Data-intensive Computing on the Cloud: Concepts, Technologies and Applications

B. Ramamurthy
bina@buffalo.edu

This talks is partially supported by National Science Foundation grants DUE: #0920335, OCI: #1041280
Presenter’s Background in cloud computing

• Bina
  o Is a PI on two current NSF* grants related to cloud computing:
    o 2009-2012: Data-Intensive computing education: CCLI Phase 2: $250K
    o 2010-2012: Cloud-enabled Evolutionary Genetics Testbed: OCI-CI-TEAM: $250K
  o Faculty at the CSE department at University at Buffalo.

*National Science Foundation
Outline of the talk

• Introduction to Data-intensive computing on the cloud
  o Technology context: multi-core, virtualization, 64-bit processors, parallel computing models, big-data storages...
  o Cloud models: IaaS (Amazon AWS), PaaS (Microsoft Azure), SaaS (Google App Engine)

• Demonstration of cloud capabilities
  o Cloud models: Demos on Amazon EC2 cloud
  o Data-intensive Computing: MapReduce

• A Certificate Program in Data-intensive Computing offered by SUNY (yes, SUNY approved)

• Questions and Answers
Introduction: A Golden Era in Computing

- Powerful multi-core processors
- General purpose graphic processors
- Superior software methodologies
- Virtualization leveraging the powerful hardware
- Explosion of domain applications
- Proliferation of devices
- Wider bandwidth for communication

2/7/2012 NCCC
Top Ten Largest Databases

Ref: http://www.focus.com/fyi/operations/10-largest-databases-in-the-world/
Top Ten Largest Databases in 2007 vs Facebook’s cluster in 2010

Top ten largest databases (2007)

Ref: http://www.focus.com/fyi/operations/10-largest-databases-in-the-world/
Big-data Challenges

• Scalability issue: large scale data, high performance computing, automation, response time, rapid prototyping, and rapid time to production
• Need to effectively address (i) ever shortening cycle of obsolescence, (ii) heterogeneity and (iii) rapid changes in requirements
• Transform data from diverse sources into intelligence and deliver intelligence to right people/user/systems
• How to store the big-data? What new computing models are needed?
• What about providing all this in a cost-effective manner?
Enter the cloud

• **Cloud computing** is Internet-based computing, whereby shared resources, software and information are provided to computers and other devices on-demand, like the electricity grid.

• The cloud computing is a culmination of numerous attempts at large scale computing with seamless access to virtually limitless resources.
  
  o on-demand computing, utility computing, ubiquitous computing, autonomic computing, platform computing, edge computing, elastic computing, **grid computing**, …
The Cloud Computing

• Cloud provides processor, software, operating systems, storage, monitoring, load balancing, clusters and other requirements as a service

• Pay as you go model of business

• When using a public cloud the model is similar to renting a property than owning one.

• An organization could also maintain a private cloud and/or use both.

• Cloud computing models:
  o platform (PaaS),
  o software (SaaS),
  o infrastructure (IaaS),
  o Services-based application programming interface (API)
Windows Azure

- Enterprise-level on-demand capacity builder
- Fabric of cycles and storage available on-request for a cost
- You have to use Azure API to work with the infrastructure offered by Microsoft
- Significant features: web role, worker role, blob storage, table and drive-storage
- Platform as a service
Google App Engine

- This is more a web interface for a development environment that offers a one stop facility for design, development and deployment Java and Python-based applications in Java, Go and Python.
- Google offers the same reliability, availability and scalability at par with Google’s own applications
- Interface is software programming based
- Comprehensive programming platform irrespective of the size (small or large)
- Signature features: templates and appspot, excellent monitoring and management console;
- Free version to explore at: http://code.google.com/appengine/
- Software as a service: Evolutionary Genetics Testbed
Amazon EC2

- Amazon EC2 is one large complex web service.
- EC2 provides an API for instantiating computing instances with any of the operating systems supported.
- It can facilitate computations through Amazon Machine Images (AMIs) for various other models.
- Signature features: S3, Cloud Management Console, MapReduce Cloud, Amazon Machine Image (AMI)
- Excellent distribution, load balancing, cloud monitoring tools
- You can explore amazon using the free account at:
  - http://aws.amazon.com/free/
Demos

• Amazon AWS: EC2 & S3 (among the many infrastructure services)
  o Archiving on the cloud,
    • Windows instance
  o Rescuing legacy applications using the cloud,
    • Windows instance
  o A three-tier enterprise application
    • Tomcat, Mysql, Web server Linux instance
    • Bitnami AMI (Amazon Machine Image)
  o A big-data application on a distributed cluster (Data-intensive computing)
    • Word count application on a cluster
    • MapReduce programming model on Hadoop Cluster
Summary

• We explored the need for data-intensive or big-data computing
• We discussed three popular cloud models that are delivered as services
• We illustrated cloud concepts and demonstrated the cloud capabilities through simple applications
• Data-intensive computing on the cloud is an essential and indispensable skill for the workforce of today and tomorrow
• UB has implemented a SUNY-wide a Certificate Program in Data-intensive Computing
References & useful links

- For miscellaneous information: [http://www.cse.buffalo.edu/~bina](http://www.cse.buffalo.edu/~bina)