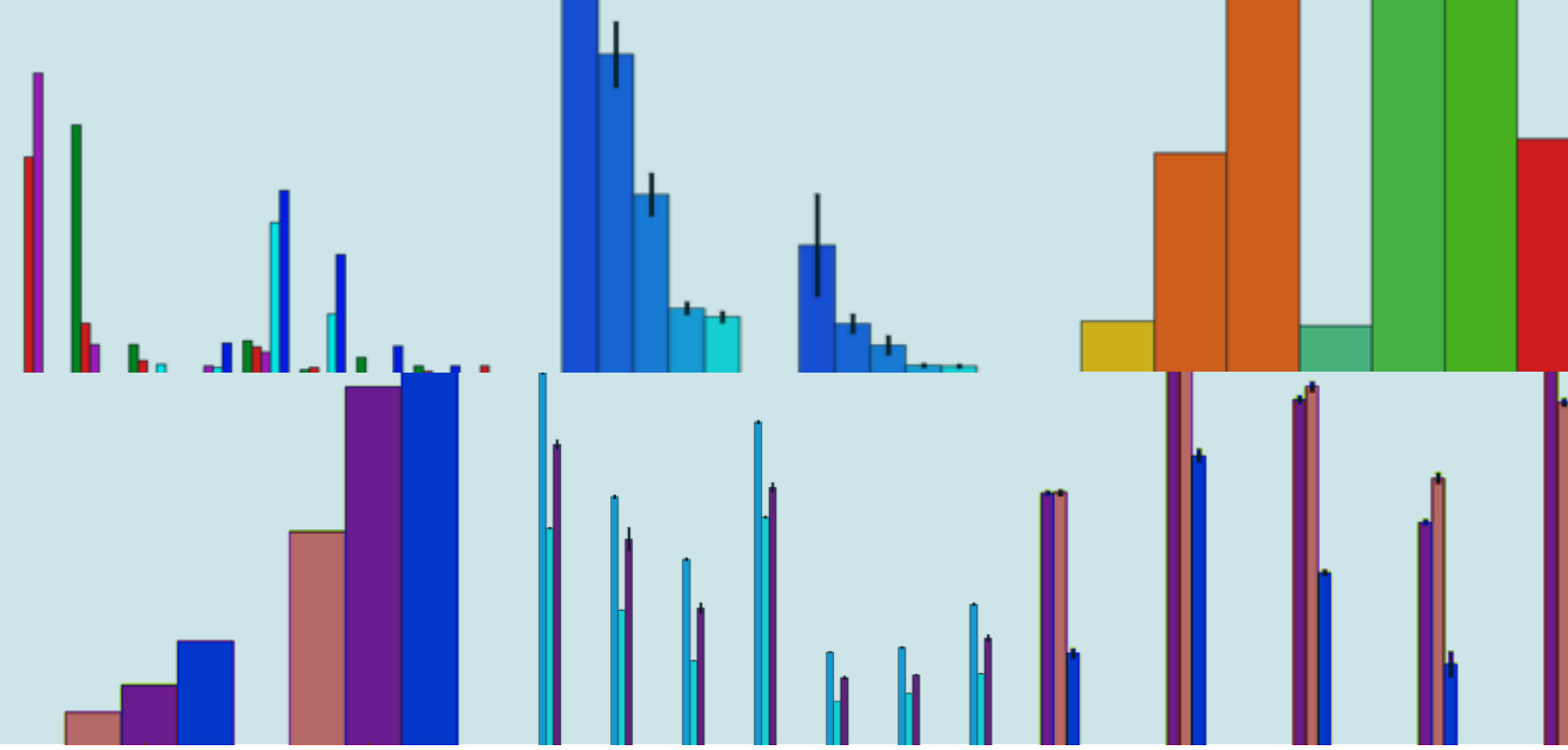


CITIZEN SCIENCE DATA FACTORY

A Distributed Data Collection Platform for Citizen Science

Appendix B: Cloud Computing Performance Testing

Prepared by

azavea

scistarter.com

The Citizen Science Data Factory report includes four parts:

- Part 1: Data Collection Platform Evaluation
- Part 2: Technology Evaluation
- Appendix A: Wireframe Designs
- Appendix B: Cloud Computing Performance Testing

It is available for download and distribution from:

- http://www.azavea.com/research/company-research/citizen-science/
- http://www.scistarter.com/research/

Funded by:

The report was funded in part by a grant from the Alfred P. Sloan Foundation. The opinions expressed herein are those of the authors and do not represent the opinion of the Foundation.

This report was researched and developed by Azavea and SciStarter, including contributions from:

- Darlene Cavalier, SciStarter
- Robert Cheetham, Azavea
- Rob Emanuele, Azavea
- Jeff Frankl, Azavea
- Mary Johnson, Azavea
- Tamara Manik-Perlman, Azavea
- Josh Marcus, Azavea
- Michael Tedeschi, Azavea

This report also benefited from input provided by both the project representatives interviewed and input from the Citizen Science Community Forum.

Azavea is a B Corporation that creates civic geospatial software and data analytics. Our mission is to apply geospatial technology for civic and social impact and to advance the state-of-the-art through research.

SciStarter brings together citizen scientists; thousands of potential projects offered by researchers and organizations; and the tools, resources, and services that enable people to find, pursue and enjoy these activities.

Published by:

Azavea
340 N 12th St, Suite 402
Philadelphia, PA 19107
Telephone: +1 215 925 2600
Web: www.azavea.com
Email: info@azavea.com

SciStarter
306 Delancey St
Philadelphia, PA 19106
Telephone: +1 267 253 1310
Web: www.scistarter.org
Email: darlene@scistarter.com

Cloud Computing and Citizen Science

This report has aimed to outline the requirements for a future citizen science data collection platform. We have included an evaluation of existing online citizen science projects, contemporary data collection technology, sensor platforms, and a UI/UX design for a future platform. We have suggested that a future data collection platform should have several technical attributes:

1. Open source
2. Leverage existing efforts
3. Scalability
4. Cost of implementation
5. Carrying cost over time

We have also suggested that this platform should support integration with sensor data, have strong visualization features, and implement some basis analysis features. There are many challenges to overcome in order to build such a platform, but we believe that one of the most significant technical challenges centers on performance and scalability.

There are several ways to interpret "web site performance" but we are referring to speed and responsiveness of a web application. Several studies of web site performance have demonstrated that speed matters. Fast, responsive web sites have better user engagement. Further, fast is different. A faster user experience enables new types of user experience. This is the core idea behind many of Google's innovations, including instant search response, predictive word completion, GoogleMaps tiles, and other features. The recent HealthCare.gov debacle has also demonstrated what happens to user engagement when a web site is unacceptably slow.

Web site speed is impacted by many factors. A web application that zips along with half a dozen users may slow with a couple of dozen users and crash completely with hundreds. Web site traffic (the number of simultaneous users and number of requests they make) is often unpredictable and "lumpy" – it varies a great deal at different times of day and different conditions. The architecture and implementation practices can have a significant impact as well - a web site that uses caching and a content delivery network is likely to be more responsive than one that does not. Network capacity can also have a significant impact on speed. Some functions, such as mapping, visualization and statistics, require far more computing power than serving up text and images. Further, as data sets become large, filtering and transforming them often becomes slower or requires special care. Finally, the design of the user interface can affect the user's perception of speed and responsiveness.

A generalized data collection platform for citizen science will encompass all of these concerns. Demand will vary a great deal based on time of year, time of day and press coverage. If it is successful, it will need to process very large data sets. If it supports sensor input, the number of data records will increase exponentially. We believe that such a platform should support visualization of the data with maps, graphs, basic statistics, and data analytics.

While it would be possible to develop a high performance infrastructure to support this type of capability, we believe leveraging contemporary cloud computing platforms will be a more sustainable approach that will support both scalability based on changing needs and lower cost than building a conventional data center.

This addendum to the report examines some of the leading cloud computing platform providers and tests processing performance with geospatial data, one of the more complex data types that will need to be leveraged in order to support mapping and spatial data processing.

What is cloud computing?

Cloud computing, according to the official NIST definition, 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." (http://www.nist.gov/itl/csd/cloud-102511.cfm) When an application is run "in the cloud", it is merely saying that the hardware and network level infrastructure is provisioned as part of an independently managed and dynamically allocated set of hardware. This infra-

structure can be managed on-site by the same organization that deploys the applications running on top of it, a "private cloud". Or the physical hardware and networking infrastructure can be managed company that makes the service available to customers to rent, a "public cloud". This report focuses on public cloud offerings that include data storage options, as well as a range of computational and networking power.

IaaS, PaaS, and SaaS

Cloud computing universe is often organized into three different types, including: infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS) and software-as-a-service (SaaS). This is a bit confusing as they all involve infrastructure, platforms and software. However, IaaS generally refers to resources – computing, storage, and networks - customers can rent to run network applications. Examples include Rackspace, Google Compute Engine, Amazon EC2 or Linode. "Platform" or PaaS offerings provide an integrated set of services - databases, web servers, development environments, auto-scaling, caching and others – that are built on top of PaaS components. Examples include Amazon Web Services, Google App Engine, and Heroku. Finally Software as a Service (SaaS) refers to actual online applications sold on a subscription or rental basis. SaaS applications are often based on PaaS or IaaS services. Examples include Gmail, Office 365, Salesforce.com, or, in the citizen science context, Crowdcrafting.

Cloud Service Concepts

Servers

Servers are rented from a public cloud service provider on an hourly basis, although in some cases servers can be reserved for more time for a discount. The servers appear to be completely isolated machines, as if you were renting a physical computer that you could log into, install software on, etc; however in reality the servers exist as one of many virtual machines running on very powerful hardware in data centers owned by the cloud service provider. Virtual machines enable one physical computer to operate several "virtual" computers (the virtual machines), with each of the virtual machines sharing the hardware resources of the host. So when you are renting a server from a cloud service with 4 CPUs, 8 GB of memory, and 800 GB of storage space, you are really renting a virtual machine that is provisioned from its host machine 4 CPU's worth of processing capacity, 8 GB of the host's memory, and a 800 GB section of the host's storage capacity. Externally, however, it seems as though the server is a completely isolated and independent machine, capable of running any software that a physical device with similar specifications could.

Object Store

Public cloud service providers often provide a storage mechanism that allows for redundant storage of high volumes of data to be accessed by servers. The redundancy of storage happens when your data is copied to multiple data centers in different geographic locations, which allows for one whole datacenter to fail and you data to remain safe and backed up. It is called 'object storage' because it is unlike a normal file system on a regular hard drive. The files, or objects, are accessed as if they were web pages - through an HTTP interface - unlike a regular file system. This enables access to the data from anywhere in the world.

Block Storage

Block storage devices act as though they were physical devices that can be hooked up to a server as an external hard drive would. They can be formatted to any file system, and can provide fast access to large amounts of data. An instance of a virtual server is usually provisioned with some storage, but the storage that is provided only lives as long as the server does, so if you "turn it off", you can lose the contents of the storage. Block storage services can be used to save files and data from server instances past the lifetime of servers. If a block storage volume is associated with the virtual server instance, the data on that volume can be saved after the server is deleted, and can be attached to another server in the future so that the data may be re-used. Block storage generally provides quicker access to data than object storage, but is somewhat less redundant.

Load Balancer

Load balancing is a technique to distribute workload across two or more servers. It can aid greatly in the number of requests that your application can handle. It also can prevent outages, because if a range of servers are serving requests, and one fails, the others can still service the requests. Load balancer offerings are unlike server offerings in the sense that they do not support the installation of any software – they only route network traffic.

Topology

A computer network topology is a description of the arrangement of systems in a computer network. When an application runs on a cloud platform, all of the servers, load balancers, and data stores used to serve that application are part of the cloud topology. The simplest topology is a single server. A more complex topology might include various application servers running in different data centers, that are accessed through an array of load balancers, which use databases run on different servers, which are backed up by different object stores. A good cloud topology design enables applications to scale well, to be resilient, and get the most performance out of the application for the price.

Other Services

Some platform companies provide many more services. We did not test them in our benchmarking, but there merit mention as they can be important differentiators between different services:

- Managed Databases
- Logging
- Caching: Files stored in memory can be served much faster than files stored on a disk drive. 1GB of memory storage is more expensive than 1 GB of disk storage, but the improvement in speed is sometimes worth the additional cost.
- Content Delivery Network (CDN): A CDN pushes frequently used files "closer" to the end user by storing them in more locations distributed around the world. The both speed up the user's perception of speed and reduce traffic on the server.
- Email and Messaging: Some providers incorporate email, SMS and phone call routing.
- Monitoring and Auto-Scaling: Some cloud services include the ability to monitor the traffic and workload of an application and when the traffic rises, to automatically add new infrastructure components (servers, storage, etc.). Conversely, when traffic falls, capacity can be reduced in order to save money.
- Machine Learning, Statistics and Prediction: Some cloud services offer specialized statistics and machine learning services that support predictive analytics.
- Analytics: Services for tracking and analyzing application usage.
- Security: Isolated virtual environments for processing sensitive or high security or data.
- Payment: Credit card payment services for e-commerce.

AWS

Amazon Web Services was launched in 2002, and steadily expanded their offerings to include a comprehensive suite of services. It is now the leader in cloud computing services and is both one of the best known and broadly used platform and infrastructure providers. Their cloud products are vast, with many server and storage options, as well as load balancers, monitoring, managed databases and more. They operate data centers located on almost every continent and are well-positioned to provide physical presence in locations, like the EU, that may require data to be processed within their jurisdiction. Large, high traffic sites such as Netflix, Reddit, Pinterest, and many more rely on AWS to run their products.

Like many providers, AWS has had several highly publicized outages, including one in December 2012 that caused many sites to not function for a large part of the US. [http://aws.amazon.com/message/680587/] In March of 2011, Reddit became unavailable for periods of the day due to a failure in the EBS product in one of Amazon's availability zones (one physical data center) [http://blog.reddit.com/2011/03/why-reddit-was-down-for-6-of-last-24.html].

Despite these outages, AWS has very consistently kept its large network of services running at a high level of availability. Further, the reliability is generally higher than can be expected with conventional hardware. As Miles Ward, a Solutions Architect at Amazon, puts it, EBS volumes "are made of real world things and as a result they are subject to potential failure." [http://youtu.be/vXkBvuAM7T4] EBS volumes " that operate with 20 GB or less of modified data since their most recent Amazon EBS snapshot can expect an annual failure rate (AFR) of between 0.1% – 0.5%, where failure refers to a complete loss of the volume. This compares with commodity hard disks that will typically fail with an AFR of around 4%, making EBS volumes 10 times more reliable than typical commodity disk drives." Failure is inevitabe in complex, distributed systems, and the best way to mitigate them it is to design systems that are fault tolerant at the application layer. In Reddit's postmortem blog post about its outages, they accept responsibility for not distributing data across availability zones, or using RAID configuration on some servers. This is to say that, although Amazon's outages have received a lot of public attention, the fact that so many high profile companies rely on AWSservices to such an extent that a few hours of issues generates a lot of press coverage suggests that outages are the exceptions that prove the rule.

OpenStack, HP and Rackspace

OpenStack is not a cloud service provider per se but a set of open source software tools that are used to manage a cloud hardware infrastructure. The open source nature of the project is a significant differentiation from Amazon Web Services. The OpenStack suite can be run on a variety of hardware, including private and public clouds, and there are a number of commercial providers that use it to manage their public cloud offerings, including: Cloudwatt, DreamCompute, eNocloud, HP, and Rackspace. The software was developed primarily by Rackspace, but it also has contributions from NASA software components developed to run their private data centers.

One argument in favor of OpenStack is the lack of vendor lock-in'. Theoretically, because Open Stack is open source software that is implemented on a number of public cloud providers, one could move cloud service providers without changing much of their cloud deployment and management tooling. For example, if you have a script that manages a set of servers on a Rackspace deployment, then if you were to choose to move your application to HP Cloud, then there would be minimal change to that script, and it should just work with the new Open Stack deployment. However point is weak on two fronts. First, switching providers will rarely be seamless and will likely require surgery to the management scripts with any change of provider.

Server names and service offerings will be different, so a script that spins up servers on Rackspace will rarely survive a move to HP intact. In general, the idea that moving an application from one cloud provider to another without making changes may not be a good idea; mapping the topology from one deployment to the other without considering the differences between the servers, storage, load balancers, etc. is a sure way to introduce performance issues. The second weakness of the "no vendor lock-in" argument is the growth of tools that support all of the major cloud platforms. For example, jclouds, an open source Java toolkit, now supports Amazon, GoGrid, Ninefold, vCloud, OpenStack and Azure. These concerns do not obviate the advantage of an open source platform. Staff skills, architecture experience and other considerations will transfer more seamlessly when the basic platform remains consistent. But the fact remains that whjen a project or a product makes a commitment to a cloud platform, it is expensive and disruptive to shift from that initial decision. The initial decision has some significant consequences.

NASA currently uses both OpenStack and AWS. NASA CIO Linda Cureton responded to questions about why NASA would not only use OpenStack, since it is software that they helped create and are contributing to, by saying "our computing strategy for cloud now, is to choose the right cloud strategy for the right environment. Everything is not OpenStack. OpenStack, I think, has a great value when you look at features that aren't necessarily there in the commercial providers. For example, if you have some very specific security requirements or if you are dealing with extremely large datasets, looking at a solution like OpenStack starts to tip the scale to that. But your run of the mill, general purpose type web servicing, a lot of commercial providers have been doing that for a while, and they do that

very effectively and so OpenStack needs to compete with that, but I think the big value for OpenStack is with their customized needs."

OpenStack seems to be a good choice if you need a powerful open source platform for running a private cloud; for example NASA and CERN are using OpenStack at their private data centers. However, as mentioned above, there are companies who use OpenStack to manage powerful public clouds. In particular HP and Rackspace have offerings that compare with AWS: similar server configurations that can be dynamically resourced by the hour through an API or web console, block storage devices that can be attached to these servers, and object store services for highly redundant storage of large amounts of data.

Platform Comparison

The following section outlines some rough cost and performance comparisons between three different services: Amazon, HP and Rackspace. Neither the cost nor the performance comparisons are ideal - the offerings from the providers are difference, and an apples-to-apples comparison is inherently difficult. The purpose of the comparison is aimed at suggesting the best platform for development and deployment of a cloud-based citizen science platform that will be based on open source software, will scale based on different levels of use and will be able to provide robust data processing and visualization capabilities.

Cost Comparison

The following is a comparison of server offerings and price from Amazon Web Services, Rackspace, HP, Google Compute Engine and Linode. The table compares the servers by RAM, number of CPU cores, and the hourly price of the server as well as listing storage capacity.

Amazon

The Amazon pricing is based on the following assumptions:

- On Demand – reserve and spot pricing is available but this is the hourly pricing
- US East region
- Linux
- EC2 Compute Unit (ECU) – One EC2 Compute Unit (ECU) provides the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor
- Instance memory specifications in GibiBytes (GiB), not Gigabytes (GB), where 1 GiB = 1.024 GB

Instance	CPU Cores	ECU	Memory (GB)	Storage	Cost/Hr	Price/CPU Core/GB
Micro Instances						
t1.micro	1		0.615	EBS Only	$0.020	0.0325
General Purpose - Current CPU						
m3.xlarge	4	13	15	2 x 40 SSD	$0.450	0.0075
m3.2xlarge	8	26	30	2 x 80 SSD	$0.900	0.0038
General Purpose - Previous CPU						
m1.small	1	1	1.7	1 x 160	$0.060	0.0353
m1.medium	1	2	3.75	1 x 410	$0.120	0.0320
m1.large	2	4	7.5	2 x 420	$0.240	0.0160
m1.xlarge	4	8	15	4 x 420	$0.480	0.0080
Compute Optimized - Current CPU						
c3.large	2	7	3.75	2 x 16 SSD	$0.150	0.0200
c3.xlarge	4	14	7.5	2 x 40 SSD	$0.300	0.0100
c3.2xlarge	8	28	15	2 x 80 SSD	$0.600	0.0050
c3.4xlarge	16	55	30	2 x 160 SSD	$1.200	0.0025
c3.8xlarge	2	108	60	2 x 320 SSD	$2.400	0.0200
Compute Optimized - Previous CPU						
c1.medium	2	5	1.7	1 x 350	$0.145	0.0426
c1.xlarge	8	20	7	4 x 420	$0.580	0.0104
cc2.8xlarge	32	88	60.5	4 x 840	$2.400	0.0012
GPU Instances - Current GPU						
g2.2xlarge	8	26	15	60 SSD	$0.650	0.0054
GPU Instances - Previous GPU						
cg1.4xlarge	16	33.5	22.5	2 x 840	$2.100	0.0058
Memory Optimized - Current CPU						
m2.xlarge	2	6.5	17.1	1 x 420	$0.410	0.0120
m2.2xlarge	4	13	34.2	1 x 850	$0.820	0.0060
m2.4xlarge	8	26	68.4	2 x 840	$1.640	0.0030
cr1.8xlarge	32	88	244	2 x 120 SSD	$3.500	0.0004
Storage Optimized - Current CPU						
i2.xlarge	4	14	30.5	1 x 800 SSD	$0.853	0.0070
i2.2xlarge	8	27	61	2 x 800 SSD	$1.705	0.0035
i2.4xlarge	16	53	122	4 x 800 SSD	$3.410	0.0017
i2.8xlarge	32	104	244	8 x 800 SSD	$6.820	0.0009
hs1.8xlarge	16	35	117	24 x 2048	$4.600	0.0025
Storage Optimized - Previous CPU						
hi1.4xlarge	16	35	60.5	2x1024 SSD	$3.100	0.0032

HP

The HP pricing is based on the following assumptions:

- On Demand
- Linux
- An HP Cloud Compute Unit (CCU) is a unit of CPU capacity that describes the amount of compute power that a virtual core has available to it. 6.5 CCUs are roughly equivalent to the minimum power of one logical core (a hardware hyper-thread) of an Intel(R) 2012 Xeon(R) 2.60 GHz CPU

Instance	CPU Cores	HP CCU	Memory (GB)	Storage	Cost/Hr	Price/CPU Core/GB
Standard Instance						
Standard Extra Small	1	1	1	20GB	$0.030	0.0300
Standard Small	2	2	2	40GB	$0.060	0.0150
Standard Medium	2	4	4	80GB	$0.120	0.0150
Standard Large	4	8	8	160GB	$0.240	0.0075
Standard XL	4	15	15	300GB	$0.450	0.0075
Standard 2XL	8	30	30	570GB	$0.900	0.0038
Standard 4XL	12	60	60	900GB	$1.800	0.0025
Standard 8XL	16	103	120	1800GB	$3.240	0.0017
High Memory Instance						
High Memory Large	4	8	16	160GB	$0.360	0.0056
High Memory XL	4	15	30	300GB	$0.680	0.0057
High Memory 2XL	4	30	60	570GB	$1.350	0.0056

Rackspace

The Rackspace pricing is based on the following assumptions:

- On Demand
- Infrastructure Service Level
- Linux
- SSD disks

Instance	CPU Cores	N/A	Memory (GB)	Storage	Cost/Hr	Price/CPU Core/GB
1GB Performance	1		1	20GB SSD	$0.040	0.0400
2GB Performance	2		2	60GB SSD	$0.080	0.0200
4GB Performance	4		4	80GB SSD	$0.160	0.0100
8GB Performance	8		8	120GB SSD	$0.320	0.0050
15GB Performance	4		15	190GB SSD	$0.768	0.0128
30GB Perforamnce	8		30	340GB SSD	$1.360	0.0057
60GB Performance	16		60	640GB SSD	$2.720	0.0028
90GB Performance	24		90	940GB SSD	$4.080	0.0019
120GB Performance	32		120	1240GB SSD	$5.440	0.0014

Google

The Google pricing is based on the following assumptions:

- On Demand
- North America
- Linux
- No disks included, so storage pricing is separate

Instance	CPU Cores	HP CCU	Memory (GB)	Storage	Cost/Hr	Price/CPU Core/GB
Shared Core						
f1-micro	1		0.6	None	$0.019	0.0317
g1-small	1		1.7	None	$0.054	0.0318
Standard						
n1-standard-1	1		3.75	None	$0.104	0.0277
n1-standard-2	2		7.8	None	$0.207	0.0133
n1-standard-4	4		15	None	$0.415	0.0069
n1-standard-8	8		30	None	$0.829	0.0035
n1-standard-16	16		60	None	$1.659	0.0017
High Memory						
n1-highmem-2	2		13	None	$0.244	0.0094
n1-highmem-4	4		26	None	$0.488	0.0047
n1-highmem-8	8		52	None	$0.975	0.0023
n1-highmem-16	16		104	None	$1.951	0.0012
High CPU						
n1-highcpu-2	2		1.5		$0.131	0.0437
n1-highcpu-4	4		3.6		$0.261	0.0181
n1-highcpu-8	8		7.2		$0.522	0.0091
n1-highcpu-16	16		14.4		$1.044	0.0045

Linode

The Linode pricing is based on the following assumptions:

- On Demand
- North America
- Linux

Instance	CPU Cores	N/A	Memory (GB)	Storage	Cost/Hr	Price/CPU Core/GB
Linode 1024	1		1	48GB	$0.278	0.2780
Linode 2048	2		2	96GB	$0.056	0.0140
Linode 4096	4		4	192GB	$0.111	0.0069
Linode 8192	8		8	348GB	$0.222	0.0035
Linode 16384	16		16	768GB	$0.444	0.0017
Linode 24576	24		24	1152GB	$0.667	0.0012
Linode 32768	32		32	1536GB	$0.889	0.0009
Linode 40960	40		40	1920GB	$1.111	0.0007

Cost Comparison Summary

While no offering matches Amazon's range of products, where there are comparable server specifications, there is almost no difference in the pricing between the different offerings. At this point in time, Amazon is clearly setting the pace on pricing and other competitive vendors are matching that pricing. Vendors are therefore engaging in non-price competition. Rackspace, for example, offers Solid State Drives (SSDs) for all of its servers as well as specific commitments to I/O operations per second (IOPS). Linode offers simple, all-inclusive pricing (which includes storage and network traffic) as well as multiple data centers around the world.

Performance

Methodology

The performance benchmarking of the different platforms was performed in February 2013 and focused on testing Amazon Web Services, HP and Rackspace. To compare performance between the similar offerings from Rackspace, HP, and Amazon servers, we executed two performance tests.

The first test runs a simple Jetty web service on each of the servers, and runs an Apache Bench benchmark, running locally, against that service. The web service has two endpoints, one performs a Hillshade operation against a raster data structure in a parallel fashion (taking advantage of all the cores of the machine); and the other is a simple weighted overlay operation that combines two raster data sets. In each case the raster size is 256x256pixels. The second test focuses iterating through array and raster data storing as either integers or double numeric data types. Google Caliper was used to run the benchmark tests for each machine instance.

The source code for the benchmarks is available at: https://github.com/lossyrob/cloud-benchmark

Benchmarking Limitations

The performance benchmarking results should be taken with a grain of salt. As outlined above, these servers are virtual machines, and do not run in isolation. Rather, they run on a host computer along with a variety of instances provisioned to other customers; the overall load on the host computer can affect the performance of a specific instance. Therefore, the results from a test done during a time of relatively low network activity in the host data center can be very different from running test during a period of time where the network and servers are under heavy load. There are additional sources of variance, such as operating system processes,, garbage collection software platform scheduling, system time logging, benchmarking software variance, to name a few. Benchmarking data should be understood as an approximation of performance. While this may appear to suggest that benchmark data should not be trusted, the various issues can be largely be overcome by running the same benchmarks multiple times, having a clear understanding of the limitations, and focusing on general trends in the data rather than the specific results of a single run.

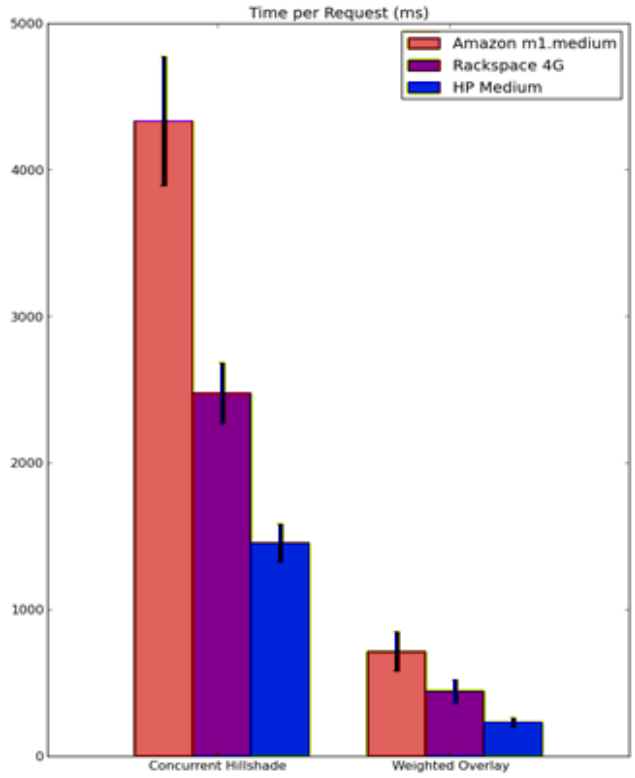
Medium Servers

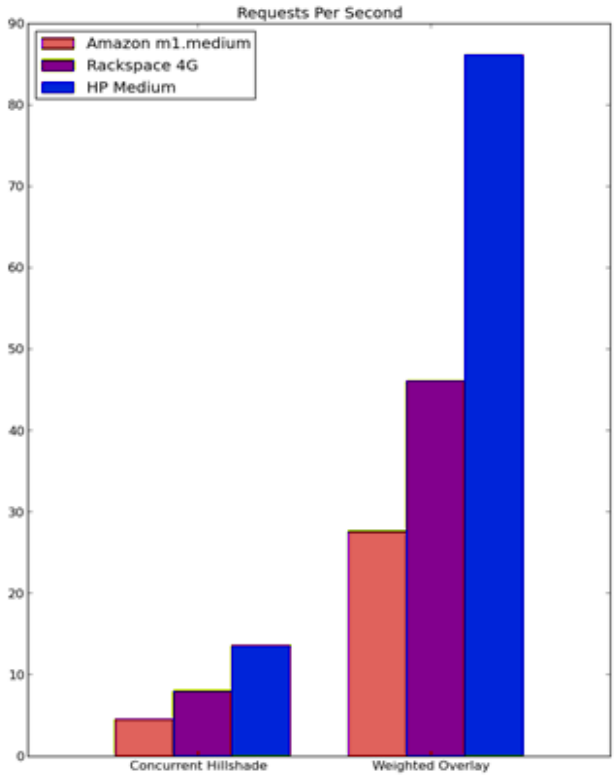
The first set of tests were run against medium sized servers. The machines used were as follows:

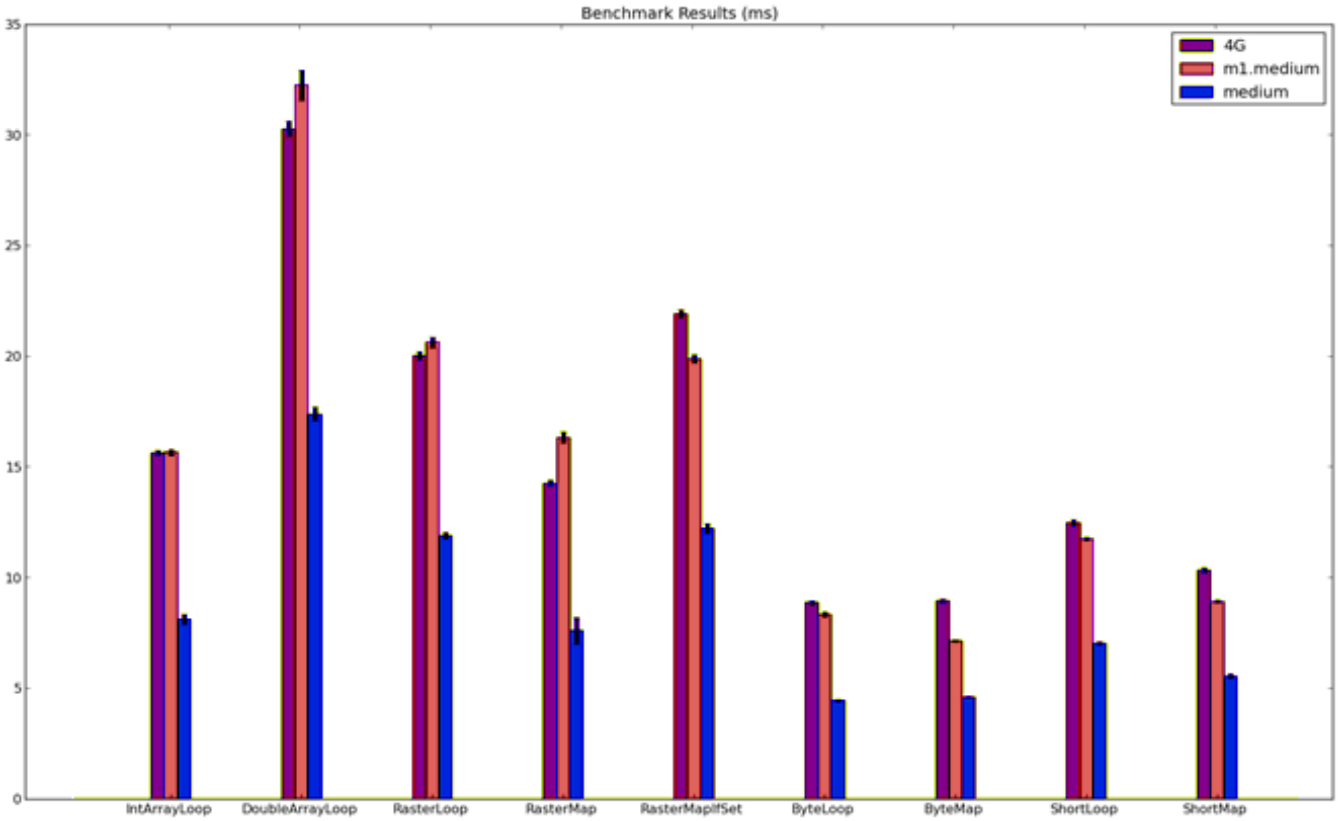
- Amazon m1.medium, 1CPU core (2 ECUs), 3.75GB RAM
- Rackspace 4GB, 4 CPU cores, 4 GB RAM, SSD disks
- HP Medium, 2 CPU cores (4 HP CCUs), 4GB RAM

The servers are not entirely equivalent; the Amazon servers have fewer CPU cores and slightly less memory; further, they are an older CPU model that is now being retired by Amazon. Finally, the Rackspace servers have SSD disks, whereas the Amazon and HP servers have conventional hard drives. However, within these limitations, the results suggest that HP servers running OpenStack have better performance than either Rackspace or Amazon for both time per request and requests per second. The graphs appear below.

In the second test the HP machines again showed better performance than both Rackspace and Amazon. However, there was almost no difference between Rackspace and Amazon machines performance.


Time per Request (ms)
Amazon m1.medium
Rackspace 4G
HP Medium
5000
4000
3000
2000
1000
0
Concurrent Hillshade
Weighted Overlay


Requests Per Second
Amazon m1.medium
Rackspace 4G
HP Medium
90
80
70
60
50
40
30
20
10
0
Concurrent Hillshade
Weighted Overlay


Benchmark Results (ms)
4G
m1.medium
medium
35
30
25
20
15
10
5
0
IntArrayLoop
DoubleArrayLoop
RasterLoop
RasterMap
RasterMapIfSet
ByteLoop
ByteMap
ShortLoop
ShortMap

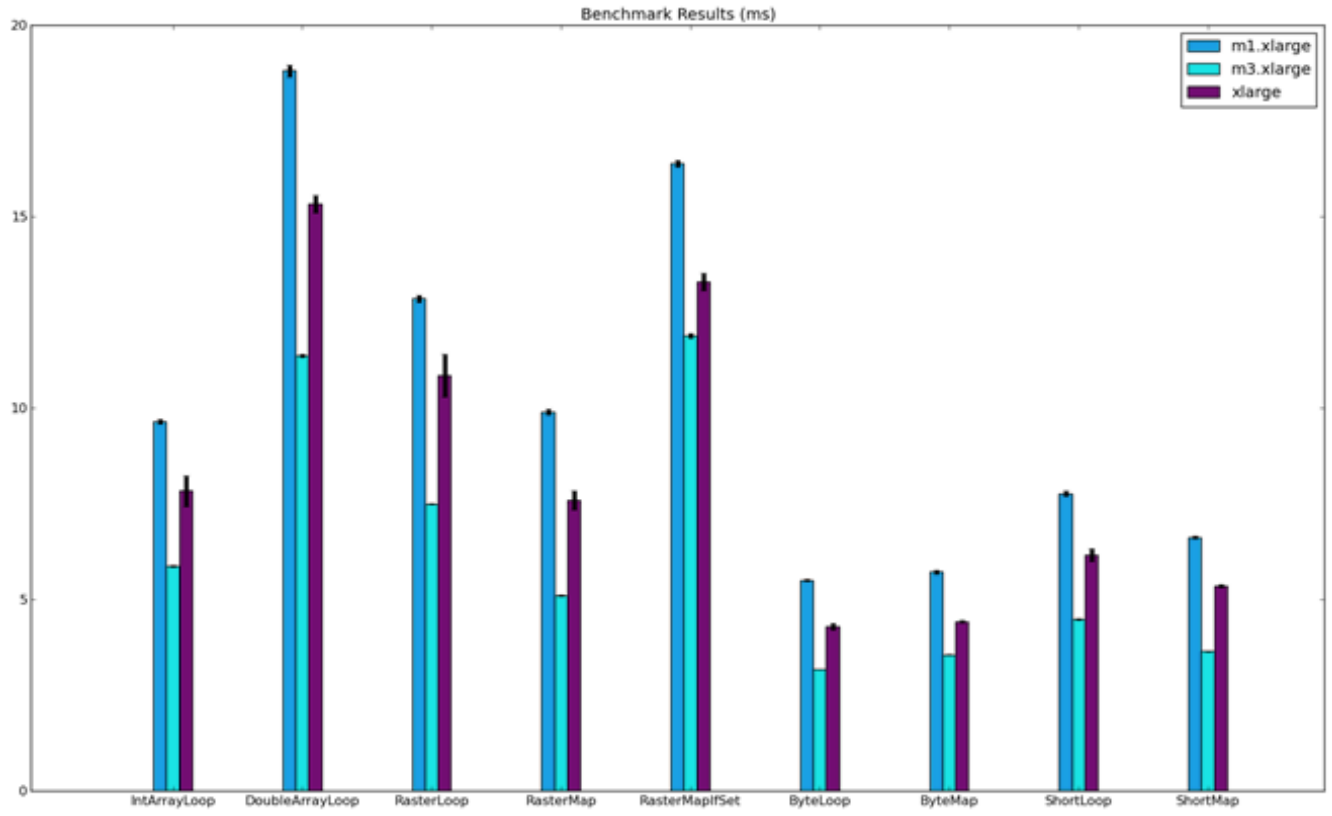
Extra Large Servers

The second set of tests were run against extra-large servers. The machines used were as follows:

- Amazon m1.xlarge, 4CPU cores (8 ECUs), 15GB RAM
- Amazon m3.xlarge, 4 CPU cores (13 ECUs), 15 GB RAM
- HP Extra Large, 4 CPU cores (15 HP CCUs), 15GB RAM

Again servers are not entirely equivalent; the Amazon m1.xlarge is an older CPU model and would be expected to perform significantly less well than the current CPU models. However, the Amazon m3.xlarge and HP Extra Large are very similar. Nonetheless, the HP servers were again measurably faster on both parts of the first test. Graphs summarizing the results appear below.

However, the second set of tests turned the tables and the Amazon m3.xlarge came out on top. The graphs appear below:



Amazon Server Size Comparison

As described above, Amazon groups its server instances into families, each with their own characteristics, strengths, and weaknesses. The amazon EC2 is a rich source of information about the specific instance types, but they are briefly summarized here:

Micro – There is currently only one type of micro instance available. It has no instance storage, and little CPU power, that is described as Amazon as "bursty". Micro instances are used when small increments of cheap CPU resources are needed. One example might be a server running diagnostics and responsible for storing and forwarding status messages from an application. Another example would be a periodic data aggregation task. Logging data would be copied to an EBS volume and a set of micro instances with EBS volumes would be started to run the job and then stopped when it's complete.

Standard – These instances are good general purpose servers, with CPU power balanced with memory resources. The other families of servers are for specialized requirements, like high CPU, high memory or other performance characteristics.

High-CPU and High-Memory – These instances provide a different CPU-to-memory ratio than the standard instances. High CPU instances are useful for tasks that require a lot more computational power relative to the amount of data; high memory instances are useful for caching servers or other tasks that require a large amount of data or files to be stored in memory.

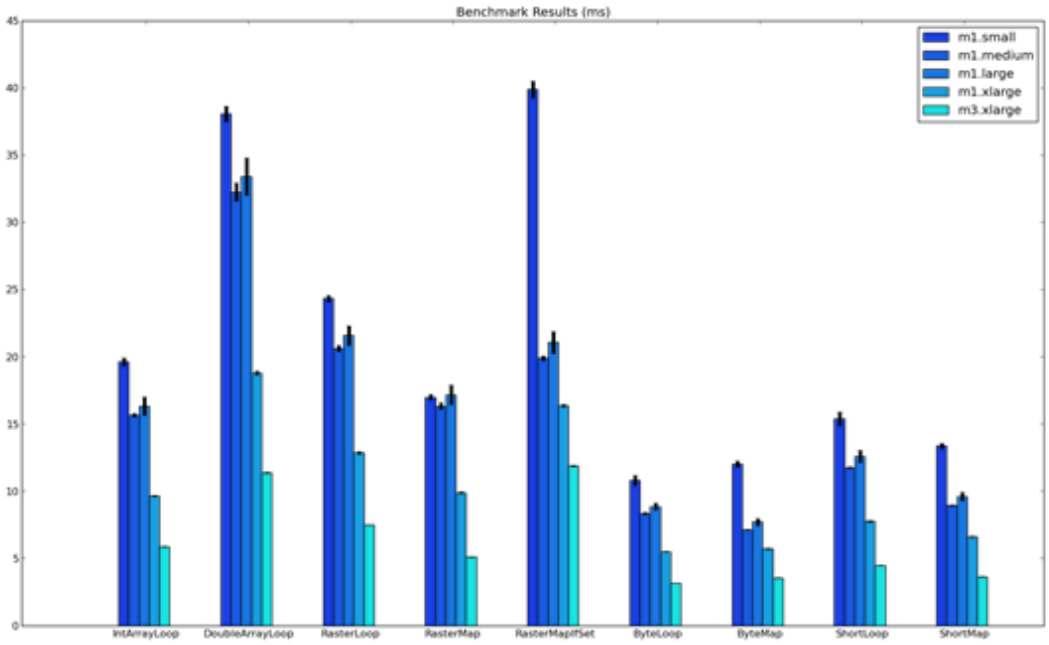
High I/O – These instances are equipped with Solid State Drives for storage and 10 Gbps network bandwidth to maximize data transfer speeds.

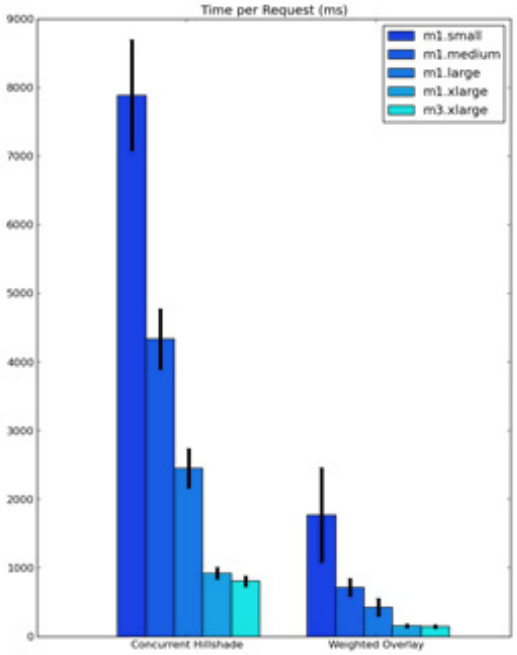
Cluster Compute – Cluster instances have a large amount of computational power and increased network bandwidth. They are most useful in High Performance Computing (HPC) applications, where large amounts of data and/or computationally expensive algorithms are distributed across a cluster of instances. There are also specialized instances with graphical processing units (GPUs) available for doing HPC on media, or for taking advantage of GPU frameworks such as CUDA or OpenCL.

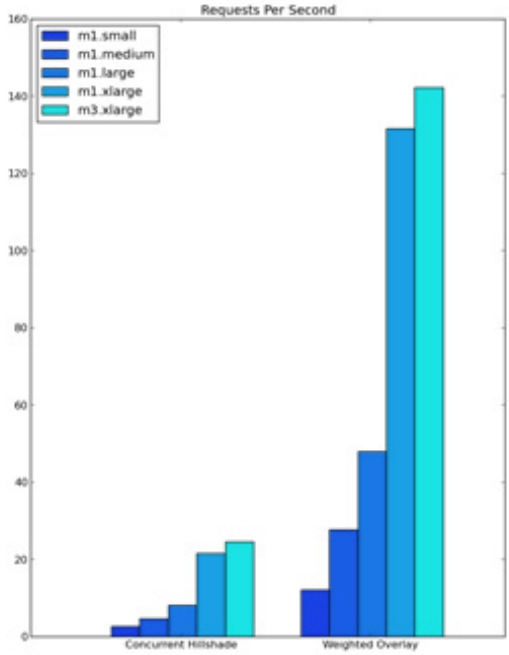
High Storage – High storage instances give an enormous amount of instance storage (48 TB), as well as increased sequential read and write speeds from the storage. In addition, with 117 GB of memory and 10 Gbps network throughput, this instance is designed to support very large data sets, though it only has 16 CPUs to work on that data.

Performance Comparison

The following benchmark results reflect the two test sets and compare first generation CPU (now obsolete) standard instance offerings, as well as one second generation instance (the M3 Extra Large Instance).







Amazon Storage Comparison

Amazon offers several types of data storage and each type has performance implications. The following describes the different types and runs the benchmarks against them.

Instance Storage

Instance storage refers to storage from disks that are physical connected to the computer hosting the instance as a virtual machine. Each instance is bundled with a specific amount of provisioned storage, dedicated to the instance. The physical disk that provides the instance storage is shared among instances of the host computer. One important difference between instance storage and the other storage types is that the storage will only last as long as the server is running. Once a server is terminated, either due to failure or by the user, all data on the instance storage is lost. Generally, any data that needs to be saved should be stored as block storage or another location.

Elastic Block Storage (EBS)

Elastic Block Storage (EBS) is a storage service that can be attached to server instances to provide permanent storage that can be provisioned separately from the actual storage instance. When an instance is terminated, the EBS volumes do not terminate with the instance, the data remains, and it can be accessed again by reattaching the storage volume to a new instance. EBS volumes are also replicated across multiple availability zones within a data center, improving the durability. This means a data center-wide EBS failure would be required to lose your data - not impossible, but fairly unlikely. Amazon's EBS offering also enables the owner to take "snapshots" of volumes, which can be used to create identical volumes at a later time. The snapshots are stored incrementally in a compressed format in S3. The S3 costs are much lower than EBS, and the owner is only charged for the amount of data that changes for subsequent snapshots. For example, if the owner creates a snapshot of a 100GB drive, and 1 GB of data changes, the snapshot will cost less than 1GB of S3 charges.

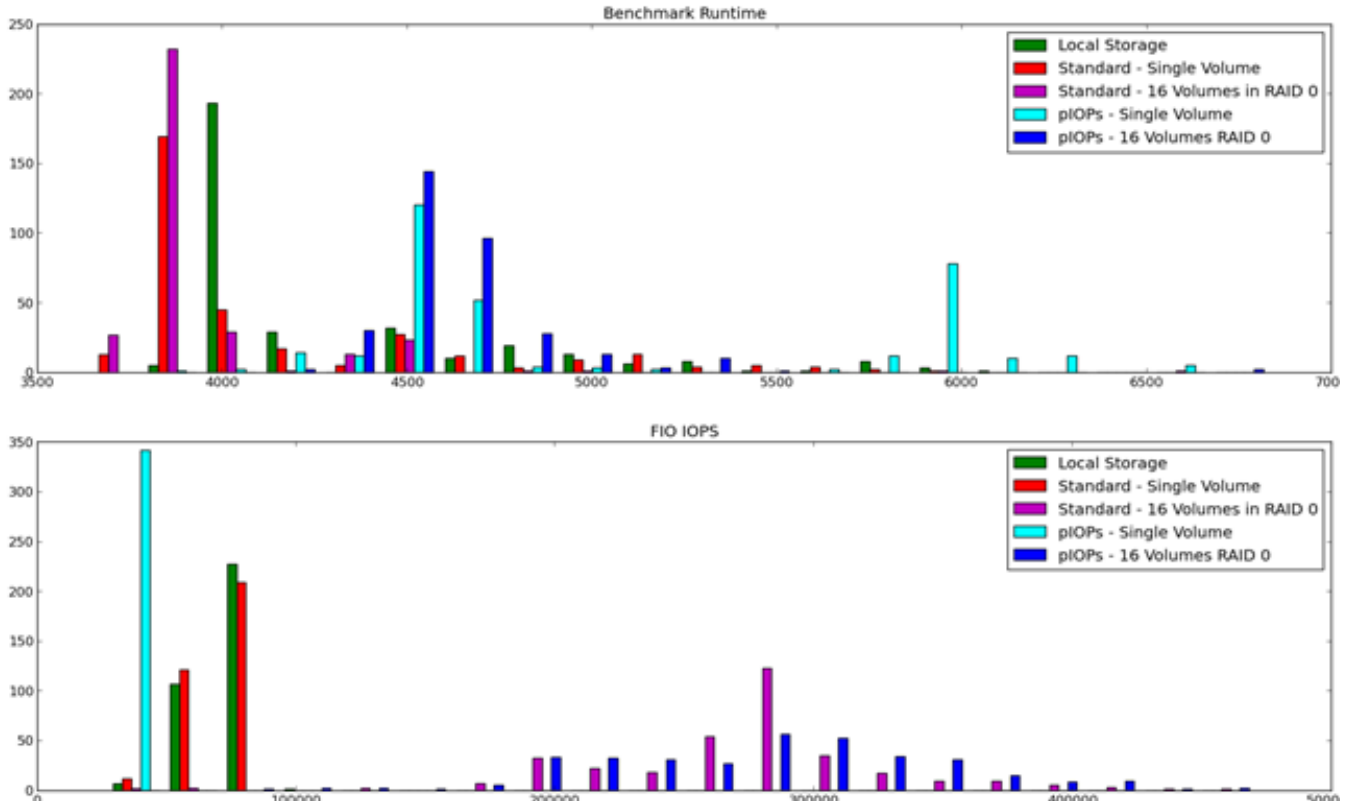
pIOPS EBS

Amazon EBS has an additional offering called "Provisioned IOPS" (pIOPS). IOPS refers to input/output (I/O) per second. Whereas standard EBS is a best-effort service, provisioned IOPS are guaranteed I/O speeds, for an extra fee. Each individual I/O transaction service time is metered. If the owner provisions 1000 IOPS for a volume, each individual I/O operation should take

1 ms. An I/O is defined at reading or writing one 16KB block. These numbers are important to understand in order to utilize pIOPS correctly. If your application does not fill the IO queue with 16 KB operations in a way that utilizes all of the provisioned I/Os per second, then you are wasting money; you still pay for the provisioned IOPs that you don't use.

Performance Comparison

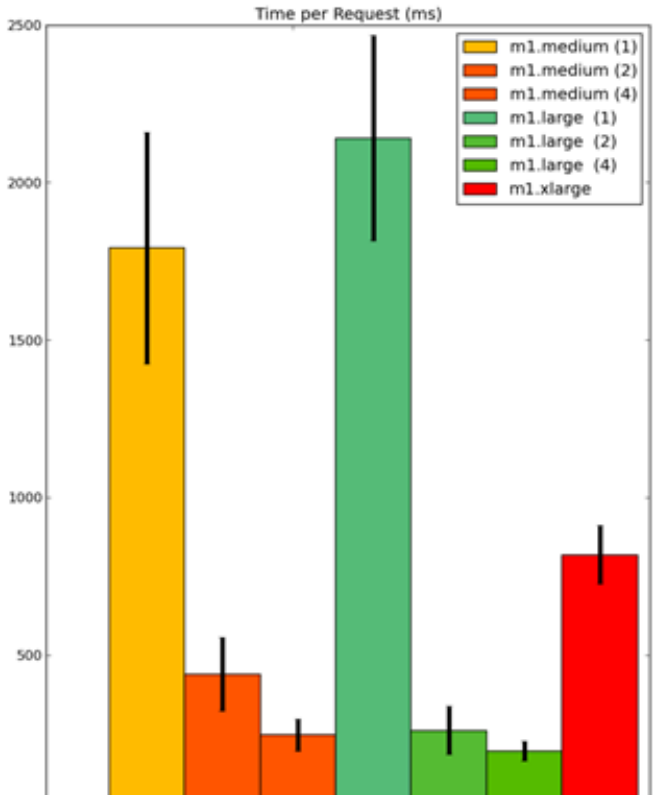
Each of the storage options has significant performance implications. The following graphs summarize the relative performance of different storage options.

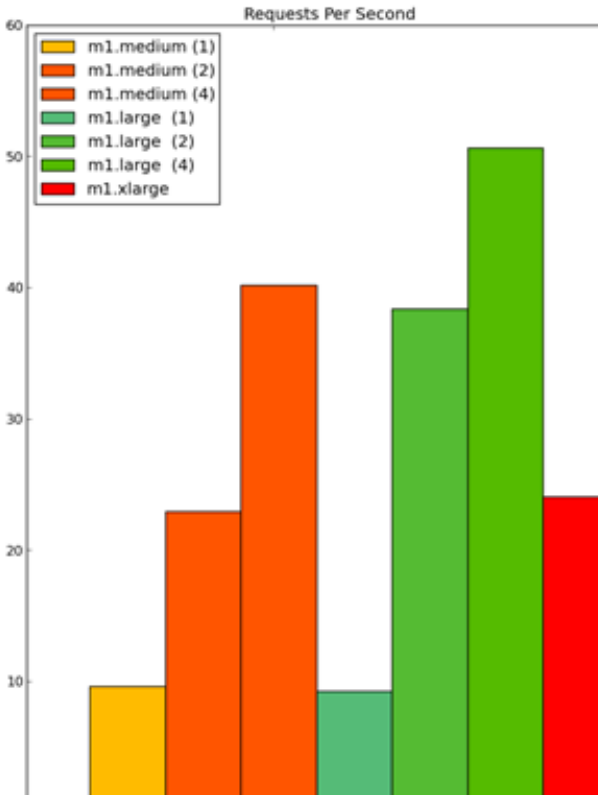


Amazon Load Balancer Comparison

Load balancers can help improve the performance and resilience of a cloud-based application. They can have a significant performance impact for large distributed cloud implementations.

The following are the results of a benchmark that measured performance of Amazon medium and large standard instances, alone and when using a load balancer in front of two and four instances. The service request returns a 256x256 PNG. The same test run on a single m1.xlarge instance is provided to give comparison. The results indicate that for certain applications, two standard medium instances behind a load balancer, while being cheaper than one standard extra-large instance, could be more performant.





Summary

Rackspace

Rackspace servers did moderately well and have comparable cost to HP and Amazon. Rackspace provides the equivalent of provisioned IOPS and solid state drives (SSD) on all of their servers. Rackspace touts its 'fanatical support'. Our experience in recent years has been that Rackspace staff work hard to live up to the tagline. However, our testing process highlighted some concerns. While small servers were slightly faster that similar Amazon servers, the Rackspace servers were much slower to start than either HP or Amazon EC2. In some instances they required hours rather than minutes and did not reliably report their status in the client dashboard. However, this issue aside, Rackspace has a broad offering that includes multiple storage offerings, monitoring, backup, email, load-balancing, DNS, backup, and queue services. On a feature comparison, Rackspace comes closest to the array of products that Amazon offers.

HP

The HP servers performed well at each size level. HP has an easy-to-use dashboard, and it was very easy to set up security groups and key pairs and get started creating instances. Instance creation was very fast, similar to Amazon EC2. Customer service, however, was conducted with solicitations made by automated phone calls and automated emails, and the contrast with Rackspace was significant. While HP offers compute, block storage, object storage, content delivery network, DNS and managed relational databases, it is also missing some critical infrastructure components, including load balancing, auto-scaling, monitoring, noSQL data stores, and distributed global data centers.

Amazon Web Services (AWS)

Amazon's cloud platform and infrastructure offerings were both the most comprehensive and easiest to manage. The offering is very mature and continues to grow and improve over time.

Amazon offers the widest range of services, including basic services such as servers (EC2), storage (S3 and EBS), and load

balancing, as well as managed databases, auto scaling, monitoring, and data processing services. Cloud platforms like Rackspace and HP offer subsets of what Amazon offers, so it is good to look to Amazon to see what other cloud platforms are going to mature into, and for cutting edge offerings. In the following section we will look more closely at Amazon's cloud offerings. Extensive training workshops, an annual conference, a partner program, and an enormous range of customers also mean that there is a large and growing ecosystem of experienced developers. This may, in fact, be the most significant argument for using Amazon Web Services. Finally, Amazon's management consoles, dashboards and billing systems were both easy to use and sophisticated.

Google Compute Engine

Google is obviously not new to cloud computing, but its Compute Engine offering has only recently been released. We did not test the performance of Google Compute Engine as it was not yet publicly available when we conducted our tests. The initial offering is not as broad as Rackspace or Amazon, but Google has enormous global reach and patience. The initial foray already offers both North America and EU data centers. The company has the capital, brand and patience to compete head-to-head with Amazon and Rackspace.

Recommendation

The HP Cloud scored well on performance benchmarks, but with a relatively narrow offering, we do not believe it offers the breadth of service that will be required for a scalable citizen science solution. Amazon and Rackspace offer the broadest set of services and currently have comparable pricing. Amazon's management tools, partners, customer ecosystem, and integrated products, however, weigh significantly in its favor.

Acknowledgements

This is Appendix B of a four part report. The full report includes the following:

- Part 1: Data Collection Platform Evaluation
- Part 2: Technology Evaluation
- Appendix A: Wireframe Designs
- Appendix B: Cloud Computing Performance Testing

This report was funded in part by a grant from the Alfred P. Sloan Foundation.

We respectfully acknowledge the individuals and projects that contributed to its development of this study; we could not have completed this document without their invaluable expertise and input. We are particularly grateful for the valuable insights provided by both the project representatives we interviewed and the input from the Citizen Science Community Forum.

These recommendations are meant to both celebrate their accomplishments and acknowledge the need for continued growth and improvement that will benefit everyone. To that end, we welcome additional comments from the citizen science community such that we may further refine this vision and move it forward toward reality.