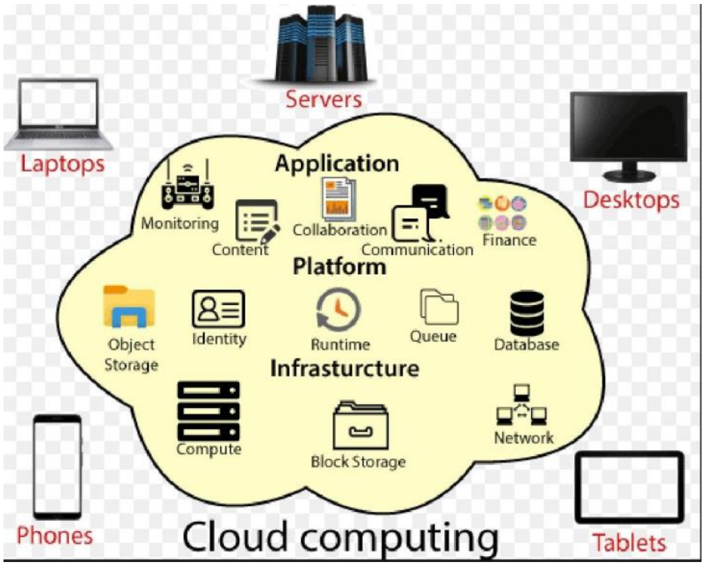
CS8791 **CLOUD COMPUTING** **L T P C**
3 0 0 3

UNIT I INTRODUCTION (9)

Introduction to Cloud Computing – Definition of Cloud – Evolution of Cloud Computing – Underlying Principles of Parallel and Distributed Computing – Cloud Characteristics – Elasticity in Cloud – On- demand Provisioning.

Cloud computing consists of three distinct types of **computing** services delivered remotely to clients via the internet. Clients typically pay a monthly or annual service fee to providers, to gain access to systems that deliver software as a service, platforms as a service and infrastructure as a service to subscribers. **Cloud Computing** is the delivery of **computing** services such as servers, storage, databases, networking, software, analytics, intelligence, and more, over the **Cloud.**



A Brief History of Cloud Computing

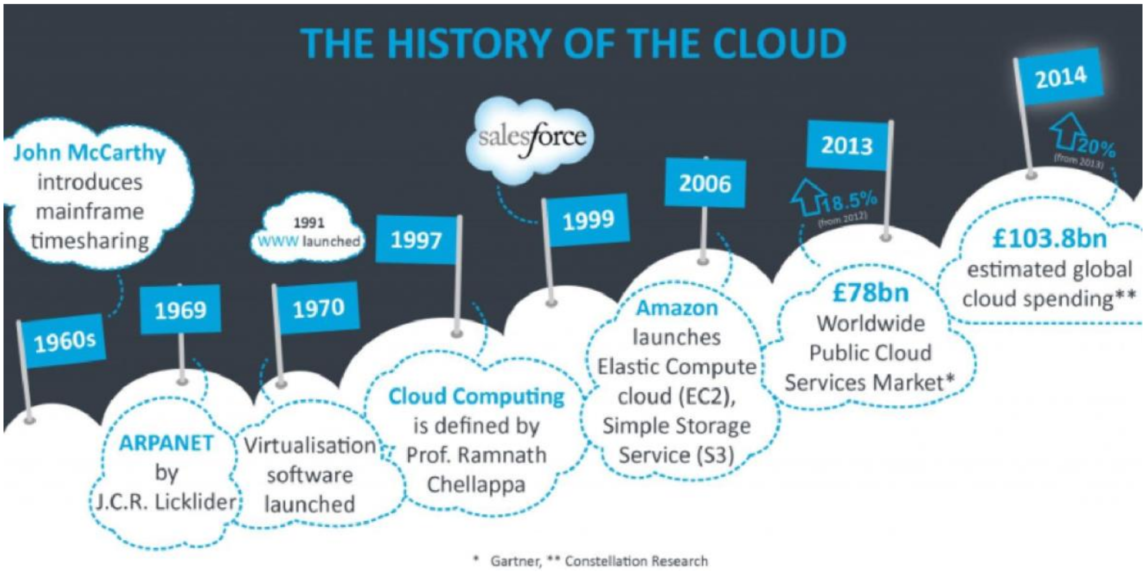
When we think of the cloud, rarely do we cast our minds back to times before the 21stcentury. After all, it’s really just been over the past decade or so that cloud computing really started to develop into the giant, omnipresent and all-powerful behemoth we know today. But the truth is that concepts of the cloud have existed for many, many years, and in fact can be traced as far back as the 1950s with mainframe computing. In those early days, mainframe computers were huge machines, and very, very expensive – too expensive to buy and maintain one for every single employee. And of course, not every single employee needed access to one at all times like they do today. As such, most organizations would purchase just one or two machines, and then implement –time-sharing‖ schedules which enabled multiple users to access the central mainframe computer from connected stations. These stations were known as –dumb terminals‖, and provided no processing power of their own. Even so, this type of shared computational power is the basic, underlying premise of cloud computing, and where it all began.

In the mid-1960s, a major advancement in cloud computing came when American computer scientist J.C.R. Licklider conceptualized an interconnected system of computers. In 1969, –Lick, as he is often known, helped develop a very primitive version of the Internet, known as the Advanced Research Projects Agency Network (ARPANET). ARPANET was the first network that allowed digital sources to be shared among computers that were not in the same physical location. Lick's vision was also for a world where everyone would be interconnected by way of computers and able to access information from anywhere. Sound familiar? Of course it does – it's the Internet as we know it, and a necessity for accessing all the benefits that the cloud realizes.

Over the decades that followed, much further advancement in cloud technology came into being. In 1972, for example, IBM released an operating system (OS) called the Virtual Machine (VM) operating system. Virtualization describes a virtual computer that acts just like a real one, with a fully-operational OS. The concept evolved with the Internet, and businesses began offering –virtual‖ private networks as a rentable service, eventually leading to the development of the modern cloud computing infrastructure in the 1990s.

Also in this decade, telecommunications companies began offering virtualized private networks, which had the same service quality as their dedicated point-to-point data connections at a reduced cost. Instead of building out physical infrastructure to allow for more users to have their own connections, telecommunications companies were now able to provide users with shared access to the same physical infrastructure.

In the early 2000s, Amazon Web Services (AWS) emerged, and Amazon launched Elastic Compute Cloud (EC2) in 2006, allowing companies and individuals to rent virtual computers through which they could use their own programs and applications. In the same year, Google launched its Google Docs services, allowing users to save, edit and transfer documents in the cloud. In 2007, IBM, Google, and several universities joined forces to develop a server farm for research projects. It was also the year that Netflix launched its video streaming service, using the cloud to stream movies and other video content into the homes and onto the computers of thousands (and eventually millions) of subscribers worldwide.


THE HISTORY OF THE CLOUD
John McCarthy introduces mainframe timesharing
1960s
1969
ARPANET by J.C.R. Licklider
1991 WWW launched
1970
Virtualisation software launched
1997
Cloud Computing is defined by Prof. Ramnath Chellappa
salesforce
1999
2006
Amazon launches Elastic Compute cloud (EC2), Simple Storage Service (S3)
2013
18.5% (from 2012)
£78bn Worldwide Public Cloud Services Market*
2014
20% (from 2013)
£103.8bn estimated global cloud spending**

* Gartner, ** Constellation Research

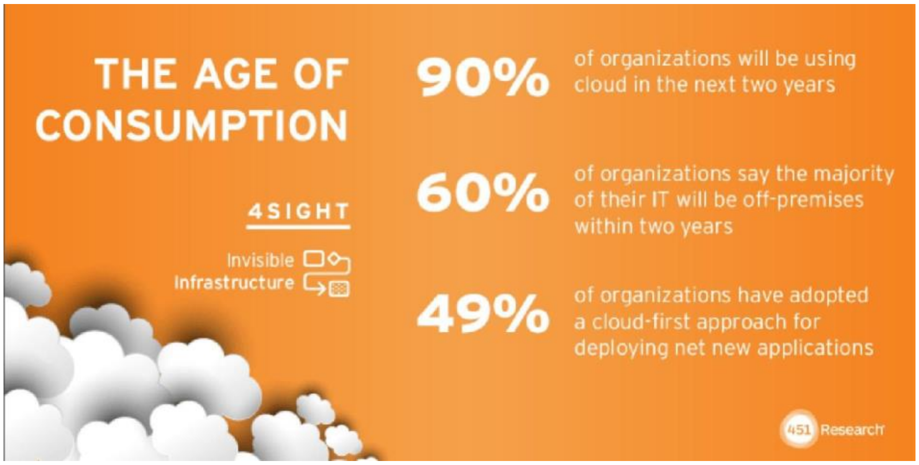
Cloud Computing Today

Over the past decade, cloud computing adoption has seen explosive growth – at both consumer and enterprise levels. Legacy software providers such as Microsoft, Oracle and Adobe have all made huge, concerted efforts to encourage users of their on-premises software offerings to upgrade to their cloud equivalents, which are usually offered on a subscription pay-as-you-go basis. At the same time, we have seen a cornucopia of cloud-native providers – such as Zendesk, Workday and ServiceNow – emerge with Software as a Service (SaaS) offerings that are (and have always been) only available in the cloud. And it's not only Software as a Service that has emerged, of course, but Platform as a Service (PaaS), Infrastructure as a Service (IaaS), Backup as a Service (BaaS) and Disaster Recovery as a Service (DRaaS) as well. Pretty much Everything as a Service (or XaaS, as it is peddled by companies like Google and Microsoft who offer such holistic resources) is now available.

Over the course of the last ten years or so, cloud computing has evolved from being something that service providers told companies they should be adopting, to the very lifeblood that runs through most modern enterprises.

As such, organizations have become increasingly accustomed to the pay-as-you-go cloud billing model, and now look upon IT purchases as a day-to-day expense, rather than a one-off investment that they will be stuck with for the foreseeable future.

In fact, 451 Research has found that 90% of organizations will be using some form of cloud computing services in the next two years, with 60% saying that the majority of their IT will be off-premise. What's more, 49% of organizations have adopted a cloud-first approach for deploying net new applications. 451 Research also predicts the cloud computing market will reach $53.3 billion in 2021 – up from $28.1 billion in 2017.



THE AGE OF CONSUMPTION

4SIGHT

Invisible Infrastructure

90% of organizations will be using cloud in the next two years

60% of organizations say the majority of their IT will be off-premises within two years

49% of organizations have adopted a cloud-first approach for deploying net new applications

451 Research

Cloud Computing Trends to Prepare For

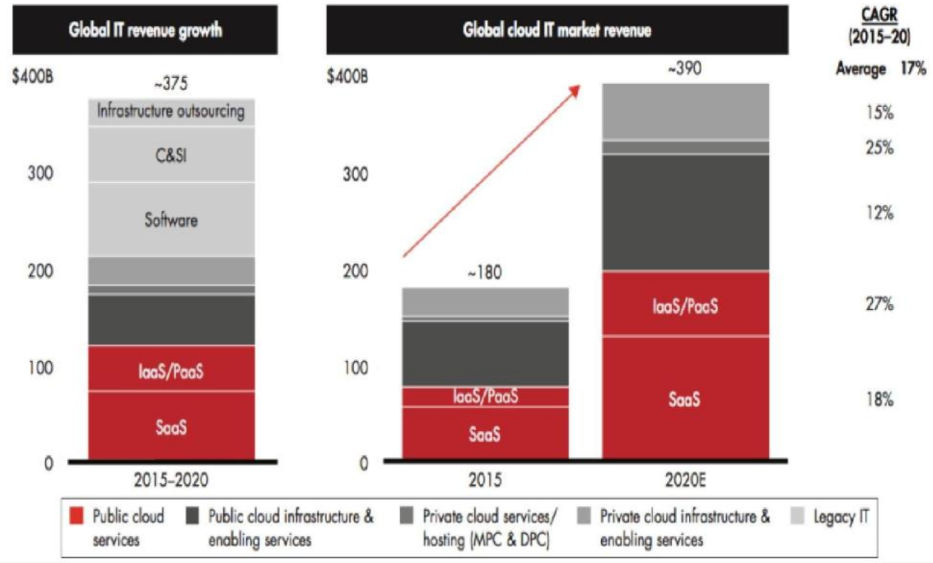
So what does this all mean for organizations and their CIOs and CTOs in the future? Where is cloud computing going? For one thing, it's clear that cloud computing has moved beyond being simply another option on the table, and has transformed into a way of doing business. Enterprises will be able to increasingly consume IT without the constraints of having to build and maintain its underlying infrastructure.

451 Research analyses also reveals that a slim majority of enterprise workloads will run off-premises IT environments by 2019, and that more than one-third of all workloads will operate in public cloud environments. In addition, research by IDC reveals that almost half of IT spending will be cloud-based in 2018, and that it will reach 60% of all IT infrastructures and 60- 70% of all software, services and technology spending by 2020. As such, it's imperative that CIOs and CTOs view cloud computing as a critical element of their organization's competitiveness, and that they explore which of their services, operations and offerings would be better served by moving them to the cloud.

Here are three cloud computing trends that CIOs and CTOs need to be prepared for in 2019.

1. **The Number of Cloud-based Services and Solutions Will Continue to Rise**

Indeed, they will. According to Bain & Company, subscription-based SaaS solutions will grow at an 18% CAGR by 2020, IaaS/PaaS at 27%, and public cloud infrastructure and enabling services at 12%. In all, cloud computing hardware, software and services are capturing 60% of all IT market growth.



2. **Hybrid and Multi-Cloud Strategies Will Play a Key Role**

Making a full cloud transformation is a challenging prospect for most enterprises. As such, hybrid and multi-cloud cloud strategies will play a key role in cloud transitions in the near future. With a multi-cloud strategy, organizations may, for example, use their own on-premises dedicated servers (private cloud) for the hosting of particularly sensitive workloads, with those that are deemed less critical hosted with a public cloud provider. Similar though distinct, a hybrid cloud solution mixes on-premises private cloud and third-party public cloud with defined orchestration between the two.

Hybrid and multi-cloud strategies are often seen as means for an organization to dip its toes into those cloudy waters, enabling a steady-paced transition with less risk and less expense. In 2019, many companies will continue such blended cloud approaches, allowing them to access the efficiency and effectiveness of public cloud computing, while still retaining central control of their most sensitive and mission-critical items. 451 Research predicts that 69% of enterprises will have either multi-cloud or hybrid cloud IT environments by 2019.

3. **Security and Compliance Will Be Critical**

As more and more businesses migrate more and more of their services and functions to the cloud, security and regulation compliance will become an increasing concern. Hackers go where the data goes, and as more data is pushed to the cloud, so too will the hackers trying to breach the system and steal valuable databases to sell on the dark web.

In May 2018, we saw the EU’s General Data Protection Regulation (GDPR) come into effect, which of course has implications for all global enterprises. Cloud computing compliance under the GDPR is not easy, and many organizations are not prepared. In fact, a recent survey from Commvault revealed that only 12% of global IT organizations understand how GDPR will affect their cloud services. And GDPR is likely only the beginning – as governments around the globe start recognizing the risks, cloud computing will no doubt start to become highly regulated.

Parallel And Distributed Computing

The simultaneous growth in availability of big data and in the number of simultaneous users on the Internet places particular pressure on the need to carry out computing tasks ―in parallel,‖ or simultaneously. Parallel and distributed computing occurs across many different topic areas in computer science, including algorithms, computer architecture, networks, operating systems, and software engineering. During the early 21st century there was explosive growth in multiprocessor design and other strategies for complex applications to run faster. Parallel and distributed computing builds on fundamental systems concepts, such as concurrency, mutual exclusion, consistency in state/memory manipulation, message-passing, and shared- memory models.

Creating a multiprocessor from a number of single CPUs requires physical links and a mechanism for communication among the processors so that they may operate in parallel. Tightly coupled multiprocessors share memory and hence may communicate by storing information in memory accessible by all processors. Loosely coupled multiprocessors, including computer networks, communicate by sending messages to each other across the physical links. Computer scientists have investigated various multiprocessor architectures. For example, the possible configurations in which hundreds or even thousands of processors may be linked together are examined to find the geometry that supports the most efficient system throughput. A much-studied topology is the hypercube, in which each processor is connected directly to some fixed number of neighbours: two for the two-dimensional square, three for the three-dimensional cube, and similarly for the higher- dimensional hypercubes. Computer scientists also investigate methods for carrying out computations on such multiprocessor machines (e.g., algorithms to make optimal use of the architecture and techniques to avoid conflicts in data transmission).

The machine-resident software that makes possible the use of a particular machine, in particular its operating system, is an integral part of this investigation. Concurrency refers to the execution of more than one procedure at the same time (perhaps with the access of shared data), either truly simultaneously (as on a multiprocessor) or in an unpredictably interleaved order. Modern programming languages such as Java include both encapsulation and features called

‑threads that allow the programmer to define the synchronization that occurs among concurrent procedures or tasks.

Two important issues in concurrency control are known as deadlocks and race conditions. Deadlock occurs when a resource held indefinitely by one process is requested by two or more other processes simultaneously. As a result, none of the processes that call for the resource can continue; they are deadlocked, waiting for the resource to be freed. An operating system can handle this situation with various prevention or detection and recovery techniques. A race condition, on the other hand, occurs when two or more concurrent processes assign a different value to a variable, and the result depends on which process assigns the variable first (or last).

Preventing deadlocks and race conditions is fundamentally important, since it ensures the integrity of the underlying application. A general prevention strategy is called process synchronization. Synchronization requires that one process wait for another to complete some operation before proceeding. For example, one process (a writer) may be writing data to a certain main memory area, while another process (a reader) may want to read data from that area. The reader and writer must be synchronized so that the writer does not overwrite existing data until the reader has processed it. Similarly, the reader should not start to read until data has been written in the area. With the advent of networks, distributed computing became feasible. A distributed computation is one that is carried out by a group of linked computers working cooperatively. Such computing usually requires a distributed operating system to manage the distributed resources. Important concerns are workload sharing, which attempts to take advantage of access to multiple computers to complete jobs faster; task migration, which supports workload sharing by efficiently distributing jobs among machines; and automatic task replication, which occurs at different sites for greater reliability.

Cloud computing's characteristics and benefits include on- demand self-service, broad *network access, and being very elastic and scalable.* As cloud computing services mature both commercially and technologically, it will be easier for companies to maximize the potential benefits. Knowing what cloud computing is and what it does, however, is just as important. The National Institute of Standards and Technology (NIST) defines cloud computing as it is known today through five particular characteristics.

1. On-demand self-service

Cloud computing resources can be provisioned without human interaction from the service provider. In other words, a manufacturing organization can provision additional computing resources as needed without going through the cloud service provider. This can be a storage space, virtual machine instances, database instances, and so on. Manufacturing organizations can use a web self-service portal as an interface to access their cloud accounts to see their cloud services, their usage, and also to provision and de-provision services as they need to.

2. Broad network access

Cloud computing resources are available over the network and can be accessed by diverse customer platforms. It other words, cloud services are available over a network—ideally high broadband communication link—such as the internet, or in the case of a private clouds it could be a local area network (LAN). Network bandwidth and latency are very important aspects of cloud computing and broad network access, because they relate to the quality of service (QoS) on the network. This is important for serving time sensitive manufacturing applications.

3. Multi-tenancy and resource pooling

Cloud computing resources are designed to support a multi-tenant model. Multi-tenancy allows multiple customers to share the same applications or the same physical infrastructure while retaining privacy and security over their information. It’s similar to people living in an apartment building, sharing the same building infrastructure but they still have their own apartments and privacy within that infrastructure. That is how cloud multi-tenancy works.

Resource pooling means that multiple customers are serviced from the same physical resources. Providers’ resource pool should be very large and flexible enough to service multiple client requirements and to provide for economy of scale. When it comes to resource pooling, resource allocation must not impact performances of critical manufacturing applications.

4. Rapid elasticity and scalability

One of the great things about cloud computing is the ability to quickly provision resources in the cloud as manufacturing organizations need them. And then to remove them when they don’t need them. Cloud computing resources can scale up or down rapidly and, in some cases, automatically, in response to business demands. It is a key feature of cloud computing. The usage, capacity, and therefore cost, can be scaled up or down with no additional contract or penalties.Elasticity is a landmark of cloud computing and it implies that manufacturing organizations can rapidly provision and de-provision any of the cloud computing resources. Rapid provisioning and de-provisioning might apply to storage or virtual machines or customer applications.

With cloud computing scalability, there is less capital expenditure on the cloud customer side. This is because as the cloud customer needs additional computing resources, they can simply provision them as needed, and they are available right away. Scalability is more planned and gradual. For instance, scalability means that manufacturing organizations are gradually planning for more capacity and of course the cloud can handle that scaling up or scaling down.

Just-in-time (JIT) service is the notion of requiring cloud elasticity either to provision more resources in the cloud or less. For example, if a manufacturing organization all of a sudden needs more computing power to perform some kind of complex calculation, this would be cloud elasticity that would be a just-in-time service. On the other hand, if the manufacturing organization needs to provision human-machine interface (HMI) tags in the database for a manufacturing project, that is not really just-in-time service, it is planned ahead of time. So it is more on the scalability side than elasticity.

Another feature available for rapid elasticity and scalability in the cloud is related to testing of manufacturing applications. If a manufacturing organization needs, for example, a few virtual machines to test a supervisory control and data acquisition (SCADA) system before they roll it out in production, they can have it up and running in minutes instead of physically ordering and waiting for hardware to be shipped. In terms of the bottom line, when manufacturing organizations need to test something in the cloud, they are paying for what they use as they use it. As long as they remember to de-provision it, they will no longer be paying for it. There is no capital expense here for computer resources. Manufacturing organizations are using the cloud provider’s investment in cloud computing resources instead. This is really useful for testing smart manufacturing solutions.

5. Measured service

Cloud computing resources usage is metered and manufacturing organizations pay accordingly for what they have used. Resource utilization can be optimized by leveraging charge-per-use capabilities. This means that cloud resource usage—whether virtual server instances that are running or storage in the cloud—gets monitored, measured and reported by the cloud service provider. The cost model is based on ―pay for what you use‖—the payment is variable based on the actual consumption by the manufacturing organization.

6. Elasticity (cloud computing)

In cloud computing, **elasticity** is defined as "the degree to which a system is able to adapt to workload changes by provisioning and de-provisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible". Elasticity is a defining characteristic that differentiates cloud computing from previously proposed computing paradigms, such as grid computing. The dynamic adaptation of capacity, e.g., by altering the use of computing resources, to meet a varying workload is called "elastic computing".

Let us illustrate elasticity through a simple example of a service provider who wants to run a website on an IaaS cloud. At moment , the website is unpopular and a single machine (most commonly a virtual machine) is sufficient to serve all web users. At moment , the website suddenly becomes popular, for example, as a result of a flash crowd, and a single machine is no longer sufficient to serve all users. Based on the number of web users simultaneously accessing the website and the resource requirements of the web server, it might be that ten machines are needed. An elastic system should immediately detect this condition and provision nine additional machines from the cloud, so as to serve all web users responsively.

At time , the website becomes unpopular again. The ten machines that are currently allocated to the website are mostly idle and a single machine would be sufficient to serve the few users who are accessing the website. An elastic system should immediately detect this condition and deprovision nine machines and release them to the cloud.

7. Cloud provisioning

Cloud provisioning refers to the processes for the deployment and integration of cloud computing services within an enterprise IT infrastructure. This is a broad term that incorporates the policies, procedures and an enterprise's objective in sourcing cloud services and solutions from a cloud service provider. Cloud provisioning primarily defines how, what and when an organization will provision cloud services. These services can be internal, public or hybrid cloud products and solutions. There are three different delivery models:

- Dynamic/On-Demand Provisioning: The customer or requesting application is provided with resources on run time.
- User Provisioning: The user/customer adds a cloud device or device themselves.
- Post-Sales/Advanced Provisioning: The customer is provided with the resource upon contract/service signup.

From a provider's standpoint, cloud provisioning can include the supply and assignment of required cloud resources to the customer. For example, the creation of virtual machines, the allocation of storage capacity and/or granting access to cloud