Intel Science & Technology Center for Cloud Computing

ISTC-CC Update

August 2015

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

(Original overview from 2011)

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. 34

Message from the PIs

Hello from ISTC-CC headquarters. This ISTC-CC Newsletter, our fourth, includes news and happenings from the last 11 months, abstracts of our many publications, and a refresher of the overall ISTC-CC research agenda. While we can't recap all that has happened in this introductory note, we do want to highlight a few things.

Greg Ganger, CMU

Phil Gibbons, Intel

As we move into the final year of ISTC-CC's five year charter, we remain thrilled with the ISTC-CC community and its growing collection of contributions. This great team continues to have major impact, within Intel and in the greater community, both in underlying ideas and technologies and in open source software systems. The larger capstone efforts are building to strong outcomes, and we expect collaborative activities among ISTC-CC and Intel researchers to continue beyond the fifth year.

As we often remark, a big part of ISTC-CC's success has been the high level of collaboration -- the individuals are stellar, but ISTC-CC is much more than the sum of its parts. Most of ISTC-CC's biggest wins have come from teams within and across the 6 participating institutions. Indeed, many of the technical papers and software artifacts involve researchers from multiple institutions... and, we're especially happy with the direct involvement of Intel collaborators and co-authors in so many cases. We also see increasing numbers of ISTC-CC students doing internships and taking full-time jobs at Intel Labs, in large part because of these collaborative experiences.

BTW, please remember to visit the ISTC-CC software page, which lists the many software releases and open source development efforts in order to make them easier to find. We continue to add to it, as we try to make our efforts ever more useful to Intel, ISTC-CC and the broader community.

As reviewed 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. Indeed, our capstone efforts involve combining activities from different areas toward larger goals.

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. Lots of energy and entire software ecosystems are growing up around both, including adoption and contributions by Intel. ISTC-CC continues to develop a range of more effective and natural abstractions for different types of non-trivial ML tasks and designing frameworks to enable them, such as consistency models based on

continued on pg. 36

Fourth Annual ISTC-CC Retreat a Success!

The ISTC-CC held its fourth annual retreat at the Intel Jones Farm campus in Hillsboro, OR on September 4 & 5, 2014. The retreat opened with a reception and dinner on Wednesday, September 3, and the main research content of the conference began on Thursday morning. The 91 attendees included faculty and students from Carnegie Mellon, Georgia Tech, Princeton, UC Berkeley & Washington, as well as 61 Intel employees. The agenda featured a keynote by Jim Blakley, of Intel, 14 research talks by faculty and students all of the five Universities, and 38 posters. By all accounts, the retreat was a big success: great interactions, lots of connections made, new insights, idea inspiration, and generally superb energy! The retreat was followed by the Board of Advisors meeting, and an additional meeting for Intel stakeholders. These meetings provided considerable positive feedback, as well as good suggestions. Full details on the retreat can be found on the ISTC-CC website. Note that the fifth ISTC-CC Retreat is scheduled for August 27 & 28, 2015 to be held again at Intel's Jones Farm campus in Oregon.

Group photo — fourth annual ISTC-CC Retreat, at the Intel Jones Farm campus in Hillsboro, OR, September 2014.

ISTC-CC Personnel

Leadership

Greg Ganger, Academic PI
Phil Gibbons, Intel PI
Executive Sponsor: Rich Uhlig, Intel
Managing Sponsor: Scott Hahn, Intel
Program Director: Jeff Parkhurst, Intel
Board of Advisors:
Curt Aubley, Intel
Randy Bryant, CMU
Jeff Chase, Duke
Jonathan Donaldson, Intel
Frans Kaashoek, MIT
Pradeep Khosla, UC San Diego

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
Michael Kaminsky, Intel
Randy Katz, Berkeley
Mike Kozuch, Intel
Ling Liu, GA Tech
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
Eric Xing, CMU
Sudhakar Yalamanchili, GA Tech

Staff

Joan Digney, Editor/Web, CMU
Olivia Vadnais, ISTC Admin. Manager, CMU

Students / Post-Docs

Yoshihisa Abe, CMU
Rachit Agarwal, UC Berkeley
Sameer Agarwal, UC Berkeley
Rachata Ausavarungnirun, CMU
Ben Blum, CMU
Kevin Kai-Wei Chang, CMU
Zhuo Chen, CMU
Anthony Chivetta, CMU
Henggang Cui, CMU
Wei Dai, CMU
Ankur Dave, UC Berkeley
Naila Farooqui, GA Tech
Kristen Gardner, CMU
Ali Ghodsi, Berkeley
Joseph Gonzalez, UC Berkeley
Samantha Gottlieb, CMU
Yan Gu, CMU
Mehgana Gupta, GA Tech
Kiryong Ha, CMU
Aaron Harlap, CMU
Liting Hu, GA Tech
Wenlu Hu, CMU
Lu Jiang, CMU
Tyler Johnson, Washington
Anuj Kalia, CMU
Anurag Khandelwal, UC Berkeley
Sudarsun Kannan, GA Tech
Samira Khan, CMU
Jin Kyu Kim, CMU
Yoongu Kim, CMU
Andy Konwinski, UC Berkeley
Guatam Kumar, UC Berkeley
Aapo Kyrola, CMU
Seunghak Lee, CMU
Mu Li, CMU
Hyeontaek Lim, CMU
Daniel Lustig, Princeton
Themistoklis Melissaris, Princeton
Justin Meza, CMU
Ishan Misra, CMU
Jun Woo Park, CMU
Gennady Pekhimenko, CMU
Ram Raghunathan, CMU
Kai Ren, CMU
Wolfgang Richter, CMU
Dipanjan Sengupta, GA Tech
Vivek Seshadri, CMU
Julian Shun, CMU
Logan Stafman, Princeton
Lavanya Subramanian, CMU
Jiaqi Tan, CMU
Brandon Taylor, CMU
Caroline Trippel, Princeton
Alexey Tumanov, CMU
Jinliang Wei, CMU
Haicheng Wu, GA Tech
Lin Xiao, CMU
Jin Xin, Princeton
Lianghong Xu, CMU
Neeraja Yadwadkar, UC Berkeley
Hobin Yoon, GA Tech
Manzil Zaheer, CMU
Qi Zhang,GA Tech
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

A sampling of significant ISTC-CC occurrences in the past 11 months.

Margaret Martonosi (Princeton) delivers her talk on "Hardware-Software Interface Issues in Heterogeneous Systems: Design, Verification, and Programming" at the 4th ISTC-CC Retreat.

2014 Quarter 3

» Alex Smola (CMU) presented a tutorial entitled "Scaling Machine Learning" at the Machine Learning Summer School in Pittsburgh, July 2014.

» Guy Blelloch (CMU) gave a series of invited lectures on teaching parallelism at Huazhong University of Science and Technology in Wuhan China, Aug. 2014.

» Karsten Schwan and Ada Gavrilovska (Georgia Tech), "Orchestrating the Execution of Distributed Applications with Virtualized Network Functions", $180,000, CISCO Corporation, donation for research, Aug. 2014.

» Onur Mutlu (CMU) presented "Error Analysis and Management for MLC NAND Flash Memory" at Flash Memory Summit 2014 (FMS), Santa Clara, CA, Aug. 2014.

» The 4th Annual ISTC-CC Retreat was held in Hillsboro, OR at the Intel Jones Farm campus.

» Phil Gibbons (IL/ISTC-CC) gave a keynote talk on "Algorithmic Challenges in M2M" at ALGO 2014, the annual federated algorithms conference in Europe, in Wroclaw, Poland, Sept. 2014.

» Karsten Schwan, PI, and Greg Eisenhauer (Georgia Tech), received Department of Energy funding ($3.2M over 3 years) for their research on "Understanding I/O Performance for Exascale Machines: Performance Understanding and Analysis for Exascale Data Management Workflows", Sept. 2014.

» Dan Siewiorek (CMU) served on the Organizing committee of the 2014 NIH/CCC Aging in Place Workshop. Sept. 2014.

» Dan Siewiorek (CMU) presented "Technologies to Support Physical Health" at the NIH/CCC Aging in Place Workshop, Sept. 11, Bethesda, MD.

» Mor Harchol Balter (CMU) received REU funding for undergraduates to do research with her.

2014 Quarter 4

» Garth Gibson (CMU) and team released IndexFS under a BSD-style license. IndexFS is middleware for scalable high-performance operations on metadata and small files that works with existing file systems such as PVFS, Lustre, and HDFS. This work won the best paper award at SC'14, Nov. 2014.

» Ion Stoica (UC Berkeley) and team released Apache Spark 1.1.1. Spark 1.1.1 includes fixes across several areas of Spark, including the core API, Streaming, PySpark, SQL, GraphX, and MLlib.

» Kiryong Ha (CMU) and Babu Pillai (IL/ISTC-CC) ported Cloudlet extensions for OpenStack to the latest Icehouse version. The new version and documentation have been published at GitHub, and new compatible base VMs for Ubuntu have been made available.

» Ling Liu (Georgia Tech) was elevated to IEEE Fellow, for "contributions to scalable Internet data management and decentralized trust management."

» Dong Zhou, Ph.D. student of Dave Andersen (CMU), interned with Ren Wang (IL/SSR/NPL).

» "Language Modeling with Power Low Rank Ensembles" by A. P. Parikh, A. Saluja, C. Dyer and E. P. Xing (CMU) received the best paper runner-up award at the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP'14), Oct. 2014.

» "Characterization and Analysis of Dynamic Parallelism in Unstructured GPU Applications" by J. Wang and S. Yalamanchili (Georgia Tech) received the best paper runner-up award at the IEEE International Symposium on Workload Characterization (IISWC'14) , Oct. 2014.

» Ling Liu (Georgia Tech) gave the key-

Year in Review

note talk at the 23rd International Conference on Software Engineering and Data Engineering (SEDE'14), Oct. 2014.

» Garth Gibson (CMU) gave a talk on "ML Systems: Experience with Machine Learning as a Distributed Systems Problem Space", University of Toronto ECE Distinguished Lecture Series, Oct. 9, 2014, Toronto, ON.

» S. Yalamanchili (Georgia Tech) presented The Era of Heterogeneity: Opportunities and Challenges, Keynote, Annual China National Computer Congress (CNCC), Oct. 2014.

» Ling Liu (Georgia Tech) was the invited speaker at IEEE the first Big Data Initiative workshop (BDIW), Oct. 1-2, 2014, Stevens Institute of Science and Technology, Hoboken, NJ.

» "Paxos Quorum Leases: Fast Reads without Sacrificing Writes" by Iulian Moraru, Dave Andersen (CMU), and Michael Kaminsky (IL/ISTC-CC) received the best paper award at the 5th ACM Symposium on Cloud Computing (SOCC'14), Nov. 2014.

» "Scaling File System Metadata Performance with Stateless Caching and Bulk Insertion" by Kai Ren, Qing Zheng, Swapnil Patil, and Garth Gibson (CMU) received the best paper award at the ACM/IEEE Int'l Conf. for High Performance Computing, Networking, Storage and Analysis (SC'14) , Nov. 2014.

» "Analyzing Log Analysis: An Empirical Study of User Log Mining" by Sara Alspaugh, Betty Beidi Chen, Jessica Lin, Archana Ganapathi, Marti A. Hearst, and Randy Katz (UC Berkeley) received the best paper award at the USENIX Large Installation System Administration Conference (LISA'14), Nov. 2014.

» Kai Ren (CMU) presented IndexFS: Scaling File System Metadata Throughput, VMWare Virtual SAN group, Nov. 3, 2014.

» Dan Siewiorek (CMU) spoke on Models of Industrial Relations, MIT-NSF Workshop: Smarter Service Systems through Innovation Partnerships and Transdisciplinary Research Nov. 20, 2014, Boston MA.

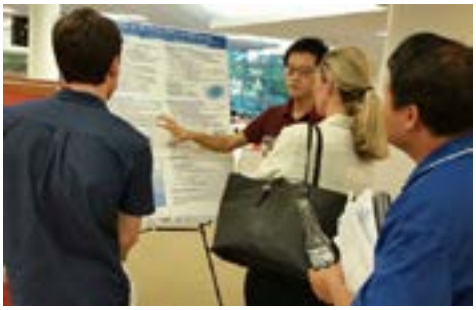
» "Balancing Context Switch Penalty and Response Time with Elastic Time Slicing" by Nagakishore Jammula, Moinuddin Qureshi, Ada Gavrilovska, Jongman Kim (Georgia Tech) received the best paper award at the 21st International Conference on High Performance Computing (HiPC'14), Dec. 2014.

» "When Twitter meets Foursquare: Tweet Location Prediction using Foursquare" by K. Lee, R. Ganti, M. Srivatsa and Li. Liu (Georgia Tech)received the Best Paper Award at the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (Mobiquitous'14), Dec. 2014.

» Garth Gibson (CMU) gave a keynote lecture on "ML Systems: Experience with Machine Learning as a Distributed Systems Problem Space" for the SEI-NSA Workshop on Predictive Analytics, Dec. 4, 2014, Linthicum, MD.

» Michael Kozuch (IL/ISTC-CC) served as Program Co-Chair for the Industry Track of Middleware'14, Dec. 2014.

» Margaret Martonosi (Princeton), Michael Pellauer (DCG/TCD), and Daniel Lustig (Princeton) were selected as best paper finalists at MICRO'14 for "PipeCheck: Specifying and Verifying Microarchitectural Enforcement of Memory Consistency Models," Dec. 2014.



Wei Dai (CMU) describes his research on "PETUUM: An ML-Centric System for Big Learning" to interested ISTC-CC retreat attendees.

2015 Quarter 1

» Dan Siewiorek (CMU) presented "Lessons from Wearable Computing and Beyond," an ACM Distinguished Lecture at Penn State, Erie PA, Behrend College, Jan. 2015.

» S.Yalamanchili (Georgia Tech) presented "Relational Processing Accelerators: From Clouds to Memory Systems," Intel, Bangalore, India, Jan. 2015.

» Justin Meza (CMU) presented "Efficient Data Mapping and Buffering Techniques for Multi-Level Cell Phase-Change Memories" at HiPEAC'15, Jan. 2015, winning the best student presentation award.

» "Data Retention in MLC NAND Flash Memory: Characterization, Optimization and Recovery" by Y. Cai, Y. Luo, E.F. Haratsch, K. Mai, and O. Mutlu (CMU) won Best Paper Runner-up at the 21st International Symposium on High-Performance Computer Architecture (HPCA'15), Feb. 2015.

» Justin Meza (CMU) was selected to receive a Google US/Canada Fellowship for his work in Systems Reliability, Feb. 2015.

» Mor Harchol-Balter (CMU) presented "Queueing with Redundant Requests: First Exact Analysis" at U.C. Berkeley, IEOR Department, Feb. 2015.

» Mor Harchol-Balter (CMU) presented "Queueing with Redundant Requests: First Exact Analysis" at the Information Theory and Applications Workshop (ITA), San Diego, CA, Feb. 2015.

» Gennady Pekhimenko (CMU) won the ACM student research competition held at ASPLOS'15, for his work entitled "Energy-Efficient Data Compression for GPU Memory Systems," March 2015.

» Margaret Martonosi (Princeton) gave an invited talk at the National Academies in Washington, DC on March 5 in the "Symposium on Continuing Innovation in Information Technology".

» Carlos Guestrin (Washington) and team released Data-Core, the open source core of GraphLab ML library.

2015 Quarter 2

» Garth Gibson (CMU) presented "Scalable parallel file system for data- and metadata-intensive workloads: On the path to a pure middleware approach to scalable storage namespaces," to the Mathematics and Computer Science Division, Argonne National Laboratory, Chicago, IL, April 17, 2015.

» Garth Gibson (CMU) presented "Bounded staleness in distributed machine learning systems: getting the right answer sooner," a talk for the University of Chicago Distinguished Lecture Series, Chicago, IL, April 16, 2015.

» M. Satyanarayanan (CMU) was the keynote speaker at WearSys 2015, the first ACM SIGMOBILE workshop on the "Wearable Systems and Applications" held in Florence, Italy and co-located with MobiSys 2015, May 2015.

continued on pg. 32

ISTC-CC News

June 13, 2015

Margaret Martonosi Awarded ACM SIGARCH/IEEE TCCA Influential ISCA Paper Award

Margaret Martonosi's paper "Wattch: A Framework for Architectural-Level Power Analysis and Optimizations," co-authored with David Brooks and Vivek Tiwari, from ISCA 2000 received ACM SIGARCH/IEEE TCCA Influential ISCA Paper Award at ISCA this year. This award recognizes the paper from the ISCA Proceedings 15 years earlier that has had the most impact on the field (in terms of research, development, products or ideas) during the intervening years.

--info from sigarch.org

May 18, 2015

Alexey Tumanov Receives ECE's Graduate Student Teaching Assistant Award

Congratulations to Alexey for receiving ECE's Outstanding Graduate Student Teaching Assistant Award for his efforts on 15-719: Advanced Cloud Computing, taught by Garth Gibson and Majd Sakr during the fall semester of 2014. In their letter of nomination Professors Sakr and Gibson cited Alexey's hard work, innovation, and commitment to student success, describing it as "unparalleled". Alexey "went way beyond the call of duty, supported the students with a pleasant constructive engagement style and built a project [that] will certainly [be] reused next year."

During the semester, Alexey developed the end-of-term course project, where the students were guided to build their own virtualized clusters and cluster schedulers on the brand new PRObE cluster called NOME. In the words of one of the students: "[Alexey was] extremely helpful and responsive. [We] had a lot of one-on-one discussions, which led to interesting insights and learning. [He] was very supportive of ideas and any issues faced. [He] strived hard to get the essence of the project into the students and drive the phases towards that goal. Probably my best project at CMU."

--with info from D. Marculescu's award presentation notes.

May 4, 2015

Best Paper Award at CCGRID '15

Congratulations to Yuzhe Tang, Arun Iyengar, Wei Tan, Liana Fong, Balaji Palanisamy, and Ling Liu for receiving the best paper award for their work on "Lightweight Indexing for Log-Structured Key-Value Stores" at the 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2015) conference. The acceptance rate at this conference is 25.7%.

April 27, 2015

Margaret Martonosi Receives Marie R. Pistilli Women in EDA Achievement Award

Margaret R. Martonosi, Hugh Trumbull Adams '35 Professor of Computer Science at Princeton University, has been selected as the Marie R. Pistilli Women in Electronic Design Automation (EDA) Achievement Award recipient for 2015. The award honors Dr. Martonosi for her technical leadership of high-impact research projects in the design of power-efficient computer architectures and sensor systems as well as her creation and organization of career development programs for women and minorities. As a highly visible woman in leadership positions at Princeton and within her professional community, Martonosi also has acted as a mentor and role model for graduate and undergraduate women students.

Martonosi is a Fellow of both IEEE and ACM. She was the 2013 recipient of the Anita Borg Institute Technical Leadership Award. She also has received the 2013 NCWIT Undergraduate Research Mentoring Award and the 2010 Princeton University Graduate Mentoring Award. In addition to many archival publications, Martonosi is an inventor on seven granted US patents, and has co-authored a technical reference book on power-aware computer architecture. She serves on the Board of Directors of the Computing Research Association (CRA).

"A leading researcher with over 160 refereed papers and seven US patents granted, a seminal figure behind computer research careers for women, and a dedicated mentor of women in technology, Dr. Martonosi is a force to be reckoned with," said Donatella Sciuto, Politecnico di Milano and chairperson of Women in Electronic Design. "We are honored to present her with the Marie Pistilli award in recognition of her notable contributions to research and technology and the impact she has made on career development programs for women and minorities."

--from www.reuters.com; more at http://www.reuters.com/

March 25, 2015

Onur Mutlu Receives Google Faculty Research Award

Congratulations to Onur on receiving a Google Faculty Research Award. Google Research Awards are one-year awards structured as unrestricted gifts to universities to support the work of world-class full-time faculty members at top universities around the world. The intent of the Google Research Awards is to support cutting-edge research in Computer Science, Engineering, and related fields. This Faculty Award is to support Professor Mutlu's research in the area of novel computer memory systems. Mutlu has been examining new memory architectures and interfaces with the goal of enabling low-cost and energy-efficient

computation near data. His related research develops both new hardware substrates and software interfaces to perform computation in or close to memory as well as software techniques that can take better advantage of such new substrates and interfaces.

--info from ECE News and google.com

March 14, 2015

First place in ACM Student Research Competition

Gennady Pekhimenko, along with co-authors Evgeny Bolotin, Mike O'Connor, Onur Mutlu, Todd C. Mowry and Stephen W. Keckler have been awarded first place in the ASPLOS Student Research Competition, Istanbul, Turkey for their work on "Energy-Efficient Data Compression for GPU Memory Systems." The ASPLOS SRC is a forum for undergraduates and graduate students to share their research results, exchange ideas, and improve their communication skills while competing for prizes. Students accepted to participate in the SRC are entitled to a travel grant (up to $500) to help cover travel expenses.

February 18, 2015

Justin Meza Receives Google US/Canada Fellowship

Justin Meza was selected to receive a Google US/Canada Fellowship for his work in Systems Reliability. Nurturing and maintaining strong relations with the academic community is a top priority at Google. The Google U.S./Canada PhD Student Fellowship Program was created to recognize outstanding graduate students doing exceptional work in computer science, related disciplines, or promising research areas.

-- info from googleresearch.blogspot.ca

February 7, 2015

Best paper runner-up at HPCA '15

Congratulations to Yu Cai, Yixin Luo, Erich F. Haratsch, Ken Mai, Onur Mutlu for receiving Best Paper Runner Up at the 21st International Symposium on High-Performance Computer Architecture (HPCA) for their paper "Data Retention in MLC NAND Flash Memory: Characterization, Optimization and Recovery."

January 30, 2015

MICRO '14 Paper an IEEE Top Pick

The MICRO 2014 paper "PipeCheck: Specifying and Verifying Microarchitectural Enforcement of Memory Consistency Models," by Daniel Lustig, Michael Pellauer and Margaret Martonosi, and originally published in the proceedings of the 47th International Symposium on Microarchitecture (MICRO'14) was selected as an annual Top Picks selection in Computer Architecture. It will appear in the IEEE Micro Special issue in May/June, 2015.

January 19, 2015

Best Student Presentation Award

Justin Meza received one of the two Best Presentation Awards at the 10th HiPEAC (High Performance and Embedded Architecture and Compilation) conference. The HiPEAC conference is the premier European forum for experts in computer architecture, programming models, compilers and operating systems for embedded and general-purpose systems. The presented paper, titled "Efficient Data Mapping and Buffering Techniques for Multi-Level Cell Phase-Change Memories", is co-authored with ECE's Onur Mutlu, alum HanBin Yoon and researchers from Google.

December 17, 2014

Best Paper Award at HiPC '14

Congratulations to Nagakishore Jammula, Moinuddin Qureshi, Ada Gavrilovska and Jongman Kim for being awarded Best Paper for their work on "Balancing Context Switch Penalty and Response Time with Elastic Time Slicing" at the 21st International Conference on High Performance Computing (HiPC '14), Goa, India.

December 13, 2014

Best Paper Nominee at Micro '14

Congratulations to Daniel Lustig, Michael Pellauer, and Margaret Martonosi for having their work "PipeCheck: Specifying and Verifying Microarchitectural Enforcement of Memory Consistency Models" nominated for Best Paper at the 47th International Symposium on Microarchitecture (MICRO).

December 2, 2014

Best Paper at Mobiquitous '14

Congratulations to Kisung Lee, Raghu Ganti, Mudhakar Srivatsa and Ling Liu on being awarded Best Paper at the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (Mobiquitous'14) for their work on "When Twitter meets Foursquare: Tweet Location Prediction using Foursquare."

December 2014

Ling Liu Named an IEEE Fellow

Ling Liu (Georgia Tech) was elevated to IEEE Fellow, for "contributions to scalable Internet data management and decentralized trust management." IEEE Fellow is a distinction reserved for select IEEE members whose extraordinary accomplishments in any of the IEEE fields of interest are deemed fitting of this prestigious grade elevation. There are now a total of 9 IEEE Fellows and 8 ACM Fellows on the ISTC-CC team.

November 16, 2014

Best Paper Award at SC '14

Congratulations to Kai Ren, Qing Zheng, Swapnil Patil, and Garth Gibson on being awarded best paper at the 2014 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'14) for their paper "IndexFS: Scaling File System Metadata Performance

continued on pg.36

Recent Publications

Succinct: Enabling Queries on Compressed Data

Rachit Agarwal, Anurag Khandelwal, Ion Stoica

Proceedings of the 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI'15), Oakland, CA, May 2015.

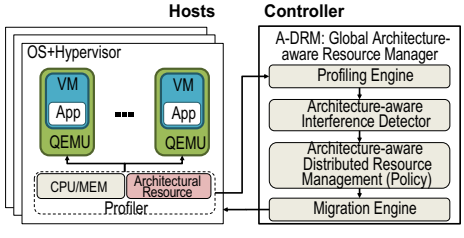
Succinct is a data store that enables efficient queries directly on a compressed representation of the input data. Succinct uses a compression technique that allows random access into the input, thus enabling efficient storage and retrieval of data. In addition, Succinct natively supports a wide range of queries including count and search of arbitrary strings, range and wildcard queries. What differentiates Succinct from previous techniques is that Succinct supports these queries without storing indexes — all the required information is embedded within the compressed representation.

Evaluation on real-world datasets show that Succinct requires an order of magnitude lower memory than systems with similar functionality. Succinct thus pushes more data in memory, and provides low query latency for a larger range of input sizes than existing systems.

A-DRM: Architecture-aware Distributed Resource Management of Virtualized Clusters

Hui Wang, Canturk Isci, Lavanya Subramanian, Jongmoo Choi, Depei Qian, Onur Mutlu

Proceedings of the 11th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE), Istanbul, Turkey, March 2015.



A-DRM prototype implementation

Virtualization technologies has been widely adopted by large-scale cloud computing platforms. These virtualized systems employ distributed resource management (DRM) to achieve high resource utilization and energy savings by dynamically migrating and consolidating virtual machines. DRM schemes usually use operating-system-level metrics, such as CPU utilization, memory capacity demand and I/O utilization, to detect and balance resource contention. However, they are oblivious to microarchitecture-level resource interference (e.g., memory bandwidth contention between different VMs running on a host), which is currently not exposed to the operating system.

We observe that the lack of visibility into microarchitecture-level resource interference significantly impacts the performance of virtualized systems. Motivated by this observation, we propose a novel architecture-aware DRM scheme (A-DRM), that takes into account microarchitecture-level resource interference when making migration decisions in a virtualized cluster. A-DRM makes use of three core techniques: 1) a profiler to monitor the microarchitecture-level resource usage behavior online for each physical host, 2) a memory bandwidth interference model to assess the interference degree among virtual machines on a host, and 3) a cost-benefit analysis to determine a candidate virtual machine and a host for migration.

Real system experiments on thirty randomly selected combinations of applications from the CPU2006, PARSEC, STREAM, NAS Parallel Benchmark suites in a four-host virtualized cluster show that A-DRM can improve performance by up to 26.55%, with an average of 9.67%, compared to traditional DRM schemes that lack visibility into microarchitecture-level resource utilization and contention.

Agility and Performance in Elastic Distributed Storage

Lianghong Xu, James Cipar, Elie Krevat, Alexey Tumanov, Nitin Gupta, Michael A. Kozuch, Gregory R. Ganger

ACM Transactions on Storage, Vol. 10, No. 4, Article 16, October 2014.

Elastic storage systems can be expanded or contracted to meet current demand, allowing servers to be turned off or used for other tasks. However, the usefulness of an elastic distributed storage system is limited by its agility: how quickly it can increase or decrease its number of servers. Due to the large amount of data they must migrate during elastic resizing, state of the art designs usually have to make painful trade-offs among performance, elasticity, and agility.

This article describes the state of the art in elastic storage and a new system, called SpringFS, that can quickly change its number of active servers, while retaining elasticity and performance goals. SpringFS uses a novel technique, termed bounded write offloading, that restricts the set of servers where writes to overloaded servers are redirected. This technique, combined with the read offloading and passive migration policies used in SpringFS, minimizes the work needed before deactivation or activation of servers. Analysis of real-world traces from Hadoop deployments at Facebook and various Cloudera customers and experiments with the SpringFS prototype confirm SpringFS's agility, show that it reduces the amount of data migrated for elastic resizing by up to two orders of magnitude, and show that it cuts the percentage of active servers required by 67–82%, outdoing state-of-the-art designs by 6–120%.

Architecting to Achieve a Billion Requests Per Second Throughput on a Single Key-Value Store Server Platform

Sheng Li, Hyeontaek Lim, Victor W. Lee, Jung Ho Ahn, Anuj Kalia, Michael Kaminsky, David G. Andersen, O. Seongil, Sukhan Lee, Pradeep Dubey

Proceedings of the 42nd Annual International Symposium on Computer Architecture (ISCA '15), June 13 - 17, 2015, Portland, OR.

Distributed in-memory key-value stores

Recent Publications

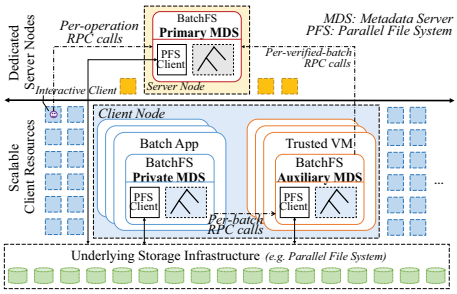
(KVSs), such as memcached, have become a critical data serving layer in modern Internet-oriented datacenter infrastructure. Their performance and efficiency directly affect the QoS of web services and the efficiency of datacenters. Traditionally, these systems have had significant overheads from inefficient network processing, OS kernel involvement, and concurrency control. Two recent research thrusts have focused upon improving key-value performance. Hardware-centric research has started to explore specialized platforms including FPGAs for KVSs; results demonstrated an order of magnitude increase in throughput and energy efficiency over stock memcached. Software-centric research revisited the KVS application to address fundamental software bottlenecks and to exploit the full potential of modern commodity hardware; these efforts too showed orders of magnitude improvement over stock memcached.

We aim at architecting high performance and efficient KVS platforms, and start with a rigorous architectural characterization across system stacks over a collection of representative KVS implementations. Our detailed full-system characterization not only identifies the critical hardware/software ingredients for high-performance KVS systems, but also leads to guided optimizations atop a recent design to achieve a record-setting throughput of 120 million requests per second (MRPS) on a single commodity server. Our implementation delivers 9.2X the performance (RPS) and 2.8X the system energy efficiency (RPS/watt) of the best-published FPGA-based claims. We craft a set of design principles for future platform architectures, and via detailed simulations demonstrate the capability of achieving a billion RPS with a single server constructed following our principles.

BatchFS: Scaling the File System Control Plane with Client-Funded Metadata Servers

Qing Zheng, Kai Ren, Garth Gibson

Proceedings of Parallel Data Storage Workshop (PDSW'14), co-located with the Int. Conference for High Performance Computing, Networking, Storage and Analysis, November 2014.



BatchFS is designed as file system metadata middleware layered on top of an existing cluster file system or an object storage platform exposing a flat namespace, which allows BatchFS to reuse the data path offered by these underlying storage substrates already optimized and tuned for maximum bandwidth. BatchFS features a client-driven metadata architecture that can shift server computation to client machines to achieve highly agile scalability.

Parallel file systems are often characterized by a layered architecture that decouples metadata management from I/O operations, allowing file systems to facilitate fast concurrent access to file contents. However, metadata intensive workloads are still likely to bottleneck at the file system control plane due to namespace synchronization, which taxes application performance through lock contention on directories, transaction serialization, and RPC overheads. In this paper, we propose a client-driven file system metadata architecture, BatchFS, that is optimized for noninteractive, or batch, workloads. To avoid metadata bottlenecks, BatchFS features a relaxed consistency model marked by lazy namespace synchronization and optimistic metadata verification. Capable of executing namespace operations on client-provisioned resources without contacting any metadata server, BatchFS clients are able to delay namespace synchronization until synchronization is really needed. Our goal in this vision paper is to handle these delayed operations securely and efficiently with metadata verification and bulk insertion. Preliminary experiments demonstrate that our client-funded metadata architecture outperforms a traditional synchronous file system by orders of magnitude.

Cloudlets: at the Leading Edge of Mobile-Cloud Convergence

M. Satyanarayanan, Z. Chen, K. Ha, W. Hu, W. Richter, P. Pillai

Proceedings of MobiCASE 2014: Sixth International Conference on Mobile Computing, Applications and Services, November 2014. Invited Paper.

As mobile computing and cloud computing converge, the sensing and interaction capabilities of mobile devices can be seamlessly fused with compute-intensive and data-intensive processing in the cloud. Cloudlets are important architectural components in this convergence, representing the middle tier of a mobile device — cloudlet — cloud hierarchy. We show how cloudlets enable a new genre of applications called cognitive assistance applications that augment human perception and cognition. We describe a plug-and-play architecture for cognitive assistance, and a proof of concept using Google Glass.

Clustering Service Networks with Entity, Attribute and Link Heterogeneity

Yang Zhou, Ling Liu, Xianqiang Bao, Kisung Lee, Calton Pu, Balaji Palanisamy, Emre Yigitoglu, Qi Zhang

IEEE 2015 International Conference on Web Services (ICWS 2015), New York, June 27-July 2, 2015.

Many popular web service networks are content-rich in terms of heterogeneous types of entities and links, associated with incomplete attributes. Clustering such heterogeneous service networks demands new clustering techniques that can handle two heterogeneity challenges: (1) multiple types of entities co-exist in the same service network with multiple attributes, and (2) links between entities have diverse types and carry different semantics. Existing heterogeneous graph clustering techniques tend to pick initial centroids uniformly at random, specify the number k of clusters in advance, and fix k during the clustering process. In this paper, we propose SERVICECLUS-

continued on pg. 10

Recent Publications

continued from pg. 9

TER, a novel heterogeneous SERVICE network CLUSTERing algorithm with four unique features. First, we incorporate various types of entity, attribute and link information into a unified distance measure. Second, we design a Discrete Steepest Descent method to naturally produce initial k and initial centroids simultaneously. Third, we propose a dynamic learning method to automatically adjust the link weights towards clustering convergence. Fourth, we develop an effective optimization strategy to identify new suitable k and k well-chosen centroids at each clustering iteration.

Edge Analytics in the Internet of Things

M. Satyanarayanan, P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, W. Hu, B. Amos

IEEE Pervasive Computing, Volume 14, Number 2, April-June 2015.

High-data-rate sensors are becoming ubiquitous in the Internet of Things. GigaSight is an Internet-scale repository of crowd-sourced video content that enforces privacy preferences and access controls. The architecture is a federated system of VM-based cloudlets that perform video analytics at the edge of the Internet.

Cuckoo Filter: Practically Better Than Bloom

Bin Fan, David G. Andersen, Michael Kaminsky, Michael D. Mitzenmacher

Proceedings of CoNEXT (CoNEXT'14), December 2014.

In many networking systems, Bloom filters are used for highspeed set membership tests. They permit a small fraction of false positive answers with very good space efficiency. However, they do not permit deletion of items from the set, and previous attempts to extend "standard" Bloom filters to support deletion all degrade either space or performance.

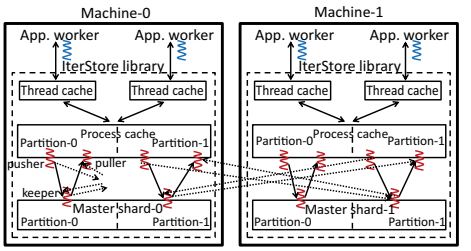
We propose a new data structure called the cuckoo filter that can replace Bloom filters for approximate set membership tests. Cuckoo filters support adding and removing items dynamically while achieving even higher performance than Bloom filters. For applications that store many items and target moderately low false positive rates, cuckoo filters have lower space overhead than space-optimized Bloom filters. Our experimental results also show that cuckoo filters outperform previous data structures that extend Bloom filters to support deletions substantially in both time and space.

Exploiting Iterativeness for Parallel ML Computations

Henggang Cui, Alexey Tumanov, Jinliang Wei, Lianghong Xu, Wei Dai, Jesse Haber-Kucharsky, Qirong Ho, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, Eric P. Xing

Proceedings of ACM Symposium on Cloud Computing 2014 (SOCC'14), November 2014.

Many large-scale machine learning (ML) applications use iterative algorithms to converge on parameter values that make the chosen model fit the input data. Often, this approach results in the same sequence of accesses to parameters repeating each iteration. This paper shows that these repeating patterns can and should be exploited to improve the efficiency of the parallel and distributed ML applications that will be a mainstay in cloud computing environments. Focusing on the increasingly popular "parameter server" approach to sharing model parameters among worker threads, we describe and demonstrate how the repeating patterns can be exploited. Examples include replacing dynamic cache and server structures with static pre-serialized structures, informing prefetch and partitioning decisions, and determining which data should be cached at each thread to avoid both contention and slow accesses to memory banks attached to other sockets. Experiments show that such exploitation reduces per-iteration time by 33–98%, for three real ML workloads, and that these improvements are robust to variation in the patterns over time.



IterStore with two partitions, running on two machines with two application threads each.

IndexFS: Scaling File System Metadata Performance with Stateless Caching and Bulk Insertion

Ren, Kai, Qing Zheng, Swapnil Patil, Garth Gibson

Proceedings of ACM/IEEE Int'l Conf. for High Performance Computing, Networking, Storage and Analysis (SC'14), November 2014. Best Paper Award.

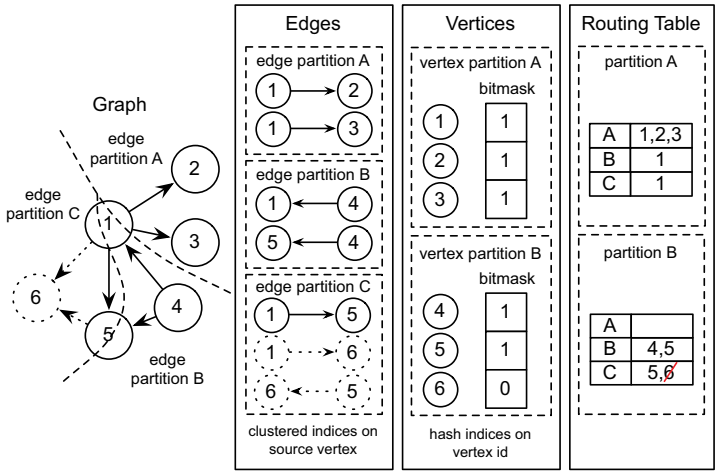
The growing size of modern storage systems is expected to exceed billions of objects, making metadata scalability critical to overall performance. Many existing distributed file systems only focus on providing highly parallel fast access to file data, and lack a scalable metadata service. In this paper, we introduce a middleware design called IndexFS that adds support to existing file systems such as PVFS, Lustre, and HDFS for scalable high-performance operations on metadata and small files. IndexFS uses a table-based architecture that incrementally partitions the namespace on a per-directory basis, preserving server and disk locality for small directories. An optimized log-structured layout is used to store metadata and small files efficiently. We also propose two client-based storm-free caching techniques: bulk namespace insertion for creation intensive workloads such as N-N checkpointing; and stateless consistent metadata caching for hot spot mitigation. By combining these techniques, we have demonstrated IndexFS scaled to 128 metadata servers. Experiments show our out-of-core metadata throughput out-performing existing solutions such as PVFS, Lustre, and HDFS by 50% to two orders of magnitude.

GraphX: Graph Processing in a Distributed Dataflow Framework

Joseph E. Gonzalez, Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, Ion Stoica

Proceedings of 11th USENIX OSDI (OSDI'14), October 2014.

In pursuit of graph processing performance, the systems community has largely abandoned general-purpose distributed dataflow frameworks in favor of specialized graph processing systems that provide tailored programming abstractions and accelerate the execution of iterative graph algorithms. In this paper we argue that many of the advantages of specialized graph processing systems can be recovered in a modern general-purpose distributed dataflow system. We introduce GraphX, an embedded graph processing framework built on top of Apache Spark, a widely used distributed dataflow system. GraphX presents a familiar composable graph abstraction that is sufficient to express existing graph APIs, yet can be implemented using only a few basic dataflow operators (e.g., join, map, group-by). To achieve performance parity with specialized graph systems, GraphX recasts graph-specific optimizations as distributed join optimizations and materialized view maintenance. By leveraging advances in distributed dataflow frameworks, GraphX brings low-cost fault tolerance to graph processing. We evaluate GraphX on real workloads and demonstrate that GraphX achieves an order of magnitude performance gain over the base dataflow framework and matches the performance of specialized graph processing systems while enabling a wider range of computation.



Distributed Graph Representation: The graph (left) is represented as a vertex and an edge collection (right). The edges are divided into three edge partitions by applying a partition function (e.g., 2D Partitioning). The vertices are partitioned by vertex id. Copartitioned with the vertices, GraphX maintains a routing table encoding the edge partitions for each vertex. If vertex 6 and adjacent edges (shown with dotted lines) are restricted from the graph (e.g., by subgraph), they are removed from the corresponding collection by updating the bitmasks thereby enabling index reuse.

Managed Communication and Consistency for Fast Data-Parallel Iterative Analytics

Jinliang Wei, Wei Dai, Aurick Qiao, Qirong Ho, Henggang Cui, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, Eric P. Xing

ACM Symposium on Cloud Computing 2015. Aug. 27 - 29, 2015, Kohala Coast, HI.

At the core of Machine Learning (ML) analytics applied to Big Data is often an expert-suggested model, whose parameters are refined by iteratively processing a training dataset until convergence. The completion time (i.e. convergence time) and quality of the learned model not only depends on the rate at which the refinements are generated but also the quality of each refinement. While data-parallel ML applications often employ a loose consistency model when updating shared model parameters to maximize parallelism, the accumulated error may seriously impact the quality of refinements and thus delay completion time, a problem that usually gets worse with scale. Although more immediate propagation of updates reduces the accumulated error, this strategy is limited by physical network bandwidth. Additionally, the performance of the widely used stochastic gradient descent (SGD) algorithm is sensitive to initial step size, simply increasing communication without adjusting the step size value accordingly fails to achieve optimal performance.

This paper presents Bösen, a system that maximizes the network communication efficiency under a given inter-machine network bandwidth budget to minimize accumulated error, while ensuring theoretical convergence guarantees for large-scale data-parallel ML applications. Furthermore, Bösen prioritizes messages that are most significant to algorithm convergence, further enhancing algorithm convergence. Finally, Bösen is the first distributed implementation of the recently presented adaptive revision algorithm, which provides orders of magnitude improvement over a carefully tuned fixed schedule of step size refinements. Experiments on two clusters with up to 1024 cores show that our mechanism significantly improves upon static communication schedules.

On Model Parallelization and Scheduling Strategies for Distributed Machine Learning

Seunghak Lee, Jin Kyu Kim, Xun Zheng, Qirong Ho, Garth Gibson, Eric Xing

Proceedings of 2014 Neural Information Processing Systems (NIPS'14), December 2014.

Distributed machine learning has typically been approached from a data parallel perspective, where big data are partitioned to multiple workers and an algorithm is executed concurrently over different data subsets under various synchronization schemes to ensure speed-up and/or correctness. A sibling problem that has received relatively less attention is how to ensure efficient and correct model parallel execution of ML algorithms, where parameters of an ML program are partitioned to dif-

continued on pg. 12

Recent Publications

continued from pg. 11

ferent workers and undergone concurrent iterative updates. We argue that model and data parallelisms impose rather different challenges for system design, algorithmic adjustment, and theoretical analysis. In this paper, we develop a system for model-parallelism, STRADS, that provides a programming abstraction for scheduling parameter updates by discovering and leveraging changing structural properties of ML programs. STRADS enables a flexible tradeoff between scheduling efficiency and fidelity to intrinsic dependencies within the models, and improves memory efficiency of distributed ML. We demonstrate the efficacy of model-parallel algorithms implemented on STRADS versus popular implementations for topic modeling, matrix factorization, and Lasso.

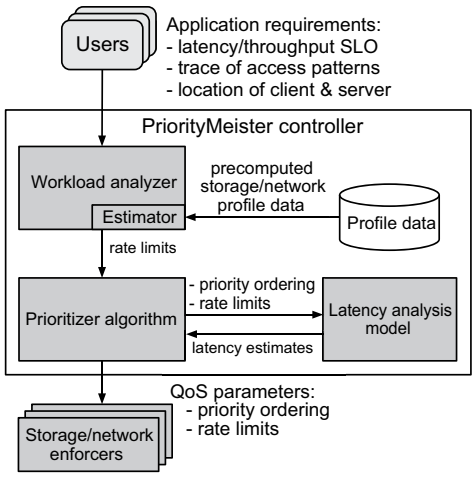
PriorityMeister: Tail Latency QoS for Shared Networked Storage

Timothy Zhu, Alexey Tumanov, Michael A. Kozuch, Mor Harchol-Balter, Gregory R. Ganger

Proceedings of ACM Symposium on Cloud Computing (SOCC'14), November 2014.

Meeting service level objectives (SLOs) for tail latency is an important and challenging open problem in cloud computing infrastructures. The challenges are exacerbated by burstiness in the workloads. This paper describes PriorityMeister–a system that employs a combination of per-workload priorities and rate limits to provide tail latency QoS for shared networked storage, even with bursty workloads.

PriorityMeister automatically and proactively configures workload priorities and rate limits across multiple stages (e.g., a shared storage stage followed by a shared network stage) to meet end-to-end tail latency SLOs. In real system experiments and under production trace workloads, PriorityMeister outperforms most recent reactive request scheduling approaches, with more workloads satisfying latency SLOs at higher latency percentiles. PriorityMeister is also robust to mis-estimation of underlying storage device



PriorityMeister controller dataflow diagram.

performance and contains the effect of misbehaving workloads.

Raising the Bar for Using GPUs in Software Packet Processing

Anuj Kalia, Dong Zhou, Michael Kaminsky, David G. Andersen

12th USENIX Symposium on Networked Systems Design and Implementation (NSDI'15), March 16-18, 2015 Santa Clara, CA.

Numerous recent research efforts have explored the use of Graphics Processing Units (GPUs) as accelerators for software-based routing and packet handling applications, typically demonstrating throughput several times higher than using legacy code on the CPU alone.

In this paper, we explore a new hypothesis about such designs: for many such applications, the benefits arise less from the GPU hardware itself as from the expression of the problem in a language such as CUDA or OpenCL that facilitates memory latency hiding and vectorization through massive concurrency. We demonstrate that in several cases, after applying a similar style of optimization to algorithm implementations, a CPU-only implementation is, in fact, more resource efficient than the version running on the GPU. To "raise the bar" for future uses of GPUs in packet processing applications, we present and evaluate a preliminary language/compiler-based framework called G-Opt that can accelerate CPU-based packet handling programs by automatically hiding memory access latency.

Reducing Latency via Redundant Requests: Exact Analysis

Kristen Gardner, Sam Zbarsky, Sherwin Doroudi, Mor Harchol-Balter, Esa Hyytiä, Alan Scheller-Wolf

Proceedings of ACM Sigmetrics/Performance 2015 Conference on Measurement and Modeling of Computer Systems (SIGMETRICS 15), Portland, OR. June 2015.

Recent computer systems research has proposed using redundant requests to reduce latency. The idea is to run a request on multiple servers and wait for the first completion (discarding all remaining copies of the request). However there is no exact analysis of systems with redundancy.

This paper presents the first exact analysis of systems with redundancy. We allow for any number of classes of redundant requests, any number of classes of non-redundant requests, any degree of redundancy, and any number of heterogeneous servers. In all cases we derive the limiting distribution on the state of the system.

In small (two or three server) systems, we derive simple forms for the distribution of response time of both the redundant classes and non-redundant classes, and we quantify the "gain" to redundant classes and "pain" to non-redundant classes caused by redundancy. We find some surprising results. First, the response time of a fully redundant class follows a simple Exponential distribution and that of the non-redundant class follows a Generalized Hyperexponential. Second, fully redundant classes are "immune" to any pain caused by other classes becoming redundant.

We also compare redundancy with other approaches for reducing latency, such as optimal probabilistic splitting of a class among servers (Opt-Split) and Join-the-Shortest-Queue (JSQ) routing of a class. We find that, in

many cases, redundancy outperforms JSQ and Opt-Split with respect to overall response time, making it an attractive solution.

Rethinking Data-Intensive Science Using Scalable Analytics Systems

Frank Austin Nothaft, Matt Massie, Timothy Danford, Zhao Zhang, Uri Laserson, Carl Yeksigian, Jey Kottalam, Arun Ahuja, Jeff Hammerbacher, Michael Linderman, Michael J. Franklin, Anthony D. Joseph, David A. Patterson

Proceedings of the 34th ACM SIGMOD International Conference on Management of Data (SIGMOD'15), May-June 2015.

"Next generation" data acquisition technologies are allowing scientists to collect exponentially more data at a lower cost. These trends are broadly impacting many scientific fields, including genomics, astronomy, and neuroscience. We can attack the problem caused by exponential data growth by applying horizontally scalable techniques from current analytics systems to accelerate scientific processing pipelines.

In this paper, we describe ADAM, an example genomics pipeline that leverages the open-source Apache Spark and Parquet systems to achieve a 28X speedup over current genomics pipelines, while reducing cost by 63%. From building this system, we were able to distill a set of techniques for implementing scientific analyses efficiently using commodity "big data" systems. To demonstrate the generality of our architecture, we then implement a scalable astronomy image processing system which achieves a 2.8-8.9X improvement over the state-of-the-art MPI-based system.

Reducing Replication Bandwidth for Distributed Document Databases

Lianghong Xu, Andrew Pavlo, Sudipta Sengupta, Jin Li, Gregory R. Ganger

ACM Symposium on Cloud Computing 2015. Aug. 27 - 29, 2015, Kohala Coast, HI.

With the rise of large-scale, Web-based applications, users are increasingly adopting a new class of document-oriented database management systems (DBMSs) that allow for rapid prototyping while also achieving scalable performance. Like for other distributed storage systems, replication is important for document DBMSs in order to guarantee availability. The network bandwidth required to keep replicas synchronized is expensive and is often a performance bottleneck. As such, there is a strong need to reduce the replication bandwidth, especially for geo-replication scenarios where wide-area network (WAN) bandwidth is limited.

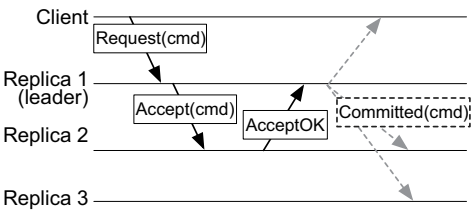
This paper presents a deduplication system called sDedup that reduces the amount of data transferred over the network for replicated document DBMSs. sDedup uses similarity-based deduplication to remove redundancy in replication data by delta encoding against similar documents selected from the entire database. It exploits key characteristics of document-oriented workloads, including small item sizes, temporal locality, and the incremental nature of document edits. Our experimental evaluation of sDedup with three real-world datasets shows that it is able to achieve up to 38× reduction in data sent over the network, significantly outperforming traditional chunk-based deduplication techniques while incurring negligible performance overhead.

Paxos Quorum Leases: Fast Reads Without Sacrificing Writes

Iulian Moraru, David G. Andersen, Michael Kaminsky

Proceedings of ACM Symposium on Cloud Computing (SOCC'14), November 2014. Best Paper Award.

This paper describes quorum leases, a new technique that allows Paxos-based systems to perform reads with high throughput and low latency. Quorum leases do not sacrifice consistency and have only a small impact on system availability and write latency. Quorum leases allow a majority of replicas to perform strongly consistent local reads, which substantially reduces read latency at those replicas (e.g., by two orders of magnitude in wide-area scenarios). Previous techniques for performing local reads in Paxos systems either (a) sacrifice consistency; (b) allow only one replica to read locally; or (c) decrease the availability of the system and increase the latency of all updates by requiring all replicas to be notified synchronously. We describe the design of quorum leases and evaluate their benefits compared to previous approaches through an implementation running in five geo-distributed Amazon EC2 datacenters.



Steady state interaction in Multi-Paxos. Asynchronous messages are represented as dashed arrows.

Scalable SPARQL Querying using Path Partitioning

Buwen Wu, Yongluan Zhou, Pingpeng Yuan, Ling Liu, Hai Jin

Proceedings of the 31st IEEE International Conference on Data Engineering (ICDE2015), April 13-16 2015, Seoul, Korea.

The emerging need for conducting complex analysis over big RDF datasets calls for scale-out solutions that can harness a computing cluster to process big RDF datasets. Queries over RDF data often involve complex self-joins, which would be very expensive to run if the data are not carefully partitioned across the cluster and hence distributed joins over massive amount of data are necessary. Existing RDF data partitioning methods can nicely localize simple queries but still need to resort to expensive distributed joins for more complex queries. In this pa-

continued on pg. 14

Recent Publications

continued from pg. 13

per, we propose a new data partitioning approach that takes use of the rich structural information in RDF datasets and minimizes the amount of data that have to be joined across different computing nodes. We conduct an extensive experimental study using two popular RDF benchmark data and one real RDF dataset that contain up to billions of RDF triples. The results indicate that our approach can produce a balanced and low redundant data partitioning scheme that can avoid or largely reduce the cost of distributed joins even for very complicated queries. In terms of query execution time, our approach can outperform the state-of-the-art methods by orders of magnitude.

SemStore: A Semantic-Preserving Distributed RDF Triple Store

Buwen Wu, Yongluan Zhou, Pingpeng Yuan, Hai Jin, Ling Liu

Proceedings of ACM International Conference on Information and Knowledge Management (CIKM'14), November 2014.

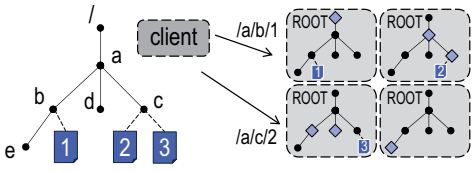
The flexibility of the RDF data model has attracted an increasing number of organizations to store their data in an RDF format. With the rapid growth of RDF datasets, we envision that it is inevitable to deploy a cluster of computing nodes to process large-scale RDF data in order to deliver desirable query performance. In this paper, we address the challenging problems of data partitioning and query optimization in a scale-out RDF engine. We identify that existing approaches only focus on using fine-grained structural information for data partitioning, and hence fail to localize many types of complex queries. We then propose a radically different approach, where a coarse-grained structure, namely Rooted Sub-Graph (RSG), is used as the partition unit. By doing so, we can capture structural information at a much greater scale and hence are able to localize many complex queries. We also propose a k-means partitioning algorithm for allocating the RSGs onto the computing nodes as well as a query optimization strategy to minimize the inter-node communication during query processing. An extensive experimental study using benchmark datasets and real dataset shows that our engine, SemStore, outperforms existing systems by orders of magnitudes in terms of query response time.

ShardFS vs. IndexFS: Replication vs. Caching Strategies for Distributed Metadata Management in Cloud Storage Systems

Lin Xiao, Kai Ren, Qing Zheng, Garth Gibson

ACM Symposium on Cloud Computing 2015. Aug. 27 - 29, 2015, Kohala Coast, HI.

The rapid growth of cloud storage systems calls for fast and scalable namespace processing. While few commercial file systems offer anything better than federating individually non-scalable namespace servers, a recent academic file system, IndexFS, demonstrates scalable namespace processing based on client caching of directory entries and permissions (directory lookup state) with no per-client state in servers. In this paper we explore explicit replication of directory lookup state in all servers as an alternative to caching this information in all clients. Both eliminate most repeated RPCs to different servers in order to resolve hierarchical permission tests. Our realization for server replicated directory lookup state, ShardFS, employs a novel file system specific hybrid optimistic and pessimistic concurrency control favoring single object transactions over distributed transactions. Our experimentation suggests that if directory lookup state mutation is a fixed fraction of operations (strong scaling for metadata), server replication does not scale as well as client caching, but if directory lookup state mutation is proportional to the number of jobs, not the number of processes per job, (weak scaling for metadata), then server replication can scale more linearly than client caching and provide lower 70 percentile response times as well.



ShardFS replicates directory lookup state to all metadata servers so every server can perform path resolution locally. File metadata and non-replicated directory metadata is stored at exactly one server determined by a hash function on the full pathname.

Scaling Distributed Machine Learning with the Parameter Server

Mu Li, David G. Andersen, Jun Woo Park, Alexander J. Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J. Shekita, Bor-Yiing Su

Proceedings of 11th USENIX OSDI (OSDI'14), October 2014.

We propose a parameter server framework for distributed machine learning problems. Both data and workloads are distributed over worker nodes, while the server nodes maintain globally shared parameters, represented as dense or sparse vectors and matrices. The framework manages asynchronous data communication between nodes, and supports flexible consistency models, elastic scalability, and continuous fault tolerance.

To demonstrate the scalability of the proposed framework, we show experimental results on petabytes of real data with billions of examples and parameters on problems ranging from Sparse Logistic Regression to Latent Dirichlet Allocation and Distributed Sketching.

Shared Memory Optimization in Virtualized Clouds

Qi Zhang, Ling Liu

IEEE 2015 International Conference on Cloud Computing. New York, June 27-July 1, 2015.

Shared memory management is widely recognized as an optimization technique in the virtualized cloud. Most current shared memory techniques allocate shared memory resources from guest VMs based on pre-defined system configurations. Such static management of shared memory not only

increases the VM memory pressure, but also limits the flexibility to balance the shared memory resources across multiple VMs running on a single host. In this paper, we present a dynamic shared memory management framework which enables multiple VMs to dynamically access the shared memory resource according to their demands. We illustrate our system design through two case studies: one aims at improving the performance of inter-domain communication while the other aims at improving VM memory swapping efficiency. We demonstrate that the dynamic shared memory mechanism not only improves the utilization of shared memory resources but also significantly enhances the performance of VM applications. Our experimental results show that by using dynamic shared memory management, we can improve the performance of inter-VM communication by up to 45 times, while mitigating the VM memory swapping overhead by up to 58%.

SMPFRAME: A Distributed Framework for Scheduled Model Parallel Machine Learning

Jin Kyu Kim, Qirong Ho, Seunghak Lee Xun Zheng, Wei Dai, Garth Gibson, Eric Xing

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-15-103, April 2015.

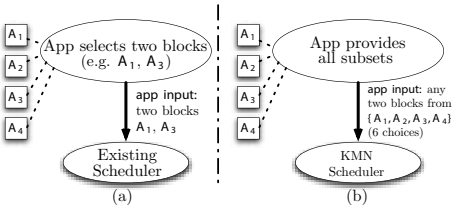
Machine learning (ML) problems commonly applied to big data by existing distributed systems share and update all ML model parameters at each machine using a partition of data—a strategy known as data-parallel. An alternative and complimentary strategy, model-parallel, partitions model parameters for non-shared parallel access and update, periodically repartitioning to facilitate communication. Model-parallelism is motivated by two challenges that data-parallelism does not usually address: (1) parameters may be dependent, thus naive concurrent updates can introduce errors that slow convergence or even cause algorithm failure; (2) model parameters converge at different rates, thus a small subset of parameters can bottleneck ML algorithm completion. We propose scheduled model parallellism (SMP), a programming approach where selection of parameters to be updated (the schedule) is explicitly separated from parameter update logic. The schedule can improve ML algorithm convergence speed by planning for parameter dependencies and uneven convergence. To support SMP at scale, we develop an archetype software framework SMPFRAME which optimizes the throughput of SMP programs, and benchmark four common ML applications written as SMP programs: LDA topic modeling, matrix factorization, sparse least-squares (Lasso) regression and sparse logistic regression. By improving ML progress per iteration through SMP programming whilst improving iteration throughput through SMPFRAME we show that SMP programs running on SMPFRAME outperform non-model-parallel ML implementations: for example, SMP LDA and SMP Lasso respectively achieve 10x and 5x faster convergence than recent, well-established baselines.

The Power of Choice in Data-Aware Cluster Scheduling

Shivaram Venkataraman, Aurojit Panda, Ganesh Ananthanarayanan, Michael J. Franklin, Ion Stoica

Proceedings of 11th USENIX OSDI (OSDI'14), October 2014.

Providing timely results in the face of rapid growth in data volumes has become important for analytical frameworks. For this reason, frameworks increasingly operate on only a subset of the input data. A key property of such sampling is that combinatorially many subsets of the input are present. We present KMN, a system that leverages these choices to perform data-aware scheduling, i.e., minimize time taken by tasks to read their inputs, for a DAG of tasks. KMN not only uses choices to co-locate tasks with their data but also percolates such combinatorial choices to downstream tasks in the DAG by launching a few additional tasks at every upstream stage. Evaluations using workloads from Facebook and Conviva on a 100-machine EC2 cluster show that KMN reduces average job duration by 81% using just 5% additional resources.



"Late binding" allows applications to specify more inputs than tasks and schedulers dynamically choose task inputs at execution time.

Using Data Transformations for Low-latency Time Series Analysis

Henggang Cui, Kimberly Keeton, Indrajit Roy, Krishnamurthy Viswanathan, Gregory R. Ganger

ACM Symposium on Cloud Computing 2015. Aug. 27 - 29, 2015, Kohala Coast, HI.

Time series analysis is commonly used when monitoring data centers, networks, weather, and even human patients. In most cases, the raw time series data is massive, from millions to billions of data points, and yet interactive analyses require low (e.g., sub-second) latency. Aperture transforms raw time series data, during ingest, into compact summarized representations that it can use to efficiently answer queries at runtime. Aperture handles a range of complex queries, from correlating hundreds of lengthy time series to predicting anomalies in the data. Aperture achieves much of its high performance by executing queries on data summaries, while providing a bound on the information lost when transforming data. By doing so, Aperture can reduce query latency as well as the data that needs to be stored and analyzed to answer a query. Our experiments on real data show that Aperture can provide one to four orders of magnitude lower query response time, while incurring only 10% ingest time overhead and less than 20% error in accuracy.

continued on pg. 16

Recent Publications

continued from pg. 15

Scheduling Multi-tenant Cloud Workloads on Accelerator-based Systems

Dipanjan Sengupta, Anshuman Goswami, Karsten Schwan

Proceedings of International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'14), November 2014.

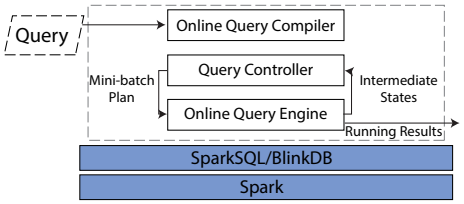
Accelerator-based systems are making rapid inroads into becoming platforms of choice for high end cloud services. There is a need therefore, to move from the current model in which high performance applications explicitly and programmatically select the GPU devices on which to run, to a dynamic model where GPUs are treated as first class schedulable entities. The Strings scheduler realizes this vision by decomposing the GPU scheduling problem into a combination of load balancing and per-device scheduling. (i) Device-level scheduling efficiently uses all of a GPU's hardware resources, including its computational and data movement engines, and (ii) load balancing goes beyond obtaining high throughput, to ensure fairness through prioritizing GPU requests that have attained least service. With its methods, Strings achieves improvements in system throughput and fairness of up to 8.70x and 13%, respectively, compared to the CUDA runtime.

G-OLA: Generalized On-Line Aggregation for Interactive Analysis on Big Data

Kai Zeng, Sameer Agarwal, Ankur Dave, Michael Armbrust, Ion Stoica

Proceedings of the 34th ACM SIGMOD International Conference on Management of Data (SIGMOD'15), May-June 2015. Demo paper.

Nearly 15 years ago, Hellerstein, Haas and Wang proposed online aggregation (OLA), a technique that allows users to (1) observe the progress of a query by showing iteratively refined approximate answers, and (2) stop the query execution once its result achieves the desired accuracy. In this demonstration, we present G-OLA, a



The system architecture of G-OLA as implemented in FluoDB.

novel mini-batch execution model that generalizes OLA to support general OLAP queries with arbitrarily nested aggregates using efficient delta maintenance techniques. We have implemented G-OLA in FluoDB, a parallel online query execution framework that is built on top of the Spark cluster computing framework that can scale to massive data sets. We will demonstrate FluoDB on a cluster of 100 machines processing roughly 10TB of real-world session logs from a video-sharing website. Using an ad optimization and an A/B testing based scenario, we will enable users to perform real-time data analysis via web-based query consoles and dashboards.

Lightweight Authentication of Freshness in Outsourced Key-Value Stores

Yuzhe Tang, Ting Wang, Ling Liu, Xin Hu, Jiyong Jang

Proceedings of 2014 Annual Computer Security Applications Conference (ACSAC'14), December 2014.

Data outsourcing offers cost-effective computing power to manage massive data streams and reliable access to data. Data owners can forward their data to clouds, and the clouds provide data mirroring, backup, and online access services to end users. However, outsourcing data to untrusted clouds requires data authenticity and query integrity to remain in the control of the data owners and users.

In this paper, we address the authenticated data-outsourcing problem specifically for multi-version key-value data that is subject to continuous updates under the constraints of data integrity, data authenticity, and "freshness" (i.e., ensuring that the value returned for a key is the latest version). We detail this problem and propose INCBM-TREE, a novel construct delivering freshness and authenticity.

Compared to existing work, we provide a solution that offers (i) lightweight signing and verification on massive data update streams for data owners and users (e.g., allowing for small memory footprint and CPU usage for a low-budget IT department), (ii) immediate authentication of data freshness, (iii) support of authentication in the presence of both real-time and historical data accesses. Extensive benchmark evaluations demonstrate that INCBM-TREE achieves higher throughput (in an order of magnitude) for data stream authentication than existing work. For data owners and end users that have limited computing power, INCBM-TREE can be a practical solution to authenticate the freshness of outsourced data while reaping the benefits of broadly available cloud services.

PipeCheck: Specifying and Verifying Microarchitectural Enforcement of Memory Consistency Models

Daniel Lustig, Michael Pellauer, Margaret Martonosi

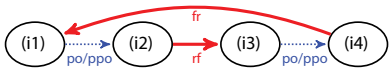
Proceedings of 47th International Symposium on Microarchitecture (MICRO'14). December, 2014. Best Paper Finalist.

We present PipeCheck, a methodology and automated tool for verifying that a particular microarchitecture correctly implements the consistency model required by its architectural specification. PipeCheck adapts the notion of a "happens before" graph from architecture-level analysis techniques to the microarchitecture space. Each node in the "microarchitecturally happens before" (μhb) graph represents not only a memory instruction, but also a particular location (e.g., pipeline stage) within the microarchitecture. Architectural specifications such as "preserved program order" are then treated as propositions to be verified, rather than simply as assumptions.

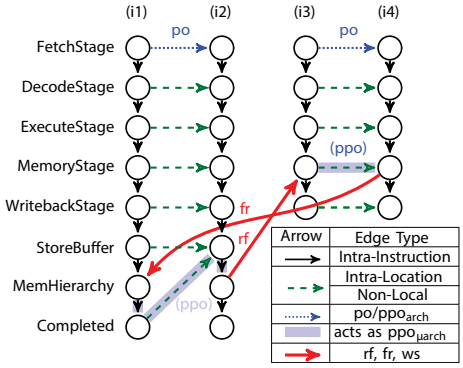
PipeCheck allows an architect to eas-

Core 0	Core 1
(i1) [x] ← 1	(i3) r1 ← [y]
(i2) [y] ← 1	(i4) r2 ← [x]
Under TSO: Forbid? r1=1, r2=0	

(a) Litmus Test Code



(b) Architecture-level analysis of one possible execution [3]. Note the presence of a cycle, indicating that this execution is forbidden.



(c) PipeCheck eliminates ppo as an assumption, and instead checks that it is replaced by calculated edge(s). In this example, the gray highlighted edges replace the ppo edges and complete a cycle.

Load→Load and Store→Store ordering litmus test.

ily and rigorously test whether a microarchitecture is stronger than, equal in strength to, or weaker than its architecturally-specified consistency model. We also specify and analyze the behavior of common microarchitectural optimizations such as speculative load reordering which technically violate formal architecture-level definitions. We evaluate PipeCheck using a library of established litmus tests on a set of open-source pipelines. Using PipeCheck, we were able to validate the largest pipeline, the OpenSPARC T2, in just minutes. We also identified a bug in the O3 pipeline of the gem5 simulator.

Deferred Lightweight Indexing for Log-Structured Key-Value Stores

Yuzhe Tang, Arun Iyengar, Wei Tan, Liana Fong, Balaji Palanisamy, Ling Liu

Proceedings of the 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid'15), May 2015, Shenzhen, Guangdong, China. Best paper award.

The recent shift towards write-intensive workload on big data (e.g., financial trading, social user-generated data streams) has pushed the proliferation of log-structured key-value stores, represented by Google's BigTable [1], Apache HBase [2] and Cassandra [3]. While providing key-based data access with a Put/Get interface, these key-value stores do not support value-based access methods, which significantly limits their applicability in modern web and database applications. In this paper, we present DELI, a DEferred Lightweight Indexing scheme on the log-structured key-value stores. To index intensively updated big data in real time, DELI aims at making the index maintenance as lightweight as possible. The key idea is to apply an append-only design for online index maintenance and to collect index garbage at carefully chosen time. DELI optimizes the performance of index garbage collection through tightly coupling its execution with a native routine process called compaction. The DELI's system design is fault-tolerant and generic (to most key-value stores); we implemented a prototype of DELI based on HBase without internal code modification. Our experiments show that the DELI offers significant performance advantage for the write-intensive index maintenance.

A Large-Scale Study of Flash Memory Errors in the Field

Justin Meza, Qiang Wu, Sanjeev Kumar, Onur Mutlu

Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS), Portland, OR, June 2015.

Servers use flash memory based solid state drives (SSDs) as a high-performance alternative to hard disk drives to store persistent data. Unfortunately, recent increases in flash density have also brought about decreases in chip-level reliability. In a data center environment, flash-based SSD failures can lead to downtime and, in the worst case, data loss. As a result, it is important to understand ash memory reliability characteristics over flash lifetime in a realistic production data center environment running modern applications and system software.

This paper presents the first large-scale study of flash-based SSD reliability in the field. We analyze data collected across a majority of flash-based solid state drives at Facebook data centers over nearly four years and many millions of operational hours in order to understand failure properties and trends of flash-based SSDs. Our study considers a variety of SSD characteristics, including: the amount of data written to and read from flash chips; how data is mapped within the SSD address space; the amount of data copied, erased, and discarded by the flash controller; and flash board temperature and bus power.

Based on our field analysis of how flash memory errors manifest when running modern workloads on modern SSDs, this paper is the first to make several major observations: (1) SSD failure rates do not increase monotonically with flash chip wear; instead they go through several distinct periods corresponding to how failures emerge and are subsequently detected, (2) the effects of read disturbance errors are not prevalent in the field, (3) sparse logical data layout across an SSD's physical address space (e.g., non-contiguous data), as measured by the amount of metadata required to track logical address translations stored in an SSD-internal DRAM buffer, can greatly affect SSD failure rate, (4) higher temperatures lead to higher failure rates, but techniques that throttle SSD operation appear to greatly reduce the negative reliability impact of higher temperatures, and (5) data written by the operating system to flash-based SSDs does not always accurately indicate the amount of wear induced on flash cells due to optimizations in the SSD controller and buffering employed in the system software. We hope that the findings of this first large-scale flash memory reliability study can inspire others to develop other publicly-available analyses and novel flash reliability solutions.

continued on pg. 18

Recent Publications

continued from pg. 17

A Top-Down Parallel Semisort

Yan Gu, Julian Shun, Yihan Sun, Guy Blelloch

ACM Symposium on Parallelism in Algorithms and Architecture. SPAA 2015, June 13 - 15, 2015.

Semisorting is the problem of reordering an input array of keys such that equal keys are contiguous but different keys are not necessarily in sorted order. Semisorting is important for collecting equal values and is widely used in practice. For example, it is the core of the MapReduce paradigm, is a key component of the database join operation, and has many other applications.

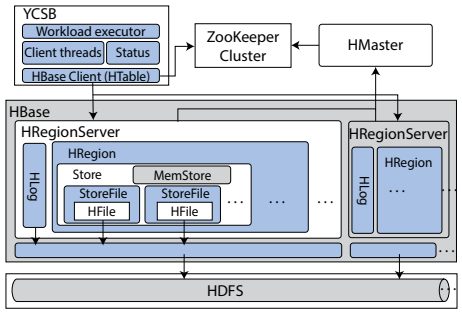
We describe a (randomized) parallel algorithm for the problem that is theoretically efficient (linear work and logarithmic depth), but is designed to be more practically efficient than previous algorithms. We use ideas from the parallel integer sorting algorithm of Rajasekaran and Reif, but instead of processing bits of a integers in a reduced range in a bottom-up fashion, we process the hashed values of keys directly top-down. We implement the algorithm and experimentally show on a variety of input distributions that it outperforms a similarly-optimized radix sort on a modern 40-core machine with hyper-threading by about a factor of 1.7-1.9, and achieves a parallel speedup of up to 38x. We discuss the various optimizations used in our implementation and present an extensive experimental analysis of its performance.

HConfig: Resource Adaptive Fast Bulk Loading in HBase

Xianqiang Bao, Ling Liu, Nong Xiao, Fang Liu, Qi Zhang, Tao Zhu

Proceedings of 10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'14), October 2014.

NoSQL (Not only SQL) data stores become a vital component in many big data computing platforms due to its inherent horizontal scalability. HBase is an open-source distributed NoSQL



HBase architecture combined with YCSB benchmark..

store that is widely used by many Internet enterprises to handle their big data computing applications (e.g. Facebook handles millions of messages each day with HBase). Optimizations that can enhance the performance of HBase are of paramount interests for big data applications that use HBase or Big Table like key-value stores. In this paper we study the problems inherent in misconfiguration of HBase clusters, including scenarios where the HBase default configurations can lead to poor performance. We develop HConfig, a semi-automated configuration manager for optimizing HBase system performance from multiple dimensions. Due to the space constraint, this paper will focus on how to improve the performance of HBase data loader using HConfig. Through this case study we will highlight the importance of resource adaptive and workload aware auto-configuration management and the design principles of HConfig. Our experiments show that the HConfig enhanced bulk loading can significantly improve the performance of HBase bulk loading jobs compared to the HBase default configuration, and achieve 2~3.7x speedup in throughput under different client threads while maintaining linear horizontal scalability.

Analyzing Log Analysis: An Empirical Study of User Log Mining

Sara Alspaugh, Betty Beidi Chen, Jessica Lin, Archana Ganapathi, Marti A. Hearst, Randy Katz

Proce. of Large Installation System Administration Conference (LISA'14), Nov. 2014. Best Student Paper Award.

We present an in-depth study of over 200K log analysis queries from Splunk, a platform for data analytics. Using these queries, we quantitatively describe log analysis behavior to inform the design of analysis tools. This study includes state machine based descriptions of typical log analysis pipelines, cluster analysis of the most common transformation types, and survey data about Splunk user roles, use cases, and skill sets. We find that log analysis primarily involves filtering, reformatting, and summarizing data and that non-technical users increasingly need data from logs to drive their decision making. We conclude with a number of suggestions for future research.

Communication-Efficient Multi-view Keyframe Extraction in Distributed Video Sensors

Shun-Hsing Ou, Yu-Chen Lu, Jui-Pin Wang, Shao-Yi Chien, Shou-De Lin, Mi-Yen Yeh, Chia-Han Lee, Phillip B. Gibbons, V. Srinivasa Somayazulu, Yen-Kuang Chen

Proceedings of IEEE Visual Communications and Image Processing Conference (VCIP'14), December 2014.

Video sensors are widely used in many applications such as security monitoring and home care. However, the growth of the number of sensors makes it impractical to stream all videos back to a central server for further processing, due to communication bandwidth and server storage constraints. Multi-view video summarization allows us to discard redundant data in the video streams taken by a group of sensors. All prior multi-view summarization methods, however, process video data in an off-line and centralized manner, which means that all videos are still required to be streamed back to the server before conducting the summarization. This paper proposes an on-line, distributed multi-view summarization system, which integrates the ideas of Maximal Marginal Relevance (MMR) and MS-Wave, a bandwidth-efficient distributed algorithm for finding k-nearest-neighbors and k-farthest-neighbors. Empirical studies show that our proposed system can discard redundant videos and keep important

keyframes as effectively as centralized approaches, while transmitting only 1/6 to 1/3 as much data.

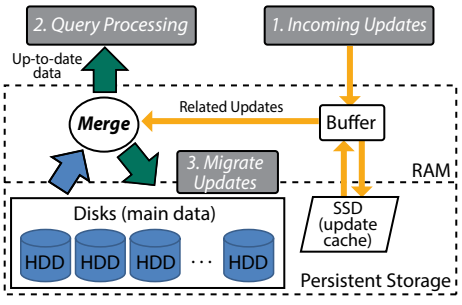
Online Updates on Data Warehouses via Judicious Use of Solid-State Storage

Manos Athanassoulis, Shimin Chen, Anastasia Ailamaki, Phillip B. Gibbons, Radu Stoica

ACM Transactions on Database Systems (TODS), March 2015.

Data warehouses have been traditionally optimized for read-only query performance, allowing only offline updates at night, essentially trading off data freshness for performance. The need for 24x7 operations in global markets and the rise of online and other quickly reacting businesses make concurrent online updates increasingly desirable. Unfortunately, state-of-the-art approaches fall short of supporting fast analysis queries over fresh data. The conventional approach of performing updates in place can dramatically slow down query performance, while prior proposals using differential updates either require large in-memory buffers or may incur significant update migration cost.

This article presents a novel approach for supporting online updates in data warehouses that overcomes the limitations of prior approaches by making judicious use of available SSDs to cache incoming updates. We model the problem of query processing with differential updates as a type of outer join between the data residing on disks and the updates residing on SSDs. We present MaSM algorithms for performing such joins and periodic migrations, with small memory footprints, low query overhead, low SSD writes, efficient in-place migration of updates, and correct ACID support. We present detailed modeling of the proposed approach, and provide proofs regarding the fundamental properties of the MaSM algorithms. Our experimentation shows that MaSM incurs only up to 7% overhead both on synthetic range scans (varying range size from 4KB to 100GB) and in a TPC-H query replay study, while also increasing the update throughput by orders of magnitude.



Framework for SSD-based differential updates.

Privacy-Preserving Multi-Keywords Search in Information Networks

Yuzhe Tang, Ling Liu

IEEE Transactions on Knowledge and Data Engineering (TKDE).

In emerging multi-domain cloud computing, it is crucially important to provide efficient search on distributed documents while preserving their owners' privacy, for which privacy preserving indexes or PPI presents a possible solution. An understudied problem for PPI techniques is how to provide differentiated privacy preservation in the face of multi-keyword document search. The differentiation is necessary as terms and phrases bear innate differences in their meanings. In this paper we present $\in$-MPPI, the first work on distributed document search with quantitative privacy preservation. In the design of $\in$-MPPI, we identified a suite of challenging problems and proposed novel solutions. For one, we formulated the quantitative privacy computation as an optimization problem that strikes a balance between privacy preservation and search efficiency. We also addressed the challenging problem of secure $\in$-MPPI construction in the multi-domain network which lacks mutual trusts between the domains. Towards a secure $\in$-MPPI construction with practical performance, we proposed techniques for improved performance of secure computations by making a novel use of secret sharing. We implemented the $\in$-MPPI construction protocol with a functioning prototype. We conducted extensive experiments to evaluate the prototype's effectiveness and efficiency based on a real-world dataset.

Reliable and Resilient Trust Management in Distributed Service Provision Networks

Zhiyuan Su, Ling Liu, Mingchu Li, Xinxin Fan, Yang Zhou

ACM Transactions on the Web, 2015.

Distributed service networks are popular platforms for service providers to offer services to consumers and for service consumers to acquire services from unknown parties. eBay and Amazon are two well-known examples of enabling and hosting such service networks to connect service providers to service consumers. Trust management is a critical component for scaling such distributed service networks to a large and growing number of participants. In this paper, we present ServiceTrust++, a feedback quality sensitive and attack resilient trust management scheme for empowering distributed service networks with effective trust management capability. Comparing with existing trust models, ServiceTrust++ has several novel features. First, we present six attack models to capture both independent and colluding attacks with malicious cliques, malicious spies and malicious camouflages. Second, we aggregate the feedback ratings based on the variances of participants' feedback behaviors and incorporate feedback similarity as weight into the local trust algorithm. Third, we compute the global trust of a participant by employing conditional trust propagation based on feedback similarity threshold. This allows ServiceTrust++ to control and prevent malicious spies and malicious camouflage peers to boost their global trust scores by manipulating the feedback ratings of good peers and by taking advantage of the uniform trust propagation. Finally, we systematically combine trust decaying strategy with threshold-value based conditional trust propagation to further strengthen the robustness of our global trust computation against sophisticated malicious feedbacks. Experimental evaluation

continued on pg. 20

Recent Publications

continued from pg. 19

with both simulation-based networks and real network dataset Epinion show that ServiceTrust++ is highly resilient against all six attack models and highly effective compared to EigenTrust, the most popular and representative trust propagation model to date.

Scaling Queries over Big RDF Graphs with Semantic Hash Partitioning

Kisung Lee, Ling Liu

Proceedings of the 40th IEEE International Conference on Very Large Databases (VLDB'14), Sept. 2014.

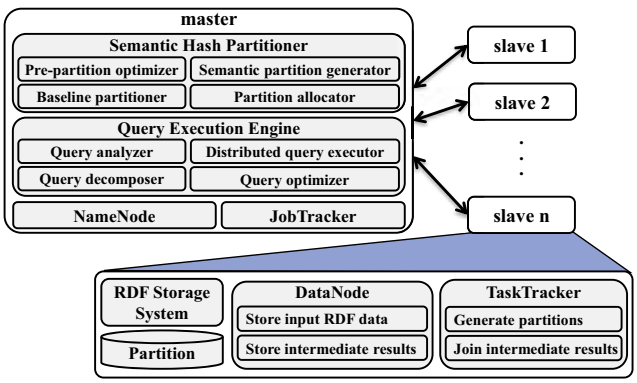
Massive volumes of big RDF data are growing beyond the performance capacity of conventional RDF data management systems operating on a single node. Applications using large RDF data demand efficient data partitioning solutions for supporting RDF data access on a cluster of compute nodes. In this paper we present a novel semantic hash partitioning approach and implement a Semantic HAsh Partitioning-Enabled distributed RDF data management system, called Shape. This paper makes three original contributions. First, the semantic hash partitioning approach we propose extends the simple hash partitioning method through direction-based triple groups and direction-based triple replications. The latter enhances the former by controlled data replication through intelligent utilization of data access locality, such that queries over big RDF graphs can be processed with zero or very small amount of inter-machine communication cost. Second, we generate locality-optimized query execution plans that are more efficient than popular multi-node RDF data management systems by effectively minimizing the inter-machine communication cost for query processing. Third but not the least, we provide a suite of locality-aware optimization techniques to further reduce the partition size and cut down on the inter-machine communication cost during distributed query processing. Experimental results show that our system scales well and can process big RDF datasets more efficiently than existing approaches.

CellIQ: Real-Time Cellular Network Analytics at Scale

Anand Padmanabha Iyer, Li Erran Li, Ion Stoica

Proceedings of the 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI'15), Oakland, CA, May 2015.

We present CellIQ, a real-time cellular network analytics system that supports rich and sophisticated analysis tasks. CellIQ is motivated by the lack of support for real-time analytics or advanced tasks such as spatio-temporal traffic hotspots and handoff sequences with performance problems in state-of-the-art systems, and the interest in such tasks by network operators. CellIQ represents cellular network data as a stream of domain specific graphs, each from a batch of data. Leveraging domain specific characteristics—the spatial and temporal locality of cellular network data—CellIQ presents a number of optimizations including geo-partitioning of input data, radius-based message broadcast, and incremental graph updates to support efficient analysis. Using data from a live cellular network and representative analytic tasks, we demonstrate that CellIQ enables fast and efficient cellular network analytics—compared to an implementation without cellular specific operators, CellIQ is 2× to 5× faster.



System architecture.

The Case for Offload Shaping

Wenlu Hu, Brandon Amos, Zhuo Chen, Kiryong Ha, Wolfgang Richter, Padmanabhan Pillai, Benjamin Gilbert, Jan Harkes, Mahadev Satyanarayanan

Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications (HotMobile'15), February 2015.

When offloading computation from a mobile device, we show that it can pay to perform additional on-device work in order to reduce the offloading workload. We call this offload shaping, and demonstrate its application at many different levels of abstraction using a variety of techniques. We show that offload shaping can produce significant reduction in resource demand, with little loss of application-level fidelity.

Value Driven Load Balancing

Sherwin Doroudi, Esa Hyytia, Mor Harchol-Balter

Performance Evaluation, vol. 79, September 2014.

To date, the study of dispatching or load balancing in server farms has primarily focused on the minimization of response time. Server farms are typically modeled by a front-end router that employs a dispatching policy to route jobs to one of several servers, with each server scheduling all the jobs in its queue via Processor-Sharing. However, the common assumption has been that all jobs are equally important or valuable, in that they are equally sensitive to delay. Our work departs from this assumption: we model each arrival as having a randomly distributed value parameter, independent of the arrival's service requirement (job size). Given such value heterogeneity, the correct metric is no longer the minimization or response time, but rather, the minimization of value-weighted response time. In this context, we ask "what is a good dispatching policy to minimize the value-weighted response time metric?" We propose a number of

new dispatching policies that are motivated by the goal of minimizing the value-weighted response time. Via a combination of exact analysis, asymptotic analysis, and simulation, we are able to deduce many unexpected results regarding dispatching.

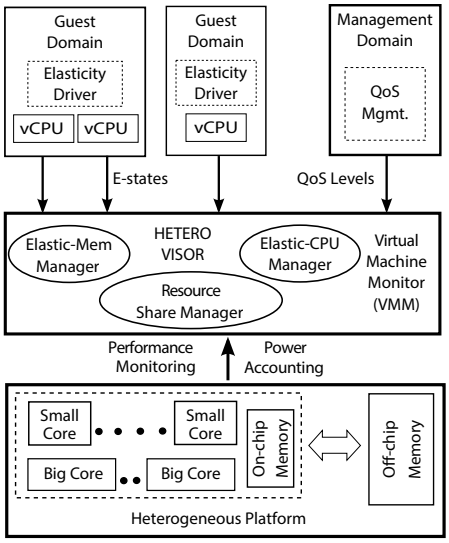
Nearly-Linear Work Parallel SDD Solvers, Low-Diameter Decomposition, and Low-Stretch Subgraphs

Guy E. Blelloch, Anupam Gupta, Ioannis Koutis, Gary L. Miller, Richard Peng, Kanat Tangwongsan

Theory of Computer Systems, Volume 55, Issue 3, October 2014.

We present the design and analysis of a nearly-linear work parallel algorithm for solving symmetric diagonally dominant (SDD) linear systems. On input an SDD n-by-n matrix A with m nonzero entries and a vector b, our algorithm computes a vector $\tilde{x}$ such that $\|\tilde{x}-A^{+}b\|_A \leq \varepsilon \cdot \|A^{+}b\|_A$ in $O(m \log^{O(1)} n \log 1/\varepsilon)$ work and $O(m^{1/3+\theta} \log 1/\varepsilon)$ depth for any $\theta > 0$, where A^{+} denotes the Moore-Penrose pseudoinverse of A. The algorithm relies on a parallel algorithm for generating low-stretch spanning trees or spanning subgraphs. To this end, we first develop a parallel decomposition algorithm that in $O(m \log^{O(1)} n)$ work and polylogarithmic depth, partitions a graph with n nodes and m edges into components with polylogarithmic diameter such that only a small fraction of the original edges are between the components. This can be used to generate low-stretch spanning trees with average stretch $O(n^{\alpha})$ in O(m logO(1) n) work and $O(n^{\alpha})$ depth for any $\alpha > 0$. Alternatively, it can be used to generate spanning subgraphs with polylogarithmic average stretch in $O(m \log^{O(1)} n)$ work and polylogarithmic depth. We apply this subgraph construction to derive a parallel linear solver.

By using this solver in known applications, our results imply improved parallel randomized algorithms for several problems, including single-source shortest paths, maximum flow, minimum-cost flow, and approximate maximum flow.



System architecture for HeteroVisor.

HeteroVisor: Exploiting Resource Heterogeneity to Enhance the Elasticity of Cloud Platforms

Vishal Gupta, Min Lee, Karsten Schwan

The 11th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE'15). March 14-15, 2015, Istanbul, Turkey.

This paper presents HeteroVisor, a heterogeneity-aware hypervisor, that exploits resource heterogeneity to enhance the elasticity of cloud systems. Introducing the notion of 'elasticity' (E) states, HeteroVisor permits applications to manage their changes in resource requirements as state transitions that implicitly move their execution among heterogeneous platform components. Masking the details of platform heterogeneity from virtual machines, the E-state abstraction allows applications to adapt their resource usage in a finegrained manner via VM-specific 'elasticity drivers' encoding VM-desired policies. The approach is explored for the heterogeneous processor and memory subsystems evolving for modern server platforms, leading to mechanisms that can manage these heterogeneous resources dynamically and as required by the different VMs being run. HeteroVisor is implemented for the Xen hypervisor, with mechanisms that go beyond core scaling to also deal with memory resources, via the online detection of hot memory pages and transparent page migration. Evaluation on an emulated heterogeneous platform uses workload traces from real-world data, demonstrating the ability to provide high on-demand performance while also reducing resource usage for these workloads.

Adversarial Active Learning

Brad Miller, Alex Kantchelian, Sadia Afroz, Rekha Bachwani, Edwin Dauber, Ling Huang, Michael Carl Tschantz, Anthony D. Joseph, J. Doug Tygar

Proceedings of 7th ACM Workshop on Artificial Intelligence and Security (AISec'14), held in conjunction with the 21st ACM Conference on Computer and Communications, November 2014.

Active learning is an area of machine learning examining strategies for allocation of nite resources, particularly human labeling efforts and to an extent feature extraction, in situations where available data exceeds available resources. In this open problem paper, we motivate the necessity of active learning in the security domain, identify problems caused by the application of present active learning techniques in adversarial settings, and propose a framework for experimentation and implementation of active learning systems in adversarial contexts. More than other contexts, adversarial contexts particularly need active learning as ongoing attempts to evade and confuse classifiers necessitate constant generation of labels for new content to keep pace with adversarial activity. Just as traditional machine learning algorithms are vulnerable to adversarial manipulation, we discuss assumptions specific to active learning that introduce additional vulnerabilities, as well as present vulnerabilities that are amplified in the active learning setting. Lastly, we present a software architecture, Security-oriented Active Learning Testbed (SALT), for the research and

continued on pg. 22

Recent Publications

continued from pg. 21

implementation of active learning applications in adversarial contexts.

Early Implementation Experience with Wearable Cognitive Assistance Applications

Zhuo Chen, Lu Jiang, Wenlu Hu, Kiryong Ha, Brandon Amos, Padmanabhan Pillai, Alex Hauptmann and Mahadev Satyanarayanan

First ACM SIGMOBILE workshop on Wearable Systems and Applications (WearSys 2015), Florence, Italy, May 2015.

A cognitive assistance application combines a wearable device such as Google Glass with cloudlet processing to provide step-by-step guidance on a complex task. In this paper, we focus on user assistance for narrow and well-defined tasks that require specialized knowledge and/or skills. We describe proof-of-concept implementations for four different tasks: assembling 2D Lego models, freehand sketching, playing ping-pong, and recommending context-relevant YouTube tutorials. We then reflect on the difficulties we faced in building these applications, and suggest future research that could simplify the creation of similar applications.

CrowdTrust: A Context-Aware Trust Model for Workers Selection in Crowdsourcing Environments

Bin Ye, Yan Wang, and Ling Liu

Proceedings of the 8th IEEE International Conference on Cloud Computing (CLOUD'15), New York, NY, June-July 2015.

On a crowdsourcing platform consisting of task publishers and workers, it is critical for a task publisher to select trustworthy workers to solve human intelligence tasks (HITs). Currently, the prevalent trust evaluation mechanism employs the overall approval rate of HITs, with which dishonest workers can easily succeed in pursuing the maximal profit by quickly giving plausible answers or counterfeiting HITs approval rates.

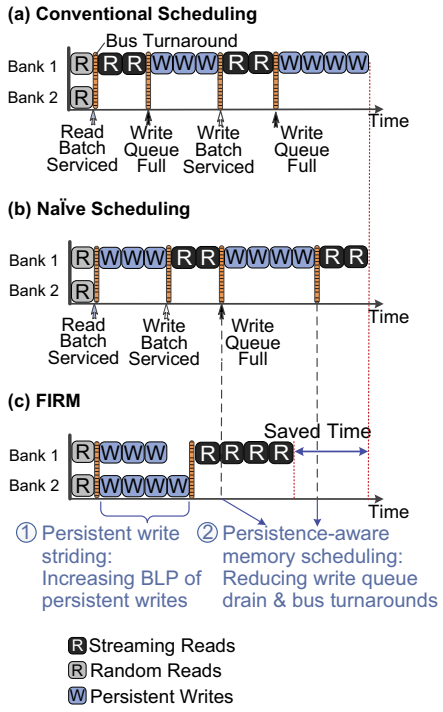
In crowdsourcing environments, a worker's trustworthiness varies in contexts, i.e. it varies in different types of tasks and different reward amounts of tasks. Thus, we propose two classifications based on task types and task reward amount respectively. On the basis of the classifications, we propose a trust evaluation model, which consists of two types of context-aware trust: task type based trust (TaTrust) and reward amount based trust (RaTrust). Then, we model trustworthy worker selection as a multi-objective combinatorial optimization problem, which is NP-hard. For solving this challenging problem, we propose an evolutionary algorithm MOWS GA based on NSGA-II. The results of experiments illustrate that our proposed trust evaluation model can effectively differentiate honest workers and dishonest workers when both of them have high overall HITs approval rates.

FIRM: Fair and High-Performance Memory Control for Persistent Memory Systems

Jishen Zhao, Onur Mutlu, Yuan Xie

Proceedings of 47th International Symposium on Microarchitecture (MICRO'14), December 2014.

Byte-addressable nonvolatile memories promise a new technology, persistent memory, which incorporates desirable attributes from both traditional main memory (byte-addressability and fast interface) and traditional storage (data persistence). To support data persistence, a persistent memory system requires sophisticated data duplication and ordering control for write requests. As a result, applications that manipulate persistent memory (persistent applications) have very different memory access characteristics than traditional (non-persistent) applications, as shown in this paper. Persistent applications introduce heavy write traffic to contiguous memory regions at a memory channel, which cannot concurrently service read and write requests, leading to memory bandwidth underutilization due to low bank-level parallelism, frequent write queue drains, and frequent bus turnarounds between reads and writes. These characteristics undermine the high-performance and



Example comparing conventional, naïve, and proposed schemes.

fairness offered by conventional memory scheduling schemes designed for non-persistent applications.

Our goal in this paper is to design a fair and high-performance memory control scheme for a persistent memory based system that runs both persistent and non-persistent applications. Our proposal, FIRM, consists of three key ideas. First, FIRM categorizes request sources as non-intensive, streaming, random and persistent, and forms batches of requests for each source. Second, FIRM strides persistent memory updates across multiple banks, thereby improving bank-level parallelism and hence memory bandwidth utilization of persistent memory accesses. Third, FIRM schedules read and write request batches from different sources in a manner that minimizes bus turnarounds and write queue drains. Our detailed evaluations show that, compared to five previous memory scheduler designs, FIRM provides significantly higher system performance and fairness.

Cuckoo Linear Algebra

Li Zhou, David G. Andersen, Mu Li, Alexander J. Smola

21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Aug 10-13, 2015, Sydney, Australia.

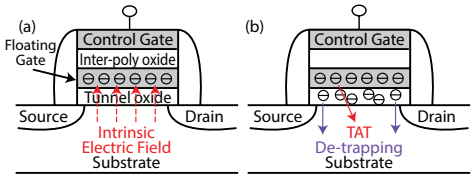
In this paper we present a novel data structure for sparse vectors based on Cuckoo hashing. It is highly memory efficient and allows for random access at near dense vector level rates. This allows us to solve sparse l1 [check] programming problems exactly and without preprocessing at a cost that is identical to dense linear algebra both in terms of memory and speed. Our approach provides a feasible alternative to the hash kernel and it excels whenever exact solutions are required, such as for feature selection.

Data Retention in MLC NAND Flash Memory: Characterization, Optimization and Recovery

Yu Cai, Yixin Luo, Erich F. Haratsch, Ken Mai, Onur Mutlu

Proceedings of the 21st International Symposium on High-Performance Computer Architecture (HPCA), Bay Area, CA, February 2015. Best paper session.

Retention errors, caused by charge leakage over time, are the dominant source of flash memory errors. Understanding, characterizing, and reducing retention errors can significantly improve NAND flash memory reliability and endurance. In this paper, we first characterize, with real 2y-nm MLC NAND flash chips, how the threshold voltage distribution of flash memory changes with different retention age – the length of time since a flash cell was programmed. We observe from our characterization results that 1) the optimal read reference voltage of a flash cell, using which the data can be read with the lowest raw bit error rate (RBER), systematically changes with its retention age, and 2) different regions of flash memory can have different retention ages, and hence different optimal read reference voltages. Based on our findings, we propose two new techniques. First, Retention Optimized Reading (ROR) adaptively learns and applies the optimal read reference voltage for each flash memory block online. The key idea of ROR is to periodically learn a tight upper bound, and from there approach the optimal read reference voltage. Our evaluations show that ROR can extend flash memory lifetime by 64% and reduce average error correction latency by 10.1%, with only 768 KB storage overhead in flash memory for a 512 GB flash-based SSD. Second, Retention Failure Recovery (RFR) recovers data with uncorrectable errors offline by identifying and probabilistically correcting flash cells with retention errors. Our evaluation shows that RFR reduces RBER by 50%, which essentially doubles the error correction capability, and thus can effectively recover data from otherwise uncorrectable flash errors.



Basics of NAND Flash Memory: (a) Cross-sectional view of a flash cell, (b) retention loss mechanisms.

GPU Performance and Power Tuning Using Regression Trees

Wenhao Jia, Elba Garza, Kelly A. Shaw, Margaret Martonosi

ACM Transactions on Architecture and Code Optimization (TACO), 12(2), June 2015.

GPU performance and power tuning is difficult, requiring extensive user expertise and time-consuming trial and error. To accelerate design tuning, statistical design space exploration methods have been proposed. This article presents Starchart, a novel design space partitioning tool that uses regression trees to approach GPU tuning problems. Improving on prior work, Starchart offers more automation in identifying key design trade-offs and models design subspaces with distinctly different behaviors. Starchart achieves good model accuracy using very few random samples: less than 0.3% of a given design space; iterative sampling can more quickly target subspaces of interest.

Adaptive-Latency DRAM: Optimizing DRAM Timing for the Common-Case

Donghyuk Lee, Yoongu Kim, Gennady Pekhimenko, Samira Khan, Vivek Seshadri, Kevin Chang, Onur Mutlu

Proceedings of the 21st International Symposium on High-Performance Computer Architecture (HPCA), Bay Area, CA, February 2015.

In current systems, memory accesses to a DRAM chip must obey a set of minimum latency restrictions specified in the DRAM standard. Such timing parameters exist to guarantee reliable operation. When deciding the timing parameters, DRAM manufacturers incorporate a very large margin as a provision against two worst-case scenarios. First, due to process variation, some outlier chips are much slower than others and cannot be operated as fast. Second, chips become slower at higher temperatures, and all chips need to operate reliably at the highest supported (i.e., worst-case) DRAM temperature (85°C). In this paper, we show that typical DRAM chips operating at typical temperatures (e.g., 55°C) are capable of providing a much smaller access latency, but are nevertheless forced to operate at the largest latency of the worst-case.

Our goal in this paper is to exploit the extra margin that is built into the DRAM timing parameters to improve performance. Using an FPGA-based testing platform, we first characterize the extra margin for 115 DRAM modules from three major manufacturers. Our results demonstrate that it is possible to reduce four of the most critical timing parameters by a minimum/maximum of 17.3%/54.8% at 55°C without sacrificing correctness. Based on this characterization, we propose Adaptive-Latency DRAM (AL-DRAM),

continued on pg. 24

Recent Publications

continued from pg. 23

a mechanism that adaptively reduces the timing parameters for DRAM modules based on the current operating condition. AL-DRAM does not require any changes to the DRAM chip or its interface.

We evaluate AL-DRAM on a real system that allows us to reconfigure the timing parameters at runtime. We show that AL-DRAM improves the performance of memory-intensive workloads by an average of 14% without introducing any errors. We discuss and show why AL-DRAM does not compromise reliability. We conclude that dynamically optimizing the DRAM timing parameters can reliably improve system performance.

ArMOR: Defending Against Consistency Model Mismatches in Heterogeneous Architectures

Daniel Lustig, Caroline Trippel, Michael Pellauer, Margaret Martonosi

42nd International Symposium on Computer Architecture (ISCA), June 2015.

Architectural heterogeneity is increasing: numerous products and studies have proven the benefits of combining cores and accelerators with varying ISAs into a single system. However, an underappreciated barrier to unlocking the full potential of heterogeneity is the need to specify and to reconcile differences in memory consistency models across layers of the hardware-software stack and among on-chip components.

This paper presents ArMOR, a framework for specifying, comparing, and translating between memory consistency models. ArMOR defines MOSTs, an architecture-independent and precise format for specifying the semantics of memory ordering requirements such as preserved program order or explicit fences. MOSTs allow any two consistency models to be directly and algorithmically compared, and they help avoid many of the pitfalls of traditional consistency model analysis. As a case study, we use ArMOR to automatically generate translation modules called shims that dynamically translate code compiled for one memory model to execute on hardware implementing a different model.

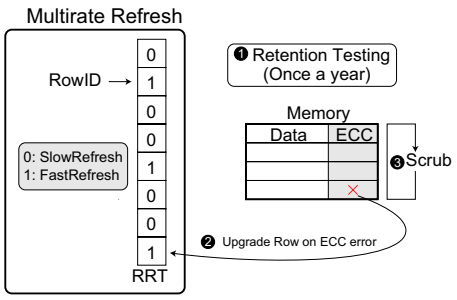
AVATAR: A Variable-Retention-Time (VRT) Aware Refresh for DRAM Systems

Moinuddin Qureshi, Dae Hyun Kim, Samira Khan, Prashant Nair, Chris Wilkerson, Onur Mutlu

Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Rio de Janeiro, Brazil, June 2015.

Multirate refresh techniques exploit the nonuniformity in retention times of DRAM cells to reduce the DRAM refresh overheads. Such techniques rely on accurate profiling of retention times of cells, and perform faster refresh only for a few rows which have cells with low retention times. Unfortunately, retention times of some cells can change at runtime due to Variable Retention Time (VRT), which makes it impractical to reliably deploy multirate refresh.

Based on experimental data from 24 DRAM chips, we develop architecture-level models for analyzing the impact of VRT. We show that simply relying on ECC DIMMs to correct VRT failures is unusable as it causes a data error once every few months. We propose AVATAR, a VRT-aware multirate refresh scheme that adaptively changes the refresh rate for different rows at runtime based on current VRT failures. AVATAR provides a time to failure in the regime of several tens of years while reducing refresh operations by 62%-72%.



Design of AVATAR.

Exploiting Compressed Block Size as an Indicator of Future Reuse

Gennady Pekhimenko, Tyler Huberty, Rui Cai, Onur Mutlu, Phillip P. Gibbons, Michael A. Kozuch, Todd C. Mowry

Proceedings of the 21st International Symposium on High-Performance Computer Architecture (HPCA), Bay Area, CA, February 2015.

We introduce a set of new Compression-Aware Management Policies (CAMP) for on-chip caches that employ data compression. Our management policies are based on two key ideas. First, we show that it is possible to build a more efficient management policy for compressed caches if the compressed block size is directly used in calculating the value (importance) of a block to the cache. This leads to Minimal-Value Eviction (MVE), a policy that evicts the cache blocks with the least value, based on both the size and the expected future reuse. Second, we show that, in some cases, compressed block size can be used as an efficient indicator of the future reuse of a cache block. We use this idea to build a new insertion policy called Size-based Insertion Policy (SIP) that dynamically prioritizes cache blocks using their compressed size as an indicator.

We compare CAMP (and its global variant G-CAMP) to prior on-chip cache management policies (both size-oblivious and size-aware) and find that our mechanisms are more effective in using compressed block size as an extra dimension in cache management decisions. Our results show that the proposed management policies (i) decrease off-chip bandwidth consumption (by 8.7% in single-core), (ii) decrease memory subsystem energy consumption (by 7.2% in single-core) for memory intensive workloads compared to the best prior mechanism, and (iii) improve performance (by 4.9%/9.0%/10.2% on average in single-/two-/four-core workload evaluations and up to 20.1%) CAMP is effective for a variety of compression algorithms and different cache designs with local and global replacement strategies.

Fast Bulk Bitwise AND and OR in DRAM

Vivek Seshadri, Kevin Hsieh, Amirali Boroumand, Donghyuk Lee, Michael A. Kozuch, Onur Mutlu, Phillip B. Gibbons, Todd C. Mowry

IEEE Computer Architecture Letters (CAL), April 2015.

Bitwise operations are an important component of modern day programming, and are used in a variety of applications such as databases. In this work, we propose a new and simple mechanism to implement bulk bitwise AND and OR operations in DRAM, which is faster and more efficient than existing mechanisms. Our mechanism exploits existing DRAM operation to perform a bitwise AND/OR of two DRAM rows completely within DRAM. The key idea is to simultaneously connect three cells to a bitline before the sense-amplification. By controlling the value of one of the cells, the sense amplifier forces the bitline to the bitwise AND or bitwise OR of the values of the other two cells. Our approach can improve the throughput of bulk bitwise AND/OR operations by 9.7X and reduce their energy consumption by 50.5X. Since our approach exploits existing DRAM operation as much as possible, it requires negligible changes to DRAM logic. We evaluate our approach using a real-world implementation of a bit-vector based index for databases. Our mechanism improves the performance of commonly-used range queries by 30% on average.

Fast Iterative Graph Computation: A Path Centric Approach

Pingpeng Yuan, Wenya Zhang, Changfeng Xie, Hai Jin, Ling Liu and Kisung Lee

Proceedings of IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'14), November 2014.

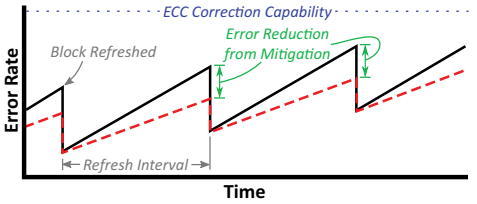
Large scale graph processing represents an interesting systems challenge due to the lack of locality. This paper presents PathGraph, a system for improving iterative graph computation on graphs with billions of edges. Our system design has three unique features: First, we model a large graph using a collection of tree-based partitions and use pathcentric computation rather than vertex-centric or edge-centric computation. Our path-centric graph parallel computation model significantly improves the memory and disk locality for iterative computation algorithms on large graphs. Second, we design a compact storage that is optimized for iterative graph parallel computation. Concretely, we use delta-compression, partition a large graph into tree-based partitions and store trees in a DFS order. By clustering highly correlated paths together, we further maximize sequential access and minimize random access on storage media. Third but not the least, we implement the path-centric computation model by using a scatter/gather programming model, which parallels the iterative computation at partition tree level and performs sequential local updates for vertices in each tree partition to improve the convergence speed. We compare PathGraph to most recent alternative graph processing systems such as GraphChi and X-Stream, and show that the path-centric approach outperforms vertex-centric and edge-centric systems on a number of graph algorithms for both in-memory and out-of-core graphs.

Read Disturb Errors in MLC NAND Flash Memory: Characterization and Mitigation

Yu Cai, Yixin Luo, Saugata Ghose, Erich F. Haratsch, Ken Mai, Onur Mutlu

Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Rio de Janeiro, Brazil, June 2015.

NAND flash memory reliability continues to degrade as the memory is scaled down and more bits are programmed per cell. A key contributor to this reduced reliability is read disturb, where a read to one row of cells impacts the threshold voltages of unread flash cells in different rows of the same block.



Exaggerated example of how read disturb mitigation reduces error rate peaks for each refresh interval. Solid black line is the unmitigated error rate, and dashed red line is the error rate after mitigation. (Note that the error rate does not include read errors introduced by reducing V_{pass}, as the unused error correction capability can tolerate errors caused by V_{pass} *Tuning*.)

Such disturbances may shift the threshold voltages of these unread cells to different logical states than originally programmed, leading to read errors that hurt endurance.

For the first time in open literature, this paper experimentally characterizes read disturb errors on state-of-the-art 2Y-nm (i.e., 20-24 nm) MLC NAND flash memory chips. Our findings (1) correlate the magnitude of threshold voltage shifts with read operation counts, (2) demonstrate how program/erase cycle count and retention age affect the read-disturb-induced error rate, and (3) identify that lowering pass-through voltage levels reduces the impact of read disturb and extend flash lifetime. Particularly, we find that the probability of read disturb errors increases with both higher wear-out and higher pass-through voltage levels.

We leverage these findings to develop two new techniques. The first technique mitigates read disturb errors by dynamically tuning the pass-through voltage on a per-block basis. Using real workload traces, our evaluations show that this technique increases flash memory endurance by an average of 21%. The second technique recovers from previously-uncorrectable flash errors by identifying and probabilistically correcting cells susceptible to read disturb errors. Our evaluations show that this recovery technique reduces the raw bit error rate by 36%.

continued on pg. 26

Recent Publications

continued from pg. 25

Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field

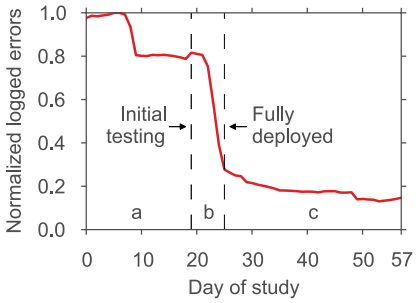
Justin Meza, Qiang Wu, Sanjeev Kumar, Onur Mutlu

Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Rio de Janeiro, Brazil, June 2015.

Computing systems use dynamic random-access memory (DRAM) as main memory. As prior works have shown, failures in DRAM devices are an important source of errors in modern servers. To reduce the effects of memory errors, error correcting codes (ECC) have been developed to help detect and correct errors when they occur. In order to develop effective techniques, including new ECC mechanisms, to combat memory errors, it is important to understand the memory reliability trends in modern systems.

In this paper, we analyze the memory errors in the entire fleet of servers at Facebook over the course of fourteen months, representing billions of device days. The systems we examine cover a wide range of devices commonly used in modern servers, with DIMMs manufactured by 4 vendors in capacities ranging from 2GB to 24GB that use the modern DDR3 communication protocol.

We observe several new reliability trends for memory systems that have not been discussed before in literature. We show that (1) memory errors follow a power-law, specifically, a Pareto distribution with decreasing hazard rate, with average error rate exceeding median error rate by around 55X; (2) non-DRAM memory failures from the memory controller and memory channel cause the majority of errors, and the hardware and software overheads to handle such errors cause a kind of denial of service attack in some servers; (3) using our detailed analysis, we provide the first evidence that more recent DRAM cell fabrication technologies (as indicated by chip density) have substantially higher failure rates, increasing by 1.8X over the previous generation; (4) DIMM architecture decisions affect memory reliability: DIMMs with fewer chips and lower transfer widths have the lowest error rates, likely due to electrical noise reduction; (5) while CPU and memory utilization do not show clear trends with respect to failure rates, workload type can influence failure rate by up to 6.5X, suggesting certain memory access patterns may induce more errors; (6) we develop a model for memory reliability and show how system design choices such as using lower density DIMMs and fewer cores per chip can reduce failure rates of a baseline server by up to 57.7%; and (7) we perform the first implementation and real-system analysis of page offlining at scale, showing that it can reduce memory error rate by 67%, and identify several real-world impediments to the technique.



The effect of page offlining on error rate. Region a shows the state of the servers before page offlining was deployed. Region b shows the state of the servers while page offlining was deployed gradually to 100% of the servers (so that any malfunctions of the deployment could be detected in a small number of machines and not all of them). Region c shows the state of the servers after page offlining was fully deployed.

Feral Concurrency Control: An Empirical Investigation of Modern Application Integrity

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

Proceedings of the 34th ACM SIGMOD International Conference on Management of Data (SIGMOD'15), May-June 2015.

The rise of data-intensive "Web 2.0" Internet services has led to a range of popular new programming frameworks that collectively embody the latest incarnation of the vision of Object-Relational Mapping (ORM) systems, albeit at unprecedented scale. In this work, we empirically investigate modern ORM-backed applications' use and disuse of database concurrency control mechanisms. Specifically, we focus our study on the common use of feral, or application-level, mechanisms for maintaining database integrity, which, across a range of ORMsystems, often take the form of declarative correctness criteria, or invariants. We quantitatively analyze the use of these mechanisms in a range of open source applications written using the Ruby on Rails ORM and find that feral invariants are the most popular means of ensuring integrity (and, by usage, are over 37 times more popular than transactions). We evaluate which of these feral invariants actually ensure integrity (by usage, up to 86:9%) and which—due to concurrency errors and lack of database support—may lead to data corruption (the remainder), which we experimentally quantify. In light of these findings, we present recommendations for database system designers for better supporting these modern ORM programming patterns, thus eliminating their adverse effects on application integrity.

Record Placement Based on Data Skew Using Solid State Drives

Jun Suzuki, Shivaram Venkataraman, Sameer Agarwal, Michael J. Franklin, Ion Stoica

Proceedings of Fifth workshop on Big Data Benchmarks, Performance Optimization, and Emerging Hardware (BPOE'14) at VLDB, September 2014.

Integrating a solid state drive (SSD) into a data store is expected to improve its I/O performance. However, there is still a large difference between the price of an SSD and a hard-disk drive (HDD). One of the methods to offset the increase in cost of consisting devices is to configure a hybrid system using both devices. In such a system, a common method to decide the placement of data records is based on reference locality, i.e., placing the

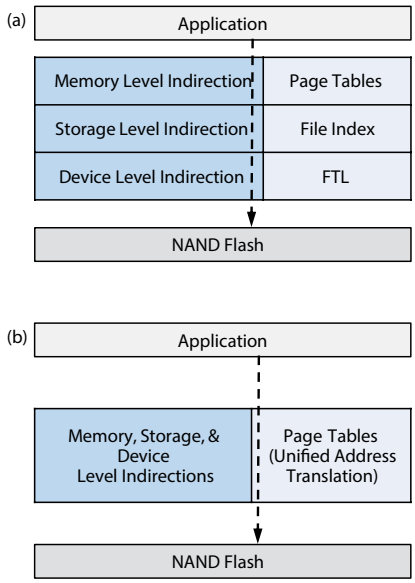
frequently accessed records in a faster SSD. In this paper, we propose an alternative that focuses on data skew by storing records with values that appear less often in an SSD while those that do more in an HDD. As we will show, this enhances the performance of fetching records using multi-dimensional indices. When records are fetched using one of the indices targeted for optimization, records stored in an SSD are likely be retrieved using random access, while those stored in an HDD using sequential access. Given the method does not rely on reference locality, its performance is stable between first and second accesses and it provides a performance gain even when a host memory is large enough to contain the entire working set of the application. Our implementation and experiments show that storing just 20% records in an SSD achieves up to 76% of the maximum reduction that would otherwise be obtained when all the records are stored in an SSD.

Unified Address Translation for Memory-Mapped SSDs with FlashMap

Jian Huang, Anirudh Badam, Moinuddin K. Qureshi, Karsten Schwan

Proceedings of the 42nd ACM International Symposium on Computer Architecture (ISCA'15), Portland, OR, June 2015.

Applications can map data on SSDs into virtual memory to transparently scale beyond DRAM capacity, permitting them to leverage high SSD capacities with few code changes. Obtaining good performance for memory-mapped SSD content, however, is hard because the virtual memory layer, the file system and the flash translation layer (FTL) perform address translations, sanity and permission checks independently from each other. We introduce FlashMap, an SSD interface that is optimized for memory-mapped SSD-files. FlashMap combines all the address translations into page tables that are used to index files and also to store the FTL-level mappings without altering the guarantees of the file system or the FTL. It uses



Comparison of (a) conventional memory-mapped SSD-file's IO stack and (b) FlashMap that combines all the address translations for mapping files on SSD into page tables.

the state in the OS memory manager and the page tables to perform sanity and permission checks respectively. By combining these layers, FlashMap reduces critical-path latency and improves DRAM caching efficiency. We find that this increases performance for applications by up to 3.32x compared to state-of-the-art SSD file-mapping mechanisms. Additionally, latency of SSD accesses reduces by up to 53.2%.

Research Problems and Opportunities in Memory Systems

Onur Mutlu, Lavanya Subramanian

Invited Article in Supercomputing Frontiers and Innovations (SUPERFRI), 2014.

The memory system is a fundamental performance and energy bottleneck in almost all computing systems. Recent system design, application, and technology trends that require more capacity, bandwidth, efficiency, and predictability out of the memory system make it an even more important system bottleneck. At the same time, DRAM technology is experiencing difficult technology scaling challenges that make the maintenance and enhancement of its capacity, energy efficiency, and reliability significantly more costly with conventional techniques.

In this article, after describing the demands and challenges faced by the memory system, we examine some promising research and design directions to overcome challenges posed by memory scaling. Specifically, we describe three major new research challenges and solution directions: 1) enabling new DRAM architectures, functions, interfaces, and better integration of the DRAM and the rest of the system (an approach we call system-DRAM co-design), 2) designing a memory system that employs emerging non-volatile memory technologies and takes advantage of multiple different technologies (i.e., hybrid memory systems), 3) providing predictable performance and QoS to applications sharing the memory system (i.e., QoS-aware memory systems). We also briefly describe our ongoing related work in combating scaling challenges of NAND flash memory.

Sequential Random Permutation, List Contraction and Tree Contraction are Highly Parallel

Julian Shun, Yan Gu, Guy E. Blelloch, Jeremy T. Fineman Phillip B. Gibbons

SODA 2015. January 4-6, 2015, San Diego, CA.

We show that simple sequential randomized iterative algorithms for random permutation, list contraction, and tree contraction are highly parallel. In particular, if iterations of the algorithms are run as soon as all of their dependencies have been resolved, the resulting computations have logarithmic depth (parallel time) with high probability. Our proofs make an interesting connection between the dependence structure of two of the problems and random binary trees. Building upon this analysis, we describe linear-work, polylogarithmic-depth algorithms for the three problems. Although asymptotically no better than the many prior

continued on pg. 28

Recent Publications

continued from pg. 27

parallel algorithms for the given problems, their advantages include very simple and fast implementations, and returning the same result as the sequential algorithm. Experiments on a 40-core machine show reasonably good performance relative to the sequential algorithms.

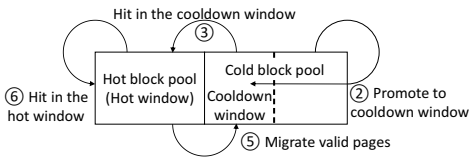
WARM: Improving NAND Flash Memory Lifetime with Write-hotness Aware Retention Management

Yixin Luo, Yu Cai, Saugata Ghose, Jongmoo Choi, Onur Mutlu

Proceedings of the 31st International Conference on Massive Storage Systems and Technologies (MSST), Santa Clara, CA, June 2015.

Increased NAND flash memory density has come at the cost of lifetime reductions. Flash lifetime can be extended by relaxing internal data retention time, the duration for which a flash cell correctly holds data. Such relaxation cannot be exposed externally to avoid altering the expected data integrity property of a flash device. Reliability mechanisms, most prominently refresh, restore the duration of data integrity, but greatly reduce the lifetime improvements from retention time relaxation by performing a large number of write operations. We find that retention time relaxation can be achieved more efficiently by exploiting heterogeneity in write-hotness, i.e., the frequency at which each page is written.

We propose WARM, a write-hotness aware retention management policy for flash memory, which identifies and physically groups together write-hot data within the flash device, allowing the flash controller to selectively perform retention time relaxation with little cost. When applied alone, WARM improves overall flash lifetime by an average of 3.24X over a conventional management policy without refresh, across a variety of real I/O workload traces. When WARM is applied together with an adaptive refresh mechanism, the average lifetime improves by 12.9X, 1.21X over adaptive refresh alone.



Write-hotness aware retention management policy overview.

The Blacklisting Memory Scheduler: Achieving High Performance and Fairness at Low Cost

Lavanya Subramanian, Donghyuk Lee, Vivek Seshadri, Harsha Rastogi, Onur Mutlu

Proceedings of 32nd IEEE International Conference on Computer Design (ICCD'14), October 2014.

In a multicore system, applications running on different cores interfere at main memory. This inter-application interference degrades overall system performance and unfairly slows down applications. Prior works have developed applicationaware memory request schedulers to tackle this problem. State-of-the-art application-aware memory request schedulers prioritize memory requests of applications that are vulnerable to interference, by ranking individual applications based on their memory access characteristics and enforcing a total rank order.

In this paper, we observe that state-of-the-art application-aware memory schedulers have two major shortcomings. First, ranking applications individually with a total order based on memory access characteristics leads to high hardware cost and complexity. Second, ranking can unfairly slow down applications that are at the bottom of the ranking stack. To overcome these shortcomings, we propose the Blacklisting Memory Scheduler (BLISS), which achieves high system performance and fairness while incurring low hardware cost and complexity. BLISS design is based on two new observations. First, we find that, to mitigate interference, it is sufficient to separate applications into only two groups, one containing applications that cause interference and another containing applications vulnerable to interference, instead of ranking individual applications with a total order. Vulnerable-to-interference group is prioritized over the interference-causing group. Second, we show that this grouping can be efficiently performed by simply counting the number of consecutive requests served from each application – an application that has a large number of consecutive requests served is dynamically classified as interference-causing.

We evaluate BLISS across a wide variety of workloads and system configurations and compare its performance and complexity with five state-of-the-art memory schedulers. Our evaluations show that BLISS achieves 5% better system performance and 25% better fairness than the best-performing previous memory scheduler while greatly reducing critical path latency and hardware area cost of the memory scheduler (by 79% and 43%, respectively).

Sorting with Asymmetric Read and Write Costs

Guy E. Blelloch, Jeremy T. Fineman, Phillip B. Gibbons, Yan Gu, Julian Shun

27th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA15), June 13 - 15, 2015, Portland, OR.

Emerging memory technologies have a significant gap between the cost, both in time and in energy, of writing to memory versus reading from memory. In this paper we present models and algorithms that account for this difference, with a focus on write-efficient sorting algorithms. First, we consider the PRAM model with asymmetric write cost, and show that sorting can be performed in O(n) writes, O(n log n) reads, and logarithmic depth (parallel time). Next, we consider a variant of the External Memory (EM) model that charges $k > 1$ for writing a block of size B to the secondary memory, and present variants of three EM sorting algorithms (multi-way merge-sort, sample sort, and heap-sort using buffer trees) that asymptotically reduce the number

of writes over the original algorithms, and perform roughly k block reads for every block write. Finally, we define a variant of the Ideal-Cache model with asymmetric write costs, and present write-efficient, cache-oblivious parallel algorithms for sorting, FFTs, and matrix multiplication. Adapting prior bounds for work-stealing and parallel-depth-first schedulers to the asymmetric setting, these yield parallel cache complexity bounds for machines with private caches or with a shared cache, respectively.

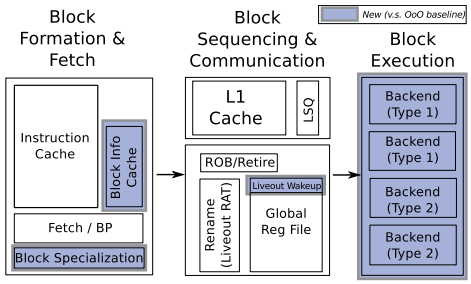
The Heterogeneous Block Architecture

Chris Fallin, Chris Wilkerson, Onur Mutlu

Proceedings of 32nd IEEE International Conference on Computer Design (ICCD'14), October 2014.

This paper makes two observations that lead to a new heterogeneous core design. First, we observe that most serial code exhibits fine-grained heterogeneity: at the scale of tens or hundreds of instructions, regions of code fit different microarchitectures better (at the same point or at different points in time). Second, we observe that by grouping contiguous regions of instructions into blocks that are executed atomically, a core can exploit this fine-grained heterogeneity: atomicity allows each block to be executed independently on its own execution backend that fits its characteristics best.

Based on these observations, we propose a fine-grained heterogeneous core design, called the heterogeneous block architecture (HBA), that combines heterogeneous execution backends into one core. HBA breaks the program into blocks of code, determines the best backend for each block, and specializes the block for that backend. As an example HBA design, we combine out-of-order, VLIW, and in-order backends, using simple heuristics to choose backends for different dynamic instruction blocks. Our extensive evaluations compare this example HBA design to multiple baseline core designs (including monolithic out-of-order, clustered out-of-order, in-order and a state-of-the-art heterogeneous core design) and show that it provides significantly better energy efficiency than all designs at similar performance.



HBA (Heterogeneous Block Architecture) overview.

Verifying Correct Microarchitectural Enforcement of Memory Consistency Models

Daniel Lustig, Michael Pellauer, Margaret Martonosi

IEEE Micro, 35 (3) (Top Picks of 2014), May-June 2015.

Memory consistency models define the rules and guarantees about the ordering and visibility of memory references on multi-threaded CPUs and systems-on-chip (SoCs). PipeCheck offers a methodology and automated tool for verifying that a particular microarchitecture correctly implements the consistency model required by its architectural specification.

Efficient Data Mapping and Buffering Techniques for Multi-Level Cell Phase-Change Memories

HanBin Yoon, Justin Meza, Naveen Muralimanohar, Norman P. Jouppi, Onur Mutlu

ACM Transactions on Architecture and Code Optimization (TACO), Vol. 11, No. 4, December 2014. Best (student) presentation award.

New phase-change memory (PCM) devices have low-access latencies (like DRAM) and high capacities (i.e., low cost per bit, like Flash). In addition to being able to scale to smaller cell sizes than DRAM, a PCM cell can also store multiple bits per cell (referred to as multilevel cell, or MLC), enabling even greater capacity per bit. However, reading and writing the different bits of data from and to an MLC PCM cell requires different amounts of time: one bit is read or written first, followed by another. Due to this asymmetric access process, the bits in an MLC PCM cell have different access latency and energy depending on which bit in the cell is being read or written.

We leverage this observation to design a new way to store and buffer data in MLC PCM devices. While traditional devices couple the bits in each cell next to one another in the address space, our key idea is to logically decouple the bits in each cell into two separate regions depending on their read/write characteristics: fast-read/slow-write bits and slow-read/fast-write bits. We propose a low-overhead hardware/software technique to predict and map data that would benefit from being in each region at runtime. In addition, we show how MLC bit decoupling provides more flexibility in the way data is buffered in the device, enabling more efficient use of existing device buffer space.

Our evaluations for a multicore system show that MLC bit decoupling improves system performance by 19.2%, memory energy efficiency by 14.4%, and thread fairness by 19.3% over a state-of-the-art MLC PCM system that couples the bits in its cells. We show that our results are consistent across a variety of workloads and system configurations.

Diffusion of Lexical Change in Social Media

J. Eisenstein, B. O'Connor, N. A. Smith, E. P. Xing

PLOS ONE, volume 9, Issue 11, November 2014.

Computer-mediated communication is driving fundamental changes in the nature of written language. We investigate these changes by statistical analysis of a dataset comprising 107

continued on pg. 30

Recent Publications

continued from pg. 29

million Twitter messages (authored by 2.7 million unique user accounts). Using a latent vector autoregressive model to aggregate across thousands of words, we identify high-level patterns in diffusion of linguistic change over the United States. Our model is robust to unpredictable changes in Twitter's sampling rate, and provides a probabilistic characterization of the relationship of macro-scale linguistic influence to a set of demographic and geographic predictors. The results of this analysis offer support for prior arguments that focus on geographical proximity and population size. However, demographic similarity – especially with regard to race – plays an even more central role, as cities with similar racial demographics are far more likely to share linguistic influence. Rather than moving towards a single unified "netspeak" dialect, language evolution in computer-mediated communication reproduces existing fault lines in spoken American English.

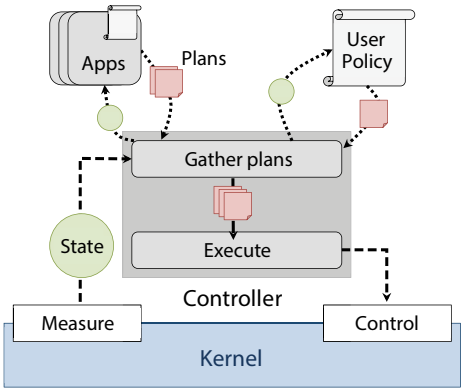
From Feast to Famine: Managing Mobile Resources Across Environments and Preferences

Robert Kiefer, Erik Nordstrom, Michael J. Freedman

Proceedings of 2nd USENIX Conference on Timely Results in Operating Systems (TRIOS'14), October 2014.

Mobile devices regularly move between feast and famine—environments that differ greatly in the capacity and cost of available network resources. Managing these resources effectively is an important aspect of a user's mobile experience. However, preferences for resource management vary across users, time, and operating conditions, and user and application interests may not align. Furthermore, today's mobile OS mechanisms are typically coarse-grained, inflexible, and scattered across system and application settings. Users must adopt a "one size fits all" solution or micro-manage their devices.

This paper introduces Tango, a platform for managing network resource usage through a programmatic model



The Tango architecture.

that expresses user and app interests ("policies"). Tango centralizes policy expression and enforcement in a controller process that monitors device state and adjusts network usage according to a user's (potentially dynamic) interests. To align interests and leverage app-specific knowledge, Tango uses a constraint model that informs apps of network limitations so they can optimize their usage. We evaluate how to design policies that account for data limits, user experience, and battery life. We demonstrate how Tango improves individual network-intensive apps like music streaming, as well as conditions when multiple apps compete for limited resources.

Fast and Accurate Mapping of Complete Genomics Reads

Donghyuk Lee, Farhad Hormozdiari, Hongyi Xin, Faraz Hach, Onur Mutlu, Can Alkan

Methods, Elsevier, October 2014.

Many recent advances in genomics and the expectations of personalized medicine are made possible thanks to power of high throughput sequencing (HTS) in sequencing large collections of human genomes. There are tens of different sequencing technologies currently available, and each HTS platform have different strengths and biases. This diversity both makes it possible to use different technologies to correct for shortcomings; but also requires to develop different algorithms for each platform due to the differences in data types and error models. The first problem to tackle in analyzing HTS data for resequencing applications is the read mapping stage, where many tools have been developed for the most popular HTS methods, but publicly available and open source aligners are still lacking for the Complete Genomics (CG) platform. Unfortunately, Burrows-Wheeler based methods are not practical for CG data due to the gapped nature of the reads generated by this method. Here we provide a sensitive read mapper (sirFAST) for the CG technology based on the seed-and-extend paradigm that can quickly map CG reads to a reference genome. We evaluate the performance and accuracy of sirFAST using both simulated and publicly available real data sets, showing high precision and recall rates.

Fast Iterative Graph Computation with Resource Aware Graph Parallel Abstractions

Yang Zhou, Ling Liu, Kisung Lee, Calton Pu, Qi Zhang

Proceedings of ACM Symposium on High-Performance Parallel and Distributed Computing (ACM HPDC 2015), Portland, Oregon, June 15-19, 2015.

Iterative computation on large graphs has challenged system research from two aspects: (1) how to conduct high performance parallel processing for both in-memory and out-of-core graphs; and (2) how to handle large graphs that exceed the resource boundary of traditional systems by resource aware graph partitioning such that it is feasible to run large-scale graph analysis on a single PC. This paper presents GraphLego, a resource adaptive graph processing system with multi-level programmable graph parallel abstractions. GraphLego is novel in three aspects: (1) we argue that vertex-centric or edge-centric graph partitioning are ineffective for parallel processing of large graphs and we introduce three alternative graph parallel abstractions to enable a large graph to be partitioned at the granularity of subgraphs by slice, strip and dice based partitioning; (2) we use

dice-based data placement algorithm to store a large graph on disk by minimizing non-sequential disk access and enabling more structured in-memory access; and (3) we dynamically determine the right level of graph parallel abstraction to maximize sequential access and minimize random access. GraphLego can run efficiently on different computers with diverse resource capacities and respond to different memory requirements by real-world graphs of different complexity. Extensive experiments show the competitiveness of GraphLego against existing representative graph processing systems, such as GraphChi, GraphLab and X-Stream.

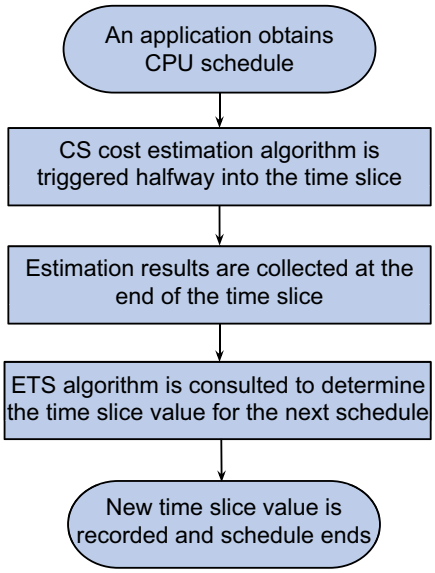
Balancing Context Switch Penalty and Response Time with Elastic Time Slicing

Nagakishore Jammula, Moinuddin Qureshi, Ada Gavrilovska, Jongman Kim

Proceedings of 21st International Conference on High Performance Computing (HiPC'14), December 2014.

Virtualization allows the platform to have increased number of logical processors by multiplexing the underlying resources across different virtual machines. The hardware resources get time shared not only between different virtual machines, but also between different workloads of the same virtual machine. An important source of performance degradation in such a scenario comes from the cache warmup penalties a workload experiences when it gets scheduled, as the working set belonging to the workload gets displaced by other concurrently running workloads. We show that a virtual machine that time switches between four workloads can cause some of the workloads a slowdown of as much as 54%. However, such performance degradation depends on the workload behavior, with some workloads experiencing negligible degradation and some severe degradation.

We propose Elastic Time Slicing (ETS) to reduce the context switch overhead for the most affected workloads. We demonstrate that by taking the workload-specific context switch overhead into consideration, the CPU scheduler can make better decisions to minimize the context switch penalty for the most affected workloads, thereby resulting in substantial performance improvements. ETS enhances performance without compromising on response time, thereby achieving dual benefits. To facilitate ETS, we develop a low-overhead hardware-based mechanism that dynamically estimates the sensitivity of a given workload to context switching. We evaluate the accuracy of the mechanism under various cache management policies and show that it is very reliable. Context switch related warmup penalties increase as optimizations are applied to address traditional cache misses. For the first time, we assess the impact of advanced replacement policies and establish that it is significant.



A high level overview of the ETS framework.

Other Interesting Papers by ISTC-CC Faculty

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

Managing GPU Concurrency in Heterogeneous Architectures. Onur Kayiran, Nachiappan Chidambaram Nachiappan, Adwait Jog, Rachata Ausavarungnirun, Mahmut T. Kandemir, Gabriel H. Loh, Onur Mutlu, and Chita R. Das. Proceedings of 47th International Symposium on Microarchitecture (MICRO'14), December 2014.

A Case for Core-Assisted Bottleneck Acceleration in GPUs: Enabling Efficient Data Compression. Nandita Vijaykumar, Gennady Pekhimenko, Adwait Jog, Abhishek Bhowmick, Rachata Ausavarungnirun, Onur Mutlu, Chita Das, Mahmut Kandemir, Todd C. Mowry. Proceedings of the 42nd International Symposium on Computer Architecture (ISCA), Portland, OR, June 2015.

A Portable Benchmark Suite for Highly Parallel Data Intensive Query Processing. I. Saeed, J. Young, and S. Yalamanchili. 2nd Workshop on Parallel Programming for Analytics Applications, held with PPoPP, February 7-11, 2015.

A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing. Junwhan Ahn, Sungpack Hong, Sungjoo Yoo, Onur Mutlu, Kiyoung Choi. Proceedings of the 42nd International Symposium on Computer Architecture (ISCA), Portland, OR, June 2015.

Cymric: A Framework for Prototyping Near-Memory Architectures. C. Kersey, H. Kim and S. Yalamanchili. Sixth Workshop on Architectural Research Prototyping (WARP), held with ISCA.

Design and Evaluation of Hierarchical Rings with Deflection Routing. Rachata Ausavarungnirun, Chris Fallin, Xiangyao Yu, Kevin Chang, Greg Nazario, Reetuparna Das, Gabriel Loh, and Onur Mutlu. Proceedings of the 26th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD'14), October 2014.

High-Performance and Lightweight Transaction Support in Flash-Based SSDs. Youyou Lu, Jiwu Shu, Jia Guo, Shuai Li, Onur Mutlu. To appear in IEEE Transactions on Computers (TC), 2015.

Main Memory Scaling: Challenges and Solution Directions. Onur Mutlu. Invited Book Chapter in More than Moore Technologies for Next Generation Computer Design, Springer, 2015.

Mitigating Prefetcher-Caused Pollution using Informed Caching Policies for Prefetched Blocks. Vivek Seshadri, Samihan Yedkar, Hongyi Xin, Onur Mutlu, Phillip P. Gibbons, Michael A. Kozuch, Todd C. Mowry. ACM Transactions on Architecture and Code Optimization (TACO), Vol. 11, No. 4, January 2015.

Page Overlays: An Enhanced Virtual

Recent Publications

continued from pg. 31

Memory Framework to Enable Fine-grained Memory Management. Vivek Seshadri, Gennady Pekhimenko, Olatunji Ruwase, Onur Mutlu, Phillip Gibbons, Michael Kozuch, Todd C. Mowry, Trishul Chilimbi. Proceedings of the 42nd International Symposium on Computer Architecture (ISCA), Portland, OR, June 2015.

PIM-Enabled Instructions: A Low-Overhead, Locality-Aware Processing-in-Memory Architecture. Junwhan Ahn, Sungjoo Yoo, Onur Mutlu, Kiyoung Choi. Proceedings of the 42nd International Symposium on Computer Architecture (ISCA), Portland, OR, June 2015.

Ramulator: A Fast and Extensible DRAM Simulator. Yoongu Kim, Weikun Yang, and Onur Mutlu. To appear in IEEE Computer Architecture Letters (CAL), 2015.

The Main Memory System: Challenges and Opportunities. Onur Mutlu, Justin Meza, and Lavanya Subramanian. Invited Article in Communications of the Korean Institute of Information Scientists and Engineers (KIISE), 2015.

LogicBlox co-author, experiments on NVIDIA Multipredicate Join Algorithms for Accelerating Relational Graph Processing on GPUs. H. Wu, D. Zinn, M. Aref, and S. Yalamanchili. Proceedings of 5th International Workshop on Accelerating Data Management Systems Using Modern Processor and Storage Architectures (ADMS'14), September 2014.

Comparative Evaluation of FPGA and ASIC Implementations of Bufferless and Buffered Routing Algorithms for On-Chip Networks. Yu Cai, Ken Mai, Onur Mutlu. Proceedings of the 16th International Symposium on Quality Electronic Design (ISQED), Santa Clara, CA, March 2015.

Control Principles and On-Chip Circuits for Active Cooling Using Integrated Superlattice-Based Thin-Film Thermoelectric Devices. Borislav Alexandrov, Owen Sullivan, William J. Song, Sudhakar Yalamanchili, Satish Kumar, and Saibal Mukhopadhyay. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Volume 22, Issue 9, September 2014.

Characterizing the Performance Effect of Trials and Rotations in Applications that use Quantum Phase Estimation. Ali JavadiAbhari, Shruti Patil, Chen-Fu Chiang, Jeff Heckey, Margaret Martonosi, and Frederic T. Chong. Proceedings of IEEE International Symposium on Workload Characterization (IISWC'14), October 2014.

A group of ISTC-CC 2014 retreat attendees gathered for a meal and discussion. From L to R, Michael Kaminsky (Intel), Dong Zhou (CMU), Phil Gibbons (Intel), Dave Andersen (CMU), Chris Ramming (Intel), Greg Ganger (CMU), M. (Satya) Satyanarayanan (CMU).

Year in Review

continued from pg. 5

» Zhuo Chen (adv. M. Satyanarayanan, CMU) presented "Early Implementation Experience with Wearable Cognitive Assistance Applications" at WearSys 2015, the first ACM SIGMOBILE workshop on the "Wearable Systems and Applications" held in Florence, Italy, May 2015.

» Dan Siewiorek (CMU) presented "Converting Mobile Sensing into Data and Data into Action," a distinguished Lecture at the Peking University, Beijing, China, May 29, 2015.

» Phil Gibbons (IL/ISTC-CC) gave a keynote talk entitled "Big Data: Scale Down, Scale Up, Scale Out" at the 29th IEEE International Parallel & Distributed Processing Symposium (IPDPS'15) in Hyderabad, India, May 28, 2015.

» Dan Siewiorek (CMU) presented "Converting Mobile Sensing into Data and Data into Action," in Shenzhen, China, May 28, 2015.

» Dan Siewiorek (CMU) spoke on "Aging in Place," at the Chinese University of Hong Kong, May 27, 2015.

» Mor Harchol Balter (CMU) gave a talk at the Stanford ISL Seminar on "Queues with Redundant Jobs: First Exact Analysis," May 21, 2015.

» Dan Siewiorek (CMU) presented "Converting Mobile Sensing into Data and Data into Action" at the Harbin Institute of Technology, Harbin, China, May 20, 2015.

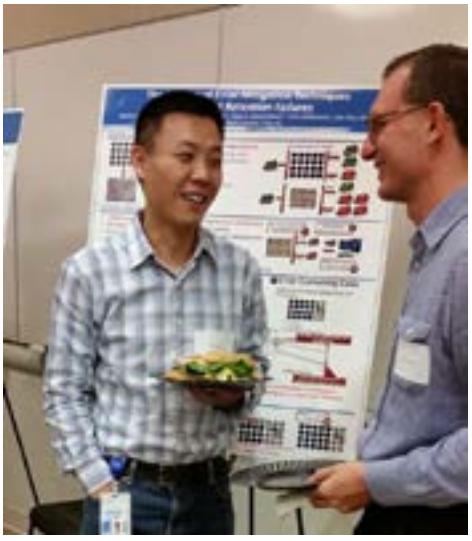
» Dan Siewiorek (CMU) gave a short course on "Wearable Computing and its Reliability," at the Harbin Institute of Technology, China, May 2015.

continued from pg. 32

- S. Yalamanchili (Georgia Tech) presented research progress in Adaptive 3D Architectures at the SRC Design Review held at Intel Portland, May 2015.
- "Deferred Lightweight Indexing for Log-Structured Key-Value Stores" by Ling Liu (Georgia Tech) and co-authors received the best paper award at the 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid'15), May 2015.
- Alexey Tumanov (CMU) won the 2015 CMU Electrical and Computer Engineering Graduate Student Teaching Assistant Award, May 2015.
- Margaret Martonosi (Princeton) and her co-authors David Brooks and Vivek Tiwari won the 2015 ACM SIGARCH/IEEE TCCA Influential ISCA Paper Award for their paper on Wattch from ISCA'00, June 2015.
- Margaret Martonosi (Princeton) received the Marie R. Pistilli Women in EDA Award at DAC'15, June 2015.
- Mor Harchol Balter (CMU) gave a tutorial talk at ISCA on "What Queueing Theory Teaches Us About Computer Systems Design," June 2015.
- Margaret Martonosi (Princeton)gave a Distinguished Lecture at University of Waterloo, June 2015.
- Phil Gibbons (IL/ISTC-CC) gave a keynote on "Living on the Edge…with only Clouds to fall back on" at the 16th IEEE International Conference on Mobile Data Management (MDM'15), June 18, 2015.
- Ling Liu (Georgia Tech) presented a keynote talk on "Data Analytics as a Service: The Next Big Challenge in Services Computing," at the 11th World Congress on Web Services (Services'15), June 30, 2015.
- Dan Siewiorek (CMU) gave a keynote talk on "Converting Mobile Sensing into Data and Data into Action," at the 16th IEEE International Conference on Mobile Data Management (MDM'15), June 16, 2015.
- Onur Mutlu (CMU) presented the keynote talk "Rethinking Memory System Design for Data-Intensive Computing" at the 11th International Workshop on Data Management on New Hardware (DaMoN), in Melbourne, Australia, June 2015.
- Kavitha Chandasekar (Georgia Tech), interned with Josh Fryman's team in DCG's Extreme Scale Technology group.
- Naila Farooqui, a Ph.D. student of Karsten Schwan (Georgia Tech), interned with Tatiana Shpeisman's team in IL/SSR/PSL.
- Tae Jun Ham, a Ph.D. student of Margaret Martonosi (Princeton), interned with Chris Hughes in IL/SSR/PCL.
- Dipanjan Sengupta, a Ph.D. student of Karsten Schwan (Georgia Tech), interned with Ted Willke's team in IL/SSR/PCL.
- Blaise Tine, a Ph.D. student of Sudhakar Yalamanchili (Georgia Tech), interned with Deborah Marr's team in IL/ADR/AAL.
- Hsiao-Yu (Fish) Tung, an M.S. student of Alex Smola (CMU), interned with Chris Hughes in IL/SSR/PCL.
- Varun Saravagi, an M.S. student at CMU, is an ISTC-CC-funded intern with Babu Pillai (IL/ISTC-CC).
- Ion Stoica (UC Berkeley) and team released Apache Spark 1.4.0, with contributions from 210 developers from 70 institutions. It brings an R API to Spark, as well as usability improvements in Spark's core engine and expansion of MLlib and Spark Streaming.
- Calton Pu (Georgia Tech) served as Program Co-Chair for the IEEE CLOUD'15 conference.

2015 Quarter 3

- Guy Blelloch (CMU) and Phil Gibbons were awarded an NSF grant for $845k over 3 years for "Write-efficient Parallel Algorithms for Emerging Memory Technologies."
- M. Satyanarayanan's (CMU) group's extensions for cloudlet-based Open Edge Computing have been integrated with the Kilo release of OpenStack. These are available via github. Details at http://elijah.cs.cmu.edu/development.html
- The parameter server project (CMU) is now part of the larger DMLC (Distributed Machine Learning in C++) framework. Libraries for deep learning, such as Caffe, Minerva and CXXNET have been integrated into the parameter server, courtesy of Mu Li's efforts. The code is available at github.com/dmlc.
- M. Satyanarayanan's (CMU) NSF Large proposal entitled "Wearable Cognitive Assistance" has just been awarded $2.8M over 4 years (Aug 2015 - Aug 2019). Co-PIs are Dan Siewiorek, Martial Hebert and Roberta Klatzky.
- Mor Harchol-Balter (CMU) gave the keynote talk at the International Conference on Distributed Computing Systems (ICDCS) on July 1, 2015 titled, "What Queueing Theory Teaches us about Distributed Computer System Design."
- Onur Mutlu (CMU) presented the keynote talk "Rethinking Memory System Design for Data-Intensive Computing" at the 15th International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS) in Samos, Greece, July 2015.
- Dan Siewiorek (CMU) presented "On You: A Story of Wearable Computing, Computer History Museum Panel" with Thad Starner, Greg Priest-Dorman, Aug. 3, 2015, Mountain View, CA.
- Onur Mutlu (CMU) presented "Read Disturb Errors in MLC NAND Flash Memory" at the Flash Memory Summit 2015 (FMS) in Santa Clara, CA, Aug. 2015.
- Mike Freedman (Princeton) served as Program Co-Chair for the ACM Symposium on Cloud Computing (SoCC'15).
- Onur Mutlu (CMU) presented the technical talk "The DRAM RowHammer Problem (and Its Reliability and Security Implications)" at various venues.
- The 5th Annual ISTC-CC Retreat will be held in Hillsboro, OR at the Intel Jones Farm campus, Aug. 26-28, 2015.



Sheng Li (intel) and Onur Mutlu (CMU) discuss "The Efficacy of Error Mitigation Techniques for DRAM Retention Failures" at an ISTC-CC Retreat '14 poster session.

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 fourth year for the Cloud Computing Center. We are seeing lots of engagement on projects within the center from Intel technology stakeholders. As we enter our final year of the center, we look forward to seeing this great research get it's due accolades (lots of Best Paper and Best Paper runner up awards during the first 6 months of this year) and collaboration with Intel reaching full fruition. Thanks to all the hard work by professors, students and our researchers embedded on the CMU campus. Here's looking forward to another successful year!

Message from the PIs

continued from pg. 2

bounded staleness and parallelization architectures based on what we call "parameter servers" (servers for widely shared ML model parameters). One of our major capstone efforts is coalescing the knowledge from these activities and experiences to lay out a taxonomy of major big-learning styles and the techniques/frameworks that best serve them. Our hope is both to bring some clarity to this still evolving problem space and to ensure that there are no major holes in the solution set, as we move deeper into the data science era.

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 continued efforts to promote platform specialization add a major challenge to the scale and dynamicity of cloud scheduling: workloads are better off when they are assigned to the "right" resources, but they are sometimes happier to get second or third choices rather than waiting for the "right" ones. The challenge is to match the right workloads to the right resources (at the right times), to maximize exploitation of resource specialization. We have devised new interfaces and automation support needed for making specialization truly effective; pulling together the different aspects of doing this, and promulgating them into real open source schedulers (with YARN being the primary target). This is the focus of our second capstone on resource management for specialization.

While much of ISTC-CC's work focuses on core cloud infrastructure, our Cloudlet efforts develop technologies for bringing parts of that core closer to the edge. Cool demonstrations have focused on cognitive assistance, which is something that demands both significant computing resources and low-latency locality-sensitive turnaround, illustrating the need and serve as strong case studies. (And, Greg continues to wait impatiently for the resulting assistance for his failing wet-ware memory!) Multi-company planning, such as a recent workshop including Intel and others, are exploring approaches to pushing these concepts into real deployments and products.

Lots of progress has been made on many other fronts, as well. As one example, our new scalable HPC/cloud storage designs for systems that involve large-scale metadata-intensive workloads have become a core part of the DoE/LANL vision for exascale scientific computing (which includes big-learning approaches). As another, our specialization research is yielding new memory system designs and new approaches to robustly exploiting heterogeneity, as well as several summer interns at Intel Labs. And ... and ... and ...

There are too many other examples of cool results, 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 over the next year and beyond.

ISTC-CC News & Awards

continued from pg. 7

with Stateless Caching and Bulk Insertion." SC'14 had 394 submissions and over 10,000 attendees.

November 9, 2014

Best Student Paper Award at LISA '14

Congratulations to Sara Alspaugh, Betty Beidi Chen, Jessica Lin, Archana Ganapathi, Marti A. Hearst, and Randy Katz on receiving the award for Best Student Paper at the 2014 USENIX Large Installation System Administration Conference (LISA'14) for their work on "Analyzing Log Analysis: An Empirical Study of User Log Mining."

November 3, 2014

Best Paper Award at SOCC '14

Congratulations to Iulian Moraru, Dave Andersen and Michael Kaminsky for receiving the best paper award at the 5th ACM Symposium on Cloud Computing (SOCC'14) for their work on "Paxos Quorum Leases: Fast Reads without Sacrificing Writes."

October 26, 2014

Best Paper Runner-up at IISWC'14

"Characterization and Analysis of Dynamic Parallelism in Unstructured GPU Applications" by J. Wang and S. Yalamanchili was a best paper runner-up at the IEEE International Symposium on Workload Characterization (IISWC'14).

October 25, 2014

Best Paper Runner-up at EMNLP '14

"Language Modeling with Power Low Rank Ensembles" by A. P. Parikh, A. Saluja, C. Dyer and E. P. Xing was a best paper runner-up at the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP'14).

August 24, 2014

Best Paper Award at KDD

Aaron Li, Amr Ahmed, Sujith Ravi, Alex Smola have won the KDD 2014 best paper prize for their work on "Reducing the sampling complexity of topic models." This paper yields over an order of magnitude speedup in sampling topic models, in particular when the amount of data is large or when the generative model is dense.