The Business of Clouds

By Guy Rosen

At the turn of the 20th century, companies stopped generating their own power and plugged into the electricity grid. In his now famous book *The Big Switch*, Nick Carr analogizes those events of a hundred years ago to the tectonic shift taking place in the technology industry today.

Just as with electricity, businesses are now turning to on-demand, mass-produced computing power as a viable alternative to maintaining their IT infrastructure in-house.

In this article, we'll try to hunt down some hard data in order to shed some light on the magnitude of this shift. We'll also take a look at why it is all so significant, examining what the cloud means for businesses and how it is fueling a new generation of tech startups.

What is the Cloud?

While the exact definition of cloud computing is subject to heated debate, we can use one of the more accepted definitions from NIST, which lays out five essential characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service. Of particular interest to us are the three service models NIST describes:

Infrastructure as a service (*IaaS*) displaces in-house servers, storage and networks by providing those resources on-demand. Instead of purchasing a server, you can now provision one within minutes and discard it when you're finished, often paying by the hour only for what you actually used. (See also "Elasticity in the Cloud," page 3, for more.)

Platform as a service (*PaaS*) adds a layer to the infrastructure, providing a platform upon which applications can be written and deployed. These platforms aim to focus the programmers on the business logic, freeing them from the worries of the physical (or virtual) infrastructure.

Software as a service (*SaaS*) refers to applications running on cloud infrastructures, typically delivered to the end user via a web browser. The end-user need not understand a thing about the underlying infrastructure or platform! This model has uprooted traditional software, which was delivered on CDs and required installation, possibly even requiring purchase of a server to run on.

The Hype

Research outfit Gartner describes cloud computing as the most hyped subject in IT today. IDC, another leading firm, estimated that cloud IT spending was at $16 billion in 2008 and would reach $42 billion by 2012. Using Google Trends, we can find more evidence of the growing interest in cloud computing by analyzing search volume for the term *cloud computing.*

It's extraordinary that a term that was virtually unheard of as recently as 2006 is now one of the hottest areas of the tech industry.

The Reality

The big question is whether cloud computing is just a lot of hot air. To add to the mystery, hard data is exceedingly hard to come by. Amazon, the largest player in the IaaS space, is deliberately vague. In its financial reports, the revenues from its IaaS service are rolled into the "other" category.

In an attempt to shed some light on de facto adoption of cloud infrastructure, I conducted some research during 2009 that tries to answer these questions.

The first study, a monthly report titled "State of the Cloud" (see **www.jackofallclouds.com/category/state-of-the-cloud/**), aims to estimate the adoption of cloud infrastructure among public web sites. It's relatively straightforward to determine whether a given site is running on cloud infrastructure, and if so, from which provider, by examining the site's DNS records as well as the ownership of its IP. Now all we need is a data set that will provide a large number of sites to run this test on. For this, we can use a site listing such as that published by marketing analytics vendor Quantcast.

Quantcast makes available a ranked list of the Internet's top million sites (see **www.quantcast.com/top-sites-1**). To complete this survey, we'll test each and every one of the sites listed and tally the total number of sites in the cloud and the total number of sites hosted on each provider.

In practice the top 500,000 of these million were used.

The caveat to this technique is that it analyzes a particular cross section of cloud usage and cannot pretend to take in its full breadth. Not included are back end use cases such as servers used for development, for research or for other internal office systems. This adoption of the cloud among enterprises and backend IT systems has been likened to the dark matter of the universe—many times larger but nearly impossible to measure directly.

For now, let's focus on the achievable and examine the results for the high-visibility category of public web sites. See Figures 1 and 2.

From this data, we can draw two main conclusions:

First, on the one hand, cloud infrastructure is in its infancy with a small slice of the overall web hosting market. On the other hand, the cloud is growing rapidly. So rapidly in fact, that Amazon EC2 alone grew 58 percent in the four months analyzed, equivalent to 294 percent annual growth.

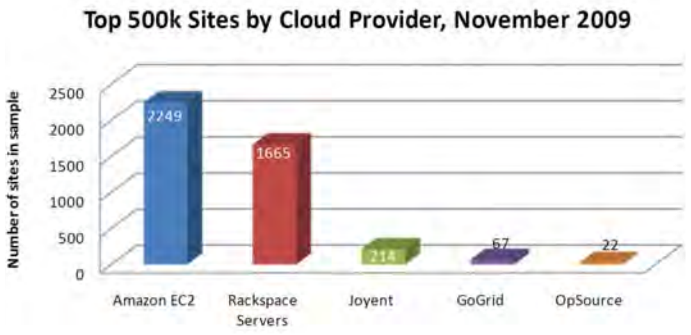

Figure 1: Amazon EC2 has a clear hold on the cloud market.

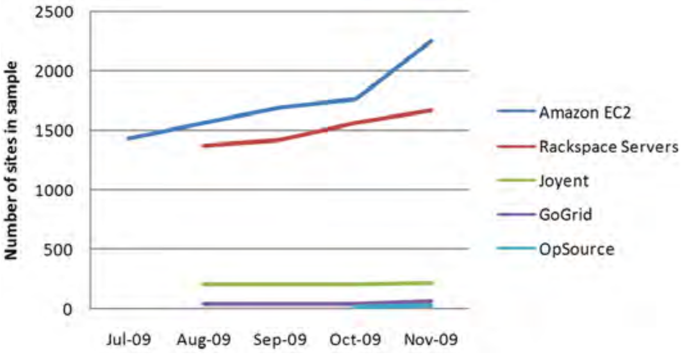

Figure 2: The top 500,000 sites by cloud provider are shown.

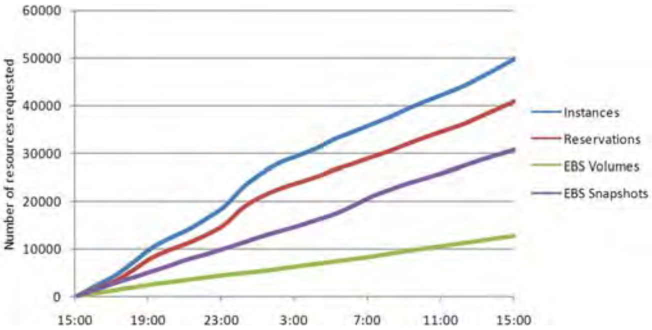

Figure 3: The chart shows resource usage of Amazon EC2 in the eastern United States in September 2009 over a 24-hour period.

Second, Amazon EC2 leads the cloud infrastructure industry by a wide margin. Amazon is reaping the rewards of being the innovating pioneer. Its first cloud service was launched as early as 2005 and the richness of its offering is at present unmatched.

The second study we'll discuss examined the overall usage of Amazon EC2, based on publicly-available data that had been overlooked. Every time you provision a resource from Amazon EC2 (for example, request to start a new server instance) that resource is assigned an ID, such as *i-31a74258*. The ID is an opaque number that is used to refer to the resource in subsequent commands.

In a simplistic scenario, that ID would be a serial number that increments each time a resource is provisioned. If that were the case, we could perform a very simple yet powerful measurement: we could provision a resource and record its ID at a certain point in time. Twenty-four hours later, we could perform the same action, again recording the ID. The difference between the two IDs would represent the number of resources provisioned within the 24-hour period.

Unfortunately, at first glance Amazon EC2's IDs appear to have no meaning at all and are certainly not trivial serial numbers. A mixture of luck and investigations began to reveal patterns in the IDs. One by one, these patterns were isolated and dissected until it was discovered that underlying the seemingly opaque ID there is, after all, an incrementing serial number.

For example, the resource ID *31a74258* can be translated to reveal the serial number *00000258*. (This process was published in detail and can be found in the blog post Anatomy of an Amazon EC2 Resource ID: **www.jackofallclouds.com/2009/09/anatomy-of-an-amazon-ec2-resource-id**.) With these serial numbers now visible, we can perform our measurement as described above. Indeed, during a 24-hour period in September 2009 the IDs for several types of resources were recorded, translated and the resource usage calculated from the differences. See Figure 3.

Over the 24-hour period observed, the quantities of resources provisioned were:

- Instances (servers): 50,242
- Reservations (atomic commands to launch instances): 41,121
- EBS volumes (network storage drives): 12,840
- EBS snapshots (snapshots of EBS volumes): 30,925.

These numbers are incredible to say the least. They show the use of Amazon EC2 to be extensive as well as dynamic. We should recall that these numbers represent the number of resources *created* and do not provide clues to how many of them exist at any given point in time, because we do not know which resources were later destroyed and when.

The above view is of a single 24-hour period. RightScale, a company that provides management services on top of IaaS, collected IDs from the logs it has stored since its inception and broadened the analysis to a much larger timeframe—almost three years (see **http://blog.rightscale.com/2009/10/05/amazon-usage-estimates**).

With this perspective, we can clearly witness the substantial growth Amazon EC2 has seen since its launch, from as little as a few hundred instances per day in the early days to today's volumes of 40,000-50,000 daily instances and more.

Is the Cloud Good for Business?

What's driving adoption is the business side of the equation. We can fold the benefits of the cloud into two primary categories: economics and focus.

The first and foremost of the cloud's benefits is cost. Informal polls among customers of IaaS suggest that economics trumps all other factors. Instead of large up-front payments, you pay as you go for what you really use. From an accounting point of view, there are no assets on the company's balance sheet: CAPEX (capital expenditure) becomes OPEX (operating expenditure), an accountant's dream!

When it comes to IT, economies of scale matter. Maintaining 10,000 servers is cheaper per server than maintaining one server alone. Simple geographic factors also come into play: whereas an on-premise server must be powered by the electricity from your local grid, cloud datacenters are often constructed near sources of low-cost electricity such as hydroelectric facilities (so the cloud is good for the environment as well). These cost savings can then be passed on to customers.

The second reason you might use the cloud is in order to focus on your core competencies and outsource everything else. What is the benefit of holding on to on-premise servers, air conditioned server rooms and enterprise software—not to mention the IT staff necessary to maintain them—when you can outsource the lot? In the new model, your company, be it a legal firm, a motor company, or a multinational bank, focuses on its core business goals. The cloud companies, in turn, focus on *their* core competency, providing better, more reliable and cheaper IT. Everyone wins.

Start-Up Companies Love the Cloud

One sector particularly boosted by cloud computing is the tech start-up space. Just a few years ago, building a web application meant you had to estimate (or *guesstimate*) the computing power and bandwidth needed and purchase the necessary equipment up front. In practice this would lead to two common scenarios:

1) Underutilization: Before that big break comes and millions come swarming to your web site, you're only using a small fraction of the resources you purchased. The rest of the computing power is sitting idle—wasted dollars.

2) Overutilization: Finally, the big break comes! Unfortunately, it's bigger than expected and the servers come crashing down under the load. To make up for this, teams scramble to set up more servers and the CEO, under pressure, authorizes the purchase of even more costly equipment. To make things worse, a few days later the surge subsides and the company is left with even more idle servers.

If there's something start-up companies don't have much of, it's money, particularly up front. Investors prefer to see results before channeling additional funds to a company. Additionally, experience shows that new companies go through a few iterations of their idea before hitting the jackpot. Under this assumption, what matters is not to succeed cheaply but to *fail* cheaply so that you have enough cash left for the next round.

Along comes cloud computing. Out goes up-front investment and in comes pay-per-use and elasticity. This elasticity—the ability to scale up as well as down—leaves the two scenarios described above as moot points. Before the big break, you provision the minimal number of required servers in the cloud and pay just for them. When the floods arrive, the cloud enables you to provision as many resources as needed to handle the load, so you pay for what you need but not a penny more. After the surge, you can scale your resources back down.

One of the best-known examples of this is a start-up company called Animoto. Animoto is a web-based service that generates animated videos based on photos and music the user provides. Video generation is a computation-intensive operation, so computing power is of the utmost importance.

At first, Animoto maintained approximately 50 active servers running on Amazon EC2, which was enough to handle the mediocre success they were seeing at the time. Then, one day, its marketing efforts on Facebook bore fruit, the application went viral, and the traffic went through the roof. Over the course of just three days, Animoto scaled up its usage to 3,500 servers. How would this have been feasible, practically or economically, before the age of cloud computing? Following the initial surge, traffic continued to spike up and down for a while. Animoto took advantage of the cloud's elasticity by scaling up and down as necessary, paying only for what they really had to.

The Animoto story illustrates the tidal change for start-ups. It's not surprising to see, therefore, that the number of such companies is consistently on the rise. If you like, cloud computing has lowered the price of buying a lottery ticket for the big game that is the startup economy. It's become so cheap to take a shot that more and more entrepreneurs are choosing the bootstrap route, starting out on their own dime. When they do seek external investment, they find that investors are forking over less and less in initial funding, out of realization that it now takes less to get a start-up off the ground.

A Bounty of Opportunity

Cloud computing isn't just an enabler for start-ups—the number of start-ups providing cloud-related services is growing rapidly, too. The colossal change in IT consumption has created a ripe opportunity for small, newly formed companies to outsmart the large, well-established, but slow-to-move incumbents.

The classic opportunity is in SaaS applications at the top of the cloud stack. The existing players are struggling to rework their traditional software offerings into the cloud paradigm. In the meantime, start-ups are infiltrating the market with low-cost, on-demand alternatives. These start-ups are enjoying both sides of the cloud equation: on the one hand the rising need for SaaS and awareness of its validity from consumers; on the other hand the availability of PaaS and IaaS which lower costs and reduce time-to-market. Examples of such organizations include Unfuddle (SaaS-based source control running on the Amazon EC2 IaaS) and FlightCaster (flight delay forcaster running on the Heroku PaaS).

The second major opportunity is down the stack. Although providing IaaS services remains the realm of established businesses, a category of enabling technologies is emerging. Users of IaaS tend to need more than what the provider offers, ranging from management and integration to security and inter-provider mechanisms. The belief among start-ups and venture capitalists alike is that there is a large market for facilitating the migration of big business into the cloud. Examples of such companies include RightScale, Elastra, and my own start-up, Vircado.

The third and final category of start-ups aims to profit from the increased competition between IaaS providers. These providers are in a constant race to widen their portfolio and lower their costs. Start-ups can innovate and be the ones to deliver that sought-after edge, in areas ranging from datacenter automation to virtualization technologies and support management. Examples in this category include Virtensys and ParaScale.

I for one am convinced that beyond the hype and excitement the world of IT is undergoing a very real period of evolution. Cloud computing is not a flash flood: it will be years before its full effect is realized.

Biography

Guy Rosen is co-founder and CEO of Vircado, a startup company in the cloud computing space. He also blogs about cloud computing at JackOfAllClouds.com*, where he publishes original research and analysis of the cloud market.*