Why Cloud Computing?
The Academic Perspective

Jimmy Lin
The iSchool (College of Information Studies)
University of Maryland

Wednesday, January 20, 2010

What is Cloud Computing?

- “Rebranding” of Web 2.0
- Utility computing
- “Everything as a service”
- Large-data processing

How much data?

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN's LHC will generate 15 PB a year (??)

How do we do it?

- Key ideas:
  - Divide-and-conquer algorithms
  - Large clusters of commodity servers
  - Models for organization computations at a massive scale
- MapReduce
  - Programming model originally developed by Google
  - Open-source Hadoop implementation led by Yahoo!

Why is this important?

- Research:
  - Need for scalable algorithms and systems
  - Machine learning: there’s no data like more data
- Education:
  - Need to train the next generation of innovators
  - Teaching students to think at “web-scale” and in parallel

Activities at Maryland

- Google/IBM Academic Cloud Computing Initiative (ACCI)
  - Started Fall 2007
  - Provides Hadoop cluster access to universities
- NSF Cluster Exploratory (CLuE Program)
  - Started Summer 2008
  - Provides funding to spur initiatives
  - Work at Maryland includes efforts in text processing and bioinformatics
- “Cloud computing” courses
  - Spring 2008: use of Google/IBM cluster
  - Fall 2008: use of Amazon’s EC2
  - Spring 2010: use of both resources

Why the cloud?

- MapReduce enables large-data processing
  - Experimental turnarounds from days to hours
  - Enables analytics broadly applicable to many industries (e.g., data warehousing, business intelligence applications)
- Industry support: access to private and public clouds
  - Move to hybrid models

When it gets cloudy...

- Cluster reliability: reliance on external resources
- Slow data transfer
- Security less of a concern (fortunately)

Questions?