REDORBIT
your universe, online

NSF Awards Millions For Cloud Computing Research

CLuE awards promote academic use o cluster computing resources on IBM/Google cloud

Today, the National Science Foundation (NSF) announced it has awarded nearly $5 million in grants to 14 universities through its Cluster Exploratory (CLuE) program to participate in the IBM/Google Cloud Computing University Initiative. The initiative will provide the computing infrastructure for leading-edge research projects that could help us better understand our planet, our bodies, and pursue the limits of the World Wide Web.

In 2007, IBM and Google announced a joint university initiative to help computer science students gain the skills they need to build cloud applications. Now, NSF is using the same infrastructure and open source methods to award CLuE grants to universities around the United States. Through this program, universities will use software and services running on an IBM/Google cloud to explore innovative research ideas in data-intensive computing. These projects cover a range of activities that could lead not only to advances in computing research, but also to significant contributions in science and engineering more broadly.

NSF awarded Cluster Exploratory (CLuE) program grants to Carnegie-Mellon University, Florida International University, the Massachusetts Institute of Technology, Purdue University, University of California-Irvine, University of California-San Diego, University of California-Santa Barbara, University of Maryland, University of Massachusetts, University of Virginia, University of Washington, University of Wisconsin, University of Utah and Yale University.

"Academic researchers have expressed a need for access to massively scaled computing infrastructure to explore radically new approaches to solving data-intensive problems. These approaches would be unthinkable using ordinary computing resources available on campuses today," Jeannette Wing, NSF's assistant director for computer & information science and engineering. "We are pleased to provide the awards to these fourteen universities, enabling researchers to engage with this emerging and novel model of computing."

"IBM is intensely focused on applying technology and science to make the world work better," said Willy Chiu, vice president, IBM Cloud Labs. "IBM is thrilled to power the groundbreaking studies taking place at these prestigious universities, and to help enable researchers and students around the world tackle some of the biggest problems of our time."

"We're pleased and excited that the CluE program will support a wide range of original research," said Alfred Spector, Google's vice president for research and special initiatives. "We're looking forward to seeing the grantees solve challenging problems across various fields through creative applications of distributed computing." The universities will run a wide range of advanced projects and explore innovative research ideas in data-intensive computing, including advancements in image processing, comparative studies of large-scale data analysis, studies and improvements to the Internet, and human genome sequencing, among others, using software and services on the IBM/Google cloud infrastructure.

Carnegie-Mellon University

Researchers at Carnegie-Mellon University are using cloud computing to characterize the topicality of web content to more effectively process web searches. Routing searches topically requires less effort than traditional searches, enabling significant computational and financial savings. The project is using the Google/IBM cluster to "crawl" the web and perform the data cleansing and pre-processing necessary to develop a web dataset of 1 billion documents to support the research. The web dataset is also being made available to the larger information retrieval community to multiply the impact of the project on that discipline.

The second research project is focused on developing the Integrated Cluster Computing Architecture (INCA) for machine translation (using computers to translate from one language to another). Open-source toolkits make it easier for new research groups to tackle the problem at lower costs, broadening participation. Unfortunately, existing toolkits have not kept up with the computing infrastructure required for modern "big data" approaches to machine translations; INCA will fill this void.

Florida International University

Florida International University (FIU) researchers are leveraging cloud computing to analyze aerial images and objects to help support disaster mitigation and environmental protection. Specifically, the CluE effort at FIU relates to its TerraFly project, which is a web-service of 40 terabytes of aerial imagery, geospatial queries and local data. Students and researchers will now be able to precisely code these images in real-time.

Massachusetts Institute of Technology, University of Wisconsin-Madison and Yale University

These three universities are using the National Science Foundation CLuE grants for a comparative study of approaches to cluster-based, large-scale data analysis. Both MapReduce and parallel database systems provide scalable data processing over hundreds to thousands of nodes, yet it's important for researchers to know the differences in performance and scalability of these two approaches to know which is more suitable when designing new data-intensive computing applications.

Purdue University

This project is investigating linguistic extensions to MapReduce abstractions for programming modern, large-scale systems, with special focus on applications that manipulate large, unstructured graphs. This will impact a broad class of scientific applications. Graphs have important utility in the social sciences (social networks), recommender systems, and business and finance (networks of transactions), among others. The specific case study targeted by the research is a comparative analysis of graph-structured biochemical networks and pathways which underlie many important problems in biology.

University of California-Irvine

In many applications, data-quality issues resulting from a variety of errors create inconsistencies in structures, representations or semantics. Simple spelling variations such as "Schwarzenegger" vs. "Schwarseneger," "Brittany Spears" vs."Britney Spears," or "PO Box" vs. "P.O. Box" are an example of this. Dealing with these issues is becoming increasingly important as the volume of data being processed increases. This project is providing support for efficient fuzzy queries on large text repositories. Supporting fuzzy queries can ultimately help applications mitigate their data quality issues because entities with different representations can be matched.

University of California-San Diego / San Diego Supercomputer Center

Researchers at the University of California, San Diego are studying how to manage and process massive spatial data sets on large-scale compute clusters. The specific test case is analysis of high-resolution topographic data sets from airborne LiDAR surveys, which are of great interest to many Earth scientists. Providing efficient access and analytic capabilities will have broad impact beyond the geosciences because the techniques are likely to be applicable to other types of large data sets.

University of California-Santa Barbara

Many of today's data-intensive application domains, including searches on social networks like Facebook and protein matching in bioinformatics, require us to answer complex queries on highly-connected data. The UCSB Massive Graphs in Clusters (MAGIC) project is focused on developing software infrastructure that can efficiently answer queries on extremely large graph datasets. The MAGIC software will provide an easy to use interface for searching and analyzing data, and manage the processing of queries to efficiently take advantage of computing resources like large datacenters.

University of Maryland-College Park

The CluE initiative is funding another machine translation project that promises to bridge the language divide in today's multi-cultural and multi-faceted society. Systems capable of converting text from one language into another have the potential to transform how diverse individuals and organizations communicate. By coupling network analysis with cross-language information retrieval techniques, the result is a richer, multilingual contextual model that will guide a machine translation system in translating different types of text. The potential broader impact of this project is no less than knowledge dissemination across language boundaries, which will serve to enrich the lives of all the world's citizens.

A second project focuses on developing parallel algorithms for analyzing the next generation of sequencing data. Scientists can now generate the rough equivalent of an entire human genome in just a few days with one single sequencing instrument. The analysis of these data is complicated by their size - a single run of a sequencing instrument yields terabytes of information, often requiring a significant scale-up of the existing computational infrastructure needed for analysis.

University of Massachusetts-Amherst

This project focuses on how researchers at the Center for Intelligent Information Retrieval (CIIR) are using the CluE infrastructure to learn more about word relationships. These relationships are not labeled explicitly in text and are quite varied; by exploiting these relationships, this project will help lead to a more effective ranking of web-retrieval results.

University of Virginia

Imagine continuously zooming into an image from your personal photo collection. Unlike with modern image processing software, however, this zoom operation would reveal details missing from the original image. For example, zooming into someone's shirt would eventually show a high-resolution image of the threads that compose it. A research team in the Department of Computer Science at the University of Virginia plans to develop techniques for intelligently enlarging a digital image that use a database of millions of on-line images to find examples of what its components look like at a higher spatial resolution.

University of Washington

Astrophysics is addressing many fundamental questions about the nature of the universe through a series of ambitious wide-field optical and infrared imaging surveys. New methodologies for analyzing and understanding petascale data sets are required to answer these questions. This research project is focused on developing new algorithms for indexing, accessing and analyzing astronomical images. This work is expected to have a broad range of applications to other data intensive fields.

University of Washington and University of Utah

This project is building a new infrastructure for computational oceanography that uses the CluE platform to allow ad hoc, longitudinal query and visualization of massive ocean simulation results at interactive speeds. This infrastructure leverages and extends two existing systems: GridFields, a library for general and efficient manipulation of simulation results; and VisTrails, a comprehensive platform for scientific workflow, collaboration, visualization, and provenance.

IBM/Google Cloud Computing University Initiative

The following resources are available from IBM and Google to these universities to leverage for their respective projects:

* A cluster of processors running an open source implementation of Google's published computing infrastructure (MapReduce and GFS from Apache's Hadoop project)

* A Creative Commons licensed university curriculum developed by Google and the University of Washington focusing on massively parallel computing techniques

* Open source software designed by IBM to help students develop programs for clusters running Hadoop. The software works with Eclipse, an open source development platform.

* Management, monitoring and dynamic resource provisioning by IBM using IBM Tivoli systems management software

---

Image Caption: A computer visualization of a river bed created using VisTrails, a system developed by University of Utah computer scientists to help scientists create high-quality visualizations. Under the CluE initiative, the University of Utah team will work with other computer scientists from the University of Washington to expand the capabilities of VisTrails and make it easier to visualize very large data sets. Credit: Juliana Freire and Claudio Silva, University of Utah

---

On the Net:

- National Science Foundation

Story from REDORBIT NEWS:
http://www.redorbit.com/news/display/?id=1676748

Published: 2009/04/24 09:34:38 CDT