Data Intensive Scalable Computing

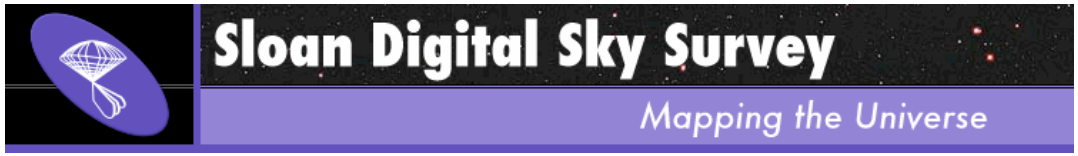
Randal E. Bryant
Carnegie Mellon University

`http://www.cs.cmu.edu/~bryant`

Examples of Big Data Sources

Wal-Mart

- 267 million items/day, sold at 6,000 stores
- HP building them 4PB data warehouse
- Mine data to manage supply chain, understand market trends, formulate pricing strategies



Sloan Digital Sky Survey

- New Mexico telescope captures 200 GB image data / day
- Latest dataset release: 10 TB, 287 million celestial objects
- SkyServer provides SQL access
- Next generation LSST even bigger

Our Data-Driven World

Science

- Data bases from astronomy, genomics, natural languages, seismic modeling, …

Humanities

- Scanned books, historic documents, …

Commerce

- Corporate sales, stock market transactions, census, airline traffic, …

Entertainment

- Internet images, Hollywood movies, MP3 files, …

Medicine

- MRI & CT scans, patient records, …

Cloud Computing Varieties

"I don't want to be a system administrator. You handle my data & applications."

- Hosted services
- Documents, web-based email, etc.
- Can access from anywhere
- Easy sharing and collaboration

"I've got terabytes of data. Tell me what they mean."

- Very large, shared data repository
- Complex analysis
- ***Data-intensive scalable computing*** **(DISC)**

CS Research Issues

Applications

- Language translation, image processing, …

Application Support

- Machine learning over very large data sets
- Web crawling

Programming

- Abstract programming models to support large-scale computation
- Distributed databases

System Design

- Error detection & recovery mechanisms
- Resource scheduling and load balancing
- Distribution and sharing of data across system

Getting Started

Goal

- Get faculty & students active in DISC

Software: Hadoop

- Open source project inspired by Google infrastructure
  - Distributed file system
  - MapReduce programming environment
- Supported and used by Yahoo
- Prototype on single machine, map onto cluster

Hardware: Rely on Kindness of Others

Press Release 08-031

NSF Partners With Google and IBM to Enhance Academic Research Opportunities

Computer science researchers at universities and colleges will be able to utilize large-scale computing cluster

February 25, 2008

Today the National Science Foundation's Computer and Information Science and Engineering (CISE) Directorate announced the creation of a strategic relationship with Google Inc. and IBM. The Cluster Exploratory (CluE) relationship will enable the academic research community to conduct experiments and test new theories and ideas using a large-scale, massively distributed computing cluster.

- **Google setting up dedicated cluster for university use**
- **Loaded with open-source software**
  - Including Hadoop
- **IBM providing additional software support**
- **NSF will determine how facility should be used.**

More Sources of Kindness

Yahoo, Carnegie Mellon Switch On Supercomputer

Submitted by David A. Utter on Mon, 11/12/2007 - 11:08.

Comment | Email | Print

The M45 supercomputer provided by Yahoo opened its ports to its partners at Carnegie Mellon University, where the initiative should help boost research that benefits the broader Internet community.

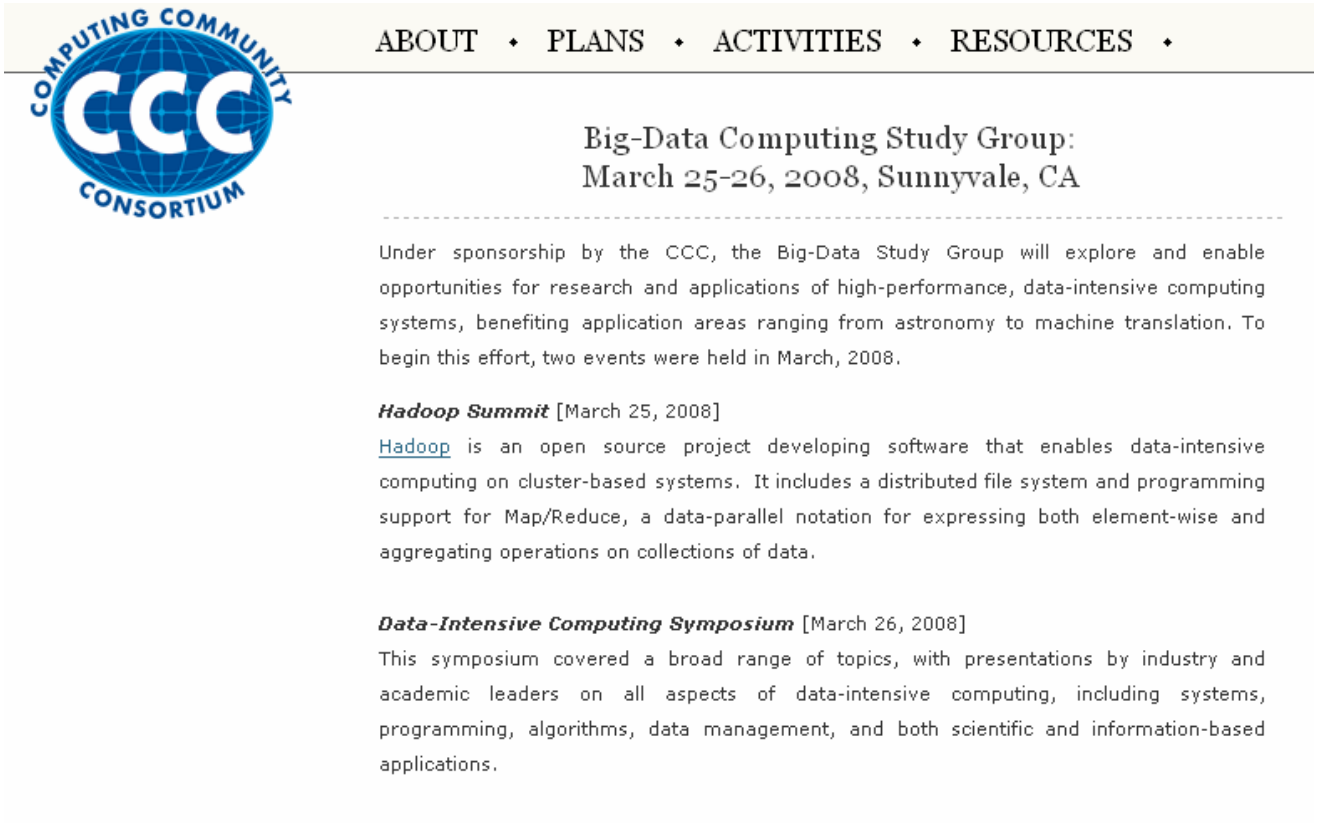
YAHOO! Carnegie Mellon

For those of you firing up the old faithful laptop for a morning of surfing, blogging, maybe a little development work, get a load of what some of the lucky geeks at Carnegie Mellon University got to play with this morning:

The M45, Yahoo's supercomputing cluster, has approximately 4,000 processors, three terabytes of memory, 1.5 petabytes of disks, and a peak performance of more than 27 trillion calculations per second (27 teraflops), placing it among the top 50 fastest supercomputers in the world.

- **Yahoo: Major supporter of Hadoop**
- **Yahoo plans to work with other universities**

Big-Data Computing Study Group



ABOUT • PLANS • ACTIVITIES • RESOURCES •

Big-Data Computing Study Group:
March 25-26, 2008, Sunnyvale, CA

Under sponsorship by the CCC, the Big-Data Study Group will explore and enable opportunities for research and applications of high-performance, data-intensive computing systems, benefiting application areas ranging from astronomy to machine translation. To begin this effort, two events were held in March, 2008.

Hadoop Summit [March 25, 2008]

Hadoop is an open source project developing software that enables data-intensive computing on cluster-based systems. It includes a distributed file system and programming support for Map/Reduce, a data-parallel notation for expressing both element-wise and aggregating operations on collections of data.

Data-Intensive Computing Symposium [March 26, 2008]

This symposium covered a broad range of topics, with presentations by industry and academic leaders on all aspects of data-intensive computing, including systems, programming, algorithms, data management, and both scientific and information-based applications.

- **Co-organized by REB & Thomas Kwan (Yahoo!)**
- **Supported by Computing Community Consortium**

BDCSG Activities

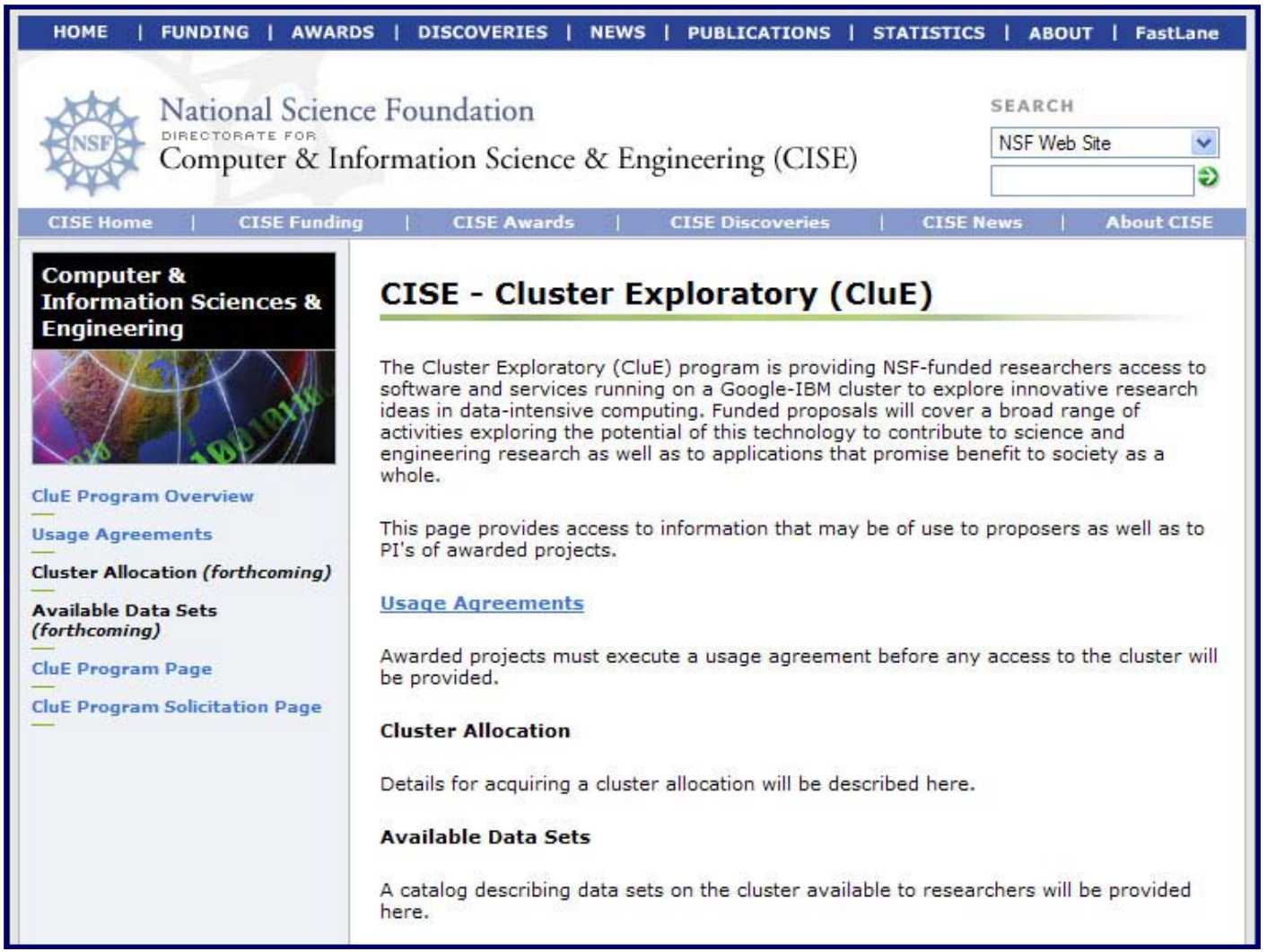
Hadoop Summit

- 350+ people showed up
- Power of Open Source

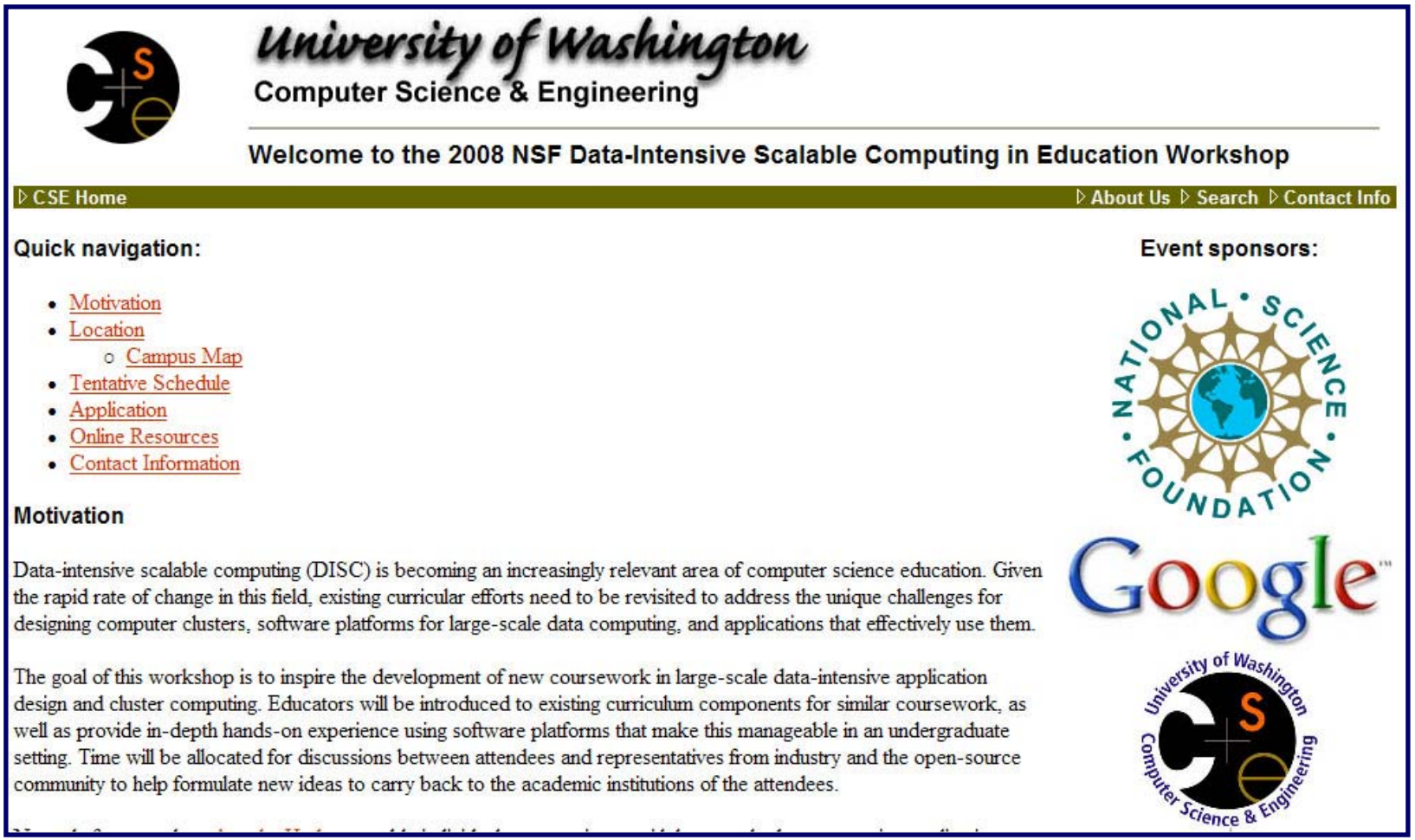
Data-Intensive Computing Symposium

- ~100 from universities, companies, govt. labs, NSF
- 14 invited speakers
  - Google, Yahoo!, Microsoft, Intel
  - CMU, UC Berkeley, Cornell, MIT, Johns Hopkins, UIUC, UW
  - NSF

NSF Involvement



Curriculum Development



University of Washington
Computer Science & Engineering

Welcome to the 2008 NSF Data-Intensive Scalable Computing in Education Workshop

CSE Home | About Us | Search | Contact Info

Quick navigation:

- Motivation
- Location
  - Campus Map
- Tentative Schedule
- Application
- Online Resources
- Contact Information

Event sponsors:

Motivation

Data-intensive scalable computing (DISC) is becoming an increasingly relevant area of computer science education. Given the rapid rate of change in this field, existing curricular efforts need to be revisited to address the unique challenges for designing computer clusters, software platforms for large-scale data computing, and applications that effectively use them.

The goal of this workshop is to inspire the development of new coursework in large-scale data-intensive application design and cluster computing. Educators will be introduced to existing curriculum components for similar coursework, as well as provide in-depth hands-on experience using software platforms that make this manageable in an undergraduate setting. Time will be allocated for discussions between attendees and representatives from industry and the open-source community to help formulate new ideas to carry back to the academic institutions of the attendees.

- **Workshop for educators July 16–18, 2008**

Christophe Bisciglia

- UW/Google
- Catalyst / instigator

Future Workshops

CCA-08:

Cloud Computing and Its Applications

HOME

More details will be provided shortly.

E.D.S.' service management center in Plano, Tex. (Photo: Electronic Data Systems)

Organizing Committee

Charlie Catlett, Argonne National Laboratory

Ian Foster, Argonne and University of Chicago

Joe Hellerstein, University of California Berkeley

October 22 and 23, 2008

Gleacher Center
Chicago, IL

Dramatic growth in data and equally rapid decline in the cost of highly integrated clusters has spurred the emergence of the data center as the platform of choice for a growing class of data-intensive applications. To encourage conversations between those developing applications, algorithms, software, and hardware for such "cloud" platforms, we are convening the first workshop on Cloud Computing and its Applications (CCA'08).

This workshop will include a mixture of invited and contributed talks on cloud computing, data intensive scalable computing, and related topics.

Topics of interest include:

- compute and storage cloud architectures and implementations
- map-reduce and its generalizations
- programming models and tools
- novel data-intensive computing applications
- data intensive scalable computing
- distributed data intensive computing

Concluding Thoughts

The World is Ready for a New Approach to Large-Scale Computing

- **Optimized for data-driven applications**
- **Technology favoring centralized facilities**
  - Storage capacity & computer power growing faster than network bandwidth

Industry is Catching on Quickly

- **Large crowd for Hadoop Summit**

University Researchers / Educators Eager to Get Involved

- **Spans wide range of CS disciplines**
- **Across multiple institutions**