Gartner Research

Publication Date: 17 June 2008 ID Number: G00156220

Cloud Computing: Defining and Describing an Emerging Phenomenon

Daryl C. Plummer, Thomas J. Bittman, Tom Austin, David W. Cearley, David Mitchell Smith

The cloud is emerging as the latest way to approach alternative delivery models for IT capabilities. It is a way of delivering IT-enabled services in the form of software, infrastructure and more. This research examines the definition of cloud computing and how it will evolve.

Key Findings

- Cloud computing is about delivering IT-enabled services.
- While cloud services do not have to be massively scaled, the best cloud services will be both scalable and elastic.
- Not all cloud services will be available or usable immediately.

Recommendations

- Begin to catalog what cloud services are available (from Amazon to Zoho) and consider what their use will mean to your (or your customers') interests.
- Demand that your technology providers explain how Web/cloud platform models will affect their offering and pricing strategies.
- Determine whether the cloud will ultimately be robust enough to meet your business goals.
- Determine when or if you would be able to offer services to the cloud or whether your role will be as a service consumer.

TABLE OF CONTENTS

Gartner

1.0 What You Need to Know

This document was revised on 24 June 2008. For more information, see the Corrections page on gartner.com.

Cloud computing is an alternative delivery and acquisition model for IT-related services. This paradigm will shift the way purchasers of IT products and services contract with vendors and the way those vendors deliver their wares. However, there are also risks and barriers to the emergence of this model. A definition of cloud computing is a style of computing where massively scalable IT-enabled capabilities are delivered 'as a service' to external customers using Internet technologies. This leads to the industrialization of IT and will alter the way many organizations deliver business services that are enabled by IT. It also leads to new ways of acquiring and using technology that not all businesses will be ready for.

2.0 Body

Throughout the history of business, delivery of shared business services has been a key enabler of growth and a way to more consistently penetrate larger and wider customer bases. For example, shared product delivery services have reduced costs for shippers and consumers, and shared customer service processes have shaped an entire era of responsive call centers.

If we go a bit further back, the advent of industrialization offered the world the ability to deliver a wide range of products at a price that was low enough to make available, to average people, products that were previously only available to the wealthy or to governments and businesses. It was only a matter of time before shared IT services gained significant penetration to foster dramatic shifts in the IT industry. The use of virtualization technologies (see Note 1), open source, service-oriented architectures and widely available computing standards, combined with the pervasiveness of the global Internet (see Note 2), is making computing-related services generally available to the world at reduced costs and massive scale. Because of this, new IT service delivery and acquisition models are emerging. One such model has been termed "cloud computing." In this research, we examine Gartner's definition of cloud computing and its place in the changing landscape of IT.

3.0 What Is Cloud Computing?

Gartner defines cloud computing as "a style of computing where massively scalable IT-enabled capabilities are delivered 'as a service' to external customers using Internet technologies."

If we break down Gartner's definition, what we find is a set of mutually supportive concepts. First and foremost is the concept of delivering services (that is, results as opposed to components). Implementation doesn't matter as long as the results of the implementation can be defined and measured in terms of a service with associated service-level requirements. Included in this concept is payment based on usage, not on physical assets. The payment can be subsidized (for example, by advertising) or paid directly by the customer. The second concept is that of massive scalability. Economies of scale reduce the cost of the service. Implicit in the idea of scalability is flexibility and low barriers to entry for customers. Third, delivery using Internet technologies implies that specific standards that are pervasive, accessible and visible in a global sense are used. Finally, these services are provided to multiple external customers, leveraging shared resources to increase the economies of scale.

4.0 Scalability vs. Elasticity

The question of scalability is often discussed in relation to the cloud. Gartner's definition mentions massive scalability but does not explicitly mention the concept of elasticity. Scale is an aspect of performance and the ability to support customer needs. The concept of elasticity is related to the ability to support those needs in large or small scale at will. The key issue with elasticity is the ability for a system to scale both in an upward direction (for example, to millions of users) and in a downward direction (for example, to one user) without disrupting the economics of the business model associated with the cloud service. Typical enterprise-class systems are scaled in the upward direction, but the cost of running that scaled complex for one or two users would be prohibitive given the cost of operating and maintaining the environment. However, global-class providers like Google, eBay or Zoho have a model that is not primarily based on the cost of the infrastructure and operations or software licenses and maintenance, but leverages ad revenue and other mechanisms to support their existence. This means that the number of users and what they are doing, while extremely relevant, can be detached from the operating economics of the cloud provider.

In addition to the issue of elasticity is the question of whether something that does not provide massive scalability is to be considered cloud or not. The definition is not intended to be a threshold that determines inclusion in the cloud model. Instead, it is intended as a guideline for the relative "cloudiness" of a particular solution. This means that massive scale is not an ultimate attribute for cloud-computing providers — It is an indicator of where they are in the relative breadth of their cloud solutions. As cloud computing evolves, a small number of massive IT providers will emerge supporting massive workloads, massive data manipulation and general-purpose services. Their differentiation will primarily be economies of scale. However, there will also be a "long tail" of midsize and even relatively small providers differentiating on leading-edge and special-purpose technologies, thriving on continued IT innovation and maintaining price pressure on even the most massive providers.

5.0 Changing How Services Are Delivered

During the past 15 years, a continuing trend toward industrialization of IT has grown in popularity. IT services delivered through hardware, software and people are becoming repeatable and usable by a wide range of customers and service providers. This is, in part, because of commoditization and standardization of technologies, virtualization and the rise of service-oriented software architectures, and, most importantly, the dramatic growth in popularity/use of the Internet and the Web. These things, taken together, constitute the basis of a discontinuity that amounts to a new opportunity to shape the relationship between those who use IT services and those who sell them.

What the discontinuity implies is that the ability to deliver specialized services in IT can now be paired with the ability to deliver those services in an industrialized and pervasive way. The reality of that implication is that users of IT-related services can focus on what the services provide to them rather than how the services are implemented or hosted. Just as utility companies sell power to subscribers and telephone companies sell voice and data services, IT services such as network security management, data center hosting or even departmental billing can now be easily delivered as a contractual service. The buying decision then shifts from buying products that enable the delivery of some function (like billing) toward contracting with someone else to deliver those functions. Certainly, this is not new, but it does represent a different model from the license-based, on-premises models that have been dominant in the IT industry for so long. Names for this type of operation have come into vogue at different times. Utility computing, software as a service (SaaS) and application service providers (ASPs) all have their places in the pantheon of

industrialized delivery models. However, none has garnered widespread acceptance as the central theme for how any and all IT-related services can be delivered to the world.

The types of IT services that can be provided range far and wide. Compute facilities (for example, Amazon's Elastic Compute Cloud [EC2]) provide computational services so that users can use CPU cycles without the need to buy computers. Storage services (for example, Amazon's Simple Storage Service [S3]) provide a way to archive data and documents without the need to continually grow farms of storage networks and servers. SaaS companies like salesforce.com offer CRM services through their multitenant shared facilities to clients to manage their customers without ever buying any software. This represents only the beginning of options for delivering complex capabilities of all kinds to businesses and individuals alike. Even client computing can be provided as a service (through hosted desktop and virtual-machine technologies), thus potentially removing the need for a PC altogether.

These things, and more, are increasingly falling under the banner of cloud computing.

6.0 Perspectives on the Cloud

The basics behind cloud computing are not new. One can nitpick the definition by asking what delivering something "as a service" means, but the central theme of the definition is that service delivery is when a provider enables a contractual relationship between itself and a consumer to deliver some capability or bit of work. This is defined not by the physical implementation but in terms of results. Because of that, the discontinuity created by this phenomenon is not one of technology implementation but one of relationships and interfaces. The relationship between the consumer and the provider of an IT service establishes the parameters of the service to be delivered. So the decision of which service to buy or how to pay for it is made based on price, performance, quality of service, trust, disaster recovery, security guarantees or even reputation, but not generally on implementation.

The concept of a cloud, just like the concept of the Internet, will benefit from the realization that only one public cloud exists. Customers external to a company are served through the public incarnation of this cloud. The idea of multiple public clouds implies that a consumer must select which cloud he or she will use. And, given that the technologies for delivering IT services pervasively must be the same for everyone in a public context, the idea that one cloud will differ from another loses steam. Services are just services ultimately, and while a customer might choose to use Google services vs. Zoho services, those services are accessed through the same mechanisms regardless of vendor. By 2011, one public cloud and many private clouds will be the commonly accepted view of where cloud services are accessible.

Just as in the early days of the Internet, the cloud definition is best-suited to include a public cloud (external, like the Internet) and private clouds (internal, like intranets). Private clouds will be used by companies that do not want to have their IT-related services available to external customers but that do want to leverage the delivery and acquisition model the cloud enables.

This has several implications:

1. The cloud is a mechanism that supports alternative delivery and acquisition models (ADMs) of multiple types (for example, grid, utility and SaaS) in that anything and everything can be offered as a service.

2. The cloud is more than just SaaS in that everything as a service (XaaS) would be a more appropriate appellation. Through the cloud, we can get services that are primarily hardware based where the software is an integral part of the delivery, not a specific value proposition in itself. Storage as a service is an example of this. SaaS is an ADM

and can be delivered through the cloud just as any other service can. To say that SaaS and the cloud were the same would be a limited view.

3. The Internet and potentially the Web are necessary for the existence today of the cloud but are not solely definitive of it. There must be a deliberate delivery of service to constitute cloud computing. (In the future, some other globally distributed network may supplant the Internet in this capacity to support cloud computing.)

4. Internal clouds can exist using Internet/Web technologies but must also include the intentional ADM of XaaS to internal customers and private partners. This is what distinguished these internal clouds from the one public cloud.

Users are already changing their buying behaviors. Although it is unlikely that they will completely abandon on-premises models or that they will buy complex mission-critical processes as services through the cloud, there will be a movement to consume services in a more cost-effective way.

Cloud is indeed an alternative form of delivery for IT-enabled services. Just like other forms of delivery before it, including Web hosting, ASPs, managed service providers (MSPs) and SaaS, it is critical that IT operations not just assess the ability of the service provider to meet business requirements but also assess the operational requirement service-level agreements (SLAs), pricing, and terms and conditions for requirements and risks. The most important thing is for expectations to be set and known upfront by all parties. Contracts should be no longer than a year to ensure that service levels and additional expectations can be built into the contract or re-evaluated as the values of the services change over time for the customers. Cloud services should not be bought by business users unbeknownst to IT if the services are mission-critical or have the potential to be mission-critical in a 12-month time frame because lack of oversight and considerations of IT operational requirements could pose unacceptable risks and costs to the business. Moreover, IT should be aware of business usage of cloud-computing services even for noncritical services to be sure that IT is meeting its customer requirements over time and not missing business requirements that result in a higher amount of "shadow IT" and increased overall business cost and risk (see Note 3).

7.0 Cloud and Related Concepts

There are many concepts that seem similar to cloud computing but are, in fact, complementary to it. If we examine a few of them, we can establish how these concepts build on one another rather than just being alternative names for the same phenomenon.

SaaS: With cloud computing gaining substantial buzz in the IT industry, some vendors have transformed their positioning from SaaS providers to cloud-computing providers without changing one element of their offerings. When comparing the two definitions, it becomes clear that cloud computing is a necessary underpinning for a provider to deliver a scalable SaaS offering to the market (see "Tutorial for Understanding the Relationship Between Cloud Computing and SaaS"). Massively scalable IT-related functions are a concept that goes directly to a provider's ability to deliver a single set of common code and data definitions, which are consumed by all contracted customers. This is fundamental for any SaaS application.

Web Platforms/Cloud Platforms: A Web platform provides programmatic access to Web-based capabilities that act as a foundation to create a composite application and/or business process. Capabilities (that is, services) may include infrastructure, applications, content, application components, business processes or ecosystem management. The programming model is built on Web-oriented architecture and principles. This is complementary to cloud computing in that the Web platform (also called a cloud platform) is used to enable the provisioning of services through Web-based architecture to the cloud. There is not one Web platform but multiple Web platforms.

Utility: Wikipedia defines utility computing as: "The packaging of computing resources, such as computation and storage, as a metered service similar to a physical public utility (such as electricity, water, natural gas, or telephone network). This system has the advantage of a low or no initial cost to acquire hardware; instead, computational resources are essentially rented. Customers with very large computations or a sudden peak in demand can also avoid the delays that would result from physically acquiring and assembling a large number of computers."

Grid: A grid is a collection of resources (owned by multiple organizations) that is coordinated to enable the resources to solve a common problem. They may be divisions of one company or distinct legal entities. Three conceptual forms of grids are computing grids (the most common today), data grids and collaboration grids.

Real-Time Infrastructure (RTI): An RTI is an IT infrastructure shared across customers, business units or applications where business policies and SLAs drive its dynamic and automatic optimization to reduce costs while increasing agility and quality of service. RTI includes the concepts of service orientation and shared resources for multiple customers. Add access through Internet technologies and an RTI is a cloud-computing service for infrastructure. Essentially, RTI, in one form or another, will be the "engine room" for cloud-computing providers.

8.0 The Demand for Choice Leads to XaaS

Pressures on businesses have long been to increase business effectiveness while either controlling or reducing cost more effectively. However, in IT departments and the IT industry, effectiveness has often been limited by the acquisition/delivery models for how those services reach the people who need to use them. That would inevitably lead to a crisis of budget versus results. This has amounted to tremendous pressure being applied to IT budgets, so many companies have looked to off-premises models to shift the expense of maintaining IT systems away from their site and toward external service providers, outsourcers and SaaS vendors. However, it would be shortsighted to focus only on shifting IT services away from on-premises implementations. Instead, the ability to choose how an IT service is delivered from on-premises data centers to private clouds, off-premises utilities or even the public cloud is most critical.

Choice is enabled through the flexibility inherent in a cloud model. The further away from the implementation details a user can get, the more freedom the provider of the IT service has in how it is delivered. This allows the provider to tailor costs, systems and quality to suit the customers and the need for an effective business model. This is the essence of service delivery. Once this has been established, the next step for many companies is to figure out how many IT capabilities can be delivered as a service. Ultimately, everything potentially becomes a service.

9.0 Common-Sense Definition

A simple common-sense definition of cloud computing could be said to be "anything but my assets." Just as in the era of network computing a network computer was "anything but a PC" and, in the client/server era, client/server was "anything but a mainframe," a common-sense definition of cloud computing as a movement of assets from bought and implemented on-site to hosted, owned and provided by someone else has merit.

10.0 Conclusion

Cloud computing is an evolving concept and will take many years to fully mature. Our definition establishes the Internet as part of that evolution and allows for the proliferation of IT services in a globally pervasive way. The relationship between buyer and seller is thus redefined, and the way in which individuals or IT departments gain value from technology is altered. Customers seeking to shift costs from on-premises implementations to delivery of services from providers who offer

SLAs and performance guarantees will benefit from a cautious approach to a cloud model. Vendors selling IT technologies and services will begin to shift to cloud platforms and cloud-based services as primary marketing messages in the next few years or run the risk of being marginalized as their products sell less to users who are buying more from cloud-computing service providers. The cloud-computing model is not just the next generation of the Internet. When organizations cross the threshold between "the Internet as a communications channel" and "the deliberate delivery of service over the Internet," everything begins to change. The Internet alone does not do that.

Note 1
The Role of Virtualization

Virtualization has been a recent hot trend in the industry, but for most organizations, virtualization has been seen as a new consolidation solution. Virtualization is much more than that. Virtualization, in its many forms, has enabled more efficient use of shared resources, increasing the economies of scale of computing. This could be seen as another form of consolidation. But virtualization also deconsolidates. The role of virtualization is to decouple two IT components that were previously integrated together. Ideally, the new interface created here can be service-oriented, allowing independence and portability. Inside of an IT organization, this could allow, for example, management and deployment of hardware independently of operating systems and applications. Taken to its extreme, this service-oriented interface enables cloud computing. Amazon's EC2, for example, enables customers to contract for raw processing power to run customer virtual machines.

Note 2
The Role of the Internet

As with all definitions, it is important to recognize that as technology options change, the specifics of the definition can also change. The Gartner definition specifically cites the Internet, whereas the Wikipedia definition (for example) does not. Currently, we feel that the Internet is the single most pervasive and globally visible network available. These characteristics are necessary to delivering IT services to a generic cloud and to a simple definition. However, as time passes, other networking options may become popular and then the specific connection to the Internet will be only one part of the definition. In fact, within companies, the use of private networks means that a cloud-style environment can be delivered without ever using Internet technologies.

Note 3
Operational Risks of Cloud Computing

Organizations that are evaluating the benefits of cloud-based services must also identify associated operational and security risks to develop compensating controls or to define use cases that contain an acceptable level of risk. Since cloud-computing environments are externally provided and shared, organizations need to evaluate risk in areas such as data integrity and privacy and need to understand issues in areas such as e-discovery, compliance and audit reporting. Compensating controls may take the form of SLAs with the provider, periodic assessment of provider capabilities, or new audit or monitoring functions.

This research is part of a set of related research pieces. See "Cloud Computing Confusion Leads to Opportunity" for an overview.

REGIONAL HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
U.S.A.
+1 203 964 0096

European Headquarters
Tamesis
The Glanty
Egham
Surrey, TW20 9AW
UNITED KINGDOM
+44 1784 431611

Asia/Pacific Headquarters
Gartner Australasia Pty. Ltd.
Level 9, 141 Walker Street
North Sydney
New South Wales 2060
AUSTRALIA
+61 2 9459 4600

Japan Headquarters
Gartner Japan Ltd.
Aobadai Hills, 6F
7-7, Aobadai, 4-chome
Meguro-ku, Tokyo 153-0042
JAPAN
+81 3 3481 3670

Latin America Headquarters
Gartner do Brazil
Av. das Nações Unidas, 12551
9° andar—World Trade Center
04578-903—São Paulo SP
BRAZIL
+55 11 3443 1509

Gartner