

## Tuan A. Cao

4104 Upson Hall  
Computer Science Department  
Cornell University, Ithaca, NY 14853

Phone: 1 607 280 9867  
tuancao@cs.cornell.edu  
<http://www.cs.cornell.edu/~tuancao/>

### RESEARCH INTERESTS

Large-scale data management with particular interests in cloud computing infrastructure

### EDUCATION

**Ph.D. Candidate**, Computer Science, Cornell University Fall 2007 - Present  
Advisor: Prof. Johannes Gehrke  
**B.E.**, Computer Science and Engineering, Pune University, India May 2005  
Grade: Distinction. Rank: 2/120 in Computer Science Department  
(Full Scholarship from Indian Government)

### EMPLOYMENT

**Summer Intern**, Google Inc., Mountain View, USA Summer 2008, Summer 2012  
**Software Engineer**, Yahoo! Inc., Bangalore, India Jul. 2006 - Mar. 2007

### PATENTS

Inventors: Nitin Gupta, Sreeram Balakrishnan, **Tuan Cao**, Jayant Madhavan, Alon Halevy  
Title: "Searching for Join Candidates" Submitted, Summer 2012

### HONORS AND AWARDS

Member of Vietnamese Delegation at the Asia Pacific ICT Awards, APICTA 2006  
Awarded 2nd Price, "Vietnamese Intelligence" Contest 2005  
Awarded 1st Price, "National Level Undergrad Software Contest", IIT Delhi, IISc Bangalore 2005  
Awarded The Best Vietnamese Student, Pune University, India 2002 - 2004  
Awarded Cultural Exchange Program (CEP) scholarship, Indian Government 2001

### TEACHING EXPERIENCES

CS 4320: Introduction to Database Systems Fall 2011  
CS 5300: Architecture of Large-Scale Information Systems Spring 2009  
CS 419: Computer Networks Spring, Fall 2008  
CS 414: Operating Systems Fall 2007

### PUBLICATIONS

- **T. Cao**, M. V. Salles, B. Sowell, Y. Yue, A. Demers, J. Gehrke, W. White. "Fast Checkpoint Recovery Algorithms for Frequently Consistent Applications." SIGMOD 2011, Athens, Greece.
- **T. Cao**, B. Sowell, M. V. Salles, A. Demers, J. Gehrke. "BRRL: A Recovery Library for Main-Memory Applications in the Cloud (Demo Paper)." SIGMOD 2011, Athens, Greece.
- G. Wang, M. V. Salles, B. Sowell, X. Wang, **T. Cao**, A. Demers, J. Gehrke, W. White. "Behavioral Simulations in MapReduce." VLDB 2010, Singapore.
- M. V. Salles, **T. Cao**, B. Sowell, A. Demers, J. Gehrke, C. Koch, W. White. "An Evaluation of CheckpointRecovery for Massively Multiplayer Online Games." VLDB 2009, Lyon France.
- H. Ballani, P. Francis, **T. Cao**, J. Wang. "Making Routers Last Longer with ViAggre." NSDI 2009, Boston USA.
- H. Ballani, P. Francis, **T. Cao**, J. Wang. "ViAggre: Making Routers Last Longer!" HotNets 2008, Vancouver Canada.
- **T. Cao**, V. Kapoor. "Transparent Network and Transport Layer Normalizer." ICTACS 2006, Vietnam.

## RESEARCH PROJECTS AT CORNELL

### **Fault Tolerance for Large-scale Cloud Computing Applications**

2008 - Present

This research takes advantage of frequent points of consistency in long-running applications to develop novel checkpoint recovery algorithms that trade additional memory for significantly lower overhead and latency. Experimental evaluations show that one of our new algorithms attains nearly constant latency and reduces overhead by more than an order of magnitude for low to medium update rates. Additionally, in a heavily loaded main-memory transaction processing system, it still reduces overhead by more than a factor of two.

In light of these findings, we have developed BRRL, a recovery library for frequently consistent main-memory applications in the cloud. BRRL differs from existing recovery libraries by providing a simple table abstraction and using schema information to optimize checkpointing. The work was published at VLDB 2009 and at SIGMOD 2011.

### **Behavioral Simulations in MapReduce**

2009 - 2010

While there has been a great deal of work on simulation tools from the high-performance computing community, behavioral simulations remain challenging to program and automatically scale in parallel environments. We have designed and implemented BRACE (Big Red Agent-based Computation Engine), which extends the MapReduce framework to process these simulations efficiently across a cluster. BRACE also includes a high-level language called BRASIL (the Big Red Agent SIMulation Language) that has object-oriented features for programming simulations. BRASIL can be compiled to a dataflow representation for automatic parallelization and optimization. The work was published at VLDB 2010.

### **Virtual Aggregation**

2007 - 2008

Virtual Aggregation is a “configuration-only” approach to shrinking the routing table on routers. It does not require any changes to router software or routing protocols and can be deployed independently and autonomously by any ISP. The key design is to use a divide-and-conquer technique that allows an ISP to modify its internal routing such that individual routers in its network only maintain a part of the global routing table. Evaluations show that Virtual Aggregation can reduce routing-table size on individual routers by one order of magnitude while imposing almost no traffic stretch and very little increase in router load. The work was published at NSDI 2009.

## INTERNSHIP PROJECTS AT GOOGLE

### **Merge Evaluation**

Fusion Tables, Summer 2012

Merge is a newly born search engine that exploits structural properties of tables to improve search quality. The objective of Merge is to provide relevant tables which add substantial information to a search query. Unlike traditional search systems, Merge provides search in the context of a table.

My internship project consists of two parts. First, I designed and implemented an evaluation framework which helps developers to improve Merge ranking mechanisms; in particular, the Merge system is benchmarked primarily using this framework. Second, I have developed a probabilistic retrieval model to rank Merge candidates, and I will continue working on it remotely with my host after the internship.

### **Caribou Analysis**

Gmail Backend, Summer 2008

My intern project was the design and implementation of a map-reduce generic framework which is capable of analyzing petabytes of data from various data centers at Google. The framework ran five times faster and required only half of the resources compared to the previous approaches.

## REFERENCES

**Johannes Gehrke**

Professor  
Dept. of Computer Science  
Cornell University  
johannes@cs.cornell.edu

**Alan Demers**

Principal Research Scientist  
Dept. of Computer Science  
Cornell University  
ademers@cs.cornell.edu

**Nitin Gupta**

Research Scientist  
Structured Data Group  
Google Inc.  
nigupta@google.com