Where is this cloud thing anyway?

By Peter van Eijk, independent IT consultant.

Introducing Cloud Performance.

CMG has a long tradition of adapting to new technologies. The cloud is no exception. Computing technology resources have always been finite and fallible, so sooner or later it becomes important to measure and understand uptime, performance, capacity and cost. Clouds are no exception.

MeasureIT introduces 'Cloud Performance', an occasional series dedicated to all things measurable in cloud computing. I will share my own results, intermixed with other contributions from within our community.

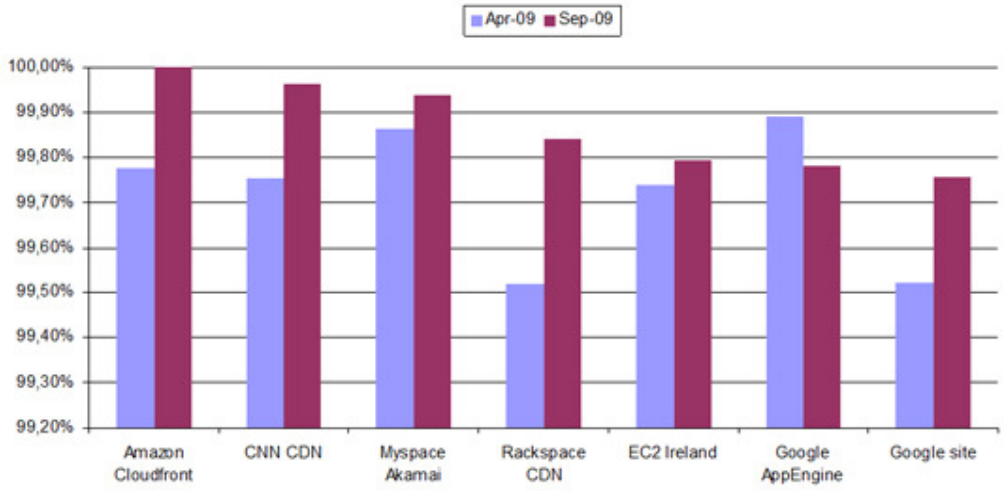
If you are considering a cloud infrastructure, your first questions should be:

- Does it exist?
- Does it scale?
- How does it grow?

This article reports on a number of measurements that we did in recent years. Using a global network of dozens of monitoring stations, we have measured worldwide uptime and proximity of a number of cloud providers.

Uptime

I looked at the uptime of the services of a number of cloud providers, such as Amazon, Akamai, and Google AppEngine. The results are in the following figure. What we see is that cloud uptime is pretty good, and it is getting better. As a matter of fact, these clouds perform better than a lot of commercial websites. Other data shows that APIs of popular websites have much worse uptimes.



Scaling and Growing

Another interesting question is: where is the cloud? Insight into cloud locations leads to insight in the robustness and scalability of cloud computing solutions. Cloud computing paradigms differ wildly, including self-service virtual machines (typically categorized under the heading 'Infrastructure as a Service'), programming platforms ('Platform as a Service'), and content distribution networks (CDN).

The promise of cloud computing is that it 'scales with your application demands' across a wide range of workload volumes. The traditional scaling model for webservices is to use loadbalancers in front of a farm of webservers. This is inherently rather local. In addition, such infrastructure is typically rather tied to a specific application. The next two models illustrate different approaches.

Google App Engine

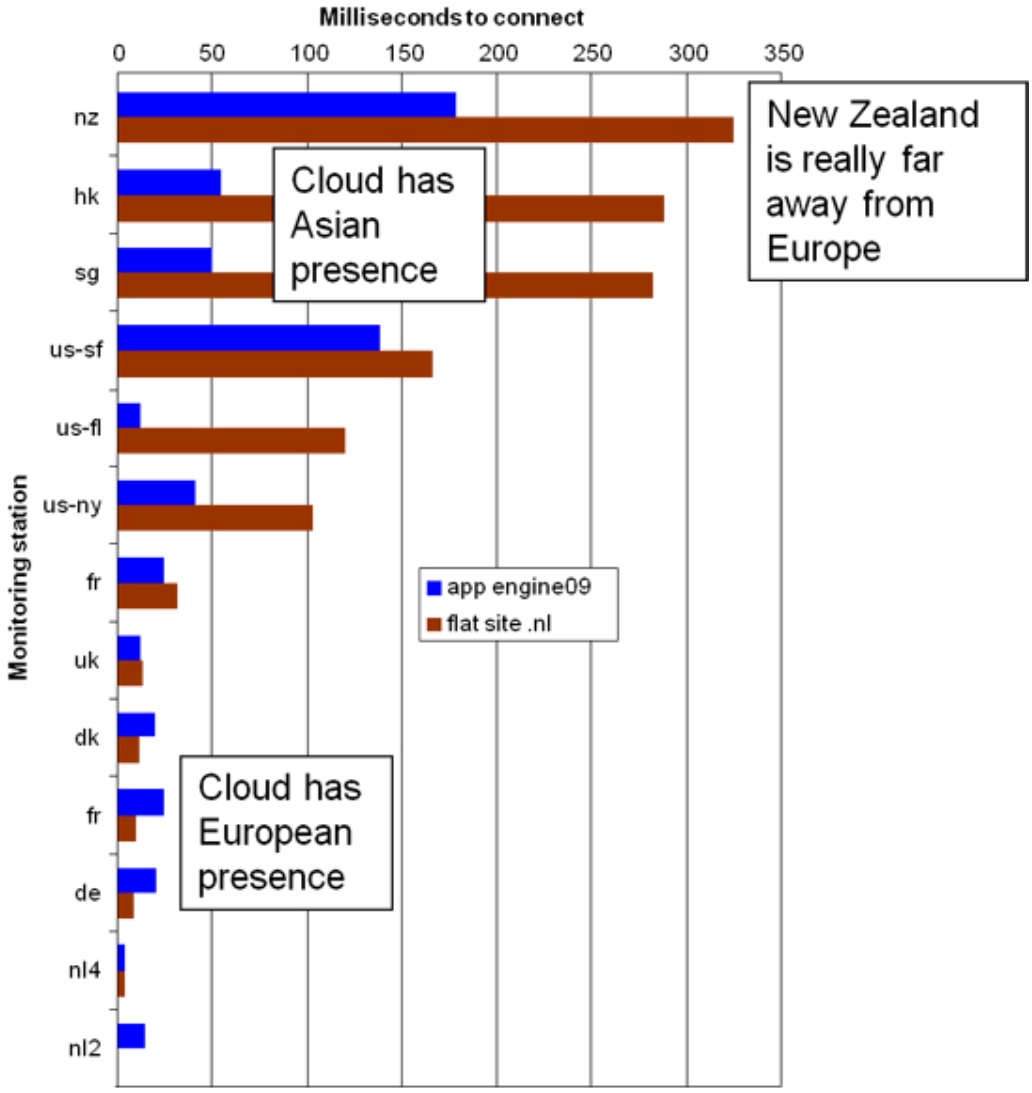
Google App Engine is a computing platform, also known as a Platform as a Service (PaaS). Application developers upload their code to this platform, where it gets executed in an unspecified location. There is a services based data storage notion, which can work in a distributed way.

We looked at the distance from the monitoring stations to a Google App Engine application. The following figure shows that the App Engine cloud is in fact spread out around the world, with data I collected in 2009.

The figure presents the time to connect to our Google App Engine application over HTTP, as well as the connect time to a regular website that is hosted in the Netherlands.

On the vertical axis we see a number of WatchMouse monitoring stations, so that each horizontal bar represents the connect time from that station to either a regular website in the Netherlands, or an application at Google Apps Engine. A few observations can be made from this data.

- New Zealand is really far away from Europe at over 300 milliseconds connect time, which equals the round trip delay. Note the speed of light is a large contributor to this delay
- The AppEngine cloud has a presence in Asia, as it is 50 milliseconds away from Hong Kong and Singapore.
- This cloud has a European presence, with less than 20 milliseconds round trip delay.
- Puzzling is the distance to San Francisco, as this is close to the headquarters of Google.
- From the graph it is clear that that cloud is in more than one place on the globe. Given the programming model of Google Apps Engine, this means that it is very scalable.
- On the average the App Engine cloud is nearer to a client than any web service that is hosted in a single location.



Our research also shows that this cloud is expanding.

Amazon CloudFront

Amazon CloudFront is a content distribution network (CDN). It redirects requests for static web content to the nearest CloudFront servers.

The summary is that CloudFront is on the average about 40-50 milliseconds away from a random point on the Internet. This is pretty good compared to a site located in e.g. New York (120 milliseconds) and is in the same league as other content distribution networks. In specific markets, it is pretty near: San Francisco: 3 milliseconds, New York: 13 milliseconds, Western Europe: 1-30 milliseconds.

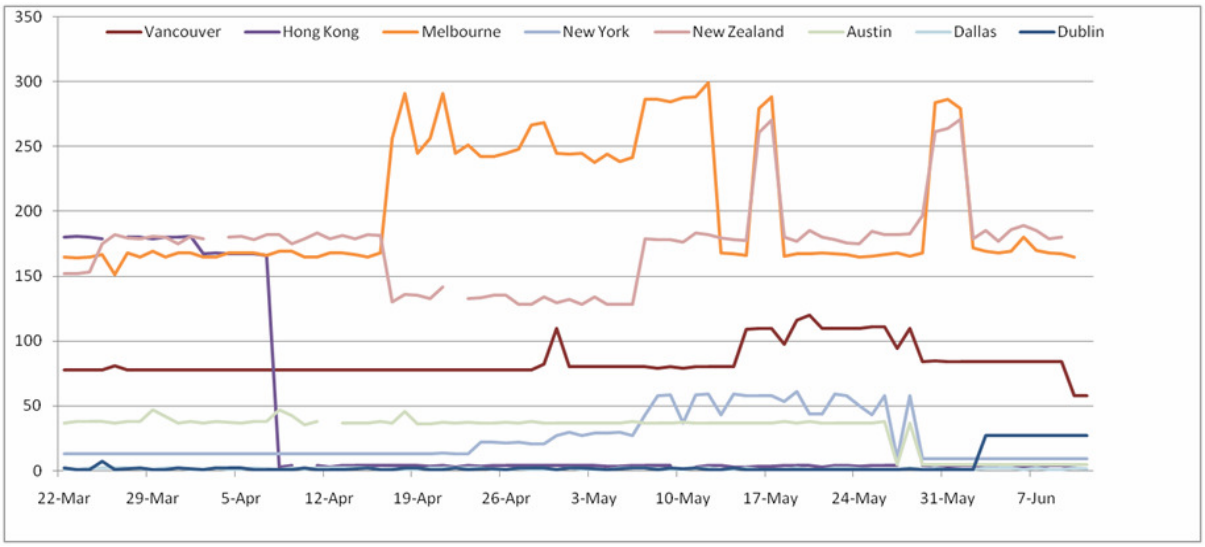
According to Amazon, CloudFront is in 16 locations, in contrast to the S3 storage service and the EC2 compute service, which have only 4 points of presence around the world (as of end 2010).

The following table gives distances (in milliseconds) from selected locations of the monitoring network to Amazon CloudFront (cities annotated with CF have CloudFront locations):

Distance	City	Country	CF
1	Amsterdam	Netherlands	CF
1	Ashburn	U.S.A.	CF
2	Santa Clara	U.S.A.	CF
2	Dallas	U.S.A.	CF
3	Hong Kong	China	CF
3	Singapore	Singapore	CF
4	Cologne	Germany	
5	Nagano	Japan	
7	Manchester	United Kingdom	
9	New York	U.S.A.	CF
11	Kuala Lumpur	Malaysia	
12	London	United Kingdom	CF
20	Padova	Italy	
27	Dublin	Ireland	CF
36	Bangkok	Thailand	
59	Mumbai	India	
75	Haifa	Israel	
154	Sydney	Australia	
176	Rio de Janeiro	Brazil	
244	Cape Town	South Africa	

As the table shows, the proximity of CloudFront is uneven around the world. A number of countries are pretty far away from the nearest CloudFront presence.

CloudFront changes its connectivity regularly, mostly for the better. An interesting data point for example is that on April 8, CloudFront created a presence near Hong Kong dropping the distance from 160 milliseconds to 4 milliseconds. The following graph gives more detail.



This data shows that maintaining good proximity in an ever changing Internet is not a trivial thing to do. See for example the ever changing proximities to New York and New Zealand.

Our research was done in collaboration with Jitscale (a cloud consultancy) and WatchMouse. Distances are measured by measuring a TCP connect to an http URL of an object provided by Cloudfront, and does not include DNS lookup.

Conclusion: so what?

The data shows that cloud computing services have left the lab a while ago, and are up and running around the world. They have progressed beyond proofs of concept and 'demoware'. This is not to say that these services are superior to every alternative. Every situation merits its own risk analysis.

The data also shows that the infrastructure underlying some of the services is in active development, and its deployment is expanding worldwide. These services scale by expanding their local distribution, not by racking up more centralized massive resources.