FALL UPDATE
PDL PACKET

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • FALL 2013
http://www.pdl.cmu.edu/

PDL CONSORTIUM MEMBERS

Actifio
American Power Conversion
EMC Corporation
Emulex
Facebook
Fusion-io
Google
Hewlett-Packard Labs
Hitachi
Huawei Technologies
Intel Corporation
Microsoft Research
NEC Laboratories
NetApp, Inc.
Oracle Corporation
Panasas
Samsung Information Systems America
Seagate Technology
Symantec Corporation
VMware, Inc.
Western Digital

CONTENTS

THE PDL PACKET

EDITOR
Joan Digney
CONTACTS
Greg Ganger
PDL Director
Bill Courtright
PDL Executive Director
Karen Lindenfelser
PDL Administrative Manager
The Parallel Data Laboratory
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
TEL 412-268-6716
FAX 412-268-3010

http://www.pdl.cmu.edu/Publications/

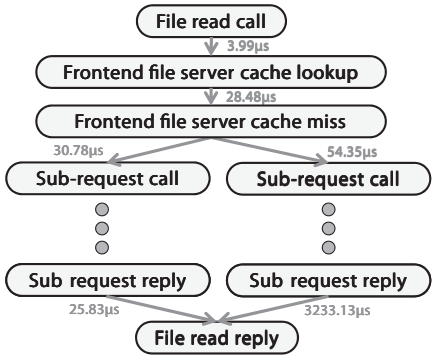
SELECTED RECENT PUBLICATIONS

Visualizing Request-flow Comparison to Aid Performance Diagnosis in Distributed Systems

Raja R. Sambasivan, Ilari Shafer, Michelle L. Mazurek & Gregory R. Ganger

IEEE Transactions on Visualization and Computer Graphics (Proceedings Information Visualization 2013), vol. 19, no. 12, Dec. 2013.

Distributed systems are complex to develop and administer, and performance problem diagnosis is particularly challenging. When performance degrades, the problem might be in any of the system's many components or could be a result of poor interactions among them. Recent research efforts have created tools that automatically localize the problem to a small number of potential culprits, but research is needed to understand what visualization techniques work best for helping distributed systems developers understand and explore their results. This paper compares the relative merits of three well-known visualization approaches (side-by-side, diff, and animation) in the context of presenting the results of one proven automated localization technique called request-flow comparison. Via a 26-person user study, which included real distributed systems developers, we identify the unique benefits that each approach provides for different problem types and usage modes.



Example request-flow graph. This graph shows the flow of a read request through a distributed storage system. Node names represent important events observed on the various components while completing the required work. Edges show latencies between these events. Fan-outs represent the start of parallel activity, and synchronization points (i.e., joins) are indicated by fan-ins. Due to space constraints, only the events observed on the frontend file server are shown.

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

Anshul Gandhi, Mor Harchol-Balter, Ram Raghunathan & Michael Kozuch

Transactions on Computer Systems, Volume 30, Issue 4, Article 14.

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

We introduce a dynamic capacity management policy, AutoScale, that greatly reduces the number of servers needed

continued on page 6

DISSERTATION ABSTRACT:
Algorithmic Engineering Towards More Efficient Key-Value Systems

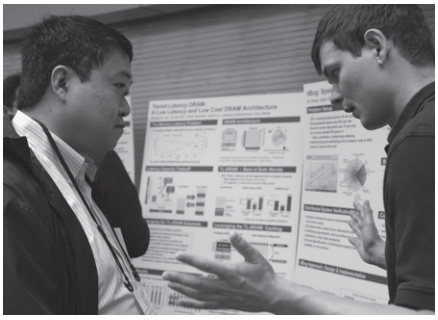
Bin Fan

Carnegie Mellon University SCS

Ph.D. Dissertation, October 24, 2013

Distributed key-value systems have been widely used as elemental components of many Internet-scale services at sites such as Amazon, Facebook and Twitter. This thesis examines a system design approach to scale existing key-value systems, both horizontally and vertically, by carefully engineering and integrating techniques that are grounded in recent theory but also informed by underlying architectures and expected workloads in practice. As a case study, we re-design FAWN-KV (i.e., a distributed key-value cluster consisting of wimpy key-value nodes) to achieve higher memory efficiency and ensure higher throughput even in the worst case.

First, to improve the worst-case throughput of a FAWN-KV system, we propose a randomized load balancing scheme that can fully utilize all the nodes regardless of their query distribution. We analytically prove and empirically demonstrate that deploying a very small but extremely fast load balancer at FAWN-KV can effectively prevent uneven or dynamic workloads creating hotspots on individual nodes. Moreover, our analysis provides service designers a mathematically tractable approach to estimate the worst-case throughput and help them avoid drastic over-provisioning in similar distributed key-value systems.



Jiri Simsa describes his research on the "Systematic Evaluation of Distributed Systems" to Jin Li (Microsoft Research) at a PDL Spring Visit Day poster session.

Second, to implement the extremely high-speed load balancer and also to improve the space efficiency of individual key-value nodes, we propose novel data structures and algorithms, including cuckoo filter, a Bloom filter replacement that is high-speed, highly compact and delete-supporting, and optimistic cuckoo hashing, a fast and space-efficient hashing scheme that scales on multiple CPUs. Both algorithms are built upon conventional cuckoo hashing but are optimized for our target architectures and workloads. Using them as building blocks, we design and implement MemC3 to serve transient data from DRAM with high throughput and low-latency retrieval, and SILT to provide cost-effective access to persistent data on flash storage with extremely small memory footprint (e.g., 0.7 bytes per entry).

DISSERTATION ABSTRACT:
Dynamic Server Provisioning for Data Center Power Management

Anshul Gandhi

Carnegie Mellon University SCS

Ph.D. Dissertation, June 2013

Data centers play an important role in today's IT infrastructure. However, their enormous power consumption makes them very expensive to operate. Sadly, much of the power used by data centers is wasted because of poor capacity management, leading to low server utilization.

In order to reduce data center power consumption, researchers have proposed several dynamic server provisioning approaches. However, there are many challenges that hinder the successful deployment of dynamic server provisioning, including: (i) unpredictability in workload demand, (ii) switching costs when setting up new servers, and (iii) unavailability of data when provisioning stateful servers. Most of the existing research in dynamic server provisioning has ignored, or carefully sidestepped, these important challenges at the expense of reduced benefits. In order to realize the full potential of dynamic server provisioning, we must overcome these associated challenges.

This thesis provides new research contributions that explicitly address the open challenges in dynamic server provisioning. We fist develop novel performance modeling tools to estimate the effect of these challenges on response time and power. In doing so, we also address several long-standing open questions in queueing theory, such as the analysis of multi-server systems with switching costs. We then present practical dynamic provisioning solutions for multi-tier data centers, including novel solutions that allow scaling the stateful caching tier and solutions that are robust to load spikes. Our implementation results using realistic workloads and request traces on a 38-server multi-tier testbed demonstrate that dynamic server provisioning can successfully meet typical response time guarantees while significantly lowering power consumption.

While this thesis focuses on server provisioning for reducing power in data centers, the ideas presented herein can also be applied to: (i) private clouds, where unneeded servers can be repurposed for "valley-filling" via batch jobs, to increase server utilization, (ii) community clouds, where unneeded servers can be given away to other groups, to increase the total throughput, and (iii) public clouds, where unneeded virtual machines can be released back to the cloud, to reduce rental costs.

continued on page 3

continued from page 3

Garth and his light-sabre lead his students in the ways of the force, or to an M.S. in Information Networking. One of the two... (Students from L to R: Fan Xiang, Chinmay Kamat, Pavan Kumar Alampalli, Praveen Kumar Ramakrishnan).

THESIS PROPOSAL:
An Automated Approach for Mitigating Service Performance Problems with Efficient Resource Allocations

Elie Krevat, ECE

August 23, 2013

Distributed and cloud computing services are increasingly built atop a preexisting infrastructure of shared services. These services have separate performance characteristics and require enough resources to support each application's service level objectives (SLOs), while preferably not wasting too many resources from overprovisioning. Changes in a service's performance are common (e.g., multiple times per day) for any number of reasons, such as from modified system configurations, hardware failures, or increased loads. Even worse, a problem in any one service can cause cascading delays across a complex web of interdependent services.

In this proposal, we describe an automated approach to mitigating such performance problems through reactive resource provisioning. When a problem occurs, we attempt to mitigate the problem in the short term by automatically assigning the right types and quantities of resources across services that can usefully apply them. Our proposed approach makes use of end-to-end request traces to determine the actual service flow and synchronicity requirements, combined with resource usage statistics to determine specific demands. This general monitoring framework is also used to discover each service's elastic scaling properties and to provide online feedback to better evaluate resource assignments.

THESIS PROPOSAL:
Scaling Metadata Service for Weak Scaling Workloads

Lin Xiao, SCS

August 5, 2013

Many file systems achieve high performance and scalability for large files instead of large number of files by striping or chunking files into data servers. However, as observed in real clusters, small files dominate the namespace, so metadata service could become the bottleneck as systems grow. The thesis is attacking the metadata scaling problem for weak scaling workloads. Metadata workloads in which file metadata operations increased much faster than directory name operations are defined as weak scaling workloads.

In this proposal, we present ShardFS, a hybrid solution to replicate directory names and sharding file metadata on oblivious metadata servers with demonstration using Hadoop file system(HDFS). We show the correctness and scalability of the system. We also focus on how to store metadata on disk to achieve similar performance as when they fit in memory. Finally we discuss the trade-offs on building metadata service with scalable distributed B-tree.

THESIS PROPOSAL:
Efficient Hypervisor Based Malware Detection

Peter Friedrich Klemperer, ECE

May 28, 2013

Recent years have seen an uptick in master boot record (MBR) based rootkits that load before the Windows operating system and subvert the operating system's own procedures. As such, MBR rootkits are difficult to counter with operating system-based antivirus software that runs at the same privilege-level as the rookits. Hypervisors operate at a higher privilege level than the guests they manage, creating a high-ground position in the host. This high-ground position can be exploited to perform security checks on the virtual machine guests where the checking software is isolated from guest-based viruses. The system proposed in this prospectus will target existing hypervisor systems to improve security with real-time, coherent memory introspection capabilities.

High performance guest memory introspection will decouple memory introspection from virtual machine guest execution, establish coherent and consistent memory views between the host and running guest, and provide intelligent memory translation to accelerate host-to-guest memory access. Existing introspection systems have provided one or two of these properties but not all three at once. This prospectus will present a new concurrent-computing approach to accelerate hypervisor based introspection of virtual machine guest memory that combines all three elements to improve performance and security.

The proposed system accelerates existing introspection systems and enables security protection techniques previously dismissed as too slow. In this prospectus, I will explain why existing introspection systems are inadequate, show how existing system performance can be improved, plan an initial prototype, and present several demonstrating applications based on that prototype. These demonstrating applications will be used to evaluate the prototype's performance and utility for supporting security applications.

continued on page 5

September 2013
Congratulations to the Guptas

Nitin and Sumedha Gupta are very happy to introduce you to their new daughter Avni. She was born on Sept 12, 2013 at 5:07 AM. She weighed 7 lb 10 oz and was 20.5 inches tall at birth. She likes to listen to conversations and spend time with her brother. Big brother Atharv loves his baby sister a lot.

July 2013
Welcoming the Most Wonderfully Perfect Grandson in the World!

Congratulations to Karen Lindenfelser, who is the proud Grandma of Landon Thomas Ziants, born on July 8, 2013 to mom and dad Julie Lindenfelser and Charles Ziants. So much love has been added to the Lindenfelser family!

July 2013
Support for Running Concurrent-write HPC Code on HDFS in PLFS

New code on running concurrent-write HPC codes on top of single-writer HDFS storage is now supported in the 2.4 release of PLFS. The code is available online at github.com/plfs.

July 2013
Call for Users: NSF PRObE 1000 Node Systems Research Testbed

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

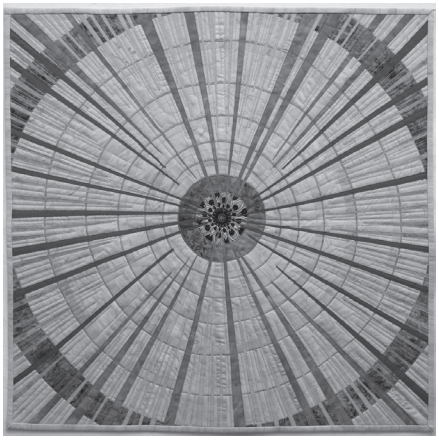
June 2013
Pavan Alampalli and Chinmay Kamat Receive Teaching Awards

Congratulations to Pavan and Chinmay on receiving their awards for Excellence as Teaching Assistants, specifically for their work with Garth and Greg on the Storage Systems class.

May 2013
Lorrie Cranor Exhibits Quilts at Pittsburgh Children's Museum

The Pittsburgh Children's Museum is hosting an exhibit of six quilts made by Associate Professor Lorrie Faith Cranor. Cranor, who has joint appointments at the Institute for Software Research, the Engineering and Public Policy Department and CyLab, is on sabbatical this semester and is a fellow at the STUDIO for Creative Inquiry. The exhibit, which is on the yellow wall opposite the "Garage" room, includes a quilt she based on an art installation called "More Light" by Dick Esterle in the museum's great dome. See photos of the exhibit and read more on her blog.

--8.5x11 News, May 9: Vol. 23, No. 41



March 2013
Honorable Mention in Qualcomm Innovation Fellowship Decision

Gennady Pekhimenko and Chris Fallin, advised by Onur Mutlu received honorable mention for their work on "Block-Based Heterogeneous Core Designs for Higher System Performance and Efficiency" in their application for the 2013 Qualcomm Innovation Fellowship. Out of 138 submitted proposals (from 15 schools), the judges selected 33 finalists from which they selected 8 winning teams. For the first time, honorable mentions were made, with honorees receiving grants equal to 50% of grants made to the winners.

--with info from qualcomm.com

March 2013
Yoongu Kim Awarded Intel Ph.D. Fellowship

Congratulations to Yoongu Kim (CMU) who was awarded a graduate fellowship

continued on page 5

continued from page 4

through the Intel Ph.D. Fellowship Program for his work on "SALP: A Novel Substrate for Enhancing Main Memory." This program seeks to find the top Ph.D. students at leading U.S. universities through a competitive and rigorous selection process and the The recipients of these awards represent the future thought leaders in their respective areas of research.

--with info from intel.com

February 2013
Gennady Pekhimenko Awarded Microsoft PhD Fellowship

Congratulations to Gennady, who received a Microsoft Research PhD Fellowship. The fellowship is a two-year program for outstanding PhD students nominated by their universities. The award covers 100% of the recipient's tuition and fees for two academic years and also includes a stipend and a conference and travel allowance. All recipients are offered the opportunity to complete an internship during the year following the award. Gennady's research focus is improving energy and performance characteristics of modern memory subsystems.

--with info from Carnegie Mellon News

continued from page 3

M.S. THESIS:
Cleaning Mechanism of Hybrid File System for Shingled Storage

Fan Xiang, INI

MS Information Security Technology and Management, April, 2013.

Continuous growth in demand for large capacity hard disk drives is driving the areal density towards the superparamagnetic limit. Shingled Magnetic Recording (SMR) increases areal density up to 2.5x by partially overlapping previously written tracks. To avoid multi-track read-modify-write penalties, data should only be appended and no random write allowed on shingled bands. To deal with this, previous work implemented a SMR-aware File System and hide sequential write property from the user. Because of the sacrifice of random write property in SMR, the SMR-aware File System faces the challenge of garbage collection. In this work, we present the cleaner design with band-level preemption and modularized cleaning policy selection. The performance of the cleaner is tested using an aging tool, which simulates file creation and deletion activity over a period time. Experiments show that the cleaner effectively recovers dead space on disk. We also conduct experiments on different cleaning policies in SMR-aware File System. A cost-benefit cleaning policy performs better than a Greedy policy. Cost-benefit takes the ages of files and write costs into account when picking a band to clean. It reduces file creation time by 80% compared with greedy. Another cleaning strategy is idle cleaning. The cleaner chooses to work during system idle time to minimize user request waiting time. It reduces time of cleaning by 30% compared with normal cleaning.

M.S. THESIS:
Metadata Optimization for Shingled Disks

Pavan K. Alampalli, INI

MS Information Networking, May 2013.

Continuous growth in demand for large capacity hard disk drives is driving disk areal density towards the superparamagnetic limit. Shingled Magnetic Recording (SMR) increases areal density by 2.5x by partially overlapping previously written tracks. To avoid multi-track read modify write penalties, data can only be appended and no random write is allowed. Previous work implemented an SMR-aware File System called ShingledFS. This work builds on the ShingledFS with the aim of optimizing metadata storage. SMR disks contain two partitions, shingled and unshingled. Shingled partitions have tracks shingled on one another and thus there is write-amplification. An unshingled partition has tracks laid out with a track gap between adjacent tracks, just as in the traditional disk. We are trading-off data density in order to achieve random-write capability in the unshingled partition by not shingling the tracks. We store ever-changing metadata about the files in the filesystem. This research explores the use of the LevelDB embedded key-value database to optimize the metadata storage. We report on the ways in which metadata can be packed into LevelDB and the resultant savings in the terms of unshingled disk space.

continued from page 1

in data centers driven by unpredictable, time-varying load, while meeting response time SLAs. AutoScale scales the data center capacity, adding or removing servers as needed. AutoScale has two key features: (i) it autonomically maintains just the right amount of spare capacity to handle bursts in the request rate; and (ii) it is robust not just to changes in the request rate of real-world traces, but also request size and server efficiency.

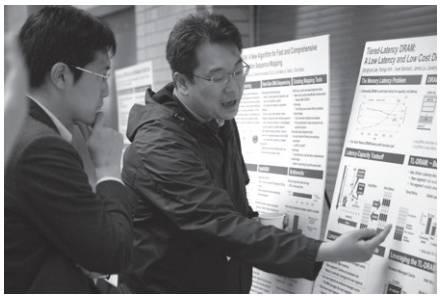
We evaluate our dynamic capacity management approach via implementation on a 38-server multi-tier data center, serving a web site of the type seen in Facebook or Amazon, with a key-value store workload. We demonstrate that AutoScale vastly improves upon existing dynamic capacity management policies with respect to meeting SLAs and robustness.

Making Problem Diagnosis Work for Large-Scale, Production Storage Systems

Michael P. Kasick, Priya Narasimhan & Kevin Harms

Proceedings of the 27th Large Installation System Administration Conference (LISA '13), Washington, DC, November 2013.

Intrepid has a very-large, production GPFS storage system consisting of 128 file servers, 32 storage controllers, 1152 disk arrays, and 11,520 total disks. In such a large system, performance problems are both inevitable and difficult to troubleshoot. We present our experiences, of taking an automated problem diagnosis approach from proof-of-concept on a 12-server testbench parallel-file-system cluster, and making it work on Intrepid's storage system. We also present a 15-month case study, of problems observed from the analysis of 624 GB of Intrepid's instrumentation data, in which we diagnose a variety of performance-related storage-system problems, in a matter of hours, as compared to the days or longer with manual approaches.



Donghyuk Lee discusses his poster "Tiered-Latency DRAM: A Low Latency and Low Cost DRAM Architecture" with Fanglu Guo of Symantec.

RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization

Vivek Seshadri, Yoongu Kim, Chris Fallin, Donghyuk Lee, Rachata Ausavarungnirun, Gennady Pekhimenko, Yixin Luo, Onur Mutlu, Michael A. Kozuch, Phillip B. Gibbons & Todd C. Mowry

Proceedings of the 46th International Symposium on Microarchitecture (MICRO), Davis, CA, December 2013.

Several system level operations trigger bulk data copy or initialization. Despite the fact that these bulk data operations do not require any computation, current systems transfer a large quantity of data back and forth on the memory channel to perform such operations. As a result, bulk data operations consume high latency, bandwidth, and energy --- degrading both system performance and energy efficiency.

In this work, we propose RowClone, a new and simple mechanism to perform bulk copy and initialization operations completely within DRAM --- eliminating the need to transfer any data over the memory channel to perform such operations. Our key observation is that DRAM can internally and efficiently transfer a large quantity of data (multiple KBs) between a row of DRAM cells and the associated row-buffer. Based on this, our primary mechanism can copy an entire row's worth of data between two rows that share a row-buffer. This mechanism, which we call the Fast Parallel Mode, can reduce the latency and energy of a bulk copy operation by 11.6x and 74.4x, respectively. To efficiently copy data across rows that do not share a row-buffer, we propose a second mode of RowClone, the Pipelined Serial Mode. RowClone requires only a 0.01% increase in DRAM chip area. We quantitatively evaluate the benefits of RowClone using fork, one of the most frequently invoked system calls, and five copy and initialization intensive applications/phases. Our results show that RowClone can significantly improve both single-core and multi-core system performance, while also significantly reducing energy consumption.

Linearly Compressed Pages: A Low-Complexity, Low-Latency Main Memory Compression Framework

Gennady Pekhimenko, Vivek Seshadri, Yoongu Kim, Hongyi Xin, Onur Mutlu, Michael A. Kozuch, Phillip B. Gibbons & Todd C. Mowry

Proceedings of the 46th International Symposium on Microarchitecture (MICRO), Davis, CA, December 2013.

Data compression is a promising approach for meeting the increasing memory capacity demands expected in future systems. Unfortunately, existing compression algorithms do not translate well when directly applied to main memory because they require the memory controller to perform non-trivial computations to locate a cache line within a compressed memory page, thereby increasing access latency and degrading system performance. Prior proposals for addressing this performance degradation problem are either costly or energy inefficient.

By leveraging the key insight that all cache lines within a page should be

continued on page 7

continued from page 6

compressed to the same size, this paper proposes a new approach to main memory compression - Linearly Compressed Pages (LCP) - that avoids the performance degradation problem without requiring costly or energy-inefficient hardware. We show that any compression algorithm can be adapted to fit the requirements of LCP, and we specifically adapt two previously proposed compression algorithms to LCP: Frequent Pattern Compression and Base-Delta-Immediate compression. Evaluations using benchmarks from SPEC CPU2006 and five server benchmarks show that our approach can significantly increase the effective memory capacity (69% on average). In addition to the capacity gains, we evaluate the benefit of transferring consecutive compressed cache lines between the memory controller and main memory. Our new mechanism considerably reduces the memory bandwidth requirements of most of the evaluated benchmarks (34% on average), and improves overall performance (6.1%/13.9%/10.7% for single-/two-/four-core workloads on average) compared to a baseline system that does not employ main memory compression. LCP also decreases energy consumed by the main memory subsystem (9.5% on average over the best prior mechanism).

There Is More Consensus in Egalitarian Parliaments

Iulian Moraru, David G. Andersen & Michael Kaminsky

Proceedings of the 24th ACM Symposium on Operating Systems Principles (SOSP'13), November 3-6, 2013, Nemacolin Woodlands Resort, Farmington, PA.

This paper describes the design and implementation of Egalitarian Paxos (EPaxos), a new distributed consensus algorithm based on Paxos. EPaxos achieves three goals: (1) optimal commit latency in the wide-area when tolerating one and two failures, under realistic conditions; (2) uniform load balancing across all replicas (thus achieving high throughput); and (3) graceful performance degradation when replicas are slow or crash. Egalitarian Paxos is to our knowledge the first protocol to simultaneously achieve all of these goals efficiently: requiring only a simple majority of replicas to be non-faulty, using a number of messages linear in the number of replicas to choose a command, and committing commands after just one communication round (one round trip) in the common case or after at most two rounds in any case. We prove Egalitarian Paxos's properties theoretically and demonstrate its advantages empirically through an implementation running on Amazon EC2.

Consistent, Durable, and Safe Memory Management for Byte-addressable Non Volatile Main Memory

Iulian Moraru, David G. Andersen, Michael Kaminsky, Niraj Tolia, Nathan Binkert & Parthasarathy Ranganathan

TRIOS: Conference on Timely Results in Operating Systems. Held in conjunction with SOSP '13. Farmington, PA, November 3, 2013.

This paper presents three building blocks for enabling the efficient and safe design of persistent data stores for emerging non-volatile memory technologies. Taking the fullest advantage of the low latency and high bandwidths of emerging memories such as phase change memory (PCM), spin torque, and memristor necessitates a serious look at placing these persistent storage technologies on the main memory bus. Doing so, however, introduces critical challenges of not sacrificing the data reliability and consistency that users demand from storage. This paper introduces techniques for (1) robust wear-aware memory allocation, (2) preventing of erroneous writes, and (3) consistency-preserving updates that are cache-efficient. We show through our evaluation that these techniques are efficiently implementable and effective by demonstrating a B+-tree implementation modified to make full use of our toolkit.

Measuring Password Guessability for an Entire University

Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay & Blase Ur

In CCS 2013: ACM Conference on Computer and Communications Security, November 2013.

Despite considerable research on passwords, empirical studies of password strength have been limited by lack of access to plaintext passwords, small data sets, and password sets specifically collected for a research study or from low-value accounts. Properties of passwords used for high-value accounts thus remain poorly understood.

We fill this gap by studying the single-sign-on passwords used by over 25,000 faculty, staff, and students at a research university with a complex password policy. Key aspects of our contributions rest on our (indirect) access to plaintext passwords. We describe our data collection methodology, particularly the many precautions we took to minimize risks to users. We then analyze how guessable the collected passwords would be during an offline attack by subjecting them to a state-of-the-art password cracking algorithm. We discover significant correlations between a number of demographic and behavioral factors and password strength. For example, we find that users associated with the computer science school make passwords more than 1.5 times as strong as those of users associated with the business school. In addition, we find that stronger passwords are correlated with

continued on page 8

continued from page 7

a higher rate of errors entering them.

We also compare the guessability and other characteristics of the passwords we analyzed to sets previously collected in controlled experiments or leaked from low-value accounts. We find more consistent similarities between the university passwords and passwords collected for research studies under similar composition policies than we do between the university passwords and subsets of passwords leaked from low-value accounts that happen to comply with the same policies.

Program Interference in MLC NAND Flash Memory: Characterization, Modeling, and Mitigation

Yu Cai, Onur Mutlu, Erich F. Haratsch & Ken Mai

Proceedings of the 31st IEEE International Conference on Computer Design (ICCD), Asheville, NC, October 2013.

As NAND flash memory continues to scale down to smaller process technology nodes, its reliability and endurance are degrading. One important source of reduced reliability is the phenomenon of program interference: when a flash cell is programmed to a value, the programming operation affects the threshold voltage of not only that cell, but also the other cells surrounding it. This interference potentially causes a surrounding cell to move to a logical state (i.e., a threshold voltage range) that is different from its original state, leading to an error when the cell is read. Understanding, characterizing, and modeling of program interference, i.e., how much the threshold voltage of a cell shifts when another cell is programmed, can enable the design of mechanisms that can effectively and efficiently predict and/or tolerate such errors.

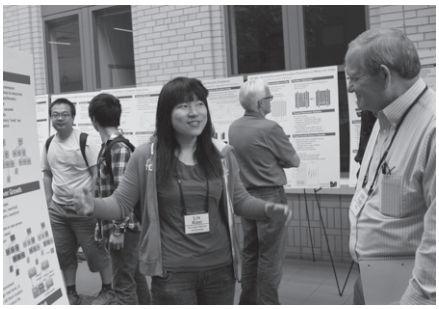
In this paper, we provide the first experimental characterization of and a realistic model for program interference in modern MLC NAND flash memory. To this end, we utilize the read-retry mechanism present in some state-of-the-art 2Y-nm (i.e., 20-24nm) flash chips to measure the changes in threshold voltage distributions of cells when a particular cell is programmed. Our results show that the amount of program interference received by a cell depends on 1) the location of the programmed cells, 2) the order in which cells are programmed, and 3) the data values of the cell that is being programmed as well as the cells surrounding it. Based on our experimental characterization, we develop a new model that predicts the amount of program interference as a function of threshold voltage values and changes in neighboring cells. We devise and evaluate one application of this model that adjusts the read reference voltage to the predicted threshold voltage distribution with the goal of minimizing erroneous reads. Our analysis shows that this new technique can reduce the raw flash bit error rate by 64% and thereby improve flash lifetime by 30%. We hope that the understanding and models developed in this paper lead to other error tolerance mechanisms for future flash memories.

Memory-Efficient GroupBy-Aggregate using Compressed Buffer Trees

Hrishikesh Amur, Wolfgang Richter, David G. Andersen, Michael Kaminsky, Karsten Schwan, Athula Balachandran & Erik Zawadzki

SoCC'13, Oct. 01–03 2013, Santa Clara, CA, USA.

Memory is rapidly becoming a precious resource in many data processing environments. This paper introduces a new data structure called a Compressed Buffer Tree (CBT). Using a combination of buffering, compression, and lazy aggregation, CBTs can improve the memory efficiency of the GroupBy-Aggregate abstraction which forms the basis of many data processing models like MapReduce and data-



Lin Xiao talks to Jerry Fredin (NetApp)about her work during a PDL Spring Visit Day poster session.

bases. We evaluate CBTs in the context of MapReduce aggregation, and show that CBTs can provide significant advantages over existing hashbased aggregation techniques: up to 2X less memory and 1.5X the throughput, at the cost of 2.5X CPU.

I/O Acceleration with Pattern Detection

Jun He, John Bent, Aaron Torres, Gary Grider, Garth Gibson, Carlos Maltzahn & Xian-He Sun

The 22nd Int. ACM Symposium on High Performance Parallel and Distributed Computing (HPDC'13), New York City, June 17-21, 2013.

The I/O bottleneck in high-performance computing is becoming worse as application data continues to grow. In this work, we explore how patterns of I/O within these applications can significantly affect the effectiveness of the underlying storage systems and how these same patterns can be utilized to improve many aspects of the I/O stack and mitigate the I/O bottleneck. We offer three main contributions in this paper. First, we develop and evaluate algorithms by which I/O patterns can be efficiently discovered and described. Second, we implement one such algorithm to reduce the metadata quantity in a virtual parallel file system by up to several orders of magnitude, thereby increasing the performance of writes and reads by up to 40 and

continued on page 9

continued from page 8

480 percent respectively. Third, we build a prototype file system with pattern-aware prefetching and evaluate it to show a 46 percent reduction in I/O latency. Finally, we believe that efficient pattern discovery and description, coupled with the observed predictability of complex patterns within many high-performance applications, offers significant potential to enable many additional I/O optimizations.

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

Garth Gibson, Gary Grider, Andree Jacobson & Wyatt Lloyd

USENIX ;login:, v 38, n 3, June 2013.

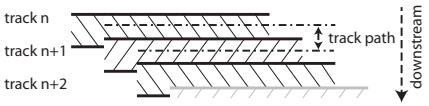
If you have ever aspired to create a software system that can harness a thousand computers and perform some impressive feat, you know the dismal prospects of finding such a cluster ready and waiting for you to make magic with it. Today, however, if you are a systems researcher and your promised feat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

Shingled Magnetic Recording: Areal Density Increase Requires New Data Management

Tim Feldman & Garth Gibson

USENIX ;login:, v 38, n 3, June 2013.

Shingled Magnetic Recording (SMR) is the next technology being deployed to increase areal density in hard disk drives (HDDs). The technology will provide the capacity growth spurt for the teens of the 21st century. SMR drives get that increased density by writing overlapping sectors, which means sectors cannot be written randomly without destroying the data in adjacent sectors. SMR drives can either maintain the current model for HDDs by performing data retention behind the scenes, or expose the underlying sector layout, so that file system developers can develop SMR-aware file systems.



Schematic of Shingled Magnetic Recording

Specialized Storage for Big Numeric Time Series

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

Proceedings of the 5th Workshop on Hot Topics in Storage and File Systems, June 2013.

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

Hadoop's Adolescence: An Analysis of Hadoop Usage in Scientific Workloads

Kai Ren, YongChul Kwon, Magdalena Balazinska & Bill Howe

Very Large Data Bases (VLDB), August, 2013.

We analyze Hadoop workloads from three different research clusters from a user-centric perspective. The goal is to better understand data scientists' use of the system and how well the use of the system matches its design. Our analysis suggests that Hadoop usage is still in its adolescence. We see underuse of Hadoop features, extensions, and tools. We see significant diversity in resource usage and application styles, including some interactive and iterative workloads, motivating new tools in the ecosystem. We also observe significant opportunities for optimizations of these workloads. We find that job customization and configuration are used in a narrow scope, suggesting the future pursuit of automatic tuning systems. Overall, we present the first user-centered measurement study of Hadoop and find significant opportunities for improving its efficient use for data scientists.

Active Disk Meets Flash: A Case for Intelligent SSDs

Sangyeun Cho, Chanik Park, Hyunok Oh, Sungchan Kim, Youngmin & Gregory R. Ganger

Proceedings of the ACM Int'l Conference on Supercomputing (ICS), Eugene, OR, June 2013.

Intelligent solid-state drives (iSSDs) allow execution of limited application functions (e.g., data filtering or aggregation) on their internal hardware resources, exploiting SSD characteristics and trends to provide large and growing performance and energy efficiency benefits. Most notably, internal flash media bandwidth can be significantly (2-4× or more) higher than the external bandwidth with which the SSD is connected to a host system, and the higher internal bandwidth can be exploited within an iSSD. Also, SSD bandwidth is projected to increase rapidly over time, creating a substantial energy cost for streaming of data to an external CPU for processing, which can be avoided via iSSD processing. This paper makes a case for iSSDs by detailing these trends, quantifying the potential benefits across a range of application activities, describing how SSD architectures could be extended cost-effectively, and demonstrating the concept with measurements of a prototype iSSD running simple data scan functions. Our analyses indicate that, with less than a 2% increase in hardware

continued on page 10

continued from page 9

cost over a traditional SSD, an iSSD can provide 2-4× performance increases and 5-27× energy efficiency gains for a range of data-intensive computations.

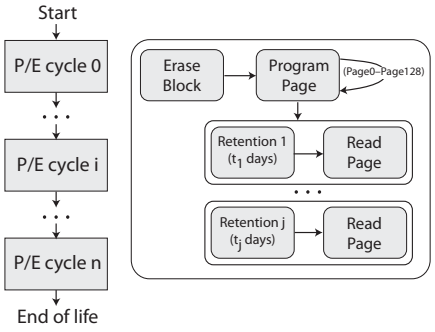
Error Analysis and Retention-Aware Error Management for NAND Flash Memory

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

Intel Technology Journal (ITJ) Special Issue on Memory Resiliency, Vol. 17, No. 1, May 2013.

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

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



NAND flash programming model for error characterization.

threshold voltage distortions that cause various errors. We conclude with a brief description of our ongoing related work in combating scaling challenges of both NAND flash memory and DRAM memory.

A Proof of Correctness for Egalitarian Paxos

Iulian Moraru, David G. Andersen & Michael Kaminsky

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-13-111. August 2013.

This paper presents a proof of correctness for Egalitarian Paxos (EPaxos), a new distributed consensus algorithm based on Paxos. EPaxos achieves three goals: (1) availability without interruption as long as a simple majority of replicas are reachable—its availability is not interrupted when replicas crash or fail to respond; (2) uniform load balancing across all replicas—no replicas experience higher load because they have special roles; and (3) optimal commit latency in the wide-area when tolerating one and two failures, under realistic conditions. Egalitarian Paxos is to our knowledge the first distributed consensus protocol to achieve all of these goals efficiently: requiring only a simple majority of replicas to be non-faulty, using a number of messages linear in the number of replicas to choose a command, and committing commands after just one communication round (one round trip) in the common case or after at most two rounds in any case.

Saving Cash by Using Less Cache

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

4th USENIX Conference on Hot Topics in Cloud Computing (HotCloud 2012). June 12-13, 2012. Boston, MA.

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

An Experimental Study of Data Retention Behavior in Modern DRAM Devices: Implications for Retention Time Profiling Mechanisms

Jamie Liu, Ben Jaiyen, Yoongu Kim, Chris Wilkerson & Onur Mutlu

Proceedings of the 40th International Symposium on Computer Architecture (ISCA), Tel-Aviv, Israel, June 2013.

DRAM cells store data in the form of charge on a capacitor. This charge leaks off over time, eventually causing data to be lost. To prevent this data loss from occurring, DRAM cells must be periodically refreshed. Unfortunately, DRAM refresh operations waste energy and also degrade system performance by interfering with memory requests. These problems are expected to worsen as DRAM density increases.

The amount of time that a DRAM cell can safely retain data without being refreshed is called the cell's retention time. In current systems, all DRAM cells are refreshed at the rate required

continued on page 11

continued from page 10

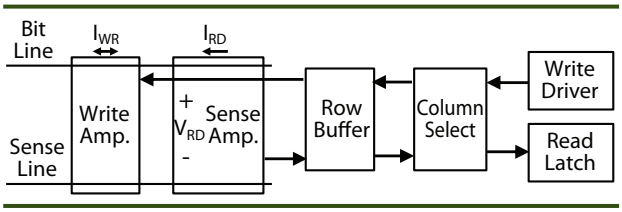
to guarantee the integrity of the cell with the shortest retention time, resulting in unnecessary refreshes for cells with longer retention times. Prior work has proposed to reduce unnecessary refreshes by exploiting differences in retention time among DRAM cells; however, such mechanisms require knowledge of each cell's retention time. In this paper, we present a comprehensive quantitative study of retention behavior in modern DRAMs. Using a temperature-controlled FPGA-based testing platform, we collect retention time information from 248 commodity DDR3 DRAM chips from five major DRAM vendors. We observe two significant phenomena: data pattern dependence, where the retention time of each DRAM cell is significantly affected by the data stored in other DRAM cells, and variable retention time, where the retention time of some DRAM cells changes unpredictably over time. We discuss possible physical explanations for these phenomena, how their magnitude may be affected by DRAM technology scaling, and their ramifications for DRAM retention time profiling mechanisms.

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

Emre Kultursay, Mahmut Kandemir, Anand Sivasubramaniam & Onur Mutlu

Proceedings of the 2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Austin, TX, April 2013.

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



The organization of the sense and write circuitry for STT-RAM bitlines. (Note the existence of separate sense amplifiers and row buffer storage.)

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

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

Proceedings of the 5th Workshop on Energy-Efficient Design (WEED), Tel-Aviv, Israel, June 2013.

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

PETAL: Preset Encoding Table Information Leakage

Jiaqi Tan & Jayvardhan Nahata

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-13-106, April 2013.

SPDY is an application-layer protocol which multiplexes multiple HTTP requests and compresses HTTP headers over a single TCP connection protected by SSL/TLS encryption. Web applications are ubiquitous, and HTTP headers carry HTTP cookies which often contain sensitive information which can result in loss of privacy if leaked. We perform a security analysis on the proposed compression scheme for the next revision of the SPDY protocol, particularly with respect to the previously disclosed CRIME attack which uses compression-based information leaks. We have identified a new information leakage in the compression scheme of the proposed

continued on page 12

continued from page 11

and previous versions of the SPDY protocol, which we call PETAL1, which exploits the use of a fixed Huffman encoding table and the lack of byte-alignment of encoded characters, and we have identified a way to recover cookies using this information leakage by exploiting the way that multiple HTTP cookies with the same name but different Path attributes are handled by current web browsers. We perform a detailed analysis of the impact of this information leakage, and find that after considering practical issues such as the byte-padded nature of network communications, our hypothesized attack only leaks less than 2-bits of information for 30-character uppercase alphanumeric strings, and does not allow a network attacker to recover meaningful amounts of information despite our discovered information leakage.

SOFTScale: Stealing Opportunistically For Transient Scaling

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

13th International Middleware Conference (Middleware 2012). Dec. 3-7, 2012, Montreal, Quebec.

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

In this paper, we propose SOFTScale, an approach to handling load spikes in multi-tier data centers without having to over-provision resources. SOFTScale works by opportunistically stealing resources from other tiers to alleviate the bottleneck tier, even when the tiers are carefully provisioned at capacity. SOFTScale is especially useful during the transient overload periods when additional capacity is being brought online. Via implementation on a 28-server multi-tier testbed, we investigate a range of possible load spikes, including an artificial doubling or tripling of load, as well as large spikes in real traces. We find that SOFTScale can meet our stringent 95th percentile response time Service Level Agreement goal of 500ms without using any additional resources even under some extreme load spikes that would normally cause the system (without SOFTScale) to exhibit response times as high as 96 seconds.

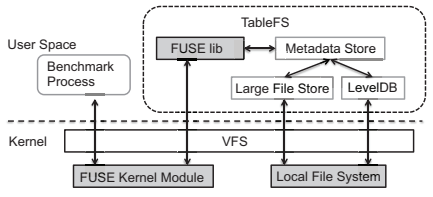
TABLEFS: Enhancing Metadata Efficiency in the Local File System

Kai Ren & Garth Gibson

2013 USENIX Annual Technical Conference, June 26-28, 2013, San Jose, CA.

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

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



The architecture of TABLEFS. A FUSE kernel module redirects file system calls from a benchmark process to TABLEFS, and TABLEFS stores objects into either LevelDB or a large file store.

Just-in-Time Provisioning for Cyber Foraging

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

MobiSys'13, June 25–28, 2013, Taipei, Taiwan.

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