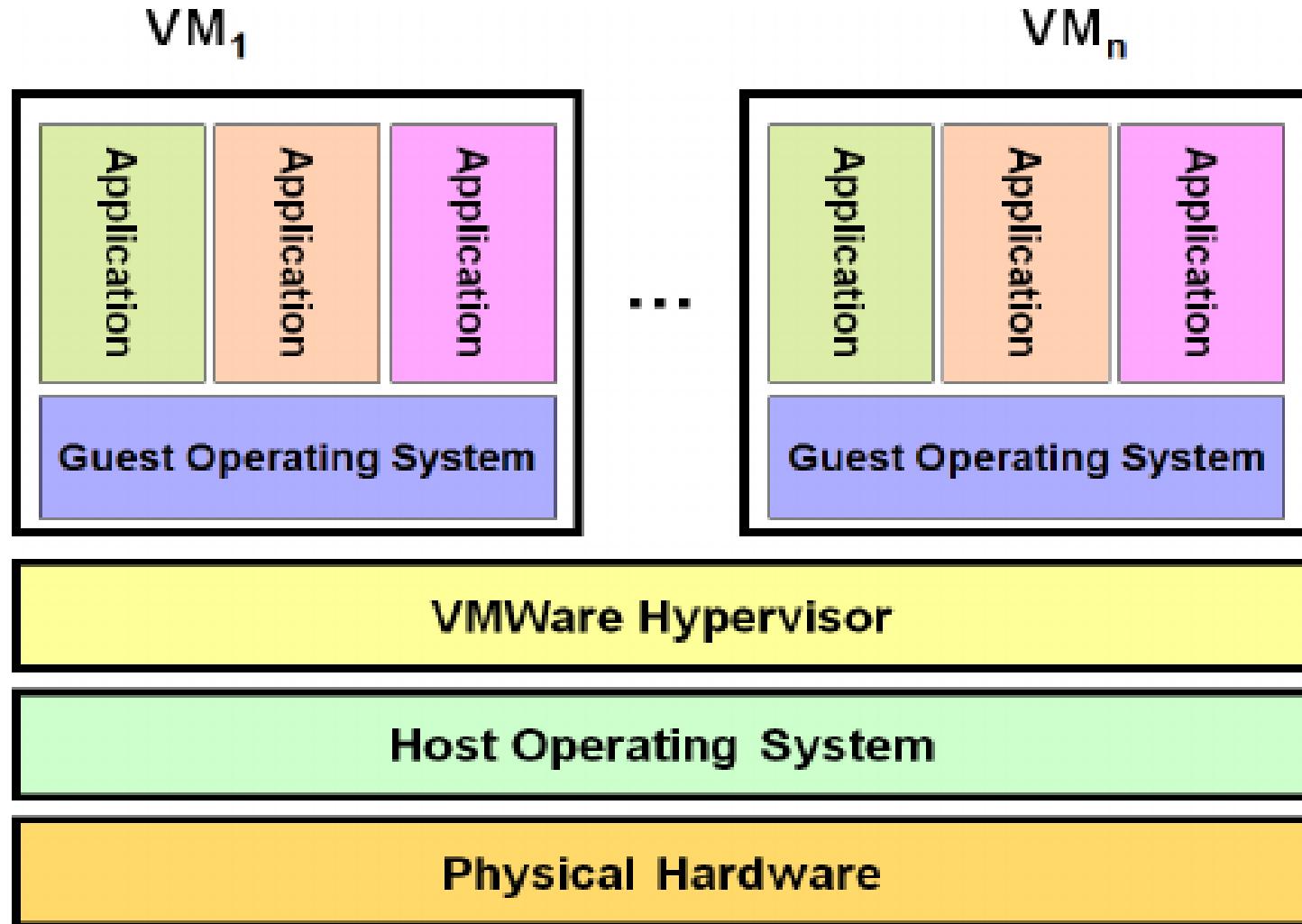
Physical server vs virtual machine



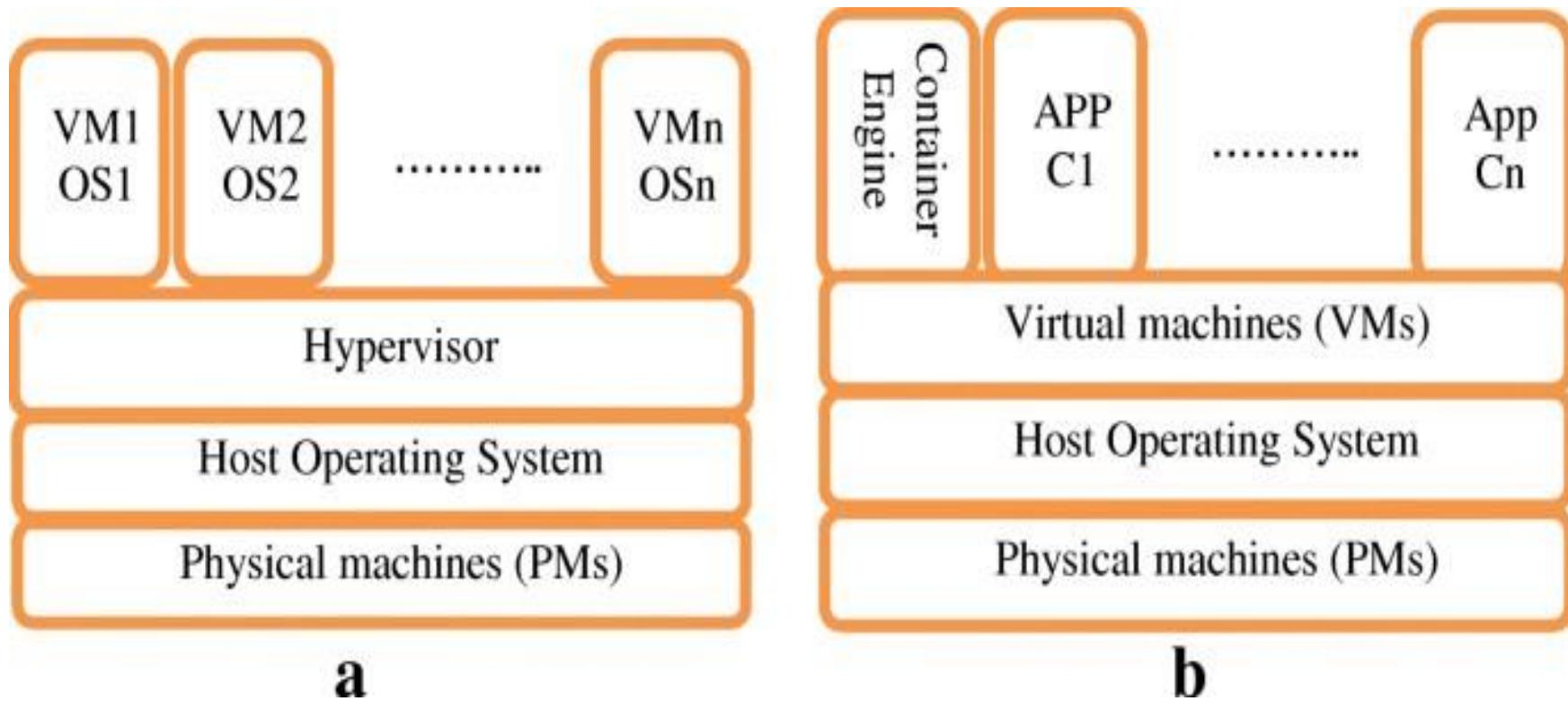
Virtual Servers vs. Physical Servers

- A physical server is a piece of equipment on which data is stored and read. This may be located onsite in your server room, or it could be stored at a colocation facility (a **data center**) with a trusted vendor.
- **Virtualization** is the act of placing multiple "virtual servers" on physical equipment. This allows physical server resources to be split between multiple workloads for maximum efficiency and cost savings.

Hosted Virtual Machine Architecture



A placement architecture for a container as a service (CaaS)



Data Center



What Is a Data Center

- A data center is a physical facility that organizations use to house their critical applications and data.
- A data center's design is based on a network of computing and storage resources that enable the delivery of shared applications and data.
- The key components of a data center design include **routers**, **switches**, **firewalls**, **storage systems**, **servers**, and **application-delivery controllers**.
- Data center infrastructure refers to the core physical or hardware-based resources and components – including all IT infrastructure devices, equipment and technologies – that comprise a data center.
- Data Center is modeled and identified in a design plan that includes a complete listing of necessary infrastructure components used to create a data center.

What defines a modern data center?

- Infrastructure has shifted from traditional **on-premises** physical servers to **virtual networks** that support applications and workloads across pools of physical infrastructure and into a multi-cloud environment.
- In this era, data exists and is connected across multiple data centers, the edge, and public and private clouds.
- The data center must be able to communicate across these multiple sites, both on-premises and in the cloud.
- Even the public cloud is a collection of data centers. When applications are hosted in the cloud, they are using data center resources from the cloud provider.

A data center infrastructure

A data center infrastructure may include:

- Servers.
- Computers.
- Networking equipment, such as routers or switches.
- Security, such as firewall or biometric security system.
- Storage, such as **storage area network (SAN)** or backup/tape storage.
- **Data center** management software/applications.

It can also include non-computing resources, such as:

- Power and cooling devices, such as air conditioners or generators
- Physical server racks/c chassis
- Cables
- Internet backbone

What are the core components of a data center?

- Data center design includes routers, switches, firewalls, storage systems, servers, and application delivery controllers. Because these components store and manage business-critical data and applications, data center security is critical in data center design. Together, they provide:
- **Network infrastructure.** This connects servers (physical and virtualized), data center services, storage, and external connectivity to end-user locations.
- **Storage infrastructure.** Data is the fuel of the modern data center. Storage systems are used to hold this valuable commodity.
- **Computing resources.** Applications are the engines of a data center. These servers provide the processing, memory, local storage, and network connectivity that drive applications.

How do data centers operate?

Data center services are typically deployed to protect the performance and integrity of the core data center components.

- **Network security appliances.** These include firewall and intrusion protection to safeguard the data center.
- **Application delivery assurance.** To maintain application performance, these mechanisms provide application resiliency and availability via automatic failover and load balancing.

What is in a data center facility?

- Data center components require significant infrastructure to support the center's hardware and software.
- These include power subsystems, uninterruptible power supplies (UPS), ventilation, cooling systems, fire suppression, backup generators, and connections to external networks.

Why are data centers important to business?

- Email and file sharing
- Productivity applications
- Customer relationship management (CRM)
- Enterprise resource planning (ERP) and databases
- Big data, artificial intelligence, and machine learning
- Virtual desktops, communications and collaboration services

PARAMETER	DATA CENTER	CLOUD
Infrastructure & Applications	Dedicated to one customer or organization	Shared across customers
Services model	Dedicated team of the customer/organization	Shared services across customers
Location	The physical location of data center can be within or outside the organisation's premises	Cloud Data center is located off-premise in service provider location
Accessed via	Dedicated and reliable WAN links like MPLS, P2P connections.	Primarily accessed over unreliable Internet
Security	More secured and better security can be implemented based on respective companies IT policy	Less Secured than traditional Data Center
Setup and upgrade time	Data Center takes much longer to be provisioned and high on operational cost	Almost available immediately based on subscription/payment by respective customer. Cloud systems can be built within moments and can be de-commissioned instantly.

PARAMETER	DATA CENTER	CLOUD
Implementation and operating cost	High since dedicated servers and supporting infrastructure needs to be provisioned.	Low since shared applications and servers are provisioned which multiple customers leverage. Hence the cost is shared across customers.
Scalability	Low on scalability. Capacity expansion of Data center requires to spend the significant amount of money to match increasing workloads	Cloud facility is highly scalable and quickly adapts to your business needs. Cloud offers unlimited capacity expansion based on vendor's products and service plans.
Reliability	Data Centers are generally less reliable than Cloud Data Centers	Cloud servers use multiple data centers in different geographical locations with appropriate backup. This provides safety from unwarranted downtime

What is MPLS in cloud ?

Multiprotocol level switching is a networking technology that routes traffic using the shortest path based on “levels”, rather than network addresses , to handle forwarding over private wide area network.

Types of data centers

- **Enterprise data centers:** These are built, owned, and operated by companies and are optimized for their end users. Most often they are housed on the corporate campus.
- **Managed services data centers:** These data centers are managed by a third party (or a managed services provider) on behalf of a company. The company leases the equipment and infrastructure instead of buying it.
- **Colocation data centers :** In colocation ("colo") data centers, a company rents space within a data center owned by others and located off company premises. The colocation data center hosts the infrastructure: building, cooling, bandwidth, security, etc., while the company provides and manages the components, including servers, storage, and firewalls.

Types of data centers

- **Cloud data centers**: In this off-premises form of data center, data and applications are hosted by a cloud services provider such as Amazon Web Services (AWS), Microsoft (Azure), or IBM Cloud or other public cloud provider.
- Essentially, a **cloud data** service is a remote version of a **data center** – located somewhere away from your company's physical premises – that lets you access your **data** through the internet.
- A **data center** traditionally refers to server hardware on your premises to store and access **data** through your local network

Acloud Data Center

- A cloud Data Center is not physically located in a particular organization's office – it's all online! When your data is stored on **cloud servers**, it automatically gets fragmented and duplicated across various locations for secure storage. In case there are any failures, your **cloud services provider** will make sure that there is a backup of your backup as well!

Benefits of cloud computing over the traditional data center

- **Low maintenance cost.** For a customer maintenance cost is almost nil. Since you are using hardware from the cloud provider's datacenter, you don't need to maintain hardware at all. Your cost is saved from geographical location cost, hardware purchase, upgrades, datacenter staff, power, facility management cost, etc. All this is bared by the cloud provider. Also, for cloud providers, this is also low since they are operating multiple clients from the same facility and hence cost is low compared to cost one has to bear when all those clients are operating from different datacenters. This is very much environment friendly too since you are reducing the need for multiple facilities to fewer ones.
- **Cheap resources.** Cloud providers have a pool of resources and from which you get assigned your share. This means cloud providers maintain and operate a large volume of resources and distribute smaller chunks to customers. This obviously reduces the cost of maintenance and operation for cloud providers and in turn provides low cost, cheap resources to customers.

Benefits of cloud computing over the traditional data center

- **Scale as per your need.** In a traditional data center you have to study and plan your capacity well in advance to finalize your hardware purchase. Once purchased you are stuck with purchased limited capacity and you can not accommodate if capacity requirement grows beyond limit before your estimated time. It again goes through planning, purchasing new hardware which is a time-consuming process. In the cloud you can scale up and scale down your computing capacity almost instantly (or way shorter in time than traditional purchase process). And don't even need to worry and follow for approvals, purchase, billing, etc things.
- **Pay as you use.** In traditional data centers whenever you buy hardware you make an investment upfront even if you don't use the full capacity of purchased hardware. In the cloud, you are billed per your use. So your expenditure on computing is optimum with your use.

Benefits of cloud computing over the traditional data center

- **The latest technology at your service.** Technology changing very fast these days. Hardware you buy today becomes obsolete in a couple of months. And if you are making huge investments in hardware, the company expects to use it at least for a couple of years. So you are stuck with the hardware you brought with a nice price tag and now way behind from its latest counterparts. Cloud provides you the latest tech always and you don't need to worry about upgrades or maintenance. All these hardware aspects are the headache of cloud providers and they take care of it in the background. As a customer, all the latest technology is at your service without any hassle.

Benefits of cloud computing over the traditional data center

- **Redundancy.** Redundancy in traditional datacenter means cost investment to build almost identical facilities of the primary. Along with it also involves cost for infrastructure which connects them. Also, on-site redundancy for power, network, etc. is also expensive and maintenance prone. When you are opting cloud, everything said previously is just vanished from your plate. Cloud at single entity level like single server, storage disk, etc is already redundant. Nothing to be done and no extra cost is being billed to you for it. For your infra design requirement if you want, you can use ready-made services provided by cloud (for redundancy) and you are all set from failures.
- **Accessibility.** With an on-premise datacenter, you have very limited connectivity mostly locally. If you want access to inside entities, you need to maintain your own VPN. Cloud services have a portal with access to almost all of their services over the web. It can be accessed from anywhere with internet. Also, if you want to opt-in for a VPN, you get a pre-configured secure VPN from your cloud provider. No need for designing and maintaining aVPN!

Cloud Infrastructure

- Cloud infrastructure refers to the hardware and software components, such as servers, storage, networking and virtualization software, that are needed to support the computing requirements of a cloud computing model.
- In addition, cloud infrastructures include a software abstraction layer that virtualizes resources and logically presents them to users through programmatic means.

What are the standards for data center infrastructure?

- The most widely adopted standard for data center design and data center infrastructure is ANSI/TIA-942. It includes standards for ANSI/TIA-942-ready certification, which ensures compliance with one of four categories of data center tiers rated for levels of redundancy and fault tolerance.
- **Tier 1: Basic site infrastructure.** A Tier 1 data center offers limited protection against physical events. It has single-capacity components and a single, **non redundant** distribution path.
- **Tier 2: Redundant-capacity component site infrastructure.** This data center offers improved protection against physical events. It has redundant-capacity components and a single, nonredundant distribution path.

What are the standards for data center infrastructure?

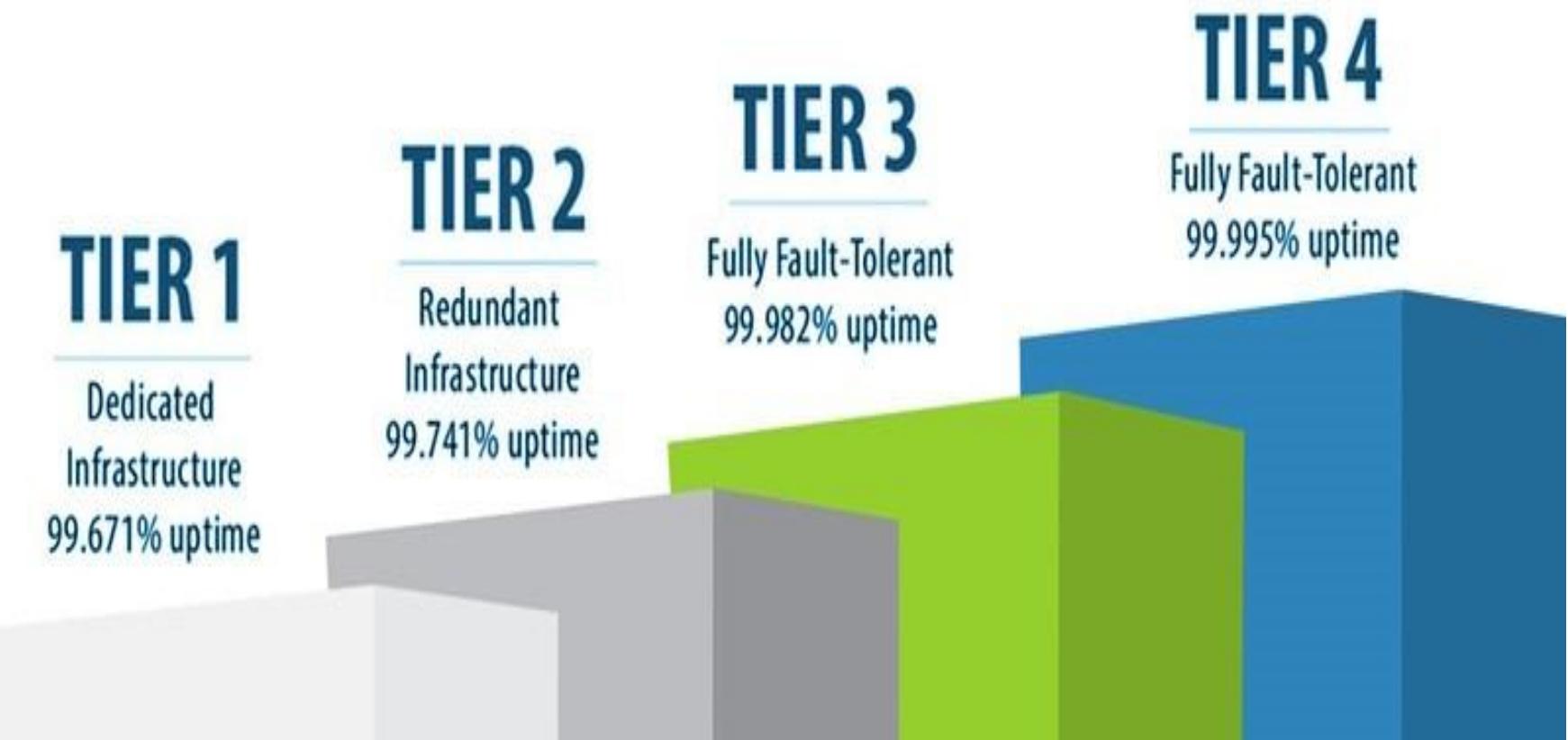
- **Tier 3: Concurrently maintainable site infrastructure.** This data center protects against virtually all physical events, providing redundant-capacity components and multiple independent distribution paths. Each component can be removed or replaced without disrupting services to end users.
- **Tier 4: Fault-tolerant site infrastructure.** This data center provides the highest levels of fault tolerance and redundancy. Redundant-capacity components and multiple independent distribution paths enable concurrent maintainability and one fault anywhere in the installation without causing downtime.

The data centers



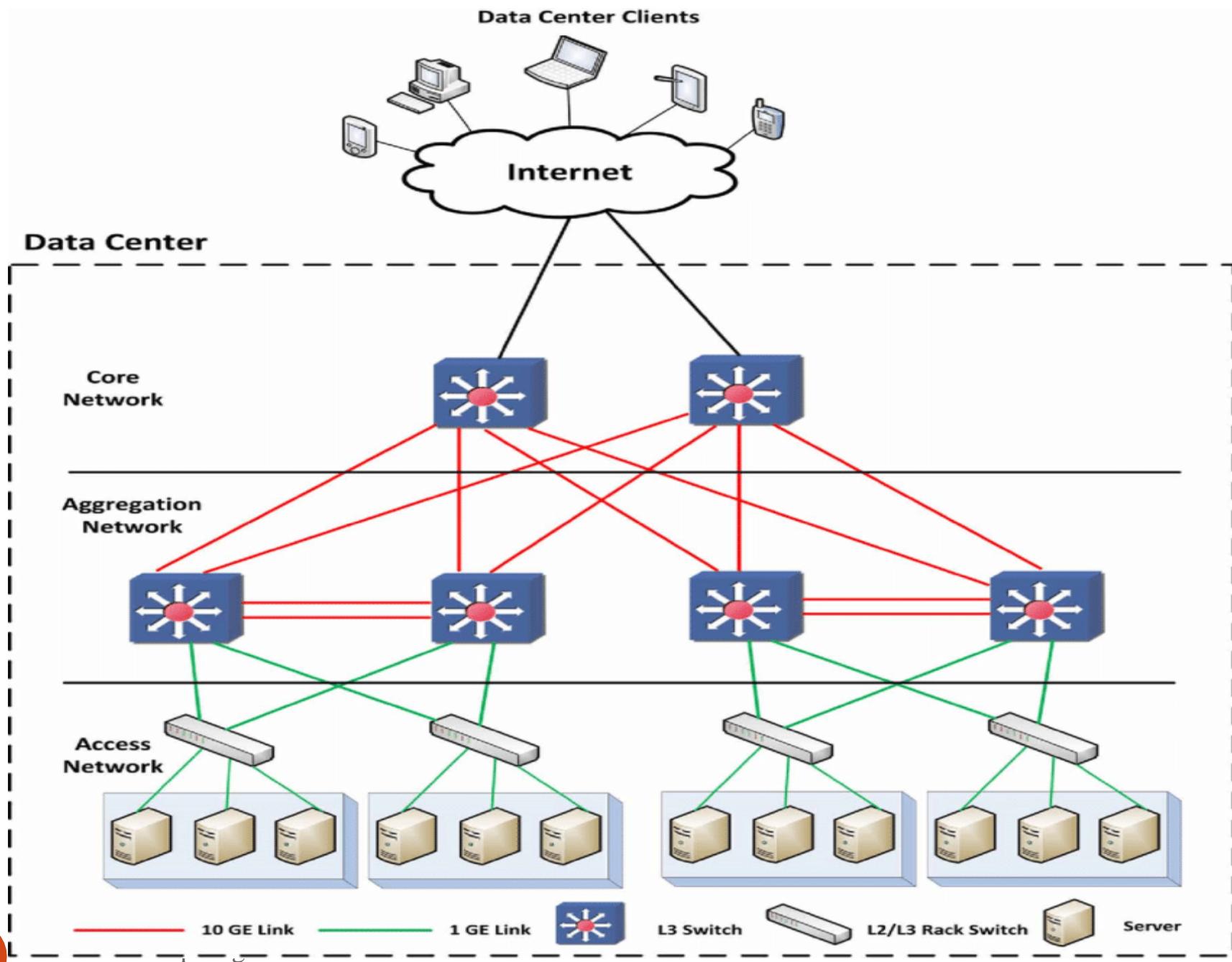
A large group of networked computer servers typically used by organizations for the remote storage, processing, or distribution of large amounts of data.

DATA CENTER TIERS



Data center networking

- **Data center networking** is the integration of a constellation of networking resources — switching, routing, load balancing, analytics, etc. — to facilitate the storage and processing of applications and data.
- Modern data center networking architectures leverage full-stack networking and security virtualization platforms that support a rich set of data services connecting everything from VMs, containers, and bare metal applications while enabling centralized management and granular security controls.



Requirements for a modern data center networking platform

- **Automation.** Achieving speed and agility in modern data centers depends greatly on automated provisioning of networking services for applications. Far faster and more reliable than a human administrator, modern networking platforms not only find the most efficient way to program a network, balance workloads, and automate time-consuming tasks, they also respond dynamically to changes in usage.
- **Consistent policies.** With modern data center networking responsible for integrating resources from edge to cloud, consistent application of policies is essential.
- **Single pane of glass.** Typically connecting resources located both on-premises, in the cloud, and at the edge, modern data center networking platforms offer centralized management from a single console.
- **Granular security.** Today's data center networking platforms often feature integrated security controls that can include micro-segmentation and IDS/IPS.
- **Global visibility.** Most data center networking platforms can display a visual representation of the network and its interconnections, which makes troubleshooting network issues much easier.

The data centers

- A data center (sometimes spelled *datacenter*) is a centralized repository, either physical or virtual, for the storage, management, and dissemination of data and information organized around a particular body of knowledge or pertaining to a particular business.
- The terms "cloud" and "data center" may sound like interchangeable technical jargon or trendy buzz words referring to the same infrastructure, but the two computing systems have less in common than the fact that they both store data.

A cloud and a data center

- The main difference between a cloud and a data center is that a cloud is an off-premise form of computing that stores data on the Internet, whereas a data center refers to on-premise hardware that stores data within an organization's local network.
- While cloud services are outsourced to third-party cloud providers who perform all updates and ongoing maintenance, data centers are typically run by an in-house IT department.

Need a cloud or a data center 1/2

- A data center is ideal for companies that need a customized, dedicated system that gives them full control over their data and equipment. Since only the company will be using the infrastructure's power, a data center is also more suitable for organizations that run many different types of applications and complex workloads.
- A data center, however, has limited capacity -- once you build a data center, you will not be able to change the amount of storage and workload it can withstand without purchasing and installing more equipment.

Need a cloud or a data center 2/2

- A cloud system is scalable to your business needs. It has potentially unlimited capacity, based on your vendor's offerings and service plans.
- One disadvantage of the cloud is that you will not have as much control as you would a data center, since a third party is managing the system.
- Furthermore, unless you have a private cloud within the company network, you will be sharing resources with other cloud users in your provider's public cloud.

Cloud vs. data center costs

- For most small businesses, the cloud is a more cost-effective option than a data center. Because you will be building an infrastructure from the ground up and will be responsible for your own maintenance and administration, a data center takes much longer to get started and can cost businesses \$10 million to \$25 million per year to operate.
- Unlike a data center, cloud computing does not require time or capital to get up and running. Instead, most cloud providers offer a range of affordable subscription plans to meet your budget and scale the service to your performance needs. Whereas data centers take time to build, depending on your provider, cloud services are available for use almost immediately after registration.

Categorizing the data centers

<https://uptimeinstitute.com/>

The Uptime Institute categorizes the **data centers** by four levels: **Tier I, II, III** and IV. These levels correspond to a certain number of guarantees on the type of hardware deployed in the **data center** to ensure redundancy. Availability: 99.67%

The Uptime Institute has categorized data centers into four hosting tiers:

- **Tier I Data Centers:** Data centers with Tier I topology offers single uplink and servers, with 99.671% uptime. However, these data centers are non-redundant, catering to basic business demands. As a result, any unwarranted failure in the capacity system thwarts the ongoing performance.
- **Tier II Data Centers:** These modern data centers have single, non-redundant path for power source. Data centers listed with this topology offers redundant capacity components to ascertain smooth access, with 99.741% network uptime.
- **Tier III Data Centers:** Equipped with redundant components with manifold power and cooling options, these data centers can efficiently and expeditiously switch to maintain backup paths ensuring 99.982% network availability.
- **Tier IV data Centers:** These are fault -tolerant **data center** having multiple power and environment control channels with activated data backup options, providing 9.995% network availability.

Categorization of data centers

- **Tier 1:** composed of a single path for power and cooling distribution, without redundant components, providing 99.671% availability.
- **Tier II:** composed of a single path for power and cooling distribution, with redundant components, providing 99.741% availability
- **Tier III:** composed of multiple active power and cooling distribution paths, but only one path active, has redundant components, and is concurrently maintainable, providing 99.982% availability
- **Tier IV:** composed of multiple active power and cooling distribution paths, has redundant components, and is fault tolerant, providing 99.995% availability

Categorizing the data centers

- Tier 1 to 4 data center is a standardized methodology used to define uptime of data center and useful for measuring: Data center performance, Investment, and ROI (return on investment)
- Tier 4 data center considered as **most robust and less prone** to failures. Tier 4 is designed to host mission critical servers and computer systems, with fully redundant subsystems (cooling, power, network links, storage etc) and compartmentalized security zones controlled by biometric access controls methods.
- Naturally, the simplest is a Tier 1 data center used by small business or shops.

What kind of upfront costs are associated with building your own data center?

- **Network connection cost** – you'll have to pay for fiber on-site from one more ISP
- **Power** – this expense accounts for 70-80 percent of the total costs of running a data center, and is also highly variable by region.
- **Data center staffing** – around-the-clock monitoring, on-site maintenance and equipment optimization requires a dedicated and responsive operations staff, and accounts as the second largest expense after power.
- **Annual facility and infrastructure maintenance** – a more unpredictable cost of a data center ranging from 3-5 percent of the initial construction cost. Repairs and additions are expected around the third year of operation.

Integration of Wireless Sensor Networks with Cloud:

- These days Wireless Sensor Networks (WSN) are integrated with cloud computing to facilitate the end users in many ways.
- It helps the end users to run several applications of various wireless sensor networks through virtualization. This integration can provide sensor-as a-service and known as sensor cloud.
- Cloud computing helps to virtualize the physical sensors and this virtualization provides end users an opportunity to execute multiple applications without having any care of numbers and types of WSN.

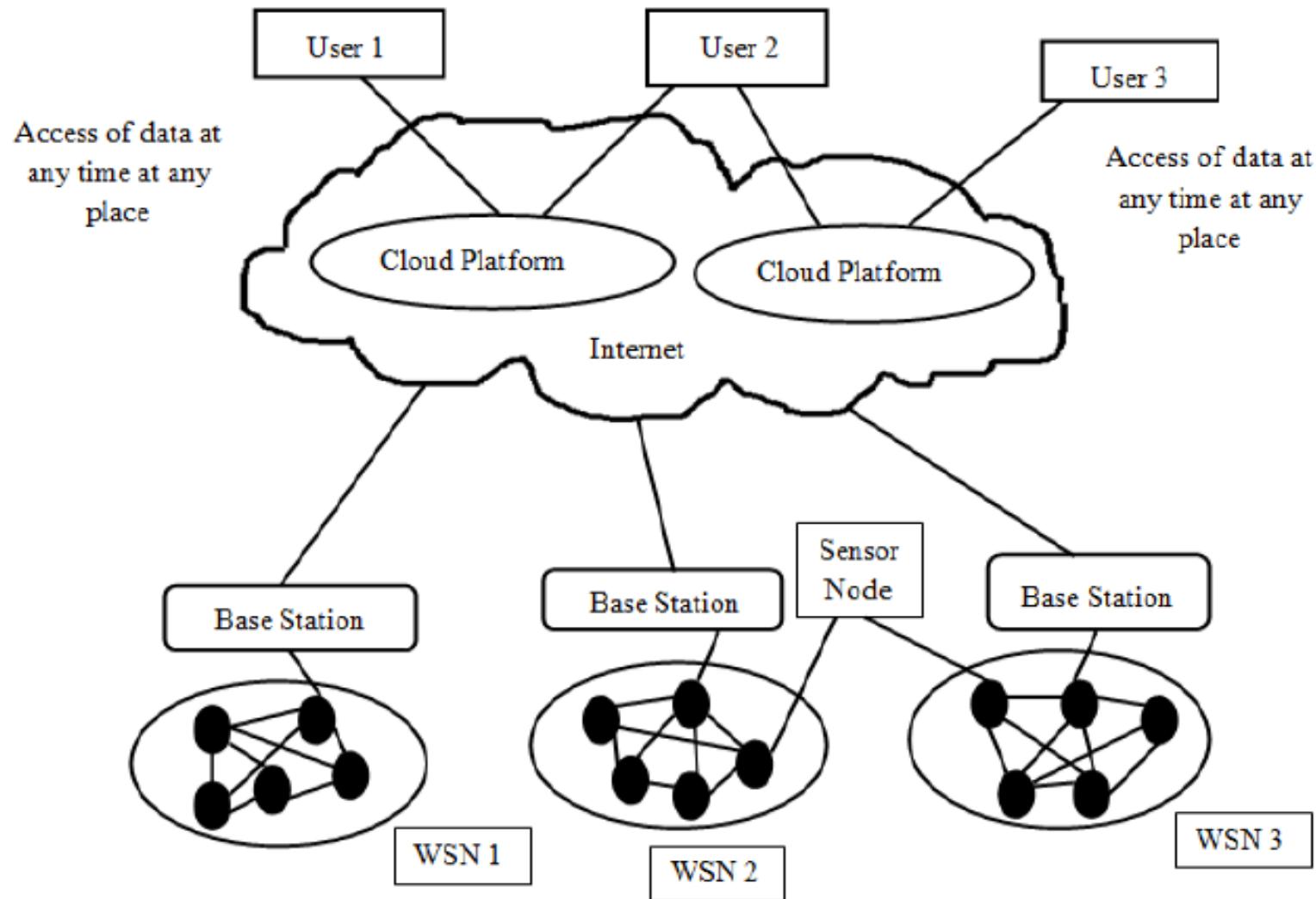


Fig 1: Integrating WSN with cloud [16]

- WSN is having several applications and being integrated with other technologies , although it is having some challenges such as energy efficiency, security, fault tolerance, scalability etc.
- An application of WSN does not use all physical sensors always. Sensor cloud is a paradigm which collects the data from physical sensors and sends it into a cloud computing infrastructure.
- This processed data is provided to the users anytime and anywhere when requested by him. This paradigm provides intelligent operation and communication of WSN by integrating it with cloud to serve the people better. Sensor cloud can be of three types: Independent SC, collaborative SC and mutual SC.
- The architecture of sensor cloud model is slightly different with the architecture of traditional WSN. In this new model, many small sensor nodes are available which collect the data.
- This data is sent to the base station or gateway and then the collected information is transmitted to the cloud. This information is stored on the cloud computing platform. Multiple users or clients can access the data from the cloud using internet.

Advantages of sensor cloud

- Increased data storage
- Increased processing power
- Dynamic provisioning of services
- Quick response time
- Flexibility
- Scalability
- Multi-tenancy
- Agility of services
- Resource optimization
- Collaboration
- Automation
- Virtualization

Issues and challenges with sensor cloud

- Storage issue
- Power issue
- Bandwidth limitation
- Massive scaling
- Real time multimedia content processing
- Authorization issues
- Security and privacy support issues
- Efficient information dissemination
- Pricing issues
- Network access management
- Resource and hardware compatibility issues
- Resource scheduling
- Resource usage policy
- Interface standardization issues
- Quality of Service

ARCHITECTURE OF SENSOR CLOUD

- The sensor cloud architecture is shown in fig 2.
- The integrating system includes many components such as

Identity Access Management Unit (IAMU),

Request Subscriber (RS),

Sub/Pub Broker

Data Processing Unit (DPU).

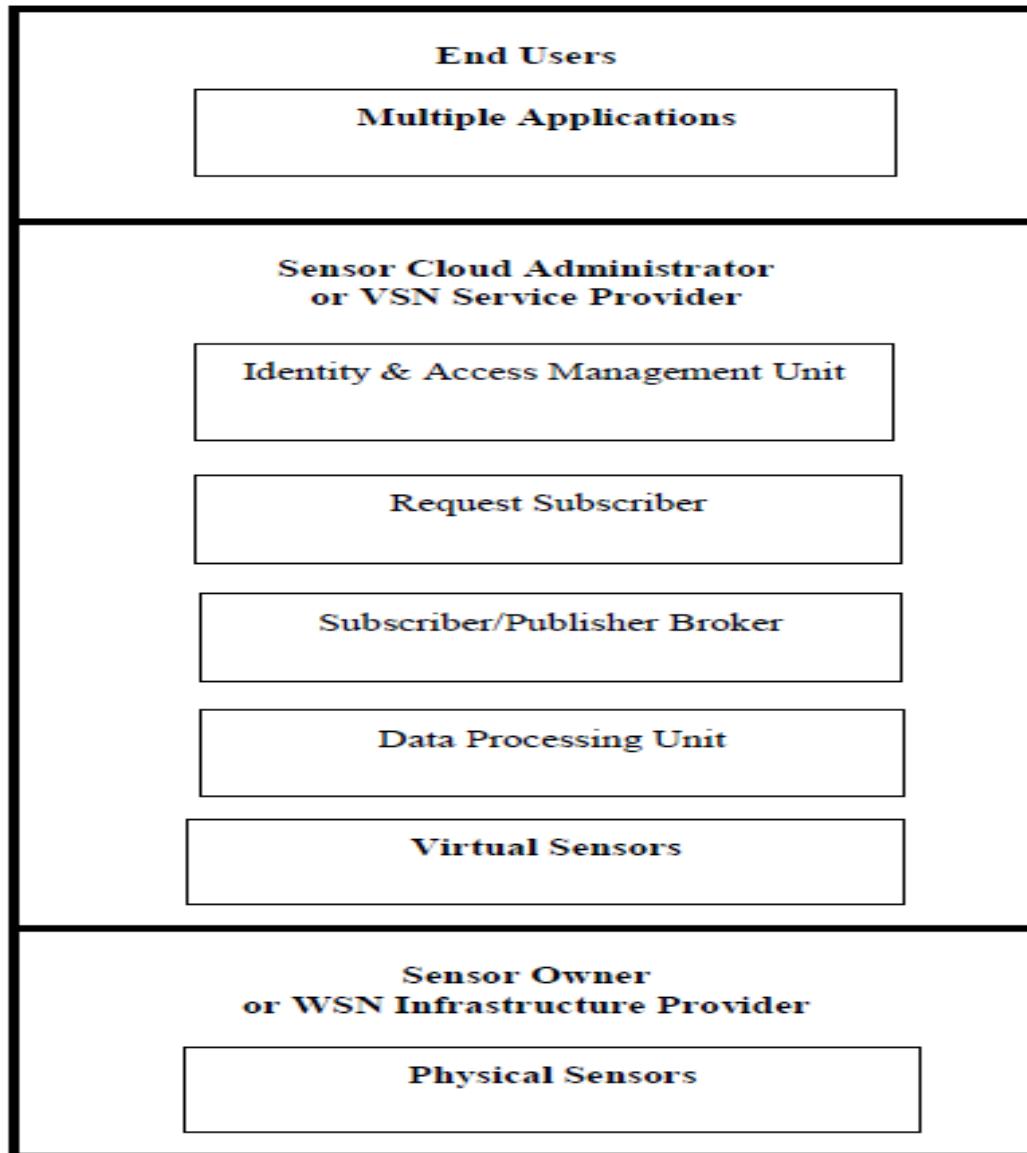


Fig 2: Architecture of Sensor Cloud

- *A. Identity Access Management Unit (IAMU):*

Basic goals of Identity Access Management Unit (IAMU) are to create authentication for consumer, to define service type and to provide the policy for access control of cloud resources. With help of IAMU consumer gets connected with provider.

- *B. Request Subscriber (RS)*

RS explains the request of users. This unit verifies the request of clients. This unit works for retrieval of data on the user's request. RS passes the request to Sub/ Pub Broker. It also implements monitoring & metering.

- *C. Subscriber/ Publisher (Sub/Pub) Broker*

Publisher submits the new data in the resource system. Data is stored in the index of the Sub/Pub Broker. When any Subscriber requests the data then RS creates the new subscription. If this subscription is matched with publisher data, then Sub/Pub Broker access the data. This system reduces resource consumption as well as complexity of the system.

- *D. Data Processing Unit (DPU)*

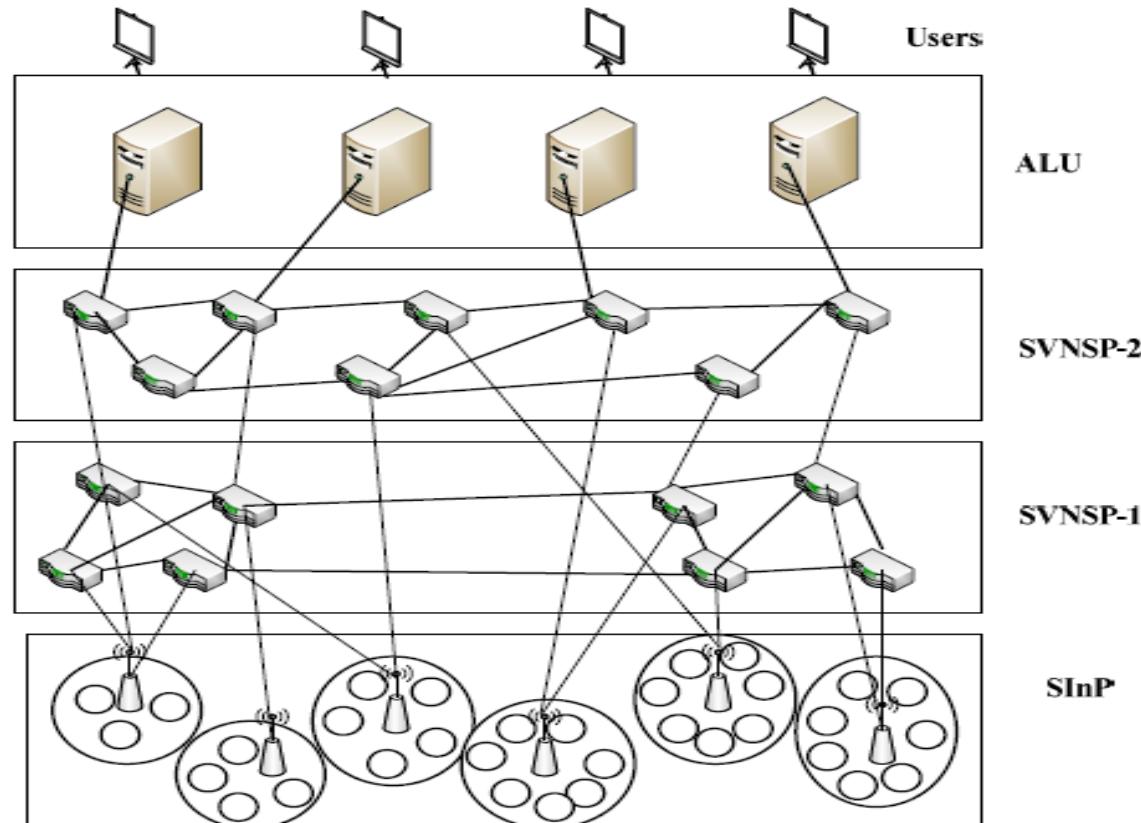
This unit is useful for processing and disseminating the outcome which is created by queries. DPU is used as database. It has collection of several tables. In these tables individual sensor readings are stored. Every sensor reading is linked with each other in the table. This unit is helpful for storing the data and also for retrieving the data from cloud.

Process between User and WSN

- User creates the account at IAMU & sends the request to it.
- IAMU defines type of service and creates request message.
- Request message is sent to the RS.
- RS creates the subscription & sends it to Sub/Pub Broker.
- Data is collected from WSNs to virtual nodes.
- Virtual nodes relate it with each application.
- DPU stores the data from virtual node.
- DPU send the index data to the Sub/Pub Broker.
- Sub/Pub Broker matches both index data with DPU and required by RS.
- If data is matched, then Sub/Pub Broker sends request to retrieve the data from DPU.
- DPU provides the requested data to the Sub/Pub Broker.
- Sub/Pub Broker forwards the data via RS and IAMU to the client.

VIRTUALIZATION IN SENSOR CLOUD

Virtualization concept facilitates to execute many applications concurrently at same sensor node. Sensor network virtualization model is shown in fig 3.



*ALU: Application Level User
SVNSP: Sensor Virtualization Network Service Provider
SInP: Sensor Infrastructure Provider*

Fig 3: Model of sensor network virtualization [15]

- This model consists of two types of providers. First provider is called Sensor Infrastructure Provider (SInP) that manages all the sensor nodes in WSN.
- Other provider is known as Sensor Virtualization Network Service Provider (SVNSP) which creates virtualization of sensor networks using multiple SInP. This virtualization of sensor networks is called Virtual Sensor Network (VSN). Application Level User (ALU) can connect to various SVNSP for different applications .
- Virtualization helps in maximum utilization of physical sensors through multiple applications running at a time. This technology provides new opportunities in various fields of applications such as battlefield surveillance, healthcare, vehicle telemetric, structural monitoring and agriculture monitoring etc.

- Virtualization in WSN can be divided in two categories as discussed below:
 - Node level virtualization
 - Network level virtualization

Node Level Virtualization

- This type of virtualization is implemented to execute multiple tasks on a physical sensor node at the same time. This is shown in fig 4.

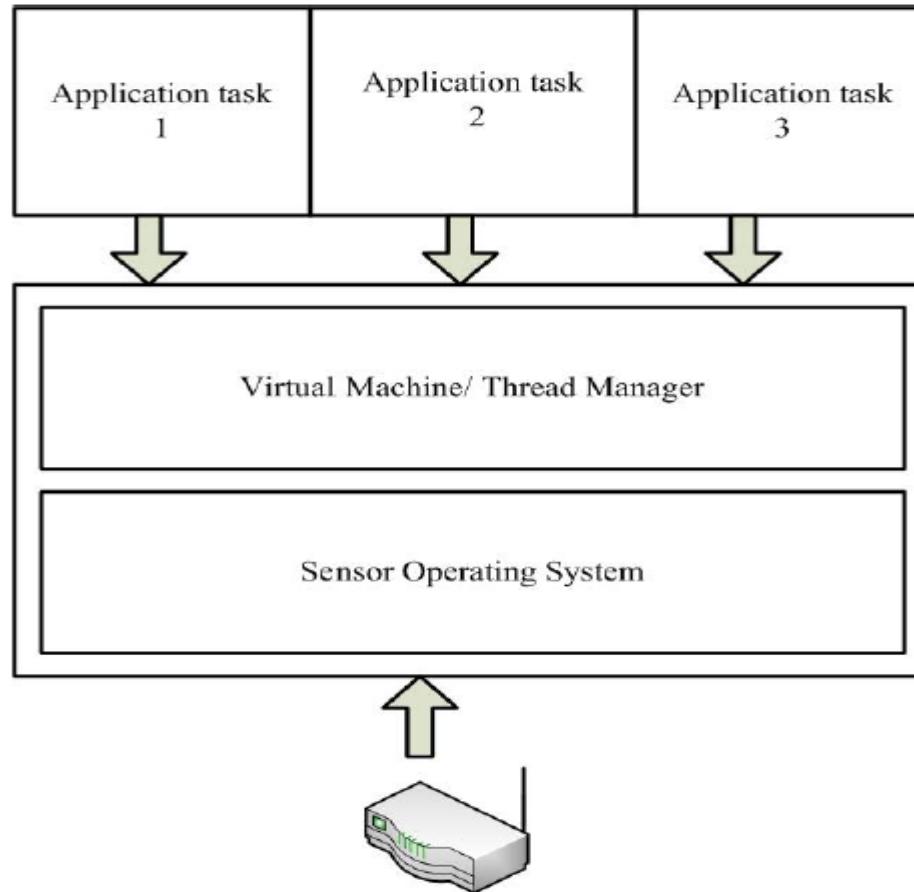


Fig 4: Model of node level virtualization

B. Network Level Virtualization

Here we create subsets of the WSN nodes with respect to various applications. Fig 5 shows network level virtualization in which logical groups of sensor nodes are created. These groups are related to the particular applications.

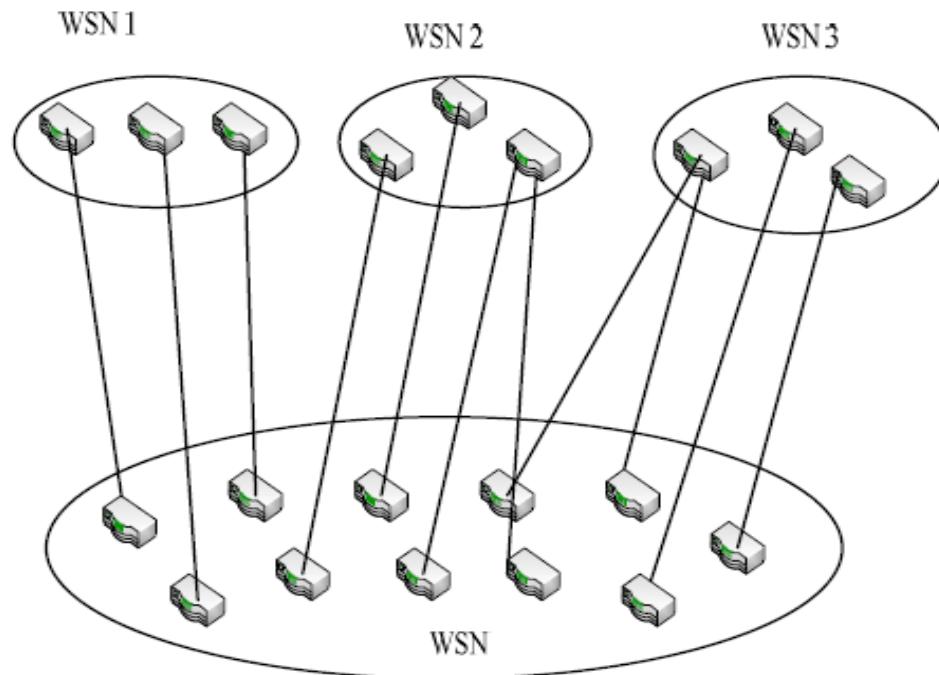


Fig 5: Model of network level virtualization

VI. APPLICATIONS OF SENSOR CLOUD

- There are several applications of sensor cloud in today's era.
- *Rock sliding & animal crossing monitoring* is one of the popular applications of sensor cloud to protect animals from rock slides in mountain area.
- When this application is deployed, an emergency signaling is used to make the animals aware of rock sliding.
- Fig 6 shows a mountain area where animals cross the road and on the mountains, there may be rock slides. In this case, there is a single physical WSN.
- Two VSNs are using this single WSN for two different applications: *monitoring rock slides* and *monitoring animal crossing*.

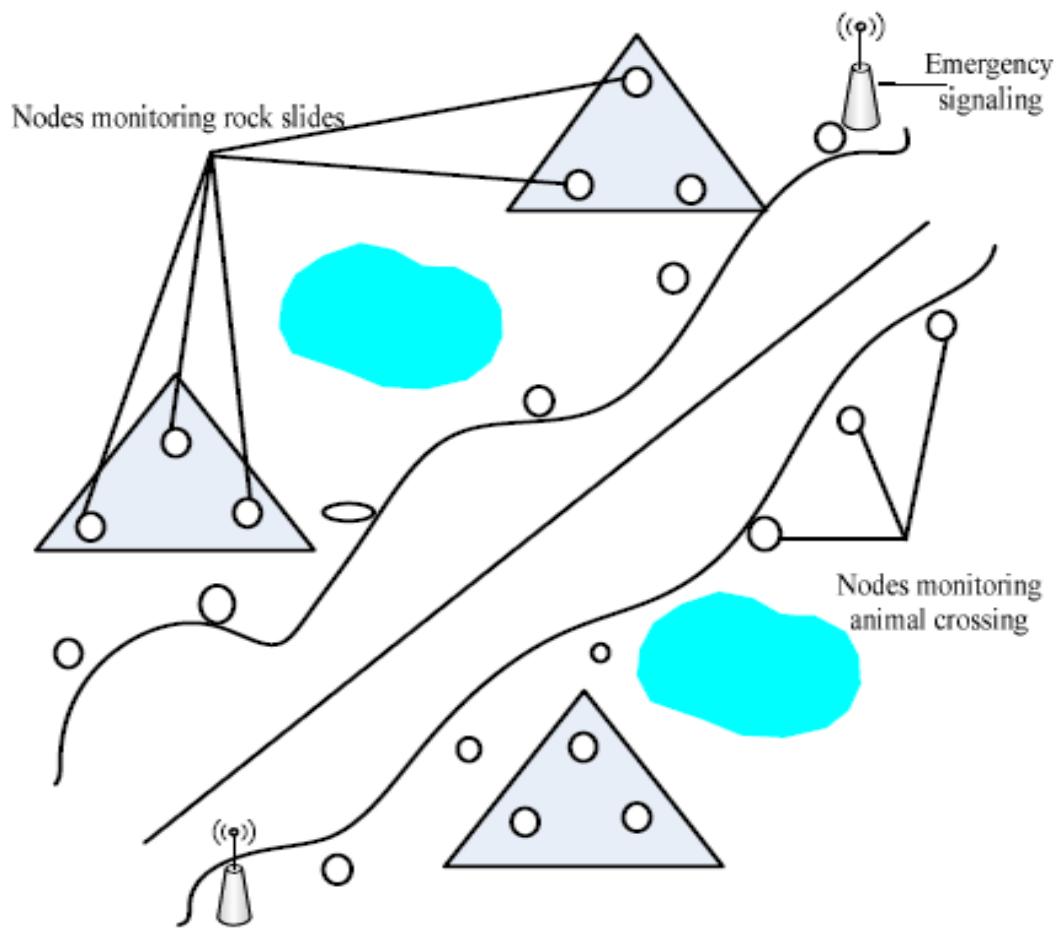


Fig 6: Rock sliding & animal crossing monitoring

Contd.

- Battlefield monitoring
- Environmental monitoring
- Disaster detection
- Smart home monitoring
- Healthcare monitoring
- Wildlife monitoring
- Weather monitoring
- Transport and vehicle telematics
- Agriculture monitoring
- Industrial monitoring
- Structural monitoring
- Target tracking applications

ISSUES AND CHALLENGES

- ***Authorization Issue-*** It is a most common issue in sensor network .If anyone can authorize sensor network without any permission of administrator then generate the authorization issue in Sensor-Cloud.
- ***Energy Issue-*** WSN and sensor-cloud are require to lot of energy and power consumption for proper working of sensor nodes and large amount of sensor data that is avoid the large data transmission.
- ***Security Issue-*** For authorization transaction and smooth working of network or infrastructure however maintain and manage integrity in daily routine. Therefore, if a sensor-cloud network system sensor data are send
- ***Storage Issue-*** Storing the sensor data at back end side is important issues in sensor cloud. Data storage issue is one of the common issues in wireless sensor network and sensor cloud.

SUBSTANCES INVOLVED IN SENSOR CLOUD

- *Owner*- Sensor cloud owner is a person who manages or maintained its own physical sensor .which is established as per the owner requirement.
- *Administrator*- Sensor cloud services are organized by the sensor administrator. It also regulates virtual sensor and the user interface.
- *End user*- It can be defined as the person who utilized the sensor cloud data for one or more application.
- *Earth observation*- In this application sensor grid analyzes, visualizes the GPS data and also collects the data from various GPS location.

ADVANTAGES OF SENSOR CLOUD

- **Analysis**-It is combination of large amount collected sensor data and sensor networks over the cloud computing prototype. It maintain cloud computing infrastructure being fascinating for various kind of analysis.
- **Scalability**-If the need of increase resources so, organization add some extra services from cloud merchant without any expenditure. That is called scalability of sensor cloud.
- **Visualization**-Sensor cloud infrastructure provides a imagination platform to be used for gathered and reacquire sensor data from different sources.
- **Collaboration**-Sensor cloud allow to share sensor data by several category of retailers therefore, the union of many physical sensor networks.

- *Increase data storage and processing*-It allocate facility to store data and excessive processing and also provide an application manage large amount of data.
- *Dynamic processing of services*-In this benefit Sensor cloud access their data from anywhere, everywhere and any time they want to access data from sensor data.
- *Flexibility*-It provides extensibility to its user then the prior computing procedure. It allows to stored and share sensor data under flexibility usage environment.
- *Quick response time*-It is concatenation of (WSN) wireless sensor network and cloud computing supply a rapid to the user. Therefore it is called real time application.
- *Automation*-Automation play an important role in sensor cloud computing. It also increases the transmission time to significant changes.

- *Multitenancy-*

It is an ability which distributes services to multiple users and share sensor cloud resources. It also provided openness of sensor data to access anywhere and everywhere.

Conclusion

- This module describes a survey on integration of WSN with cloud. People can use the sensors of multiple WSN for various applications with help of virtualization and cloud.
- Virtualization facilitates in creating virtual sensor networks from various physical sensors which helps the cloud to provide sensor-as a-service to its end users. In this way, this integration helps to optimize the usefulness of the sensor networks as well as the cloud.