

NJIT

The logo features the letters 'NJIT' in a large, white, serif font. A thick, white, curved line starts under the 'J' and sweeps upwards and to the right, ending under the 'T'.

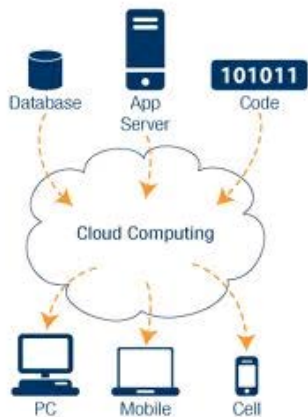
New Jersey's Science &
Technology University

THE EDGE IN KNOWLEDGE

Course Overview

Cloud Computing at a Glance

- **Utility computing**: our data and applications are hosted somewhere in the Internet (“in the cloud”)
 - Most services we access over the Internet are in the cloud (e.g., Google, Amazon, Yahoo)



Cloud infrastructure = Data centers with 100,000's servers

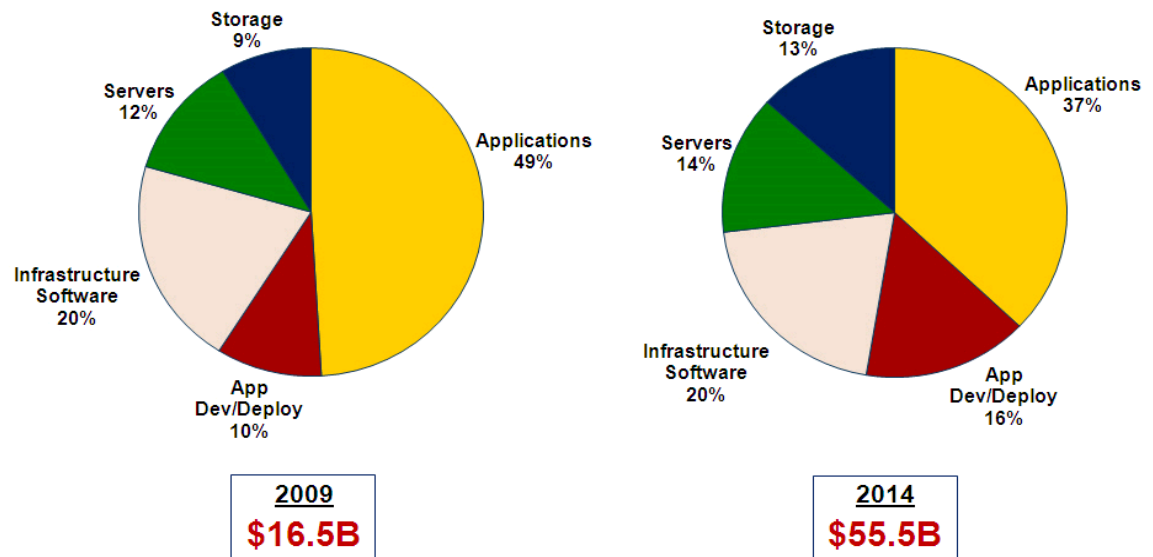
- Benefits:
 - Providers: economies of scale by having many users sharing the same infrastructure
 - Consumers: reduced cost and overhead

Tech Predictions

■ Gartner's 2011 Strategic Technologies

1. Cloud Computing
2. Mobile Applications
3. Social Computing
4. ...

Worldwide Public IT Cloud Services* Spending (\$B)
by Offering Category
2009, 2014



Source: IDC, June 2010

* Includes spending on Applications, Application Development & Deployment Software, Systems Infrastructure Software, Server capacity and Storage capacity provided via the public Cloud Services delivery model.

Course Goals

- Deep understanding of fundamental concepts in cloud computing
- Learn how to use existing commercial cloud computing infrastructures
- Learn how to implement cloud applications using novel programming models for parallel computing

Course Material

- The systems, programming models, and architectural concepts covered in this course have been created in the past 10 years
 - Changing the way we do distributed computing (or even computing in general)
 - Have started to become mainstream in the past 6-7 years
 - Are here to stay for many years
 - Used by big companies such as Google, Amazon, Microsoft, Yahoo, etc.

Course Outline (1)

- Context for this class: brief history of distributed systems and parallel computing
- Cloud computing
 - Overview
 - Three commercial cloud platforms
 - Amazon Web Services (AWS)
 - Google AppEngine
 - Windows Azure
 - Learn how to set up and manage the cloud (AWS)

Course Outline (2)

- Programming models & frameworks for parallel programming in cloud computing
 - Google's MapReduce & Apaches' Hadoop (open source implementation of MapReduce)
 - Yahoo's Pig Latin
 - Microsoft's Dryad & DryadLINQ
 - Assume large amounts of distributed data that need to be processed as fast as possible
 - Provide high-level API to describe parallel programs
 - Provide support for task scheduling & fault-tolerance

Course Outline (3)

- Distributed storage systems
 - Google File System: a scalable distributed file system for data intensive applications
 - Google's BigTable: a distributed, persistent data structure to store (key, value) pairs - multidimensional map
 - Amazon's Dynamo: Conceptually similar to BigTable, but focused on high availability

Course Outline (4)

- Virtualization
 - Architecture of VmWare and Xen virtual machines monitors: **allow multiple OSs to run concurrently on the same machine, while providing users with the illusion of owning a physical machine**
 - Virtual machine migration: useful for load balancing, fault-tolerance, or having your virtual computer follow you as you move from one place to another

Course Outline (5)

- Cloud security
 - Information leakage in cloud computing
- Multi-core architectures
 - All computers are multi-core now
 - Why? Heat dissipation and other physical limitations for sequential machines
 - Operating systems for multicore

Programming Assignment

- On Amazon Web Services
 - Create Hadoop AML (i.e., bundle of OS + Hadoop to be installed in the cloud)
 - Implement two Hadoop applications
 - Hadoop programming is done in Java over Linux

Project (1)

- **Group project:** design, implement, and measure performance of cloud applications
 - Applications deal with large amounts of data (Gbytes) and are useful in real-life
 - Analyze scalability as function of the cluster size and data size
 - Study performance degradation in the presence of node failures
 - **Hadoop on Amazon Web Services platform**
- Get familiar with Amazon public data sets (or think about other interesting data sets)
 - <http://aws.amazon.com/publicdatasets/>
 - Astronomy, Biology, Chemistry, Climate, Economics, Encyclopedic, Geographic, Mathematics

Project (2)

- The general requirements will be presented after the midterm
- You will work in groups of 3
- Outline presentation
 - You will choose your project idea
 - Presentation: who, what, why, how
- Final project presentation will be uploaded and discussed during the last week of the semester
 - Provides details for design & implementation
 - Shows performance evaluation graphs & conclusions

Paper Presentation

- Groups of two students (and rarely individual students) will create a power point presentation for one paper
 - Email me the slides at least 3 days before the paper is due to be presented; I will provide feedback
 - Papers are already posted and must be presented during the assigned week/module
 - Prepare a 45-minutes power point presentation
- Significant extra-credit is given for participation in the discussions
- Check the additional information that will be posted on how to read and present research papers

Grading

- 25% Midterm Exam
 - Covers the first 3 modules. Closed book (i.e., papers, notes)
- 25% Final Exam
 - Only the material covered after midterm. Closed book (papers, notes)
- 10% Homework
 - There will be 4 homework assigned to prepare you for the midterm and final (with similar type of questions)
 - The solutions will be posted after the due date
- 15% Programming assignment
- 15% Project
- 10% Paper presentation

NJIT

THE EDGE IN KNOWLEDGE