CHAPTER - 2
LITERATURE REVIEW

2.1 HISTORY OF CLOUD COMPUTING

The history of computing began with the invention of abacus, though the era of modern computing started in 1941, with the first electromechanical computer “Z3”. It was the world's first programmable, fully automatic computing machine. John Mauchley and J. Presper Eckert built the ENIAC (Electronic Numerical Integrator and Computer) four years later. ENIAC was the first electronic general-purpose computer, programmed to solve a full range of computing problems (Shurkin, 1996) [36].

The sixties and seventies showed the mainframe computers, used by various organizations and business corporations. These computers were used to solve critical applications, process bulk data, accomplish resource planning, and undergo a myriad financial transactions. In 1947 transistors and later on in 1969 microprocessors came into existence, which boosted the invention of present day computers. With the advent of commercial personal computers by IBM in the early 1980s, the computer industry was totally revolutionized and it led to commercial success of organized data processing. As a result various other computer manufacturers developed their own PCs compatible with IBM PCs. Only apple computers Inc., developed its own non-IBM compatible computers, the *Apple* and *Macintosh*.

Cloud computing got its boost by the development and popularity of the internet, which was invented in 1969 by Advanced Research Projects Agency (ARPA) as a communication network that could survived a nuclear war. Keeping internet in mind, operating systems and application software were modified and accordingly were the attached peripheral devices, also stored information of the computer.

With the invention of Email and World Wide Web, combined with the fast networking technologies like ADSL and Ethernet, the networking became easily available almost everywhere (Freiberger & Swaine, 2000) [13]. The high bandwidth and technologies like Java and Ajax or Web Services have allowed the development of highly interactive websites and then around the year 2000 the deployment of whole applications through Internet browsers, which was popular under the name of ‘Software-as-a-Service’ (Finch, 2006) [11]. In analogy with the provision of software through the Internet, computing power was also delivered through a network. Grid Computing was well established since the beginning of the nineties in the academic field (Foster & Kesselman, 2003) [12].

With the deployment of internet browsers, grid computing established itself in the field of academics in 1990s.

Another milestone was reached in 2007 when Google began building large data centers for the benefit of students, so that they could tap the internet and to program and research from remote(Brodkin, 2007) [5]. “This was a new model in which computing chores could move from individual desktops or from computer centers and could be handled as services over the internet” (Bohm, M. and Lohar, S. 2007) [2,25]. This concept was called cloud computing. Cloud computing is a modern and amazing evaluation and gives rise to many trends.

Cloud computing represents a development and large faction of many trends (Buyya, 2009) [6]. Amazon came out with the first public cloud computing services and shocked the IT giants like IBM, and SUN, who were also working on the same concept. Amazon used the cloud computing service with its online book store in 1995 and diversified into CDs and DVDs and other form of digital media, jewelry, further apparel, grocery, automotive parts and accessories, apart from computer hardware and software. Amazon emphasized on designing a secure financial transaction processing system, to ensure that data of their customers and retail partners could not be compromised.

2.2 DEFINITIONS OF CLOUD COMPUTING

Cloud computing is a new evolving concept, which has no general definition but several individual and group researchers have authored it differently:

In “A Break in the Cloud: Toward a Cloud Definition”, a White Paper published for the ACM Computer Communication Reviews, the authors found over twenty distinct definitions of cloud computing.(Vaquero, Rodero-Merino, Caceres, & Lindner, 2009) [39].

Gartner: “Cloud computing is a style of computing where massively scalable IT-related capabilities are provided as a service across the Internet to multiple external customers” [16].

Rhoton (2010): “The cloud is IT as a Service, delivered by the IT resources that are independent of location” [33].

Forrester: “A pool of abstracted highly scalable, and managed infrastructure capable of hosting end-customer applications and billed by consumption” this definition is quoted in the book “Creating Business Agility: How Convergence of Cloud, Social, Mobile, Video and Big Data enables competitive Advantages” [17].

Cloud is basically a large group of virtualized resources which is easily usable and accessible (such as software, hardware, services and development platforms). This group of virtualized resources can be used by a pay-per-use model in which SLAs are offered by the Infrastructure Providers, and also guarantees best resource utilization.

Rhoton (2010) also said that cloud computing definitions include different elements of the infrastructure but do not address every single aspect that anyone has associated with cloud computing [33]. Consumers like Facebook and Google use identity management websites to maintain personal data and information in social network profiles, post pictures in flickr and use services of websites like LinkedIn to disclose professional associations. Similarly, cloud services use massive amount of computer server power, storage and network bandwidth at very low prices. Cloud provides the user a configurable control over part of the data center, the capability to control a server in the data center, consequently, to run the program that he/she has selected as said by Charles Babcock, in 2010 [1].

Barrie Sosinsky in his book “Cloud Computing Bible” (2011) defined cloud computing as, “Cloud computing is an abstraction based on the notion of pooling physical resources and presenting them as a virtual resource. Cloud computing can come in many different types, and the services and applications that run on cloud may or may not be delivered by a cloud service provider” [38].

Linthicum D. S. further defined [24] Cloud computing: “Cloud computing is the ability to provide IT resources over the internet. This includes storage services, database services, information services, testing services, security services.”

US National Institute of Standards and Technology (NIST) has issued a normative definition of the cloud computing in 2011, which says: “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” [32].

2.2.1 Characteristics of Cloud Computing

Five essential characteristics of cloud computing given by NIST and depicted in Figure 2.1 is as follows [32]:

A. On-demand self-service: Computing capabilities such as network storage and server time can be determined and selected automatically by the consumer as and when needed without human interaction with each of the service provider.

B. Broad network access: All services are accessible over the network and can be accessed through varied platforms such as laptops, PDAs and mobile phones.

C. Resource pooling: Multi-tenancy model is used for resource pooling and serving multiple customers simultaneously; a variety of virtual and physical resources are dynamically allocated and reallocated as per customer requirement. These resources are location independent and customer normally has no control or knowledge of the exact site of the provided resources.

D. Rapid elasticity: Services can be quickly and elastically provisioned. Scaling out and scaling in can be done automatically in some cases. The resources existing for provisioning often appear to be unconstrained to the consumer, and can be purchased in any quantity as per actual need at any time.

E. Measured Service: Cloud system possesses the capability to automatically monitor the resource use. It provides a metering functionality, at some level of abstraction. Transparency of utilized services can be maintained by monitoring and reporting resource usage.

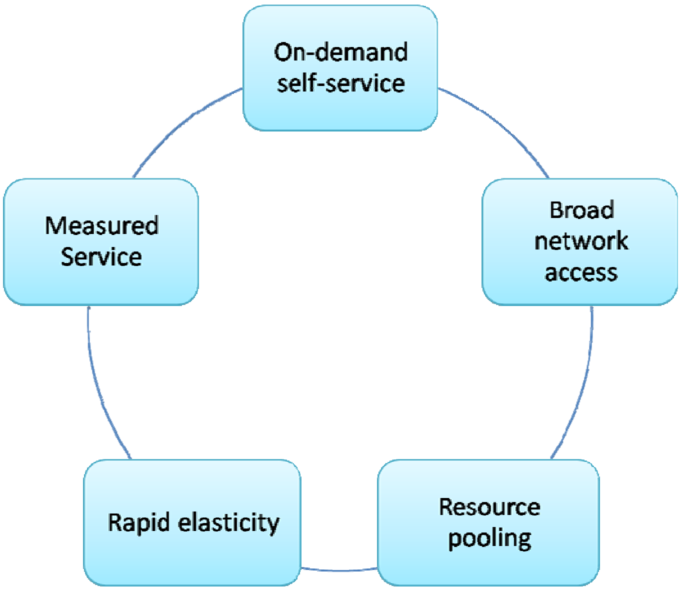

Figure 2.1: Characteristics of Cloud Computing

2.2.2 Deployment Model for Cloud Computing

NIST proposed four models of cloud services deployment, which are [32]:

A. Private cloud: In this model of cloud, infrastructure is private for an organization. It may be on premise or off premise, may be managed by the organization itself or by third party.

B. Community cloud: The cloud infrastructure is shared for several organizations and shared by them. It is for a specific group of people that has common motives and concerns such as operations, security concerns, and strategy. It may be managed by the organizations or by a third party and may exist on premise or off premise.

C. Public cloud: This type of cloud infrastructure is owned by an enterprise, which is selling cloud services and is provided to general public.

D. Hybrid cloud: This type of cloud infrastructure provides data and application portability by bundling together two or more cloud models by technology specifications.

The Figure 2.2 below tells us about the different location that cloud can come in from.

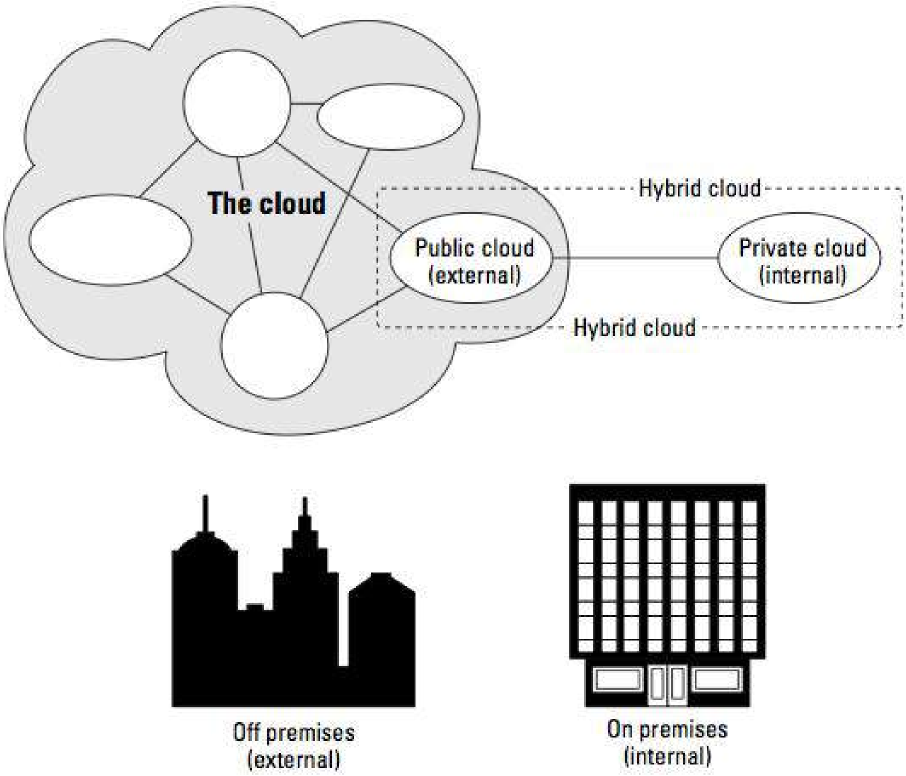

Figure 2.2: Deployment Models of cloud by Sosinsky

Cloud computing does not only offer packaged solutions but also focuses strongly towards services orientation.

2.2.3 Cloud Service Models

NIST presented in the same document defining cloud computing, three service models [32]:

A. Cloud Software as a Service (SaaS): SaaS provides capabilities to the customer to use the application software of the third party provider deployed on a cloud infrastructure. The software applications can be accessed from a wide range of client devices through a web browser or any other thin client interface (e.g., web-based email). The user does not bother or control the basic cloud infrastructure including servers, network, operating systems, and storage. Even customer is not concerned about the individual application capabilities, and user-specific application configuration settings.

B. Cloud Platform as a Service (PaaS): In this model, consumer is provided with the capabilities to set up and deploy applications created by him onto the cloud infrastructure or can deploy acquired applications supported by the provider of cloud services. The user does not bother to manage or control the basic cloud infrastructure which includes servers, operating systems, network, or storage, but has full control over the deployed applications and hosting environment configurations.

C. Cloud Infrastructure as a Service (IaaS): This service model provisions the hardware requirements of the consumer such as networks, storage, CPU and other basic computing resources where the user is able to deploy and run software as per their requirement, which can include operating systems and application software. The consumer does not manage or control the fundamental cloud infrastructure but has control over operating systems; deployed applications, storage, and some degree of control on selecting networking components (e.g., host firewalls).

A classical example of Software as a Service is Salesforce.com, Google App Engine and Microsoft Windows Azure are the providers of Platform as a Service, and one of the best examples of Infrastructure as a Service is Amazon Elastic Compute Cloud (EC2).

Although, this classification has been most commonly used and universally accepted in the related literature. But, some other service models for specific components or categories depending on their understanding or portfolio are also quoted in the literature, such as: XaaS, (Everything as a Service). 'Business Process as a Service' is one such example which is a layer on-top of SaaS where intricate business processes are provided to consumers that may cover multiple applications as stated by Breiter, Spatzier, & Behrend in 2011 [4].

These three service model are classified as "SPI Model" (Software as a service, Platform as a service, and Infrastructure as a service) by Rhoton [33]. In Figure 2.3 below Rhoton showed that cloud services vary according to their flexibility and degree of optimization and he called it SPI Tradeoffs.

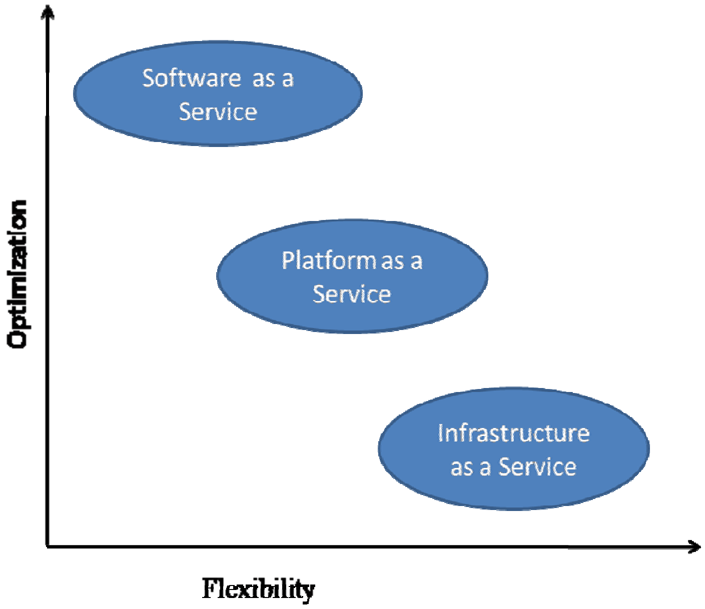

Figure 2.3: Tradeoffs between three services given by Rhoton

2.3 FOUNDATION OF CLOUD COMPUTING

In order to understand cloud computing it is essential to understand the key concept that forms its foundation and are important to its development and varied implementation.

2.3.1 Autonomic Computing

The term ‘Autonomic Computing’ was initially introduced by Paul Horn, Senior Vice President at IBM in 2001 at the National Academy of Engineers (NAE) at Harvard University (Chess & Kephart, 2003) [8]. He describes autonomic computing as the building and designing of “computing systems capable of running themselves, adjusting to varying circumstances, and preparing their resources to handle most efficiently the workloads we put upon them. These autonomic systems must anticipate the needs and allow users to concentrate on what they want to accomplish rather than figuring how to rig premise the computing systems to get them there” (IBM, 2001) [18].

IBM defined, characteristics of such a high-level system by following eight key elements:

A. Comprehensive and specific knowledge about all its components must be maintained.

B. Able to self-configure to suit changeable and possibly unpredictable conditions.

C. Constantly monitor itself for optimal functioning.

D. Self-healing ability to find alternate ways to function when problems encountered.

E. Capability to detect threats and guard itself from them.

F. Adaptable to changing environmental conditions.

G. Adhere to open standards rather than proprietary technologies.

H. It must anticipate demand while remaining transparent to the user.

Hence autonomic computing has put a great impact in the development of cloud computing. One of the essential features of cloud computing is “optimizing costs of operating and maintaining infrastructure and business functionality”, which could not be implemented without a self regulating system as in autonomic computing.

2.3.2 Utility Computing

This is not a new concept, but a computing resource to customer who can pay for these resources. Computers providing public utility to the customers is something like the telephone public utility open to public who can pay for the metered services when need it. The objective of utility computing is merely to use services effectively while reducing associated cost. Thus like software, hardware and network bandwidth is acquired as a product, utility computing is also a computing service at an initial low cost.

Service Level Agreement (SLA) between the consumer and service provider is one of the major problems as stated by Sokolov in 2009 [37].Quality of services and measure of guaranteed uptime are most important and crucial issues rather than SLAs. Uptime requirements are generally 99.99% but cloud providers have not yet been ready to meet the desired level of quality expectations (McKinsey & Company, 2009) [26].

2.3.3 Grid Computing

Myerson said in 2009 that Cloud computing has its origin at grid computing, which provides on-demand resource provisioning [28]. Grid computing use and interconnect many computers for problem solving through highly parallel computation.

“Grid computing may be confined to solving scientific problems which require a large number of computer processing cycles and needs an access to a large amount of data (Rhoton, 2010,) [33].

In their paper “Grid vs. Cloud – A Technology Comparison” (2011), Brandic and Dustdar defined Grid computing as a source of concepts and tools for the provision of High Performance Computing (HPC) resources, and applications as services that may be added on-demand and transparently. The primary aim of grid computing is to provide access to HPC infrastructure on demand by implementing standardized protocols and services. One prime application area of the Grid is transparent use of interconnected computational resources for scientific and large-scale application (Brandic & Dustdar, 2011) [3].

Broad differences between Grids and Clouds were summarized by Brandic and Dustdar (2011) as:

- Business Models: Business models in Grid are usually based on bilateral opinion between academic institutions while, clouds requires more differentiated business prevision of services. The different business models may vary from resource providers to businesses that attempt to run a mixed approach, that is they allow users to create their own services and also offer their own services(example: Microsoft Azzure, SUN NI Grid).
- Resource Management: Another major difference between Grid and Cloud is in method of Resource Management, since Grid is based on batch systems, and cloud computing uses virtualization.
- Resource Provision Models: In Cloud computing based on use SLAs, compliance and trust management while, Grid resource provision models are based on virtual organization where associations are established offline.
- Resource Availability: The resource sharing relies in the best effort manner in Grid, resources may not be available and sometimes there are plenty. Cloud relies on massive elasticity. Cloud has to balance between wasting resources due to virtualization overhead, and standby modes of devices on one hand, and pooling of resources on the other. This facilitates efficient consumption of resources and reduction of energy consumption.

2.3.4 Service-Oriented Architecture (SOA)

SOA disaggregates the information technology background into unassociated coupled functional primitives called services. These services employ actions and may be used by variety of different business applications (Rhoton, 2010) [33].

SOA enhances the agility aspects to architecture, which allows us to manage with system changes using a configuration layer rather than continuously having to develop these systems (Linthicum, 2009) [24].

Cloud also provides the IT resources and computing power on demand like hosting of data, process and services which is similar to what is suggested in SOA.

2.4 CLOUD COMPUTING AND TECHNICAL EDUCATION

2.4.1 Legacy Systems in Technical Education

By the early 2000s, application integration was among the top IT issues in the organizations; enterprise suit model was the only option for integration. If the integration indeed exists - is a costly, time-consuming endeavor. These applications are often sliced in their own business units at the institution, making it difficult to share information, report on a single source of truth and minimize the errors associated with duplicated data entry. On top of all this, each application has its own unique user interface and login requirements.

The integrations for these disparate systems are highly complex and require significant IT resources to maintain and update. Upgrades need to be carefully synchronized to ensure a seamless flow of information.

While the best-of-breed model slowly made its way into the technical education space, the leading ERP vendors were simultaneously trying to extend their products to fill those previously-mentioned gaps. Many organizations that adopted the best-of-breed approach found it challenging to incorporate new requirements, business processes or initiatives into their web of applications. Instead, they opted to deploy a fully integrated, single source suite of applications for core functionality, with a few ancillary best-of-breed solutions for specialized needs.

It is possible to improve efficiency and effectiveness of educational services within technical education with the advent of advanced communication technologies (ICT's).

Irani (2003) [19], said that organizations can be benefitted by Enterprise Application Integration (EAI), which includes assisting in business process integration, e-service based transformation, supporting mutual decision-making, abridged integration cost and delivering flexible, and maintainable integrated Information Technology (IT) infrastructures. (Samtani and Sadhwani, 2004) [35], also defined EIA as a process of building an integrated infrastructure linking disparate systems, applications, and data sources within a corporate enterprise. EAI was viewed by (Linthicum, 2004) [24] as the unrestricted sharing of information between two or more enterprise applications and business processes. EAI is a solution to intra/inter-organizational system and processes integration (Lam, 2005) [23].

As mentioned by Lam W, Venky Shankararaman (2007) [22], enterprise integration can solve the following business problems:

- Aggregation of Information: Organizing, aggregating, and presenting information from various IT sources in single view.
- Single point of data entry: Single interface for data entry in place of multiple manual data entry forms.
- Process inefficiency: Elimination of manual inefficiencies and reduction in the time and effort required to complete business processes.
- Web channel integration: Enabling direct web based access for the existing business systems to the customers and partners.
- Supplier integration and supply chain optimization.

While integrating a business process of large companies, innovative system must be provided and coupled with the existing system to provide required information and services. This could help glean business integration during the integration of information systems.

According to the 2013 Survey of Chief Information Officers given by the Leadership Board for CIOs (LBCIO), 83 percent of the respondents use “ERP vendor-supplied solutions today for their core administrative applications,” defined as financials, student systems, human resources and advancement. Only 13 percent use best-of-breed, homegrown, open-source or a hybrid combination. It appears that ERP has held its ground.

Truly effective ERP systems consist of a broad set of functional capabilities, all tightly unified for easier data sharing, reporting, navigation, maintenance and upgrades. Rather than an amalgam of acquired and afterthought functionality, the ideal system has fewer integration points to ensure consistency, flexibility, usability and ease of maintenance.

2.4.2 Comparison of Cloud and Own Infrastructure

Table 2.1: Comparison of cloud and own infrastructure

S.No.	Factors	Owned Infrastructure	Cloud
1	Implementation	Higher cost, Higher risk	Lower cost, faster to deploy
2	Software	Huge capital cost licensing model	Only operational cost Use based subscription model
3	Updates/Upgrades	Additional cost and risk	Part of subscription costs
4	Maintenance	An additional cost of 15-25 % license fees per year	No additional cost
5	Infrastructure:	Infrastructure cost and risk are at customer’s end	Infrastructure outsourced to the provider; No one time setup cost only subscription fee is applicable
6	Security/Contingency	Data Privacy/security	Data Privacy/security
7	Backup and Recovery	Backup and data recovery managed by customer at his cost and risk	Managed by cloud service provider as an integral part of services and only at built in subscription costs
8	Training	Extensive training and higher cost	Limited training low cost

Thus educational institutions which deploy cloud technology are at an advantage to use latest innovation, fast implementations and immediate updates. The added benefits of cloud are package integration, low operating cost, better service level and comprehensive security.

2.4.3 Factors affecting use of Cloud Computing in Technical Education

Cloud computing addresses four main issues and concerns in technical education

A. Integration and Security

The largest challenge for IT department in technical educational institutions is the optimum and successful integration of services. Integration challenges increase with the availability in the new services in cloud computing. Another key challenge is the security issue, such as the security of the facility where data is stored, security of data transport and the reliability of the provider.

B. Risk and Compliance Issues

Cloud computing raises a number of questions about privacy protection, data security and privacy issues, with regard to shared storage. For example, in a shared service environment an institution cannot control where its information is stored and how or by whom it is accessed. This is a big risk factor how does one manage those risks? How to protect against risks to data security, integrity, and availability, vendor lock-in, and security holes?

C. Governance Questions

Universities should consider cloud computing as an emerging technology and phenomenon that will require sensible care from decision makers and stakeholders on the campus, and must be convinced and embrace cloud computing as an important viable option for running the institute. Certain business functions could be sourced from cloud computing and would require a more complex decision making approach.

D. IT Staffing Implications

The IT personnel can be directly impacted by changes and use of cloud computing. IT staffing level should also discern the benefits of the cloud computing for the skills and experience required by IT employees for going forward. IT department of technical educational institutions required a different set of skill sets and training and have greater understanding of outsourcing and contractual issues and practices. Although there are too many issues to adopt technology but it's a reliable and popular technology (Nilam, Baldev Singh - 2014) [29].

2.5 TECHNOLOGY DIFFUSION

2.5.1 Implementation Gap and Gartner's Hype Cycle

Some people say that cloud computing is a temporary hype that will disappear after some time, since it is rather difficult to predict future, so before moving a business to the cloud, it would be wise to investigate about this new technology, its facts and the actual investment. It would be good to explore its commercial viability and see if cloud computing will pay off.

Hype cycle is a term used to include the tendency of a new technology to gain an exaggerated interest, before they are actually deployed and used in commercial environment. Gartner was the first to describe this hype cycle in information technology research and their application. Hype cycle methodology represents a graphical representation of technology application and an insight to manage its deployment in solving real business problems, thus developing new opportunities.

Gartner Hype Cycle consists of following five phases as explained by graph of Figure 2.4 [14]:

- **Technology Trigger:** When a potential technology gets its breakthrough by the media triggers publicity and concept stories. Sometimes products may not exist and commercial viability is difficult to prove.
- **Peak of Inflated Expectations:** Many success stories and also failure statistics are being produced by early publicity. Some companies may proceed accordingly and some may do not.
- **Trough of Disillusionment**: If the technology implementation fails to match the expectations, the interest declines and as a result, disillusionment is clearly to be seen as the through shows. The technology producers may fail or if they improvise their product or technology, only than people would invest in their product.
- **Slope of Enlightenmen**t: The slope shows how the technology can be useful for the enterprise and it takes shapes when the second and third generation

improvised products came out from the technology product. Yet, some unadventurous companies remain watchful.

- **Plateau of Productivity:** This phenomenon occurs when the technology products viability is clearly defined and main stream adoption starts. Market accepts the relevance of technology and its applicability starts paying off.

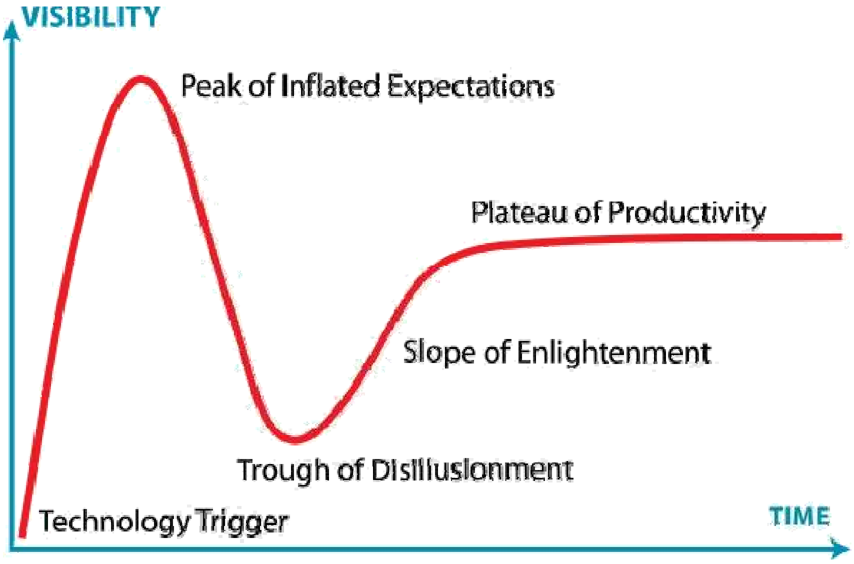

Figure 2.4: Hype Cycle given by Gartner

Gartner had Evaluated Maturity of 1,800 Technologies in his report of 2010[14].Cloud computing was placed at the end of the Peak of Inflated Expectations but it was predicted that cloud computing will become main stream in five years. (Gartner, Inc, 2010) [14].

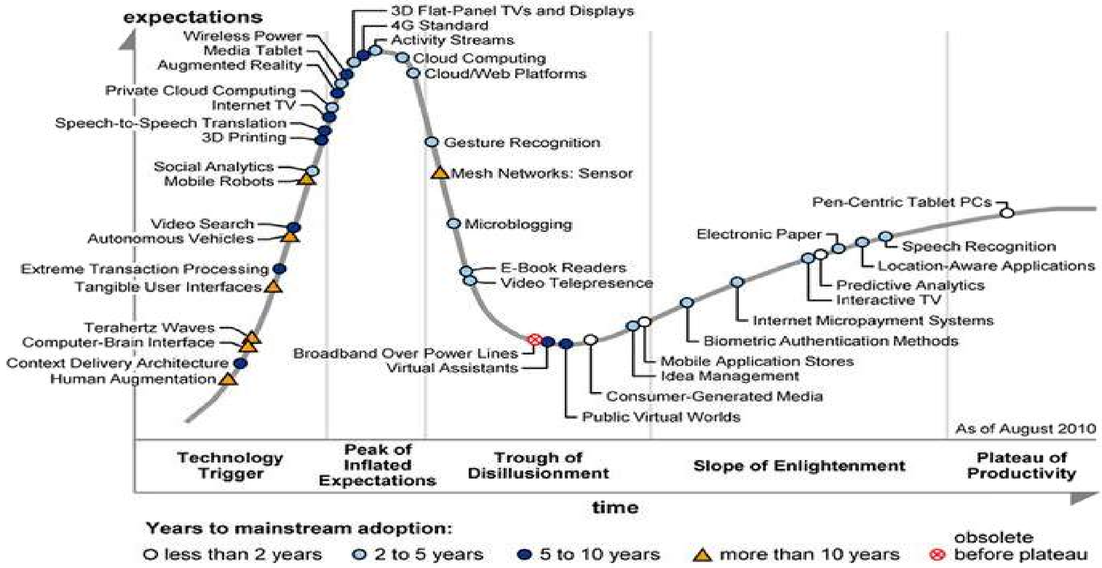

Figure 2.5: Hype Cycle of Emerging Technologies given by Gartner

Figure 2.5 helps us to understand the hype around cloud computing, and at the same time we can examine the historical behavior cycle of new technologies.

Geoffrey Moore slightly modified the “Technology adoption lifecycle” theory in his book, ‘Crossing the Chasm’ (2002) [27]. Geoffrey Moors technology adoption curve theory bifurcate the target market according to the speed with which they switch over to new technologies, thus giving rise to five categories of these segments shown in the figure below:

- Innovators: Innovators are pioneers of new technology they offer methods, solutions to address social and economic problem in a different way. They are youngsters who are willing to take risks, and are very clear about their financial goals.
- Early Adopters: These are the individuals having advanced education who are ready to adopt the innovations just after the innovators.
- Early Majority: Third category of individuals which takes significantly more time to adopt the technology than the innovators. Early majority fall in the group of early adopters, though they are slow in the adoption process.
- Late Majority: In this category individuals in this category are uncertain in their approach of technology evaluation. They will adopt the innovation rather cautiously and wait for the average member to adopt the technology.
- Laggards: This market segments is very necessary, and are last to adopt the innovation.

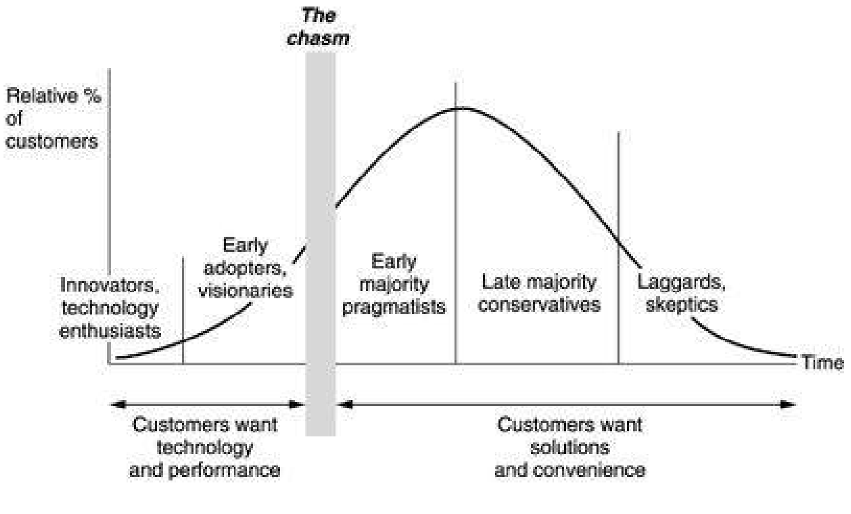
The chasm
Relative % of customers
Innovators, technology enthusiasts
Early adopters, visionaries
Early majority pragmatists
Late majority conservatives
Laggards, skeptics
Time
Customers want technology and performance
Customers want solutions and convenience

Figure 2.6: Technology Adoption Curve given by Moore

Some companies adopt the new technologies and gain competitive advantage by showcasing their adoption. Others are wary of new technologies and may not risk their business, outside their core competency (Rhoton, 2010) [33].

Moore in 2002 said that there is a gap between the early adopter and early majority categories for discontinuous or disruptive innovations, (See Figure 2.6 Technology Adoption Curve given by Moore) [27].

This transition becomes difficult between early adopters and early majority, as is the gap between visionaries and pragmatists. The technologies which can bridge this gap will create a 'bandwagon effect' and the product becomes ubiquitous in that momentum others will die in the beginning.

However, Rhoton (2010) has observed that organizations which are technology savvy adopt new developments faster than those organizations that are wary of new technologies and developments and remain aloof and distant form them [33].

In cloud computing the outsourcing and subcontracting moves large portions of technology systems outside the control and responsibility of the company. This is only good for those companies or organizations which are pleased to see their fixed IT cost reduced, and they preferred to focus on non-technical aspects of their business (BV Pranay Kumar, 2013 and Rhoton, 2010) [7,33].