Charles Zheng

Introduction to "Cloud Computing"

Stats 290
March 11, 2015

Review of Parallel/Distributed Computing

- Physical speed limit for sequential computing
- Only choice: parallelization
- Types of parallelization:
  - Multi-core on the same computer (`multicore`)
  - Simple network of workstations (`snow`)
  - Cluster computing (`batch`)

What is the cloud?

- Computationally, same paradigm as cluster computing
  - *"I don't understand what we would do differently in the light of cloud computing other than change the wording of some of our ads"* — Larry Ellison, Oracle
- Pricing model: Own/Rent → **Pay-as-you-go**
- Consequences:
  - Individuals can choose to create a "**personal cluster**"
  - Large-scale **interactivity** enabled due to "cost associativity"

Types of Cloud Services

- Low-level : Utility computing
- Mid-level: Hosting
- High level: Applications

Types of Cloud Services

- **Low-level : Utility computing**
  - Users launch servers in the cloud
  - Users specify memory, networking, CPUs
  - Users act as sysadmins
  - Just like using your own computer though SSH
- Examples:
  - Amazon Elastic Compute Cloud (EC2)
  - Google ComputeEngine

Types of Cloud Services

- **Mid-level: Hosting**
  - Platform for hosting specific types of applications written in specific programming language
  - Hosting service act as sysadmins to manage scaling
  - Examples:
    - Traditional web hosting (http, PHP, Ruby on Rails, etc)
    - Microsoft Azure (.NET Framework)
    - Amazon Elastic MapReduce (Apache Hadoop)

Types of Cloud Services

- **High level: Applications**
  - Users use an existing application on the cloud
  - Examples:
    - Search engines
    - Dropbox
    - Google documents
    - cloud versions: Adobe Photoshop, Microsoft office

Weaknesses of cloud computing

- Latency between virtual machines
  - Supercomputers still dominate in scientific computing, e.g. weather simulations
- Transfer costs for large datasets
  - Cheaper and faster to send 100 TB on a hard drive via FedEx
- Obstacles for reproducible research
  - Streaming data: what if I can't archive the stream?
  - Dependence on cloud provider: what if they discontinue their service?

Utility Computing for Individuals

The story of the personal computer

- Historically, all computing was done on **mainframes**
  - Many users share on mainframe
  - Users log in using a terminal
- Division between admins and users
- "Home computer revolution" predicted in 1970's but did not become mainstream until 1988
- Now you can administer your own personal computer…

Utility computing

- Attractive due to *cost associativity*
  - 100 hours using 1 computer = 1 hour using 100 computers
    - Costs are now comparable
- Increasing availability of tools to streamline configuration
  - Scripts to auto. launch and configure are common
- Conveniently use new frameworks like Hadoop or Spark

Personal Cloud Computing

- Cheaper to buy computing hours/pay-by-hour software than invest in hardware and personal software licenses
  - Exemplified by Google Chromebook
- Now you can administer your own cluster…
  - A daunting task, but the process is becoming more streamlined
  - Necessary for cutting-edge paradigms (like Spark)
- Take responsibility of your own security
  - Pro: No more headaches connecting to organization network
  - Con: Nobody watching out for you over your shoulder

Elements of Utility Computing

- A billable, password-protected account
  - Secret access codes for programmatic access
- Resources
  - Data storage, Virtual instances, Machine images
  - SSH key pairs
- Configuration
  - Security groups, Geographic service zone
- Interface
  - Online dashboard
  - Programmatic APIs

A simple workflow using Amazon EC2

- (Demo in class)

Managing your cloud

- Run scripts on your own machine
- Scripts can launch, configure, run, collect, and cleanup
- (see boto demonstration in IPython notebook)

Interactive Computing in the Cloud

Interfaces

- One option: Handle GUI on client side
  - E.g. Get results of computation from Rserve, then display in R
- Other option: Web sockets
  - Launch a web application from the cloud, then access it locally
  - RStudio server, IPython notebook, 0xdata (demos)

What can interactivity do for you?

- Adjust your experiments on the fly
  - Video: Spark streaming for neuroscience
- Scale up exploratory data analysis
- Probe for weaknesses in your methods using simulations
- [Your startup idea here]

Conclusion

- Cloud computing: mostly the same hardware and software but now priced by usage
- Key feature is cost associativity: pay the same for 100 hours on 1 machine or 1 hour on 100 machines in parallel
- It is becoming easier and easier for individuals without specialized training to administer their own “personal cluster”: try it!

Friday's lecture: Spark tutorial

- You will be given access to a Linux instance on EC2
- Learn how to
  - use the Hadoop filesystem
  - launch IPython notebooks
  - run Spark jobs from your browser