Summary
This workshop will introduce the building blocks of large-scale cloud computing to the design automation community. Tremendous advances have been made in the creation of new EDA technologies to design complex multi-billion transistor chips. However, EDA software tools are still built as monolithic software components that only recently have started building parallelism into the software, with multi-threaded implementations on multi-core processor systems. In contrast, the software systems run by companies such as Yahoo, Google and Facebook employ building blocks that are designed from the ground-up to be highly distributed and scalable. These building blocks perform distributed computation (Map-Reduce, Hadoop), distributed key-value databases (Membase, NoSQL), distributed storage (Hadoop Distributed File System, Google File System), and run on hundreds to thousands of machines.

The workshop will start with an overview on the evolution of scalable systems which have evolved into massive clusters with commodity components and the new paradigm of Data-Intensive Scalable Computing. It will be followed by presentations on Hadoop, an open source implementation of Map-Reduce, Membase, an open source distributed, key-value database management system, and distributed file systems that provide a storage framework for these large distributed data-intensive applications. The presentations will cover these technologies and their applications to inspire innovative ways of building new design automation solutions.

Organizers
Manish Pandey - Synopsys, Inc., Mountain View, CA
Andreas Kuehlmann - Coverity, Inc., San Francisco, CA

Speakers
Dr. Randal E. Bryant - Carnegie Mellon University, Pittsburgh, PA
Dr. Kazi A. Zaman - Yahoo!, Inc., Sunnyvale, CA
Frank Weigel - Couchbase, Mountain View, CA

Agenda
1:00pm Workshop Introduction – Dr. Manish Pandey
1:10pm Data-Intensive Scalable Computing – Dr. Randal E. Bryant
1:40pm Q&A
1:50pm Evolving Hadoop -- Dr. Kazi A. Zaman
2:40pm Q&A
2:50pm Scaling Data in the Cloud – Frank Weigel
3:30pm Q&A
3:40pm Concluding Remarks
Presentation Abstracts

Title: Data-Intensive Scalable Computing  
Speaker: Dr. Randal E. Bryant

Abstract: Web search engines have become fixtures in our society, but few people realize that they are actually publicly accessible supercomputing systems, where a single query can unleash the power of several hundred processors operating on a data set of over 200 terabytes. With Internet search, computing has risen to entirely new levels of scale, especially in terms of the sizes of the data sets involved. Google and its competitors have created a new class of large-scale computer systems, which we label "Data-Intensive Scalable Computer" (DISC) systems. DISC systems differ from conventional supercomputers in their focus is on data: they acquire and maintain continually changing data sets, in addition to performing large-scale computations over the data. DISC points the way to new ways of organizing large-scale computing systems to be more robust, scalable, and cost effective than are current high-performance computing systems.

Programs for DISC systems must be written in ways that allows them to be executed in a loosely-coupled asynchronous environment, such as the Map/Reduce framework pioneered by Google. Although Map/Reduce has surprisingly broad applicability, a richer set of programming languages and models is required to realize the full potential of DISC.

Title: Evolving Hadoop  
Speaker: Dr. Kazi A. Zaman

Abstract: In this talk, we trace the evolution of Apache Hadoop, a framework for large scale distributed computation, from its roots in search technology to being a highly scalable multitenant service powering enterprise class data applications. Apache Hadoop has two main components. HDFS, the Hadoop Distributed File System, designed to reliably store petabytes of data on commodity hardware and the Map Reduce programming framework. We cover the architecture of these two components with a focus on recent work in the area of scalability. We also examine how Hadoop is used for a diverse set of application classes including Data Pipelines/ETL, Machine Learning (e.g. Spam Filtering), Warehousing/Analytics and content personalization.

Title: Scaling Data in the Cloud  
Speaker: Frank Weigel

Abstract: The economics of cloud/distributed computing are compelling, and in many industries increasing numbers of application developers and app owners are embracing the model. A critical consideration in the move towards cloud systems is how to manage and scale your data across a distributed infrastructure. Typically, RDBMS technology is problematic for the cloud because of its centralized, "scale up" model, which simply doesn’t provide the scalability and performance required by most modern software systems. New technologies have recently emerged that provide a “scale out” data model that delivers the elasticity and flexibility needed for distributed environments. In this session, Frank Weigel focuses on elastic data management solutions (those that scale-out), which hold the promise of enabling a fully-automated cloud that can seamlessly scale both application logic and the data behind the application. He will also provide real-world examples of companies that are using these technologies to achieve the scalability and performance needed to serve millions of users with their cloud-based applications.
Dr. Randal E. Bryant

Dr. Randal E. Bryant is Dean of the Carnegie Mellon University School of Computer Science. He has been on the faculty at Carnegie Mellon since 1984, starting as an Assistant Professor and progressing to his current rank of University Professor of Computer Science. He also holds a courtesy appointment in the Electrical and Computer Engineering Department.

Much of Dr. Bryant's research has focused on methods for formally verifying digital hardware, and more recently some forms of software. His 1986 paper on symbolic Boolean manipulation using Ordered Binary Decision Diagrams (BDDs) has one of the highest citation count of any publication in the computer science literature. More recently, he has become interested in the opportunities and challenges presented by computer systems working with very large data sets.

Dr. Bryant has received widespread recognition for his work. He is a fellow of the IEEE and the ACM, as well as a member of the National Academy of Engineering and the American Academy of Arts and Sciences. His awards include the 2007 IEEE Piore Award, the 1997 ACM Kanellakis Theory and Practice Award (shared with Edmund M. Clarke, Ken McMillan, and Allen Emerson) for contributing to the development of symbolic model checking, as well as the 1989 IEEE W.R.G. Baker Prize for the best paper appearing in any IEEE publication during the preceding year. In the field of electronic design automation, he has won both the IEEE/CEDA Phil Kaufman Award and the ACM/IEEE A. Richard Newton technical impact award.

Dr. Bryant received his B.S. in Applied Mathematics from the University of Michigan in 1973, and his PhD from MIT in 1981. He was on the faculty at Caltech from 1981 to 1984.

Dr. Kazi A. Zaman

Dr. Kazi A. Zaman is the Director of Engineering for Hadoop. He is responsible for managing the engineering team at Yahoo! working on Apache Hadoop (HDFS and Map Reduce). Hadoop runs on over 40,000 machines at Yahoo! for a diverse set of business critical applications ranging from anti spam algorithms for mail and content recommendation algorithms for the Yahoo! front page to warehousing and analytic applications processing petabyes of data. Previously he worked as a Staff Engineer at Siebel Systems where he worked on the core query processing capabilities of the Siebel Analytics Server (now the Oracle BIEE server). He received his bachelor's degree in Computer Science & Engineering from IIT Kharagpur and subsequently completed his Ph.D. in Computer Science at Columbia University.

Frank Weigel

Frank Weigel is Director of Product Management at Couchbase, the company formed by the recent merger of Membase and CouchOne. He is responsible for the company's core NoSQL database products, and works with customers and users around the world to understand emerging requirements for low-latency, scalable data stores. He started his career 10 years ago in software engineering, primarily focused on dynamic binary translators, before crossing to the dark side and telling others what their code should do.