Optimizing IT Services using Cloud, Data Center, and Adherence to Compliance Standards

EMC Proven Professional Knowledge Sharing 2011

Amrith Raj Radhakrishnan
IT Analyst
Tata Consultancy Services Limited
amrith.radhakrishnan@tcs.com
amrithraj.r@gmail.com

Table of Contents

Table of Figures

Disclaimer: The views, processes, or methodologies published in this article are those of the author. They do not necessarily reflect EMC Corporation's views, processes, or methodologies.

Preface

About this article

This article provides insight to the paradigm shift currently taking place in the information technology industry. Topics include cloud computing, virtualization, and ‘green’ data centers. We will discuss:

- options available for managing the information technology infrastructure
- problems caused by data growth and how to address them
- problems related to cloud computing, including security

Compliance has never been a topic well understood when we talk about technology. Often considered outside the scope of technology, we will see how ITIL, COBIT, and Val IT help proper governance of information technology companies. We will also see how government regulations such as SOX, PCI-DSS, HIPAA, and GLBA have influenced the IT industry.

The discussion in this article is not just limited to IT, but includes other areas which are directly or indirectly relevant to Information Technology. It provides the ways an IT solution has to be chosen and applied to a new environment as well as a solution where the need for transformation is required.

How is this article organized?

This article has four major sections:

1. Introduction: Focuses on understanding the backgroud and importance of information technolgy and its evolution, the Internet, data growth, and classification of businesses according to size; small, medium, and large enterprises.
2. Cloud Computing considerations: Focuses on cloud computing, its implementation and management, and who, when, and how a company should approach cloud computing, IT regulations, and security concerns, and apply best practices.
3. Building efficient data centers: Focuses on developing an efficient green data center, a detailed analysis of virtualization options in the industry, tiers in the data center, processes, steps, and best practices.
4. IT Compliance and Controls: Understanding compliance and regulations, various governance frameworks, concepts and practices, and considerations while applying ITIL, COBIT, and Val IT.

Executive Summary

Today, data drive the world. The success of business depends on keeping data secure and accessible so it can be put to work as business information. In the last few years, the IT industry has seen many changes with a number of new, emerging technologies. The most recent—cloud computing—is now a reality and evolutions in various hardware, software platforms, and the data center have been seen in the IT industry. Businesses focus on excelling in their core business rather than focusing on IT. This is only possible if IT is agile, robust, and reliable.

Businesses should not view IT as a matter of concern but a solution to its problems. Finding a solution to a major IT problem is a big challenge. But the challenge becomes bigger when we have multiple solutions with less knowledge of the solution that we need to apply. New technologies and options have emerged recently, contributing to even greater confusion. Multiple options make it difficult to select the right IT solution that will transform their business. Should a business build, manage, and support its own IT infrastructure? Should it host its hardware resources in a data center provider in the form of co-location? Should it use resources from a cloud service provider or even outsource the IT and the services to a third party firm? The confusion gets deeper as we apply each solution.

Each solution has its advantages and disadvantages. A good feature of a solution might be excellent for one type of business but might be inappropriate for another. How do we select one? With the features provided by cloud computing, every company will be keen to harness its power. It is a fact that data centers are one of the highest consumers of electricity. The need to reduce carbon emissions is also rising due to ‘green’ compliance. It is not just to comply with the policies and regulations but also to reduce the rising electricity bills of the data center.

So, how do we make a decision? How do we decide what is good for an enterprise to run the business effectively? How do we adopt an IT solution to solve business problems without causing a chain-reaction of complex issues after implementing a wrong one? Will adding more servers and storage solve the problem? How do we govern our IT infrastructure? How do we apply existing service management tools? How to transform an existing IT infrastructure which is currently suffering from problems to a long lasting, strong, robust setup? How to get insight into new technologies that can transform the IT for the business? All these questions and more will be answered in this article. A multi-dimensional view of the current IT problems and its solutions will also be provided.

In summary, the article will explain how IT can deliver more services using cloud computing, data center, server, I/O virtualization, converged infrastructure, enterprise storage, apply proper governance and management, and more. The article takes a business-oriented view with respect to IT and provides a detailed analysis explaining its applications across various types of enterprises. The objective is that IT should deliver the best to a business by leveraging the cloud, data center, and compliance.

1. Introduction

"We can't solve problems by using the same kind of thinking we used when creating them".

(Albert Einstein).

Information technology continually evolves. It has changed rapidly from mainframes, to open systems, to a client-server model, to today's cloud computing. The problems that businesses face today with data growth, competition, and compliance require that information technology be agile and reliable. Businesses depend on information technology for its sustainability. In business, making the right IT decision can mean the difference between success and failure.

Enterprises are looking forward to applying new technologies such as cloud computing, virtualization, converged infrastructure, and more to excel in their core business. Applying a solution has always been a daunting task. A proof-of-concept to demonstrate feasibility is still the best method of testing a solution. However, when applied in a real-time environment, the solution might not perform as expected. For business to excel, we should be able to get a real picture of these technologies and regulations which will help transform their businesses.

1.1 Evolving Information Technology

Traditionally, IT infrastructure has meant the data center of the organization. The data center—facilities that primarily contain electronic equipment used for data processing, data storage, and communications networking—was usually an integral part of the organization which does all the computing. Its efficiency was assessed by its effectiveness in processing and storing a vast amount of data. It wasn't complex to run a data center because the amount of data handled was less. Then, IT hadn't proven its capability to its fullest.

As our economy has shifted from paper-based to digital information management, data centers have become common and essential to the functioning of business, communications, academic, and governmental systems. Data centers are found in nearly every sector of the economy: financial services, media, high-tech, universities, government institutions. Many others use and operate data centers to aid business processes, information management, and communications functions.

In the mid-1990s, the Internet emerged as the largest interconnected computer network. The Internet turned out to be one of the most effective forms of communication, enabling companies to reach existing and potential customers. It was usually represented by a cloud, indicating 'all the other stuff'. Meanwhile, users were interested with only the website they surf, not how the website arrived at their Internet browser. Email service in the mid-1990s used static HTML

pages wherein the page refreshed each and every time a request operation was performed. Now, there are Web 2.0 services which can do multi-tasking within the web page without the need to refresh the page.

A familiar example is Facebook, which has quintupled in size to a network that touches more than 500 million users. Another is the telecommunications industry. More than 4 billion people around the world now use cell phones, and for 450 million of those people, the Web is a fully mobile experience.

There is also a change in the way IT is deployed, such as virtualization and cloud computing. This rapid change in technology raises serious questions for executives about how to help their companies capitalize on the transformation. For senior executives, merely understanding the upcoming trends isn't enough. They need to think strategically about how to adapt management and organizational structures to meet these new demands.

Also, data center growth–driven by escalating growth in data creation–continues to be one of the biggest challenges of the IT infrastructure. Although usually ignored, proper IT governance is also one of the focuses that will help IT deliver its best. We will see the importance of these topics and how it is affecting businesses in the following sections.

Six Steps of Delivery Model Evolution

1. **Centralized Computing** – IT services were limited to a small number of people within the organization. Large companies had to acquire the computer room, set up a private network, install the hardware, and bring the expertise to manage all these. The computer would usually be a mainframe which would have to be sliced to cater to the needs of specific projects.
2. **Time Sharing** – Share the IT services with other companies.
3. **Service Bureaus** – The time-sharing concept commercialized to become today's Service Providers.
4. **Outsourcing** – When the Internet expanded, companies dedicated a part or all of its IT operation to a third-party company specializing in IT, thus reducing costs.
5. **Application Service Providers** – When part of an operation didn't require to stay in-house to meet reliability requirements, companies started to outsource specific applications such as its email service or company website.
6. **Software as a Service (SaaS)** – The latest evolution.

1.2 Data center growth

Data growth is associated with increased cost related to hardware, software, associated maintenance, administration, and services. Managing this data growth is the biggest hardware infrastructure challenge the data center faces[1]. Data growth leads to 'data center growth' which increases the amount of hardware required. Increased hardware in the data center increases power requirements. More cooling infrastructure is required to keep the hardware functional, which in turn consume more power.

Some of the factors that have contributed to data growth are:

- Digitisation
- Social networking
- Email
- Research content
- User files
- Compliance and regulations
- Backup, data retention, and much more!

Consider the following facts about data growth:

- The New York Stock Exchange generates about one terabyte of new trade data per day.[2]
- Facebook hosts approximately 10 billion photos, taking up one petabyte of storage.[3]
- Ancestry.com, the genealogy site, stores around 2.5 petabytes of data.[4]
- The Internet Archive stores around 2 petabytes of data, and is growing at a rate of 20 terabytes per month.[5]
- The Large Hadron Collider near Geneva, Switzerland, will produce about 15 petabytes of data per year.[6]

[1] Gartner Press release:"Gartner Survey Shows Data Growth as the Largest Data Center Infrastructure Challenge" http://www.gartner.com/it/page.jsp?id=1460213
[2] http://intelligent-enterprise.informationweek.com/showArticle.jhtml;jsessionid=VYVZYLR0VKQUXQE1GHRSKHWATMY32JVN?articleID=207800705
[3] http://mashable.com/2008/10/15/facebook-10-billion-photos/
4 http://blog.familytreemagazine.com/insider/Inside+Ancestrycoms+TopSecret+Data+Center.aspx
[5] http://www.archive.org/about/faqs.php
[6] http://www.interactions.org/cms/?pid=1027032

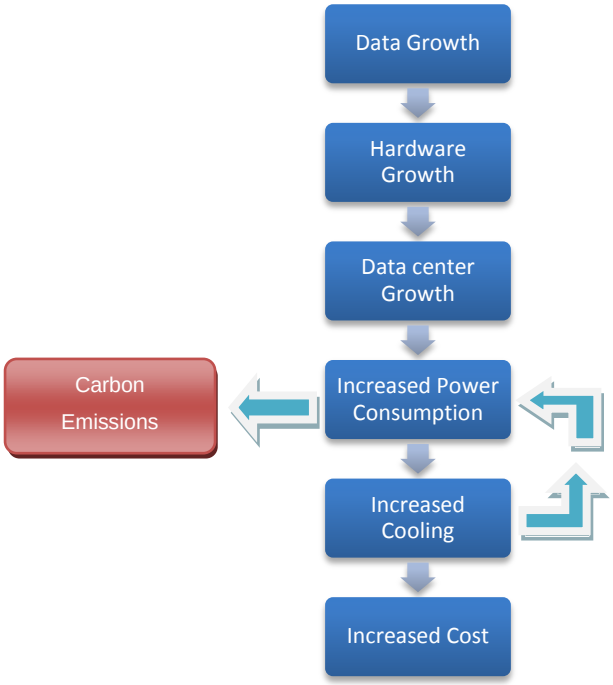


Figure 1 - Data Growth leading to Carbon Emissions and Increased Cost

The U.S. data center industry is in the midst of a major growth period stimulated by increasing demand for data processing and storage. This demand is driven by several factors, including but not limited to:

- Increased use of electronic transactions in financial services, such as online banking and electronic trading
- Growing use of Internet communication and entertainment
- Shift to electronic medical records for healthcare
- Growth in global commerce and services
- Adoption of satellite navigation and electronic shipment tracking in transportation

Other important trends contributing to data center growth in the government sector include:

- Use of the Internet to publish government information
- Government regulations requiring digital records retention
- Enhanced disaster recovery requirements
- Emergency, health, and safety services
- Information security and national security
- Digital provision of government services (e.g. e-filing of taxes and USPS online tracking)
- High performance scientific computing

1.3 Need for Green

In the U.S. alone, power consumption by data centers and servers is projected to grow to 100 million MWh by 2011, requiring the power equivalent of ten new coal-fired or nuclear power plants[7]. As data center power consumption grows, so do carbon dioxide (CO_2) emissions. Gartner estimates that data centers currently generate 23 percent of all emissions produced by the Information and Communications Technology industry[8], a figure that continues to trend upward (Figure 2).

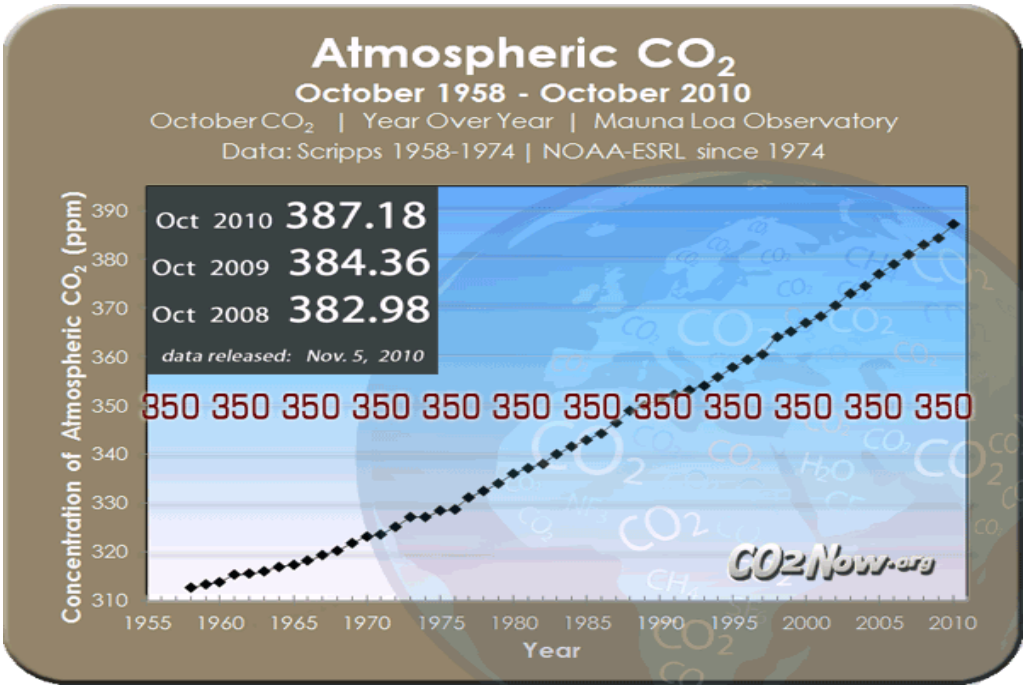


Figure 2 : Growth of worldwide Atmospheric $C0_2$ [9]

[7] "Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431," EPA ENERGY STAR Program, August 2, 2007, at 58. http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf

[8] "Gartner: Data Centres Account for 23% of Global ICT CO2 Emissions," November 5, 2007, http://engineers.ihs.com/news/gartner-datacentre-co2.htm

[9] Image source: http://co2now.org/

According to the U.S. Environmental Protection Agency (EPA), U.S. data centers were estimated to produce 44.4 million metric tons (MMT) of CO2 emissions in 2007 and, based on historical trends, will produce more than 79 MMT of CO_2 in 2011. To put this figure in perspective, it is approximately one half of the total carbon emissions of the entire airline industry, according to McKinsey. Data center emissions are projected to actually exceed those of the airlines by 2020[10].

Given the growing use of power-hungry hardware in the data center, these trends appear incontrovertible. While server processing capability is becoming cheaper, in accordance with Moore's Law, power consumption is increasing at such a rate as to mitigate the resulting hardware cost improvements.

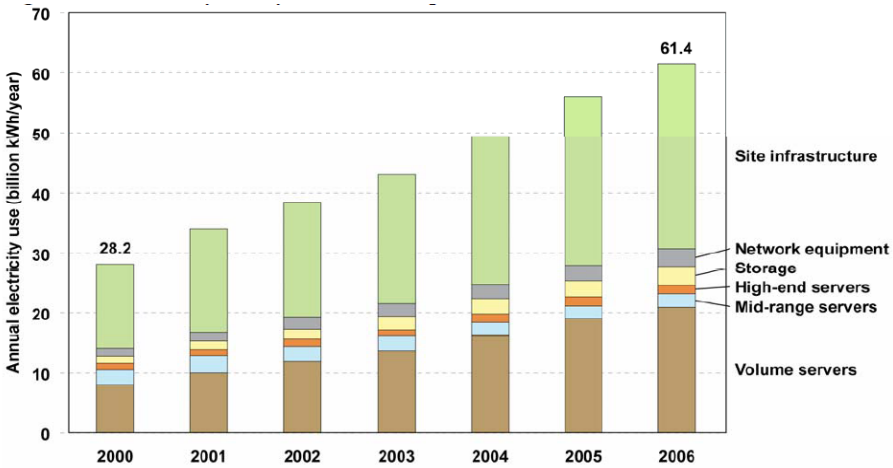


Figure 3 - IT Data center energy use by component[11]

Over the past several years, growth in data center carbon emissions has even sparked debate among legislators regarding whether there should be mandated energy usage reductions or improvements in data center energy efficiency. In December of 2006, through Public Law 109-341, the U.S. Congress issued a request to the EPA requiring a study of the growth trends associated with data centers and servers, and how the problem could be best addressed.

In its subsequent report, the EPA raised four important consequences of data center energy growth:

[10] Walaika Haskins, "Data Centers May Spew More Carbon than Airlines by 2020," TechNewsWorld, May 1, 2008, http://www.technewsworld.com/story/62840.html
[11] "Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431," EPA ENERGY STAR Program, August 2, 2007, page 21

- Increased energy costs for business and government
- Increased emissions, including greenhouse gases, from electricity generation
- Increased strain on the existing power grid to meet the increased electricity demand
- Increased capital costs for expansion of data center capacity and construction of new data centers

For these reasons, there has been mounting interest in opportunities for energy efficiency in this sector. To its credit, the IT industry is actively investigating and developing solutions, such as power-managed servers and adaptive cooling. We shall see how we achieve this in section 3 of this article.

1.4 IT Costs

It is quite clear that data growth leads to increased cost. This includes capital and operational costs. Historically, IT leaders have felt that technology needs are underfunded. But last year, due to the global economic crisis, concerns about funding shifted to how to reduce technology budgets, or increase the revenue from technology-related endeavors, while at the same time enhancing effectiveness.

It is also clear that technology budgets will remain flat (or even decrease) for the foreseeable future as higher education institutions continue to grapple with the effects and after-effects of the downturn in the economy. Given this reality, now may be the time to turn this issue completely on its head. That is, rather than bemoaning insufficient funding for present and planned IT services and initiatives and trying to find the best ways to seek increased funds, perhaps the time has come for IT leaders to accept the level of funding or technology as a given and begin to work with others on campus to determine what services can be offered within the allocated budget. Rather than seeing the perceived (or real) lack of funding as a problem, perhaps IT leaders can see it as an opportunity to engage with other campus leaders in a meaningful discussion about priorities. By turning lemons into lemonade, IT leaders can embrace the current fiscal climate and use it as a catalyst to begin a conversation—one that is good to have at any time but that is especially important now.

Critical questions for funding IT include the following:

- Has the IT organization worked to create a multi-year operating and capital budget, however imperfect, so that both the IT organization and the broader campus can understand and effectively plan within fiscal realities?

- Has IT leadership strategically leveraged the economic downturn to create opportunities for meaningful discussions of collaboration and cooperation across the campus?
- Has the IT organization engaged with the campus broadly to better understand how others perceive their technology needs and to learn more about which services they value most highly?
- Does the governance structure provide a framework for engaging in a meaningful conversation about institutional priorities for information technology, given limited financial resources?
- Are IT leaders prepared to make the cognitive leap from identifying strategies to obtain additional funds to seeking to create consensus about what can be done given a stable (and limited) level of funding?

1.5 Compliance and Governance

When we talk about technologies, Compliance and Governance are two topics that haven't received the required attention needed. As businesses move from paper to electronics, the regulations applicable in the geography require the IT department to comply with it. Compliance is typically defined as the necessary mandatory response to an authorized third party, such as a government regulator. However, it can also be:

- A voluntary response to a trade association or vertical industry body to adopt common practices that make it easier for customers to work with the industry as opposed to a substitute industry
- A response to a mandated industry standard, such as PCI DSS (Payment Card Industry Data Security Standard) for the handling of information such as that related to credit card transactions
- An intentional response to protect an enterprise against lawsuits
- A voluntary response to follow good practices to protect intellectual assets (e.g. patents or trade secrets)

Note that compliance and governance are not the same. Governance deals with the processes and systems designed to ensure the proper accountability for the conduct of a company's business. In that sense, governance is broader than compliance. Governance, risk management, and compliance (GRC)—although they can be described separately, individually, and distinctly—are interrelated and overlap, so it is important to integrate a GRC framework in which all three are considered simultaneously. This places the focus on what needs to be done rather than on how to divide responsibilities among each of the three GRC pillars. While

organizations may want to consider compliance before governance (such as complying with a specific regulation without a formal governance structure in place), the inverse should be the case. Governance is concerned with the overall conduct of an organization, whereas compliance only results in constraints on that governance. (G.Hill)

IT governance is needed to ensure that IT investments generate value/reward and mitigate IT-associated risks, avoiding failure. IT is central to organizational success—effective and efficient delivery of services and goods—especially when IT is designed to bring about change in an organization. This change process, commonly referred to as "business transformation," is now the prime enabler of new business models in both the private and public sectors. Business transformation offers many rewards, but it also has the potential for many risks which may disrupt operations and have unintended consequences. The dilemma becomes how to balance risk and rewards when using IT to enable organizational change.

2. Cloud Computing Considerations

Cloud computing, the hottest buzzword in the technology industry, has topped Gartner's Top 10 Strategic technologies consecutively in 2011[12] and 2010[13]. Still, it is not well understood or defined as it is still an evolving paradigm. As with any new technologies that emerge, there are risks and issues such as security, trust, privacy, availability, replication, data integrity, regulatory requirements, and others. Rushing to adopt cloud computing without proper analysis is not a good decision. However, ignoring cloud computing altogether because of its security and other issues might not be a good idea, either.

2.1 In-house/On-premises model

Unless there is a requirement to install a new application and utilize it to its fullest, the storage and server setup in a typical environment without the cloud (onsite) would have an existing server which would continue to do the same thing as it did yesterday. Now, if it is required to have additional applications or if this application needs to be moved to another 'virtualized' infrastructure, a 'change request' is required. The change request has to go through its processes and has to be approved by all the stakeholders and a 'change committee'.

[12] Gartner press release: "Identifies the Top 10 Strategic Technologies for 2011" October 19, 2010 http://www.gartner.com/it/page.jsp?id=1454221
[13] Gartner press release: "Identifies the Top 10 Strategic Technologies for 2010" October 20, 2009 http://www.gartner.com/it/page.jsp?id=1210613

The chances of the change being approved depend on various factors and, if it is approved, will be implemented provided all the required resources are available.

In a couple of months, the server requires additional storage. Another change request is required for this and it will be appropriately dealt with depending on the availability of the storage. In case the storage is not available, the storage folks will request the vendor to add additional disks and finally add the storage to the server.

The organization might have outsourced all or part of its IT operations to an IT service provider. The outsourcer might provide a dedicated team for this organization and its environment would be as stable as the organization itself.

Managing all the above is not easy. Points to note in this model include:

1. **Increased Total Cost of Ownership and initial cost:** acquiring, installing, and managing the resources
2. **Increased operational expenditure:** IT staff, power and cooling requirements, pay for even what you don't use everyday
3. **Flexibility**: Rapid changes based on business needs is not possible
4. **Infrastructure Complexity**
5. **Painful System Upgrades:** Software patches and hardware refresh
6. **IT Management/Governance**: Following all proper applied frameworks and processes
7. **Security/Data Control**: Data is highly secured and integrity is maintained, fewer dependencies
8. **Regulatory Compliance**: Much easier to manage and ensure that all compliance requirements are stricly in place
9. **Performance**: Can expect better performance than any other model because everything is in-house with fewer dependencies

2.2 Businesses need services, not technologies

"I need the application up and running all the time". This seems to be the usual phrase from the business to IT. Businesses require its services to be available all the time. While we in IT may get excited by talk of virtualized cloud services and ITIL frameworks, the people impacted by these technologies usually cannot care less about the fancy verbal footwork; they simply want to know how their working lives will be improved by what we are peddling. People in the business just want services. They want the right technology-enabled services that help them get their job done and broaden their horizons into new opportunities.

The traditional role of IT has been to deliver and manage technology-based capabilities that are an aggregation of assets such as applications, databases, networks, storage, and servers. IT is typically organized around the various technology domains, taking in new work via project requests and moving it through a plan-build-run lifecycle. This delivery-oriented, technologic-centric approach to IT has inherent latency built into the model, which has, in turn, created an ever-increasing tension between itself and the business it serves.

Cloud computing is an IT offering which solves old problems quickly, cost-effectively, and with higher quality results.

2.3 Definition

There are several definitions of cloud computing from various organizations and research units. To understand it simpler and its uses, National Institute of Standards and Technology (NIST) defines[14]:

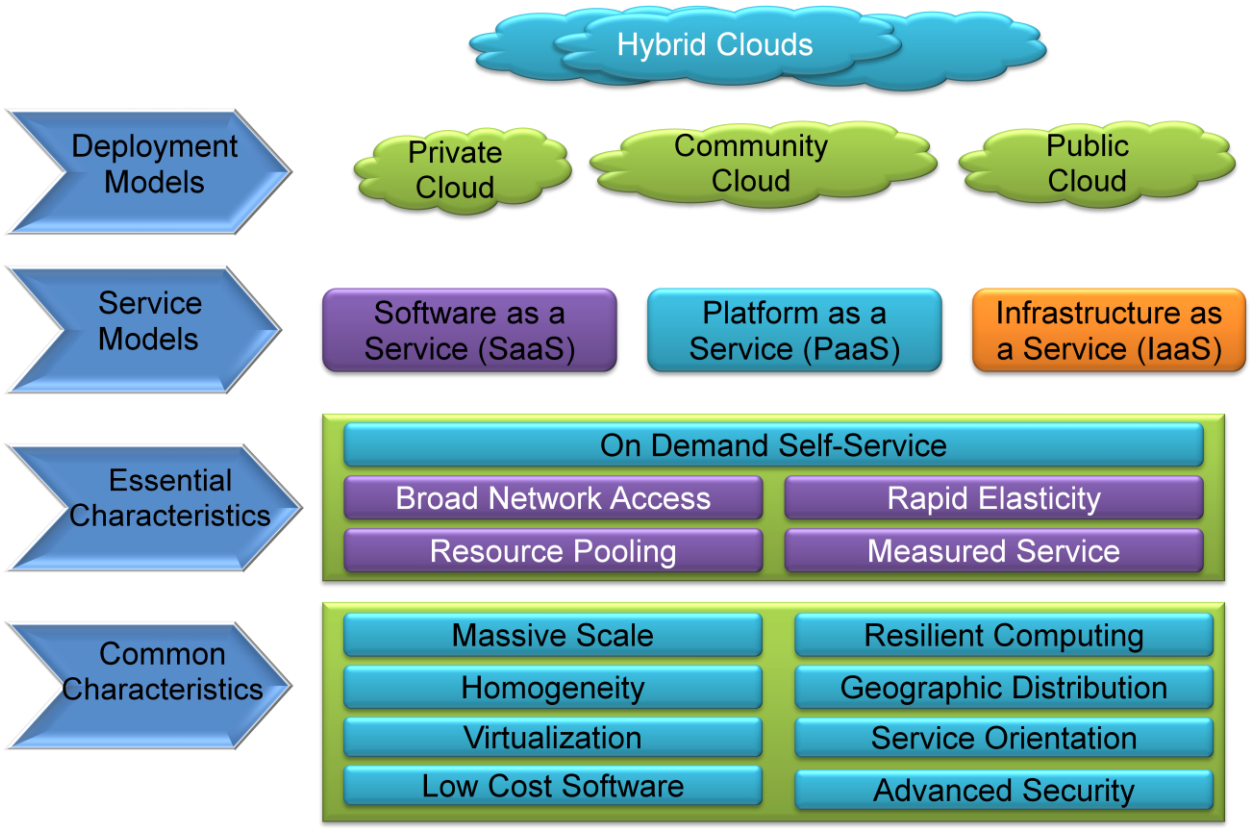


Figure 4 - Cloud Definition Framework

[14] The NIST definition of Cloud computing, Authors: Peter Mell and Tim Grance, csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc

Cloud computing is a model for enabling 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. This cloud model promotes availability and is composed of five essential characteristics**,** three service models, and four deployment models.

The five characteristics

- 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 (e.g. mobile phones, laptops, and PDAs).
- Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that 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 data center). 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 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). Resource usage can be monitored, controlled, and reported providing transparency for both the provider and consumer of the utilized service.

Delivery Models

- Cloud Software as a Service (SaaS): The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client

interface, such as a web browser (e.g. web-based email). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings. Salesforce.com is an example of SaaS.

- Cloud Platform as a Service (PaaS): The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations. Google App Engine and Microsoft Azure are examples of PaaS.
- Cloud Infrastructure as a Service (IaaS): The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g. host firewalls). Amazon Web Services – Elastic Compute Cloud, Rackspace, and GoGrid are examples of IaaS.

Deployment Models

- Private cloud: The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on-premise or off-premise.
- Community cloud: The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g. mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on-premise or off-premise.
- Public cloud: The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.
- Hybrid cloud: The cloud infrastructure is composed of two or more clouds (private, community, or public) that remain unique entities but are bound together

by standardized or proprietary technology that enables data and application portability (e.g. cloud bursting for load-balancing between clouds).

Multi-Tenancy

Multi-tenancy in cloud service models implies a need for policy-driven enforcement, segmentation, isolation, governance, service levels, and chargeback/billing models for different consumer constituencies. Consumers might utilize a public cloud provider's service offerings or actually be from the same organization, such as different business units rather than distinct organizational entities, but would still share infrastructure.

From a provider's perspective, multi-tenancy suggests an architectural and design approach to enable economies of scale, availability, management, segmentation, isolation, and operational efficiency by leveraging shared infrastructure, data, metadata, services, and applications across many different consumers.

Multi-tenancy can also take on different definitions depending upon the provider's cloud service model as much as it may entail enabling the capabilities described above at the infrastructure, database, or application levels. An example would be the difference between an IaaS and SaaS multi-tenant implementation.

Cloud deployment models place different importance on multi-tenancy. However, even in the case of a private cloud, a single organization may have a multitude of third-party consultants and contractors, as well as a desire for a high degree of logical separation between business units. Thus, multi-tenancy concerns should always be considered.

2.4 Importance of Cloud Computing

There are instances where the cloud computing offering proves to be significantly better than setting up the IT on your own. Small organizations can move to cloud quickly, unlike the big players. With the benefits of any new technologies, cloud computing presents risks and issues such as security, loss of privacy, regulatory compliance, data replication, coherency and erosion of integrity, application sprawl, dependencies, and others. Due to these factors, no large organization would want to rush into cloud computing. On the other hand, these factors are of less importance to a young company and hence they gain more by moving to the cloud quickly.

Large organizations are seriously considering cloud computing as its platform for IT or they are moving some of its operations to the cloud. Cloud computing offers access to completely

different levels of scale and economics in terms of the ability to scale very rapidly and to operate IT systems more cost-effectively than previously possible.

A poll conducted by a research firm showed the top three benefits of deploying cloud computing:

1. **Easy to deploy**: Faster delivery of service. Increased time to value and time to market.
2. **Pay as you go**: Reduction in cost. Reduced CapEx versus OpEx. Competitive costs.
3. **Less in-house IT staff**: IT department transformation, more focus on innovation.

Recent economic troubles required IT to be quick to respond to the requirements of businesses. Cloud computing enables businesses to achieve this easily (e.g. a retail firm whose business typically increases during the Christmas holidays can acquire IT-as-a-service instantly from a Cloud service provider) (**Easy to deploy**), and pay only for the period it uses the service (**Pay as you go**). Management of the cloud and the IT services offered will be undertaken by the Cloud service provider, enabling the retail firm's employees to focus on innovating the business rather than the IT (**Less in-house IT staff**).

Managing an environment which grows from a few servers to thousands of servers will become more difficult without proper processes, procedures, and tools. Often, large organizations will have best practice frameworks such as ITIL and COBIT to govern, manage, and secure these large and complex environments. Cloud computing helps organizations simplify their management by applying the solutions quickly with less TCO.

2.5 Impact of Cloud Computing

This section describes the impact of cloud computing on various types of consumers. The consumers can be categorized into:

1. Individual Consumers
2. Individual Businesses
3. Start-ups/Young Businesses
4. Small and Medium-Size Businesses(SMB)
5. Enterprise Businesses

2.5.1 Individual Consumers

Cloud computing is already used by tech-savvy users who are inspired by the success of websites such as Facebook, Napster, MySpace, and so on to start their own similar business. Students who require various computing resources use some form of cloud computing.

Building several servers from a laptop is now possible thanks to cloud computing. This isn't virtualization of the laptop's hardware resources but rather, installing, configuring, accessing, and managing a variety of 'assets' using the Internet. The best example is Amazon Web Services (AWS) which allows any user with an Internet connection to set up their own IT infrastructure by paying for only its usage.

Any user having access to the Internet will have an email account that is hosted on a cloud. A slightly advanced user might use cloud computing in the form of social networking (e.g. Facebook, MySpace, Orkut, LinkedIn), Web albums to share photos (e.g. Google Picasa, Flickr, Snapfish), and store documents on the web (e.g. Google Docs, Zoho). All these services are actually hosted in some form of public cloud.

2.5.2 Individual Businesses

Some users focus on developing businesses using the cloud, instead of just using it as service. Inspired by the low entry costs for cloud services, technically savvy consumers are now using cloud-based tools to develop their businesses and to market or sell their products. The expectation is that software should be nearly free of charge, and that users should pay only for additional services or extra capacity. Consumers can host a website to attract customers, use eBay or Craigslist to sell and market individual items, use virtual marketing to spread the word, place ads with search engine providers, engage with online banks to manage funds, supervise online accountancy services to manage finances, and use office assistants to book trips and arrange appointments. All of this computing power can reside in the cloud.

2.5.3 Start-ups

By definition, a startup company (or startup) has a limited operating history. These companies, generally newly created, are in a phase of development and research for markets. When entrepreneurs wish to start their businesses, IT would have lower priority than the core business which is still evolving. The focus is more on marketing, research, and development of the product. Once in place, the entrepreneur could simply opt for hosting the IT required for this business on the cloud and adopt the pay-as-you-grow approach. A company would demonstrate scalability by implementing a robust enterprise resource planning (ERP) solution and hosting it on the premises. Currently, a more common approach is to outsource the majority of IT and maintain a lean IT shop. The challenge now becomes getting locked into provider contracts and the levels of service that the Cloud Service Provider will face. Critical success factors are the ability to scale the infrastructure as volume increases, and rapidly modify the service for new product lines, channels, markets, or business models. One potential model is a mixed model

based on the classic definition of core and context, with control for context maintained internally. The evolution depends on the interoperability across platforms that are internal or are in the cloud. Start-ups have less legacy data and fewer processes and applications than established companies, and they pioneer some of the cloud computing services for an integrated business.

2.5.4 Small and Medium-Size Businesses

There may be as many definitions of small and medium-sized businesses (SMB) as there are definitions of cloud computing. Often, the SMB category is defined by revenue, but when discussing technology requirements, it's equally important to think about the number of products, number of channels, countries of operation, and integration of the supply chain with third parties. In short, saying something is a "small business" is a measure of the business' complexity. Many small businesses grow through acquisition, or are born as a spin-off from a larger business.

The SMB age is a critical component in understanding the maturity and entrenchment of legacy processes and data. The requirements for data security and privacy are no less onerous than for a larger enterprise. One generalization about SMBs is that their IT departments are smaller, and are therefore less diverse in skills and knowledge than those of larger enterprise businesses.

Significant IT projects can become difficult to justify and investment in IT can decline, IT infrastructure become outdated, and the IT group can have difficulty responding to business needs in a timely manner. Decision making in a SMB is often concentrated among fewer individuals than in a larger enterprise. Depending on the specific scenario, the SMB environment has some essential characteristics that can accelerate growth in the broad use of cloud computing. We may see complex SMBs as the vanguard of cloud computing with no in-house infrastructure and IT services delivered from a combination of Cloud Service Providers.

2.5.5 Enterprise Businesses

Mature enterprise businesses are broadening their use of cloud-enabled computing. At a minimum, this could mean allowing users to access services beyond the corporate firewall. Broader use of cloud services includes using knowledge tools to support personal productivity, such as online research or travel services. Companies may use corporate applications, such as employee work surveys that use the company's directory to populate broad characteristics but that don't include personally identifiable information. Mature businesses adopting cloud computing may also use cloud applications in business-critical departments and functions, such

as Salesforce.com applications, document management, purchasing, and logistics. In these cases, the users access applications and store in the cloud data that includes personal and sensitive information. In evaluating an application run in-house or a cloud-based service, security and privacy concerns could trump costs. An important consideration is data redundancy between Cloud Service Providers and the traditional enterprise applications. Vendor lock-in to a proprietary architecture or solution would kill the cost, flexibility, and extensibility arguments. The compelling argument for a cloud solution is time to market, where a cloud application is the only feasible alternative, given cost and time constraints.

2.6 Governing the Cloud

As assets move out of the data center, an organization's control over the asset diminishes while the Cloud Service Provider's control increases. As SaaS keeps 'all the other stuff' behind the end user, organizations have minimal control over it while Cloud Service Providers have more. On the other hand, in an IaaS deployment model, organizations have full control over the infrastructure. The key point to note here is, with any model, organizations should develop and enforce proper frameworks and service level agreements (SLA) with the Cloud Service Provider to ensure all the contractual agreements are met. Figure 5 illustrates the impact of cloud computing on the Governance structure of an IT organization.

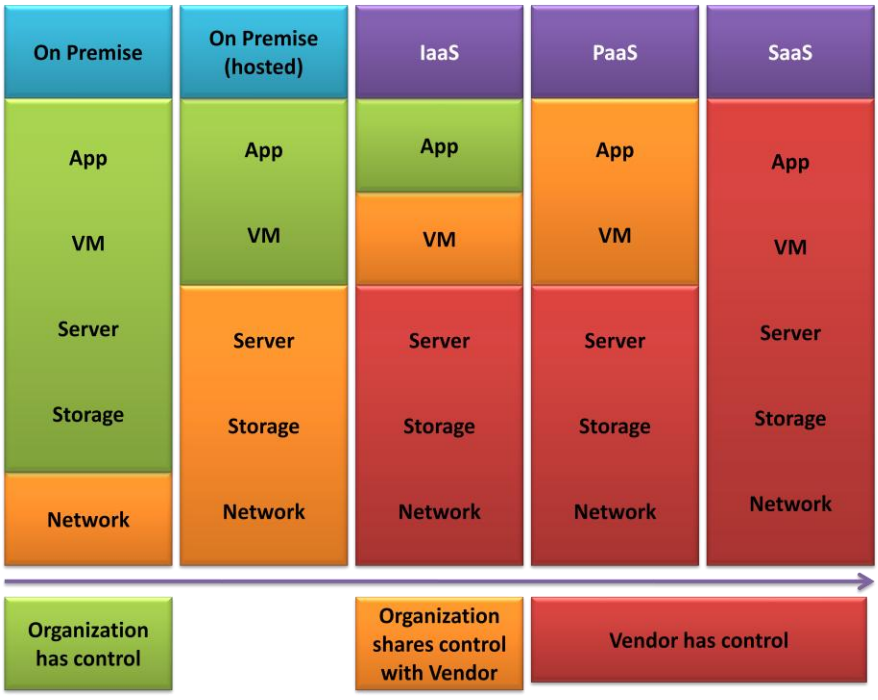


Figure 5 - Impact of cloud computing on the Governance structure of an IT organization

2.7 Risks in the Cloud

When resources move beyond internal boundaries to the cloud, accessing valuable data assets must be secure. As most cloud services are accessed via the Internet, there exists a huge risk to the data in transit. The risk is not just with the data path through which the data travels but also the location where each customer's data is stored. Using virtualization, several virtual machines would be co-located on the same physical server. A Cloud Service Provider must ensure that each customer's data is secure and isolated from others.

One of the known high profile 'risks' in the cloud was seen recently when Wikileaks used Amazon Web Services to host its website due to heavy attack on its old servers. This could have been the best example of when cloud can be used in a crisis. Privacy and confidentiality would have been showcased had this been completely successful.

However, Amazon stopped hosting the Wikileaks website citing the 'content' that is hosted on it inappropriate[15][16]. This raises several questions related to privacy and confidentiality of the data being hosted on the cloud. From a cloud perspective, the service provider denied hosting the asset because of the type of content hosted by the customer. That might imply that service providers can actually look into the customer's data.

Various security aspects must be considered before moving to the cloud. Enterprise customers must check all the terms and conditions with the service provider to ensure there are no loopholes, especially data security. The most important security aspects that customers need to look at are mentioned in detail below.

2.7.1 Confidentiality

When data moves out of the enterprise to the public cloud, the data is physically shared with other customers apart from the enterprise. Only logically, the data is separated from other customers.

Enterprises must ensure that the data in any form is encrypted. It is well known that the data in transit is always encrypted (familiar examples include Secure Socket Layer and Transport Level Security). However, point to multi-point encryption for data in transit is not as readily available as point-to-point encryption. This might bring issues when the enterprise scales and the data in transit are exposed to the risk of data theft.

---

[15] "Facing Lieberman Boycott, Amazon Ousts WikiLeaks", antiwar.com December 01, 2010, http://news.antiwar.com/2010/12/01/facing-lieberman-boycott-amazon-ousts-wikileaks/

[16] http://itknowledgeexchange.techtarget.com/mobile-cloud-view/wikileaks-shut-down-provides-lesson-for-cloud-buyers/

Data at rest possibly might not be encrypted. Enterprises must ensure in the contractual agreements the responsibility of encrypting the data at rest with the enterprise itself or with the Cloud Service Provider.

Cryptographic key management in the enterprise today is a failed model of proprietary 'solutions' that is not effective, nor is it scalable. If we expect key management to work in the cloud, then we need a new model. Fortunately, Organization for the Advancement of Structured Information Standards (OASIS), and specifically, its Key Management Interoperability Protocol (KMIP) Technical Committee (TC) is working on an improved model—unified cloud management. While KMIP will certainly improve enterprise key management, such an improvement is still not good enough to scale to cloud computing.[17]

Cloud computing needs federated key management. Similar to federated identity management, federated key management is needed for inter-(cross) enterprise and cloud use. The primary issues that need to be addressed with key management for cloud computing are not only interoperability, which KMIP is addressing, but also scalability, which KMIP is not addressing.

The lack of a viable key management model available today for cloud computing is a major security and operational issue. A scalable key management model is required for cloud computing use to effectively scale.

2.7.2 Data Integrity

One area that enterprises often overlook is data integrity. Cloud storage is a very important part of enterprise cloud computing. We continue to assume that data stored in storage arrays are completely flawless. However, research[18] conducted by CERN outlined the existence of low level data corruptions caused by various factors. The most reliable RAID configuration, RAID-5, doesn't actually help much for data corruption[19]. There is some effort to reduce data corruption, but it is very unlikely that it will disappear completely.

Various audits, such as Sarbanes–Oxley, and DS11 of COBIT – Manage Data, require that organizations have the ability to track data as it moves through the enterprise to ensure data integrity. Using public cloud, the Cloud Service Provider must be able to assure data lineage to

[17] "Key Management in the Cloud" By Tim Mather, January 7, 2010. http://broadcast.oreilly.com/2010/01/key-management-in-the-cloud.html

[18] "Data integrity", Bernd Panzer-Steindel, CERN/IT, 2007. http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797

[19] RAID controllers don't check the 'parity' when reading data from RAID 5 file systems. In principle the RAID controller should report problems on the disk level to the OS, but this seems not always to be the case. The controller allows to run the 'verify' command which reads all data from disk and re-calculates the RAID5 checksum and corrects the discovered bad RAID5 blocks (block size is 64 KB). However, it does not have a notion of what is 'correct' data from the user point of view.

customer and auditor satisfaction. Cloud Service Providers must be able to provide satisfactory results when it comes to information accuracy, completeness, and authenticity.

2.7.3 Data Availability

When Enterprises move from in-house infrastructure to clouds, there is a need to ensure that the data is available to authorized users at any given time. Data availability doesn't just include online data but backup data as well. Cloud Service Providers must ensure proper data backups (if data is backed up to the cloud). The data available must be reliable (data integrity) and also must be secured (encrypted – confidentiality).

As data is no longer physically separated from data belonging to other customers, Cloud Service Providers must guarantee that the data used by a customer will in no way be accessed or used by any other customer.
Data Sanitization/Remanence must be done in a similar way as done while disposing of in-house data. This as well is an important aspect for data availability in the cloud.

2.7.4 Identity and Access Management

Identity and Access Management (IAM) will play a key role for customers adopting cloud services apart from security and compliance. Managing IAM remains one of the greatest challenges facing IT today. While an enterprise may be able to leverage several cloud computing services without a good IAM strategy, in the long run, extending an organization's identity services into the cloud is a necessary precursor toward strategic use of on-demand computing services. To support aggressive adoption of an admittedly immature cloud-based IAM, companies should have a good understanding of the capabilities of their Cloud Computing providers, including:

1. **Identity Provisioning:** Provisioning today is proprietary to the Cloud Service Provider. Automating provisioning for users to access the cloud is a painful task. Enterprises that have invested in user management processes within an enterprise will seek to extend those processes and practice to cloud services.
2. **Authentication:** Strong authentication is available only through delegation and federation. When organizations start to utilize cloud services, authenticating users in a trustworthy and manageable manner is a vital requirement. Organizations must address authentication-related challenges such as credential management, strong authentication (typically defined as multi-factor authentication), delegated authentication, and managing trust across all types of cloud services.

3. **Federation:** In a cloud computing environment, Federated Identity Management plays a vital role in enabling organizations to authenticate their users of cloud services using the organization's chosen identity provider (IdP). In that context, exchanging identity attributes between the service provider and the IdP in a secure way is also an important requirement. Organizations considering federated identity management in the cloud should understand the various challenges and possible solutions to address those challenges with respect to identity lifecycle management and available authentication methods to protect confidentiality and integrity, while also supporting non-repudiation.
4. **Authorization & user profile management:** The requirements for user profiles and access control policy vary depending on whether the user is acting on their own behalf (such as a consumer) or as a member of an organization (such as an employer, university, hospital, or other enterprise). The access control requirements in SPI environments include establishing trusted user profile and policy information, using it to control access within the cloud service, and doing this in an auditable way.

2.8 Threats to Cloud Computing

There are numerous issues that need to be considered when organizations decide to move to the cloud. Organizations as well as Cloud Service Providers must thoroughly understand the risks discussed above before making a decision. According to a document released by the Cloud Security Alliance, the top threats to cloud computing include[20]:

2.8.1 Abuse and Nefarious Use of Cloud Computing

IaaS providers offer their customers the illusion of unlimited compute, network, and storage capacity—often coupled with a 'frictionless' registration process where anyone with a valid credit card can register and immediately begin using cloud services. Some providers even offer free limited trial periods. By abusing the relative anonymity behind these registration and usage models, spammers, malicious code authors, and other criminals have been able to conduct their activities with relative impunity. While PaaS providers have traditionally suffered most from these kinds of attacks, recent evidence shows that hackers have begun to target IaaS vendors as well. Future areas of concern include password and key cracking, DDOS, launching dynamic attack points, hosting malicious data, botnet command and control, building rainbow tables and CAPTCHA solving farms.

[20] Cloud Security Alliance "Top Threats to Cloud Computing" Version 1.0 (2010).

Examples

IaaS offerings have hosted the Zeus botnet, InfoStealer Trojan horses, and downloads for Microsoft Office and Adobe PDF exploits. Additionally, botnets have used IaaS servers for command and control functions. Spam continues to be a problem—as a defensive measure, entire blocks of IaaS network addresses have been publicly blacklisted.

Remediation

- Stricter initial registration and validation processes.
- Enhanced credit card fraud monitoring and coordination.
- Comprehensive introspection of customer network traffic.
- Monitoring public blacklists for one's own network blocks.

Impact

Criminals continue to leverage new technologies to extend their reach, avoid detection, and improve the effectiveness of their activities. Cloud computing providers are actively targeted, partially because their relatively weak registration systems facilitate anonymity, and providers' fraud detection capabilities are limited.

Impacted service models: IaaS and PaaS

2.8.2 Insecure Interfaces and APIs Description

Cloud computing providers expose a set of software interfaces or APIs that customers use to manage and interact with cloud services. Provisioning, management, orchestration, and monitoring are all performed using these interfaces. The security and availability of general cloud services is dependent upon the security of these basic APIs. From authentication and access control to encryption and activity monitoring, these interfaces must be designed to protect against both accidental and malicious attempts to circumvent policy. Furthermore, organizations and third parties often build upon these interfaces to offer value-added services to their customers. This introduces the complexity of the new layered API; it also increases risk, as organizations may be required to relinquish their credentials to third parties in order to enable their agency.

Examples

Anonymous access and/or reusable tokens or passwords, clear-text authentication or transmission of content, inflexible access controls or improper authorizations, limited monitoring and logging capabilities, unknown service or API dependencies.

Remediation

- Analyze the security model of cloud provider interfaces.
- Ensure strong authentication and access controls are implemented in concert with encrypted transmission.
- Understand the dependency chain associated with the API.

Impact

While most providers strive to ensure security is well integrated into their service models, it is critical for consumers of those services to understand the security implications associated with the use, management, orchestration, and monitoring of cloud services. Reliance on a weak set of interfaces and APIs exposes organizations to a variety of security issues related to confidentiality, integrity, availability, and accountability.

Impacted service models: IaaS, PaaS, and SaaS

2.8.3 Malicious Insiders

The threat of a malicious insider is well-known to most organizations. This threat is amplified for consumers of cloud services by the convergence of IT services and customers under a single management domain, combined with a general lack of transparency into provider process and procedure. For example, a provider may not reveal how it grants employees access to physical and virtual assets, how it monitors these employees, or how it analyzes and reports on policy compliance. To complicate matters, there is often little or no visibility into the hiring standards and practices for cloud employees. This kind of situation clearly creates an attractive opportunity for an adversary—ranging from the hobbyist hacker, to organized crime, to corporate espionage, or even nation-state sponsored intrusion. The level of access granted could enable such an adversary to harvest confidential data or gain complete control over the cloud services with little or no risk of detection.

Examples

No public examples are available at this time.

Remediation

- Enforce strict supply chain management and conduct a comprehensive supplier assessment.
- Specify human resource requirements as part of legal contracts.
- Require transparency into overall information security and management practices, as well as compliance reporting.
- Determine security breach notification processes.

Impact

The impact that malicious insiders can have on an organization is considerable, given their level of access and ability to infiltrate organizations and assets. Brand damage, financial impact, and productivity losses are just some of the ways a malicious insider can affect an operation. As organizations adopt cloud services, the human element takes on an even more profound importance. Therefore, it is critical that consumers of cloud services understand what providers are doing to detect and defend against the malicious insider threat.

Impacted service models: IaaS, PaaS, and SaaS

2.8.4 Shared Technology Issues

IaaS vendors deliver their services in a scalable manner by sharing infrastructure. Often, the underlying components that make up this infrastructure (e.g. CPU caches, GPUs, and so on) were not designed to offer strong isolation properties for a multi-tenant architecture. To address this gap, a virtualization hypervisor mediates access between guest operating systems and the physical compute resources. Still, even hypervisors have exhibited flaws that enable guest operating systems to gain inappropriate levels of control or influence on the underlying platform. An in-depth defense strategy is recommended, and should include compute, storage, and network security enforcement and monitoring. Strong compartmentalization should be employed to ensure that individual customers do not impact the operations of other tenants running on the same cloud provider. Customers should not have access to any other tenant's actual or residual data, network traffic, etc.

Examples

- Joanna Rutkowska's Red and Blue Pill exploits
- Kortchinksy's CloudBurst presentations

Remediation

- Implement security best practices for installation/configuration.
- Monitor environment for unauthorized changes/activity.
- Promote strong authentication and access control for administrative access and operations.
- Enforce service level agreements for patching and vulnerability remediation.
- Conduct vulnerability scanning and configuration audits.

Impact

Attacks have surfaced in recent years that target the shared technology inside cloud computing

environments. Disk partitions, CPU caches, GPUs, and other shared elements were never designed for strong compartmentalization. As a result, attackers focus on how to impact the operations of other cloud customers, and how to gain unauthorized access to data.

Impacted service models: IaaS

2.8.5 Data Loss or Leakage

There are many ways to compromise data. Deleting or altering records without a backup of the original content is an obvious example. Unlinking a record from a larger context may render it unrecoverable, as can storage on unreliable media. Loss of an encoding key may result in effective destruction. Finally, unauthorized parties must be prevented from gaining access to sensitive data. The threat of data compromise increases in the cloud due to the number of and interactions between risks and challenges which are either unique to cloud, or more dangerous because of the architectural or operational characteristics of the cloud environment.

Examples

Insufficient authentication, authorization, and audit (AAA) controls; inconsistent use of encryption and software keys; operational failures; persistence and remanence challenges: disposal challenges; risk of association; jurisdiction and political issues; data center reliability; and disaster recovery.

Remediation

- Implement strong API access control.
- Encrypt and protect integrity of data in transit.
- Analyze data protection at both design and run time.
- Implement strong key generation, storage and management, and destruction practices.
- Contractually demand providers wipe persistent media before it is released into the pool.
- Contractually specify provider backup and retention strategies.

Impact

Data loss or leakage can have a devastating impact on a business. Beyond the damage to one's brand and reputation, a loss could significantly impact employee, partner, and customer morale and trust. Loss of core intellectual property could have competitive and financial implications. Worse still, depending upon the data that is lost or leaked, there might be compliance violations and legal ramifications.

Impacted service models: IaaS, PaaS, and SaaS

2.8.6 Account or Service Hijacking

Account or service hijacking is not new. Attack methods such as phishing, fraud, and exploitation of software vulnerabilities still achieve results. Credentials and passwords are often reused, which amplifies the impact of such attacks. Cloud solutions add a new threat to the landscape. If an attacker gains access to your credentials, they can eavesdrop on your activities and transactions, manipulate data, return falsified information, and redirect your clients to illegitimate sites. Your account or service instances may become a new base for the attacker. From here, they may leverage the power of your reputation to launch subsequent attacks.

Examples

No public examples are available.

Remediation

- Prohibit sharing of account credentials between users and services.
- Leverage strong two-factor authentication techniques where possible.
- Employ proactive monitoring to detect unauthorized activity.
- Understand cloud provider security policies and SLAs.

Impact

Account and service hijacking, usually with stolen credentials, remains a top threat. With stolen credentials, attackers can often access critical areas of deployed cloud computing services, allowing them to compromise the confidentiality, integrity, and availability of those services. Organizations should be aware of these techniques as well as common in-depth defense protection strategies to contain the damage (and possible litigation).

Impacted service models: IaaS, PaaS, and SaaS

2.8.7 Unknown Risk Profile

One of the tenets of cloud computing is the reduction of hardware and software ownership and maintenance to allow companies to focus on their core business strengths. This has clear financial and operational benefits, which must be weighed carefully against the contradictory security concerns. This is further complicated by the fact that cloud deployments are driven by anticipated benefits by groups who may lose track of the security ramifications. Versions of software, code updates, security practices, vulnerability profiles, intrusion attempts, and security design, are all important factors for estimating your company's security posture. Information about who is sharing your infrastructure may be pertinent, in addition to network intrusion logs, redirection attempts and/or successes, and other logs. Security by obscurity may be low effort,

but it can result in unknown exposures. It may also impair the in-depth analysis required by highly controlled or regulated operational areas.

Examples

- IRS asked Amazon EC2 to perform a C&A; Amazon refused. http://news.qualys.com/newsblog/forrester-cloud-computingqa.html
- Heartland Data Breach: Heartland's payment processing systems were using known-vulnerable software and actually infected, but Heartland was "willing to do only the bare minimum and comply with state laws instead of taking the extra effort to notify every single customer, regardless of law, about whether their data has been stolen." http://www.pcworld.com/article/158038/heartland_has_no_heart_for_violated_customers.html

Remediation

- Disclosure of applicable logs and data.
- Partial/full disclosure of infrastructure details (e.g. patch levels, firewalls, and so on).
- Monitoring and alerting on necessary information.

Impact

When adopting a cloud service, the features and functionality may be well advertised, but what about details or compliance of the internal security procedures, configuration hardening, patching, auditing, and logging? How are your data and related logs stored and who has access to them? What information, if any, will the vendor disclose in the event of a security incident? Often such questions are not clearly answered or are overlooked, leaving customers with an unknown risk profile that may include serious threats.

Impacted service models: IaaS, PaaS, and SaaS

Possible Security Benefits

- Centralized data.
- Segmented data and applications.
- Better logging/accountability.
- Standardized images for asset deployment.
- Better resilience to attack and streamlining incident response.
- More streamlined audit and compliance.
- Better visibility to process.
- Faster deployment of applications, services, etc.

2.9 Migration to the Cloud

Almost all services can leverage cloud computing; some are suitable but some are not. Some companies host everything on the cloud, some don't. Not everything requires the same level of security. Applications and services that need to be within the firewall must not be exposed to the Internet.

Some things to consider:

- Understand your services.
- Perform the due diligence. Perform audits.
- Where is it hosted? How is it hosted?
- Prefer private clouds if compliance requires.
- Focus on security.
- We still outsource our IT operations after a lot of verification. Perform the same due diligence with the SaaS provider.

Enterprises might not move all of their IT assets (applications, platforms, and infrastructure) which are currently hosted in their data centers or co-located facilities. There will be certain critical applications that need to stay in-house but cloud vendors are now capable of hosting almost anything that was previously deployed in-house.

IT managers would love to move some of their applications to the cloud to reduce costs. It is their opportunity to remove capital and management expenses and to focus on pure business. However, this can't happen overnight and it requires proper assessment and planning.

Not all applications can be migrated to the cloud easily. New applications 'compatible' with the cloud in the data center can move more easily than applications that have various dependencies and that are integrated with other systems, tools, and processes. Enterprises see these applications as a potential service that need not to be in-house. Moving these applications to the cloud can be very complicated and time consuming.

There are several steps and strategies that need to be considered before migrating to the cloud. Here, we are considering an example where the enterprise decides to migrate some of their assets (application, infrastructure) to the cloud for various benefits.

The whole process of migration to the cloud can be broadly divided into two phases. The first phase, which is the most important phase, involves assessment of the IT assets which will be

moved to the cloud. The next phase–migration phase–involves the actual process of movement of the IT asset to the cloud.

2.10 Cloud Migration - Assessment Phase

Assessing the existing infrastructure is the most time-consuming and important phase while moving to the cloud. Various factors have to be considered during this phase. It is during this phase that the organization realizes the facts about its IT operations and the benefits of deploying cloud computing.

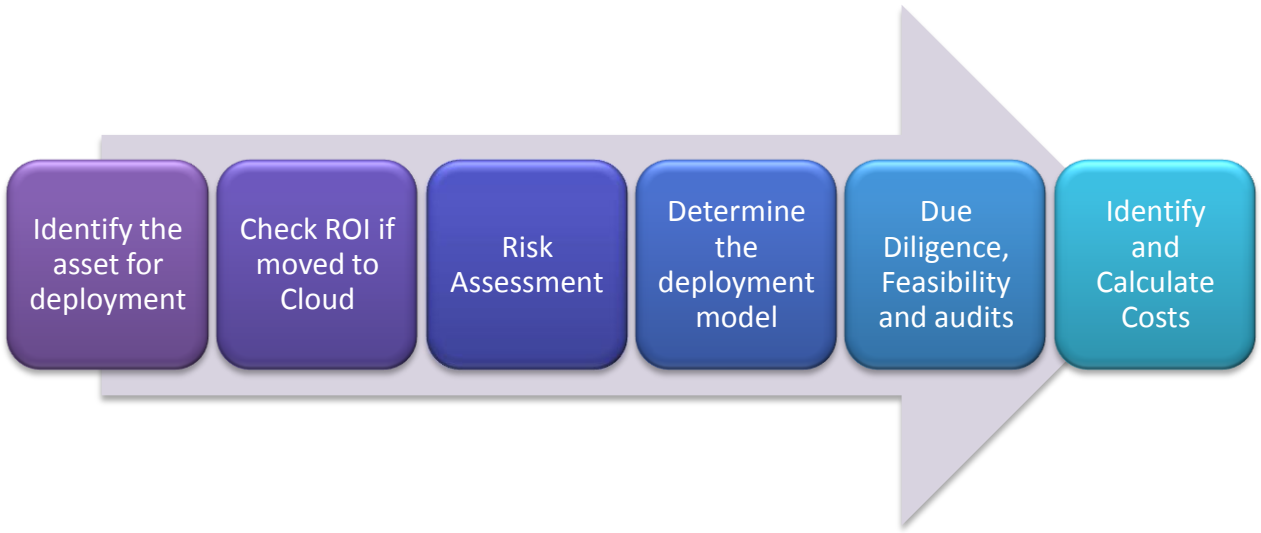


Figure 6 – Cloud Migration: Assessment Phase

2.10.1 Identify the asset for cloud deployment

Today, almost everything in IT can be moved to the cloud. Organizations must decide on what needs to move to the cloud (for a new deployment, organizations must check on which asset is best if deployed in the cloud). Assets supported by the cloud can be simply classified into:

1. Data
2. Application/Services/Funtions/Processes

This means that the assets capable of being deployed are either information or services.

Each organization's environment and requirement are different. Making an asset checklist for each category of the organization is not practical. There are certain assets in some organizations that are not as important as in others, e.g. an insurance company is really concerned with the number of years that particular data is retained; on the other hand, a retail company wants to have its system availability to be very high while placing low importance to

the retention period of the data. With that in mind, each organization has to identify those assets that can potentially be moved to the cloud.

Some of the applications that can be moved to the cloud include:

1. **Email:** Email is the lifeblood of many organizations, and as a result, many companies are not willing to let go of it. That is understandable. But hosted email providers have moved beyond the days of packing 5,000 mailboxes belonging to 300 accounts onto a cheap computer running with a basic POP3/SMTP setup. While basic email service is still out there, you can get hosted Exchange services from a variety of vendors as well as some upscale, non-Exchange offerings. Email architecture has become quite standardized, and there is really no value-add to keeping it inside your firewall other than mitigating regulatory concerns.
2. **Conferencing software:** Setting up and maintaining conferencing software is not fun. To make matters worse, when it is down, it needs to be restored in a hurry. Like email, there is zero benefit to locating this within your firewall. Also like email, the setup and configuration can be complex enough to require an expert, unless you don't mind tying up a staff member for a few days. For a low monthly or yearly fee, this weight can be off your shoulders. No one will notice or mind and your staff can move on to other tasks.
3. **CRM:** The decision to outsource CRM can be scary. After all, like email, CRM is where so many of the company's crown jewels are stored. But there are no technical benefits or advantages to having CRM in-house. It's a fairly low bandwidth application with maintenance overhead you do not need. In addition, the licensing of many CRM systems can be a hassle. Moving to a hosted CRM system can free you to spend more time on more important issues.
4. **Web hosting:** Hosted Web space used to be as awful as hosted email, unless the organization were willing to spend big bucks on a dedicated server. Many vendors have shifted to (or offer) a virtualized hosting environment, which has dramatically increased uptime, reduced security risks, and allowed them to provide much more open and direct access to the servers. This is great news, especially for companies with custom applications that require a deployment path beyond copying some files over.
5. **Development test labs**: Building and maintaining test environments for software developers is a major undertaking. To do it right, you need all sorts of permutations of operating systems, patches, and relevant applications. You could easily find yourself with nearly 100 test beds for a simple Web application, for example. Why do this to yourself when there are quality vendors out there who already have these test systems

set up or that allow you to configure them with point-and-click ease? And you can safely give the keys to the development staff and know that they can't permanently mangle the test systems, too.

6. **Video hosting:** Many companies would block video hosting sites under the assumption that they were only for games, or the real fear of having ads placed on your videos, or compromised video quality. Now, the big name sites have upgraded their quality and few companies block them because there is plenty of legitimate use. In addition, some sites allow you to pay a fairly low charge to give you more control over your video, such as deciding where it can appear and enhancing its quality.
7. **Email security:** Even if you do not put your email with a hosted vendor, you will want to look at having a third party perform your anti-spam and antivirus duties, even if it's only as a first-line of defense. If you look at how much incoming email is spam, you'll see that you can reduce your bandwidth needs dramatically by allowing a third party to perform an initial scan of your email. It will also allow you to have far fewer email servers. Speaking from personal experience, even a small company's email servers and network can be overwhelmed by incoming spam. Getting a good spam scanner outside the network can make a night-and-day difference.
8. **Common application components:** There is always the perpetual "build" vs. "buy" question for development projects, but the cloud adds a new wrinkle. Many functions that used to be the purview of components or libraries you could buy are now being made available as Web services, typically billed on a per-usage basis. Often, these services take a variety of lower-level functions and roll them into a complete, discrete offering. You would be surprised at how many of these Web services are available, and depending upon your usage scenario, it could make sense to use them instead of going it alone.
9. **Basic office applications:** If you need the full power of the Microsoft Office suite, by all means, this isn't for you. But if you are one of the many organizations that use only a small fraction of the Office feature set, it may make sense to look at one of the new crop of online Office replacements (or even Microsoft's online version of Office).
10. **Batch processing applications**: One type of application that shines in the cloud are batch processing applications, such as data warehouses. As long as you can get the data needed into the cloud without disrupting your operations (such as "seeding" it with the shipment of physical media or synchronization over time), the ability to rapidly scale capacity in the cloud can result in tremendous savings. For example, if you need 15 servers' worth of computing capacity for a once-per-week job, do you really want to have

15 servers sitting idle in your server room waiting for that? Or would you rather just modify the task to start 15 cloud instances, run the process, and shut them down again? In a scenario like this, it is clear that cloud computing can deliver significant advantages.

The assets that can be moved are not just limited to applications; an Enterprise can move its development platform and infrastructure to the cloud as well. Examples of the platforms a company can move are listed below:

1. **Application development platform:** This is one that is already used by many companies. Developing applications using the cloud is now faster and easier. One of the best features is that they include many application stacks and other resources used by developers. Concerns still exist with some customers related to compatability of applications developed in the traditional way and applications developed in the cloud. Cloud Service Providers are working on developing standards to ensure uniformity. From a business perspective, application development in PaaS has the potential to use cloud rather than using the traditional way.
2. **Collaboration Platform:** As this is an emerging category in computer software, colloboration or federation is gaining popularity in the research community. Communities however are seriously looking into colloboration provided that service providers show maturity. Any large company would have various departments writing a lot of reports and documents. Most of the time, the whole end product relies on the platform on which they work. Remember, this is not the application but the platform that enables the application. Enterprises use cloud collaboration platform to manage schedules, contact list, projects, reports, marketing materials, expense reports, budgets, financial statements, presentations, and so on.
3. **Database**: This can be synonymous with application plaftorm but is another ‘platform’ to have applications run on it. Cloud Service Providers provide various options for enterprise customers to run their applications on database engines. A large enterprise might look into this option only after thorough analysis and security assessment, but small and young companies already use databases based on cloud services to manage their data.

Infrastructure is one area where an organization would lose its ‘control’ if moved to cloud. While assessing the infrastructure, not all can be moved to the cloud but there would be some data in certain infrastructure that would perform better if moved to the cloud. Candidates include:

1. **Network:** Almost all organizations have their network services and operations outsourced to a cloud (network) service provider. This trend would further grow within the organization. Unlike any other potential asset, networking will prominently stay in the cloud.
2. **Data Backup:** Most organizations continue to back up their critical data to tape and vault them to a safe location. Considering the importance of data and the difficulty in managing them in the in-house data centers, backup to the cloud can play a key role in improving the organization's infrastructure. Some companies might not be able to use cloud services to back up their data due to regulations that mandate them to keep their data within their data centers.
3. **Compute services:** Servers that are required for development, testing, hosting CRM, intranet websites, and so forth can use cloud-based compute services. Companies need to invest in services important to their customers. Internal operations that depend on IT can be moved to cloud to save on capital expenses. As we have seen earlier, the cost of purchasing, installing, operating, and managing hardware (servers) only grows because of added operational expenses involved in power and cooling.
4. **Storage Services:** Most critical data (data of the organization's customers) can stay in-house, if regulations require. Meanwhile, data which need not be on expensive storage arrays should be moved to the cloud to reduce costs. Unlike any other hardware resource, storage once consumed can never be recycled easily. Companies have shown interest to move not all but some of their to cloud-based storage services. Data archiving in the cloud is also a solution in which a company can invest as that data is not frequently used.

2.10.2 Check Return on Investment and Total Cost of Ownership

Some assets are best kept in-house rather than in the cloud, but with the assessment completed in the previous step, companies should be ready to go further to the next stage.

After an organization successfully shortlists the assets, it is necessary to know if the *migration* will benefit in the long run. The tangible value of benefits? Profits. Profits that will result in successful movement must be calculated before making a decision to go to the next step. There are various additional factors that actually constitute the return of investment (ROI) which makes enterprises look beyond the features of the asset if it was moved to the cloud.

It will be difficult to move applications that are hosted on legacy platforms and which are no longer supported or which can no longer be upgraded, to the cloud. Some applications are

required to remain on old platforms/hardware; hence, the migration might turn out to be an engineering project which could last anywhere between several months to years. With that in mind, in the long run, the value of an asset if moved to the cloud needs to be thoroughly calculated.

Reduced operational expense is the most significant return that can expected. Calculations will include savings due to power and cooling of the hardware involved. The expense of managing the infrastructure, in-house support, and staffing must also be considered. Inflation, price rise, and expenses from the Cloud Service Provider must be noted as well.

2.10.3 Risk Assessment

Refer to the 'risks in cloud' section discussed previously and check how it is going to affect the environment. Some risks might not apply for a particular environment but security issues such as confidentiality, integrity, and availability must be checked thoroughly. Check strict regulations which are applicable to the organization. Regulations must not be compromised for saving costs as it will have serious consequences if not properly adhered to.

2.10.4 Selecting the Deployment Model

This phase will be the result of the conclusions made from the above assessment phases. The risk assessment of a selected asset would imply the possible cloud deployment model. For example, if the company assesses the lower risks and the ROI of having CRM, web hosting, and development platforms reside in the cloud, the company should conclude to use the public cloud. If the risk of placing (customer) data in the cloud increases for various reasons, private clouds must be used instead of public. Certain applications or services that are common to certain organizations will use community clouds.

2.10.5 Due diligence, feasibility, and audits

In some aspects, due diligence can be synonymous with assessment. During the assessment stage, various documents must be properly organized and managed which must be checked and re-checked to have an effective insight of the new services (Cloud) that is being procured. Auditing the Cloud Service Provider's track record for data center operations, security, and privacy, and asking to review such audits is important. Feasibility of the entire migration and the impact of moving to the cloud is one of the final stages of the assessment. In this step, the problems that might occur during the migration and after successful transition to the cloud are discussed and documented. Back-out plans and remediation steps required in case of a failure must be documented.

2.10.6 Identify and Calculate Costs

The final step in the assessment phase is to calculate the expenses involved for a company to leverage cloud services. Some cloud costs might actually be geared to deal with temporary spikes in compute load rather than moving an entire infrastructure out of the data center. The actual expenses would include:

- Migration costs – In cases of migrating an existing asset to the cloud.
- Decommission expenses – Costs involved in completely decommissioning the old asset.
- One time installation costs – Theoretically, nil. But there would be some expenses involved in the procurement of the resources (in the case of private clouds) and expenses for installing the cloud applications.
- Compute costs per month and/or costs throughout the life cycle.
- Some costs are hourly based. Calculate the costs that will incur throughout the lifecycle of the project.
- Scale-out expenses – It's the expenses that must be added in case the cloud infrastructure needs to be expanded for sudden or seasonal spikes.
- Other expenses.

2.11 Proof of concept and migration phase

The step of validating a new solution in an environment will vary among organizations. As mentioned before, some organization might not require performing all the phases mentioned above. Organizations that need to move their existing asset to the cloud have to do quite a lot of pre-checks but others have to see the benefits and savings on applying cloud services.

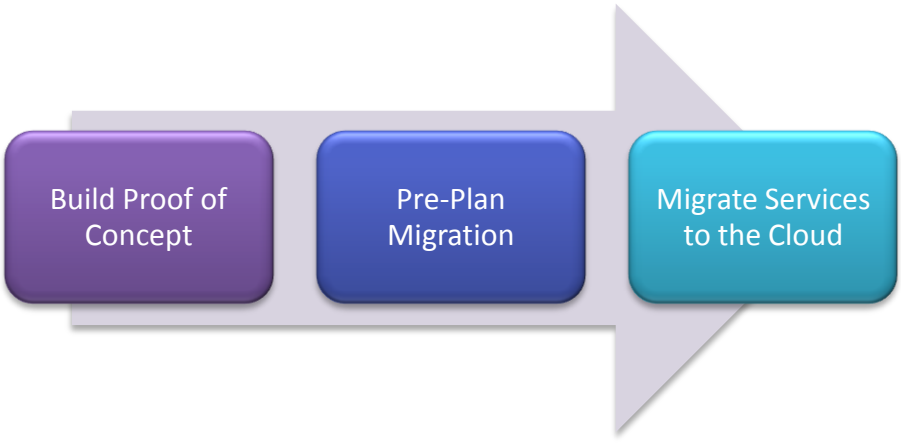


Figure 7 - Cloud Migration: Proof of Concept and Migration Phase

Unlike the assessment phase, once the organization has made a decision to migrate its assets on to the cloud, the steps involved will be fewer and straightforward. It has been assumed that during the final stages of assessment, the organization has shortlisted a few cloud vendors for the implementation.

2.11.1 Proof of Concept

One of the first steps is to check the functionality of the asset being moved to the cloud. Install the application/asset using the help of best practices provided by the service provider. Try to use the asset as used in a production environment. Performing various stress tests on the application is recommended.

For example, in a Web application migration, try developing models of all the components in the architecture (including database, web application, load balancer, web server) with minimal data. If other components that are required to develop the application are also cloud-based derivates (such as database, web application, and load balancer), various other testers are required to test them.

An organization would learn how to use cloud services as they would have a hands-on experience in this phase. Organizations would actually see how the cloud works in their environment and could help train their employees on using it as well.

2.11.2 Pre-plan and Migration

Enterprises at this stage will have a clear picture of what they are going to see post-migration to the cloud. Proper migration timelines need to be framed. A hybrid migration strategy can also be used where only a part of applications will be moved to the cloud and the rest will remain in-house. A hybrid approach is low-risk compared to a full migration. Rather than moving an entire asset at once, parts can be moved and optimized one at a time.

Architects need to ask the following questions:

- What amount of data/application/services are being moved?
- How much does the Cloud Serivce Provider support and at what scale will it grow?
- How many different teams will be involved?
- What is the back out plan?
- How long (in months) will the in-house asset/application be required after the migration?
- What is the time required for migration?

Appropriate teams must be contacted for advice, issues, and recommendations to the plans. Proper back out plans have to be developed in case the migration does not succeed.

Organizations not migrating but deploying a new cloud solution wouldn't have too many tasks to perform with respect to migration. Instead, they will have a straightforward approach to apply the solution after a successful proof of concept.

2.12 Leveraging the Cloud

After you have migrated your application to the cloud, run the necessary tests, and confirmed that everything is working as expected, it is advisable to invest time and resources to determine how to leverage additional benefits of the cloud. Questions that can be asked at this stage are:

- Now that I have migrated existing services, what else can I do to leverage the elasticity and scalability benefits that the cloud promises?
- What do I need to do differently to implement scalability to the assets?
- How can I take advantage of some of the other advanced features and services?
- How can I automate processes so it is easier to maintain and manage my assets in the cloud?
- What do I need to do specifically in my cloud application so that it can restore itself to original state in the event of a failure (hardware or software)?

2.13 Optimizing the Cloud

Especially with cloud computing which is improving and evolving every day, there is a scope of improvement and optimization in the organization that has to be applied. This is an evolving process in the organization. The cloud vendor must be involved in most of the technical decisions which would be taken.

- There might be slight performance degradation for the cloud computing services on some of the assets. The Cloud Service Provider must be made aware of any technical issues. Check the vendor's knowledge base.
- Automate manual processes.
- Improve cloud goverance – apply the same ITIL and COBIT frameworks.
- Review existing assets for migration to the cloud.
- Control costs by applying lean practices.

3. Efficient Data Centers

Although cloud computing appears to be the next big thing, data centers still exist and will certainly not disappear in the near future. Several large Enterprises have valid reasons to have their data in-house such as security, regulations, management, and reliability. As explained in the introduction, data centers are one of the highest producers of carbon emissions. Enterprises are constantly trying to keep up with regulations as well as competition in the market. The design and efficiency of a data center is very important today for corporate sustainability. We will see how various options available today help enterprises grow and sustain in today's highly competitive market.

Building an efficient, green data center requires increasing the utilization of existing and new IT resources, containing and improving power and cooling equipment use and requirements, reducing the data center's physical infrastructure footprint, reducing IT administrative and maintenance costs, and optimizing staff productivity.

The data center is an important asset for large and small businesses alike. As investment in a data center is very high, the expectation from it is very high as well. Data centers need to deliver more than what they used to deliver several years ago. Some business expectations are:

- High availability
- Reduced Total Cost of Ownership
- Increased Return on Investment
- Easy maintenance and administration
- Lower operational costs
- Lower electricity costs – less power consumption
- Business Continuity
- Disaster Recovery

In short, the three important objectives of a business are:

- Increased Revenue
- Reduced Cost
- Better utilization of assets

3.1 Time to transform the data center

Transforming an existing data center is not a decision that companies make overnight. The problems that pile up due to lack of processes, standards, and increasing costs make IT

managers realize the need for a strategic plan to ensure efficiency and sustainability. An enterprise will know that it is time to transform the data center when it experiences problems such as complexity, cost, and capacity:

- Complexity
  - Lack of standards – No proper frameworks, goverance, and IT management.
  - Large amount of non-shared assets – More expensive resources with poor efficiency.
  - Increasing business needs – Company expansion, data growth, and need for availability.
  - Proliferation of infrastructure – Tougher to manage IT assets.
- Cost
  - Increasing energy costs as discussed in previous section.
  - Underutilization of resources – Lack of virtualization.
  - Real estate and technology management.
- Capacity
  - Data center space constraints.
  - Difficulty supporting business innovation.
  - Performance issues.

Failure to adopt modular standardization as a design strategy for Network-Critical Physical Infrastructure (NCPI) is costly on all fronts: unnecessary expense, avoidable downtime, and lost business opportunity.

3.2 Considerations for Efficient Data Centers

Transforming an existing data center and building a new data center are two different complex projects. In this section we will discuss the key considerations while designing and transforming a data center. Some generic considerations are:

- **Location:** Data centers need not be close to the location of the business. Instead, they should be at a location which is safe and secure. Safe from natural disasters such as earthquake, tornado, tsumai, and so forth. Keeping the rising cost of real estate in mind, a location must be selected where there is sufficient electricity and network connectivity.
- **Disaster Recovery Site:** A geographically remote location from the primary data center must be chosed to ensure availability of services in case of a catastrophic disaster in the primary data center. Select a site with different geographical characteristics than the primary site.

The crucial parts of the data center are the physical infrastructure. The important elements are:

- **Power:** Elements that make up the power infrastructure include the electrical service entrance of the building, main distribution, generator(s), UPS systems and batteries, surge protection, transformers, distribution panels, and circuit breakers.
- **Cooling:** Cooling systems required to successfully remove heat from the data center include computer room air conditioners (CRACs) and their associated subsystems—chillers, cooling towers, condensers, ductwork, pump packages, piping—and any rack- or row-level cooling or air distribution devices.
- **Racks and physical structure:** There are many systems that could be considered part of the physical structure of a data center. The most critical elements are the IT racks housing the IT equipment, plus physical room elements such as dropped ceilings and floors (both raised floors and concrete "slab" floors).
- **Management:** Management spans all facilities elements. To have reliable facilities, it is important to have visibility to all of the components of the physical infrastructure. Management includes systems such as building management systems (BMS), network management systems (NMS), element managers, and other monitoring hardware and software.
- **Cabling:** The cabling infrastructure encompasses all data cables that are part of the data center as well as the power cables necessary to deliver power to all of the loads.
- **Security & fire protection:** Subsystems included here are physical security devices at the room and rack level and fire detection/suppression systems.

All of these components must be integrated into a seamless end-to-end system supported by a system-wide management system to ensure proper data center operation and management.

3.3 Data Center Tiers

The TIA-942: Data Center Standards Overview[21] describes the requirements for the data center infrastructure. The simplest is a Tier 1 data center, which is basically a server room, following basic guidelines for the installation of computer systems. The most stringent level is a Tier 4 data center, which is designed to host mission-critical computer systems, with fully redundant subsystems and compartmentalized security zones controlled by biometric access control methods. Another consideration is the placement of the data center in a subterranean context, for data security as well as environmental considerations such as cooling requirements.

[21] "TIA-942 Data Center Standards Overview" http://www.adc.com/Attachment/1270711929361/102264AE.pdf

The four levels are defined and copyrighted by the Uptime Institute, a Santa Fe, New Mexico-based think tank and professional services organization. The levels describe the availability of data from the hardware at a location. The higher the tier, the greater the accessibility. The levels are:

Tier	Requirement
1 – Basic: 99.671% Availability	▪ Susceptible to disruptions from both planned and unplanned activity. ▪ Single path for power and cooling distribution, no redundant components. ▪ May or may not have a raised floor, UPS, or generator. ▪ Takes three months to implement. ▪ Annual downtime of 28.8 hours. ▪ Must be shut down completely for perform preventivbe maintenance.
2 – Redundant Components: 99.741% Availability	▪ Less susceptible to disruption from both planned and unplanned activity. ▪ Single path for power and cooling disruption, includes redundant components(N+1). ▪ Includes raised floor, UPS, and generator. ▪ Takes three to six monts to implement. ▪ Annual downtime of 22.0 hours ▪ Maintenance of power path and other parts of the infrastrucutre require a processing shutdown.
3 – Concurrently Maintainable: 99.982% Availability	▪ Enables planned activity without disrupting computer hardware operation, but unplanned events will still cause disruptiuon. ▪ Multiple power and cooling distribution paths but with only one path active, includes redundant components (N+1). ▪ Takes 15 to 20 months to implement. ▪ Annual Downtime of 1.6 hours. ▪ Includes raised floor and sufficient capacity and distribution to carry load on one path while performing maintenance on the other.
4 – Fault Tolerance: 99.999% Availability	▪ Planned activity does not disrupt critical load and data center can sustain at least one worst-case unplanned event with no critical impact. ▪ Multiple active power and cooling distribution paths, includes redundant components (2 (N+1), i.e 2 UPS each with N+1 redundancy). ▪ Takes 15 to 20 months to implement. ▪ Annual Downtime of 0.4 hours.

3.4 Data Center Consolidation

No data center is perfect and identical; each have its own uniqueness and distinguished features. Some parameters that are required to identify each environment are:

Checklist:

- Number of data centers.
- Data center Environment (size)
    - Closet (1-2 racks)
    - Computer or communications room (3-10 racks/cabinets)
    - Mid-size data center, computer or communications room (10-100 racks)
    - Large Data Center (100+ racks)
- Data center Environment (location)
    - Disaster Recovery (DR)
    - Redundant computer/data center
    - Main data center
    - Regular office space
- Floor size (Sq/ft)
- Ceiling type (Dropped(false)/none)
- Floor (raised floor/hard floor)
- Server count in the room(s), in the network(s)
- Switch count in the room(s), in the network(s)
- Rack count in the room(s), in the network(s)
- Number of desktops in the network(s)
- Laptops in the network(s)
- Disaster Recovery site capability
- Number of single point of failure devices
- Cooling systems (CRAC and design)
- Maintenance, monitoring, management, and support capability
- Data center audits/certification (based on various standards such as SS507, ISO/IEC-24762, ANSI/TIA-942, ISO/IEC20000, etc.)

Large enterprises with multiple data centers across geographies must look into data center consolidation immediately. Apart from cost savings there are other benefits that can be realized by consolidating the infrastructure in the data centers. Key benefits of data center consolidation are:

Activity	Objectives	Savings
Physical Consolidation	▪ Reduction in the number of equipment ▪ Reduction in the number of data center sites	▪ Facility Costs ▪ Equipment (Hardware and Software) Cost ▪ Operational expenses (Administration, Staffing, etc.)
Equipment Standardization	▪ Replacement of existing (legacy) hardware platforms with a selected standard	▪ Hardware costs ▪ Hardware maintenance ▪ Administration costs
Server, Storage, and Data base consolidation	▪ Optimize service applications per service platform ▪ Replace many low-end servers with high-end servers ▪ Share backup, replication, and data recovery platforms	▪ Save on number of servers ▪ Save costs on number of storage/backup devices ▪ Hardware maintenance ▪ Storage and backup staff
Application Consolidation	▪ Standardization of applications used across the business	▪ Number of servers ▪ Software maintenance

Table 1 – Key Benefits of Data Center Consolidation

3.5 Equipment Placement and Cooling

Equipment location in the data center plays a crucial role in power consumption. Not all hardware equipment has the same power consumption and cooling requirements. Designers need to find out which racks dissipate higher heat and place them in only certain locations. Uneven scattering of equipment will increase power consumption due to increased cooling requirements.

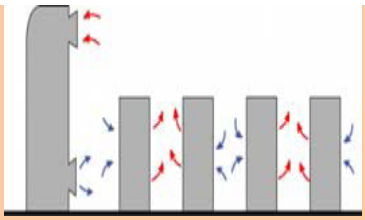
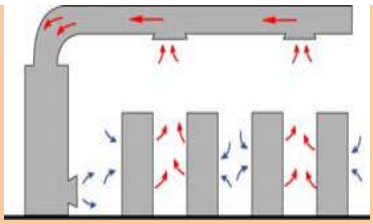
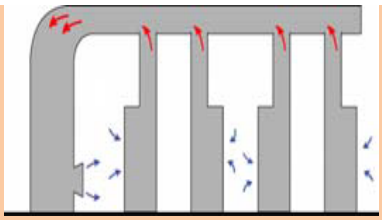
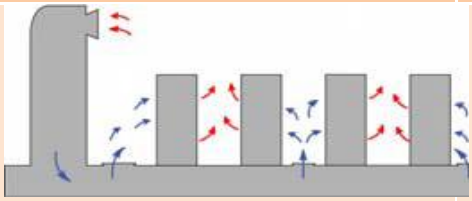
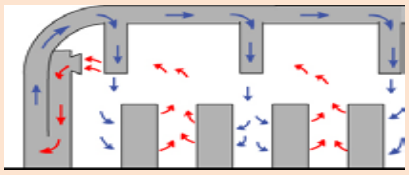
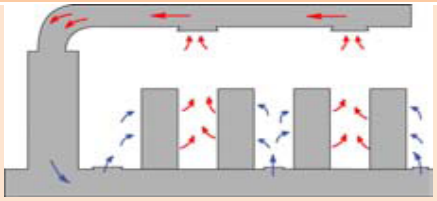
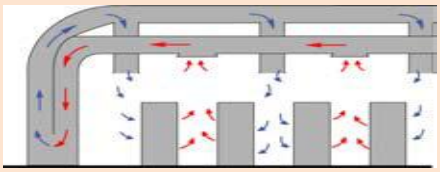
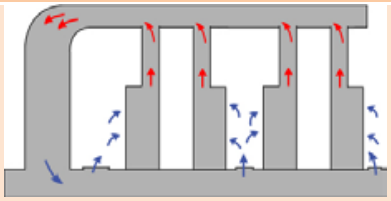
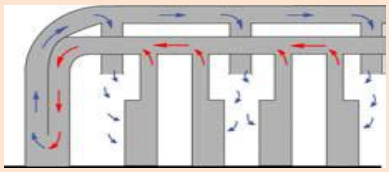
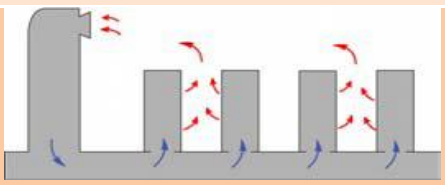
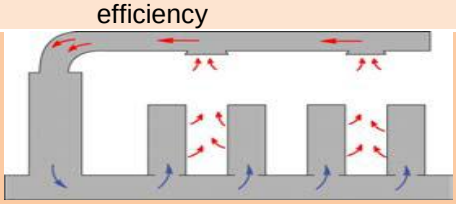
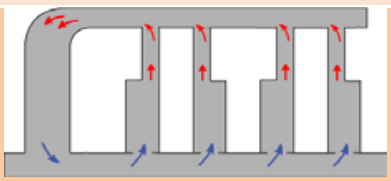
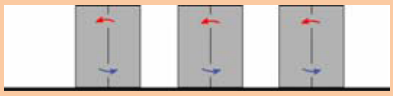
	Flooded Return	Locally Ducted Return	Fully Ducted Return
Flooded Supply	 **Small LAN rooms < 40kW (Hard Floor)** ▪ Simple installation ▪ Low cost ▪ Cools up to 3kW per rack	 **General use** ▪ Cools racks to 3kW ▪ No raised floor needed ▪ Low cost / ease of install	 **Hot rack problem solver** ▪ Cools racks to 8kW ▪ Retrofittable (vendor specific) ▪ No raised floor needed ▪ Increased CRAC efficiencies
Locally Ducted Supply	 Raised Floor environments  Hard Floor environments **General use** ▪ Cools racks to 3kW	 Raised Floor environments  Hard Floor environments **General use** ▪ Cools racks to 5kW ▪ High performance/High efficiency	 Raised Floor environments  Hard Floor environments **Hot rack problem solver** ▪ Cools racks to 8kW ▪ Retrofittable (vendor specific)
Fully Ducted Supply	 **General use** ▪ Enclosures/mainframes with vertical airflow ▪ Raised floor environments with poor static pressure	 **General use: mainframes** ▪ Enclosures / mainframes with vertical airflow ▪ Raised floor environments with poor static pressure	 Raised Floor environments  Hard floor environments **Hot rack problem solver** ▪ Cools racks up to 15kW specialized installation

Table 2 - The nine types of cooling systems[22]

[22] "Air Distribution Architecture Options for Mission Critical Facilities" By Neil Rasmussen, APC White Paper #55

Separating the racks which dissipate higher heat with lower dissipating racks will reduce hot spots in the data center. All types of devices do not dissipate heat in the same directions–Front-to-Back and Side-to-Side. Hardware devices have to be sorted according to their:

- Hardware type – Rack mount Servers, Blade Servers
- Hardware type contd. – Storage, Mainframe
- Hardware type contd. – Switches, Routers
- Power Requirement/Consumption
- Cooling Requirement
- Cooling system direction – Front-to-Back or Side-to-Side

Now, that you have the hardware sorted, determine the existing data center type or the data center currently being developed. Use the check list given above. Ensure the data is as correct as possible.

The three possibilities of supply and return cooling distribution systems are Flooded, Locally Ducted, and Fully Ducted. Each one can be the supply or return or a combination of both in the data center cooling system. Hence, there can be nine types of cooling systems as shown in Table 2. In general, the cost and complexity of the cooling system is lowest at the top and left of Table 2, and increases for the types that are down and to the right as the complexity of the ducting system increases.

In most cases, the preferred method for building data centers is to use a hard floor. Contrary to popular belief, cooling methods for hard floor installations can provide the same or better capabilities and performance as the raised floor. Figure 9 shows a properly planned data center. The goal while planning one should be to separate the hot and cold air effectively and reduce the cooling requirements.

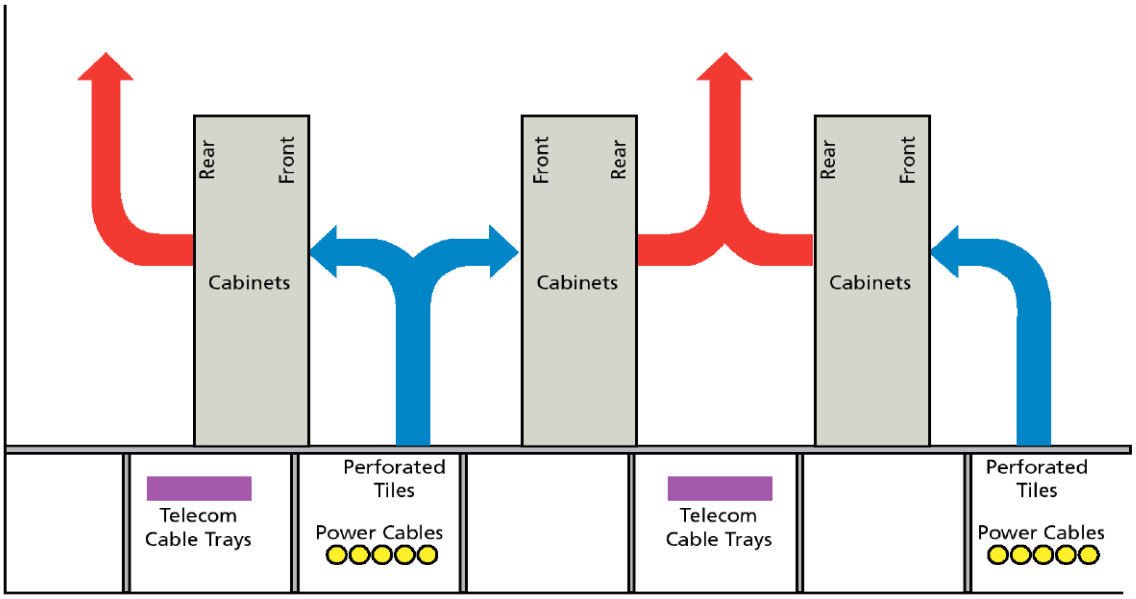


Figure 8 - Illustration of a planned data center with optimal equipment placement and hot aisle/cold aisle cooling

In general, Fully Ducted supply or Fully Ducted return is used for providing cooling to enclosures operating at power levels in the range of 5-15kW. Since enclosures drawing 5-15kW typically represent a small fraction of the enclosures in a data center, this method is typically used in combination with simpler methods. The use of Fully Ducted design applied only when and where needed allows data centers to be designed to average heat load but still be able to handle high density enclosures when needed.

3.5.1 Airflow: Front to Back and Side to Side

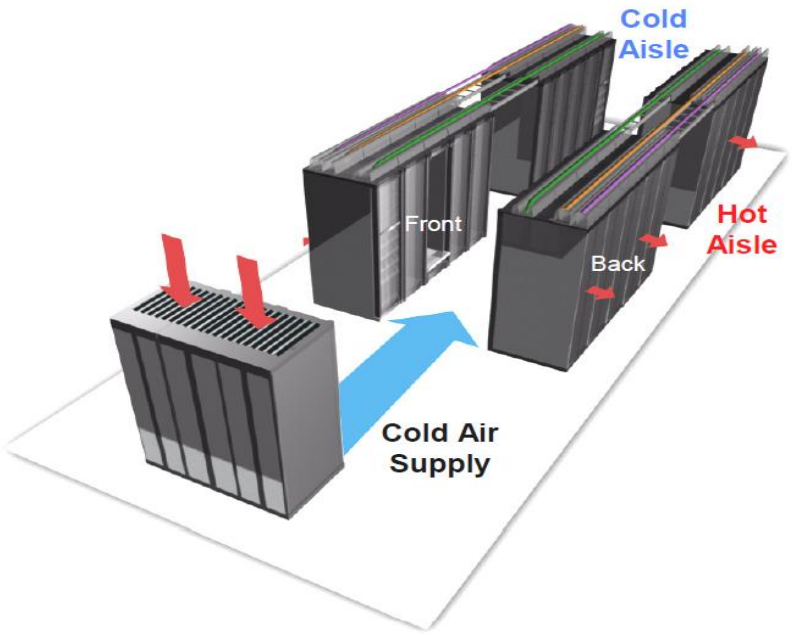


Figure 9 - Typical Hot aisle/Cold aisle cooling method

There are two directions that air flows through the rack. Figure 8 shows a typical data center using a hot and cold aisle cooling method. The vast majority of servers and rack mount storage use front-to-back air flow. Hot and cold air is separated to enhance cooling efficiency. However, many types of switches and routers are constrained by their design and require side-to-side air flow.

Regardless of the equipment's air flow design, it is essential that an adequate amount of cool air is received. If this cool air supply is not met, the availability of equipment and business processes they support will suffer since the life of an electronic device is directly related to the operating temperature. Based on various factors, the side-to-side cooling method varies depending on the environment.

The recommended methods for side-to-side cooling are:

1. **Side air distribution:** Requires vertical rack space for air turning device. Floor space is optimized by increasing power density per rack. Higher efficiency. Lower cooling cost, fewer racks required, and optimized floor space. Recommended for high density environments.
2. **Supplement fans**: Better floor space utilization. Better efficieny based on fan location. Higher costs. Recommended for hot aisle/cold aisle enclousures.

3. **Low density enclosure:** Consumes more floor space, Higher efficiency, lower cooling costs, higher cost per rack, less equipment per rack requiring more racks for equipment. Recommnded for isolated racks in low density environments.
4. **Open frame racks with increased inter-rack spacing:** Consumes more floor space, Higher cooling costs. Low cost per rack, less equipment per rack requiring more racks for equipment. Recommended for isolated racks in a low density environment.

3.6 Optimizing the Core infrastructure

The core infrastructure of a data center is the physical hardware–servers, storage, and network equipment. These are the resources that serve the purpose of the data center. Selecting and managing the equipment properly is the key driving factor for an efficient data center.

Equipment must be selected with a balance of performance and cost. The cost does not imply the equipment hardware cost. There will also be recurring costs based on the operational expenses of the hardware which need to be considered. Rising power and cooling costs require the hardware to be used efficiently by leveraging virtualization and consolidation.

3.7 Server virtualization and consolidation

The economic crisis and a need to "do more with less" drive more businesses toward virtualization. As adoption increases, enthusiastic supporters generate buzz about the technology, which drives even more people toward it. A strong virtualization market benefits everyone by spurring innovation and making the technology more powerful and easier to use. Server virtualization still remains one of the best ways to effectively utilize data center hardware. Important reasons why data centers are seriously considering virtualization are:

- Greater application availability
- Faster/simpler server management
- Faster/simpler DR and Business Continuity
- Server consolidation

Virtualization and consolidation work together toward one common goal–reduce cost and increase efficiency. The business case to justify server consolidation is:

- Reduced:
    - Servers
    - NICs and/or HBAs
    - Cables

- Switch ports
- Switches
- Rack space
- Floor space
- Power
- Cooling
- CapEx and OpEx

Apart from these features there are other aspects of virtualization that have made it an important tool in data center transformation. The advantages and disadvantages are:

Advantages of Virtualization

- **Facilities management**
  - Saving data center space
  - Hardware cost saving
  - Reduced energy bills
- **Security/Business Continuity**
  - Easy to back up the complete image
  - Disaster Recovery/Business Continuity
  - Virtual appliance might enhance security
  - Virtual desktops provide enhanced control over security
- **Resource management**
  - Simplifies chargeback systems
  - Optimizes use of existing hardware resources
  - Faster deployment of new logical servers
  - Moving logical servers between hardware
  - More flexible infrastructure
- **Management**
  - More, smaller applications logically separated from each other
  - Fewer servers
  - Eases hardware and software maintenance

Disadvantages of Virtualization

- **Security/Risks**
  - Internal Resistance – Application owners resisting control over their existing hardware.

- New, relatively unproven technology, tools, and processes – The impact and implications of the (new) changes due to introduction of the technology must be clearly understood in the environment.
- Re-organization/consolidation of servers and applications.
- Loss of logical servers – Unless the images are well managed, there is always a possibility of accidental deletion.
- Consolidated data centers – Although the advantages outweigh the disadvantages, it must be noted that it creates a single point of failure unless there is a proper Business Continuity/Disaster Recovery plan. The consolidated data center must have adequate supporting infrastructure to handle the workloads of multiple data centers.
- Real time or near real time requirements: Time lag of 5-10 seconds can be seen if the Virtual Machine is heavily utilized. Not a big problem, but can be a concern if the system requires real-time response[23].

- **Performance**
  - Resource consumption overhead
  - Resource allocation
  - Bottlenecks/Queuing delays
    - CPU
    - Memory constraints
    - Network constraints

    Most of these constraints can be managed by adding more resources and/or reorganizing which virtual servers are running on which hosts.

- **Management**
  - Support – Although the ability to isolate applications through virtualized appliances can simplify support, the disadvantage is that many software vendors do not fully support virtualized environments or else require that the issue be reproduced in a non-virtualized environment.
  - Multiple servers on one physical server – Having multiple virtual machines on a single physical machine can introduce the potential for security vulnerabilities.
  - Increased failure impact – a single HW failure now affects multiple logical/virtual servers, any reboot/maintenance to the host system interrupts all guest OS, however this can be

[23] Further reading on time delays: http://www.vmware.com/resources/techresources/238 and the VMware KB1420 concerning "Linux Guest Timing" http://kb.vmware.com/selfservice/microsites/search.do?cmd=displayKC&externalId=1420

managed using “hot migration” technologies (such as VMotion or Live Migration), bottlenecks like CPU cycles, memory capacity, and network.

- Specialized skill set is required.

3.7.1 Virtualization Considerations

- **Identifying the *right* servers to virtualize:** Servers which can be virtulized include: Test and development servers, servers requiring low performance – lightly used servers, quality assurance, software support where it’s important to be able to quickly and easily reproduce a relatively large number of environments, demo scenarios, and servers with predictable resource consumption profiles. This will help in planning the distribution of work for virtualized servers.
- **Identifying the servers which should *not* be virtualized:** Although all servers are eligible for virtualization, there are certain situations where the potential risks far outweigh any advantages that might be gained. A well known example is the database servers which rarely have any advantages if virtualized. Database server utilization is better improved by employing multiple database instances or an application virtualization-type server such as Citrix that already have virtualization involved in it. There are other examples which vary by environment.
- **Follow vendor’s best practices:** It is highly recommended that you carefully read the performance tuning and best practices documents and white papers that are available from host system vendors. For example, the “VMware tuning Best Practices for ESX Server 3” document, which is available from VMware (http://www.vmware.com/pdf/vi_performance_tuning.pdf) focuses primarily on VMware’s ESX server. However, many of its performance tips are applicable to other solutions as well.
- **Review in-house applications:** Poorly written applications often consume multiple resources. Therefore, whenever possible, review in-house applications to determine if they can be enhanced to better use available resources. For example, a database engine given poorly written SQL will likely put unnecessary stress on CPU, network, disk I/O, and memory. CA CEM and Introscope can rapidly identify resource bottlenecks in web applications

3.8 I/O Virtualization

3.8.1 Fabric Sprawl

Networks are the nervous system in the data center. They interconnect every component in the data center including the power and CRAC systems. Servers use the network links the most for three important needs–local area network (LAN), storage area network (SAN), and inter-process communications. The total Ethernet and Fibre Channel cables and associated switch ports exceed more than four per server in most of the environments. The investment required for the Network Interface Cards (NICs), Host Bus Adapters (HBAs), switch, and cables is costly in large environments. Management complexity caused due to this increase in networking components involved is referred to as Fabric Sprawl. Using various I/O virtualization technologies, it is now possible to reduce the number of networking components, solving Fabric Sprawl.

3.8.2 An Introduction to I/O Virtualization

Input/output (I/O) virtualization is a methodology to simplify management, lower costs, and improve server performance in enterprise environments. I/O virtualization environments are created by abstracting the upper layer protocols from the physical connections.

The technology enables one physical adapter card to appear as multiple virtual network interface cards (vNICs) and virtual host bus adapters (vHBAs). Virtual NICs and HBAs function as conventional NICs and HBAs, and are designed to be compatible with existing operating systems, hypervisors, and applications. They appear as normal cards to networking resources (LANs and SANs).

In the physical view, virtual I/O replaces a server's multiple I/O cables with a single cable that provides a shared transport for all network and storage connections. That cable (or commonly two cables for redundancy) connects to an external device, which then provides connections to the data center networks

Small data centers with moderate performance requirements should continue to use iSCSI storage for easy management. Large data centers using different protocols for storage and local area network must seriously consider I/O virtualization to improve costs and management. Data centers which are newly being built should do an analysis for which virtualization to choose.

There are three options for I/O virtualization in the data center; InfiniBand (IBA), Converged Enhanced Ethernet (CEE), and Multi-root Input/Output Virtualization (MRIOV).

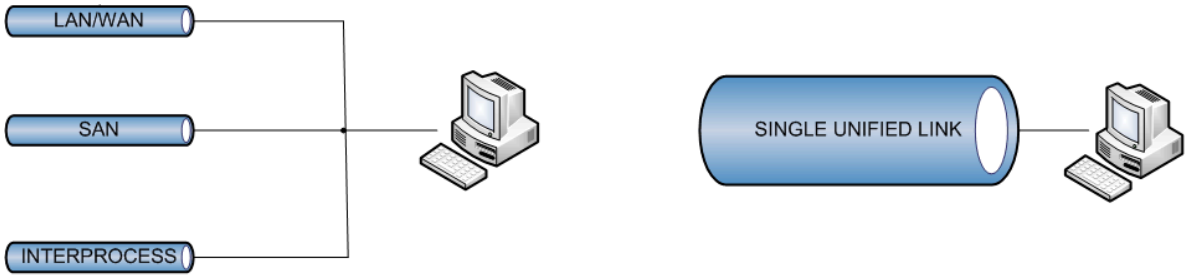


Figure 10 - Virtualized I/O in a Single Link against multiple separate links[24]

3.8.3 InfiniBand (IBA)

InfiniBand is a switched fabric communications link for high performance computing and enterprise data centers. Designed to be scalable, its features include high throughput, low latency, quality of service (QoS), and failover. The InfiniBand architecture specification defines a connection between processor nodes and high performance I/O nodes such as storage devices.

3.8.3.1 InfiniBand Terms

- **HCA –** Host Channel Adapter, an adapter similar to HBA
- **TCA –** Target Channel Adapter, storage or gateway adapater card
- **Shared I/O Gateway –** Same as I/O Virtualization, IBA to IP, FC, iSCSI gateway
- **RDMA –** Remote Direct Memory Access, lowest latency memory-to-memory transfers
- **HPCC –** High Performance Compute Clusters**,** large nodal clusters
- **IBA Director –** Large port count five 9s switch, 288 to 864 port switches

3.8.3.2 InfiniBand Suppliers

- **Cisco:** Switches, Directors, and HCAs
- **Mellanox:** Switches, Directors, and HCAs
- **Qlogic:** Switches, Directors, and HCAs
- **Voltaire:** Switches, Directors, HCAs, and end-to-end full management
- **Xsigo:** Switches and HCAs

3.8.3.3 InfiniBand Implementation

In recent years, InfiniBand has been increasingly adopted in Enterprise data centers, for example Oracle Exadata and Exalogic Machines, financial sectors, cloud computing (an InfiniBand-based system won the best of VMWorld for Cloud Computing) and more. InfiniBand has been mostly used for high performance clustering computer cluster applications. A number

[24] This concept of aggregating the IO is used in Data Center Bridging (DCE/CEE). InfiniBand and MRIOV are also similar concepts.

of the TOP500 supercomputers have used InfiniBand including the former[25] reigning fastest supercomputer, the IBM Roadrunner. In another example of InfiniBand use within high performance computing, the Cray XD1 uses built-in Mellanox InfiniBand switches to create a fabric between HyperTransport-connected Opteron-based compute nodes.

Usage: High Performance Computing and Enterprise Data centers

3.8.4 Data Center Bridging

Many computer communications improvements have occurred in the data center. Ethernet, the primary network protocol, is designed to be a best-effort network protocol that may drop packets when the network or devices are busy. Further, the complex nature of Transmission Control Protocol (TCP) causes CPU overhead, which can impact performance, Hence, a new standard has been developed that either extends the existing set of Ethernet protocols or emulate the connectivity offered by Ethernet protocols.

Data Center Bridging is the name used by IEEE standards group for enhanced/converged/data center Ethernet. This is the next-generation Ethernet standard that aggregates the various input/output communication protocols into a *reliable lossless* Ethernet channel.

Other Names:

- **DCE (Data Centre Ethernet):** Cisco's trademarked term for their implementation
- **CEE (Convergence Enhanced Ethernet):** IBM's trademarked term for their implementation

Terminology:

- **FCoE –** Fibre Channel over Ethernet, FC frames encapsulated in Ethernet packets
- **iSCSI –** SCSI mapped in TCP/IP
- **iWARP –** RDMA on Ethernet, required for HPC Clusters
- **CNA –** Converged Network Adapaters, Concurrent FCoE, iSCSI, iWARP, and TCP/IP on 10GbE NIC
- **10G TOE –** 10G TCP Offload engine

[25] "Two rival supercomputers duke it out for top spot" - Stephen Lawson, http://news.idg.no/cw/art.cfm?id=FB70C2C5-1A64-6A71-CEEA6C17D51B1E3C

3.8.4.1 FCoE

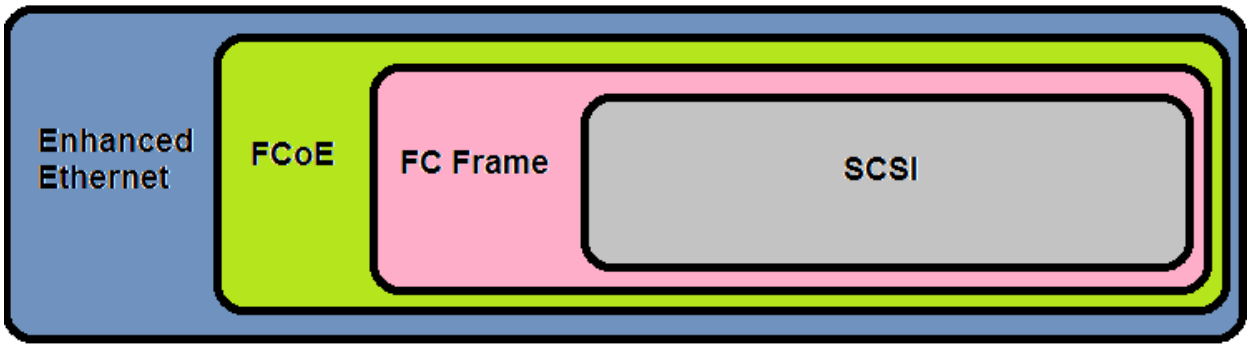


Figure 11 - FCoE Frame

- Layers FC frames directly over Ethernet (requires jumbo/mini-jumbo)
- Replacing lowest level of FC with Ethernet, T11 FC-BB-5 standard (Fibre Channel standards group)
- Requires convergence enhancements to Ethernet
    - Max size of FC packets requires Ethernet Jumbo (or mini-jumbo) frames
    - FC-like no-drop behaviour in face of congestion
        - Requires new “Priority-based Flow control” protocol
    - To control traffic interferences
        - Requires new “Enhanced Transmission Selection” protocol
    - To verify the protocol support across the end points
        - Requires new “Data Center Bridging exchange” protocol

Adaptation

The main application of FCoE is in data center SANs. FCoE has particular application in data centers due to the cabling reduction it makes possible, as well as in server virtualization applications, which often require many physical I/O connections per server.

With FCoE, network (IP) and storage (SAN) data traffic can be consolidated using a single network. This consolidation can:

- Reduce the number of network interface cards required to connect to disparate storage and IP networks
- Reduce the number of cables and switches
- Reduce power and cooling costs

DCB Suppliers: Cisco, Brocade, Emulex, Qlogic, Mellanox

3.8.5 Multi-Root Input/Output Virtualization

A new technology that is receiving growing interest among technology enthusiasts on I/O virtualization is based on PCI express. Multi-Root I/O Virtualization is a new high performance, low-latency, low-cost solution that helps transform the networking of the data center at a much lower cost than InfiniBand and DCB.

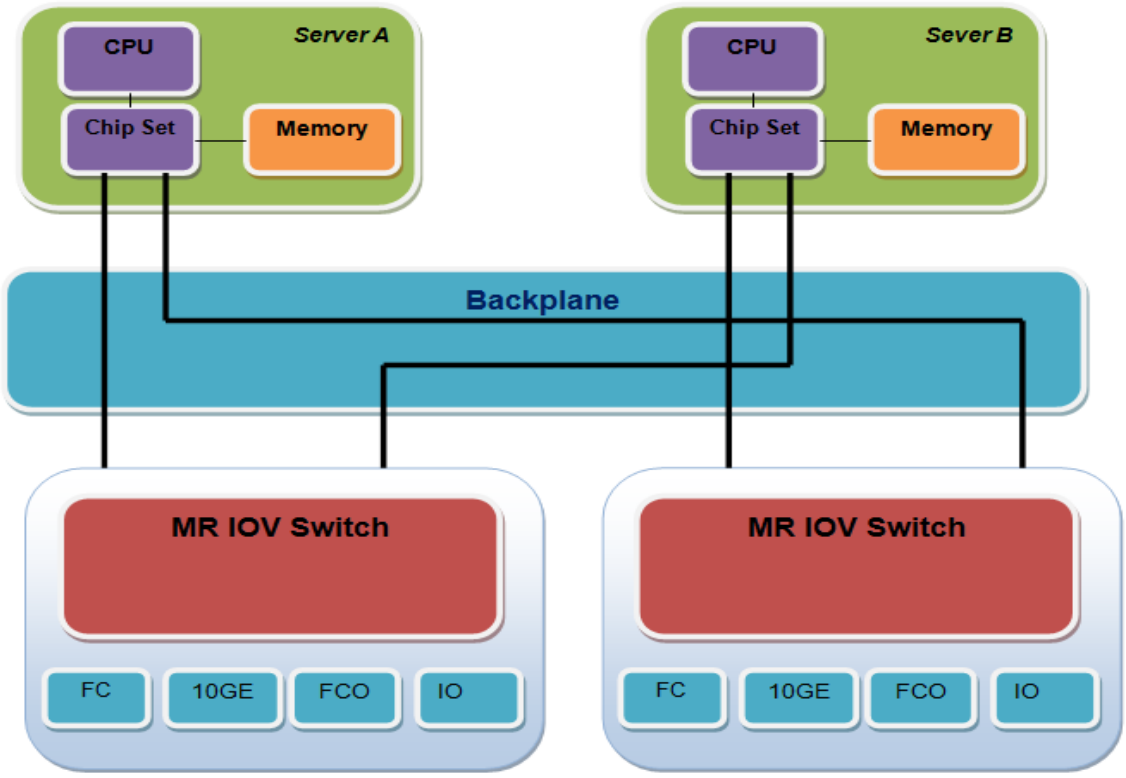


Figure 12 - Blade Servers and MR IOV

Each server using a single PCI Express (PCIe) cable connects directly to the server's native PCIe bus. This removes the need to deploy any new network in the data center and reduces I/O power consumption, capital equipment costs, number of cables, server management complexity, and operational expenses by more than 60 percent. This helps consolidate, virtualize, and share the major types of server networking and storage connectivity including Ethernet, Fibre Channel, FCoE, iSCSI, and SAS/SATA and results in providing the industry's best I/O price/performance and lowest energy consumption for deploying network and storage connectivity to servers.

Features

- Simpler
  - Fewer storage and network points and switches
- Shared I/O – Higher Utilization
  - 75 percent I/O cost reduction
  - 60 percent I/O power reduction
  - 35 percent rack space reduction
- Scalable I/O Bandwidth
  - On demand 1-40Gbps
  - No additional cost
  - Reduced I/O adapaters and cables
- Enhanced Functionality
  - Shared memory
  - IPC and PCIe speeds up to 40Gbps
- Reduced Operational Expenses
  - Simplified management
  - Server, OS, application, network, and switch transparent
  - High availability
  - Changes without physical touch
- Reduced Capital Expenses
  - Smaller, denser server
  - Fewer components
  - Fewer failures
  - Fewer opportunities for human error

Bottom line: MRIOV is flexible, providing high I/O utilization using standardized, open technologies (PCIe) at low cost.

MR IOV Suppliers: Aprius, Virentsys

3.8.6 I/O Virtualization Comparison

	Present situation (No Virtualization)	MRIOV (PCI Express)	InfiniBand Solutions	FCoE Solutions
Utilization	Very Low ~ 15%	**Very High ~80%**	Low	OK
Reliability	Neutral	**High**	-	OK
IO Performance	10Gb	**80Gb**	20Gb	10Gb
TCO	High	**Low**	High	High
Management	Poor	**Best**	Best	OK
I/O Cost	High	**Low**	High	High

Table 3 - Comparison of I/O Virtualization

Issues with I/O Virtualization

InfiniBand: InfiniBand is primarily utilized for HPC, Not a large install base in Enterprise data centers. Only a few storage vendors have native IBA interfaces.

DCB: Technology is new and initial investment is expensive. Requires new adapters, switches, and cables to be effective.

MRIOV: Applies only in rack and blade system (PCI), very few vendors, immature, all required hardware must be provided by the server vendors.

Recommendations: Current data center networking is difficult to manage and the operational expenses involved are very high. I/O virtualization has the potential to improve network performance and ease management. Enterprises have to decide which type of I/O virtualization is suitable for their environment. Small but new data centers must consider MRIOV as their networking technology. Until MRIOV matures, FCoE is the alternative for Enterprise data centers unless the company requires high performance computing for which InfiniBand will be suitable.

3.9 Optimizing Storage

Contrary to the previous belief where storage once used cannot be used again, technologies today have improved wherein we can effectively utilize storage, reducing cost. Although not implemented in every data center, they have the potential to efficiently utilize the storage resource. With digitization and data growth, storage continues to grow exponentially. Although the growth cannot be controlled, enterprises must look into alternatives that will help reduce the

amount of physical space required to store the data. Thankfully, there are technologies that will help achieve this and more. A few of them are mentioned below:

- Storage Virtualization
- Thin Provisioning
- Deduplication
- Automated Storage Tiering
- Intelligent Power Management
- Solid State Devices
- Writable Snapshots

Before implementing or transforming an existing storage estate, a set of questions need to be answered, similar to the data center consolidation checklist. This will help to assess the infrastructure and understand the specific needs of the environment.

3.9.1 Requirement Gathering: Storage Checklist

- Type of business – Banking, Insurance, Manufaturing, Retail, Technology, etc.
- Critical applications and expected performance
- Number of data centers with an Enterprise storage array
  - Storage vendors in each data center
  - Number of years since the last refresh cycle
  - Connectivity between the data centers – Multiple locations with no connectivity or all interconnected data centers
- Number of data centers that exceed the storage usage or more than half a Petabyte
- Distance between the data centers
- Business Continuity and Disaster Recovery requirements
- Recovery Time Objective (RTO) and Recovery Point Objective (RPO)
- Replication requirements
- Storage growth requirements
  - Bandwidth requirements
- Data retention requirements
  - Rentention period
  - Recovery time (based on RTO) – Use of a VTL
  - Tape Vaulting

3.9.2 Storage Virtualization

Unlike server virtualization where the hardware is sliced to emulate 'sliced' operating systems, storage virtualization is the aggregation of various storage arrays as one storage array. Data centers using multiple storage arrays of different vendors must consider storage virtualization to reduce complexity and ease manageability.

3.9.2.1 Storage Virtualization recommendations

- Select the virtualization appliance that has features required for your environment.
  - Example: Consider an environment having 2 Symmetrix® DMX-4, 2 HDS USP, 1 IBM DS4300, and 2 NetApp 3040. Manging various storage arrays through various software tools and multi-pathing softwares is going to be a dauting task. Storage virtualization would help in this case where the virtualization appliance abstracts storage from all the storage arrays as if they were from a single array and helps ease storage management. The best appliance to be used can vary with the requirements of the data center. Proper analysis of the appliance is recommended before implementing.
- Not all environments require storage virtualization.
  - Example: Consider an environment having 4 Symmetrix DMX-3 and 1 Hitachi AMS 2000 used for archiving data and 2 IBM DS4300 for development and testing. Manageability of such an environment is best without adding a virtualization layer. While the Symmetrix is used for production applications, the growth and requirements to manage in AMS and the DS4300 will be minimal.
- Consider using when multiple storage vendors are used in your environment.
  - Storage vendors might endorse the features provided by the appliance over the features built into the primary storage array. Although the features supersede the requirement, it might not be worth it to invest in the appliance. Instead, maybe upgrade the storage array to a better one.

Storage Virtualization Products: EMC Invista™, IBM SVC, NetApp V-Series, Hitachi USPV/VM, BlueArc Titan/Mercury, HP XP (OEM of Hitachi), XIOTech ISE Age, etc.

3.9.2.2 Thin Provisioning

Thin Provisioning is a storage virtualization feature that emulates logical units (LUNs) to appear as a large capacity but actually use only a small portion of the physical disk. This helps optimize utilization of available resources.

Storage Pools: A concept similar to a RAID group, storage pools are a collection of disks on which LUNs are created. Pools are dedicated for pool (thick and thin) LUNs.

Thin LUNs: LUNs that present more storage to an application than is physically allocated. Thin LUNs are created from the pools.

Thick LUNs: LUNs that are created from pools and not the traditional LUNs.

Advantages of Thin Provisioning

- Provides best space efficiency
- Easy set up and management
- Minimal host impact
- Energy and capital savings
- Applications where space consumption is difficult to forecast
- Automated storage tiering
- Applications with moderate performance requirements
- Taking advantage of data efficiency services like compression and space reclaimation

However, some applications require the data to be on normal LUNs. Use RAID groups and traditional LUNs:

- When milliseconds of performance are critical
- For the best and most predictable performance
- For precise data placement
- For physical separation of data
- When you are not as concerned about efficiency

Thick LUNs with storage pools must be used for:

- Applications that require very good performance
- Easy set up and management
- Automated storage tiering
- Minimal host impact

Recommendations

Proper considerations must be made before choosing the LUN type which will be used for certain applications. Due to the storage-on-demand feature of thin LUNs, not all application environments are well suited to thin LUNs. In general, it is a best practice to use thin-friendly

applications; that is, applications that do not pre-allocate all the storage space during initialization and that reuse deleted space before consuming additional storage.

3.9.2.3 Deduplication

Data deduplication or Single Instancing essentially refers to the elimination of redundant data. In the deduplication process, duplicate data is deleted, leaving only one copy (single instance) of the data to be stored. However, indexing of all data is still retained should that data ever be required.

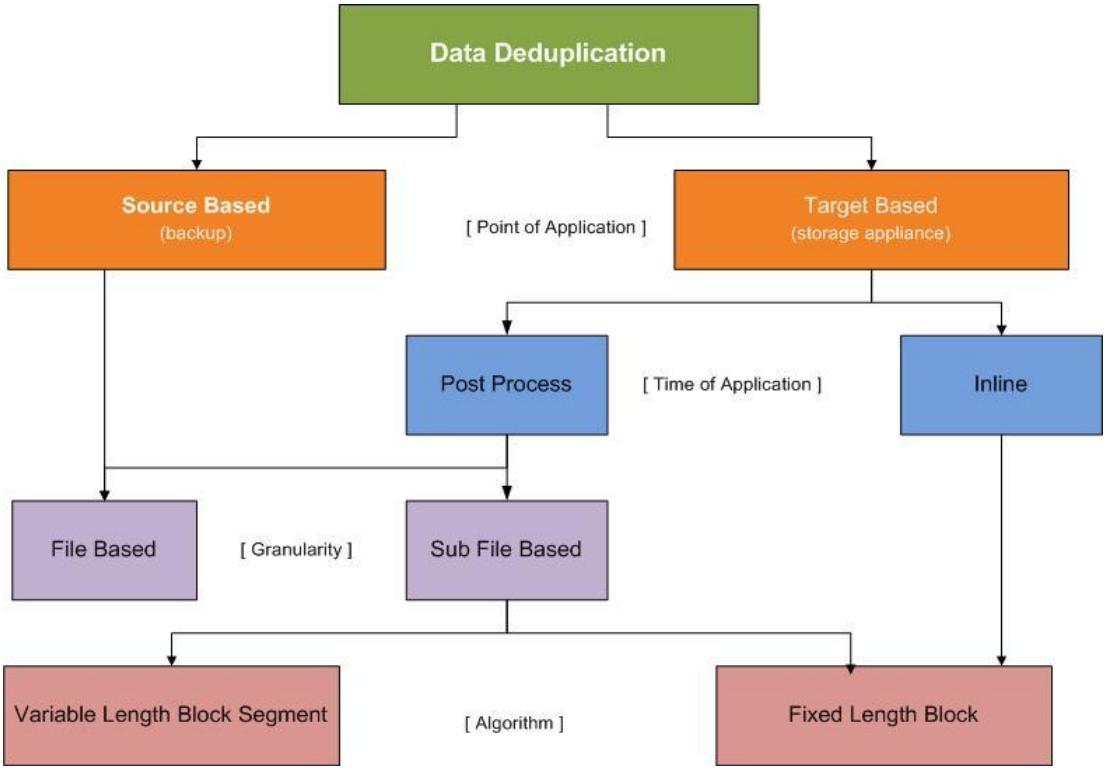


Figure 13 – Target- Vs Source-based deduplication[26]

Terms

Target-based deduplication acts on the target data storage media. In this case, the client is unmodified and not aware of any deduplication. The deduplication engine can be embedded in the hardware array, which can be used as a NAS/SAN device with deduplication capabilities. Alternatively, it can also be offered as an independent software or hardware appliance which

[26] Image source: http://blog.druva.com/category/enterprise-data-protection/

acts as intermediary between backup server and storage arrays. In both cases it improves only the storage utilization.

On the other hand, **Source-based deduplication** acts on the data at the source before it's moved. A deduplication-aware backup agent is installed on the client which backs up only unique data. The result is improved bandwidth and storage utilization. But, this imposes additional computational load on the backup client.

Inline vs Post-process Deduplication

In target-based deduplication, the deduplication engine can either process data for duplicates in real time (i.e. as and when it's sent to target) or after it's been stored in the target storage.

The former is called **inline deduplication**. The obvious advantages are:

- Increase in overall efficiency as data is only passed and processed once
- The processed data is instantaneously available for post-storage processes such as recovery and replication reducing the RPO and RTO window

The disadvantages are:

- Decrease in write throughput
- Extent of deduplication is less – only the fixed-length block deduplication approach can be used

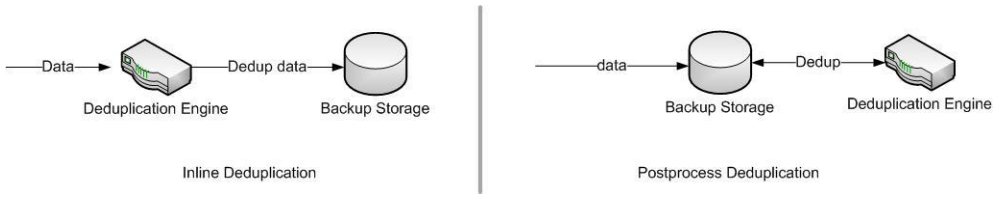


Figure 14 - Inline vs Post-process deduplication

Post-process deduplication asynchronously acts on the stored data. It has the exact opposite effect on advantages and disadvantages of the *inline deduplication* listed above.

ROI Benefits

Each organization has a capacity to generate data. The extent of savings depends upon–but is not directly proportional to–the number of applications or end users generating data. Overall, the deduplication savings depend upon:

- Number of applications or end users generating data
- Total data
- Daily change in data
- Type of data (emails/documents/media, etc.)
- Backup policy (weekly-full – daily-incremental or daily-full)
- Retention period (90 days, one year, etc.)
- Deduplication technology in place

The actual benefits of deduplication are realized once the same dataset is processed multiple times over a span of time for weekly/daily backups. This is especially true for *variable length data segment* technology which has a much better capability for dealing with arbitrary byte insertions.

Recommendations

- Deduplication saves the physical storage space but the process of deduplicating has an impact on application performance. Select the type of deduplication depending on the environment and performance requirements.
- Deduplication consumes system resources and can alter the data layout on disk. Due to the application's I/O pattern and the effect of deduplication on the data layout, the read and write I/O performance can vary. The space savings and the performance impact depend on the application and the data contents. Performance impact due to deduplication must be carefully considered and measured in a test setup and taken into sizing considerations before deploying deduplication in performance-sensitive solutions. If there is a small amount of new data, run deduplication infrequently, because there's no benefit in running it frequently in such a case, and it consumes system resources. How often you run it depends on the rate of change of the data in the flexible volume.
- The more concurrent deduplication processes you're running, the more system resources are consumed. Given the previous two items, the best option is to do one of the following:

- Use the proper modes available in the storage device so that deduplication runs only when significant additional data has been written to each volume.
- Stagger the deduplication schedule for the flexible volumes so that it runs on alternate days, reducing the possibility of running too many concurrent sessions.
- Run deduplication manually.

- If Snapshot copies are required, run deduplication before creating them to minimize the amount of data before the data gets locked in to the copies. Make sure that deduplication has completed before creating the copy. Snapshot copies created on a flexible volume before deduplication is given a chance to complete on that flexible volume could result in lower space savings.
- For deduplication to run properly, you need to leave some free space for the deduplication metadata.

3.9.3 Replication Selection recommendations

- Business requirements drive the replication options. Companies that run their businesses around the clock need to have their IT running without any disruptions, Synchronous replication is recommended for these type of data centers. However, not all data centers can leverage synchronous replication because of its limitations in distances[27]. Business must have their data centers within a metropolitan area to enable synchronous replication. Synchronous replications cause performance penalties because of the write that is required on the remote storage array for every write that happens to the primary storage. It is important to note this as performance-hungry applications might have some issues due to this.
- Companies which can tolerate a small amount of data loss can opt for asynchronous replications. Also, companies whose data centers are located at distances exceeding the limit of synchronous replication have to rely on asynchronous replication. Performance is not impacted by this kind of replication as the data is written on the remote storage array at a later point of time.
- Semi-synchronous replication writes are considered complete as soon as local storage acknowledges it and a remote server acknowledges that it has received the write either into memory or to a dedicated log file. The actual remote write is not performed immediately but is performed asynchronously, resulting in better performance than synchronous replication but with increased risk of the remote write failing. The

[27] DWDM and CWDM are the technologies used for multiplexing optical signals across long distances. Above 100 km (even 140km with repeaters), synchronous replication is not recommended due to high latency.

application of semi-synchronous replication varies with vendor offerings. They are equivalent to asynhronous replications. Enterprises must read the fine print to understand the differences between both and the requirement in the environment.

3.9.4 Best practices in Storage Capacity Management

- Keep a spreadsheet to track the storage being provisioned. Although the reports from change management utilities would generate reports, those would be difficult to understand the requested and provisioned data. Some key data that must be noted while provisioning storage are:
  - Amount of storage requested
  - Tier of the storage and its location with respect to the array
  - Application/Server
  - Requestor's contact details
  - Request/Change/Ticket number
- Monitor the capacity by using proper reporting tools such as EMC ControlCenter®, HDS Device Manager, etc. Be aware of the storage that is growing. Generate graphs to understand the growth rate and its projections.
- Check for the exisiting amount of disks and their utilization and conclude the required amount of buffer space.
- Monitor Fibre channel port usage:
  - Number of free ports on the director/switches
  - Number of free linecard slots in the director

3.9.4.1 Conclusion: Optimizing Storage

Storage is an ever-growing entity in the data center that has to be managed properly. New technologies available must be chosen appropriately for sustainability. Although cost of storage is decreasing while the data grows, the concern is managing this vast amount of storage. Thin provisioning, deduplication, and storage virtualization are a few technologies that can be used to reduce the physical storage footprint and also ease management. Enterprises need to evaluate which storage solution provides the best return on investment.

3.10 Converged Infrastructure

A converged infrastructure is a server infrastructure solution that combines data center physical resources (server processing, storage networking, Ethernet switching, power sources) into a single "resource pool" that can be allocated, aggregated, disaggregated, sized, configured,

repurposed, and maintained to meet the needs of the business. Other terms used to describe this type of infrastructure are "fabric-based" and "unified computing".

Cisco's Unified Computing System (UCS), HP Converged Infrastructure, and the Private Cloud solution from the virtual computing environment (VCE) coalition–Acadia are the players in Unified Computing.

Cisco UCS promises to minimize the management of servers. It uses FCoE as its communication medium for both storage and server traffic (I/O virtualization). The x86-based computing systems are tuned to work with VMware utilizing the features of server virtualization. All communication paths are via Cisco's Nexus switch.

HP's Converged data center provides broader coverage of the data center managing not just the server components but also the network and storage components.

The major components of HP Converged Infrastructure are:

- **Flex Fabric (Converged Networking):** Intelligent resilient networking framework
- **Virtual Resource Pools (Compute and Storage):** Pools of servers including ProLiant, and Integrity Servers and Blade systems. The storage pools consist of EVA, SVSP, P4000, and XP9000.
- **Data center Smart Grid (Power and Capacity management):** Thermal logic, Insight control, and Environment Edge.
- **Matrix Operating Environment (Orchestration):** Insight controls and dynamics and SIM.

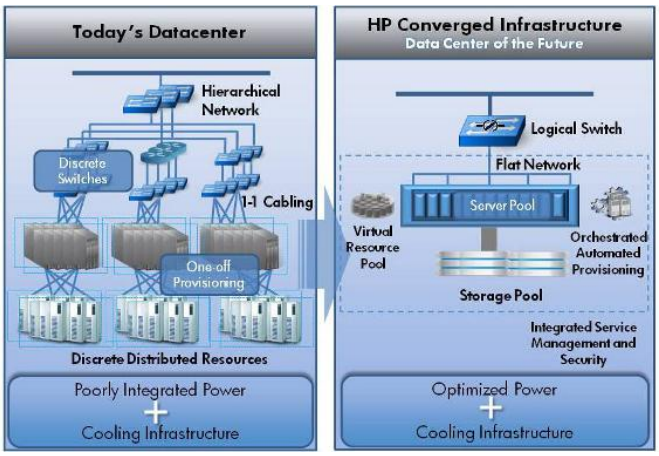


Figure 15 - HP's Converged Infrastructure[28]

[28] Image source : http://h18004.www1.hp.com/products/solutions/converged/ci-ra.html

Acadia is the VCE coalition formed by Cisco, EMC, and VMware. Although a private cloud solution, Acadia uses Vblock™ infrastructure solutions using storage, networking, and compute by leveraging technologies from EMC, Cisco, and VMware. VCE's Vblock Infrastructure Packages deliver a complete IT infrastructure that integrates best-of-breed virtualization, networking, compute, storage, security, and management technologies. The three companies have invested in an industry-first collaborative delivery of seamless customer support with end-to-end vendor accountability.

3.11 Open Source Solutions

Community developed open source tools available today provide IT solutions at low costs. Large Enterprises would not consider using open source tools due to lack of support to the products. However, there are various open source solutions available for today's technology challenges which are hard to ignore.

Small and newly-formed organizations can leverage open source for managing most of their IT environment. Large organizations cannot ignore open source solutions because some solutions are available only in the open source community. In this section we will go through the available open source solutions at a glance for the small, medium, and large organizations.

The following list of open source solutions is not an extensive list. There are other solutions available which can be applied on various businesses as well. This list is an example showing the capabilities of open source solutions that can be applied in certain businesses.

3.11.1 Operating System – Servers and Workstations

Microsoft Windows dominates market share in desktop operating system (OS) usage[29]. This is the most used OS because of ease of use and integration of the directory services. Linux, on the other hand, is powerful and almost free of viruses. Linux is not deployed in large enterprises due to several reasons including innumerable and growing distributions, lack of support (although some distribution provide), users familiar with Microsoft tools, less powerful productivity tools (Open Office vs. Microsoft Office), incompatible application software, delays due to community development, and more. However, small organizations are leveraging Open source operating systems to be used by their employees.

[29] Microsoft Windows has a market share of 88.21% in desktop operating systems across the world. Source: http://gs.statcounter.com/#os-ww-yearly-2008-2010

Small organizations can customize the Linux OS to the specific needs of the users and use it. Ubuntu, the most popular and most used[30] Linux OS can be customized for users to use in that department. In the event that specific services need to be installed, the company can install other operating systems (non-Linux) for other needs that can't be satisfied by a Linux-based operating system.

Advantages

- Huge costs svaings as most Linux-based operating systems and applications are free
- Open source – Can be customized according to the needs of the organization
- Security – Linux-based operating systems are safe and secure
- Virus free – They are virtually virus-free
- Can use a wide variety of community developed applications

Disadvantages

- Multiple distribution creating lots of confusion among users
- Lack of familiarily of the operating system with users
- Not all applications and system software have proper support and patches
- Compatability issues with exisiting setup
- Application development for certain custom-based applications
- Although most hardware are supported, there might be issues with drivers and hardware compatability

Although Linux isn't 'popular', the cost savings if a small company leverages the desktop operating system on Linux is phenomenal. In the server market, applications running on a Unix-based operating system (AIX, HP-UX) can move to industry-standard processor architecture (x86) and run a Linux-based server operating system.

3.11.2 Storage

Gluster Storage Platform

Gluster Storage Platform is an open source clustered storage solution. The software is a powerful and flexible solution that simplifies the task of managing unstructured file data whether you have a few terabytes of storage or multiple petabytes. Gluster Storage Platform integrates the file system, an operating system layer, and a web-based management interface and installer. Installation takes less than 10 minutes. It supports native GlusterFS, CIFS, and NFS

[30] Top Ten Distributions. http://distrowatch.com/dwres.php?resource=major

access protocols and InfiniBand, GigE, and 10GigE network interconnects. You can get started with just a single system configured as a standalone NAS server, and scale from there as needed.

The core of Gluster Storage Platform is the GlusterFS file system. The Storage Platform integrates GlusterFS with an operating systems layer and a graphical user interface (GUI) and installer. Gluster Storage Platform installs directly on any industry-standard x64 server creating a fully functional storage server. GlusterFS may be deployed independently as a user space application on supported operating systems for certain use cases or for access to advanced configuration options. Regardless of deployment method, Gluster is fully POSIX-compliant and installs on industry-standard hardware providing highly scalable, easy to use storage with the superior economics of commodity hardware plus open source software.

Deploying Gluster

Gluster is scale-out NAS for unstructured file data that is applicable to multiple use cases. GlusterFS is modular and can be configured and optimized for a wide range of workloads. The global namespace aggregates disk and memory into a single pool of resources with flexible back-end disk options, supporting direct-attached, JBOD, or SAN storage. Storage servers can be added or removed without service disruption, enabling storage to grow or shrink on-the-fly in the most dynamic environments. Gluster solutions are often deployed in the following scenarios:

Scalable NAS: Gluster provides up to petabytes of file storage in a single namespace for data centers that need scalability and performance that traditional NAS and SAN systems can't provide. Storage servers are added in building block fashion to meet capacity and performance requirements; compute, I/O, and disk resources can be added independently. Gluster is a lower cost alternative to SAN storage, providing comparable reliability and performance at significantly lower cost.

Virtual Storage for Virtual Machines: Gluster is a scalable NAS solution, designed for file storage; Virtual Machine (VM) images are files that are efficiently stored on Gluster. With the ability to scale to petabytes and spread I/O across multiple storage servers, Gluster eliminates the primary problems from which traditional VM storage solutions suffer. Unlike SAN storage for VMs, Gluster provides a single mount-point that thousands of VMs can share, offering simplified management and lower cost.

Cloud Storage: Gluster is well suited to both public and private clouds. Enterprises build centralized storage pools that are allocated on a per-volume basis to end user customers.

Gluster easily deploys on public cloud platforms such as Amazon and Rackspace to aggregate block storage into a single pool that can be shared across many clients. Gluster is POSIX-compliant so the interface abstracts, cloud vendor APIs, and applications do not need to be modified.

FreeNAS is a FreeBSD-based free NAS server, supporting: CIFS (Samba), FTP, NFS, rsync, AFP protocols, iSCSI, S.M.A.R.T., local user authentication, and software RAID (0,1,5), with a web-based configuration interface. The operating system can be installed on any 386/IA-32 and x86-64 based platform.

Small businesses can get maximum benefits from FreeNAS as it can replace traditional Windows servers which are not optimized for file sharing[31]. The most important feature of FreeNAS is its lightweight OS. It's fast to boot, easy to install, and provides a very user-friendly web-based interface.

Openfiler is a network storage operating system, fronted by a web-based management user interface. With the features we built into Openfiler, you can take advantage of file-based NAS and block-based SAN functionality in a single cohesive framework.

Any industry standard x86 or x86/64 server can be converted into a powerful multi-protocol network storage appliance, replete with an intuitive browser-based management interface, in as little as 15 minutes. File-based storage networking protocols such as CIFS and NFS ensure cross-platform compatibility in homogeneous networks–with client support for Windows, Linux, and Unix. Fibre channel and iSCSI target features provide excellent integration capabilities for virtualization environments such as Xen and VMware.

iSCSI target functionality is especially useful for enterprise applications such as Microsoft Exchange server integration, Oracle 10g RAC backend storage, or video surveillance and disk-to-disk backup.

IET (iSCSI Enterprise Target) is an open source iSCSI target with professional features that works well in enterprise environments under real workload, and is scalable and versatile enough to meet the challenge of future storage needs and developments. Although not a full-fledged operating system, the package can be installed on any standard hardware with few basic Linux components. This provides flexibility for businesses to build their own iSCSI target solution.

[31] Windows has an OS optimized for NAS – Windows Storage Server 2008. It does have better features than FreeNAS. Large businesses would prefer a NAS appliance than a file server, using FreeNAS would be a better alternative for small to medium enterprises.

Apache Hadoop

Apache Hadoop is a software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of nodes and petabytes of data. Hadoop provides a reliable shared storage and analysis system. The Hadoop framework transparently provides applications both reliability and data motion. Hadoop implements a computational paradigm named MapReduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system (HDFS) that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both MapReduce and the distributed file system are designed so that node failures are automatically handled by the framework.

3.12 Section Conclusion

The data center is the organization's most important IT asset. Managing the data center is a daunting task if it is not properly designed. IT architects and managers must consider all options–especially the power and cooling–to transform the data center into a robust, agile, and reliable data center.

Data center transformation approaches vary from enterprise to enterprise. The common goals remain the same; reduce costs and increase efficiency. Businesses must look into the alternate technologies that would perform the same operation at a better value.

For businesses to remain viable, the data center must be built to not just deliver what it promises but also to ensure it is designed for future growth. Enterprises must adhere to the regulations and compliance requirements applicable to the geography where the data center exists. These regulations not only are a requirement but enable the business to showcase its capability.

4. IT Compliance and Controls

Compliance means conforming to a rule, such as a specification, policy, standard, or law. Regulatory compliance describes the goal that corporations or public agencies aspire to in their efforts to ensure that personnel are aware of and take steps to comply with relevant laws and regulations. Due to the increasing number of regulations and need for operational transparency, organizations are increasingly adopting the use of consolidated and harmonized sets of compliance controls.

This approach is used to ensure that all necessary governance requirements can be met without the unnecessary duplication of effort and activity from resources. Compliance and controls are concepts that we don't hear everyday due to focus on technology. Organizations should be aware of the terms and adhere to them. Government or industry standards are other reasons why IT services needs to comply with regulations.

Whether a network professional, a storage engineer, or a CEO, compliance holds a different meaning to different stake holders. Regardless, the objective of compliance is to reduce vulnerabilities, avoid penalties, or meet internal obligations. Compliance violations start with $100 fines and can go all the way up to $250,000 and 10 years in prison (HIPPA, 1996). Compliance is now a requirement for any organization.

4.1 Why be compliant?

It's no longer a question if a compliant solution is required. Rather, the question is what is the best fit that can be applied. Every organization, regardless of their size, must comply with a wide range of rules and regulations. The only constant is that the range keeps getting wider. There are three main reasons why corporate IT needs to be compliant:

1) Government regulation
2) Industry standard
3) Technical requirement

Various corporate scandals have highlighted the need for stronger compliance and regulations for publicly listed companies. Although it is a mandatory requirement in large organizations, smaller companies are becoming compliant to showcase their capabilities. A government regulation is a mandatory requirement that a company in that particular sector HAS to comply with. Failing to comply might result in terminating certain services to that organization, being blacklisted, or losing the company's face value with its shareholders.

Let's look at some of the government regulations that are important to an organization. Some might not apply to certain organizations.

Some of the regulations, standards, and frameworks are listed below:

- Sarbanes-Oxley Act of 2002 (SOX)
- Payment Card Industry Data Security Standard (PCI DSS)
- Health Insurance Portability and Accountability Act (HIPAA)
- Gramm-Leach-Bliley Act (GLB Act)

- ISO 27000 Series of Standards
- Information Technology Infrastructure Library (ITIL)
- Control Objectives for Information and related Technology (COBIT)
- Val IT

4.1.1 Sarbanes-Oxley Act of 2002

As a result of the financial scandals at major Fortune 100 companies in 2001, the United States Congress enacted the Sarbanes-Oxley Act (SOX) of 2002. This act affects how public companies report financials, and significantly impacts IT. SOX compliance requires more than documentation and/or establishment of financial controls; it also requires the assessment of a company's IT infrastructure, operations, and personnel. Unfortunately, the requirements of the Sarbanes-Oxley Act of 2002 do not scale based on the size or revenue of a company. Small to medium-sized companies (IT department) will face unique challenges, both budgetary and with personnel, in their effort to comply with the Sarbanes-Oxley Act of 2002.

Coverage: Financial Industry

What Will SOX Accomplish?

There continues to be much controversy and debate about the effectiveness of SOX. Although most people who are aware of the requirements to comply with SOX (Section 404) believe the intention was good, there exists controversy over whether the existing 302 reporting requirements are sufficient.

If you read Sections 302 and 404, you may see similarities, and subsequently, why a controversy may exist as to whether (Section 404) SOX requirements and compliance were necessary. An example of Sections 302 and 404 is shown below, as they pertain to a company's executive management assertions.

Section 302

In accordance with Section 302, executive management of a public company:

1) Are responsible for establishing and maintaining internal controls
2) Have designed such internal controls to ensure that material information relating to the issuer and its consolidated subsidiaries is made known to such officers by others within those entities, particularly during the period in which the periodic reports are being prepared

Section 404

In accordance with Section 404, executive management of a public company:

1) Is responsible for establishing and maintaining an adequate internal control structure and procedures for financial reporting
2) Must report the effectiveness of the internal control structure and procedures

4.1.2 Payment Card Industry Data Security Standard

The Payment Card Industry Data Security Standard (PCI DSS) is a worldwide information security standard defined by the Payment Card Industry Security Standards Council. The standard was created to help payment card industry organizations that process card payments prevent credit card fraud through increased controls around data and its exposure to compromise. The standard applies to all organizations that hold, process, or exchange cardholder information from any card branded with the logo of one of the card brands.

Validation of compliance can be performed either internally or externally, depending on the volume of card transactions the organization handles, but regardless of the size of the organization, compliance must be assessed annually. Organizations handling large volumes of transactions must have their compliance assessed by an independent assessor known as a Qualified Security Assessor (QSA), while companies handling smaller volumes have the option of demonstrating compliance via a Self-Assessment Questionnaire (SAQ). In some regions, these SAQs still require signoff by a QSA for submission.

Coverage

- Any company that accepts credit cards must be PCI compliant
- Three types of companies that do this: Level 1, Level 2, and Level 3 merchants; levels are based upon the number of credit card transactions

Requirement Standards

- Build and maintain a secure network
    1) Install and maintain a firewall configuration to protect cardholder data
    2) Do not use vendor-supplied defaults for system passwords and other security parameters
- Protect Cardholder Data
    3) Protect stored cardholder data

4) Encrypt transmission of cardholder data across open, public networks

- Maintain a Vulnerability Management Program
    5) Use and regularly update anti-virus software on all systems commonly affected by malware
    6) Develop and maintain secure systems and applications
- Implement Strong Access Control Measures
    7) Restrict access to cardholder data by a business' need-to-know
    8) Assign a Unique ID to each person with computer access
    9) Restrict physical access to cardholder data
- Regularly monitor and test the network
    10) Track and monitor all access to network resources and cardholder data
    11) Regularly test security systems and processes
- Maintain an information security policy

4.1.3 Health Insurance Portability and Accountability Act

The Health Insurance Portability and Accountability Act (HIPAA) of 1996 (P.L.104-191) was enacted by the U.S. Congress in 1996. It was originally sponsored by Sen. Edward Kennedy (D-Mass.) and Sen. Nancy Kassebaum (R-Kan.). According to the Centers for Medicare and Medicaid Services (CMS) website, Title I of HIPAA protects health insurance coverage for workers and their families when they change or lose their jobs. Title II of HIPAA, known as the Administrative Simplification (AS) provisions, requires the establishment of national standards for electronic health care transactions and national identifiers for providers, health insurance plans, and employers.

Coverage: Hospitals, Medical centers, Medical insurance companies

The regulation is broken down into two titles:

1. Title I: Health Care Access, Portability, and Renewability
2. Title II: Preventing Health Care Fraud and Abuse; Administrative Simplification; Medical Liability Reform. Per the requirements of Title II, the HHS has promulgated five rules regarding Administrative Simplification:
    a. Privacy Rule
    b. Transactions and Code Sets Rule
    c. Security Rule
    d. Unique Identifiers Rule (National Provider Identifier)

e. Enforcement Rule

Two of the Title II rules are of the most interest to IT services: the **Privacy Rule** and the **Security Rule.**

An excerpt from the Technical safeguards under the Security Rule is shown below. The listed indicate the technical aspects of the regulation that an IT manager must be aware.

Technical Safeguards – controlling access to computer systems and enabling covered entities to protect communications containing PHI transmitted electronically over open networks from being intercepted by anyone other than the intended recipient.

- ✓ Information systems housing PHI must be protected from intrusion. When information flows over open networks, some form of encryption must be utilized. If closed systems/networks are utilized, existing access controls are considered sufficient and encryption is optional.
- ✓ Each covered entity is responsible for ensuring that the data within its systems has not been changed or erased in an unauthorized manner.
- ✓ Data corroboration, including the use of check sum, double-keying, message authentication, and digital signature may be used to ensure data integrity.
- ✓ Covered entities must also authenticate entities with which they communicate. Authentication consists of corroborating that an entity is who it claims to be. Examples of corroboration include: password systems, two or three-way handshakes, telephone callback, and token systems.
- ✓ Covered entities must make documentation of their HIPAA practices available to the Government to determine compliance.
- ✓ In addition to policies and procedures and access to records, information technology documentation should also include a written record of all configuration settings on the components of the network because these components are complex, configurable, and always changing.
- ✓ Documented risk analysis and risk management programs are required. Covered entities must carefully consider the risks of their operations as they implement systems to comply with the Act. (The requirement of risk analysis and risk management implies that the Act's security requirements are a minimum standard and places responsibility on covered entities to take all reasonable precautions necessary to prevent PHI from being used for non-health purposes.)

4.1.4 Gramm-Leach-Bliley Act

The Gramm-Leach-Bliley Act (GLB Act), also known as the Financial Modernization Act of 1999, is a federal law enacted in the United States to control the ways that financial institutions deal with the private information of individuals. The Act consists of three sections:

1) The Financial Privacy Rule, which regulates the collection and disclosure of private financial information
2) The Safeguards Rule, which stipulates that financial institutions must implement security programs to protect such information
3) The Pre-texting provisions, which prohibit the practice of pre-texting (accessing private information using false pretenses)

The Act also requires financial institutions to give customers written privacy notices that explain their information-sharing practices.

An excerpt of the regulation[32] is shown below:

"Each bank shall implement a comprehensive written information security program that includes administrative, technical, and physical safeguards appropriate to the size and complexity of the bank and the nature and scope of its activities. While all parts of the bank are not required to implement a uniform set of policies, all elements of the information security program must be coordinated".

A bank's information security program shall be designed to:

1) Ensure the security and confidentiality of customer information
2) Protect against any anticipated threats or hazards to the security or integrity of such information
3) Protect against unauthorized access to or use of such information that could result in substantial harm or inconvenience to any customer

Coverage: Financial Institutions

[32] "Interagency Guidelines Establishing Standards for Safeguarding Customer Information and Rescission of Year 2000 Standards for Safety and Soundness", page 19
http://www.ffiec.gov/exam/InfoBase/documents/02-joi-safeguard_customer_info_final_rule-010201.pdf

4.2 Achieving the Goals

Now that we know the important regulations that require organizational compliance, we need to know various standards and frameworks that are available and can be applied. In this section we will look at how various IT governance frameworks help the organization reach compliance and also how they benefit the organization.

IT Governance is a framework for the leadership, organizational structures, and business processes, standards, and compliance to these standards, which ensure that the organization's IT supports and enables the achievement of its strategies and objectives.

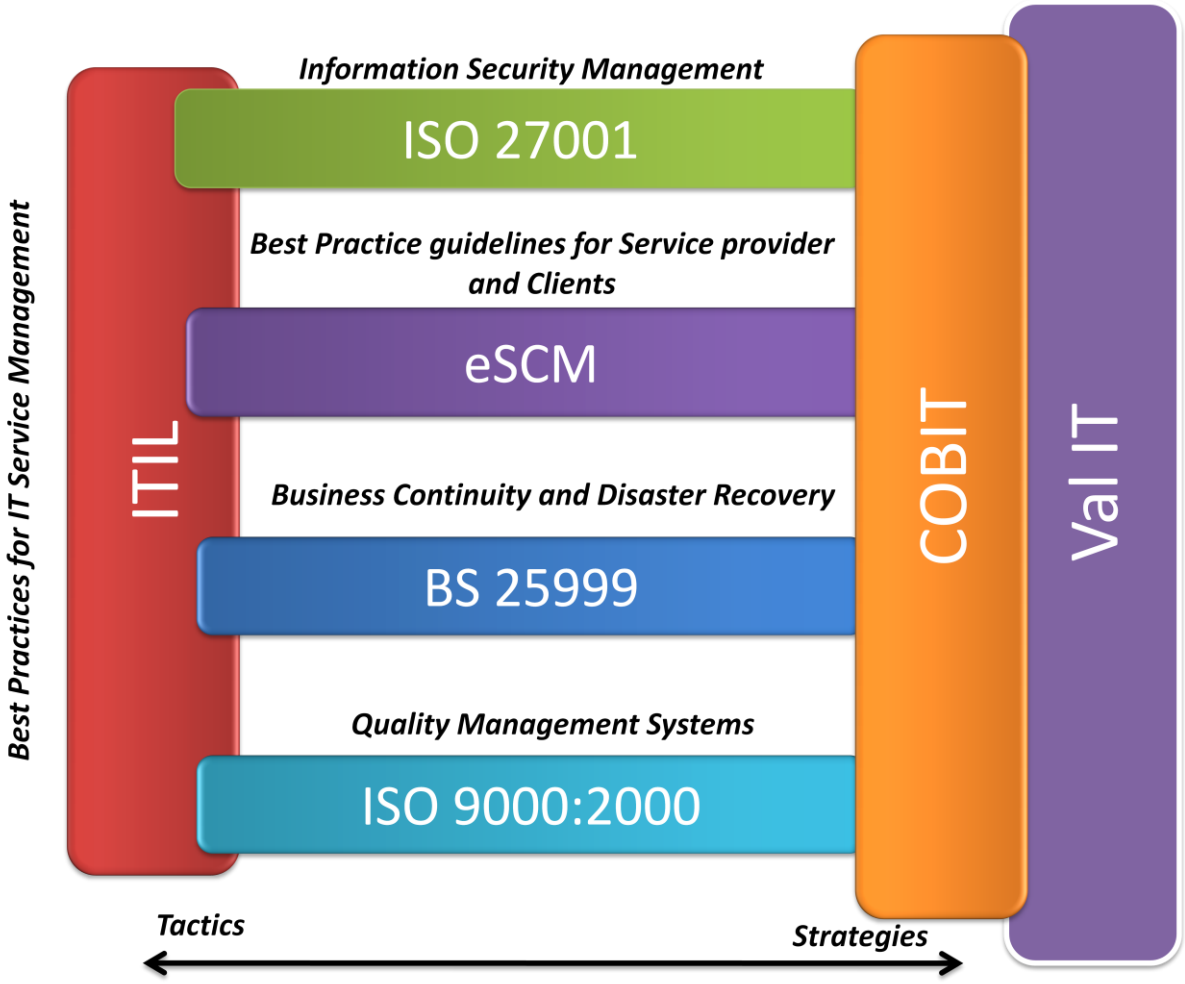


Figure 16 – A Comprehensive IT Governance Framework

There are various frameworks available that are used in an organization to achieve its goals. While these achieve the organizational goals, they also help in complying with various

regulations and standards. Some of the frameworks restrict themselves as to what is to be done. However, there are frameworks that detail the process of an activity.

Figure 16 shows the IT governance frameworks consisting of various frameworks and standards that help the organization achieve their goal. For IT governance to be effective, Val IT and COBIT must be integrated with the rest of the frameworks. Val IT and COBIT are the strategies, whereas ITIL and the standards are the operations or tactics involved in achieving those.

4.2.1 Information Technology Infrastructure Library

Information Technology Infrastructure Library (ITIL) is the best practice framework of concepts and practices for IT service management, IT development, and IT operations. ITIL is essentially a series of documents that are used to aid the implementation of a lifecycle framework for IT Service Management. This customizable framework defines how Service Management is applied within an organization. It is aligned with the international standard, ISO 20000.

4.2.1.1 Basics

ITIL is organized into five titles that revolve around the service life cycle:

1. Service Strategy
2. Service Design
3. Service Transition
4. Service Operation
5. Continual Service Improvement

These in turn describe a closed loop feedback system that provides feedback throughout all areas of the lifecycle. The volumes continue to provide a framework of best practice disciplines that enable IT Services to be provided effectively.

4.2.1.2 How ITIL helps organizations

Organizations can benefit from ITIL in several important ways:

- Quantifiable Return On Investment (ROI) for IT expenditure
- Closer linkage from business drivers into IT investments
- Creation of more dynamic and flexible ITSM models
- IT service performance being driven more directly by its value to the business
- Greater utilization of IT assets
- IT services become more customer-focused

- The quality and cost of IT services are better managed
- The IT organization develops a clearer structure and becomes more efficient
- IT changes are easier to manage
- There is a uniform frame of reference for internal communication about IT
- IT procedures are standardized and integrated
- Demonstrable and auditable performance measurements are defined

4.2.1.3 Recommendations

Although all large organizations follow ITIL processes and guidelines, there is a lot of scope of improvement for the processes already in place (Continual Service Improvement). However, small and medium sized organizations as well as some large organizations do not use ITIL for their processes and activities. They should consider implementing it immediately as the advantages of having ITIL processes will reduce IT expenditure and improve IT efficiency.

One of the disciplines of IT Service Management is change management. Most vulnerabilities or service outages could be prevented if an 'alteration' or 'change' was managed properly. As an example, a server's registry could've been modified by the administrator while checking another unrelated problem causing the state of the server to no longer be in the recommended state or, in the case of a 'Non-standard configuration', it might attract other issues. On the other hand, an important security update installed after proper consent and verifying the impact of the update from the stakeholders helps the server to be in a 'New standard configuration'. Change management helps keep track of the alterations and ensures the alteration to proceed only on a requirement basis.

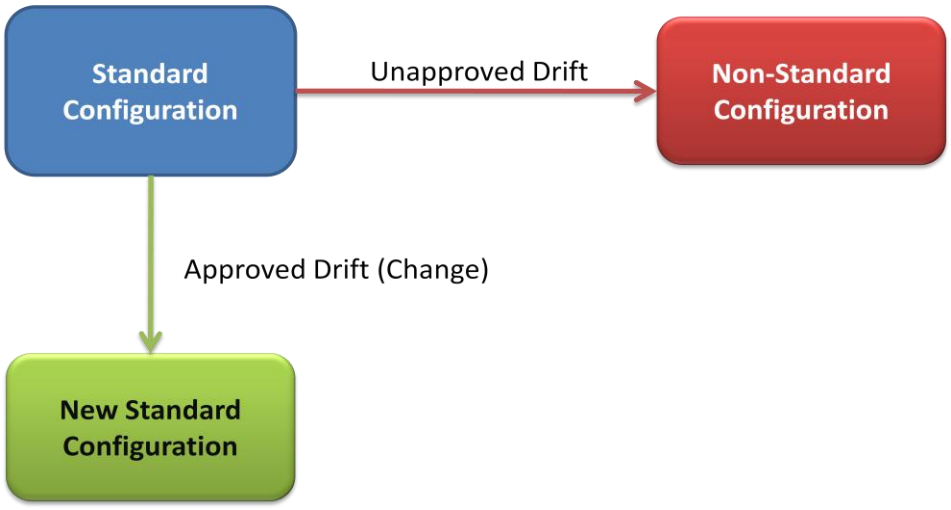


Figure 17 - Importance of change management (unapproved drift causes vulnerabilities)

Companies who outsource their IT operations completely to third-party service providers must ensure that the service provider use ITIL processes for their operations. This applies even for those companies who have partially outsourced their operations.

4.2.2 eSourcing Capability Model

The eSourcing Capability Model (eSCM) is a framework developed to improve the relationship between IT Services providers and their customers. eSCM is broadly classified into:

1. eSourcing Capability Model for Service Providers (eSCM-SP)
2. eSourcing Capability Model for Client Organizations (eSCM-CL)

eSCM-SP is a “best practices” capability model with three purposes:

1) To give service providers guidance that will help them improve their capability across the sourcing life-cycle
2) To provide clients with an objective means of evaluating the capability of service providers
3) To offer service providers a standard to use when differentiating themselves from competitors

eSCM-CL serves the following two purposes:

1) To give client organization guidance that will help them to improve their sourcing capability
2) To provide client organization with an objective means of evaluating their sourcing capability

4.2.2.1 How eSCM helps organizations

eSCM can benefit the organization in the following ways:

- Helps service providers of IT-enabled services provide a reference model and capability
- Helps service providers develop and improve their ability
- Improved service from service providers
- Assists in contract negotiations
- Enables transition of resources and reverse transition at contract completion
- Methods of buyers to appraise the capabilities of outsourcers
- Provides a level structure taking into account changes in business needs over the life of the contract

- Assists client organizations' guidance in improving their capability across the sourcing life-cycle
- Provides clients an objective means of evaluating their capability

4.2.3 ISO/IEC 27001

ISO/IEC 27001, part of the growing ISO/IEC 27000 family of standards, is an Information Security Management System (ISMS) standard published in October 2005 by the International Organization for Standardization (ISO) and the International Electro-technical Commission (IEC). The standard recognizes the importance of information as a valuable asset of an organization. 'Information' includes that kept on computer systems, paper transmitted via post, email, and so forth.

The main aims of ISO 27001 are to maintain:

1. Confidentiality
2. Integrity
3. Availability

Benefits of achieving ISO27001

- Demonstrates a commitment to Information Security throughtout the organization
- Heighted customer confidence
- Demonstrates credibility and that laws and regulations are followed
- Can realise cost savings via reduced security breaches
- Proof of security to third parties (for clients, partners, and legal purposes)
- Competitive advantage: 'documented quality' by an independent authority
- Cost reductions through transparent, optimized structures
- Security becomes an integral part of business processes
- Knowledge and monitoring of the IT risks and residual IT risks
- Documentation of structures and processes
- Increased employee awareness of security
- Evaluation of the organization's processes from a security point of view
- Prioritizing the security of the business operations: business continuity management
- Globally recognized standard
- Potential reduction in insurance premiums
- Referencing the IT process management standard (ITIL) to ISO 27001
- Seamless transition from ISO 27001 in management systems to ISO 9000.

4.2.4 BS 25999 – Business Continuity Management System Standard

British Standard Institution's British Standard (BS) 25999 takes the idea of a business continuity plan and broadens its scope to encompass the entire enterprise. BS 25999 focuses the organization's objectives, making the development of a Business Continuity Management System (BCMS) more efficient and its implementation more effective. Creating a BCMS involves not only the establishment and implementation of a plan to protect employees, processes, and supply chains, but also the continual improvement of plans in order to maintain a system vital to the organization's resilience.

4.2.4.1 Benefits of achieving BS25999

- Provides a common framework, based on international good practice, to manage business continuity
- Proactively improves your resilience when faced with disruptions to your ability to achieve key objectives
- Provides a rehearsed method of restoring your ability to supply critical products and services to an agreed level and timeframe following a disruption
- Delivers a proven response for managing a disruption
- Helps protect and enhance your reputation and brand
- Opens new markets and helps win new business
- Enables a clearer understanding of how your whole organization works and can help identify opportunities for improvement
- Demonstrates that applicable laws and regulations are being observed
- Creates an opportunity to reduce the burden of internal and external BCM audits and may reduce business interruption insurance premiums

4.2.5 COBIT and Val IT

COBIT and Val IT are used by executives to deliver IT governance frameworks. The rest of the frameworks which we have discussed are required by the IT operations and architecture. The metrics, performance, and indicators from these frameworks feed as input for COBIT and Val IT.

COBIT is an IT governance framework and supporting toolset that allows managers to bridge the gap between control requirements, technical issues, and business risks. COBIT enables clear policy development and good practice for IT control throughout organizations. COBIT emphasizes regulatory compliance, helps organizations increase the value attained from IT, enables alignment, and simplifies implementation of the COBIT framework.

4.2.5.1 Benefits of implementing COBIT

- Improves IT efficiency and effectiveness
- Helps IT understand the needs of the business
- Puts practices in place to meet the business needs as efficiently as possible
- Ensures alignment of business and IT
- Helps executives understand and manage IT investments throughout their lifecycle
- A common language for executives, management, and IT professionals
- A better understanding of how the business and IT can work together for successful delivery of IT initiatives
- Improved efficiency and optimization of cost
- Reduced operational risk
- Clear policy development
- More efficient and successful audits
- Clear ownership and responsibilities, based on process orientation

Val IT is a framework that focuses on value delivery, one of the five main areas of the governance of IT, and ensures that IT-enabled investments are managed through their full lifecycle.

4.2.5.2 Benefits of implementing Val IT

- Fosters the partnership between IT and the rest of the business
- Assists the board and executive management in understanding and carrying out their roles related to IT-enabled business investments
- Helps enterprises make better decisions on where to invest in business change
- Provides a common language for executives, business management, and IT professionals to ensure IT-related investments are in line with business strategy
- Increased ROI for projects
- Business value is generated
- IT-enabled investments are managed through their full lifecycle
- Increased value of technology investments, leading to business growth
- Reduced costs resulting from inefficient investments
- Better decisions are made on where to invest in business change

4.2.6 Compliance and Controls Conclusion

Organizations are required to comply with regulations and standards for various reasons. These regulations are not just mandatory; they also help customers understand the organization and its capability. IT governance and standards are a set of frameworks that helps organizations achieve those goals.

IT governance is required in medium and large scale organizations. Proper implementation helps the customer achieve value apart from just being compliant. Some useful points for compliance are:

- **Prepare to be compliant**: Keep current regarding compliance statutes, case law, regulations, and industry standards and support industry-wide compliance practices
- **Prevent non-compliance**: Through affiliate and advertiser application process
- **Protect your integrity**: Identify compliance issues and take appropriate action

Conclusion

Cloud computing and technologies such as virtualization and converged infrastructure are now available to organizations to harness its capabilities. The choices you make will depend on what is appropriate for your organization. Any decision must take into consideration factors such as revenue, industry type, and maturity of the industry.

Small and medium size companies must look at options of moving their infrastructure to the cloud almost immediately. This would reduce capital and operational expenses. As well, large companies must seriously consider the potential that the cloud offers and look for opportunities where this can be implemented.

Data centers have to be efficient to comply with ‘green’ regulations as well as to reduce costs due to rising electricity expenditure. Organizations must perform power and cooling, server, and storage utilization assessments and remove inefficient resources from the data center, followed by implementation of a powerful and comprehensive plan for infrastructure optimization.

IT governance frameworks transform the IT delivery model and help organizations generate revenue. The need for compliance has guided the way for organizations to efficiently utilize the IT they own. Various frameworks and standards can be used to achieve compliance as well as generate value out of IT.

APPENDIX A: List of Cloud Platforms, Providers, and Enablers

Cloud computing infrastructure and solution provider	
3Tera	AppLogic grid OS used as cloud computing platform by service providers and enterprises
Appistry	Cloud computing middleware. Enables easily scalable cloud computing in the enterprise.
Cassatt computing resources	Cassatt Active Response platform enables administrators to set policies to power physical and virtual servers safely on and off and pool their
CloudHan	Cloud tech and infrastructure consultant, in China.
CloudScale Networks	Cloud enabler. Currently in private ALPHA only
Joyent	Cloud Infrastructure (Accelerators), and consulting for developers and enterprise.
nScaled, Inc	Cloud related services such as Migrations, Deployment, Planning, Consulting
Qlayer	Provides software for data centers that enables cloud computing, support VSAN, VLAN, VPDC, currently support VMware ESX.
Skytap	IaaS service optimized for QA, Training, Demo, and Ops Testing. Supports VMware, Xen hypervisors & Windows, Linux & Solaris OS guests.
Webscale Solutions	IT Strategy and Consulting on Cloud computing. Specialize in ROI investigations of CC. a CC provider evaluation framework and Enterprise Cloud Roadmap development.
Cloud computing infrastructure provider	
Agathon Group	Cloud provider. Services include highly available VPS, virtual private datacenters, and ready to use LAMP stacks. Self-service ordering. Custom development and managed services available.
Amazon Web Services	Amazon EC2/S3 (Hardware-a-a-Service & Cloud Storage)
CohesiveFTconfigs) with deployment to many virtual and cloud environs.	CohesiveFT Elastic Server FactoryWebservice for assembling full application stacks (contextualization, custom apps, middleware, on top of base.
ElasticHosts	UK-based instant, on-demand servers in the cloud.
Flexiscale	An instant provisioner of web servers with some advanced features like auto-scaling coming soon.
GoGrid	Instant, on-demand servers offering “control in the cloud”. Deploy Windows/Linux servers via web-interface in minutes.
GridLayerservers	Cloud Provider. A service by Layered Technologies that delivers Virtual Private Data Centers and virtual private servers from grids of commodity.
LayeredTechnologies	Cloud Provider. Provider of on-demand hosting and cloud and utility computing solutions through its brand, GridLayer.
ReliaCloud	Deployed within a robust and resilient virtualization environment and architected to maximize uptime and performance. Free benefits include high availability, load balancing, robust APIs, and persistent servers.
Mosso	Rackspace’s cloud hosting service

Newservers	Instant provisioning of web servers either Windows or Linux.
Plura Processing	On-demand infrastructure for high-performance computing.
Cloud computing PaaS provider	
Aptana Cloud	Elastic Elastic Application Cloud™ featuring fully stacked and integrated PHP app engines, Ajax/Jaxer app engines, and soon Ruby on Rails app.
Bungee Connect	Provides end-to-end tools and systems required to develop, deploy and host web applications (Platform as a Service).
Coherence	Oracle Coherence Data Grid for EC2 and other cloud platforms.
Force.com	Salesforce.com's application development platform (PaaS).
GigaSpaces	Middleware for the cloud, "cloudware".
Google App Engine	(PaaS) Now support python.
Heroku	Ruby on Rails in their Cloud.
Morph Labs	Fully managed, open, elastically-scalable, end-to-end deployment and delivery platform for Ruby on Rails and Java (Jetty, JRuby, Groovy, and Grails) web applications. Leverages AWS, but completely abstracts details and complexities from developers.
Intuit Partner Platform (IPP)	Platform as a Service (PaaS) from Intuit.
Qrimp	An AJAX-based PaaS
RightScale	RightScale provides a platform and expertise that enable companies to create scalable web applications running on Amazon's Web Services that are reliable, easy to manage, and cost less.
Stax	Java Platform as a Service.
Cloud computing based service provider	
CAM Solutions	Monitoring-as-a-Service(TM).
CloudStatus	CloudEnabler. Real-time performance trending of cloud infrastructure (currently AWS).
DATASiSAR	Cloud Computing technology based consulting & IT Services provider.
Kaavo's IMOD	An easy to use online application.
Microsoft Mesh	
Nasstar	SaaS provider. Business-grade Hosted Desktop service, UK market leaders.
Nirvanix	Cloud Storage.
TrustSaaS	Uptime monitoring and alerting service ('SaaS Weather Report') for Software as a Service (SaaS) run by an independent third party.
UtilityStatus	Utility Computing Platform for SaaS charged in elapsed CPU time running on EC2.

Semantic computing Cloud service provider	
ThoughtExpress	Generic Enterprise Management Service based in semantics supported by semantic computing cloud to perform enterprise information processing to deliver: BPM, BI, enterprise modeling, and semantic human interface without the need to program.
Cloud Security Consultants and Overlay Network Providers	
CohesiveFT	CohesiveFT's VPN-Cubed products are virtual firewallls, switches, hubs, and routers that are used to build overlay networks in clouds, across clouds, and to connect enterprise data centers to public clouds.
Cloud End-Points	
XPack	A dedicated cloud endpoint from Moderro Technologies. A solid state, power saving, VESA mountable desktop appliance with custom desktop environment designed for web applications.

APPENDIX B: Glossary

Application programming interface(API)	An Application Programming Interface (API) is a particular set of rules and specifications that a software program can follow to access and make use of the services and resources provided by another particular software program that implements that API. It serves as an interface between different software programs and facilitates their interaction, similar to the way the user interface facilitates interaction between humans and computers.
Carbon Emissions	Polluting carbon substances released into the atmosphere.
Change Request	A petition for modifying the behavior of a system due to normal business changes or because there is a bug in the system.
Cloud, Cloud Computing	Cloud computing is a general term for anything that involves delivering hosted services over the Internet. These services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS).
Compliance	Compliance is either a state of being in accordance with established guidelines, specifications, or legislation or the process of becoming so.
Data center	A data center (sometimes spelled datacenter) is a centralized repository, either physical or virtual, for the storage, management, and dissemination of data and information organized around a particular body of knowledge or pertaining to a particular business.
Deduplication	Data deduplication (often called "intelligent compression" or "single-instance storage") is a method of reducing storage needs by eliminating redundant data. Only one unique instance of the data is actually retained on storage media, such as disk or tape. Redundant data is replaced with a pointer to the unique data copy.
Fibre Channel	Fibre Channel is a technology for transmitting data between computer devices at data rates of up to 4 Gbps (and 10 Gbps in the near future). Fibre Channel is especially suited for connecting computer servers to shared storage devices and for interconnecting storage controllers and drives.
Governance	Governance describes the mechanisms an organization uses to ensure that its constituents follow its established processes and policies. It is the primary means of maintaining oversight and accountability in a loosely coupled organizational structure.
In-house/On premises	The IT infrastructure managed by the organization utilizing it and is behind the organizations firewall.
iSCSI	iSCSI is Internet SCSI (Small Computer System Interface), an Internet Protocol (IP)-based storage networking standard for linking data storage facilities, developed by the Internet Engineering Task Force (IETF). By carrying SCSI commands over IP networks, iSCSI is used to facilitate data transfers over intranets and to manage storage over long distances.
IT Asset	IT asset is any company-owned information, system, or hardware that is used in the course of business activities.

IT infrastructure	All of the hardware, software, networks, facilities, etc., that are required to develop, test, deliver, monitor, control, or support IT Services. The term IT Infrastructure includes all of the Information Technology but not the associated people, processes, and documentation.
Mainframe	A mainframe is a high-performance computer used for large-scale computing purposes that require greater availability and security than a smaller-scale machine can offer.
Open Source	Open source refers to any program whose source code is made available for use or modification as users or other developers see fit. Open source software is usually developed as a public collaboration and made freely available.
Open Systems	One of a class of computers and associated software that provides some combination of interoperability, portability, and open software standards, particularly Unix and Unix-like systems.
Operating expenses (OPEX)	A category of expenditure that a business incurs as a result of performing its normal business operations. One of the typical responsibilities that management must contend with is determining how low operating expenses can be reduced without significantly affecting the firm's ability to compete with its competitors.
Return on investment (ROI)	A measure of a corporation's profitability, equal to a fiscal year's income divided by common stock and preferred stock equity plus long-term debt. ROI measures how effectively the firm uses its capital to generate profit; the higher the ROI, the better.
Service Provider	A company that provides a specific service or services.
Snapshot	A copy made of a disk drive at a specific moment in time. Snapshots are useful for backing up data at different intervals, which allows information to be recovered from different periods of time.
Total Cost of ownership (TCO)	Total of direct capital investment in hardware and software plus indirect costs of installation, training, repairs, downtime, technical support, and upgrading. Also called cost of ownership or ownership cost.

APPENDIX C: References

1. John W. Rittinghouse and James F. Ransome; Cloud Computing Implementation, Management, and Security
2. Anthony T. Velte, Toby J. Velte and Robert Elsenpeter; Cloud Computing: A Practical Approach
3. Cloud Computing Use cases whitepaper by Cloud Computing Use case discussion group
4. Top Threats to Cloud Computing V1.0, Prepared by the Cloud Security Alliance,March 2010
5. Gartner Press release: “Gartner Survey Shows Data Growth as the Largest Data Center Infrastructure Challenge”
6. Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431, U.S. Environmental Protection Agency, EPA ENERGY STAR Program
7. Security Guidance for Critical Areas of Focus in Cloud Computing V2.1, Prepared by the Cloud Security Alliance
8. Cloud Storage Security with a Focus on CDMI by Eric A. Hibbard, SNIA Education
9. Cloud Security and Privacy by Tim Mather, Subra Kumaraswamy, and Shahed Latif
10. 10 applications you can move to the cloud – Justin James, Nov 4th 2010, TechRepublic Blogs,
11. Air Distribution Architecture Options for Mission Critical Facilities, APC Whitepaper #55 By Neil Rasmussen
12. Cooling Options for Rack Equipment with Side to Side Airflow, APC Whitepaper # 50 by Neil Rasmussen
13. Estimating a Data Center's Electrical Carbon Footprint, APC Whitepaper # 66, by Dennis Bouley
14. Ten Cooling Solutions to Support High-Density Server Deployment, APC Whitepaper # 42 by Peter Hannaford
15. Telecommunications Infrastructure Standard for Data Centers, ANSI/TIA - 942, Chris DiMinico
16. Multi Root I/O Virtualization and its Potential to consolidate I/O Infrastructures, FSC TEC-Team, Bernhard Schrader
17. Standardization and Modularity in Network-Critical Physical Infrastructure, APC White Paper #116, By Suzanne Niles

18. Analysis of EPA Report to Congress (Law 109-431) By Greg Schulz
19. Hot Aisle vs. Cold Aisle Containment, APC Whitepaper#135 by John Niemann
20. The Efficient, Green Data Center, EMC Whitepaper by Raza Syed
21. New Normal of Datacenter, HCL Whitepaper
22. Data center Standards Overview, ADC Whitepaper
23. Data Center Bridging Version 1.0, November 2008 by Steve Garrison, Val Oliva, Gary Lee and Robert Hays
24. IO Virtualization, Whitepaper by Virtensys
25. PCI Express and IOV: Maximizing Multi-Processor Systems by Shreyas Shah, PLX Technology, Inc
26. Storage Virtualization Seminar, TechTarget Storage Media by Marc Staimer
27. en.wikipedia.org for various terms and definitions
28. Various other whitepapers, press releases and magazines