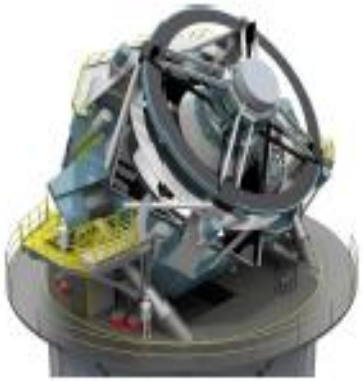
Cloud Computing

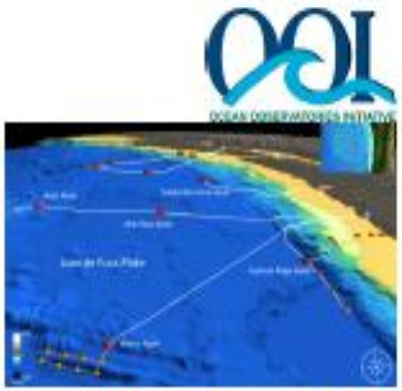
Computational Requirements for Intelligent Systems

- Need to run faster
  - Need to run large data (especially SIMD)
  - What is “large”?
  - Two arrays of 1,000,000 elements can be added in ~0.015s on the current CPU
  - 10,000,000,000 elements takes ~3.5 minutes
- Run experiments in
  - days instead of months,
  - hours instead of days,
  - minutes instead of hours
- Quickly iterate over designs and parameters of deep learning networks

Data-driven Discovery in Science



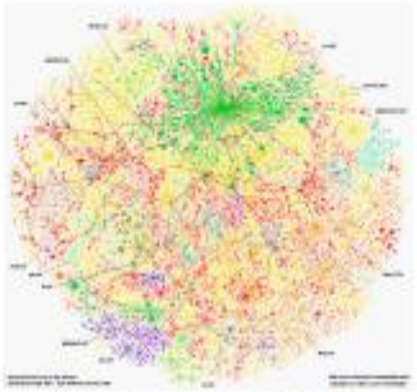
Astronomy: LSST



Oceanography: OOI

Physics: LHC

Biology: Sequencing



Sociology: The Web

Internet of Things

Economics: POS Terminals

Neuroscience: EEG, fMRI

Credit: Ed Lazowska, Univ. of WA

Need for speed up

1. Number of calculations done per unit time
   Floating Point **OP**erations/**s**econd

   FLOP/s ➔ 1990s MFLOP/s ➔ GFLOP/s ➔ TFLOP/s

   Currently: 100's of PFLOP/s

   Near future: Exa FLOP/s

2. Data I/O (Storage ⬅➔ Main Memory)
   - Read speeds and write speeds
   - Transfer speeds

- NEED FOR SPEED➔ NEED FOR PARALLEL

Conventional vs Hybrid compute systems

- **Conventional**: CPU only
  - multiple threads on a single machine : e.g. OpenMP
  - Multiple processes on a single machine or across machines: e.g. MPI
- **Hybrid**: Using "accelerator" also
  - "master-slave" –CPU drives GPU
  - CPU should not be idle while GPU is computing.
  - Cannot assume GPU is always faster than CPU

Computing Infrastructure

- Large scale, distributed, heterogeneous, multicore, accelerators, deep storage hierarchies, experimental systems

TOP 5 MACHINES (TOP500.ORG Feb 2022)

PRESENTED BY

FIND OUT MORE AT

top500.org

NOVEMBER 2021			SITE	COUNTRY	CORES	RMAX PFLOP/S	POWER MW
1	Fugaku	Fujitsu A64FX (48C, 2.2GHz), Tofu Interconnect D	RIKEN R-CCS	Japan	7,630,848	442.0	29.9
2	Summit	IBM POWER9 (22C, 3.07GHz), NVIDIA Volta GV100 (80C), Dual-Rail Mellanox EDR Infiniband	DOE/SC/ORNL	USA	2,414,592	148.6	10.1
3	Sierra	IBM POWER9 (22C, 3.1GHz), NVIDIA Tesla V100 (80C), Dual-Rail Mellanox EDR Infiniband	DOE/NNSA/LLNL	USA	1,572,480	94.6	7.44
4	Sunway TaihuLight	Shenwei SW26010 (260C, 1.45 GHz) Custom Interconnect	NSCC in Wuxi	China	10,649,600	93.0	15.4
5	Perlmutter	HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10 (274 GB)	LBNL	USA	761,856	70.9	2.58

https://top500.org/lists/top500/2021/11/

Transportation

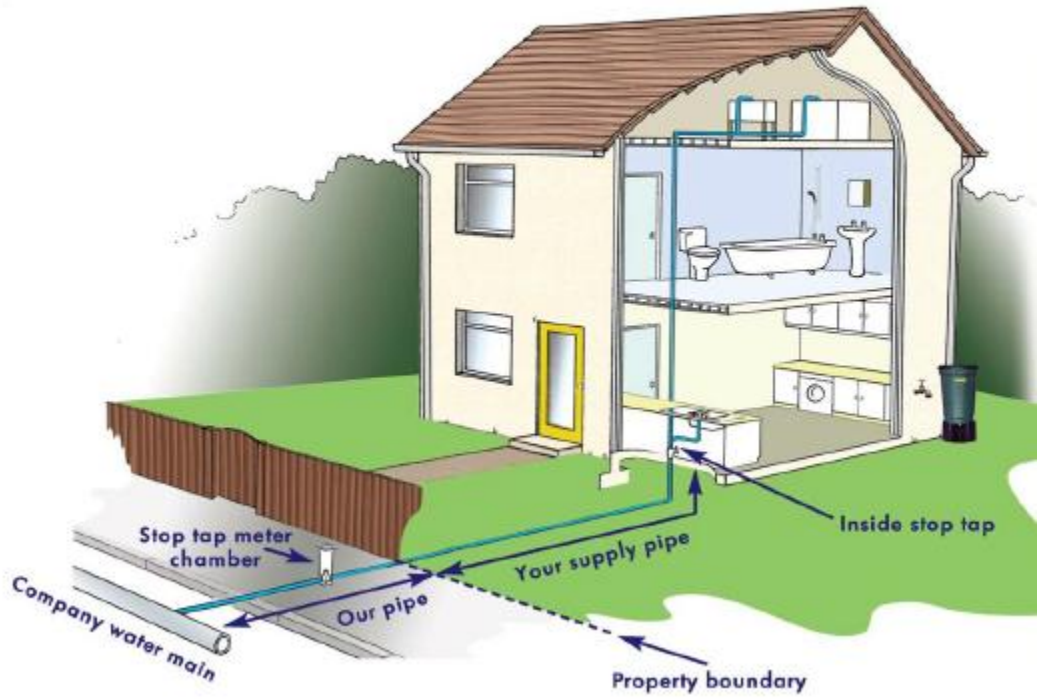
- Which one should you pick?
- Should you buy/rent?

Cloud Computing

- **Power/ heat/electricity/water supply to your home**

Before

Now



Think of it as Internet Computing

- Computation done over the internet

Enabled through:

- High Bandwidth and High Speed Internet
- Utility Computing
- Virtualization
- .....

Why Cloud Computing?

- **Large-Scale Data-Intensive Applications**
- **Flexibility**
- **Scalability**
- **Customized to current needs:**
  - Hardware
  - Software
- **Effect: Reduce Cost**
  - Reduce Maintenance
  - High Utilization
  - High Availability
  - Reduced Carbon Footprint

CO_2 Footprint

- Consolidation of servers
- Higher utilization
- Reduced power usage

Source: Knowable Magazine

Stop sending 'thank you' emails to slow down global warming

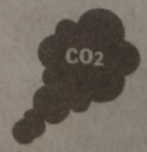
Every email you send consumes electricity and adds a tiny amount of carbon dioxide to the atmosphere. Since there are so many of us, and each one sends many emails a day, our combined email carbon footprint is enormous. The easiest way to trim it is to stop sending useless emails, such as the 'thank you' messages you send to your boss reflexively. Here's what a British survey found:

1 email=1g of CO_2

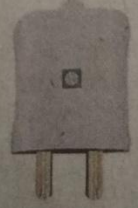
64m
Unnecessary emails sent daily in the UK

71%
Britons won't mind not receiving Thank You mails

16,433 tonnes
CO_2 emissions cut if each Briton sends 1 less email a day



That's equivalent to cancelling 81,152 London-Madrid flights, or 88,270 Delhi-Mumbai flights



Source: World Economic Forum

Flexibility

- **Software**: Any software platform
- **Access**: access resources from any machine connected to the Internet
- **Deploy infrastructure from anywhere at anytime**
  Software controls infrastructure

Scalability

- **Instant**
- **Control via software**
  - Add/cancel/rebuild resources instantly
- **Start small**, then scale your resources up/down as you need
- Illusion of **infinite resources available on demand.**

Customization

- Everything in your wish list
  - Software platforms
  - Storage
  - Network bandwidth
  - Speed

Cost

- Pay-as-you-go model
- Small/medium size companies can tap the infrastructure of corporate giants.
  - Time to service/market
  - No upfront cost

Availability and Reliability

- **Availability**
  - Having access to software, platform, infrastructure from anywhere at any time
  - All you **need is a device connected to the internet**
- **Reliability**
  - The **system's fault tolerance is managed by the cloud providers** and users no longer need to worry about it.

Maintenance

- **Reduce** the size of a client's IT department
- Is the **responsibility of the cloud vendor**
- Responsibility includes:
  - Software updates,
  - Security patches,
  - System backup,
  - Monitoring system's health,
  - Etc.

Utilization

Consolidation of a large number of resources

- CPU cycles
- Storage
- Network Bandwidth

Drawbacks

- Security
- Privacy
- Vendor lock-in
- Network-dependent
- Migration

Why call it “Cloud computing”?

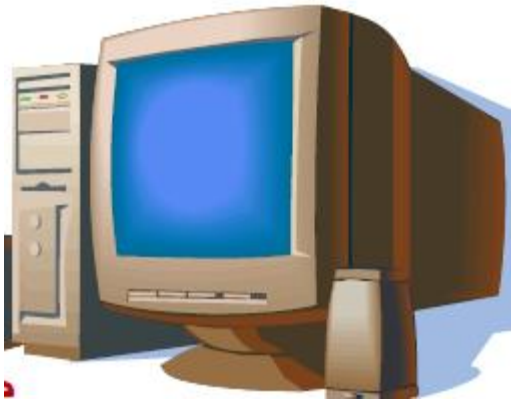
- Wikipedia: “**The term derives from the fact that most technology diagrams depict the Internet or IP availability by using a drawing of a cloud**."

Existing Computing Paradigms

- Personal Computing
- Reconfigurable Computing
- Parallel Computing
- Distributed Computing
- Ubiquitous Computing
- Autonomic Computing
- Super Computing
- Grid Computing
- Cluster Computing
- Utility Computing
- Pervasive Computing
- Mobile Computing
- Cloud Computing

Personal Computing

- Personal computing system
- Local software installation, maintenance
- Local system maintenance
- Customizable to user needs
- Very low utilization
- High up-front cost



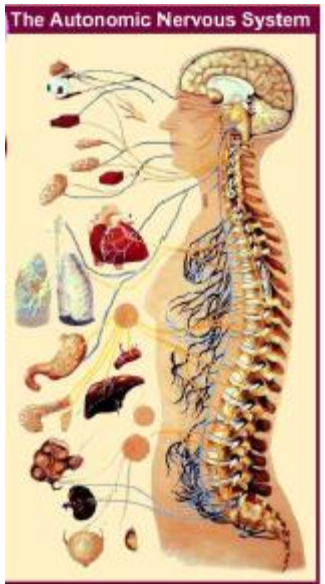
Reconfigurable Computing

Field Programmable Gate Arrays (FPGAs)Reprogrammable Hardware

- Can exploit parallel code
- Slow programming time (ms)
- Power hungry

Autonomic Computing

- Motivation: rapidly growing complexity of integrating, managing and operating computer systems
- Introduced by IBM in 2001
- Inspired by Human **Autonomic Nervous System ANS**



Mobile Computing

- You can use computing technology on the move
- Since 1990's
- Intermittent (not continuous) connectivity
- Limited Bandwidth

Utility Computing

Water, gas, and electricity are provided to every home and business as commodity services.

- get connected to the utility companies, "public" infrastructure
- get these utility services on-demand
- pay-as-you use

Utility Computing is doing same for computing resources (**processing power, bandwidth, data storage, and enterprise software services**)

Distributed Computing

- Distributed System: multiple autonomous computers connected through a communication network
- The system has a distributed memory where each processor has its private memory.
- Using distributed systems to solve large problems.
- Information exchanged using communication models, ex: MPI

Distributed Computing (Cluster Computing)

Characteristics:

- tightly coupled computers
- single system image
- Centralized Job management & scheduling system

- Better performance and availability and more cost-effectiveness over single computer with same capabilities
- Since 1987

Distributed Computing (Grid Computing)

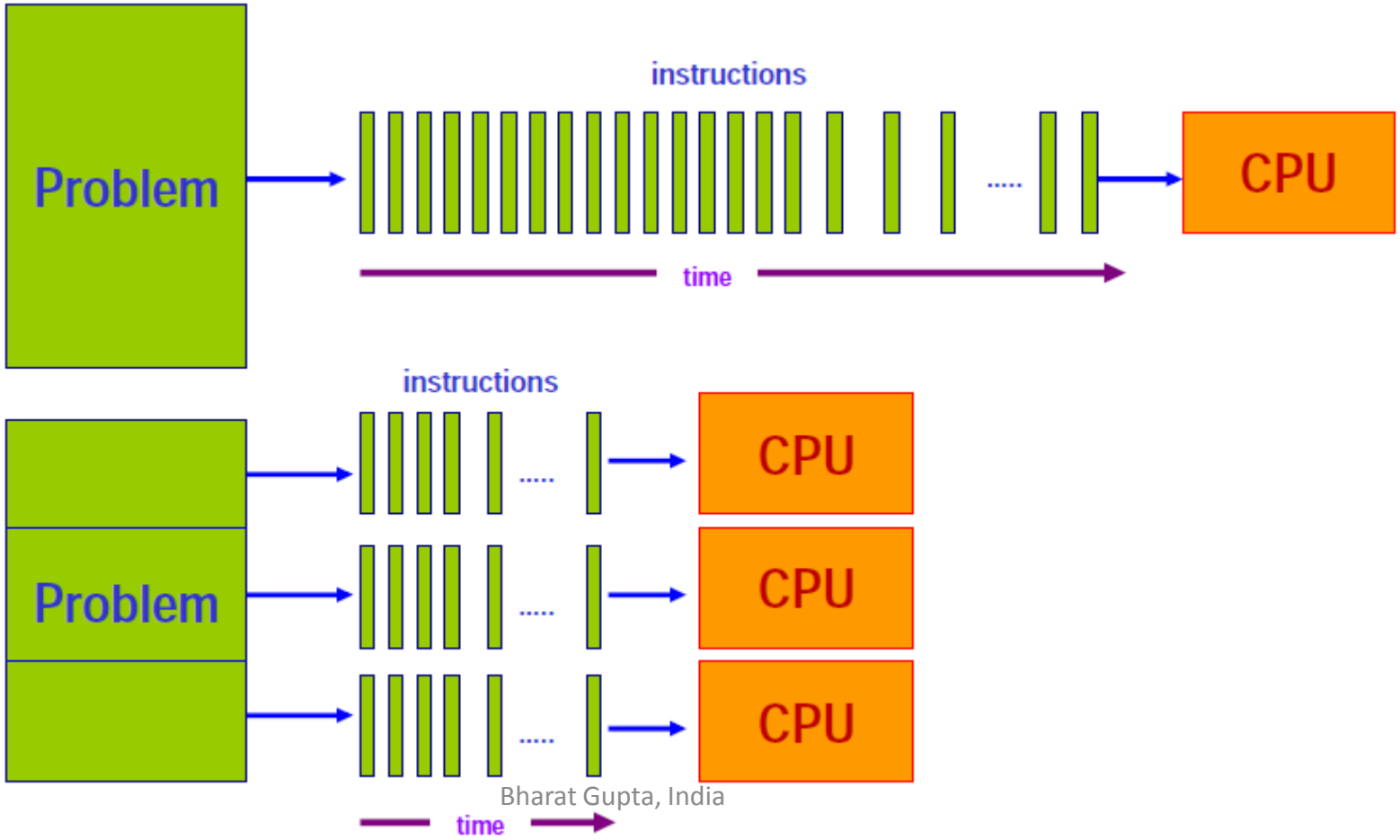
- According to Gartner, "**a grid is a collection of resources owned by multiple organizations that is coordinated to allow them to solve a common problem**."

  Characteristics:

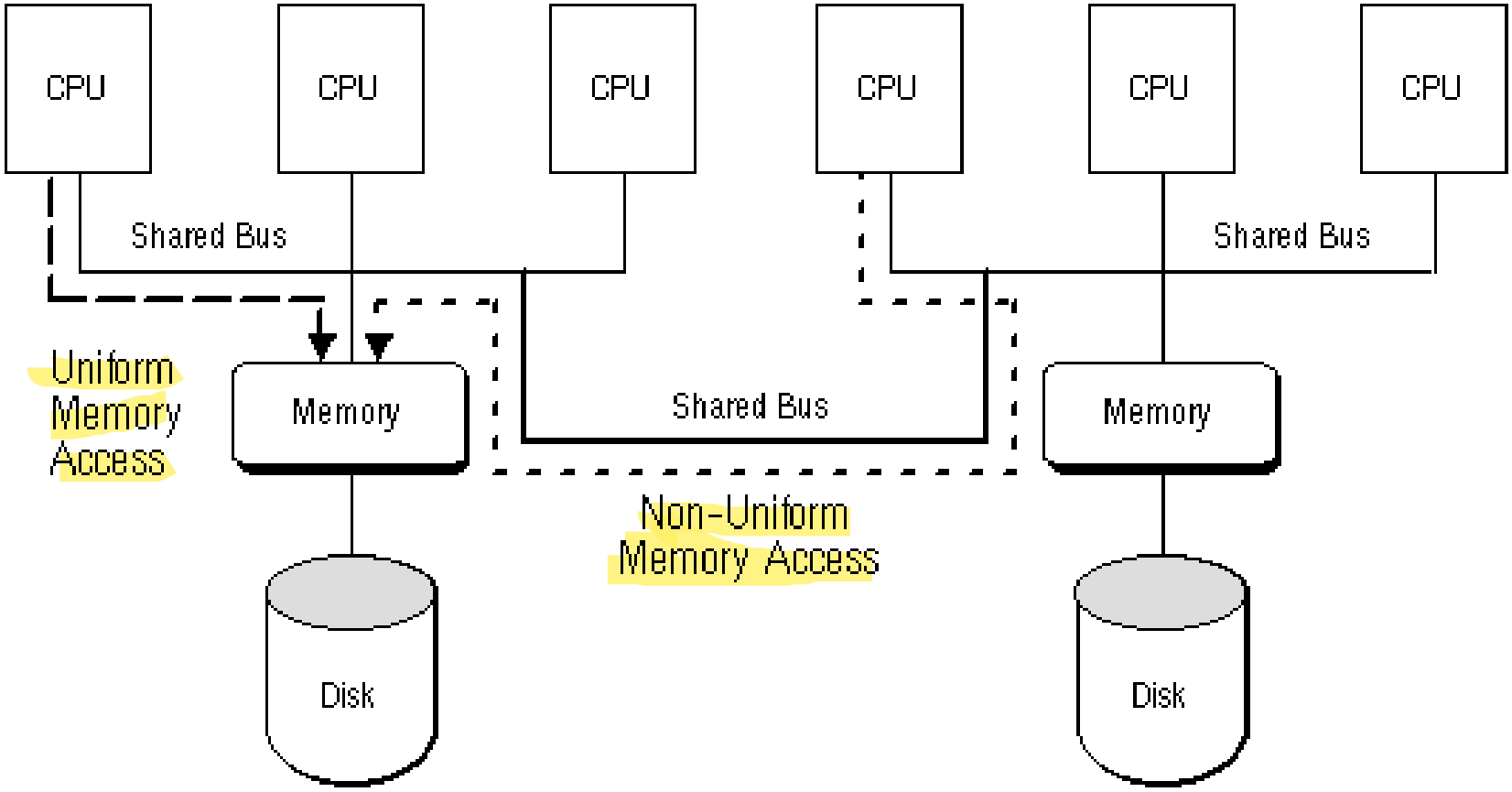
  - Loosely coupled
  - **No** Single System Image
  - Distributed Job Management & scheduling
- Originated early 1990's

Parallel Computing

- Calculations of large problems are divided into smaller parts and carried out **simultaneously/concurrently** on different processors.
- All have access to a shared memory that is used to exchange information between processors



- **Uniform memory access (UMA)** is a shared memory architecture used in parallel computing



Super Computing

- Thousands of processors
  - Used for compute-intensive problems
  - Days instead of Years
- Introduced in the 1960's

Ubiquitous Computing

- **Ubiquitous**: "seeming to be in all places"
- Ubiquitous computing can occur using any device, in any location, and in any format.
- The underlying technologies to support ubiquitous computing include
  - Internet,
  - advanced middleware,
  - operating system,
  - mobile code,
  - sensors,
  - microprocessors,
  - new I/O and user interfaces,
  - networks,
  - mobile protocols,
  - location and positioning and new materials,
  - etc.

Pervasive Computing

- Pervasive: "present or noticeable in every part of a thing or place"
- Pervasive computing is the trend of embedding computational capability (**generally in the form of microprocessors**) into everyday objects to make them effectively communicate and perform useful tasks in a way that minimizes the end user's need to interact with computers as computers.
- Information processing engaged in everyday's activities and objects.
- Goal is to make devices smart

A Cloud is

- **Data center hardware and software :** that the vendors use to offer the computing resources and services

Cloud Computing Characteristics

Common Characteristics:

- Massive Scale
- Resilient Computing
- Homogeneity
- Geographic Distribution
- Virtualization
- Service Orientation
- Low Cost Software
- Advanced Security

Essential Characteristics:

- On Demand Self-Service
- Broad Network Access
- Rapid Elasticity
- Resource Pooling
- Measured Service

Essential Cloud Characteristics

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

- **On-demand self-service:** A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed **automatically without requiring human interaction** with each service's provider.
- **Broad network access:** Capabilities are available over the network and accessed through standard mechanisms that promote use by **heterogeneous thin or thick client platforms** .
- **Resource pooling:** There is a sense of **location independence,** the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter).
  - **Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.**

- **Rapid elasticity:** Capabilities can be **rapidly** and elastically provisioned, in some cases automatically, to quickly scale out and **rapidly released** to quickly scale in.
  - To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in **any quantity at any time.**
- **Measured Service:** Cloud systems automatically control and optimize resource use by leveraging a metering capability at **some level of abstraction appropriate to the type of service** (e.g., storage, processing, bandwidth, and active user accounts).

- **Scalability:** Infrastructure capacity allows for traffic spikes and minimizes delays.
- **Resiliency:** Cloud providers have mirrored solutions to minimize downtime in the event of a disaster.
- **Homogeneity**: No matter which cloud provider and architecture an organization uses, an open cloud will make it easy for them to work with other groups, even if those other groups choose different providers and architectures.

Enabling Technologies

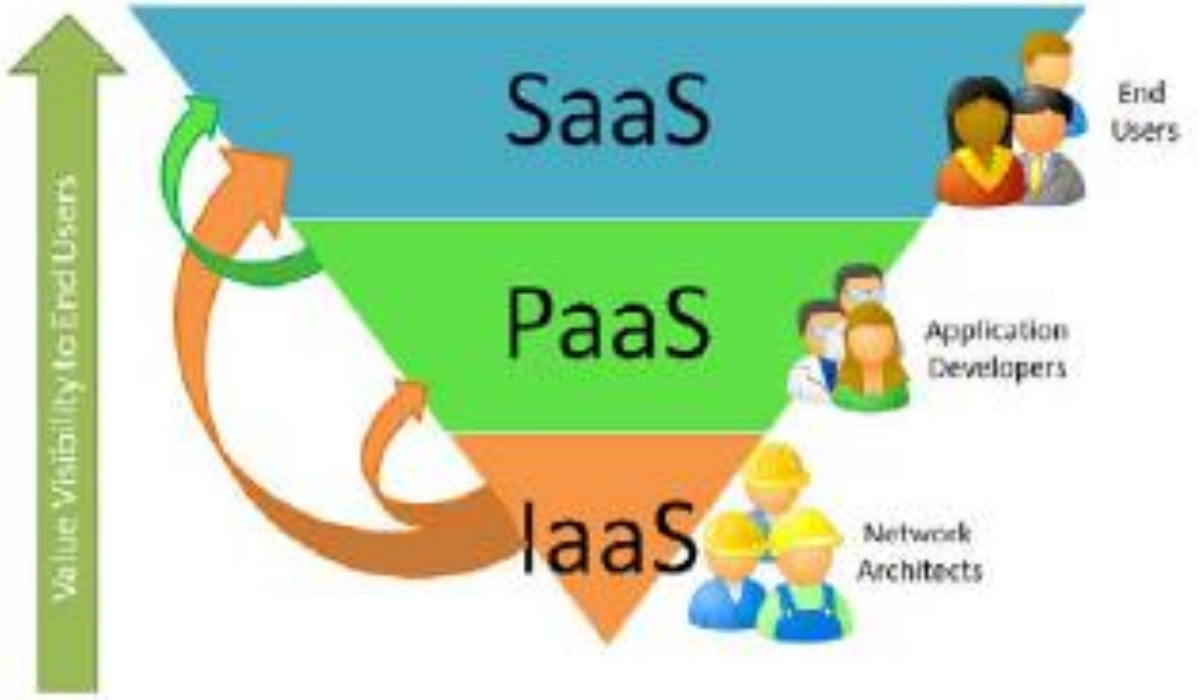
- Virtualization
- Web 2.0
- Distributed Storage
- Distributed Computing
- Utility Computing
- Network Bandwidth & Latency
- Fault-Tolerant Systems

Cloud Computing Services

- **Software as a Service (SaaS)**
  - Use provider/vendor **applications** over a network through browser
- **Platform as a Service (PaaS)**
  - Delivery of a computing **platform for custom software development** as a service
- **Infrastructure as a Service (IaaS)**
  - Deliver of computer infrastructure as a service

- **The list continues to grow (XAAS)…**

Cloud Computing Services

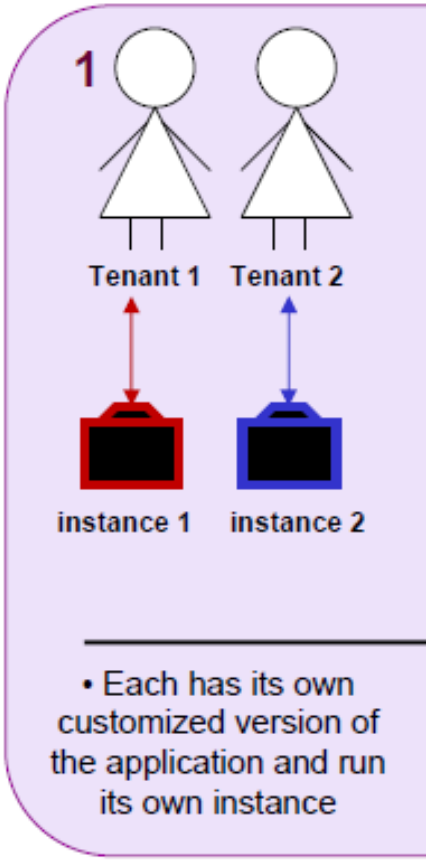


SaaS

- Started around 1999
- Application is licensed to a customer as a service on demand
- Software Delivery Model:
  - Hosted on the vendor's web servers
  - Downloaded at the consumer's device and disabled when on-demand contract is over

SaaS architecture

- **Distinguishing attributes: configurability, multi-tenant efficiency, scalability**



SaaS Examples

- ***Basecamp*** is the leading web-based project management and collaboration tool. To-dos, files, messages, schedules, and milestones
- **Salesforce**: CRM (world #1)
- **Quickbase** : QuickBase is a low-code platform for citizen development for building, customizing and connecting scalable, secure cloud applications mapped to unique business challenges without compromising IT governance and control.

PaaS

- Delivery of an **integrated computing platform** (to build/test/deploy custom apps) & solution stack as a service.
- Deploy your applications & don't worry about buying & managing the underlying hardware and software layers.

PaaS Examples

Etelos

LongJump

IaaS

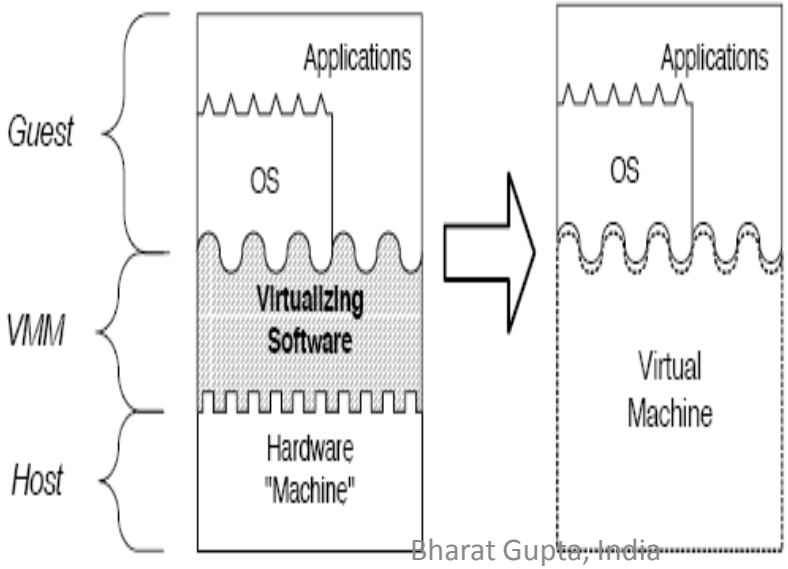
- **Delivery of computer infrastructure** (typically platform virtualization environment) **as a service**
- Buy resources
  - Servers
  - Software
  - Data center space
  - Network equipment as fully outsourced services

Virtual Machines

- **Virtual machine adds software to a physical machine to give it the appearance of a different platform or multiple platforms.**
- Eliminate real machine constraint
  - Increases portability and flexibility
- Benefits
  - Cross platform compatibility
  - Increase Security
  - Enhance Performance
  - Simplify software migration

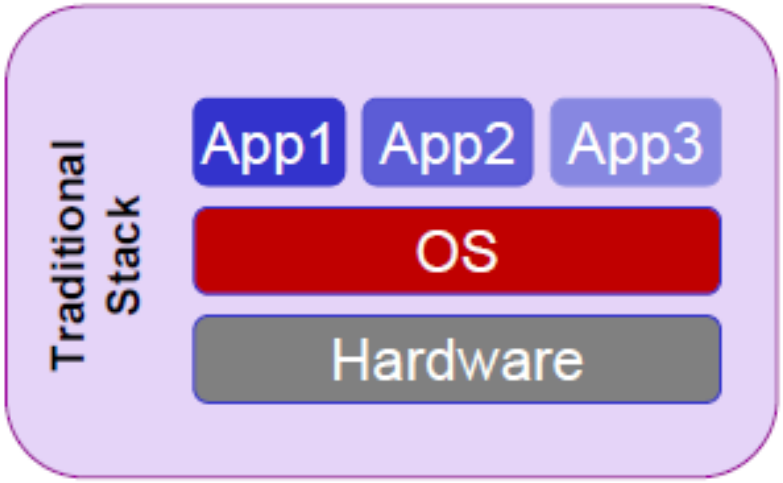
Virtual Machine Basics

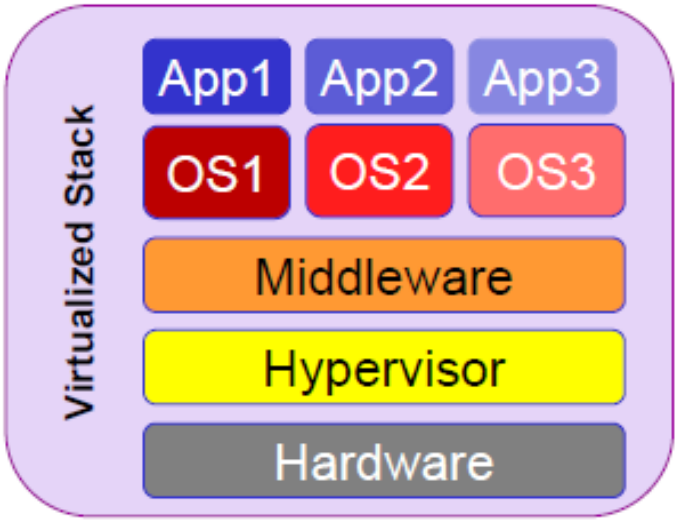
- Virtual software placed between underlying machine and conventional software
- Virtualization process involves:
  - Mapping of virtual resources (registers and memory) to real hardware resources
  - Using real machine instructions to carry out the actions specified by the virtual machine instructions



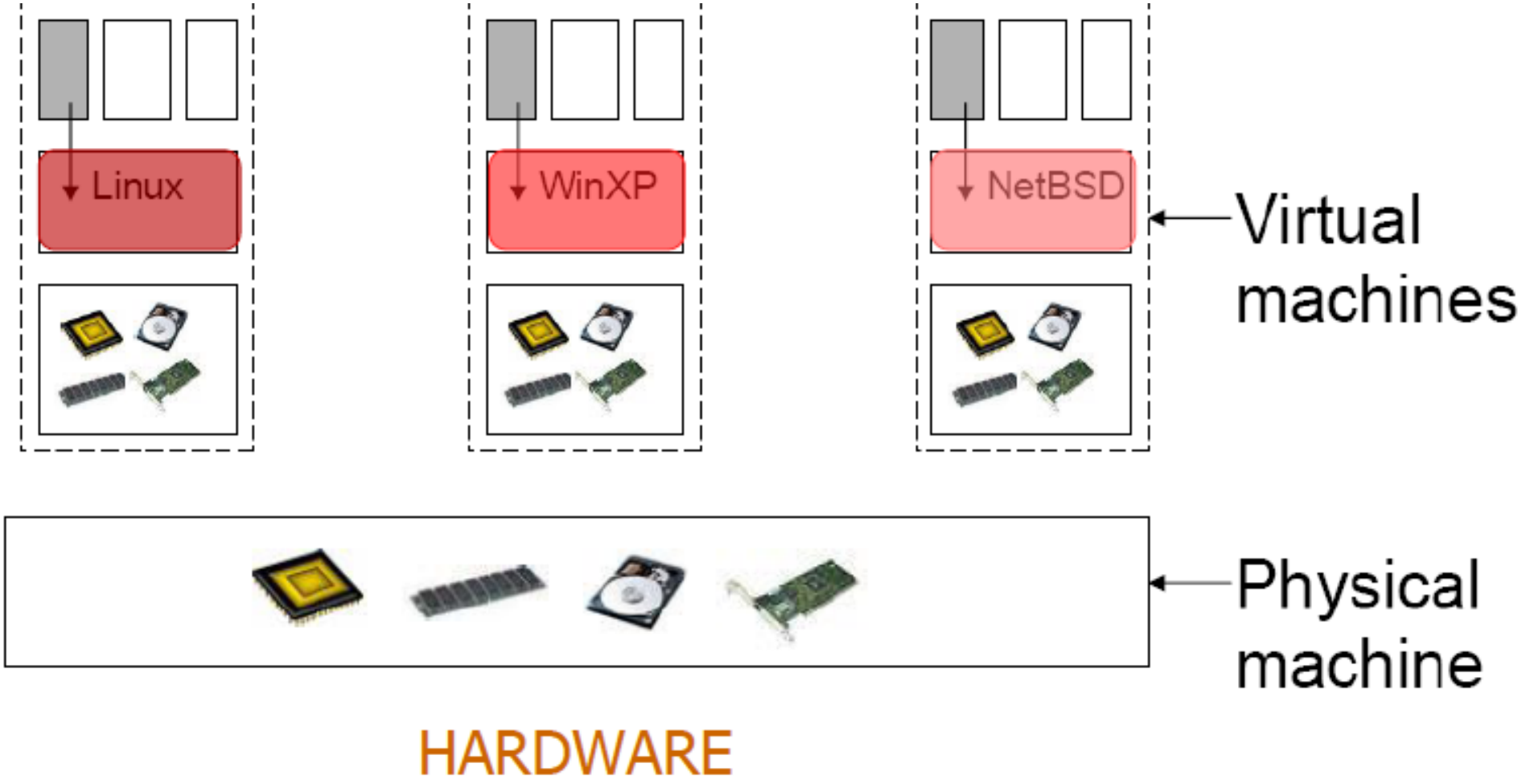
IaaS

- **Virtualization Technology is a major enabler of IaaS**
  - It's a path to share IT resource pools: Web servers, storage, data, network, software and databases.
  - Higher utilization rates





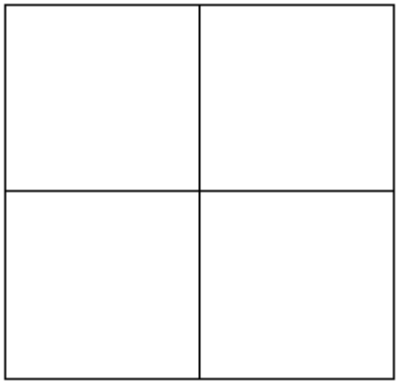
IaaS

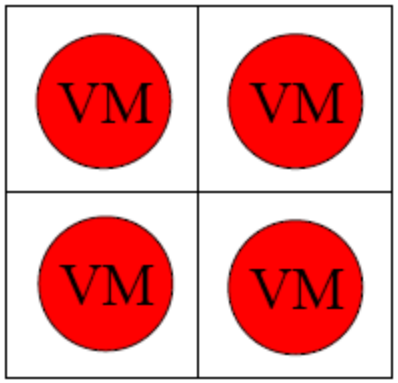

Linux
WinXP
NetBSD
Virtual machines
Physical machine
HARDWARE

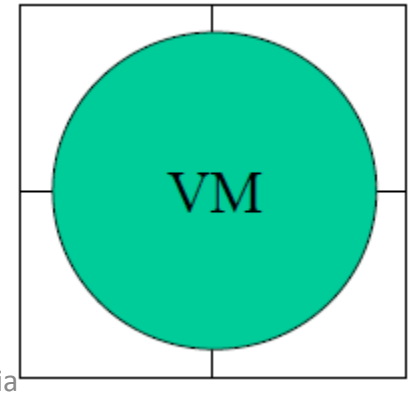
IaaS

- **Granularity of VMs:** Multi-core processors

Quad Core:

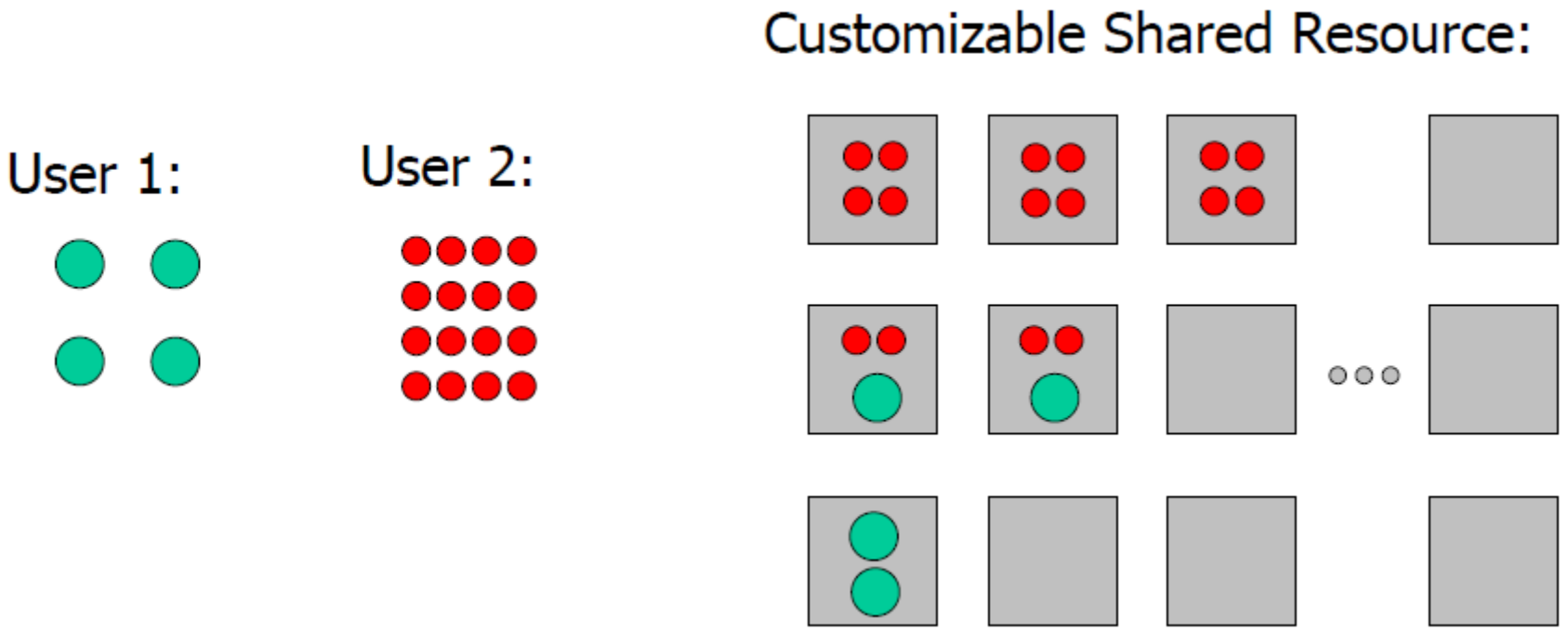




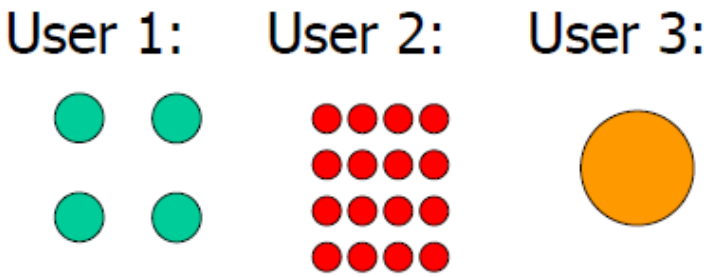


Resource sharing

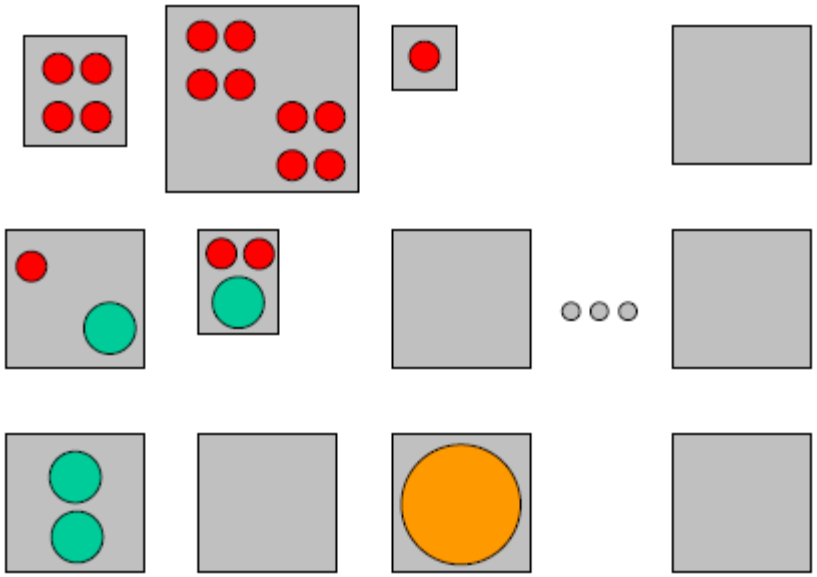
- Offering computing resources as a service through:
  - Virtualization
  - Dynamic provisioning



Heterogeneous Physical Resources



Customizable **Shared Heterogeneous Resource**



More (XaaS)

- **Data as a service (DaaS)**
- **Hardware (HaaS)**
- **.....**

Cloud Storage

- Several large Web companies are now exploiting the fact that they have data storage capacity that can be hired out to others.

- Amazon's Elastic Compute Cloud (EC2) and Simple Storage Solution (S3) are well known examples

Cloud Differences in Perspectives of Providers, Vendors, and Users

Cloud Players	IaaS	PaaS	SaaS
IT administrators/cloud providers	Monitor SLAs	Monitor SLAs and enable service platforms	Monitor SLAs and deploy software
Software developers (vendors)	To deploy and store data	Enabling platforms via configurators and APIs	Develop and deploy software
End users or business users	To deploy and store data	To develop and test Web software	Use business software

Storage Services in Three Cloud Computing Systems

Storage System	Features
GFS: Google File System	Very large sustainable reading and writing bandwidth, mostly continuous accessing instead of random accessing. The programming interface is similar to that of the POSIX file system accessing interface.
HDFS: Hadoop Distributed File System	The open source clone of GFS. Written in Java. The programming interfaces are similar to POSIX but not identical.
Amazon S3 and EBS	S3 is used for retrieving and storing data from/to remote servers. EBS is built on top of S3 for using virtual disks in running EC2 instances.

Five Major Cloud Platforms and Their Service Offerings

Model	IBM	Amazon	Google	Microsoft	Salesforce
PaaS	BlueCloud, WCA, RC2		App Engine (GAE)	Windows Azure	Force.com
IaaS	Ensembles	AWS		Windows Azure	
SaaS	Lotus Live		Gmail, Docs	.NET service, Dynamic CRM	Online CRM, Gifttag
Virtualization		OS and Xen	Application Container	OS level/ Hypel-V	
Service Offerings	SOA, B2, TSAM, RAD, Web 2.0	EC2, S3, SQS, SimpleDB	GFS, Chubby, BigTable, MapReduce	Live, SQL Hotmail	Apex, visual force, record security
Security Features	WebSphere2 and PowerVM tuned for protection	PKI, VPN, EBS to recover from failure	Chubby locks for security enforcement	Replicated data, rule-based access control	Admin./record security, uses metadata API
User Interfaces		EC2 command-line tools	Web-based admin. console	Windows Azure portal	
Web API	Yes	Yes	Yes	Yes	Yes
Programming Support	AMI		Python	.NET Framework	

Note: *WCA: WebSphere CloudBurst Appliance; RC2: Research Compute Cloud; RAD: Rational Application Developer; SOA: Service-Oriented Architecture; TSAM: Tivoli Service Automation Manager; EC2: Elastic Compute Cloud; S3: Simple Storage Service; SQS: Simple Queue Service; GAE: Google App Engine; AWS: Amazon Web Services; SQL: Structured Query Language; EBS: Elastic Block Store; CRM: Consumer Relationship Management.*

Infrastructure as a service (IaaS)

- Most basic cloud service model
- Cloud providers offer computers, as physical or more often as virtual machines, and other resources.
- Virtual machines are run as guests by a hypervisor, such as **Xen or KVM.**

Some IaaS Offerings from Public Clouds

Table 1. Worldwide IaaS Public Cloud Services Market Share, 2019-2020 (Millions of U.S. Dollars)

Company	2020 Revenue	2020 Market Share (%)	2019 Revenue	2019 Market Share (%)	2019-2020 Growth (%)
Amazon	26,201	40.8	20,365	44.6	28.7
Microsoft	12,658	19.7	7,950	17.4	59.2
Alibaba	6,117	9.5	4,004	8.8	52.8
Google	3,932	6.1	2,367	5.2	66.1
Huawei	2,672	4.2	882	1.9	202.8
Others	12,706	19.8	10,115	22.1	25.6
Total	**64,286**	**100.0**	**45,684**	**100.0**	**40.7**

Source: Gartner (June 2021)

Platform as a service (PaaS)

- Cloud providers deliver a **computing platform** typically **including operating system, programming language execution environment, database, and web server.**
- Examples of PaaS include: Amazon Elastic Beanstalk, Cloud Foundry, Heroku, Force.com, EngineYard, Mendix, Google App Engine, Microsoft Azure and OrangeScape.

PaaS Offerings from Public Clouds

Software as a service (SaaS)

- Cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients.
- Examples of SaaS include: Google Apps,Quickbooks Online, Limelight Video Platform, Salesforce.com, and Microsoft Office 365.