# The Structure of the New IT Frontier:  Cloud Computing – Part I

Dexter Duncan[1], Xingchen Chu[2], Christian Vecchiola[2], and Rajkumar Buyya[1, 2]

[1] Manjrasoft Pty Ltd, Melbourne, Australia
Email: {dexter, raj}@manjrasoft.com

[2] **Clou**d Computing and **D**istributed **S**ystems (CLOUDS) Laboratory
Department of Computer Science and Software Engineering
The University of Melbourne, Australia
Email: {xchu, csve, raj }@csse.unimelb.edu.au

The latest craze in the IT (Information Technology) sector is Cloud Computing.  Some call it the new frontier.  The Wild West is here, bringing with it opportunities for new companies.  Conferences on this topic are already in their second or third year. Companies in this space are attracting venture capital funds even during the Global Financial Crisis.  IDC estimates the market size for Cloud Computing to be at USD $42 Billion by 2012, with an annual growth rate of 27%. Extending this to 2020, the market size for Cloud Computing stands at an astronomical $284 Billion. With the growing shift in application software development from PC to Data Centers and Clouds, we expect that this opportunity will increase even higher.  With these projections, so many vendors and service providers have got into the game and so much press activity has taken place. Gartner's 2009 IT Hype Cycle for Emerging Technologies, released this July, places "Cloud Computing" at the peak of the "hype cycle". The lead adopter phase is over and a period of uncertainty, innovation and consolidation has begun as companies find practical uses for Cloud Computing.



Cloud computing is often called the 5th Utility [2].   Utilities such as water, gas, and electricity are fundamental in our daily life and are exploited on a pay per use basis. The existing infrastructures deliver these services almost anywhere and anytime.  The usage of these utilities is charged, according to different policies, to the end user.  The same idea of utility is applied to computing and a consistent shift towards this approach is seen with the proliferation of Cloud Computing.

Cloud Computing promises to deliver on-demand IT resources on a pay-per-use basis by dynamically growing or shrinking the virtualized resources and providing them as a service to the users over the internet. This is aimed at a global market with massive demands in real time, ranging from the end users that host their personal documents on the Internet, to enterprises outsourcing their entire IT infrastructure to external data centers. The approach of Cloud Computing makes the notion of IT as a fifth utility a reality: not only computing and storage resources are delivered on demand but the entire stack of computing is offloaded to the Cloud.

CIOs should ask: how does Cloud Computing help my organization?  Can this new commoditized infrastructure give me a competitive advantage in delivering services or does it provide new flexibility? Many of the advantages of Cloud Computing are a culmination of advantages derived from the multiple

technologies that drive it – such as Grids, Clusters, Virtualization, Utility (e.g. pay-per-use), Service Oriented Architecture and Quality of Service needs.

Some of the most famous names in Cloud Computing are giants such as IBM, Microsoft, Google and Sun Microsystems (now part of Oracle). In addition, a number of IT vendors, including HP, Dell, EMC (VMWare), Cisco, Citrix, RedHat, AT&T, Unisys, Oracle, and small to mid-size start-ups have made product and service announcements addressing the Cloud Computing space. Even Amazon, the bookseller, opened up its datacenters for business by selling compute and storage "on-demand" and includes this new service as one of its best revenue growth areas.
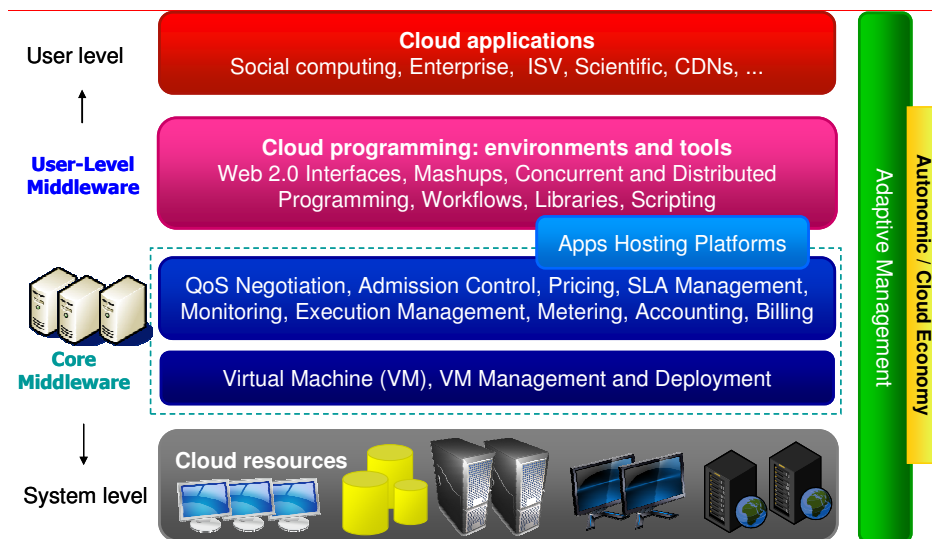
*What is Cloud Computing and who are the teams playing in this space?*

The CEO of Microsoft, Steve Balmer, reportedly told a reporter in New York in March 2009, "Anything that has been a server needs to be a service." His statement is a good summary of Microsoft's Azure strategy, but only partly covers the definition of Cloud Computing. A new framework race is on that allows enterprises to manage based on the demand based commodity infrastructure. The framework needs to make it easy to manage, provision, monitor, secure and change resources when needed or based on pre-determined service levels that match the applications.

## Cloud Computing Reference Model

*Cloud Computing is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on a service-level agreement [1].* According to this definition, Cloud Computing refers to the practice of delivering software and infrastructure as a service on a "pay as you go" basis.

Figure 1 illustrates the layered architecture of the Cloud Computing. Generally, the Cloud architecture defines 4 distinct layers from physical hardware to end user applications. Physical resources including spare desktop machines, clusters and datacenters form the lowest level of the stack, on top of which the virtual infrastructure is deployed. Infrastructure supporting large scale Cloud deployments are more likely datacenters hosting hundreds or thousands of machines, while small scale, in house Clouds provide a more heterogeneous scenario where the idle CPU cycles of spare desktop machines are used to leverage the compute workload. This level provides the "horse power" of the Cloud.



**Figure 1:** Cloud Computing layered architecture.

The physical infrastructure is managed by the core middleware layer whose objective is to provide an appropriate run time environment for applications and to maximize the use of the physical resources. In order to provide advanced services, such as application isolation, quality of service, and sandboxing, the core middleware rely on virtualization technologies. Among the different solutions for virtualization, hardware level virtualization and programming language level virtualization are the most popular. Hardware level virtualization guarantees complete isolation of applications and a fine partitioning of the physical resources, such as memory and CPU, by means of virtual machines. Programming level virtualization provides sandboxing and managed execution for applications developed with a specific technology or programming language. On top of this, the core middleware provides a wide set of services that assist service providers in delivering a professional and commercial service to end users. These services include: negotiation of the quality of service, admission control, execution management, monitoring, accounting, and billing.

Together with the physical infrastructure the core middleware represents the platform on top of which the applications are deployed in the Cloud. It is very rare to have direct access to this layer. More commonly, the services delivered by the core middleware are accessed through a user level middleware. This provides environments and tools simplifying the development and the deployment of applications in the Cloud: web 2.0 interfaces, command line tools, libraries, and programming languages. The user level middleware constitutes the access point of applications to the Cloud.

The Cloud Computing model introduces several benefits for applications and enterprises. The adaptive management of the Cloud allows applications to scale on demand according to their needs: applications dynamically acquire more resource to host their services to handle peak workloads and release when the load decreases. Enterprises do not have to build for the peak capacity anymore, but they instead provision as many resources as they need, for the time they need, and when they need. Moreover, by moving their IT infrastructure into the Cloud, enterprises reduce their administration and maintenance costs. This opportunity becomes even more appealing for startups, who have limited cash, but scale their IT infrastructure as their business grows. This model is also convenient for service providers that aim to maximize the revenue from their physical infrastructure. Besides the most common "pay as you go" strategy more effective pricing policies are devised according to the specific services delivered to the end user. The use of virtualization technologies allows a fine control over the resources and the services that are made available at runtime for applications. This introduces the opportunity of adopting various pricing models that benefit both the customers and the vendors.

The model endorsed by Cloud Computing provides the capability of moving the execution of applications on a distributed infrastructure that, in case of public clouds, belongs to third parties. While this model is certainly convenient, it also brings additional issues from a legal and a security point of view. For example, the infrastructure constituting the Computing Cloud is made of datacenters and clusters located in different countries where different laws for digital content apply. The same application is considered legal or illegal according to the where is hosted. In addition, privacy and confidentiality of data depends on the location of its storage. For example, confidentiality of accounts in a bank located in Switzerland may not be guaranteed by the use of data center located in United States. In order to address this issue some Cloud Computing vendors include the geographic location of the hosting as a parameter of the service level agreement made with the customer. For example, Amazon EC2 provides the concept of availability zones that identify the location of the datacenters where applications are hosted. Users have access to different availability zones and decide where to host their applications. Security issues are also being addressed, for example with Amazon's new VPN (Virtual Private Network) service, the huge issue of security suddenly looks like a small obstacle for enterprises to join the cloud. Since Cloud Computing is still in its infancy the solutions devised to address these issues are fundamental for wide adoption to take place.

**Structure of Cloud Computing**

Cloud computing, as depicted in Figure 2, encompasses so many aspects of computing (from hardware to software) that a single solution is not able to provide it all. More likely, specific solutions address the user needs and are successful in delivering IT resources as a real utility. By leveraging different types of services provided by Cloud Computing, it is useful to satisfy the needs of everyone. Generally, it incorporates the combination of the following:

- *Infrastructure as a Service (IaaS):* delivers the computer infrastructure, typically a virtualized computer as a service. The end user has full controls over the virtualized computer instance, and can customize the instance accordingly. The virtualization technology is used to provide multi-tenancy and isolation to the users as different virtual instances may be allocated to a single physical machine. Unlike purchasing the physical servers, *IaaS* is charged on a utility basis depending on the consumption of the resources.
- *Platform as a Service (PaaS):* delivers a computing platform and solution stack as a service. It hides all the complexity of managing the underlying hardware, provides all the facilities required to support the complete lifecycle of building and deploying web applications and services entirely from the internet.
- *Software as a Service (SaaS):* is a model of software deployment where in a provider delivers its software as a service to be used by customers on demand.
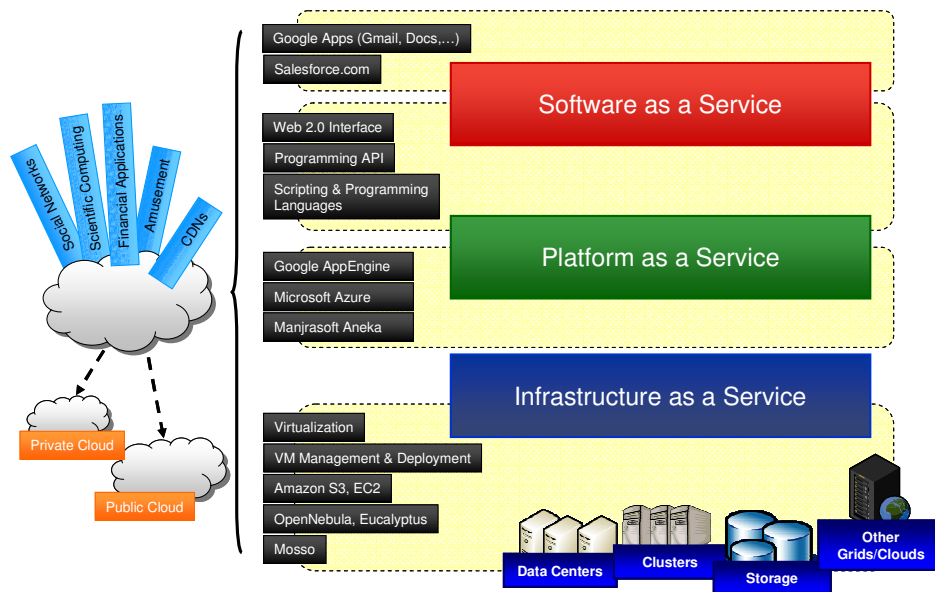


**Figure 2.** Cloud Computing architecture.

**Samples of Infrastructure as a Service (IaaS)**

The biggest name in the IaaS space is also the most unlikely name and a good example of how business is turned on its head by leveraging investment in technology.  Amazon.com launched Elastic Compute Cloud (EC2) three years ago and is now the king of offering a pay-as-you-go elastic hosting service.  This type of service is foundation for turning capital expense (CapEx) into an Operating Expense (OpEx) on corporate balance sheets.

IaaS solutions provide users with physical or virtual resources that satisfy the requirements of the user applications in terms of CPU, memory, operating system and storage. These and other QoS parameters are established through a Service Level Agreement (SLA) between the customer and the service provider. Examples of this approach are Amazon EC2[1] and S3[2], Mosso[3], Nirvanix, Go Grid and Engine Yard. Nirvanix has a subscription based service, GoGrid provides infrastructure as a Utility but similar to the enterprise datacenter, and Engine Yard is a Ruby on Rails (Java) hosting company.

On his blog[4], Werner Vogels, Amazon VP & CTO, writes about what CIOs around the world are telling him:   "They [CIOs] have bought into the cloud as a target for a significant portion of their services,

---

[1] http://aws.amazon.com/ec2/

[2] http://aws.amazon.com/s3/

[3] http://www.mosso.com/

[4] http://www.allthingsdistributed.com/2009/08/amazon_virtual_private_cloud.html

as the benefits are too obvious to ignore, and most expect that their transition will be a continuous process. They would accelerate the adoption of cloud services if they could access a form of cloud that would give them the best of both worlds: the flexibility and cost-effectiveness of accessing a virtually infinite pool of resources without owning it, while being able to integrate those resources into their existing datacenter environments such that they could continue to leverage existing investments in their management and control infrastructure."

Virtualization is one of the key drivers that make IaaS elastic and most CIOs are already aware of the big names involved: EMC (VMWare) and Citrix Systems (XenServer). The tsunami of this commoditization of infrastructure has temporarily been a focus of many, with new opportunities rapidly becoming apparent as the wave of change recedes back into the background. This commoditization of infrastructure (both physical and virtual machines) and the on-demand nature of the service now present new opportunities to improve the performance of services and solutions that either did not exist or were not apparent in previous IT platform iterations. Configuring servers and provisioning VMs is now the domain of giants like Amazon, GoGrid, Mosso, Microsoft and others that have perfected the efficiencies and art of offering infrastructure as a service.

Even though many enterprises are still developing their data centers around virtualization, the window of innovation over the next few years is now clearly focused above this commodity layer. This is evident from many of the applications and automation solutions that are springing up(built on-top of the physical and VM commodity layer). You are not going to make your enterprise hardware cheaper than the big boys operating in the IaaS space. As Google project manager, Rajen Sheth, rightly writes in a Google blog-site, "[…Where] can I find the same economics, ease of maintenance, and pace of innovation that is inherent in the cloud?" This means the real place of innovation and growth is in Platform as a Service and Software as a Service. Expect to see traditional Infrastructure and virtualization companies listed above playing aggressively in the Platform area through in-house development or acquisitions. The recent VMWare purchase of Springsource and Cloud Foundry is an early example. Even Cisco, which sells infrastructure to enterprises and hosting companies, has recently invested in Majitek, a company that "(enables) traditional business capabilities to be delivered as a service, on demand, anywhere, anytime."


## Exploring Platform as a Service (PaaS)

Platform as a Service (PaaS) market size is growing at 160% annual growth rate (according to IDC, Gartner, AMR and William Blair & Co) and expected to reach to just around USD $9 to 10 Billion in 2012. This is by far the highest growth area in the cloud.

Cloud Computing platforms possess characteristics of both Clusters and Grids, with their own special attributes and capabilities such as strong support for virtualization, dynamic provisioning of Web Services, and strong support for creating third party, value added services by building on Cloud compute, storage, and application services. Thus, Clouds provide services to users without reference to the infrastructure on which they are hosted. It is important for the platform to be flexible, easily customizable, and extensible while being service oriented to provide unique SLA/QoS & Web Services management, security and dynamic resourcing environments. This includes the ability to remotely control, monitor and dynamically change single and groups of nodes, which is not only important to simplify the management, but also to identify and remove bottlenecks.

Many of the existing middleware platforms are proprietary and tied to their infrastructure or applications' service such as Google's App Engine[5] and SalesForce's Force.com. Most of the non-proprietary ones are Java/Linux based, with a few that sit on .NET framework. PaaS solutions, such as Google AppEngine and Microsoft Azure[6] provide users with a development platform for creating distributed applications that can automatically scale on demand.

Outside those mentioned above, other companies providing Platform as a Service are: 3Tera, Apprenda, DataSynapse, RightScale, Manjrasoft, Univa UD, Elastra and Enomaly. RightScale recently launched a service that works across multiple public clouds at the same time, abstracting the providers and giving the end-user a single super-management console to configure and provision new virtual machines.

---

[5] http://code.google.com/appengine/docs/whatisgoogleappengine.html
[6] http://www.microsoft.com/azure/

Univa UD offers HPC as a service.  Apprenda's SaaSGrid offers a platform that enables Software as a Service for your .NET enterprise applications.  Quoting from their website, "SaaSGrid allows you to manage both utility and dedicated compute resources as a single compute mesh, giving you the most powerful and flexible solution for delivering your applications as true Software as a Service." Similarly, Manjrasoft's Aneka virtualizes and accelerates your applications and dynamically provisions resources across private, public or hybrid networks for .NET environments.

## Software as a Service (SaaS)

Almost half of the projected $42 billion opportunity is in software applications.  Most CIOs enter Cloud Computing through Software as a Service project, because of the scalability, flexibility and cost savings involved.  Whether it is finding the new gold nugget of information with data mining, analytics with databases, saving time on rendering large design or image files, or more complex financial analysis and life science applications, the opportunity to build and accelerate your applications based on on-demand resources has never been greater. Other applications include CRM, ERP, accounting, scheduling, automated billing, content management, HR management and so on; IT organizations are often seen as implementing critical areas of business strategy, critical in wringing out more profitability or creating a competitive advantage with solutions mirroring business needs.

Examples of Software as a Service are Salesforce.com[7] and Clarizen.com[8], which respectively provide on line CRM and project management services.  Google and Zoho provide desktop applications which are hosted in the cloud and are replacing the traditional desktop based Microsoft Office software.

## Private vs. Public Clouds

Some applications are so strategic and central to the business, they will never move to the public cloud. However, even those may want to have a disaster recovery plan or virtual private network that leverages resources in the public cloud.   Regarding private clouds, Amazon CTO says in his Blog: "Virtualization and increased automation may give them some improvements in utilization, but they would still be holding the capital, and the operational cost would still be significantly higher."  He points out the main advantage of public clouds is managing resources as an operational expense rather a capital one.  For some CIOs, an operational expense option is not a good plan, but they are building private clouds.

Most CIOs need to ensure their private data center network evolves to interwork with the new "commoditized" public cloud offerings. This allows them to maximize their existing investment and opens up opportunity for a future "unified design" where you easily move your data and share compute resources in a federated model either for purposes of managing peak load (cloud bursting), minimizing time and money (brokering) or as a disaster recovery solution.   This means adopting many of the characteristics of the Cloud such as Virtualization, Web Services, Failure Management, Provisioning on Demand, Discovery Services, SLA based, etc. which are a combination of the typical Cluster and Grid characteristics.   This gives the dynamic elastic infrastructure an on-demand, self-serve, self-managed, internet enabled, and consumption based structure needed to take advantage of hybrid or inter-cloud offerings.  Many enterprises are focused on building "private clouds" until some of the security, control and terms of service of public cloud computing are clear. IBM is a good example of this.

In a future article we will give more detail on offerings from some major players and discuss the trend toward Market Oriented Clouds.

---

[7] http://www.salesforce.com
[8] http://www.clarenz.com

**References and Useful Pointers**

1. R. Buyya, C. S. Yeo, and S. Venugopal, Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities, Keynote Paper, Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications (HPCC 2008, IEEE CS Press, Los Alamitos, CA, USA), Sept. 25-27, 2008, Dalian, China.

2. R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility, Future Generation Computer Systems, Volume 25, Number 6, Pages: 599-616, ISSN: 0167-739X, Elsevier Science, Amsterdam, The Netherlands, June 2009.

3. For a list of the "Top 150 companies in Cloud Computing", Check out this list from *Virtualization Journal:* http://virtualization.sys-con.com/node/770174

4. Another list in *Datamation* covers "85 companies shaping the emerging cloud" : http://itmanagement.earthweb.com/features/article.php/12297_3835941_1/85-Cloud-Computing-Vendors-Shaping-the-Emerging-Cloud.htm