Global Knowledge®

Expert Reference Series of White Papers

# Types of Cloud Deployments

# Types of Cloud Deployments

John Hales, Global Knowledge VMware, SDN, and SoftLayer Instructor, A+, Network+, CTT+, MCSE, MCDBA, MOUS, VCP, VCAP, VCI, EMCSA

## Introduction

This white paper will discuss the basics of cloud computing, including a brief discussion on the location of the resources, followed by a review of the characteristics of cloud computing and the types (models) available. We will also briefly compare and contrast the various models.

This document is the first in a series of white papers that will discuss each of these cloud-computing models in further detail, with a separate document for each type. If you are already familiar with cloud computing, you may wish to skip this white paper and jump directly to the particular type(s) you are interested in.

## Cloud Computing Locations

While not very relevant to the cloud-computing models available (as each model is available at any of the possible locations), the locations nevertheless will be mentioned in this and future white papers and thus will be briefly defined here.

The National Institute of Standards and Technology (NIST), an arm of the US federal government), has defined much of what cloud computing is (at least to them, but as they are a standards organization, many others have followed their definitions). We will refer to them throughout this series of white papers for consistency.
Note that NIST doesn't call them cloud locations, but rather "Deployment Models."

### Public

The public location means that the resources (servers, storage, networking, and/or applications) you will be accessing are usually located on the Internet (hence publicly available and the name of this type). This is not always true as there are some specialized networks (such as those used by the government) that may have restricted access, but for the most part, the resources you want to access are reached via the Internet.

The broader definition is that the resources are owned by a third party (the cloud provider) which is rented (either by directly paying or via ads you are shown) in some fashion from them. The resources are located at one or more datacenters of the provider.

### Private

The private location means that the resources are (usually) owned by and accessible through a private network, but in any case always for the exclusive use by a single entity or company. Typically, the idea is that an IT department at an organization owns the resources and makes them available to employees of the company in the various ways that cloud-computing offers. This doesn't have to be the case, as a third party could own them and make them accessible to just that organization.

One of the biggest advantages is that the company owns, or at least controls, all the resources and can optimize them any way they wish and deploy them much more quickly than traditional methods provided; the (potential) downside is that the company must purchase all the resources. Another advantage of this model is that the company has complete control over all the security aspects of the deployment.

### Hybrid

Hybrid is simply some combination of the previous two locations, where some resources are located within the organization's datacenters and some are accessed publicly. This doesn't have to be the case, as it is possible to federate several private clouds or public clouds as well, but this is a far less common scenario.

Use cases for this model include the following:

- Development and testing, where resources can be quickly provisioned as needed and just as quickly deprovisioned when the project is complete. Sensitive company data may be stored in the private cloud onsite.

- Cloud bursting, where the normal load is handled by the company's own resources, but during period of peak demand (such as during the holidays for an e-commerce site), when the company's own resources are fully utilized, additional capacity is rented as needed to maintain desired performance levels. The advantage is that it is generally cheaper to own something than to rent or lease it if it will be used most or all of the time, but if needed for a short duration, renting is cheaper. This provides the best of both worlds, minimizing the total cost required to meet required performance levels.

- Backup/Disaster Recovery, where data may be kept onsite, but backed up offsite somewhere, similar to the way that tapes used to be shipped offsite. It can also be used for companies that need a disaster recovery location, but only have a single datacenter for all their resources and need someplace they can run temporarily in the event of an emergency, much like companies like Sun Guard (now Sun Guard Availability Services) provided in physical datacenters in the past. In other words, they kept servers in a datacenter that could be powered up in the event of a disaster. These servers were available to multiple customers.

## Characteristics of Cloud Computing

What is cloud computing? At its most basic, it is accessing resources over a network, usually the Internet. But cloud computing is really much more than that. There are five basics of cloud computing as defined by NIST, as follows:

- Rapid elasticity.
- Measured service.
- On-demand, self-service.
- Broad network access.
- Resource pooling.

These characteristics apply, no matter which cloud type you select, but you could have what most would consider a cloud-computing environment without meeting all five criteria. Let's begin with a brief review of each component.

### Rapid Elasticity

This is probably one of the most distinguishing characteristics of a cloud deployment vs. a virtualization project or any other IT project. What this means is that resources (especially computer resources, namely CPU and memory, but to a lesser extent storage and networking resources as well) can be scaled up or down very quickly to match demand. The most common example of this occurring in the physical world is power generation, which changes dynamically, almost instantly, to the changing needs for power by everyone on the grid. It appears seamless to us as consumers, but requires a lot of technology and monitoring controls to make it appear to be so easy.

So it is in the cloud-computing world. There are a lot of monitoring requirements to see changes to demand in real time (or near real time) if you are trying to scale based on the demand for resources. When more people come to an e-commerce site, for example, additional web servers may be needed to handle the load and keep latency low.

There is another use of rapid elasticity as well, namely not scaling based on end-user demand for resources, but the ability to scale based on projects that need resources quickly, such as adding a new research project, or cleaning up after a project has ended. This may also involve adding and/or removing resources quickly.

Why is this so important? Because it minimizes costs by providing resources Just In Time (JIT), instead of overprovisioning in case resources might be needed in the future or to handle relatively brief peaks, such as handling the huge increase in e-commerce traffic around the holidays at the end of the year that typically only lasts for a month. It also means you can stop paying for resources you don't need as soon as you're done with them.

It is important to note in this regard that to consumers, the resources appear virtually unlimited, but as a resource provider, care must be taken to ensure that sufficient capacity is available, and more is added as needed. Going back to the power generation example previously cited, if enough capacity doesn't exist to handle peak demand, either brownouts or rolling blackouts are instituted.

## Measured Service

The idea of the service being measured is that usage is measured and thus can be billed. In traditional computing, you purchase a server and then have full use of it for "free" for its lifetime (excluding power, maintenance, etc.). This is known as a capital expenditure (CapEx). This money is paid up front, so if you no longer need the server in a few months, for example, the money is already spent.

On the other hand, with cloud computing, you rent the resources you use by the hour, month, etc. There is no upfront CapEx cost, just an ongoing operating expenditure (OpEx) for as long as you need the device. For this scheme to work, usage needs to be measured. This may be flat fee based on a provisioned number of CPUs, gigabytes of RAM, etc., or even more granular with actual gigahertz of CPU time or gigabytes of RAM consumed. The same holds true for network and storage usage, and may extend further to software licenses, graphics cards, or whatever the provider wants to make available. Depending on the type of cloud computing being consumed, it could be time on a computing platform, not at actual server, or even something as simple as an email account.

Note that while it is measured, there may or may not be an actual bill to the user who requested the services of the provider. In public cloud scenarios, you are almost always billed (unless you are on a trial period or some other temporary, special arrangement or are at least shown ads or otherwise providing a revenue stream to the provider [such as with "free" email accounts]). In private cloud scenarios, however, the consumer may be billed for the resources (bill back, in other words an in-house transfer of funds between departments) or they may just be used internally to track how resources are consumed (show back, in other words no money changes hands, but IT can show where the budget is spent).

One quick note on the provider side: measuring is important for providers as well to understand demand and capacity and to plan for expected peak demand periods. It allows them to understand what is in use, report on it, and potentially control usage.

## On-Demand, Self-Service

This is one of the most important parts of cloud computing. The idea is that users can create and delete servers, storage, and networking without a formal approval process and lengthy delay by IT, including time to procure equipment, install it, and get it configured for the user, which often takes months. With cloud computing, the process may be as short as a few tens of minutes. The user can create and delete what they want, when they want, without human interaction.

Note that we have been discussing servers, but other resources can also be provisioned, such as an email account, or a web-hosting platform, depending on the type of computing being used.

### Broad Network Access

Another defining aspect of cloud computing is that there is wide access to the network used to access the resources. For example, in typical cloud-computing environments, this means network access is usually Internet-based, the most pervasive network on earth. Even in private cloud scenarios, most users in the company will have access to the resources (depending on security needs of course).

One of the enabling characteristics of cloud computing is widely available, fast network connections. Without this fast access from almost any device in the world in almost any place in the world, cloud computing would be limited to very few, at least a few very large organizations and governments. Another major enabling characteristic is the wide variety of devices that can be connected to networks and used with cloud computing, including desktops, laptops, tablets, and even smart phones.

### Resource Pooling

The idea of resource pooling is that resources can be combined and shared. For example, in the shared email account example mentioned earlier, you don't have an entire server to yourself; it is shared with many other users; in the aggregate, across all of the provider's servers, you are sharing resources with millions of others. Even if you get an entire server to yourself, you are still sharing network access at least, and often storage access as well, with others located in the same datacenter. Resource pooling provides the economies of scale that makes cloud computing economically viable. It also makes possible the rapid elasticity previously described.

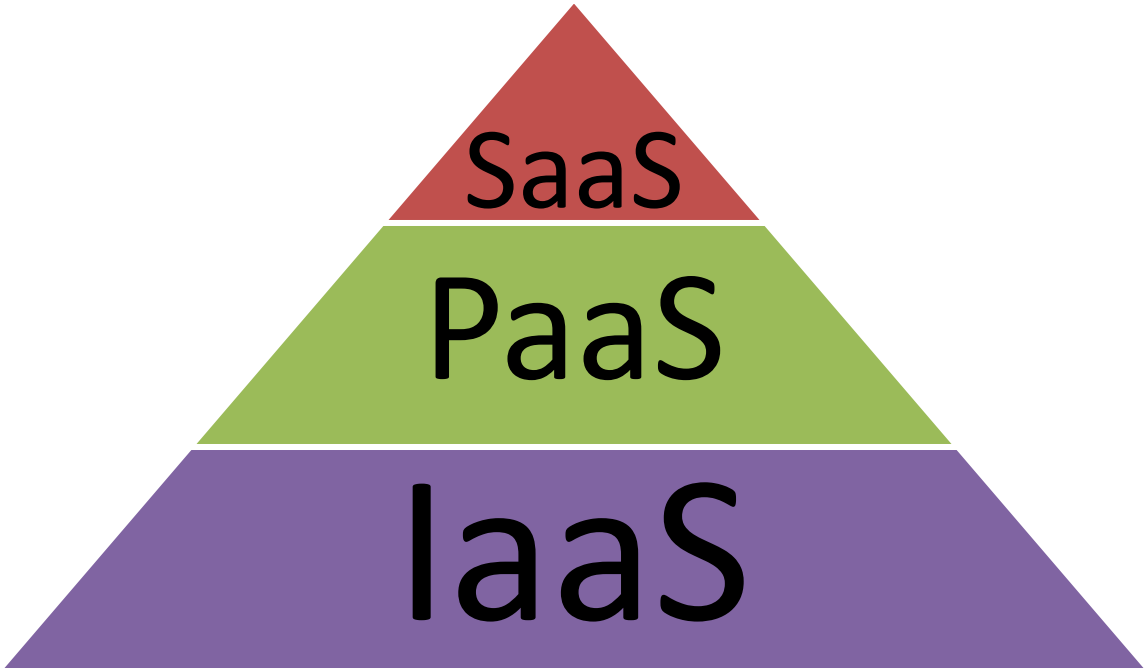
## Types of Cloud Computing

All of the forgoing material serves as background material and a prelude to this section, which is simply a foretaste of the series of white papers designed to go into the types of cloud computing in detail. While many people are familiar with the foregoing material, and maybe even are familiar with the basics listed below, fewer have in-depth knowledge of each of these areas, and thus each will have an entire white paper dedicated to them to discuss some common offerings from various vendors in each type of cloud. Each white paper will have an in-depth discussion of how to implement the particular cloud type (not vendor specific step-by-step directions, but an understanding of the process of implementing the specific cloud type), as well as the considerations that need to be understood when that type is implemented.

This section, then, provides the foundational understanding to take advantage of the rest of the white papers in this series.

Note that in the NIST definition, they call these types of clouds "Service Models."

Before we begin, a simple diagram will illustrate the relationships between the three most common types and will set the stage for the discussion of each type.




## Infrastructure as a Service (IaaS)

Drawn as the base of the pyramid, this is where infrastructure (physical or virtual servers, networking, and storage) is provisioned and the user has complete control over all aspects (subject to the offerings of the provider), including network speed, number and speed of CPUs, amount of RAM, etc. The user can configure the operating system, applications, etc. In short, it is much like deploying a physical or virtual server on premises today, except it may not be on premises and you don't pay for it all up front.

The advantage is complete control, while the disadvantage is complete control, meaning that the user is responsible for sizing, installing, and maintaining operating systems and applications, backing up the systems, etc.

In this model, networking is always shared in some fashion on the cloud provider's infrastructure with other consumers, and storage often is (though doesn't have to be) shared.

## Platform as a Service (PaaS)

PaaS is the next layer up in the pyramid, as the consumer doesn't determine server size, storage (at least directly — you may be able to choose how much you want), or networking, or even the operating system installed like they can with IaaS. Instead, the consumer uses the resources provisioned by the provider, and can utilize any programming language, utility, or tool provided by them to deploy their own applications. The consumer can thus focus on their application and not worry about all the underlying infrastructure, backups (potentially), etc. Be sure to read the terms and conditions and Service Level Agreements (SLAs) provided by the cloud provider to ensure that your needs are met, especially with regard to any backups, uptime guarantees, etc.

## Software as a Service (SaaS)

At the top of the pyramid is SaaS. With SaaS, you don't choose any of the server, storage, or networking specifics, and you don't choose an operating system, or even a language to develop new tools in. Rather, you get an application, such as email, a Customer Relationship Management (CRM) application, or a word processor. You can't choose anything other than settings allowed by the application, possibly including things like the amount of storage desired, default font, or things of an application nature. This is the most familiar scenario to most people today, even if they don't know these applications are SaaS applications, as we use them every day. Things like Gmail and Map quest are ubiquitous today.

Note that SaaS applications can be accessed via a browser where the great majority of the processing is done on the cloud provider's infrastructure, or via an application, such as one that runs on Android or iOS, or even a Windows or PC application, where the processing is shared between the cloud provider's infrastructure and the consumer's device. In either case, data is typically stored on the cloud provider's infrastructure.

## Compare / Contrast Cloud Types

As just described, there are three primary cloud types and each is best suited to specific use cases, many of which will be described in the individual white papers in this series in the appropriate sections. If you just want to use an application, then SaaS is for you, while if you want to develop a program, website, etc., than PaaS is for you. On the other hand, if you want maximum control over all aspects of the deployment (subject to the provider's offered capabilities), then choose IaaS.

There are a wide range of offerings in each broad category as well, and these will also be discussed in the relevant related white papers.

## Conclusion

Cloud computing is a big force in IT today, and it isn't going away. In fact, cloud adoption is going up geometrically, both for end users (think apps on your phone or tablet) as well as for organizations of all sizes. In fact, many smaller organizations may not have any on-premises infrastructure at all, other than networking infrastructure to get connected to the cloud. With this transformation in IT, it behooves all of us in the industry to understand it and adapt or risk being out of a job, like punch card operators. We need to understand the strengths and weaknesses of each of the location and types of cloud offerings available, and make use of them as appropriate, based on business objectives, regulatory requirements, etc. That is the ultimate purpose of this series of white papers, and we hope it helps you to plan and prepare for the future that is rapidly coming upon us.

For more information on projected growth rates, expected balances between types of cloud and locations of cloud delivered services, etc., check out this white paper by Cisco: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.pdf.

## Learn More

Learn more about how you can improve productivity, enhance efficiency, and sharpen your competitive edge through training.

- Cloud Essentials
- Cloud and Virtualization Essentials

Visit **www.globalknowledge.com** or call **1-800-COURSES (1-800-268-7737)** to speak with a Global Knowledge training advisor.

## About the Author

John Hales, VCP, VCP-DT, VCAP-DCA, VCI, is a VMware, SDN, and SoftLayer instructor at Global Knowledge. He is a lead instructor for the SoftLayer offerings worldwide. John is also the author of many books, from involved technical books from Sybex to exam preparation books, to many quick reference guides from BarCharts, in addition to custom courseware for individual customers. John has various certifications, including the VMware VCA-DCV, VCA-DT, VCA-Cloud, VCP, VCP-DT, VCAP-DCA, VCI, and VCI Level 2, the Microsoft MCSE, MCDBA, MOUS, and MCT, the EMC EMCSA (EMC Storage Administrator), and the CompTIA A+, Network+, and CTT+. John lives with his wife and children in Sunrise, Florida.