SCTE CABLE-TEC EXPO'13
OCTOBER 21-24 / ATLANTA, GA

SCTE Society of Cable Telecommunications Engineers

Technology Evolution Driving Cloud Solutions

Cloud: as-a-Service Delivery

A Technical Paper prepared for the Society of Cable Telecommunications Engineers
By

David M. Grimes
Chief Technology Officer
NaviSite, a Time Warner Cable Company
125 Elwood Davis Rd.
315-634-9333
dgrimes@navisite.com

Overview

Just a few years ago the term cloud entered the IT jargon and quickly dominated much of the discussion within the industry. The problem, though, was a lack of clarity as to what cloud actually meant. One of the first widely accepted definitions is now more precisely referred to as Infrastructure as a Service, or IaaS. Indeed Amazon EC2 - the Elastic Compute Cloud - was one of the earliest manifestations of what we now commonly call cloud computing.

But the concept of cloud has been around for much longer than just these past few years, and the term itself has been co-opted by another well-established delivery model: Software as a Service, or SaaS. One of the epitomic players in this space, Salesforce.com now uses the term cloud pervasively in its branding and marketing. We see a similar co-option across a variety of other technologies such as Desktop, Backup, Storage and Disaster Recovery. Most providers of these services liberally categorize their offering as "cloud services", and cloud has become synonymous with the "as-a-Service" delivery model.

So what about cloud is so compelling that it has become the word of choice to describe this model of consuming technology? Perhaps more importantly, what key technology innovations have contributed to this transformation? In this paper, we will look at the technology evolution which has enabled the current set of "as-a-Service" offerings, the capabilities and limitations of the current state of technology, and what is on the horizon.

Most of the foundational advancements have come in the form of the realization of Moore’s Law which states:

"The number of transistors incorporated in a chip will approximately double every 24 months."

-- Gordon Moore, Intel co-founder [1]

Foundational Technology Advancements

Although Moore's prediction was specific to transistor density, it has been more generally interpreted to mean "performance will double every 24 months." For much of the 4 decades since Moore made his prediction, that has indeed been the case. Recently, though, this doubling of performance has not come in the form of improvements to single threaded workloads, rather it has come in the form of multi-core CPUs. While many workloads do not benefit from massively multi-core CPUs, virtualization definitely does! The continued increases in core count, combined with enhancements specifically aimed to improve virtualization performance (AMD-V, VT-x) have led to massive server consolidation. Competitively priced 2-socket systems can be readily configured with 10-core processors presenting 40 logical CPUs to the hypervisor. Given it is not uncommon to see best practices suggesting manageable oversubscription ratios of 3 to 1 or higher, it is quite practical to expect VM-to-host consolidation ratios of 50+ to 1.

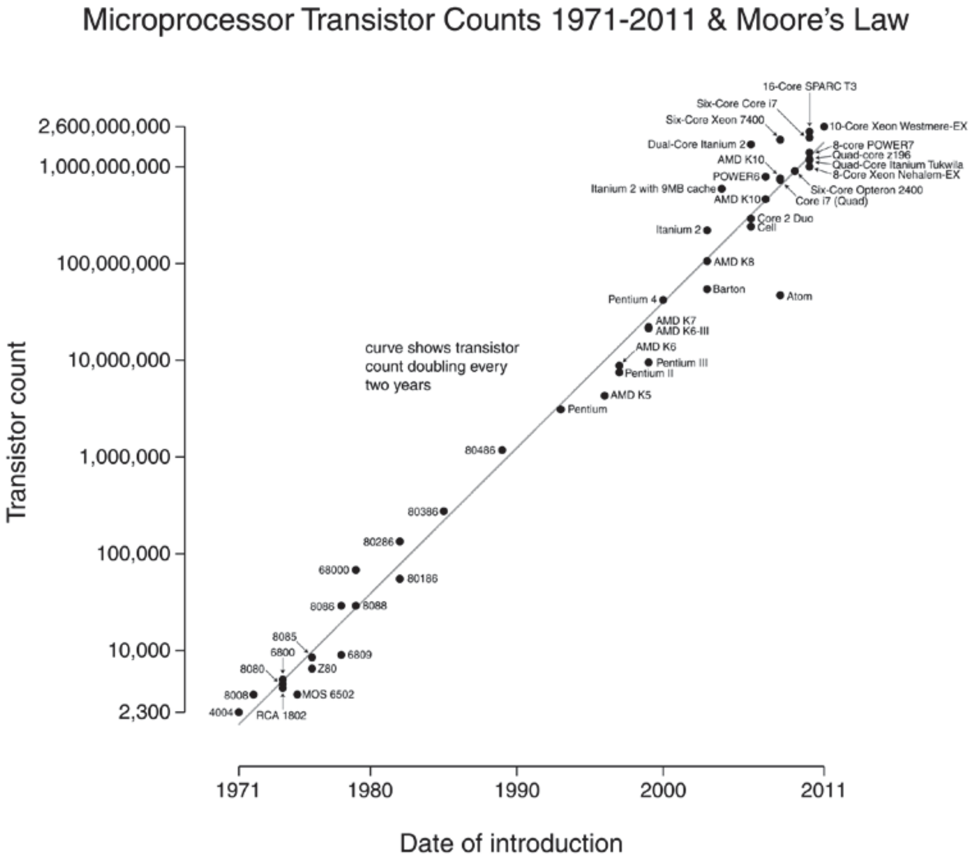

Figure 1 – Microprocessor Transistor Counts 1971-2011 & Moore's Law

While these numbers are practical from a CPU perspective, they also generate demand for ever larger RAM configurations. Thankfully, RAM has also loosely obeyed Moore's law, and it is now common to find 2-socket server blades with 192 or even 384 GB of RAM. These configurations are based on commodity priced components, not the "high end" memory which historically commands a nearly 2 to 1 or higher price premium. Much larger RAM configurations are available (1TB of RAM anyone?), as are varied CPU configurations including 4-socket and 8-socket options. However, a large virtualization infrastructure running a heterogeneous workload is most cost effectively handled by configurations which fall in to the "sweet spot". Specifically configurations which are based on commodity components that maintain an average GB of RAM to logical core ratio of roughly 6-12 are most effective. Higher ratios tend to lead to too much oversubscription of CPU (vCPU to pCPU ratio gets too high, often leading to performance issues), whereas lower ratios are often RAM starved.

These advances in CPU capabilities and RAM capacities have been accompanied by a shift toward very high density (in terms of power per square foot and rack unit) form factors. While discrete servers in a 1U or 2U form factor are still quite common, many virtualization infrastructures leverage very high density chassis/blade configurations incorporating as many as 18 high performance blades into 5U. Many of these chassis/blade solutions leverage a shared uplink network design to minimize northbound port density requirements. Network capacities are yet another area where Moore's law has been loosely followed, and 10+Gbit Ethernet is now commonplace both at the datacenter core and often at the distribution layer. While much of the network traffic in a highly virtualized environment is East⇔West traffic (traffic between VMs, often contained within the backplane or L2 domain), different deployment topologies can lead to varying North⇔South traffic (traffic in or out of the virtual environment) volumes. In most heterogeneous environments, however, today's underlying network capacities are more than sufficient to handle the traffic even with the theoretical choke points of a shared uplink design.

So far we've covered CPU, RAM and network, but there is a 4th pillar of technology which completes the foundation of IT infrastructure: storage. Storage technologies have kept pace with advances in technology more broadly, but primarily in the form of capacity. Individual drives in the "big and slow" 3.5" SATA class are now available in 4TB capacities, and the more popular 2.5" SAS drives typically used where IO performance is a concern are available in sizes up to 600GB (15k RPM), or even 900GB (10k RPM). The issue, though, is that spinning media storage has not undergone a similar transformation when it comes to performance. Access latencies, IOPS, and maximum transfer rates today are not materially different than they were a few years ago. Although individual drive performance has not changed significantly, enterprise class storage arrays have continued to advance and largely meet the needs of the highly consolidated compute environment discussed previously. Additionally, similar to virtualization-improving CPU extensions, collaboration between virtualization software vendors and storage vendors are providing capabilities which optimize a variety of use cases. One such example of this is VAAI (VMware vSphere Storage API – Array Integration) [2]. This is an API which allows, among other things, the hypervisor to offload common tasks to the array itself, freeing up CPU cycles, reducing bandwidth on the SAN, and enabling these common tasks to complete much more quickly – sometimes in order of magnitude terms.

In addition to these advances, another technology has been making inroads in the storage market, specifically in the context of virtualization and consolidation: NAND flash. Flash augmented spinning media arrays, as well as 100% flash solutions are now available at a price point which, while still a multiple of traditional spinning media, is becoming viable. The price premium of flash storage is somewhat mitigated in the service provider world due to the ability to purchase and operate at a large scale, resulting in buying leverage not typically available to the small enterprise. Flash storage has a number of benefits including low latency IO, significantly higher IOPS per unit, lower power consumption, and higher potential total density per RU. Even with these upsides, flash storage still represents a relatively small portion of the storage market. Cost is still the primary inhibitor, but also the reality that not all workloads actually need the capabilities of flash storage. It does, however, provide a significant step forward for use cases not well served by even the best available spinning media solutions.

As we can see, there have been significant advances in the underlying “core four” (CPU, RAM, storage, and network) technologies which form the basis of almost all IT infrastructure. These advances have led to massive consolidation enabled by virtualization. While this allows dramatically more efficient use of these resources, it also creates an environment with a much larger fault domain. While likelihood of individual component failure is not inherently higher, and is often lower today than it has been in the past, the impact of failure can be much broader. This reality demands that platform architects take this into account when designing a virtual infrastructure. Not only should component level resilience be taken into account (redundant PSUs, redundant connectivity and associated configuration such as LACP, SAN multi-pathing, etc.) to eliminate single points of failure, but the fault domain itself should be well understood and well defined. Configuration drift over time and lack of proactive maintenance and adherence to reference architecture plans are common challenges in the long term operations of a large virtual infrastructure.

Cloud Evolved: XaaS

These advancements in core technologies have prompted a transformation in how IT services are delivered. While this transformation is still in its early stages, and is in many ways a continuing evolution of the trends we've seen toward outsourcing, the "as a service" delivery model has gained momentum. What specific aspects of IT have been moving in this direction?

Infrastructure as a Service, or IaaS, is one of the areas which has been most notably effected. While it is true that traditional methods still dominate, in percentage terms, with existing IT deployments there is considerable acceleration in the adoption of IaaS. IaaS itself is enabled through a combination of the foundational technology advancements discussed previously in addition to innovation in virtualization software. With a few exceptions, modern hardware is simply too powerful to be effectively utilized by a single workload. Virtualization enables enterprises and service providers, operating at scale, to consolidate and more fully utilize the underlying horsepower of these systems. This increased utilization drives efficiency, in turn optimizing cost.

But cloud, and by extension the "as a service" model, is more than just virtualization. What then is cloud? The National Institute of Standards and Technology (NIST), has developed a reference definition which states:

"Cloud computing is a model for enabling ubiquitous, 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 is composed of five essential characteristics, three service models, and four deployment models."

--National Institute of Standards and Technology [3]

Let's take a look at the five essential characteristics, and consider what each offers in terms of value proposition relative to traditional methods:

On-demand self-service

This characteristic enables the consumer of the services to provision (and release) resources at will, without human interaction. Whether these resources are compute resources for server workloads, storage resources, or even hosted virtual desktops, what historically took days or weeks is now possible in minutes.

Broad network access

This characteristic captures the fact that cloud resources are typically hosted in a 3rd party datacenter and accessed by a variety of end user computing devices over the network. In order to ensure access is ubiquitous, cloud based solutions adhere to a variety of delivery standards. With continued improvements in both last mile wired connectivity, as well as high speed wireless access from sophisticated mobile devices, historical problems associated with latency and bandwidth are no longer barriers to adoption.

Resource Pooling

Due to the rough adherence to Moore's Law relative to increases in capacity across the “core four” resources (compute, memory, storage, and network), it is possible to leverage a large shared pool of resources across many tenants. While this does allow service providers to optimize their infrastructure and drive efficiency, the multi-tenancy has been one of the most significant barriers to adoption in the enterprise. Thankfully, through a variety of resource specific mechanisms, proper tenant isolation is manageable.

Rapid elasticity

While capacity planning is always a good practice, the burden of capacity management falls to the service provider – not the tenant – in a cloud based infrastructure. Individual tenants can add or remove resources as needed (including via automation) without concerning themselves about the underlying physical infrastructure. In contrast, the service provider is able to manage capacity at an aggregate level where individual tenant ebb and flow is a statistically insignificant impact to the larger trend.

Measured service

Cloud enables a transition from a CAPEX based model to an OPEX based model, and costs scale with the business. Rather than overbuying infrastructure in an attempt to guarantee future needs are met, and then revisiting periodically during a “refresh cycle”, the consumer of cloud services can think more in terms of “just in time” resources, and pay only for what they use.

Conclusion

While in many ways evolutionary in nature, cloud is revolutionary in the way it enables business agility. The five key characteristics of cloud can be truly transformative to a business. Let's look at an example to see how:

Consider an online retail business focused on a relatively narrow vertical market in a relatively narrow geography. In a traditional infrastructure model, if this business wanted to test adjacent vertical markets, it would likely need to build a business plan, secure and deploy infrastructure (justified by the business plan), and finally launch the new service. There are two possible outcomes – either the entry into the adjacent market is successful, or it isn't. If it isn't successful, the investment in infrastructure was a waste. True, the infrastructure could be repurposed, but more likely this infrastructure requirement would have been a prohibitive barrier to attempt the project in the first place. The alternate outcome, success, is equally challenging. If the initial investment in infrastructure was minimalist, which is likely given the relatively high cost, it may not be able to handle the load associated with real success. Growing that infrastructure would take time as well due to the length of traditional procurement cycles.

Alternatively, a cloud-based infrastructure is superior in both outcomes. Initial investment in infrastructure is low, since cloud is an OPEX based model. So, what in many cases in traditional models causes a project to not even get off the ground is of limited concern when choosing cloud. In the case the project fails, simply decommission the cloud infrastructure, and eliminate any ongoing cost. In the case of success, the cloud based infrastructure can be scaled up to accommodate load as needed. In short cloud enables true business agility, and can transform a business and the way it approaches the market.

The past several years have seen a continued rapid pace of evolution within the core technologies that service IT, as well as accelerated innovation that has enabled an emerging class of "as a service" offerings. Some of these offerings are still in their infancy, but there is already evidence that we've begun to "cross the chasm" [4] that lies between visionaries and the pragmatists. Technologies that are widely available today, as well as those just on the horizon, will ultimately make it easier for people to realize the benefits derived from the five key characteristics of cloud. When engineered and operated at scale by a service provider, the tenant can reap these benefits without being encumbered by many of the underlying technical complexities.

One thing is clear; while the term cloud may be as overused today as it has been since it came into the discussion, there is no doubt it has spawned a new way of thinking about IT.

Bibliography

(1) [http://www.intel.com/content/www/us/en/history/museum-gordon-moore-law.html]

(2) [http://www.vmware.com/files/pdf/techpaper/VMware-vSphere-Storage-API-Array-integration.pdf]

(3) [http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf]

(4) Crossing the Chasm, [http://www.geoffreyamoore.com/books-by-geoffrey-moore/]

Abbreviations and Acronyms

Abbreviation	Description
Acronym	Description