# The Cloud: What is it?

**Sample from Microsoft's "Enterprise Cloud Strategy 2nd Edition by Barry Briggs and Eduardo Kassner**

**Distributed by: CCG**

PUBLISHED BY
Microsoft Press
A division of Microsoft Corporation
One Microsoft Way
Redmond, Washington 98052-6399



Microsoft Press books are available through booksellers and distributors worldwide. If you need support related to this book, email Microsoft Press Support at mspinput@microsoft.com. Please tell us what you think of this book at http://aka.ms/tellpress.



**Acquisitions Editor:** Janine Patrick
**Developmental Editor:** Bob Russell, Octal Publishing, Inc.
**Editorial Production:** Dianne Russell, Octal Publishing, Inc.
**Copyeditor:** Bob Russell

# The cloud: what is it?

Like any new technology, the cloud comes with a whole new set of terms, acronyms and abbreviations. Nevertheless, it's important to understand the different forms of cloud computing in order to make the right decisions about how to use it. In this chapter, we examine the ways in which the cloud manifests itself and how you can employ each in enterprise computing.

So, what is the cloud?

At its core, the cloud physically consists of millions of servers distributed across multiple, very large datacenters strategically located all over the world. All cloud providers use custom-designed server hardware that is focused on reducing cost, improving environment footprint, and, of course, providing the greatest compute capability.

The datacenters in which the servers are contained are themselves designed for maximum efficiency and minimal environmental impact; considerable research goes into making datacenters as "green" as possible. For example, Microsoft's Quincy, Washington datacenter is located next to a hydroelectric facility on the Columbia River; it is emblematic of how cloud providers take advantage of local opportunities to reduce their carbon footprints. Elsewhere, datacenters in cooler climates use ambient air rather than air conditioning systems to reduce electric consumption. Some providers use wind power, and others use cheaper nonpotable water in the air conditioning systems where that is necessary.

A key measure of datacenter efficiency is called *Power Usage Effectiveness* (PUE), which measures how power coming into the datacenter is used. A perfect PUE score is 1.0, meaning that all of the power goes to the computing equipment (formally, PUE is defined as total facility energy used divided by that consumed by IT equipment). Traditional enterprise datacenters typically realize a PUE of 2.0, meaning that half of the incoming energy is used by noncomputing equipment such as air conditioning, lighting, and so forth. Cloud datacenters are now achieving PUE ratings of 1.1 and even less. This is the result of significant investment and innovation on the part of the cloud providers.

# Public, Private, and Hybrid Clouds

In the following subsections, we define and examine the three major cloud models.

## Private Cloud

The first set of definitions we'll discuss is the distinction between "private" and "public" clouds.

The term *private cloud* is often misused; some will say it is the same as a traditional on-premises datacenter. In fact, they are very different. In the traditional on-premises model, IT departments purchase hardware as applications need them, and often this year's servers will look and behave very differently from last year's. Moreover, IT departments traditionally maintain a mix of hardware and software, ranging from mainframe to PC server, with a variety of operating systems, databases, and other system software. All of this effectively prevents the notion of on-demand computing, which is the essence of the cloud.

In a private cloud, technologies specific to the cloud model are hosted in an on-premises datacenter, with large numbers of commodity hardware running identical system software: in other words, a "cloud" that belongs to you. Private clouds can be useful because they can implement a technology stack that is consistent with the public cloud. This might be necessary in scenarios for which certain applications or data cannot be moved off premises (we discuss reasons for not moving to the public cloud in Chapter 6.)

However, private clouds are of very limited utility. They do not provide the cost savings and efficiencies that the public cloud can, because private clouds require a significant capital expense budget and an operations staff; thus, they remain on your company's balance sheet. Moreover, individual companies cannot achieve the aforementioned economies of scale of a public cloud provider, so their costs are proportionately higher.

## Public Cloud

A *public cloud*, which is the primary focus of this book, is built, managed, and maintained by a large technology vendor that makes computing, storage, and software available on a rental basis. The leading public cloud vendors have datacenters all over the world with literally millions of servers available for use. Customers (enterprises) can either take advantage of applications that already exist in the cloud or they can upload their own proprietary applications, and, as we shall see, there are a number of ways in which applications can physically exist in the cloud but appear to be private to the enterprise corporate network.

## Hybrid Cloud

Often, an enterprise will want to keep some of its applications on-premises while moving others to the public cloud. Of course, it is desirable that all of these applications continue to run as they did previously; that is, as if they were all still local and on the same network. When some applications are in the cloud and some are on-premises, this is termed a *hybrid cloud*. Every enterprise will have a hybrid cloud at some time, even if they plan to eventually move all of their applications off-premises, there will be a time during the transition when some applications have moved and others have not: a hybrid model.

To securely connect the two environments, multiple solutions exist. You can set up a Virtual Private Network (VPN), which makes cloud applications appear to be on the same internal network as the enterprise. You can set up VPNs on a per-application basis or, with a hardware device, for the entire corporate ecosystem.

Alternatively, enterprises can purchase through their telecom provider a dedicated line linking the corporate datacenter with the cloud; bandwidth can be purchased as needed. This solution is preferable when it is desired to keep all traffic off of the public Internet or when substantially higher bandwidth is required. However, it of course entails additional cost.

# Hyperscale makes computing an on-demand service

With the cloud, computing can operate at *hyperscale,* meaning that computing resources scale with the demand placed on them. Hyperscale computing requires the ready availability of whatever computing capabilities you need, whenever you need them. Thus, if you need 10,000 servers for an overnight big data analytics job, but only for a few hours, you'll have them, and then you can release them back when finished. Hyperscale also implies the notion of configurability (and reconfigurability) at scale. Today, a given server might be allocated to a particular real-time application with very high Service-Level Agreement (SLA); tomorrow, it might be assigned a background task with a very different SLA, all at the request of the consumer of cloud functions.

Hyperscale also means that computing capability can be accessed from anywhere in the world with similar latency, which in turn means that cloud providers must build enormous datacenters all over the planet (which they have). The global scale of the public cloud in turn provides any number of new capabilities, such as the ability to do geo-distribution of data and to do cross-region failover, to name a just two.

The potential, then, of hyperscale computing—its features and its economics—far exceeds that of any enterprise datacenter.

Because of this incredible global scale, computing can be provided *as a service,* meaning that the cloud offers a set of capabilities that enterprises can rent and use for a period of time, add on to as more capability is needed, and then discontinue when no longer needed. Of course, as we've noted, this model is analogous to other commonly used services such as telecom, electricity, and so on: you pay for what you use and no more.

# "As a service"

As we've said, in the cloud, computing is made available as a service, and there are three predominant application models for cloud computing. Let's take a closer look at each of them.

## Infrastructure as a service

With the infrastructure as a service (IaaS—pronounced "eye-as") model, you are renting only the server hardware and a small amount of software (the hypervisor) to host your application's virtual machine (VM), where the VM consists of the operating system, associated system software, and the application itself. IaaS means that VMs are simply *moved* from on-premises to the cloud. Figure 2-1 illustrates that many operating systems and applications can coexist on a cloud server. A thin piece of code called a *hypervisor* ensures that each one runs in a timely and efficient fashion.

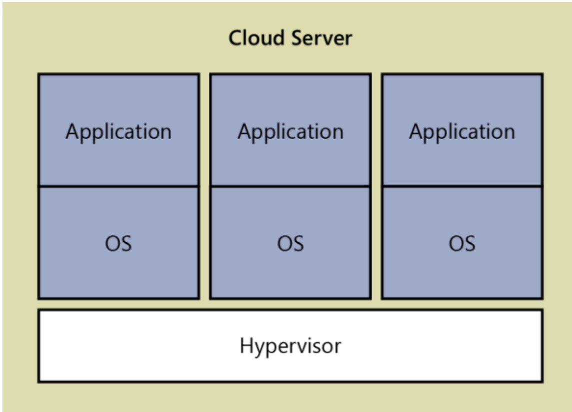


**Figure 2-1:** Infrastructure as a service

In other words, you supply—and maintain—the pieces highlighted in blue in Figure 2-1.

This is the easiest and fastest migration strategy; it offers many benefits, including cost savings. However, it still means that your operations staff will need to perform such tasks as patch management, updates, and upgrades. Nevertheless, IaaS is one of the most common cloud deployment patterns to date because it reduces the time between purchasing and deployment to almost nothing. Additionally, because it is the most similar to how IT operates today, it provides an easy onboarding ramp for your current IT culture and processes. As we shall see, the bulk of migration especially in the early phases of cloud adoption is to IaaS.

## Platform as a service

In platform as a service (PaaS—pronounced "pahz"), the cloud provider maintains all system software, removing the burden of upgrades and patches from the IT department. In a PaaS deployment model (Figure 2-2), all that the enterprise needs to focus on is deploying its code on the PaaS machines; the cloud provider ensures that operating systems, database software, integration software, and other features are maintained, kept up to date, and achieve a high SLA.

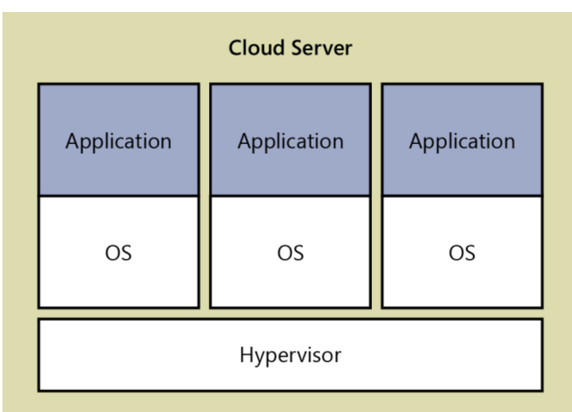


**Figure 2-2:** Platform as a service

Note that in Figure 2-2 the pieces in blue—the parts that the user must supply and maintain—consist *only* of the application.

PaaS provides IT departments with important benefits, most important among them being the cost savings associated with reduced or eliminated maintenance of system software and other rote functions. However, PaaS usually implies some redesign of the application in order to best take advantage of the model.

## Software as a service

In software as a service (SaaS—pronounced "sass"), you simply rent an application from a vendor, such as Microsoft Office 365 for email and productivity. This is by far the most cost-effective of all the options because typically the only work involved for the IT department is provisioning users and data and, perhaps, integrating the application with single sign-on (SSO). Typically, SaaS applications are used for functions that are not considered business-differentiating, for which custom or customized applications encode the competitively differentiating business models and rules.

As we discuss further in Chapter 6, when choosing how to move functionality to the cloud, you should always be on the lookout for opportunities to use SaaS-based applications. Usually, they will provide you with the highest return on investment.

## Containers

Containers—which lie somewhere between IaaS and PaaS on the "as-a-service" spectrum—are a means by which applications can share a single instance of an operating system, as illustrated in Figure 2-3. This provides the appropriate isolation and security guarantees preventing applications from "stepping" on one another. Because starting a containerized application typically does not involve loading and initializing an entire VM with an operating system, container startup can be very fast, so scale-up and scale-down can be very performant.

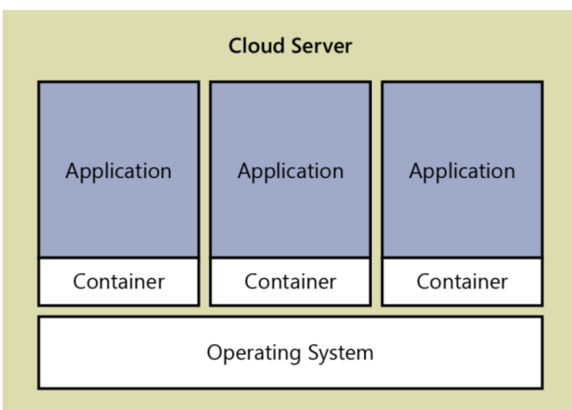


**Figure 2-3:** Container architecture

Containers have many advantages. Often it is possible to package an application with few or little changes to run within a container. Having created containers, it's often useful to deploy multiple copies for scale or resiliency reasons. A related technology, *orchestration,* can help automate the process of deploying many copies of many different applications or components to a *cluster* of servers. We discuss all of this, including tradeoffs, in more detail in Chapter 10.

# “As-a-service,” compared

Figure 2-4 compares the various “as-a-service” technologies with on-premises computing. The items in blue represent components or software that the enterprise (you) are responsible for maintaining; the items in orange are the responsibility of the cloud provider.

As you can see, for an on-premises datacenter, the enterprise is fully responsible for everything, from the datacenter’s operation, the facilities, electricity and air conditioning, all the way through the application. As the migration to the cloud progresses, more and more of these expenses are borne by the cloud provider.

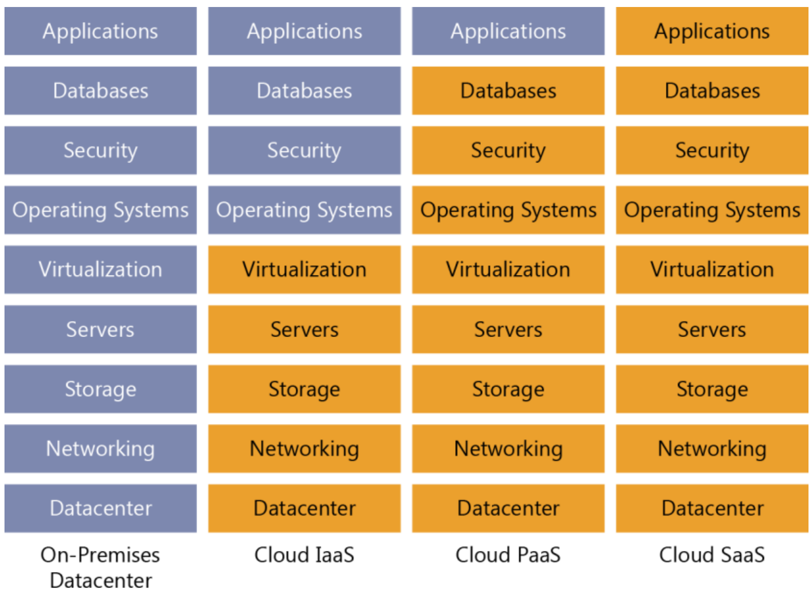


**Figure 2-4:** “As-a-service” compared

CCG is a cloud, data, and analytics company which enables enterprises to do more with their data. By partnering with CCG, you'll gain access to an in-depth cloud strategy and roadmap, custom migration and optimization guidance from our cloud engineers, and 24/7 access to a cloud management team. Learn more about CCG's cloud offers by contacting a Cloud Strategy and Migration Expert or by visiting us on our Cloud Strategy and Migration solution page.

# Contact CCG:


www.ccganalytics.com
info@ccganalytics.com

(813) 265-3239