Amazon Cloud Storage (S3) Background and Experience Information

Description: Amazon Simple Storage Service (S3)

Information: http://aws.amazon.com/s3/

Background: Amazon provides cloud-based storage through their S3 system. This is a scalable infrastructure that stores data at multiple centers within the Amazon.com infrastructure. In addition, Amazon provides computational services through EC2.

API Information

Here is the link to the API documentation:

http://docs.amazonwebservices.com/AmazonS3/latest/API/

Here is another link to further documentation:

http://aws.amazon.com/documentation/s3/

REST Example:

http://www.anyexample.com/programming/php/uploading_files_to_amazon_s3_with_rest_api.xml

The API uses Buckets, Objects, Keys, and Operations. An object has four parts: value, key, metadata, and an access control policy. Objects are stored in buckets. The API is available in both REST and SOAP protocol. Responses are formatted in XML.

Limitations

Amazon has a maximum Object size of 5TB.

Moving Data to the Cloud

In addition to online, Amazon has the following capabilities

They also have a feature called AWS Import/Export which provides the following:

* Data Migration – If you have data you need to upload into the AWS cloud for the first time, AWS Import/Export is often much faster than transferring that data via the Internet.
* Content Distribution – Send data to your customers on portable storage devices.

* Direct Data Interchange – If you regularly receive content on portable storage devices from your business associates, you can have them send it directly to AWS for import into your Amazon S3 buckets.
* Offsite Backup – Send full or incremental backups to Amazon S3 for reliable and redundant offsite storage.
* Disaster Recovery – In the event you need to quickly retrieve a large backup stored in Amazon S3, use AWS Import/Export to transfer the data to a portable storage device and deliver it to your site.

This is basically you sending them your data on physical storage media instead of paying the IN transfer costs. This could be useful if you have a decent amount of data and want to avoid the higher costs associated with data transfer.

http://aws.amazon.com/importexport/

Pricing

Their pricing depends on storage used, data transfer in/out, PUT/COPY/POST/LIST requests, and GET requests. Each of these variables has an associated cost.

http://aws.amazon.com/s3/

If you scroll to the middle you can see the cost associated with each variable.

They also have a simple calculator available that will give you the monthly cost in real numbers:

http://calculator.s3.amazonaws.com/calc5.html

Here are some further points:

1. You pay $0.15 per gigabyte stored, on a monthly basis. If you store truly massive amounts of data (upwards of 50 terabytes), you get a small per-gigabyte discount. If you have 1GB sitting on S3 for a year, you will pay $1.80 for the entire year.

2. You then pay for the amount of data transfer each month. Pricing starts at $0.17 per gigabyte transferred (different per region used). (If you are able to have more than 10TB worth of data transfer each month, you will pay the first discount tier's rates, $0.13/GB.)

3. There are very minimal charges for file management requests such as COPY, POST, and LIST. They end up being "$0.01 per 1,000 PUT, COPY, POST, or LIST requests" and "$0.01 per 10,000 GET and all other requests," with fees waived for DELETE requests.

4. Prices vary slightly depending on the geographic locations of the datacenters. According to the FAQ page, it doesn't matter where you live: Anybody can use S3. It's the location of specific data that determines price.

Experiences

The following describes the pros/cons in working with Amazon S3 based on input from 3 projects at JPL:

- Lunar Mapping and Modeling Project
- DesdynI mission
- Early Detection Research Network

PROS:

1. Highly redundant/scalable.

2. Decent security controlled by ACL.

3. You can store your data by in different regions (US west,US east, EU, or asia pacific). You can move or replicate your data across regions for
even further data access, availability, or redundancy.

4. Low latency (usually GigE out) or as we noticed to JPL from Amazon

5. Multiple access methods outside of using an AMI (Amazon Machine Image). You can use a REST-style HTTP interface, SOAP interface, or even BitTorrent to access your data (some of this requires coder skills).

6. If you host a full "OS" instance and add a data storage volume using EBS to your virtual machine you can then leverage a VPC (Virtual Private Connection) between your site and Amazon for enhanced security.

7. If you need a CDN (Content Delivery Network) you can easily integrate your data with Amazon CloudFront.

8. Excellent S3 FireFox Organizer utility for Amazon S3. This Firefox add-on allows you to interact directly with Amazon S3 using a FTP style GUI inside your browser. Not really relevant for something large scale but still interesting.

9. Best of all..you don't need to worry about hardware purchases,
technical skills involved in developing a high-end SAN/NAS solution or making sure hardware warranty is up to date. This is all done for you when using Amazon S3.

10. Ability to create signed URLs that expire after a specified timeout so people can't reuse the same URL to re-download data.

11. Can mount it as a subdirectory in Unix using S3FS (http://code.google.com/p/s3fs/)

12. Amazon just bumped the maximum file size from 5 gigabytes to 5 terabytes last Thursday. That's been preventing us from uploading full sized LMMP images there, but doesn't look to be a problem anymore.

CONS:

1. It can get expensive. It of course depends on your usage and node requirements. Their is a cost per GB stored in S3, cost per GB transfer out, and cost per GB transfer in. They do not charge for doing a local "region" only copy between Amazon S3 region. The cost will also decrease per GB if your volume increases.

2. I don't think we can put ITAR data in the cloud...

3. Vendor lock-in.

4. You do need some coding skills if you want to interact directly with Amazon S3. You will need to use their API which is a bit of work IMHO. I am not a guru programmer so it could be simple for other folks to use (to get around this you can use an Amazon virtual machine + EBS).

5. Believe it or not..but they have experienced downtime over multiple regions in Amazon S3. They have been better recently but it still happens:

http://www.readwriteweb.com/archives/more_amazon_s3_downtime.php

6. This could be a con but I consider not having physical control over your data a bit worrisome. I would always recommend (at least for now) to have a tape or local copy of any data you place in the cloud.

7. I have noticed variable speeds in terms of transfers between Amazon S3 and local systems (at JPL or home). It will sometimes be super fast and other times a bit on the "slow" side. It is still super fast overall but 10 MB/s vs 15 MB/s on some transfers is still a noticeable difference. I believe this is because of the "shared" infrastructure that Amazon employs and if you have someone on that network fabric sending/receiving a ton of data it potentially could impact your speed. I haven't been able to fully confirm how Amazon designs their storage infrastructure (that is a closely guarded secret for obvious reasons) but I would imagine they only have so much bandwith available per "rack".

8. Slow to upload and delete a lot of small files (under 1 megabyte, like image tiles) due to the overhead from the HTTP protocol.