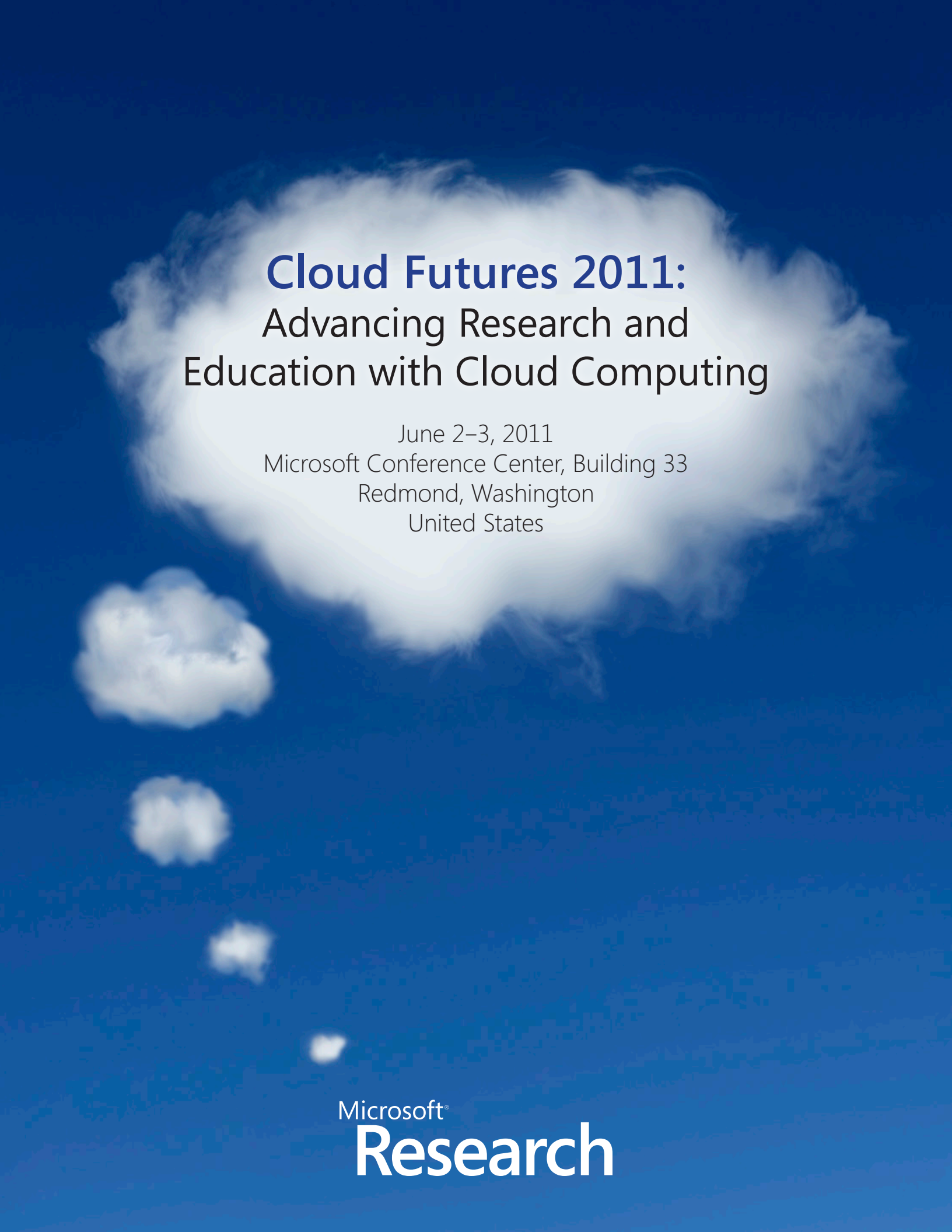
Cloud Futures 2011:
Advancing Research and Education with Cloud Computing

June 2–3, 2011
Microsoft Conference Center, Building 33
Redmond, Washington
United States



Microsoft® Research

Cloud Futures 2011

Advancing Research and Education with Cloud Computing

June 2–3, 2011
Microsoft Conference Center, Building 33
Redmond, Washington
United States

The Cloud Futures Workshop series brings together thought leaders from academia, industry, and government to discuss the role of cloud computing across a variety of research and educational areas—including computer science, earth sciences, healthcare, humanities, life sciences, and social sciences. Presentations and discussions will highlight how new techniques, software platforms, and methods of research and teaching in the cloud may solve distinct challenges arising in those diverse areas.

About the Workshop

Cloud computing is an exciting platform for research and education. It has the potential to advance scientific and technological progress by making data and computing resources readily available at unprecedented economy of scale and nearly infinite scalability. To realize the full promise of cloud computing for research and education, however, we must think about the cloud as a holistic platform for creating new services, new experiences, and new methods to pursue research, teaching, and scholarly communication. This goal presents a broad range of interesting questions.

This workshop will include presentations that illustrate the role of cloud computing across a variety of research and educational areas—including computer science, engineering, earth sciences, healthcare, humanities, life sciences, and social sciences. Attendees will have the opportunity to learn how new techniques and methods of research in the cloud may solve distinct challenges arising in those diverse areas.

http://research.microsoft.com/cloudfutures2011

Microsoft Research

Agenda

THURSDAY, JUNE 2, 2011

7:45AM–8:30AM Transportation from Bellevue Westin Hotel to Microsoft Conference Center

8:45AM–9:00AM Welcome and Conference Logistics

9:00AM–10:00AM **KEYNOTE** | **Cascade**

- **The Cloud Will Change Everything**
  – Jim Larus, Microsoft Research, eXtreme Computing Group

10:00AM–10:30AM Coffee break

10:30AM–12:00PM **SYSTEMS 1** | **Cascade**

- **Cloud, HPC, or Hybrid: A Case Study Involving Satellite Image Processing**
  – Marty Humphrey, University of Virginia
- **Classical and Iterative MapReduce on Azure**
  – Judy Qiu, Thilina Gunarathne, Geoffrey Fox, Pervasive Technology Institute and School of Informatics and Computing, Indiana University
- **Data Semantics Aware Clouds for High-Performance Analytics Systems**
  – Jun Wang, the University of Central Florida

EDUCATION | **St. Helens**

- **Bringing the Cloud to the Classroom Using the Microsoft Imagine Cup**
  – James L. Parrish, Jr., University of Arkansas at Little Rock
- **Introducing Cloud Computing into STEM Undergraduate Curriculum Using Microsoft Azure**
  – Bina Ramamurthy, University at Buffalo
- **InstantLab—The Cloud as Operating System Teaching Platform**
  – Alexander Schmidt and Andreas Polze, Hasso Plattner Institute at University of Potsdam

12:00PM–1:00PM Lunch

1:00PM–2:00PM **KEYNOTE** | **Cascade**

- **Making Sense at Scale with Algorithms, Machines and People**
  – Michael J. Franklin, AMPLAB Director University of California, Berkeley

2:00PM–3:30PM **SYSTEMS 2** | **Cascade**

- **Relational Data Markets in the Cloud: Challenges and Opportunities**
  – Magdalena Balazinska, University of Washington
- **Into the Blue: Streaming Data Processing in the Cloud**
  – Stephen Wong, Rice University
- **Stork Data Scheduler for Windows Azure**
  – Tevfik Kosar, State University of New York (SUNY) at Buffalo

WINDOWS AZURE TUTORIAL, Part 1 | **St. Helens**

- **Windows Azure Distilled**
  – Krishna Kumar, Windows Azure Academic Lead, Microsoft Corporation

3:30PM–4:00PM Break

4:00PM–5:30PM **APPLICATIONS 1** | **Cascade**

- **Using the Cloud to Model and Manage Large Watershed Systems**
  – Marty Humphrey, University of Virginia
- **Large Scale Prediction of Transcription Factor Binding Sites for Gene Regulation using Cloud Computing**
  – Zhengchang Su, University of North Carolina
- **University of Southern California, Clever Transportation Project (USC2T)**
  – Barak Fishbain, University of Southern California

WINDOWS AZURE TUTORIAL, Part 2 | **St. Helens**

- **Windows Azure Distilled**
  – Krishna Kumar, Windows Azure Academic Lead, Microsoft Corporation

5:30PM–6:30PM Shuttle to Bellevue Westin Hotel

6:30PM–9:00PM Reception and dinner at Palomino, in Bellevue (www.palomino.com)

FRIDAY, JUNE 3, 2011

7:45AM–8:30AM Transportation from Bellevue Westin Hotel to Microsoft Conference Center

8:30AM–9:00AM Breakfast | Rainier

9:00AM–10:00AM **KEYNOTE** | **Cascade**

- **Inside Windows Azure, Microsoft's Cloud OS**
  – Mark Russinovich, Technical Fellow, Windows Azure Fabric, Microsoft Corporation

10:00AM–10:30AM Break

10:30AM–12:00PM **SYSTEMS 3** | **St. Helens**

- **Multi-Cloud and Cloud-Desktop Coordination Made Simplified by GXP on Azure**
  – Kenjiro Taura and Ting Chen, University of Tokyo
- **Running Large Workflows in the Cloud**
  – Paul Watson, Newcastle University
- **Towards Enabling Mid-Scale Geo-Science Experiments Through Microsoft Trident and Windows Azure**
  – Eran Chinthaka Withana and Beth Plale, Indiana University

APPLICATIONS 2 | **Cascade**

- **Network Display and Sharing System for Large Cultural Heritage Objects via Cloud Computing Environments**
  – Yasuhide Okamoto, University of Tokyo
- **Deep Natural Language Processing for Improving a Search Engine Infrastructure using Windows Azure**
  – Daisuke Kawahara, Kyoto University
- **A Secure Multi-Party Collaboration System for Australian Coal Supply Chains**
  – Shiping Chen and Chen Wang, CSIRO ICT Centre

12:00PM–1:00PM Lunch | Rainier

1:00PM–2:30PM **APPLICATIONS 3** | **St. Helens**

- **Beekeeping Enters the Cloud**
  – James T. Wilkes, Appalachian State University
- **Scaling Document Clustering on the Cloud**
  – Rob Gillen, Oak Ridge National Laboratory
- **Microsoft Azure: The Ultimate Flexible Enterprise-Level Solution**
  – Janet L. Bailey, University of Arkansas at Little Rock

SECURITY AND SOFTWARE DEVELOPMENT | **Cascade**

- **Risk Assessment and Cloud Strategy Development**
  – Barbara Endicott-Popovsky, University of Washington and Kirsten Ferguson-Boucher, Aberystwyth University, Wales
- **Cloud Forensics: an Overview**
  – Keyun Ruan, University College Dublin
- **Cloud Based Product Development**
  – Joris Poort, Harvard Business School

2:30PM–3:00PM Break

3:00PM–4:30PM **SYSTEMS 4** | **St. Helens**

- **Expanding the Horizons of Cloud Computing Beyond the Data Center**
  – Jon Weissman and Abhishek Chandra, University of Minnesota
- **URSA: Scalable Load Balancing and Power Management in Cluster Storage Systems**
  – Seung-won Hwang, Pohang University of Science and Technology (POSTECH)
- **Achieving Energy Efficient Computing by Jointly Scheduling Services and Batch Jobs in Virtualized Environments**
  – Tajana Simunic Rosing, University of California, San Diego

APPLICATIONS 4 | **Cascade**

- **High Through-Put, Low Impedance e-Science on Microsoft Azure**
  – David Abramson, Monash University
- **Abstractions for Life-Science Applications on Clouds**
  – Shantenu Jha, Andre Luckow, Rutgers University
- **The Parallelization of Geoscience Apps at C3L with Azure**
  – Craig Mudge, University of Adelaide

4:30PM–5:30PM **CLOUD RESEARCH PANEL** | **Cascade**

– **Chair:** Ed Lazowska

Table of Contents

Table of Contents continued

SECURITY AND SOFTWARE DEVELOPMENT

SYSTEMS

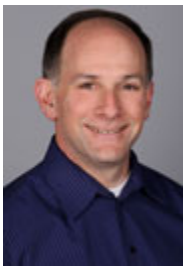
WINDOWS AZURE TUTORIAL

The Cloud Will Change Everything

Jim Larus, Microsoft Research, eXtreme Computing Group

"Cloud computing" is fast on its way to becoming a meaningless, oversold marketing slogan. In the midst of this hype, it is easy to overlook the fundamental change that is occurring. Computation, which used to be confined to the machine beside your desk, is increasingly centralized in vast shared facilities and at the same time liberated by battery-powered, wireless devices. Performance, security, and reliability are no longer problems that can be considered in isolation – the wires and software connecting pieces offer more challenges and opportunities than components themselves. The eXtreme Computing Group (XCG) in Microsoft Research is taking a holistic approach to research in this area, by bring together researchers and developers with expertise in data center design, computer architecture, operating systems, computer security, programming language, mobile computation, and user interfaces to tackle the challenges of cloud computing.

BIOGRAPHY:



James Larus *is a Director of the eXtreme Computing Group (XCG) in Microsoft Research. Larus has been an active contributor to the programming languages, compiler, and computer architecture communities. He joined Microsoft Research as a Senior Researcher in 1998 to start and lead the Software Productivity Tools (SPT) group, which developed and applied a variety of innovative techniques in static program analysis and constructed tools that found defects (bugs) in software. He also started and lead the Singularity project, which pioneered a new architecture for safe system code. Before joining Microsoft, Larus was an Assistant and Associate Professor of Computer Science at the University of Wisconsin-Madison. Larus became an ACM Fellow in 2006.*

Making Sense at Scale with Algorithms, Machines and People

Michael J. Franklin, AMPLAB Director, University of California, Berkeley

The creation, analysis, and dissemination of data have become profoundly democratized. Social networks spanning 100's of millions of users enable instantaneous discussion, debate, and information sharing. Streams of tweets, blogs, photos, and videos identify breaking events faster and in more detail than ever before. Deep, on-line datasets enable analysis of previously unreachable information. This sea change is the result of a confluence of Information Technology advances such as: intensively networked systems, cloud computing, social computing, and pervasive devices and communication.

The key challenge is that the massive scale and diversity of this continuous flood of information breaks our existing technologies. State-of-the-art Machine Learning algorithms do not scale to massive data sets. Existing data analytics frameworks cope poorly with incomplete and dirty data and cannot process heterogeneous multi-format information. Current large-scale processing architectures struggle with diversity of programming models and job types and do not support the rapid marshalling and unmarshalling of resources to solve specific problems. All of these limitations lead to a Scalability Dilemma: beyond a point, our current systems tend to perform worse as they are given more data, more processing resources, and involve more people — exactly the opposite of what should happen.

To address these issues, a group of us from machine learning, systems, databases, and networking just started a new five-year research effort called the AMPLab, where AMP stands for "Algorithms, Machines, and People". AMPLab envisions a world where massive data, cloud computing, communication and people resources can be continually, flexibly and dynamically be brought to bear on a range of hard problems by huge numbers of people connected to the cloud via mobile and other client devices of increasing power and sophistication. The Founding Sponsors of the AMPLab are Google and SAP. Amazon Web Services, Cloudera, Ericsson, Huawei, IBM, Intel, Microsoft, NEC, NetApp, Splunk, and VMWare are also sponsors.

In this talk, Michael Franklin will give an overview of the AMPLab motivation and research agenda and discuss several of our initial projects. One such project, CrowdDB, is developing infrastructure to support hybrid cloud/crowd query answering systems - leveraging the very different skills, and performance, reliability, and cost characteristics of large groups of machines and large groups of people.

BIOGRAPHY:

Michael Franklin *is a Professor of Computer Science at UC Berkeley, focusing on new approaches for data management and data analysis. He is a director of the Algorithms, Machines and People Laboratory (AMPLab), a multidisciplinary lab working at the intersection of Machine Learning, Large-Scale Data Management and Crowdsourcing. He founded and is CTO of Truviso, Inc. a real-time data analytics company that enables customers to quickly make sense of diverse, high-speed, continuous streams of information. He is a Fellow of the Association for Computing Machinery, and a recipient of the National Science Foundation CAREER award and the ACM SIGMOD "Test of Time" award.*

He is currently serving as a committee member on the National Academy of Science study on Analysis of Massive Data, and is an editorial board member for the ACM Journal of Data Quality. He is a Fellow of the Association for Computing Machinery, and a recipient of the National Science Foundation CAREER award and the ACM SIGMOD "Test of Time" award. He received his Ph.D. from the University of Wisconsin in 1993.

Inside Windows Azure, Microsoft's Cloud OS

Mark Russinovich, Technical Fellow, Windows Azure Fabric, Microsoft Corporation

Mark Russinovich, working on the Windows Azure team, presents an under the hood tour of the internals of Microsoft's new cloud OS. Windows Azure is that it delivers a platform that enables developers to quickly develop elastic, highly available web service applications. Mark will explain how Windows Azure implements the foundation for such applications, covering datacenter architecture, cloud OS architecture, and what goes on behind the scenes when Windows Azure deploys a service, a machine fails or comes online and a role fails. Mark will also give some insight into the direction of Windows Azure's future evolution.

BIOGRAPHY

Mark Russinovich *is a Technical Fellow in the Windows Azure group at Microsoft working on Microsoft's datacenter operating system. He is a widely recognized expert in Windows operating system internals as well as operating system security and design. He is author of the recently published cyberthriller Zero Day, co-author of the Microsoft Press Windows Internals books, and co-author of the forthcoming Sysinternals Administrator's Reference. Russinovich joined Microsoft in 2006 when Microsoft acquired Winternals Software, the company he cofounded in 1996, as well as Sysinternals, where he authors and publishers dozens of popular Windows administration and diagnostic utilities. He is a featured speaker at major industry conferences including Microsoft's TechEd, WinHEC, and Professional Developers Conference.*

Using the Cloud to Model and Manage Large Watershed Systems

Jon Goodall, University of South Carolina
Marty Humphrey, University of Virginia

Understanding hydrologic systems at the scale of large watersheds is of critical importance to society when faced with extreme events such as floods and droughts, or with minimizing human impacts on water quality. Climate change and increasing population are further complicating watershed-scale prediction by placing additional stress and uncertainty on future hydrologic system conditions. New data collection and management approaches are allowing models to capture water flow through built and natural environments at an increasing level of detail. A significant barrier to advancing hydrologic science and water resource management is insufficient computational infrastructure to leverage these existing and future data resources within simulation models.

We have recently been awarded a National Science Foundation (NSF) "Computing in the Cloud" grant to advance hydrologic science and water resource management by leveraging cloud computing for modeling large watershed systems. We will use Windows Azure in three ways. First, we will create a cloud-enabled hydrologic model. Second, we will improve the process of hydrologic model parameterization by creating cloud-based data processing workflows. Third, in Windows Azure, we will apply the model and data processing tool to a large watershed in order to address a relevant hydrologic research question related to quantifying impacts of climate change on water resources.

In this talk, we first give an overview of our project. Then, we present the specifics of how we intend to use Windows Azure and discuss issues that we anticipate will arise as we pursue the goals of our research project.

BIOGRAPHIES

Jon Goodall *is an Assistant Professor in the Department of Civil and Environmental Engineering at the University of South Carolina (2007 – present). He received his Ph.D. at the University of Texas at Austin, Civil Engineering in 2005. Before joining the University of South Carolina, Dr. Goodall was an Assistant Professor of the Practice at Duke University (2005-2007). Dr. Goodall's area of specialization is water resource engineering with emphasis on surface water hydrology. His research interests include the application of Geographic Information Systems (GIS) to hydrology, watershed modeling and management, and the design of cyberinfrastructure to advance hydrologic science and water resource management.*

Marty Humphrey *is an Associate Professor in the Department of Computer Science at the University of Virginia. He received a B.S. and M.S. degree in Electrical Engineering from Clarkson University in 1986 and 1989, respectively. He received his Ph.D. degree in computer science from the University of Massachusetts in 1996. He has been on the faculty of the University of Virginia since 1988. Dr. Humphrey's research interests include the use of Cloud computing for cancer research and for environmental research.*

Large Scale Prediction of Transcription Factor Binding Sites for Gene Regulation using Cloud Computing

Zhengchang Su, University of North Carolina

Although tremendous advances have been made in identifying the gene-coding DNA sequences in bacterial genomes using computational methods, our understanding of regulatory DNA sequences is very limited due to the lack of efficient computational methods for predicting them. Regulatory sequences specify when, how much, and where the genes should be expressed in the cell through their interactions with small proteins called transcription factors (TFs). Therefore, identifying these sequences, also called TF binding sites (TFBS), in a genome is as important as identifying gene-coding sequences for understanding the biology of the cell. Rapid recent advances in genome sequencing technology are dramatically reducing the time and cost of sequencing a genome. Over 1,500 bacterial genomes have been sequenced and this number is rising exponentially. Our very limited understanding of the gene regulatory systems in sequenced prokaryotic genomes has largely hindered our understanding of their biology and applications in renewable energy production and environment protection as well as the prevention of the diseases they cause. To fill in this gap, we have recently developed an efficient and accurate algorithm for predicting TFBSs in a group of related genomes, and have parallelized it on an in-house cluster using MPI. Although this algorithm can potentially predict TFBSs in a few thousand genomes, its capability will soon be dwarfed by the sequencing of hundreds of thousands genomes as a result of the on going world-wide efforts to sample various microbiomes using new sequencing technologies. Cloud computing holds promise to overcome the computational and storage challenges for predicting TFBSs in all sequenced genomes in the future. I will present our preliminary results to port our algorithm on the Microsoft Azure Cloud Platform as an attempt to achieve such a goal.

BIOGRAPHY

Dr. Zhengchang Su *is an assistant professor at the Department of Bioinformatics and Genomics, the University of North Carolina at Charlotte (UNCC). Dr. Su received his Ph.D. in biophysics from the University of Alabama at Birmingham in 2000. Before joining the faculty of UNCC, he was a research scientist at University of Georgia in Athens and a postdoctoral fellow at Oak Ridge National Labs. Dr. Su's recent research focuses on developing algorithms and software to understand gene transcriptional regulation in bacteria and animals.*

University of Southern California, Clever Transportation Project (USC2T)

Barak Fishbain, University of Southern California

Creating a large, dynamic database that integrates transportation data from multiple sources is unprecedented. While there are many examples of university based programs that archive data from the state highway system (e.g. University of California, University of Virginia) and use the data for traffic management research, typically these databases consist of one type of transportation data and are small in size. The University of Southern California (USC) with Microsoft have taken the great challenge to develop the data management and analytics means that will constitute the building of a large scale transportation data warehouse for heterogeneous transportation data (e.g., traffic flows, recorded by loop detectors; police reports; videos and images; and operational public transit data, such as passenger counts or buses' locations), which is collected from all transportation agencies in Southern California.

The collection of the traffic data, its refinement as well as the required geostreaming queries are implemented on Microsoft's StreamInsight. Then the data is saved to Azure Tables storage space. In order to allow processing of queries to such large data set efficiently, the data is aggregated to support predefined set of spatial and temporal queries. A StreamInsight server, which resides on the Azure AppFabric, handles this aggregation process. The sketches are then written to SQLAzure. Other than infrastructural benefits, this eliminates communication and data transfer costs to a cloud storage platform.

The wide and flexible range of storage offers, made by Windows and SQL Azure, as well as the desire to guarantee reduced system cost for end corporate users, call for finding a cost efficient blend of storage infrastructures which are well tuned for consistency and concurrency requirements of our data intensive sensor network application. Utilizing this whole set of technologies can bring even more profit to the data owner, in case a decision is made to make it available through the Microsoft Data Market.

BIOGRAPHY

Barak Fishbain *is the Associate Director of University of Southern California's Integrated Media Systems Center (IMSC) and Research Associate at the Viterbi School of Engineering. His research focuses on pattern recognition in multidimensional data sets. This includes traffic safety and traffic data, sensor networks, images, videos and medical imaging. Fishbain earned his PhD at Tel-Aviv University's School of Electrical Engineering (2008) where he also holds an MS in Electrical Engineering (2004). His BS degree in electrical engineering is from Technion – Israel's Institute of Technology (1998).*

Network Display and Sharing System for Large Cultural Heritage Objects via Cloud Computing Environments

Yasuhide Okamoto, University of Tokyo

Recent advances in sensing and software technologies enable us to obtain large-scale, yet fine 3D mesh models of cultural heritage assets such as old statues, buildings, and entire excavation sites. By using latest laser range sensors specialized for various environments, and sophisticated parallel computing techniques, we can construct high definition 3D models of those precious objects easily and quickly. Such archived data can be utilized for many kinds of purposes, such as archaeology, education, and entertainment. However, such large 3D models cannot be accessed interactively from general users through the Internet because of the limitation of the computer performance and network bandwidth. Cloud computing technology is a solution that can process a very large amount of information without adding to each client user's processing cost. In this presentation, we propose an interactive display and sharing system for large 3D mesh models, stored on a remote environment through a network of relatively small capacity machines, based on the cloud computing concept. Our system uses both model- and image-based rendering methods for efficient load balance between a server and clients. On the server, the 3D models are rendered by the model-based method using a hierarchical data structure with Level of Detail (LOD). On the client, an arbitrary view is constructed by using a novel image-based method, referred to as the Grid-Lumigraph, which blends colors from sampling images received from the server. The resulting rendering system can efficiently render any images in real time. We implemented the system and evaluated the rendering and data transferring performance.

BIOGRAPHY

Yasuhide Okamoto *is a postdoctoral scholar of Electrical Engineering and Computer Science, University of California, Berkeley, and a postdoctoral fellow of Institute of Industrial Science, the University of Tokyo, Japan. He received his PhD degree in information science and technology from the University of Tokyo in 2010. He was a project research fellow of Institute of Industrial Science, the University of Tokyo from April 2010 for one year, and currently he is visiting University of California, Berkeley under Young Researcher Overseas Visits Program by JSPS (Japan Society for the Promotion of Science). His research interests are in Computer graphics and Computer vision, in particular, huge mesh rendering, virtual and mixed reality technologies, and human computer interface.*

Deep Natural Language Processing for Improving a Search Engine Infrastructure using Windows Azure

Daisuke Kawahara, Kyoto University

In this talk, I will introduce a search engine infrastructure, TSUBAKI, which is based on deep Natural Language Processing, and describe our experiences of using 10,000 CPU cores on Windows Azure for deepening the indices of TSUBAKI.

While most conventional search engines register only words to their indices, TSUBAKI provides a framework that indexes synonym relations, hypernym-hyponym relations, dependency/case/ellipsis relations and so forth. These indices enable TSUBAKI to capture the semantic matching between a given query and documents more precisely and flexibly.

Among these various relations, case/ellipsis relations have not been indexed in a large scale. This is mainly because the speed of these analyses is not fast enough due to the necessity of referring to a large database of predicate-argument patterns, which is automatically acquired from a large corpus. Thus, we apply case/ellipsis analysis to a huge Web corpus by using the cloud computing environment of Azure.

To apply case/ellipsis analysis to millions of Web pages of TSUBAKI in a practical time, it is necessary to use the order of 10,000 CPU cores. To use 10,000 CPU cores on Azure, it is required to create 29 hosted services of 350 CPU cores due to the limitation of Azure. Each service analyzes Web pages using a master/worker model. In this model, a master manages workers, which concurrently apply our analysis to an input file. This concurrent execution does not need reciprocal communication and can be performed independently.

We implemented the above framework and tested our analysis on 1x350, 2x350 and 8x350 step by step. Finally, we confirmed that we can obtain 29x350 CPU cores and execute our analysis on these CPU cores. Our future work is to run the online search process of TSUBAKI on Azure in addition to the offline indexing process.

BIOGRAPHY

Daisuke Kawahara *received his B.S. and M.S. in Electronic Science and Engineering from Kyoto University in 1997 and 1999, respectively. He obtained his Ph.D. in Informatics from Kyoto University in 2005. He is currently an associate professor of the Graduate School of Informatics at Kyoto University. His research interests include natural language processing, knowledge acquisition and web mining.*

A Secure Multi-party Collaboration System for Australian Coal Supply Chains

Shiping Chen and Chen Wang, CSIRO ICT Centre

A typical transport supply chain in Australian coal industry involves multiple independent business entities that own different resources. For example, the Hunter Valley Coal Chain (HVCC) is the Worlds largest coal export operation, which consists of 35 coal mines owned by 14 producers, 24 load points, 28 trains run by 2 rail operators, tracks own by 2 operators and 2 coal port terminals. The producers and operators are independent and they contract with each other to ensure the transportation resources (trucks, trains, rails, ports etc.) for shipping the coal. As individually negotiated contracts may not lead to an optimal (sometimes even not feasible) resource usage in the whole coal chain, a coordinator's role is needed to efficiently run the coal chain. However, there are obstacles to establish such a role because participants need to make some sensitive information available to the coordinator. The current coal chain practices do not guarantee that the information won't flow to the competitors of the information provider and convincing miners and operators to share information is difficult.

We are developing a cloud-based system to address this issue. The coordinator is hosted in an external party managed platform (offered by an infrastructure as a service) provider such as Microsoft Azure. The system ensures that shared information is only used by the coordinator for the purpose specified by participants. The system states are monitored by participants to ensure compliance. The system requires the implementation of high throughput logging and runtime anomaly detection mechanism. We will report our progress in this talk.

BIOGRAPHIES

Dr. Shiping Chen *received his PhD degree in computer science from the University of New South Wales (UNSW), Australia. He is a senior research scientist at CSIRO ICT Centre working on middleware and distributed systems. He also holds an honorary associate with the University of Sydney through co-supervising PhD & Master's students. He is actively involved in service computing research community through publications and PC member services (WWW, ICSOC, ICWS, SCC, etc.). His current research interests include software architecture, secure data storage, and compliance assurance technologies. He is a member of the IEEE.*

Dr. Chen Wang *received his PhD from Nanjing University, China. He is a senior research scientist in CSIRO ICT Centre, Australia. His research interests are primarily in the area of distributed, parallel and recently trustworthy systems. His recent work focuses on accountability of outsourced services and resource management of cloud computing and the smart grid.*

Beekeeping Enters the Cloud

James T. Wilkes, Appalachian State University

Utilizing cloud computing to address recent declining bee populations is the focus of this talk. Since the winter of 2006 – 2007, overwintering honeybee colonies in the US have died in large numbers. Affected beekeepers span the entire spectrum of the industry: migratory beekeepers to stationary beekeepers; and commercial beekeepers, part-time beekeepers, to backyard beekeepers. If the observed losses continue, they threaten not only the livelihoods of beekeepers who manage bees, but the livelihood of farmers who require bees to pollinate their crops and the food supply from those crops.

Hive Tracks, www.hivetracks.com , and The Bee Informed Partnership www.beeinformed.org ,are two current projects using cloud based solutions to help beekeepers make better hive management decisions. Hive Tracks is a web based record keeping service for beekeepers to assist them in efficient and effective management of their honeybee colonies. The Bee Informed Partnership is a survey based project that uses beekeepers' real world experience to solve beekeepers' real world problems. Through several tiers of increasingly detailed data collection on hive management practices and bee health, this project will adapt the tools developed by human epidemiologists to study honey bee colony health. Its focus will be to identify management practices that keep colonies alive and findings will be shared rapidly, transparently, and in ways that will enable beekeepers to make informed individualized decisions. Data collection, entry, analysis, and dissemination will all be done in the cloud. An overview of the two projects along with a vision for their future growth will be presented.

BIOGRAPHY

Dr. James Wilkes *is a Professor of Computer Science and Chair of the Department of Computer Science at Appalachian State University where he has been a faculty member for 19 years. He has also been around bees his whole life. His father kept several stands of bees in his backyard after purchasing his first bee package through mail order from Sears in 1964. Dr. Wilkes started keeping bees in the year 2000 with three hives and now manages about ninety hives from his family farm, Faith Mountain Farm, www.faithmtnfarm.com in Creston, NC. Dr. Wilkes is cofounder of Hive Tracks along with a fellow beekeeper and software engineer friend. He is the technical lead for the Bee Informed Partnership.*

Scaling Document Clustering on the Cloud

Rob Gillen, Oak Ridge National Laboratory

Cloud computing has gained significant popularity over the past few years and introduces a number of new and compelling capabilities as a computational platform. Beyond the well-established benefits such as the massive scalability of both compute on demand and storage on demand, cloud computing offers the ability to think differently about the problems we aretrying to solve. Rather than facing constraints of fixed limitations of a target computational platform, we are able to develop codes and algorithms that can make intelligent use of the resources available. A particular code may begin to solve a problem on a modest set of hardware and then, as the incoming data set grows or solution evolves, issue calls to the underlying cloud infrastructure to allocate the appropriate increase in hardware. The software can then reconfigure its control structures to accommodate the newly acquired hardware and resume the task of solving the problem at hand.

This talk will discuss early progress in the application of the above techniques to a document clustering algorithm developed at Oak Ridge National Laboratory. The original codes used to solve this problem utilize a memory resident non-binary tree, which causes the problem size to be limited by the amount of physical ram in the machine. The application of the cloud and the techniques described in this talk are allowing the algorithm to span multiple machines in a self-scaling, fault tolerant manner. This talk will provide initial results, lessons learned, future work, as well as a brief comments on utilizing various cloud vendors for the same code base.

BIOGRAPHY

Rob Gillen *is currently a member of the research staff at Oak Ridge National Laboratory where he is working with the Computer Science Research Group (part of the Computer Science and Mathematics Division and the Computing and Computational Sciences Directorate). His research focus is cloud computing and similar distributed technologies. Prior to coming to ORNL, he obtained over 8 years of industry experience working in large-scale datacenters and building systems for provisioning and maintaining large-scale systems software and services. Rob is an active user of Amazon Web Services, Windows Azure, and various private cloud technologies. Rob is a Windows Azure MVP.*

Microsoft Azure: The Ultimate Flexible Enterprise-Level Solution

Janet L. Bailey, University of Arkansas and
Bradley K. Jensen, Principal Academic Relationship Manager, Microsoft

The Walmart Realty Division at Bentonville, Arkansas was the early adopter of cloud computing inside of the corporation. In exploring new capabilities and looking for a viable cloud alternative to provide the capacity and capabilities needed by the corporation Dr. Janet Bailey, her UALR College of Business MIS students, and Dr. Bradley Jensen, Microsoft Principal Academic Relationship Manager, were asked for assistance in evaluating the value and fit of Microsoft Azure for Walmart's enterprise-level cloud computing requirements. To this end, research was conducted to identify the needs and assess the fit. To demonstrate Azure's capabilities to the Senior Executives sample operational demos for the PC and mobile environment were developed and deployed to Azure with a goal of demonstrating the potential benefits of housing transactional data on Azure. As a major global player, Walmart's applications are international in scope and require accessibility anytime, anywhere.

This presentation will report the results of the 3 month study conducted. Specifically, the presentation will focus on how the study was conducted and the resulting value to student participants, the list of enterprise-level requirements identified during the research, analysis of Azure's capability to meet those requirements, and a brief demo of the web and Phone 7 demo applications created. An explanation of why Azure was ultimately selected over other cloud providers will also be presented.

BIOGRAPHIES

Dr. Janet Bailey, *Associate Professor of MIS, University of Arkansas at Little Rock, possesses 20 years of industry experience in needs assessment, systems analysis and design, programming, network installation, and security. She holds a Ph.D. in Business Computer Information Systems with a minor in human factors. She is the recipient of the 2001 and 2009 UALR College of Business' Academic Excellence Award in Teaching. Dr. Bailey is a member of the Microsoft Enterprise Consortium Board of Directors, UALR Student AITP Advisor, and Microsoft Imagine Cup mentor for three National Finalists (2009, 2011) and two World Finalists (2009, 2010). Since Fall 2009 her students have collectively deployed over 500 applications to Azure as part of their curriculum. In recognition of her leadership, Walmart Corporation requested she and her students deliver an Azure proof-of-concept study to Walmart Senior Executives.*

Dr. Bradley Jensen, *a Principal Academic Relationship Manager with Microsoft, brings a wealth of relationship, teaching, research, product marketing, sales, and domestic/international executive experience. As a former faculty member, he has demonstrated capabilities in Curriculum Development, Course Design and Delivery, Research, Consulting, Document Management, Publishing, Telecommunications, Aerospace, Legal, and Commercial Property Management. Dr. Jensen's proven accomplishments are in Information Security, Business Intelligence, Software Development, Project Management, Strategic Alliances, E-Commerce, Strategic Marketing, P & L Management, Team Building, Consultative Selling, Cloud Computing, and Mobile Development. He serves as co-chair of the Board of Directors for Texas Business and Education Consortium STEM, co-chair of the Microsoft Enterprise Consortium Board of Directors, and a member of numerous MIS, CS, and industry Advisory Boards.*

High Through-Put, Low Impedance e-Science on Microsoft Azure

David Abramson, Monash University

In this project we expose Nimrod as a scientific service in Windows Azure, and implement two applications of significance. Nimrod is a toolkit that supports design patterns for coarse-grained data and task parallelism. It supports large-scale parameter sweeps and searches using complex time consuming applications. Based on a generic platform for legacy applications, it also provides a workflow and optimisation framework for scientific studies. Traditionally, Nimrod has used the Grid to provide computational resources. Overall, it has enabled numerous e-Science applications in important domains, including Chemistry and Physics, Medical and Life Sciences, Engineering and Design, Economics and Finance, Environmental Science, and Earth Sciences and Astronomy etc. Nimrod not only provides a framework that allows users to express the built-in parallelism in their applications, but it also enables users to take advantage of the large-scale distributed resources in the Cloud without worrying about difficult issues such as scalability, fault tolerance, performance, resource management and the different abstraction layers of Cloud.

We are currently porting the Nimrod "Agent" to Azure, which is responsible for spawning computations on Cloud resources. It actually represents the minimum amount of Nimrod that needs to run under Azure. After that, we will implement a runtime system outside Azure, including a job scheduler, monitor, resource manager and Web portal. To demonstrate the effectiveness of the solution we plan two significant case studies. One, Ceniza, in conjunction with Instituto Tecnológico de Costa Rica, models ash dispersion from a volcanic eruption. This time consuming computation is important in planning infrastructure in urban areas affected by volcanic ash. Another, iGrid, with University of Queensland in the area of energy economics, investigates the role of smart-grids and electrical-generation distribution in lowering Australia's carbon emissions. Again, this computationally expensive application will allow the comparison of different generation strategies, including mixes of solar, wind and traditional capabilities. Together these applications will stress the Nimrod implementation on Azure, but will also provide interesting scientific results.

BIOGRAPHY

Professor David Abramson *has been involved in computer architecture and high performance computing research since 1979. Previous to joining Monash University in 1997, he has held appointments at Griffith University, CSIRO, and RMIT. At CSIRO he was the program leader of the Division of Information Technology High Performance Computing Program, and was also an adjunct Associate Professor at RMIT in Melbourne. He served as a program manager and chief investigator in the Co-operative Research Centre for Intelligent Decisions Systems and the Co-operative Research Centre for Enterprise Distributed Systems.*

Abramson is currently an ARC Professorial Fellow; Professor of Computer Science in the Faculty of Information Technology at Monash University, Australia, and science director of the Monash e-Research Centre. He is a fellow of the Association for Computing Machinery (ACM) and the Academy of Science and Technological Engineering (ATSE), and a member of the IEEE.

Abramson's current interests are in high performance computer systems design and software engineering tools for programming parallel and distributed supercomputers

Abstractions for Life-Science Applications on Clouds

Shantenu Jha and Andre Luckow, Rutgers University

Although multiple enterprise applications have made effective use of Clouds, the barriers to the development of large-scale high-performance scientific applications that can utilize Clouds in an extensible and scalable fashion remain high. Developing logically and/or physically distributed applications at large-scales is fundamentally a difficult undertaking; developing them for emerging platforms such as Clouds is made harder due to the lack of commonly acceptable and widely used programming models and application-level abstractions. Additionally, several of the challenges—such as coordination of multiple components, that face scientific applications on traditional distributed environments such as Grids, carry-over to cloud environments.

Based upon a recently submitted paper to ACM Computing Surveys http://www.cct.lsu.edu/~sjha/select_publications/dpa_surveypaper.pdf we present a comprehensive analysis of the usage of production distributed infrastructure, explore abstractions that exist as well as those that should be provided to better support the effective development, deployment, and execution of large-scale (scientific) high-performance applications. The advantages of such abstractions vary from better utilization of underlying resources to the greater flexibility in the deployment of applications, as well as lowered development effort. We are extending the analytical framework for logically and physically distributed applications, that was developed in the paper referenced above, to focus on Programming and Systems abstractions for Dynamic, Distributed and Data-Intensive Applications (See: http://wiki.esi.ac.uk/3DPAS).

Building upon this understanding, we are developing and implementing multiple abstractions, and applying them to support a range of life-science applications, for example, next-generation gene sequencing analytics and ensemble-based computing for effective drug design. We present some initial implementation and proof of effectiveness of these abstractions on Azure and FutureGrid/XD. The ultimate goal of our work is to enable a wide range of existing, emerging and novel life-science applications to better and effectively utilize cloud platforms.

BIOGRAPHIES

Shantenu Jha *is an Assistant Professor at Rutgers University. He is also a member of the Graduate Faculty, School of Informatics, Edinburgh and a visiting researcher at UCL & e-Science Institute, University of Edinburgh. He has more than 60 publications at the interface of Computer Science, Computational Science and Cyberinfrastructre Development. For select publications see: http://www.cct.lsu.edu/~sjha/select_publications/ Jha is the PI of the SAGA (http://saga.cct.lsu.edu) Project and the UK-EPSRC & NSF co-funded 3DPAS (Distributed dynamical data-intensive programming abstractions and systems) research theme. He was the PI of the just concluded NSF-funded $2M Cybertools project. He is the lead author of a book to be published by Wiley in Summer of 2011 on "Abstractions for Distributed Applications and Systems: A Computational Science Perspective".*

Dr. Andre Luckow *works full-time for the Innovation Division of BMW in Germany, and is also part-time researcher with Shantenu Jha.*

The Parallelization of Geoscience Apps at C3L with Azure

Craig Mudge, University of Adelaide

Subsurface imaging deeper than two kilometres is our main interest – for discovery of minerals, hot dry rocks for geothermal mining, aquifer management, and carbon sequestration. Magnetotellurics (MT) is a relatively new geophysical technique pioneered by our collaborators, whereby naturally occurring electromagnetic waves are used as source fields for imaging the Earth's electrical resistivity structure at depths ranging from tens of meters to hundreds of kilometres. The random and weak nature of the excitation requires substantial processing.

We are building an Azure cloud environment as our first collaboration with geophysicists[1] to reduce the run time of a 3D inversion of Gawler Craton data from weeks to days. This paper discusses parallelisation of the existing inversion procedure, written as a sequential FORTRAN program, by using a parallel programming cloud framework to allocate independent computations across tens of machines. The optimisation method used in the inversion permits distribution of the workload across frequencies. Forward modeling iterations can be unwound and allocated across processors to further reduce execution time. We break down the 5.5 months effort into learning MT, FORTRAN parallelisation, and building a Web application. The next stage of our collaboration will make our rapid inversion program accessible as a cloud application, available to explorers, worldwide.

Motivation: As the information products of national labs and government agencies are increasingly used to make investment decisions, and as input to national policy, scientists realise that their products need to be dynamic and to encourage analysis. Our rapid inversion is our first attempt at modeling on demand for policy makers, or explorers, to analyse information products for robustness and to extend them more easily for further analysis and interpretation. Cloud computing with its universality of access, pay-per-use model, and massive storage is well matched to this need. At the discovery end, clever deployment of cloud computing in the field can combine massive historical data with real time sensing and modeling to provide unparalleled productivity increases in exploration.

BIOGRAPHY

*The Collaborative Cloud Computing Lab (C3L), at the University of Adelaide, was established by **Craig Mudge** in July 2010. This "No-machines lab" is the first Australian eScience centre enabled by cloud computing, and follows the Jim Gray Fourth Paradigm model of coupling science and computer science for advances afforded by massive data discovery. It follows his Chairmanship of a Working Group on Cloud Computing in the Australian Academy of Technological Sciences and Engineering. Mudge returned to Australia in 2005 after ten years in Silicon Valley where he led the Xerox PARC computer science lab. He founded Austek Microsystems based on a breakthrough design technology developed at CSIRO. Its products included signal processing chips that formed the basis of Cochlear and radio astronomy breakthroughs and the world's first single-chip cache. He was an engineer at DEC and led micro-chip research at CSIRO, establishing Fabless Chip Design in Australia. Mudge received his Ph.D. in computer science from the University of North Carolina at Chapel Hill and an undergraduate degree from the ANU. He co-authored "Computer Engineering" with Gordon Bell, published over sixty papers, holds six patents, and taught at Caltech and CMU.*

[1]Including Graham Heinson and Stephan Thiel in the Department of Geology and Geophysics in the School of Earth and Environmental Sciences

Bringing the Cloud to the Classroom Using the Microsoft Imagine Cup

James L. Parrish, Jr., University of Arkansas at Little Rock
Bradley K. Jensen, Principal Academic Relationship Manager, Microsoft
Patricia Day, University of Arkansas at Little Rock

Organizations of all types are adopting cloud computing technologies to gain operational effectiveness and strategic advantage in their environments. As such, it seems appropriate that students in technology related disciplines be exposed to cloud computing in their academic programs. Teaching cloud computing from a purely technological perspective is one way to expose students to cloud computing, however it does not educate students on how to recognize and exploit opportunities to leverage cloud computing to obtain operational and strategic benefits. Likewise, teaching cloud computing from a purely business perspective does not allow students to gain technical knowledge of cloud computing technology and does not expose them to the full breadth of opportunities that can be exploited using cloud computing technology.

To overcome these challenges, we propose teaching cloud computing from a problem solving perspective using the Microsoft Imagine Cup Software Design Competition as a framework to bring the cloud into the classroom. The Microsoft Imagine Cup allows students to think of innovative ways to utilize cloud computing and then gives them the opportunity to apply this thinking to the creation of an information system that runs on the Microsoft Azure cloud computing platform. In our presentation, we will cover how and when cloud computing should be covered in the classroom, provide a brief introduction to the Imagine Cup and discuss its utility as a framework for teaching students about cloud computing, and share our experiences using the Imagine Cup as a way to teach students how to think innovatively about the use of cloud computing and how to turn that thinking into a working information system. Examples of past student Imagine Cup projects will be presented to illustrate the breadth of the organizational opportunities identified by students that can be exploited using cloud computing.

BIOGRAPHY:

Dr. James L. Parrish, Jr. *is an Assistant Professor of MIS at the University of Arkansas at Little Rock and President of Infoventure Systems Consulting. He is an active researcher in the areas of cloud computing, knowledge management, and information systems security and has his research published in several international journals and conference proceedings. In addition to this research, he has demonstrated experience in the areas of teaching information systems at the graduate and undergraduate levels, information systems course development, IT management in the public and private sectors, and IT consulting. Dr. Parrish has a passion for getting students excited about information systems and has been extensively involved in mentoring student teams in the Microsoft Imagine Cup competition. He has mentored several teams over the past 3 years and was the team mentor for Team MedRX, whose cloud-based medical research application was awarded the Cloud Computing Achievement Award at the 2010 U.S. Imagine Cup finals.*

Dr. Bradley Jensen, *a Principal Academic Relationship Manager with Microsoft, brings a wealth of relationship, teaching, research, product marketing, sales, and domestic/international executive experience. As a former faculty member, he has demonstrated capabilities in Curriculum Development, Course Design and Delivery, Research, Consulting, Document Management, Publishing, Telecommunications, Aerospace, Legal, and Commercial Property Management. Dr. Jensen's proven accomplishments are in Information Security, Business Intelligence, Software Development, Project Management, Strategic Alliances, E-Commerce, Strategic Marketing, P & L Management, Team Building, Consultative Selling, Cloud Computing, and Mobile Development. He serves as co-chair of the Board of Directors for Texas Business and Education Consortium STEM, co-chair of the Microsoft Enterprise Consortium Board of Directors, and a member of numerous MIS, CS, and industry Advisory Boards.*

Patricia Day *has recently received her MS in MIS at the University of Arkansas in Little Rock. While in graduate school, she had the opportunity to work with Microsoft Azure in the classroom setting. She is currently working on an embedded project for the 2011 Microsoft Imagine cup competition that uses SQL Azure to store data. An earlier version of this project was invited to the 2010 Imagine Cup World Finals in Warsaw, Poland. Patricia was a Microsoft Student Partner at UALR during 2010-2011 and has taught computer graphics and drafting classes and Southern Arkansas University Tech, National Park Community College and Ouachita Technical College. Prior to starting graduate school, she was a computer drafter and owned her own drafting business.*

Introducing Cloud Computing into STEM Undergraduate Curriculum Using Microsoft Azure

Dr. Bina Ramamurthy, CSE Department, University at Buffalo

We are witnessing a golden era in computing with a mature Internet, highly powerful multi-core at the processor level, dedicated graphics processors, superior software methodologies, virtualization that leverages the powerful hardware, availability of wider bandwidth for communication, and proliferation of devices. Above all, an explosion of data-intensive and compute-intensive applications is fueling the growth of newer computing models. The large-scale data and computational needs of these models have resulted in the emergence of cloud computing. The cloud is being promoted as the next generation computer and it is imperative that we introduce it to students at all levels of our curriculum. We provide practical hands-on modules to introduce cloud-computing that can be easily adopted into existing courses in Science, Technology, Engineering and Mathematics (STEM) courses.

We examined all the three prominent cloud models, Amazon Elastic Compute Cloud, Google App Engine and Microsoft Azure as the framework for teaching parallelism. While all the models are powerful in their own respect, Azure abstracts all the infrastructural details to provide an intuitive model for large-volume enterprise-level application development. Our goal is to focus on teaching specific core competencies that have broad impact across STEM courses. These competencies include parallelism, algorithms for knowledge discovery, services-oriented design, large-scale data-intensive analytics, design for high-performance and automatic load balancing, large-scale storage and monitoring for access control and performance. Some of the influential features of Azure include the development environment offered by Visual Studio with an easy access to production environment on the cloud, the role-based application modules in web-role and worker-role supporting the ease of cloud-application development, persistence store offered by AzureSQL, blob storage for multi-media data, REST-based web services, and AppFabric for composite application development. We will present the design and development of an application Name2Face to demonstrate the perfect fit of Azure for teaching cloud computing.

BIOGRAPHY:

Dr. Bina Ramamurthy *is a faculty member at University at Buffalo, Computer Science and Engineering Department. She has been involved in the computer systems research, development and teaching for the past two decades. Her current research is in the area of distributed systems with emphasis on data-intensive computing and cyber-infrastructures. She has been the principal investigator on National Science Foundation funded projects in the area of grid-computing, embedded systems and currently in the area of data-intensive computing and in evolutionary biology on the cloud (Project Pop!World). She has directed more than 50 graduate projects and is well-published in these areas. She has given numerous invited presentations at prominent conferences in the areas of grid-computing and cloud computing.*

Bina Ramamurthy received the B.E. (Honors) in Electronics and Communication from Guindy Engineering College, Madras University, India, and the Ph.D. in Computer Engineering (1997) from the University at Buffalo, Buffalo, NY.

InstantLab—The Cloud as Operating System Teaching Platform

Alexander Schmidt and Andreas Polze, Operating Systems and Middleware Group, Hasso-Plattner-Institute at University of Potsdam

More than 5 years of experience with the Windows Research Kernel in teaching operating system graduate classes have revealed several issues: the infrastructure needed for conducting those experiments is an ever-changing target. On the one hand, hardware has to be maintained and updated appropriately; on the other hand, the variety of platform versions we had to deal with is constantly increasing. To remedy these issues we propose migrating these experiments into the Cloud as the Silver Bullet for teaching operating systems.

InstantLab is our approach for conducting operating system experiments – such as extending the kernel with new system calls – in the cloud, which relieves faculty and students from the maintenance, versioning, and configuration overhead. In the talk we present our experiences with using the Cloud as a platform for our service, which also includes several drawbacks and limitations we faced with today's Cloud implementations. We will also give a demonstration of our platform and present early evaluation results. Finally, our experiments will be made available on Windows Azure in a pre-packaged format.

BIOGRAPHY:

Dipl.-Inf. Alexander Schmidt *studied computer science at the Chemnitz University of Technology where he graduated and received his diploma. In 2006 Alexander Schmidt joined the Operating Systems and Middleware group at Hasso-Plattner-Institut (HPI) as a Ph.D. student. His main research focus is in the area of application monitoring and especially in the operating system context.*

At HPI, Alexander is involved in teaching operating systems courses as well as the Windows Research Kernel project. He contributes to the Windows Monitoring Kernel, an efficient event-logging infrastructure for monitoring arbitrary applications based on Windows systems as well as the NTrace tool for dynamically instrumenting applications at function boundaries. As part of his thesis, he created the KStruct OS kernel inspection framework, which focuses on consistently accessing shared data structures while the OS is running.

Alexander has been a Summer intern with Microsoft in Redmond in 2008 and 2009.

Prof. Dr. Andreas Polze *is the Operating Systems and Middleware Professor at the Hasso Plattner Institute for Software Engineering at University Potsdam, Germany. He is also the head of the Ph.D. school on "Service-Oriented Systems Engineering" at HPI. Andreas received a doctoral degree from Freie University Berlin, Germany, in 1994 and a habilitation degree from Humboldt University Berlin in 2001, both in Computer Science.*

At HPI, his current research focuses on architectures of operating systems, on component-based middleware, as well as on predictable distributed and cloud computing. Andreas Polze was visiting scientist with the Dynamic Systems Unit at Software Engineering Institute, at Carnegie Mellon University, Pittsburgh, USA, were he worked on real-time computing on standard middleware (CORBA) and with the Real-Time Systems Laboratory at University of Illinois, Urbana-Champaign. Andreas has acted as work component leader and member of scientific board in the 6th framework European Integration project "Adaptive Services Grid". Work in ASG has strong links to the Web Services community and industrial standardization efforts.

His current research interests include Predictable Service Computing, Adaptive System Configuration, and End-to-End Service Availability for standard middleware platforms. He is member of the GI and the IEEE. He currently is member of the program committees of ISORC (Intl. Symp. On Object-Oriented Real-Time Computing) and WORDS (Workshop on Real-Time Dependable Systems). Andreas Polze has (co-) authored more than 60 papers in scientific journals and conference proceedings. He has contributed to five books.

Together with Mark Russinovich and David Solomon, Andreas Polze is one of the co-authors of the Windows Curriculum Resource Kit (CRK), the top-download at the Microsoft faculty resource center. Current projects are centered around the Windows Research Kernel (WRK). Andreas Polze has been funded through the Rotor-I and Rotor -II projects. He received a Phoenix Direct Funding award in 2007 for his research on Phoenix for Real-time Robotics and Process Control.

Risk Assessment and Cloud Strategy Development

Barbara Endicott-Popovsky, University of Washington
Kirsten Ferguson-Boucher, Aberystwyth University, Wales

Little actual research has been undertaken to formally assess the impact of Cloud Computing on professional information management practice, a particular concern for research universities, such as the University of Washington, considering research data storage in the cloud. In 2010, Aberystwyth University, Wales, concluded a cloud research project, commissioned by the UK Archives and Records Association, designed to 1) identify key legal, technological, and organizational issues related to storing assets in a virtual environment, 2) develop good practice guidelines, and 3) define a set of criteria for selecting cloud services.

A major outcome of this research was the Cloud Computing Toolkit (March 2011) that provides a thorough approach to risk assessment and cloud strategy development. Among other things, the toolkit proposes that cloud users, with their cloud service providers, explore such things as: information management practices, legal and regulatory compliance, contract cost, monitoring, auditing and reporting, exit strategy, security, availability management and resource provisioning, incident response, identity and access management, and business continuity.

A research collaboration has begun between Aberystwyth University and the Center for Information Assurance and Cybersecurity at the University of Washington to extend the Toolkit to an Information Governance and Assurance approach to the collection and use of research data in the cloud.

Research includes:

- Implications for recordkeeping principles and practice.
- Implications for long-term preservation of digital material.
- Development of models for engaging professional bodies into a more active role.
- Development of education artifacts for incorporation in information science curricula.
- Enhancement of Toolkit guidance and standards.

There are both opportunities and threats in the cloud. Universities face serious challenges in cloud storage of research data. We expect our research to illuminate this area of cloud computing adoption.

BIOGRAPHIES

Barbara Endicott-Popovsky, Ph.D., *is Director for the Center of Information Assurance and Cybersecurity at the University of Washington, Academic Director for the Masters in Infrastructure Planning and Management in the Urban Planning Department of the School of Built Environments and a Research Associate Professor with the Information School. Her academic career follows a 20-year career in industry in executive management. Her research interests include enterprise-wide information systems security, forensic-ready networks, the science of digital forensics and secure coding practices. She is a member of the American Academy of Forensic Scientists.*

Kirsten Ferguson-Boucher, M.A., MScEcon., *is Lecturer in Records Management, Aberystwyth University, Wales, United Kingdom, and Co-PI for the Cloud Computing Research Project--UK Archives and Records Association, as well as Co-Developer of the Cloud Computing Toolkit. Her work has attracted broad interest within the UK and the European Union. Her research interests include cloud computing records management, digital records forensics.*

Cloud Forensics: an Overview

Keyun Ruan, University College Dublin

Cloud forensics is a new area on cybercrime investigation in the Cloud. Cloud forensics face tremendous difficulties at the moment, such as jurisdiction, multi-tenancy, evidence segregation, missing terms in SLA and lack of awareness. In this talk, Keyun Ruan will first give a general introduction on cloud forensics including its state of art, and the three dimensions, i.e., technical, organizational and legal dimensions of cloud forensics. She will also share the results of a recent survey "Cloud Forensics and Critical Criteria for Cloud Forensic Capability" towards digital forensics experts and practitioners worldwide which has received 187 responses. Lastly she will discuss the suggested key terms to be included in the SLA between CSP and cloud customers on cloud forensics.

References:

Mark Pollitt, Former Head of FBI

BIOGRAPHY

Keyun Ruan *is a thought leader worldwide on cloud forensics research based in Dublin, Ireland. She is one of the first researchers who identified the area of cloud forensics. She is the author of several important academic publications on cloud forensics and the editor of book "Cloud Forensics and Cybercrime: Applications of Investigative Processes" to be published by IGI Global in 2012. She is co-founder of a San Francisco based consulting firm Cloud Forensic Network which is building world's largest expert network for cloud forensic services and tools. She has previously spoken at Cloud Expo 2010 in Prague and Silicon Valley and various other conferences.*

For more of her bio please visit Ruankeyun.com/bio

Cloud Based Product Development

Joris Poort, Harvard Business School

Cloud computing offers significant benefits to the development of computer aided engineering (CAE) product development including a decrease in development timelines, improvement of product performance, and significant reductions in overall development cost. The complete benefits of high performance cloud computing in product development environments for manufacturing and high tech firms will be realized through both the integration of disciplinary analysis software and numerical optimization tools to leverage the nearly infinite compute resources that have become accessible through the cloud.

Academic research and industry applications have both shown that adapting product development methods for high performance cloud computing opens up a breadth of new opportunities. Specifically, cloud based product development provides numerous benefits to the development of high tech products including:

Reduced Development Timelines
- Elimination of computational bottlenecks
- Dynamic scaling of computational resources
- Reducing iterations needed for convergence

Improved Performance
- Capturing, understanding, and leveraging design interdependencies
- Broader exploration of design space
- Elimination of errors

Reduced Cost
- Engineering time savings
- Capital expenditure reductions on computing resources
- Direct pay-as-you-go cost allocations

Improved Visibility and Control
- Increased oversight, visibility, and control
- Increased ability to measure progress

Benefits from cloud computing for product development are realized through:

Cloud Databases
- Dynamic real-time cloud databases to store and share data
- Standardized model data relationships to capture the interdependencies and relationships between data in the cloud

Cloud Analysis & Processes
- CAE disciplinary tools architected for the cloud to compute off-the-shelf analysis for various disciplines
- Numerical tools to perform computation, optimization, and designs of experiments in the cloud

Cloud Computing
- High-performance cluster computing in the cloud, dynamically scalable

Cloud computing has fundamentally changed traditional product development and addresses some of the key issues faced in engineering product development in aerospace and automotive industries.

BIOGRAPHY

Joris Poort *holds a B.S. magna cum laude in Mechanical Engineering with a Minor in Applied Mathematics from the University of Michigan, a M.S. magna cum laude in Aeronautics and Astronautics from the University of Washington, and an M.B.A. from Harvard Business School. Professionally, Joris spent over four years at Boeing on the 787 program using high performance computing for the optimization of aircraft wings. Joris' work has resulted in numerous research papers, hundreds of millions in dollars in cost savings, and magnitudes of improvement in Boeing's product development timelines. Joris has also spent time with McKinsey & Company supporting high-tech product development engagements. Most recently while at Harvard, Joris has been researching and developing cloud computing based product development solutions for aerospace and automotive companies. Specifically, Joris' research focuses on the integration of different analysis disciplines through cloud computing platforms reducing development time and improving performance of complex high-tech products.*

Cloud, HPC, or Hybrid: A Case Study Involving Satellite Image Processing

Marty Humphrey, University of Virginia

While the potential benefits of the cloud are significant for many areas of information technology, much more study must take place to more precisely determine the value for scientific applications. For example, for many scientific applications, a key benefit of higher availability that can come from a cloud deployment may be unimportant. Rather, for those many scientific applications that might just need compute cycles, is it is better to pursue a cloud deployment or just buy a few more cheap nodes for their new (or already-existing) local cluster?

We perform a concrete analysis of the value of a cloud design and deployment for a particular large-scale application we have created to process and derive scientific results from the MODIS satellite system. This analysis is unique in many respects, particularly that we designed the cloud application ("BESS") from first principles (in Windows Azure) and then only later did we "port" it to a local cluster environment (and finally a hybrid "cloud-bursting" version). By comparing on dimensions of platform-dependent first-version issues, debugging, fault tolerance, correctness, economics, usability, and run-time speed, we provide general guidance for the broad scientific community.

BIOGRAPHY:

Marty Humphrey *is an Associate Professor in the Department of Computer Science at the University of Virginia. He received a B.S. and M.S. degree in Electrical Engineering from Clarkson University in 1986 and 1989, respectively. He received a Ph.D. degree in computer science from the University of Massachusetts in 1996. He has been on the faculty of the University of Virginia since 1988.*

Classical and Iterative MapReduce on Azure

Geoffrey Fox, Indiana University

We describe experiences on Azure, Amazon and academic systems (FutureGrid) on different applications (mostly from the Life Sciences) using various implementations of MapReduce. In particular we discuss MapReduceRoles4Azure that implements MapReduce on Azure building on Azure Queues for task scheduling; Azure Blob storage for input, output and intermediate data storage; Azure Tables for meta-data storage and monitoring. MRRoles4Azure supports a combiner step; dynamically scaling up and down of the compute resources; Web based monitoring console; testing and deployment using Azure local development fabric. We show that Azure gets excellent performance on parallel data intensive applications not requiring the microsecond latency of classic MPI-based simulations.

There exists many data analytics as well as scientific computation algorithms that rely on iterative computations, where each iterative step can be easily specified as a MapReduce computation. Twister4Azure extends the MRRoles4Azure to support such iterative MapReduce executions, drawing lessons from the Java Twister iterative MapReduce framework that we introduced earlier in thesis of Jaliya Ekanayake. Iterative extensions include a merge step; in-memory caching of static data (between iterations); cache aware hybrid scheduling using Azure Queues as well as a bulletin board (special table). Twister4Azure and MRRoles4Azure offer the familiar MapReduce programming model with fault tolerance features similar to traditional MapReduce and a decentralized control model without a master node implying no single point of failure. We test on data mining algorithms applied to Metagenomics and requiring parallel linear algebra in their compute intensive kernel.
MRRoles4Azure and an alpha version of Twister4Azure can be downloaded from http://salsahpc.indiana.edu/mapreduceroles4azure. Twister can be downloaded at http://www.iterativemapreduce.org/ .

BIOGRAPHY:

Geoffrey Fox *received a Ph.D. in Theoretical Physics from Cambridge University and is now Distinguished Professor of Informatics and Computing, and Physics at Indiana University where he is director of the Digital Science Center and Associate Dean for Research and Graduate Studies at the School of Informatics and Computing. He previously held positions at Caltech, Syracuse University and Florida State University. He has supervised the PhD of 62 students and published over 600 papers in physics and computer science. He currently works in applying computer science to Bioinformatics, Defense, Earthquake and Ice-sheet Science, Particle Physics and Chemical Informatics. He is principal investigator of FutureGrid – a national facility to enable development of new approaches to computing. He is involved in several projects to enhance the capabilities of Minority Serving Institutions.*

Data Semantics Aware Clouds for High-performance Analytics

Jun Wang, the University of Central Florida

Today's cutting-edge research deals with the increasing volume and complexity of data produced by ultra-scale simulations, high resolution scientific equipment and experiments. Representatives include analytics- and simulation- driven applications such as astrophysics data analysis, bioinformatics, etc. In these fields, scientists are dealing with large amounts of data and processing (analyzing) them to explore new concepts and ideas.

Many scientists are exploring the possibilities of deploying applications with large scale of data on cloud computing platforms such as Windows Azure. Recently, the successful deployment of eScience applications on clouds motivates us to deploy HPC analytics applications to the cloud, especially MapReduce enabled. The reason behind this lies in a fact that eScience applications and HPC analytics applications share some important features: tera-scale or peta-scale data size and high cost to run on single or several supercomputers or large platforms. However, HPC analytics applications bear some distinct characteristics such as complex data access patterns, interest locality, and which pose new challenges to its adoption of clouds. 1) There exists a mismatch between the high-performance requirement of HPC analytics applications on large-scale data and the tolled computation and I/O capacity in the cloud platforms due to virtualization technology; 2) There exists a "data semantics"gap between the way data was stored by the cloud platforms and the way data will be accessed by the HPC analytics applications; 3) Emerging parallel data processing models impose new design and implementation issues on the clouds, including MapReduce based and non-MapReduce based. However, current solutions do not deal well with these challenges and have several limitations. First, we examine the cloud platforms supporting Hadoop/MapReduce, such as Amazon EC2 and Open Cirrus. Although the Hadoop/MapReduce framework supports applications which operate on very large data sets, it is not designed for semantics-based HPC analytics and will suffer with limited performance. Second, we examine the cloud platforms not supporting Hadoop/MapReduce; specifically, we examine Windows Azure. Lastly, co-locating computation and storage is the most salient feature in MapReduce. Data are often prepared beforehand and evenly stored at local node where a compute task can be launched. Unfortunately, current cloud architectures and applications make it hard to implement good co-location.

We have been developing a data semantics aware framework to support HPC analytics applications on clouds. We envision better solutions in distributed and parallel file and storage systems, high-performance parallel programming APIs for big data processing, and interoperable data conversion middleware are imperatively needed. Our approach will gear toward data-semantics aware big scientific data processing. This infrastructure consists of three components; 1) a MapReduce API with data semantics awareness (MARS) to develop high-performance MapReduce applications, 2) a translation layer equipped with data-semantics aware HPC interfaces (HIAS), and 3) a neural network based Data-Affinity-Aware data placement scheme (NDAFA).

BIOGRAPHY:

Dr. Jun Wang *is a Charles N. Millican Faculty Fellow and Assistant Professor of Computer Engineering at the University of Central Floirda, Orlando, FL, USA. He received his Ph.D. in Computer Science and Engineering from University of Cincinnati in 2002. He is the recipient of National Science Foundation Early Career Award 2009 and Department of Energy Early Career Principal Investigator Award 2005. He has authored over 60 publications in premier journals such as IEEE Transactions on Computers, IEEE Transactions on Parallel and Distributed Systems, and leading HPC and systems conferences such as IPDPS, HPDC, EuroSys, ICS, Middleware, FAST. He has conducted extensive research in the areas of Computer Systems and High Performance Computing. His specific research interests include massive storage and file System in local, distributed and parallel systems environment. He has graduated 5 Ph.D. students who upon their graduations were employed by major US IT corporations (e.g., Google, Microsoft, etc). He serves on Editorial board for the International Journal of Parallel, Emergent and Distributes Systems (IJPEDS).*

Into the Blue: Streaming Data Processing in the Cloud

Christopher Alme, Christopher Nunu, Dennis Qian,
Stanley Roberts and Stephen Wong
Rice University

Most cloud data processing utilizes a "batch-mode" where processing requests are submitted and results returned. The elasticity of the cloud is utilized to scale the system to handle larger or smaller numbers of processing requests. Very few cloud applications are designed to handle continuous streams of data [See, for instance, refs. 1, 2, 3] and none can handle multiple, independent data inputs with multiple independent outputs. Existing systems tend to be based on MapReduce/Hadoop, require custom programming for each application and are difficult to modify, reconfigure, extend or re-use in new situations. To address these issues, we will present the preliminary implementation of an Azure cloud application that features multiple simultaneous, independent, real-time input and output endpoints. The system also designed to use drop-in processing modules that the user assembles into a processing graph to perform the desired operations. A processing graph may have multiple input endpoints of differing types to simultaneously gather information from a wide variety of sources. The user assembles the graph by selecting a module from a library and specifying the desired input and output connections. Process allocation, connection mechanics and data synchronization are handled transparently. Multiple independent output endpoints are also supported, enabling the user to simultaneously extract different processing results from a single processing graph. The user can modify, reconfigure and extend the graph without re-deploying or even stopping the system. System configurations are stored in the cloud and cloud-enabled fault tolerance is supported. Applications for this technology include environmental sensor monitoring, real-time aircraft/vehicle tracking, highway monitoring and process monitoring.

[1]STREAM project (http://www.streamproject.eu) : query-based real-time processing

[2]HStreaming (http://hstreaming.com): Hadoop in the cloud

[3]Logothetis and Yocum, "Ad-hoc Data Processing in the Cloud" (Proc. VLDB Endowment , 1:2, pp. 1472-1475, 2008, http://cseweb.ucsd.edu/~kyocum/pubs/mortar_vldb08.pdf): MapReduce-based system

BIOGRAPHIES:

Christopher Alme, Christopher Nunu, Dennis Qian and Stanley Roberts *are the Fall 2010 COMP410 team at Rice University. COMP410 is an undergraduate software engineering course that utilizes a "discovery mode" pedagogical style where the class works as a team to design and implement a cutting edge enterprise-class software project.*

Alme, Nunu, Qian and Roberts *are all CS majors headed off to the Azure Cloud and Identity groups at Microsoft, Mobile Development at Google, and the OpenWorks team at Halliburton, respectively.*

Stephen Wong *was originally trained in physics (Ph.D. in semiconductor physics and Howard Hughes Fellow, M.I.T., 1998), including a year at Bell Labs with future Nobel laureate and Energy Secretary, Steven Chu. After a stint at Hughes Research Laboratories, in 1993 he switched to academia and software consulting for Eastman Kodak. Since 1998, he has been completely devoted to CS teaching and research, the last 10 years at Rice University.*

Stork Data Scheduler for Windows Azure

Tevfik Kosar, State University of New York (SUNY) at Buffalo

Applications and experiments in all areas of science are becoming increasingly complex and more demanding in terms of their computational and data requirements. Some applications generate data volumes reaching hundreds of terabytes and even petabytes. Sharing, disseminating, and analyzing these petascale data sets becomes a big challenge especially when distributed resources are used. For this purpose, we have introduced the Stork data scheduler which focuses on planning, scheduling, monitoring and management of data placement tasks and application-level end-to-end optimization of networked I/O for petascale distributed applications. Stork is considered one of the very first examples of data placement scheduling and it has been very actively used in many application areas including coastal hazard prediction, reservoir uncertainty analysis, digital sky imaging, educational video processing, numerical relativity, and multiscale computational fluid
dynamics resulting in breakthrough research.

As part of our recent NSF CiC Award, we are further developing and enhancing the Stork Data Scheduler to support the Windows Azure cloud computing environment, in order to mitigate the data handling bottleneck in data intensive cloud computing applications as well. The Stork Data Scheduler for Azure will make a distinctive contribution to cloud computing environments because it implements techniques specific to queuing, scheduling, and optimization of data placement jobs; provides high reliability in data transfers; and creates a level of abstraction between the user applications and the underlying data transfer and storage resources via a modular, uniform interface. Stork data scheduler for Azure will provide enhanced functionality for cloud computing such as data aggregation and connection caching, peer-to-peer and streamed data management; early error detection, classification, and recovery in data transfers; scheduled storage management; optimal protocol tuning; and end-to-end performance prediction services. Stork for Azure will change how domain scientists perform their research by rapidly facilitating sharing of experience, raw data, and results in cloud computing environments.

BIOGRAPHY

Tevfik Kosar *is an Associate Professor of Computer Science and Engineering at the State University of New York (SUNY) at Buffalo. Kosar has received his Ph.D. degree in Computer Science from University of Wisconsin-Madison in 2005. His main research interests lie in the cross-section of petascale distributed systems, eScience, Grids, Clouds, and collaborative computing with a focus on large-scale data-intensive distributed applications. He is the primary designer and developer of the Stork distributed data scheduling system which has been adopted by many national and international institutions, and the lead investigator of the state-wide PetaShare distributed storage network in Louisiana. He has published more than fifty academic papers in leading journals and conferences. Kosar is recipient of NSF CAREER Award, LSU Rainmaker Award, LSU Flagship Faculty Award, Baton Rouge Business Report's Top 40 Under 40 Award, 1012 Corridor's Young Scientist Award, College of Basic Science's Research Award, and CCT Faculty of the Year Award.*

Relational Data Markets in the Cloud: Challenges and Opportunities

Magdalena Balazinska, University of Washington

Cloud-computing is transforming many aspects of data management. Most recently, the cloud is seeing the emergence of digital markets for data (raw and derived) and associated services. One example is the Windows Azure DataMarket.

These markets, however, are still in their infancy. The economic and algorithmic principles guiding the pricing of data, data products, and the services that deliver them are largely unexplored. Existing pricing frameworks are simplistic and can exhibit unexpected and undesirable properties leading to, for example, arbitrage situations, fairness violations, and unpredictability. Further, the technology to facilitate these cloud-based data markets and enforce pricing policies is underdeveloped.

In this talk, we present early work at the University of Washington related to building relational data markets in the cloud. We first discuss how one can support fine-grained data pricing efficiently, what fine-grained data pricing enables, and the properties that different pricing schemes entail. We then discuss several tools that we are currently building in support of a cloud data market. These tools include a Pricing Advisor that helps content providers price their data, a Market Monitor that observes current prices and searches for pre-defined property violations such as arbitrage opportunities, a Comparison Shopper that assess and compares the cleanliness and information content of various datasets, and more.

BIOGRAPHY

Magdalena Balazinska *is an Assistant Professor in the department of Computer Science and Engineering at the University of Washington. Magdalena's research interests are broadly in the fields of databases and distributed systems. Her current research focuses on data intensive scalable computing, sensor and scientific data management, and cloud computing. Magdalena holds a PhD from the Massachusetts Institute of Technology (2006). She is a Microsoft Research New Faculty Fellow (2007), received an NSF CAREER Award (2009), a 10-year most influental paper award (2010), an HP Labs Research Innovation Award (2009-2011), a Rogel Faculty Support Award (2006), and a Microsoft Research Graduate Fellowship (2003-2005).*

Multi-Cloud and Cloud-Desktop Coordination Made Simplified by GXP on Azure

Kenjiro Taura and Ting Chen, University of Tokyo

In this talk, we will describe our latest development effort, GXP on Azure. GXP on Azure provides users/developers with an easy-to-use environment for executing workflows on Azure. Underneath it is a more general interface for remote process invocation and management, which makes it an attractive infrastructure for cloud-backed desktop computing. Initially, GXP was designed as a parallel shell that connects multiple administrative domains. Specifically, it accommodates resources having various interfaces (SSH and diverse batch schedulers), provides users with a uniform and fast/interactive interface to these diverse resources, and frees users from installation burdens on remote hosts, which usually do not share a file system with the user's desktop. GXP later evolved to a workflow execution system that queues and dispatches commands to remote hosts.

For workflow description, we have so far used the make. A general interface between the frontend language and execution engine (scheduler and dispatcher) has been defined and frontends based on other scripting languages are underway. Thanks to these initial design scope of harnessing resources across multiple administrative domains, it can naturally evolve to a platform encompassing cloud resources, with GXP on Azure being a specific example. It benefits end users by giving a workflow execution environment and a description language familiar to them and portable across a range of resource types. We are specifically targeting natural language processing tools and workflows developed by Tsujii et al., which we have so far been running on Linux clusters of 8,192 cores. In this talk, I will brief the relevant aspects of GXP design and talk about implementation of GXP on Azure in some depth. I also touch upon on-going research efforts towards an easy-to-use and high performance framework for data intensive computing, with GXP being one of its main components.

BIOGRAPHY

Kenjiro Taura *is associate professor at Department of Information and Communication Engineering, University of Tokyo. He received his B.S., M.S., and DSc degrees from University of Tokyo in 1992, 1994, and 1997. His major research interests include parallel/distributed computing and programming languages.*

Running Large Workflows in the Cloud

Paul Watson, Newcastle University

Workflows have become fundamental to e-Research. In this talk we will describe our experiences of running very large workflows on the Azure cloud. This includes workflows that take over 3 weeks to run on up to 100 nodes.

We have been working with chemists who use QSAR (Quantitative Structure-Activity Relationships) to mine experimental data for patterns that relate the chemical structure of a drug to its activity. If a successful QSAR model can be derived from the data then it can be used to focus new chemical synthesis, dramatically reducing both cost and time.

A novel approach– the Discovery Bus – is used to automate the mining process, but it is computationally intensive; for one set of input data that became available, it was predicted that it would take five years to process on the original single-server implementation. We have therefore utilised cloud computing to build a scalable solution that generates results in days rather than years. This exploits the opportunities for parallelism inherent in the Discovery Bus; for example, multiple models are independently generated before the best is selected.

Discovery Bus computations are represented as workflows, and different approaches to achieving efficient scalability have been explored. In the original, all the potential parallelism inherent in the workflow was exposed to the scheduler. Unfortunately, the granularity of the parallelism was small, and so co-ordination and communication costs reduced efficiency. A recent alternative enables the user to identify those sub-workflows that should be considered as potential units of parallelism. The higher granularity results in increased efficiency.

The talk will describe our experiences in running very large workflows on Azure, comparing the results from the different approaches. In the largest run carried out, this involves efficiently scheduling 280K sub-workflows containing a total of over 2 million service executions, on up to 100 nodes.

BIOGRAPHY

Paul Watson *is Professor of Computer Science and Director of the Digital Institute at Newcastle University, UK. He also directs the RCUK Digital Economy Hub on Social Inclusion through the Digital Economy. He graduated in 1983 with a BSc in Computer Engineering from Manchester University, followed by a PhD in 1986. In the 80s, as a Lecturer at Manchester University, he was a designer of the Alvey Flagship and Esprit EDS systems. From 1990-5 he worked for ICL as a system designer of the Goldrush MegaServer parallel database server, which was released as a product in 1994.*

In August 1995 he moved to Newcastle University, to continue his research on scalable information management. His current focus is on the design of Cloud Computing platforms to support e-research; this includes applications in chemistry, healthcare and activity recognition.

Towards Enabling Mid-Scale Geo-Science Experiments Through Microsoft Trident and Windows Azure

Eran Chinthaka Withana and Beth Plale, Indiana University

Cloud computing's unique model of computation can greatly benefit the scientific endeavor, particularly in support of ensemble runs. We are extending existing efforts in building and executing meteorology models on Windows HPC Server and orchestrated through the Trident Scientific Workflow Workbench and using components of the Linked Environments for Atmospheric Discovery II (LEAD II) cyberinfrastructure. In collaboration with atmospheric researchers at University of Miami, we are using Windows Azure to carry out numerical storm surge ensemble runs from which storm surge predictions can be made. We are working with a probabilistic ensemble execution of the Sea, Lake and Overland Surges from Hurricanes (SLOSH) storm surge prediction model. Since a typical SLOSH simulation only takes a few minutes to run on a medium-sized workstation, the generation of ensemble products that require 15,000 instances of SLOSH becomes possible on the cloud without requiring statistical approximations. The individual instances are independent so a high degree of scale-up can be achieved. We focus in particular on fault tolerant ensemble execution using Azure worker roles, orchestrated through Trident, with downstream workflow management of the data results. A single instance of a SLOSH ensemble run takes 3 GB of input and generates 8 GB of output, making it well suited to data center execution.

In this talk we discuss the framework that enables a researcher to use cloud computing resources to run ensemble simulation studies. We build on our experiences in successful deployment of the Weather Research Forecast (WRF) model, WPS, and GRADS visualization toolkit on Windows HPC Server and Windows Azure. The WRF demonstration was done using the VM role support built into Windows Azure. We will discuss the infrastructure we have built to utilize Trident for large scale workflows, including Sigiri, our framework for managing interactions with Grids and Clouds.

BIOGRAPHIES

Eran Chinthaka Withana *is a PhD Candidate in School of Informatics and Computing , Indiana University working with Professor Beth Plale in Data To Insight center. His research work involves enabling user inspired management of scientific jobs in grids and cloud computing resources. He is a member of the LEAD II team involved in enabling geo-science experiments in cloud computing resources. He was also involved in delivering timely weather forecasts for the meteorologists working on Vortex2 experiment in summer 2010, using LEAD II cyberinfrastructure.*

Beth Plale *directs the Data To Insight Center at Indiana University. Dr. Plale is Professor of computer science in the School of Informatics and Computing. Prior to joining Indiana University, Dr. Plale held a Postdoc in the College of Computing at Georgia Institute of Technology. Plale's Ph.D. is in computer science from State University of New York Binghamton. Her research is in data provenance and metadata, digital preservation of scientific data, workflow systems in e-Science, and complex events processing. She is an ACM Senior Member, IEEE Member, and recipient of the distinguished DOE Early Career Award.*

Expanding the Horizons of Cloud Computing Beyond the Data Center

Jon Weissman and Abhishek Chandra, University of Minnesota

In this talk, we present the Minnesota vision of cloud computing and the research challenges that we are addressing in our group. We are pursuing three areas that we believe will drive the evolution of the cloud and enable it to move beyond the current data center computing model. First, we describe how the power of edge resources can play an important role in the evolving cloud ecosystem to support in-situ data processing, locality-awareness, and privacy. We present a new cloud architecture called Nebula that is fully complementary and synergistic with the commercial cloud.

Second, the emergence of global services spanning multiple data-centers poses new challenges for efficient service execution and delivery. We describe our work with a global multi-data-center MapReduce/Hadoop service that we are building. One aspect of this work is to extend the rack-awareness of HDFS to data-center awareness. This work is a first-step towards general support for data-centric global cloud services.

Lastly, we describe our vision of a user-centric cloud in support of mobile hand-held devices that responds to user context, preferences, and device resource availability, to improve performance, reliability, and fidelity. Our user-centric cloud model can dynamically out-source data and computation to the cloud opportunistically. This work is also exploring implicit sharing patterns across mobile users in the cloud based on social ties and shared interests/preferences.

Preliminary results for running applications in real clouds for all three areas will be presented.

BIOGRAPHY

Jon Weissman *is a leading researcher in the area of high performance distributed computing. His involvement dates back to the influential Legion project at the University of Virginia during his Ph.D. He is currently an Associate Professor of Computer Science at the University of Minnesota where he leads the Distributed Computing Systems Group. His current research interests are in cloud computing, Grid computing, distributed systems, high performance computing, resource management, reliability, and e-science applications. He works primarily at the boundary between applications and systems. He received his B.S. degree from Carnegie-Mellon University in 1984, and his M.S. and Ph.D. degrees from the University of Virginia in 1989 and 1995, respectively, all in computer science. He is a senior member of the IEEE and awardee of the NSF CAREER Award (1995).*

Abhishek Chandra *is an Associate Professor in the Department of Computer Science and Engineering at the University of Minnesota. His research interests are in the areas of Operating Systems, Distributed Systems, and Computer Networks, with recent focus on large-scale distributed systems, such as Clouds and Grids. He received his B.Tech. degree in Computer Science and Engineering from IIT Kanpur, India, and M.S. and PhD degrees in Computer Science from the University of Massachusetts Amherst. He is a recipient of the NSF CAREER Award, his PhD dissertation titled "Resource Allocation for Self-Managing Servers" was nominated for the ACM Dissertation Award, and he has been a co-author on two Best Paper/Student Paper Awards.*

URSA: Scalable Load Balancing and Power Management in Cluster Storage Systems

Seung-won Hwang, Pohang University of Science and Technology (POSTECH)

Enterprise and cloud data centers are comprised of tens of thousands of servers providing megabytes of storage to a large number of users and applications. At such a scale, these storage systems face two key challenges: (a) hot-spots due to the dynamic popularity of stored objects and (b) high operational costs due to power and cooling. Existing storage solutions, however, are unsuitable to address these challenges because of the large number of servers and data objects. In this talk, we describe the design, implementation, and evaluation of URSA, which scales to enterprise data centers and minimizes latency and bandwidth costs during system reconfiguration. Toward this goal, URSA formulates an optimization problem that selects a subset of objects from hot-spots and performs topology-aware migration to minimize reconfiguration costs. We also show that the same reconfiguration techniques can reduce power costs. Our evaluation shows URSA achieves cost-effective load balancing, is time-responsive in computing placement decisions, e.g., about two minutes for 10k nodes and 10M objects, and provides effective power savings. This is joint work with Gae-won You (POSTECH), Navendu Jain (Microsoft Research), and Hua-jun Zeng (Microsoft).

BIOGRAPHY

Seung-won Hwang *is an associate professor of Computer Science and Engineering at Pohang University of Science and Technology (POSTECH), Korea. Prior to joining POSTECH, she received her Ph.D. in computer science from University of Illinois at Urbana-Champaign. Her research lies in web-scale data management, published in major international journals and conferences, including ACM TODS, IEEE TKDE, SIGMOD, VLDB, SIGKDD, and ICDE.*

Achieving Energy Efficient Computing by Jointly Scheduling Services and Batch Jobs in Virtualized Environments

Tajana Simunic Rosing, University of California, San Diego

In a future where data will not only come from classical computing systems as it does today, but also from millions of sensors and mobile devices, the need for energy-efficient large-scale data computation will explode. Virtualized data centers facilitate higher resource utilization and energy efficiency through consolidation. However, mixing services oriented workloads with batch jobs is typically avoided due to complex interactions and widely different quality of service (QoS) requirements, resulting in low utilizations per server and high overall energy costs. We show that the state of the art techniques that focus on consolidating homogeneous workloads do not scale well for resource management with heterogeneous workloads such as services and throughput oriented batch jobs. This happens due to the contrasting requirements of these workloads in terms of both computation and QoS. We demonstrate that co-scheduling services and batch jobs provides significant opportunity for improved energy efficiency as each tends to stress different aspects of the host hardware. A new unified metric, qMIPS/Watt, is defined, that quantifies the combined efficiency in terms of work done per Joule of the heterogeneous workload combination, derated by the negative effect batch jobs have on the ability of the service jobs to meet their performance goals. This metric is used by our VM resource management framework, called Themis, to maximize energy- efficient throughput of the latter without sacrificing the service guarantees of the former by utilizing a provably stable controller. Themis' policy has been tested on state of the art servers. It outperforms prior proposed policies by up to 70% on average in work done per Joule, while meeting the response time requirements of services, and keeping average throughput of batch jobs only 7% lower than running them on a separate server.

BIOGRAPHY

Tajana Šimunic Rosing *is currently an Associate Professor in Computer Science Department at UCSD. Her research interests are energy efficient computing, embedded and wireless systems. Tajana's current work is focused on developing energy efficient scheduling policies for virtualized server environments and on energy efficiency in population area healthcare networks. From 1998 until 2005 she was a full time research scientist at HP Labs while also leading research efforts at Stanford University. She finished her PhD in EE in 2001 at Stanford, concurrently with finishing her Masters in Engineering Management. Her PhD topic was dynamic management of power consumption. Prior to pursuing the PhD, she worked as a senior design engineer at Altera Corporation. Her MS thesis topic at University of Arizona was high-speed interconnect and driver-receiver circuit design. She has served at a number of Technical Paper Committees, and is currently an Associate Editor of IEEE Transactions on Mobile Computing.*

Windows Azure Distilled

Krishna Kumar, Windows Azure Academic Lead, Microsoft Corporation

Join us in this tutorial for a dive deep into the Windows Azure platform features that attendees vote for at the beginning of the session. We will start off by examining the various building blocks, features and tooling by building a simple example that encompasses the entire Azure platform, including how to get started with and make the most of Windows Azure's data storage, compute and relational database services. Next, through a series of "scientific app" demos we'll review a variety of features of the platform and patterns that can help you move your on-premise applications to the cloud or enhance them with cloud-based capabilities. Finally, we will discuss and prototype any specific application archetypes that arise during the audience interaction.

Expect a no-fluff, all-stuff discussion jam-packed with interesting demos.

BIOGRAPHY

Krishna Kumar *is Senior Academic Relations Manager at Microsoft where he partners with faculty and researchers around their teaching and research needs. His professional interests lie in the field of parallel and distributed enterprise computing with special emphasis on the cloud, especially around Windows Azure - Microsoft's cloud platform. He leads the national academic effort around Azure by working with early adopters around cloud research and curriculum incorporation. Krishna has been with Microsoft for over 10 years in various roles working with multiple developer focused technologies.*