Intel Science & Technology Center for Cloud Computing

ISTC-CC Update

September 2013

www.istc-cc.cmu.edu

Table of Contents

Carnegie Mellon University

Georgia Tech

intel

PRINCETON UNIVERSITY

UC Berkeley

UNIVERSITY of WASHINGTON

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 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. 32

Message from the PIs

Hello from ISTC-CC headquarters. This is our second ISTC-CC Newsletter, which provides an overview of ISTC-CC, highlights and news from the last 15 months, and abstracts of our many publications.

Greg Ganger, CMU

Phil Gibbons, Intel

Given the incredible caliber of the ISTC-CC community, it's impossible to recap everything in this introductory note, but we will touch on some particularly noteworthy items.

First, a bit more bragging about the team. The collection of researchers contributing to ISTC-CC's vision and activities is special, not only for their individual capabilities but also for their collaborative nature. ISTC-CC quickly evolved as a community, and it continues to grow as new researchers join in our efforts. There continue to be many collaborative activities across areas and across institutions, both those officially part of ISTC-CC and others. We've still not had the elusive 5-institution technical publication, but there are many 2- and 3-institution papers. We will hold our third ISTC-CC Retreat in early November 2013, and we again expect 100 participants from Intel and the universities to discuss ISTC-CC research.

Before discussing some of major research items, we want to draw attention to a recent addition to the ISTC-CC website: the ISTC-CC benchmarks page. As a service to the broader research community, as well as to increase sharing within the ISTC-CC community, we created a listing of benchmarks that we have found useful, including several new ones created and released by ISTC-CC researchers. We continue to add to it, and pointers to (or suggestions for) new ones are welcome, as we try to make it ever more useful to Intel, ISTC-CC and the broader community.

As described in the ISTC-CC overview article, we continue to describe the overall ISTC-CC agenda in terms of four inter-related "pillars"—specialization, automation, big data, to the edge—designed to enable cloud computing infrastructures that provide a strong foundation for future cloud computing. (We're guiltily proud of the pillar metaphor. ;)) But, the categorization is for agenda presentation purposes only, as the activities increasingly 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.

One area where ISTC-CC impact has been huge is something we call "big learning systems": (new) frameworks for supporting efficient Big Data analytics based on advanced machine learning (ML) algorithms. In particular, ISTC-CC's GraphLab and Spark have become very popular open source systems in addition to changing mindsets on the right way to enable ML on Big Data. In contrast to first-generation Big Data abstractions like map-reduce (e.g., as implemented in Hadoop), which are good for simple tasks like filtering or sorting large datasets, ISTC-CC is identifying more natural abstractions for different types of non-trivial ML tasks and designing frameworks to enable them. The result is both better productivity and more efficient execution...sometimes orders of magnitude more efficient! Both GraphLab and Spark now have large and active user and developer communities, regularly highlighted at events like the Hadoop

continued on pg. 35

Second Annual ISTC-CC Retreat a Success!

The ISTC-CC held its second annual retreat in Pittsburgh on November 29 & 30, 2012. The 105 attendees included 62 from CMU (18 faculty/staff & 44 students), 5 from Georgia Tech (3 faculty & 2 students), 4 from Princeton (2 Faculty & 2 students), 3 from UC Berkeley (2 faculty & 1 student), and 1 each from Washington, Penn State, and Brown, as well as 28 Intel employees. The agenda featured welcoming remarks by Wen-Hann Wang (Intel Labs, Executive Sponsor for ISTC-CC) and keynotes by Balint Fleischer of Intel, who spoke on "Clumpy Data Centers" and Das Kamhout , also of Intel (IT Cloud Lead), who talked about "Experiences from Intel IT." The agenda also featured 12 research talks by faculty and students from all four Universities, 1 research collaboration talk from Intel Labs by Ted Wilke, 5 breakout groups, and 53 posters. By all accounts, the retreat was a big success! 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. There was also a "Madness Session", hosted by Michael Kaminsky, in which key research ideas, tools/testbeds and gaps were identified. Full details on the retreat can be found on the ISTC-CC website, and the third ISTC-CC Retreat is scheduled for November 7-8, 2013.

Group photo — second annual ISTC-CC Retreat, November 2012.

ISTC-CC Personnel

Leadership

Greg Ganger, Academic PI
Phil Gibbons, Intel PI
Executive Sponsor: Rich Uhlig, Intel
Program Director: Jeff Parkhurst, Intel
Board of Advisors:
Randy Bryant, CMU
Jeff Chase, Duke
Balint Fleisher, Intel
Frans Kaashoek, MIT
Pradeep Khosla, UC San Diego
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, U. Washington
Mor Harchol-Balter, CMU
Anthony Joseph, Berkeley
Randy Katz, Berkeley
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
Alex Smola, CMU
Ion Stoica, Berkeley
Matthew Wolf, GA Tech
Sudhakar Yalamanchili, GA Tech
Eric Xing, CMU

Staff

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

Students / Post-Docs

Yoshihisa Abe, CMU
Sameer Agarwal, Berkeley
Hrishikesh Amur, GA Tech
Michael Ashley-Rollman, CMU
Rachata Ausavarungnirun, CMU
Ben Blum, CMU
Joseph Bradley, CMU
Kevin Kai-Wei Chang, CMU
Zhuo Chen, CMU
Anthony Chivetta, CMU
Jim Cipar, CMU
Henggang Cui, CMU
Chris Fallin, CMU
Bin Fan, CMU
Anshul Gandhi, CMU
Kristen Gardner, CMU
Elmer Garduno, CMU
Ali Ghodsi, Berkeley
Joseph Gonzalez, CMU
Michelle Goodstein, CMU
Samantha Gottlieb, CMU
Haijie Gu, CMU
Kiryong Ha, CMU
Qirong Ho, CMU
Liting Hu, GA Tech
Ben Jaiyen, CMU
Deeptal Jayasinghe, GA Tech
Wenhao Jia, Princeton
Lu Jiang, CMU
Tyler Johnson, Washington
Sudarsun Kannan, GA Tech
Mike Kasick, CMU
Deby Katz, CMU
Samira Khan, CMU
Jin Kyu Kim, CMU
Yoongu Kim, CMU
Andy Konwinski, Berkeley
Elie Krevat, CMU
Guatam Kumar, Berkeley
Aapo Kyrola, CMU
Mu Li, 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
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
Lavanya Subramanian, CMU
Anand Suresh, CMU
Jiaqi Tan, CMU
Wittawat Tantisiriroj, CMU
Alexey Tumanov, CMU
Vijay Vasudevan, CMU
Chengwei Wang, GA Tech
Yifan Wang, CMU
Jinliang Wei, CMU
Haicheng Wu, GA Tech
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
Jishen Zhao, CMU
Dong Zhou, 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.

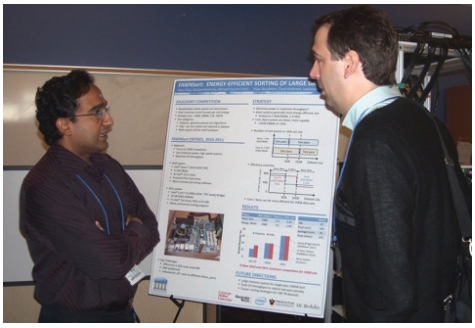
Automation: Addresses cloud's particular automation challenges, focusing on order of magnitude efficiency gains from smart resource allocation/scheduling and greatly improved problem diagnosis capabilities.

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 social media).

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 in the past 15 months.



Babu Pillai and Paolo Narvaez (both of Intel) discuss FAWNSort at and ISTC-CC Retreat poster session.

2013 Quarter 3

» Margaret Martonosi (Princeton) was named the recipient of this year's Anita Borg Institute ABIE Technical Leadership Award.

» Karsten Schwan, Matt Wolf (GA Tech), and colleagues were recipients of an R&D Magazine 2013 R&D100 award, for developing ADIOS, one of the top 100 technology products of the year.

» Ling Liu (GA Tech), Bhuvan Bamba, Kun-Lung Wu, Bugra Gedik and Ling Liu won Best Paper Award in IEEE ICWS 2013 for "SLIM: A Scalable Location-Sensitive Information Monitoring Service'', Proceedings of the 20th IEEE International Conference on Web Services (ICWS'13), June-July 2013.

» This quarter saw $2.8M in Amplifying Funding from NSF: Four new NSF grants were awarded to Dave Andersen (CMU), Michael Kaminsky (Intel), Mor Harchol-Balter (CMU), Sudha Yalamanchili (GA Tech), Guy Blelloch (CMU) and Phil Gibbons (Intel).

» Thanks to support from Intel, Garth Gibson and Greg Ganger (CMU) developed and launched a new grad course "Advanced Cloud Computing" (www.cs.cmu.edu/~15719).

» Three Georgia Tech Ph.D. students worked at Intel as summer interns in Summer 2013: Alex Merritt worked on topics in display virtualization, Sudarsun Kannan worked on topics in non-volatile memory, and Dipanjan Sengupta worked on asynchronous parallelism.

» Ling Liu (GA Tech) presented a keynote address "Towards Big Data Analytics as a Service: Exploring Reuse Opportunities" at the 14th IEEE International Conference on Information Reuse and Integration (IRI'13), San Francisco, CA, August 2013.

» Garth Gibson (CMU) presented a keynote talk "Modern Storage Systems: Revolution or Evolution" to the ASME 2013 Conference on Information Storage and Processing Systems (ISPS'13), Santa Clara, CA, June 2013.

» M. Satyanarayanan (CMU) presented a keynote talk at CompArch 2013 conference, Vancouver, BC, June 2013.

» Greg Ganger (CMU) gave a keynote talk at EMC University Day on "Exploiting Bounded Staleness for High-Performance 'Big Data' Systems" in June, which introduced some ISTC-CC projects.

» Phil Gibbons (Intel) presented an invited retrospective talk "25 Years of SPAA" at the 25th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA'13), Montreal, Canada, July 2013.

» Mor Harchol-Balter (CMU) gave invited talks on "Power Management in Data Centers: Theory and Practice" at CMU (Tepper School of Business and Computer Science), at University of Southern California (Computer Science), at Georgia Tech (Computer Science), and at INFORMS-Applied Probability Society.

» Ling Liu and Calton Pu (both GA Tech) gave invited talks at the Big Data Summer Camp in Peking, China, July 2013.

» Calton Pu (GA Tech) gave an Invited talk in Big Data Summer Camp in Peking, China, July 2013.

» S. Yalamanchili (GA Tech) gave an invited talk, "Scaling Data Warehousing Applications using GPUs," at the Second International Workshop on Performance Analysis of Workload Optimized Systems, April 2013.

Year in Review

» Onur Mutlu (CMU) gave an invited talk "Memory Scaling: A Systems Architecture Perspective" at MemCon 2013, Santa Clara, CA, August 2013. He also gave invited talks at Bogazici University, Barcelona Supercomputing Center, and INRIA.

» Jeff Parkhurst (Intel) gave a talk at IDF on "Beyond Hadoop MapReduce: Processing Big Data" with Ted Willke and Chris Black (Intel) as co-presenters.

2013 Quarter 2

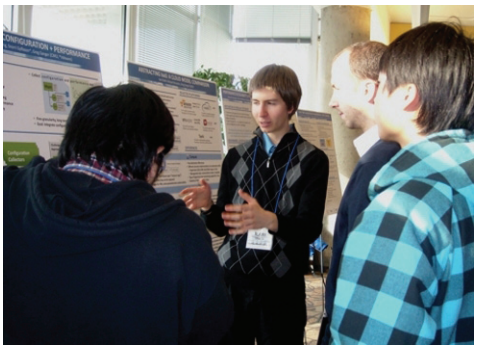
» Shicong Meng (GA Tech) won the 2012 SPEC Distinguished Dissertation Award for his Ph.D. dissertation "Monitoring-as-a-Service in the Cloud."

» Yoongu Kim (CMU) and Daniel Lustig (Princeton) were awarded Intel Graduate Fellowships.

» Wolf Richter (CMU), Haicheng Wu (GA Tech), and Chris Fallin/Gennady Pekhimenko (CMU) were awarded fellowships from IBM, NVIDIA, and Qualcomm, respectively.

» Mor Harchol-Balter (CMU) served as General Chair for Sigmetrics '13.

» Jai Dayal, Karsten Schwan, Jay Lofstead, Matthew Wolf, Scott Klasky, Hasan Abbasi, Norbert Podhorszki, Greg Eisenhauer and Fang Zhen (GA Tech) won the best paper award at the International Workshop on High Performance Data Intensive Computing (HPDIC'13), a satellite workshop of IPDPS'13, for their paper "I/O Containers: Managing the Data Analytics and Visualization Pipelines of High End Codes."



Ilari Shafer, CMU, describes his research on vQuery to a group of ISTC-CC Retreat attendees.

» The ISTC-CC created a benchmark page www.istc-cc.cmu.edu/research/benchmarks/ for cloud-related workloads.

» Daniel P. Siewiorek (CMU) has been named director of the Quality of Life Technology (QoLT) Center, a National Science Foundation Engineering Research Center.

» Margaret Martonosi (Princeton) received the AT&T / NCWIT Undergraduate Research Mentoring Award.

» Karsten Schwan and Vanish Talwar (Georgia Tech) organized a research track and panel session on 'Management of Big Data Systems' for the ICAC conference in San Jose, June 2013.

» S. Yalamanchili (Georgia Tech) presented "New Workloads and New Rules — New Challenges," an invited talk at the Second International Workshop on Performance Analysis of Workload Optimized Systems, April 2013.

» Onur Mutlu (CMU) gave an invited talk on "Memory Scaling: A Systems Architecture Perspective," at the 5th IMW, Monterey, CA, May 2013.

» Guy Blelloch (CMU) gave an invited talk on "Big Data on Small Machines" at the "Big Data Analytics" workshop in Cambridge, UK in May.

» Guy Blelloch (CMU) gave an invited talk on "Internally Deterministic Parallel Algorithms" at Workshop on Determinism and Correctness in Parallel Programming, in Houston in March.

» Mahadev Satyanarayanan (Satya) (CMU) gave a well-attended talk on cloudlets at IBM Research on May 21.

» Garth Gibson presented a keynote talk to the 2013 ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC'13), New York, NY, "Concurrent Write Sharing: Overcoming the Bane of File Systems."

» Padmanabhan Pillai, "Rethinking Cloud Architecture for Mobile Multimedia Applications," at Georgia Tech, CERCS IAB meeting, April 2013.

» Jeff Parkhurst spoke at Cisco Big Data research summit on "Big Data Vision and University Research Overview".

2013 Quarter 1

» Onur Mutlu (CMU) has been appointed the Dr. William D. and Nancy W. Strecker Early Career Professor.

» David Andersen (CMU) and Michael Kaminsky (Intel) were awarded the Allen Newell Award for Research Excellence, given annually by the School of Computer Science at Carnegie Mellon University, for their work on the FAWN project.

» Aapo Kyrola (CMU PhD student) was awarded a 2013 Graduate Fellowship from the VMWare Academic Program to support his work on large-scale machine learning and graph computation.

» Gennady Pekhimenko (CMU PhD Student) was awarded a 2013 Microsoft Research PhD Fellowship to support his work on computer architecture.

» Julian Shun (CMU PhD student) was awarded a 2013 Facebook Fellowship to support his work on parallel computing.

» Phil Gibbons (Intel Labs), in his role as Editor-in-Chief, successfully launched the ACM Transactions on Parallel Computing. He also accepted an invitation to the Editorial Board of the newly launched IEEE Transactions on Cloud Computing.

» Wyatt Lloyd (Princeton PhD Student) presented "Stronger Consistency and Semantics for Geo-Replicated Storage" at Georgia Tech, Penn State, ATT Research, Univ. of Waterloo, USC, Univ. of Toronto, and at University College London. His talk is based on ISTC-CC funded work.

2012 Quarter 4

» Garth Gibson (CMU) elected as ACM Fellow for contributions to the performance and reliability of storage systems.

» Ion Stoica (UC Berkeley) elected as ACM Fellow for contributions to networking, distributed systems and cloud computing.

continued on pg. 34

ISTC-CC News

SEPT 16, 2013

Schwan and Wolf Co-recipients of the 2013 R&D100 Award

Called the "Oscars of Innovation", the R&D 100 Awards recognize and celebrate the top 100 technology products of the year. Past winners have included sophisticated testing equipment, innovative new materials, chemistry breakthroughs, biomedical products, consumer items, and high-energy physics. The R&D 100 Awards span industry, academia, and government-sponsored research.

Karsten Schwan and Matt Wolf (GA Tech) are co-recipients of R&D Magazine's 2013 R&D100 Award for Information Technologies for development of the ADIOS system for high performance I/O. The project was led by a research team at Oak Ridge National Labs, in collaboration with additional researchers at Rutgers Univ. and Sandia Labs.

-- Info from www.rdmag.com

JULY 31, 2013

Margaret Martonosi Awarded ABIE Prize for Technical Leadership

The Anita Borg Institute (ABI), a non-profit organization focused on the advancement of women in computer science and engineering, announced the winners of the 2013 Grace Hopper Celebration ABIE Awards. The ABIE Awards give the community of women technologists a chance to honor leaders in the categories of technology leadership, social impact, change agent, education, and emerging leader. Winners are nominated by their peers, and chosen by a panel of fellow technologists and past ABIE Award winners. This year, ABI received a record number of nominations for distinguished technologists in every category.

The ABIE Technical Leadership Award has been granted to Dr. Margaret Martonosi, Hugh Trumbull Adams '35 Professor of Computer Science at Princeton University. She is one of the foremost researchers in power-efficient computer architectures. Her work has greatly shaped computing's response to the grand challenge of power dissipation. In recent years she has also developed mobile sensing systems and mobile networking technologies specifically suited to the developing world.

-- Info from Marketwired

JULY 17, 2013

Karsten Schwan Reports on Management of Big Data Panel

ISTC-CC faculty Karsten Schwan organized a panel on Management of Big Data Systems at the International Conference on Autonomic Computing in San Jose in June 2103, featuring speakers from many of the key 'big data' companies in the U.S. The well-attended panel's charge was as follows: "New challenges for managing 'big data' applications arise from new usage models now being envisioned and/or pursued by researchers and practitioners. Such usage models include 'spot market' pricing for virtual machines at Amazon, 'fast data' processing for online data queries, and real-time analytics for the expected exabyte-level outputs of future high performance machines. Coupled with such use cases are new opportunities derived from the potentially immeasurably large collective data volumes captured by end devices like smartphones, wearables, and others. The purpose of this panel is to identify and discuss the 'management' issues that arise for new cloud usage models and big data applications, and to describe new problems the community should investigate. A desired outcome is to find issues common to these multiple usages and environments, and to discover and investigate cross-cutting solutions."

JULY 7, 2013

New ISTC-CC Benchmarks Page

ISTC-CC researchers rely extensively on benchmarks for evaluating novel cloud computing infrastructure concepts and systems. To increase sharing within the ISTC community, as well as the wider community, we maintain a list of benchmarks that we have constructed and shared as well as a number of others that we have found useful.

JULY 1, 2013

GraphLab 2 Workshop a Success

The second Graph Lab Workshop was held in San Francisco on July 1, 2013 and was a great success. Over 570 participants attended from academia and industry to discuss challenges of applied large scale machine learning. Among the presentations given during the one day workshop were three from ISTC-CC members. Prof. Carlos Guestrin (GraphLab Inc. & University of Washington) spoke on "Graphs at Scale with GraphLab." Dr. Theodore Willke (Intel Labs) gave an invited talk on "Intel GraphBuilder 2.0", and Aapo Kyrola (CMU) talked about "What can you do with GraphChi-- what's new?"

JULY 1, 2013

Call for Users: NSF PRObE 1000 Node Systems Research Testbed

Garth Gibson's long-running effort, in collaboration with Gary Grider of LANL and the New Mexico Consortium, to make a large-scale testbed available for systems researchers has come to fruition. See www.pdl.cmu.edu/PRObE/participate.shtml to find out how to apply for 1000 nodes for systems research experiments, via NSF's PROBe.

JUNE 21, 2013

Georgia Tech Grad Student wins 2012 SPEC Distinguished Dissertation Award

Congratulations to Shicong Meng (nominated by Professor Ling Liu from the Georgia Institute of Technology) who has been awarded the 2012 SPEC Distinguished Dissertation Award, for hew resesarch on "Monitoring-as-a-

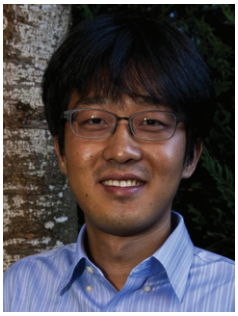
Service in the Cloud." For the award, the Research Group of the Standard Performance Evaluation Corporation (SPEC) selects a Ph.D. student whose thesis is regarded to be an exceptional, innovative contribution in the scope of the SPEC Research Group for the second time. Nine nominations were submitted and reviewed by the selection committee. The criteria for the selection are the overall contribution in terms of scientific originality, practical relevance, impact, and quality of the presentation. The winner will receive $1000, which will be awarded at the ICPE 2013 International Conference in Prague, Czech Republic.

-- with info from the SPEC Research Group News Room

MARCH 2013

Kim and Lustig Awarded Intel Ph.D. Fellowships

Congratulations to Yoongu Kim (CMU) and Daniel Lustig (Princeton) who were awarded graduate fellowships through the Intel Ph.D. Fellowship Program. This program seeks to find the top Ph.D. students at leading U.S. universities through a competitive and rigorous selection process.



MARCH 2013

Wolfgang Richter Awarded IBM Ph.D. Fellowship

Congratulations to Wolf Richter, who will be receiving an IBM Ph.D. Fellowship. The IBM Ph.D. Fellowship Awards Program is a worldwide program, which honors exceptional Ph.D. students who have an interest in solving problems of interest to IBM and which are fundamental to innovation including, innovative software, new types of computers, technology, and interdisciplinary projects that create social and business value. The 2013-2014 Fellowship begins in the fall semester of 2013 and covers the academic year; an associated internship may be a summer assignment in 2013 or 2014.

FEB 7, 2013

Onur Mutlu has been appointed the Dr. William D. and Nancy W. Strecker Early Career Professor

Onur Mutlu has been appointed the Dr. William D. and Nancy W. Strecker Early Career Professor in the top-ranked Department of Electrical and Computer Engineering. Mutlu, who directs the SAFARI research group at CMU, says his group is researching how to make computing platforms and chips — from those used in mobile systems to those used in large-scale data centers — much more energy-efficient, predictable, robust and capable. A major focus of his group is developing microprocessors, computer memories and platforms that can efficiently, quickly and reliably store, manipulate and communicate massive amounts of data. To do this, Mutlu's group is rethinking how computer memory should be designed.

-- Carnegie Mellon 8.5x11 News, Feb. 7, 2013

FEB 7, 2013

Aapo Kyrölä Wins VMWare Graduate Fellowship

Aapo Kyrölä, a Ph.D. student in the Computer Science Department, has been awarded a 2013 Graduate Fellowship from the VMWare Academic Program to support his work on large-scale machine learning and graph computation.

He is one of three recipients this year of the fellowship, which begins in September and includes a stipend of $35,000, plus full tuition and fees. The fellowship supports outstanding students who are pursuing research related to VMWare's business interests, such as core machine virtualization and work related to cloud computing.

Kyrölä's main research project is GraphChi, a disk-based system for computing efficiently on graphs with billions of edges, such as social networks or web graphs. By using a novel Parallel Sliding Windows algorithm, GraphChi is able to execute many advanced data mining, machine learning and recommendation algorithms using just a single consumer-level computer. GraphChi is part of the GraphLab project. Guy Blelloch, professor of computer science, and Carlos Guestrin, now at the University of Washington, are his advisers.

-- Carnegie Mellon 8.5x11 News, Feb. 7, 2013

FEB 7, 2013

Gennady Pekhimenko Receives Microsoft Research Ph.D. Fellowship

Gennady Pekhimenko, a Ph.D. student in the Computer Science Department, is among 12 students of U.S. universities who are recipients of 2013 Microsoft Research Ph.D. Fellowships.

Pekhimenko's research focus is improving energy and performance characteristics of modern memory subsystems. In particular, he is applying new compression algorithms to on-chip caches and main memory to provide higher capacity with minimal hardware changes to existing designs.

The two-year fellowship covers all tuition and fees for the 2013-14 and 2014-15 academic years and includes a travel allowance, the offer of a paid internship, and a $28,000 annual stipend.

-- Carnegie Mellon 8.5x11 News, Feb. 7, 2013

continued on pg. 36

Recent Publications

Enhanced Monitoring-as-a-Service for Effective Cloud Management

Shicong Meng, Ling Liu

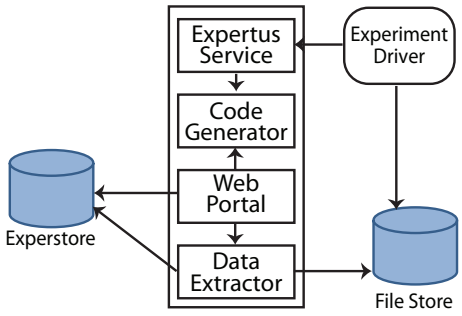
IEEE Transactions on Computers, 62(9), September 2013.

This paper introduces the concept of monitoring-as-a-service (MaaS), its main components, and a suite of key functional requirements of MaaS in Cloud. We argue that MaaS should support not only the conventional state monitoring capabilities, such as instantaneous violation detection, periodical state monitoring and single tenant monitoring, but also performance-enhanced functionalities that can optimize on monitoring cost, scalability, and the effectiveness of monitoring service consolidation and isolation. In this paper we present three enhanced MaaS capabilities and show that window based state monitoring is not only more resilient to noises and outliers, but also saves considerable communication cost. Similarly, violation-likelihood based state monitoring can dynamically adjust monitoring intensity based on the likelihood of detecting important events, leading to significant gain in monitoring service consolidation. Finally, multi-tenancy support in state monitoring allows multiple Cloud users to enjoy MaaS with improved performance and efficiency at more affordable cost. We perform extensive experiments in an emulated Cloud environment with real world system and network traces. The experimental results suggest that our MaaS framework achieves significant lower monitoring cost, higher scalability and better multitenancy performance.

An Automated Approach to Create, Store, and Analyze Large-scale Experimental Data in Clouds

Deepal Jayashinghe, Josh Kimball, Siddarth Choudhary, Tao Zhu, Calton Pu

14th IEEE International Conference on Information Reuse and Integration (IRI'13), August 2013.



Key Components of Automated Infrastructure

The flexibility and scalability of computing clouds make them an attractive application migration target; yet, the cloud remains a black-box for the most part. In particular, their opacity impedes the efficient but necessary testing and tuning prior to moving new applications into the cloud. A natural and presumably unbiased approach to reveal the cloud's complexity is to collect significant performance data by conducting more experimental studies. However, conducting large-scale system experiments is particularly challenging because of the practical difficulties that arise during experimental deployment, configuration, execution and data processing. In this paper we address some of these challenges through Expertus – a flexible automation framework we have developed to create, store and analyze large-scale experimental measurement data. We create performance data by automating the measurement processes for large-scale experimentation, including: the application deployment, configuration, workload execution and data collection processes. We have automated the processing of heterogeneous data as well as the storage of it in a data warehouse, which we have specifically designed for housing measurement data. Finally, we have developed a rich web portal to navigate, statistically analyze and visualize the collected data. Expertus combines template-driven code generation techniques with aspect-oriented programming concepts to generate the necessary resources to fully automate the experiment measurement process. In Expertus, a researcher provides only the high-level description about the experiment, and the framework does everything else. At the end, the researcher can graphically navigate and process the data in the web portal.

Leveraging Endpoint Flexibility in Data-Intensive Clusters

Mosharaf Chowdhury, Srikanth Kandula, Ion Stoica

ACM SIGCOMM 2013 Conference (SIGCOMM'13), August 2013.

Many applications do not constrain the destinations of their network transfers. New opportunities emerge when such transfers contribute a large amount of network bytes. By choosing the endpoints to avoid congested links, completion times of these transfers, as well as that of others without similar flexibility can be improved. In this paper, we focus on leveraging the flexibility in replica placement during writes to cluster file systems (CFSes), which account for almost half of all cross-rack traffic in data-intensive clusters. The replicas of a CFS write can be placed in any subset of machines as long as they are in multiple fault domains and ensure a balanced use of storage throughout the cluster. We study CFS interactions with the cluster network, analyze optimizations for replica placement, and propose Sinbad -- a system that identifies imbalance and adapts replica destinations to navigate around congested links. Experiments on EC2 and trace-driven simulations show that block writes complete 1.3× (respectively, 1.58×) faster as the network becomes more balanced. As a collateral benefit, end-to-end completion times of data- intensive jobs improve as well. Sinbad does so with little impact on the long-term storage balance.

A General Bootstrap Performance Diagnostic

Ariel Kleiner, Ameet Talwalkar, Sameer Agarwal, Ion Stoica, Michael I. Jordan

19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'13), August 2013.

As datasets become larger, more complex, and more available to diverse groups of analysts, it would be quite

Recent Publications

useful to be able to automatically and generically assess the quality of estimates, much as we are able to automatically train and evaluate predictive models such as classifiers. However, despite the fundamental importance of estimator quality assessment in data analysis, this task has eluded highly automatic solutions. While the bootstrap provides perhaps the most promising step in this direction, its level of automation is limited by the difficulty of evaluating its finite sample performance and even its asymptotic consistency. Thus, we present here a general diagnostic procedure which directly and automatically evaluates the accuracy of the bootstrap's outputs, determining whether or not the bootstrap is performing satisfactorily when applied to a given dataset and estimator. We show that our proposed diagnostic is effective via an extensive empirical evaluation on a variety of estimators and simulated and real datasets, including a real-world query workload from Conviva, Inc. involving 1.7TB of data (i.e., approximately 0.5 billion data points).

A Short Primer on Causal Consistency

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

USENIX;login, 38(4), August 2013.

The growing prevalence of geo-distributed services that span multiple geographically separate locations has triggered a resurgence of research on consistency for distributed storage. The CAP theorem and other earlier results prove that no distributed storage system can simultaneously provide all desirable properties—e.g., CAP shows this for strong Consistency, Availability, and Partition tolerance—and some must be sacrificed to enable others. In this article, we suggest causal consistency represents an excellent point in this tradeoff space; it is compatible with strong performance and liveness properties while being far easier to reason about than the previously-settled-for choice: "eventual" consistency.

Cuckoo Filter: Better Than Bloom

Bin Fan, David G. Andersen, Michael Kaminsky

USENIX;login, 38 (4), August 2013.

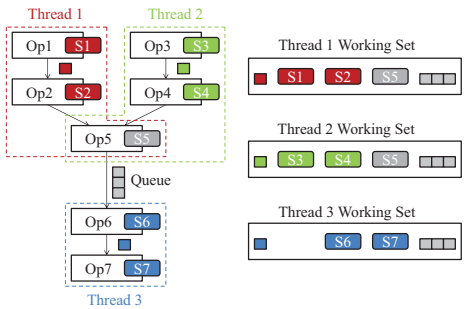
High-speed approximate set-membership tests are critical for many applications, and Bloom filters are used widely in practice, but do not support deletion. In this article, we describe a new data structure called the cuckoo filter that can replace Bloom filters for many approximate set-membership test applications. Cuckoo filters allow adding and removing items dynamically while achieving higher lookup performance, and also use less space than conventional, non-deletion-supporting Bloom filters for applications that require low false positive rates ($\epsilon < 3\%$).

Cache Topology Aware Mapping of Stream Processing Applications onto CMPs

Fang Zheng, Chitra Venkatramani, Karsten Schwan, Rohit Wagle

33rd IEEE International Conference on Distributed Computing Systems (ICDCS'13), July 2013.

Data Stream Processing is an important class of data intensive applications in the "Big Data" era. Chip Multi-Processors (CMPs) are the standard hosting platforms in modern data centers. Gaining high performance for stream processing applications on CMPs is therefore of great interest. Since the performance of stream processing applications largely depends on their effective use of the complex cache structure present on CMPs, this paper proposes the StreamMap approach for tuning streaming applications' use of cache. Our major idea is to map application threads to CPU cores to facilitate data sharing AND mitigate memory resource contention among threads in a holistic manner. Applying StreamMap to the IBM's System S middleware leads to improvements of up to 1.8× in the performance of realistic applications over standard Linux OS scheduler on three different CMP platforms.



Cache Behavior of Streaming Programs

Volley: Violation Likelihood Based State Monitoring for Datacenters

Shicong Meng, Arun Iyengar, Isabelle Rouvellou, Ling Liu

33rd IEEE Int. Conf. on Distributed Computing Systems (ICDCS'13), July 2013.

Distributed state monitoring plays a critical role in Cloud datacenter management. One fundamental problem in distributed state monitoring is to minimize the monitoring cost while maximizing the monitoring accuracy at the same time. In this paper, we present Volley, a violation likelihood based approach for efficient distributed state monitoring in Cloud datacenters. Volley achieves both efficiency and accuracy with a flexible monitoring framework which uses dynamic monitoring intervals determined by the likelihood of detecting state violations. Volley consists of three unique techniques. It utilizes efficient node-level adaptation algorithms that minimize monitoring cost with controlled accuracy. Volley also employs a distributed scheme that coordinates the adaptation on multiple monitor nodes of the same task for optimal tasklevel efficiency. Furthermore, it enables multi-task level cost reduction by exploring state correlation among monitoring tasks. We perform extensive experiments to evaluate Volley with system, network and application monitoring tasks in a virtualized datacenter environment. Our results show that Volley can reduce considerable monitoring cost and still deliver

continued on pg. 10

Recent Publications

continued from pg. 9

user specified monitoring accuracy under various scenarios.

M/G/k with Staggered Setup

Anshul Gandhi, Mor Harchol-Balter

OR Letters, 41(4), July 2013.

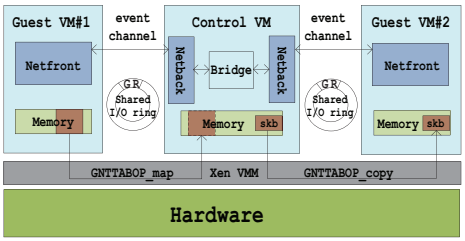
We consider the M/G/k/staggered-setup, where idle servers are turned off to save cost, necessitating a setup time for turning a server back on; however, at most one server may be in setup mode at any time. We show that, for exponentially distributed setup times, the response time of an M/G/k/staggered-setup approximately decomposes into the sum of the response time for an M/G/k and the setup time, where the approximation is nearly exact. This generalizes a prior decomposition result for an M/M/k/staggered-setup.

Residency Aware Inter-VM Communication in Virtualized Cloud: Performance Measurement and Analysis

Qi Zhang, Ling Liu, Yi Ren, Kisung Lee, Yuzhe Tang, Xu Zhao, Yang Zhou

6th IEEE International Conference on Cloud Computing (CLOUD'13), June-July, 2013.

A known problem for virtualized cloud data centers is the inter-VM communication inefficiency for data transfer between co-resident VMs. Several engineering efforts have been made on building a shared memory based channel between co-resident VMs. The implementations differ in terms of whether user/program transparency, OS kernel transparency or VMM transparency is supported. However, none of existing works has engaged in an in-depth measurement study with quantitative and qualitative analysis on performance improvements as well as tradeoffs introduced by such a residency-aware inter-VM communication mechanism. In this paper we present an extensive experimental study, aiming at addressing a number of fundamental issues and providing deeper insights regarding the design of a shared memory channel for co-resident VMs. Example questions include how much performance gains can a residency-aware shared memory inter-VM communication mechanism provide under different mixtures of local and remote network I/O workloads, what overhead will the residence-awareness detection and communication channel switch introduce over the remote inter-VM communication, what factors may exert significant impact on the throughput and latency performance of such a shared memory channel. We believe that this measurement study not only helps system developers to gain valuable lessons and generate new ideas to further improve the inter-VM communication performance. It also offers new opportunities for cloud service providers to deploy their services more efficiently and for cloud service consumers to improve the performance of their application systems running in the Cloud.



Xen architecture overview.

Secure Cloud Storage Service with An Efficient DOKS Protocol

ZhengTao Jiang, Ling Liu

IEEE 10th International Conference on Services Computing (SCC'13), June-July 2013.

Storage services based on public clouds provide customers with elastic storage and on-demand accessibility. However, moving data to remote cloud storage also raises privacy concerns. Cryptographic cloud storage and search over encrypted data have attracted attentions from both industry and academics. In this paper, we present a new approach to constructing efficient oblivious keyword search (OKS) protocol, which permits fast search (i.e., sub-linear time) and relatively short ciphertext, while providing provably strong privacy for both users and cloud storage service providers. Previous OKS protocols have ciphertext size linear in the number of keywords, which consume much storage space and relatively long searching time. We formally define a Disjunctively Oblivious Keyword Search (DOKS) protocol realizing oblivious keyword search with the ciphertext size constant in size of keywords, which is significantly less than that of previous OKS protocols. Our approach improves both the privacy and efficiency of existing OKS protocols. With DOKS, adversary cannot distinguish two search keywords submitted by users, and cannot know the relations between ciphertext of documents and search keywords. A search keyword cannot be reused by adversaries. Users can get the matching documents without revealing statistical information on search keywords.

Performance Overhead Among Three Hypervisors: An Experimental Study using Hadoop Benchmarks

Jack Li, Qingyang Wang, Deepal Jayasinghe, Junhee Park, Tao Zhu, Calton Pu

2nd IEEE International Congress on Big Data (BigData'13), June-July 2013.

Hypervisors are widely used in cloud environments and their impact on application performance has been a topic of significant research and practical interest. We conduct experimental measurements of several benchmarks using Hadoop MapReduce to evaluate and compare the performance impact of three popular hypervisors: a commercial hypervisor, Xen, and KVM. We found that differences in the workload type (CPU or I/O intensive), workload size and VM placement yielded significant performance differences among the hypervisors. In our study, we used the three hypervisors to run several MapReduce benchmarks such as Word Count, TestDSFIO, and TeraSort and further validated our observed hypothesis using microbenchmarks. We observed for CPU-bound benchmark, the performance difference between the three hypervisors was negligible; however, significant performance variations were seen for I/O-bound

benchmarks. Moreover, adding more virtual machines on the same physical host degraded the performance on all three hypervisors, yet we observed different degradation trends amongst them. Concretely, the commercial hypervisor is 46% faster at TestDFSIO Write than KVM, but 49% slower in the TeraSort benchmark. In addition, increasing the workload size for TeraSort yielded completion times for CVM that were two times that of Xen and KVM. The performance differences shown between the hypervisors suggests that further analysis and consideration of hypervisors is needed in the future when deploying applications to cloud environments.

Who Is Your Neighbor: Net I/O Performance Interference in Virtualized Clouds

Xing Pu, Ling Liu, Yiduo Mei, Sankaran Sivathanu, Younggyun Koh, Calton Pu, Yuanda Cao

IEEE Transactions on Services Computing, vol 6, 2013.

User-perceived performance continues to be the most important QoS indicator in cloud-based data centers today. Effective allocation of virtual machines (VMs) to handle both CPU intensive and I/O intensive workloads is a crucial performance management capability in virtualized clouds. Although a fair amount of researches have dedicated to measuring and scheduling jobs among VMs, there still lacks of in-depth understanding of performance factors that impact the efficiency and effectiveness of resource multiplexing and scheduling among VMs. In this paper, we present the experimental research on performance interference in parallel processing of CPU-intensive and network-intensive workloads on Xen virtual machine monitor (VMM). Based on our study, we conclude with five key findings which are critical for effective performance management and tuning in virtualized clouds. First, colocating network-intensive workloads in isolated VMs incurs high overheads of switches and events in Dom0 and VMM. Second, colocating CPU-intensive workloads in isolated VMs incurs high CPU contention due to fast I/O processing in I/O channel. Third, running CPU-intensive and network-intensive workloads in conjunction incurs the least resource contention, delivering higher aggregate performance. Fourth, performance of network-intensive workload is insensitive to CPU assignment among VMs, whereas adaptive CPU assignment among VMs is critical to CPU-intensive workload. The more CPUs pinned on Dom0 the worse performance is achieved by CPU-intensive workload. Last, due to fast I/O processing in I/O channel, limitation on grant table is a potential bottleneck in Xen. We argue that identifying the factors that impact the total demand of exchanged memory pages is important to the in-depth understanding of interference costs in Dom0 and VMM.

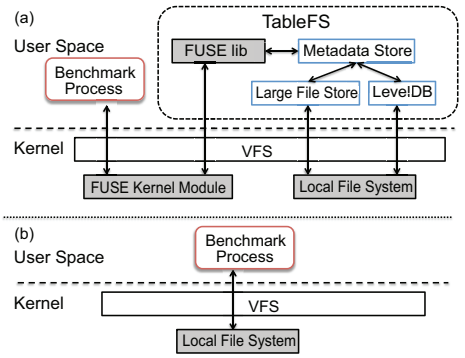
TABLEFS: Enhancing Metadata Efficiency in the Local File System

Kai Ren,Garth Gibson

2013 USENIX Annual Technical Conference (ATC'13), June 2013.

File systems that manage magnetic disks have long recognized the importance of sequential allocation and large transfer sizes for file data. Fast random access has dominated metadata lookup data structures with increasing use of B-trees on-disk. Yet our experiments with workloads dominated by metadata and small file access indicate that even sophisticated local disk file systems like Ext4, XFS and Btrfs leave a lot of opportunity for performance improvement in workloads dominated by metadata and small files.

In this paper we present a stacked file system, TABLEFS, which uses another local file system as an object store. TABLEFS organizes all metadata into a single sparse table backed on disk using a Log-Structured Merge (LSM) tree, LevelDB in our experiments. By stacking, TABLEFS asks only for efficient large file allocation and access from the underlying local file system. By using an LSM tree, TABLEFS ensures metadata is written to disk in large, non-overwrite, sorted and indexed



(a) The architecture of TABLEFS. A FUSE kernel module redirects file system calls from a benchmark process to TABLEFS, and TABLEFS stores objects into either LevelDB or a large file store. (b) When we benchmark a local file system, there is no FUSE overhead to be paid.

logs. Even an inefficient FUSE based user level implementation of TABLEFS can perform comparably to Ext4, XFS and Btrfs on data-intensive benchmarks, and can outperform them by 50% to as much as 1000% for metadata-intensive workloads. Such promising performance results from TABLEFS suggest that local disk file systems can be significantly improved by more aggressive aggregation and batching of metadata updates.

Exact Analysis of the M/M/k/setup Class of Markov Chains via Recursive Renewal Reward

Anshul Gandhi, Sherwin Doroudi, Mor Harchol-Balter, Alan Scheller-Wolf

ACM SIGMETRICS 2013 Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'13), June 2013.

The M/M/k/setup model, where there is a penalty for turning servers on, is common in data centers, call centers and manufacturing systems. Setup costs take the form of a time delay, and sometimes there is additionally a power penalty, as in the case of data centers. While the M/M/1/setup was exactly analyzed in 1964, no exact analysis exists to date for the M/M/k/setup with k > 1. In this paper we provide the rst exact, closed-form analysis for the M/M/k/setup and some of its

continued on pg. 12

Recent Publications

continued from pg. 11

important variants including systems in which idle servers delay for a period of time before turning oor can be put to sleep. Our analysis is made possible by our development of a new technique, Recursive Renewal Reward (RRR), for solving Markov chains with a repeating structure. RRR uses ideas from renewal reward theory and busy period analysis to obtain closed-form expressions for metrics of interest such as the transform of time in system and the transform of power consumed by the system. The simplicity, intuitiveness, and versatility of RRR makes it useful for analyzing Markov chains far beyond the M/M/k/setup. In general, RRR should be used to reduce the analysis of any 2-dimensional Markov chain which is innite in at most one dimension and repeating to the problem of solving a system of polynomial equations. In the case where all transitions in the repeating portion of the Markov chain are skip-free and all up/down arrows are unidirectional, the resulting system of equations will yield a closed-form solution.

Scalable Crowd-Sourcing of Video from Mobile Devices

Pieter Simoens, Yu Xiao, Padmanabhan Pillai, Zhuo Chen, Kiryong Ha, Mahadev Satyanarayanan

11th International Conference on Mobile Systems, Applications, and Services (MobiSys'13), June 2013.

We propose a scalable Internet system for continuous collection of crowd-sourced video from devices such as Google Glass. Our hybrid cloud architecture, GigaSight, is effectively a Content Delivery Network (CDN) in reverse. It achieves scalability by decentralizing the collection infrastructure using cloudlets based on virtual machines (VMs). Based on time, location, and content, privacy sensitive information is automatically removed from the video. This process, which we refer to as denaturing, is executed in a user-specific VM on the cloudlet. Users can perform content-based searches on the total catalog of denatured videos. Our experiments reveal the bottlenecks for video upload, denaturing, indexing, and contentbased search. They also provide insight on how parameters such as frame rate and resolution impact scalability.

Just-in-Time Provisioning for Cyber Foraging

Kiryong Ha, Padmanabhan Pillai, Wolfgang Richter, Yoshihisa Abe, Mahadev Satyanarayanan

11th International Conference on Mobile Systems, Applications, and Services (MobiSys'13), June 2013.

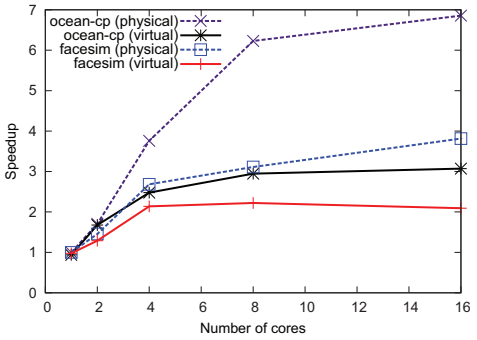
Cloud offload is an important technique in mobile computing. VM-based cloudlets have been proposed as offload sites for the resource-intensive and latency-sensitive computations typically associated with mobile multimedia applications. Since cloud offload relies on precisely-configured back-end software, it is difficult to support at global scale across cloudlets in multiple domains. To address this problem, we describe just-in-time (JIT) provisioning of cloudlets under the control of an associated mobile device. Using a suite of five representative mobile applications, we demonstrate a prototype system that is capable of provisioning a cloudlet with a non-trivial VM image in 10 seconds. This speed is achieved through dynamic VM synthesis and a series of optimizations to aggressively reduce transfer costs and startup latency.

A Hidden Cost of Virtualization when Scaling Multicore Applications

Xiaoning Ding, Michael A. Kozuch, Phillip B. Gibbons

5th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud'13), June 2013.

As the number of cores in a multicore node increases in accordance with Moore's law, the question arises as to whether there are any "hidden" costs of a cloud's virtualized environment when scaling applications to take advantage of larger core counts. This paper identifies one such cost, resulting in up to a 583% slowdown as the multicore application is scaled. Surprisingly, these slowdowns arise even when the application's VM has dedicated use of the underlying physical hardware and does not use emulated resources. Our preliminary findings indicate that the source of the slowdowns is the intervention from the VMM during synchronization-induced idling in the application, guest OS, or supporting libraries. We survey several possible mitigations, and report preliminary findings on the use of "idleness consolidation" and "IPI-free wakeup" as a partial mitigation.



Speedups of ocean-cp and facesim benchmarks from the SPLASH-2X and PARSEC-3.0 suites, on the virtual machine and physical machine varying the number of cores.

Shark: SQL and Rich Analytics at Scale

Reynold S. Xin, Josh Rosen, Matei Zaharia, Michael J. Franklin, Scott Shenker, Ion Stoica

ACM SIGMOD International Conference on Management of Data (SIGMOD'13), June 2013.

Shark is a new data analysis system that marries query processing with complex analytics on large clusters. It leverages a novel distributed memory abstraction to provide a unified engine that can run SQL queries and sophisticated analytics functions (e.g. iterative machine learning) at scale, and efficiently recovers from failures mid-query. This allows Shark to run SQL queries up to 100X faster than Apache Hive, and machine learning programs more than 100X faster than Hadoop. Unlike previous systems, Shark shows that it is possible to achieve these speedups while retaining a MapReduce-like ex-

ecution engine, and the fine-grained fault tolerance properties that such engine provides. It extends such an engine in several ways, including column-oriented in-memory storage and dynamic mid-query replanning, to effectively execute SQL. The result is a system that matches the speedups reported for MPP analytic databases over MapReduce, while offering fault tolerance properties and complex analytics capabilities that they lack.

GraphX: A Resilient Distributed Graph System on Spark

Reynold S. Xin, Joseph E. Gonzalez, Michael J. Franklin, Ion Stoica

First International Workshop on Graph Data Management Experiences and Systems (GRADES'13), June 2013.

From social networks to targeted advertising, big graphs capture the structure in data and are central to recent advances in machine learning and data mining. Unfortunately, directly applying existing data-parallel tools to graph computation tasks can be cumbersome and inefficient. The need for intuitive, scalable tools for graph computation has lead to the development of new graph-parallel systems (e.g. Pregel, PowerGraph) which are designed to efficiently execute graph algorithms. Unfortunately, these new graph-parallel systems do not address the challenges of graph construction and transformation which are often just as problematic as the subsequent computation. Furthermore, existing graph-parallel systems provide limited fault-tolerance and support for interactive data mining.

We introduce GraphX, which combines the advantages of both data-parallel and graph-parallel systems by efficiently expressing graph computation within the Spark data-parallel framework. We leverage new ideas in distributed graph representation to efficiently distribute graphs as tabular data-structures. Similarly, we leverage advances in data-flow systems to exploit in-memory computation and fault-tolerance. We provide powerful new operations to simplify graph construction and transformation. Using these primitives we implement the PowerGraph and Pregel abstractions in less than 20 lines of code. Finally, by exploiting the Scala foundation of Spark, we enable users to interactively load, transform, and compute on massive graphs.

Bolt-on Causal Consistency

Peter Bailis, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica

ACM SIGMOD International Conference on Management of Data (SIGMOD'13), June 2013.

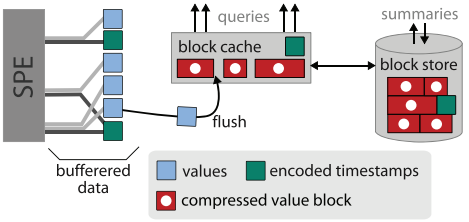
We consider the problem of separating consistency-related safety properties from availability and durability in distributed data stores via the application of a "bolt-on" shim layer that upgrades the safety of an underlying general-purpose data store. This shim provides the same consistency guarantees atop a wide range of widely deployed but often inflexible stores. As causal consistency is one of the strongest consistency models that remain available during system partitions, we develop a shim layer that upgrades eventually consistent stores to provide convergent causal consistency. Accordingly, we leverage widely deployed eventually consistent infrastructure as a common substrate for providing causal guarantees. We describe algorithms and shim implementations that are suitable for a large class of application-level causality relationships and evaluate our techniques using an existing, production-ready data store and with real-world explicit causality relationships.

Specialized Storage for Big Numeric Time Series

Ilari Shafer, Raja R. Sambasivan, Anthony Rowe, Gregory R. Ganger

5th Usenix Workshop on Hot Topics in Storage and File Systems (HotStorage'13), June 2013.

Numeric time series data has unique storage requirements and access patterns that can benefit from specialized support, given its importance in Big Data analyses. Popular frameworks and databases focus on addressing other needs, making them a suboptimal fit. This paper describes the support needed for numeric time series, suggests an architecture for efficient time series storage, and illustrates its potential for satisfying key requirements.



Possible design for numeric time series archival. Streams are separated into timestamps and values, buffered, and written back to a store of compressed blocks. Not shown are metadata storage, higher-level query architecture, or possible distribution across nodes. A stream processing engine (SPE) emphasizes that streaming queries on incoming data are not within scope and can be handled separately.

Multi-tenancy on GPGPU-based Servers

Dipanjan Sengupta, Raghavendra Belapure, Karsten Schwan

7th International Workshop on Virtualization Technologies in Distributed Computing (VTDC'13), June 2013.

While GPUs have become prominent both in high performance computing and in online or cloud services, they still appear as explicitly selected 'devices' rather than as first class schedulable entities that can be efficiently shared by diverse server applications. To combat the consequent likely under-utilization of GPUs when used in modern server or cloud settings, we propose 'Rain', a system level abstraction for GPU "hyperthreading" that makes it possible to efficiently utilize GPUs without compromising fairness among multiple tenant applications. Rain uses a multi-level GPU scheduler that decomposes the scheduling problem into a combination of load balancing and per-device scheduling. Implemented by overriding applications' standard GPU selection calls, Rain operates without the need for appli-

continued on pg. 14

Recent Publications

continued from pg. 13

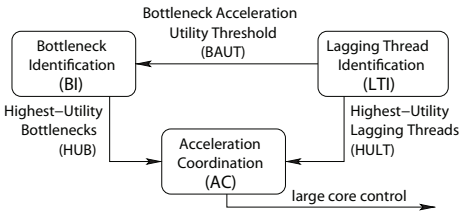
cation modification, making possible GPU scheduling methods that include prioritizing certain jobs, guaranteeing fair shares of GPU resources, and/or favoring jobs with least attained GPU services. GPU multi-tenancy via Rain is evaluated with server workloads using a wide variety of CUDA SDK and Rodinia suite benchmarks, on a multi-GPU, multi-core machine typifying future high end server machines. Averaged over ten applications, GPU multi-tenancy on a smaller scale server platform results in application speedups of up to 1.73x compared to their traditional implementation with NVIDIA's CUDA runtime. Averaged over 25 pairs of short and long running applications, on an emulated larger scale server machine, multi-tenancy results in system throughput improvements of up to 6.71x, and in 43% and 29.3% improvements in fairness compared to using the CUDA runtime and a naïve fair-share scheduler.

Utility-Based Acceleration of Multithreaded Applications on Asymmetric CMPs

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

40th ACM International Symposium on Computer Architecture (ISCA'13), June 2013.

Asymmetric Chip Multiprocessors (ACMPs) are becoming a reality. ACMPs can speed up parallel applications if they can identify and accelerate code segments that are critical for performance. Proposals already exist for using coarse- grained thread scheduling and fine-grained bottleneck acceleration. Unfortunately, there have been no proposals of- fered thus far to decide which code segments to accelerate in cases where both coarse-grained thread scheduling and fine-grained bottleneck acceleration could have value. This paper proposes Utility-Based Acceleration of Multithreaded Applications on Asymmetric CMPs (UBA), a cooperative software/hardware mechanism for identifying and accelerat- ing the most likely critical code segments from a set of multithreaded applications running on an ACMP. The key idea is a new Utility of Acceleration metric that quantifies the performance benefit of accelerating a bottleneck or a thread by taking into account both the criticality and the expected speedup. UBA outperforms the best of two state-of-the-art mechanisms by 11% for single application workloads and by 7% for two-application workloads on an ACMP with 52 small cores and 3 large cores.



Block diagram of UBA.

A Case for Efficient Hardware-Software Cooperative Coordinated Management of Storage and Memory

Justin Meza, Yixin Luo, Samira Khan, Jishen Zhao, Yuan Xie, Onur Mutlu

5th Workshop on Energy-Efficient Design (WEED'13), June 2013.

Most applications manipulate persistent data, yet traditional systems decouple data manipulation from persistence in a two-level storage model. Programming languages and system software manipulate data in one set of formats in volatile main memory (DRAM) using a load/store interface, while storage systems maintain persistence in another set of formats in non-volatile memories, such as Flash and hard disk drives in traditional systems, using a file system interface. Unfortunately, such an approach suffers from the system performance and energy overheads of locating data, moving data, and translating data between the different formats of these two levels of storage that are accessed via two vastly different interfaces. Yet today, new non-volatile memory (NVM) technologies show the promise of storage capacity and endurance similar to or better than Flash at latencies comparable to DRAM, making them prime candidates for providing applications a persistent single-level store with a single load/store interface to access all system data. Our key insight is that in future systems equipped with NVM, the energy consumed executing operating system and file system code to access persistent data in traditional systems becomes an increasingly large contributor to total energy. The goal of this work is to explore the design of a Persistent Memory Manager that coordinates the management of memory and storage under a single hardware unit in a single address space. Our initial simulation-based exploration shows that such a system with a persistent memory can improve energy efficiency and performance by eliminating the instructions and data movement traditionally used to perform I/O operations.

Orchestrated Scheduling and Prefetching for GPGPUs

Adwait Jog, Onur Kayiran, Asit K. Mishra, Mahmut T. Kandemir, Onur Mutlu, Ravishankar Iyer, Chita R. Das

40th ACM International Symposium on Computer Architecture (ISCA'13), June 2013.

In this paper, we present techniques that coordinate the thread scheduling and prefetching decisions in a General Purpose Graphics Processing Unit (GPGPU) architecture to better tolerate long memory latencies. We demonstrate that existing warp scheduling policies in GPGPU architectures are unable to effectively incorporate data prefetching. The main reason is that they schedule consecutive warps, which are likely to access nearby cache blocks and thus prefetch accurately for one another, back-to-back in consecutive cycles. This either 1) causes prefetches to be generated by a warp too close to the time their corresponding addresses are actually demanded by another warp, or 2) requires sophisticated prefetcher designs to correctly predict the addresses required by a future "far-ahead" warp while executing the current warp.

We propose a new prefetch-aware warp scheduling policy that overcomes these problems. The key idea is to

separate in time the scheduling of consecutive warps such that they are not executed back-to-back. We show that this policy not only enables a simple prefetcher to be effective in tolerating memory latencies but also improves memory bank parallelism, even when prefetching is not employed. Experimental evaluations across a diverse set of applications on a 30-core simulated GPGPU platform demonstrate that the prefetch-aware warp scheduler provides 25% and 7% average performance improvement over baselines that employ prefetching in conjunction with, respectively, the commonlyemployed round-robin scheduler or the recently-proposed two-level warp scheduler. Moreover, when prefetching is not employed, the prefetch-aware warp scheduler provides higher performance than both of these baseline schedulers as it better exploits memory bank parallelism.

Space-Efficient, High-Performance Rank & Select Structures

Dong Zhou, David G. Andersen, Michael Kaminsky

12th International Symposium on Experimental Algorithms (SEA'13), June 2013.

Rank & select data structures are one of the fundamental building blocks for many modern succinct data structures. With the continued growth of massive-scale information services, the space efficiency of succinct data structures is becoming increasingly attractive in practice. In this paper, we re-examine the design of rank & select data structures from the bottom up, applying an architectural perspective to optimize their operation. We present our results in the form of a recipe for constructing space and time efficient rank & select data structures for a given hardware architecture. By adopting a cache-centric design approach, our rank & select structures impose space overhead as low as the most space-efficient, but slower, prior designs—only 3.2% and 0.39% extra space respectively—while offering performance competitive with the highestperformance prior designs.

A Heterogeneous Multiple Network-on-Chip Design: An Application-Aware Approach

Asit K. Mishra, Onur Mutlu, Chita R. Das

50th Design Automation Conference (DAC'13), June 2013.

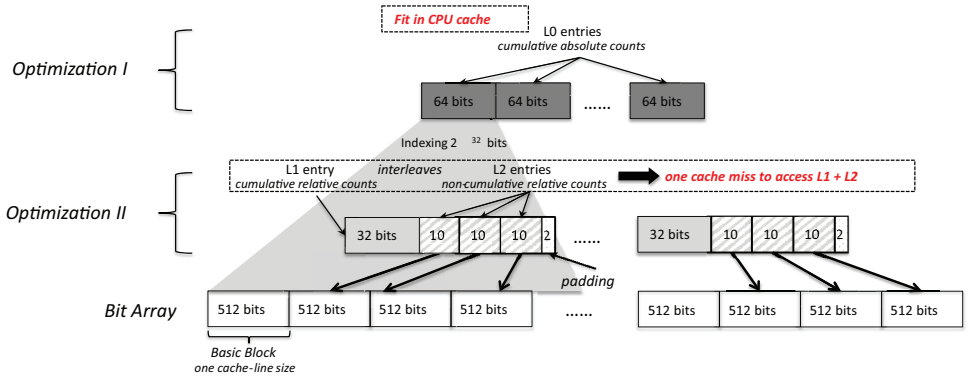
Current network-on-chip designs in chip-multiprocessors are agnostic to application requirements and hence are provisioned for the general case, leading to wasted energy and performance. We observe that applications can generally be classified as either network bandwidth-sensitive or latency-sensitive. We propose the use of two separate networks on chip, where one network is optimized for bandwidth and the other for latency, and the steering of applications to the appropriate network. We further observe that not all bandwidth (latency) sensitive applications are equally sensitive to network bandwidth (latency). Hence, within each network, we prioritize packets based on the relative sensitivity of the applications they belong to. We introduce two metrics, network episode height and length, as proxies to estimate bandwidth and latency sensitivity, to classify and rank applications. Our evaluations show that the resulting heterogeneous two-network design can provide significant energy savings and performance improvements across a variety of workloads compared to a single one-size-fits-all single network and homogeneous multiple networks.

Software-Controlled Transparent Management of Heterogeneous Memory Resources in Virtualized Systems

Min Lee, Vishal Gupta, Karsten Schwan

8th ACM SIGPLAN Workshop on Memory Systems Performance and Correctness (MSPC'13), June 2013.

This paper presents a software-controlled technique for managing the heterogeneous memory resources of next generation multicore platforms with fast 3D die-stacked memory and additional slow off-chip memory. Implemented for virtualized server systems, the technique detects the 'hot' pages critical to program performance in order to then maintain them in the scarce fast 3D memory resources. Challenges overcome for the technique's implementation include the need to minimize its runtime overheads, the lack of hypervisor-level direct visibility into the memory access behavior of guest virtual machines, and the need to make page migration transparent to guests. This paper presents hypervisor-level mechanisms that (i) build a page access history of virtual machines, by periodically scanning pagetable access bits and (ii) intercept guest page table operations to create mirrored page-tables and enable guest-transparent page migration. The methods are implemented in the Xen hypervisor and evaluated on a larger scale multicore platform. The resulting ability to characterize the memory behavior of representative server workloads demonstrates the feasibility of software-managed heterogeneous memory resources.



Our Rank Structure

continued on pg. 16

Recent Publications

continued from pg. 15

Automating the Debugging of Datacenter Applications with ADDA

Cristian Zamfir, Gautam Altekar, Ion Stoica

43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'13), June 2013.

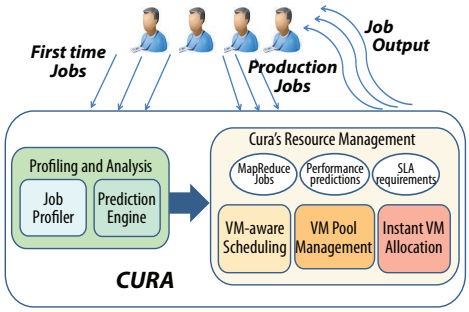
Debugging data-intensive distributed applications running in datacenters is complex and time-consuming because developers do not have practical ways of deterministically replaying failed executions. The reason why building such tools is hard is that non-determinism that may be tolerable on a single node is exacerbated in large clusters of interacting nodes, and datacenter applications produce terabytes of intermediate data exchanged by nodes, thus making full input recording infeasible. We present ADDA, a replay-debugging system for datacenters that has lower recording and storage overhead than existing systems. ADDA is based on two techniques: First, ADDA provides control plane determinism, leveraging our observation that many typical datacenter applications consist of a separate "control plane" and "data plane", and most bugs reside in the former. Second, ADDA does not record "data plane" inputs, instead it synthesizes them during replay, starting from the application's external inputs, which are typically persisted in append-only storage for reasons unrelated to debugging. We evaluate ADDA and show that it deterministically replays real-world failures in Hypertable and Memcached.

Cura: A Cost-optimized Model for MapReduce in a Cloud

Balaji Palanisamy, Aameek Singh, Ling Liu, Bryan Langston

27th IEEE International Parallel & Distributed Processing Symposium (IPDPS'13), May 2013.

On-chip heterogeneity has become key to balancing performance and power constraints, resulting in disparate (functionally overlapping but not equivalent) cores on a single die. Requiring developers to deal with such heterogeneity can impede adoption through increased programming effort and result in cross-platform incompatibility. To evolve systems software toward dynamically accommodating heterogeneity, this paper develops and evaluates the kinship approach and metric for mapping workloads to heterogeneous cores. For this metric, we provide a model and online methods for maximizing utility in terms of performance, power, or latency, to automatically choose the task-to-resource mappings best able to use the different features of heterogeneous cores. Such online scheduling at bounded cost is realized with a hypervisor-level implementation that is evaluated on a variety of actual, experimental heterogeneous platforms. These evaluations demonstrate the both general applicability and the utility of kinship-based scheduling, accommodating dynamic workloads with available resources and scaling both with the number of processes and with different types/configurations of compute resources. Performance improvements with kinship-based scheduling are obtained for runs across multiple generations of heterogeneous platforms.



Cura: System Architecture

Kinship: Resource Management for Performance and Functionally Asymmetric Platforms

Vishakha Gupta, Rob Knauerhase, Paul Brett, Karsten Schwan

ACM International Conference on Computing Frontiers (CF'13), May 2013.

On-chip heterogeneity has become key to balancing performance and power constraints, resulting in disparate (functionally overlapping but not equivalent) cores on a single die. Requiring developers to deal with such heterogeneity can impede adoption through increased programming effort and result in cross-platform incompatibility. To evolve systems software toward dynamically accommodating heterogeneity, this paper develops and evaluates the kinship approach and metric for mapping workloads to heterogeneous cores. For this metric, we provide a model and online methods for maximizing utility in terms of performance, power, or latency, to automatically choose the task-to-resource mappings best able to use the different features of heterogeneous cores. Such online scheduling at bounded cost is realized with a hypervisor-level implementation that is evaluated on a variety of actual, experimental heterogeneous platforms. These evaluations demonstrate the both general applicability and the utility of kinship-based scheduling, accommodating dynamic workloads with available resources and scaling both with the number of processes and with different types/configurations of compute resources. Performance improvements with kinship-based scheduling are obtained for runs across multiple generations of heterogeneous platforms.

FlexIO: I/O Middleware for Location-Flexible Scientific Data Analytics

Fang Zheng, Hongbo Zou, Greg Eisenhauer, Karsten Schwan, Matthew Wolf, Jai Dayal, Tuan-Anh Nguyen, Jianting Cao, Hasan Abbasi, Scott Klasky, Norbert Podhorszki, Hongfeng Yu

27th IEEE International Parallel and Distributed Processing Symposium (IPDPS'13), May 2013.

Increasingly severe I/O bottlenecks on High-End Computing machines are prompting scientists to process simulation output data while simulations are running and before placing data on disk – "in situ" and/or "in-transit". There are several options in placing in-

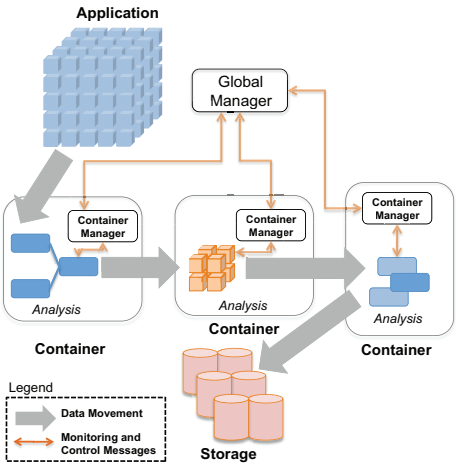
situ data analytics along the I/O path: on compute nodes, on staging nodes dedicated to analytics, or after data is stored on persistent storage. Different placements have different impact on end to end performance and cost. The consequence is a need for flexibility in the location of in situ data analytics. The FlexIO facility described in this paper supports flexible placement of in situ analytics, by offering simple abstractions and methods that help developers exploit the opportunities and trade-offs in performing analytics at different levels of the I/O hierarchy. Experimental results with several large-scale scientific applications demonstrate the importance of flexibility in analytics placement.

I/O Containers: Managing the Data Analytics and Visualization Pipelines of High End Codes

Jai Dayal, Karsten Schwan, Jay Lofstead, Matthew Wolf, Scott Klasky, Hasan Abbasi, Norbert Podhorszki, Greg Eisenhauer, Fang Zhen

International Workshop on High Performance Data Intensive Computing (HPDIC'13), with IPDPS 2013, May 2013. Best paper.

Lack of I/O scalability is known to cause measurable slowdowns for large-scale scientific applications running on high end machines. This is prompting researchers to devise 'I/O staging' methods in which outputs are processed via online analysis and visualization methods to support desired science outcomes. Organized as online workflows and carried out in I/O pipelines, these analysis components run concurrently with science simulations, often using a smaller set of nodes on the high end machine termed 'staging areas'. This paper presents a new approach to dealing with several challenges arising for such online analytics, including: how to efficiently run multiple analytics components on staging area resources providing them with the levels of end-to-end performance they need and how to manage staging resources when analytics actions change due to user or data-dependent behavior. Our approach designs and implements middleware constructs that delineate and manage I/O pipeline resources called 'I/O Containers'. Experimental evaluations of containers with realistic scientific applications demonstrate the feasibility and utility of the approach.



I/O Pipeline Showing I/O Containers

When Cycles Are Cheap, Some Tables Can Be Huge

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

14th Workshop on Hot Topics in Operating Systems (HotOS'13), May 2013.

The goal of this paper is to raise a new question: What changes in operating systems and networks if it were feasible to have a (type of) lookup table that supported billions, or hundreds of billions, of entries, using only a few bits per entry. We do so by showing that the progress of Moore's law, continuing to give more and more transistors per chip, makes it possible to apply formerly ludicrous amounts of brute-force parallel computation to find spacesavings opportunities.

We make two primary observations: First, that some applications can tolerate getting an incorrect answer from the table if they query for a key that is not in the table. For these applications, we can discard the keys entirely, using storage space only for the values. Further, for some applications, the value is not arbitrary. If the range of output values is small, we can instead view the problem as one of set separation. These two observations allow us to shrink the size of the mapping by brute force searching for a "perfect mapping" from inputs to outputs that (1) does not store the input keys; and (2) avoids collisions (and thus the related storage). Our preliminary results show that we can reduce memory consumption by an order of magnitude compared to traditional hash tables while providing competitive or better lookup performance.

Solving the Straggler Problem with Bounded Staleness

James Cipar, Qirong Ho, Jin Kyu Kim, Seunghak Lee, Gregory R. Ganger, Garth Gibson, Kimberly Keeton, Eric Xing

14th Workshop on Hot Topics in Operating Systems (HotOS'13), May 2013.

Many important applications fall into the broad class of iterative convergent algorithms. Parallel implementations of these algorithms are naturally expressed using the Bulk Synchronous Parallel (BSP) model of computation. However, implementations using BSP are plagued by the straggler problem, where every transient slowdown of any given thread can delay all other threads. This paper presents the Stale Synchronous Parallel (SSP) model as a generalization of BSP that preserves many of its advantages, while avoiding the straggler problem. Algorithms using SSP can execute efficiently, even with significant delays in some threads, addressing the oft-faced straggler problem.

Making Every Bit Count in Wide-Area Analytics

Ariel Rabkin, Matvey Arye, Siddhartha Sen, Vivek Pai, Michael J. Freedman

14th Workshop on Hot Topics in Operating Systems (HotOS'13), May 2013.

Many data sets, such as system logs, are generated from widely distributed locations. Current distributed systems often discard this data because they lack the ability to backhaul it efficiently, or to do anything meaningful with it at the distributed sites. This leads to

continued on pg. 18

Recent Publications

continued from pg. 17

lost functionality, efficiency, and business opportunities. The problem with traditional backhaul approaches is that they are slow and costly, and require analysts to define the data they are interested in up-front. We propose a new architecture that stores data at the edge (i.e., near where it is generated) and supports rich real-time and historical queries on this data, while adjusting data quality to cope with the vagaries of wide-area bandwidth. In essence, this design transforms a distributed data collection system into a distributed data analysis system, where decisions about collection do not preclude decisions about analysis.

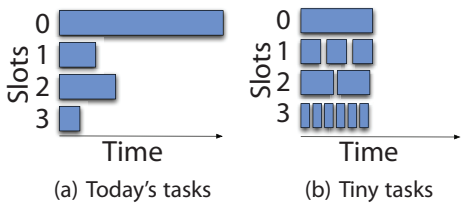
The Case for Tiny Tasks in Compute Clusters

Kay Ousterhout, Aurojit Panda, Joshua Rosen, Shivaram Venkataraman, Reynold Xin, Sylvia Ratnasamy, Scott Shenker, Ion Stoica

14th Workshop on Hot Topics in Operating Systems (HotOS'13), May 2013.

We argue for breaking data-parallel jobs in compute clusters into tiny tasks that each complete in hundreds of milliseconds. Tiny tasks avoid the need for complex skew mitigation techniques: by breaking a large job into millions of tiny tasks, work will be evenly spread over available resources by the scheduler. Furthermore, tiny tasks alleviate long wait times seen in today's clusters for interactive jobs: even large batch jobs can be split into small tasks that finish quickly. We demonstrate a 5.2x improvement in response times due to the use of smaller tasks.

In current data-parallel computing frameworks, high task launch overheads and scalability limitations prevent users from running short tasks. Recent research has addressed many of these bottlenecks; we discuss remaining challenges and propose a task execution framework that can efficiently support tiny tasks.



Tasks for a single job in a 4-slot cluster. With tiny tasks, work is allocated to machines at fine timegranularity, mitigating the effect of stragglers and allowing the job to complete more quickly.

HAT, Not CAP: Towards Highly Available Transactions

Peter Bailis, Alan Fekete, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica

14th Workshop on Hot Topics in Operating Systems (HotOS'13), May 2013.

While the CAP Theorem is often interpreted to preclude the availability of transactions in a partition-prone environment, we show that highly available systems can provide useful transactional semantics, often matching those of today's ACID databases. We propose Highly Available Transactions (HATs) that are available in the presence of partitions. HATs support many desirable ACID guarantees for arbitrary transactional sequences of read and write operations and permit low-latency operation.

Fairness and Isolation in Multi-Tenant Storage as Optimization Decomposition

David Shue, Michael J. Freedman, Anees Shaikh

ACM SIGOPS Operating System Review (OSR), 2013.

Shared storage services enjoy wide adoption in commercial clouds. But most systems today provide weak performance isolation and fairness between tenants, if at all. Most approaches to multi-tenant resource allocation are based either on per-VM allocations or hard rate limits that assume uniform workloads to achieve high utilization. Instead, Pisces, our system for shared key-value storage, achieves datacenter-wide per-tenant performance isolation and fairness.

Pisces achieves per-tenant weighted fair sharing of system resources across the entire shared service, even when partitions belonging to different tenants are co-located and when demand for different partitions is skewed or time-varying. The focus of this paper is to highlight the optimization model that motivates the decomposition of Pisces's fair sharing problem into four complementary mechanisms—partition placement, weight allocation, replica selection, and weighted fair queuing—that operate on different time-scales to provide system-wide max-min fairness. An evaluation of our Pisces storage prototype achieves nearly ideal (0.98 Min- Max Ratio) fair sharing, strong performance isolation, and robustness to skew and shifts in tenant demand.

PRObE: A Thousand-Node Experimental Cluster for Computer Systems Research

Garth Gibson, Gary Grider, Andree Jacobson, Wyatt Lloyd

USENIX ;login:, 2013.

If you have ever aspired to create a software system that can harness a thousand computers and perform some impressive feat, you know the dismal prospects of finding such a cluster ready and waiting for you to make magic with it. Today, however, if you are a systems researcher and your promised feat is impressive enough, there is such a resource available online: PRObE. This article is an introduction to and call for proposals for use of the PRObE facilities.

Performance Analysis of Network I/O Workloads in Virtualized Data Centers

Yiduo Mei, Ling Liu, Xing Pu, Sankaran Sivathanu, Xiaoshe Dong

IEEE Transactions on Service Computing, 2013.

Server consolidation and application consolidation through virtualization are key performance optimizations in cloud-based service delivery industry. In this paper, we argue that it is important for both cloud consumers and cloud providers to understand the various factors that may have significant impact on the performance of applications running in a virtualized cloud. This

paper presents an extensive performance study of network I/O workloads in a virtualized cloud environment. We first show that current implementation of virtual machine monitor (VMM) does not provide sufficient performance isolation to guarantee the effectiveness of resource sharing across multiple virtual machine instances (VMs) running on a single physical host machine, especially when applications running on neighboring VMs are competing for computing and communication resources. Then we study a set of representative workloads in cloud-based data centers, which compete for either CPU or network I/O resources, and present the detailed analysis on different factors that can impact the throughput performance and resource sharing effectiveness. For example, we analyze the cost and the benefit of running idle VM instances on a physical host where some applications are hosted concurrently. We also present an in-depth discussion on the performance impact of colocating applications that compete for either CPU or network I/O resources. Finally, we analyze the impact of different CPU resource scheduling strategies and different workload rates on the performance of applications running on different VMs hosted by the same physical machine.

Distance-Aware Bloom Filters: Enabling Collaborative Search for Efficient Resource Discovery

Yiming Zhang, Ling Liu

Future Generation Computer Systems, 2013.

Resource discovery in large-scale Peer-to-Peer (P2P) networks is challenging due to lack of effective methods for guiding queries. Based on the observation that the effectiveness of P2P resource discovery is determined by the utilization of hints, i.e., a summary of where the resources are, scattered in the network, in this paper we propose the distance-aware bloom filters (DABF) that disseminate hint information to faraway nodes by decaying BFs with different deterministic masks. Based on DABF, we design a novel Collaborative P2P Search (CPS) mechanism, which supports intelligent message behaviours including suspend, resume, terminate, move, reside, dispatch, notify and order. The effectiveness of our proposals is demonstrated through theoretical analysis and extensive simulations, in which we observed a remarkable reduction in search latency over previous approaches.

Computing Infrastructure for Big Data Processing

Ling Liu

Frontiers of Computer Science, 2013.

With computing systems transforming from single-processor devices to the ubiquitous and networked devices and the datacenter-scale computing in the cloud, the parallelism has become ubiquitous at many levels. At micro level, parallelisms are being explored from the underlying circuits, to pipelining and instruction level parallelism on multi-cores or many cores on a chip as well as in a machine. From macro level, parallelisms are being promoted from multiple machines on a rack, many racks in a data center, to the globally shared infrastructure of the Internet. With the push of big data, we are entering a new era of parallel computing driven by novel and ground breaking research innovation on elastic parallelism and scalability. In this article, we will give an overview of computing infrastructure for big data processing, focusing on architectural, storage and networking challenges of supporting big data analysis. We will briefly discuss emerging computing infrastructure and technologies that are promising for improving data parallelism, task parallelism and encouraging vertical and horizontal computation parallelism.

Error Analysis and Retention-Aware Error Management for NAND Flash Memory

Yu Cai, Gulay Yalcin, Onur Mutlu, Erich F. Haratsch, Adrian Cristal, Osman Unsal, Ken Mai

Intel Technology Journal (ITJ) Special Issue on Memory Resiliency, 2013.

With continued scaling of NAND flash memory process technology and multiple bits programmed per cell, NAND flash reliability and endurance are degrading. In our research, we experimentally measure, characterize, analyze, and model error patterns in nanoscale flash memories. Based on the understanding developed using real flash memory chips, we design techniques for more efficient and effective error management than traditionally used costly error correction codes.

In this article, we summarize our major error characterization results and mitigation techniques for NAND flash memory. We first provide a characterization of errors that occur in 30- to 40-nm flash memories, showing that retention errors, caused due to flash cells leaking charge over time, are the dominant source of errors. Second, we describe retention-aware error management techniques that aim to mitigate retention errors. The key idea is to periodically read, correct, and reprogram (in-place) or remap the stored data before it accumulates more retention errors than can be corrected by simple ECC. Third, we briefly touch upon our recent work that characterizes the distribution of the threshold voltages across different cells in a modern 20- to 24-nm flash memory, with the hope that such a characterization can enable the design of more effective and efficient error correction mechanisms to combat threshold voltage distortions that cause various errors. We conclude with a brief description of our ongoing related work in combating scaling challenges of both NAND flash memory and DRAM memory.

MemC3: Compact and Concurrent MemCache with Dumber Caching and Smarter Hashing

Bin Fan, David G. Andersen, Michael Kaminsky

10th Symposium on Networked Systems Design and Implementation (NSDI'13), April 2013.

This paper presents a set of architecturally and workload-inspired algorith-

continued on pg. 20

Recent Publications

continued from pg. 19

mic and engineering improvements to the popular Memcached system that substantially improve both its memory efficiency and throughput. These techniques—optimistic cuckoo hashing, a compact LRU-approximating eviction algorithm based upon CLOCK, and comprehensive implementation of optimistic locking—enable the resulting system to use 30% less memory for small key-value pairs, and serve up to 3x as many queries per second over the network. We have implemented these modifications in a system we call MemC3—Memcached with CLOCK and Concurrent Cuckoo hashing—but believe that they also apply more generally to many of today's read-intensive, highly concurrent networked storage and caching systems.

Stronger Semantics for Low-Latency Geo-Replicated Storage

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

10th Symposium on Networked Systems Design and Implementation (NSDI'13), April 2013.

We present the first scalable, geo-replicated storage system that guarantees low latency, offers a rich data model, and provides "stronger" semantics. Namely, all client requests are satisfied in the local datacenter in which they arise; the system efficiently supports useful data model abstractions such as column families and counter columns; and clients can access data in a causal-lyconsistent fashion with read-only and write-only transactional support, even for keys spread across many servers.

The primary contributions of this work are enabling scalable causal consistency for the complex columnfamily data model, as well as novel, non-blocking algorithms for both read-only and write-only transactions. Our evaluation shows that our system, Eiger, achieves low latency (single-ms), has throughput competitive with eventually-consistent and non-transactional Cassandra (less than 7% overhead for one of Facebook's real-world workloads), and scales out to large clusters almost linearly (averaging 96% increases up to 128 server clusters).

Effective Straggler Mitigation: Attack of the Clones

Ganesh Ananthanarayanan, Ali Ghodsi, Scott Shenker, Ion Stoica

10th Symposium on Networked Systems Design and Implementation (NSDI'13), April 2013.

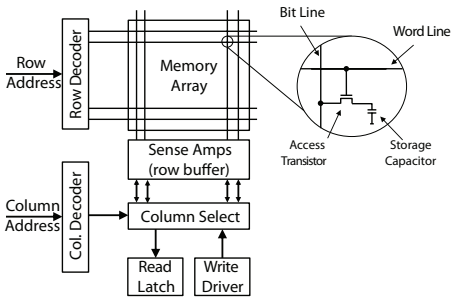
Small jobs, that are typically run for interactive data analyses in datacenters, continue to be plagued by disproportionately long-running tasks called stragglers. In the production clusters at Facebook and Microsoft Bing, even after applying state-of-the-art straggler mitigation techniques, these latency sensitive jobs have stragglers that are on average 8 times slower than themedian task in that job. Such stragglers increase the average job duration by 47%. This is because current mitigation techniques all involve an element of waiting and speculation. We instead propose full cloning of small jobs, avoiding waiting and speculation altogether. Cloning of small jobs only marginally increases utilization because workloads show that while the majority of jobs are small, they only consume a small fraction of the resources. The main challenge of cloning is, however, that extra clones can cause contention for intermediate data. We use a technique, delay assignment, which efficiently avoids such contention. Evaluation of our system, Dolly, using production workloads shows that the small jobs speedup by 34% to 46% after state-of-the-artmitigation techniques have been applied, using just 5% extra resources for cloning.

Evaluating STT-RAM as an Energy-Efficient Main Memory Alternative

Emre Kultursay, Mahmut Kandemir, Anand Sivasubramaniam, Onur Mutlu

IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'13), April 2013.

In this paper, we explore the possibility of using STT-RAM technology to completely replace DRAM in main memory. Our goal is to make STT-RAM performance comparable to DRAM while



DRAM bank organization.

providing substantial power savings. Towards this goal, we first analyze the performance and energy of STTRAM, and then identify key optimizations that can be employed to improve its characteristics. Specifically, using partial write and row buffer write bypass, we show that STT-RAM main memory performance and energy can be significantly improved. Our experiments indicate that an optimized, equal capacity STTRAM main memory can provide performance comparable to DRAM main memory, with an average 60% reduction in main memory energy.

The Impact of Mobile Multimedia Applications on Data Center Consolidation

Kiryong Ha, Padmanabhan Pillai, Grace Lewis, Soumya Simanta, Sarah Clinch, Nigel Davies, Mahadev Satyanarayanan

IEEE International Conference on Cloud Engineering (IC2E'13), March 2013.

The convergence of mobile computing and cloud computing enables new multimedia applications that are both resource-intensive and interaction-intensive. For these applications, end-to-end network bandwidth and latency matter greatly when cloud resources are used to augment the computational power and battery life of a mobile device.We first present quantitative evidence that this crucial design consideration to meet interactive performance criteria limits data center consolidation. We then describe an architectural solution that is a seamless extension of today's cloud computing infrastructure.

Pisces achieves per-tenant weighted fair sharing of system resources across the entire shared service, even when partitions belonging to different tenants are co-located and when demand for different partitions is skewed or time-varying. The focus of this paper is to highlight the optimization model that motivates the decomposition of Pisces's fair sharing problem into four complementary mechanisms -- partition placement, weight allocation, replica selection, and weighted fair queuing -- that operate on different time-scales to provide system-wide max-min fairness. An evaluation of our Pisces storage prototype achieves nearly ideal (0:98 Min-Max Ratio) fair sharing, strong performance isolation, and robustness to skew and shifts in tenant demand.

Asymmetry-aware Execution Placement on Manycore Chips

Alexey Tumanov, Joshua Wise, Onur Mutlu, Gregory R. Ganger

In Proc. of the 3rd Workshop on Systems for Future Multicore Architectures (SFMA'13), EuroSys'13, Apr. 14-17, 2013, Prague, Czech Republic.

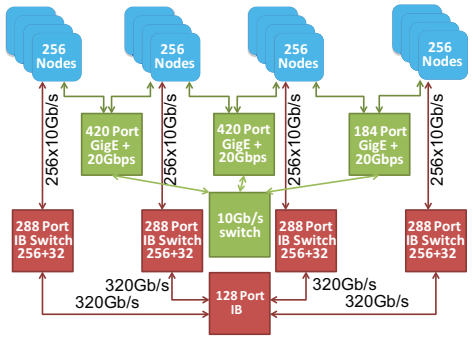
Network-on-chip based manycore systems with multiple memory controllers on a chip are gaining prevalence. Among other research considerations, placing an increasing number of cores on a chip creates a type of resource access asymmetries that didn't exist before. A common assumption of uniform or hierarchical memory controller access no longer holds. In this paper, we report on our experience with memory access asymmetries in a real manycore processor, the implications and extent of the problem they pose, and one potential thread placement solution that mitigates them. Our user-space scheduler harvests memory controller usage information generated in kernel space on a per process basis and enables thread placement decisions informed by threads' historical physical memory usage patterns. Results reveal a clear need for low-overhead, per-process memory controller hardware counters and show improved benchmark and application performance with a memory controller usage-aware execution placement policy.

PRObE: A Thousand-Node Experimental Cluster for Computer Systems Research

Garth Gibson, Gary Grider, Andree Jacobson, Wyatt Lloyd

USENIX ;login:, 2013.

If you have ever aspired to create a software system that can harness a thousand computers and perform some impressive feat, you know the dismal prospects of finding such a cluster ready and waiting for you to make magic with it. Today, however, if you are a systems researcher and your promised feat is impressive enough, there is such a resource available online: PRObE. This article is an introduction to and call for proposals for use of the PRObE facilities.



Block diagram of PRObE's Kodiak cluster

Extracting Useful Computation From Error-Prone Processors for Streaming Applications

Yavuz Yetim, Margaret Martonosi, Sharad Malik

Design, Automation & Test in Europe Conference (DATE'13), March 2013.

As semiconductor fabrics scale closer to fundamental physical limits, their reliability is decreasing due to process variation, noise margin effects, aging effects, and increased susceptibility to soft errors. Reliability can be regained through redundancy, error checking with recovery, voltage scaling and other means, but these techniques impose area/energy costs. Since some applications (e.g. media) can tolerate limited computation errors and still provide useful results, error-tolerant computation models have been explored, with both the application and computation fabric having stochastic characteristics. Stochastic computation has, however, largely focused on application-specific hardware solutions, and is not general enough to handle arbitrary bit errors that impact memory addressing or control in processors.

In response, this paper addresses requirements for errortolerant execution by proposing and evaluating techniques for running error-tolerant software on a general-purpose processor built from an unreliable fabric. We study the minimum errorprotection required, from a microarchitecture perspective, to still produce useful results at the application output. Even with random errors as frequent as every 250μs, our proposed design allows JPEG and MP3 benchmarks to sustain good output quality—14dB and 7dB respectively. Overall, this work establishes the potential for error-tolerant single-threaded execution, and details its required hardware/system support.

Performance Analysis of Network I/O Workloads in Virtualized Data Centers

Yiduo Mei, Ling Liu, Xing Pu, Sankaran Sivathanu, Xiaoshe Dong

IEEE Transactions on Service Computing, 2013.

Server consolidation and application consolidation through virtualization are key performance optimizations in cloud-based service delivery industry. In this paper, we argue that it is important for both cloud consumers and cloud providers to understand the various factors that may have significant impact on the performance of applications running in a virtualized cloud. This paper presents an extensive performance study of network I/O workloads in a virtualized cloud environment. We first show that current implementation of virtual machine monitor (VMM) does not provide sufficient performance isolation to guarantee the effectiveness of resource sharing across multiple vir-

continued on pg. 22

Recent Publications

continued from pg. 21

tual machine instances (VMs) running on a single physical host machine, especially when applications running on neighboring VMs are competing for computing and communication resources. Then we study a set of representative workloads in cloud-based data centers, which compete for either CPU or network I/O resources, and present the detailed analysis on different factors that can impact the throughput performance and resource sharing effectiveness. For example, we analyze the cost and the benefit of running idle VM instances on a physical host where some applications are hosted concurrently. We also present an in-depth discussion on the performance impact of colocating applications that compete for either CPU or network I/O resources. Finally, we analyze the impact of different CPU resource scheduling strategies and different workload rates on the performance of applications running on different VMs hosted by the same physical machine.

Distance-Aware Bloom Filters: Enabling Collaborative Search for Efficient Resource Discovery

Yiming Zhang, Ling Liu

Future Generation Computer Systems, 2013.

Resource discovery in large-scale Peer-to-Peer (P2P) networks is challenging due to lack of effective methods for guiding queries. Based on the observation that the effectiveness of P2P resource discovery is determined by the utilization of hints, i.e., a summary of where the resources are, scattered in the network, in this paper we propose the distance-aware bloom filters (DABF) that disseminate hint information to faraway nodes by decaying BFs with different deterministic masks. Based on DABF, we design a novel Collaborative P2P Search (CPS) mechanism, which supports intelligent message behaviours including suspend, resume, terminate, move, reside, dispatch, notify and order. The effectiveness of our proposals is demonstrated through theoretical analysis and extensive simulations, in which we observed a remarkable reduction in search latency over previous approaches.

Computing Infrastructure for Big Data Processing

Ling Liu

Frontiers of Computer Science, 2013.

With computing systems transforming from single-processor devices to the ubiquitous and networked devices and the datacenter-scale computing in the cloud, the parallelism has become ubiquitous at many levels. At micro level, parallelisms are being explored from the underlying circuits, to pipelining and instruction level parallelism on multi-cores or many cores on a chip as well as in a machine. From macro level, parallelisms are being promoted from multiple machines on a rack, many racks in a data center, to the globally shared infrastructure of the Internet. With the push of big data, we are entering a new era of parallel computing driven by novel and ground breaking research innovation on elastic parallelism and scalability. In this article, we will give an overview of computing infrastructure for big data processing, focusing on architectural, storage and networking challenges of supporting big data analysis. We will briefly discuss emerging computing infrastructure and technologies that are promising for improving data parallelism, task parallelism and encouraging vertical and horizontal computation parallelism.

Ligra: A Lightweight Graph Processing Framework for Shared-Memory

Julian Shun,Guy Blelloch

18th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP'13), February 2013.

There has been significant recent interest in parallel frameworks for processing graphs due to their applicability in studying social networks, the Web graph, networks in biology, and unstructured meshes in scientific simulation. Due to the desire to process large graphs, these systems have emphasized the ability to run on distributed memory machines. Today, however, a single multicore server can support more than a terabyte of memory, which can fit graphs with tens or even hundreds of billions of edges. Furthermore, for graph algorithms, shared-memory multicores are generally significantly more efficient on a per core, per dollar, and per joule basis than distributed memory systems, and shared-memory algorithms tend to be simpler than their distributed counterparts.

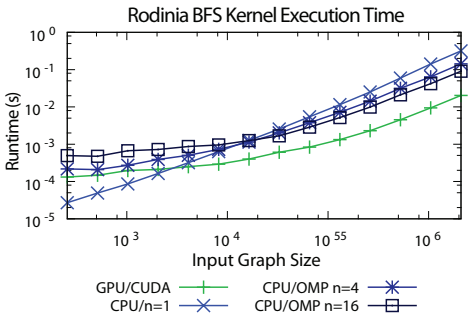
In this paper, we present a lightweight graph processing framework that is specific for shared-memory parallel/multicore machines, which makes graph traversal algorithms easy to write. The framework has two very simple routines, one for mapping over edges and one for mapping over vertices. Our routines can be applied to any subset of the vertices, which makes the framework useful for many graph traversal algorithms that operate on subsets of the vertices. Based on recent ideas used in a very fast algorithm for breadth-first search (BFS), our routines automatically adapt to the density of vertex sets. We implement several algorithms in this framework, including BFS, graph radii estimation, graph connectivity, betweenness centrality, PageRank and single-source shortest paths. Our algorithms expressed using this framework are very simple and concise, and perform almost as well as highly optimized code. Furthermore, they get good speedups on a 40-core machine and are significantly more efficient than previously reported results using graph frameworks on machines with many more cores.

Reducing GPU Offload Latency Via Fine-Grained CPU-GPU Synchronization

Dan Lustig, Margaret Martonosi

International Symposium on High-Performance Computer Architecture (HPCA'13), February, 2013.

GPUs are seeing increasingly widespread use for general purpose computation due to their excellent performance for highly-parallel, throughput-oriented applications. For many workloads, however, the performance benefits of offloading are hindered by the large and unpredictable



Runtime of breadth-first search [8] for varying input set sizes and hardware. The GPU is a discrete NVIDIA GTX 580 running CUDA and the CPU is an Intel Xeon X7560 running sequential code or a multithreaded OpenMP version. As input size changes, the type of hardware providing the fastest runtime changes: a single-threaded CPU is fastest for smallest inputs, and the discrete NVIDIA GPU is fastest at large sizes.

overheads of launching GPU kernels and of transferring data between CPU and GPU.

This paper proposes and evaluates hardware and software support for reducing overheads and improving data latency predictability when offloading computation to GPUs. We first characterize program execution using real-system measurements to highlight the degree to which kernel launch and data transfer are major sources of overhead. We then propose a scheme of full-empty bits to track when regions of data have been transferred. This dependency tracking is fast, efficient, and fine-grained, mitigating much of the latency uncertainty and cost of offloading in current systems. On top of these fullempty bits, we build APIs that allow for early kernel launch and proactive data returns. These techniques enable faster kernel completion, while correctness remains guaranteed by the full/empty bits.

Taken together, these techniques can both greatly improve GPU application performance and broaden the space of applications for which GPUs are beneficial. In particular, across a set of seven diverse benchmarks that make use of our support, the mean improvement in runtime is 26%.

MISE: Providing Performance Predictability and Improving Fairness in Shared Main Memory Systems

Lavanya Subramanian, Vivek Seshadri, Yoongu Kim, Ben Jaiyen, Onur Mutlu

International Symposium on High-Performance Computer Architecture (HPCA'13), February, 2013.

Applications running concurrently on a multicore system interfere with each other at the main memory. This interference can slow down diUerent applications diUerently. Accurately estimating the slowdown of each application in such a system can enable mechanisms that can enforce quality-of-service. While much prior work has focused on mitigating the performance degradation due to inter-application interference, there is little work on estimating slowdown of individual applications in a multi-programmed environment. Our goal in this work is to build such an estimation scheme.

To this end, we present our simple Memory-Interferenceinduced Slowdown Estimation (MISE) model that estimates slowdowns caused by memory interference. We build our model based on two observations. First, the performance of a memorybound application is roughly proportional to the rate at which its memory requests are served, suggesting that request-servicerate can be used as a proxy for performance. Second, when an application's requests are prioritized over all other applications' requests, the application experiences very little interference from other applications. This provides a means for estimating the uninterfered request-service-rate of an application while it is run alongside other applications. Using the above observations, our model estimates the slowdown of an application as the ratio of its uninterfered and interfered request service rates. We propose simple changes to the above model to estimate the slowdown of non-memory-bound applications.

We demonstrate the eUectiveness of our model by developing two new memory scheduling schemes: 1) one that provides soft quality-of-service guarantees and 2) another that explicitly attempts to minimize maximum slowdown (i.e., unfairness) in the system. Evaluations show that our techniques perform signi Vcantly better than state-of-the-art memory scheduling approaches to address the above problems.

Application-to-Core Mapping Policies to Reduce Memory System Interference in Multi-Core Systems

Reetuparna Das, Rachata Ausavarungnirun, Onur Mutlu, Akhilesh Kumar, Mani Azimi

International Symposium on High-Performance Computer Architecture (HPCA'13), February, 2013.

Future many-core processors are likely to concurrently execute a large number of diverse applications. How these ap- plications are mapped to cores largely determines the interference between these applications in critical shared hardware resources. This paper proposes new application-to-core mapping policies to improve system performance by reducing inter-application interference in the on-chip network and memory controllers. The major new ideas of our policies are to: 1) map network-latency-sensitive applications to separate parts of the network from network-bandwidth-intensive applications such that the former can make fast progress without heavy interference from the latter, 2) map those applications that benefit more from being closer to the memory controllers close to these resources.

Our evaluations show that, averaged over 128 multiprogrammed workloads of 35 different benchmarks running on a 64-core system, our final application-to-core mapping policy improves system throughput by 16.7% over a state-of-the-art baseline, while also reducing system unfairness by 22.4% and average interconnect power consumption by 52.3%.

continued on pg. 24

Recent Publications

continued from pg. 23

Lowering Barriers to Large-scale Mobile Crowdsensing

Yu Xiao, Pieter Simoens, Padmanabhan Pillai, Kiryong Ha, Mahadev Satyanarayanan

14th ACM International Workshop on Mobile Computing Systems and Applications (HotMobile'13), February 2013.

Mobile crowdsensing is becoming a vital technique for environment monitoring, infrastructure management, and social computing. However, deploying mobile crowdsensing applications in large-scale environments is not a trivial task. It creates a tremendous burden on application developers as well as mobile users. In this paper we try to reveal the barriers hampering the scale-up of mobile crowdsensing applications, and to offer our initial thoughts on the potential solutions to lowering the barriers.

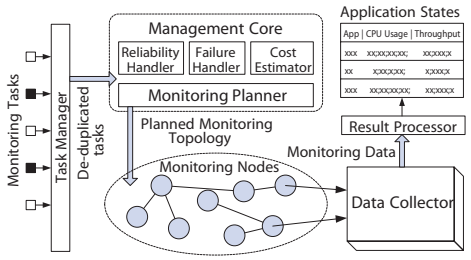
Resource-Aware Application State Monitoring

Shicong Meng, Srinivas R. Kashyap, Chitra Venkatramani, Ling Liu

IEEE Transactions on Parallel and Distributed Systems (TPDS'13), December 2012.

The increasing popularity of large-scale distributed applications in datacenters has led to the growing demand of distributed application state monitoring. These application state monitoring tasks often involve collecting values of various status attributes from a large number of nodes. One challenge in such large-scale application state monitoring is to organize nodes into a monitoring overlay that achieves monitoring scalability and cost-effectiveness at the same time.

In this paper, we present REMO, a REsource-aware application state MOnitoring system, to address the challenge of monitoring overlay construction. REMO distinguishes itself from existing works in several key aspects. First, it jointly considers inter-task costsharing opportunities and node-level resource constraints. Furthermore, it explicitly models the per-message processing overhead which can be substantial but is often ignored by previous works. Second, REMO produces a forest of optimized monitoring trees through iterations of two phases. One phase explores cost-sharing opportunities between tasks, and the other refines the tree with resource-sensitive construction schemes. Finally, REMO also employs an adaptive algorithm that balances the benefits and costs of overlay adaptation. This is particularly useful for large systems with constantly changing monitoring tasks. Moreover, we enhance REMO in terms of both performance and applicability with a series of optimization and extension techniques. We perform extensive experiments including deploying REMO on a BlueGene/P rack running IBM's large-scale distributed streaming system - System S. Using REMO in the context of collecting over 200 monitoring tasks for an application deployed across 200 nodes results in a 35%-45% decrease in the percentage error of collected attributes compared to existing schemes.



A High-Level System Model

Elastic Resource Allocation in Datacenters: Gremlins in the Management Plane

Mukil Kesavan, Ada Gavrilovska, Karsten Schwan

VMware Technical Journal, December 2012.

Virtualization has simplified the management of datacenter infrastructures and enabled new services that can benefit both customers and providers. From a provider perspective, one of the key services in a virtualized datacenter is elastic allocation of resources to workloads, using a combination of virtual machine migration and per-server work-conserving scheduling. Known challenges to elastic resource allocation include scalability, hardware heterogeneity, hard and soft virtual machine placement constraints, resource partitions, and others. This paper describes an additional challenge, which is the need for IT management to consider two design constraints that are systemic to large-scale deployments: failures in management operations and high variability in cost. The paper first illustrates these challenges, using data collected from a 700-server datacenter running a hierarchical resource management system built on the VMware vSphere platform. Next, it articulates and demonstrates methods for dealing with cost variability and failures, with a goal of improving management effectiveness. The methods make dynamic tradeoffs between management accuracy compared to overheads, within constraints imposed by observed failure behavior and cost variability.

Theia: Visual Signatures for Problem Diagnosis in Large Hadoop Clusters

Elmer Garduno, Soila P. Kavulya, Jiaqi Tan, Rajeev Gandhi, Priya Narasimhan

USENIX Large Installation System Administration Conference (LISA'12), December 2012. Best Student Paper Award.

Diagnosing performance problems in large distributed systems can be daunting as the copious volume of monitoring information available can obscure the root-cause of the problem. Automated diagnosis tools help narrow down the possible root-causes—however, these tools are not perfect thereby motivating the need for visualization tools that allow users to explore their data and gain insight on the root-cause. In this paper we describe Theia, a visualization tool that analyzes application-level logs in a Hadoop cluster, and generates visual signatures of each job's performance. These visual signatures provide compact representations of task durations,

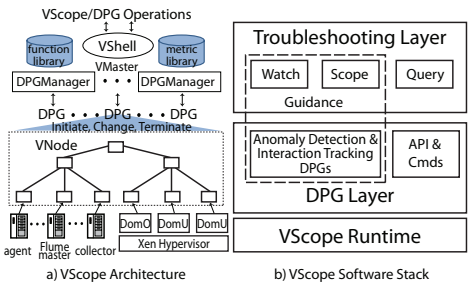
task status, and data consumption by jobs. We demonstrate the utility of Theia on real incidents experienced by users on a production Hadoop cluster.

VScope: Middleware for Troubleshooting Time-Sensitive Data Center Applications

Chengwei Wang, Infantdani Abel Rayan, Greg Eisenhauer, Karsten Schwan, Vanish Talwar, Matthew Wolf, Chad Huneycutt

ACM Middleware (Middleware'12), December 2012.

Data-Intensive infrastructures are increasingly used for on-line processing of live data to guide operations and decision making. VScope is a exible monitoring and analysis middleware for troubleshooting such large-scale, time-sensitive, multi-tier applications. With VScope, lightweight anomaly detection and interaction tracking methods can be run continuously throughout an application's execution. The runtime events generated by these methods can then initiate more detailed and heavier weight analyses which are dynamically deployed in the places where they may be most likely fruitful for root cause diagnosis and mitigation. We comprehensively evaluate VScope prototype in a virtualized data center environment with over 1000 virtual machines (VMs), and apply VScope to a representative on-line log processing application. Experimental results show that VScope can deploy and operate a variety of on-line analytics functions and metrics with a few seconds at large scale. Compared to traditional logging approaches, VScope based troubleshooting has substantially lower perturbation and generates much smaller log data volumes. It can also resolve complex cross-tier or cross-software-level issues unsolvable solely by application-level or per-tier mechanisms.



VScope System Design

TABLEFS: Embedding a NoSQL Database inside the Local File System

Kai Ren, Garth Gibson

1st Storage System, Hard Disk and Solid State Technologies Summit, IEEE Asia-Pacific Magnetic Recording Conference (APMRC), November 2012.

File systems that manage magnetic disks have long recognized the importance of sequential allocation and large transfer sizes for file data. Fast random access has dominated metadata lookup data structures with increasingly use of B-trees on-disk. For updates, on-disk data structures are increasingly non-overwrite, copy-on-write, log-like and deferred. Yet our experiments with workloads dominated by metadata and small file access indicate that even sophisticated local disk file systems like Ext4, XFS and BTRFS leaves a lot of opportunity for performance improvement in workloads dominated by metadata and small files.

In this paper we present a simple stacked file system, TableFS, which uses another local file system as an object store and organizes all metadata into a single sparse table backed on-disk using a Log-Structured Merge (LSM) tree, LevelDB in our experiments. By stacking, TableFS asks only for efficient large file allocation and access from the local file system. By using an LSM tree, TableFS ensures metadata is written to disk in large, non-overwrite, sorted and indexed logs, and inherits a compaction algorithm. Even an inefficient FUSE based user level implementation of TableFS can perform comparably to Ext4, XFS and BTRFS on simple dataintensive benchmarks, and can outperform them by 50% to as much as 1000% for a metadata-intensive query/update workload on data-free files. Such promising performance results from TableFS suggest that local disk file systems can be significantly improved by much more aggressive aggregation and batching of metadata updates.

Performance Isolation and Fairness for Multi-Tenant Cloud Storage

David Shue, Michael J. Freedman, Anees Shaikh

Proc. Symposium on Operating Systems Design and Implementation (OSDI '12), October 2012.

Shared storage services enjoy wide adoption in commercial clouds. But most systems today provide weak performance isolation and fairness between tenants, if at all. Misbehaving or high-demand tenants can overload the shared service and disrupt other well-behaved tenants, leading to unpredictable performance and violating SLAs.

This paper presents Pisces, a system for achieving datacenter-wide per-tenant performance isolation and fairness in shared key-value storage. Today's approaches for multi-tenant resource allocation are based either on per-VM allocations or hard rate limits that assume uniform workloads to achieve high utilization. Pisces achieves per-tenant weighted fair shares (or minimal rates) of the aggregate resources of the shared service, even when different tenants' partitions are co-located and when demand for different partitions is skewed, time-varying, or bottlenecked by different server resources. Pisces does so by decomposing the fair sharing problem into a combination of four complementary mechanisms—partition placement, weight allocation, replica selection, and weighted fair queuing—that operate on different time-scales and combine to provide system-wide max-min fairness.

An evaluation of our Pisces storage prototype achieves nearly ideal (0.99 Min-Max Ratio) weighted fair sharing, strong performance isolation, and robustness to skew and shifts in tenant demand. These properties are achieved with minimal overhead (<3%), even when running at high utilization (more than 400,000 requests/second/server for 10B requests).

continued on pg. 26

Recent Publications

continued from pg. 25

A Case for Scaling HPC Metadata Performance through De-specialization

Swapnil Patil, Kai Ren, Garth Gibson

Proc. of the Seventh Parallel Data Storage Workshop (PDSW12), co-located with the Int. Conference for High Performance Computing, Networking, Storage and Analysis (SC12), November 2012.

We envision a scalable metadata service with two goals. The first goal – evolution, not revolution – emphasizes the need for a solution that adds new support to existing cluster file systems that lack a scalable metadata path. Although newer cluster file systems, including Google's Colossus file system [9], OrangeFS [16], UCSC's Ceph [27] and Copernicus [12], promise a distributed metadata service, it is undesirable to replace existing cluster file systems running in large production environments just because their metadata path does not provide the desired scalability or the desired functionality. Several large cluster file system installations, such as Panasas PanFS running at LANL [28] and PVFS running on Argonne BG/P [1], [21], can benefit from a solution that provides, for instance, distributed directory support that does not require any modifications to the running cluster file system. The second goal – generality and de-specialization – promises a fully, distributed and scalable metadata service that performs well for ingest, lookups, and scans. In particular, all metadata, including directory entries, i-nodes and block management, should be stored in one structure; this is different from today's file systems that use specialized on-disk structures for each type of metadata.

SOFTScale: Stealing Opportunistically For Transient Scaling

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

ACM Middleware (Middleware'12), December 2012.

Dynamic capacity provisioning is a well studied approach to handling gradual changes in data center load. However, abrupt spikes in load are still problematic in that the work in the system rises very quickly during the setup time needed to turn on additional capacity. Performance can be severely affected even if it takes only 5 seconds to bring additional capacity online.

In this paper, we propose SOFTScale, an approach to handling load spikes in multi-tier data centers without having to over-provision resources. SOFTScale works by opportunistically stealing resources from other tiers to alleviate the bottleneck tier, even when the tiers are carefully provisioned at capacity. SOFTScale is especially useful during the transient overload periods when additional capacity is being brought online.

Via implementation on a 28-server multi-tier testbed, we investigate a range of possible load spikes, including an artificial doubling or tripling of load, as well as large spikes in real traces. We find that SOFTScale can meet our stringent 95th percentile response time Service Level Agreement goal of 500ms without using any additional resources even under some extreme load spikes that would normally cause the system (without SOFTScale) to exhibit response times as high as 96 seconds.

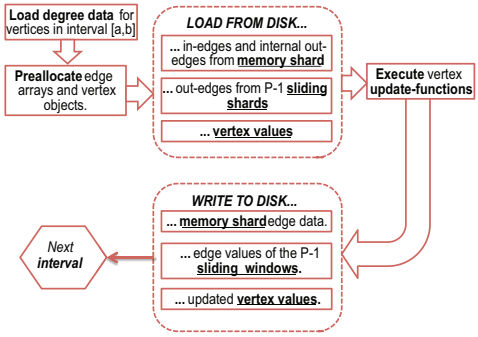
GraphChi: Large-Scale Graph Computation on Just a PC

Aapo Kyrola, Guy Blelloch, Carlos Guestrin

Symposium on Operating Systems Design and Implementation (OSDI'12), October 2012.

Current systems for graph computation require a distributed computing cluster to handle very large real-world problems, such as analysis on social networks or the web graph. While distributed computational resources have become more accessible, developing distributed graph algorithms still remains challenging, especially to non-experts.

In this work, we present GraphChi, a disk-based system for computing efficiently on graphs with billions of edges. By using a well-known method to break large graphs into small parts, and a novel parallel sliding windows method, GraphChi is able to execute several advanced data mining, graph mining, and machine learning algorithms on very large graphs, using just a single consumer-level computer. We further extend GraphChi to support graphs that evolve over time, and demonstrate that, on a single computer, GraphChi can process over one hundred thousand graph updates per second, while simultaneously performing computation. We show, through experiments and theoretical analysis, that GraphChi performs well on both SSDs and rotational hard drives.



Main execution flow. Sequence of operations for processing one execution interval with GraphChi.

By repeating experiments reported for existing distributed systems, we show that, with only fraction of the resources, GraphChi can solve the same problems in very reasonable time. Our work makes large-scale graph computation available to anyone with a modern PC.

AutoScale: Dynamic, Robust Capacity Management for Multi-Tier Data Centers

Anshul Gandhi, Mor Harchol-Balter, Ram Raghunathan, Mike Kozuch

ACM Transactions on Computer Systems (TOCS) vol. 30, No. 4, Article 14, 2012.

Energy costs for data centers continue to rise, already exceeding $15 billion yearly. Sadly much of this power is wasted. Servers are only busy 10–30% of the time on average, but they are often left on, while idle, utilizing 60%

or more of peak power when in the idle state. We introduce a dynamic capacity management policy, AutoScale, that greatly reduces the number of servers needed in data centers driven by unpredictable, time-varying load, while meeting response time SLAs. AutoScale scales the data center capacity, adding or removing servers as needed. AutoScale has two key features: (i) it autonomically maintains just the right amount of spare capacity to handle bursts in the request rate; and (ii) it is robust not just to changes in the request rate of real-world traces, but also request size and server efficiency. We evaluate our dynamic capacity management approach via implementation on a 38-server multi-tier data center, serving a web site of the type seen in Facebook or Amazon, with a key-value store workload. We demonstrate that AutoScale vastly improves upon existing dynamic capacity management policies with respect to meeting SLAs and robustness.

PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs

Joseph E. Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, Carlos Guestrin

Symposium on Operating Systems Design and Implementation (OSDI'12), October 2012.

Large-scale graph-structured computation is central to tasks ranging from targeted advertising to natural language processing and has led to the development of several graph-parallel abstractions including Pregel and GraphLab. However, the natural graphs commonly found in the real-world have highly skewed power-law degree distributions, which challenge the assumptions made by these abstractions, limiting performance and scalability.

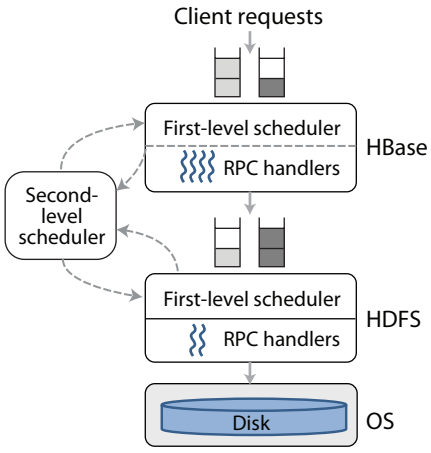
In this paper, we characterize the challenges of computation on natural graphs in the context of existing graphparallel abstractions. We then introduce the PowerGraph abstraction which exploits the internal structure of graph programs to address these challenges. Leveraging the PowerGraph abstraction we introduce a new approach to distributed graph placement and representation that exploits the structure of power-law graphs. We provide a detailed analysis and experimental evaluation comparing PowerGraph to two popular graph-parallel systems. Finally, we describe three different implementation strategies for PowerGraph and discuss their relative merits with empirical evaluations on large-scale real-world problems demonstrating order of magnitude gains.

Cake: Enabling High-level SLOs on Shared Storage Systems

A. Wang, S. Venkataraman, S. Alspaugh, R. H. Katz, I. Stoica

ACM Symposium on Cloud Computing (SOCC'12), October 2012.

Cake is a coordinated, multi-resource scheduler for shared distributed storage environments with the goal of achieving both high throughput and bounded latency. Cake uses a two-level scheduling scheme to enforce high-level service-level objectives (SLOs). Firstlevel schedulers control consumption of resources such as disk and CPU. These schedulers (1) provide mechanisms for differentiated scheduling, (2) split large requests into smaller chunks, and (3) limit the number of outstanding device requests, which together allow for effective control over multi-resource consumption within the storage system. Cake's second-level scheduler coordinates the first-level schedulers to map high-level SLO requirements into actual scheduling parameters. These parameters are dynamically adjusted over time to enforce high-level performance specifications for changing workloads. We evaluate Cake using multiple workloads derived from real-world traces. Our results show that Cake allows application programmers to explore the latency vs. throughput trade-off by setting different high-level performance requirements on their workloads. Furthermore, we show that using Cake has concrete economic and business advantages, reducing provisioning costs by up to 50% for a consolidated workload and reducing the completion time of an analytics cycle by up to 40%.



Architecture of the Cake software stack on a single storage node. Cake adds first-level schedulers to the RPC layers of HBase and HDFS. The first-level schedulers are coordinated by Cake's SLO-aware second-level scheduler.

Heterogeneity and Dynamicity of Clouds at Scale: Google Trace

C. Reiss, A. Tumanoy, G. R. Ganger, R. H. Katz, M. A. Kozuch

ACM Symposium on Cloud Computing (SOCC'12), October 2012.

To better understand the challenges in developing effective cloudbased resource schedulers, we analyze the first publicly available trace data from a sizable multi-purpose cluster. The most notable workload characteristic is heterogeneity: in resource types (e.g., cores:RAM per machine) and their usage (e.g., duration and resources needed). Such heterogeneity reduces the effectiveness of traditional slot- and core-based scheduling. Furthermore, some tasks are constrained as to the kind of machine types they can use, increasing the complexity of resource assignment and complicating task migration. The workload is also highly dynamic, varying over time and most workload features, and is driven by many short jobs that demand quick scheduling decisions. While few simplifying assumptions apply, we find that many longer-running jobs have relatively stable resource utilizations, which can help adaptive resource schedulers.

continued on pg. 28

Recent Publications

continued from pg. 27

Using Vector Interfaces to Deliver Millions of IOPS from a Networked Key-value Storage Server

Vijay Vasudevan, Michael Kaminsky, David Andersen

ACM Symposium on Cloud Computing (SOCC'12), October 2012.

The performance of non-volatile memories (NVM) has grown by a factor of 100 during the last several years: Flash devices today are capable of over 1 million I/Os per second. Unfortunately, this incredible growth has put strain on software storage systems looking to extract their full potential.

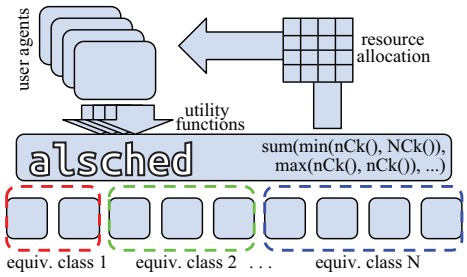
To address this increasing software-I/O gap, we propose using vector interfaces in high-performance networked systems. Vector interfaces organize requests and computation in a distributed system into collections of similar but independent units of work, thereby providing opportunities to amortize and eliminate the redundant work common in many high-performance systems. By integrating vector interfaces into storage and RPC components, we demonstrate that a single key-value storage server can provide 1.6 million requests per second with a median latency below one millisecond, over fourteen times greater than the same software absent the use of vector interfaces. We show that pervasively applying vector interfaces is necessary to achieve this potential and describe how to compose these interfaces together to ensure that vectors of work are propagated throughout a distributed system.

alsched: Algebraic Scheduling of Mixed Workloads in Heterogeneous Clouds

Alexey Tumanov, James Cipar, Michael Kozuch, Gregory Ganger

ACM Symposium on Cloud Computing (SOCC'12), October 2012.

As cloud resources and applications grow more heterogeneous, allocating the right resources to different tenants' activities increasingly depends upon understanding tradeoffs regarding their individual behaviors. One may require a specific amount of RAM, another may benefit from a GPU, and a third may benefit from executing on the same rack as a fourth. This paper promotes the need for and an approach for accommodating diverse tenant needs, based on having resource requests indicate any soft (i.e., when certain resource types would be better, but are not mandatory) and hard constraints in the form of composable utility functions. A scheduler that accepts such requests can then maximize overall utility, perhaps weighted by priorities, taking into account application specifics. Experiments with a prototype scheduler, called alsched, demonstrate that support for soft constraints is important for efficiency in multi-purpose clouds and that composable utility functions can provide it.



alsched System Model

The Potential Dangers of Causal Consistency and an Explicit Solution

Peter Bailis, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica

ACM Symposium on Cloud Computing (SOCC'12), October 2012.

Causal consistency is the strongest consistency model that is available in the presence of partitions and provides useful semantics for human-facing distributed services. Here, we expose its serious and inherent scalability limitations due to write propagation requirements and traditional dependency tracking mechanisms. As an alternative to classic potential causality, we advocate the use of explicit causality, or applicationdefined happens-before relations. Explicit causality, a subset of potential causality, tracks only relevant dependencies and reduces several of the potential dangers of causal consistency.

True Elasticity in Multi-Tenant Clusters through Amoeba

Ganesh Anantharanayanan, Christopher Douglas, Raghu Ramakrishnan, Sriram Rao, Ion Stoica

ACM Symposium on Cloud Computing (SOCC'12), October 2012.

Data-intensive computing (DISC) frameworks scale by partitioning a job across a set of fault-tolerant tasks, then diffus- ing those tasks across large clusters. Multi-tenanted clusters must accommodate service-level objectives (SLO) in their resource model, often expressed as a maximum latency for allocating the desired set of resources to every job. When jobs are partitioned into tasks statically, a cluster cannot meet its SLOs while maintaining both high utilization and efficiency. Ideally, we want to give resources to jobs when they are free but would expect to reclaim them instanta- neously when new jobs arrive, without losing work. DISC frameworks do not support such elasticity because interrupting running tasks incurs high overheads. Amoeba enables lightweight elasticity in DISC frameworks by identifying points at which running tasks of over-provisioned jobs can be safely exited, committing their outputs, and spawning new tasks for the remaining work. Effectively, tasks of DISC jobs are now sized dynamically in response to global resource scarcity or abundance. Simulation and deployment of our prototype shows that Amoeba speeds up jobs by 32% without compromising utilization or efficiency.

Net-Cohort: Detecting and Managing VM Ensembles in Virtualized Data Centers

Liting Hu, Karsten Schwan, Ajay Gulati, Junjie Zhang, Chengwei Wang

Proceedings of 2012 IEEE International Conference on Autonomic Computing (ICAC'12), September 2012.

Bi-section bandwidth is a critical re-

source in today's data centers because of the high cost and limited bandwidth of higher-level network switches and routers. This problem is aggravated in virtualized environments where a set of virtual machines, jointly implementing some service, may run across multiple L2 hops. Since data center administrators typically do not have visibility into such sets of communicating VMs, this can cause inter-VM traffic to traverse bottlenecked network paths. To address this problem, we present 'NetCohort', which offers lightweight system-level techniques to (1) discover VM ensembles and (2) collect information about intra-ensemble VM interactions. Net- Cohort can dynamically identify ensembles to manipulate entire services/applications rather than individual VMs, and to support VM placement engines in co-locating communicating VMs in order to reduce the consumption of bi-section bandwidth. An implementation of Net-Cohort on a Xen-based system with 15 hosts and 225 VMs shows that its methods can detect VM ensembles at low cost and with about 90.0% accuracy. Placements based on ensemble information provided by Net- Cohort can result in an up to 385% improvement in application throughput for a RUBiS instance, a 56.4% improvement in application throughput for a Hadoop instance, and a 12.76 times improvement in quality of service for a SIPp instance.

Project Hoover: Auto-Scaling Streaming Map-Reduce Applications

Rajalakshmi Ramesh, Liting Hu, Karsten Schwan

MDBS Workshop at ICAC Conference, September 2012.

Real-time data processing frameworks like S4 and Flume have become scalable and reliable solutions for acquiring, moving, and processing voluminous amounts of data continuously produced by large numbers of online sources. Yet these frameworks lack the elasticity to horizontally scale-up or scale-down their based on current rates of input events and desired event processing latencies. The Project Hoover middleware provides distributed methods for measuring, aggregating, and analyzing the performance of distributed Flume components, thereby enabling online configuration changes to meet varying processing demands. Experimental evaluations with a sample Flume data processing code show Hoover's approach to be capable of dynamically and continuously monitoring Flume performance, demonstrating that such data can be used to right-size the number of Flume collectors according to different log production rates.

Chrysalis Analysis: Incorporating Synchronization Arcs in Dataflow-Analysis-based Parallel Monitoring

Michelle Goodstein, Shimin Chen, Phillip B. Gibbons, Michael Kozuch, Todd Mowry

International Conference on Parallel Architectures and Compilation Techniques (PACT'12), September 2012.

Software lifeguards, or tools that monitor applications at runtime, are an effective way of identifying program errors and security exploits. Parallel programs are susceptible to a wider range of possible errors than sequential programs, making them even more in need of online monitoring. Unfortunately, monitoring parallel applications is difficult due to inter-thread data dependences. In prior work, we introduced a new software framework for online parallel program monitoring inspired by dataflow analysis, called Butterfly Analysis. Butterfly Analysis uses bounded windows of uncertainty to model the finite upper bound on delay between when an instruction is issued and when all its effects are visible throughout the system. While Butterfly Analysis offers many advantages, it ignored one key source of ordering information which affected its false positive rate: explicit software synchronization, and the corresponding high-level happens-before arcs.

In this work we introduce Chrysalis Analysis, which extends the Butterfly Analysis framework to incorporate explicit happensbefore arcs resulting from high-level synchronization within a monitored program. We show how to adapt two standard dataflow analysis techniques and two memory and security lifeguards to Chrysalis Analysis, using novel techniques for dealing with the many complexities introduced by happens-before arcs. Our security tool implementation shows that Chrysalis Analysis matches the key advantages of Butterfly Analysis—parallel monitoring, no detailed inter-thread data dependence tracking, no strong memory consistency requirements, and no missed errors—while significantly reducing the number of false positives.

Interactive Analytical Processing in Big Data Systems: A Cross-Industry Study of MapReduce Workloads

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

38th International Very Large Databases Conference (VLDB'12), August 2012.

Within the past few years, organizations in diverse industries have adopted MapReduce-based systems for large-scale data processing. Along with these new users, important new workloads have emerged which feature many small, short, and increasingly interactive jobs in addition to the large, long-running batch jobs for which MapReduce was originally designed. As interactive, large-scale query processing is a strength of the RDBMS community, it is important that lessons from that field be carried over and applied where possible in this new domain. However, these new workloads have not yet been described in the literature. We fill this gap with an empirical analysis of MapReduce traces from six separate business-critical deployments inside Facebook and at Cloudera customers in e-commerce, telecommunications, media, and retail. Our key contribution is a characterization of new MapReduce workloads which are driven in part by interactive analysis, and which make heavy use of query-like programming frameworks on top of MapReduce. These workloads display diverse behaviors which invalidate prior assumptions about MapReduce such as uniform data access, regular diurnal

continued on pg. 30

Recent Publications

continued from pg. 29

patterns, and prevalence of large jobs. A secondary contribution is a first step towards creating a TPC-like data processing benchmark for MapReduce.

Application-to-Core Mapping Policies to Reduce Memory Interference in Multi-Core Systems

Reetuparna Das, Rachata Ausavarungnirun, Onur Mutlu, Akhilesh Kumar, Mani Azimi

Proceedings of the 21st ACM International Conference on Parallel Architectures and Compilation Techniques (PACT'12) Poster Session, September 2012.

How applications running on a many-core system are mapped to cores largely determines the interference between these applications in critical shared resources. This paper proposes application-to-core mapping policies to improve system performance by reducing inter-application interference in the on-chip network and memory controllers. The major new ideas of our policies are to: 1) map network-latency-sensitive applications to separate parts of the network from network-bandwidth-intensive applications such that the former can make fast progress without heavy interference from the latter, 2) map those applications that benefit more from being closer to the memory controllers close to these resources. Our evaluations show that both ideas significantly improve system throughput, fairness and interconnect power efficiency.

RainMon: An Integrated Approach to Mining Bursty Timeseries Monitoring Data

Ilari Shafer, Kai Ren, Vishnu Boddeti, Yashihisa Abe, Greg Ganger, Christos Faloutsos

Proc. 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'12), August 2012.

Metrics like disk activity and network trac are widespread sources of diagnosis and monitoring information in datacenters and networks. However, as the scale of these systems increases, examining the raw data yields diminishing insight. We present RainMon, a novel end-to-end approach for mining timeseries monitoring data designed to handle its size and unique characteristics. Our system is able to (a) mine large, bursty, real-world monitoring data, (b) find significant trends and anomalies in the data, (c) compress the raw data effectively, and (d) estimate trends to make forecasts. Furthermore, RainMon integrates the full analysis process from data storage to the user interface to provide accessible long-term diagnosis. We apply RainMon to three real-world datasets from production systems and show its utility in discovering anomalous machines and time periods.

Xerxes: Distributed Load Generator for Cloud-scale Experimentation

Mukil Kesavan, Ada Gavrilovska, Karsten Schwan

7th OpenCirrus Summit, June 2012.

With the growing acceptance of cloud computing as a viable computing paradigm, a number of research and real-life dynamic cloud-scale resource allocation and management systems have been developed over the last few years. An important problem facing system developers is the evaluation of such systems at scale. In this paper we present the design of a distributed load generation framework, Xerxes, that can generate appropriate resource load patterns across varying datacenter scales, thereby representing various cloud load scenarios. Toward this end, we first characterize the resource consumption of four distributed cloud applications that represent some of the most widely used classes of applications in the cloud. We then demonstrate how, using Xerxes, these patterns can be directly replayed at scale, potentially even beyond what is easily achievable through application reconfiguration. Furthermore, Xerxes allows for additional parameter manipulation and exploration of a wide range of load scenarios. Finally, we demonstrate the ability to use Xerxes with publicly available datacenter traces which can be replayed across datacenters with different configurations. Our experiments are conducted on a 700-node 2800-core private cloud datacenter, virtualized with the VMware vSphere virtualization stack. The benefits of such a microbenchmark for cloud-scale experimentation include: (i) decoupling load scaling from application logic, (ii) reslience to faults and failures, since applications tend to crash altogether when some components fail, particularly at scales, and (iii) ease of testing and the ability to understand system behavior in a variety of actual or anticipated scenarios.

Other Interesting Papers by ISTC-CC Faculty

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

Starchart: Hardware/Software Optimization Using Recursive Partitioning Regression Trees. Wenhao Jia, Kelly A. Shaw, Margaret Martonosi. 22nd ACM International Conference on Parallel Architectures and Compilation Techniques (PACT'13), September 2013.

Road-Network Aware Trajectory Clustering: Integrating Locality, Flow and Density. Binh Han, Ling Liu, Edward R. Omiecinski. IEEE Transactions on Mobile Computing, September 2013.

Social Influence Based Clustering of Heterogeneous Information Networks. Yang Zhou, Ling Liu. 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'13), August 2013.

TripleBit: a Fast and Compact System for Large Scale RDF Data. Pingpeng Yuan, Pu Liu, Buwen Wu, Hai Jin, Wenya Zhang, Ling Liu. 39th International Conference on Very Large Databases (VLDB'13), August 2013.

Developing a Predictive Model of Quality of Experience for Internet Video. Athula Balachandran, Vyas Sekar, Aditya Akella, Srinivasan Seshan, Ion Stoica, Hui Zhang. ACM SIGCOMM 2013 Conference (SIGCOMM'13), August 2013.

Recent Publications

Reducing Contention Through Priority Updates. Julian Shun, Guy E. Blelloch, Jeremy T. Fineman, Phillip B. Gibbons. 25th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA'13), July 2013.

Efficient BVH Construction via Approximate Agglomerative Clustering. Yan Gu, Yong He, Kayvon Fatahalian, Guy E. Blelloch. ACM High Performance Graphics (HPG'13), July 2013.

Program-Centric Cost Models for Locality. Guy E. Blelloch, Jeremy Fineman, Phillip B. Gibbons, Harsha Vardhan Simhadri. 8th ACM SIGPLAN Workshop on Memory Systems Performance and Correctness (MSPC'13), June 2013.

An Experimental Study of Data Retention Behavior in Modern DRAM Devices: Implications for Retention Time Profiling Mechanisms. Jamie Liu, Ben Jaiyen, Yoongu Kim, Chris Wilkerson, Onur Mutlu. 40th ACM International Symposium on Computer Architecture (ISCA'13), June 2013.

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.

Threshold Voltage Distribution in MLC NAND Flash Memory: Characterization, Analysis and Modeling. Yu Cai, Erich F. Haratsch, Onur Mutlu, and Ken Mai, Design, Automation, and Test in Europe Conference (DATE'13), March 2013.

Relational Algorithms for Multi-Bulk-Synchronous Processors (short paper). G. Diamos, H. Wu, J. Wang, A. Lele, and S. Yalamanchili, 18th Symposium on Principles and Practice of Parallel Programming (PPoPP'13), February 2013.

Error Analysis and Retention-Aware Error Management for NAND Flash Memory. Yu Cai, Gulay Yalcin, Onur Mutlu, Erich F. Haratsch, Adrian Cristal, Osman Unsal, and Ken Mai, Intel Technology Journal (ITJ) Special Issue on Memory Resiliency, 2013.

Reducing GPU Offload Latency Via Fine-Grained CPU-GPU Synchronization. Dan Lustig and Margaret Martonosi, International Symposium on High-Performance Computer Architecture (HPCA'13), February, 2013.

Tiered-Latency DRAM: A Low Latency and Low Cost DRAM Architecture. Donghyuk Lee, Yoongu Kim, Vivek Seshadri, Jamie Liu, Lavanya Subramanian, and Onur Mutlu, International Symposium on High-Performance Computer Architecture (HPCA'13), February, 2013.

TLB Improvements for CMPs: Inter-Core Cooperative Prefetchers and Shared Last-Level TLBs. Dan Lustig, Abhishek Bhattacharjee, Margaret Martonosi, ACM Transactions on Architecture and Compiler Optimization (TACO'13), January, 2013.

Discovering Structure in Unstructured I/O. Jun He, John Bent, Aaron Torres, Gary Grider, Garth Gibson, Carlos Maltzahn, and Xian-He Sun. Proc. of the Seventh Parallel Data Storage Workshop (PDSW12), co-located with the Int. Conference for High Performance Computing, Networking, Storage and Analysis (SC12), November 2012.

DCast: Sustaining Collaboration in Overlay Multicast despite Rational Collusion. Haifeng Yu, Phillip B. Gibbons, Chenwei Shi. (CCS'12), October 2012.

HAT: Heterogeneous Adaptive Throttling for On-Chip Networks. Kevin Chang, Rachata Ausavarungnirun, Chris Fallin, Onur Mutlu. Proceedings of the 24th International Symposium on Computer Architecture and High Performance Computing, October 2012.

Generation Smartphone. Dan Siewiorek. IEEE Spectrum, vol. 49, no. 9, September, 2012, pp. 54-58.

Row Buffer Locality Aware Caching Policies for Hybrid Memories. HanBin Yoon, Justin Meza, Rachata Ausavarungnirun, Rachael Harding, Onur Mutlu. Proceedings of the 30th IEEE International Conference on Computer Design (ICCD'12), September 2012. Best paper award (in Computer Systems and Applications track).

Flash Correct-and-Refresh: Retention-Aware Error Management for Increased Flash Memory Lifetime. Yu Cai, Gulay Yalcin, Onur Mutlu, Eric Haratsch, Adrian Cristal, Osman Unsal, Ken Mai. Proceedings of the 30th IEEE International Conference on Computer Design (ICCD'12), September 2012.

A Case for Small Row Buffers in Non-Volatile Main Memories. Justin Meza, Jing Li, and Onur Mutlu. Proceedings of the 30th IEEE International Conference on Computer Design (ICCD'12) Poster Session, September 2012.

Base-Delta-Immediate Compression: A Practical Data Compression Mechanism for On-Chip Caches. Gennady Pekhimenko, Vivek Seshadri, Onur Mutlu, Todd C. Mowry, Phillip B. Gibbons, Michael A. Kozuch. International Conference on Parallel Architectures and Compilation Techniques (PACT'12), September 2012.

The Evicted-Address Filter: A Unified Mechanism to Address Both Cache Pollution and Thrashing. Vivek Seshadri, Onur Mutlu, Todd C Mowry, Michael A Kozuch. International Conference on Parallel Architectures and Compilation Techniques (PACT'12), September 2012.

Wearable Computers. D. Siewiorek, Smailagic, A., Starner, T. Chapter 12 in The Human-Computer Interaction Handbook, Fundamentals, Evolving Technologies, and Emerging Applications, Third Edition, J. A. Jacko (ed), CRC Press, pp. 273-296.

On-Chip Networks from a Networking Perspective: Congestion and Scalability in Many-core Interconnects. George Nychis, Chris Fallin, Thomas Moscibroda, Onur Mutlu, Srinivasan Seshan. Proceedings of the 2012 ACM SIGCOMM Conference (SIGCOMM'12), August 2012.

The Cost of Fault Tolerance in Multi-Party Communication Complexity. Binbin Chen, Haifeng Yu, Yuda Zhao, Phillip B. Gibbons. Proc. 31st ACM Symposium on Principles of Distributed Computing (PODC'12), July 2012.

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 platforms based on emerging technologies like non-volatile memory and specialized cores.
» 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 two 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 and frameworks as well as specialized computing platforms.
» Problem diagnosis and mitigation: exploring new techniques for effectively diagnosing and mitigating problems given the anticipated scale and complexity increases coming with future cloud computing.

Pillar 3: Big Data

"Big Data analytics" refers to a rapidly growing style of computing characterized by its reliance on large and often dynamically growing datasets. 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 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 Big Data activities to eventually dominate future cloud computing.

We envision future cloud computing infrastructures that efficiently and effectively support Big Data analytics. 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 Big Data analytics in future cloud computing infrastructures. Three particular areas of focus will be:

» "Big Learning" frameworks and systems that more effectively accommodate the advanced machine learning algorithms and interactive processing that will characterize much of next generation Big Data analytics. This includes a focused effort on Big Learning for genome analysis.
» 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.
» Understanding Big Data applications, creating classifications and benchmarks to represent them, and providing support for programmers building them.

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 and data span client devices, edge-local cloud resources, and core cloud resources, as discussed above.
» Addressing edge connectivity issues by creating effective data staging and caching techniques that mitigate reliance on expensive and robust Internet uplinks/downlinks for clients, while preserving data consistency requirements.
» 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 great second year for the Cloud Computing Center. I continue to see deep engagement between many of the projects and our Intel Researchers on the CMU campus. The research work at the center continues to make great strides and along with it, garnering the attention of a variety of groups at Intel. 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

- Kai Li (Princeton) received ACM SIGOPS Hall of Fame award for his paper "Memory Coherence in Shared Virtual Memory Systems" in ACM TOCS.
- Onur Mutlu (CMU) and co-authors won the Best Paper Award at ICCD'12 for their work on "Row Buffer Locality Aware Caching Policies for Hybrid Memories."
- Elmer Garduno, Soila Kavulya, Jiaqi Tan, Rajeev Gandhi, and Priya Narasimhan (CMU) won the Best Student Paper Award at USENIX LISA'12 for their work on failure-diagnosis and visualization for Hadoop.
- Priya Narasimhan (CMU) served as program co-chair of Middleware'12. Onur Mutlu (CMU) served as program co-chair for Micro'12. Margaret Martonosi (Princeton) is serving as program chair for ISCA'13. Karsten Schwan agreed to be program co-chair for Middleware'13. Jeff Parkhurst agreed to be conference co-chair for the Intel Big Data/Hadoop summit at Jones Farm in February 2013.
- Garth Gibson was awarded a Los Alamos National Laboratory Outstanding Innovation, 2011 Technology Transfer, Distinguished Copyright for the Parallel Log-structured File System (PLFS).
- Wyatt Lloyd won the Princeton Wu Prize for Excellence.
- Satya (CMU) was awarded two new grants: from IMSL (Institute for Museum & Library Sciences), Gloriana St. Clair and Mahadev Satyanarayanan (PIs), for "Olive: A Digital Archive for Executable Content", and from the Sloan Foundation, Mahadev Satyanarayanan and Gloriana St. Clair (PIs), for "Storing and Archiving Executable Content: the Olive Project".
- The Second Annual ISTC-CC Retreat hosted 62 attendees from CMU (18 faculty/staff & 44 students), 28 from Intel, 5 from Georgia Tech (3 faculty & 2 students), 4 from Princeton (2 Faculty & 2 students), 3 from UC Berkeley (2 faculty & 1 student), and 1 each from Washington, Penn State, and Brown. The agenda featured welcoming remarks by Wen-Hann Wang (Intel Labs, Executive Sponsor for ISTC-CC), keynotes by Balint Fleischer (General Manager, Intel Data Center Group) and Das Kamhout (Intel IT Cloud Lead), 12 research talks by faculty and students from all four Universities, a research collaboration talk by Ted Willke (Intel Labs / SAL), 5 breakout groups, and 53 posters.
- Vijay Vasudevan's (CMU) Ph.D thesis on FAWN was one of two theses nominated by Carnegie Mellon for submission to ACM's Outstanding Dissertation Award competition.
- Garth Gibson (CMU) gave the keynote talk at the Storage System, Hard Disk and Solid State Technologies Summit, co-located with the Asia-Pacific Magnetic Recording Conference (APMRC), Singapore, November 1.
- Ada Gavrilovska (Georgia Tech) presented "Experiences from Fighting the Cloud Management Wars at Scale" at the VMware Research Symposium, September 26.
- Ada Gavrilovska(Georgia Tech) presented "OpenCirrus as a datacenter research instrument" at the NSF workshop on Instrumentation-as-a-Service for Computer and Information Science and Engineering.
- Karsten Schwan(Georgia Tech) presented "Managing Applications on Exascale Machines" at the DOE ExaOS workshop in Washington DC,
- Dan Siewiorek (CMU) gave a keynote talk "Overview of the Quality of Life Technology Center: Six Years of Progress," at Fourth International Symposium on QoLT, Incheon, Korea, October 31.
- Dan Siewiorek (CMU) gave a keynote talk "Human System Interaction in QoLT" at the Fourth International Symposium on Rehabilitation Research, Incheon, Korea, November 1.
- Dan Siewiorek (CMU) gave an invited talk on Virtual Coaches in Health Care at Sangnam Institute of Management, Yonsie University, Seoul, Korea, November 2.
- Hyeontaek Lim (CMU grad student) gave a talk at Facebook on the SILT work.
- Phil Gibbons (Intel Labs) spoke on "Trumping the Multicore Memory Hierarchy with Hi-Spade," at Georgia Tech CSE Seminar, November 7.
- Michael Kozuch (Intel Labs), "Don't Overspend in the Datacenter: Dynamic Resource Scaling Techniques for Internet Services," at the Penn State CSE Colloquium.
- Mike Freedman (Princeton) presented "Performance Isolation and Fairness for Multi-Tennant Cloud Storage" at UC Berkeley, Systems Seminar, December 4.

Justin Meza describes his research on "Row Buffer Locality-Aware Caching Policy for Hybrid Memories" to an ISTC-CC Retreat guest.

2012 Quarter 3

- Mike Freedman (Princeton) received a Presidential Early Career Award for Scientists and Engineers (PECASE) for "efforts in designing, building, and prototyping a modern, highly scalable, replicated storage cloud system that provides strong robustness guarantees."
- Ling Liu (GA Tech) and her co-authors received the best paper award at CLOUD'12 for their work on "Reliable State Monitoring in Cloud Datacenters."
- Ling Liu (GA Tech) served as general chair of VLDB 2012 held in Istanbul, Turkey, August 2012.
- Onur Mutlu (CMU) received 3 awards: (i) an Intel Early Career Faculty Honor Program Award, (ii) an IBM Faculty Partnership Award, and (iii) an HP Labs Innovation Research Program Award.
- Ada Gavrilovska (GA Tech) was awarded an NSF grant for her work on cloud power management.

- ISTC-CC presented an overview talk, 12 posters, and 3 demos at Intel's University Collaboration Office "Showcase" in Santa Clara on June 27.
- M. Satyanarayanan (CMU) served as Program Chair for the 3rd ACM Asia-Pacific Systems Workshop (APSys'12) held in Seoul, S. Korea on July 23-24. Frans Kaashoek was the keynote speaker at the workshop.
- Michael Kozuch is General Chair of the Eighth Open Cirrus Summit, a workshop to be held in conjunction with the International Conference on Autonomic Computing (ICAC'12).
- Guy Blelloch (CMU) and Phil Gibbons (Intel Labs) co-organized an NSF Workshop on Research Directions in the Principles of Parallel Computing.
- Wen-Hann Wang (ISTC-CC Sponsoring Executive) gave a keynote talk on "Powering the Cloud Computing of the Future" at the OpenCirrus summit in Beijing, June 2012.
- Phil Gibbons (Intel Labs) gave three talks on multi-core computing at the Madalgo Summer School on Algorithms for Modern Parallel and Distributed Models in Aarhus, Denmark.
- S. Yalamanchili (GA Tech) gave an invited talk "Scalable Resource Composition in a Flat World," at the 1st Workshop in Unconventional Cluster Architectures and Applications, 2012.
- G. Hsieh, A. Kerr, H. Kim, J. Lee, N. Lakshminarayana, S. Li, A. Rodrigues, and S. Yalamanchili (GA Tech) gave a tutorial at the IEEE International Symposium on Computer Architecture entitled "Ocelot and SST-Macsim," June 2012.
- Ling Liu (GA Tech) presented three invited talks on Big Data and Big Data Analytics at (i) Tokyo Univ. June 2012, (ii) Keynote in the 2012 Big Data workshop, July 2012, organized by Big Data Summer School in RUC, Peking, and (iii) Keynote in the VLDB 2012 PhD workshop, August 2012 Istanbul, Turkey.
- Calton Pu (GA Tech) presented his ISTC cloud research at several international sites: Renmin University (Beijing, China), Huazhong Inst. Sci. Tech. (Wuhan, China), Univ. Tokyo (Tokyo, Japan), and Data Eng. Workshop (Nagoya, Japan).
- Onur Mutlu (CMU) gave a distinguished lecture entitled "Scaling the Main Memory System in the Many-Core Era," at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, June 2012. Onur also gave a number of talks at Intel Labs during his summertime visit, including "Architecting and Exploiting Asymmetry in Multi-Core Architectures" at Intel Archfest (HIllsboro, OR), August 2012.

Message from the PIs

continued from pg. 2

summit. Intel Labs also became active in GraphLab, including developing and releasing the GraphBuilder tools for helping developers use it. One more brag: 570 people from academia and industry attended the second GraphLab Workshop this past summer.

Another area where major progress is being made is on resource scheduling in clouds, and especially on a topic induced by the cross-section of three ISTC-CC pillars: scheduling for specialization. Our recent studies of cloud workloads, such as our analysis of the cluster traces from Google's internal private cloud (another 3-institution effort!), expose some of the challenges faced: extreme heterogeneity and dynamicity in the workloads. Our continued efforts to promote platform specialization add another: workloads are better off when they are assigned to the "right" resources, but they are often happier to get second or third choices rather than waiting for the "right" ones. Our new approaches to scheduling are creating the interfaces and automation support needed to make specialization truly effective, in the live.

Continuing with our cross-pillar theme, our scheduling efforts are also enabling cooperative scheduling of resources among multiple big-learning systems. Our Mesos system has become another active open source project, used in real production clusters, such as at Twitter. Our continuing research is exploring how to combine the different scheduling aspects into a solution that addresses all of these issues at once.

As one more (of many!) example activity of growing impact, our Cloudlet project has progressed to the point of gaining interest within Intel from two of Intel's business units. A joint gathering of our team and the Intel folks explored potential Intel directions for running with the cloudlet idea. And, of course, we continue to explore the best ways of realizing cloudlets and the applications that could exploit them.

Lots of progress has been made on many other fronts, as well. As one quick example, our Egalitarian Paxos protocols are enabling much more efficient wide-area data replication, such as between cloud data centers. As another, we are finding that we can merge stream processing and batch processing computations with an approach we call discretized streams. Our diagnosis research has made great strides, and our most recent graduate from the project will be joining Intel Labs. 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 News

continued from pg. 7

FEB 5, 2013

ISTC-CC Researchers Awarded Allen Newell Award for Research Excellence!

Congratulations to David Andersen and Michael Kaminsky for winning the Allen Newell Award for Research Excellence!

The Allen Newell Award for Research Excellence is awarded (roughly) annually by the School of Computer Science at Carnegie Mellon. It recognizes an outstanding body of work that epitomizes Allen Newell's research style as expressed in his words: "Good science responds to real phenomena or real problems. Good science is in the details. Good science makes a difference."

David and Michael won the award for "Energy-efficient Data Intensive Computing" for their FAWN project, which is part of the Intel Science & Technology Center for Cloud Computing. FAWN (Fast Array of Wimpy Nodes) demonstrates how many low-power (e.g., Atom) nodes with SSDs provides significant energy-efficiency gains for important cloud workloads such as key-value stores, and how to redesign system software to get the maximum benefit from such platforms.

DEC 12, 2012

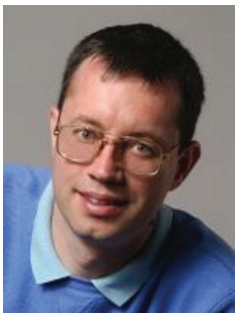
CMU Group Wins Best Student Paper Award at USENIX LISA 2012

Congratulations to Soila Pertet Kavulya and co-authors Elmer Garduno, Jiaqi Tan, Rajeev Gandhi, and Priya Narasimhan, all of CMU, for winning Best Student Paper at 26th USENIX Large Installation System Administration Conference (LISA'12), Dec 9-14, SanDiego, CA for their work on visualization for failure diagnosis in the paper "Theia: Visual Signatures for Problem Diagnosis in Large Haddop Clusters."

DEC 11, 2012

Garth Gibson & Ion Stoica Named ACM Fellows

Congratulations to Garth Gibson (CMU) and Ion Stoica (UC Berkeley), who have been named Class of 2012 ACM Fellows. Garth was cited for "contributions to the performance and reliability of storage systems." Ion Stoica was named for "contributions to networking, distributed systems, and cloud computing." The full list of awardees may be found at www.acm.org/press-room/news-releases/2012/fellows-2012.



DEC 8, 2012

Collaborative Intel / ISTC-CC Open Source Code Release: Graphbuilder

In an example of how the collaborative efforts of Intel and the members of the ISTC-CC can produce important technology advancements, Intel has released beta open source software, called GraphBuilder, to help data scientists in industry and academia to rapidly develop new applications that draw insights from Big Data. GraphBuilder is the first scalable open source library to take large data sets and construct them into "Graphs," web-like structures that outline relationships among data. For a more in-depth explanation of Graphlab and constructing graphs from Big Data, see Ted Willke's blog on the subject (http://blogs.intel.com/intellabs/2012/12/06/graphbuilder/).

-- with info from Ted Willke's Intel Blog and the Intel Newsroom

NOV 28, 2012

Onur Mutlu Wins Intel Early Career Award For Innovative Research

Carnegie Mellon University's Onur Mutlu has received the prestigious 2012 Intel Early Career Faculty Award for outstanding research and educational contributions in the field of computer architecture. Intel's Early Career Faculty Honor Program award provides financial and networking support to those faculty members early in their careers who show great promise as future academic leaders in disruptive computing technologies. The program helps promote the careers of promising faculty members and fosters long-term collaborative relationships with senior technical leaders at Intel.

Mutlu's current research focuses on new memory architectures and technologies to make computers store and manipulate data more efficiently and reliably.

Mutlu's research has received several other prestigious recognitions, including numerous best paper awards and "Top Pick" paper selections by the Institute of Electrical and Electronics Engineers (IEEE) Micro journal. In 2011, he received the Young Computer Architect Award from IEEE Computer Society's Technical Committee on Computer Architecture. And in 2012, CMU's College of Engineering recognized him with the George Tallman Ladd Research Award.

--expanded article in CMU News, Nov. 28, 2012

NOV 12, 2012

Sloan Foundation Grant for CMU's Olive Project

It has been announced by the Sloan Foundation that they are giving CMU $400K over two years to develop a production-quality open-source implementation of " Olive." Olive is the project name for VM-based archiving of executable content in the cloud. Combined with a similar grant from the Institute for Museum and Library Science, this gives the project $800K over two years. The grant will support the technical development of a platform for archiving executable content and the environment in which it runs, as well as a plan for the institutionalization and ongoing sustainability of such an archive. The archived VM images will be executed "close to the user"—at a native Linux client, or in a cloudlet with one-hop VNC to iOS/Android/Windows tablet or smartphone.