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
Seagate Technology
Symantec Corporation
VMware, Inc.
Yahoo! Labs

PDL Packet

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • SPRING 2010
http://www.pdl.cmu.edu/

FAWN: A Fast Array of Wimpy Nodes

David Andersen & Joan Digney

Large-scale data-intensive applications, such as high-performance key-value storage systems, are growing in both size and importance; they now are critical parts of major Internet services such as Amazon, LinkedIn, and Facebook. The workloads these systems support are I/O intensive, massively parallel, require large clusters to support them, and the size of objects stored is typically small.

The clusters that serve these workloads must provide both high performance and low cost operation. Unfortunately, small-object random-access workloads are particularly ill-served by conventional disk-based or memory-based clusters. We began the FAWN project asking how we could overcome problems such as the poor seek performance of disks and the huge energy draw of DRAM-based clusters, where power costs may comprise up to half of the three-year total cost of owning a computer. Could we build a cost-effective cluster for data-intensive workloads that uses a fraction of the power required by a conventional architecture, but that still meets the same capacity, availability, throughput, and latency requirements?

The FAWN architecture—a Fast Array of Wimpy Nodes—couples low-power, efficient embedded CPUs with flash storage to provide efficient, fast, and cost-effective access to large, random-access data. FAWN creates a well matched system architecture around flash, which is faster than disk, cheaper than DRAM, and consumes less power than either. Each node can use the full capacity of the flash without memory or bus bottlenecks, without wasted excess capability.

To efficiently run data-intensive applications, FAWN uses "wimpy" processors selected to reduce I/O-induced idle cycles while maintaining high performance. Because CPU power consumption grows super-linearly with speed, a FAWN clus-

continued on page 16

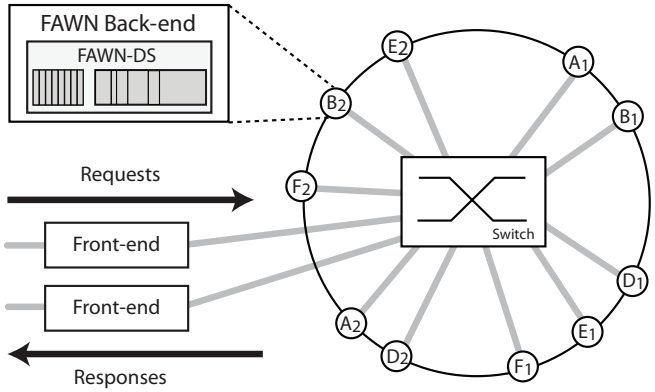
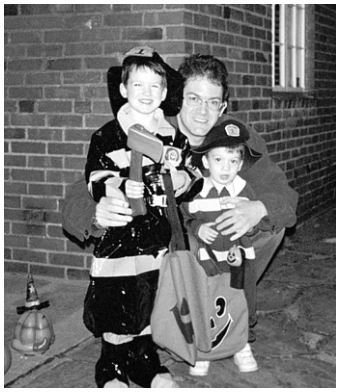


Figure 1: FAWN-KV Architecture.



FROM THE DIRECTOR'S CHAIR

Greg Ganger

Hello from fabulous Pittsburgh!

It has been another great year for the Parallel Data Lab, my tenth as PDL Director. Some highlights include deployment of two cloud computing testbeds, awards for several researchers and papers, and many students graduating and joining PDL companies. Along the way, excellent progress has been made on continuing PDL projects, and of course many papers have been published describing research results. Let me highlight a few things and briefly wax nostalgic about having been PDL Director for 10 years.

The FAWN (Fast Array of Wimpy Nodes) project has really blossomed over the last year. Led by Prof. David Andersen, this 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. Several prototypes have been built and demonstrated, and a paper describing the FAWN key-value store was named Best Paper at SOSP 2009. An article in this newsletter describes FAWN in a bit more detail.

PDL's foray into data-intensive computing (DISC) has many other components, as well. We are also exploring the efficiency of popular DISC frameworks, such as Google's MapReduce and the open source Hadoop system. Interestingly, even for highly tuned benchmark results reported by Google and Yahoo!, these frameworks are far less efficient than could be hoped—they generally utilize 3–8 times more resources (computers) than required to achieve their given job completion times, even ignoring the FAWN-suggested gains from using more efficient hardware. We are exploring more efficient approaches and also data distribution algorithms for allowing power-proportional scaling of the cluster size dedicated to particular data-intensive computations. In addition to efficiency, we are exploring higher-level frameworks and mechanisms for extremely large-scale metadata services, such as cloud databases and huge GIGA+ directories.

Greatly complementing our experimental explorations of data-intensive computing are our deployment-based explorations of cloud computing. Over the last year, we have deployed two clusters in the Data Center Observatory (DCO) for use as cloud computing infrastructures, as parts of the broad OpenCloud and OpenCirrus open cloud computing testbeds. One is set up as a Hadoop service used by various scientists that mine large quantities of data, and the other 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 being heavily instrumented to provide us with deep insight into the usage patterns and efficiencies of such clouds. Both are also being used as test environments for improved tools and algorithms for managing and using cloud computing infrastructures.

PDL's Perspective system for distributed home/consumer storage has been deployed in multiple student homes as well as in a lounge and two offices at CMU. The research has turned toward understanding and providing for the complex access control requirements when storage becomes easily shared across home/personal devices. Several user studies have been conducted to better understand

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.

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
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 Stroucken
Spencer Whitman
Charlene Zang

VISITING RESEARCHER

Likun Liu

GRADUATE STUDENTS

Yoshihisa Abe
Robson Cordeiro
Jim Cipar
Debabrata Dash
Tudor Dumitraş
Bin Fu
Anshul Gandhi
Varun Gupta
Wesley Jin
Mike Kasick
Soila Kavulya
Elie Krevat
Karan Kumar
Patrick Lanigan
Yuan Liang
Hyeontaek Lim
Michelle Mazurek
Nathan Mickulicz
Iulian Moraru
Ippokratis Pandis
Luca Parolini
Swapnil Patil
Adam Pennington
Amar Phanishayee
Milo Polte
Kai Ren
Wolfgang Richter
Raja Sambasivan
Koushik Sampath
Rich Shay
Jiri Simsa
Shafeeq Sinnamohideen
Wittawat Tantisiriroj
Vijay Vasudevan
Pedro Vaz de Melo
Gaurav Veda
Matthew Wachs
Lin Xiao
Lianghong Xu

FROM THE DIRECTOR'S CHAIR

users' needs and current practices, including some described in recent CHI and SOUPS papers. Based on those studies, a novel policy-based access control mechanism is being developed for this new domain of distributed storage.

Our explorations of automation for large-scale storage also continue, both in the Self-* Storage project and in efforts inspired by it (e.g., the cloud computing deployments discussed above). Much of the ongoing effort focuses on the very difficult challenges involved with automating aspects of problem diagnosis. It is clear that there will be no silver bullet here, and PDL research is probing a number of complementary paths. For example, one approach being explored is comparison of resource utilization statistics across servers that should be receiving roughly equal workloads—divergence of one from the crowd implicates it as a likely source of performance problems. Another approach being explored is 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. These and other machine learning tools are being applied to significantly decrease the lack of guidance facing humans seeking to diagnose problems.

Many other ongoing PDL projects are also producing cool results, especially for creation of scalable and fault-tolerant cluster-based storage. For example, our DiskReduce approach to providing space-efficient redundancy for DISC storage (which originally supported only replication) has now been adopted by and integrated into the open source Hadoop file system. Our work on making it easy to scale metadata services to multiple servers without complex consistency protocols is ready for prime-time. We have developed a new protocol (called Zzyzx) that provides unprecedented efficiency and scalability for Byzantine fault-tolerant services, providing a scheme for metadata to complement the fault-tolerant storage scheme we recently developed. This newsletter and the PDL website offer more details and additional research highlights.

I was shocked, while working on this PDL Director's letter, to realize that this is my tenth letter and thus marked my tenth year. I'm always overwhelmed by the accomplishments of the PDL students and staff, and it's a pleasure to work with them. Thinking about the many dozens of students and staff who have made PDL such a special place, including 16 completed PhDs that I have advised, I choke up a bit. Their impact and technical contributions are too numerous to list, and both they and their work continue to lead as many of them are now major players at PDL companies. Ten years later, and it is still a joy and an honor to be a part of PDL.

Greg discusses storage with Joe Tucek of HP and Sarun Savetsila, a CMU student at the PDL Retreat.

Ippokratis Pandis explains his research on Speculative Lock Inheritance to Wei Hu of Oracle at the PDL Retreat.

YEAR IN REVIEW

May 2010

- 12th Annual PDL Spring Industry Visit Day.
- Swapnil Patil proposed his Ph.D research titled "Scalability, Usability and Applicability of Massive File System Directories."
- Michelle Mazurek will be interning under Eno Thereska at MSR-Cambridge this summer.
- Mike Kasick will be interning with IBM Almaden this summer.
- Raja Sambasivan is interning at Google Pittsburgh this summer, continuing his work on performance problem diagnosis using end-to-end traces.
- Elie Krevat and Jim Cipar will be interning with HP Labs.
- Lin Xiao will be interning with Google.
- Kai Ren will be interning at Facebook from May through August, joining their data infrastructure team to do projects related to Hive or Hadoop.
- Wittawat Tantisiriroj is interning with Yahoo! in Sunnyvale, CA.
- Soila Kavulya presented "An Analysis of Traces from a Production MapReduce Cluster" at the 10th Symposium on Cluster, Cloud and Grid Computing (CCGrid 2010) in Melbourne, Australia.
- Garth Gibson presented "Managing the Coming Data Deluge; File Systems Panel" at The Future of Large Scale Computing Symposium in Stanford, CA.

April 2010

- Matthew Wachs proposed his Ph.D. research on "Improving Bandwidth Guarantees for Storage Workloads with Performance Insulation."
- Jiaqi Tan presented "Kahuna: Problem Diagnosis for MapReduce-based Cloud Computing Environments" at NOMS 2010 in Osaka, Japan.
- Michelle Mazurek presented "Access Control for Home Data Sharing: Attitudes, Needs and Practices" at the 2010 Conference on Human Factors in Computing Systems, in Atlanta, GA.

March 2010

- Garth Gibson discussed "Developing Systems for Scale: Experience with Faster-than-Moore's-Law HPC Storage Systems Growth" at the Exascale Evaluation and Research Techniques Workshop (EXERT10) in Pittsburgh, PA.
- Garth Gibson presented "Extreme Scale IO: On the Road to Exascale" and "Panasas @ Petascale" at the NSF Extreme Scale IO Workshop in Austin TX.

February 2010

- Mike Kasick spoke at FAST 2010 in San Jose, CA, presenting "Black-Box Problem Diagnosis in Parallel File Systems."
- Lorrie Cranor testified at a Congressional hearing on Feb 24 on the Collection and Use of Location Information for Commercial Purposes.

January 2010

- Michelle Mazurek received an NSF IGERT fellowship through the Carnegie Mellon Usable Privacy and Security (CUPS) Lab.
- Greg Ganger talked about "Open Cirrus at Carnegie Mellon University" and Garth Gibson discussed "DiskReduce v2.0 for HDFS" at the OpenCirrus Summit in Sunnyvale, CA.

December 2009

- U Kang and Babis Tsourakakis received the Best Applications Paper Award (runner up), at ICDM'09 in Miami, FL, for their paper "PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations."
- Kee-Tee (Lawrence) Tan gave his Master's Thesis presentation on "Joulesort on a Low-power CPU-GPU Hybrid Architecture."
- Lei Li proposed his Ph.D. thesis research titled "Fast Algorithms for Time Series Mining."

November 2009

- 17th Annual Parallel Data Lab Workshop & Retreat
- Milo Polte presented "…And Eat It Too: High Read Performance in Write-Optimized HPC I/O Middleware File Formats" at the 4th SC Petascale Data Storage Workshop held in Portland, OR.
- Wittawat Tantisiriroj presented "DiskReduce: RAID for Data-Intensive Scalable Computing" at the 4th

continued on page 19

Greg valiantly took several cream pies in the face, including a few from his son William, on CMU's annual Pi-A-Professor Day, which is sponsored by CMU's student branch of the National Society of Black Engineers.

Visual, Log-based Causal Tracing for Performance Debugging of MapReduce Systems

Tan, Kavulya, Gandhi & Narasimhan

30th IEEE International Conference on Distributed Computing Systems (ICDCS) 2010, Genoa, Italy, June 2010.

The distributed nature and large scale of MapReduce programs and systems poses two challenges in using existing profiling and debugging tools to understand MapReduce programs. Existing tools produce too much information because of the large scale of MapReduce programs, and they do not expose program behaviors in terms of Maps and Reduces. We have developed a novel non-intrusive log-analysis technique which extracts state-machine views of the control- and dataflows in MapReduce behavior from the native logs of Hadoop MapReduce systems, and it synthesizes these views to create a unified, causal view of MapReduce program behavior. This technique enables us to visualize MapReduce programs in terms of MapReduce-specific behaviors, aiding operators in reasoning about and debugging performance problems in MapReduce systems. We validate our technique and visualizations using a real-world workload, showing how to understand the structure and performance behavior of MapReduce jobs, and diagnose injected performance problems reproduced from real-world problems.

BEMC: A Searchable, Compressed Representation for Large Seismic Wavefields

López, Ramírez-Guzmán, Bielak & O'Hallaron

22nd Int. Conf on Scientific and Statistical Database Management (SSDBM'10), Heidelberg, Germany, June 30 - July 2, 2010.

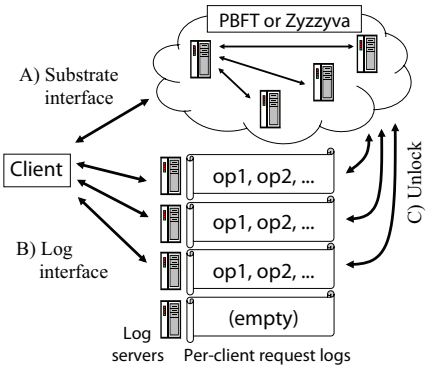
State-of-the-art numerical solvers in Earth Sciences produce multi terabyte datasets per execution. Operating on increasingly larger datasets becomes challenging due to insufficient data bandwidth. Queries result in difficult to handle I/O access patterns. BEMC is a new mechanism that allows querying and processing wavefields in the compressed representation. This approach combines well-known spatial-indexing techniques with novel compressed representations, thus reducing I/O bandwidth requirements. A new compression approach based on boundary integral representations exploits properties of the simulated domain. Frequency domain representation further compresses the data by eliminating temporal redundancy found in wave propagation data. This representation enables the transformation of a large I/O workload into a massively-parallel CPU-intensive computation. Queries to this representation result in largely sequential I/O accesses. Although, decompression places heavy demands on the CPU, it exhibits parallelism well-suited for many-core processors. We evaluate our approach in the context of data analysis for the Earth Sciences datasets.

Zzyzx: Scalable Fault Tolerance Through Byzantine Locking

Hendricks, Sinnamohideen, Ganger & Reiter

Proceedings of the 40th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. Chicago, Illinois, June 2010.

Zzyzx is a Byzantine fault-tolerant replicated state machine protocol that outperforms prior approaches and provides near-linear throughput scaling. Using a new technique called Byzantine Locking, Zzyzx allows a client to extract state from an underlying replicated state machine and access it via a second protocol specialized for use by a single client. This second protocol requires just one roundtrip and 2 f+1 responsive servers—compared to Zyzzyva, this results



Zzyzx components. The execution of Zzyzx can be divided into three subprotocols: A) If a client has not locked the objects needed for an operation, the client uses a substrate protocol such as PBFT or Zyzzyva. B) If a client holds locks for all objects touched by an operation, the client uses the log interface. C) If a client tries to access an object for which another client holds a lock, the unlock subprotocol is run.

in 39–43% lower response times and a factor of 2.2–2.9× higher throughput. Furthermore, the extracted state can be transferred to other servers, allowing non-overlapping sets of servers to manage different state. Thus, Zzyzx allows throughput to be scaled by adding servers when concurrent data sharing is not common. When data sharing is common, performance can match that of the underlying replicated state machine protocol.

DiscFinder: A Data-intensive Scalable Cluster Finder for Astrophysics

Fu, Ren, López, Fink & Gibson

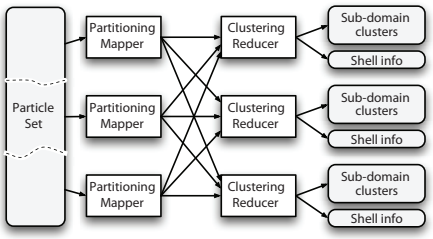
Proceedings of the ACM International Symposium on High Performance Distributed Computing (HPDC), Chicago, IL. June, 2010.

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 galax-

continued on page 6

RECENT PUBLICATIONS

continued from page 5



DiscFinder Partition and Clustering stages. This is the central MapReduce job in the DiscFinder pipeline.

ies. DiscFinder runs on commodity compute clusters and scales to large datasets with billions of particles. It is designed to operate on databases 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.

A Transparently-Scalable Metadata Service for the Ursa Minor Storage System

Sinnamohideen, Sambasivan, Hendricks, Liu & Ganger

Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-10-102. March 2010.

This technical report describes the design and implementation of the Ursa Minor Metadata Service. Like many prior direct-access file systems, Ursa Minor provides for scalable data access — adding storage servers provides a proportional increase in data throughput. Unlike most previous systems, it also provides for scaling metadata throughput by adding metadata servers. Scaling metadata is more challenging than scaling data because, unlike data operations, a single metadata operation may involve items served by different metadata servers. Existing systems that handle such operations do so using relatively complex distributed transaction protocols. Ursa Minor takes a novel approach by reusing metadata migration, an existing feature normally used to support load balancing, to implement multi-server operations. Additionally, Ursa Minor uses an object-ID assignment scheme that minimizes the occurrence of multi-server operations. The combination of these approaches allows us to implement a desired feature with less complexity than alternative methods and with minimal performance penalty (within 1% of optimal in common cases).

An Analysis of Traces from a Production MapReduce Cluster

Kavulya, Tan, Gandhi & Narasimhan

10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2010). May 17-20, 2010, Melbourne, Victoria, Australia.

MapReduce is a programming paradigm for parallel processing that is increasingly being used for data-intensive applications in cloud computing environments. An understanding of the characteristics of workloads running in MapReduce environments benefits both the service providers in the cloud and users: the service provider can use this knowledge to make better scheduling decisions, while the user can learn what aspects of their jobs impact performance. This paper analyzes 10 months of MapReduce logs from the M45 supercomputing cluster which Yahoo! made freely available to select universities for systems research. We characterized resource utilization patterns, job patterns, and sources of failures. We use an instance-based learning technique that exploits temporal locality to predict job completion times from historical data and identify potential performance problems in our dataset.

Diagnosing Performance Problems by Visualizing and Comparing System Behaviours

Sambasivan, Zheng, Krevat, Whitman, Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-103, February 2010.

Spectroscope is a new toolset aimed at assisting developers with the long-standing challenge of performance debugging in distributed systems. To do so, it mines end-to-end traces of request processing within and across components. Using Spectroscope, developers can visualize and compare system behaviours between two periods or system versions, identifying and ranking various changes in the flow or timing of request processing. Examples of how Spectroscope has been used to diagnose real performance problems seen in a distributed storage system are presented, and Spectroscope's primary assumptions and algorithms are evaluated.

Robust and Flexible Power-proportional Storage

Amur, Cipar, Gupta, Ganger, Kozuch & Schwan

ACM Symposium on Cloud Computing (SOCC). June 10-11, 2010, Indianapolis, IN.

Power-proportional cluster-based storage is an important component

continued on page 7

For our 2009 PDL Workshop and Retreat we moved to the beautiful, historic Bedford Springs Resort in Bedford Springs, PA.

continued from page 6

of an overall cloud computing infrastructure. With it, substantial subsets of nodes in the storage cluster can be turned off to save power during periods of low utilization. Rabbit is a distributed file system that arranges its data-layout to provide ideal power-proportionality down to very low minimum number of powered-up nodes (enough to store a primary replica of available datasets). Rabbit addresses the node failure rates of large-scale clusters with data layouts that minimize the number of nodes that must be powered-up if a primary fails. Rabbit also allows different datasets to use different subsets of nodes as a building block for interference avoidance when the infrastructure is shared by multiple tenants. Experiments with a Rabbit prototype demonstrate its power-proportionality, and simulation experiments demonstrate its properties at scale.

Kahuna: Problem Diagnosis for MapReduce-Based Cloud Computing Environments

Tan, Pan, Marinelli, Kavulya, Gandhi & Narasimhan

Proceedings of the 12th IEEE/IFIP Network Operations and Management Symposium (NOMS) 2010, Osaka, Japan, Apr 2010.

We present Kahuna, an approach that aims to diagnose performance problems in MapReduce systems. Central to Kahuna's approach is our insight on peer-similarity, that nodes behave alike in the absence of performance problems, and that a node that behaves differently is the likely culprit of a performance problem. We present applications of Kahuna's insight in techniques and their algorithms to statistically compare blackbox (OS-level performance metrics) and white-box (Hadoop log statistics) data across the different nodes of a MapReduce cluster, in order to identify the faulty node(s). We also present empirical evidence of our peer-similarity observations from the 4000-processor Yahoo! M45 Hadoop cluster. In addition, we demonstrate Kahuna's effectiveness through experimental evaluation of two algorithms for a number of reported performance problems, on four different workloads in a 100-node Hadoop cluster running on Amazon's EC2 infrastructure.

Standardizing Privacy Notices: An Online Study of the Nutrition Label Approach

Kelley, Cesca, Bresee & L.F. Cranor

28th ACM Conference on Human Factors in Computing Systems (CHI2010). Atlanta, GA, April 2010.

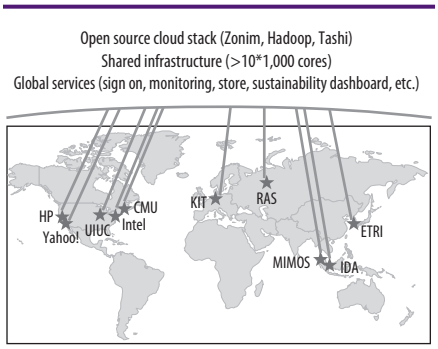
Earlier work has shown that consumers cannot effectively find information in privacy policies and that they do not enjoy using them. In our previous research on nutrition labeling and other similar consumer information design processes we developed a standardized table format for privacy policies. We compared this standardized format, and two short variants (one tabular, one text) with the current status quo: full text natural language policies and layered policies. We conducted an online user study of 789 participants to test if these three more intentionally designed, standardized privacy policy formats, assisted by consumer education, can benefit consumers. Our results show that providing standardized privacy policy presentations can have significant positive effects on accuracy of information finding, overall speed, and reader enjoyment with privacy policies.

Open Cirrus: A Global Cloud Computing Testbed

Avetisyan, Campbell, Gupta, Heath, Ko, Ganger, Kozuch, O'Hallaron, Kunze, Kwen, Lai, Lyons, Milojicic, Lee, Soh, Ming, Lake & Namgoong

IEEE Computer, April 2010.

Open Cirrus is a cloud computing testbed that, unlike existing alternatives,



Open Cirrus testbed. Each of the 10 current sites consists of a cluster with at least 1,000 cores and associated storage. The testbed offers a cloud stack consisting of physical and virtual machines and global services such as sign-on, monitoring, storage, and job submission.

federates distributed data centers. It aims to spur innovation in systems and applications research and catalyze development of an open source service stack for the cloud.

Access Control for Home Data Sharing: Attitudes, Needs and Practices

Mazurek, Arsenault, Bresee, Gupta, Ion, Johns, Lee, Liang, Olsen, Salmon, Shay, Vaniea, Bauer, L.F. Cranor, Ganger & Reiter

28th ACM Conference on Human Factors in Computing Systems (CHI2010). Atlanta, GA, April 2010.

As digital content becomes more prevalent in the home, nontechnical users are increasingly interested in sharing that content with others and accessing it from multiple devices. Not much is known about how these users think about controlling access to this data. To better understand this, we conducted semi-structured, in-situ interviews with 33 users in 15 households. We found that users create ad-hoc access-control mechanisms that do not always work; that their ideal policies are complex and multi-dimensional; that a priori policy specification is often insufficient; and

continued on page20

May 2010

Bruno Sinopoli Receives NSF Career Award

Carnegie Mellon University's Bruno Sinopoli has received a five-year, $400,000 grant from the National Science Foundation to develop computer tools for securing and controlling cyber-physical systems.

"I am honored to receive this award which will help me continue investigating tools and methodologies to design and analyze cyber-physical and networked embedded systems," said Sinopoli, an assistant professor of electrical and computer engineering and a researcher at Carnegie Mellon CyLab.

Sinopoli said his goal is to set new standards for the robustness and security of critical infrastructures, such as power, gas and water distribution networks, transportation systems and other physical structures. "While critical infrastructure can greatly benefit from the extensive use of information and communication technologies to improve safety and performance indices, their integration raises issues of reliability and security. In this project I want to address these concerns."

-- from CMU Press Release May 11, 2010

May 2010

Vijay Vasudevan Wins Yahoo! Research Award

Vijay Vasudevan is among three Carnegie Mellon Ph.D. students who have been selected by Yahoo! as winners in the 2010 Key Scientific Challenges program for their research proposals on the future of the Internet. Vijay submitted a successful proposal for Green Computing. The winners receive $5,000 each in unrestricted seed funding for their research and will also have the opportunity to work closely with some of the world's most well-known and experienced Internet scientists at Yahoo! Labs to advance their research over the next several months. In September he will have the opportunity to present and defend his findings to peers and Yahoo! Labs leaders in a structured workshop. As was the case last year when Yahoo! inaugurated the program, no other university had as many winners in the program as Carnegie Mellon.

-- from CMU Press Release, May 5, 2010.

April 2010

Congratulations Andy and Nikki

Dr. Andrew J. Klosterman and Mrs. Nikki L. Klosterman were wed at the Schenley Park Cafe & Visitors Center on April 17, 2010 in sight of the location of their first "real" date when Nikki met Andy sitting in a hammock at the top of Flagstaff Hill.

Nikki is a Child Life Assistant, spending her time seeing to the needs of oncology patients at Children's Hospital of Pittsburgh of UPMC. Andy is working for Avere Systems, Inc. in Pittsburgh, which is building performance acceleration appliances for network attached storage. They are purchasing a home in Ben Avon, PA, just down the Ohio River from downtown.

March 2010

Brandon Salmon and Mary Moran Wed

Brandon and Mary were wed on Saturday, March 27, 2010 at the Washington, D.C. LDS Temple. A reception followed at the Fraser Gallery where they held an Iron Chef Competition, inviting their guests to contribute an appetizer using the color orange as the secret ingredient.

They are currently living in San Francisco, where Brandon works for Tintri Systems. Best wishes to the happy couple!

February 2010

Lorrie Cranor Addresses Congressional Subcommittees About Privacy Issues and Location-Based Services

CMU's Lorrie F. Cranor discussed the risks and benefits of online services that collect and use location information at joint meetings of the U.S. Congressional Subcommittee on Commerce, Trade and Consumer Protection and the Subcommittee on Communication and Technology Wednesday, Feb. 24, 2010 in Washington, D.C.

Increasingly popular location-based services allow Internet users to share their location with friends, track employees or children, or receive information based on current geographic location. GPS and other technology built into cell phones and laptop computers allows people to be located automatically, often to within a few hundred feet. However, there is growing concern about the invasive nature of this technology, according to

continued on page 9

continued from page 8

Cranor, an associate professor of computer science and engineering and public policy at Carnegie Mellon.

"Due to the way cellular technology works, for example, the widespread use of cell phones enables round-the-clock surveillance of citizens. It is important that the storage of individual location data be minimized and protections be put in place to limit when it can be disclosed to the government," said Cranor, who has conducted several studies about privacy issues and location-sharing technologies.

Another cause for concern is the lack of accessibility to privacy controls on a variety of location-sharing applications. During a recent evaluation of 84 location-sharing applications, Cranor's team found that "the majority of those privacy controls are not easily accessible from the main page or home page of the application itself."

"Only 18 of the 84 services we reviewed this month mentioned privacy controls or security on the front page of their Web site," Cranor said. "In most cases, it is almost impossible to find out what a service is going to do with your location information without signing up for the service and trying it out."

In addition, Cranor's team found many location-based services had no privacy policies posted on their Web sites, and those that did post policies often made no mention of location information. A report on the Carnegie Mellon location sharing study is available online.

-- from Carnegie Mellon University Press Release Feb. 23, 2010

February 2010

Carnegie Mellon Joins Open Cirrus Test Bed For Advancing Cloud Computing Research

Carnegie Mellon University's School of Computer Science is the latest research institution to host a site as part of Open Cirrus(tm), a global, open-source test bed for the advancement of cloud computing research and education. The computing cluster, housed in Carnegie Mellon's Data Center Observatory, will provide resources for Carnegie Mellon faculty and other researchers worldwide. Open Cirrus was launched in 2008 by HP, Intel and Yahoo! to promote open collaboration among industry, academia and governments on data-intensive computing.

"Having a facility like this and being able to participate in Open Cirrus will provide us with unprecedented opportunities for research and education on Internet-scale computing," said Randal E. Bryant, dean of the School of Computer Science. "We see applications well beyond those being pursued by industry today, including astronomy, neuroscience, and knowledge extraction and representation, and we will be able to delve more deeply into the design of the system itself."

Greg Ganger, professor of electrical and computer engineering and director of Carnegie Mellon's Parallel Data Lab, said the new computing cluster, which has 159 servers and 1,165 processing cores, was made possible by Intel's generous donation of CPUs and money. The cluster has 2.4 trillion bytes, or terabytes, of memory and almost 900 terabytes of storage. A contribution by APC of power management and cooling systems also was crucial for building and operating the cluster. Like other sites in Open Cirrus, the computing cluster will be made available to researchers worldwide later this year.

Ganger said much of the research at the Carnegie Mellon site likely will focus on the university's strengths — how to make the cloud computing infrastructure faster, more reliable and more energy efficient and how to use the cloud in innovative ways for new applications. "This site embodies our commitment to the collaborative, open-source research environment that Open Cirrus promotes and to aggressively pursuing cloud computing research on this campus," he said.

-- from Carnegie Mellon University Press Release Feb. 15, 2010

February 2010

Priya Narasimhan Receives Benjamin Richard Teare Teaching Award

Congratulations to Priya Narasimhan, associate professor of Electrical and Computer Engineering, who has received the Benjamin Richard Teare Teaching Award. This award is made to a faculty member within the Carnegie Institute of Technology in recognition of excellence in engineering education. The basis for selection is excellence in engineering education in the areas of teaching and/or educational innovation and educational leadership.

Priya is the award's 2009-2010 recipient in recognition of her efforts in transforming the undergraduate Embedded Systems capstone design course in Electrical and Computer Engineering, and for her passion, dedication, and high performance in teaching. In addition, she has introduced many community-based projects, such as development of assistive technologies for the visually impaired, which have provided great motivation for the students and have raised student accomplishments to very high levels.

January 2010

Ganger Next Holder of the Stephen J. Jatras Professorship in Electrical and Computer Engineering

PDL Director and Professor of ECE, Greg Ganger will be the next recipient of the Stephen J. Jatras Professorship in

continued on page 10

continued from page 9

Electrical and Computer Engineering. This chair was established in 1997 and has been previously held by Mark H. Kryder (first recipient, 1997) and Rob A. Rutenbar (second recipient, 2001).Stephen J. Jatras (EE '47) retired as chairman of the Telex Corporation. He was a Life Trustee of Carnegie Mellon, having served on the Board of Trustees since 1976, and co-chaired the ECE Advisory Board from its inception in 1992. The recipient of several alumni awards and a number of humanitarian awards for charitable work, Jatras died in January 2000.

December 2009

U Kang and Babis Tsourakakis: ICDM'09 Best Application Paper Award Runner Up

SCS graduate students U Kang and Babis Tsourakakis attracted the Best Applications Paper Award (runner up), at the International Conference on Data Mining (ICDM'09), held this year in Miami, Florida, for their paper "PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations" by U Kang, Charalampos (Babis) Tsourakakis, and Christos Faloutsos, ICDM 2009, Miami FL.

The paper was selected from among 70 accepted papers, out of a total of 786 submissions, and it showed how to use 'hadoop' and Yahoo's M45 machine, to analyze one of the largest publicly available graphs (over 100Gb). Moreover, the paper has been invited for fast-track possible publication to the KAIS journal.

November 2009

Tudor Dumitraş Awarded John Vlissides Award

Tudor Dumitraş was awarded the prestigious John Vlissides Award at the 2009 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA). This award is given to the doctoral student showing significant promise in applied software research and the most potential for having a significant impact on the practice of software development.

Dumitraş, who also won the ACM Student Research Competition at OOPSLA 2009, is advised by Priya Narasimhan, associate professor of electrical and computer engineering. His research is currently focused on dependable, online upgrades in distributed systems.

--CMU 8.5x11 News, Nov. 12, 2009

October 2009

Best Paper Award from SOSP'09 in Big Sky, Montana!

Huge congratulations to Amar Phanishayee, Jason Franklin, Lawrence Tan, Vijay Vasudevan and Dave Andersen on their best paper award at the 22nd ACM Symposium on Operating Systems Principles (SOSP '09). Their paper "FAWN: A Fast Array of Wimpy Nodes" presents a new cluster architecture for low-power data-intensive computing.

October 2009

Adrian Perrig Wins Award for Innovative Cybersecurity Research

Adrian Perrig was awarded a Security 7 Award from Information Security magazine for innovative cybersecurity research in academia. Perrig, technical director of Carnegie Mellon CyLab, a professor in the departments of Electrical and Computer Engineering and Engineering and Public Policy, and the School of Computer Science, will be recognized in the magazine's October issue. The magazine's editor, Michael S. Mimoso, said the awards recognize the achievements of security practitioners and researchers in a variety of industries, including education.

--CMU 8.5x11 News Oct 15, 2009

August 2009

Cranor Receives NSF Funding for Interdisciplinary Doctoral Program in Privacy & Security

Associate Professor Lorrie Cranor and her colleagues received a five-year, $3 million grant from the National Science Foundation (NSF) to establish a Ph.D. program in usable privacy and security. "Carnegie Mellon's CyLab Usable Privacy and Security (CUPS) Doctoral Training Program will offer Ph.D. students a new cross-disciplinary training experience that helps them produce solutions to ongoing tensions between security, privacy and usability," said Cranor, associate professor in the Institute for Software Research, the Department of Engineering and Public Policy and Carnegie Mellon CyLab. She noted that students will be actively involved in Carnegie Mellon's broad usable privacy and security research, which spans three major approaches: finding ways to build systems that "just work" without involving humans in security-critical functions; finding ways of making secure systems intuitive and easy to use; and finding ways to effectively teach humans how to perform security-critical tasks. For more on the new program, including a list of core faculty, visit http://www.cmu.edu/news/archive/2009/August/aug25_doctoralprogram.shtml.

--CMU 8.5x11 News, August 27, 2009

August 2009

Priya follows up the YinzCam with iBurgh

Pittsburgh is the first U.S. city with its own iPhone app. iBurgh, developed by Priya Narasimhan and her research group, allows users to take a picture of civic problems such as potholes, graffiti or other hazards and directly upload them, accompanied by a GPS location, to city council and other municipal administration authorities for review. Pittsburgh iPhone users can find the application at the App Store on their phones or at http://appshopper.com/utilities/iburgh.

Previous to the iBurgh app, Priya and her

continued on page 11

continued from page 10

group launched the YinzCam, another mobile phone app which allows hockey fans to view replays and alternate action angles at Pittsburgh Penguins hockey games on their phones or other handheld WiFi devices. And to top off all Priya's good news, YinzCam made Network World's top 10 list of sports innovations to love! (http://www.networkworld.com/slideshows/2009/081809-sports-technologies.html# -- slide 10).

August 2009

Nikos Hardavellas Appointed to June & Donald Brewer Chair of EE/CS at Northwestern

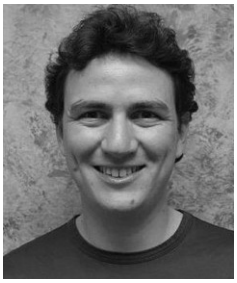
Congratulations to Nikos, soon to be June and Donald Brewer Assistant Professor of Electrical Engineering and Computer Science at Northwestern University. He has been appointed to the endowed chair for a two-year period, September 1, 2009 to August 31, 2011. Along with the title and honor, Prof. Hardavellas will receive a discretionary fund for each of the two years. This chair is awarded to Northwestern University's very best young faculty in the McCormick School of Engineering.

July 2009

Carlos Guestrin Wins Presidential Early Career Award

Carlos Guestrin, the Finmeccanica Assistant Professor of Computer Science and Machine Learning, has won a Presidential Early Career Award for Scientists and Engineers (PECASE), the highest honor bestowed by the U.S. government on scientists and engineers beginning their careers. He was nominated by the Department of Defense, which recognized him last year with the Office of Naval Research's Young Investigator Award.

The PECASE program recognizes 100 scientists and engineers who show exceptional potential for leadership at the frontiers of knowledge. "These extraordinarily gifted young scientists and engineers represent the best in our country," President Obama said. "With their talent, creativity and dedication, I am confident that they will lead their fields in new breakthroughs and discoveries and help us use science and technology to lift up our nation and our world."



Guestrin's long-term research interest is developing efficient algorithms and methods for designing, analyzing and controlling complex real-world systems. A painter, Guestrin also explores the intersection of computer science and art. Last semester, he and Visiting Art Professor Osman Khan co-taught "New Media Installation: Art That Learns," an interdisciplinary class in which students created interactive installations that incorporated the learning ability of computers (http://www.youtube.com/watch?v=ey9bZJOidHg). For more on the PECASE award and Guestrin's other honors, visit http://www.cmu.edu/news/archive/2009/July/july10_guestrinaward.shtml

--CMU 8.5x11 News, July 16, 2009

June 2009

Greg Ganger Earns Prestigious HP Innovation Research Award

Greg Ganger, a professor of electrical and computer engineering and director of the Parallel Data Lab, is among 60 recipients worldwide who received 2009 HP Innovation Research Awards. The award encourages open collaboration with HP Labs for mutually beneficial, high-impact research.

Ganger, who also received an HP Innovation Lab Award in 2008, will lead a research initiative in collaboration with HP Labs focused on data storage infrastructure issues, based on his winning proposal "Toward Scalable Self-Storage." Ganger was chosen from a group of nearly 300 applicants from more than 140 universities in 29 countries on a range of topics within the eight high-impact research themes at HP labs - analytics, cloud computing, content transformation, digital commercial print, immersive interaction, information management, intelligent infrastructure and sustainability.

"This award recognizes the ongoing innovative work that Carnegie Mellon professors bring to all collaborative research efforts," said Mark S. Kamlet, Carnegie Mellon provost and senior vice president. "We are proud of their accomplishments and the vital impact their research will have for a variety of industry sectors."

--CMU 8.5x11 News, June 17, 2009

June 2009

Jure Leskovek Wins ACM Doctoral Dissertation Award

Jure Leskovec won the prestigious 2009 SIGKDD Doctoral Dissertation Award from the Association of Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining for his thesis "Dynamics of Large Networks." He was advised by School of Computer Science Professor Christos Faloutsos, who also advised the 2008 runner-up Jimeng Sun. Leskovec will present a short summary of his work at the SIGKDD Conference in Paris on Sunday, June 28.

--CMU 8.5x11 News, June 4, 2009

Michael Stroucken photographs an assignment to verify its completion to his professor while attending the Retreat.

DISSERTATION ABSTRACT:
File System Virtual Appliances

Michael Abd-El-Malek

Carnegie Mellon University
ECE Ph.D. Dissertation, CMU-PDL-09-109, August 4, 2009.

Implementing and maintaining file systems is painful. OS functionality is notoriously difficult to develop and debug, and file systems are more so than most because of their size and interactions with other OS components. In-kernel file systems must adhere to a large number of internal OS interfaces. Though difficult during initial file system development, these dependencies particularly complicate porting a file system to different OSs or even across OS versions.

This dissertation describes an architecture that addresses the file system portability problem. Virtual machines are used to decouple the OS on which a file system runs from the OS on which user applications run. The file system is distributed as a file system virtual appliance (FSVA), a virtual machine running the file system developers' preferred OS (version). Users runs their applications in a separate virtual machine, using their preferred OS (version).

An FSVA design and implementation is described that maintains file system semantics with few, if any, code changes. This is achieved by sending all file system operations from the user OS to the FSVA. A unified buffer cache is maintained by using shared memory between the user OS and FSVA and by letting the user OS control the FSVA's buffer cache size. Features such as resource isolation and security are maintained through a single FSVA-per-user-OS design. Virtual machine migration is supported by simultaneously migrating a user OS and FSVA(s), maintaining shared memory mappings and live migration's low downtime.

Several case studies demonstrate FSVAs' effectiveness in providing OS-independent file system implementations. Measurements show that FSVA overheads on different workloads vary from 0–40%. The main overhead source is the communication latency between the user OS and FSVA. If a processor core is dedicated to an FSVA, a power-efficient polling mechanism reduces the overheads to 0--10%. Alternatively, relaxing the FSVA design goals by handling the frequent access-control file system checks in the user OS leads to similar overhead reductions as polling, but without the need for an additional core.

DISSERTATION ABSTRACT:
Efficient Byzantine Fault Tolerance for Scalable Storage and Services

James Vincent Hendricks

Carnegie Mellon University
SCS Ph.D. Dissertation, CMU-CS-09-146, July 16, 2009.

Distributed systems experience and should tolerate faults beyond simple component crashes as such systems grow in size and importance. Unfortunately, tolerating arbitrary faults, also known as Byzantine faults, poses several challenges to system designers, often limiting performance, requiring additional hardware, or both. This dissertation presents new protocols that provide substantially better performance than previously demonstrated. The Byzantine fault-tolerant erasure-coded block storage protocol proposed in this thesis provides 40% higher write throughput than the best prior approach. The Byzantine fault-tolerant replicated state machine provides a factor of 2.2--2.9 times higher throughput than the best prior approach. Furthermore, the protocols presented in this dissertation require 25--33% fewer responsive servers than the nearest competitors. To enable these results, this dissertation introduces several new techniques, including homomorphic fingerprinting, partial encoding, and Byzantine Locking, that provide unprecedented scalability, higher throughput, lower latency, and lower computational overhead. This dissertation also considers new methods for analyzing the correctness of distributed systems in the presence of faulty clients. Distributed services and storage systems built using these techniques can provide Byzantine fault tolerance in a more efficient, higher performance, and more scalable manner than previously thought possible.

DISSERTATION ABSTRACT:
Delayed Instantiation Bulk Operations for Management of Distributed, Object-based Storage Systems

Andrew J. Klosterman

Carnegie Mellon University
ECE Ph.D. Dissertation, CMU-PDL-09-108, August 17, 2009.

The basic distributed, object-based storage system model lacks features for storage management. This work presents and analyzes a strategy for using existing facilities to implement atomic operations on sets of objects. These bulk operations form the basis for managing snapshots (read-only copies) and forks (read-write copies) of portions of the storage system. Specifically, we propose to leverage the access control capabilities, and annotations at the metadata server, to allow for selective clone and delete operations on sets of objects.

In order to act upon a set of objects, a bulk operation follows these steps. First, the metadata server accepts the operation, contacts the storage nodes to revoke outstanding capabilities on the set of objects, and retains a record of the operation and the affected set of objects. At this point, clients can make no changes to existing objects since any capabilities they hold will be rejected by storage nodes. Second, when clients subsequently contact the

continued on page 13

continued from page 12

metadata server to access affected objects (e.g., acquire fresh capabilities), any records of bulk operations are consulted. Finding that a client is accessing an affected object, the metadata server will take the necessary steps to enact the uninstantiated operation before responding to the client request. This eventual enforcement of operation semantics ensures compliance with the operation's intent but delays the corresponding work until the next client access. With appropriate background instantiation, the work of instantiating bulk operations can be hidden from clients.

In this dissertation, we present algorithms suitable for performing bulk operations over distributed objects using m-of-n encodings. The core logic is concentrated at the metadata server, with minimal support at clients and storage nodes. We quantify the overheads associated with the implementation and describe schemes for mitigating them. We demonstrate the use of bulk operations to create snapshots in an NFS server running atop distributed, object-based storage.

DISSERTATION ABSTRACT:
Chip Multiprocessors for Server Workloads

Nikos Hardavellas

Carnegie Mellon University
SCS Ph.D. Dissertation, CMU-CS-09-150, July 2009.

We stand on the cusp of the giga-scale era of chip integration. Technological advancements in semiconductor fabrication yield ever-smaller and faster devices, enabling billion-transistor chips with multi-gigahertz clock frequencies. To utilize the abundant transistors on chip, modern processors pack an exponentially increasing number of cores on chip, multi-megabyte caches, and large interconnects to facilitate intra-chip data transfers. However, the growing on-chip resources do not directly translate into a commensurate increase in performance. Rather, they come at the cost of increased on-chip data access latency, while thermal considerations and pin constraints limit the parallelism that a multicore chip can support.

To mitigate the increasing on-chip data access latency, cache blocks on chip should be placed close to the cores that use them. We observe that cache access patterns can be classified at run time into distinct classes with different on-chip block placement requirements. Based on this observation, we propose Reactive NUCA (R-NUCA), a distributed cache design which reacts to the class of each access to place blocks close to the requesting cores. We then explore the design space of physically-constrained multicore processors, and find that future multicores should utilize low-operational-power transistors even for time-critical components (e.g., cores) to ease the power wall, employ novel on-chip block placement techniques to utilize efficiently large caches, while techniques like 3D-stacked memory can mitigate the off-chip bandwidth constraint even for peak-performance designs. Moving forward, we find that heterogeneous multicores hold great promise in improving designs even further.

DISSERTATION ABSTRACT:
Putting Home Data Management into Perspective

Brandon Watts Salmon

Carnegie Mellon University
ECE Ph.D. Dissertation, CMU-PDL-09-113, August 17, 2009.

Distributed storage is coming home. An increasing number of home and personal electronic devices create, use, and display digitized forms of music, images, videos, as well as more conventional files (e.g., nancial records and contact lists). In-home networks enable these devices to communicate, and a variety of device-specific and datatype-specific tools are emerging. The transition to digital homes gives exciting new capabilities to users, but it also makes them responsible for administration tasks usually handled by dedicated professionals in other settings. It is unclear that traditional data management practices will work for "normal people" reluctant to put time into administration.

This dissertation presents a number of studies of the way home users deal with their storage. One intriguing finding of these studies is that home users rarely organize and access their data via traditional folder-based naming. Usually, they do so based on data attributes. Computing researchers have long talked about attribute-based data navigation, while continuing to use folder-based approaches. However, users of home and personal storage live it. Popular interfaces (e.g., iTunes, iPhoto, and even drop-down lists of recently-opened Word documents) allow users to navigate file collections via attributes like publisher-provided metadata, extracted keywords, and date/time. In contrast, the abstractions provided by filesystems and associated tools for managing files have remained tightly tied to namespaces built on folders.

To correct the disconnect between semantic data access and folder-based replica management, this dissertation presents a new primitive that I call a "view", as a replacement for the traditional volume abstraction. A view is a compact description of a set of files, expressed much like a search query, and a device on which that data should be stored. For example, one view might be "all files with type=music and artist=Beatles stored on Liz's iPod" and another "all files with owner=Liz stored on Liz's laptop". Each device participating in a view-based filesystem maintains and publishes one or more views to describe the files that it stores. A view-based filesystem ensures that any file that matches a view will eventually be stored on the device named

continued on page 14

continued from page 13

in the view. Since views describe sets of files using the same attribute-based style as users' other tools, view-based management replica management should be easier than folder-based file management.

In this dissertation I present the design of Perspective, a view-based filesystem, and Insight, a set of view-based management tools. User studies, deployments and benchmarks using these prototypes show that view-based management simplifies some important tasks for non-technical users and can be supported efficiently by a distributed filesystem.

THESIS ABSTRACT:
Log-based Approaches to Characterizing and Diagnosing MapReduce Systems

Jiaqi Tan

Carnegie Mellon University SCS Master's Thesis, CMU-CS-09-143, July 2009.

MapReduce programs and systems are large-scale, highly distributed and parallel, consisting of many interdependent Map and Reduce tasks executing simultaneously on potentially large numbers of cluster nodes. They typically process large datasets and run for long durations. Thus, diagnosing failures in MapReduce programs is challenging due to their scale. This renders traditional time-based Service-Level Objectives ineffective. Hence, even detecting whether a MapReduce program is suffering from a performance problem is difficult. Tools for debugging and profiling traditional programs are not suitable for MapReduce programs, as they generate too much information at the scale of MapReduce programs, do not fully expose the distributed interdependencies, and do not expose information at the MapReduce level of abstraction. Hadoop, the open-source implementation of MapReduce, natively generates logs that record the system's execution, with low overheads. From these logs, we can extract state-machine views of Hadoop's execution, and we can synthesize these views to create a single unified, causal, distributed control-flow and data-flow view of MapReduce program behavior. This state-machine view enables us to diagnose problems in MapReduce systems. We can also generate visualizations of MapReduce programs in combinations of the time, space, and volume dimensions of their behavior that can aid users in reasoning about and debugging performance problems. We evaluate our diagnosis algorithm based on these state-machine views on synthetically injected faults on Hadoop clusters on Amazon's EC2 infrastructure. Several examples illustrate how our visualization tools were used to optimize application performance on the production M45 Hadoop cluster.

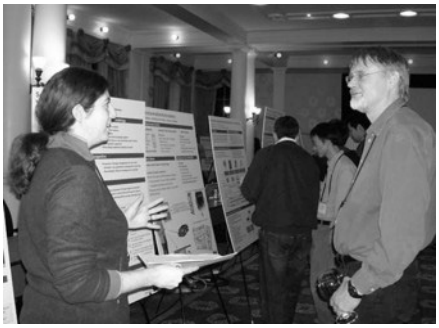
THESIS ABSTRACT:
The Blind Men and the Elephant: Piecing Together Hadoop for Diagnosis

Xinghao Pan

Carnegie Mellon University SCS Master's Thesis, CMU-CS-09-135, May 2009.

Google's MapReduce framework enables distributed, data-intensive, parallel applications by decomposing a massive job into smaller (Map and Reduce) tasks and a massive data-set into smaller partitions, such that each task processes a different partition in parallel. However, performance problems in a distributed MapReduce system can be hard to diagnose and to localize to a specific node or a set of nodes. On the other hand, the structure of large number of nodes performing similar tasks naturally affords us opportunities for observing the system from multiple viewpoints.

We present a "Blind Men and the Elephant" (BliMeE) framework in which we exploit this structure, and demonstrate how problems in a MapReduce system can be diagnose by corroborating the multiple viewpoints. More specifically, we present algorithms within the BliMeE framework based on OS-level performance counters, on white-box metrics extracted from logs, and on application-level heartbeats. We show that our BliMeE algorithms are able to capture a variety of faults including resource hogs and application hangs, and to localize the fault to subsets of slave nodes in the MapReduce system.



Michelle Mazurek discusses her research on home storage with John Wilkes of Google at the PDL Retreat.

In addition, we discuss how the diagnostic algorithms' outcomes can be further synthesized in a repeated application of the BliMeE approach. We present a simple supervised learning technique which allows us to identify a fault if it has been previously observed.

THESIS ABSTRACT:
Joulesort On A Low-Power CPU-GPU Hybrid Architecture

Kee-Tee (Lawrence) Tan

Carnegie Mellon University ECE Master's Thesis, December 7, 2009.

This paper analyses the energy efficiency of a low-power CPU-GPU hybrid architecture. We evaluate the NVIDIA Ion architecture, which couples an Intel Atom processor with an integrated GPU that has an order of magnitude fewer processors compared to traditional discrete

continued on page 15

continued from page 14

GPUs. We attempt to create a system that balances computation and I/O capabilities by attaching flash storage that allows sequential access to data with very high throughput. To evaluate this architecture, we implemented a Joulesort candidate that can sort in excess of 18000 records per Joule. We discuss the techniques used to ensure that the work is distributed between the CPU and the GPU so as to fully utilize system resources. We also analyse the different components in this system and attempt to identify the bottlenecks, which will help guide future work using such an architecture.

THESIS PROPOSAL:
Scalability, Usability and Applicability of Massive File System Directories

Swapnil Patil, SCS

May 10, 2010

Over the last decade file system evolution has favored scaling for large files instead of scaling for large number of files. And two forces are calling to change this: workloads that generate large number of small I/O accesses at high speeds and the dramatically increasing application-level parallelism. Data-intensive applications in high-performance computing (HPC) have a growing need for POSIX-like file systems for trillions of files and directories with billions of files—and most large-scale file systems are ill-equipped to meet these new requirements.

This thesis proposes to understand the tradeoffs in scaling traditional file system directories to store billions of files and sustain hundreds of thousands of concurrent mutations per second. We will also focus on the challenges in using such large mutating directories with the existing programming API and production file system implementations. Finally, we will talk about how we can apply scalable file system directories to provide generalized abstractions that simplify the development of data management infrastructure for non-HPC environments such as Internet services.

THESIS PROPOSAL:
Fast Algorithms for Time Series Mining

Lei Li, SCS

December 1, 2009

Time series data arise in numerous applications, such as motion capture, computer network monitoring, data center monitoring, environmental monitoring and many more. Finding patterns in such collections of sequences is crucial for leveraging them to solve real-world, domain specific problems, for example, to build humanoid robots, to detect pollution in drinking water, and to identify intrusion in computer networks.

The central theme of our work is to answer the question: how to find interesting and unexpected patterns in large time series? In this proposal, we focus on fast algorithms on mining large collections of co-evolving time series, with or with out missing values. We will present three pieces of our current work: natural stitching of human motions, time series mining and summarization with missing values, and a parallel learning algorithm for the underlying model, Linear Dynamical Systems (LDS). Algorithms proposed in these work allow us to obtain meaningful patterns effectively and efficiently, and subsequently to perform various mining tasks including forecasting, compression, and segmentation for co-evolving time series, even with missing values. Furthermore, we apply our algorithms to solve practical problems including recovering occlusions in human motion capture, and generating natural motions by stitching together carefully chosen pairs of candidates. We also proposed a parallel learning algorithm for LDS to fully utilize the power of multicore/multiprocessors, which will serve as a corner stone of many applications and algorithms for time series. All our algorithms scale linearly with respect to the length of sequences, and outperform the competitors often by large factors.

Based on aforementioned work, we propose to attack a number of interesting problems in mining time series data, which can be categorized into two classes: (a) without missing values: including feature extraction, indexing, clustering and data stream monitoring; (b) with missing values: mining under domain constraints, like bone-length constraints in motion capture sequences. Potential applications of these proposed work include occlusion recovery for motion capture, fast retrieval of similar sequences in a large database, and anomaly detection in sensor data and network traffics.

THESIS PROPOSAL:
Improving bandwidth guarantees for storage workloads with performance insulation

Matthew Wachs, SCS

April 8, 2010

Many storage workloads do not need the level of performance afforded by a dedicated storage system, but do need predictable and controllable performance. This makes sharing a storage system among multiple such workloads or clients appealing, if quality-of-service guarantees can be made. Unfortunately, inter-workload interference, such as a reduction of locality when multiple request streams are interleaved, can result in dramatic loss of efficiency and performance.

Performance insulation is a storage system property where each workload sharing the system is assigned a fraction of disk time and receives nearly that fraction of its standalone (dedicated-disk) bandwidth. Because there is usually some overhead caused by shar-

continued on page 17

continued from page 1

ter's slower CPUs are able to dedicate more transistors to basic operations and execute significantly more instructions per Joule than their faster counterparts: multi-GHz superscalar quad-core processors can execute approximately 100 million instructions per Joule, assuming all cores are active and avoid stalls or mispredictions. Lower-frequency in-order CPUs can provide over 1 billion instructions per Joule—an order of magnitude more efficient while still running at 1/2 the frequency.

While FAWN is based upon fundamental hardware efficiencies, our experience suggests that applications must be tailored to operate in the memory and CPU-constrained environment before they can reap these efficiency benefits. We have explored several designs in random access and throughput-intensive workloads. Below, we explain briefly how we designed our key-value store, FAWN-KV, to operate in this environment and to take advantage of the fast Flash storage on our nodes.

Figure 1 gives an overview of the entire FAWN-KV system. FAWN-KV begins with a log-structured per-node data-store to serialize writes and make them fast on flash. It then uses this log structure as the basis for chain replication between cluster nodes, providing reliability and strong consistency, while ensuring that all maintenance operations—including failure handling and node insertion—require only efficient bulk sequential reads and writes.

Client requests enter the system at one of several front-ends. The front-end nodes forward the request to the back-end FAWN-KV node responsible for serving that particular key. The back-end node serves the request from its FAWN-DS log-structured per-node datastore and returns the result to the front-end (which in turn replies to the client). Writes proceed similarly.

FAWN-DS is a log-structured key-value store. Each store contains values for the key range associated with one virtual ID. It acts to clients like a disk-based hash table that supports Store, Lookup, and Delete. It is designed specifically to perform well on flash storage and to operate within the constrained DRAM available on wimpy nodes: all writes to the datastore are sequential, and reads require a single random access.

FAWN-DS is designed specifically to perform well on flash storage and to operate within the constrained DRAM available on wimpy nodes: all writes to the datastore are sequential, and reads require a single random access. To map a key to a value FAWN-DS uses an in-memory (DRAM) Hash Index to map 160-bit keys to a value stored in the Data Log. It stores only a fragment of the actual key in memory to find a location in the log; it then reads the full key (and the value) from the log and verifies that the key it read was, in fact, the correct key. The design trades a small and configurable chance of requiring two reads from flash (we set it to roughly 1 in 32,768 accesses) for drastically reduced memory requirements (only six bytes of DRAM per key-value pair).

This append-only data log provides the basis for replication and strong consistency using chain replication between nodes (Fig. 2). Data is distributed across nodes using consistent hashing, with data split into contiguous ranges on disk such that all replication and node insertion operations involve only a fully in-order traversal of the subset of data that must be copied to a new node. Together with the log structure, these properties combine to provide fast failover and fast node insertion, and they minimize the time the affected datastore's key range is locked during such operations—for a single node failure and recovery, the affected key range is blocked for at most 100 milliseconds.

The large number of back-end FAWN-KV storage nodes are organized into a ring using consistent hashing. Keys are mapped to the node that follows the key in the ring (its successor). To balance load and reduce failover times, each physical node joins the ring as a small number (V) of virtual nodes, each virtual node representing a virtual ID (VID) in the ring space. Each physical

continued on page 17

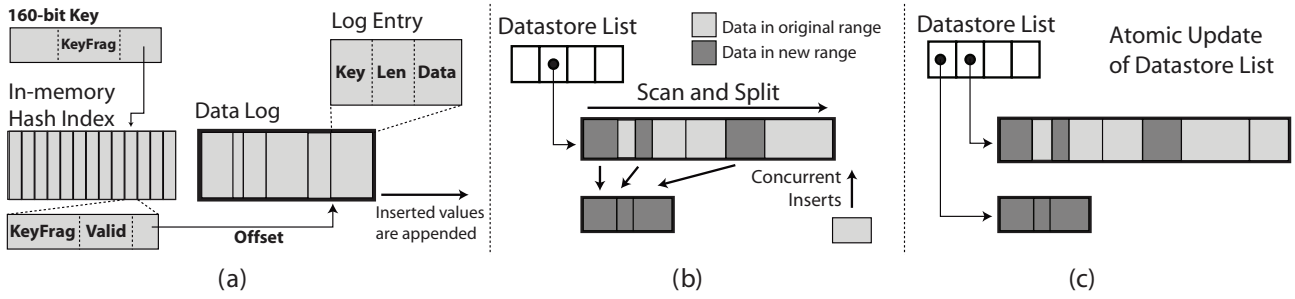


Figure 2: (a) FAWN-DS appends writes to the end of the Data Log. (b) Split requires a sequential scan of the data region, transferring out-of-range entries to the new store. (c) After scan is complete, the datastore list is atomically updated to add the new store. Compaction of the original store will clean up out-of-range entries.

continued from page 16

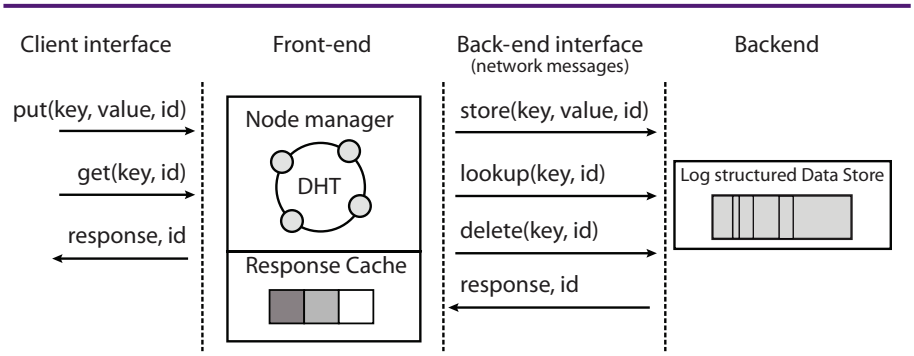


Figure 3: FAWN-KV Interfaces—Front-ends manage backends, route requests, and cache responses. Back-ends use FAWN-DS to store key-value pairs.

node is thus responsible for V different (non-contiguous) key ranges. The data associated with each virtual ID is stored on flash using FAWN-DS. Figure 3 depicts FAWN-KV request processing. Client applications send requests to front-ends using a standard put/get interface. Front-ends send the request to the back-end node that owns the key space for the request. The back-end node satisfies the request using its FAWN-DS and replies to the front-ends.

To date, we have built three prototype FAWN clusters. Our latest is coming online as we write this article in late spring 2010, consisting of twenty Intel Atom-based nodes running at 1.6GHz. Our evaluation of FAWN-KV uses our second prototype, made of twenty-one 500 MHz embedded CPUs. Each node can serve up to 1300 256-byte queries per second, exploiting nearly all of the raw I/O capability of their attached flash devices, and consumes under 5W when network and support hardware is taken into account. The FAWN cluster achieves 364 queries per Joule—two orders of magnitude better than traditional disk-based clusters, demonstrating that the FAWN architecture has significant potential for many I/O-intensive workloads.

For an in-depth explanation of the FAWN architecture and evaluation of the performance of the FAWN system, please see our 2009 SOSP paper "FAWN: A Fast Array of Wimpy Nodes" [1].

[1] FAWN: A Fast Array of Wimpy Nodes. David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, Vijay Vasudevan. Proceedings of the 22nd ACM Symposium on Operating Systems Principles (SOSP'09), Big Sky, MT., Oct, 2009.

DISSERTATIONS & PROPOSALS

continued from page 15

ing, there could be a drop in efficiency; but a system providing performance insulation provides a lower bound on relative efficiency at all times, called the R-value. An experimental storage system server called Argon confirms that performance insulation can be achieved in practice for R-values of 0.8-0.9.

While Argon and performance insulation provide a limit on loss of efficiency, many storage workloads also need performance guarantees. For instance, a video streaming application may need no less than 6 MB/s of bandwidth to draw frames at the appropriate rate; it is more straightforward to express this directly as a performance guarantee than indirectly as an efficiency guarantee. To ensure performance guarantees will consistently be met, the appropriate amount of resources need to be reserved and a feasibility test must be performed over the set of workloads assigned to a system -- the process of admission control. Unfortunately, due to the physical characteristics of hard disk drives (mechanical positioning times), the amount of resources needed to provide a given level of bandwidth can vary by more than an order of magnitude, depending on the access pattern. Additional, significant variability and uncertainty in resource requirements are caused by inter-workload interference.

Though much of the difficulty of the admission control problem is fundamental, storage systems with the property of performance insulation strictly limit inter-workload interference, removing the major source of "artificial" complexity in making appropriate reservations. The proposed thesis will compare Argon, a system with performance insulation, to systems proposed in the literature that do not explicitly manage efficiency. We will build mechanisms for providing and managing bandwidth guarantees on top of Argon, and implement basic admission control schemes in our system. Using various combinations of workloads, we will experimentally confirm that efficiency is higher and that admission control is more accurate and effective in Argon, making bandwidth guarantees more reliable and efficient.

RECENT PROGRESS IN CLAYTRONICS

Seth Copen Goldstein & The Claytronics Group

One way to view claytronics is as a massively distributed system where the resource-constrained compute nodes are connected by a sparse, constantly-changing network where physical locality is important. Such systems are notoriously hard to program, but are also becoming increasingly commonplace and more important: e.g., sensor networks, peer-to-peer applications on the Internet, and even manycore processors.

In our effort to learn how to control claytronics, we developed two programming languages *LDP* and *Meld* that are oriented towards programming the system as *a whole* rather than as individual nodes. These new programming languages could be viewed as *coordination* languages. LDP is a reactive programming language that allows a programmer to specify actions to be taken out on sub-groups of the system which satisfy certain distributed properties. For example, one could write a simple rule to look at the load on every group of three connected processors and have the heaviest-loaded processor within that group transfer work to the most lightly-loaded of the three. Meld is a forward-chaining logic programming language that can construct programs which sense the state of the system, and in the process of proving the facts that make up a program, can also alter the state of the system through side-effecting facts. For claytronics, the side-effecting facts can be used to change the shape of the system. We are in the process of showing how Meld can be used to program more traditional concurrent systems: e.g., multicore processors. In this environment the side-effecting facts are used to coordinate the parallel program: e.g., distribute work or change the layout of data structures.

A prototype 1mm diameter tube housing a claytronics control-die.

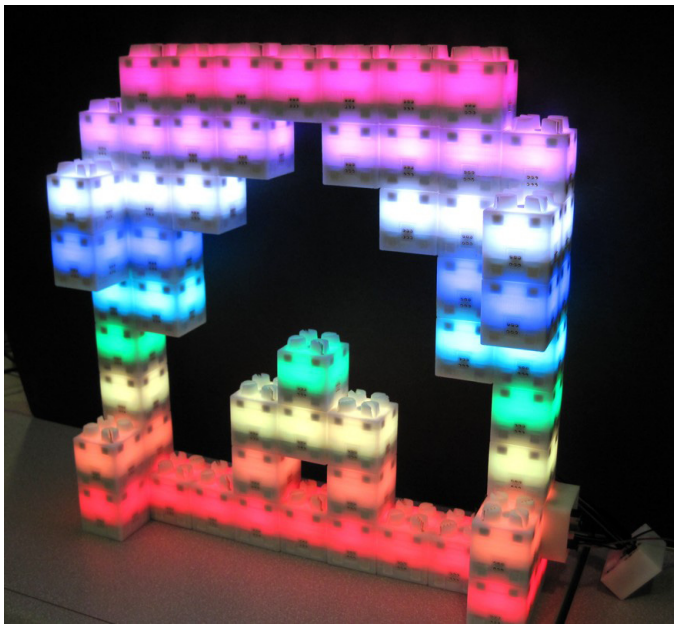


Figure 2: A 58-block BlinkyBlock system running an LDP program.

Until recently, both LDP and Meld have only been used on our claytronics simulator. We are still in the process of creating the individual units of a claytronics system through a combination of photolithography and MEMS-based directed self-assembly. Figure 1 shows a recently created 1mm diameter tube with a control-die inside. To test our programming ideas on real hardware, we have also developed *BlinkyBlocks*, a distributed system based on 4mm-cubed blocks. Each block has a processor, a multi-color LED, an accelerometer, and a 6 port-network switch. When connected together, the blocks form a mesh-network. The blocks run both LDP and MELD programs and enable us to test our programs on real hardware. A 58-block system running an LDP program is shown in Figure 2. We invite you to come and play with the BlinkyBlock system in our lab (GHC 7114) or to download the development system and simulator (www.cs.cmu.edu/claytronics/blinkyblocks).

It is clear that future computing systems will rely more and more heavily on concurrency, whether they are multicore processors, large-scale clusters, or programmable matter. Hence parallel execution will be the main method for meeting future computing needs. Unfortunately, writing correct and efficient concurrent programs (for either distributed or parallel systems) is notoriously difficult. We hope that hardware systems such as the BlinkyBlocks and languages like LDP and Meld will offer a platform for understanding how to program concurrent systems more easily.

Mike Kasick discusses "Black-Box Problem Diagnosis in Parallel File Systems" at the 17th PDL Workshop & Retreat.

continued from page 4

SC Petascale Data Storage Workshop held in Portland, OR.

- Xinghao Pan presented the paper "Blind Men and the Elephant: Piecing Together Hadoop for Diagnosis" at ISSRE 2009 in Mysuru, India.

October 2009

- Dave Andersen and his group received the best paper award at SOSP '09 for their paper "FAWN: A Fast Array of Wimpy Nodes" from SOSP'09 in Big Sky, MT. Vijay Vasudevan presented the paper.
- Garth Gibson discussed "Understanding and Maturing the Data-Intensive Scalable Computing Storage Substrate" at the 2009 Microsoft eScience Workshop in Pittsburgh, PA.

September 2009

- Greg Ganger presented "Towards Self-* Storage" at NEC Labs.
- Garth Gibson presented "Cloud Storage and Parallel File Systems" to the Storage Networking Industry Association's Storage Developer Conference (SDC09), Santa Clara, CA.

August 2009

- Garth Gibson presented "Update on Petascale Data Storage Institute" at HEC FSIO 2009 in Arlington, VA. Greg Ganger attended.
- Michael Abd-El-Malek successfully defended his Ph.D. dissertation on "File System Virtual Appliances" and is enjoying life at Google in California.
- Andrew J. Klosterman completed his Ph.D with the defense of his research on "Delayed Instantiation Bulk Operations for Management of Distributed, Object-based Storage Systems" and has gone on to work with Avere Systems here in Pittsburgh.
- Brandon Salmon successfully defended his Ph.D. research on "Putting Home Data Management into Perspective." He is now with Tintri Systems in San Francisco.
- Nikos Hardavellas has been appointed to the June & Donald Brewer Chair of EE/CS at Northwestern University. He defended his Ph.D. research on "Chip Multiprocessors for Server Workloads" in July.
- Vijay Vasudevan presented "Safe and Effective Fine-grained TCP Retransmissions for Datacenter Communication" at SIGCOMM'09 in Barcelona, Spain.

July 2009

- James Hendricks successfully defended his Ph.D. Dissertation on "Efficient Byzantine Fault Tolerance for Scalable Storage and Services" and is now working with Google.
- Jiaqi Tan completed an internship with the Cloud Computing & Grid Infrastructure Group at Yahoo! in Sunnyvale, CA from May to July 2009, working with Mac Yang to develop diagnosis tools for Hadoop.
- Jiaqi Tan presented his Master's thesis titled "Log-based Approaches to Characterizing and Diagnosing MapReduce Systems." He is now with DSO National Laboratories in Singapore.
- Greg Ganger presented "Towards Self-* Storage" in the EMC Innovation Network Lecture Series.

June 2009

- Jiaqi presented the paper "Ganesha: Black-Box Diagnosis for MapReduce Systems" at HotMetrics '09 in Seattle, WA.
- Jure Leskovec won the 2009 SIGKDD Doctoral Dissertation Award from ACM's Special Interest Group on Knowledge Discovery and Data Mining for his thesis "Dynamics of Large Networks." Jure is now an Assistant Professor of Machine Learning at Stanford.
- Greg Ganger spoke at a Technical Exchange on a visit to APC in Providence, RI.

Bin Fu presents his research on "Astronomy Application of Map-Reduce: A Distributed Friends-of-Friends Algorithm" at the PDL Retreat.

- Garth Gibson presented "Hot-Cloud Panel: A Storage Viewpoint" and "In Search of an API for Scalable File Systems: Under the Table or Above it?" at the Workshop on Hot Topics in Cloud Computing, San Diego, CA.

May 2009

- 11th Annual PDL Spring Industry Visit Day.
- Xinghao Pan presented his Master's research titled "The Blind Men and the Elephant: Piecing Together Hadoop for Diagnosis." He is now with DSO National Laboratories in Singapore.
- Greg Ganger attended HEC FSIO 2009 in Arlington, VA.
- Swapnil Patil interned with Microsoft Research in Redmond, WA.
- Jiri Simsa interned with Microsoft Research in Cambridge, UK.
- Garth Gibson presented "Directions for TDMR System Architecture: Synergies with SSDs" at the IEEE International Symposium on Magnetics (INTERMAG09) in Sacramento CA.

continued from page 7

that people's mental models of access control and security are often misaligned with current systems. We detail these findings and present a set of associated guidelines for designing usable access-control systems for the home environment.

Black-Box Problem Diagnosis in Parallel File Systems

Kasick, Tan, Gandhi & Narasimhan

Proceedings of the 8th USENIX Conference on File and Storage Technologies(FAST '10), San Jose, CA, February 2010.

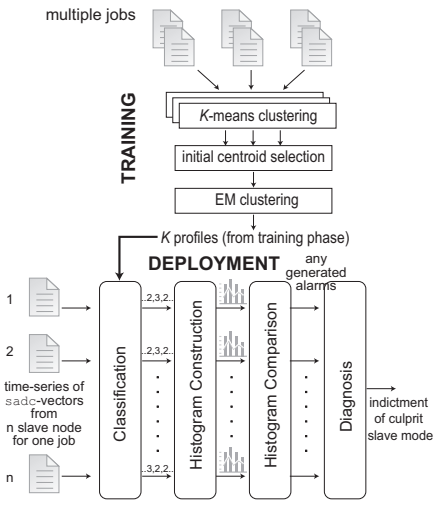
We focus on automatically diagnosing different performance problems in parallel file systems by identifying, gathering and analyzing OS-level, black-box performance metrics on every node in the cluster. Our peer comparison diagnosis approach compares the statistical attributes of these metrics across I/O servers, to identify the faulty node. We develop a root-cause analysis procedure that further analyzes the affected metrics to pinpoint the faulty resource (storage or network), and demonstrate that this approach works commonly across stripe-based parallel file systems. We demonstrate our approach for realistic storage and network problems injected into three different file-system benchmarks (dd, IOzone, and Post-Mark), in both PVFS and Lustre clusters.

Ganesha: Black-Box Diagnosis for MapReduce Systems

Pan, Tan, Kavulya, Gandhi & Narasimhan

Workshop on Hot Topics in Measurement & Modeling of Computer Systems (HotMetrics), Seattle, WA, June 2009.

Ganesha aims to diagnose faults transparently (in a black-box manner) in MapReduce systems, by analyzing OS-level metrics. Ganesha's approach is based on peer-symmetry under fault-free conditions, and can diagnose faults that manifest asymmetrically at nodes within a MapReduce system. We evaluate Ganesha by diagnosing Hadoop problems for the Gridmix Hadoop benchmark on 10-node and 50-node MapReduce clusters on Amazon's EC2. We also candidly highlight faults that escape Ganesha's diagnosis.



Ganesha's approach.

Radius Plots for Mining Terabyte Scale Graphs: Algorithms, Patterns, and Observations

Kang, Tsourakakis, Appel, Faloutsos & Jure Leskovec

SIAM International Conference on Data Mining (SDM) 2010, Columbus, Ohio, USA.

Given large, multi-million node graphs (e.g., Facebook, web-crawls, etc.), how do they evolve over time? How are they connected? What are the central nodes and the outliers of the graphs? We show that the Radius Plot (pdf of node radii) can answer these questions. However, computing the Radius Plot is prohibitively expensive for graphs reaching the planetary scale. There are two major contributions in this paper: (a) We propose HADI (HAdoop DIameter and radii estimator), a carefully designed and fine-tuned algorithm to compute the diameter of massive graphs, that runs on the top of the HADOOP /MAPREDUCE system, with excellent scale-up on the number of available machines (b) We run HADI on several real world datasets including YahooWeb (6B edges, 1/8 of a Terabyte), one of the largest public graphs ever analyzed. Thanks to HADI, we report fascinating patterns on large networks, like the surprisingly small effective diameter, the multi-modal/bi-modal shape of the Radius Plot, and its palindrome motion over time.

PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations

Kang, Tsourakakis & Faloutsos

IEEE International Conference on Data Mining (ICDM) 2009, Miami, Florida, USA. Best Application Paper (runner-up).

In this paper, we describe PEGASUS, an open source Peta Graph Mining library which performs typical graph mining tasks such as computing the diameter of the graph, computing the radius of each node and finding the connected components. As the size of graphs reaches several Giga-, Tera- or Peta-bytes, the necessity for such a library grows too. To the best of our knowledge, PEGASUS is the first such library, implemented on the top of the HADOOP platform, the open source version of MAPREDUCE. Many graph mining operations (PageRank, spectral clustering, diameter estimation, connected components etc.) are essentially a repeated matrix-vector multiplication. In this paper we describe a very important primitive for PEGASUS, called GIM-V (Generalized Iterated Matrix-Vector multiplication). GIM-V is highly optimized, achieving (a) good scale-up on the number of available machines (b) linear running time on the number of edges, and (c) more than 5 times faster performance over the non-optimized version of GIM-V. Our experiments

continued on page 21

continued from page 20

ran on M45, one of the top 50 supercomputers in the world. We report our findings on several real graphs, including one of the largest publicly available Web Graphs, thanks to Yahoo!, with 6.7 billion edges.

Blind Men and the Elephant: Piecing Together Hadoop for Diagnosis

Pan, Tan, Kavulya, Gandhi & Narasimhan

20th IEEE International Symposium on Software Reliability Engineering (ISSRE), Industrial Track, Mysuru, India, Nov 2009.

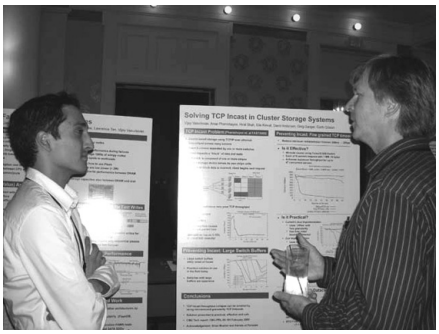
Google's MapReduce framework enables distributed, data-intensive, parallel applications by decomposing a massive job into smaller (Map and Reduce) tasks and a massive data-set into smaller partitions, such that each task processes a different partition in parallel. However, performance problems in a distributed MapReduce system can be hard to diagnose and to localize to a specific node or a set of nodes. On the other hand, the structure of large number of nodes performing similar tasks naturally affords us opportunities for observing the system from multiple viewpoints.

We present a "Blind Men and the Elephant" (Blimey) framework in which we exploit this structure, and demonstrate how problems in a MapReduce system can be diagnosed by corroborating the multiple viewpoints. More specifically, we present algorithms within the Blimey framework based on OS-level performance counters, on white-box metrics extracted from logs, and on application-level heartbeats. We show that our Blimey algorithms are able to capture a variety of faults including resource hogs and application hangs, and to localize the fault to subsets of slave nodes in the MapReduce system. In addition, we discuss how the diagnostic algorithms' outcomes can be further synthesized in a repeated application of the Blimey approach. We present a simple supervised learning technique which allows us to identify a fault if it has been previously observed.

…And Eat It Too: High Read Performance in Write-optimized HPC I/O Middleware File Formats

Polte, Lofstead, Bent, Gibson, Klasky, Liu, Parashar, Podhorszki, Schwan, Wingate & Wolf

4th Petascale Data Storage Workshop held in conjunction with Supercomputing '09, November 15, 2009. Portland, Oregon. As HPC applications run on increasingly high process counts on larger and larger machines, both the frequency of checkpoints needed for fault tolerance and the resolution and size of Data Analysis Dumps are expected to increase proportionally. In order to maintain an acceptable ratio of time spent performing useful computation work to time spent performing I/O, write bandwidth to the underlying storage system must increase proportionally to this increase in the checkpoint and computation size. Unfortunately, popular scientific self-describing file formats such as netCDF and HDF5 are designed with a focus on portability and exibility. Extra care and careful crafting of the output structure and API calls is required to optimize for write performance using these APIs. To provide sufficient write bandwidth to continue to support the demands of scientific applications, the HPC community has developed a number of I/O middleware layers, that structure output into write-optimized file formats. However, the obvious concern with any write optimized file format would be a corresponding penalty on reads. In the log-structured filesystem, for example, a file generated by random writes could be written efficiently, but reading the file back sequentially later would result in very poor performance. Simulation results require efficient read-back for visualization and analytics, and though most checkpoint files are never



Vijay Vasudevan discusses his poster on "Solving TCP Incast in Cluster Storage Systems" with Dan Dahle of Intel at the 2009 PDL Retreat & Workshop

used, the efficiency of a restart is very important in the face of inevitable failures. The utility of write speed improving middleware would be greatly diminished if it sacrificed acceptable read performance. In this paper we examine the read performance of two write-optimized middleware layers on large parallel machines and compare it to reading data natively in popular file formats.

Co-scheduling of Disk Head Time in Cluster-based Storage

Wachs & Ganger

28th International Symposium On Reliable Distributed Systems September 27-30, 2009. Niagara Falls, New York, U.S.A.

Disk timeslicing is a promising technique for storage performance insulation. To work with cluster-based storage, however, timeslices associated with striped data must be co-scheduled on the corresponding servers. This paper describes algorithms for determining global timeslice schedules and mechanisms for coordinating the independent server activities. Experiments with a prototype show that, combined, they can provide performance insulation for workloads sharing a storage cluster—each workload realizes a configured minimum effi-

continued on page 22

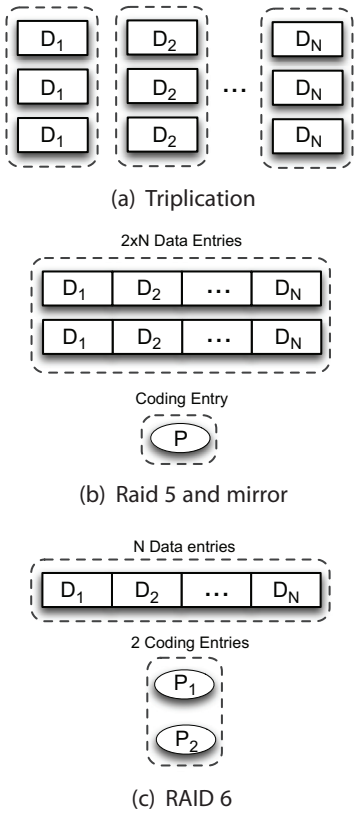
continued from page 21

ciency within its timeslices regardless of the activities of the other workloads.

DiskReduce: RAID for Data-Intensive Scalable Computing

Fan, Tantisiriroj, Xiao & Gibson

4th Petascale Data Storage Workshop held in conjunction with Supercomputing '09, November 15, 2009. Portland, Oregon. Data-intensive file systems, developed for Internet services and popular in cloud computing, provide high reliability and availability by replicating data, typically three copies of everything. Alternatively high performance computing, which has comparable scale, and smaller scale enterprise storage systems get similar tolerance for multiple failures from lower overhead erasure encoding, or RAID, organizations. DiskReduce is a modification of the Hadoop distributed file system HDFS) enabling asynchronous compression of initially triplicated data down to RAID-class redundancy overheads. In addition to increasing a cluster's storage capacity as seen by its users by up to a factor of three, DiskReduce can delay encoding long enough to deliver the performance benefits of multiple data copies.



Codewords providing protection against double node failures.

Understanding and Maturing the Data-Intensive Scalable Computing Storage Substrate

Gibson, Fan, Patil, Polte, Tantisiriroj & Xiao

Microsoft Research eScience Workshop 2009, Pittsburgh, PA, October 16-17, 2009.

Modern science has available to it, and is more productively pursued with, massive amounts of data, typically either gathered from sensors or output from some simulation or processing. The table below shows a sampling of data sets that a few scientists at Carnegie Mellon University have available to them or intend to construct soon. Data Intensive Scalable Computing (DISC) couples computational resources with the data storage and access capabilities to handle massive data science quickly and efficiently. Our topic in this extended abstract is the effectiveness of the data intensive file systems embedded in a DISC system. We are interested in understanding the differences between data intensive file system implementations and high performance computing (HPC) parallel file system implementations. Both are used at comparable scale and speed. Beyond feature inclusions, which we expect to evolve as data intensive file systems see wider use, we find that performance does not need to be vastly different. A big source of difference is seen in their approaches to data failure tolerance: replication in DISC file systems versus RAID in HPC parallel file systems. We address the inclusion of RAID in a DISC file system to dramatically increase the effective capacity available to users. This work is part of a larger effort to mature and optimize DISC infrastructure services.

PLFS: A Checkpoint Filesystem for Parallel Applications

Bent, Gibson, Grider, McClelland, Nowoczynski, Nunez, Polte & Wingate

Supercomputing '09, November 15, 2009. Portland, Oregon.

Parallel applications running across thousands of processors must protect themselves from inevitable system failures. Many applications insulate themselves from failures by checkpointing. For many applications, checkpointing into a shared single file is most convenient. With such an approach, the size of writes are often small and not aligned with file system boundaries. Unfortunately for these applications, this preferred data layout results in pathologically poor performance from the underlying file system which is optimized for large, aligned writes to non-shared files. To address this fundamental mismatch, we have developed a virtual parallel log structured file system, PLFS. PLFS remaps an application's preferred data layout into one which is optimized for the underlying file system. Through testing on PanFS, Lustre, and GPFS, we have seen that this layer of indirection and reorganization can reduce checkpoint time by an order of magnitude for several important benchmarks and real applications without any application modification.

No Downtime for Data Conversions: Rethinking Hot Upgrades

Dumitraş & Narasimhan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-106. July 2009.

Unavailability in enterprise systems is usually the result of planned events,

continued on page 23

continued from page 22

such as upgrades, rather than failures. Major system upgrades entail complex data conversions that are difficult to perform on the fly, in the face of live workloads. Minimizing the downtime imposed by such conversions is a time-intensive and error-prone manual process. We present Imago, a system that aims to simplify the upgrade process, and we show that it can eliminate all the causes of planned downtime recorded during the upgrade history of one of the ten most popular websites. Building on the lessons learned from past research on live upgrades in middleware systems, Imago trades off a need for additional storage resources for the ability to perform end-to-end, enterprise upgrades online, with minimal application-specific knowledge.

Safe and Effective Fine-grained TCP Retransmissions for Datacenter Communication

Vasudevan, Phanishayee, Shah, Krevat, Andersen, Ganger, Gibson & Mueller

SIGCOMM'09, August 17–21, 2009, Barcelona, Spain.

This paper presents a practical solution to a problem facing high-fan-in, high-bandwidth synchronized TCP workloads in datacenter Ethernets|the TCP incast problem. In these networks, receivers can experience a drastic reduction in application throughput when simultaneously requesting data from many servers using TCP. Inbound data overfills small switch buffers, leading to TCP timeouts lasting hundreds of milliseconds. For many datacenter workloads that have a barrier synchronization requirement (e.g., filesystem reads and parallel data-intensive queries), throughput is reduced by up to 90%. For latency-sensitive applications, TCP timeouts in the datacenter impose delays of hundreds of milliseconds in networks with round-trip-times in microseconds. Our practical solution uses high-resolution timers to enable microsecond-granularity

Michelle Mazurek presents "Access Control at Home: Attitudes, Needs, Practices" at the PDL Retreat.

TCP timeouts. We demonstrate that this technique is effective in avoiding TCP incast collapse in simulation and in real-world experiments. We show that eliminating the minimum retransmission timeout bound is safe for all environments, including the wide-area.

FAWN: A Fast Array of Wimpy Nodes

Andersen, Franklin, Kaminsky, Phanishayee, Tan & Vasudevan

Proceedings of the 22nd ACM Symposium on Operating Systems Principles (SOSP 2009), Big Sky, MT. October 2009. Best Paper Award!

This paper presents a new cluster architecture for low-power data-intensive computing. FAWN couples low-power embedded CPUs to small amounts of local flash storage, and balances computation and I/O capabilities to enable efficient, massively parallel access to data. The key contributions of this paper are the principles of the FAWN architecture and the design and implementation of FAWN-KV—a consistent, replicated, highly available, and high-performance key-value storage system built on a FAWN prototype. Our design centers around purely log-structured datastores that provide the basis for high performance on flash storage, as well as for replication and consistency obtained using chain replication on a consistent hashing ring. Our evaluation demonstrates that FAWN clusters can handle roughly 350 key-value queries per Joule of energy—two orders of magnitude more than a disk-based system.

Mochi: Visual Log-Analysis Based Tools for Debugging Hadoop

Tan, Pan, Kavulya, Gandhi & Narasimhan

Workshop on Hot Topics in Cloud Computing (HotCloud '09), San Diego, CA, on June 15, 2009. Mochi, a new visual, log-analysis based debugging tool correlates Hadoop's behavior in space, time and volume, and extracts a causal, unified control- and data-flow model of Hadoop across the nodes of a cluster. Mochi's analysis produces visualizations of Hadoop's behavior using which users can reason about and debug performance issues. We provide examples of Mochi's value in revealing a Hadoop job's structure, in optimizing real-world workloads, and in identifying anomalous Hadoop behavior, on the Yahoo! M45 Hadoop cluster.

Directions for Shingled-Write and Two-Dimensional Magnetic Recording System Architectures: Synergies with Solid-State Disks.

Gibson & Polte

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-104. May 2009.

Shingled-writing and two-dimensional magnetic recording, TDMR, will change core characteristics of magnetic disk operation and require systems software be adapted appropriately. Because a band of adjacent tracks overlap one another, they must be written

continued on page 24

continued from page 23

in a specific order. Once overlapped, a track cannot be updated in place, because the tracks overlapping it will be overwritten by the update. If this behavior is exposed to operating systems directly, there will be very low acceptance of these products. However, disk controller software can emulate full compliance with existing interfaces, and may be able to mask almost all performance implications as well.

In Search of an API for Scalable File Systems: Under the Table or Above It?

Patil, Gibson, Ganger, López, Polte, Tantisiroj & Xiao

USENIX HotCloud Workshop 2009. June 2009, San Diego CA.

"Big Data" is everywhere – both the IT industry and the scientific computing community are routinely handling terabytes to petabytes of data [24]. This preponderance of data has fueled the development of data-intensive scalable computing (DISC) systems that manage, process and store massive data-sets in a distributed manner. For example, Google and Yahoo have built their respective Internet services stack to distribute processing (MapReduce and Hadoop), to program computation (Sawzall and Pig) and to store the structured output data (Bigtable and HBase). Both these stacks are layered on their respective distributed file systems, GoogleFS [12] and Hadoop distributed FS [15], that are designed "from scratch" to deliver high performance primarily for their anticipated DISC workloads.

However, cluster file systems have been used by the high performance computing (HPC) community at even larger scales for more than a decade. These cluster file systems, including IBM GPFS [28], Panasas PanFS [34], PVFS [26] and Lustre [21], are required to meet the scalability demands of highly parallel I/O access patterns generated by scientific applications that execute simultaneously on tens to hundreds of thousands of nodes. Thus, given the importance of scalable storage to both the DISC and the HPC world, we take a step back and ask ourselves if we are at a point where we can distill the key commonalities of these scalable file systems.

This is not a paper about engineering yet another "right" file system or database, but rather about how do we evolve the most dominant data storage API – the file system interface–to provide the right abstraction for both DISC and HPC applications. What structures should be added to the file system to enable highly scalable and highly concurrent storage? Our goal is not to define the API calls per se, but to identify the file system abstractions that should be exposed to programmers to make their applications more powerful and portable. This paper highlights two such abstractions. First, we show how commodity large-scale file systems can support distributed data processing enabled by the Hadoop/MapReduce style of parallel programming frameworks. And second, we argue for an abstraction that supports indexing and searching based on extensible attributes, by interpreting BigTable [6] as a file system with a filtered directory scan interface.

System-Call Based Problem Diagnosis for PVFS

Kasick, Bare, Marinelli, Tan, Gandhi, Narasimhan

Proceedings of the 5th Workshop on Hot Topics in System Dependability (HotDep '09). Lisbon, Portugal. June 2009.

We present a syscall-based approach to automatically diagnose performance problems, server-to-client propagated errors, and server crash/hang problems in PVFS. Our approach compares the statistical and semantic attributes of syscalls across PVFS servers in order to diagnose the culprit server, under these problems, for different file-system benchmarks–dd, PostMark and IOzone–in a PVFS cluster.

PDL Workshop and Retreat 2009.