

The Compelling Economics of Cloud Computing

A Joyent White Paper

It's More Than Just Hardware and Manpower Costs

You're a CIO. If you had a dollar for every time someone asked you about your cloud strategy, you'd probably be retired by now. Well, maybe we are exaggerating. But not by much.

The buzz around cloud computing is deafening, between media coverage, IT industry analyst assessments of this nascent field, and the rapidly growing list of conferences and panels addressing the cloud. To date, analysis of the true benefits of moving critical computing functions into the cloud remains largely framed by a straight-up comparison between buying servers, renting portions of servers at co-location facilities, and buying computing and storage capacity in a public cloud. The results of this comparison are fairly predictable. Buying cloud capacity is cheaper than renting a part of a server is cheaper than buying your own hardware. A secondary and common benefit is the ease of scaling an application up or down in the cloud.

But there are other key elements to any shopping decision for a cloud deployment, not only related to cloud versus co-lo versus buy, but also within the diverse range of cloud providers themselves. Further, in any Capital Expense (CapEx) or Operating Expense (OpEx) calculations, it's important to consider capabilities that are truly unique to cloud-based deployments that may not show up in traditional CapEx and OpEx calculations. So here's a quick rundown.

What is the Cloud? A Quick Primer

Chances are you know this already. But just in case, here is the definition of cloud computing from the National Institutes of Standards (NIST). This is our preferred definition:

“Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics:

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

It comes in three service models:

- Cloud Software as a Service (SaaS)
- Cloud Platform as a Service (PaaS)
- Cloud Infrastructure as a Service (IaaS)

Finally, it is delivered in four deployment models:

- Private cloud
- Community cloud
- Public cloud
- Hybrid cloud

The key concepts above are powerful, customizable, convenient, cheap, on-demand, rapidly scalable, and configurable. In fact, the cloud is not really a new thing. All Web-based email applications are in the cloud. The Internet itself is a type of cloud. What's different now is that, for the first time ever, organizations are using on-demand computing power to replace an entire application stack with no concern for where the actual computers running the stack are located.

Naturally, servers, load-balancers, operating systems, routers and other key pieces of technology all power the public cloud and provide the baseline computational power. But buying computing power in the cloud could mean that a particular virtual computer

is located in Boston, Tokyo, London or Sao Paulo, either simultaneously or sequentially and in various combinations of the four at different times of day or night. The physical cord to the data center remains, but the intermediary layer of virtualization of computing resources has made that cord far less important. There is still the potential for huge failures - witness what happened to Amazon Elastic Compute Cloud in April 2011. But over all, the capability to cut the chord, save money, gain flexibility, and future-proof an application or site makes cloud computing so enticing.

Why the Cloud Rocks

Pure CapEx and OpEx Utilization Comparisons

Think of it as the difference between a timeshare, a vacation condo shared with some friends, and a standalone vacation home that you own yourself. If you own the vacation home, you have to maintain it and pay for it. If the roof leaks, you pay for it. Ad nauseum. You bear all marginal costs, even if you only use it twice a year. This is the same as owning your own hardware and running your own data center and computing infrastructure.

If you share a vacation condo with a few friends, you have a co-location facility. It's a bit better than owning outright yourself, but chances are the condo remains empty some of the time and the maintenance costs are still heavy.

With a vacation timeshare, lots of other people help pay for the upkeep of the unit and it's almost always occupied. Plus, you can enjoy a more luxurious place, like one with a pool, than if you were paying entirely yourself. Like a timeshare, the cloud allows lots of people to share computing capacity on an as-needed basis, allowing them to ratchet up or dial back computing resources when desired. This type of arrangement can reduce an organization's capital expenses and operating expenses substantially. Microsoft pegged the number at 80%.¹ A 2011 study by IT consultancy Gartner, Inc. of CIOs in the United Kingdom placed cloud savings estimates on IT spending at 50%.²

¹ "Microsoft study shows 80% savings by using the cloud", NetworkWorld, March 10, 2011.
Robert Mullins

² "Cloud computing to save tech budgets", Silicon.com, January 24, 2011. Nick Heath.



Photo credits: <http://bitly.com/m1iAml>

In CapEx, this reduction cuts across the entire data center infrastructure including servers, storage arrays, software licenses (when needed), routers, and load-balancers. On the OpEx side, costs shared by cloud deployments include sys admins, hardware engineers, network engineers, facilities management, electricity, fire protection, and insurance or local and state taxes on facilities. There are other hidden OpEx costs that a cloud instance can eliminate such as purchasing and acquisition overhead, asset insurance, and business interruption planning and software.

Any cloud provider worth its salt has baked business interruption planning into the mix. Except....

Not All Clouds Are Equal

Some Fail. Some Cost a Lot More.

If you are a CIO who is cloud shopping, you have certainly read about the massive outage of Amazon Elastic Compute cloud (EC2) service in April 2011. EC2 is by far the most widely used cloud service. In fact, many cloud-based development and deployment platforms, such as the Ruby-on-Rails Platform-as-a-Service company Heroku, rely entirely on EC2. But as users of EC2 quickly learned, despite massive redundancy and enormous collective computing resources, public clouds can go down. With the Amazon outage, a bottleneck inherent to the architecture of EC2's storage arrays was exacerbated when a hardware failure occurred. This added more pressure to the bottleneck, causing a death spiral as servers and storage arrays searched around wildly for the proper resources to complete assigned tasks.

However, not all public clouds have these sorts of bottlenecks and it is possible to design a cloud architecture in such a way as to utilize numerous smaller input and output pathways and eliminate architectural bottlenecks while actually improving both stability and response times. The point here is that architectural differences matter a lot.

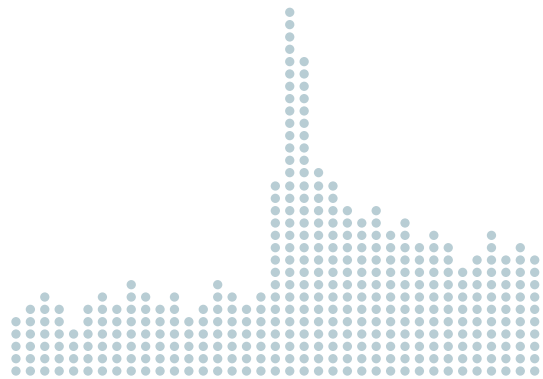
A second key piece to why not all clouds are alike is the actual structure of the virtualization on the hardware. EC2 and many other clouds segment servers and storage arrays into "virtual machines"—meaning hardware emulation. This adds an extra layer of abstraction that creates additional computational drag on the system and results in both lower capacity and slower response times. Other cloud providers have chosen to build a virtual operating system (OS) and virtual software stack rather than emulate the hardware. Eliminating this layer of abstraction has tremendous benefits. Virtual OS clouds can be more than ten times faster than clouds built on virtual hardware architectures.³ This efficiency allows CIOs to purchase less capacity (and spend less money) to achieve the same results. Eliminating the hardware portion eliminates another failure point and level of complexity, further enhancing stability.

So how can this impact true spending tallies? Owning the hardware and software outright and running the data center in-house will generally result in a spend between eight and ten percent of total revenue for technology intensive businesses. Using co-location and renting server space can bring that down to between five and eight percent. Running computing capacity in EC2-like versions of the public cloud will cost between two and three percent of total annual revenues. So there is a tremendous difference not only between owning outright and co-location, but also between different types of cloud deployments.

Insuring Against Unexpected Spikes

Things the Cloud Can Do That Physical Ownership Can't

With more than one million daily active users of its games around the world, Digital Chocolate is a large, successful Zynga-like social games company with a loyal fan base. Social games are incredibly viral and demand can spike very quickly; double-digit increases in required compute power over the course of minutes are not uncommon. That demand can reverse and ebb just as quickly. Adding and removing capacity - scalability - is essential to maintaining user experience and to keeping IT infrastructure costs under control. From the start, the Digital Chocolate team realized that to keep its business running it needed to run in the Cloud. By being in the Cloud, it could quickly augment capacity in response to minute-by-minute and hour-by-hour changes in user traffic and demand.



Such behavior would simply not be possible outside the Cloud. Buying and bringing online a new server in real-time can't happen. Even if one is ready and waiting in the rack, it takes a few minutes (at a minimum) to bring it onto the network. Then that server is yours forever and will likely never be perfectly optimized. Virality in most online media creates wildly uneven demand periods, even within the course of a day. Likewise, seasonal demand cycles are also quite common in many verticals. In eCommerce, for example, over 50% of all demand is concentrated in the holiday

season. The operating costs of maintaining that server, whether it's running at five percent capacity or 100 percent capacity, are equal. You still need a systems administrator to patch the box and deal with software license issues. You still need a network engineer to make sure the server is properly configured for the network.

While buying a fraction of a server in a co-location facility is a better value for uneven demand cycles, it still presents the same problems of higher costs resulting from poor resource optimization across both the CapEx and OpEx realms. Contracts for co-location facilities run three years and, again, cannot happen in real-time. Systems administrators have to configure new servers and business people have to sign agreements. You may not own the hardware, but you own all the associated costs and an inflexible three year lease. And none of this happens very quickly.

With Cloud deployments, it's possible for Digital Chocolate only to buy the compute power required for the big game launch or in response to a positive blog review, throttling up capacity on-demand and then throttling back down. So for CapEx optimization, reliance on Cloud computing insures Digital Chocolate against sharp spikes when the entire state of New York decides to log on to "Galaxy Life". On the OpEx side, the insurance is similar. The Cloud allows companies to ramp-up capacity without adding technical staff or IT professionals. In the current hiring environment, where IT professionals demand a fast growing premium, controlling headcount is perhaps even more valuable than controlling hardware or capital costs.

Make Sure Your Insurance is Not Merely Notional

A word to the wise here. Just putting your infrastructure in the cloud is not enough to guarantee insurance against traffic spikes. To ramp up capacity quickly on complex software stacks, extensive preparation is required for big applications running on EC2. It's not just as easy as flipping a switch and can take real systems administrators time to build-out additional capacity - unless this need has been anticipated and arranged for in advance. That said, the best cloud providers target minimum performance thresholds rather than maximum performance thresholds. Minimum performance

thresholds basically mean that the promised compute capacity is merely a minimum service level that will expand on demand in case of a big workload burst. The expansion will not be infinite, but should be enough to survive most decent sized spikes and provide a buffer while additional compute resources are assigned or configured.

Does your cloud provider have a minimum performance or a maximum performance approach to its customers? This is a good question to ask because minimum thresholds reflect a willingness to increase capacity first and ask questions later—exactly what you want your cloud provider to do in a pinch. The whole purpose of being in the cloud (and buying virality insurance) is not to have to think about it. Minimum performance thresholds are like buying a Porsche that runs great at 55 MPH but you know can peg it in a pinch, no questions asked. Maximum performance thresholds are more akin to buying a Camry or an Accord that runs well most of the time, but, for top performance, requires some modifications that don't happen automatically or in real-time.

How the Cloud Amps Up Innovation

LinkedIn is one of the dominant social networking sites and is now a mature Silicon Valley company. Yet LinkedIn manages to launch an astonishing array of new products in a short span of time, almost in defiance of normal organizational gravity. How? LinkedIn leverages cloud deployments to cheaply stand up new products and turbo-charge innovation.

The cost of standing up an entirely new application in the cloud is so much cheaper and easier than in a traditional datacenter. There is no need to buy new servers, rent co-location capacity, or hire IT staff. Likewise, cloud deployments are very easy to dismantle without cost penalties when new products simply don't work out. The broader implication for innovative companies is, rather than worry about buying huge amounts of hardware to launch a new product, or signing rigid co-location agreements, it's far easier now to build lots and lots of new applications, try them out on the Internet, and then either scale them up or shut them down.

For companies like LinkedIn, the ability to try out lots of different things is a key part of their innovation strategy. In other words, dramatic decreases in CapEx and OpEx for a new idea mean lots more new ideas get created and hopefully some are home runs.

One caveat is in order here. Just because it's cloud doesn't mean it's push button simple to bring up a new application. Rather, the mileage varies widely between cloud providers in terms of ease of use and speed at which it is possible to stand up an innovative application. So if this is a key factor in your decision, make sure you have a complete grasp of the software and IT architecture requirements required of your team to stand up quickly in the cloud.

Conclusion

The cloud is on-demand computing power that is easy to scale up or down and is not physically tied to a single or even two or three locations. Using the cloud can dramatically cut CapEx and OpEx for any company that deploys a lot of technology. It can do so in a number of ways:

- First, **the cloud saves you money** by allowing organizations to pay for compute capacity costs and operating costs (such as IT staff and facilities costs) on fractional and on-demand bases that better map to the actual compute demand curve. Not all clouds are equal, however, so the architecture of the cloud is a critical factor in determining stability and reliability. Also, software differentiation between cloud providers can mean that some clouds are slower than others. Running on less efficient clouds costs more because more compute capacity and storage are required.
- Second, **putting key infrastructure in the cloud buys “virality insurance”** by making it possible to call-up additional compute capacity in an instant and scale instantly to meet demand. This is increasingly a must-have feature for businesses due to the inherent uneven demand that is a reality for so many businesses using the Internet to run key parts of their operations.
- Third, **cloud computing makes innovation easy** by making it so much cheaper and easier to try many more new products and applications. This means more new ideas are tried and penalties for failure are small, lowering obstacles to innovation.

Cloud deployment costs far less up-front than putting in place physical server capacity or renting shared server capacity for a three year term. So having a cloud strategy alone is not enough. Having a smart cloud strategy will make just about everything in your IT infrastructure easier to manage, while reducing CapEx and OpEx costs by a factor of ten and turbo-charging innovation.

For more information, please visit www.Joyent.com.