# 5 Cloud Computing

## 5.1 Introduction

*"If computers of the kind I have advocated become the computers of the future, then computing may someday be organised as public utility just as the telephone system is a public utility. The computer utility could become the basis of a new and important industry." – John McCarthy 1961*[1].

McCarthy's vision took almost half a century to realise. Today, computing as a public utility takes the form of cloud computing and epitomise how businesses demand IT be delivered. The National Institute of Standards and Technology (NIST) defines Cloud Computing as "*a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."*

Cloud computing has certainly captured a great deal of attention and following over the past three years. IT giants like HP, IBM, Intel and Microsoft have evolved their business strategies around cloud computing. Traditional software houses like Microsoft and SAP are offering their highly successful software suites as cloud services to address growing customer demand for utility based charging and collaboration. "cloud computing" is reiterated on a daily basis in both IT and business news. The theme was mentioned in varying degrees in 71 of 77 Gartner's hype cycles for 2011 with 3 of them dedicated to different aspect of cloud computing. Forrester Research estimated the global market size for cloud computing to be US$241billion by 2020. Public cloud will constitute two-thirds of that market size, or US$159.3billion.

## 5.2 Market Trends

### 5.2.1 Sheer Volume of Content Transferred

The sheer volume of content to be transferred, due to ubiquity of access and demand for media, will strain any unoptimised content delivery mechanism. Since online presence is critical to businesses today, the need to deliver contents quickly and in time to their customers will continue to be a priority.

Nokia Siemens Networks predicts that broadband data will increase 1,000 fold by 2020. Cicso also projected 90% compound annual growth rate (CAGR) of global traffic for video and 65% for data through 2016. Both projections outpace Moores Law so the delivery of content in the long term cannot be fulfilled simply by buying better servers. Content delivery and server utilisation must be optimised to satisfy that demand for information.

The types of content transferred will also expand with different sources like intelligent, networked devices (Internet of Things) and contextual data, to processed information and interactive media (e.g. Internet TV). By 2014, 50% of all data would have been processed by cloud. Cloud computing and the global distribution of public clouds can provide that last mile delivery.

[1] Harold Abelson. Architects of the Information Society Thirty Five Years of the Laboratory for Computer Science at MIT. U.S.A.: Wonder Book Publishing; 1999.

### 5.2.2 Demand for “Instant Performance”

Social analytics is increasingly common as a business strategy to better understand and respond to the customer as an individual. The volume of data adds to the wealth of electronic records waiting to be analysed and combined for intelligence and better insights into the business. Businesses demand such insights with increasing immediacy.

Demand for “instant performance” extends beyond the processing of data. Where IT could take months to provision a single server, it now faces competition from cloud and other hosting providers that can do the same in minutes. It is not uncommon to hear stories of business units bypassing IT because public hosting and cloud providers can fulfil their requirements quicker.

Casual supercomputers fulfill both the demand for faster turnaround time and a reduction of the cost of ownership. These compute behemoths are assembled solely for the compute task, are flexible in the amount of performance required, and are released once the task is completed. Cycle Computing made news with its 30,000 core cluster at a cost of US$1,279/hour. The cluster was constructed on-demand on Amazon Web Services and was torn down eight hours later.

### 5.2.3 Seamless Experience from Personal Cloud and BYOD

In Singapore, 30% of organisations have implemented cloud computing to address the need arising from consumer devices. The ubiquity of an untethered Internet and affordability of powerful, easy-to-use devices has driven productivity as executives stay connected to the company anytime, anywhere and using any device.

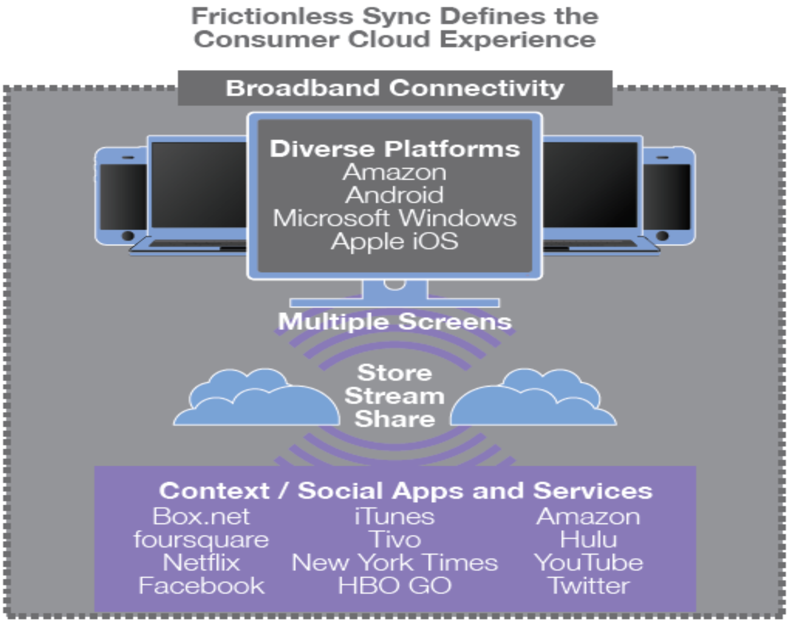


**Figure 1**: Frictionless Sync Defines the Consumer Cloud Experience

As with the proliferation of the Internet, the advent of personal clouds represents a tectonic shift in the personalisation of IT. The seamless experience and integration of data, social and communication services into a single smart device will be the new benchmark for corporate IT offerings.

Smarter mobile devices are replacing the laptop in providing a more convenient way to access information. It is not uncommon today to find business executives preferring the more mobile tablet devices to a full-featured laptop. Consistency of data across different devices will be a challenge IT departments need to resolve.

Today, Android devices pull configuration and personal data from a user's Google account. In a similar fashion, IOS devices share information using Apple's iCloud; user data stored in dropboxes can be accessed from any smart device. Automatic synchronisation of data across personal devices dilutes lockdown to a single device. Expectation of the same convenience will force IT to rethink what personal computing entails. Experience with the benefits of such personal cloud services will increase both demand and acceptance of such services in corporations. Cloud services provide a natural avenue for enterprises to support the resulting myriad of devices and services.

Advances both in technology and processes will address the current concerns and shortcomings of cloud computing, e.g. reliability, interoperability, security and privacy, service level enforceability, and predictability of charges. Such integration of technology with process and policies, together with adoption of cloud standards, will position cloud computing for wider appeal.

Insatiable compute demand of an ever data hungry world brought about by ubiquitous and faster connectivity, the demand for “instant performance”, faster, more sophisticated analytics and more connected businesses, and extrapolation of the seamless integration provided by personal clouds with the personalisation of IT will stimulate cloud adoption.

### 5.2.4 An Arduous Journey toward Cloud Computing

Cloud computing does not just waive the cover charges. It outlines the underlying architectures upon which services are designed and applies equally to utility computing and internal corporate data centres. Cloud computing evolved from a myriad of technologies including autonomic computing, grid computing, multi-tenancy, service oriented architecture (SOA) and (network, server and storage) virtualisation. It abstracts design details from the cloud user, presenting compute as an on-demand service. This section reminisces on the passage of the vision of “Compute as a Utility” to its realisation as “Cloud Computing”.

Since McCarthy's lecture in 1961, several notable attempts have been made to redefine how compute is used and delivered. The first virtual machine appeared with IBM's CP/CPM that was productised as VM/370 in 1972. Back then, the issue of multi-tenancy was addressed either as an integrated time-shared system with sophisticated segregation of privileges between users, or with each user having his own computer walled within a virtual machine.

By the mid 1980s, general interest in compute as a utility dipped to a low with the advent of affordable personal workstations that fulfilled most compute demands. Grid computing, coined in early 1990s, revived the concept that “compute should be accessible like an electric power grid”. Landmark projects like SETI@home and Globus Toolkit in 1999 and 1998 respectively laid the groundwork for tapping unused compute resources and synchronising these compute jobs. Grid computing was the precursor to coordinating compute in the cloud.

By 2000, grid computing had taken off in research and development (R&D). IT vendors like IBM, HP and Sun Microsystems started offering grid computing services. Most notably, Sun Microsystem started marketing Sun Cloud[2] for US$1 per CPU/hour. This was the first time compute was available commercially as a utility on a global scale. In a controversial article published in the Harvard Business Review in 2003, Nicholas Carr declared, "IT does not matter". He posited that IT was becoming commoditised and would be delivered like other utility services.

Virtualisation technologies started gaining traction in 2005 as a means to improve data centre efficiencies by consolidating workloads. Network, server, and storage virtualisation providers collaborated to deploy autonomous technologies that enabled rapid provisioning of services in a virtualised data centre environment. This paved the way for the internal corporate data centre to transit to cloud computing.

In 2006, Amazon launched its Elastic Compute cloud (EC2) and Storage (S3) services that offered compute and storage rental with two distinct features. These services use a pricing model that charged "per use", and services were provisioned (and released) within minutes of payment. Computing was now accessible as a utility.

Amazon was soon followed by Google Apps that offered an office suite as a service, Google App Engine that provides a J2EE platform charged by CPU cycle, Microsoft's Azure platform service, and a flurry of web hosting providers like Rackspace and 1and1[3]. There was just as much interest in providing software that enables enterprises to run cloud services in their internal data centres. Examples include Joyent's SmartOS, 3tera's AppLogic, Microsoft's Windows Azure, and OpenStack.

In this surge of technologies, it is easy to forget that cloud computing is not entirely a technology play. Many other adjacent advances, especially in IT management, are necessary for a successful cloud deployment. Efficiencies gained must be balanced against the loss of control. The relationship between IT organisations, their technology partners and IT's customers, must be managed through an effective IT governance structure.

After an arduous 50 years of attempts with varying degree of success, John McCarthy's vision to organise compute as a public utility has finally been realised with cloud computing. The pervasion of IT into businesses and personal lives conduced the demand and reliance on IT, and the corresponding availability of IT expertise, into a global desire for compute on demand. Additionally cloud computing allows data centre owners to realise many of its benefits by facilitating clouds to be built within their facilities. A vision born when computing resources were scarce and costly will be tested against today's demand for increasingly instant results.

[2] Wikipedia. Sun Cloud. [Online] Available from: http://en.wikipedia.org/wiki/Sun_Cloud [Accessed 9th July 2012].

[3] Linda Leung. More Web Hosts Moving to the Cloud. [Online] Available from: http://www.datacenterknowledge.com/archives/2010/01/27/more-web-hosts-moving-to-the-cloud/ [Accessed 9th July 2012].

### 5.2.5 Cloud Computing Today

The National Institute of Standards and Technology (NIST) defines cloud computing as *"a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."*

### 5.2.6 Cloud Service Models

NIST broadly categorises clouds into three service models:

- **Cloud Software as a Service (SaaS).** Consumers are given access to the provider's applications that runs on a cloud infrastructure. Examples of SaaS include Google's GMail, Microsoft 365 and Salesforce.com. Consumers of these SaaS access the applications using a variety of clients such as a Web browser or even a mobile application. Management of infrastructure, operating environment, platform services, and application configuration are left to the cloud provider.

- **Cloud Platform as a Service (PaaS).** Consumers are given access to platform on which they can develop their custom applications (or host acquired applications). Google AppEngine, Microsoft Azure and Force.com are examples of PaaS. Consumers of PaaS launch their applications using the specific programming platforms supported by the specific PaaS. The PaaS provider takes care of delivering the programming platform and all underlying software and hardware infrastructure.

- **Cloud Infrastructure as a Service (IaaS).** Consumers are given an operating system instance on which they can install software and set up arbitrary services and applications. The IaaS provider takes care of the server hardware and network, usually using a virtualised environment. Responsibility of maintaining the operating system usually falls on the consumer.

The division of responsibilities between the provider and consumer for each of these service models compared against a virtualised traditional IT environment is illustrated in **Error! Reference source not found.**.

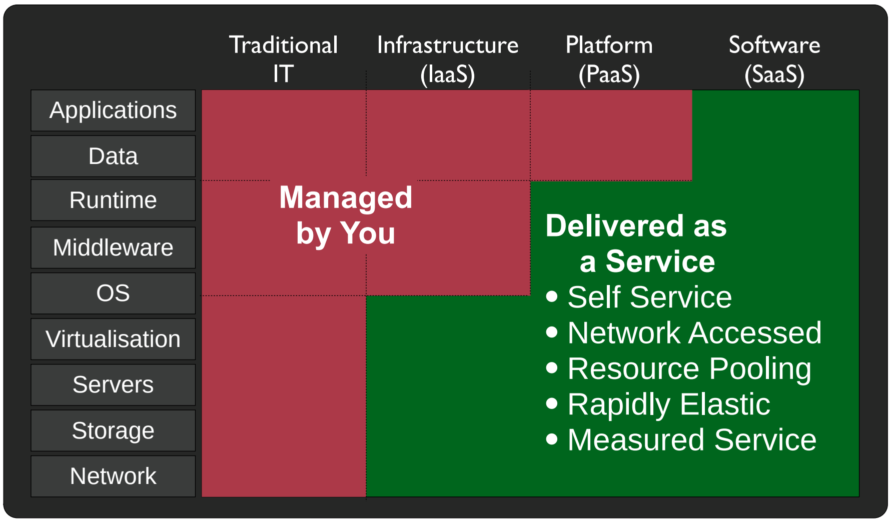


**Figure 2**: Division of Responsibility by Service Models

Beyond these three service models, numerous cloud services have emerged. Notable cloud offerings include Security as a Service, Data as a Service, Desktop as a Service, Storage as a Service, Communications as a Service, Database Platform as a Service and Service Delivery Platform as a Service. These offerings may be viewed as a specific implementation of the three Service Models depending on the level of involvement of the cloud service consumer.

### 5.2.7 Cloud Deployment Models

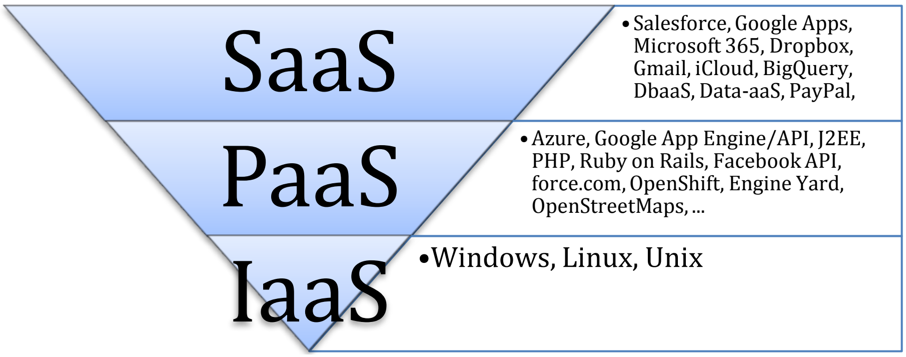
Cloud deployment affects the scale and hence, efficiency, of the cloud implementation.

- **Private Cloud** is a cloud infrastructure operated solely for a single organisation. Such single tenant clouds may be managed by the organisation or a third party and may be hosted within the organisation's premises or in a third party data centre.
- **Public Cloud** is a cloud infrastructure operated by a cloud provider that is available for public consumption. These multi-tenant clouds serve a variety of customers and usually enjoy the largest scale and utilisation efficiency. Amazon Web Services and Microsoft Azure are two well-known public cloud providers.
- **Community Cloud** is a public cloud infrastructure serving a specific industry or community that share a common trait or set of concerns (e.g. security and compliance requirements, or a certain common application). An example is Sita's ATI Cloud that provides airline employees online access to infrastructure, desktop and other services.
- **Hybrid Clouds** are clouds that deployed across two or more cloud deployment models. Successful hybrid cloud implementation requires integration that enables data and application portability between the different cloud services. The most common hybrid clouds are composed of private and public clouds where workload is overflowed from the private cloud into the public cloud.

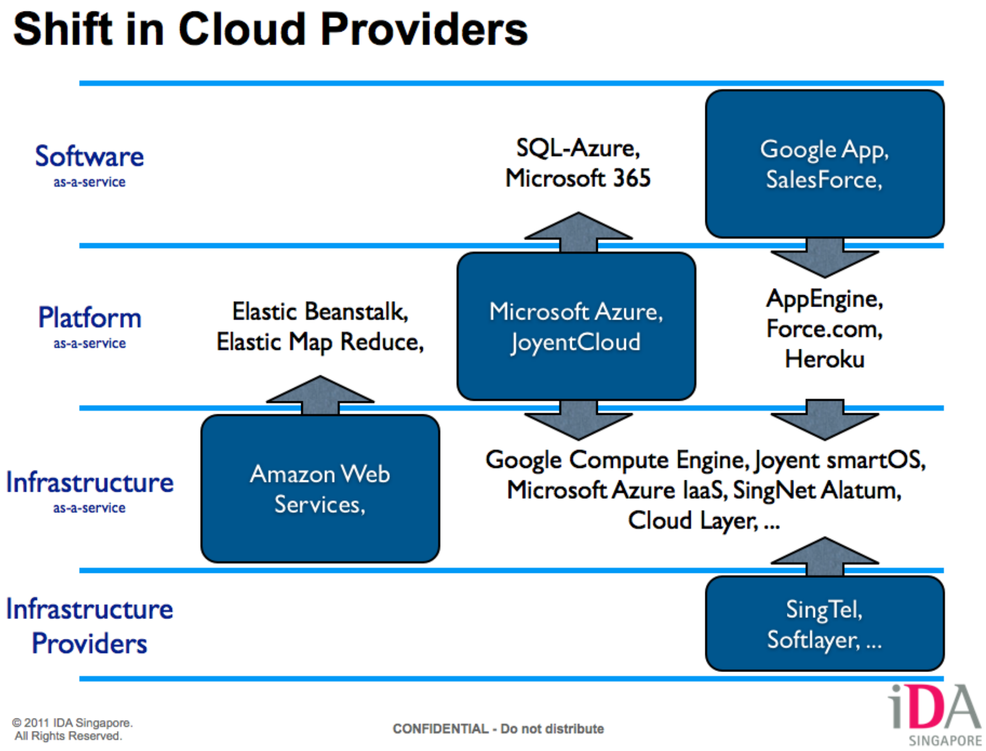
### 5.2.8 What is available in the Public Cloud today

Most private clouds today are IaaS although enterprises who have standardised their technology architectures may provide PaaS to their developers. The public cloud is experiencing tremendous growth and is forecasted to represent two-thirds of the market by 2020.

It is not hard to imagine that the diversity of offerings from public cloud providers has “exploded” from a handful of IaaS to hundreds of PaaS to thousands of SaaS. The consolidation of operating systems has gravitated most cloud providers toward variants of Microsoft Windows or Linux, with a handful delivering other Unix variants (e.g., Oracle’s Solaris).


SaaS
• Salesforce, Google Apps, Microsoft 365, Dropbox, Gmail, iCloud, BigQuery, DbaaS, Data-aaS, PayPal,
PaaS
• Azure, Google App Engine/API, J2EE, PHP, Ruby on Rails, Facebook API, force.com, OpenShift, Engine Yard, OpenStreetMaps, ...
IaaS
•Windows, Linux, Unix

The PaaS offerings consist of either popular middleware platforms (e.g. Microsoft’s .NET framework, J2EE, and Rails), or programmatic extensions from successful SaaS offerings (e.g force.com, and Google App Engine).



Shift in Cloud Providers

Software as-a-service

SQL-Azure, Microsoft 365

Google App, SalesForce,

Platform as-a-service

Elastic Beanstalk, Elastic Map Reduce,

Microsoft Azure, JoyentCloud

AppEngine, Force.com, Heroku

Infrastructure as-a-service

Amazon Web Services,

Google Compute Engine, Joyent smartOS, Microsoft Azure IaaS, SingNet Alatum, Cloud Layer, ...

Infrastructure Providers

SingTel, Softlayer, ...

© 2011 IDA Singapore. All Rights Reserved.

CONFIDENTIAL - Do not distribute

iDA SINGAPORE

A noticeable trend is for public cloud vendors to expand beyond their traditional service models into adjacent service models as depicted below:

- Salesforce.com expanded from a SaaS provider for CRM to provide programmatic application programming interface (API) through force.com, and later acquired Heroku PaaS in 2011;
- Google expanded from a SaaS provider (Google App) who expanded into PaaS with its Google App Engine in 2008, and then into IaaS with Google Compute Engine in 2012;
- Amazon expanded from a very successful IaaS into platforms with its Elastic Beanstalk; and
- Microsoft's Azure expanded from PaaS into IaaS. It is also pursuing enterprises with its Microsoft 365 SaaS solution.

Apparent from the above examples, the expansion to adjacent service model builds on the existing success of the cloud providers. The expanded service model is tightly integrated vertically with the provider's existing services. Elastic Beanstalk from Amazon offers its customers new capabilities and the added convenience of standard platforms. Google and Salesforce's expansion into PaaS provides a programmatic interface that affords their customers better flexibility.

## 5.3 Cloud Economics: Scale and Elasticity

Notwithstanding technological implementations, the advent of cloud computing changes the economic landscape of computing by availing compute on an unprecedented scale. The utility prices of cloud computing are becoming the benchmark for enterprise IT forcing corporations to rethink the value proposition of running their IT infrastructure.

### 5.3.1 Scale Matters

A large cloud setup takes advantage of significant economies of scale in three areas:

- **Supply-side savings** in cost per server;
- **Demand-side aggregation** increases utilisation by smoothing overall compute demand variability; and
- **Multi-tenancy efficiency** distributes application management and server costs to more tenants.

Consider Microsoft's 700,000 square-foot Chicago Cloud Data Centre. The facility currently houses 224,000 servers in 112 forty-foot containers, and has a maximum capacity for 300,000 servers. This US$500 million facility is operated by a skeleton crew of only 45.

On such astronomical scale, Microsoft reduces the costs of power both by negotiating a favourable bulk purchase price and by locating the facility to reduce cooling power loads using ambient air. Furthermore, standardisation and automation allowed Microsoft to operate the entire facility with a minimal crew reducing labour cost. New servers are ordered by containers of 1,800 to 2,500 servers allowing Microsoft to enjoy massive discounts over smaller buyers.

Single-tenant environments achieved an average utilisation of less than 20%. Compute resources are generally provisioned for peak demands and fail to sustain a high utilisation because of the random nature of workload, and time or seasonal peaks depending on the organisation’s locality or industry. In addition, workloads differ greatly in resource profiles making it even harder to optimise resource utilisation.

Scale vastly improves the utilisation of available compute resources. Large public cloud operators are able to maintain an average utilisation of around 90% without violating their service level agreements (SLAs). Operating a massive cloud allows the cloud provider to aggregate compute demands from various geographical diverse clients into a shared compute resource pool. This aggregation soothes the variability of compute demands from individual clients by offsetting peak demand from one client with low demands from others. Further aggregation can occur if clients span different industries with different seasonal demands. The very act of consolidating all these demands pushes utilisation beyond what can be achieved on a smaller scale.

Finally, multi-tenancy of a centrally managed application or platform improves cost efficiency. Management costs can be distributed across a large number of customers instead of being borne per customer in a single-tenant environment. Hosting more application instances amortises server overhead and the resulting savings can be passed on to customers.

Supply-side savings reduces the cost of ownership per unit of server as the cloud provider gains bargaining power because of their scale. Demand-side aggregation and multi-tenancy efficiencies optimise the total costs of ownership further by drastically improving utilisation. Large public cloud providers with 100,000 servers enjoy an 80% lower total cost of ownership (TCO) compared to their smaller counterparts with 1,000 servers (See **Error! Reference source not found.**).
The ubiquity of the Internet has extended the reach of cloud providers to a global market of cloud consumers. The aggregated compute requirements from this customer pool are multi-fold that of any single enterprise. This massive consolidation of compute allows cloud providers to operate at unprecedented efficiencies and benefit from an equally unprecedented economy of scale.

## 5.3.2 Elasticity Changes the Game

A key benefit of cloud computing is its ability to match computing resources closely to workload by adjusting resources at a pre-defined granularity (e.g. adding or removing server instances with EC2). This is accomplished within minutes for a cloud environment, compared to days or weeks in traditional IT, allowing for very fine grain adjustments that accomplish unprecedented utilisation.

This shift of compute cost from a capital expense to operating expense has a far ranging effect on cloud consumer behaviour. Already, examples of how industry leaders are changing their game with elasticity abound:

- In cloud computing, 1,000 CPU-hour costs the same if you use 1 CPU for 1,000 hours, or 1,000 CPU for an hour. Pixar Animation Studios takes advantage of this aspect of elasticity to reduce turnaround time needed for rendering their movie.
- When Zynga launched FarmVille, it expected only 200,000 daily active users within the first two months. Instead, the game proved a runaway success and gathered 1 million users every month. Zynga’s choice to implement the game on an elastic

cloud saved them from certain embarrassment and allowed their game to scale as the number of users grew rapidly.

- Cycle Computing built a 30,472 core cluster on 3,809 AWS compute instances by salvaging unused "spot instances". The "super computer" lasted seven hours and at its peak costs only US$1,279 per hour.
- The New York Times used 100 Amazon EC2 instances to process 4TB of TIFF into 11 million PDF documents within 24 hours and under US$240.

The above examples are the quintessence of leveraging cloud elasticity to avoid high upfront investment in hardware, reduce recurring costs of over-provisioned network bandwidth, reduce turnaround time to deal with unexpected spikes in compute requirements, and manage costs as demand falls. Cloud elasticity ultimately reduces business exposure and increases business agility to market changes.

Whereas scale affects the bottomline of business by bringing down costs, Cloud elasticity impacts both the bottomline by using IT budgets more efficiently and the topline of business by allowing low-risk experimentation. In the long term, IT agility enabled by cloud computing will transform the business landscape.

## 5.4 <u>The Future with Cloud Computing</u>

*In "The Big Switch", Nicolas Carr pointed to the business of supplying computing services over the Internet from MySpace, Facebook, Flickr, to YouTube, Google Docs, and Box to describe his vision of a "World Wide Computer":*
*"All these services hit at the revolutionary potential of the new computing grid and the information utilities that run on it. In the years ahead, more and more of the information-processing tasks that we rely on, at home and at work, will be handled by big data centres located out on the Internet. … If the electric dynamo was the machine that fashioned twentieth century society – that made us who we are – the information dynamo is the machine that will fashion the new society of the twenty-first century." – The Big Switch, Nicolas Carr.*

The machine will deliver a compute experience that is seamless and integrated. The experience extrapolates from the current packaged cloud services like the iCloud, Dropbox and Google Search. In this computing nirvana, the machine automagically coordinates data from different sources, invokes software services from different software providers, and orchestrates hardware and software compute platforms in the most optimised fashion. Security policies on data and intermediate results will impose constraints on the orchestrated services. Service levels will be directly translated from business requirements and will be used to identify resources necessary to deliver the required performance. Workload and usage profiles can be automatically learnt and fed back into the resource placement strategy.

Overarching these components is the cloud brokerage service that coordinates the resources and automatically reconfigures these resources to best match the security and service level specifications of the user. The broker leverages common programmatic interfaces to query the status of each cloud and to deploy tasks into different clouds. Such brokers will be a service in the "World Wide Computer".

Implementation trends today provide a glimpse of this future. The recent interest in Big Data saw cloud vendors like Amazon and Microsoft deliver turnkey Hadoop services in their respective clouds. The industry can look forward to more turnkey services that will increase the lure of cloud computing beyond just IaaS.

Such a vision may still be decades away but the various pieces are coming together today. The following table illustrates the current maturity of each resource:

| ***Resource/ Component*** | ***Current State/ Noteworthy work*** | ***Challenges*** |
|---|---|---|
| ***Services***<br>*Services will be the way application and software will be presented in the machine. Services will be physically detached from the data and compute. The cloud customer will be able to invoke any software components from different software providers either as a pre-programmed package or in an ad-hoc fashion* | *Software is available as a turnkey VM.*<br>*Different compute needs like Hadoop and databases are becoming available as a service (e.g. EMR)*<br>*Business Software as a service e.g., Google App is gaining acceptance*<br>*Consolidation of OS platforms and common libraries* | *Software requires specific platforms*<br>*Specific schema is needed due to presumption in data*<br>*SaaS tend to be single monolithic blocks that is difficult to integrate to other services*<br>*Different charging model and "cloud readiness" of enterprise software*<br>*Most SaaS deployments tend to ignore security frameworks and requires different handling* |
| ***Data***<br>*Data will be available to software services in a variety of formats with rich description of purpose and intent. Such data can be structured or unstructured and without restriction on its proximity to the task* | *Common data store is available within the cloud provider's environment but not inter-cloud.*<br>*Some common formats like XML, JSON, and SQL like access API*<br>*Distributed file systems with common namespace, e.g. HDFS, Global FS.* | *Rich data description*<br>*Common Access methods*<br>*Data distribution*<br>*Location of data is generally not well controlled*<br>*Removal of data cannot be guaranteed* |
| ***Compute***<br>*Compute will be orchestrated per task and optimised on the fly based on the location of the data, application software, and security specification. The cloud will automatically figure out the appropriate platform and sizing to achieve the specified compute response time and service level* | *IaaS is generally well defined and understood*<br>*Automatic and live migration of compute is increasingly adopted.* | *Platforms are still evolving and fluid*<br><br>*Current restrictions around interoperability and performance across long distances.* |
| ***Networks***<br>*Networks will be orchestrated around security policies and service level requirements. A subnet as we know it today,* | *The Internet now connects most of the world and permeates through fixed and mobile connections.* | *Continued demands for higher speeds and better reliability especially across longer distances* |

| | | |
|---|---|---|
| *will span multiple cloud providers* | *IPv6 provides addresses to the next wave of devices and ensure end-to-end connectivity*<br>*Lossless networks simplify connectivity by converging both storage and network.* | *Software defined networks are still not widely adopted.*<br>*Lack of controls on performance of partner networks.* |

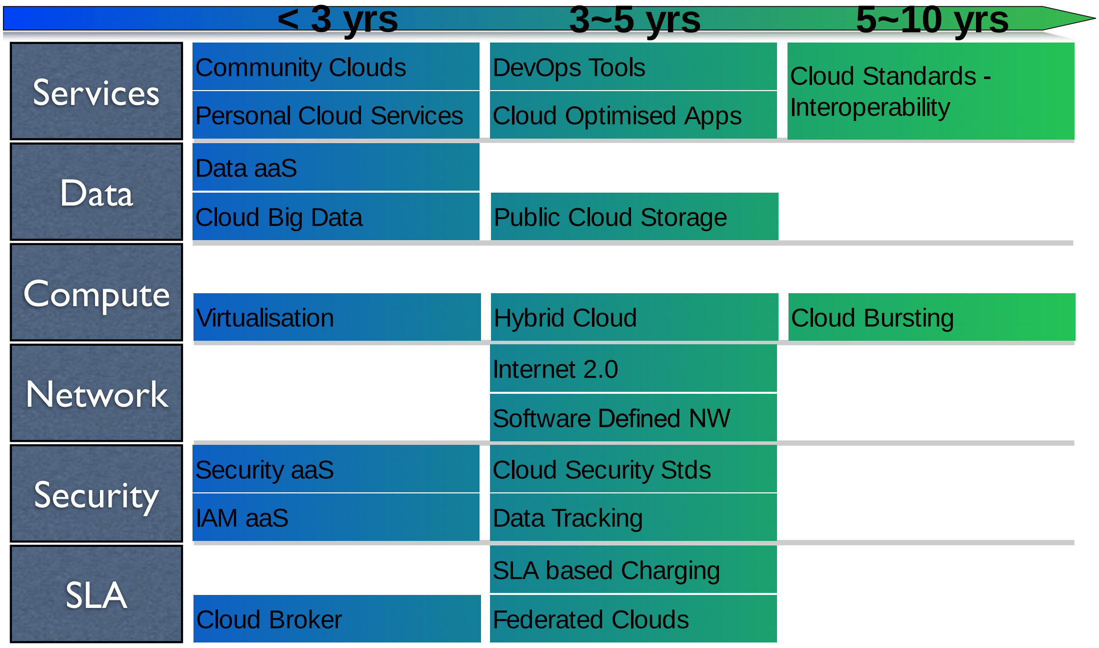
| ***Resource/ Component*** *(cont.)* | ***Current State/ Noteworthy work*** *(cont.)* | ***Challenges*** *(cont.)* |
|---|---|---|
| ***Security***<br>*Security policies that prescribes confidentiality, integrity and availability.*<br>*Legal boundaries will be described and enforced* | *Standards are still evolving and generally immature*<br>*Disaster Recover practices are adopting cloud* | *Walled boundaries for data (legal or classification)*<br>*Programmatic way to tag security classification of cloud providers*<br>*Security policies, practices and responsibility divide of user/provider is uncertain* |
| ***Service Level***<br>*Service Level requirements will identify compute resources needed to deliver the required performance, based on the specific workload profile* | *Standards are still evolving and generally immature*<br>*SLA is generally specified by cloud provider*<br>*Responsibility to mitigate outage is usually left to cloud user.* | *There need to be better ability to control SLA from a service perspective (e.g. coordinate across cloud providers to reach required SLA)*<br>*Fine grain performance requirements (e.g. response time) that can be observed* |

Most of the challenges identified above have been addressed to a limited degree especially within a more established cloud provider environment. The most notable examples are Google's integration of Google Compute Engine (IaaS), Google App Engine (PaaS) and its application services like Google Drive, Google Fusion Tables and Google Map Services. Single cloud provider implementations alleviate the pain of coordinating multiple cloud services.

Nicolas Carr's vision of a "World Wide Computer" points us to an interoperable Cloud of Clouds. A Cloud that marries services, compute and networks from different cloud providers, and that automatically arrange resources available to meet security and service levels requirements.

## 5.5 Technology Outlook

In line with the vision of a seamless, automatic and service-oriented cloud, several technologies in the pipelines are ripe for adoption in the foreseeable future.



|  | < 3 yrs | 3~5 yrs | 5~10 yrs |
|---|---|---|---|
| Services | Community Clouds<br>Personal Cloud Services | DevOps Tools<br>Cloud Optimised Apps | Cloud Standards - Interoperability |
| Data | Data aaS<br>Cloud Big Data | Public Cloud Storage |  |
| Compute | Virtualisation | Hybrid Cloud | Cloud Bursting |
| Network |  | Internet 2.0<br>Software Defined NW |  |
| Security | Security aaS<br>IAM aaS | Cloud Security Stds<br>Data Tracking |  |
| SLA | Cloud Broker | SLA based Charging<br>Federated Clouds |  |

**Figure 3**: Technology Adoption Map

The Technology Adoption table above illustrates technologies according to the resources and requirements identified earlier. Adoption is generally defined as the moment when accelerated deployment of the specific technology is undertaken by its target consumers. The timeline of adoption is estimated as relatively immediate (less than three years), in the mid term (three to five years) and in the longer term (five to 10 years).

The following sections are organised according to this timeline and discusses each technology in greater detail.

### 5.5.1 Less than three years

Most of the technologies that will be adopted within the next three years are already showing signs of maturing today. Examples of deployment of these technologies already exist even though they might not have hit the main stream.

#### 5.5.1.1 Community Clouds

**Resource/Requirement:** Service

NIST defined the community cloud as a deployment model. A community cloud infrastructure is shared by several organisations and supports a specific community that has shared concerns (e.g. mission, security requirements, policy, and compliance considerations). A community cloud will target a limited set of organisations or individuals.

The key to understanding the community cloud is to comprehend its ability to address the concerns of the community it serves in unity. Compliance of such clouds is scalable and applied to all organisations adopting it. As a result, such cloud deployments will appeal to organisations with well defined security and compliance requirements, e.g. government clouds.

Today, pockets of community clouds have already evolved. Notably, the Singapore Government's G-Cloud aims to address common compute needs of local government agencies by providing different security levels that closely match the tiered architecture commonly found in Web services provided by the respective agencies. Another example is SITA's ATI Cloud. It aims to provide a set of applications to airline customers while meeting compliance and standards required of the airline industry.

While today's community cloud may leverage public cloud infrastructures, private clouds remains the cornerstone of most implementations. Meeting the compliance requirements of specific communities are the primary reason for such deployments. Delivering these services using private clouds mitigate many of the challenges posed by public clouds.

Multi-tenancy of community clouds will create value for the cloud. Sharing of data between different members of the community can bring about huge benefits that was not possible in the previously silo'ed approach. One such example is the health care community cloud. Shared health data provides convenience to patients and allows pre-emptive health management. The universal view of health data improves the management of clusters of contagion. While it is possible to share data without cloud, building the service around a community cloud prevents the natural fragmentation of data early in development. The cloud implementation also provides for better capacity management and growth.

Central to the implementation, however, is the fact that the community cloud provider is really a brokerage service that has mapped the compliance requirements into a resource pool. With the extended capabilities of cloud brokerage services, smaller communities will be able to provide their unique cloud offerings.

**Enablers**

Many organisations operating within a well-defined vertical already share information using some SOA frameworks. The movement to community cloud typically involves an industry leader establishing a consortium that will move specific operations into the cloud. Examples of such leaders include SITA and Changi Airport.

Community clouds leverage on high commonality in data, tools, compliance, and security requirements, enhancing collaboration and coordination of business processes between participating organisations.

### 5.5.1.2 Personal Cloud Services

The proliferation of smarter devices and ubiquity of mobile connectivity enabled these personal devices to participate in delivering cloud services. Indeed, personal cloud services extends availability of data in the cloud, including personal data (e.g. contacts and personal preferences); messaging services (e.g. email and electronic chat); and media (e.g. music and videos), into the personal devices to create a seamless experience regardless of where or how it is accessed.

Google's Android and Apples IOS devices are connected to their user's Google and Apple account to provide automatic data synchronisation. This decouples the user's data from their devices making the latter a portal to the data that is stored in the cloud.

The resultant experience changes the way we work and play. A calendar SaaS may provide synchronisation to a personal device. When the device participates in SaaS, location information from the device can augment information about a meeting to provide a context-aware reminder. The travel time to the meeting venue can take into account traffic information and the user can be informed when it is time to leave for the next meeting and the best route given the traffic conditions.

**Enablers**

Personal cloud services are enabled by a ubiquitous mobile Internet and smarter consumer devices. The miniaturisation of sensors enabled a myriad of sensors to be packed into the mobile device. These sensors range from simple magnetic field sensors to multi-dimensional accelerometer to sophisticated GPS systems. The device software provides access to these sensors through applications.

### 5.5.1.3 Virtualisation

Virtualisation technologies started maturing as far back as 2005 but only saw mainstream adoption in Singapore as late as 2010. This class of technologies will continue to gain traction in enterprise datacentres and with hosting providers, and is a prerequisite of cloud computing.

Virtualisation is not restricted to the well-discussed server virtualisation though and has found its place in various aspects of enterprise computing for many years now. Storage virtualisation is a matured technology that was restricted to high-end storage systems like the Hitachi Data Systems 9900 and has since trickled down to lower cost storage systems. In addition, network equipment has provided virtualisation in the form of virtual local area networks (VLANs) since early 2000.

Advances in virtualisation management software provide a tight integration between storage, compute and network. VLANs are pushed into the hypervisor layer of the virtualised environment to improve operational efficiency by allowing virtual machines (VMs) from different subnets to coexist in a single physical host. Network administration is automated with the hypervisor provisioning the correct VLAN when the VM is started. At the same time, storage is provisioned as VMs are created. Software images are cloned instantaneously from a reference pool.

Virtualisation is precursor and a key building component of cloud computing. There is a growing trend for software vendors to deliver their software in a standard VM instance (e.g. in OVF format). Examples of traditional software delivered as virtual appliances are Oracle Database Templates, Ruby on Rails, SugarCRM, and even security products like Barracuda firewall. Cloud vendors can provide turnkey instances of these appliances that their users can simply fire up, avoiding the complexity of installation and setup and hence, reducing support cost.

A noteworthy trend today is the availing of converged virtualised architectures where hardware providing compute, storage, and network, and software providing virtualisation and management are delivered as an integrated solution. The entire stack of software and hardware are prequalified to interoperate with each other. Such a unified architecture greatly speeds up deployment and operational management.

Another virtualisation technology experiencing growth in the cloud is Virtual Desktop Infrastructure (VDI). These virtual desktop instances enjoy ease of management, better security and rapid provisioning compared to their physical counterparts. Korea Telecom (now known as KT Corporation) took advantage of their proximity to enterprise customers and offered VDI services hosted in their cloud for US$25 per user per month.

Server and storage virtualisation are well adopted as of this writing.

### 5.5.1.4 Security-as-a-Service

Security-as-a-Service is the delivery of cloud-based services that aids to improve security or governance of the consuming organisation. Security services ranges from Web or e-mail content filters, file encryption and storage, and logfile analysis, to intrusion detection, vulnerability assessments, and disaster recovery, to security policy and governance enforcement frameworks. The services may be consumed in another cloud, in an enterprise data centre, or directly in end-user devices.

Cloud-based security controls services are not new. Services have already been in mainstream adoption for years and include e-mail and distributed denial of service (ddos) detection and prevention. The benefits of delivering these services in the cloud, compared to their enterprise counterpart, and are their efficiency, effectiveness and flexibility. The specialisation of task to deliver a specific service and the massive amount of data collected from their clients allow cloud security vendors to detect new threats more accurately and react more quickly than most local installations.

Providing Security-as-a-service is a natural extension to existing security services that are already outsourced. Services like penetration testing and vulnerability assessment already used automated tools. The tools today cover an entire range of services including site monitoring, application and network vulnerability testing and reporting services. Security providers can readily maintain these tools “as-a-service”.

Finally, adoption of cloud and standardisation because of virtualised servers has created an opportunity that allows disaster recovery sites to differ in setup from production site. Key services can be provisioned and restored in a cloud service. Disaster Recovery as-a-service (DRaaS) reduces time-to-restore using ready servers and eliminates the cost to maintain a duplicate set of systems to recover to. Furthermore, recovery procedures can be tested and practised without affecting production infrastructure at the cost of provisioning the systems just for the duration of the practice.

**Enablers**

Security is the top concern of CIOs and the biggest inhibitor to cloud adoption. The broad nature of Security-as-a-service offerings means that adoption varies greatly depending on specific security applications. Cloud-based implementations like DDOS detection and prevention, secure e-mail and Web gateways, and cloud-based application security testing are examples of very established cloud-based security services.

**Inhibitors**

Concerns are elevated by lack of transparency in the cloud and fears about leakage of information via covert channels when a peer is compromised in the multi-tenant environment. One such attack was demonstrated when a researcher successfully located his chosen target compute instance in Amazon Web Service by iteratively launching compute instances until a co-located instance is found. He proceeded to take the target compute instance down[4]. Such inherent risk in multi-tenant environment prohibits adoption of Security as-a-service for sensitive data.

### 5.5.1.5 Identity/Access Management-as-a-Service

Identity and Access Management (IAM) as a Service (IAMaaS) refers to SaaS forms of IAM. IAMaaS requires minimal, and often no, enterprise on-premise presence of hardware and software.

IAMaaS is critical for the federation of identity across multiple cloud providers. Insofar as best of breed selection of cloud is concerned, an organisation may select multiple cloud providers that specialise in their specific task. This entails either maintaining different identities for individual services, or having the cloud services use a single identity. IAMaaS provides the coordination required to share identity across the services.

The adoption of IAMaaS will be driven both by adoption of a myriad of cloud services and consolidation of islands of identity as a result of mergers and acquisitions, or where identity was just not centrally coordinated.

**Enablers**

IAMaaS could become the holy grail of a global federated identity. The growing support for authentication and authorisation protocols like OpenID and OAuth presents an opportunity to realise a distributed global identity system. Typical use cases are when Web-based services re-use the authentication services of their more popular counterparts (e.g. identifying a user by their facebook account) and relieve its users from the need to remember another password.

[4] Thomas Ristenpart. Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds. [Online] Available from: http://dl.acm.org/citation.cfm?id=1653687 [Accessed 9th July 2012].

Single corporate identities will also simplify security controls especially in a global implementation where identities must be kept synchronised. These centrally managed identities allows for faster detection and management of identity compromises.

IAMaaS also provides an opportunity for implementers to institute a corporate-wide identity and access policy that can then determine the authentication and authorisation architecture.

**Inhibitors**

Existing IAM solutions may hinder implementation of cloud-based IAM. Implementation and technical differences aside, many identity solutions are local to individual departments and will require tremendous effort to sanitise for corporate wide usage. Existing applications may not be compatible with the new authentication and authorisation mechanism and may need to be updated.

Finally, governance or policies over the location of IAM data is hindering massive adoption of IAMaaS.

### 5.5.1.6 Cloud Brokers

A cloud broker is an entity that manages the use, performance and delivery of cloud services and negotiates relationships between cloud providers and cloud consumers.

As cloud computing evolves, cloud providers seek to differentiate their services by creating a myriad of offerings with varying features. For instance, IaaS evolved to cater to different workloads. The complexity of selecting the appropriate service explodes with PaaS and SaaS offerings and when different services are deployed.

Cloud brokerage services are increasing in demand as the integration of cloud services becomes too complex for cloud consumers. Cloud brokers bridge the gap between a cloud consumer needing to become an expert in all cloud services, and a cloud provider needing to provide customised services to the consumer. The goal of a cloud broker is to tailor cloud services to the specific requirements of a consumer.

A cloud broker provides services in the three forms:

- **Service Intermediation** where the cloud broker enhances the service from a cloud provider by improving specific capability and providing value-added services;
- **Service Aggregation** where the cloud broker combines and integrates services from one or more cloud providers into one or more services;
- **Service Arbitrage** where the cloud broker aggregates services from multiple cloud providers but retains the flexibility to re-evaluate and select services from a different cloud provider.

A cloud broker may enhance the original cloud offering by improving service level and security performance, and add value with centralised billing, reporting or identity management feature. Cloud brokerage may be delivered as a software, appliance, platform or even as a cloud service. Cloud brokers may consumer services from another cloud broker.

The basis of automated cloud brokers are cloud control API. These can range from APIs provided by the various cloud implementations like AWS's libraries for the Ruby, PHP or Perl, to APIs that work across different cloud implementations like Apache's DeltaCloud. Feature sets supported by each API will continue to grow affording greater monitoring and control to cloud brokers and consequently, cloud consumers.

Cloud brokers can be organised, but are not limited to, around technology verticals, geographic regions, or industry verticals.

**Enablers**

Owing their business models, cloud brokers tend to rapidly adapt to changes in both demand and supply in the cloud market.

The key value of cloud brokers will be in the federation of various cloud services and enforcement of security and service level requirements. Operating between cloud consumer and providers, cloud brokers are in the best position to orchestrate compute, network and data resources such that consumer requirements are met.

Finally, community clouds and other special purpose clouds can be implemented via a cloud broker.

**Inhibitors**

The multi-layered approach for cloud brokers abstracts both complexity and specific agreements that can be formed between a cloud provider and consumer. This abstraction leaves the cloud consumer without direct access to the cloud provider and forces the cloud consumer to rely on the cloud broker to ensure conformance of a cloud provider to fulfill any compliance requirements.

The ability of cloud brokers to layer over each other necessitates a common service level and security understanding. This is particularly important where industry-wide policies need to be enforced.

## 5.5.2 Three to five years

The following technologies are expected to be adopted within the next three to five years. They tend to be fairly established technically but are pending the final environmental or social push to amass final adoption.

### 5.5.2.1 Development-Operations (DevOp) Tools

DevOps (a portmanteau of development and operations) is a software development method that stresses communication, collaboration and integration between software developers and IT professionals. DevOps is a response to the growing awareness that there is a disconnect between development and operations activities.

Lee Thompson, former Chief Technologist of E*Trade Financial, described the relationship between developers and operations as a "Wall of Confusion". This wall is caused by a combination of conflicting motivations, process and tooling. Whereas development is focussed on introducing change, operations desire stability and tend to be change-averse.

The impact of this contention between operations and development results in release cycles delays, and longer downtime of services and other inefficiencies. Such effects are immediately felt by businesses whose business is reliant on rapid innovation and continued delivery of IT services as a competitive edge.

As a methodology, DevOps aligns the roles of development and operations under the context of shared business objectives. It introduces a set of tools and processes that reconnects both roles by enabling agile development at the organisation level. DevOps allows fast and responsive, yet stable, operations that can be kept in sync with the pace of innovation coming out of the development process.

Cloud computing provided exactly the environment for DevOps to flourish. APIs that extend cloud infrastructure operations are integrated directly into developer tools. The ability to clone environments at a low cost and to tear down such environments when no longer needed enhances the developer's ability to test their codes in exactly the same environment as operations. Quality assurance processes like unit and functional testing can be integrated into the deployment cycles.

DevOps tools are readily available today. Broadly, these tools allow the creation of platforms that improve communication and integration between development and operations.

Puppet and Chef are two tools that can be integrated into the DevOps process to allow rapid deployment in the cloud environment. These operations oriented tools allow substantial scripting capabilities in the form of "recipes" where repeatable deployment is possible. Such recipes enable developers to replicate production environments and operations to test new production environments. Sharing such recipes allow developers to better test their code, and operations to involve developers when rolling out new environments.

Automated build tools have progressed into continuous integration tools. Hudson (previously Jenkins), Cruisecontrol and Bamboo are examples of such tools that integrate source code management, build, self-test, and deployment. While the process is development focussed, operations can be involved by providing their consideration and by expanding the automated test plans with operational requirements. This integration allows development to identify integration problems early in their development process.

DevOps tools also include monitoring tools like Graphite, Kibana and Logstash. Such log analysis tools will be crucial for development to understand the performance of their codes especially in a distributed deployment environment like the cloud.

**Enablers**

DevOps has its roots in agile methodologies. For IT organisations, DevOps provides a streamlined release deployment cycle. It shortens deployment cycle and provides faster feedback from operations to development. Such agile approaches extended into operations improve overall stability and control and thus greatly reduces the risk of deployment.

Operational data from the monitoring tools, and the involvement of operations in testing and replication of production environments, contribute to development by cutting down time spent on reworking designs that are not optimised for the production environment. The monitoring tools will further inform operations of workload patterns of the new codes so they can react accordingly. Finally, data from these tools provides the basis for both operations and development to investigate any issues with respect to the deployment.

**Inhibitors**

Adoption of DevOps is still very limited in traditional enterprises. The adoption of DevOps tools is usually not a result of enterprise-wide architecture but an effort by individuals or certain groups that “straddle” operations and development.

DevOps tools are just part of the methodology. The full potential of DevOps requires strong support of company policies. Lack of cross training between development and operations is another inhibitor.

Finally, organisations that already enjoy stable operations may find that their operational processes may not have taken DevOps into consideration. Subsequently, there will be a strong reluctance to change what has always worked. Ultimately, the move to DevOps and proper adoption of DevOps tools requires the re-alignment of both development and operations toward business goals. It will be a huge undertaking to change culture, processes and tools.

### 5.5.2.2 Cloud Optimised Applications

Applications are traditionally designed on run in a single-threaded environment. Given the correct platform, such applications can be migrated into a cloud without modification. Cloud computing, however, presents many features that can be exploited to improve scalability and reliability.

A cloud-optimised application interfaces directly with the programmatic interface of the underlying cloud, be it private, public or hybrid, to take advantage of its scale and elasticity. The design goals of such applications can be to achieve reliability beyond the cloud provider, extreme and automatic scalability, or cost efficiency.

Netflix’s video delivery service is an example of a cloud-optimised application that is designed to survive disruption of the underlying cloud provider. In April 2011, the design was tested when an Amazon Web Services (AWS) data centre experienced a major service disruption. The application was able to continue serving Netflix’s customers from unaffected AWS data centres.

An application will need to be designed to scale horizontally to take advantage of cloud elasticity. This is particularly applicable today where demand can be unpredictable, as in the case of Zynga's FarmVille.

Finally, cloud optimised applications can achieve cost optimisation by monitoring utilisation and directly controlling resource allocation. This is accomplished by the rich set of Cloud APIs that provide both control and monitoring capabilities. Applications can directly request for additional resources or release unused resources according to resource allocation objectives.

**Enablers**

High profile failures and disruptions of cloud providers have created more attention that mission critical applications hosted in the cloud must be "designed for failure". As more business functions migrate into the cloud, IT will establish guidelines to define when and where the various practices of cloud optimised application design should be applied.

PaaS providers, like Google App Engine, hides the complexity of "autoscaling" to provide horizontal scaling, automatic failover and other fault tolerant features. Popular programming frameworks like Spring, Rails and Node.js are evolving to take advantage of underlying cloud architecture and simplify the development of cloud optimised applications.

**Inhibitors**

Cloud optimised applications are currently adopted by cutting-edge Web startups and a handful of organisations with sufficient business reasons (e.g. cost savings) to invest in such optimisation. These implementations, e.g. by Netflix, are tightly integrated to the underlying cloud provider and generally cannot be re-used for other cloud providers.

Horizontal scaling of applications mandates the use of techniques similar to designing distributed applications and an understanding of the interactions between the various components that may be geographically distributed. Instrumentation tools for such systems remain scarce and currently need to be built into the application.

Cloud APIs are still evolving and adoption is hindered by risk of being locked into a particular cloud provider.

### 5.5.2.3 Public Cloud Storage

Cloud storage is a storage utility that offers storage as a cloud-style service. Public cloud storage is aimed at enterprises rather than the consumer community, and excludes SaaS storage that does not integrate with other applications.

Exponential growth of corporate data, partly caused by ingestion of more external information than expected, created a gap between the need to store such data and the cost of doing so. Public cloud storage fills this gap with is its pay-per-use property where IT does not need to accurately provision for future growth.

Public cloud storage services will be integrated with enterprise applications that are not mission-critical. The services also present an alternative for archival and backup.

**Enablers**

Private cloud storage systems that deliver file services through established network file protocols like the Network File System and SMBFS already exists. These protocols are very established and well supported by most operating systems. Meanwhile, technologies that improve provider efficiency like data de-duplication, thin provisioning and on-the-fly encryption are already implemented in many storage devices.

These systems avoids the security challenge that using public cloud storage poses but introduces the users to the concepts and technologies behind public cloud storage and eases future transitions.

**Inhibitors**

Factors governing the use of public cloud storage are the main inhibitors to this technology. This includes SLAs, data location, transport costs and ingestion fees, and disaster recovery service options. Moreover, the cost of using such a service may be highly variable if the service is offered directly to business users.

### 5.5.2.4 Hybrid Clouds

NIST defined hybrid clouds as clouds that are deployed across two or more cloud deployment models (private, public, and community). Successful hybrid cloud implementation requires integration that enables data and application portability between the different cloud services. The most common hybrid clouds are composed of private and public clouds where workload is overflowed from the private cloud into the public cloud.

Most companies with existing infrastructure have undergone or at least are planning for the transition toward virtualisation and then, toward private clouds. Private clouds do not permit the same scalability that public clouds do. For specific workloads, hybrid clouds provide the possibility to augment private clouds with the scalability of public clouds, or the compliance provided by community clouds.

The boundaries between deployment models in a hybrid cloud can be drawn in different ways:

1. **Spill over loads** – This is the most common and best understood use of hybrid clouds where workloads above a certain threshold are directed into a secondary cloud. This usually occurs between the enterprise IT's private cloud and an external public cloud. The internal environment specific to the service is frequently duplicated in the external cloud so that resources can be dynamically added when demand exceeds the capacity of internal systems.

2. **Cloud service composition** – The service is organised such that part of it runs on internal systems, and other parts may be delivered from external cloud environments. There are on-going data exchanges and process coordination

between the separated parts. An example is the use of mashup capabilities provided by external vendors, e.g. the integration of rainfall data with OneMap.sg's geospatial service.

3. **Administrative control** – The service is divided along the administrative jurisdiction of respective cloud owners. This boundary frequently occurs when the final service requires coordination from the different cloud owners and is especially common in business-to-business (B2B) setups sharing a community cloud.

4. **Joint security and management** – Security and/or management processes and tools are applied to the creation and operation of internal systems and external cloud services. An example is when IT uses provisioning services from a public cloud and extends administrative control to the cloud provider.

Ultimately, the hybrid cloud is a deployment model that allows a cloud user to arrive at the optimal balance of cloud deployment strategy for any applicable service.

**Enablers**

The hybrid cloud is a preamble to the vision of a global cloud of clouds and is deployed when IT needs to evolve beyond the existing deployment model.

Implementing a private cloud is the most common strategy for most enterprises' forays into cloud computing as it re-uses existing IT infrastructure. Such enterprises can choose to expand their existing IT infrastructures, or explore a hybrid deployment where workloads can spill over to a trusted public cloud provider.

The growth of community clouds will also inspire growth in hybrid clouds as enterprises streamline their IT operations to leverage the former.

**Inhibitors**

The security concerns about the cloud have to be managed together with the participating cloud providers. Hybrid clouds today are largely silos that perform specific services and tend to be less aggressive when collaborating with partner clouds. This is particularly true since most hybrids involves integrating processes from the enterprise's own private cloud.

The tight integration and multi-partite ownership makes identifying the root cause of any failure challenging as it may be met with massive political pressure and finger pointing. The fear of such responses may delay the use of hybrid clouds for mission critical systems as companies approach to deployment model with more caution.

### 5.5.2.5 Software Defined Networking (SDN)

IDC expects there will be 50,000 10GbE ports shipped by 2015 – a six-fold increase over 2011. A brief look at 2010 recall the intrusion of non-traditional players like HP, Dell and IBM, and new entrants like Arista Networks, Plexxi and Big Switch Networks into the arena of established network providers like Cisco, Juniper and Alcatel-Lucent. Low cost networking silicon has a part to play but the centre of the play is the emerging software ecosystem that promises the possibility of a cycle of innovation in the stagnating market.

SDN allows the operator to manage the network as a unified abstraction. Conceptually, it introduces a layer that decouples network services from physical connections. This enables network entities to migrate between physical interfaces without changing identities or adapting to specific routing or topological designs. SDN decouples network control (learning and forwarding decisions) from network topology (junctions, interfaces and how they peer).

In the area of cloud computing, SDN affords greater flexibility to both cloud provider and consumer. Early adopters of cloud technologies will be aware of issues where VLANs restricted the mobility of cloud-enabled services still to persist today. SDN can free cloud services from having to be aware of specific details of network plumbing, or logical network states.

**Enablers**

SDN abstracts the intelligence of the network into a central control that may coordinate networks that may span continents. Cloud adoption supports the development and adoption of SDNbecause of their distributed nature and the need to standardise cloud deployments.

Although Openflow started as project at Stanford University, the protocol and its control concepts are widely discussed in the network industry. The interest is related to the possibility of centralising control of network deployments that may already exists in multiple sites.

Finally, the availability of merchant network chips allows new entrants into the previously lucrative network market. The value proposition of such network providers lies in their low-cost physical connectivity. The concept of SDN plays well with these devices.

**Inhibitors**

Incumbent network deployments tend to be extremely complex. Years of layering to patch problems have created networks that work miraculously. There is a certain amount of untangling before a sensible SDN policy can be applied globally.
Further network deployments tend to be distributed to the local IT department even though they might conform to some equipment standards. This means a corporate network is an aggregate of many different network policies that are optimised for each department and geographic region.

### 5.5.2.6 Cloud Security Standards

Security is a key inhibitor to cloud adoption. While many companies recognises the benefits that migrating to the cloud can have on both topline and bottomline, the general lack of transparency of cloud implementations and security processes hinders the use of cloud for critical information.

Cloud security standards will provide a common platform for understanding and evaluating the security controls. These standards improve the cloud computing landscape in two fronts:

1. Providing a common platform on which different cloud providers can be evaluated is instrumental to the further adoption of cloud computing. An organisation looking to adopt cloud computing can use the standards as a starting point to evaluate the suitability of a particular cloud provider. More advanced industries will build on the standard to mandate necessary controls that can help the potential cloud user satisfy regulatory compliance requirements for that specific industry.

2. Cloud providers will be encouraged to improve their internal security controls and processes as a result of availing direct comparison between providers. Even where providers are not able to release the full details about the controls they have instated, compliance with specific security standards, especially with the certification from independent audits, will help differentiate different cloud offerings. Over the long term, this can only have a positive impact on the level of confidence in the cloud computing.

To date, there have been several noteworthy efforts to define cloud security standards.

Locally, the IT Standards Committee has published ***TR30: Technical Reference for Virtualisation Security for Servers***, and ***TR31: Security and Service Level Guidelines for the Usage of Public Cloud Computing Services***. TR30 arms enterprise infocomm personnel, cloud service providers, cloud users and buyers with a set of guidelines and best practices to address security risks posed by virtualisation based on compute hypervisors. TR31 provides security and service level guidelines to be considered by users seeking public SaaS and IaaS cloud services.

In the USA, the Federal Risk and Authorization Management Programme (FedRAMP) was launched to ensure consistency of security practices across cloud providers providing services to the US government. Its aim is to accelerate the adoption of cloud computing by creating transparent standards and processes for security authorisation and allowing US agencies to leverage the same authorisation on a government-wide scale. US General Services Administration (GSA) that is running FedRAMP started accepting certification applications in June 2012 and aims to authorise three Cloud Providers by end of 2012.

Internationally, the Cloud Security Alliance (CSA) published its third Security Guidance for Critical Areas of Cloud Computing. It provides a practical and actionable road map to managers wanting to adopt the cloud paradigm without compromising IT security. This third version extends on the previous versions with practical recommendations and requirements that can be measured and audited. CSA has also expanded its work to address concerns surrounding cloud privacy, cloud controls, data governance, and trusted cloud.

#### 5.5.2.7 Data Tracking

Data tracking is the ability to trace data transactions as they traverse a cloud service. With the increasing adoption of cloud computing, concerns regarding security, storage and transfer of data within and out of cloud computing environments increases as well. This is worsened by the different legislation and policies of the countries the data can finally reside or pass through. For example, most will be aware of the on-going debate over the potential loss of privacy and confidentiality of data stored on US-hosted cloud service because of US Patriot Act.

Data tracking provides an audit trail on the movement and access of data within a participating cloud environment. Such a trail will provide transparency and accountability of data in cloud environments and eventually, better visibility and confidence among cloud consumers.

There are many uses for such a technology:

1. To identify and alert consumers of data leakage
2. To identify and potentially prevent movement of data across legislation boundaries
3. To provide forensic support of a compromised system
4. To function as part of a Digital Rights Management (DRM) framework

Sophisticated event triggers can be set up to extend this technology to systems that can benefit from tracking data movements.

### 5.5.2.8 SLA-based Charging

SLAs can contain numerous service performance metrics with different service level objectives (SLOs) as determined by business needs. SLO can measure the reliability or performance level of a service.

Cloud computing and its underlying principle of resource pooling makes it difficult to pin point the root causes of service disruption or performance issues because of the complex interaction between workload demands using the same resource. Today, cloud providers dictate the SLAs around reliability of the cloud service and the compensation terms when failing to meet the SLAs.

High profile outages like the Amazon Web Services outage in April 2011 and Gmail outage a year later highlighted the need to pay attention the reliability of even the largest cloud providers. Services like Cloudkick have since evolved to provide independent monitoring of various cloud providers.

The reliability and resilience of cloud providers are just one part of SLA. The other measures the performance of the cloud service and is particularly important for PaaS and SaaS. In the latter metric, the cloud consumer will have specific performance objectives that the cloud service must meet. For example, the service level objective may be measured against average time taken to perform a particular transaction. The parallel today is how IaaS providers charge a different price depending on the specifications of the virtual machine delivered.

In August 2012, AWS added a feature to its Elastic Block Storage service that provisions Input-Output per second (IOPS) to the cloud consumer. The feature is designed to deliver predictable performance for workloads with highly intensive I/O demands, such as database applications. The cloud consumer determines the size and performance, and leaves the specific implementation to AWS.

This development highlights that the cloud consumer is not just interested in paying for “time in warm irons” but is increasingly demanding a direct delivery of performance. For demanding customers, charging can be central around SLA terms and can be defined in much higher resolution. Instead of determining the specification of an IaaS instance beforehand, the consumer may be allowed to pay different prices depending on the time taken to complete specific compute transactions.

**Enablers**

The SLA management framework can be accomplished through cloud brokers even when the cloud provider does not provide a matching SLA. The cloud broker acts as a service intermediary that enhances the offerings either of a single cloud provider to translate the SLA of the provider to match the SLA requirement of the consumer, or aggregates services from multiple cloud providers to achieve the SLAs.

Cloud monitoring services are gaining traction and are evolving from information congregation channels to provide automatic callbacks for subscribed events. Such a mechanism allows cloud brokers, or advanced cloud consumers, to automatically trigger corrective actions depending on the gravity of the triggering event.

**Inhibitors**

The industry will take a while to converge around a set of SLA terms as cloud brokers continue to push new and more innovative ways match reliability terms from providers to consumers.

Performance-based objectives are very much in their infancy. IaaS providers are currently happy expanding their offerings based on the approximate specifications of the VM offered. The consumers are still left to bridge the promised specifications to their desired performance. The development of provisioned IOPS may nudge the industry, especially PaaS and SaaS providers, in the correct direction.

## 5.5.3 Five to ten years

The technology landscape changes extremely rapidly and it is indeed a challenge to imagine what the landscape may be like more than five years ahead. The technologies listed here are critical to achieve the vision of cloud of clouds.

### 5.5.3.1 Cloud Standards – Interoperability

As more cloud providers offer their interpretation of cloud computing, standards bodies like ITU-T and ISO/IEC/JTC1 SC38[5] undertake the behemoth task to standardise cloud architecture.

This collaborative effort aims to produce a cloud architecture standard that will allow competitive differences between vendors while maintaining enough functional stability for interoperability between implementations. Standardised cloud architecture lays the foundation for cloud interoperability that, in turn, enables cloud bursting.

[5] ISO/IEC/JTC1 SC38: Distributed application platforms and services (DAPS) [Online] Available from: http://www.iso.org/iso/standards_development/technical_committees/other_bodies/iso_technical_committee.htm?commid=601355 [Accessed 9th July 2012].

Further, there are also other efforts that standardise more specific aspects of cloud operations. Delta cloud is an implementation of a cloud API capable of controlling multiple cloud providers. Even VMWare's Open Virtualisation Format (OVF) has been accepted by the Distributed Management Task Force (DMTF), an industry group of 160 member companies and technology organisations across 43 countries, and lays the foundation for how VMs can be transferred across compatible clouds.

The cloud industry is moving very rapidly to meet the many new requirements of cloud consumers. As standards may take some time to be developed and adopted, we will have to take a longer term view of when the standards will be generally adopted.

### 5.5.3.2 Cloud Bursting

Cloud bursting refers to the ability of a cloud to use another cloud-based service. This ability is useful to tide over a period of peak workload or an extraordinary event that has not been catered for in the original cloud. Workload can cloud-burst from a private or public cloud to another private or public cloud.

This capability adds flexibility and agility across all cloud deployment models and are not restricted to hybrid clouds. A cloud can burst across to another cloud using similar boundaries defined for hybrid clouds. This is frequently due to spillover compute loads but a cloud can also burst to take advantage of other partner's cloud properties, e.g. proximity to the end user or availability of special resources. A key attribute about busting is it happens automatically and tends to be transparent to the cloud consumer (e.g. cloud bursting between cloud providers).

While cloud bursting is typically driven by need or spillover of compute loads, availability of other resources, e.g. data and networks must be arranged before the bursting can be effective.

Cloud bursting is mainly discussed in the context of IaaS but can be extended to PaaS and SaaS. This can be implemented either by spilling over the backend IaaS in the shorter term, or the features can be built directly into PaaS and SaaS in the future. The former is achieved by design techniques like Cloud Optimised Applications and can be achieved today. The latter will require more collaboration between cloud providers and the extension of currently available semantics but will provide a more seamless and transparent experience to the cloud consumer.

**Enablers**

Cross cloud platform controls are becoming available. At the same time, monitoring tools like Ganglia are expanding monitoring capabilities to provide cloud specific features, and can monitor performance and capacity of different participating cloud implementations.

Techniques that can migrate IaaS workloads like VMware's Vmotion are already commonly deployed.

There is also increased work in standards that can ensure sharing of data (DMTF) and compute (OVF). Finally, SDN will allow cloud providers to extend their networks into borrowed cloud services.

Techniques that can eventually support cloud bursting are already being deployed by cloud providers. Google App Engine is one example where the PaaS automatically scales the resources depending on the user's workload. This all happens within Google's cloud but the service is delivered across Google's many data centres worldwide.

**Inhibitors**

Cloud bursting can span on-premise IT private clouds and multiple public cloud providers. Provisions for security and compliance must be enforced at different layers of cloud bursting to ensure ultimate compliance of the cloud consumer. Such controls are currently not yet popular even in public clouds. Furthermore, a consistent semantics must be achieved before cloud bursting on a large scale can occur.

While a lot of work on standards is underway, the question that remains as to where the interests of the established cloud providers will prevail and subsequently whether these standards will be widely adopted. With the cloud industry moving at such a rapid pace and with the cloud providers innovating within their respective offerings, the industry might take a while to arrive at a common ground.

Finally, there are still some technical difficulties around cloud bursting. High-speed data transfer or the means to deliver specific data sets to the partner clouds, and synchronisation of data especially across longer distances are technical inhibitors that remain unsolved.

## 5.6 Market Applications/Opportunities

Cloud computing brings aboard many challenges to implementation but also presents much larger opportunities. The key principles of massive scalability, cost efficiency through scale and elasticity, common platform, and standardised practices provides many opportunities across several industry sectors.

Companies are already employing cloud services as a means to control rising cost and to stay focus on their core business activities. Stories of businesses moving to SaaS like Salesforce, Google Apps and Office 365 are based around the concerns of the bottom-line. The next wave of adoption will be driven by businesses exploring to expand their top-line and to improve collaboration with their partners.

### 5.6.1 Cloud Business Drivers

#### 5.6.1.1 Healthcare

Cloud computing provides a way for health organisations to reduce costs, simplify management and improve services in a safe and secure manner. Cloud computing in healthcare extends beyond the traditional services like email, communications, meeting and collaboration into more direct applications of ICT such as online health and wellness, management and sharing of patient records, and tracking of disease patterns.

According to research by MarketsandMarkets, the cloud computing market in health care will grow to US$5.4billion by 2017 at CAGR of 20.5% from 2012 to 2017[6].
An example of cloud computing in the healthcare industry is IHiS healthcare cloud.[7] The cloud service promotes better collaboration within the health care industry and can extend into personal health management. The cloud implementation also addresses issues surrounding the solution's ability to expand and scale its services as demands picks up.

IHiS's approach leverages the best of cloud's offerings, i.e. its cost efficiency in providing scale and elasticity, to align their investments to the demands of the service. The novelty though is IHiS's consideration of partner establishments in its implementation. This potentially allows its partners to use the healthcare cloud to promote long term care and lifestyle changes towards developing a healthier nation.

#### 5.6.1.2 Public Land Transportation

The public land transportation is a complex system of many modes of transports, e.g. buses, trains and taxis, and includes the participation of several transport operators. The system attempts to meet commuter demands that may be routine, (e.g. getting to work), seasonal, or ad-hoc in nature. On top of usage patterns, the system needs to contend with the additional complexity of road blockages, traffic flow, special events that may re-route traffic, and many other operational issues.

With the advent of GPS tracking and electronic payment systems, a wealth of data is waiting to be processed, and that presents a chance to streamline the operations of a single operator. A cloud implementation can help to optimise the operations of public transportation as a whole by incorporating data from multiple transport providers.

Online information can also be made available to commuters. Such information can be combined with personal information, e.g. meeting time and venue, of the commuter to inform him of the best transportation route to his destination.

In the long term, such location and destination information may also integrate into the future public transport system where the modes of transports are optimised on the fly to adjust frequency, or even routes according to the demand.

[6] MarketsandMarkets. Healthcare Cloud Computing (Clinical, EMR, SaaS, Private, Public, Hybrid) Market – Global Trends, Challenges, Opportunities & Forecasts (2012 – 2017). [Online] Available from: http://www.marketsandmarkets.com/Market-Reports/cloud-computing-healthcare-market-347.html [Accessed 9th July 2012].

[7] Integrated Health Information Systems. Transforming Healthcare with IT Innovation. [Online] Available from: http://www.ihis.com.sg/LinkClick.aspx?fileticket=B5-1rkCA8fo%3D&tabid=58 [Accessed 9th July 2012].