Cloud Computing

Daniel Watrous

Management Information Systems

Northwest Nazarene University

Cloud Computing

Definition of Cloud Computing

Cloud computing may have different meanings depending on perspective. The term cloud computing is frequently used by businesses in consumer directed communications. This can lead many consumers of online services to refer to the cloud as a collection of services with specific functions, like email, social networking and even monitoring of markets. A common term for these services is Software as a Service (SaaS) (Armbrust et al., 2010).

From a software developer perspective, the term cloud computing is more likely to be associated with a particular approach to deploying a software application. Cloud in this case refers to virtual compute resources of various types. Various levels of abstraction are available when discussing cloud computing. On one end is Infrastructure as a Service (IaaS), which is comprised of raw compute resources. On the other end is Platform as a Service (PaaS), which is made up of a proprietary stack of integrated services and components (Armbrust et al., 2010).

Infrastructure as a Service

The most flexible offering in the cloud space is Infrastructure as a Service. This generally refers to a scenario where the service provides compute processors, memory and persistent storage. The consumer, who is most often a developer in this case, is free to install any operating system, in any configuration and run any software that he likes. An IaaS consumer is also responsible for security, scaling, maintaining and otherwise caring for the computing resources.

Platform as a Service

Somewhat less flexible are Platform as a Service offerings. In this scenario, a service provides a type of compute container. Available memory, storage and compute processors are often concealed behind an abstraction layer. This allows the platform to take care of details like adding additional capacity, storage and even distributing capacity between datacenters for reliability. This is generally a managed platform, which means that securing and patching the operating system and other stack elements are taken care of by the platform provider.

Software as a Service

Software as a Service generally refers to cloud computing from the perspective of the end user. The developer of the software that is consumed as a service is free to use a PaaS, an IaaS or any other method for deploying his software. From the consumer's perspective, however, the service is simply available and responsive. There is no requirement imposed on the end user to have specialized hardware or to physically or geographically accommodate the service. From the end user perspective, the service is always available from anywhere.

Captial vs. Operating Expense

From the business operations perspective, there is a key financial difference between a traditional approach to provisioning computing resources and the cloud approaches mentioned above. Cloud consumers pay as they go and for only the resources they actually use, much like a utility such as water or electricity. This is represented on the finances of the company as an operating expense. When compute resources are procured directly and hosted in a physical facility, that must be depreciated over time. Accurate assessment of capacity needs is less important in the cloud, but very important when procuring and managing fixed compute resources.

A Brief History of Computing

The evolution of computing shares some similarities with electricity and other modern utilities. For example, when electric production technology was new, there were no transmission lines or committed energy sources that could be converted into electricity, like coal. Decades passed before the infrastructure became available to bring electricity to every home. Even more time passed before the use of electronic devices became ubiquitous, such as electric motors in manufacturing equipment. Existing infrastructure investment within enterprises slowed the adoption of the newer technology. However, over time, the disruptive nature of electricity being carried to each home and business was tremendous. As the number of consumers increased, so did production. Economies of scale in electrical production coupled with broader adoption brought prices down (Manyika et al., 2013).

Computing has evolved in similar ways. Many years were required to standardize hardware and software interfaces. As the process of sharing data between different systems improved, the perceived utility of using computers to communicate increased. There have been several shifts from centralized to distributed computing. The latest trend toward cloud computing is a type of centralization that makes cloud computing resources available to any consumer.

Advancements in technology, such as the decrease in feature size for silicon production, have increased overall capacity of computing resources, while decreasing their physical size. Equally important is the development of ever faster network infrastructure which connects computing devices. The ability to move large amounts of data has made it possible to centralize computing resources, since the data can be quickly transferred to them.

Economics of Compute Capacity

Another similarity between cloud computing and other utility models is that the amount of compute capacity available is finite. It’s true that from the point of view of the consumer, there is infinite capacity, from the cloud provider’s perspective, there is limited capital available. Resource mismatches like this have been seen with other utilities (Buyya, Yeo, & Venugopal, 2008). For example, when a consumer powers on an electrical device, it is expected to work. In other words, the consumer is not generally concerned with whether or not there is any electricity available, like he might think about milk in his refrigerator or gasoline in his automobile. Still, limited availability of electricity has resulted in brown-outs in some large metropolitan areas when demand exceeded available production capacity. Limited water supplies can have a similar effect on agricultural production. Providers of cloud services must then find the most profitable and sustainable way to provide their service based on the inherently finite nature of it. Two approaches to pricing of cloud are common. The first is to treat cloud resources as a commodity. The second is allow market demand to set the prices in a form of an auction (Zhang, Zhu, & Boutaba, 2011).

Resource Utilization Models

Commodity. Under the commodity model of pricing, the price is set in a way similar to how gasoline is priced. For a given resource type, there is a fixed price for all consumers which may fluctuate over time to respond to changes in cost to provide the service and demand for the service. Changes in price can be slow relative to the market. Price points are also more susceptible to competitive pressures. When the resource offering is largely normalized so that one cloud offering functions the same as another, opportunities to differentiate are less common, which results in thinner margins as competition increases (Zhang et al., 2011).

Auction. Another approach is to sell services based on a type of auction. In this scenario, consumers bid on available compute resources. Cloud providers allocate resources to the highest bidder. Naturally, this pricing model more closely tracks with consumer demand and can bring a greater profit when demand peaks. From a cloud consumer perspective, this model presents challenges when it comes to budgeting for anticipated demand.

To date, the commodity pricing model has been the primary method used by the majority of cloud providers. As the market matures and pricing fluctuates more frequently, either in an auction scenario or based on related commodity pricing, such as energy, it's possible that cloud compute resources could be purchased and traded using mechanisms similar to how futures function in commodity markets.

Energy

The cost of energy to power data centers has been estimated at 20% of the total cost of running a data center (Zhang et al., 2011). Naturally this significant connection to world energy markets has the potential to impact profits and availability. It also raises questions about energy surcharges during times when energy is scarce. For large consumers of cloud compute resources, energy prices are likely to become a critical factor in their cloud strategy.

Capacity and Innovation

Technological innovations that decrease space and energy requirements have the potential to make a big difference to the bottom line of both providers and consumers of cloud computing resources. One such innovation introduced by Hewlett Packard is their new moonshot system (Packard, 2014). The moonshot system transitions away from traditional server technology and instead starts with a chassis that can receive server cartridges of various types, even custom cartridges. Each cartridge can have general or specialized functions. This makes it possible for a single hardware chassis to accommodate both IaaS and PaaS functions, even when the platform service becomes very specialized. HP claims that this new approach decreases space requirements by 80% and energy consumption by up to 89%.

In order for cloud computing to be profitable for both the provider and the consumer, it’s necessary to look beyond pricing to global energy markets and technological innovations that will reduce reliance on energy and increase utility of existing data center space.

Security and Privacy

An increasing number of cloud applications deal in some way with sensitive consumer information. In some cases this consumer information is further protected by government regulation and statutes. The result is a complex legal and technical environment that needs to balance capacity and development resources against consumer and regulatory expectations for privacy and security. At least one author has observed that achieving 100% secure compute resources is virtually impossible (Kim, 2009). Consumer expectations, however, don’t always reflect the reality in terms of complexity and time required to maintain secure environments.

Some surveys have identified confidentiality, authentication and authorization, integrity, access control, etc. as some of the major aspects when considering security (Dubey & Verma, 2013). These same surveys have gone further to identify aspects of trust based on consumer expectations. Security and privacy go beyond the technology and require careful consideration of human perceptions and expectations. It is likely that legislation will emerge to establish some of these guidelines and norms as the cloud becomes a more prevalent resource for businesses that handle sensitive user information.

Big Data

In keeping with the adage that knowledge is power, many modern businesses and research initiatives are finding that access to large amounts of digital data have the potential to provide key competitive advantage and further research efforts. As the amount of available digital data increases, so do the computing requirements to process it. This is frequently referred to as big data.

Some have suggested that data becomes big data when the amount of data that is being processed exceeds the capacity of a single computer. In some cases, hundreds or thousands of computers must be interconnected in order to hold and process large amounts of data. Traditional methods for storing and analyzing this data have also come up short.

Google's Problem: Web Scale

Google is a good illustration of the big data example. As the Internet or World Wide Web has grown in popularity, various means have been devised to provide access to the various web resources. Some initial efforts were an extension of what the Yellow Pages did for businesses. Paper books were compiled and distributed that contained links to thousands of websites. Unlike the Yellow Pages, which documented address and contact information for physical businesses, the Internet provided a type of virtual property that was much easier to setup and take down and change. As a result, the useful life span of a printed directory was extremely short.

Google and other companies decided that search was a preferred way to locate resources on the Internet. However, they went a step further and took on the challenge of creating an index that included every page on the Internet. As an example, the Yellow Pages may have a single listing in its published directory for a given company, regardless of how many physical locations, employees or telephone numbers that company had. Google's approach was to index every page, which meant that a company website with 100,000 pages would have 100,000 entries in Google's index. As Google began to crawl the web, hundreds of millions of pages of content were encountered. Current estimates put the total number of indexed web pages over four billion (Size, 2014). Not only did this volume of data exceed what could be stored on a single machine, but the available data storage tools, such as relational databases, were insufficient to facilitate access to the indexed data after it had been processed.

Google's Solution

To solve this, Google pioneered three technologies that have become a backbone of most large scale cloud computing today. These involve storage, computation and access. A principle design characteristic that Google has followed in developing all of it's technological foundation is to build systems in a way that can leverage commodity hardware. In other words, they wanted to use inexpensive servers rather than buying expensive high end systems.

Distributed File System. The first problem Google needed to solve was how to store all the data they were collecting as they crawled the Internet. Since no single computer was large enough, it became necessary to link several servers together and store a portion of the information on each. Reliability was key to the success of the storage solution so that if one server failed, no data was lost and processing work was not interrupted. The result was the Google Distributed File system, which runs today on thousands of computers containing enormous amounts of data (Ghemawat, Gobioff, & Leung, 2003).

From a business perspective, this solution provided Google with the option to grow storage capacity over time. Because they avoided a complex vendor specific solution, they were free to purchase new servers from any source, or even build their own. This flexibility enabled them to meet growth demands while maintaining a predictable budget.

Map Reduce. With data now available, Google required a way to process it all and derive some meaning from it. In order to accomplish this, Google embraced the use of a programming paradigm known as Map Reduce. The concept is simple to map data in certain ways to extract elements that may be meaningful. As this mapping process is performed across many different pieces of data, the results are reduced down to a different format. An example of this could be to map each user visible word from a web page and reduce them in to count buckets. This produces an alternative view into the document which shows how many times each user visible word appears. In order to accomplish this, Google created a distributed Map Reduce system (Dean & Ghemawat, 2004).

There are some key business and technical benefits that come from this approach. One is that the complexities of processing large amounts of data remain scoped as smaller more manageable chunks. A single page from the Internet may be mapped and reduced into dozens or hundreds of different formats. Some map reduce jobs may even operate on intermediary formats. With this approach they are able to keep each processing step isolated from the others, which results in a more resilient system. The nature of map reduce work also produces chunks small enough to run on a single machine, which makes it possible to easily distribute the work load to as many machines as necessary to accomplish the desired processing in an acceptable time.

Big Table. The final component of Google's primary architecture is a system that enables the storage of indexed data for fast retrieval. Traditional databases suffered from some constraints that made them unsuitable for this task. They did not cluster well, which impacted scalability. They struggled to manage indexes on the scale required by Google. They imposed structure requirements that were not consistent with the open nature of Google's objective. In other words, web pages were not standard. It wasn't possible to extract a common set of data from all web pages and store them in a traditional database table. What Google required as an unstructured environment where each element in the index could take on a form that more closely resembled the page on the Internet that was represented by it. As a result they created a flexible, distributable datastore technology which they call Big Table (Chang et al., 2006).

Some key business advantages to this technology included the absence of licensing costs for external database technologies. As with other parts of their technology stack, Big Table would easily distribute across inexpensive servers and requests for data could be balanced across available systems. Finally, Google was empowered to create a close relationship between data in the index and the origin web pages that to which that data corresponded.

As public cloud offerings emerged, Google provided this internal infrastructure was a platform to developers (a PaaS). Other groups and companies took Google's published research and created open source systems, such as Hadoop, which provided a similar stack. Today, it's possible to buy infrastructure (IaaS) and setup a similar large scale processing system, or even buy large scale processing as a service (SaaS).

Conclusion

Cloud computing is a shift from the traditional approach of buying physical computing resources. Compute capacity is made available on demand and consumers pay for only what they use. From the user perspective, this compute capacity is infinite and available on demand, much like utilities, such as water and electricity. From the cloud provider perspective, capacity is finite and there are various external considerations which factor into the cost to provide the service, including energy costs. The biggest benefit coming to many businesses is the ability to process and analyze increasingly large and complex sets of data. The companies that manage their knowledge assets well and can derive meaningful conclusions from that data have a key competitive advantage. As shown in the case of Google, a cloud style approach to fluid and flexible compute capacity, even when the servers are managed internally, brings scale and manageability benefits to large enterprises over older models where large, complex systems were purchased from vendors. Alongside all the upsides to cloud computing there are potential risks related to security and privacy as more businesses purchase compute and storage from external vendors. Questions of ownership of data and access control will be critical to businesses and governmental agencies as they takes steps to leverage the power of the cloud in the future.

References

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., et al. (2010). A view of cloud computing. *Communications of the ACM*.

Buyya, R., Yeo, C. S., & Venugopal, S. (2008). Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. *High Performance Computing and Communications*.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., et al. (2006). Bigtable: A distributed storage system for structured data. *OSDI*.

Dean, J., & Ghemawat, S. (2004). Mapreduce: Simplied data processing on large clusters. *OSDI*.

Dubey, S. K., & Verma, A. (2013). Security and privacy in cloud computing: A survey. *Semantics Knowledge and Grid*, *2*(6), 123-125.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The google file system. *SOSP*.

Kim, W. (2009). Cloud computing: Today and tomorrow. *Journal of Object Technology*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity* (Tech. Rep.). McKinsey Global Institute.

Manyika, J., Chui, M., Bughin, J., Dobbs, R., Bisson, P., & Marrs, A. (2013). *Disruptive technologies: Advances that will transform life, business, and the global economy* (Tech. Rep.). McKinsey Global Institute.

Packard, H. (2014). *Hp moonshot system.* Available from `http://www.hp.com/go/moonshot`

Size, W. W. W. (2014). World wide web size. Available from `http://www.worldwidewebsize.com/`

Zhang, Q., Zhu, Q., & Boutaba, R. (2011). Dynamic resource allocation for spot markets in cloud computing environments. *Utility and Cloud Computing*, 178-185.