AVASANT

Unraveling the Cloud: Architecture, Adoption & Performance Optimization

May 2013

The cloud has pretty much disrupted the de facto standard: the client/server and enterprise computing models, which displaced the mainframes .As a concept, cloud continues to be defined and redefined. The primary benefit of somewhat reduced cost and scalability is fast giving way to increased flexibility, creating new business offerings and establishing deeper customer relationships. Cloud vendors maintain a lower charge out rate for their services in an effort to deliver higher economic value to their customers. This downward spiral in prices of cloud services and continuous assessment to identify new usage, to further boost utilization, makes for a very compelling business case. The final goal would be to enable any organization to create and offer cloud computing services on standard hardware. This paper elaborates on the dos and don'ts of harnessing the cloud and to optimize the business metrics most important to their organizations.

Cloud Computing Defined

There are several ways cloud computing is defined , but as per National Institute of Standards & Technology (NIST), USA , "Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources. These resources: networks, servers, storage, applications and services, can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models."

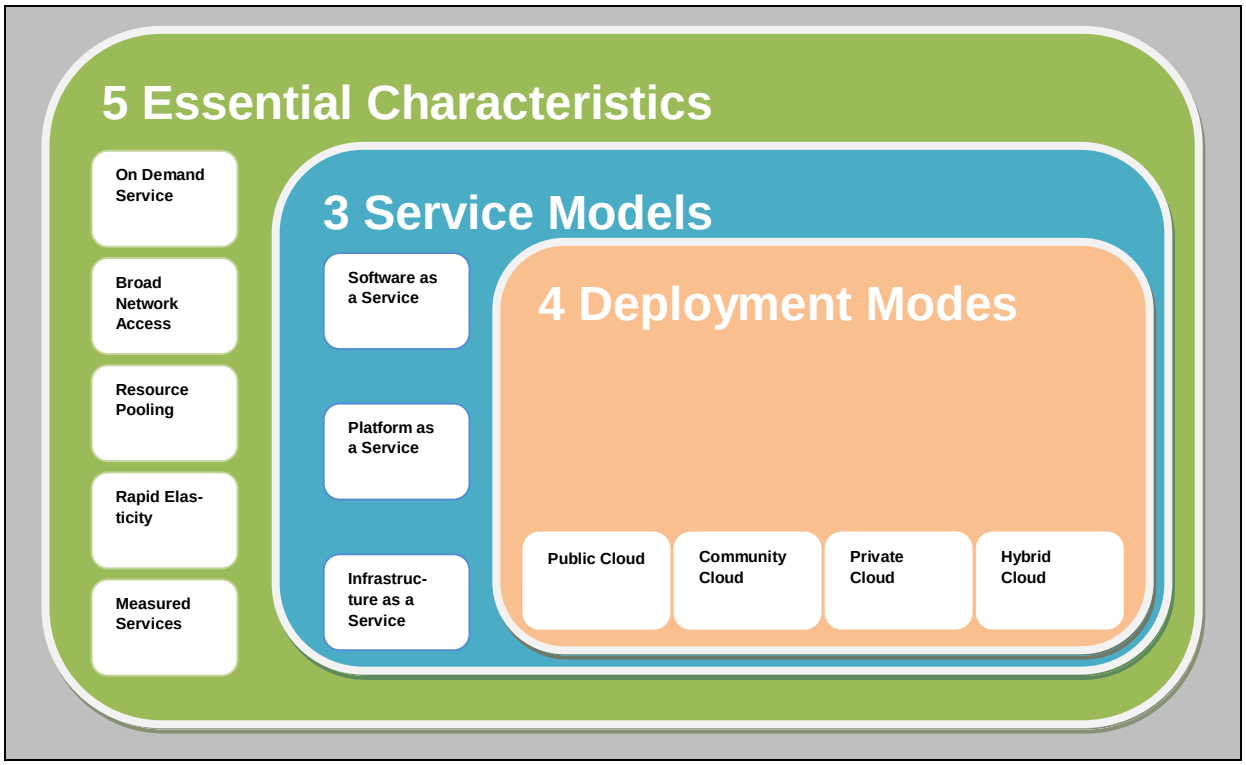


Figure 1 NIST Definition of Cloud Computing

Essential characteristics of cloud computing as defined by NIST are

1. **On-demand self-service**
2. **Broad network access**:
3. **Resource pooling**
4. **Rapid elasticity**
5. **Measured service**

Three service modes in cloud computing are Software as a Service (SaaS) **which allows provider's deployed applications to be used,** Platform as a Service (PaaS) **where providers gives languages,tools and libraries** to develop and deploy new application**,** Infrastructure as a Service (IaaS) where providers let consumer use the computing resources like processing ,storage ,network to deploy its IT environment on the cloud.

IaaS gives consumers control over operating systems, storage, deployed applications, and certain select networking components like firewalls and load balancers. BPaaS provides processes for employee benefit management, business travel, procurement or IT-centric processes such as software development and testing. SaaS is essentially about providing content services (e.g. video-on-demand) and higher value network services (e.g.VoIP), besides providing applications as a service (e.g. Sales force automation). PaaS involves deployment of consumer created applications using programming languages and tools supported by the provider (e.g. Java, Python, and .Net).

Deployment models

Private cloud: The cloud infrastructure is for exclusive use by a single organization which may comprise of multiple consumers (business units). The infrastructure may be owned, managed, and operated by the organization, a third party, or both and it may exist on or off-premises.

Community cloud: The cloud infrastructure is for a specific group of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations The infrastructure may be jointly owned, managed, and operated by one or more organizations within the group, a third party, or both and it may exist on or off-premises.

Public cloud: The cloud infrastructure is provisioned for use by public. Businesses, academic institutions, or government organizations own, manage and operate the cloud infrastructure existing on the premises of the cloud provider.

Hybrid cloud The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

Cloud Reference Architecture

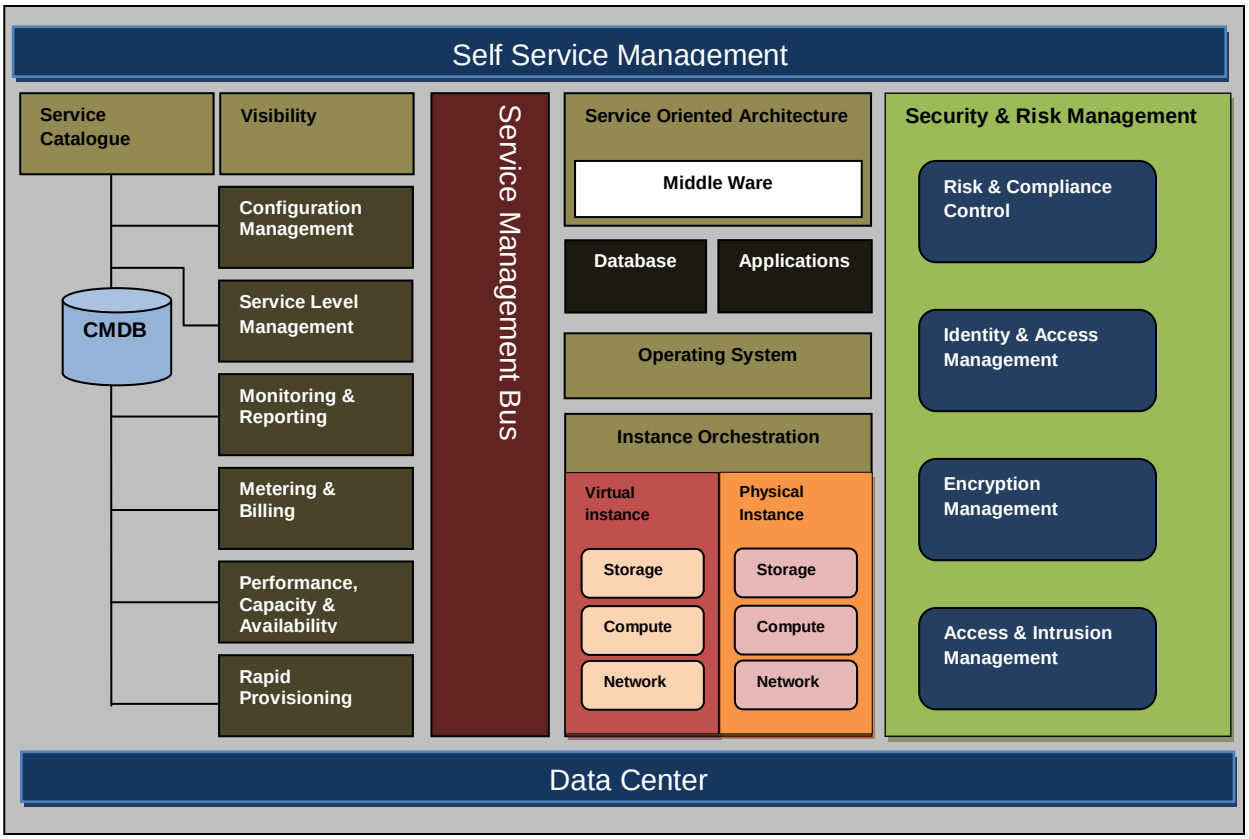


Figure 2 HCL Cloud Reference Architecture Model

Technology giant HCL gave the cloud reference architecture model, which is based on the Service Oriented Architecture (SOA). While they say, the concept may be following rulebook, the implementation could be flexible based on client requirements.

For all cloud services there is software required to implement cloud service specifics: for IaaS, there are typically hypervisors installed on the managed hardware infrastructure, for PaaS there would be a multi-tenancy enabled middleware platform, for SaaS a (multi-tenancy enabled) end-user application and for BPaaS a multi-tenancy enabled business process engine.

The common cloud management platform (CCMP), composed of Base Support Services (BSS) & Operational Support Services (OSS), is defined as a general platform to support the management of any category of cloud services across I/P/S/BPaaS.

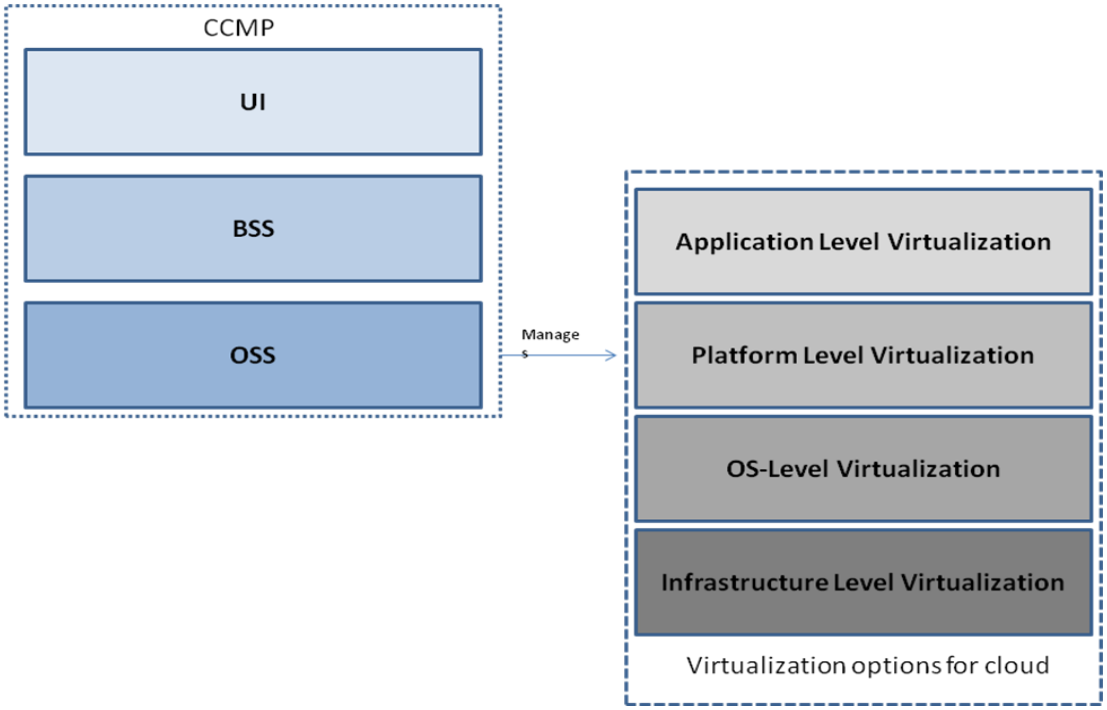


Figure 3 CCMP supports any level of virtualization

As an example, the billing service of the CCMP BSS must be usable to conduct billing for the consumption of virtual machines (IaaS). A multi-tenancy capable middleware platform is required to determine level of collaboration services like Lotus Live (SaaS). The OSS represents the set of operational management/technical-related services exposed by the CCMP. Many management domains in the OSS are encountered in traditionally managed data centers (e.g. monitoring & event management, provisioning, incident & problem management etc.) while other components are new and automated for use in the cloud based environment. (e.g., service automation, image lifecycle management). For example, in case of a failure of a physical server in a data center an incident gets recorded and a ticket is raised and assigned to an admin. If the admin fails to resolve, the incident is escalated. In contrast, for VMaaS cloud services, the virtual machines, which have been running on a bad physical machine, are automatically transitioned to another physical machine. This traditional approach guarantees a high quality of service but is rather costly. In the cloud scenario, the broken hardware/hypervisor would be disabled and when users start their virtual machine, again it is brought up on one of the remaining hypervisors. All faulty hardware is repaired on a periodic basis e.g., weekly, fortnightly vs. repairing it immediately, whenever it fails. Due to the automation, the cost of administration is also reduced. The above dramatically increases the efficiency by leveraging new technology approaches and building on high degrees of standardization and automation.

All CCMP functions can be used for the consistent management of cloud services, virtualized on any level - hypervisor, operating system, platform or application level virtualization. The ideal case from a cost optimization and economies-of-scale perspective is to use as much as shared CCMP OSS/BSS functionality across multiple cloud services. Minimal variance on the infrastructure side is critical for enabling the high degrees of automation and economies of scale, which are base characteristics of any cloud. Total homogeneity is the ideal case but there will be cloud installations with a few variants. The above architectural principles serve as a guideline for creating or comprehending any cloud environment**.**

Which way is the Cloud drifting?

Trends in cloud adoption

A recent survey conducted by CRN of around 200 executives identified that an overwhelming 85% were already using cloud computing services and another 13% planned on getting on to the platform soon. It lays to rest any questions about the acceptability and adoption of cloud computing by enterprises. Out of those who are already using the service, 90% of the installations are less than 5 years old. This is understandable, as Cloud is a new concept and organizations have been cautious in its uptake. Mail services, office productivity tools and collaboration tools have been one of the most common applications that have moved to the Cloud as a SaaS. Other software services, which are being used by organizations, including small & medium enterprises, include- Time Sheet & Expense Management, Travel Planning & Booking, Expense Management, Sales Force Automation etc.

Cloud services in demand

Majority of the businesses (nearly 55%) are using cloud to meet the software, storage and infrastructure requirements. The demand for PaaS is muted compared to other services. While the demand for Infrastructure, Software and Storage are on par, it is expected that rate of growth in demand for storage in cloud will surpass the growth rate for other services.

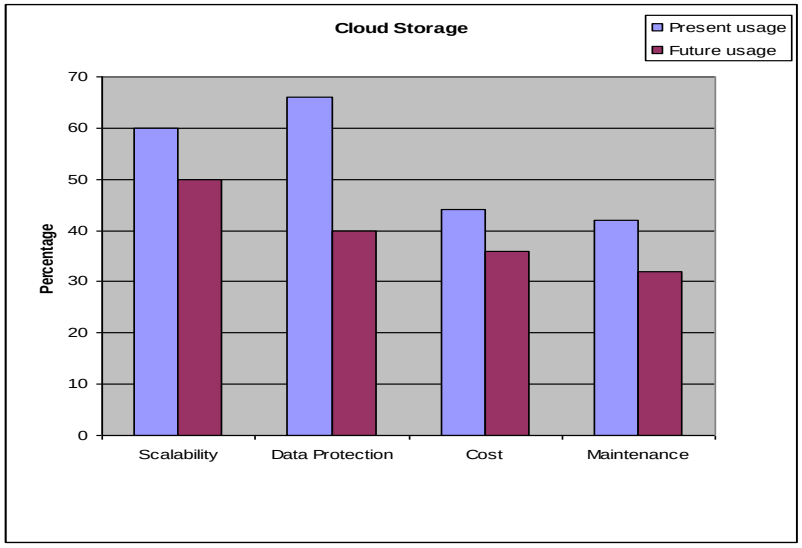


Figure 4 Reason for using cloud storage

Scalability is one of the main drivers for the anticipated growth in the demand for cloud storage. Customer views it as providing the maximum value addition. The survey revealed that current users of Cloud are very satisfied with data security and disaster recovery in the cloud, though users are more apprehensive of these features. Data recovery and easy archiving was positive experience for those who have deployed cloud. The biggest drawback to adopting cloud storage was loss of data control.

Customers are shifting towards deploying more traditional business applications to cloud, instead of just sticking to non-critical applications as per Forrester Research. As acceptability of cloud increases, developers are bringing more data that are sensitive into the cloud, like customer information. Core intellectual property data and business transactions have also started moving to the cloud.

Keeping an eye on optimal service levels

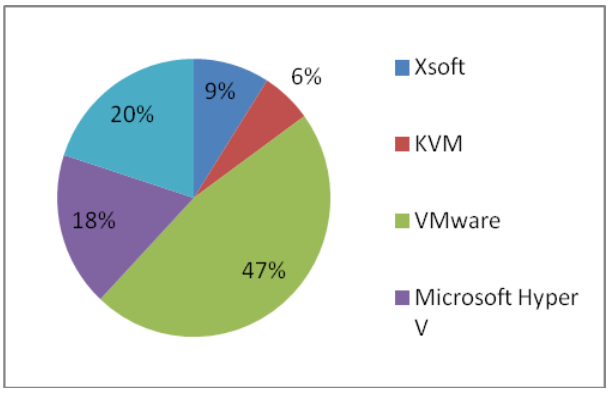
A very pertinent case of cloud outages can be demonstrated by taking the example of Amazon Web Services (AWS). The largest public cloud provider's parent arm is its biggest customer. AWS has experienced many outages but its architecture is such that usually Amazon.com, the retail site, is not impacted. Amazons latest earnings report showed that it makes about $10.8 billion per quarter, or about $118 million per day and $4.9 million per hour. Therefore, for every hour of downtime that Amazon experiences, it could stand to lose close to $5 million. Downtime occurs in every situation whether the servers are with the cloud service providers or in a local data center. It is strongly advised to correctly estimate the cost of outages as well as the proper amount to spend on failover services and only then enter into service level agreements (SLA). This makes the provider of cloud service understand the revenue its client company loses at times of outages.

Customers with stringent security needs can opt for additional premium services such as cryptographic key management, intrusion detection services, application firewall, advanced logging and reporting typically offered through the activation of software. The above overheads, if deemed as a necessity for data protection, can change the comparative economics between cloud based and on-premises IT solutions. Time slicing by service providers using multi tenancy will keep the cost low as compared to having the same features in an on-premises environment but then again in the context of security buyers need to decide for themselves whether sharing the capability among multiple customers is adequate for their needs.

Public, Private, Community, Hybrid or Openstack?

Amazon is a public cloud architecture accounting for 90% of public cloud capacity. VMware is a private one with 90% of enterprise virtualization clouds being on it while Openstack is an industry consortium and the primary open source alternative to the above two. However, the war for Cloud Infrastructure is heating up. The move is towards having open source platforms.

Demand for Virtualization Platform



Demand for Cloud Platform

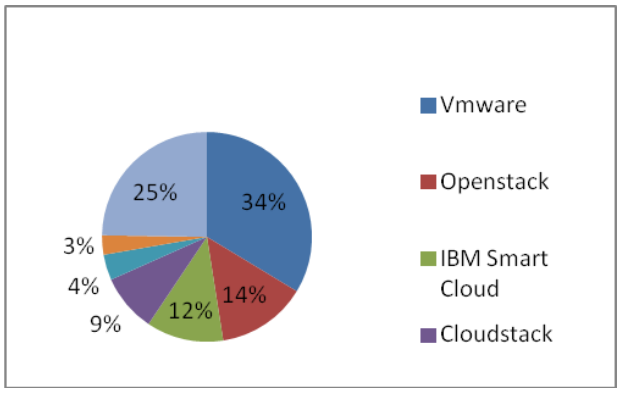


Figure 5 Cloud Infrastructure Usages

Source: Everest Group

The easiest way of using cloud computing is to obtain free of charge infrastructure resources from the public cloud and it is the one which offers maximum economies of scale. Businesses and software development teams at large enterprises and service provider organizations are going around their own IT operations to access faster public cloud computing infrastructure, creating a so-called “shadow IT” group. Interest and traction are centered less on legacy infrastructure and application management and more on leveraging public and increasingly private cloud computing to improve time to market, efficiency and eventually total cost of operation.

As public cloud has to address a large number of customers, the focus is on scalability. Public cloud hosts application that has mass demand and the infrastructure on which application runs are build out of industry standard building blocks. At the other end of the spectrum, private cloud is more technology intensive and could host specialized applications on advance technology infrastructure.

Due to the huge volume of users, public cloud focuses on standard contracts with standard billing models. Microsoft, Google, Amazon and Salesforce have high availability infrastructure resources, various application development environments (Azure, Google App Engine, and Force.com) and complete software solutions (Google Apps, Microsoft 365, Dynamics, Salesforce CRM). Public cloud provides the full range of cloud benefits. Long-term deployment of public cloud computing should be strategically evaluated to check if this would lead to a vendor lockdown.

Some organizations prefer private cloud, which provides businesses the benefit of more efficient hardware, automatic provisioning of resources, all with the security of the system being on the company’s premises with native security features in place. Primary motive behind using a private cloud varies from disaster recovery and resilience to making use of existing in house hardware, control over configuration and how it operates or simply the flexibility of choosing between rent and buy at any given time and then moving back and forth. Key business differentiator happens at the PaaS level in private clouds, where one builds new applications specifically written as per the need that can scale dynamically. The level of speed, flexibility and volume of booking resources in a private cloud cannot compete with that of a public cloud.

Hybrid clouds combine the low cost services from public clouds with data protection and reliable applications from private clouds. It enables joint data and application exchange using several cloud forms, which remain logically separated. A community cloud comprises of several organizations with similar interests sharing the infrastructure resources.

Openstack promotes open standards that aim to end customer reliance on singles cloud provider’s proprietary code. Their ultimate goal is to enable any organization to create and offer cloud-computing services that run on standard hardware. It is also viewed as the kernel for cloud operations, on which vendors can build all sorts of software to run on in the cloud. It is in sharp contrast to an organizations approach of laboring to create the workarounds that are required to use a closed cloud but it also dramatically increases reliance upon the community maintaining or improving the open source code. Large-scale production deployment of OpenStack is growing among large enterprise, service-provider and converged customers.

Cloud Adoption-Drivers & Risk

In an enterprise, there are two types of Cloud buyers namely Corporate IT and Business. The main buying considerations for Corporate IT is to ensure control, security, performance and scalability. They have preference for dedicated private cloud – hosted, managed or on-premises. On the other hand, the drivers for Business buyers are new capabilities, flexibility, ease of use and time to market. These buyers prefer shared public cloud model.

It is important that irrespective of the choice of various types of cloud implementation or varying drivers for its implementation, a robust risk management framework be in place to assess the business risks of the decision to transition to cloud. NIST's risk management framework is one of the effective tools to deploy for this purpose.

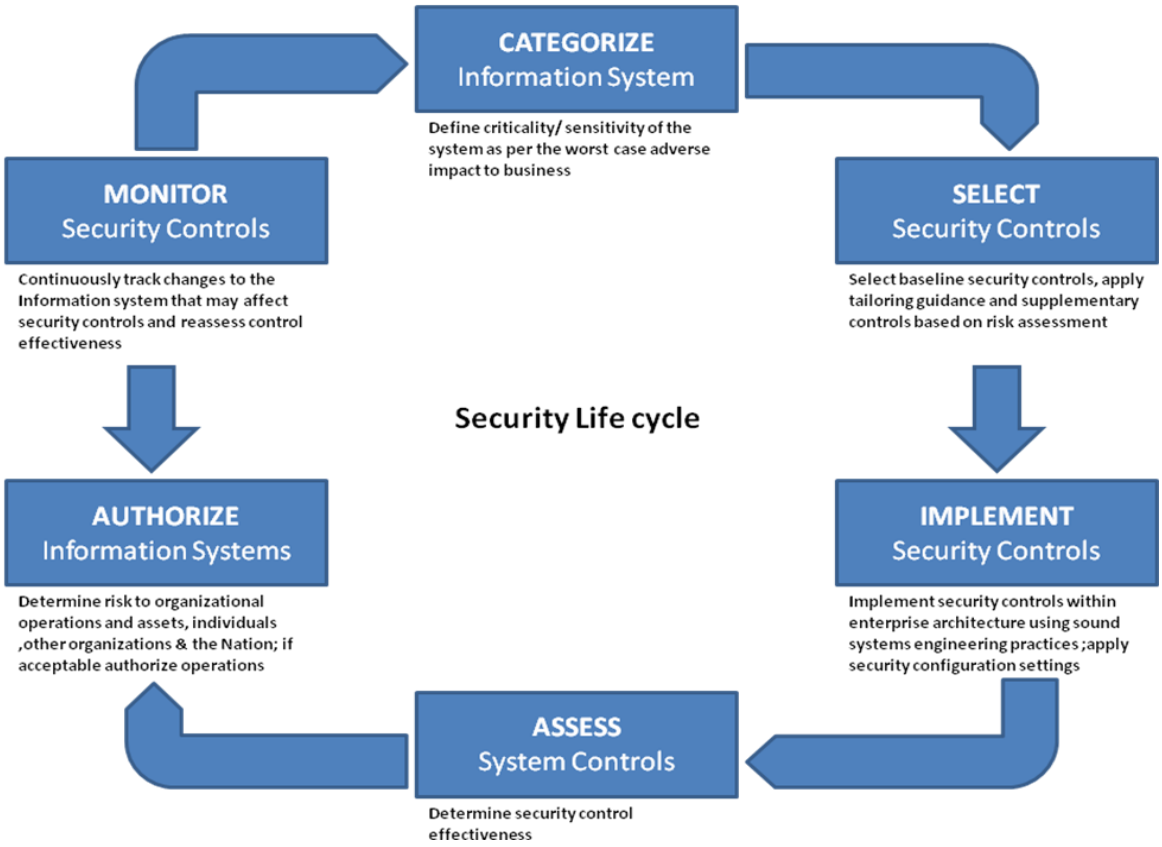


Figure 6. NIST Risk Management Framework

Performance Optimization

Cloud users should first establish working requirements in terms of the metrics important to their organization. Metrics like cost, customer satisfaction and speed of delivery, allow for quantitative evaluation of a strategy. The key question is how the cloud operation could be made more efficient to further enhance its impact on business. Cockcroft, the chief IT architect of Netflix, said "The switch to Amazon allowed Netflix to try using large data center resources, fail at it without paying a heavy penalty in unused gear because it was only rented by the hour, then try again." Such observations would generally hold true while they use any other cloud service provider and should serve as a useful yardstick of cost optimization. Some of the factors that need to be considered to ensure an optimized architecture and minimized cost areas are:

1. Study traffic patterns

It is important to study the traffic pattern, placement of content servers, distribution of virtual machines- to enhance performance and reducing latencies.

2. Utilize reserved instances to cut cost

Service providers offers a price advantage to customers who sign up for reserved instances by paying an upfront fee for guaranteed access to a defined amount of service for a year. It is advisable for users to determine which workloads tend to operate at a steady state and then identify what that state is. By identifying steady-state workloads, customers know how much of a reserved instance they should buy and can move workloads onto the lower-cost server type to gain significant savings. Most customers end up running the more expensive, on-demand mode. This is due to companies not keeping track of their steady state usage. It is imperative to provision for a little more that steady state usage as reserved instances so that occasional peaks do not trigger firing up of on-demand instances.

3. Disaster Planning

Cloud users procure huge number of reserved instances from their provider to cope with natural calamities that stop operations. These instances give them a great fall back plan in case of any calamities. One region serving as the fail-over for the other provides their customers a guarantee of continued operations in most eventualities. Customer may not use the reserved instances lot of times, but if disaster strikes, that capacity will be taken from someone else and given to them. That is possible because providers as AWS sells spot instances that go for substantially under-market prices until a pre-paid customer needs them.

4. Use idle reserve capacity for development and testing

Peaks in user demand or disaster recovery is a rarity. To further optimize capacity, customers put their reserve instances to good use by performing their development and testing on them. This significantly cut their cloud bill by incapacitating the need for a separate set of expensive on-demand account.

5. Consolidate accounts to gain discounts

It is advisable for users to consolidate their cloud bill under their company's name, instead of having many individuals and departments run their own accounts. Amazon discounts services for large users, but it does not offer discounts to multiple independent accounts. A customer becomes eligible for a low-priced billing tier if the accounts are consolidated.

Optimizing cloud services and keeping budgets under control also requires few management disciplines, namely:

- Deploying and running applications in an uncontrolled fashion.
- Practice of over provisioning in the data center being carried over to cloud. Typical data centers have CPU utilization ranging from 8% to 12%. For cloud services, one should aim for CPU utilization of 17% to 20%. Server instances requisitioned should not be underutilized.
- Server instances should also not be overused as that will lead to customers experiencing delays in getting response or some traffic will not be able to access the business cloud application, leading to business loss.

Overutilization by purchasing fewer number of instances, underutilization of memory and computing along with provisioning for excess capacity based on faulty demand forecasting are some of the pitfalls to avoid. Tagging resources and keeping users informed with regular updates, seeking explanations when usage goes up are some of the proactive ways of keeping costs down. Budgeting for weekend usages for development and testing should help in keeping the organization immune to overage charges.

In the public cloud space, there is intense competition between the likes of Google, Amazon and Microsoft. The price for various types of services like one-demand virtual machines and storage is very dynamic and continuously falling. Providers are continuously evolving its pricing structure by throwing in value added services. In such a scenario, for a discerning user, it is important to negotiate good deals. It is important to have an optimized and flexible structure, which can be periodically renegotiated to drive value through the organization. It is important to have good advisory support for negotiating a good cloud services contract for optimal results.

Concluding Remarks

- Cloud computing offers a plethora of new things, a wealth of options and opportunities of paramount importance to the user like cost reductions, increase in flexibility, top-line business growth and acquired freedom which enables innovation despite shrinking IT budgets.
- Cloud service providers need to focus more on delivering business value rather competing solely on pricing.
- Private cloud still remains the preferred choice but open standards are fast gaining traction.
- The types of applications deployed in public clouds are shifting from less critical to more traditional business applications.
- Developers are showing a strong preference for public clouds that give them access to far greater elasticity in terms of resource allocation.
- There is an increase in customer records and core business transactions happening over the public cloud.
- Security issues continue to be a big barrier but if cost and speed of application deployment are primary concerns, there is no better solution than the public cloud.
- Due to constant cloud evolution, optimum usage of resources needs to be relooked periodically. It calls for a concerted effort to keep management, development teams and service providers on the same page.
- Overall businesses are extremely positive about cloud computing and have very high expectations of it. It is foreseen that lack of confidence businesses currently have towards large public cloud providers in terms of data confidentiality and legal concerns giving way to a more symbiotic and mutually beneficial relationship.

About the Authors

Dr. Pradeep K. Mukherji is the President and Managing Partner at Avasant APAC and Africa based in Mumbai, India

Paritosh Sharma is a Consultant at Avasant and based in Mumbai, India

About Avasant

With its global headquarters in Los Angeles, California, Avasant is a top management consulting, research, and events firm servicing global clients across the public, private, and non-profit sectors. Our talented team of consultants, lawyers and technologists average over 20 years of industry-honed experience and have conducted 1,000+ engagements in over 40 countries worldwide. Avasant drives customer value through the use of our proprietary consulting and advisory methods, which have been refined over decades of 'real-world' transaction and engagement experience. The combination of our world-class resources allows Avasant to yield superior business outcomes in three primary domains: Strategic Sourcing, Technology Optimization and Globalization Consulting.

For more details about Avasant's services and capabilities, Please visit our website at: **www.avasant.com**

Visit us for a complete list of research papers on: **www.avasant.com/research**

AVASANT

3601 N Aviation Boulevard, Suite 3000, Manhattan Beach, CA, 90266, USA
Tel: 310-643-3030
Fax: 310-643-3033
Email: contactus@avasant.com

Global Headquarters	Los Angeles
North America	Austin | New York | Chicago | San Diego | San Francisco
Asia	Beijing | Bangalore |Mumbai | Tokyo
Europe	London
Middle East & Africa	Accra

For any further communications, please contact: marketing@avasant.com