Intel Science & Technology Center for Cloud Computing

ISTC-CC Update

Summer 2012

www.istc-cc.cmu.edu

Table of Contents

Carnegie Mellon University

Georgia Tech

intel

PRINCETON UNIVERSITY

UC Berkeley

ISTC-CC Research Overview

Cloud computing has become a source of enormous buzz and excitement, promising great reductions in the effort of establishing new applications and services, increases in the efficiency of operating them, and improvements in the ability to share data and services. Indeed, we believe that cloud computing has a bright future and envision a future in which nearly all storage and computing is done via cloud computing resources. But, realizing the promise of cloud computing will require an enormous amount of research and development across a broad array of topics.

ISTC-CC was established to address a critical part of the needed advancement: underlying cloud infrastructure technologies to serve as a robust, efficient foundation for cloud applications. The ISTC-CC research agenda is organized into four inter-related research "pillars" (themes) architected to create a strong foundation for cloud computing of the future:

The four ISTC-CC pillars will provide a strong foundation for cloud computing of the future, delivering cloud's promised benefits to the broad collection of applications and services that will rely on it.

Pillar 1: Specialization

Driving greater efficiency is a significant global challenge for cloud datacenters. Current approaches to cloud deployment, especially for increasingly popular private clouds, follow traditional data center practices of identifying a single server architecture and avoiding heterogeneity as much as possible. IT staff have long followed such practices to reduce administration complexity—homogeneity yields uniformity, simplifying many aspects of maintenance, such as load balancing, inventory, diagnosis, repair, and so on. Current best practice tries to find a configuration that is suitable for all potential uses of a given infrastructure.

Unfortunately, there is no single server configuration that is best, or close to best, for all applications. Some applications are computation-heavy, needing powerful CPUs

continued on pg. 30

Message from the PIs

Hello from ISTC-CC headquarters. With this first ISTC-CC Newsletter, we are hoping to give folks an overview of ISTC-CC and highlights of its first year.

There's a lot to share, which is why the newsletter is so long. In fact, ISTC-CC researchers have made so much impact, on so many fronts, that it's difficult to believe that it was launched (officially) only a year ago. As we pulled together and merged the publications, news items, highlights, and awards, we were reminded (hit over the head?) of the incredibly strong community of researchers that make up ISTC-CC. Just "wow".

In fact, the word "community" is worth calling out, as ISTC-CC has quickly become much more than the sum of its (very strong) parts. For example, there have been a number of ISTC-CC activities leading to papers with authors from multiple participating institutions, and every institution has contributed to at least one such paper. In fact, every institution has been part of a tri-institution paper, with authors from Intel and two universities. No 5-institution publication yet, though. Showing the collaborative ISTC-CC spirit, however, there have also been many joint papers with researchers from other ISTCs, other companies, and other universities. The first ISTC-CC Retreat (see the callout below) brought together over 115 technical folks from Intel and the universities to discuss ISTC-CC research, so we expect such collaborations to be the norm.

Greg Ganger, CMU

Phil Gibbons, Intel

We won't try to recap everything, in this introductory note, but we will highlight a few things. As described in the research overview article, the ISTC-CC agenda is composed of four interrelated "pillars" designed to enable cloud computing infrastructures that provide a strong foundation for future cloud computing. But that's for agenda presentation purposes, as a lot of the activities span pillars, such as scheduling (automation) of multiple data-intensive frameworks (big data) across heterogeneous (specialized) cluster resources. In any case, we've had great progress on a lot of fronts.

Perhaps the most visible early impact has come in the area of new frameworks for supporting efficient Big Data computing based on advanced machine learning (ML) algorithms. In particular, ISTC-CC's GraphLab and Spark activities have both taken off in a big way. In contrast to first generation Big Data abstractions like map-reduce (e.g., as implemented by Hadoop), which are good for simple tasks like filtering or sorting large datasets, these new programming frameworks provide programmers with more natural abstractions for expressing nontrivial ML tasks. The result is both better productivity and more efficient execution... sometimes orders of magnitude more efficient! Both are now usable open source systems, demonstrated at the recent Intel UCO Showcase, and Ted Willke's group at Intel Labs (IL/CSR/SAL) has become active GraphLab contributors. Both have also gained substantial user communities, as illustrated by Spark being presented at the June 2012 Hadoop Summit and over 350 people attending the July 2012 GraphLab workshop.

Another highly fruitful area has been specialization, where our continuing FAWN work flourishes and draws attention, and not just because the W ("wimpy") induces grins. For example, for the third consecutive year, ISTC-CC

continued on pg. 29

First Annual ISTC-CC Retreat a Success!

By all accounts, the First Annual ISTC-CC Retreat was a great success! Attendees included 35 from Intel and 81 from the four universities. The agenda featured keynotes by Jason Waxman and Rich Uhlig of Intel, 12 research talks by faculty and students from all four Universities, 12 breakout groups, and 57 posters. The talks, poster sessions and breakouts provided a tremendous opportunity for attendees to interact, find collaboration opportunities, and exchange early stage ideas and feedback on many ISTC-CC research projects. Faculty and students made key connections across institutions that should greatly benefit the projects going forward. Indeed, substantial collaboration within ISTC-CC grew out of the first ISTC-CC Retreat, including multi-institution activities and papers. There was a "Madness Session", hosted by Michael Kaminsky, in which key research ideas, tools/testbeds and gaps were identified. Leaders from three other ISTCs (visual, secure, and embedded) also attended, leading to promising cross-center collaboration discussions. It was also noted that the ISTC-CC research community is larger than just Intel and the hub and spoke schools. In fact, several posters and presentations acknowledged collaborators in ISTC-CC's open research from other companies, including Microsoft, AT&T Research, VMware, HP Labs and Google, as well as other schools, including Lancaster University (UK), EPFL (Switzerland) and Rice University (Houston). The full retreat details can be found on the ISTC-CC website, and the second ISTC-CC Retreat is scheduled for November 29-30, 2012.

Group photo -- first annual ISTC-CC Retreat, December 2011.

ISTC-CC Personnel

Leadership

Greg Ganger (Academic PI)
Phil Gibbons (Intel PI)
Executive Sponsor: Wen Hann Wang (CSR)
Managing Director: Rich Uhlig (CSR-SAL)
Program Director: Jeff Parkhurst (UCO)
Board of Advisors:
Randy Bryant (CMU)
Balint Fleisher (Intel)
Frans Kaashoek (MIT)
Pradeep Khosla (UC San Diego)
Rich Uhlig (Intel)
Wen-Hann Wang (Intel)
Jason Waxman (Intel)

Faculty

David Andersen, CMU
Guy Blelloch, CMU
Greg Eisenhauer, GA Tech
Mike Freedman, Princeton
Greg Ganger, CMU
Ada Gavrilovska, GA Tech
Phillip Gibbons, Intel
Garth Gibson, CMU
Carlos Guestrin, CMU
Mor Harchol-Balter, CMU
Alex Hauptmann, CMU
Anthony Joseph, Berkeley
Randy Katz, Berkeley
Kai Li, Princeton
Ling Liu, GA Tech
Michael Kaminsky, Intel
Mike Kozuch, Intel
Margaret Martonosi, Princeton
Todd Mowry, CMU
Onur Mutlu, CMU
Priya Narasimhan, CMU
Padmanabhan (Babu) Pillai, Intel
Calton Pu, GA Tech
Mahadev (Satya) Satyanarayanan, CMU
Karsten Schwan, GA Tech
Dan Siewiorek, CMU
Ion Stoica, Berkeley
Matthew Wolf, GA Tech
Sudhakar Yalamanchili, GA Tech

Staff

Joan Digney, Editor/Web, CMU
Jennifer Gabig, ISTC Admin. Manager, CMU
Michael Stroucken, Admin/Programmer, CMU

Students / Post-Docs

Yoshihisa Abe, CMU
Sameer Agarwal, Berkeley
Hrishikesh Amur, GA Tech
Michael Ashley-Rollman, CMU
Ben Blum, CMU
Kevin Kai-Wei Chang, CMU
Anthony Chivetta, CMU
James Cipar, CMU
Xiaoning Ding, Intel
Bin Fan, CMU
Anshul Gandhi, CMU
Elmer Garduno, CMU
Ali Ghodsi, Berkeley
Joseph Gonzalez, CMU
Michelle Goodstein, CMU
Kiryong Ha, CMU
Liting Hu, GA Tech
Ben Jaiyen, CMU
Deeptal Jayasinghe, GA Tech
Wenhao Jia, Princeton
Lu Jiang, CMU
Mike Kasick, CMU
Soila Kavulya, CMU
Yoongu Kim, CMU
Andy Konwinski, Berkeley
Elie Krevat, CMU
Guatam Kumar, Berkeley
Aapo Kyrola, CMU
Hyeontaek Lim, CMU
Guimin Lin, CMU
Jamie Liu, CMU
Wyatt Lloyd, Princeton
Yucheng Low, CMU
Daniel Lustig, Princeton
Shrikant Mether, CMU
Justin Meza, CMU
Lulian Moraru, CMU
Balaji Palanisamy, GA Tech
Swapnil Patil, CMU
Gennady Pekhimenko, CMU
Amar Phanishayee, CMU
Kai Ren, CMU
Wolfgang Richter, CMU
Tunji Ruwase, CMU
Raja Sambasivan, CMU
Vivek Seshadri, CMU
Ilari Shafer, CMU
Jainam Shah, CMU
Yixiao Shen, CMU
Bin Sheng, CMU
Julian Shun, CMU
Harsha Vardhan Simhadri, CMU
Jiri Simsa, CMU
Anand Suresh, CMU
Wittawat Tantisiriroj, CMU
Alexey Tumanov, CMU
Vijay Vasudevan, CMU
Chengwei Wang, GA Tech
Yifan Wang, CMU
Lin Xiao, CMU
Jin Xin, Princeton
Lianghong Xu, CMU
Ozlem Bilgir Yetim, Princeton
Hanbin Yoon, CMU
Hobin Yoon, GA Tech
Jeff Young, GA Tech
Matei Zaharia, Berkeley
Xu Zhang, CMU
Timothy Zhu, CMU

The ISTC-CC Update

The Newsletter for the Intel Science and Technology Center for Cloud Computing

Carnegie Mellon University
ISTC-CC
CIC 4th Floor
4720 Forbes Avenue
Pittsburgh, PA 15213
T (412) 268-2476

EDITOR

Joan Digney

The ISTC-CC Update provides an update on ISTC-CC activities to increase awareness in the research community.

THE ISTC-CC LOGO

ISTC logo embodies its mission, having four inter-related research pillars (themes) architected to create a strong foundation for cloud computing of the future.

The research agenda of the ISTC-CC is composed of the following four themes.

Specialization: Explores specialization as a primary means for order of magnitude improvements in efficiency (e.g., energy), including use of emerging technologies like non-volatile memory and specialized cores.

Automation: Addresses cloud's particular automation challenges, focusing on order of magnitude efficiency gains from smart resource allocation/scheduling and greatly improved problem diagnosis capabilities.

Big Data: Addresses the critical need for cloud computing to extend beyond traditional big data usage (primarily, search) to efficiently and effectively support Big Data analytics, including the continuous ingest, integration, and exploitation of live data feeds (e.g., video or twitter).

To the Edge: Explores new frameworks for edge/cloud cooperation that can efficiently and effectively exploit billions of context-aware clients and enable cloud-assisted client applications whose execution spans client devices, edge-local cloud resources, and core cloud resources.

Year in Review

This section lists a sampling of significant ISTC-CC occurrences from the last year.

Participants at the First Annual ISTC-CC retreat take part in a breakout session to brainstorm new ideas.

June 2012

» Mike Freedman received a Presidential Early Career Award for Scientists and Engineers (PECASE), in 2011 for "efforts in designing, building, and prototyping a modern, highly scalable, replicated storage cloud system that provides strong robustness guarantees..."

» Guy Blelloch and Phil Gibbons co-organized an NSF Workshop on Research Directions in the Principles of Parallel Computing.

» Michael Kozuch was Co-Chair for OpenCirrus Summit.

» Ling Liu chaired the Cloud Computing track for ICDCS'12. She will also be the general chair of VLDB 2012 to be held in Istanbul, Turkey, Aug 27-Aug 31, 2012.

» Priya Narasimhan will serve as a Program Chair for the 2012 ACM/IFIP/IEEE International Conference on Middleware.

» Calton Pu was awarded the Honorary Chair of the IEEE 8th World Congress on Services (Services 2012).

» ISTC-CC Sponsoring Executive Wen-Hann Wang gave a keynote talk on "Powering the Cloud Computing of the Future" at the OpenCirrus summit in Beijing.

» Margaret Martonosi was appointed the Hugh Trumbull Adams '35 Professor of Computer Science.

» Ling Liu and her co-authors received the Best Paper award at CLOUD'12 for their work on "Reliable State Monitoring in Cloud Datacenters".

» Ling Liu presented "Social Influence based Data Analytics" at the University of Tokyo.

» Ling Liu presented "Connecting Big Data with Big Data Analytics" at FIRST Organization Japan.

» Carlos Guestrin gave GraphLab presentations at MIT, UW, Penn, and Sony.

» Garth Gibson is a member of the Technical Advisory Board for a large new research activity called "VISION Cloud: Virtualised Storage Services Foundation for the Future Internet."

» Matei Zaharia described "Spark and Shark: High-Speed In-Memory Analytics over Hadoop and Hive Data" at the Hadoop Summit. A number of other ISTC-CC folks participated in the Hadoop Summit as well.

May 2012

» A team from the ISTC for Cloud Computing—Babu Pillai, Michael Kaminsky, Mike Kozuch and Dave Andersen—were announced winners in 3 categories of the 2012 JouleSort competition.

» Onur Mutlu and co-authors' paper "MinBD: Minimally-Buffered Deflection Routing for Energy-Efficient Interconnect" nominated for the best paper award at NOCS'12.

» Mor Harchol-Balter was awarded an NSF grant amplifying funding.

» Onur Mutlu was awarded an NSF grant on Rearchitecting the Memory Hierarchy.

April 2012

» M. Satyanarayanan gave a distinguished lecture in the Computer Science Department at the University of Illinois at Urbana-Champaign.

» Ion Stoica and co-authors' SPARK paper named best paper at NSDI'12.

» Ling Liu received IEEE's 2012 Technical Achievement Award "for pioneering contributions to novel internet data management and decentralized trust management."

Year in Review

March 2012

» Karsten Schwan presented "Virtualized Manycore Systems" at the Microsoft Cambridge Systems Workshop.

» Onur Mutlu presented "Bottleneck Identification and Scheduling in Multithreaded Applications," at ASPLOS'12, London, UK.

» Kai Li was elected to the Nat'l Academy of Engineering (NAE) "for advances in data storage and distributed computer systems." Kai joins Randy Katz and Dan Siewiorek as ISTC-CC faculty who are NAE members.

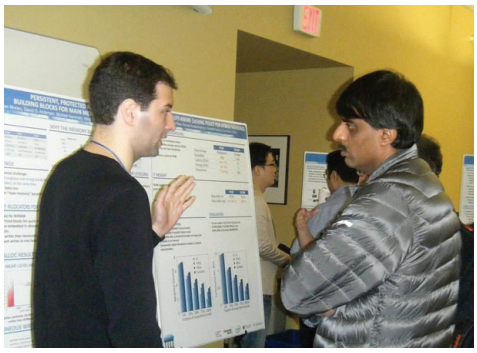
» Randy Katz was appointed to the Board of Advisors of ConteXtream Inc., a leading provider of cloud-scale network virtualization.

» Sudha Yalamanchili received an NVIDIA Corporation Professor Partnership Award. for his work on "Oncilla – Optimizing Data Warehousing Applications for Accelerator Clouds."

February 2012

» Onur Mutlu received the George Tallman Ladd Research Award. The G.T. Ladd award is made to a faculty member within the Carnegie Institute of Technology in recognition of outstanding research and professional accomplishments and potential.

» Garth Gibson presented "Recent Work in Storage Systems for Big-Data" at the NetApp CTO's Distinguished Lecture Series, in Sunnyvale, CA.



Iulian Moraru (CMU) discusses his work on "Persistent, Protected and Cached: Building Blocks for Main Memory Data Stores" with Prashant Chandra (Intel) at an ISTC-CC Retreat poster session.

» Priya Narasimhan was awarded Global Pittsburgh's 2012 International Bridge Award to recognize her entrepreneurial and technology transition efforts.

» Onur Mutlu presented "Some Opportunities and Obstacles in Cross-Layer and Cross-Component (Power) Management," at the NSF Workshop on Cross-Layer Power Optimization and Management (NSF CPOM), Los Angeles, CA.

» Justin Rattner, Intel CTO, visited ISTC-CC headquarters at Carnegie Mellon.

January 2012

» Guy Blelloch gave the keynote talk at ALENEX'12 in Kyoto, Japan on "Problem Based Benchmarks."

» Michael Kozuch was named to Editorial Board of the International Journal of Cloud Computing. The inaugural issue came out this month

» Garth Gibson's proposal for ISR/NSA Science of Security: Systematic Testing of Distributed and Multithreaded Systems at Scale was successful.

» Garth Gibson received a LANL/CMU Institute grant for Reliable High Performance Information Technology (IRHPIT) award. He also received an equipment award from NSF for a High Core Count Computer, as part of the NSF Parallel Reconfigurable Observational Environment for Data Intensive Supercomputing and High End Computing (PRObE).

» Michael Kaminsky began serving as General Chair for SOSP'13.

» Phil Gibbons presented "Intel Science and Technology Center for Cloud Computing" at the Intel-NTU CCC Champion-PI meeting in Taipei, Taiwan.

» Elmer Garduno presented "Using Visual Signatures for Hadoop Diagnosis" at Intel Hadoop/Big-Data Summit Portland, OR.

» Sudha Yalamanchili received a grant from AMD Corp. for his work on "Developing a Heterogeneous Architecture Simulator and Memory System Design Exploration and Extending Ocelot to Support OpenCL."

December 2011

» The first annual ISTC-CC Retreat was held, Dec. 8-9, on CMU campus.

» Jason Waxman (Intel) gave a keynote on "Opportunities in Cloud Computing: Discussion at ISTC Summit" at the first annual ISTC-CC Retreat.

» Rich Uhlig gave a keynote on "Optimizing for the Cloud: Tech Trends, Testbeds and Working Together" at the first annual ISTC-CC Retreat.

» There were 8 additional presentations by students and 6 by ISTC-CC faculty, at the ISTC-CC Retreat, as well as two research poster sessions.

» Randy Katz received a Public Service Medal (Pingat Bakti Masyarakat) from the President of Singapore.

November 2011

» Michael Kozuch presented key observations from Intel's Open Cirrus effort at the "Support for Experimental Computer Science Workshop" at SC'11 in Seattle, WA.

» M. Satyanarayanan gave a keynote talk at the Corning Workshop on Cloud Computing, November 2011. He was the only academic speaker; the others were from Microsoft, IBM, Arista Networks and Myoonet.

» Much of our ISTC-CC research was presented at the PDL Retreat, both as talks (15) and posters. The PDL Retreat was attended by 45 technical leaders from 20 companies, including Intel, Microsoft, Google, Facebook, VMware, EMC, HP, and Oracle. These attendees provide feedback on the research ideas, offer assistance (e.g., data center traces from Google and anecdotes from many), and help to promulgate the work by word of mouth and as technology transfer avenues.

» Ling Liu (Georgia Tech) gave a keynote on cloud computing at the Financial Services conference in Busan, Korea.

» A tutorial "Heterogeneous Computing with GPU Ocelot", was presented by Georgia Tech at the IEEE International Symposium on Workload Characterization, in Austin, TX.

» Garth Gibson presented "Future

continued on pg. 32

ISTC-CC News

July 20, 2012

First GraphLab Workshop on Large-scale Machine Learning

The recent First GraphLab Workshop on Large-scale Machine Learning brought together industry and academic professionals to explore the state-of-the-art on the development of machine-learning techniques for working with huge data sets. The GraphLab Workshop included about 320 participants and 15 talks and demonstrations on systems, abstractions, languages, and algorithms for large-scale data analysis. GraphLab is particularly suited to problems with dependencies in the data, which cannot be easily or efficiently separated into independent subproblems. The workshop also included the release of GraphLab 2.1, an updated abstraction that increases the scalability of GraphLab and GraphChi, which is able to solve Web-scale problems on a single personal computer. Several of the workshop's talks included announcements on new big data developments, including Intel's Ted Willke's announcement of the development of GraphBuilder, which uses Hadoop to overcome the gap between unstructured data and the formation of the data's graph of dependencies. The workshop also featured several short discussions led by participants from Yahoo!, Twitter, Stanford University, Netflix, Pandora, IBM, and One Kings Lane.

July 18, 2012

Endowed Professorship for Margaret Martonosi

We are pleased to announce that Margaret Martonosi has been chosen the Hugh Trumbull Adams '35 Professor of Computer Science at Princeton University. Margaret's research interests include computer architectures and the hardware/software interface, particularly power-efficient systems, and embedded systems issues in mobile networks.

May 25, 2012

ISTC-CC Team wins JouleSort Competition!

A team from the ISTC for Cloud Computing—Babu Pillai, Michael Kaminsky, Mike Kozuch (Intel Labs), and Dave Andersen (CMU)—were announced winners in 3 categories of the 2012 JouleSort competition, setting new records for fewest joules needed to sort 10^8, 10^9, and 10^{10} records. The team used an Intel Core i7-2700K desktop processor, coupled with 16 Intel 710 Series SSDs to beat existing energy efficiency records in the 10GB, 100GB, and 1TB categories by 2.6% (their record from last year), 33%, and 729%, respectively.

May 15, 2012

Priya's Student wins Alumni Award for Undergraduate Excellence in CS

We are pleased to announce that this year's recipient of the Alumni Award for Undergraduate Excellence in Computer Science is Nikhil Khadke, for his work entitled "Transparent System Call Based Performance Debugging for Cloud Computing." Nikhil is advised by Priya Narasimhan.

April 24, 2012

Ling Liu 2012 Technical Achievement Award Recipient

Ling Liu, a Professor in the School of Computer Science at Georgia Institute of Technology, has been awarded a Technical Acheivement Award "for pioneering contributions to novel internet data management and decentralized trust management."

The award is presented for outstanding and innovative contributions to the fields of computer and information science and engineering or computer technology, usually within the past ten, and not more than fifteen years. Ling's research interests are in the areas of database systems, distributed computing, and Internet data management and data mining with the focus on performance, availability, fault tolerance, security and privacy.

-- IEEE Computer Society News

March 30, 2012

Garth Gibson & Randy Katz Receive 2012 Jean-Claude Laprie Award in Dependable Computing

Garth Gibson and Randy Katz have been awarded the 2012 Jean-Claude Laprie Award in Dependable Computing, Industrial/Commercial Product Impact Category, by the IFIP Working Group 10.4 on Dependable Computing and Fault Tolerance. The award is for outstanding papers published at least 10 years ago that have significantly influenced the theory and/or practice of dependable computing, and given for "A Case for Redundant Arrays of Inexpensive Disks (RAID)," by D.A. Patterson, G.A. Gibson, and R.H. Katz, Proc. of 1988 ACM SIGMOD Int. Conf. on Management of Data, June 2, 1998. The groundbreaking paper introduced the concept of RAID, which rapidly became the common configuration paradigm for disks at all but the very low end of the server market. Its impact is primarily to industry where RAID was a truly disruptive technology. The RAID levels as defined in this paper persist to the present day. The paper familiarized development engineers who didn't normally work in the area of High Availability or Fault Tolerance with the concepts of improving reliability and availability through redundancy.

-- http://www.dependability.org/articles/laprie/laprie2012.html

February 23, 2012

Onur Mutlu receives George Tallman Ladd award

Onur Mutlu, assistant professor of electrical and computer engineering has received the George Tallman Ladd

ISTC-CC News

Research Award. The G.T. Ladd award is made to a faculty member within the Carnegie Institute of Technology in recognition of outstanding research and professional accomplishments and potential. The award is in the form of a memento and an honorarium. More than one award may be made in each year and is made based on excellence in research as measured by scholarly publications, research program development, development of funding, and awards and other recognition. Congratulations Onur!

February 13, 2012

Kai Li Elected to NAE

Congratulations to Kai Li, a professor of computer science at the School of Engineering and Applied Science, who has been elected as a member of the National Academy of Engineering, one of the highest professional honors among engineers. Li, the Paul M. and Marsha R. Wythes Professor in Computer Science, was selected for his contributions in the fields of data storage and distributed computer systems from among 66 new members and 10 foreign associates.

-- Princeton School of Engineering News

February 2012

Michelle Mazurek and Hyeontaek Lim Facebook Fellowship Winners!

Two ISTC-CC students have been named winners of a 2012-2013 Facebook Fellowship (12 awarded). Hyeontaek is working to improve the resource efficiency of distributed systems. He hopes to deliver more affordable data-intensive computing, facilitating future innovations for large-scale Internet services. Michelle is researching ways to let users share their content accurately and quickly, secure in the knowledge that only the right people will see it.

The fellowship program began in 2010 to "foster ties to the academic community and support the research of promising computer science Ph.D. students." Each student will be granted full tuition, and stipends for tuition, conference travel and a new computer.

-- Info from http://on.fb.me/wjiCZZ

December 25, 2011

ISTC-CC: CMU's Early Distributed File Systems Work Laid Foundation For Cloud Storage

Many years later, CMU's seminal work on distributed file systems continues to have huge impact on the field. From AFS to Coda, CMU's work has laid a foundation atop which some of the most popular cloud backup and sharing systems, such as Dropbox and iCloud, are built. Check out this Wired article for more.

-- Wired, December 25, 2011

December 8, 2011

Guy Blelloch Named An ACM Fellow

Congratulations to Guy Blelloch, who has been made an ACM Fellow for his contributions to parallel computing. The ACM Fellows Program was established by Council in 1993 to recognize and honor outstanding ACM members for their achievements in computer science and information technology and for their significant contributions to the mission of the ACM. The ACM Fellows serve as distinguished colleagues to whom the ACM and its members look for guidance and leadership as the world of information technology evolves.

There are 6 other ISTC-CC faculty who are ACM Fellows: Dan Siewiorek (CMU, inducted 1994), Randy Katz (UC Berkeley, 1996), Kai Li (Princeton, 1998), M. Satyanaryanan (CMU, 2002), Phil Gibbons (Intel, 2006), Margaret Martonosi (Princeton, 2009).

October 25, 2011

Garth Gibson's & Randy Katz's 1988 RAID Paper Enters SIGOPS Hall Of Fame

We are very pleased to announce that Garth Gibson's original RAID paper from SIGMOD 1988 — "A Case for Redundant Array of Inexpensive Disks" by Patterson, Gibson and Katz — was one of the four papers to be honored as a 2011 SIGOPS Hall of Fame Award paper. The award was made at the 23rd ACM Symposium on Operating Systems Principles (SOSP), October 23-26, 2011, Cascais, Portugal.

The SIGOPS Hall of Fame Award was instituted in 2005 to recognize the most influential Operating Systems papers that were published at least ten years in the past. The Hall of Fame Award Committee consists of past program chairs from SOSP, OSDI, EuroSys, past Weiser and Turing Award winners from the SIGOPS community, and representatives of each of the Hall of Fame Award papers.

October 20, 2011

ISTC-CC Officially Launched

The Cloud Computing ISTC had an opening ceremony on October 19th at CMU with Wen-Hann Wang (Executive Sponsor) and Rich Uhlig (Managing Sponsor) in attendance along with Jared Cohon (CMU President) and Mark Kamlet (CMU Provost). Also in attendance were Limor Fix, Matt Haycock, Vida Ilderem, and Ravi Iyer. Other notables on the CMU side included Randy Bryant (Dean of the school of Computer Science), Ed Schlesinger (Dept Chair of Electrical and Computer Engineering) and Pradeep Khosla (Dean of the College of Engineering). The ISTC-CC was announced by the media on August 3, 2011 and officially began operation on September 1, 2011.

August 3, 2011

Intel Labs Invests $30M in the Future of Cloud and Embedded Computing with the Opening of Latest Intel Science and Technology Centers

Aimed at shaping the future of cloud computing and how increasing numbers of everyday devices will add computing capabilities, Intel Labs announced the latest Intel Science and Technology Centers (ISTC) for Cloud Computing Research and for Embed-

continued on pg. 29

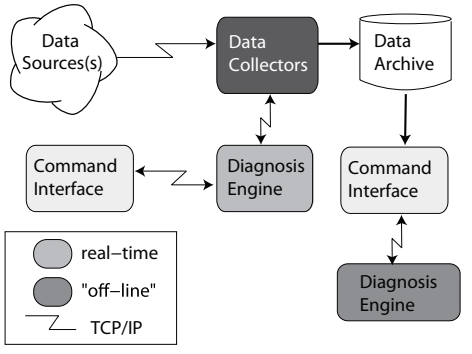
Recent Publications

Draco: Statistical Diagnosis of Chronic Problems in Large Distributed Systems

Soila Pertet Kavulya, Kaustubh Joshi, Matti Hiltunen, Scott Daniels, Rajeev Gandhi, Priya Narasimhan

IEEE/IFIP Conference on Dependable Systems and Networks (DSN'12), June 2012.

Chronics are recurrent problems that often fly under the radar of operations teams because they do not affect enough users or service invocations to set off alarm thresholds. In contrast with major outages that are rare, often have a single cause, and as a result are relatively easy to detect and diagnose quickly, chronic problems are elusive because they are often triggered by complex conditions, persist in a system for days or weeks, and coexist with other problems active at the same time. In this paper, we present Draco, a scalable engine to diagnose chronics that addresses these issues by using a "top-down" approach that starts by heuristically identifying user interactions that are likely to have failed, e.g., dropped calls, and drills down to identify groups of properties that best explain the difference between failed and successful interactions by using a scalable Bayesian learner. We have deployed Draco in production for the VoIP operations of a major ISP. In addition to providing examples of chronics that Draco has helped identify, we show via a comprehensive evaluation on production data that Draco provided 97% coverage, had fewer than 4% false positives, and outperformed state-of-the-art diagnostic techniques by up to 56% for complex chronics.



Draco's flexible architecture supports multiple data sources, and the diagnosis engines can run in either real-time or offline mode.

Reliable State Monitoring in Cloud Datacenters

Shicong Meng, Arun Iyengar, Isabelle Rouvellou, Ling Liu, Kisung Lee, Balaji Palanisamy

Proceedings of the 2012 IEEE CLOUD Conference (CLOUD'12), June 2012.

State monitoring is widely used for detecting critical events and abnormalities of distributed systems. As the scale of such systems grows and the degree of workload consolidation increases in Cloud datacenters, node failures and performance interferences, especially transient ones, become the norm rather than the exception. Hence, distributed state monitoring tasks are often exposed to impaired communication caused by such dynamics on different nodes. Unfortunately, existing distributed state monitoring approaches are often designed under the assumption of always online distributed monitoring nodes and reliable inter-node communication. As a result, these approaches often produce misleading results which in turn introduce various problems to Cloud users who rely on state monitoring results to perform automatic management tasks such as auto-scaling.

This paper introduces a new state monitoring approach that tackles this challenge by exposing and handling communication dynamics such as message delay and loss in Cloud monitoring environments. Our approach delivers two distinct features. First, it quantitatively estimates the accuracy of monitoring results to capture uncertainties introduced by messaging dynamics. This feature helps users to distinguish trustworthy monitoring results from ones heavily deviated from the truth, yet significantly improves monitoring utility compared with simple techniques that invalidate all monitoring results generated with the presence of messaging dynamics. Second, our approach also adapts to non-transient messaging issues by reconfiguring distributed monitoring algorithms to minimize monitoring errors. Our experimental results show that, even under severe message loss and delay, our approach consistently improves monitoring accuracy, and when applied to Cloud application auto-scaling, outperforms existing state monitoring techniques in terms of the ability to correctly trigger dynamic provisioning.

Expertus: A Generator Approach to Automate Performance Testing in IaaS Clouds

Deepal Jayasinghe, Galen Swint, Simon Malkowski, Jack Li, Qingyang Wang, Calton Pu

Proceedings of the 2012 IEEE CLOUD Conference (CLOUD'12), June 2012.

Cloud computing is an emerging technology paradigm that revolutionizes the computing landscape by providing on-demand delivery of software, platform, and infrastructure over the Internet. Yet, architecting, deploying, and configuring enterprise applications to run well on modern clouds remains a challenge due to associated complexities and non-trivial implications. The natural and presumably unbiased approach to these questions is thorough testing before moving applications to production settings. However, thorough testing of enterprise applications on modern clouds is cumbersome and error-prone due to a large number of relevant scenarios and difficulties in testing process. We address some of these challenges through Expertus—a flexible code generation framework for automated performance testing of distributed applications in Infrastructure as a Service (IaaS) clouds. Expertus uses a multi-pass compiler approach and leverages template-driven code generation to modularly incorporate different software applications on IaaS clouds. Expertus automatically handles complex configuration dependencies of software applications and significantly reduces human errors associated with manual approaches for software configuration and testing. To date, Expertus has been used to study three distributed applications on five IaaS clouds with over 10,000 different hardware, software, and virtualization configurations. The flexibility and extensibility of Expertus and our

Recent Publications

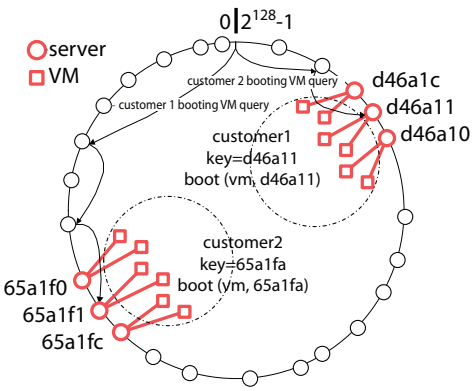
own experience on using it shows that new clouds, applications, and software packages can easily be incorporated.

v-Bundle: Flexible Group Resource Offerings in Clouds

Liting Hu, Kyung Dong Ryu, Dilma Da Silva, Karsten Schwan

ICDCS'12, June 2012.

Traditional Infrastructure-as-a-Service offerings provide customers with large numbers of fixed-size virtual machine (VM) instances with resource allocations that are designed to meet application demands. With application demands varying over time, cloud providers gain efficiencies through resource consolidation and over-commitment. For cloud customers, however, this leads to inefficient use of the cloud resources they have purchased. To address cloud customers' dynamic application requirements, we present a new cloud resource offering, called v-Bundle, which makes flexible the exchange of resource capacity among multiple VM instances belonging to the same customer. Specifically targeting network resources, for each customer application, we first use DHT-based techniques to achieve an initial VM placement that minimizes its use of the datacenter network's bi-section bandwidth. When VMs' networking requirements change, the customer can then use v-Bundle to trade the networking resources allocated to her application. v-Bundle maintains information about network resources with any-cast tree-based methods implemented as extensions of the Pastry pub-sub core.



v-Bundle's topology-aware VM Placement.

Experimental evaluations show that the approach can scale well to thousands of hosts and VMs, and that v-Bundle can provide customers with better bandwidth utilization and improved application quality of service through borrowing extra bandwidth when needed, at no additional cost in terms of the total resources allocated to the customer.

Characterizing and Improving the Use of Demand-Fetched Caches in GPUs

Wenhao Jia, Kelly A. Shaw, Margaret Martonosi

Proceedings of the 26th International Conference on Supercomputing (ICS '12), June 2012.

Initially introduced as special-purpose accelerators for games and graphics code, graphics processing units (GPUs) have emerged as widely-used high-performance parallel computing platforms. GPUs traditionally provided only software-managed local memories (or scratchpads) instead of demand-fetched caches. Increasingly, however, GPUs are being used in broader application domains where memory access patterns are both harder to analyze and harder to manage in software-controlled caches.

In response, GPU vendors have included sizable demand-fetched caches in recent chip designs. Nonetheless, several problems remain. First, since these hardware caches are quite new and highly-configurable, it can be difficult to know when and how to use them; they sometimes degrade performance instead of improving it. Second, since GPU programming is quite distinct from general-purpose programming, application programmers do not yet have solid intuition about which memory reference patterns are amenable to demand-fetched caches. In response, this paper characterizes application performance on GPUs with caches and provides a taxonomy for reasoning about different types of access patterns and locality. Based on this taxonomy, we present an algorithm which can be automated and applied at compile-time to identify an application's memory access patterns and to use that information to intelligently configure cache usage to improve application performance. Experiments on real GPU systems show that our algorithm reliably predicts when GPU caches will help or hurt performance. Compared to always passively turning caches on, our method can increase the average benefit of caches from 5.8% to 18.0% for applications that have significant performance sensitivity to caching.

Brief Announcement: The Problem Based Benchmark Suite

Julian Shun, Guy E. Blelloch, Jeremy T. Fineman, Phillip B. Gibbons, Aapo Kyrola, Harsha Vardhan Simhadri, Kanat Tangwongsan

Proceedings of the 24th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA'12), June 2012.

This announcement describes the problem based benchmark suite (PBBS). PBBS is a set of benchmarks designed for comparing parallel algorithmic approaches, parallel programming language styles, and machine architectures across a broad set of problems. Each benchmark is defined concretely in terms of a problem specification and a set of input distributions. No requirements are made in terms of algorithmic approach, programming language, or machine architecture. The goal of the benchmarks is not only to compare runtimes, but also to be able to compare code and other aspects of an implementation (e.g., portability, robustness, determinism, and generality). As such the code for an implementation of a benchmark is as important as its runtime, and the public PBBS repository will include both code and performance results.

The benchmarks are designed to make it easy for others to try their own implementations, or to add new benchmark problems. Each benchmark problem includes the problem specification, the specification of input and output file formats, default input generators, test codes that check the correctness of the output for a given input, driver code

continued on pg. 10

Recent Publications

continued from pg. 9

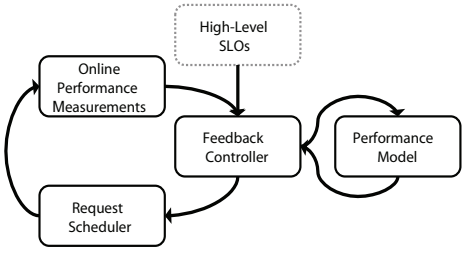
that can be linked with implementations, a baseline sequential implementation, a baseline multicore implementation, and scripts for running timings (and checks) and outputting the results in a standard format. The current suite includes the following problems: integer sort, comparison sort, remove duplicates, dictionary, breadth first search, spanning forest, minimum spanning forest, maximal independent set, maximal matching, K-nearest neighbors, Delaunay triangulation, convex hull, suffix arrays, n-body, and ray casting. For each problem, we report the performance of our baseline multicore implementation on a 40-core machine.

Sweet Storage SLOs with Frosting

Andrew Wang, Shivaram Venkataraman, Sara Alspaugh, Ion Stoica, Randy Katz

HotCloud'12, Boston, MA, June 2012.

Modern datacenters support a large number of applications with diverse performance requirements. These performance requirements are expressed at the application layer as high-level service-level objectives (SLOs). However, large-scale distributed storage systems are unaware of these high-level SLOs. This lack of awareness results in poor performance when workloads from multiple applications are consolidated onto the same storage cluster to increase utilization. In this paper, we argue that because SLOs are expressed at a high level, a high-level control mechanism is required. This is in contrast to existing approaches, which use block- or disk-level mechanisms. These require manual translation of high-level requirements into low-level parameters. We present Frosting, a request scheduling layer on top of a distributed storage system that allows application programmers to specify their high-level SLOs directly. Frosting improves over the state-of-the-art by automatically translating high-level SLOs into internal scheduling parameters and uses feedback control to adapt these parameters to changes in the workload. Our preliminary results demonstrate that our overlay approach can multiplex both latency-sensitive and batch applications to increase utilization, while still maintaining a 100ms 99th percentile latency SLO for latency sensitive clients.



Frosting design: High-level SLOs are translated by the feedback controller into parameter settings in the request scheduler.

Commodity Converged Fabrics for Global Address Spaces in Accelerator Clouds

J. Young and S. Yalamanchili

IEEE International Conference on High Performance Computing Communications (HPCC'12), June 2012.

Hardware support for Global Address Spaces (GAS) has previously focused on providing efficient access across remote memories, typically using custom interconnects or high-level software layers. New technologies, such as Extoll, HyperShare, and NumaConnect now allow for cheaper ways to build GAS support into the data center, thus making high-performance coherent and non-coherent remote memory access available for standard data center applications.

At the same time, data center designers are currently experimenting with a greater use of accelerators like GPUs to enhance traditionally CPU-oriented processes, such as data warehousing queries for in-core databases. However, there are very few workable approaches for these accelerator clusters that both use commodity interconnects and also support simple multi-node programming models, such as GAS.

We propose a new commodity-based approach for supporting non-coherent GAS in accelerator clouds using the Hyper-Transport Consortium's HyperTransport over Ethernet (HToE) specification. This work details a system model for using HToE for accelerated data warehousing applications and investigates potential bottlenecks and design optimizations for an HToE network adapter, or HyperTransport Ethernet Adapter (HTEA).

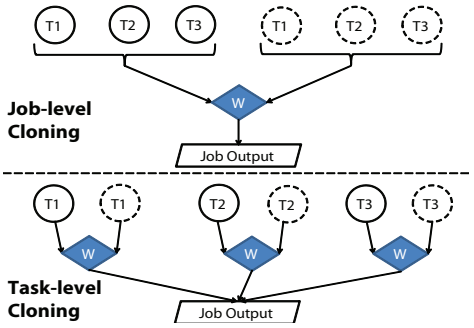
Using a detailed network simulator model and timing measured for queries run on high-end GPUs [34], we find that the addition of wider de-encapsulation pipelines and the use of bulk acknowledgments in the HTEA can improve overall throughput and reduce latency for multiple senders using a common accelerator. Furthermore, we show that the bandwidth of one receiving HTEA can vary from 2.8 Gbps to 24.45 Gbps, depending on the optimizations used, and the inter-HTEA latency for one packet is 1,480 ns. A brief analysis of the path from remote memory to accelerators also demonstrates that the bandwidth of today's GPUs can easily handle a stream-based computation model using HToE.

Why Let Resources Idle? Aggressive Cloning of Jobs with Dolly

Ganesh Ananthanarayanan, Ali Ghodsi, Scott Shenker, Ion Stoica

HotCloud'12, Boston, MA, June 2012.

Despite prior research on outlier mitigation, our analysis of jobs from the Facebook cluster shows that outliers still occur, especially in small jobs. Small jobs are particularly sensitive to long-running outlier tasks because of their interactive nature. Outlier mitigation strategies rely on comparing different tasks of the same job and launching speculative copies for the slower tasks. However, small jobs execute all their tasks simultaneously, thereby not providing sufficient time to observe and compare tasks. Building on the observation that clusters are underutilized, we take speculation to its logical extreme—run full clones of jobs to mitigate the effect of outliers. The heavy-tail distribution of job sizes implies that we can impact most jobs without using much resources. Trace-



Job-level and task-level cloning for a job with 3 tasks with and 2 clones. The rhombus with "W" picks the winner between the clones, i.e., the earliest.

driven simulations show that average completion time of all the small jobs improves by 47% using cloning, at the cost of just 3% extra resources.

Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters

Matei Zaharia, Tathagata Das, Haoyuan Li, Scott Shenker, Ion Stoica

HotCloud'12, Boston, MA, June 2012.

Many important "big data" applications need to process data arriving in real time. However, current programming models for distributed stream processing are relatively low-level, often leaving the user to worry about consistency of state across the system and fault recovery. Furthermore, the models that provide fault recovery do so in an expensive manner, requiring either hot replication or long recovery times. We propose a new programming model, discretized streams (D-Streams), that offers a high-level functional programming API, strong consistency, and efficient fault recovery. D-Streams support a new recovery mechanism that improves efficiency over the traditional replication and upstream backup solutions in streaming databases: parallel recovery of lost state across the cluster. We have prototyped D-Streams in an extension to the Spark cluster computing framework called Spark Streaming, which lets users seamlessly intermix streaming, batch and interactive queries.

A Case for Performance-Centric Network Allocation

Gautam Kumar, Mosharaf Chowdhury, Sylvia Ratnasamy, Ion Stoica

HotCloud'12, Boston, MA, June 2012.

We consider the problem of allocating network resources across applications in a private cluster running data-parallel frameworks. Our primary observation is that these applications have different communication requirements and thus require different support from the network to effectively parallelize. We argue that network resources should be shared in a performance-centric fashion that aids parallelism and allows developers to reason about the overall performance of their applications. This paper tries to address the question of whether/how fairness-centric proposals relate to a performance-centric approach for different communication patterns common in these frameworks and engages in a quest for a unified mechanism to share the network in such settings.

Staged Memory Scheduling: Achieving High Performance and Scalability in Heterogeneous Systems

Rachata Ausavarungnirun, Gabriel Loh, Kevin Chang, Lavanya Subramanian, Onur Mutlu

Proceedings of the 39th International Symposium on Computer Architecture (ISCA'12), Portland, OR, June 2012.

When multiple processor (CPU) cores and a GPU integrated together on the same chip share the off-chip main memory, requests from the GPU can heavily interfere with requests from the CPU cores, leading to low system performance and starvation of CPU cores. Unfortunately, state-of-the-art application-aware memory scheduling algorithms are ineffective at solving this problem at low complexity due to the large amount of GPU traffic. A large and costly request buffer is needed to provide these algorithms with enough visibility across the global request stream, requiring relatively complex hardware implementations.

This paper proposes a fundamentally new approach that decouples the memory controller's three primary tasks into three significantly simpler structures that together improve system performance and fairness, especially in integrated CPU-GPU systems. Our three-stage memory controller first groups requests based on row-buffer locality. This grouping allows the second stage to focus only on inter-application request scheduling. These two stages enforce high-level policies regarding performance and fairness, and therefore the last stage consists of simple per-bank FIFO queues (no further command reordering within each bank) and straightforward logic that deals only with low-level DRAM commands and timing.

We evaluate the design trade-offs involved in our Staged Memory Scheduler (SMS) and compare it against three state-of-the-art memory controller designs. Our evaluations show that SMS improves CPU performance without degrading GPU frame rate beyond a generally acceptable level, while being significantly less complex to implement than previous application-aware schedulers. Furthermore, SMS can be configured by the system software to prioritize the CPU or the GPU at varying levels to address different performance needs.

Saving Cash by Using Less Cache

Anshul Gandhi, Timothy Zhu, Mor Harchol-Balter, Michael Kozuch

HotCloud'12, Boston, MA, June 2012.

Everyone loves a large caching tier in their multitier cloud-based web service because it both alleviates database load and provides lower request latencies. Even when load drops severely, administrators are reluctant to scale down their caching tier. This paper makes the case that (i) scaling down the caching tier is viable with respect to performance, and (ii) the savings are potentially huge; e.g., a 4x drop in load can result in 90% savings in the caching tier size.

continued on pg. 12

Recent Publications

continued from pg. 11

Automated Diagnosis without Predictability is a Recipe for Failure

Raja R. Sambasivan, Gregory R. Ganger

Proceedings of the 4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud '12), June 2012, Boston, MA.

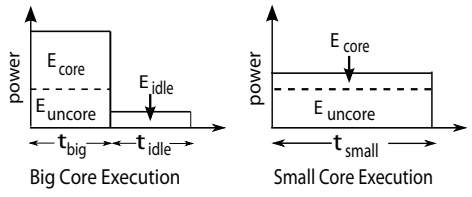
Automated management is critical to the success of cloud computing, given its scale and complexity. But, most systems do not satisfy one of the key properties required for automation: predictability, which in turn relies upon low variance. Most automation tools are not effective when variance is consistently high. Using automated performance diagnosis as a concrete example, this position paper argues that for automation to become a reality, system builders must treat variance as an important metric and make conscious decisions about where to reduce it. To help with this task, we describe a framework for reasoning about sources of variance in distributed systems and describe an example tool for helping identify them.

The Forgotten 'Uncore': On the Energy-Efficiency of Heterogeneous Cores

Vishal Gupta, Paul Brett, David Koufaty, Dheeraj Reddy, Scott Hahn, Karsten Schwan, Ganapati Srinivasa

Short Paper, Usenix ATC'12, Boston, MA, June 2012.

Heterogeneous multicore processors (HMPs), consisting of cores with different performance/power characteristics, have been proposed to deliver higher energy efficiency than symmetric multicores. This paper investigates the opportunities and limitations in using HMPs to gain energy-efficiency. Unlike previous work focused on server systems, we focus on the client workloads typically seen in modern end-user devices. Further, beyond considering core power usage, we also consider the 'uncore' subsystem shared by all cores, which in modern platforms, is an increasingly important contributor to total SoC power. Experimental evaluations use client applications and usage scenarios seen on mobile devices and a unique testbed comprised of heterogeneous cores, with results that highlight the need for uncore-awareness and uncore scalability to maximize intended efficiency gains from heterogeneous cores.



Effect of uncore power on the energy efficiency of heterogeneous cores.

Evaluating the Need for Complexity in Energy-Aware Management for Cloud Platforms

Pooja Ghumre, Junwei Li, Mukil Kesavan, Ada Gavrilovska, Karsten Schwan

Greenmetrics'12, in conjunction with Sigmetrics, London, UK, June 2012.

In order to curtail the continuous increase in power consumption of modern datacenters, researchers are responding with sophisticated energy-aware workload management methods. This increases the complexity and cost of the management operation, and may lead to increases in failure rates. The goal of this paper is to illustrate that there exists considerable diversity in the effectiveness of different, potentially 'smarter' workload management methods depending on the target metric or the characteristics of the workload being managed. We conduct experiments on a datacenter prototype platform, virtualized with the VMware vSphere software, and using representative cloud applications { a distributed key-value store and a map-reduce computation. We observe that, on our testbed, different workload placement decisions may be quite effective for some metrics, but may lead to only marginal impact on others. In particular, we are considering the impact on energy-sensitive metrics, such as power or temperature, as corresponding energy-aware management methods typically come with greater complexity due to fact that they must consider the complex energy consumption trends of various components in the cloud infrastructure. We show that for certain applications, such costs can be avoided, as different management policies and placement decisions have marginal impact on the target metric. The objective is to understand whether for certain classes of applications, and/or application configurations, it is necessary, or it is possible, to avoid the use of complex management methods.

Are Sleep States Effective in Data Centers?

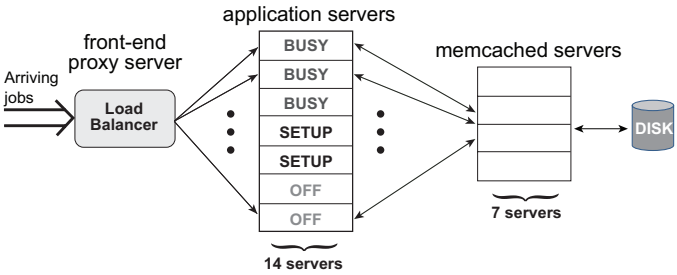
Anshul Gandhi, Michael Kozuch, Mor Harchol-Balter

Proceedings of the International Green Computing Conference (IGCC'12), San Jose, CA, June 2012.

While sleep states have existed for mobile devices and workstations for some time, these sleep states have not been incorporated into most of the servers in today's data centers. High setup times make data center administrators fearful of any form of dynamic power management, whereby servers are suspended or shut down when load drops. This general reluctance has stalled research into whether there might be some feasible sleep state (with sufficiently low setup overhead and/or sufficiently low power) that would actually be beneficial in data centers.

This paper investigates the regime of sleep states that would be advantageous in data centers. We consider the benefits of sleep states across three orthogonal dimensions: (i) the variability in the workload trace, (ii) the type of dynamic power management policy employed, and (iii) the size of the data center.

Our implementation results on a 24-server multi-tier testbed indicate that under many traces, sleep states greatly enhance dynamic power management. In fact, given the right sleep states, even a naive policy that simply tries to match capacity with demand,



Sleep state in servers: experimental setup.

can be very effective. By contrast, we characterize certain types of traces for which even the "best" sleep state under consideration is ineffective. Our simulation results suggest that sleep states are even more beneficial for larger data centers.

A Cyber-Physical Integrated System for Application Performance and Energy Management in Data Centers

Hui Chen, Pengcheng Xiong, Ada Gavrilovska, Karsten Schwan and Chengzhong Xu

Third International Green Computing Conference (IGCC'12), San Jose, CA, June 2012.

Both performance and energy cost are important concerns for current data center operators. Traditionally, however, IT and mechanical engineers have separately optimized the cyber vs. physical aspects of data center operations. In contrast, the work presented in this paper jointly considers both the IT - cyber - and the physical systems in data centers, the eventual goal being to develop performance and power management techniques that holistically operate to control the entire complex of data center installations. Toward this end, we propose a balance of payments model for holistic power and performance management. As an example of coordinated data center management system, the energy-aware cyber-physical system (EaCPS) uses an application controller on the cyber side to guarantee application performance, and on the physical side, it utilizes electric current-aware capacity management (CACM) to smartly place executables to reduce the energy consumption of each chassis present in a data center rack. A web application, representative of a multi-tier web site, is used to evaluate the performance of the controller on the cyber side, the CACM control on the physical side, and of the holistic EaCPS methods in a mid-size, instrumented data center. Results indicate that coordinated EaCPS outperforms the cyber and physical control modules working separately.

Understanding TCP Incast and Its Implications for Big Data Workloads

Y. Chen, R. Griffith, D. Zats, A. D. Joseph, R. H. Katz

USENIX ;login;, 37(3), June 2012.

This article develops and validates a quantitative model that accurately predicts the onset of incast and TCP behavior both before and after. This article also investigates how incast affects the Apache Hadoop implementation of MapReduce, an important example of a big data application. We close the article by reflecting on some technology and data analysis trends surrounding big data, speculate on how these trends interact with incast, and make recommendations for datacenter operators.

Kilo-NOC: A Heterogeneous Network-on-Chip Architecture for Scalability and Service Guarantees

Boris Grot, Joel Hestness, Stephen W. Keckler, and Onur Mutlu

IEEE Micro, Special Issue: Micro's Top Picks from 2011 Computer Architecture Conferences, May-June 2012.

Today's chip-level multiprocessors (CMPs) feature up to a hundred discrete cores, and with increasing levels of integration, CMPs with hundreds of cores, cache tiles, and specialized accelerators are anticipated in the near future. In this paper, we propose and evaluate technologies to enable networks-on-chip (NOCs) to support a thousand connected components (Kilo-NOC) with high area and energy efficiency, good performance, and strong quality-of-service (QOS) guarantees. Our analysis shows that QOS support burdens the network with high area and energy costs. In response, we propose a new lightweight topology-aware QOS architecture that provides service guarantees for applications such as consolidated servers on CMPs and real-time SOCs. Unlike prior NOC quality-of-service proposals which require QOS support at every network node, our scheme restricts the extent of hardware support to portions of the die, reducing router complexity in the rest of the chip. We further improve network area- and energy-efficiency through a novel flow control mechanism that enables a single-network, low-cost elastic buffer implementation. Together, these techniques yield a heterogeneous Kilo-NOC architecture that consumes 45% less area and 29% less power than a state-of-the-art QOS-enabled NOC without these features.

Dynamic Management of Resources and Workloads for RDBMS in Cloud: a Control-theoretic Approach

Pengcheng Xiong

Proc. of ACM SIGMOD/PODS PhD Symposium, Scottsdale, AZ, May 2012.

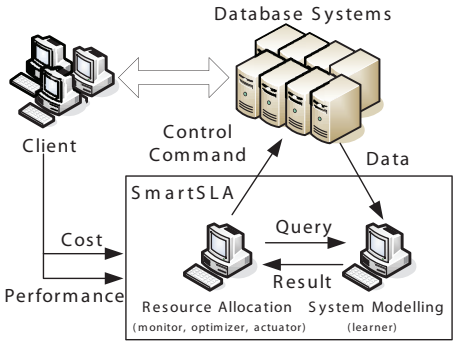
As cloud computing environments become explosively popular, dealing with unpredictable changes, uncertainties, and disturbances in both systems and environments turns out to be one of the major challenges facing the concurrent computing industry. My research goal is to dynamically manage resources and workloads for RDBMS in cloud computing environments in order to achieve "better performance but lower cost", i.e., better service level compliance but lower consumption of virtualized computing resource(s).

Nowadays, although control theory offers a principled way to deal with the challenge based on feedback mecha-

continued on pg. 14

Recent Publications

continued from pg. 13



The architecture of our testbed.

nisms, a controller is typically designed based on the system designer's domain knowledge and intuition instead of the behavior of the system being controlled. My research approach is based on the essence of control theory but transcends state-of-the-art control-theoretic approaches by leveraging interdisciplinary areas, especially from machine learning. While machine learning is often viewed merely as a toolbox that can be deployed for many data-centric problems, my research makes efforts to incorporate machine learning as a full- edged engineering discipline into control-theoretic approaches for realizing my research goal.

My PhD thesis work implements two solid systems by leveraging machine learning techniques, namely, ActiveSLA and SmartSLA. ActiveSLA is an automatic controller featuring risk assessment admission control to obtain the most profitable service-level compliance. SmartSLA is an automatic controller featuring cost-sensitive adaptation to achieve the lowest total cost. The experimental results show that both of the two systems outperform the state-of-the-art methods.

Dynamic Compilation of Data Parallel Kernels for Vector Processors

A. Kerr, G. Diamos, S. Yalamanchili

International Symposium on Code Generation and Optimization, April 2012.

Modern processors enjoy augmented throughput and power efficiency through specialized functional units leveraged via instruction set extensions. These functional units accelerate performance for specific types of operations but must be programmed explicitly. Moreover, applications targeting these specialized units will not take advantage of future ISA extensions and tend not to be portable across multiple ISAs. As architecture designers increasingly rely on heterogeneity for performance improvements, the challenges of leveraging specialized functional units will only become more critical. In particular, exploiting software parallelism without sacrificing portability across the spectrum of commodity and multi-core SIMD processors remains elusive.

This work applies dynamic compilation to explicitly data-parallel kernels and describes a set of program transformations that efficiently compile bulk-synchronous scalar kernels for SIMD functional units while tolerating control-flow divergence. It is agnostic to specific features of ISAs, and performance scalability is expected from 2-wide to arbitrary-width vector units. This technique is evaluated with existing workloads originally targeting GPU computing. A microbenchmark written in CUDA achieving near peak throughput on a GPU achieves over 90% peak throughput on an Intel Sandybridge. Speedups for real-world applications running on CPUs featuring SSE4 achieve up to 3.9x over current state of the art heterogeneous compilers for data-parallel workloads.

Interactive Use of Cloud Services: Amazon SQS and S3

Hobin Yoon, Jim Donahue, Ada Gavrilovska, Karsten Schwan

12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid'12), Ottawa, ON, Canada, May 2012.

Interactive use of cloud services is of keen interest to science end users, including for storing and accessing shared data sets. This paper evaluates the viability of interactively using two important cloud services offered by Amazon: SQS (Simple Queue Service) and S3 (Simple Storage Service). Specifically, we first measure the send-to-receive message latencies of SQS and then determine and devise rate controls to obtain suitable latencies and latency variations. Second, for S3, when transferring data into the cloud, we determine that increased parallelism in Transfer Manager can significantly improve upload performance, achieving up to 4 times improvements with careful elimination of upload bottlenecks.

Optimizing Data Warehousing Applications for GPUs using Kernel Fusion/Fission

H. Wu, G. Diamos, J. Wang, H. Cadambi, S. Yalamanchili, S. Chakradhar

Workshop on Multicore and GPU Programming Models, Languages and Compilers, May 2012.

Data warehousing applications represent an emergent application arena that requires the processing of relational queries and computations over massive amounts of data. Modern general purpose GPUs are high core count architectures that potentially offer substantial improvements in throughput for these applications. However, there are significant challenges that arise due to the overheads of data movement through the memory hierarchy and between the GPU and host CPU. This paper proposes a set of compiler optimizations to address these challenges.

Inspired in part by loop fusion/fission optimizations in the scientific computing community, we propose kernel fusion and kernel fission. Kernel fusion fuses the code bodies of two GPU kernels to i) eliminate redundant operations across dependent kernels, ii) reduce data movement between GPU registers and GPU memory, iii) reduce data movement between GPU memory and CPU memory, and iv) improve spatial and temporal locality of memory references. Kernel fission partitions a kernel into segments such that segment computations and data transfers between the GPU and host CPU can be overlapped. Fusion and fission can also be applied concurrently to a set of kernels. We empirically evaluate the benefits of fusion/fission on relational algebra operators drawn from

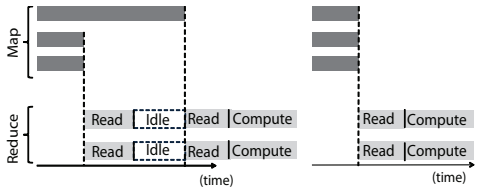
the TPC-H benchmark suite. All kernels are implemented in CUDA and the experiments are performed with NVIDIA Fermi GPUs. In general, we observed data throughput improvements ranging from 13.1% to 41.4% for the SELECT operator and queries Q1 and Q21 in the TPC-H benchmark suite. We present key insights, lessons learned, and opportunities for further improvements.

PACMan: Coordinated Memory Caching for Parallel Jobs

Ganesh Anantharanayanan, Ali Ghodsi, Andrew Wang, Dhruba Borthakur, Srikanth Kandula, Scott Shenker, Ion Stoica

NSDI '12, San Jose, CA, April 2012.

Data-intensive analytics on large clusters is important for modern Internet services. As machines in these clusters have large memories, in-memory caching of inputs is an effective way to speed up these analytics jobs. The key challenge, however, is that these jobs run multiple tasks in parallel and a job is sped up only when inputs of all such parallel tasks are cached. Indeed, a single task whose input is not cached can slow down the entire job. To meet this "all-or-nothing" property, we have built PACMan, a caching service that coordinates access to the distributed caches. This coordination is essential to improve job completion times and cluster efficiency. To this end, we have implemented two cache replacement policies on top of PACMan's coordinated infrastructure – LIFE that minimizes average completion time by evicting large incomplete inputs, and LFU-F that maximizes cluster efficiency by evicting less frequently accessed inputs. Evaluations on production workloads from Facebook and Microsoft Bing show that PACMan reduces average completion time of jobs by 53%and 51%(small interactive jobs improve by 77%), and improves efficiency of the cluster by 47% and 54%, respectively.



All-or-nothing property matters for efficiency. In this example of a job with 3 map tasks and 2 reduce tasks, even if one map task is delayed (due to lack of memory locality), reduce tasks idle and hurt efficiency.

MinBD: Minimally-Buffered Deflection Routing for Energy-Efficient Interconnect

Chris Fallin, Greg Nazario, Xiangyao Yu, Kevin Chang, Rachata Ausavarungnirun, Onur Mutlu

Proceedings of the 6th ACM/IEEE International Symposium on Networks on Chip (NOCS'12), Lyngby, Denmark, May 2012.

A conventional Network-on-Chip (NoC) router uses input buffers to store in-flight packets. These buffers improve performance, but consume significant power. It is possible to bypass these buffers when they are empty, reducing dynamic power, but static buffer power, and dynamic power when buffers are utilized, remain. To improve energy efficiency, bufferless deflection routing removes input buffers, and instead uses deflection (misrouting) to resolve contention. However, at high network load, deflections cause unnecessary network hops, wasting power and reducing performance.

In this work, we propose a new NoC router design called the minimally-buffered deflection (MinBD) router. This router combines deflection routing with a small "side buffer," which is much smaller than conventional input buffers. A MinBD router places some network traffic that would have otherwise been deflected in this side buffer, reducing deflections significantly. The router buffers only a fraction of traffic, thus making more efficient use of buffer space than a router that holds every flit in its input buffers. We evaluate MinBD against input-buffered routers of various sizes that implement buffer bypassing, a bufferless router, and a hybrid design, and show that MinBD is more energy-efficient than all prior designs, and has performance that approaches the conventional input-buffered router with area and power close to the bufferless router.

Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing

Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica

NSDI '12, San Jose, CA, April, 2012.

We present Resilient Distributed Datasets (RDDs), a distributed memory abstraction that lets programmers perform in-memory computations on large clusters in a fault-tolerant manner. RDDs are motivated by two types of applications that current computing frameworks handle inefficiently: iterative algorithms and interactive data mining tools. In both cases, keeping data in memory can improve performance by an order of magnitude. To achieve fault tolerance efficiently, RDDs provide a restricted form of shared memory, based on coarse-grained transformations rather than fine-grained updates to shared state. However, we show that RDDs are expressive enough to capture a wide class of computations, including recent specialized programming models for iterative jobs, such as Pregel, and new applications that these models do not capture. We have implemented RDDs in a system called Spark, which we evaluate through a variety of user applications and benchmarks.

Jettison: Efficient Idle Desktop Consolidation with Partial VM Migration

Nilton Bila, Eyal de Lara, Kaustubh Joshi, H. Andres Lagar-Cavilla, Matti Hiltunen, M. Satyanarayanan

EuroSys'12. Bern, Switzerland, April 2012.

Idle desktop systems are frequently left powered, often because of applications that maintain network presence or to enable potential remote access. Unfortunately, an idle PC consumes up to 60% of its peak power. Solutions

continued on pg. 16

Recent Publications

continued from pg. 15

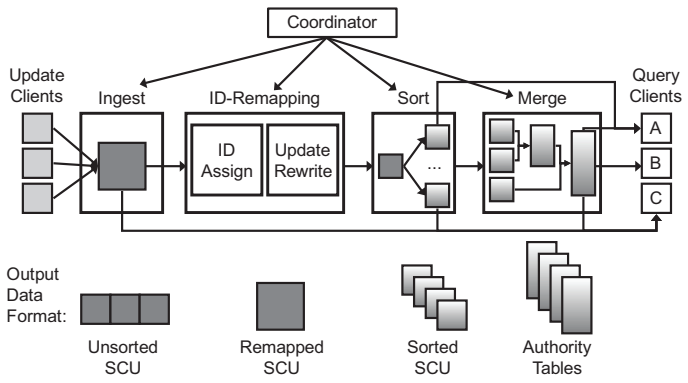
have been proposed that perform consolidation of idle desktop virtual machines. However, desktop VMs are often large requiring gigabytes of memory. Consolidating such VMs, creates bulk network transfers lasting in the order of minutes, and utilizes server memory inefficiently. When multiple VMs migrate simultaneously, each VM's experienced migration latency grows, and this limits the use of VM consolidation to environments in which only a few daily migrations are expected for each VM. This paper introduces Partial VM Migration, a technique that transparently migrates only the working set of an idle VM. Jettison, our partial VM migration prototype, can deliver 85% to 104% of the energy savings of full VM migration, while using less than 10% as much network resources, and providing migration latencies that are two to three orders of magnitude smaller.

LazyBase: Trading Freshness for Performance in a Scalable Database

James Cipar, Greg Ganger, Kimberly Keeton, Charles B. Morrey III, Craig A. N. Soules, Alistair Veitch

EuroSys'12. Bern, Switzerland, April 2012.

The LazyBase scalable database system is specialized for the growing class of data analysis applications that extract knowledge from large, rapidly changing data sets. It provides the scalability of popular NoSQL systems without the query-time complexity associated with their eventual consistency models, offering a clear consistency model and explicit per-query control over the trade-off between latency and result freshness. With an architecture designed around batching and pipelining of updates, LazyBase simultaneously ingests atomic batches of updates at a very high throughput and offers quick read queries to a stale-but-consistent version of the data. Although slightly stale results are sufficient for many analysis queries, fully up-to-date results can be obtained when necessary by also scanning updates still in the pipeline. Compared to the Cassandra NoSQL system, Lazy- Base provides 4X–5X faster update throughput and 4X faster read query throughput for range queries while remaining competitive for point queries. We demonstrate LazyBase's tradeoff between query latency and result freshness as well as the benefits of its consistency model. We also demonstrate specific cases where Cassandra's consistency model is weaker than LazyBase's.



Lazybase pipeline.

Reoptimizing Data Parallel Computing

Sameer Agarwal, Srikanth Kandula, Nico Bruno, Ming-Chuan Wu, Ion Stoica, Jingren Zhou

9th USENIX Symposium on Networked Systems Design and Implementation (NSDI), San Jose, CA, April 2012.

Performant execution of data-parallel jobs needs good execution plans. Certain properties of the code, the data, and the interaction between them are crucial to generate these plans. Yet, these properties are difficult to estimate due to the highly distributed nature of these frameworks, the freedom that allows users to specify arbitrary code as operations on the data, and since jobs in modern clusters have evolved beyond single map and reduce phases to logical graphs of operations. Using fixed a priori estimates of these properties to choose execution plans, as modern systems do, leads to poor performance in several instances. We present RoPE, a first step towards re-optimizing data-parallel jobs. RoPE collects certain code and data properties by piggybacking on job execution. It adapts execution plans by feeding these properties to a query optimizer. We show how this improves the future invocations of the same (and similar) jobs and characterize the scenarios of benefit. Experiments on Bing's production clusters show up to 2.0× improvement across response time for production jobs at the 75th percentile while using 1.5× fewer resources.

Enabling Efficient and Scalable Hybrid Memories Using Fine-Granularity DRAM Cache Management

Justin Meza, Jichuan Chang, HanBin Yoon, Onur Mutlu, Parthasarathy Ranganathan

IEEE Computer Architecture Letters (CAL), March 2012.

Hybrid main memories composed of DRAM as a cache to scalable non-volatile memories such as phase-change memory (PCM) can provide much larger storage capacity than traditional main memories. A key challenge for enabling high-performance and scalable hybrid memories, though, is efficiently managing the metadata (e.g., tags) for data cached in DRAM at a fine granularity. Based on the observation that storing metadata off-chip in the same row as their data exploits DRAM row buffer locality, this paper reduces the overhead of fine-granularity DRAM caches by only caching the metadata for recently accessed rows on-chip using a small buffer. Leveraging the flexibility and efficiency of such a fine-granularity DRAM cache, we also develop an adaptive policy to choose the best granularity when migrating data into DRAM. On a hybrid memory with a 512MB DRAM cache, our proposal using an 8KB on-chip buffer can achieve within 6% of the performance of, and 18% better energy efficiency than, a conventional 8MB SRAM metadata store, even when the energy

overhead due to large SRAM metadata storage is not considered.

Towards Understanding Heterogeneous Clouds at Scale: Google Trace Analysis

Charles Reiss, Alexey Tumanov, Gregory R. Ganger, Randy H. Katz, Michael A. Kozuch

Intel Science and Technology Center for Cloud Computing Technical Report ISTC-CC-TR-12-101, April 2012.

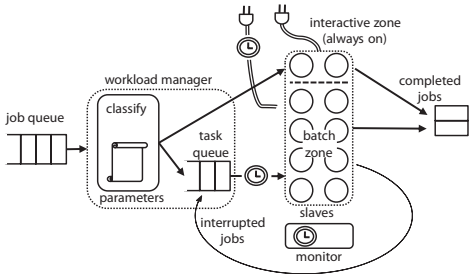
With the emergence of large, heterogeneous, shared computing clusters, their efficient use by mixed distributed workloads and tenants remains an important challenge. Unfortunately, little data has been available about such workloads and clusters. This paper analyzes a recent Google release of scheduler request and utilization data across a large (12500+) general-purpose compute cluster over 29 days. We characterize cluster resource requests, their distribution, and the actual resource utilization. Unlike previous scheduler traces we are aware of, this one includes diverse workloads – from large web services to large CPU-intensive batch programs – and permits comparison of actual resource utilization with the user-supplied resource estimates available to the cluster resource scheduler. We observe some under-utilization despite over-commitment of resources, difficulty of scheduling high-priority tasks that specify constraints, and lack of dynamic adjustments to user allocation requests despite the apparent availability of this feature in the scheduler.

Energy Efficiency for Large-Scale MapReduce Workloads with Significant Interactive Analysis

Y. Chen, S. Alspaugh, D. Borthakur, R. H. Katz

EuroSys'12. Bern, Switzerland, April 2012.

MapReduce workloads have evolved to include increasing amounts of time-sensitive, interactive data analysis; we refer to such workloads as MapReduce with Interactive Analysis (MIA). Such workloads run on large clusters, whose size and cost make energy efficiency a critical concern. Prior works on MapReduce energy efficiency have not yet considered this workload class. Increasing hardware utilization helps improve efficiency, but is challenging to achieve for MIA workloads. These concerns lead us to develop BEEMR (Berkeley Energy Efficient MapReduce), an energy efficient MapReduce workload manager motivated by empirical analysis of real-life MIA traces at Facebook. The key insight is that although MIA clusters host huge data volumes, the interactive jobs operate on a small fraction of the data, and thus can be served by a small pool of dedicated machines; the less time-sensitive jobs can run on the rest of the cluster in a batch fashion. BEEMR achieves 40-50% energy savings under tight design constraints, and represents a first step towards improving energy efficiency for an increasingly important class of datacenter workloads.



The BEEMR workload manager (i.e., job tracker) classifies each job into one of three classes which determines which cluster zone will service the job. Interactive jobs are serviced in the interactive zone, while batchable and interruptible jobs are serviced in the batch zone. Energy savings come from aggregating jobs in the batch zone to achieve high utilization, executing them in regular batches, and then transitioning machines in the batch zone to a low-power state when the batch completes.

Bottleneck Identification and Scheduling in Multithreaded Applications

José A. Joao, M. Aater Suleman, Onur Mutlu, Yale N. Patt

ASPLOS'12 March 2012, London, UK.

Performance of multithreaded applications is limited by a variety of bottlenecks, e.g. critical sections, barriers and slow pipeline stages. These bottlenecks serialize execution, waste valuable execution cycles, and limit scalability of applications. This paper proposes Bottleneck Identification and Scheduling (BIS), a cooperative software-hardware mechanism to identify and accelerate the most critical bottlenecks. BIS identifies which bottlenecks are likely to reduce performance by measuring the number of cycles threads have to wait for each bottleneck, and accelerates those bottlenecks using one or more fast cores on an Asymmetric Chip Multi-Processor (ACMP). Unlike previous work that targets specific bottlenecks, BIS can identify and accelerate bottlenecks regardless of their type. We compare BIS to four previous approaches and show that it outperforms the best of them by 15% on average. BIS' performance improvement increases as the number of cores and the number of fast cores in the system increase.

Region Scheduling: Efficiently Using the Cache Architectures via Page-level Affinity

Min Lee, Karsten Schwan

ASPLOS'12, March 2012, London, England, UK.

The performance of modern many-core platforms strongly depends on the effectiveness of using their complex cache and memory structures. This indicates the need for a memory-centric approach to platform scheduling, in which it is the locations of memory blocks in caches rather than CPU idleness that determines where application processes are run. Using the term 'memory region' to denote the current set of physical memory pages actively used by an application, this paper presents and evaluates region-based scheduling methods for multicore platforms. This involves (i) continuously and at runtime identifying the memory regions used by executable entities, and their sizes, (ii) mapping these regions to caches to match performance goals, and (iii) maintaining region to cache mappings by ensuring that en-

continued on pg. 18

Recent Publications

continued from pg. 17

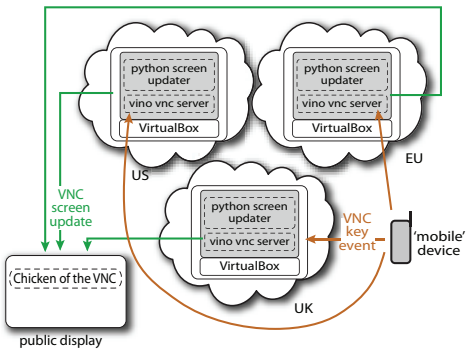
tities run on processors with direct access to the caches containing their regions. Region scheduling can implement policies that (i) offer improved performance to applications by 'unifying' the multiple caches present on the underlying physical machine and/or by 'balancing' cache usage to take maximum advantage of available cache space, (ii) better isolate applications from each other, particularly when their performance is strongly affected by cache availability, and also (iii) take advantage of standard scheduling and CPU-based load balancing when regioning is ineffective. The paper describes region scheduling and its system-level implementation and evaluates its performance with micro-benchmarks and representative multi-core applications. Single applications see performance improvements of up to 15% with region scheduling, and we observe 40% latency improvements when a platform is shared by multiple applications. Superior isolation is shown to be particularly important for cache-sensitive or real-time codes.

How Close is Close Enough? Understanding the Role of Cloudlets in Supporting Display Appropriation by Mobile Users

Sarah Clinch, Jan Harkes, Adrian Friday, Nigel Davies, Mahadev Satyanarayanan

2012 IEEE International Conference on Pervasive Computing and Communications (PerCom 2012). Lugano, Switzerland, March 2012.

Transient use of displays by mobile users was prophesied two decades ago. Today, convergence of a range of technologies enable the realization of this vision. For researchers in this space, one key question is where to physically locate the application for which the display has been appropriated. The emergence of cloud and cloudlet computing has increased the range of possible locations. In this paper we focus on understanding the extent to which application location impacts user experience when appropriating displays. We describe a usage model in which public displays can be appropriated to support spontaneous use of interactive applications, present an example architecture based on cloudlets, and explore how application location impacts user experience.



Benchmarking architecture.

Kineograph: Taking the Pulse of a Fast-Changing and Connected World

Raymond Cheng, Ji Hong , Aapo Kyrola, Youshan Miao, Xuetian Weng, Ming Wu, Fan Yang, Lidong Zhou, Feng Zhao, Enhong Chen

EuroSys'12. Bern, Switzerland, April 2012.

Kineograph is a distributed system that takes a stream of incoming data to construct a continuously changing graph, which captures the relationships that exist in the data feed. As a computing platform, Kineograph further supports graph-mining algorithms to extract timely insights from the fast-changing graph structure. To accommodate graph- mining algorithms that assume a static underlying graph, Kineograph creates a series of consistent snapshots, using a novel and efficient epoch commit protocol. To keep up with continuous updates on the graph, Kineograph includes an incremental graph-computation engine. We have developed three applications on top of Kineograph to analyze Twitter data: user ranking, approximate shortest paths, and controversial topic detection. For these applications, Kineograph takes a live Twitter data feed and maintains a graph of edges between all users and hashtags. Our evaluation shows that with 40 machines processing 100K tweets per second, Kineograph is able to continuously compute global properties, such as user ranks, with less than 2.5-minute timeliness guarantees. This rate of traffic is more than 10 times the reported peak rate of Twitter as of October 2011.

Stargazer: Automated Regression-Based GPU Design Space Exploration

Wenhao Jia, Kelly A. Shaw, Margaret Martonosi

Proceedings of 2012 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'12), New Brunswick, NJ, April 2012.

Graphics processing units (GPUs) are of increasing interest because they offer massive parallelism for high-throughput computing. While GPUs promise high peak performance, their challenge is a less-familiar programming model with more complex and irregular performance trade-offs than traditional CPUs or CMPs. In particular, modest changes in software or hardware characteristics can lead to large or unpredictable changes in performance. In response to these challenges, our work proposes, evaluates, and offers usage examples of Stargazer, an automated GPU performance exploration frame-work based on stepwise regression modeling. Stargazer sparsely and randomly samples parameter values from a full GPU design space and simulates these designs. Then, our automated stepwise algorithm uses these sampled simulations to build a performance estimator that identifies the most significant architectural parameters and their interactions. The result is an application-specific performance model which can accurately predict program runtime for any point in the design space. Because very few initial performance samples are required relative to the extremely large design space, our method can drastically reduce simulation time in GPU studies. For example, we used Stargazer to explore a design space of nearly 1 million possibilities by sampling only 300 designs. For 11 GPU applications, we were able to estimate their runtime with less than 1.1% aver-

age error. In addition, we demonstrate several usage scenarios of Stargazer.

Benchmarking Next Generation Hardware Platforms: An Experimental Approach

Vishakha Gupta, Adit Ranadive, Ada Gavrilovska, Karsten Schwan

3rd Workshop on SoCs, Heterogeneous Architectures and Workloads (SHAW-3), held in conjunction with HPCA-18. February 2012, New Orleans, Louisiana, USA.

Heterogeneous multi-cores–platforms comprised of both general purpose and accelerator cores—are becoming increasingly common. Further, with processor designs in which there are many cores on a chip, a recent trend is to include functional and performance asymmetries to balance their power usage vs. performance requirements. Coupled with this trend in CPUs is the development of high end interconnects providing low latency and high throughput communication. Understanding the utility of such next generation platforms for future datacenter workloads requires investigations that evaluate the combined effects on workload of (1) processing units, (2) interconnect, and (3) usage models. For benchmarks, then, this requires functionality that makes it possible to easily yet separately vary different benchmark attributes that affect the performance observed for application-relevant metrics like throughput, end-to-end latency, and the effects on both due to the presence of other concurrently running applications. To obtain these properties, benchmarks must be designed to test different and varying, rather than fixed, combinations of factors pertaining to their processing and communication behavior and their respective usage patterns (e.g., degree of burstiness). The 'Nectere' benchmarking framework is intended for understanding and evaluating next generation multicore platforms under varying workload conditions. This paper demonstrates two specific benchmarks constructed with Nectere: (1) a financial benchmark posing low-latency challenges, and (2) an image processing benchmark with high throughput expectations. Benchmark characteristics can be varied along dimensions that include their relative usage of heterogeneous processors, like CPUs vs. graphics processors (GPUs), and their use of the interconnect through variations in data sizes and communication rates. With Nectere, one can create a mix of workloads to study the effects of consolidation, and one can create both single- and multi-node versions of these benchmarks. Results presented in the paper evaluate workload ability or inability to share resources like GPUs or network interconnects, and the effects of such sharing on applications running in consolidated systems.

Lynx: A Dynamic Instrumentation System for Data-Parallel Applications on GPGPU Architectures

Naila Farooqui, Andrew Kerr, Greg Eisenhauer, Karsten Schwan, Sudhakar Yalamanchili

2012 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'12), New Brunswick, NJ, April 2012.

As parallel execution platforms continue to proliferate, there is a growing need for real-time introspection tools to provide insight into platform behavior for performance debugging, correctness checks, and to drive effective resource management schemes. To address this need, we present the Lynx dynamic instrumentation system. Lynx provides the capability to write instrumentation routines that are (1) selective, instrumenting only what is needed, (2) transparent, without changes to the applications' source code, (3) customizable, and (4) efficient. Lynx is embedded into the broader GPU Ocelot system, which provides run-time code generation of CUDA programs for heterogeneous architectures. This paper describes (1) the Lynx framework and implementation, (2) its language constructs geared to the Single Instruction Multiple Data (SIMD) model of data-parallel programming used in current general-purpose GPU (GPGPU) based systems, and (3) useful performance metrics described via Lynx's instrumentation language that provide insights into the design of effective instrumentation routines for GPGPU systems. The paper concludes with a comparative analysis of Lynx with existing GPU profiling tools and a quantitative assessment of Lynx's instrumentation performance, providing insights into optimization opportunities for running instrumented GPU kernels.

Reducing Memory Interference in Multicore Systems via Application-Aware Memory Channel Partitioning

Sai Prashanth Muralidhara, Lavanya Subramanian, Onur Mutlu, Mahmut Kandemir, Thomas Moscibroda

The 44th International Symposium on Microarchitecture, Porto Alegre, Brazil, December 2011.

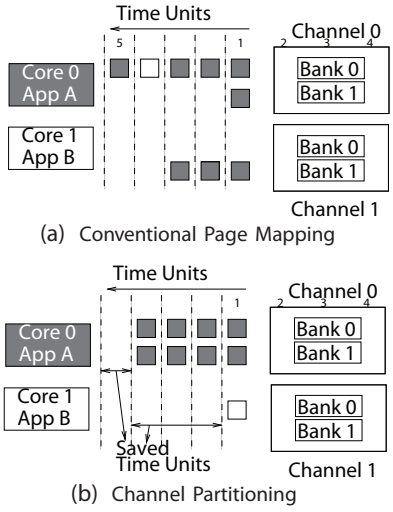
Main memory is a major shared resource among cores in a multicore system. If the interference between different applications' memory requests is not controlled effectively, system performance can degrade significantly. Previous work aimed to mitigate the problem of interference between applications by changing the scheduling policy in the memory controller, i.e., by prioritizing memory requests from applications in a way that benefits system performance.

In this paper, we first present an alternative approach to reducing inter-application interference in the memory system: application-aware memory channel partitioning (MCP). The idea is to map the data of applications that are likely to severely interfere with each other to different memory channels. The key principles are to partition onto separate channels 1) the data of light (memory non-intensive) and heavy (memory-intensive) applications, 2) the data of applications with low and high row-buffer locality. Second, we observe that interference can be further reduced with a combination of memory channel partitioning and scheduling, which we call integrated memory partitioning and scheduling (IMPS). The key idea is to 1) always prioritize very light

continued on pg. 20

Recent Publications

continued from pg. 19



Conceptual example showing benefits of mapping data of low and high memory-intensity applications to separate channels.

applications in the memory scheduler since such applications cause negligible interference to others, 2) use MCP to reduce interference among the remaining applications.

We evaluate MCP and IMPS on a variety of multiprogrammed workloads and system configurations and compare them to four previously proposed state-of-the-art memory scheduling policies. Averaged over 240 workloads on a 24-core system with 4 memory channels, MCP improves system throughput by 7.1% over an application-unaware memory scheduler and 1% over the previous best scheduler, while avoiding modifications to existing memory schedulers. IMPS improves system throughput by 11.1% over an application-unaware scheduler and 5% over the previous best scheduler, while incurring much lower hardware complexity than the latter.

Improving GPU Performance via Large Warps and Two-Level Warp Scheduling

Veynu Narasiman, Chang Joo Lee, Michael Shebanow, Rustam Miftakhutdinov, Onur Mutlu, Yale Patt

The 44th Internat'l Symp. on Microarchitecture, Porto Alegre, Brazil, December 2011.

Due to their massive computational power, graphics processing units (GPUs) have become a popular platform for executing general purpose parallel applications. GPU programming models allow the programmer to create thousands of threads, each executing the same computing kernel. GPUs exploit this parallelism in two ways. First, threads are grouped into fixed-size SIMD batches known as warps, and second, many such warps are concurrently executed on a single GPU core. Despite these techniques, the computational resources on GPU cores are still underutilized, resulting in performance far short of what could be delivered. Two reasons for this are conditional branch instructions and stalls due to long latency operations.

To improve GPU performance, computational resources must be more effectively utilized. To accomplish this, we propose two independent ideas: the large warp microarchitecture and two-level warp scheduling. We show that when combined, our mechanisms improve performance by 19.1% over traditional GPU cores for a wide variety of general purpose parallel applications that heretofore have not been able to fully exploit the available resources of the GPU chip.

SHiP: Signature-Based Hit Predictor for High Performance Caching

Carole-Jean Wu, Aamer Jaleel, William Hasenplaugh, Margaret Martonosi, Simon Steely Jr., Joel Emer

The 44th Annual IEEE/ACM International Symposium on Microarchitecture, December 2011.

The shared last-level caches in CMPs play an important role in improving application performance and reducing off-chip memory bandwidth requirements. In order to use LLCs more efficiently, recent research has shown that changing the re-reference prediction on cache insertions and cache hits can significantly improve cache performance. A fundamental challenge, however, is how to best predict the re-reference pattern of an incoming cache line.

This paper shows that cache performance can be improved by correlating the re-reference behavior of a cache line with a unique signature. We investigate the use of memory region, program counter, and instruction sequence history based signatures. We also propose a novel Signature-based Hit Predictor (SHiP) to learn the re-reference behavior of cache lines belonging to each signature.

Overall, we find that SHiP offers substantial improvements over the baseline LRU replacement and state-of-the-art replacement policy proposals. On average, SHiP improves sequential and multiprogrammed application performance by roughly 10% and 12% over LRU replacement, respectively. Compared to recent replacement policy proposals such as Seg-LRU and SDBP, SHiP nearly doubles the performance gains while requiring less hardware overhead.

PACMan: Prefetch-Aware Cache Management for High Performance Caching

Carole-Jean Wu, Aamer Jaleel, Margaret Martonosi, Simon Steely Jr., Joel Emer

The 44th Annual IEEE/ACM International Symposium on Microarchitecture, December 2011.

Hardware prefetching and last-level cache (LLC) management are two independent mechanisms to mitigate the growing latency to memory. However, the interaction between LLC management and hardware prefetching has received very little attention. This paper characterizes the performance of state-of-the-art LLC management policies in the presence and absence of hardware prefetching. Although prefetching improves performance by fetching useful data in advance, it can interact with LLC management policies to introduce application performance variability. This variability stems from the fact that current replacement policies treat prefetch and demand requests identically.

In order to provide better and more predictable performance, we propose Prefetch-Aware Cache Management

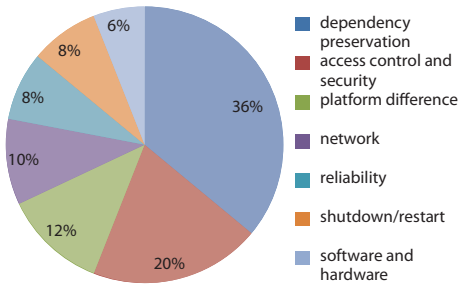
(PACMan). PACMan dynamically estimates and mitigates the degree of prefetch-induced cache interference by modifying the cache insertion and hit promotion policies to treat demand and prefetch requests differently. Across a variety of emerging workloads, we show that PACMan eliminates the performance variability in state-of-the-art replacement policies under the influence of prefetching. In fact, PACMan improves performance consistently across multimedia, games, server, and SPEC CPU2006 workloads by an average of 21.9% over the baseline LRU policy. For multiprogrammed workloads, on a 4-core CMP, PACMan improves performance by 21.5% on average.

Why Do Migrations Fail and What Can We Do about It

Gong Zhang,Ling Liu

Usenix 25th Large Installation System Administration Conference, Boston, MA, December 2011.

This paper investigates the main causes that make the application migration to Cloud complicated and error-prone through two case studies. We first discuss the typical configuration errors in each migration case study based on our error categorization model, which classifies the configuration errors into seven categories. Then we describe the common installation errors across both case studies. By analyzing operator errors in our case studies for migrating applications to cloud, we present the design of CloudMig, a semi-automated migration validation system with two unique characteristics. First, we develop a continual query (CQ) based configuration policy checking system, which facilitate operators to weave important configuration constraints into CQ-based policies and periodically run these policies to monitor the configuration changes and detect and alert the possible configuration constraints violations. Second, CloudMig combines the CQ based policy checking with the template based installation automation to help operators reduce the installation errors and increase the correctness assurance of application migration. Our experiments show that CloudMig can effectively detect a majority of the configuration errors in the migration process.



Overall migration error distribution. The legend lists the error types in the decreasing frequency order.

SIMD Re-convergence at Thread Frontiers

Gregory Diamos, Andrew Kerr, Haicheng Wu, and Sudhakar Yalamanchili, Benjamin Ashbaugh, Subramaniam Maiyuran

The 44th Annual IEEE/ACM International Symposium on Microarchitecture, December 2011.

Hardware and compiler techniques for mapping data-parallel programs with divergent control flow to SIMD architectures have recently enabled the emergence of new GPGPU programming models such as CUDA, OpenCL, and DirectX Compute. The impact of branch divergence can be quite different depending upon whether the program's control flow is structured or unstructured. In this paper, we show that unstructured control flow occurs frequently in applications and can lead to significant code expansion when executed using existing approaches for handling branch divergence.

This paper proposes a new technique for automatically mapping arbitrary control flow onto SIMD processors that relies on a concept of a Thread Frontier, which is a bounded region of the program containing all threads that have branched away from the current warp. This technique is evaluated on a GPU emulator configured to model i) a commodity GPU (Intel Sandybridge), and ii) custom hardware support not realized in current GPU architectures. It is shown that this new technique performs identically to the best existing method for structured control flow, and re-converges at the earliest possible point when executing unstructured control flow. This leads to i) between 1:5 – 633:2% reductions in dynamic instruction counts for several real applications, ii) simplification of the compilation process, and iii) ability to efficiently add high level unstructured programming constructs (e.g., exceptions) to existing data-parallel languages.

Purlieus: Locality-aware Resource Allocation for MapReduce in a Cloud

Balaji Palanisamy, Aameek Singh, Ling Liu, Bhushan Jain

ACM/IEEE International Conference on SuperComputing (SC2011), Seattle WA, November 2011.

The large amount of energy consumed by Internet services represents significant and fast-growing financial and environmental costs. Increasingly, services are exploring dynamic methods to minimize energy costs while respecting their service-level agreements (SLAs). Furthermore, it will soon be important for these services to manage their usage of "brown energy" (produced via carbon-intensive means) relative to renewable or "green" energy. This paper introduces a general, optimization-based framework for enabling multi-data-center services to manage their brown energy consumption and leverage green energy, while respecting their SLAs and minimizing energy costs. Based on the framework, we propose a policy for request distribution across the data centers. Our policy can be used to abide by caps on brown energy consumption, such as those that might arise from Kyoto-style carbon limits, from corporate pledges on carbon-neutrality, or from limits imposed on services to encourage brown energy conservation. We evaluate our framework and policy extensively through simulations and real experiments. Our results show how our policy allows a service to trade off consumption and cost. For example, using our policy, the service can reduce

continued on pg. 22

Recent Publications

continued from pg. 21

brown energy consumption by 24% for only a 10% increase in cost, while still abiding by SLAs.

On the Duality of Data-intensive File System Design: Reconciling HDFS and PVFS

Wittawat Tantisiriroj, Swapnil Patil, Garth Gibson, Seung Woo Son, Samuel J. Lang, Robert B. Ross

SC11, November 2011, Seattle, Washington USA.

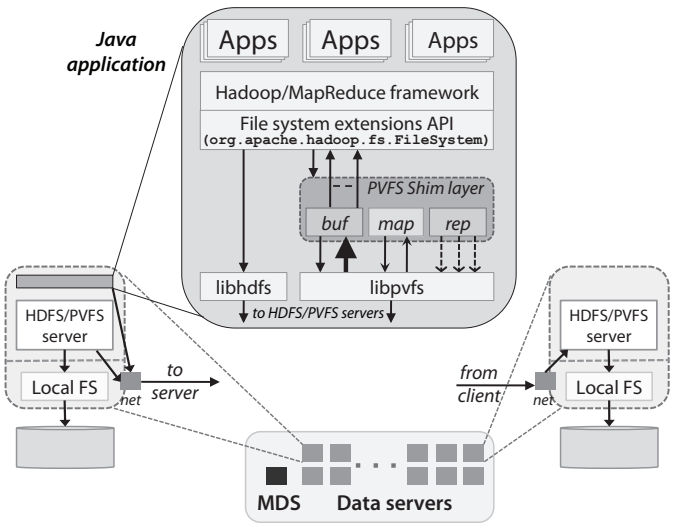
Data-intensive applications fall into two computing styles: Internet services (cloud computing) or high-performance computing (HPC). In both categories, the underlying file system is a key component for scalable application performance. In this paper, we explore the similarities and differences between PVFS, a parallel file system used in HPC at large scale, and HDFS, the primary storage system used in cloud computing with Hadoop. We integrate PVFS into Hadoop and compare its performance to HDFS using a set of data-intensive computing benchmarks. We study how HDFS-specific optimizations can be matched using PVFS and how consistency, durability, and persistence tradeoffs made by these file systems affect application performance. We show how to embed multiple replicas into a PVFS file, including a mapping with a complete copy local to the writing client, to emulate HDFS's file layout policies. We also highlight implementation issues with HDFS's dependence on disk bandwidth and benefits from pipelined replication.

YCSB++: Benchmarking and Performance Debugging Advanced Features in Scalable Table Stores

Swapnil Patil, Milo Polte, Kai Ren, Wittawat Tantisiriroj, Lin Xiao, Julio Lopez, Garth Gibson, Adam Fuchs, Billie Rinaldi

Proc. of the 2nd ACM Symposium on Cloud Computing (SOCC '11), October 2011, Cascais, Portugal.

Inspired by Google's BigTable, a variety of scalable, semistructured, weak-semantic table stores have been developed and optimized for different priorities such as query speed, ingest speed, availability, and interactivity. As these systems mature, performance benchmarking will advance from measuring the rate of simple workloads to understanding and debugging the performance of advanced features such as ingest speed-up techniques and function shipping filters from client to servers. This paper describes YCSB++, a set of extensions to the Yahoo! Cloud Serving Benchmark (YCSB) to improve performance understanding and debugging of these advanced features. YCSB++ includes multi-tester coordination for increased load and eventual consistency measurement, multi-phase workloads to quantify the consequences of work deferment and the benefits of anticipatory configuration optimization such as B-tree pre-splitting or bulk loading, and abstract APIs for explicit incorporation of advanced features in benchmark tests. To enhance performance debugging, we customized an existing cluster monitoring tool to gather the internal statistics of YCSB++, table stores, system services like HDFS, and operating systems, and to offer easy post-test correlation and reporting of performance behaviors. YCSB++ features are illustrated in case studies of two BigTable-like table stores, Apache HBase and Accumulo, developed to emphasize high ingest rates and fine-grained security.



Hadoop-PVFS Shim Layer: The shim layer allows Hadoop to use PVFS in place of HDFS. This layer has three responsibilities: to perform readahead buering ('buf' module), to expose data layout mapping to Hadoop (`map' module) and to emulate replication ('rep' module).

Small Cache, Big Effect: Provable Load Balancing for Randomly Partitioned Cluster Services

Bin Fan, Hyeontaek Lim, David G. Andersen, Michael Kaminsky

Proceedings of ACM Symposium on Cloud Computing (SOCC'11), Cascais, Portugal, October 2011.

Load balancing requests across a cluster of back-end servers is critical for avoiding performance bottlenecks and meeting service-level objectives (SLOs) in large-scale cloud computing services. This paper shows how a small, fast popularity-based front-end cache can ensure load balancing for an important class of such services; furthermore, we prove an O(n log n) lower-bound on the necessary cache size and show that this size depends only on the total number of back-end nodes n, not the number of items stored in the system. We validate our analysis through simulation and empirical results running a key-value storage system on an 85-node cluster.

Collaborating with Executable Content Across Space and Time

Mahadev Satyanarayanan, Vasanth Bala, Gloriana St. Clair, Erika Linke

7th International Conference on Collaborative Computing, Orlando, Florida, USA, October 2011.

Executable content is of growing importance in many domains. How does one share and archive such content at

Internet-scale for spatial and temporal collaboration? Spatial collaboration refers to the classic concept of user collaboration: two or more users who are at different Internet locations performing a task using shared context. Temporal collaboration refers to the archiving of context by one user and use of that context by another user, possibly many years or decades later. The term "shared context" has typically meant shared documents or a shared workspace such as a whiteboard. However, executable content forces us to think differently. Just specifying a standardized data format is not sufficient; one has to accurately reproduce computation. We observe that the precise encapsulation of computing state provided by a virtual machine (VM) may help us solve this problem. We can cope with large VM size through a streaming mechanism that demand fetches memory and disk state during execution. Based on our positive initial experience with VMs for archiving execution state, we propose the creation of Olive, an Internet ecosystem of curated VM image collections.

Precomputing Possible Configuration Error Diagnoses

A. Rabkin, R. H. Katz

IEEE/ACM International Conference on Automated Software Engineering, Lawrence, KS, November 2011.

Complex software packages, particularly systems software, often require substantial customization before being used. Small mistakes in configuration can lead to hard-to-diagnose error messages. We demonstrate how to build a map from each program point to the options that might cause an error at that point. This can aid users in troubleshooting these errors without any need to install or use additional tools. Our approach relies on static dataflow analysis, meaning all the analysis is done in advance. We evaluate our work in detail on two substantial systems, Hadoop and the JChord program analysis toolkit, using failure injection and also by using log messages as a source of labeled program points. When logs and stack traces are available, they can be incorporated into the analysis. This reduces the number of false positives by nearly a factor of four for Hadoop, at the cost of approximately one minute's work per unique query.

SILT: A Memory-Efficient, High-Performance Key-Value Store

Hyeontaek Lim, Bin Fan, David G. Andersen, Michael Kaminsky

Proceedings of 23nd ACM Symposium on Operating Systems Principles (SOSP'11), Cascais, Portugal, October 2011.

SILT (Small Index Large Table) is a memory-efficient, high-performance key-value store system based on flash storage that scales to serve billions of key-value items on a single node. It requires only 0.7 bytes of DRAM per entry and retrieves key/value pairs using on average 1.01 flash reads each. SILT combines new algorithmic and systems techniques to balance the use of memory, storage, and computation. Our contributions include: (1) the design of three basic key-value stores each with a different emphasis on memory-efficiency and write-friendliness; (2) synthesis of the basic key-value stores to build a SILT key-value store system; and (3) an analytical model for tuning system parameters carefully to meet the needs of different workloads. SILT requires one to two orders of magnitude less memory to provide comparable throughput to current high-performance key-value systems on a commodity desktop system with flash storage.

Don't Settle for Eventual: Stronger Consistency for Wide-Area Storage with COPS

Wyatt Lloyd, Michael J. Freedman, Michael Kaminsky, David G. Andersen

Proceedings of 23nd ACM Symp. on Operating Systems Principles (SOSP'11), Cascais, Portugal, October 2011.

Geo-replicated, distributed data stores that support complex online applications, such as social networks, must provide an "always-on" experience where operations always complete with low latency. Today's systems often sacrifice strong consistency to achieve these goals, exposing inconsistencies to their clients and necessitating complex application logic. In this paper, we identify and define a consistency model—causal consistency with convergent conflict handling, or causal+—that is the strongest achieved under these constraints.

We present the design and implementation of COPS, a key-value store that delivers this consistency model across the wide-area. A key contribution of COPS is its scalability, which can enforce causal dependencies between keys stored across an entire cluster, rather than a single server like previous systems. The central approach in COPS is tracking and explicitly checking whether causal dependencies between keys are satisfied in the local cluster before exposing writes. Further, in COPS-GT, we introduce get transactions in order to obtain a consistent view of multiple keys without locking or blocking. Our evaluation shows that COPS completes operations in less than a millisecond, provides throughput similar to previous systems when using one server per cluster, and scales well as we increase the number of servers in each cluster. It also shows that COPS-GT provides similar latency, throughput, and scaling to COPS for common workloads.

Design Implications for Enterprise Storage Systems via Multi-Dimensional Trace Analysis

Y. Chen, K. Srinivasan, G. Goodson, R. Katz

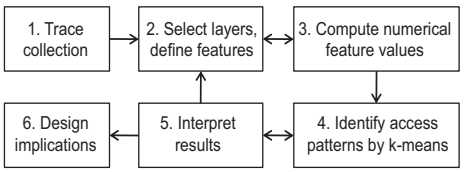
In 23nd ACM Symposium on Operating Systems Principles, Cascais, Portugal, October 2011.

Enterprise storage systems are facing enormous challenges due to increasing growth and heterogeneity of the data stored. Designing future storage systems requires comprehensive insights that existing trace analysis methods are ill-equipped to supply. In this paper, we seek to provide such insights by using a new methodology that lev-

continued on pg. 24

Recent Publications

continued from pg. 24



Methodology overview. The two-way arrows and the loop from Step 2 through Step 5 indicate our many iterations between the steps.

erages an objective, multi-dimensional statistical technique to extract data access patterns from network storage system traces. We apply our method on two large-scale real-world production network storage system traces to obtain comprehensive access patterns and design insights at user, application, file, and directory levels. We derive simple, easily implementable, threshold-based design optimizations that enable efficient data placement and capacity optimization strategies for servers, consolidation policies for clients, and improved caching performance for both.

State Monitoring in Cloud Datacenters

Shicong Meng, Ling Liu, Ting Wang

IEEE Transactions on Knowledge and Data Engineering, Special Issue on Cloud Data Management. September 2011.

Monitoring global states of a distributed cloud application is a critical functionality for cloud datacenter management. State monitoring requires meeting two demanding objectives: high level of correctness, which ensures zero or low error rate, and high communication efficiency, which demands minimal communication cost in detecting state updates. Most existing work follows an instantaneous model which triggers state alerts whenever a constraint is violated. This model may cause frequent and unnecessary alerts due to momentary value bursts and outliers. Countermeasures of such alerts may further cause problematic operations. In this paper, we present a WIndow-based StatE monitoring (WISE) framework for efficiently managing cloud applications. Window-based state monitoring reports alerts only when state violation is continuous within a time window. We show that it is not only more resilient to value bursts and outliers, but also able to save considerable communication when implemented in a distributed manner based on four technical contributions. First, we present the architectural design and deployment options for window-based state monitoring with centralized parameter tuning. Second, we develop a new distributed parameter tuning scheme enabling WISE to scale to much more monitoring nodes as each node tunes its monitoring parameters reactively without global information. Third, we introduce two optimization techniques, including their design rationale, correctness and usage model, to further reduce the communication cost. Finally, we provide an in-depth empirical study of the scalability of WISE, and evaluate the improvement brought by the distributed tuning scheme and the two performance optimizations. Our results show that WISE reduces communication by 50-90 percent compared with instantaneous monitoring approaches, and the improved WISE gains a clear scalability advantage over its centralized version.

The Case for Sleep States in Servers

Anshul Gandhi, Mor Harchol-Balter, Michael Kozuch

In 4th ACM Workshop on Power-Aware Computing and Systems, Cascais, Portugal, October 2011.

While sleep states have existed for mobile devices and workstations for some time, these sleep states have largely not been incorporated into the servers in today's data centers.

Chip designers have been unmotivated to design sleep states because data center administrators haven't expressed any desire to have them. High setup times make administrators fearful of any form of dynamic power management, whereby servers are suspended or shut down when load drops. This general reluctance has stalled research into whether there might be some feasible sleep state (with sufficiently low setup overhead and/or sufficiently low power) that would actually be beneficial in data centers.

This paper uses both experimentation and theory to investigate the regime of sleep states that should be advantageous in data centers. Implementation experiments involve a 24-server multi-tier testbed, serving a web site of the type seen in Facebook or Amazon with key-value workload and a range of hypothetical sleep states. Analytical modeling is used to understand the effect of scaling up to larger data centers. The goal of this research is to encourage data center administrators to consider dynamic power management and to spur chip designers to develop useful sleep states for servers.

ResourceExchange: Latency-Aware Scheduling in Virtualized Environments with High Performance Fabrics

Adit Ranadive, Ada Gavrilovska, Karsten Schwan

IEEE Cluster'11, Austin, TX, September 2011.

Virtualized infrastructures have seen strong acceptance in data center systems and applications, but have not yet seen adoptance for latency-sensitive codes which require I/O to arrive predictability, or response times to be generated within certain timeliness guarantees. Examples of such applications include certain classes of parallel HPC codes, server systems performing phone call or multimedia delivery, or financial services in electronic trading platforms, like ICE and CME.

In this paper, we argue that the use of high-performance, VMM-bypass capable devices can help create the virtualized infrastructures needed for the latency-sensitive applications listed above. However, to enable consolidation, problems to be solved go beyond efficient I/O virtualization, and include dealing with the shared use of I/O and compute resource, in ways that minimize or eliminate interference. Toward this end, we describe ResEx – a resource management approach for virtualized RDMA-based platforms which

incorporates concepts from supply-demand theory and congestion pricing to dynamically control the allocation of CPU and I/O resources of guest VMs. ResEx and its mechanisms and abstractions allow multiple 'pricing policies' to be deployed on these types of virtualized platforms, including such which reduce interference and enhance isolation by identifying and taxing VMs responsible for resource congestion. While the main ideas behind ResEx are more general, the design presented in this paper is specific for InfiniBand RDMA-based virtualized platforms due to the use of asynchronous monitoring needed to determine the VMs' I/O usage, and the methods to establish the trading rate for the underlying CPU and I/O resources. The latter is particularly necessary since the hypervisor's only mechanism to control I/O usage is by making appropriate adjustments in the VM's CPU resources.

The experimental evaluation of our solution uses InfiniBand platforms virtualized with the open source Xen hypervisor, and an RDMA-based latency-sensitive benchmark, BenchEx, based on a model of a financial trading platform. The results demonstrate the utility of the ResEx approach in making RDMA-based virtualized platforms more manageable and better suited for hosting even latency-sensitive workloads. ResEx can reduce the latency interference by as much as 30% in some cases as shown.

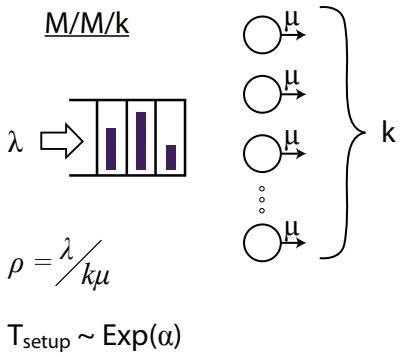
How Data Center Size Impacts the Effectiveness of Dynamic Power Management

Anshul Gandhi, Mor Harchol-Balter

49th Annual Allerton Conference on Communication, Control, and Computing, September 2011.

Power consumption accounts for a significant portion of a data center's operating expenses. Sadly, much of this power is wasted by servers that are left on even when there is no work to do.

Dynamic power management aims to reduce power wastage in data centers by turning servers off when they are not needed. However, turning a server back on requires a setup time, which can adversely affect system performance. Thus, it is not obvious whether dynamic power management should be employed in a data center.



Our M/M/k model for analyzing the effectiveness of dynamic power management in data centers.

In this paper, we analyze the effectiveness of dynamic power management in data centers under an M/M/k model via Matrix-analytic methods. We find that the effectiveness of even the simplest dynamic power management policy increases with the data center size, surpassing static power management when the number of servers exceeds 50, under realistic setup costs and server utilizations. Furthermore, we find that a small enhancement to traditional dynamic power management, involving delaying the time until a server turns off, can yield benefits over static power management even for data center sizes as small as 4 servers.

Spatially-aware Optimization of Energy Consumption in Consolidated Datacenter Systems

Hui Chen, Pramod Kumar, Mukil Kesavan, Karsten Schwan, Ada Gavrilovska, Yogendra Joshi

InterPACK2011, Portland, OR, July 2011.

Energy efficiency in data center operation depends on many factors, including power distribution, thermal load and consequent cooling costs, and IT management in terms of how and where IT load is placed and moved under changing request loads. Current methods provided by vendors consolidate IT loads onto the smallest number of machines needed to meet application requirements. This paper's goal is to gain further improvements in energy efficiency by also making such methods 'spatially aware', so that load is placed onto machines in ways that respect the efficiency of both cooling and power usage, across and within racks. To help implement spatially aware load placement, we propose a model-based reinforcement learning method to learn and then predict the thermal distribution of different placements for incoming workloads. The method is trained with actual data captured in a fully instrumented data center facility. Experimental results showing notable differences in total power consumption for representative application loads indicate the utility of a two-level spatially-aware workload management (SpAWM) technique in which (i) load is distributed across racks in ways that recognize differences in cooling efficiencies and (ii) within racks, load is distributed so as to take into account cooling effectiveness due to local air flow. The technique is being implemented using online methods that continuously monitor current power and resource usage within and across racks, sense BladeCenter-level inlet temperatures, understand and manage IT load according to an environment's thermal map. Specifically, at data center level, monitoring informs SpAWM about power usage and thermal distribution across racks. At rack-level, SpAWM workload distribution is based on power caps provided by maximum inlet temperatures determined by CRAC speeds and supply air temperature. SpAWM can be realized as a set of management methods running in VMWare's ESXServer virtualization infrastructure. Its use has the potential of attaining up to 32% improvements on the CRAC supply temperature requirement compared to non-spatially aware techniques, which can lower the inlet temperature 2~3°C, that is to say we can increase the CRAC supply temperature 2~3°C to save nearly 13% -18% cooling energy.

continued on pg. 26

Recent Publications

continued from pg. 25

Variations in Performance and Scalability when Migrating n-Tier Applications to Different Clouds

Deepal Jayasinghe, Simon Malkowski, Qingyang Wang, Jack Li, Pengcheng Xiong, Calton Pu

CLOUD 2011, July 2011. Washington, DC. [Best Paper Award].

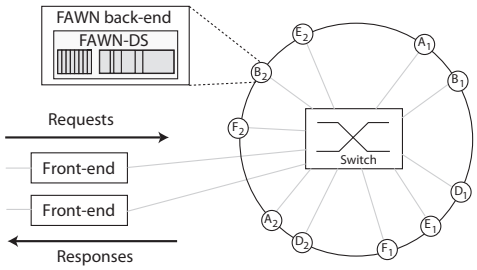
The increasing popularity of computing clouds continues to drive both industry and research to provide answers to a large variety of new and challenging questions. We aim to answer some of these questions by evaluating performance and scalability when an n-tier application is migrated from a traditional datacenter environment to an IaaS cloud. We used a representative n-tier macro-benchmark (RUBBoS) and compared its performance and scalability in three different testbeds: Amazon EC2, Open Cirrus (an open scientific research cloud), and Emulab (academic research testbed). Interestingly, we found that the best-performing configuration in Emulab can become the worst-performing configuration in EC2. Subsequently, we identified the bottleneck components, high context switch overhead and network driver processing overhead, to be at the system level. These overhead problems were confirmed at a finer granularity through micro-benchmark experiments that measure component performance directly. We describe concrete alternative approaches as practical solutions for resolving these problems.

FAWN: A Fast Array of Wimpy Nodes

David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, Vijay Vasudevan

In Communications of the ACM (CACM), Vol. 54, No. 7, pp. 101-109, July 2011.

This paper presents a fast array of wimpy nodes—FAWN—an approach for achieving low-power data-intensive datacenter computing. FAWN couples low-power processors to small amounts of local flash storage, balanc-



FAWN-KV architecture.

ing computation and I/O capabilities. FAWN optimizes for per node energy efficiency to enable efficient, massively parallel access to data. The key contributions of this paper are the principles of the FAWN approach and the design and implementation of FAWN-KV—a consistent, replicated, highly available, and high-performance key-value storage system built on a FAWN prototype. Our design centers around purely log-structured datastores that provide the basis for high performance on flash storage, as well as for replication and consistency obtained using chain replication on a consistent hashing ring. Our evaluation demonstrates that FAWN clusters can handle roughly 350 key-value queries per Joule of energy—two orders of magnitude more than a disk-based system.

GraphLab: A Distributed Framework for Machine Learning in the Cloud

Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin

arXiv:1107.0922v1 [cs.LG] 5 July 2011.

Machine Learning (ML) techniques are indispensable in a wide range of fields. Unfortunately, the exponential increase of dataset sizes are rapidly extending the runtime of sequential algorithms and threatening to slow future progress in ML. With the promise of affordable large-scale parallel computing, Cloud systems offer a viable platform to resolve the computational challenges in ML. However, designing and implementing efficient, provably correct distributed ML algorithms is often prohibitively challenging. To enable ML researchers to easily and efficiently use parallel systems, we introduced the GraphLab abstraction which is designed to represent the computational patterns in ML algorithms while permitting efficient parallel and distributed implementations.

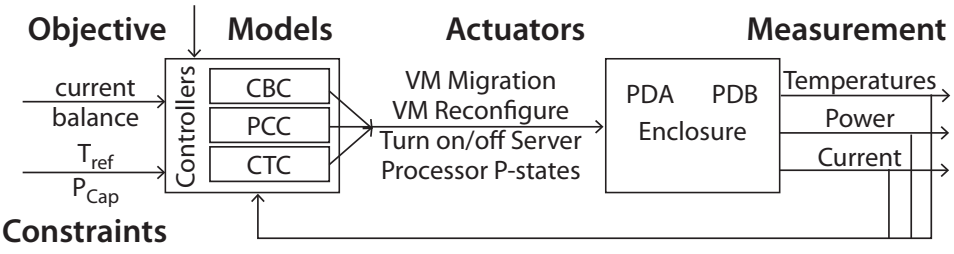
In this paper we provide a formal description of the GraphLab parallel abstraction and present an efficient distributed implementation. We conduct a comprehensive evaluation of GraphLab on three state-of-the-art ML algorithms using real large-scale data and a 64 node EC2 cluster of 512 processors. We find that GraphLab achieves orders of magnitude performance gains over Hadoop while performing comparably or superior to hand-tuned MPI implementations.

CACM: Current-aware Capacity Management in Consolidated Server Enclosures

Hui Chen, Ada Gavrilovska, Karsten Schwan, Mukil Kesavan

Second International Green Computing Conference, Work-in-Progress, Orlando, FL, July 2011.

Using virtualization to consolidate servers is a routine method for reducing power consumption in data centers. Current practice, however, assumes homogeneous servers that operate in a homogeneous physical environment. Experimental evidence collected in our mid-size, fully instrumented data center challenges those assumptions, by finding that chassis construction can significantly influence cooling power usage. In particular, the multiple power domains in a single chassis can have different levels of power efficiency, and further, power consumption is affected by the differences in electrical current levels across these two domains. This paper describes experiments designed to validate these facts, followed by a proposed current-aware capacity management system (CACM) that controls resource allocation across power domains by periodically migrating virtual machines among servers. The method



CACM control system.

not only fully accounts for the influence of current difference between the two domains, but also enforces power caps and safety levels for node temperature levels. Comparisons with industry-standard techniques that are not aware of physical constraints show that current-awareness can improve performance as well as power consumption, with about 16% in energy savings. Such savings indicate the utility of adding physical awareness to the ways in which IT systems are managed.

Recipes for Baking Black Forest Databases: Building and Querying Black Hole Merger Trees from Cosmological Simulations

J. López, C. Degraf, T. DiMatteo, B. Fu, E. Fink, G. Gibson

23rd Scientific and Statistical Database Management Conference (SSDBM'11), July 2011.

Large-scale N-body simulations play an important role in advancing our understanding of the formation and evolution of large structures in the universe. These computations require a large number of particles, in the order of 10-100 of billions, to realistically model phenomena such as the formation of galaxies. Among these particles, black holes play a dominant role on the formation of these structure. The properties of the black holes need to be assembled in merger tree histories to model the process where two or more black holes merge to form a larger one. In the past, these analyses have been carried out with custom approaches that no longer scale to the size of black hole datasets produced by current cosmological simulations. We present algorithms and strategies to store, in relational databases (RDBMS), a forest of black hole merger trees. We implemented this approach and present results with datasets containing 0.5 billion time series records belonging to over 2 million black holes. We demonstrate that this is a feasible approach to support interactive analysis and enables flexible exploration of black hole forest datasets.

Other Interesting Papers by ISTC-CC Faculty

See http://www.istc-cc.cmu.edu/publications/index.shtml

Fine-Grained Access Control of Personal Data. Ting Wang, Mudhakar Srivatsa and Ling Liu. Proceedings of the 2012 ACM Symposium on Access Control Models and Technologies (SACMAT), Newark, USA, June 2012.

NEAT: Road Network Aware Trajectory Clustering. Binh Han, Ling Liu and Edward Omiecinski, ICDCS'12, June 2012.

Human Mobility Modeling at Metropolitan Scales. S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, 10th ACM Int'l Conf. on Mobile Systems, Applications, and Services (MobiSys'12), June 2012.

Scaling Spatial Alarm Services on Road Networks. Kisung Lee, Ling Liu, Shicong Meng, Balaji Palanisamy. Proceedings of IEEE Int. Conf. on Web Services (ICWS 2012), Honolulu, Hawaii, USA, June 2012.

StockMarket Volatility Prediction: A Service-Oriented Multi-Kernel Learning Approach. Feng Wang, Ling Liu, and Chenxiao Dou, Proceedings of IEEE Int. Conf on Service Computing (SCC'12), June 2012.

Exact and Approximate Computation of a Histogram of Pairwise Distances between Astronomical Objects. Bin Fu, Eugene Fink, Garth Gibson and Jaime Carbonell, First Workshop on High Performance Computing in Astronomy (AstroHPC'12), June 2012.

Greedy Sequential Maximal Independent Set and Matching are Parallel on Average. Guy E. Blelloch, Jeremy Fineman, and Julian Shun, Proc. 24th ACM Symp. on Parallelism in Algorithms and Architectures (SPAA'12), June 2012.

Parallel and I/O Efficient Algorithms for Set Covering Problems. Guy Blelloch, Harsha Vardhan Simhadri and Kanat Tangwongsan, Proceedings of the 24th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA'12), June 2012.

Parallel Probabilistic Tree Embeddings, k-Median, and Buy-at-Bulk Network Design. Guy Blelloch, Anupam Gupta, and Kanat Tangwongsan, Proceedings of the 24th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA'12), June 2012.

A Scalable Server for 3D Metaverses. Ewen Cheslack-Postava, Tahir Azim, Behram F.T. Mistree, Daniel Reiter Horn, Jeff Terrace, Philip Levis, and Michael J. Freedman, Usenix ATC'12, Boston, MA, June 2012.

RAIDR: Retention-Aware Intelligent DRAM Refresh. Jamie Liu, Benjamin

continued on pg. 28

From L to R, Greg Ganger, CMU, Garth Gibson, CMU, Dave Andersen, CMU, Frans Kaashoek, MIT, and Ion Stoica, Berkeley, at the ISTC-CC Retreat.

Recent Publications

continued from pg. 27

Jaiyen, Richard Veras, and Onur Mutlu, Proc. of the 39th Int'l Symposium on Computer Architecture (ISCA'12), Portland, OR, June 2012.

A Case for Exploiting Subarray-Level Parallelism (SALP) in DRAM. Yoongu Kim, Vivek Seshadri, Donghyuk Lee, Jamie Liu, and Onur Mutlu, Proceedings of the 39th International Symposium on Computer Architecture (ISCA'12), Portland, OR, June 2012.

High-Confidence Near-Duplicate Image Detection. Wei Dong, Zhe Wang, and Kai Li. ACM International Conference on Multimedia Retrieval (ICMR'12), Hong Kong, June 2012.

BWS: Balanced Work Stealing for Time-Sharing Multicores. Kaibo Wang, Xiaoning Ding, Phillip B. Gibbons, and Xiaodong Zhang, EuroSys'12. Bern,Switzerland, April 2012.

A Petri Net Approach to Mediation-aided Composition of Web Services. YanHua Du, Xitong Li, Pengcheng Xiong, IEEE Transactions on Automation Science and Engineering (TASE), 9(2), April 2012.

Microscopic Social Influence. Ting Wang, Mudhakar Srivatsa, Dakshi Agrawal and Ling Liu. Proceedings of 2012 SIAM International Conference on Data Mining. April 2012. Anaheim, CA.

Error Patterns in MLC NAND Flash Memory: Measurement, Characterization, and Analysis. Yu Cai, Erich F. Haratsch, Onur Mutlu, Ken Mai. Design,

Swapnil Patil (CMU) presents his research on "Performance debugging scalable table stores" at the ISTC-CC Retreat.

Automation, and Test in Europe Conference (DATE), March 2012, Dresden, Germany.

Internally Deterministic Parallel Algorithms Can Be Fast. Guy E. Blelloch, Jeremy T. Fineman, Phillip B. Gibbons, Julian Shun. Sigplan/ACM Conference on Principles and Practices of Parallel Programming (PPoPP), February 2012, New Orleans, LA.

Characterization and Transformation of Unstructured Control Flow in Bulk Synchronous GPU Applications.H. Wu, G. Diamos, J. Wang, S. Li, and S. Yalamanchili, International Journal of High Performance Computing Applications, Vol. 26, February 2012.

Parallel Application Memory Scheduling. Eiman Ebrahimi, Rustam Miftakhutdinov, Chris Fallin, Chang Joo Lee, Onur Mutlu, and Yale N. Patt. The 44th Int'l Symposium on Microarchitecture, Porto Alegre, Brazil, December 2011.

User Preferences for Indicator And Feedback Modalities: A preliminary survey study for developing a coaching system to facilitate wheelchair power seat function usage. H-Y. Liu, G. Grindle, F-C. Chuang, A. Kelleher, R. Cooper, D. Siewiorek, A. Smailagic, R. Cooper. IEEE Pervasive Computing, Vol. 10, October-December 2011.

Prefetch-Aware Memory Controllers. Chang Joo Lee, Onur Mutlu, Veynu Narasiman, and Yale N. Patt, IEEE Transactions on Computers, 60(10), October 2011.

Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks. Hamid Hajabdolali Bazzaz, Malveeka Tewari, Guohui Wang, George Porter, T. S. Eugene Ng, David G. Andersen, Michael Kaminsky, Michael A. Kozuch, Amin Vahdat. Proc. of ACM Symposium on Cloud Computing (SOCC'11), Cascais, Portugal, October 2011.

IdleChat: Enabling High Bandwidth Real-time Applications in Residential Broadband Networks. Ramya Raghavendra, Michael Kaminsky, Konstantina Papagiannaki, Srinivasan Seshan, and Elizabeth Belding. In ACM SIGMOBILE

Mei Chen, ISTC-EC, Karsten Schwasn, Georgia Tech, and Greg Ganger, CMU, discuss cloud computing at the ISTC-CC Retreat

Mobile Computing and Communications Review (MC2R) (invited), July 2011.

MODELZ: Monitoring, Detection, and Analysis of Energy-Greedy Anomalies in Mobile Handsets. Hahnsang Kim, Kang G. Shin, and Padmanabhan Pillai, IEEE Transactions on Mobile Computing, July 2011.

Summary of Publications

(not an exhaustive list)

ASPLOS'12 - 2 papers

CLOUD'12 - 2 papers

EuroSys'12 - 5 papers

HotCloud'12 - 6 papers

ICDCS'12 - 2 papers

IGCC'12 - 2 papers

ISCA'12 - 3 papers

ISPASS'12 - 2 papers

MICRO'11 - 6 papers

NSDI'12 - 3 papers

SC'11 - 2 papers

SOCC'11 - 3 papers

SOSP'11 - 3 papers

SPAA'12 - 3 papers

Usenix ATC'12 - 2 papers

ISTC-CC News

continued from pg. 7

ded Computing, both headquartered at Carnegie Mellon University.

The ISTC for Cloud Computing forms a new cloud computing research community that broadens Intel's "Cloud 2015" vision with new ideas from top academic researchers, and includes research that extends and improves on Intel's existing cloud computing initiatives. The center combines top researchers from Carnegie Mellon University, Georgia Institute of Technology, University of California Berkeley, Princeton University, and Intel. The researchers will explore technology that will have has important future implications for the cloud, including built-in application optimization, more efficient and effective support of big data analytics on massive amounts of online data, and making the cloud more distributed and localized by extending cloud capabilities to the network edge and even to client devices.

In the future, these capabilities could enable a digital personal handler via a device wired into your glasses that sees what you see, to constantly pull data from the cloud and whisper information to you during the day -- telling you who people are, where to buy an item you just saw, or how to adjust your plans when something new comes up.

-- Intel News Room, by Connie Brown

September 20, 2011

2011 Intel PhD Fellowship Winners Announced

Congratulations to Georgia Tech student Priyanka Tembey, co-advised by Drs. Schwan and Gavrilovska, and CMU student Michelle Goodstein, advised by Dr. Mowry, who have been awarded Intel PhD Fellowships.

The fellowship program was started in the early 90s by Gordon Moore to recognize and honor top students for their leading edge research in areas that would benefit the mankind; it was open to all fields of research. Today's program keeps that focus and also places an emphasis on developing students who are well aware of issues facing the Semiconductor, High Tech/IT fields. Every winning student is assigned a technical mentor in Intel who is also a leader in their field. Students are encouraged to work through their mentor and develop a deep, understanding of the technical issues facing the industry and be on the forefront of solving the technical challenges that lie ahead.

-- from Research@Intel Blogs

Message from the PIs

continued from pg. 2

researchers produced award-winning results in the annual JouleSort competition, which measures how energy-efficiently various sorting tasks can be completed. In addition to being the most interesting ISTC-CC artifacts to look at, for most visitors, the FAWN testbeds continue to serve as canaries for sussing out efficiency challenges that will face new technologies, such as the index sizes and OS overheads that could significantly impede Flash and memory-class storage. Fortunately, by exposing them and then tackling them, ISTC-CC researchers are paving the way.

The comment about FAWN's visual appeal highlights an interesting challenge that ISTC-CC faces: cloud computing is meant to be used, not seen. As such, attention-holding demos that showcase ISTC-CC research are tough to devise. "See how much easier it is to build an app?" and "believe me, it used less energy" are tough sells for lay-folk, and even the techie struggles to get excited when shown a new programming abstraction. As such, while we had multiple demos at the UCO Showcase, we contributed in the Intel media showcase primarily by describing research, answering questions, and collaborating with ISTC-EC researchers who built atop one of our cloud computing testbeds. Such is the life of the (digital) plumber.

The primary exception to this rule arises in some "To the Edge" activities. In particular, our cloud-assisted perception explorations include such things as adaptive computation placement aimed at minimizing latency to improve the user experience. So, for example, a device with a high-speed connection to a nearby cloud data center might be best-served to exploit it by offloading computation, while one without such a connection might avoid it and do computing work locally instead. ISTC-CC researchers showcased just such a system/scenario with a visual simulation application, during the UCO showcase.

Lots of progress has been made on many other fronts, as well. As one quick example, our elastic service-sizing research is creating and proving effective policies for adaptive use of power states in servers. As another, new data storage and processing schemes are being developed to support scalable data management with consistency that fits application needs. Our cluster scheduling research is developing new approaches to supporting multiple simultaneous Big Data frameworks (e.g., Hadoop, Spark, and GraphLab, all at once) across a heterogeneous collection of specialized platforms... and, the challenges involved in doing so are sparking great discussions across collaborating ISTC-CC institutions. And... and... and...

There are too many other examples of cool first-year outcomes, but the news items and paper abstracts throughout this newsletter provide a broader overview. Of course, all of the papers can be found via the ISTC-CC website and the ISTC-CC researchers are happy to discuss their work. We hope you enjoy the newsletter, and we look forward to sharing ISTC-CC's successes in the months and years to come.

ISTC-CC Research Overview

continued from pg. 1

and little I/O bandwidth, while others are I/O-bound and involve large amounts of random I/O requests. Some are memory-limited, while others process data in streams (from storage or over the network) with little need for RAM. And, some may have characteristics that can exploit particular hardware assists, such as GPUs, encryption accelerators, and so on. A multi-purpose cloud could easily see a mix of all of these varied application types, and a lowest-common-denominator type configuration will fall far short of best-case efficiency.

We believe that specialization is crucial to achieving the best efficiency—in computer systems, as in any large-scale system (including society), specialization is fundamental to efficiency. Future cloud computing infrastructures will benefit from this concept, purposefully including mixes of different platforms specialized for different classes of applications. Instead of using a single platform configuration to serve all applications, each application (and/or application phase, and/or application component) can be run on available servers that most closely match its particular characteristics. We believe that such an approach can provide order-of-magnitude efficiency gains, where appropriate specialization is applied, while retaining the economies of scale and elastic resource allocation promised by cloud computing.

Additional platforms under consideration include lightweight nodes (such as nodes that use Intel® Atom processors), heterogeneous many-core architectures, and CPUs with integrated graphics, with varied memory, interconnect and storage configurations/technologies. Realizing this vision will require a number of inter-related research activities:

» Understanding important application classes, the trade-offs between them, and formulating specializations to optimize performance.
» Exploring the impact of new technologies like non-volatile memory (NAND flash, phase change memory, etc.).
» Creating algorithms and frameworks for exploiting such specializations.
» Programming applications so that they are adaptable to different platform characteristics, to maximize the benefits of specialization within clouds regardless of the platforms they offer.

In addition, the heterogeneity inherent to this vision will also require new automation approaches.

Pillar 2: Automation

As computer complexity has grown and system costs have shrunk, operational costs have become a significant factor in the total cost of ownership. Moreover, cloud computing raises the stakes, making the challenges tougher while simultaneously promising benefits that can only be achieved if those challenges are met. Operational costs include human administration, downtime-induced losses, and energy usage. Administration expenses arise from the broad collection of management tasks, including planning and deployment, data protection, problem diagnosis and repair, performance tuning, software upgrades, and so on. Most of these become more difficult with cloud computing, as the scale increases, the workloads run on a given infrastructure become more varied and opaque, workloads mix more (inviting interference), and pre-knowledge of user demands becomes rare rather than expected. And, of course, our introduction of specialization (Pillar 1) aims to take advantage of platforms tailored to particular workloads.

Automation is the key to driving down operational costs. With effective automation, any given IT staff can manage much larger infrastructures. Automation can also reduce losses related to downtime, both by eliminating failures induced by human error (the largest source of failures) and by reducing diagnosis and recovery times, increasing availability. Automation can significantly improve energy efficiency, both by ensuring the right (specialized) platform is used for each application, by improving server utilization, and by actively powering down hardware when it is not needed.

Within this broad pillar, ISTC-CC research will tackle key automation challenges related to efficiency, productivity and robustness, with three primary focus areas:

» Resource scheduling and task placement: devising mechanisms and policies for maximizing several goals including energy efficiency, interference avoidance, and data availability and locality. Such scheduling must accommodate diverse mixes of workloads as well as specialized computing platforms.
» Devising automated tools for software upgrade management, runtime correctness checking, and programmer productivity that are sufficiently low overhead to be used with production code at scale.
» Problem diagnosis: exploring new techniques for diagnosing problems effectively given the anticipated scale and complexity increases coming with future cloud computing.

Pillar 3: Big Data

Data-intensive scalable computing (DISC) refers to a rapidly growing style of computing characterized by its reliance on large and often dynamically growing datasets ("BigData"). With massive amounts of data arising from such diverse sources as telescope imagery, medical records, online transaction records, checkout stands and web pages, many researchers and practitioners are discovering that statistical models extracted from data collections promise major advances in science, health care, business efficiencies, and information access. In fact, in domain after domain, statistical approaches are quickly bypassing expertise-based approaches in terms of efficacy and robustness.

The shift toward DISC and Big Data analytics pervades large-scale computer usage, from the sciences (e.g., genome sequencing) to business intelligence (e.g., workflow optimization) to data warehousing (e.g., recommendation systems) to medicine (e.g., diagnosis) to Internet services (e.g., social network analysis) and so on. Based on this shift, and their resource demands relative to more traditional activities, we expect DISC and Big Data activities to eventually dominate future cloud computing.

We envision future cloud computing

infrastructures that efficiently and effectively support DISC analytics on Big Data. This requires programming and execution frameworks that provide efficiency to programmers (in terms of effort to construct and run analytics activities) and the infrastructure (in terms of resources required for given work). In addition to static data corpuses, some analytics will focus partially or entirely on live data feeds (e.g., video or social networks), involving the continuous ingest, integration, and exploitation of new observation data.

ISTC-CC research will devise new frameworks for supporting DISC analytics of Big Data in future cloud computing infrastructures. Three particular areas of focus will be:

» Understanding DISC applications, creating classifications and benchmarks to represent them, and providing support for programmers building them.
» Frameworks that more effectively accommodate the advanced machine learning algorithms and interactive processing that will characterize much of next generation DISC analytics.
» Cloud databases for huge, distributed data corpuses supporting efficient processing and adaptive use of indices. This focus includes supporting datasets that are continuously updated by live feeds, requiring efficient ingest, appropriate consistency models, and use of incremental results.

Note that these efforts each involve aspects of Automation, and that Big Data applications represent one or more classes for which Specialization is likely warranted. The aspects related to live data feeds, which often originate from client devices and social media applications, lead us into the last pillar.

Pillar 4: To the Edge

Future cloud computing will be a combination of public and private clouds, or hybrid clouds, but will also extend beyond large datacenters that power cloud computing to include billions of clients and edge devices. This includes networking components in select locations and mobile devices closely associated with their users that will be directly involved in many "cloud" activities. These devices will not only use remote cloud resources, as with today's offerings, but they will also contribute to them. Although they offer limited resources of their own, edge devices do serve as bridges to the physical world with sensors, actuators, and "context" that would not otherwise be available. Such physical-world resources and content will be among the most valuable in the cloud.

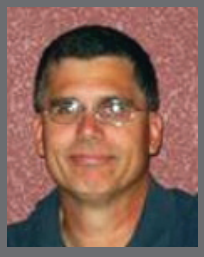
Effective cloud computing support for edge devices must actively consider location as a first-class and non-fungible property. Location becomes important in several ways. First, sensor data (e.g., video) should be understood in the context of the location (and time, etc.) at which it was captured; this is particularly relevant for applications that seek to pool sensor data from multiple edge devices at a common location. Second, many cloud applications used with edge devices will be interactive in nature, making connectivity and latency critical issues; devices do not always have good connectivity to wide-area networks and communication over long distances increases latency.

We envision future cloud computing infrastructures that adaptively and agilely distribute functionality among core cloud resources (i.e., backend data centers), edge-local cloud resources (e.g., servers in coffee shops, sports arenas, campus buildings, waiting rooms, hotel lobbies, etc.),and edge devices (e.g., mobile handhelds, tablets, netbooks, laptops, and wearables). This requires programming and execution frameworks that allow resource-intensive software components to run in any of these locations, based on location, connectivity, and resource availability. It also requires the ability to rapidly combine information captured at one or more edge devices with other such information and core resources (including data repositories) without losing critical location context.

ISTC-CC research will devise new frameworks for edge/cloud cooperation. Three focus areas will be:

» Enabling and effectively supporting applications whose execution spans client devices, edge-local cloud resources, and core cloud resources, as discussed above.
» Addressing edge connectivity issues by creating new ways to mitigate reliance on expensive and robust Internet uplinks for clients.
» Exploring edge architectures, such as resource-poor edge connection points vs. more capable edge-local servers, and platforms for supporting cloud-at-the edge applications.

Program Director's Corner



Jeff Parkhurst, Intel

It has been a fantastic first year for the Cloud Computing Center. I have seen deep engagement between many of the projects and our Intel Researchers on the CMU campus. The research work at the center has also drawn the attention of many of our researchers on Intel's main campuses, with them engaged on a variety of projects including GraphLab, some of the Hetero-core work, along with the research into Edgelets. We are always looking to expand this type of engagement and I am happy to facilitate this. If you are an ISTC funded university researcher or an Intel employee looking to engage, please contact me at jeff.parkhurst@intel.com. Here's looking forward to another successful year!

Year in Review

continued from pg. 5

Needs, Obstacles and Technologies for Storage," 7th Kavli Futures Symposium, Scalable Energy-Efficient Data Centers and Clouds, Santa Barbara, CA.

October 2011

- ISTC-CC's official launch ceremony was held on October 19.
- Randy Katz's ASE'11 paper "Precomputing possible configuration error diagnoses" received the ACM SIGSOFT Distinguished Paper Award.
- Greg Ganger presented "Overview of ISTC-CC" at the Open Cirrus summit.
- Garth Gibson and Randy Katz receive the 2012 Jean-Claude Laprie Award in Dependable Computing, Industrial/Commercial Product Impact Category, for "A Case for Redundant Arrays of Inexpensive Disks (RAID)."
- Guy Blelloch was made an ACM Fellow for his contributions to parallel computing. There are 6 other ISTC-CC faculty who are ACM Fellows: Dan Siewiorek (CMU, inducted 1994), Randy Katz (UC Berkeley, 1996), Kai Li (Princeton, 1998), M. Satyanaryanan (CMU, 2002), Phil Gibbons (Intel, 2006), and Margaret Martonosi (Princeton, 2009).
- Garth Gibson's original RAID paper from SIGMOD 1988—"A Case for Redundant Array of Inexpensive Disks" by Patterson, Gibson and Katz—was one of the four papers to be honored as a 2011 SIGOPS Hall of Fame Award paper.
- The 6th OpenCirrus summit was held at Georgia Tech, organized by ISTC-CC and including many ISTC-CC talks.
- Georgia Tech student Priyanka Tembey, co-advised by Drs. Schwan and Gavrilovska, was awarded an Intel PhD Fellowship. CMU student Michelle Goodstein, advised by Dr. Mowry, was also awarded an Intel PhD Fellowship.
- Ada Gavrilovska was awarded an NSF grant, via the EAGER program, amplifying funding for the "To the Edge" pillar.
- M. Satyanarayanan and Dan Sieworek were awarded an NSF grant amplifying funding for the "To the Edge" pillar.
- Dan Siewiorek was also awarded an NSF SBIR grant amplifying funding.
- Greg Ganger was awarded an NSF grant amplifying funding.
- 7 ISTC-CC grad students gave presentations in Cascais, Portugal at the co-located conferences SOSP '11 (Oct. 23-26) and SOCC '11 (Oct 26-28).
- M. Satyanarayanan (Satya) gave an IBM Centennial Lecture in Austin, TX. He also gave a Cray Lecture at the University of Minnesota.
- M. Satyanarayanan gave a keynote talk "Collaborating with Executable Content Across Space and Time," at the 7th International Conference on Collaborative Computing, Orlando, FL.
- Garth Gibson presented "Scalable Table Stores: Tools for Understanding Advanced Key-Value Systems for Hadoop" at the SNIA Storage Developer Conference, Santa Clara, CA.

September 2011

- The ISTC-CC officially began operation September 1, 2011.
- Karsten Schwan and Ada Gavrilovska, received a grant from Samsung Electronics, covering Sept. 2011 - Aug. 2012 for their work on "Software Platform for Heterogeneous Multicore Platforms: from End Devices to Clouds."
- Dan Siewiorek gave a keynote talk on "Virtual Coaches in Health Care: A Vision of the Future" at the National WIC Association, 2011 Technology Conference, Pittsburgh, PA.
- Phil Gibbons gave a distinguished lecture at Cornell University on the Hi-Spade project.

August 2011

- The ISTC-CC was officially announced by Intel to the media, August 3, 2011.

July 2011

- Calton Pu won best paper award in CLOUD 2011 for "Variations in Performance and Scalability when Migrating n-Tier Applications to Different Clouds."
- Mahadev (Satya) Satyanarayanan was awarded the SIGMOBILE 2010 Outstanding Contributions Award "for pioneering a wide spectrum of technologies in support of disconnected and weakly connected mobile clients" at Mobisys 2011.
- David Andersen was named a 2011 Sloan Foundation Fellow.
- Greg Ganger was awarded the Stephen J. Jatras Professorship in Electrical and Computer Engineering at Carnegie Mellon for cutting-edge work in computer systems.
- Onur Mutlu has earned the inaugural IEEE Computer Society Technical Committee on Computer Architecture's Young Computer Architect Award "in recognition of outstanding contributions in the field of computer architecture in both research and education."
- Onur Mutlu's co-authored paper was selected for the IEEE Micro special issue on the Top Picks of the 2011 Computer Architecture conferences.
- Garth Gibson presented "BigData Storage Systems: Large Datasets in Astrophysics and Cosmology" at ICiS 2011, Park City, UT.
- Dan Siewiorek was awarded the SIGMOBILE 2010 Outstanding Contributions Award "for pioneering fundamental contributions to wearable and context-aware computing" at Mobisys 2011.

Greg Ganger introduces ISTC-CC Retreat keynote speaker Rich Uhlig (intel). Rich gave a talk on "Optimizing for the Cloud: Tech Trends, Testbeds and Working Together."