acmqueue Cloud Computing: An Overview

A summary of important cloud-computing issues distilled from ACM CTO Roundtables

Probably more than anything we've seen in IT since the invention of timesharing or the introduction of the PC, cloud computing represents a paradigm shift in the delivery architecture of information services. This overview presents some of the key topics discussed during the ACM Cloud Computing and Virtualization CTO Roundtables of 2008. While not intended to replace the in-depth roundtable discussions, the overview summarizes the fundamental issues generally agreed upon by the panels and should help readers to assess the applicability of cloud computing to their application areas.—*Mache Creeger, Moderator, CTO Roundtable Series*

WHAT IS CLOUD COMPUTING?

Google, Yahoo, Amazon, and others have built large, purpose-built architectures to support their applications and taught the rest of the world how to do massively scalable architectures to support compute, storage, and application services.

Cloud computing is about moving services, computation and/or data—for cost and business advantage—off-site to an internal or external, location-transparent, centralized facility or contractor. By making data available in the cloud, it can be more easily and ubiquitously accessed, often at much lower cost, increasing its value by enabling opportunities for enhanced collaboration, integration, and analysis on a shared common platform.

Cloud computing can be divided into three areas:

- SaaS (software-as-a-service). WAN-enabled application services (e.g., Google Apps, Salesforce.com, WebEx)
- PaaS (platform-as-a-service). Foundational elements to develop new applications (e.g., Coghead, Google Application Engine)
- IaaS (infrastructure-as-a-service). Providing computational and storage infrastructure in a centralized, location-transparent service (e.g., Amazon)

Enabling technologies. The following precursor technologies enabled cloud computing as it exists today:

- SaaS
- Inexpensive storage
- Inexpensive and plentiful client CPU bandwidth to support significant client computation
- Sophisticated client algorithms, including HTML, CSS, AJAX, REST
- Client broadband
- SOA (service-oriented architectures)
- Large infrastructure implementations from Google, Yahoo, Amazon, and others that provided real-world, massively scalable, distributed computing
- Commercial virtualization

CAPEX VS. OPEX TRADEOFF

In the past, developing an application service required a large CapEx (capital expense) to build infrastructure for peak service demand before deployment. The risk of a service's success combined with the operational requirement of a large CapEx investment severely restricted funding. Cloud computing addresses this problem by allowing expenses to track closely with resource use, thus following income rather than having to purchase for peak capacity before income is realized. Running application services on a cloud platform accomplishes this in three fundamental ways:

It moves CapEx to OpEx (operational expense), closely correlating expenses with resource use.

It allows service owners to eliminate significant system-administration head count by avoiding the need for internally purchased servers.

It smooths the path to service scaling by not requiring the CapEx-intensive architectural changes needed to scale up service capacity in the event of service success.

Because the cost of deploying new services is much lower and expenses track real usage, businesses can develop and deploy more services without fear of writing off huge capital investments for dedicated infrastructure that may never be needed. While start-ups are more focused on cost, enterprises are equally focused on flexibility to make required service changes and achieve maximum agility. Some Silicon Valley start-ups are able to go completely without infrastructure, instead using outside services for e-mail, Internet, phone, and source control. This allows the start-up to focus all of its resources on its core differentiating efforts.

BENEFITS

Large-scale multitenancy achieves significant economic advantage. Sharing the resources and purchasing power of very large-scale multitenant data centers provides an economic advantage. As an example, a major engineering services company's current internal cost to provide a gigabyte of managed storage is $3.75 per month, while Amazon charges 10 to 15 cents per month. Initially, ISP charges at the company were $3,500 per megabyte per month. After examining the cost structure of companies such as YouTube, the engineering services company assumed that YouTube's costs were in the teens. Taking advantage of network peering arrangements and consolidating the company's interfaces to a place close to the ISP's POP (point-of-presence) have brought costs down to YouTube levels.

Broad use of virtualization has also significantly reduced the company's data-center CapEx. Prior to virtualization, server utilization was between 2 and 3 percent and total data-center floor space was around 35,000 square feet. With virtualization in widespread use, server utilization is up to 80 percent and server consolidation has shrunk the square footage of the data center to 1,000 square feet.

As more cloud service vendors become available, computing and storage will become a true commodity with fine-grained pricing models, complete with arbitrage opportunities, similar to other commodities such as natural gas and electric power. Under a cloud model, pricing is based on direct storage use and/or the number of CPU cycles expended. It frees service owners from coarser-grained pricing models based on the commitment of whole servers or storage units.

Transforming high fixed-capital costs to low variable expenses. Setting up an internal cloud within a company provides an efficient service platform while placing a limit on internal capital expenditures for IT infrastructure. An external cloud service provider can supply overflow service capacity when demand increases beyond internal capacity.

Previously, companies had to operate servers for projects even though they might never be invoked—such as servicing warranties. A cost of $800 to $1,000 per month is not unreasonable to have a server idle on the data-center floor. By moving these projects to an external IaaS vendor, those functions can be placed in the cloud and the service run only when required, at pennies per CPU hour. In this way, companies can transform what was a high fixed cost into a very low variable one.

Flexibility. For large enterprises, the ease of deploying a full service set without having to set up base infrastructure to support it can be even more attractive than cost savings. Bechtel must set up new engineering centers with very little notice worldwide. Using internal cloud IT resources, Bechtel can now set up these centers to be fully functional within 30 days.

Smoother scalability path. For application architectures that easily scale with added hardware and infrastructure resources, cloud computing allows for many single services to scale over a wide demand range. Animoto's service started with 50 instances on Amazon. Because of its popularity, it was able to meet soaring demand and scale to 3,500 instances within three days.

It is not a given, however, that all application architectures scale that easily. Databases are a good example of hard-to-scale applications—hence, the widespread use of programs such as Amazon's SimpleDB.

Self-service IT infrastructure. Cloud-computing service models are often self-service, even in internal models. Previously, you had to partner with IT to develop your application, provide an execution platform, and run it. Now, much like Amazon, IT departments define use policies for automated platform and infrastructure services with line-of-business-owners developing applications on their own to meet those requirements.

Severely reduced disaster recovery cost. Most SMBs (small- to medium-size businesses) make no investment in DR (disaster recovery). By enabling VMs (virtual machines) to be sent to the cloud for access only when needed, virtualization becomes a cost-effective DR mechanism. Typical DR costs are 2N (twice the cost of the infrastructure). With a cloud-based model, true DR is available for 1.05N, a significant savings. Additionally, because external cloud service providers replicate their data, even the loss of one or two data centers will not result in lost data.

Common application platform enables third parties to add value. While telcos are moving to cloud platforms for cost effectiveness, they also see opportunities resulting from a common application platform. By allowing third parties to use their platforms, telcos can deploy services that either extend the telco's services or operate independently.

Increased automation. Amazon sees automation as a significant benefit of a cloud services model. Moving into the cloud requires a much higher level of automation because moving off-premises eliminates on-call system administrators.

Release from ABI and operating-system dependencies and restrictions. Amazon also sees cloud computing as a way of releasing data centers from the need to support the ABI (application binary interface) and operating-system requirements of key applications. With EC2, Amazon provides five popular VMs to choose from: three flavors of Linux, OpenSolaris, and Windows Server. Its only concern is effectively running the VMs; it does not have to be involved with the VM's internal operations.

MapReduce enables new services. Although not the most cost-efficient way of providing data-warehouse functionality, MapReduce's use of a large parallel-processing resource has enabled a number of companies to provide cloud-based data-warehousing services. This frees customers from having to invest in large specialty hardware purchases for small service requirements. MapReduce is expected to enable additional service types that were once limited to dedicated hardware.

USE CASES

- An international financial exchange paid for the development of a large service. It hosted data in the cloud and ran the application on the client's desktop. All operations were on a pay-as-you-go basis. This is an example of a very low initial investment required to make a commercial service operational.
- Shazam is a start-up company whose service executes on the Apple iPod. It samples songs being played on the radio, matches the songs to a library in the cloud, and returns a link to purchase that song on the iPod. It is an example of a smart device coupled with cloud-based computation and storage.
- Animoto, hosted on Amazon, was able to track demand of its service and scale up from 50 instances to 3,500 instances over a three-day period.
- A national newspaper wanted to place scanned images covering a 60-year period online. After being repeatedly turned down by the CIO for the use of six servers, the newspaper moved four terabytes into S3, ran all the software over a weekend on EC2 for $25, and launched its product.
- A major international auto-race organizer supports special race Web sites that provide live streaming video and realtime technical information. Previously, it would retain an ISP, acquire massive server power, and hire 500 engineers to baby-sit the servers at the ISP to institute server failover manually. When it moved to EC2, the savings in server rental were not that big, but it did realize orders of magnitude in personnel cost savings.
- Mogulus streams 120,000 live TV channels over the Internet and owns no hardware except for the laptops it uses. It did all of the election coverage for most of the large media sites. Its CEO states that he could not be in business without IaaS.

DISTANCE IMPLICATIONS BETWEEN COMPUTATION AND DATA

How you deal with the distance between computation and data depends heavily on application requirements. If you need to minimize expensive bandwidth, then you should find a way to keep the two in proximity. In cases where bandwidth is expensive and the distances cannot be shortened, it may make sense to download an extract of the data to work on it locally. Longer term, it would be best if developers could write the application in such a way that it would dynamically adjust its data-access mechanisms in response to the operational context (bandwidth cost, bandwidth latency, security, legal data location requirements, etc.).

DATA SECURITY

A common concern about using an external cloud service provider is that it will make data less secure. Because of the wide quality range of corporate IT security, trusting information assets to a recognized cloud service provider could very easily increase the security of those assets. Given that many corporate data centers struggle to fund, architect, and staff a complete security architecture, and that cloud service providers provide IT infrastructure as their primary business and competence, clouds could possibly increase security for the majority of their users. Moreover, 75 to 80 percent of intellectual property breaches are a result of attacks made inside the company, which would not impact a decision to use clouds one way or the other.

ADVICE

- IT should establish a revenue metric to show the cost effectiveness of its service-delivery infrastructure. Bechtel used the inverse of the hours it took to complete its corporate projects. By increasing infrastructure utilization, it lowered its cost per unit of output by 55 percent—increasing capacity and overall satisfaction.
- Although most CIOs and CTOs are interested in seeing if a cloud-based service model is a more efficient IT architecture, initial adoption is usually bottom-up and based on pragmatic business needs. Often the CIO is the last person in the adoption chain.
- In deploying services for your company, make an effort first to buy those services before building them on your own. As in all successful businesses, do not get caught up in building and supporting infrastructure that is not core to your business.
- If an application is performed trillions of times per day, anything with even a remote probability of failure is a certainty. Application developers must be trained to accept failure as inevitable and design for it.
- Cloud computing represents a shift in power in IT away from those who control capital resources to the users and developers who employ self service to provision their own applications.

  Often IT people are not rational and will resist losing control of power and budget. This can be a significant roadblock in converting to a cloud-based service architecture. Change management can often be the majority of the effort in converting to a cloud-oriented service architecture (e.g., at Bechtel it has been around 80 percent of the effort) as it takes time for people to move outside their career and risk comfort zones. A cloud service model places traditional IT skills at risk, and those people need to be transitioned from managing the physical infrastructure (such as storage or processing) to managing IT policies and service-level guarantees.
- When the load varies widely, a cloud-computing service model excels. For services that impose well-defined loads, it usually is more cost effective to make the capital investment for an internal platform (e.g., running your own Microsoft Exchange server on Amazon is not a good idea).
- By restructuring its Internet services to be similar to YouTube and placing interfaces closer to ISP POPs, Bechtel was able over several years to decrease latency by 50 percent and reduce Internet charges by several orders of magnitude. Bechtel currently pays $10 to $20 per megabyte per month.

UNANSWERED QUESTIONS

- Are data-ingestion services able to take physical delivery of a large amount of media for transfer to the cloud?
- Are appropriate data-location choices provided to the application so that users can comply with applicable law? Depending on the laws in force, data-location compliance can be quite complex and require sophisticated abstractions.

LOVE IT, HATE IT? LET US KNOW

feedback@queue.acm.org