TIP Publication Date

June 18, 2013

Understand Your Cloud Provider Strengths

Applications Need to Take Advantage

Not all clouds are made equal. Let's get that out of the way first - there are lots of variables in underlying hardware, hypervisors, memory, security measures, and management overhead to name just a few. These differences could make for interesting variations in performance. That is the topic I would like to talk about here.

At one of our High Performance Computing projects there was a need for a typical cloud characteristic, namely elasticity - the all-important feature of 'grabbing' more resources when you need them and releasing them when you are finished. For this particular task we needed about 50 NVidia M2050 (or higher) GPU's for a couple of days per month to run a month- end simulation report, a report even with 50 GPU's could take over 72 hours to complete. Red hot number crunching for a real time variable Annuity Hedging platform.

At the time, we engaged with Amazon because we were interested in their GPU VPC offering which came out in July of 2012. As part of our Proof of Concept approach when running performance tests with the standard NVidia CUDA utilities, we noticed that the NVidia M2050 cards in the Amazon cloud performed on average at 70% of our own NVidia M2050's that we were running and managing at our Co-location facility. To be sure, these tests run on the card GPU's itself, and therefore the networking infrastructure is not a factor here.

1st Mental Note "In general, do not assume that the performance (of applications) in the cloud matches your own environment, virtualized or not".

This was a Cloud to Earth moment to me; if we did not have our own hardware to test against, we would not have known there was a performance hit in the first place.

A 30% less performance would mean we needed to get 15 more GPU's to get the same compute capacity. That is a lot of dollars, believe me, or even worse, we would miss the contractually obligated service level to produce these reports within a certain number of days.

Amazon worked diligently with our technical staff to find the root cause, and eventually eliminated most of the performance differential.

This particular instance made me wonder if this was just a fluke (the 30% performance hit), or is there a more system wide trend among cloud service providers we need to be aware of? For example:

- Independent research[1] has shown that over a 30 day period in 2012 – Amazon EC2 performed on or above average in Virtual CPU and RAM performance, and the Disk performance and Virtual CPU are stable to highly stable, as compared to others in the IaaS space.

- This same research also pointed out that both VPN and Disk perform below average for IaaS, and RAM and VPN connection performance are found to be unstable.

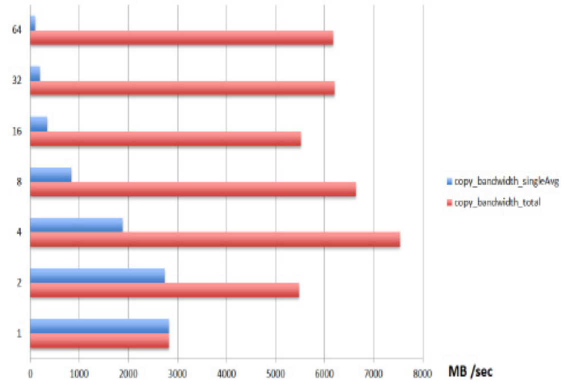
- In another example using a private cloud, running OpenStack Diablo release on top of Ubuntu 10.10[2], STREAM benchmark stress testing showed a diminishing bandwidth per machine instance as the number of instances increases, and variability in overall bandwidth (see diagram on next page)

- Boot times of instances within a single provider can vary significantly, Rackspace UK outperformed Rackspace US by about 3x, and by about .5 X when compared to aws-us-west (m1.large)[3]

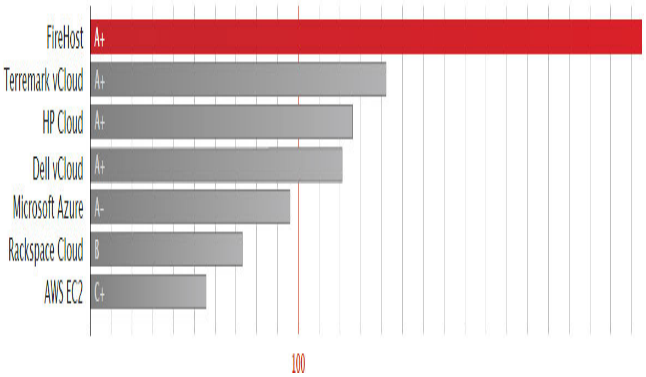
- Firehost storage performance showed double or better performance against Terremark, AWS, Azure, etc.[4]

2nd Mental Note: "Know the characteristics of the application(s) you move to the cloud, understand memory, disk, storage requirements, how fast instances need to come on-line, bandwidth consumption, and equally important, when does performance degradation occur?"

[1] CloudSpectator – Amazon EC2 report 2012.
[2] Fair Benchmarking for Cloud Computing Systems – University of Surrey, UK
[3] Fair Benchmarking for Cloud Computing Systems – University of Surrey, UK
[4] Firehost benchmarking 2013 – Vendor Commissioned



Diminishing bandwidth per machine instance as the number of instances increases



Storage performance using Fio, compilebench, IOzone, postmark, tiobench, aio-stress

Depending on what you are looking for, I listed three websites, two of which offer specific visualizations of Cloud performance CloudHarmony & Cloudsleuth , and the third, OpenBenchMarking.org features benchmark results, tons of information but rather unpleasant to navigate:

- Cloudsleuth offers visualization of response times and availability for Cloud provider data centers, overlaid onto a World map. However, no specific benchmarks are reported.

- CloudHarmony has been monitoring the availability of Cloud services (38 in total) since late 2009. Providers monitored range from those offering IaaS such as GoGrid, Rackspace and Amazon Web Services, and PaaS services such as Microsoft Azure and Google App Engine.

- OpenBenchMarking.org offers information from benchmark runs based around the Phoronix Test Suite. Test results are available for various systems, including those of AWS and Rackspace. However, presentation of results can lack information and is not readily comparable.

The table below has a small selection of useful benchmarking tools that are being used in the field today;

Memory IO	STREAM, ramspeed, CacheBench, Geekbench
CPU	LINPACK
Disk IO	Bonnie++, IOzone
Application Compression	Bzip2
Network	Iperf, MPPTEST, scp, curl, wget
Storage	Fio, compilebench, IOzone, postmark, tiobench, aio-stress

Conclusion

The increased visibility and publicly available metrics being generated by third parties, have driven the large global cloud services providers to improved performance & stability numbers in most metric reports – in order to obtain the highest levels of performance for your application(s) you need to map out the application characteristics, the optimal performance points, and understand application degradation points, and overlay these against the business performance and availability requirements. Understanding your applications will allow you to select the best cloud provider. It becomes clear that companies will have to select a myriad of cloud providers to ensure optimal performance for their business critical application, since there is no One Size Fits All.

This TIP was written by Ton Roelandse, who specializes in Applied Cloud Technologies and other cool topics. Ton welcomes comments and discussion on this topic and can be reached at ton.roelandse@trexin.com.