Computational Infrastructure

Sean Wilkinson
University of Alabama at Birmingham

Problem

You know Big Data need Big Computers, but now you are asking yourself questions such as

How do I get one?
How do I use it?
How can I afford it?

How do I get one?

- Buy a “real” workstation
  - Pay up front
  - Full control over every nitty, gritty detail
  - Full responsibility if the machine dies
- Lease “cloud” resources
  - Pay as you go
  - Significantly less control (sometimes zero)
  - Guaranteed performance and availability

Should I use the cloud?

Probably!

Unless you are using your workstation at full capacity for more than 2 hours a day, you will save money by using the cloud.

Thus, we’ll focus on cloud computing ☺

Which cloud should I use?

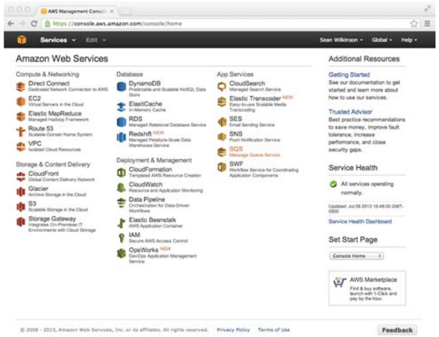
The major players right now are

- Amazon
- Google
- Microsoft
- OpenStack (Dell, HP, IBM, Rackspace, and others)

Currently, I recommend Amazon, but keep an eye on OpenStack because CERN is behind it.

How do I use it?

1. Get an Amazon account.
2. Log in to https://console.aws.amazon.com/.
3. Select the service you want to use ☺



Amazon Web Services

Compute & Networking
Direct Connect — Dedicated Network Connection to AWS
EC2 — Virtual Servers in the Cloud
Elastic MapReduce — Managed Hadoop Framework
Route 53 — Scalable Domain Name System
VPC — Isolated Cloud Resources

Storage & Content Delivery
CloudFront — Global Content Delivery Network
Glacier — Archive Storage in the Cloud
S3 — Scalable Storage in the Cloud
Storage Gateway — Integrates On-Premises IT Environments with Cloud Storage

Database
DynamoDB — Predictable and Scalable NoSQL Data Store
ElastiCache — In-Memory Cache
RDS — Managed Relational Database Service
Redshift NEW — Managed Petabyte-Scale Data Warehouse Service

Deployment & Management
CloudFormation — Templated AWS Resource Creation
CloudWatch — Resource and Application Monitoring
Data Pipeline — Orchestration for Data-Driven Workflows
Elastic Beanstalk — AWS Application Container
IAM — Secure AWS Access Control
OpsWorks NEW — DevOps Application Management Service

App Services
CloudSearch — Managed Search Service
Elastic Transcoder NEW — Easy-to-use Scalable Media Transcoding
SES — Email Sending Service
SNS — Push Notification Service
SQS — Message Queue Service
SWF — Workflow Service for Coordinating Application Components

Additional Resources

Getting Started — See our documentation to get started and learn more about how to use our services.

Trusted Advisor — Best practice recommendations to save money, improve fault tolerance, increase performance, and close security gaps.

Service Health

All services operating normally.

Updated: Jul 08 2013 16:48:00 GMT-0500

Service Health Dashboard

Set Start Page

AWS Marketplace — Find & buy software, launch with 1-Click and pay by the hour.

© 2008 - 2013, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Relevant Services

- Elastic Compute Cloud (EC2)
  - Create and destroy virtual machines instantly
  - Customize machines if you like that sort of thing, or use images from Amazon Marketplace if you don’t (https://aws.amazon.com/marketplace/)
  - Pay only for what you use
  - Buy a Top500 machine for twenty minutes, then throw it away – EC2 makes HPC disposable! (http://goo.gl/KqqCa)

Relevant Services

- Elastic Block Store (EBS)
  - Create and destroy virtual hard drives instantly
  - Configurable from 1 gigabyte to 1 terabyte
  - EC2 instances can mount these as hard drives, but they perform more like network drives
  - Pay for what you use, plus the size of the provisioned storage (you pay until you destroy it)
  - Tricky to upload data to it directly

Relevant Services

- Simple Storage Service (S3)
  - Create “buckets” for files and folders with full access control and web publishing
  - Files can be up to 5 terabytes each, and there is no limit to the number of objects you can store.
  - 99.999999999% durability over a given year!
  - Everything can be set up using only a web browser, but you can also automate using Amazon’s SDK (http://aws.amazon.com/tools/)

Relevant Services

- Glacier
  - Create “vaults” for immutable “archives” with full access control and web publishing
  - Archives are often TAR or ZIP files to save money
  - Archives can be up to 40 terabytes each, and there is no limit to the number of archives you can store within a vault.
  - Think of this as the cloud version of a tape drive.
  - Same durability, but really slow retrieval vs. S3

Data Transfer

- You have three options:
  - For “small” files and fast networks, you can upload data directly from your web browser or use Amazon’s SDK to upload to S3 or Glacier.
  - The SDK’s Multipart Upload API can be used for parallel and streaming uploads to S3 and Glacier.
  - Physically mail your hard drive(s) to Amazon to upload to EBS, S3, or Glacier. Avoid this if you can! (http://aws.amazon.com/importexport/)

Proposed Solution

1. Upload data to S3 using the Multipart Upload API in Amazon’s SDK.
2. Create a cc2.8xlarge instance on EC2 using an image that has Xiuxia Du’s ADAP installed.
3. Process data on EC2 and save results to S3.
4. Archive data and/or results into Glacier.
5. Destroy unnecessary EC2 and S3 resources.

(Note: I haven’t actually tested this yet.)

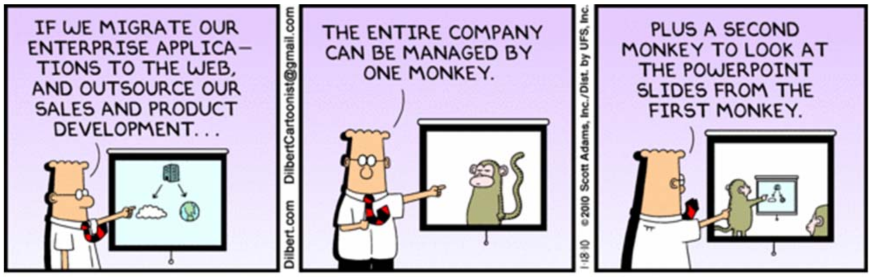
Other clouds

- Google’s offerings may outperform Amazon’s, but you have to apply for an account, and they still haven’t replied to me after 2 months …
- Microsoft Azure is strong but relatively new.
- Joyent’s new “Manta” may change the game!
- The OpenStack frees your software from vendor lock-in. CERN and Rackspace recently partnered to power computations for the LHC.

Conflict of Interest Disclosure

I own a small amount of stock in Rackspace.

Thank you!


IF WE MIGRATE OUR ENTERPRISE APPLICA– TIONS TO THE WEB, AND OUTSOURCE OUR SALES AND PRODUCT DEVELOPMENT. . .
Dilbert.com DilbertCartoonist@gmail.com
THE ENTIRE COMPANY CAN BE MANAGED BY ONE MONKEY.
1-18-10 ©2010 Scott Adams, Inc./Dist. by UFS, Inc.
PLUS A SECOND MONKEY TO LOOK AT THE POWERPOINT SLIDES FROM THE FIRST MONKEY.