GP STRATEGIES™

White Paper:

Cloud Computing Characteristics Are Key

by

Christopher Olive
Chief Architect
GP Strategies™ Corporation

What is cloud computing?

Cloud computing remains the buzzword winner of the current technology paradigm. The National Institute of Standards and Technology (NIST) defines cloud computing broadly as "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (for example, networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."[1] NIST further describes cloud along three axes: characteristics, deployment model, and service model. This provides a rather bewildering matrix of possible implementations of cloud. However, by focusing on the characteristics that NIST defines, GP Strategies can see how close to an "ideal" cloud implementation your organization can perform. One central element to note about the NIST definition is that:

> *Your implementation does not need to have all of the identified characteristics to be considered cloud computing!*

To reground our understanding of the cloud, GP Strategies discusses the overall cloud implementation and how it works with enterprise-level applications in its white paper, "IaaS and Enterprise Hosting: A Match Made in the Hosted Cloud,"[2] .

Several short key points to remember:

- Cloud is not virtualization. That is a separate topic. However, many of the underlying benefits of cloud computing would be impossible to achieve without virtualization, so it is safe to say that cloud computing requires virtualization.
- Cloud isn't really a technology at all. It is a manner of service delivery and its associated billing.

[1] "The NIST Definition of Cloud Computing." AUTHORS: Peter Mell and Tim Grance. Version 15, 10-7-09.

[2] "IaaS and Enterprise Hosting: A Match Made in the Hosted Cloud." AUTHORS: Christopher McHugh, Christopher Olive, and Andrew Wortman. 7-16-10. (http://www.gpstrategies.com/common/downloads/wpIaaS.pdf)

While a full discussion of virtualization is significantly outside the scope of this paper, in terms of cloud computing, it is important to understand the basics:

- When running applications on traditional physical hardware, it is common for most of that hardware's capacity to lay unused for the majority of the time.
- Virtualization inserts a layer of abstraction (the hypervisor) between the hardware and operating systems.
- This hypervisor allows multiple instances of an operating system to share the same physical hardware and portions out the requested compute power (for example, CPU, disk, RAM, etc.) as needed.

Characteristics shape your migration

To help highlight this portion of cloud computing, consider the characteristics identified by NIST in their definition paper:

1. On-demand self-service
2. Broad network access
3. Resource pooling
4. Rapid elasticity
5. Metered service metrics

This section analyzes these characteristics and identifies the salient features of each to determine the role that they may play in your company's migration to the cloud.

On-demand self-service

The first characteristic is one of the easiest to define. It simply requires these two things to be true:

1. The service must be always available (or some reasonable approximation of always).
2. The service received must be modifiable by the client organization without contacting the hosting provider.

It is the second that is typically the most difficult to meet. While the public providers like Amazon, Google, and Microsoft have this facet, smaller niche providers typically do not. This is more likely when the provider is also supplying services or managed hosting for the application itself, especially during an Infrastructure as a Service (IaaS)-type scenario. In enterprise application hosting scenarios, there are also potential contractual issues to consider when decreasing (or even possibly increasing) capacity without interacting with the vendor.

It is important to determine these issues before making changes to your own environment. If your service/uptime SLAs require a certain level of hardware to support, remember to ensure that you do not compromise them by unduly changing the capacity available to your applications.

Broad network access

In this context, broad network access means that the hosted application should be reachable via nearly any network-based appliance. These can include, but are not limited to, the following:

- Laptop
- Desktop
- Smartphone
- Tablet device

Broad network access is typically accomplished by using the built-in web browser for the device, as it is one of the most ubiquitous clients available. It is not the only client, as the virtual desktop (beyond the scope of this document) requires more specialized software, but it is one of the most commonly selected clients. The advantage of this setup is that client devices can be much less powerful as "thin-clients" rather than "fat-clients" (read this as needing to install software on the client to make it work).

In the early '90s, companies were migrating away from "dumb terminals" that were directly connected to the heavy iron of a mainframe or minicomputer, and going with desktop-based, PC-type machines. This is exactly the opposite of what we see here, as the pendulum swings back in favor of lighter clients, with the cloud replacing the big iron.

Resource pooling

Resource pooling is the concept that multiple organizations can share the underlying physical cloud infrastructure. This allows significantly greater purchasing power for these companies because they can typically obtain access to a larger pool of resources rather than procuring the physical or virtual infrastructure themselves. Typically, user organizations of similar security levels or needs are grouped together on a particular community cloud offering (all federal organizations, all pharmaceutical organizations, all general availability organizations, as examples, live on separate physical cloud infrastructures).

This is typically one of the more difficult paradigm shifts for security professionals. The concept of shared infrastructure (at any level) is not considered as being secure. Since the underlying infrastructure is shared in this scenario, alarm bells usually go off in the heads of the assigned security personnel. However, the hypervisor takes steps to isolate the virtual machines running on it, and there are several layered approaches that can added. This approach combines both new and traditional security measures, some of which are highlighted as follows:

1. Hypervisor products can usually include other plug-in objects to assist in security. VMWare and Microsoft make these for their own products, usually folded into the core offering, and other third party tools can be used as well.
2. Adding in Network Intrusion Detection software (like SNORT) to listen on the traffic between virtual machines (VMs) is critical, as they do not usually communicate on the external network topology like traditional servers.
3. Adding in VM-based virus scanning to look at the underlying physical files allows more cost-effective anti-virus protection.
4. Updating policies and procedures to address the additional steps necessary for employees to deploy cloud and virtualization is a key component to successful resource pooling.

With proper protocols and procedures, security risks can be mitigated and allow even higher level security operations in the cloud.

The full range of security steps necessary is outside the scope of this document, but adoption of virtualization by security personnel in the federal space is increasing with such research as the CDW Server Virtualization Life Cycle Report showing that, while adoption is slower than commercial progress, federal IT personnel are embracing virtualization and the cloud.

Rapid elasticity

Rapid elasticity is (nearly) exactly what it says on the tin. This is the ability to handle spikes in usage at least semi-automatically. While this is something you could technically obtain with physical hardware, the turnaround time necessary for implementation typically pushes that solution outside the bounds of the definition of the word “rapid.”

For example, if the application typically sees between 1,000 and 2,000 users a day, but at certain times of the week/month/year there are more users, you usually have to provision the application to handle the spikes in usage. This allows you to handle these sudden (or perhaps not so sudden) upticks in usage.

With a cloud-based application, you do not need to account for these spikes as widely. Typically, you would provision your application service for the usual level of concurrence. Then, one of two things might happen:

1. There is an anticipated spike in usage. For example, you have an application that has to be used by everyone in your organization to handle training during a three-month window.
2. A sudden, unanticipated surge in usage within a small time period occurs. Perhaps you are delivering information that happens to go viral across the larger Internet.

Prior to the option of a cloud-based application, meeting unanticipated demand (and even anticipated demand) required far more planning and cost. You either had to have servers sitting fallow, waiting to be swapped into the pool, or pay for overcapacity to handle the surges.

With cloud services, you can simply plan with your hosting provider to increase capacity during anticipated peaks, with the new (likely virtual) servers provisioned and deprovisioned for you and not remaining inactive during the rest of the year. Alternately, during a period of unanticipated load, your provider may have configured your system to automatically grow when usage reaches a certain threshold, creating and adding virtual servers to your service with a set of scripts, and then removing them in the same way when the demand hits a certain floor. The second is a more advanced cloud offering and is not available with very many providers; however, it is becoming more of an option.

Metered usage

Metered usage is also a straightforward idea: you only pay the hosting provider for the resources you consume. This makes IT more of a utility service you pay for as opposed to the traditional cost models, where you might pay some dollar figure a month to host X number of servers. However, when metered usage is applied to more complex deployments of applications, it can become muddled.

The sticking point comes when one tries to define use. Simply put, metered usage may be the concept of “$X per minute or hour your server is powered on,” and this is how most public cloud providers operate. This is an issue when one is looking to host an application that is needed 24x7, when there is no significant cost advantage to having a cloud-hosted solution (there are technical and service advantages, but we are isolating cost for this example).

From here, one may consider moving to a more granular charging fee from the hosting provider. But then we have to consider what use is. Is it sending or receiving an email in a cloud-hosted Exchange service? Is it per report run for an analytical business intelligence application? Or perhaps it's per learning completion in a Learning Management System (LMS). We then have to consider the different weights; should an email with a 2MB attachment "cost" the same as a simple text one? Should a simple report be equal to a financial end-of-year statement report? How would we accommodate other types of users and charge appropriately for an LMS? These situations can complicate the overall seeming ease of this characteristic.

So where does that leave us? Is the cloud ready for enterprise hosting?

There is no actual simple answer. The slightly less simple answer is "Yes, but with some caveats." Breaking down those considerations by characteristic, we have the following:

1. **On-demand self-service**
   This is likely to be the least adopted of the characteristics in an enterprise-class hosted application. Typically, you want to leave the sizing considerations for the application to the experts and leave them alone. At the very least, having direct access and making changes to the configuration of the hosted environment should be very carefully considered.
2. **Broad network access**
   This is more a function of the application being hosted. If we consider that nearly all enterprise applications are now web-based or have a web-based component, this is satisfied.
3. **Resource pooling**
   Security needs are likely to be the bottleneck here. Depending on your security organization's comfort level with "sharing" equipment, there may be some accreditation meetings that need to happen before they will sign off.
4. **Rapid elasticity**
   This is one of the key characteristics if your application experiences spikes in usage. This does not need to be fully automated, but it should be relatively easy to provision additional servers if you have anticipated heavy usage. You may want to consider a base "plan" of servers and a payment model to handle spikes.
5. **Metered usage**
   This goes hand in hand with elasticity. Begin with a base configuration to handle the majority of your time, and have a billing model set up with your provider to allow for unanticipated (or short-term anticipated) pay-per-unit (whether that unit be per server per hour or something else).

Three steps to reach the cloud

So, to consider the cloud (public, private, or somewhere in between), remember:

1. Get your security team's buy-in as early as possible in your research phase.
2. Settle on a cost model that works for both you and your provider. Consider the base cost + usage for spikes, or find something else that is equitable.
3. Realize that you may not need "true" cloud computing, but some hybrid of the characteristics most important to your project.

While cloud computing is indeed the next frontier in enterprise computing, careful planning and examination will make the journey easier. Involve all your stakeholders, both business and technical, to increase the return you get from making the transition.

About the Author

Mr. Olive is the Chief Architect for the GP Strategies Enterprise Technology Solutions group. He oversees the research and development efforts there and serves as technical and architectural oversight for several clients. Mr. Olive has extensive knowledge of the technical challenges facing enterprise hosting clients today. ■

For more information contact
Christopher Olive
Chief Architect
colive@gpstrategies.com

GP Strategies Corporation
6095 Marshalee Drive, Suite 300
Elkridge, MD 21075 USA

GP STRATEGIES™
Knowledge. Performance. Impact.

gpstrategies.com
1.888.843.4784
info@gpstrategies.com