PARALLEL DATA LABORATORY
CARNEGIE MELLON UNIVERSITY

PDL Packet

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • SPRING 2011
http://www.pdl.cmu.edu/

AN INFORMAL PUBLICATION FROM ACADEMIA'S PREMIERE STORAGE SYSTEMS RESEARCH CENTER DEVOTED TO ADVANCING THE STATE OF THE ART IN STORAGE AND INFORMATION INFRASTRUCTURES.

CONTENTS

PDL CONSORTIUM MEMBERS

American Power Corporation
EMC Corporation
Facebook
Google
Hewlett-Packard Labs
Hitachi, Ltd.
IBM Corporation
Intel Corporation
LSI Corporation
Microsoft Research
NEC Laboratories
NetApp, Inc.
Oracle Corporation
Panasas
Riverbed
Samsung Information Systems America
Seagate Technology
STEC, Inc.
Symantec Corporation
VMware, Inc.

Principles of Operation for Shingled Disk Devices

Garth Gibson

Solid-state storage devices are dramatically changing the playing field for durable storage systems. Still, for at least the next decade most stored information will be magnetically recorded on hard disks because of the small size of magnetically recorded bits and the low cost for a multi-terabyte device. Consumer expectations for ever larger capacity magnetic disk drives and the economics of the disk drive marketplace call for an annual aggressive increase in areal density (recently up to 40% per year). Maintaining this rate of increase in areal density in the face of the impending Superparamagnetic Limit (the density at which a bit written with current materials and writing mechanisms has too few magnetic grains to resist random grain orientation flips), is the core challenge of magnetic recording technologists today, and Shingled Magnetic Recording (SMR) is a leading technique for driving the areal density of magnetic disk drives through 1–10 terabit/inch^2 in the coming decade [Wood00].

Shingled Magnetic Recording is based on partially overlapping, or shingling, adjacent tracks. As illustrated in Figure 1, shingled tracks overlap previously written tracks, exploiting the easier task of reading thin tracks than of writing thin tracks. SMR allows significantly more engineering freedom in the design

continued on page 18

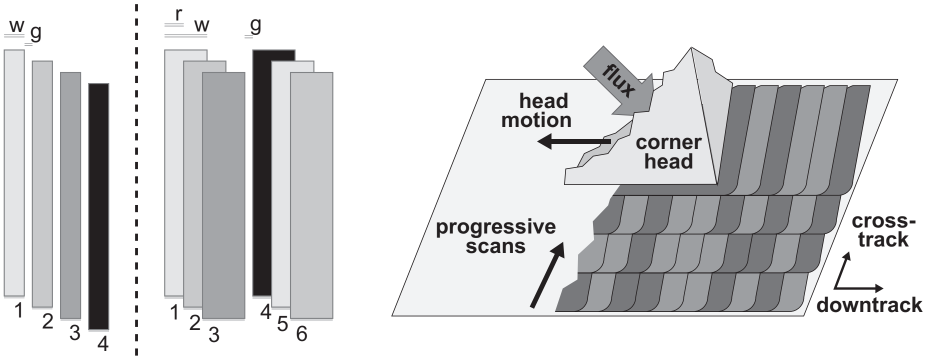


Figure 1: Conventional non-shingled writing, on the left, writes data in non-overlapping tracks, width w, with non-magnetized "guard regions", width g, between tracks. Shingled writing, on the right, leaves residual tracks, width r, less than the width of the written track, width w', before it writes the next track, largely overlapping adjacent tracks. This increases tracks per inch, increasing areal density without significantly changing materials or write head design.

FROM THE DIRECTOR'S CHAIR

Greg Ganger

Hello from fabulous Pittsburgh!

It has been another great year for the Parallel Data Lab. Some highlights include new projects starting, exciting new results on existing projects, awards for several researchers and papers, and even some PDL testimony before Congress. Along the way, many students graduated and joined PDL Consortium companies, new students and faculty have joined PDL, and many papers have been published. Let me highlight a few things.

Perhaps PDL's biggest "growth area" centers on cloud computing, as discussed further in my short article in this newsletter. For me, personally, the most interesting thing was testifying before the U.S. Congress about the benefits and risks of cloud computing; a new experience for me, to be sure. The emergence of cloud computing is an exciting development, promising great efficiency benefits and the long-sought notion of utility computing. And, lots of research will be needed to realize that promise.

PDL has a lot of active research addressing technology advances needed for cloud computing, as well as deployments to gain first-hand experience. Our OpenCloud and OpenCirrus clouds, deployed last year, continue to provide resources for real users and to provide us with invaluable Hadoop logs, instrumentation data, and case studies. We are also deploying a third cloud in the Data Center Observatory (DCO) based on VMware's new cloud computing software and vCloud APIs, with VMware's help and generous hardware donations from HP, Intel, and Samsung. These efforts are helping inform new and continuing research into in-cloud data-intensive computing, automated problem diagnosis, and other topics.

Another exciting development in our field has been the rapid changes in the underlying technologies on which we build. Of course, Flash storage has emerged and is being exploited in many ways (e.g., see the FAWN project). And, there is great excitement around forthcoming non-volatile RAM technologies, like PCM and memristors. But, there are also some interesting changes to our stalwart technology: the disk drive. For example, new manufacturing approaches have eliminated performance uniformity among batches of disks—each disk now provides slightly different streaming bandwidth, varying by as much as 20% within a batch of 20-50 disks. As another example, near-future recording approaches will force a large trade-off between storage capacity and support for over-writing of individual disk blocks; see Garth's article about shingled disks for more.

We continue to create and explore system support for data-intensive computing (DISC). As just one example, we are exploring ways in which to converge cloud databases and huge-scale parallel file systems—the two are currently evolving separately, but are moving toward similar solutions. Both would benefit, conceptually and practically (e.g., shared software), from creating high-level frameworks and mechanisms that work for both. Our ongoing GIGA+ project, which seeks to support massive directories, offers one such example, using mechanisms much like cloud databases. In continuing work, we are exploring use of such mechanisms both for large-scale metadata services in DISC systems and common high-ingest support for such services and cloud databases generally.

The FAWN (Fast Array of Wimpy Nodes) project continues to generate exciting results, including winning the 2010 10GB JouleSort benchmark competition

THE PDL PACKET

The Parallel Data Laboratory
School of Computer Science
Department of ECE
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
VOICE 412•268•6716
FAX 412•268•3010

PUBLISHER
Greg Ganger

EDITOR
Joan Digney

The PDL Packet is published once per year to update members of the PDL Consortium. A pdf version resides in the Publications section of the PDL Web pages and may be freely distributed. Contributions are welcome.

THE PDL LOGO

Skibo Castle and the lands that comprise its estate are located in the Kyle of Sutherland in the northeastern part of Scotland. Both 'Skibo' and 'Sutherland' are names whose roots are from Old Norse, the language spoken by the Vikings who began washing ashore regularly in the late ninth century. The word 'Skibo' fascinates etymologists, who are unable to agree on its original meaning. All agree that 'bo' is the Old Norse for 'land' or 'place,' but they argue whether 'ski' means 'ships' or 'peace' or 'fairy hill.'

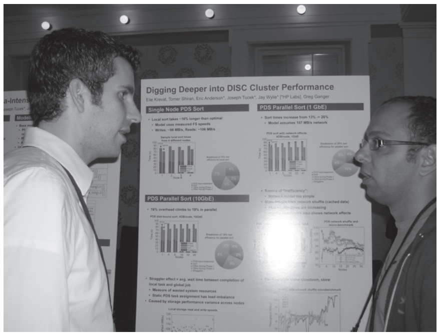
Although the earliest version of Skibo seems to be lost in the mists of time, it was most likely some kind of fortified building erected by the Norsemen. The present-day castle was built by a bishop of the Roman Catholic Church. Andrew Carnegie, after making his fortune, bought it in 1898 to serve as his summer home. In 1980, his daughter, Margaret, donated Skibo to a trust that later sold the estate. It is presently being run as a luxury hotel.

with one of their prototypes. Led by Prof. David Andersen, the FAWN project explores new cluster architectures that can provide data-intensive computing with order of magnitude improvements in energy efficiency. A FAWN cluster uses large collections of embedded processors and Flash memory, rather than smaller collections of high-end servers and disks, providing the same scalability and maximum performance levels while consuming up to one-tenth the power. New prototypes are being built with more specialized components, and there are even some real users now.

We continue to focus a lot of attention on problem diagnosis and use of automation in distributed systems, including DISC systems and large-scale storage. It is clear that there will be no silver bullet here, and PDL research is probing a number of complementary paths. One promising approach involves comparison of request flow graphs, obtained from detailed on-line tracing of work in the system, across problem and non-problem periods—changes in how given request types are serviced can localize and help explain performance problems in a system. Such tracing is increasingly available in real systems, such as throughout Google's systems. We are also exploring the many other instrumentation sources, such as time-series resource utilization and application logs, and how they can be combined to better deduce where things go wrong. Such data, combined with carefully chosen machine learning algorithms, can decrease the lack of guidance facing humans seeking to diagnose problems.

Many other ongoing PDL projects are also producing cool results. For example, we have created new data distribution algorithms for allowing elastic sizing of DISC system size, creating the potential for cloud environments efficiently supporting both data analytics and other activities. We are finding that technologies like Phase-Change Memory (PCM) can be used to replace substantial fractions of DRAM in a system, resulting in higher energy efficiency without loss in performance or robustness. Based on results from our user studies, we are developing a novel policy-based access control mechanism for distributed home storage based on attribute-based naming. This newsletter and the PDL website offer more details and additional research highlights.

I'm always overwhelmed by the accomplishments of the PDL students and staff, and it's a pleasure to work with them. As always, their accomplishments point at great things to come.



Elie Krevat discusses his research on DISC Cluster Performance with PDL Alumni Michael Abd-El-Malek (Google) at a Retreat poster session.

Garth presents a work-in-progress talk on "Shingled Writing for Magnetic Disk Drives" to interested listeners at the 2010 PDL Workshop & Retreat.

PARALLEL DATA LABORATORY

FACULTY

Greg Ganger (pdl director)
412•268•1297
ganger@ece.cmu.edu

Anastasia Ailamaki
David Andersen
Lujo Bauer
Chuck Cranor
Lorrie Cranor
Christos Faloutsos
Eugene Fink
Rajeev Gandhi
Garth Gibson
Seth Copen Goldstein
Carlos Guestrin
Mor Harchol-Balter
Bruce Krogh
Julio López
Todd Mowry
Onur Mutlu
Priya Narasimhan
David O'Hallaron
Adrian Perrig
Mike Reiter
M. Satyanarayanan
Srinivasan Seshan
Bruno Sinopoli
Hui Zhang

STAFF MEMBERS

Bill Courtright 412•268•5485
(pdl executive director) wcourtright@cmu.edu
Karen Lindenfelser, 412•268•6716
(pdl administrative manager) karen@ece.cmu.edu
Joan Digney
Mitch Franzos
Nitin Gupta
Manish Prasad
Michael Strouchen
Charlene Zang

VISITING RESEARCHER

Sangyeun Cho
Chanik Park

GRADUATE STUDENTS

Yoshihisa Abe
Rachata Ausavarungnirun
Jim Cipar
Bin Fan
Bin Fu
Anshul Gandhi
Fan Guo
Varun Gupta
Wesley Jin
Mike Kasick
Soila Kavulya
Elie Krevat
Patrick Lanigan
Lei Li
Yuan Liang
Hyeontaek Lim
Michelle Mazurek
Justin Meza
Iulian Moraru
Richard Munz
Ippokratis Pandis
Luca Parolini
Swapnil Patil
Adam Pennington
Amar Phanishayee
Milo Polte
Kai Ren
Wolfgang Richter
Raja Sambasivan
Vivek Seshadri
Ilari Shafer
Jiri Simsa
Anand Suresh
Wittawat Tantisiriroj
Alexey Tumanov
Vijay Vasudevan
Gaurav Veda
Matthew Wachs
Yifan Wang
Lin Xiao
Lianghong Xu
HanBin Yoon

YEAR IN REVIEW

April 2011

- 13th Annual PDL Spring Industry Visit Day.
- Wolfgang Richter will be interning at IBM TJ Watson with the virtualization group headed by Vas Bala.
- Jiri Simsa will be interning with Google in California this summer.
- Ilari Shafer will be interning with VMware in Palo Alto, CA this coming summer.
- Alexey Tumanov will be interning at MSR Silicon Valley this summer.
- Lianghong Xu will be interning at VMware Boston this summer.

March 2011

- Adrian Perrig received the 2011 Carnegie Institute of Technology Benjamin Richard Teare, Jr. Teaching Award.
- Raja presented “Diagnosing Performance Changes by Comparing Request Flows” at NSDI 11 in Boston.

February 2011

- Reception for Greg Ganger to celebrate his being awarded the Stephen J. Jatras Professorship in Electrical and Computer Engineering (2010).
- Three of Onur Mutlu’s papers were chosen as 2010’s most significant computer architecture papers by IEEE Micro: “Aergia: Exploiting Packet Latency Slack in On-Chip Networks”, “Data Marshaling for Multicore Architectures” and “Thread Cluster Memory Scheduling: Exploiting Differences in Memory Access Behavior” (co-authored with Mor Harchol-Balter).

Greg and his wife Jenny at the Jatras Chair Reception in February.

- David Andersen was named a 2011 Sloan Foundation Fellow.
- Mukesh Agrawal defended his Ph.D. research “Spare a Little Change? Towards a 5-Nines Internet in 250 Lines of Code.”
- Swapnil Patil presented” Scale and Concurrency of GIGA+: File System Directories with Millions of Files” at USENIX FAST 2011.
- Greg and Garth hosted a well attended PDL Alumni reunion/reception at FAST 11 in San Jose.

December 2010

- Christos Faloutsos was named an ACM Fellow.
- Tudor A. Dumitraş defended his Ph.D. dissertation on “Improving the Dependability of Distributed Systems.”

November 2010

- 18th Annual Parallel Data Lab Workshop & Retreat
- Greg Ganger was named an IEEE Fellow.
- Ryan Johnson defended his Ph.D. dissertation on “Scalable Storage Managers for the Multicore Era.”
- Debabrata Dash defended his Ph.D. research on “Automated Physical Design: A Combinatorial Optimization Approach.”

October 2010

- Mike Kasick presented “Behavior-Based Problem Localization for Parallel File Systems” at HotDep ‘10 in Vancouver, BC.
- Jiri Simsa presented “dBug: Systematic Evaluation of Distributed Systems at the 5th Int. Workshop on Systems Software Verification (SSV’10) in Vancouver BC.
- Tudor Dumitraş presented “To Upgrade or Not to Upgrade: Impact of Online Upgrades across Multiple Administrative Domains” at ACM Onward! in Reno Nevada
- Lorrie Cranor was a panelist at the Alta Associates’ Executive Women’s Forum National Conference, Oct. 20-22, in Scottsdale, Ariz. She gave a talk on “Information Security, Privacy & Risk Management: From Research to Practice.”
- M. Satyanarayanan (Satya), was honored with the SIGMOBILE 2010 Outstanding Contributions Award “for his pioneering a wide spectrum of technologies in support of disconnected and weakly connected mobile clients.”

September 2010

- Greg participated in the 2010 VMworld conference for cloud computing in San Francisco.
- Christos Faloutsos was awarded the SIGCOMM Test of Time award for his paper “On the Power Law Relationships of the Internet Topology.”

July 2010

- Greg Ganger earned a 2010 HP Innovation Research Award.
- The FAWN team won the 2010 Joule Sort Challenge.
- Greg Ganger testified to the U.S. House Committee on Oversight and Government Reform and the Subcommittee on Government Management, Organization and Procurement, discussing the benefits and risks of using cloud computing.

June 2010

- Christos Faloutsos was presented the 2010 ACM SIGKDD Innovation Award.

continued on page 22

Exploring Reactive Access Control

Mazurek, Klemperer, Shay, Takabi, Bauer & L. Cranor

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

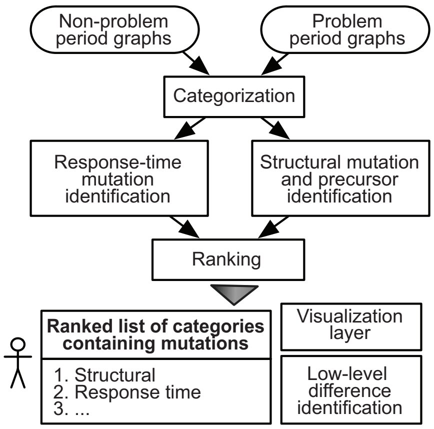
As users store and share more digital content at home, access control becomes increasingly important. One promising approach for helping non-expert users create accurate access policies is reactive policy creation, in which users can update their policy dynamically in response to access requests that would not otherwise succeed. An earlier study suggested reactive policy creation might be a good fit for file access control at home. To test this, we conducted an experience-sampling study in which participants used a simulated reactive access-control system for a week. Our results bolster the case for reactive policy creation as one mode by which home users specify access-control policy. We found both quantitative and qualitative evidence of dynamic, situational policies that are hard to implement using traditional models but that reactive policy creation can facilitate. While we found some clear disadvantages to the reactive model, they do not seem insurmountable.

Diagnosing Performance Changes by Comparing Request Flows

Sambasivan, Zheng, DeRosa, Krevat, Whitman, Strouken, Wang, Xu & Ganger

8th USENIX Symposium on Networked Systems Design and Implementation (NSDI'11). March 30 - April 1, 2011. Boston, MA.

The causes of performance changes in a distributed system often elude even its developers. This paper develops a new technique for gaining insight into such changes: comparing request flows from two executions (e.g., of two system versions or time periods). Building on end-to-end request-flow tracing within and across components, algorithms are described for identifying and ranking changes in the flow and/or timing of request processing. The implementation of these algorithms in a tool called Spectroscope is evaluated. Six case studies are presented of using Spectroscope to diagnose performance changes in a distributed storage service caused by code changes, configuration modifications, and component degradations, demonstrating the value and efficacy of comparing request flows. Preliminary experiences of using Spectroscope to diagnose performance changes within select Google services are also presented.


Non-problem period graphs
Problem period graphs
Categorization
Response-time mutation identification
Structural mutation and precursor identification
Ranking
Ranked list of categories containing mutations
1. Structural
2. Response time
3. ...
Visualization layer
Low-level difference identification

Spectroscope's workflow for comparing request flows. First, Spectroscope groups requests from both periods into categories. Second, it identifies which categories contain mutations or precursors. Third, it ranks mutation categories according to their expected contribution to the performance change. Developers are presented this ranked list. Visualizations of mutations and their precursors can be shown. Also, low-level differences can be identified for them.

Exertion-based Billing for Cloud Storage Access

Wachs, Xu, Kanevsky & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-11-105. March 2011.

Charging for cloud storage must account for two costs: the cost of the capacity used and the cost of access to that capacity. For the cost of access, current systems focus on the work requested, such as data transferred or I/O operations completed, rather than the exertion (i.e., effort/resources expended) to complete that work. But, the provider's cost is based on the exertion, and the exertion for a given amount of work can vary dramatically based on characteristics of the workload, making current charging models unfair to tenants, provider, or both. This paper argues for exertion-based metrics, such as disk time, for the access cost component of cloud storage billing. It also discusses challenges in supporting fair and predictable exertion accounting, such as significant inter-workload interference effects for storage access, and a performance insulation approach to addressing them.

Thread Cluster Memory Scheduling: Exploiting Differences in Memory Access Behavior

Kim, Papamichael, Mutlu & Harchol-Balter

Proceedings of the 43rd International Symposium on Microarchitecture (MICRO), Atlanta, GA, December 2010.

In a modern chip-multiprocessor system, memory is a shared resource among multiple concurrently executing threads. The memory scheduling algorithm should resolve memory contention by arbitrating memory access in such a way that competing threads progress at a relatively fast and even pace, resulting in high system throughput and fairness. Previously proposed memory scheduling algorithms are predominantly optimized for only one of these objectives: no scheduling algorithm provides the best system throughput and best fairness at the same time.

This paper presents a new memory scheduling algorithm that addresses system throughput and fairness separately with the goal of achieving the

continued on page 6

continued from page 5

best of both. The main idea is to divide threads into two separate clusters and employ different memory request scheduling policies in each cluster. Our proposal, Thread Cluster Memory scheduling (TCM), dynamically groups threads with similar memory access behavior into either the latency-sensitive (memory-non-intensive) or the bandwidth-sensitive (memory-intensive) cluster. TCM introduces three major ideas for prioritization: 1) we prioritize the latency-sensitive cluster over the bandwidth-sensitive cluster to improve system throughput; 2) we introduce a "niceness" metric that captures a thread's propensity to interfere with other threads; 3) we use niceness to periodically shuffle the priority order of the threads in the bandwidth-sensitive cluster to provide fair access to each thread in a way that reduces inter-thread interference. On the one hand, prioritizing memory-non-intensive threads significantly improves system throughput without degrading fairness, because such "light" threads only use a small fraction of the total available memory bandwidth. On the other hand, shuffling the priority order of memory-intensive threads improves fairness because it ensures no thread is disproportionately slowed down or starved.

We evaluate TCM on a wide variety of multiprogrammed workloads and compare its performance to four previously proposed scheduling algorithms, finding that TCM achieves both the best system throughput and fairness. Averaged over 96 workloads on a 24-core system with 4 memory channels, TCM improves system throughput and reduces maximum slowdown by 4.6%/38.6% compared to ATLAS (previous work providing the best system throughput) and 7.6%/4.6% compared to PAR-BS (previous work providing the best fairness).

Garth Gibson (R) discusses storage with Dave Anderson of Seagate (L) and John Bent of LANL (C) on a walk at the 2010 PDL Workshop and Retreat at the Bedford Springs Resort in Bedford Springs, PA.

Automation Without Predictability is a Recipe for Failure

Sambasivan & Ganger

Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-11-101, January 2011.

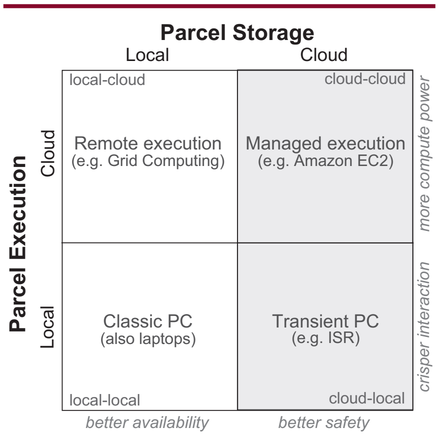
Automated management seems a must, as distributed systems and datacenters continue to grow in scale and complexity. But, automation of performance problem diagnosis and tuning relies upon predictability, which in turn relies upon low variance—most automation tools aren't effective when variance is regularly high. This paper argues that, for automation to become a reality, system builders must treat variance as an important metric and make conscious decisions about where to reduce it. To help with this task, we describe a framework for understanding sources of variance and describe an example tool for helping identify them.

The Case for Content Search of VM Clouds

Satyanarayanan, Richter, Ammons, Harkes & Goode

34th Annual IEEE Computer Software and Applications Conference Workshops (COMPSACW), July 19-23, 2010, Seoul, Korea.

The success of cloud computing can lead to large, centralized collections of virtual machine (VM) images. The ability to interactively search these VM images at a high semantic level emerges as an important capability. This paper examines the opportunities and challenges in creating such a search capability, and presents early evidence of its feasibility.



Taxonomy of VM-based Cloud Computing.

Of Passwords and People: Measuring the Effect of Password-Composition Policies

Komanduri, Shay, Kelley, Mazurek, Bauer, Christin, L. Cranor & Egelman

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Text-based passwords are the most common mechanism for authenticating humans to computer systems. To prevent users from picking passwords that are too easy for an adversary to guess, system administrators adopt password-composition policies (e.g., requiring passwords to contain symbols and numbers). Unfortunately, little is known about the relationship between password-composition policies and the strength of the resulting passwords, or about the behavior of users (e.g., writing down passwords) in response to different policies. We present a large-scale study that investigates password strength, user behavior, and user sentiment across four password-composition policies. We characterize the predictability of passwords by calculating their entropy, and find that a number of commonly held beliefs about password composition and

continued on page 7

continued from page 6

strength are inaccurate. We correlate our results with user behavior and sentiment to produce several recommendations for password-composition policies that result in strong passwords without unduly burdening users.

Thread Cluster Memory Improving Storage Bandwidth Guarantees with Performance Insulation

Wachs & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-113, October 2010.

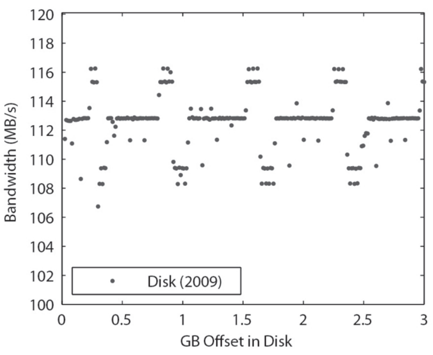
Workloads that share a storage system should achieve predictable, controllable performance despite the activities of other workloads. One desirable way of expressing performance goals is as bandwidth guarantees. Unfortunately, storage bandwidth is difficult to allocate and manage among workloads, because total system capacity depends on both the workloads' access patterns and on any interference between them. This report demonstrates a new approach to supporting soft bandwidth guarantees, building on explicit performance insulation that bounds interference among workloads and its effect on performance and total system capacity. Combining dynamic disk head timeslicing and slack assignment, this approach eliminates almost all avoidable guarantee violations, leaving just those fundamental ones faced by individual workloads whose locality change too significantly. Experiments with a prototype show an order-of-magnitude decrease in the number of guarantee violations compared to traditional token-bucket based throttling.

Disks Are Like Snowflakes: No Two Are Alike

Krevat, Tucek, & Ganger

Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-11-102, February 2011.

Gone are the days of homogeneous sets of disks. Even disks of a given batch, of the same make and model, will have significantly different bandwidths. This paper describes the disk technology trends responsible for the now-inherent heterogeneity of multi-disk systems and disk-based clusters, provides measurements quantifying it, and discusses its implications for system designers.



A closer look at intra-disk behavior with adaptive zoning. A smaller block size of 12 MB for the 2009-era disks clearly distinguishes between bandwidth differences across disk heads and surfaces.

Applying Simple Performance Models to Understand Inefficiencies in Data-Intensive Computing

Krevat, Shiran, Anderson, Tucek, Wylie & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-11-103. February 2011.

New programming frameworks for scale-out parallel analysis, such as MapReduce and Hadoop, have become a cornerstone for exploiting large datasets. However, there has been little analysis of how these systems perform relative to the capabilities of the hardware on which they run. This paper describes a simple analytical model that predicts the theoretic ideal performance of a parallel dataflow system. The model exposes the inefficiency of popular scale-out systems, which take 3–13× longer to complete jobs than the hardware should allow, even in well-tuned systems used to achieve record-breaking benchmark results. Using a simplified dataflow processing tool called Parallel DataSeries, we show that the model's ideal can be approached (i.e., that it is not wildly optimistic), coming within 10–14% of the model's prediction. Moreover, guided by the model, we present analysis of inefficiencies which exposes issues in both the disk and networking subsystems that will be faced by any DISC system built atop standard OS and networking services.

Recipes for Baking Black Forest Databases

López, Degraf, DiMatteo, Fu, Fink & Gibson

Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-11-104, February 2011.

Large-scale N-body simulations play an important role in advancing our understanding of the formation and evolution of large structures in the universe. These computations require a large number of particles, in the order of 10-100 of billions, to realistically model phenomena such as the formation of galaxies. Among these particles, black holes play a dominant role on the formation of these structure. Computational cosmologists are interested in the analysis of black hole properties throughout the simulation with high temporal resolution. The properties of the black holes need to be assembled in merger tree histories to model the process where two or more black holes merge to form a larger one. In the past this analysis has been carried out with custom approaches that no longer scales to the size of black hole datasets produced by current cosmological simulations. We present a set of algorithms and a strategy to represent and store a forest of merger trees for black holes in relational databases (RDBMS). We implemented this approach and present results with datasets containing 0.5 billion time

continued on page 23

March 2011

Adrian Perrig to Receive Teaching Award

Congratulations to Adrian who has been named a recipient of the 2011 Carnegie Institute of Technology Benjamin Richard Teare, Jr. Teaching Award. Adrian is a Professor in the Departments of Electrical & Computer Engineering, Engineering & Public Policy and Computer Science.

March 2011

PDL Alum, Jure Leskovec, Named on IEEE "AI's 10 to Watch" List

Four of the 10 most promising young scientists working today in the field of artificial intelligence are either Carnegie Mellon University faculty members or have recently earned their PhDs in computer science at CMU, according to the editors of IEEE Intelligent Systems magazine. The magazine compiles a list of these outstanding researchers, called "AI's 10 to Watch," every two years. One of these is PDL Alum Jure Leskovec, who earned his PhD in computational and statistical learning in 2008 and is now an assistant professor of computer science at Stanford University, uses large-scale data-mining and machine learning techniques to analyze the structure and evolution of the Internet.

-- excerpted from CMU News Brief, March 9, 2011 by Brian Spice

February 2011

Three Mutlu Papers Named Top Picks

Three research papers co-authored by Assistant Professor of ECE Onur Mutlu have been published in a collection of 2010's most significant computer architecture papers, as chosen by IEEE Micro magazine. At the beginning of each year, the leading IEEE periodical in computer architecture and design selects 10–12 "Top Pick" computer architecture papers of the past year based on the publication's novelty and potential for long-term impact. Mutlu's papers — among the 11 selected for 2010 — tackle the problem of designing more scalable and efficient multicore systems.

"Thread Cluster Memory Scheduling: Exploiting Differences in Memory Access Behavior" provides an application-aware memory access scheduling algorithm that maximizes system throughput and fairness at the same time. The main idea is to divide threads into two separate clusters and employ different memory request scheduling policies in each cluster such that the needs of different kinds of threads are served separately. This paper was co-authored by Yoongu Kim, Michael Papamichael and Mor Harchol-Balter. It originally appeared in MICRO-43, 2010.

"Aergia: Exploiting Packet Latency Slack in On-Chip Networks," highlights more efficient on-chip communication mechanisms. It introduces methods that identify messages critical for system performance in multicore systems and develops scheduling policies that take advantage of this information. The paper—co-written by Mutlu, Reetuparna Das (Intel), Chita Das (Penn State University), and Thomas Moscibroda (Microsoft Research)—originally appeared at the 37th International Symposium on Computer Architecture (ISCA).

Some of his students parodied Greg's advisorial style in a skit presented at the Jatras chair reception. Here is Elie Krevat as Greg.

"Data Marshaling for Multicore Architectures" develops hardware/software cooperative methods to reduce the performance overhead of remotely executing a code segment in a multicore system. The paper, co-written by Aater Suleman, Jose Joao, Khubaib, and Yale Patt, also originally appeared at the 37th ISCA.

-- with info from ECE News, Feb. 28, 2011

February 2011

Sloan Fellowships for 2011 Announced

Congratulations to David Andersen has been named a 2011 Sloan Foundation Fellow. The complete list of awardees is available at http://www.sloan.org/fellowships/page/21.

February 2011

Gregory Ganger Earns ECE Professorship For Expertise In Computer Systems

Gregory R. Ganger was awarded the Stephen J. Jatras Professorship in Electrical and Computer Engineering for cutting-edge work in computer systems. The professorship is named for the late Stephen J. Jatras (E'47), former chairman of the Telex Corp. and a leader in a variety of academic, civic and community organizations stretching from Pittsburgh to Tulsa, Okla.

Ganger, who recently testified in Washington, D.C., about the risks and benefits of cloud computing, is internationally recognized for his work in computer systems, such as storage systems, distributed systems and operating systems.

Since 2001, Ganger has served as director of the Parallel Data Lab (PDL),

continued on page 9

continued from page 8

where he is collaborating with HP labs on a research initiative focused on cloud computing issues through the prestigious HP Labs Innovation Program. More than 50 students, staff and faculty contribute to PDL research activities, and 19 of the top companies sponsor and participate in the ongoing work.

"Greg is an outstanding researcher, educator and academic leader. His work addresses fundamental engineering challenges, and solves important problems even while he builds unique systems and organizations. He is a wonderful example of the spirit of Carnegie Mellon's culture of collaboration," said Ed Schlesinger, head of Carnegie Mellon's Department of Electrical and Computer Engineering.

-- 8.5x11 News, Feb. 10, 2011 - Vol. 21, No. 30

December 2010

Faloutsos Named ACM Fellow

We are delighted to announce that Christos Faloutsos has been selected to be an ACM Fellow "for contributions to data mining, indexing, fractals, and power laws." The full list of 2010 ACM Fellows can be found at http://www.acm.org/news/featured/fellows-2010

From acm.org/news: "These men and women have made advances in technology and contributions to the computing community that are meeting the dynamic demands of the 21st century," said ACM President Alain Chesnais. "Their ability to think critically and solve problems creatively is enabling great advances on an international scale. The selection of this year's Fellows reflects broad international representation of the highest achievements in computing, which are advancing the quality of life throughout society."

ACM will formally recognize the 2010 Fellows at its annual Awards Banquet on June 4, 2011, in San Jose, California.

November 2010

Swapnil Patil Places First in ACM Student Research Competition

Congratulations to Swapnil Patil, who was awarded first place at the ACM Student Research Competition at the ACM Supercomputing Conference 2010 held in New Orleans, LA. The ACM Student Research Competition (SRC), sponsored by Microsoft Research, offers a unique forum for undergraduate and graduate students to present their original research at well-known ACM sponsored and co-sponsored conferences before a panel of judges and attendees.

November 2010

Greg Ganger Named an IEEE Fellow

ECE Professor Greg Ganger has been named an IEEE fellow, 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.

Ganger, the Stephen J. Jatras Professor of ECE and computer science and director of the university's Parallel Data Lab, was honored "for contributions to metadata integrity in file systems." Ganger has a broad range of research interests in computer systems, including operating systems, storage/file systems, security, networking and distributed systems. He is particularly interested in developing new ways to structure computer systems to address technology changes and enable new applications. As director of the Parallel Data Lab, Ganger leads projects in areas like storage system architecture, storage security, file systems, disk characterization and server implementation. He recently earned an HP Innovation Award — the third of his career — that will enable him to work with HP Labs on a research initiative focused on cloud computing, a topic on which he recently testified in Washington, D.C.

-- from ECE News Online, Nov. 27, 2010

October 2010

PDL Team Wins 3rd in Open Source Software Challenge

An SCS team that included several PDL members won 3rd position (silver prize) in the 'Open Source Software World Challenge 2010', (http://project.oss.kr/english/) from among 26 competing submissions. The award includes $2000, plus travel expenses to accept the prize.

The team consisted of graduate SCS students Mr. U Kang and Mr. Polo Chau, and advisor Christos Faloutsos, with several more contributors, listed on the project web site - www.cs.cmu.edu/~pegasus The Pegasus system is able to mine billion-node graphs, using parallelism and specifically, 'hadoop'. Code, documentation, instructions video and related papers are on the web site.

October 2010

Lorrie Cranor Information Security Panelist

Dena Haritos Tsamitis, director of the Information Networking Institute and director of education, training and outreach at Carnegie Mellon CyLab, and Lorrie Faith Cranor, director of the CyLab Usable Privacy and Security Laboratory and associate professor of computer science and electrical and

continued on page 10

continued from page 9

computer engineering, are panelists at the Alta Associates' Executive Women's Forum National Conference, Oct. 20-22, in Scottsdale, Ariz. Tsamitis will moderate and Cranor will be a panelist for the panel talk titled "Information Security, Privacy & Risk Management: From Research to Practice." Tsamitis also will be a panelist for the talk "Re-thinking Social Networking — The Good, the Bad, and the Enablement." The Executive Women's Forum is an annual gathering of executive-level women in the areas of information security, risk management, governance, compliance, IT audit and privacy.

-- from 8.5x11 News Oct. 21, 2010

October 2010

Cranor Company Receives Grant

Wombat Security Technologies, a CMU spinoff, recently was awarded a $750,000 Small Business Innovation Grant from the U.S. Air Force as a phase II grant to develop software for cybersecurity awareness and training. The company was founded by School of Computer Science Professors Jason Hong, Norman Sadeh and Lorrie Cranor.

-- from 8.5x11 News Oct. 21, 2010

October 2010

Satya Wins 2010 SIGMOBILE Award

Congratulations to M. Satyanarayanan (Satya), who has been honored with the SIGMOBILE 2010 Outstanding Contributions Award "for his pioneering a wide spectrum of technologies in support of disconnected and weakly connected mobile clients." He joins an illustrious group of previous winners: http://www.sigmobile.org/awards/oca.html The SIGMOBILE Outstanding Contribution Award is given for significant and lasting contributions to the research on mobile computing and communications and wireless networking.

September 2010

NSF Project to Make Internet Secure and Smart

Carnegie Mellon Computer Science and Electrical and Computer Engineering Professor Peter Steenkiste is leading a three-year, $7.1 million effort sponsored by the National Science Foundation (NSF) to develop a next-generation network architecture that fixes security and reliability deficiencies now threatening the viability of the Internet. The eXpressive Internet Architecture (XIA) Project, one of four new projects funded through the Future Internet Architecture Program of the NSF's Computer and Information Science and Engineering (CISE) Directorate, will include intrinsic security features so that users can be assured that the websites they access and the documents they download are legitimate.

In addition to Steenkiste who is the principal investigator, other CMU faculty members working on the project, including several members of the PDL, are David Andersen, David Feinberg, Srinivasan Seshan and Hui Zhang of the Computer Science Department; CyLab technical director Adrian Perrig; Sara Kiesler of the Human-Computer Interaction Institute; and Jon Peha and Marvin Sirbu of the Engineering and Public Policy Department.

--CMU 8.5x11 News Sept. 2, 2010

September 2010

Faloutsos Wins SIGCOMM 2010 Test of Time Award

Congratulations to Christos and his co-authors (brothers Michalis Faloutsos and Petros Faloutsos) for winning the SIGCOMM Test of Time award for their paper "On the Power Law Relationships of the Internet Topology." The ACM SIGCOMM Test of Time Award recognizes papers published 10 to 12 years in the past in Computer Communication Review or any SIGCOMM sponsored or co-sponsored conference that is deemed to be an outstanding paper whose contents are still a vibrant and useful contribution today. Here is a link to the abstract of the 1999 paper.

July 2010

Best Paper Award at PAKDD 2010

School of Computer Science Ph.D. students Leman Akoglu and Mary McGlohon received the "best paper" award in late June at the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2010). The paper, "OddBall: Spotting Anomalies in Weighted Graphs," by Akoglu, McGlohon and Professor Christos Faloutsos, gives fast algorithms to spot strange nodes in large social networks. The paper was selected among 412 submissions, and 42 accepted papers.

--CMU 8.5x11 News July 15, 2010

July 2010

Gregory Ganger Earns 2010 HP Innovation Research Award

CMU's Gregory Ganger was one of more than 60 recipients worldwide to receive the 2010 HP Innovation Award, which is designed to encourage open collaboration with HP labs resulting in mutually beneficial, high-impact research.

Ganger, a professor of electrical and computer engineering and director of the Parallel Data Lab (PDL) at Carnegie Mellon, will collaborate with HP labs on a research initiative focused on cloud computing issues. This is Ganger's second HP Innovation Award. He received his first HP Innovation Award in 2008 for research involving scalable and self-managing data storage systems.

HP reviewed more than 300 submissions from individuals at 202 univer-

continued on page 11

continued from page 10

sities in 36 countries. Ganger said the award will deepen and strengthen the PDL's long-standing ties with HP and with outstanding researchers globally. The PDL continues to work on solutions to critical problems of storage system design, implementation and evaluation.

"This is a wonderful award for Greg and his team because it recognizes the innovative, collaborative research excellence so endemic to the Parallel Data Lab," said Mark S. Kamlet, executive vice president and provost at Carnegie Mellon. "We applaud their dedication and energy in streamlining ubiquitous cloud computing use."

"The annual HP Labs Innovation Program is an ideal platform for HP to initiate highly innovative projects with leading researchers in universities worldwide. The collaborative effort between HP and these universities has delivered breakthroughs in areas such as cloud computing, optical computing and nano-materials — fundamental enablers of the next generation of products and services for communities around the globe," said Rich Friedrich, director of strategy and innovation at HP.

--CMU Press Release July 8, 2010

July 2010

Gregory Ganger Testifies in Washington About Benefits and Risks of Using Cloud Computing

In testimony to the U.S. House Committee on Oversight and Government Reform and the Subcommittee on Government Management, Organization and Procurement, Gregory Ganger discussed the benefits and risks of using cloud computing.

Ganger, head of Carnegie Mellon's Parallel Data Lab and a professor in the Department of Electrical and Computer Engineering, said that cloud computing has the potential to provide large efficiency improvements for federal information technology (IT) functions. Cloud computing refers to computing that is based on the Internet, which allows computer users to share software, databases and other services that are provided or managed by other parties over the Web. This contrasts with personal computing, where all data storage and processing occurs within the user's computer and uses software loaded onto that computer.

Ganger recommended to federal officials that the government support both standardization and research/experimentation efforts in the pursuit of cloud computing's potential. He also noted that moving federal IT "to the cloud" will require significant technical and change management training for IT staff and managers as well as explicit information and effort sharing across a broad swath of federal agencies considering the use of cloud computing.

"Cloud computing is an exciting realization of a long-sought concept: computing as a utility. Pursuing judicious use for federal IT functions is important, given the large potential benefits," Ganger said.

--Carnegie Mellon Media Notification, June 30, 2010

July 2010

FAWN Team Wins 2010 JouleSort Challenge

Congratulations to Vijay Vasudevan, Lawrence Tan, David Andersen of Carnegie Mellon University, and Michael Kaminsky, Michael A. Kozuch, Padmanabhan Pillai of Intel Labs Pittsburgh for winning the 2010 JouleSort (energy-efficient sort) for the 108 records category. Using FAWNSort on the following equipment (Intel Xeon L3426 1.86GHz, 12GB RAM, Nsort, Fusion-io ioDrive (80GB), 4 x Intel X25-E (3 x 32GB, 1 x 64GB)), they achieved 44,900 records sorted/joule. Medals are awarded each year at ACM SIGMOD. More information on the challenge, including the rules, may be found at http://sortbenchmark.org/.

June 2010

Christos Faloutsos Receives 2010 ACM SIGKDD Innovation Award

Congratulations to Christos Faloutsos, who is the winner of the 2010 ACM SIGKDD Innovations Award. The Innovation Award recognizes one individual or one group of collaborators whose outstanding technical innovations in the field of Knowledge Discovery and Data Mining have had a lasting impact in advancing the theory and practice of the field. The contributions must have significantly influenced the direction of research and development of the field or transferred to practice in significant and innovative ways and/or enabled the development of commercial systems.

Christos, a Professor of Computer Science and Electrical and Computer Engineering, focuses his research on data Mining for graphs and streams, fractals, self-similarity and power laws, indexing and data mining for video, biological and medical databases, and data base performance evaluation (data placement, workload characterization).

May 2010

Lorrie Cranor Expert on Privacy Issues in Advertising Panel

Lorrie Cranor, associate professor of computer science and engineering and public policy, discussed the privacy issues swirling around the technical mechanics of online advertising as part of a panel of experts in Washington, D.C., sponsored by The Progress & Freedom Foundation. Read more about the discussion at http://www.cmu.edu/news/archive/2010/May/may21_onlineadvertisingprivacy.shtml.

-- CMU 8.5x11 News May 27, 2010

DISSERTATION ABSTRACT: Spare a Little Change? Towards a 5-Nines Internet in 250 Lines of Code

Mukesh Agrawal

Carnegie Mellon University SCS Ph.D. Dissertation, Feb. 14, 2011

From its beginnings as a single link between two research institutions in 1969, the Internet has grown in size and scope, to become a global internetwork connecting over 700 million computers, and 1.7 billion users. No longer a niche facility for scientific collaboration, the Internet now touches the lives of the world's population, irrespective of their occupation or geography. It is used by people the world over,to pay bills, read the news, listen to music, watch videos, telephone or videoconference friends and family, and much more. The Internet is the premier communications network of our age.

Unfortunately, however, there are some respects in which the Internet lags the networks it replaces. In particular, with respect to reliability, the Internet falls far short of the Public Switched Telephone Network which proceeded it. Whereas the PSTN sought, and often delivered the vaunted "five nines" of reliability, the Internet struggles to compete. As for the cause of this reliability shortfall, available evidence indicates that much of the shortfall is due to the unreliability of IP routers themselves.

Given the importance of a reliable Internet to contemporary society,vendors and researchers have proposed a number of solutions to either improve the reliability of individual IP routers, or to make networks more resilient to the unavailability of a single router. While having some promise, these existing solutions face significant obstacles to widespread deployment. Thus, in this dissertation, we endeavor to find or construct a practical, readily deployable, method for mitigating the outages caused by IP routers.

To achieve our goal, we take inspiration from previous proposals, which advocated the use of link migration. These proposals improve network resilience, by moving links away from a failed (or failing)router, to an in-service router. To understand the constraints of a practical solution, and resolve the limitations of previous proposals, we conduct extensive experimentation, and study source code and protocol specifications. Using the insights produced by these studies,we construct a practical, readily deployable migration solution with sub-second outage times.

DISSERTATION ABSTRACT: Improving the Dependability of Distributed Systems

Tudor A. Dumitraş

Carnegie Mellon University ECE Ph.D. Dissertation, Dec., 2010

Traditional fault-tolerance mechanisms concentrate almost entirely on responding to, avoiding, or tolerating unexpected faults or security violations. However, scheduled events, such as *software upgrades*, account for most of the system unavailability and often introduce data corruption or latent errors. Through two empirical studies, this dissertation identifies the *leading causes of upgrade failure*–breaking hidden dependencies–*and of planned downtime*–complex data conversions–in distributed enterprise systems. These findings represent the foundation of a new benchmark for software-upgrade dependability.

This dissertation further introduces the *AIR properties*–Atomicity, Isolation and Runtime-testing–required for improving the dependability of distributed systems that undergo major software upgrades. The AIR properties are realized in Imago, a system designed to reduce both planned and unplanned downtime by upgrading distributed systems end-to-end. Imago builds upon the idea of isolating the production system from the upgrade operations, in order to avoid breaking hidden dependencies and to decouple the data conversions from the normal system operation. Imago includes novel mechanisms, such as providing a parallel universe for the new version, performing data conversions opportunistically, intercepting the live workload at the ingress and egress points or executing an atomic switchover to the new version, which allow it to deliver the AIR properties.

Imago harnesses opportunities provided by the emerging cloud-computing technologies, by trading resource overhead (needed by the parallel universe) for an improved dependability of the software upgrades. This approach separates the functional aspects of the upgrade from the mechanisms for online upgrade, enabling an *upgrade-as-a-service* model. This dissertation also describes techniques for assessing the impact of software upgrades, in order to reason about the implications of *relaxing the AIR guarantees*.

DISSERTATION ABSTRACT: Scalable Storage Managers for the Multicore Era

Ryan Johnson

Carnegie Mellon University ECE Ph.D. Dissertation, Nov. 16, 2010

Database Management Systems provide a crucial underpinning for today's information-driven world, providing users with efficient and up-to-date access to huge volumes of data. In order to keep pace with exploding data volumes and increasingly sophisticated processing of that data, database engines must exploit fully the underlying hardware. Recent shifts in computer architecture have led to the rise of multicore designs which depend on parallelism for performance, with the result that today's software must provide exponentially-increasing parallelism to keep the underlying hardware busy.

continued on page 13

continued from page 12

Though database workloads have the advantage of high concurrency both within and between requests, bottlenecks internal to the database engine itself prevent it from converting concurrency into sufficient parallelism, especially in transaction processing workloads. In this thesis, we show how to move the database engine off the critical path, proving that database engines can achieve the scalability needed to exploit today's parallel hardware.

We identify three key areas for achieving this goal. First, performance optimizations must focus first on the critical path. Every serial computation is a liability, while single-thread performance and even the total amount of work performed by a computation are secondary concerns. Second, in order to remain scalable as hardware parallelism continues to increase, bottlenecks must be eliminated – not just reduced – by removing the source of the contention. Finally, we identify scheduling as a critical area for current and future systems. Improper scheduling can increase artificially the length of the critical path in the system, while effective scheduling can eliminate many bottlenecks by changing access patterns and improving regularity in the system. We demonstrate the effectiveness of the above approaches by applying them to state-of-the-art database systems running on highly parallel multicore hardware. These results also generalize to the wider software community, as concerns with critical path, bottlenecks, and scheduling arise in every software domain. Finally, this work demonstrates that many of the remaining challenges in achieving database engine scalability lie with scheduling, suggesting a path toward scalability-enhancing scheduling techniques.

Michelle Mazurek discusses her research on Home Storage with Erik Riedel (EMC) at a PDL Retreat poster session.

DISSERTATION ABSTRACT: Automated Physical Design: A Combinatorial Optimization Approach

Debabrata Dash

Carnegie Mellon University SCS Ph.D. Dissertation, Nov. 15, 2010

One of the most challenging tasks for the database administrator is physically designing the database to attain optimal performance for a given workload. Physical design requires selection of an optimal set of design features from a vast search space. State-of-the-art database design tools rely on the query optimizer for comparing between physical design alternatives, and search for the optimal set of features. Although it provides an appropriate cost model for physical design, query optimization is a computationally expensive process. The significant time consumed by optimizer invocations poses serious performance limitations for physical design tools, causing long running times, especially for large problem instances. Apart from affecting the performance, the overhead also limits the physical design tools from searching the space thoroughly, thus forcing them to prune away the search space to find solutions within a reasonable time. So far it has been impossible to remove query optimization overhead without sacrificing cost estimation precision. Inaccuracies in query cost estimation are detrimental to the quality of physical design algorithms, as they increase the chances of "missing" good designs and consequently selecting sub-optimal ones. Precision loss and the resulting reduction in solution quality is particularly undesirable and it is the reason the query optimizer is used in the first place. In this thesis, we demonstrate that for the physical design problem, the costs returned by the optimizer contain an intuitive mathematical structure. By utilizing this structure, the physical design problem, can be converted to a compact convex optimization problem with integer variables and solved efficiently using mature off-the-shelf solvers. We demonstrate the effectiveness of the approach by finding near-optimal physical design for workloads containing thousands of queries and thousands of candidate design alternatives. In a more complex scenario, the exact queries run by the DBMS vary over time, therefore the optimal physical design for the queries also vary over time. We devise several online algorithms with guaranteed competitive bounds to address the physical design problem for such a workload. We demonstrate that the online algorithms provide significant speedups while imposing reasonable overhead on the system.

DISSERTATION ABSTRACT: Reusing Dynamic Redistribution to Eliminate Cross-Server Operations and Maintain Semantics while Scaling Storage Systems

Shafeeq Sinnamohideen

Carnegie Mellon University SCS Ph.D. Dissertation, July 2, 2010

Distributed file systems that scale by partitioning files and directories among a collection of servers inevitably encounter cross-server operations. A common example is a {rename} that moves a file from a directory managed by one server to a directory managed by another. Systems that provide the same semantics for cross-server operations as for those that do not span servers traditionally implement dedicated protocols for these rare operations. This thesis explores an alternate ap-

continued on page 14

continued from page 13

proach, with simplicity as a goal, that exploits the existence of dynamic redistribution functionality (e.g., for load balancing, incorporation of new servers, and so on). When a client request would involve files on multiple servers, the system can redistribute those files onto one server and have it service the request. Although such redistribution is more expensive than a dedicated cross-server protocol, analyses of file system traces indicates that such operations are extremely rare in file system workloads. Therefore, when dynamic redistribution functionality exists in the system, cross-server operations can be handled with very little additional implementation complexity and a small performance penalty.

THESIS PROPOSAL: Mining and Modeling Real Graphs: Patterns, Generators, Anomalies, and Tools

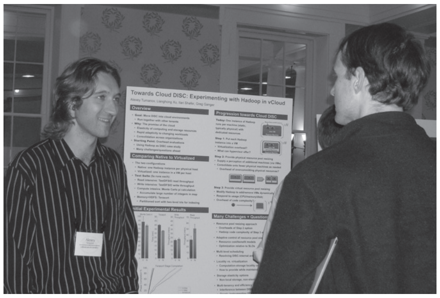
Leman Akoglu, SCS

March 15, 2011

Real graphs (a.k.a. network data) are omnipresent, existing in many domains in the form of social and information networks, collaboration networks, academic citation networks, customer-product networks, road networks, and many more. Not only that real graphs have become ubiquitous but they have also grown in size over the past few decades reaching terabytes of storage. As a result, extracting meaningful and useful knowledge from graph data effectively and efficiently has become very crucial and challenging.

In this thesis, we study real world graphs across many domains from political campaign donation networks to phone call networks and focus on mining weighted and temporal patterns in particular. Using the discovered patterns observed in real graphs, we develop models of network generation that mimic these as well as previously established patterns in weighted, time-evolving graphs. Finally, we apply our findings on several real-world problems such as anomaly detection and recommendation systems. Our completed work can be organized as follows: (1) several surprising patterns in weighted and time-evolving graphs, such as the "fortification effect" (e.g. the more donors a candidate has, the super-linearly more money he/she will raise), connected components, and patterns of human communications; (2) generators such as the RTG model based on "random typing" that mimics eleven of the properties that real, weighted graphs exhibit; and (3) several case studies, including our application of OddBall to the anomaly detection problem in graph data which we applied on seven real graphs, and our application of ValuePick to the problem of integrating "value" into the traditional recommendation process.

Based on our current work, we propose to attack a number of interesting problems in mining graph data. Here, we will extend our study on anomaly detection (1) for time-evolving graphs as well as (2) for graphs with attributes. Potential applications of this part include fraud detection in financial transaction networks that change frequently over time and email networks for which structural (links) as well as semantic (content) information is available. In addition, we propose to develop efficient methods for the node-proximity (a.k.a. k-NN) query processing problem in possibly disk-resident as well as time-evolving graphs, which has commercial value in applications such as recommender systems.



Alexey Tumanov discusses his poster on vCloud with Dave Andersen at the PDL Retreat and Workshop.

THESIS PROPOSAL: Propagation and Immunization on Networks: Theory and Tools

B. Aditya Prakash, SCS

March 15, 2011

Given a who-contacts-whom network and a virus propagation model, can we predict if there will be an epidemic? Which are the best nodes to immunize to slow down and prevent an epidemic as soon as possible? These are central and important questions in surprisingly diverse areas: epidemiology and public health, product marketing, information dissemination etc. Studying the relation of the underlying network to propagation is a vital tool for understanding the spread of contagions which ultimately helps in devising effective control policies. The sudden explosion of large datasets has allowed us to uncover many properties of networks which deviate from standard frameworks and models. As a result, there is a need for better understanding of dynamic processes like propagation on such complex networks.

In this thesis, we focus on theory and algorithms for propagation processes on static and dynamic networks and temporal evolution of data. The core of the thesis concerns itself with virus propagation models on arbitrary, and time-varying networks. The main contributions so far include (a) result that the tipping point or the so-called "epidemic threshold" for all standard models and any graph depends on the largest eigenvalue of the connectivity matrix; and (b) the design of fast and near-optimal algorithms for immunization, again for arbitrary, and time-varying networks. We also present tools to automatically find patterns and anomalies in BGP (Border Gateway Protocol) routing updates,

continued on page 15

continued from page 14

helping us devise simple mathematical models exemplifying router instability propagation. Finally, we present a novel time-series clustering algorithm which can efficiently extract effective and interpretable features allowing us, among other applications, to group different router behavior.

We plan to tackle problems in the same broad areas as before including designing algorithms for more realistic scenarios of immunization (like fractional asymmetric immunization) and application to existing simulation systems.

THESIS PROPOSAL: Algorithms for Large-Scale Astronomical Problems

Bin Fu, SCS

January 27, 2011

Modern astronomical and cosmological datasets are getting larger and larger, including billions of astronomical objects and taking up terabytes of disk space. However, many classical astrophysics applications do not scale to such data volumes, which raises the question: Can we use modern computer science techniques to help astrophysicists analyze large datasets?

In order to answer the question, we have applied distributed computing techniques and developed algorithms to provide fast scalable solutions. In this report we introduce our initial works on three astrophysics applications:

❖ We have developed a distributed version of the Friends-of-Friends technique, which is a standard astronomical application for analyzing clusters of galaxies. The distributed procedure can process tens of billions of objects, which makes it sufficiently powerful for modern astronomical datasets and cosmological simulations.

❖ The computation of correlation functions is a standard cosmological application for analyzing the distribution of matter in the universe. We have studied several approaches to this problem and developed an approximation procedure based on a combination of these approaches, which scales to massive datasets.

❖ When astronomers analyze telescope images, they match the observed objects to the catalog. We have developed a matching procedure that maintains a catalog with billions of objects and processes millions of observed objects per second.

I propose to extend our existing solutions, as well as address several new problems.

THESIS PROPOSAL: Models and Control Strategies for Data Center Energy Efficiency

Luca Parolini, ECE

November 29, 2010

Data center power consumption has significantly increased in the last decade. The Environmental Protection Agency (EPA), indicates that data center power consumption doubled from 2000 to 2006, reaching a value of 60 TWh/year (Tera Watt hour / year). Historical trends suggest another doubling by 2011 that would lead the total power consumption at 120 TWh/year. The peak power consumption of data centers was 7 GW in 2006, equivalent to the output of 15 baseload power plants and it was predicted to reach 12 GW by 2011, requiring the construction of 10 new power plants.

Many strategies to reduce the data center power consumption have been proposed in the literature and some of them have already been implemented. The proposed strategies treat the information-technology (IT) management problem and the facility management problem separately: workload is distributed to the servers to meet performance objectives under the assumption that the cooling system will remove heat as required; the cooling system responds to the thermal load generated by the servers (and other equipment in the data center) through thermostatic control. We call such strategies uncoordinated strategies.

The fundamental problem with uncoordinated strategies is that they cannot take into account the trade-off between allocating the workload to energy-efficient servers and allocating the workload to efficiently cooled servers. This thesis addresses this shortcoming by implementing a cyber-physical approach to the modeling and control of data centers so that the coupling between the computational (IT) and physical (facility) subsystems is explicitly taken into consideration.

The novelty of the proposed management approach resides in the joint dynamic control of both the computational and the physical resources of a data center. Such control approach and the ability to predict future total data center power consumption make it possible the development of new metrics, where the profit induced by executing user workload is weighted against the total cost of powering the data center. Preliminary results show that the proposed control approach can lead to larger income for data center operators than other approaches that do not consider the interactions between the computational and the physical subsystems.

THESIS PROPOSAL: Performance Modeling for Data Center Power Management

Anshul Gandhi, SCS

November 5, 2010

Data centers play an important role in today's IT infrastructure. Government organizations, hospitals, financial trading firms, and major IT companies such as Google, Amazon, IBM, and HP, all rely on data centers for their daily business activities. However, the enormous energy consumption in data

continued on page 16

continued from page 15

centers makes them very expensive to operate. On the one hand, it is desirable to limit the number of servers to reduce power consumption, but on the other hand, obtaining good response times requires having many servers available. This is one of many examples of the power-performance tradeoff faced by data centers today. Thus, an important concern is how to efficiently manage the power-performance tradeoff in data centers.

In this thesis, we propose to design and implement power-management policies for data centers that optimize the power-performance tradeoff. Specifically, we propose to address several important, yet unanswered questions in data center power management, including: (i) How many servers are needed to handle the incoming load? (ii) When should servers be turned on/turned off/left idle/put to sleep? (iii) At what frequencies should servers be run? (iv) What policy should be used to route jobs to servers? In order to answer the above questions, we follow a two-pronged approach consisting of:

1. Performance Modeling: Performance modeling is a useful tool for analyzing the behavior of large computer systems, and it has been traditionally used to predict and improve system performance. However, power necessitates the development of new models and novel analysis involving multiple CPU operating frequencies, multiple server states (busy, idle, sleep and off) and the various setup costs involved in transitioning between server states. Thus, we propose to come up with new queueing-theoretic models that will allow us to analyze the various power-performance tradeoffs in data centers.

PDL Alums Chris Lumb (Data Domain)-L and John Bucy (Google)-R enjoy reconnecting with old friends at the PDL reunion held in San Jose during the FAST '11 conference.

2. Implementation and Experimental Evaluation: While our proposed analysis will guide us in optimizing the power-performance tradeoff, it cannot completely model today's complex data centers. Thus, we propose to implement and experimentally evaluate our proposed policies on an experimental test bed. This requires figuring out the right experimental setup and workload suite for evaluation. Finally, based on our implementation results, we plan to modify our proposed policies in order to tailor them as practical solutions for real-world data centers.

THESIS PROPOSAL:
Improving Datacenter Energy Efficiency using a Fast Array of Wimpy Nodes

Vijay Vasudevan, SCS

October 12, 2010

Energy has become an increasingly large financial and scaling burden for computing. With the increasing demand for and scale of Data-Intensive Scalable Computing (DISC), the costs of running large data centers are becoming dominated by power and cooling. In this thesis we propose to help reduce the energy consumed by large-scale computing by using a FAWN: A Fast Array of Wimpy Nodes. FAWN is an approach to building datacenters using low-cost, low-power hardware devices that are individually optimized for energy efficiency (performance/watt) rather than raw performance alone. FAWN nodes are individually resource-constrained, motivating the development of distributed systems software with efficient processing, low memory consumption, and careful use of flash storage.

In this proposal, we investigate the applicability of FAWN to data-intensive workloads. First, we present FAWN-KV: a deep study into building a distributed key-value storage system on a FAWN prototype. We then present a broader classification and workload analysis showing when FAWN can be more energy-efficient, and under what conditions that wimpy nodes perform poorly. Based on our experiences building software for FAWN, we finish by presenting Storage Click: a software architecture for providing efficient processing of remote, small storage objects.

THESIS PROPOSAL:
Stochastic Models and Analysis for Resource Management in Server Farms

Varun Gupta, SCS

October 8, 2010

Server farms are popular architectures for computing infrastructures such as supercomputing centers, data centers and web server farms. As server farms become larger and their workloads more complex, designing efficient policies for managing the resources in server farms via trial-and-error becomes intractable. It is hard to predict the exact effect of various parameters on performance. Stochastic modeling and analysis techniques allow us to understand the performance of such complex systems and to guide design of policies to optimize the performance. However, most existing models of server farms are motivated by telephone networks, inventory management systems, and call centers. Modeling assumptions which hold for these problem domains are not accurate for computing server farms.

There are numerous gaps between traditional models of multi-server systems and how today's server farms operate. To cite a few: (i) Unlike call durations, supercomputing jobs and

continued on page 17

continued from page 16

file sizes have high variance in service requirements and this critically affects the optimality and performance of scheduling policies. (ii) Most existing analysis of server farms focuses on the First-Come-First-Served (FCFS) scheduling discipline, while time sharing servers (e.g., web and database servers) are better modeled by the Processor-Sharing (PS) scheduling discipline. (iii) Time sharing systems typically exhibit thrashing (resource contention) which limits the achievable concurrency level, but traditional models of time sharing systems ignore this fundamental phenomenon. (iv) Recently, minimizing energy consumption has become an important metric in managing server farms. State-of-the-art servers come with multiple knobs to control energy consumption, but traditional queueing models don't take the metric of energy consumption or these control knobs into account.

In this thesis we attempt to bridge some of these gaps by bringing the stochastic modeling and analysis literature closer to the realities of today's compute server farms. We introduce new models for computing server farms, develop stochastic analysis techniques to evaluate their performance, and propose resource management policies to optimize their performance.

PDL FORECAST: CLOUDY IN A GOOD WAY

Greg Ganger

Cloud computing has become a source of enormous buzz and attention, promising great reductions in the effort of establishing new applications/services, increases in the efficiency of operating them, and improvements in ability to share data and services. Despite the hype and energy around it, though, cloud computing is in its nascent stage.

Realizing the great promise of cloud computing will require an enormous amount of research and development across a broad array of topics. PDL is heavily involved in many of those topics, particularly the ones related to storage and data-processing infrastructure, and the scope of PDL's involvement is growing.

Some current activities are illustrated in the various summaries throughout the PDL Packet and in the project descriptions on the PDL web site. For example, we are exploring new approaches to dynamic sizing of data-intensive computing (DISC) systems, like Hadoop and Google's MapReduce+GFS, which have storage spread across the same nodes as the computation.

We are exploring new approaches to performance insulation and QoS for storage services shared by multiple tenants. We are exploring efficient and highly scalable DISC systems, which will play a major role in future cloud computing. And, the FAWN (Fast Array of Wimpy Nodes) project continues to explore specialized platforms for energy-efficient execution of specific workloads.

To complement our explorations of specific questions, we deployed two clusters last year for use as cloud computing infrastructures. One (labelled "OpenCloud") is set up as a Hadoop service used by various scientists that mine large quantities of data. The other (labelled "OpenCirrus") is set up as a virtual machine based service (based on the open source Tashi cloud computing software) used by various researchers that need computation for their work. Both are heavily instrumented to provide us with insight into the usage patterns and efficiencies of such clouds. Both also represent test environments for improved tools and algorithms for managing and using cloud computing infrastructures.

Over the last year, we've also begun a collaboration with VMware around deploying a cloud computing infrastructure based on their software that implements the proposed vCloud API standard. VMware is providing the software, as well as expertise, and we will collaborate on such issues as instrumentation and automation. The initial deployment is based on a small collection of blade servers, donated by HP, Intel, and Samsung. Those same companies have committed to donating a substantial state-of-the-art private cloud configuration capable of serving a broad collection of research computing activities and making for an interesting contrast to the inexpensive clusters based on open-source software.

We are also exploring with Intel researchers and other academics the possibility of a new cloud computing research center that might be funded by Intel and homed at Carnegie Mellon. Although it is still in the planning stage, such a center would focus on underlying infrastructure technologies for cloud computing. As such, there would be substantial overlap in research activities (and people) with PDL, amplifying both. I visualize the situation as a Venn diagram, where PDL and the new center would be unique and yet overlapping, each with some topics shared and others not. This is an exciting possibility, and we hope to experience and share the great things that it would enable.

continued from page 1

of the write head. Conventional write heads write a narrow track with sharply defined edges on both sides, tolerant of the skew introduced by different angular positions (inner to outer diameters). Shingling means that the track written can be wider, allowing a stronger write field, and only one edge needs to be sharp.

Shingling Geometry Model

Inspired by Amer's model [Amer10] for disk density sacrificed by dedicating a fraction, f, of each surface to unshingled tracks, we offer the following model for increased density by shingled writing. In this model, illustrated in Figure 1, conventional recording writes a track of width w nanometers, with guard gaps between tracks of g nanometers, typically much smaller than w. Alternatively shingled writing, with its wider tracks of w' nanometers, achieves a sharper edge so that the adjacent, down-band track may be as close as r nanometers. Adding a division of tracks into (1-f) shingled and (f) unshingled, gives an areal density increase factor, as a function of tracks per band, B (B=3 in Figure 1), of:

$$\text{Areal Density Increase Factor} = S(f/S' + B(1-f)/(S'+B-1))$$

$$\text{where } S = (w+g)/r,$$

$$\text{and } S' = (w'+g)/r$$

The best increase in areal density to result from switching from conventional to shingled magnetic recording, using this mod-el, would be (w+g)/r. One recent study suggested that this could be as large as 2.25 [Tagawa09], if technologists can achieve shingled write heads with very sharp edges.

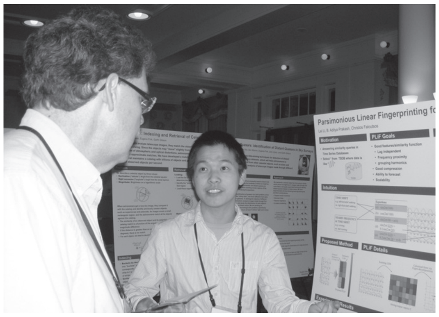
One interpretation of the model is that if residual track width is not much smaller than conventional track width, very large band sizes will be needed to achieve significant increases in areal density. Conversely, if residual track width can be much smaller than conventional track width, then relatively small band sizes can be quite effective.

Shingled Writing Systems Issues

SMR offers few technology and manufacturing challenges, but does impose significant restrictions on where data can be written without incurring multi-track read-modify-write penalties. Specifically, rewriting a sector on a track that has been shingled over cannot be done without overwriting subsequent ("down-band") tracks. As a result, data on a shingled disk will be organized into bands of shingled tracks with a non-shingled guard region between bands.

In an invited presentation to a magnetics research conference three years ago, magnetic technologists asked us: "can system software cope with shingled writing?" [Gibson09]. A straightforward answer is to say that read-modify-write of entire bands on every small random write would yield such a reduction in write performance that such a product would be difficult to use. But the flash translation layer (FTL) in modern SSDs has the same problem (large erase blocks and slow erase functions), and this has been overcome by remapping logical block addresses to flash pages, overprovisioning of flash capacity and background cleaning/defragmenting of logically overwritten values.

With FTLs in mind, one approach to integrating shingled writing into magnetic (hard) disk drives is to transparently emulate the full range


Lei Li discusses his research on Parsimonious Linear Fingerprinting with Craig Everheart (NetApp) at the PDL Retreat & Workshop.

of non-shingled disk operations in a FTL-like embedded controller, a "Shingle Translation Layer" (STL) [Gibson09]. STL firmware could, at one extreme, inexpensively implement slow read-modify-write for essentially all writes to the shingled disk. At the other extreme it could expansively remap the physical location of written data to avoid read-modify-write, dynamically defining band boundaries, employing large write-back caches and overprovisioned disk capacity to hide a background garbage collection and defragmentation process. STL transparency is to be prized, but the complexity of achieving "performance transparency" is so challenging that non-transparent interfaces should be considered too.

If a shingled disk is to have an interface different from that of current SSDs and HDDs, determining what this new interface should be is a key challenge for systems software experts. Possible options include:

- ❖ No change – full emulation of traditional HDD operations embedded in the shingled disk,
- ❖ Large sector disks – if sectors are logically tens to hundreds of megabytes and fixed, then bands

continued on page 19

continued from page 18

will always be logically erased and written consecutively,

- Append only fixed sized bands – sequential writing of bands is achieved by systems software logging changes and per-forming garbage collection at a higher level,
- Separate unshingled region – a small portion of the disk can be used for bands of one track only, eliminating down-band adjacent tracks, and allowing small random writes to be unpenalized,
- A data management and hints interface – band geometry discovery, creation/deletion of unshingled bands, etc., may enhance performance,
- An object storage interface – SCSI has a command set for access by object ID with variable length data per object – widely used for scalable storage systems in high-performance computing, and
- A virtual tape model – if the entire device can be shingled into one band and higher level software creates band gaps (tape file marks and gaps) dynamically as needed, then maximal density is available to higher level software.

Raja, looking like he doesn't know where to go next with his research. Hopefully, advice from our industry guests at the PDL Retreat helped point him in the right direction.

Key aspects in many of these will be the explicit exposure of the band size, alignment and placement to higher-level software and good support for "usable" band sizes.

Our premise is that a simple and inexpensive STL be embedded in each shingled disk, and that the system model of shingled disks be extended to facilitate system software to optimize shingling without fighting or reverse engineering the STL. Specifically, system software should:

- tell the STL when overwriting a down-band sector is free of consequences,
- be aware and in control of bands of shingled tracks so it can avoid writing in the middle of a previously written band,
- conceptualize writing to shingled bands as overwriting consecutively or appending to a log, and
- perform necessary data remapping and garbage collection (rather than forcing the STL to do this).

The selection of the right SMR system interface needs research, discussion and, in general, airing among systems software experts now, before the magnetic disk technologists finalize the principles of operation for shingled magnetic recording.

With the appropriate interface and system model, we believe that shingled disks can achieve marketing targets for geometric decreases in cost per bit, and system software can fully exploit the underlying shingled architecture without reverse engineering complex STL behaviors.

For more details, please see [Gibson11].

References

[Amer10] Amer, A., D.D.E. Long, E.L. Miller, J.-F. Paris, T. Swarz S.J., "Design Issues for a Shingled Write Disk System," 26th IEEE (MSST2010) Symposium on Massive Storage Systems and Technologies, Lake Tahoe, NV, May 2010.

[Gibson11] Gibson, G., G. Ganger, "Principles of Operation for Shingled Disk Devices," Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-11-107, April 2011.

[Gibson09] Gibson, G., M. Polte, "Directions for shingled-write and two-dimensional magnetic recording system architectures: Synergies with solid-state disks," Carnegie Mellon University Parallel Data Lab Technical Report, CMU-PDL-09-104, May 2009. Also FA-06 presentation in INTERMAG 2009, Sacramento, CA, May 2009.

[Wood00] Wood, R., "Feasibility of Magnetic Recording at 1 Tbit/inch^2," IEEE Transactions on Magnetics, v36, n1, 2000.

[Tagawa09] Tagawa, L., M. Williams, "Shingle-Write Technology and Gain Estimation," FA-02 presentation in INTERMAG 2009, Sacramento, CA, May 2009.

Mark your calendars for the 19th Annual Parallel Data Lab Workshop and Retreat, to be held at Bedford Springs Resort in Bedford Springs, PA, from November 7 to 9, 2011.

ENABLING PHASE CHANGE MEMORY AS MAIN MEMORY

Onur Mutlu & Joan Digney

Existing main memory systems are built using DRAM (dynamic random access memory) storage technology. While DRAM memories can be constructed with high bandwidth and low latency, the technology is facing significant challenges. First, the demand for main memory capacity and bandwidth is increasing, with the increasing number of cores placed on a single chip, data-intensive applications demanding more data, and the increasing need/trend for consolidation of many applications on a single system. Second, power and energy consumption of DRAM-based main memory is becoming a significant concern: DRAM memory consumes power even when idle and needs periodic refresh of volatile data cells. Third, the scaling of DRAM technology to smaller feature sizes is becoming increasingly difficult. As a result, DRAM alone will likely be inefficient and insufficient in building the main memory hierarchy of future systems. Our goal is to rethink the main memory hierarchy in the presence of these challenges and explore the potential of new memory technologies to replace or augment DRAM. We aim to also explore the potential of alternative DRAM architectures to allow better scaling of DRAM and making DRAM a more efficient component of the memory hierarchy.

Non-volatile memory/storage (NVM) technologies such as Flash, Phase Change Memory (PCM), and magnetic memory (MRAM) are promising due to their anticipated capacity benefits, non-volatility, and zero idle energy. Unfortunately, these emerging memory technologies have serious shortcomings compared to DRAM: 1) they are significantly slower to access, 2) they have very low endurance, 3) they have very high write latency and write energy. Our goal is to redesign the memory hierarchy to overcome these challenges and exploit the new opportunities of NVM technologies. We are rethinking the entire virtual memory design and main memory system to integrate PCM as a fundamental main memory component, with the goal of designing a significantly more energy-efficient, cheaper, scalable, high-capacity, and more capable memory/storage system. To this end, we have so far developed new architectural techniques for designing PCM chips and memory subsystems to improve performance and efficiency of PCM-based main memory. We briefly summarize our preliminary ideas that were published at the 36th Annual International Symposium on Computer Architecture [Lee, ISCA09] and Communications of the ACM [Lee, CACM10].

In order to bring PCM technology within competitive range of DRAM, we propose:

❖ Buffer Reorganization: We reexamine buffer organization in PCM chips. Narrow buffers in PCM chips mitigate high energy PCM writes. Multiple row buffers exploit locality to coalesce writes, hiding their latency and reducing their energy. Effective PCM buffering reduces application execution time from 1.6× to 1.2× and memory array energy from 2.2× to 1.0×, relative to DRAM-based systems [Lee, ISCA10][Lee, CACM10].

❖ Partial Writes: We propose partial writes, which track data modifications and write only modified cache lines or words to the PCM array. We expect write coalescing and partial writes to deliver an average memory module lifetime of 11.2 years. PCM endurance is expected to improve by four orders of magnitude when scaled to 32nm.

Collectively, these results suggest PCM could be a viable DRAM alternative, with architectural solutions providing competitive performance, comparable energy, and feasible lifetimes [Lee, CACM10].

PCM Technology

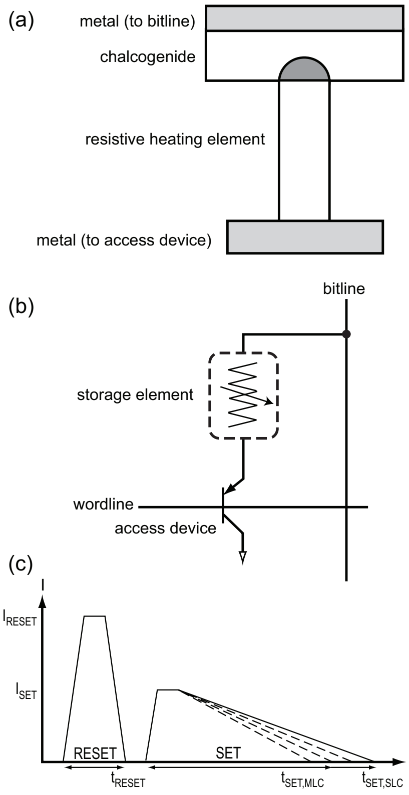


Figure 1: Phase change memory. (a) Storage element with heating resistor and chalcogenide between electrodes. (b) Cell structure with storage element and BJT access device. (c) Reset to an amorphous, high resistance state with a high, short current pulse. Set to a crystalline, low resistance state with moderate, long current pulse. Slope of set current ramp down determines the state in MLC.

As shown in Figure 1a, the PCM storage element is comprised of two metal electrodes separated by a resistive heater and a chalcogenide, the phase change material. $Ge_2Sb_2Te_5$ (GST) is most commonly used. Figure 1b shows that PCM cells are 1T/1R devices, comprised of the resistive storage element and an access transistor. Access is typically controlled by one of three devices: field-effect transistor (FET), bipolar junction transistor (BJT), or diode. Phase changes are induced by injecting current into the resistor junction and heating the chalcogenide. Current and voltage characteristics of the chalcogenide are identical regardless of its initial phase. The amplitude and width of the injected current pulse

continued on page 21

continued from page 20

determine the programmed state as shown in Figure 1c.

Operation: The access transistor injects current into the storage material and thermally induces phase change, which is detected as a programmed resistance during reads. Logical data values are captured by the resistivity of the chalcogenide. A high, short current pulse increases resistivity by abruptly discontinuing current, quickly quenching heat generation, and freezing the chalcogenide into an amorphous state (i.e., reset). A moderate, long current pulse reduces resistivity by ramping down current, gradually cooling the chalcogenide, and inducing crystal growth (i.e., set). Set latency determines write performance and reset energy determines write power. Writes are the primary wear mechanism in PCM as the thermal expansion and contraction degrades the electrode-storage contact. Write endurance, the number of writes performed before the cell cannot be programmed reliably, ranges from 1E+04 to 1E+09.

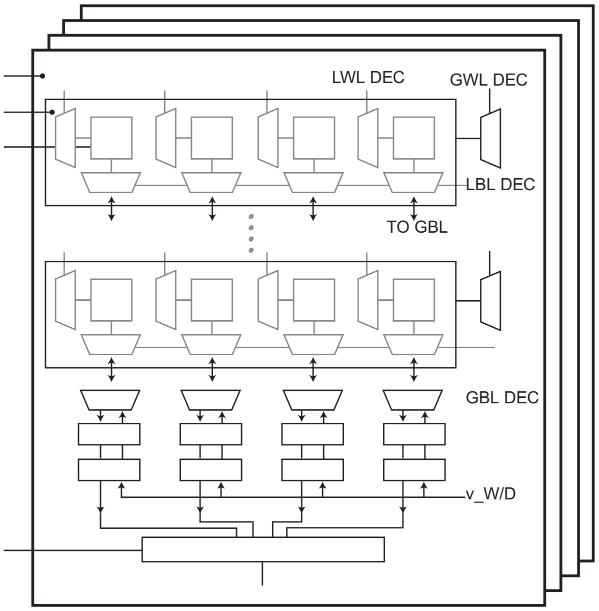


Figure 2. Array architecture. A hierarchical memory organization includes banks, blocks, and subblocks with local, global decoding for row, column addresses. Sense amplifiers (S/A) and word drivers (W/D) are multiplexed across blocks.

Architecting A DRAM Alternative

To enable PCM for practical use as main memory in general-purpose systems, we must close the delay and energy gap between PCM and DRAM. Nondestructive PCM reads help mitigate underlying delay and energy disadvantages by default. We seek to eliminate the remaining PCM-DRAM differences with architectural solutions.

Array Architecture: As shown in Figure 2, PCM cells might be hierarchically organized into banks, blocks, and subblocks. Sense amplifiers detect the change in bitline state when a memory row is accessed. Choice of bitline sense amplifiers affects array read access time. Voltage sense amplifiers are cross-coupled inverters which require differential discharging of bitline capacitances.

In DRAM, sense amplifiers both sense and buffer data using cross-coupled inverters; PCM architecture separates sensing and buffering. Sense amplifiers drive banks of explicit latches providing greater flexibility in row buffer organization by enabling multiple buffered rows. However, these latches incur area overheads. Separate sensing and buffering enables multiplexed sense amplifiers, which in turn enables buffer widths narrower than the array width. Buffer width is a critical design parameter, determining the number of expensive current sense amplifiers required.

Buffer Reorganization: To be a viable DRAM alternative, buffer organizations must hide long PCM latencies, while minimizing PCM energy costs. To achieve area neutrality across buffer organizations, we consider narrower buffers and additional buffer rows. The number of sense amplifiers decreases linearly with buffer width, significantly reducing area as fewer large circuits are required. We utilize this area by implementing multiple rows with latches much smaller than the removed sense amplifiers. Narrow widths reduce PCM write energy but negatively impact spatial locality, opportunities for write coalescing, and application performance. However, these penalties may be mitigated by the additional buffer rows.

Buffer reorganizations impact the degree of exploited locality and energy costs associated with array reads and writes. Our results in [Lee, CACM10] show that reorganizing a single, wide buffer into multiple, narrow buffers reduce both energy costs and delay. We consider buffer widths ranging from the original 2048B (common in DRAM) to 64B, which is the line size of the lowest level cache. We consider buffer rows ranging from the original single row (standard in DRAM) to a maximum of 32 rows. Among all evaluated buffer organizations, we observe a knee that minimizes both energy and delay; this organization uses four 512B-wide buffers to reduce PCM delay, energy disadvantages from 1.6×, 2.2× to more modest 1.2×, 1.0×.

Partial Writes: In addition to architecting PCM to offer competitive delay and energy characteristics relative to DRAM, we must also consider PCM wear. To mitigate wear, we propose partial writes, reducing the number of writes to the PCM array by tracking modified data from the L1 cache to the memory banks. When a buffered row is evicted and contents written to the PCM array, only modified data is

continued on page 22

ENABLING PHASE CHANGE MEMORY AS MAIN MEMORY

continued from page 21

written. Partial writes are supported by adding a modest amount of cache state to reduce the number of bits written. At the dirty word granularity, 4B modifications are tracked beginning at the L1 cache with 8b per 32B L1 line and propagated to the L2 cache with 16b per 64B L2 line. Hardware overhead is 3.1% of each cache line when tracking dirty words.

Partial writes, combined with an effective buffer organization, increase memory module lifetimes to a degree that makes PCM in main memory feasible.

Lifetime: Our ISCA 2009 paper develops an analytical model to estimate PCM lifetime as a function of different granularity of partial writes, PCM capacity, and application characteristics [Lee, ISCA2009]. An unmodified PCM architecture (with a single 2048B-wide row buffer and no partial writes) has an average module lifetime of approximately 1050 hours. Reorganized buffers and partial writes together lead to significant endurance gains, increasing average module lifetime to 11.2 years [Lee, CACM2010].

Limitations and Current Work: Our analysis in previous work has considered a limited set of memory intensive parallel workloads and a limited set of possible PCM parameters (in terms of latency, energy, and endurance). Since PCM is an evolving technology, architectural techniques need to perform well with a variety of chip parameters. In our current work, we are examining a larger set of data intensive workloads and a wider variety of possible PCM parameters.

Conclusion

The proposed memory architecture lays the foundation for exploiting PCM scalability and nonvolatility in main memory. Scalability implies lower main memory energy, greater write endurance, lower cost, and higher capacity. Furthermore, nonvolatile main memories can fundamentally change the landscape of computing. For example, system boot/hibernate can be perceived as instantaneous; application and system checkpointing can become inexpensive; file systems can provide stronger safety guarantees. Thus, we take a step toward a new memory hierarchy with deep implications across the hardware–software interface.

References

[Lee CACM10] "Phase Change Memory Architecture and the Quest for Scalability," Benjamin C. Lee, Engin Ipek, Onur Mutlu, Doug Burger. Communications of the ACM (CACM), Research Highlight, Vol. 53, No. 7, pages 99-106, July 2010.

[Lee ISCA09] Architecting Phase Change Memory as a Scalable DRAM Alternative. Benjamin C. Lee, Engin Ipek, Onur Mutlu, Doug Burger. Proceedings of the 36th International Symposium on Computer Architecture (ISCA), pages 2-13, Austin, TX, June 2009.

and file systems
designing systems with new storage te
home/ nal data management and s

Greg opens the show with his PDL Research Overview talk, and welcomes our Consortium visitors to the 18th Annual Workshop and Retreat, held at CMU and Bedford Springs, PA.

YEAR IN REVIEW

continued from page 4

- Leman Akoglu was awarded Best Paper for "OddBall: Spotting Anomalies in Weighted Graphs" at PAKDD 2010 in Hyderabad, India.
- Michelle Mazurek interned under Eno Thereska at MSR-Cambridge.
- Mike Kasick interned with IBM Almaden.
- Raja Sambasivan interned at Google Pittsburgh, continuing his work on performance problem diagnosis using end-to-end traces.
- Elie Krevat and Jim Cipar interned with HP Labs.
- Lin Xiao interned with Google.
- Wittawat Tantisiriroj interned with Yahoo! in Sunnyvale, CA.
- Ilari Shafer interned with Microsoft in Redmond, WA.

May 2010

- Lorrie Cranor discussed privacy issues with regard to the technical mechanics of online advertising as part of a panel of experts in Washington, D.C., sponsored by The Progress & Freedom Foundation.
- Bruno Sinopoli awarded a 5 year NSF Career Award grant.
- Kai Ren interned at Facebook from May through August, joining their data infrastructure team to do projects related to Hive or Hadoop.
- 12th Annual PDL Spring Industry Visit Day.

continued from page 7

series records belonging to over 2 million black holes. We demonstrate that this is a feasible approach to support interactive analysis and enables flexible exploration of black hole forest datasets.

Behavior-Based Problem Localization for Parallel File Systems

Kasick, Gandhi & Narasimhan

HotDep '10. October 3, 2010, Vancouver, BC, Canada.

We present a behavior-based problem-diagnosis approach for PVFS that analyzes a novel source of instrumentation—CPU instruction-pointer samples and function-call traces—to localize the faulty server and to enable root-cause analysis of the resource at fault. We validate our approach by injecting realistic storage and network problems into three different workloads (dd, IO-zone, and PostMark) on a PVFS cluster.

Otus: Resource Attribution and Metrics Correlation in Data-Intensive Clusters

Ren, Lopez &Gibson

MapReduce Workshop (MAPREDUCE'11), June 8, 2011 HPDC'2011, San Jose, CA, USA.

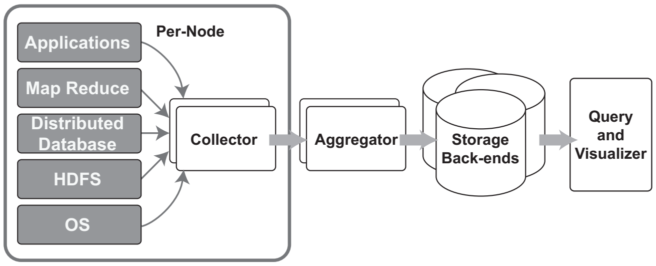
Frameworks for large scale data-intensive applications, such as Hadoop, Dryad, have gained tremendous popularity. Understanding the resource requirements of these frameworks and the performance characteristics of distributed applications is inherently difficult. In this paper, we describe the design and implementation of Otus, a monitoring tool that focuses on resource attribution and metrics correlation. Otus monitors resources utilized by different services and applications in the cluster, correlates various time-series metrics data and provides a visualization system to analyze performance data under multiple views. Our experience of using Otus in a production cluster suggests its effectiveness of helping users and cluster administrators with application performance analysis and troubleshooting.

To Upgrade or Not to Upgrade: Impact of Online Upgrades across Multiple Administrative Domains

Dumitraş, Tilevich & Narasimhan

ACM Onward! Conference, Oct. 2010, Reno, NV.

Online software upgrades are often plagued by runtime behaviors that are poorly understood and difficult to ascertain. For example, the interactions among multiple versions of the software expose the system to race conditions that can introduce latent errors or data corruption. Moreover, industry trends suggest that online upgrades are currently needed in large-scale enterprise systems, which often span multiple administrative domains (e.g., Web 2.0 applications that rely on AJAX client-side code or systems that lease cloud computing resources). In such systems, the enterprise does not control all the tiers of the system and cannot coordinate the upgrade process, making existing techniques inadequate to prevent mixed-version races. In this paper, we present an analytical framework for impact assessment, which allows system administrators to directly compare the risk of following an online-upgrade plan with the risk of delaying or canceling the upgrade. We also describe an executable model that implements our formal impact assessment and enables a systematic approach for deciding whether an online upgrade is appropriate. Our model provides a method of last resort for avoiding undesirable program behaviors, in situations where mixed-version races cannot be avoided through other technical means.



The architecture of the Otus monitoring system.

Behavior-Based Problem Localization for Parallel File Systems

Kasick, Gandhi & Narasimhan

HotDep '10. October 3, 2010, Vancouver, BC, Canada.

We present a behavior-based problem-diagnosis approach for PVFS that analyzes a novel source of instrumentation—CPU instruction-pointer samples and function-call traces—to localize the faulty server and to enable root-cause analysis of the resource at fault. We validate our approach by injecting realistic storage and network problems into three different workloads (dd, IO-zone, and PostMark) on a PVFS cluster.

Token Attempt: The Misrepresentation of Website Privacy Policies through the Misuse of P3P Compact Policy Tokens

Leon, L. Cranor, McDonald & McGuire

Cylab Technical Report CMU-CyLab-10-014, September 10, 2010.

Platform for Privacy Preferences (P3P) compact policies (CPs) are a collection of three-character and four-character tokens that summarize a website's privacy policy pertaining to cookies. User agents, including Microsoft's

continued on page 24

continued from page 23

Internet Explorer (IE) web browser, use CPs to evaluate websites' data collection practices and allow, reject, or modify cookies based on sites' privacy practices. CPs can provide a technical means to enforce users' privacy preferences if CPs accurately reflect websites' practices. Through automated analysis we can identify CPs that are erroneous due to syntax errors or semantic conflicts. We collected CPs from 33,139 websites and detected errors in 11,176 of them, including 134 TRUSTe-certified websites and 21 of the top 100 most-visited sites. Our work identifies potentially misleading practices by web administrators, as well as common accidental mistakes. We found thousands of sites using identical invalid CPs that had been recommended as workarounds for IE cookie blocking. Other sites had CPs with typos in their tokens, or other errors. 98% of invalid CPs resulted in cookies remaining unblocked by IE under it's default cookie settings. It appears that large numbers of websites that use CPs are misrepresenting their privacy practices, thus misleading users and rendering privacy protection tools ineffective. Unless regulators use their authority to take action against companies that provide erroneous machine-readable policies, users will be unable to rely on these policies.

Parsimonious Linear Fingerprinting for Time Series

Li, Prakash & Faloutsos

Proceedings of the VLDB Endowment, Vol. 3, No. 1, September 2010.

We study the problem of mining and summarizing multiple time series effectively and efficiently. We propose PLiF, a novel method to discover essential characteristics ("fingerprints"), by exploiting the joint dynamics in numerical sequences. Our ¯fingerprinting method has the following benefits: (a) it leads to interpretable features; (b) it is versatile: PLiF enables numerous mining tasks, including clustering, compression, visualization, forecasting, and segmentation, matching top competitors in each task; and (c) it is fast and scalable, with linear complexity on the length of the sequences.

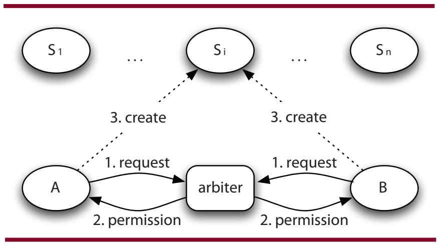
We did experiments on both synthetic and real datasets, including human motion capture data (17MB of human motions), sensor data (166 sensors), and network router traffic data (18 million raw updates over 2 years). Despite its generality, PLiF outperforms the top clustering methods on clustering; the top compression methods on compression (3 times better reconstruction error, for the same compression ratio); it gives meaningful visualization and at the same time, enjoys a linear scale-up.

dBug: Systematic Evaluation of Distributed Systems

Simsa, Bryant & Gibson

5th Int. Workshop on Systems Software Verification (SSV'10), co-located with 9th USENIX Symp. On Operating Systems Design and Implementation (OSDI'10), Vancouver BC, October 2010.

This paper presents the design, implementation and evaluation of "dBug" – a tool that leverages manual instrumentation for systematic evaluation of distributed and concurrent systems. Specifically, for a given distributed concurrent system, its initial state and a workload, the dBug tool systematically explores possible orders in which concurrent events triggered by the workload can happen. Further, dBug optionally uses the partial order reduction mechanism to avoid exploration of equivalent orders. Provided with a correctness check, the dBug tool is able to verify that all possible serializations of a given concurrent workload execute correctly. Upon encountering an error, the tool produces a trace that can be replayed to investigate the error.



Steps taken to send a message: 1) An agent requests permission from the arbiter, 2) The arbiter grants the permission, 3) The agent sends the message.

Greg congratulates Erik Riedel as he names Erik a "PDL Distinguished Alumni." Erik graduated from CMU and the PDL in 1999 and is now Senior Director of Technology & Architecture at EMC.

We applied the dBug tool to two distributed systems – the Parallel Virtual File System (PVFS) implemented in C and the FAWN-based key-value storage (FAWN-KV) implemented in C++. In particular, we integrated both systems with dBug to expose the non-determinism due to concurrency. This mechanism was used to verify that the result of concurrent execution of a number of basic operations from a fixed initial state meets the high-level specification of PVFS and FAWN-KV. The experimental evidence shows that the dBug tool is capable of systematically exploring behaviors of a distributed system in a modular, practical, and effective manner.

OddBall: Spotting Anomalies in Weighted Graphs

Akoglu, McGlohon & Faloutsos

PAKDD 2010, Hyderabad, India, 21-24 June 2010. Best Paper Award.

continued on page 25

continued from page 24

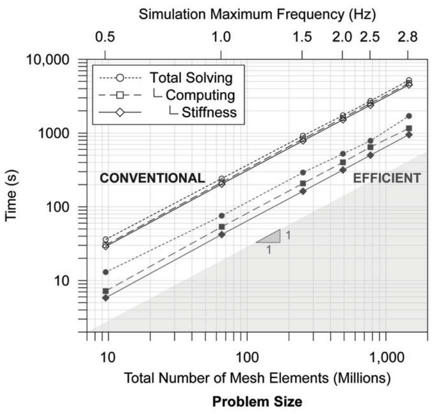
Given a large, weighted graph, how can we find anomalies? Which rules should be violated, before we label a node as an anomaly? We propose the OddBall algorithm, to find such nodes. The contributions are the following: (a) we discover several new rules (power laws) in density, weights, ranks and eigenvalues that seem to govern the so-called "neighborhood sub-graphs" and we show how to use these rules for anomaly detection; (b) we carefully choose features, and design OddBall, so that it is scalable and it can work unsupervised (no user-defined constants) and (c) we report experiments on many real graphs with up to 1.6 million nodes, where OddBall indeed spots unusual nodes that agree with intuition.

Speeding Up Finite Element Wave Propagation for Large-Scale Earthquake Simulations

Taborda, López, Karaoglu, Urbanic & Bielak

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-109. October 2010.

This paper describes the implementation and performance of a new approach to finite element earthquake simulations that represents a speedup factor of 3x in the total solving time employed by Hercules—the octree-based earthquake simulator developed by the Quake Group at Carnegie Mellon University. This gain derives from applying an efficient method for computing the stiffness contribution at the core of the solving algorithm for the discretized equations of motion. This efficient method is about 5 times faster than our previous conventional implementation. We evaluate the performance and scalability of the new implementation through numerical experiments with the 2008 Chino Hills earthquake under various problem sizes and resource conditions on up to 98K CPU cores, obtaining excellent results. These experiments required simulations with up to 11.6 billion mesh elements. The newly obtained efficiency reveals that other areas in Hercules, such as inter-processor communication, waiting time, and additional computing processes become more critical, and that improvements in these areas will result in significant enhancement in overall performance. This latest advance has enormous implications for saving CPU hours and catapults the potential of Hercules to target larger and more realistic problems, taking full advantage of the new generation of petascale supercomputers.



Scaling for the fixed-size resource case. The problem size varies across executions (X axis), while the number of CPU cores is fixed at 1032. The Y axis shows the elapsed wall-clock time.

Optimality Analysis of Energy-Performance Trade-offs for Server Farm Management

Gandhi, Gupta, Harchol-Balter & Kozuch

28th International Symposium on Computer Performance, Modeling, Measurements, and Evaluation (Performance 2010) Namur, Belgium, November 2010.

A central question in designing server farms today is how to efficiently provision the number of servers to extract the best performance under unpredictable demand patterns while not wasting energy. While one would like to turn servers off when they become idle to save energy, the large setup cost (both, in terms of setup time and energy penalty) needed to switch the server back on can adversely affect performance. The problem is made more complex by the fact that today's servers provide multiple sleep or standby states which trade off the setup cost with the power consumed while the server is 'sleeping'. With so many controls, finding the optimal server farm management policy is an almost intractable problem – how many servers should be on at any given time, how many should be off, and how many should be in some sleep state?

In this paper, we employ the popular metric of Energy-Response time Product (ERP) to capture the energy-performance tradeoff, and present the first theoretical results on the optimality of server farm management policies. For a stationary demand pattern, we prove that there exists a very small, natural class of policies that always contains the optimal policy for a single server, and conjecture it to contain a near-optimal policy for multi-server systems. For time-varying demand patterns, we propose a simple, traffic-oblivious policy and provide analytical and empirical evidence for its near-optimality.

More than Skin Deep: Measuring Effects of the Underlying Model on Access-Control System Usability

Reeder, Bauer, L. Cranor, Reiter & Vaniea

CHI 2011, May 7-12, 2011, Vancouver, BC, Canada.

In access-control systems, policy rules conflict when they prescribe different decisions (ALLOW or DENY) for the same access. We present the results of a user study that demonstrates the significant impact of conflict-resolution method on policy-authoring usability.

continued on page 26

continued from page 25

In our study of 54 participants, varying the conflict-resolution method yielded statistically significant differences in accuracy in five of the six tasks we tested, including differences in accuracy rates of up to 78%. Our results suggest that a conflict-resolution method favoring rules of smaller scope over rules of larger scope is more usable than the Microsoft Windows operating system's method of favoring deny rules over allow rules. Perhaps more importantly, our results demonstrate that even seemingly small changes to a system's semantics can fundamentally affect the system's usability in ways that are beyond the power of user interfaces to correct.

Scale and Concurrency in GIGA+: File System Directories with Millions of Files

Patil & Gibson

Proceedings of the 9th USENIX Conference on File and Storage Technologies (FAST '11), San Jose CA, February 2011.

We examine the problem of scalable file system directories, motivated by data-intensive applications requiring millions to billions of small files to be ingested in a single directory at rates of hundreds of thousands of file creates every second. We introduce a POSIX-compliant scalable directory design, GIGA+, that distributes directory entries over a cluster of server nodes that make only local, independent decisions about migration. GIGA+ uses two tenets, asynchrony and inconsistency, to: (1) partition the index among all servers without any synchronization or serialization, and (2) minimize stale and inconsistent mapping state at the clients. Applications are provided traditional strong data consistency semantics, and cluster growth requires minimal directory entry migration. We have built and demonstrated that the GIGA+ approach scales better than existing distributed directory implementations, delivers a sustained throughput of more than 98,000 file creates per second on a 32-server cluster, and balances load more efficiently than consistent hashing.

pWalrus: Towards Better Integration of Parallel File Systems into Cloud Storage

Abe & Gibson

Workshop on Interfaces and Abstractions for Scientific Data Storage (IASDS10), Heraklion, Greece, September 2010.

Amazon S3-style storage is an attractive option for clouds that provides data access over HTTP/HTTPS. At the same time, parallel file systems are an essential component in privately owned clusters that enable highly scalable data intensive computing. In this work, we take advantage of both of those storage options, and propose pWalrus, a storage service layer that integrates parallel file systems effectively into cloud storage. Essentially, it exposes the mapping between S3 objects and backing files stored in an underlying parallel file system, and allows users to selectively use the S3 interface and direct access to the files. We describe the architecture of pWalrus, and present preliminary results showing its potential to exploit the performance and scalability of parallel file systems.

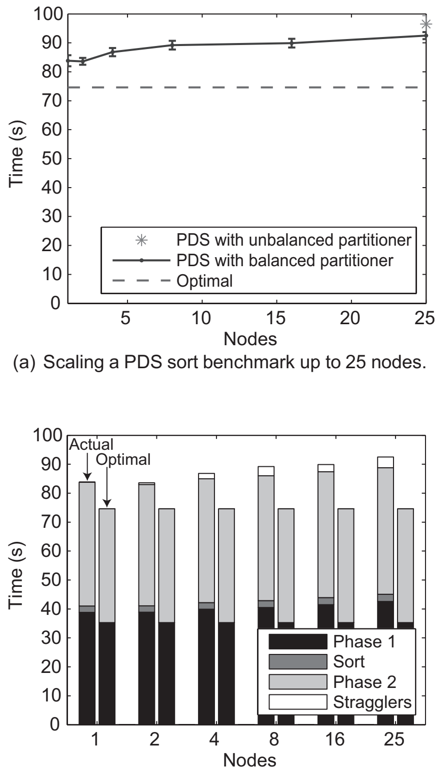
Applying Performance Models to Understand Data-intensive Computing Efficiency

Krevat, Shiran, Anderson, Tucek, J. Wylie & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-108. May 2010.

New programming frameworks for scale-out parallel analysis, such as MapReduce and Hadoop, have become a cornerstone for exploiting large datasets. However, there has been little analysis of how these systems perform relative to the capabilities of the hardware on which they run. This paper describes a simple analytical model that predicts the optimal performance of a parallel dataflow system. The model exposes the inefficiency of popular scale-out systems, which take 3—13× longer to complete jobs

continued on page 27



(a) Scaling a PDS sort benchmark up to 25 nodes.

(b) Time breakdown.

Using Parallel DataSeries to sort up to 100 GB, it is possible to approach within 12-24% of the optimal sort times as predicted by our performance model. PDS scales well for an in-memory sort with 4 GB per node up to 25 nodes in (a), although there is a small time increase starting around 4 nodes due to network effects. Also shown for the 25 node case is the performance of our older, unbalanced partitioner, which had an additional 6% performance overhead from optimal. A breakdown of time in (b) shows that the time increases at scale are mostly in the first phase of a map-reduce dataflow, which includes the network data shuffle, and in the time nodes spend waiting for stragglers due to effects of skew.

continued from page 26

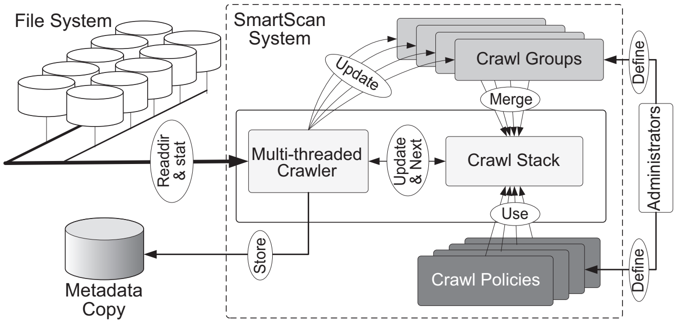
than the hardware should allow, even in well-tuned systems used to achieve record-breaking benchmark results. To validate the sanity of our model, we present small-scale experiments with Hadoop and a simplified dataflow processing tool called Parallel DataSeries. Parallel DataSeries achieves performance close to the analytic optimal, showing that the model is realistic and that large improvements in the efficiency of parallel analytics are possible.

SmartScan: Efficient Metadata Crawl for Storage Management Metadata Querying in Large File Systems

Liu, Xu, Wu, Yang & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-112, October 2010.

SmartScan is a metadata crawl tool that exploits patterns in metadata changes to significantly improve the efficiency of support for file-system-wide metadata querying, an important tool for administrators. Usually, support for metadata queries is provided by databases populated and refreshed by calling stat() on every file in the file system. For large file systems, where such storage management tools are most needed, it can take many hours to complete each scan, even if only a small percentage of the files have changed. To address this issue, we identify patterns in metadata changes that can be exploited to restrict scanning to the small subsets of directories that have recently had modified files or that have high variation in file change times. Experiments with using SmartScan on production file systems show that exploiting metadata change patterns can reduce the time needed to refresh the metadata database by one or two orders with minimal loss of freshness.



Architecture of SmartScan. SmartScan uses only the standard file system interface to gather required information, and thus doesn't require any modification to the file system itself.

Principles of Operation for Shingled Disk Devices

Gibson & Ganger

Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-11-107, April 2011.

A leading strategy for driving the areal density of magnetic disk drives through 1 – 10 terabit/inch^2 (the coming decade) is to shingle (partially overlap) adjacent tracks, imposing significant restrictions on where data can be written without incurring multi-track read-modify-write penalties. These restrictions and penalties can be 1) fully hidden from system software using techniques familiar in NAND Flash disks; 2) minimally exposed as multi-track, shingled bands of predetermined size that can be read normally, but only appended to or trimmed (erased); or 3) maximally exposed as dynamically sized bands of shingles separated by guard regions of previously erased tracks, allowing maximal capacity to be obtained by the most sophisticated system software. While the latter options require significant changes in system software, there is a rich history of demonstrations of log-structured file systems that should be able to do this, and a profusion of write-once cloud storage systems that could provide the economic "killer application".

Now is a very good time for systems software experts to take interest and weigh in as magnetic disk technologists are experimenting and prototyping shingled disks. Experience shows that changes in the interface model for magnetic disks can take decades to change (for example, 512B to 4096B sectors) unless device vendors and systems software developers work together toward a mutually desired principles of operation.

Diagnosing Performance Changes by Comparing System Behaviours

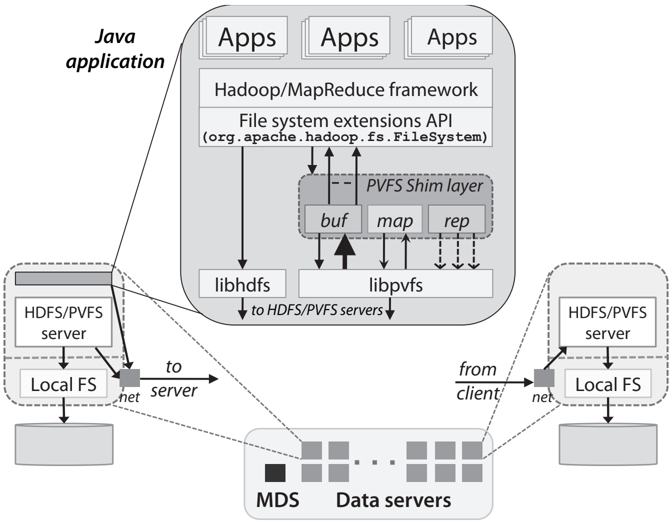
Sambasivan, Zheng, Krevat, Whitman, Stroucken, Wang, Xu & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-107. July 2010.

The causes of performance changes in a distributed system often elude even its developers. We develop a new technique for gaining insight into such changes: comparing system behaviours from two executions (e.g., of two system versions or time periods). Building on end-to-end request flow tracing within and across components, algorithms are described for identifying and ranking changes in the flow and/or timing of request processing. The implementation of these algorithms in a tool called Spectroscope is described and evaluated. Five case studies are presented of using Spectroscope to diagnose performance changes in a distributed storage system caused by code changes and configuration modifications, demonstrating the value and efficacy of comparing system behaviours.

continued on page 28

continued from page 27



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

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

Tantisiriroj, Patil, Gibson, Son, Lang & Ross

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-11-108, April 2011.

Data-intensive applications fall into two computing styles: Internet services (cloud computing) or high-performance computing (HPC). In both categories, the underlying file system is a key component in scalable application performance. In this paper we explore the similarities and differences between PVFS, a parallel file system used in HPC at large scale, and HDFS, the primary storage system used in cloud computing with Hadoop.

We integrate PVFS into Hadoop and evaluate performance versus HDFS for a set of data-intensive computing benchmarks. We study how HDFS-specific optimizations can be matched using PVFS and how consistency, durability, and persistence tradeoffs made by these file systems affect application performance. We show how to embed multiple replicas into a PVFS file, including a mapping with a complete copy local to the writing client, to emulate HDFS's file layout policies. We also highlight implementation issues with HDFS's dependence on disk bandwidth and benefits from pipelined replication.

DiscFinder: A Data-Intensive Scalable Cluster Finder for Astrophysics

Fu, López, Fink, Ren & Gibson

Proceedings of the ACM International Symposium on High Performance Distributed Computing (HPDC), Chicago, IL. June, 2010.

DiscFinder is a scalable, distributed, data-intensive group finder for analyzing observation and simulation astrophysics datasets. Group finding is a form of clustering used in astrophysics for identifying large-scale structures such as galaxies and clusters of galaxies. DiscFinder runs on commodity compute clusters and scales to large datasets with billions of particles. It is designed to operate on datasets that are much larger than the aggregate memory available in the computers where it executes. As a proof-of-concept we have implemented DiscFinder as an application on top of the Hadoop framework. DiscFinder has been used to cluster the largest open-science cosmology simulation datasets containing as many as 14.7 billion particles. We evaluate its performance and scaling properties and describe the performed optimization.

PDL Workshop and Retreat 2010.