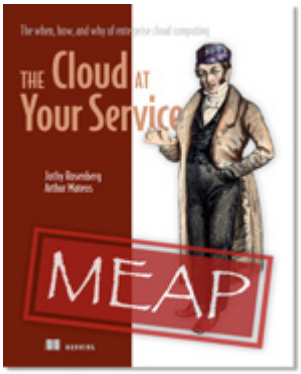
MANNING PUBLICATIONS

What is Cloud Computing?

Green Paper from



The Cloud at Your Service

EARLY ACCESS EDITION

The when, how, and why of enterprise cloud computing

Jothy Rosenberg and Arthur Mateos
MEAP Release: February 2010
Softbound print: Summer 2010 | 200 pages
ISBN: 9781935182528

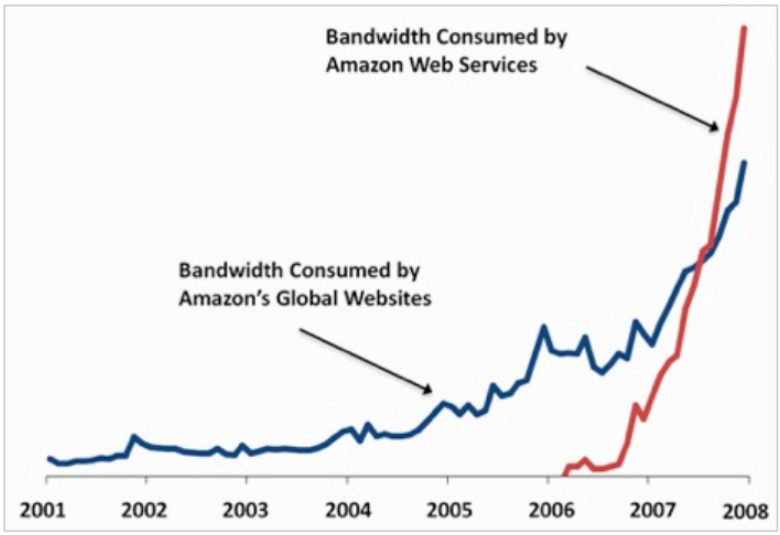
This green paper is taken from the book The Cloud at Your Service from Manning Publications. The authors discuss Cloud technology as a power behind the "flattening of the world." They demystify the concept through current examples and lay out a categorization and taxonomy of associated terms. For the table of contents, the author forum, and other resources, go to http://manning.com/rosenberg/.

Have you asked or been asked what Cloud Computing is? Before we go any further, let's answer that question. We define Cloud Computing as computing services that are offered by a third party, are available for use when needed, and can be scaled dynamically in response to a changing need. Cloud Computing represents a departure in the way IT systems are developed, operated, and managed. On the economic front there are potentially tremendous cost savings and much more potential flexibility than was ever before possible. Adoption of Cloud Computing has the potential of providing enormous economic benefit as well as greater flexibility and agility.

Cloud Computing is being written and spoken about not just in IT journals and at IT conferences but in mainstream business magazines and in the mass media. It may win the prize for the most over-hyped concept IT has ever had. Prior terms in this over-hyped category include Service-Oriented Architectures, Application Service Providers, and Artificial Intelligence, to name just a few. We hope to cut through the hype and not add to it. To accomplish that, we will not just repeat what you have been hearing but give you a framework to really understand the concept and its importance.

So what is driving all of this hype? It might be easy to simply chalk it up to analysts and other prognosticators trying to promote their services, to vendors trying to play up their capabilities to demonstrate their thought leadership in the market, or to authors trying to sell new books. However, a good deal of what is fueling the Cloud mania and all of the great expectations are the facts on the ground. Software developers around the world are beginning to use these services. Within the first 18 months, the first public cloud offering from Amazon attracted over 500,000 customers. As Figure 1 from Amazon's web site shows, the bandwidth consumed by their Cloud

quickly eclipsed that used by their online store. As the old adage goes, "Where there's smoke, there's fire," so clearly something is driving the rapid uptake in usage from a cold start in mid-2006.



Source: Amazon

Figure 1 Amazon originally deployed a large IT infrastructure to support its global e-commerce platform. In less than eighteen months after making the platform available as a cloud service to external users, its usage as measured by amount of bandwidth consumed outstripped bandwidth used internally.

Like the previous technology shifts, such as the move from Mainframes to Client-Server to the Internet, Cloud Computing will have major implications on the IT business. We'll provide you with the background and perspective to understand how it can be effectively used as a component of your overall IT portfolio.

Before we dive into that, one question lots of people ask but few answer is where the term *Cloud* came from in the first place. The answer is that, for over a decade, when people drew pictures of their applications or infrastructure and whenever they got to the point where they included the Internet, everyone just drew a picture of a cloud. Figure 2 shows a classic example.

Figure 2 A picture of a cloud is a ubiquitous representation of the Internet and is used almost universally in discussions or drawings of computer architecture.

All that meant was that there were anonymous people sitting at browsers accessing the Internet, and somehow their browser visited your site and began to access your infrastructure and applications. So from "somewhere out there" you got visitors who could become users who might buy products or services from you. Unlike internal customers to whom you might have normally been providing IT applications and services, this constituency existed "somewhere else," outside of your firewall, and hence outside of your domain of control. The image of a cloud was simply a very good way to represent this vast potential base of anonymous users coming from the Internet. Those users had to log in somewhere to get access to the Internet, from some PC, and technically each one had to have an Internet Service Provider (ISP) who might be a telecom company, their employer, or a dedicated Internet access company (e.g., AOL). Each ISP had to have a bank of machines which people could access and which, in turn, had access to the Internet. Simply put, the earliest concept of *The Cloud* was large aggregations of computers with access to the Internet accessed by people through their browser. The concept has remained surprisingly true to that early concept but has evolved and matured in very important way.

Cloud Computing as a term—and not just as a diagram artifact—came into common use within the IT community over the last couple of years. Initially, Cloud Computing strictly encompassed the use of third-party hardware resources with a low level of interaction. In this sense, Cloud Computing was really just an extension of hosting where computers used by your organization physically resided in a third party company's facilities alongside the computers other organizations paid to use. Increasingly, the definition of Cloud Computing is growing to encompass other related concepts.

There are two ways that the basic definition of Cloud Computing has expanded and evolved. The first relates to the layer in the OSI[1] stack with which one interacts; third-party providers are now offering richer sets of functionality that go well beyond the raw hardware and disk resources originally being offered. The development of Cloud Computing is also increasingly being carried out within an organization's firewall, with large enterprises creating Private Clouds and making these resources available to their users. Since you are a decision-maker or at least an influencer, we need to address why you should care about the Cloud now and in the future. And, since the current world of Cloud Computing is very diverse, we believe it is important that we lay out a categorization and taxonomy of terms.

Defining Cloud Computing

We can summarize the key concepts of Cloud Computing as:

- A pool of computing resources available for use when needed
- Resources that can be used as necessary and only when necessary
- Dynamically scalable according to need

These concepts, of course, are not entirely new concepts in IT, and, as it turns out, Cloud Computing did not emerge fully formed in mid-2005. Instead, it can be seen as the result of a natural evolution of hardware, network, and software trends that took place over several decades. As the simple example above showed, the economics behind the development and operation of a new application is drastically changed through Cloud Computing.

The five main principles that define Cloud Computing, and which are all required for us to call it Cloud Computing are summarized in Table 1.

[1] OSI: Open Systems Interconnection model divides network architecture into seven layers which, from top to bottom, are the Application, Presentation, Session, Transport, Network, Data-Link, and Physical Layers

Main principles	
Off premises	Someone else owns the hardware assets
Virtualization	High utilization of assets
Elasticity	Dynamic scale without capex
Automation	Build, deploy, configure, provision, move all without manual intervention
Metered billing	Per usage business model; pay for what you use

Table 1 The five main principles of Cloud Computing represent the key elements necessary for a service to be classified as a Cloud Computing service.

We will now lay out and discuss these principles in concrete terms.

Off premises hardware assets

The first characteristic of cloud computing is that it utilizes the computing assets of an external resource. If we consider our original corporate website example, there are three basic operational deployment options. In our previous example the website was migrated from one of those options (managed hosting) to another option (the cloud). The third option is the self-hosting option, in where the organization either owns or leases data center facilities and runs the application on hardware that it owns. This was actually the situation for the example website two years before the switch to the cloud. As is the case with many companies that have a reasonably sized investment in data center assets, this same corporate website was hosted by the company within its own facilities.

The shift from self-hosted IT to outsourced IT resources, which Cloud Computing (and also managed hosting) embodies, have important economic implications. The two primary implications are a shift of capital expenses (CAPEX) to operational expenses (OPEX), and the potential reduction in OPEX associated with operating the infrastructure. The shift from CAPEX to OPEX means a lowering of the financial barrier for the initiation of a new project. In the self-hosted model, budget must be allocated and then spent for the purchase of hardware and software licenses. This fixed cost must be committed and paid for through depreciation of the assets over several years whether or not the project is successful. In an outsourced model, the barrier for starting a new project is much smaller. In the case of managed hosting, the startup fees are typically equivalent to one month's operational cost, and one must typically commit to only one year of costs upfront. Typically, the one-year cost is roughly the same or slightly lower than the CAPEX cost for an equivalent project; however, this is offset by the reduced OPEX required to operate the infrastructure. In a cloud model, there are typically no initial startup fees. In fact, one can sign up, enter a credit card number, and start using Cloud services literally in less time than it would take to read this green paper.

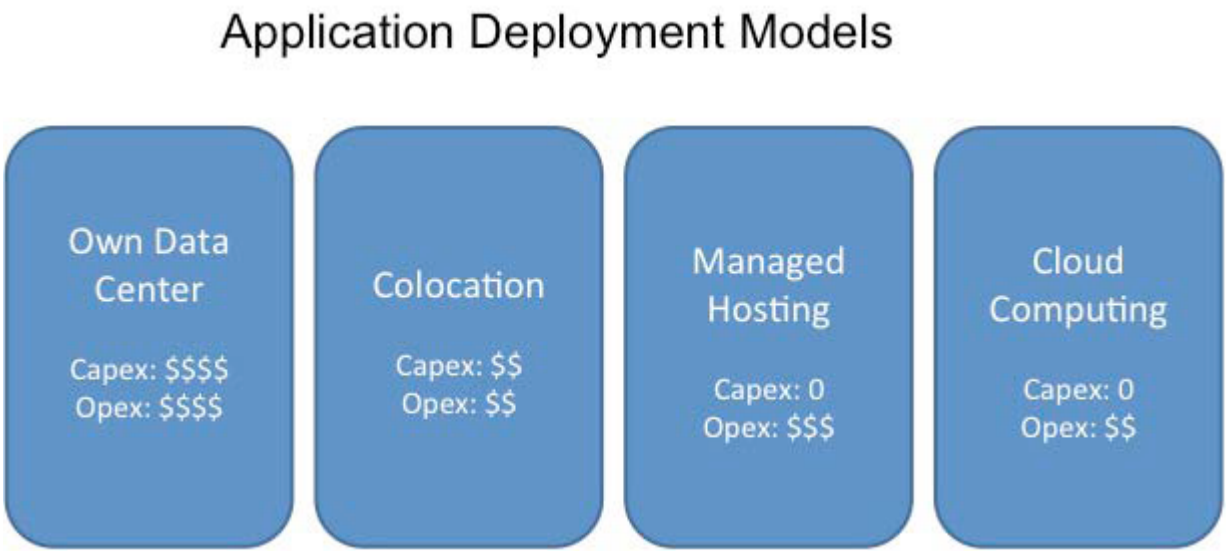


Figure 3 IT organizations have several alternatives for hosting applications. The choice of deployment model has different implications for the amount of CAPEX (upfront capital expenditure) and to OPEX (ongoing operational costs). This number of $ signs represents the relative level of CAPEX and OPEX involved with the choice of deployment model.

The drastic difference in economics that we saw in our example is due to the fact that the cost structures for Cloud infrastructures are vastly better than those one can obtain in other models (see Figure 3 for comparison). There are several reasons for the economies of scale, but the primary drivers are related to the simple economics of volume. Walmart and Costco can buy consumer goods at a price point much lower than you or I could because of the quantities of goods that they can buy in bulk. In the world of computing, these "goods" are computing, storage, power, and network capacity.

Virtualization of compute resources

The scale of Cloud infrastructures can be enormous, based on thousands of servers. Each server takes up physical space and uses significant power and cooling so getting high utilization out of each and every server is vital to keep costs low. The recent technological breakthrough that enabled high utilization on commodity hardware—and which is the single biggest factor behind Cloud being a recent IT phenomenon—is virtualization. Each server is partitioned into many virtual servers each one of which itself acts like a server that can run an operating system and a full complement of applications. Virtualized servers are the primary units that can be consumed as needed. These virtualized servers constitute a large pool of resources available to be used when required.

Elasticity as resource demands grow and shrink

The fact that this large pool of resources exists enables a concept known as "elasticity." Elasticity refers to the ability to dynamically change how much resource is consumed in response to how much is needed. Typical applications require a base level of resources under normal, steady-state conditions, but under peak load conditions need more resource. In a non-cloud world, one would have to build sufficient capacity to not only perform adequately under baseline load conditions but also to handle peak load scenarios with sufficiently good performance. In the case of a self-hosted model, this means over-provisioning the amount of hardware for a given allocation. In the case of a managed hosting deployment, one can start with a small set of resources and grow as the requirements of the application grow, but the typical time to provision a new set of dedicated hardware resources takes weeks.

Automation of new resource deployment

For a cloud deployed application, new instances can be provisioned on an as needed basis and these resources brought online in a matter of minutes. Once the peak demand has ebbed, and the additional resources are no

longer needed, these virtual instances can be brought offline. The only additional cost is for the hours that those instances were in use and active.

Metered billing for pay-as-you-go

The fifth distinguishing characteristic of Cloud computing is a metered billing model. In the case of managed hosting as we just mentioned, there typically is an initial startup fee and the requirement to enter into a commitment to purchase services for an entire year. The cloud model breaks that economic barrier because it is a pay-as-you-go model. There is no required commitment to an annual contract or to a specific level of consumption. Typically, one can allocate resources as needed and simply pay for them on an hourly basis. The removal of the economic barrier extends not only to projects being run by an IT organization but is being used by innumerable entrepreneurs today starting new businesses. Instead of needing to raise capital as they might have in the past, these brave souls have available to them the ability to utilize vast quantities of compute resource, for pennies per hour. This has drastically changed the playing field and allowed the little guy to be on equal footing with the largest corporations.

A Cloud Taxonomy

In the earliest days of commercially practicable computing, computer resources were scarce, and the primary model for their use was much like a utility. The sense however was different from the sense of utility that Cloud Computing offers today. It was more akin to the community well during a drought. Members of the community had access and were allocated a fixed amount of water. In the case of Cloud Computing today we have returned to the notion of computing being available as a utility, but one where there is no longer scarcity.

The Cloud Movement was actually presaged by the shift in business model that has taken over in the software industry that commenced at the turn of the century toward Software-as-a-Service (SaaS). In the SaaS model, the traditional enterprise license model was turned on its head and instead of having a large upfront capital investment, with SaaS, software could be purchased in a pay-as-you-go manner, with costs scaling with usage. There is no need to provision hardware and software; instead, the services were turned on when needed. This evolved into a next kind of offering that one could call hardware-as-a-service. With this capability one can build one's one applications but gain the benefit of SaaS payment models.

In the Hardware-as-a-Service model, instead of paying for use of the software, the charge is specifically for the use of the datacenter hardware and software running virtualization on its servers and providing virtual servers for rent to the public who pay as they go just for what they use.

What exactly is the big deal about all this? It fundamentally boils down to a different economic landscape. It can be thought of as a "flattening of the world," or the democratization of innovation. In former times, one of the barriers to starting a software-oriented business was the capital needed to purchase hardware to make a go at it. Back in the early days, one typically needed to sell a reasonable chunk of a new venture to financial backers such as angel investors or venture capitalists. Now, thousands of software businesses are being created using no more than ingenuity and a few hundred dollars in Cloud services charges.

Within enterprises, the capital hurdle for new projects and initiatives has also been significantly lowered. Software can be designed, written, and tested without the purchase of a single computer. Once the application is ready for production, users can simply be pointed to an instance of the application that was developed in the Cloud that is now running in the Cloud. A tricky part of operationalizing a new application was having a good handle on demand and the required capacity to service that demand in a performant manner. Buying too much hardware meant a waste of capital resources and servers doing nothing more than consuming power. Buying too little hardware risked underpowering the system and ending up with unhappy users and potentially lost sales and customers due to slow, unreliable performance.

The cloud offers the illusion of infinite resources, available on demand. The guessing game on how many users need to be supported and how scalable the application is no longer needs to be played. If only one server is needed during non-peak utilization times, but one hundred are needed during peak times, this is not a problem. In the world of the cloud you end up paying for only the resources that you use, when you use them. This is the revolutionary change: the ability to handle scale without paying a premium. This is the realm of true utility computing where resource utilization mirrors the way we consume electricity or water.

There are several ways to classify Cloud Computing, in this book we present a taxonomy where Cloud Services are described generically as X-as-a-Service, where X can take on values such as Hardware, Infrastructure, Platform, Framework, Application and even Application Infrastructure. Vendors are not in agreement with what these designations mean nor are they consistent in describing themselves as belonging to these categories. Despite this, we will reproduce one interesting hierarchy that illustrates the use of these terms, with representative vendors (some at this point historical) populating the diagram (see Figure 4).

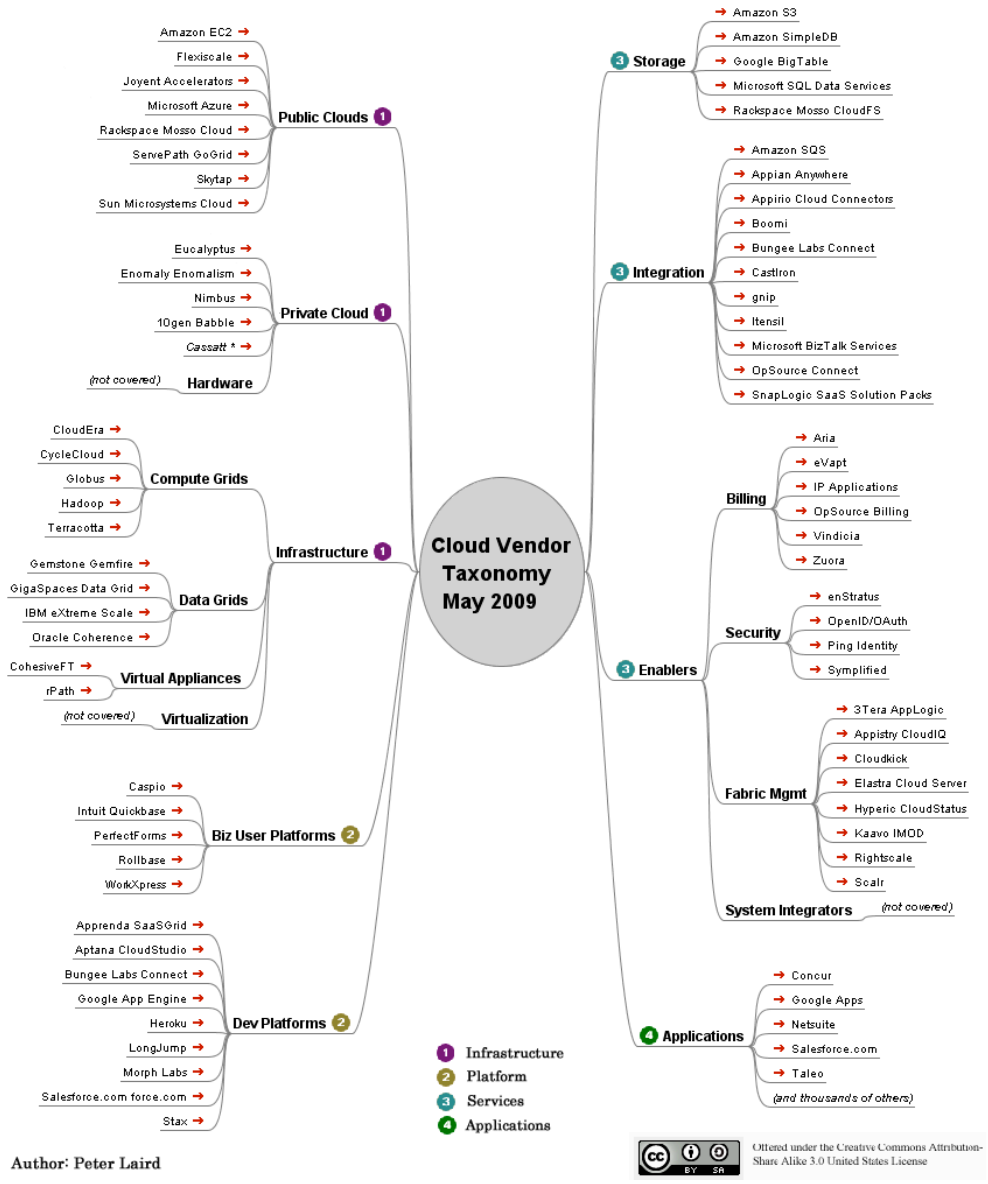


Figure 4 Cloud technologies are evolving as various vendors attempt to provide services populating the Cloud ecosystem. These services run the gamut from the hardware systems used to build cloud infrastructure to integration services and cloud-based applications. Courtesy Peter Laird http://peterlaird.blogspot.com.

Perhaps more useful, however, is to use a more simplified representation of the Cloud sprawl that can serve to highlight important aspects and key characteristics of different kinds of cloud offerings (see Figure 5).

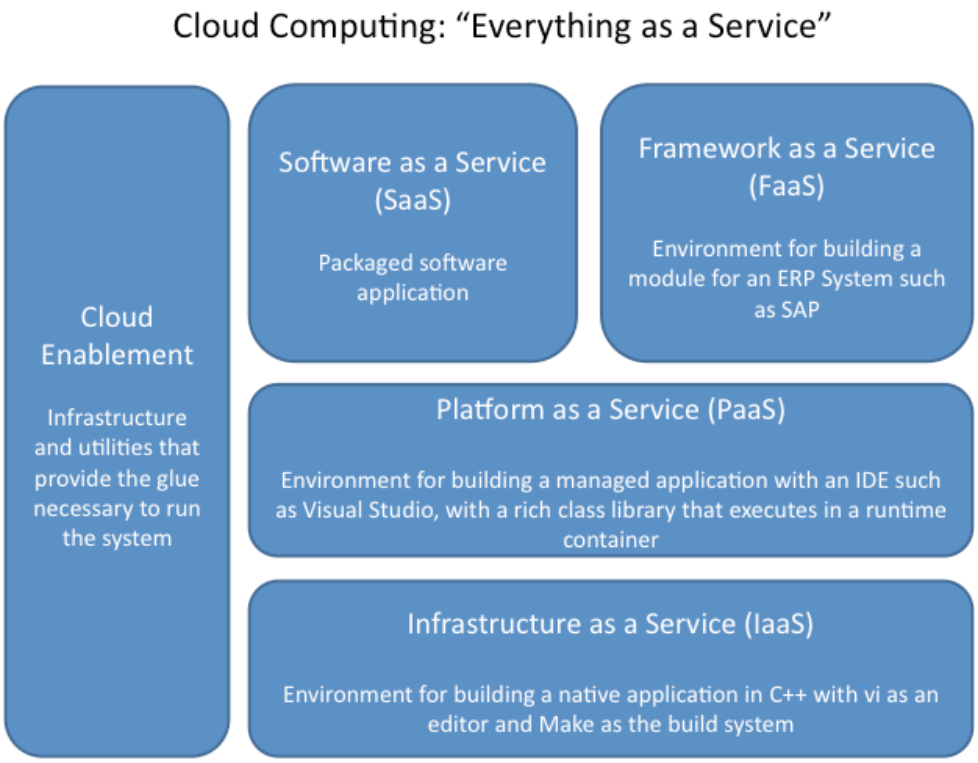


Figure 5 In the X-as-a-Service taxonomy cloud services are classified by the level of pre-packaging offered to the consumer of the specific service. An IaaS provides computing capabilities in the rawest form and hence offers the greatest flexibility. At the highest layers, there is less flexibility, but also less complexity to be managed.

What does XaaS mean generically? It means on-demand, requiring little or no capital expenditure. It means consumable remotely and across any mode of access over the Internet, and in a metered billing model.

Infrastructure as a Service (IaaS)

IAAS EXAMPLE: AMAZON ELASTIC COMPUTE CLOUD (EC2)

The lowest level of X as a Service is known as Infrastructure as a Service (IaaS) or, sometimes, Hardware as a Service (HaaS). A user of IaaS is operating at the lowest level of granularity available and with the least amount of pre-packaged functionality. An IaaS provider supplies virtual machine images of different operating system various flavors. These images can be tailored by the developer to run any custom or packaged application that can run natively on the chosen OS and be saved for a particular purpose. Instances of these virtual machine images can be brought online and used to run the application when needed. Use of these images are typically metered and charged in hour-long increments. Storage and bandwidth are also consumable commodities in an IaaS environment, with storage typically charged per Gb per month and bandwidth charged for transit into and out of the system. IaaS provides great flexibility and control over the Cloud resources being consumed, but typically with more work required of the developer to operate effectively in the environment.

Platform as a Service (PaaS)

PAAS EXAMPLES: GOOGLE APPENGINE, MICROSOFT AZURE

A Platform as a Service (PaaS) has some similarities to IaaS in that the fundamental billing quantities are somewhat similar. Consumption of CPU, bandwidth, and storage operates under similar models. The main difference between IaaS and PaaS is that there is less interaction required with the "bare metal" of the system. You do not need to interact directly and administer the virtual OSs directly. Instead, you can let the platform abstract that interaction away and concentrate specifically on writing the application. This simplification generally comes at

the cost of less flexibility and the requirement of coding in the specific languages supported by the particular PaaS provider.

Software as a Service (SaaS) and Framework as a Service (FaaS)

SAAS EXAMPLE: SALESFORCE.COM; FAAS EXAMPLE: FORCE.COM

Software as a Service (SaaS), as described earlier refers to services and applications that are available on an on-demand basis. A Framework as a Service (FaaS) is an environment that is adjunct to a SaaS offering and allows developers to extend the pre-built functionality of the SaaS applications. They are useful specifically for augmenting and enhancing the capabilities of the base SaaS system and can be used for creating either custom, specialized applications for a specific organization or general-purpose applications that can be made available to any customer of the SaaS offering. Like a PaaS environment, a developer in a FaaS environment is constrained to use the specific languages and APIs provided by the FaaS.

In addition to the classifications we just discussed, there are also some concepts that are important to introduce relative to kinds of Clouds as it relates to who owns and can utilize the resources. Private clouds are a variant of generic cloud computing where internal datacenter resources of an enterprise or organization are not made available to the general public. The public clouds of providers such as Amazon or Google were originally used as private clouds. If there are enough users within an organization, and there is enough overall capacity, a private cloud implementation can behave very much like a public cloud, albeit on a reduced scale. There has been a tremendous amount of capital investment in datacenter resources over the past decade, and one of the important movements will be the re-orienting of these assets toward cloud usage models. Hybrid clouds are a combination of private and public clouds. They might be used in cases where the capacity of a private cloud is exhausted and excess capacity needs to be provisioned elsewhere.

Summary

Cloud Computing represents a departure in the way IT systems are developed, operated, and managed. On the economic front there are potentially tremendous cost savings and much more potential flexibility than was ever before possible. Adoption of Cloud Computing has the potential of providing enormous economic benefit as well as greater flexibility and agility. We defined the Cloud as computing services that are offered by a third-party, are available for use when needed, and can be scaled dynamically in response to a changing need. Finally, we looked at a simple taxonomy that should help us understand the various flavors of cloud offerings available in the market today.