Is there a cloud in your future?
Applications of “cloud computing” to Web-scale problems

Proposal for a “wildcard” session, iConference 2008

Organizer:
Jimmy Lin (jimmylin@umd.edu)
Assistant Professor
College of Information Studies
University of Maryland, College Park

1. Background

IBM and Google recently committed a total of $30 million over two years to an initiative on “cloud computing”, in collaboration with six universities across the country (see references). They are: Berkeley, Carnegie Mellon, MIT, Stanford, the University of Maryland, and the University of Washington. I am the leader of this initiative at the University of Maryland, and to my knowledge the only participant from an iSchool (the rest are lead by faculty in computer science departments).

“Cloud computing” refers to technology for exploiting large computer clusters to tackle “Web-scale” information processing problems, where immense quantities of data make traditional sequential processing impractical. Specifically, this initiative focuses on Google’s MapReduce programming paradigm, which was specifically designed for processing extremely large data sets (and indeed used by Google itself for much of its production operations). Programs written in the MapReduce functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program’s execution across a set of machines, handling machine failures, and managing the required inter-machine communication. Hadoop is an open-source implementation of the MapReduce framework.

As a part of this initiative, IBM and Google are making Hadoop clusters available to the university collaborators, with the simultaneous goal of advancing research and education. For the past two months, the Computational Linguistics and Information Processing (CLIP) Lab at the University of Maryland has been actively exploiting this resource for research in natural language processing and information retrieval.

The exponential explosion of information on the Web and in easily accessible digital formats forces us to think “outside the box” when tackling data-intensive “Web-scale” problems. Researchers must think and analyze data at a massively parallel scale or face the prospect of being relegated to work on “toy” problems. “Cloud computing” could potentially provide the infrastructure that allows researchers to tackle “Web-scale” challenges at a reasonable cost. From an educational point of view, the ability to think about problems in terms of parallel processing algorithms will become a critical skill in tomorrow’s work force. “Cloud computing” is an emerging technology that iSchools cannot afford to ignore.

2. Goals

- To introduce the iSchool community to “cloud computing” and the MapReduce framework
- To provide the iSchool community an overview of research and education efforts currently underway
- To begin a discussion on the implications of “cloud computing” to research and education in iSchools

3. Proposed Format

I propose a 60 minute session structured in the following manner:

Overview of cloud computing [30 minutes]
(I will deliver this presentation)

- Description of the MapReduce framework.
- Discussion of the types of data-intensive information processing applications that MapReduce was designed for. I take “information processing” to broadly encompass information retrieval, natural language processing, text mining, social network analysis, etc.,
- A short demo of Hadoop, the open-source implementation of MapReduce.
- Overview of research projects at the University of Maryland that exploit this resource.
- Overview of the educational efforts associated with this initiative.

Panel commentary [15 minutes]

- Three panelists will be invited to share their views on what “cloud computing” could mean to research and education in iSchools.

Open discussion [15 minutes]

- The floor will be open to questions from the audience, and I will moderate a discussion.

3. Participants

The following panelists have been invited:

- John M. Unsworth, Dean and Professor, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. (participation confirmed)
- Another faculty from either an iSchool or from one of universities involved in the IBM/Google collaboration (to be arranged).
- Representative from IBM (to be arranged).

4. References

Selected media mentions of the IBM/Google initiative:

- New York Times: Google and I.B.M. Join in ‘Cloud Computing’ Research (October 8, 2007)
  http://www.nytimes.com/2007/10/08/technology/08cloud.html
- University of Maryland Press Release (October 8, 2007)
  http://www.newsdesk.umd.edu/culture/release.cfm?ArticleID=1515
- Washington Post: Maryland Joins Megacomputer ‘Cloud’ Project (October 9, 2007)
  http://www.washingtonpost.com/wp-dyn/content/article/2007/10/08/AR2007100801521.html

Relevant technical articles:

Luiz Andre Barroso, Jeffrey Dean, and Urs Holzle. (2003) Web Search for a Planet: The Google Cluster Architecture. *IEEE Micro*, 23(2):22-28.

Jeffrey Dean and Sanjay Ghemawat. (2004) MapReduce: Simplified Data Processing on Large Clusters. *Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI 2006).*

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, and Robert Gruber. (2006) Bigtable: A Distributed Storage System for Structured Data. *Proceedings of the 7th Symposium on Operating System Design and Implementation (OSDI 2006).*